Video Quality Assessment for Spatio-Temporal Resolution Adaptive Coding

Hanwei Zhu, *Graduate Student Member, IEEE*, Baoliang Chen, *Member, IEEE*, Lingyu Zhu, *Student Member, IEEE*, Peilin Chen, Linqi Song, *Senior Member, IEEE*, and Shiqi Wang, *Senior Member, IEEE*

***Abstract*—Spatio-temporal resolution adaptive (STRA) coding has been repeatedly proven to be a promising way to improve coding efficiency and reduce coding complexity. The wide consensus is that the optimal subsampled resolution and frame rate should be governed by so-called generalized rate-distortion performance based on the ultimately perceived distortion. However, it is non-trivial to accurately predict the quality of reconstructed videos due to the fact that the distortion originates from both subsampling and compression. To address this issue, we propose a novel video quality assessment model that is fully aware of the information available in downsampled videos for compression, such as resolution and frame rate. More specifically, the proposed model relies on quality-aware spatial features that are extracted by an image quality fine-tuned backbone. Subsequently, the spatio-temporal quality is modeled based on the transformer encoder, which is adaptive to the downsampling spatial and temporal resolutions. This enables the transformer encoder to produce discriminative features that capture long-range temporal dependencies related to the current context. The quality score, which is the output of the transformer encoder, thus reflects both the influence of the subsampling and compression. We conduct extensive experiments that demonstrate the superiority of the proposed model over state-of-the-art methods on four subsampling and compression video quality datasets. Furthermore, we apply the proposed model to bitrate ladder optimization, leading to a perceptual-aware spatial and temporal downsampling strategy that yields promising bitrate savings. The source codes of the proposed model will be publicly available at https://github.com/h4nwei/STRA-VQA.**

***Index Terms*—Full-reference video quality assessment, spatio-temporal resolution adaptive coding, vision transformer.**

I. Introduction

THE long-standing problem of video coding is how to achieve a good balance of video quality and bandwidth. A series of video compression standards, including the H.264/AVC [1], HEVC [2], AVS [3], VVC [4], VP9 [5], and AV1 [6], have been developed recently. On top of these standards, spatio-temporal resolution adaptive (STRA) coding evolves to accommodate diverse display devices, contents, and bandwidths [7]–[9]. The video communication flowchart is shown in Fig. 1, where a variety of pre-processing schemes, *i.e.*, spatial resolution and frame rate downsampling, have been performed before the video compression. Then, the subsampled videos are compressed using different codecs with varying bitrates. The upsampling methods are subsequently leveraged for post-processing the decoded videos.

One of the main technical challenges of the STRA video coding technique is to optimize perceptual quality under a given bit rate constraint, necessitating the selection of appropriate spatial resolution and frame rate based on a precise video quality assessment (VQA) model. Lin *et al*. made the pioneering attempt to study the video quality of spatial resolution rescaling and compression [10]. Subsequently, more studies considered 4K videos as the source, and different spatial rescaling methods and video encoders have been utilized to resample and compress the videos [11], [12]. In addition, researchers also examined the relationship between the video frame rate and visual quality based upon different video compression methods [12]–[14]. However, while individual studies regarding spatial or temporal subsampling facilitate an understanding of how spatial or temporal rescaling affects visual quality, such conclusions cannot be straightforwardly applied to joint subsampling and compression because of the dramatically different distortion patterns. Thus, to mitigate such limitations, Rao *et al*. carried out a small-scale subjective study for the spatially and temporally scaled videos that are further compressed by the H.264/AVC codec [15]. Furthermore, Lee *et al*. increased the number of source videos for a more comprehensive subjective quality assessment, and the HEVC encoder has been adopted to compress the video with various bitrates, resolutions, and frame rates [16].

However, on-demand subjective quality assessment is cumbersome and expensive, making it impractical to evaluate all possible visual stimuli. Therefore, the development of objective quality models that can accurately predict the visual quality of reconstructed videos is highly desirable. Although numerous VQA methods have been developed, there are only a few works dedicated to VQA of videos that undergo initial subsampling (*i.e.*, spatial and temporal downsampling) and subsequent compression. Lee *et al*. computed the natural video statistics of the differences of neighboring frames to build a full-reference video quality assessment (FR-VQA) model for the STRA coding videos [17]. However, the utilization of hand-crafted features impedes satisfactory VQA performance. Additionally, bitstream-based VQA models demonstrate that

This work was supported in part by the Shenzhen Science and Technology Program under Project JCYJ20220530140816037, in part by the Hong Kong Research Grants Council General Research Fund Projects 11203220 and 11217823, in part by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA), and in part by the Innovation and Technology Fund Project GHP/044/21SZ and PRP/059/20FX. (Corresponding author: *Shiqi Wang.*)

Hanwei Zhu, Baoliang Chen, Lingyu Zhu, Peilin Chen, and Linqi Song are with the Department of Computer Science, City University of Hong Kong, Hong Kong, China. (e-mail: {hanwei.zhu, blchen6-c, lingyzhu-c}@my.cityu.edu.hk, {plchen3, linqi.song}@cityu.edu.hk).

Shiqi Wang is with the Department of Computer Science, City University of Hong Kong, Hong Kong, China, and also with the Shenzhen Research Institute, City University of Hong Kong, Shenzhen, China (e-mail: shiqwang@cityu.edu.hk).

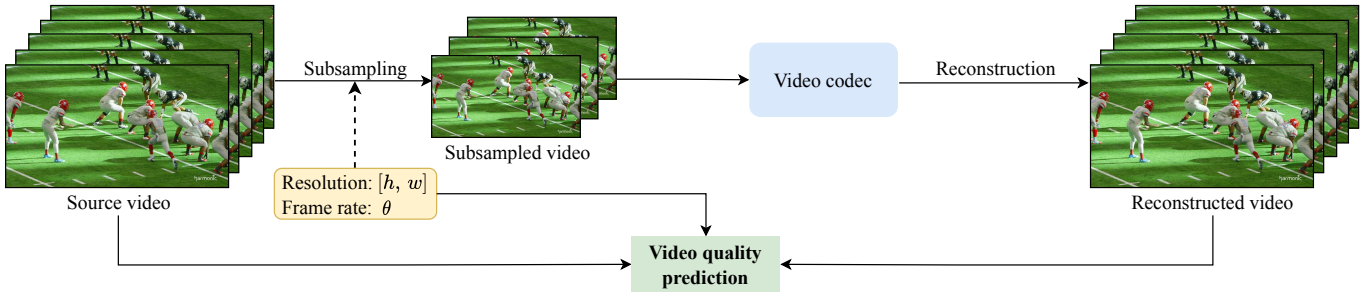

Fig. 1. The pipeline of the STRA coding with the proposed VQA model. The source videos are first downsampled by different spatial resolutions and frame rates, and then the codec with varying bitrates can be used to compress the subsampled videos. Upsampling can subsequently be applied to restore the spatial and temporal resolutions. The proposed FR-VQA model is used to predict the video quality given the source video, the reconstructed video, and the information including the spatial resolution (denoted as $[h, w]$) and frame rate (denoted as θ).

metadata information, such as spatial resolutions, frame rates, and bitrates, contains valuable knowledge for inferring quality scores [18]–[20]. Ramachandra *et al*. employed machine learning to combine metadata (resolution and frame rate) features and video features, resulting in a hybrid FR-VQA model [21]. Lee *et al*. also empirically showed that spatial resolutions and frame rates of subsampled videos in STRA coding are positively correlated with the mean opinion score (MOS) [16].

In this work, we develop the first deep learning-based FR-VQA model for STRA coding, called STRA-VQA. With the *metadata-level information*, we devise an adaptive weight prediction module (WPM). It takes the metadata information of the subsampled video as input and generates adaptive weights that contain explicit prior knowledge regarding the distortion patterns. With the *signal-level information*, the spatial domain features that can accurately reflect the spatial distortion are extracted [22]. The features lay the foundation for the proposed STRA-VQA model. Then, we develop an adaptive weight transformer encoder with the WPM to aggregate the quality-aware spatial features, which not only produces discriminative features adapted to the reconstructed video but also captures the long-range temporal quality dependencies. We conduct extensive experiments to demonstrate that STRA-VQA significantly outperforms state-of-the-art VQA methods on four subsampling and compression VQA datasets. Moreover, we apply the proposed model to bitrate ladder optimization, demonstrating promising bitrate savings compared with the existing classical quality models [23], [24]. Overall, the major contributions of this paper are summarized as follows.

- We present the deep learning-based FR-VQA model for STRA coding that achieves accurate quality assessment for the videos that are first subsampled and then compressed. The proposed model enjoys high accuracy since the distinct information on video characteristics has been incorporated into a WPM module for quality prediction.
- We devise a transformer network using the WPM module for the FR-VQA model. The unique design produces discriminative features that adapt to the input videos and successfully capture long-range temporal quality dependencies, leading to superior VQA performance.
- We optimize the proposed perceptually calibrated FR-VQA model for bitrate ladder generation in STRA coding. Experimental results show a significant reduction in bitrate at the same quality level compared with state-of-the-art methods.

II. Related Works

In this section, we first introduce the pixel-based VQA models that rely on the raw pixel data for quality assessment. Then, we provide a review of the bitstream-based VQA models that leverage the information from coding bitstreams as the dominant or supplementary information. Finally, we briefly describe the existing STRA coding algorithms.

A. Pixel-Based Objective VQA Models

With the rapid evolution of VQA datasets, numerous VQA models have been developed, including traditional [24]–[27] and deep learning-based methods [28]–[33]. Averaging the quality scores calculated by the image quality assessment (IQA) models [23], [34]–[38], such as PSNR, structural similarity index (SSIM) [23], and most apparent distortion (MAD) [34], is one of the most straightforward ways to obtain video quality. However, such IQA models do not consider motion information, which is an essential factor in VQA. As such, researchers began to incorporate different temporal modules into the IQA models, aiming to build more robust VQA models. Seshadrinathan *et al*. combined the temporal quality computed by optical flow with the SSIM-inspired spatial quality, yielding the motion-based video integrity evaluation index (MOVIE) [25]. The optical flow has also been widely used to model the temporal distortions in VQA [39], [40]. 3D-SSIM performed the structural similarity evaluation in local cube blocks [41]. The video multi-method assessment fusion (VMAF) utilized the frame difference of the distorted video to model motion information [24]. In addition, natural scene statistics (NSS) of spatial and temporal entropic differences in the bandpass domain played another vital role in video quality prediction [42], [43]. Furthermore, advanced deep learning techniques were adopted to design FR-VQA models, such as the convolutional neural aggregation network [29], 3D-CNN [30], and the long short-term memory (LSTM) network [44]. Hou *et al*. employed swin transformer blocks for spatio-temporal information extraction and proposed a dedicated perceptual quality measure for video frame interpolation [45]. Feng *et al*. utilized a transformer-based

network architecture with a two-stage training methodology to overcome difficulties such as lacking training data [46]. Apart from the general FR-VQA models, several works were explicitly designed to measure the objective quality of resolution and frame rate variation. The frame rate quality metric (FRQM) was proposed as a wavelet domain model to measure the quality difference owning to frame rate scaling [28]. Nasiri *et al*. applied the local phase correlation of complex wavelet coefficients to evaluate the motion smoothness of the video with varying frame rates [47]. Lee *et al*. computed the natural video statistics of divisively normalized differences of neighboring frames to model complex spatial and temporal scaling distortions [17]. However, the utilization of hand-crafted features shows that there could be room for improvement in achieving promising VQA performance [17]. As such, in this work, we aim to quantify the complicated distortion generated by STRA coding using a knowledge-driven approach with a specifically designed transformer architecture, which exhibits a strong capability to capture the long-range spatio-temporal quality dependencies.

B. Bitstream-Based Objective VQA Models

Bitstream-based VQA models assess the quality of a video by analyzing information from the encoded bitstream even without video signals [48]. One of the representative models is the adaptive video quality model based on the bitstream information (AVQBits) model, which is a versatile, bitstream-based video quality model that can be applied in several applications, such as video service monitoring, evaluation of video encoding quality, gaming video quality of experience, and even omnidirectional video quality [21]. The ITU-T P.1204 series of standards also include a no-reference bitstream-based model (*i.e*., P.1204.3) that has access to encoded bitstream information [49]. In addition, different curve-fitting bitstream models [50], [51] and machine-learning bitstream algorithms [52], [53] have been proposed to predict video quality. These models can provide accurate predictions of video quality and can be used in various applications [18], [19]. It is worth noting that the metadata-based quality model, which relies on resolution, frame rate, and bitrate, has been characterized as a lightweight variant of bitstream-based models [54]. Ramachandra *et al*. enhanced existing pixel-based VQA models by incorporating metadata extracted from the encoding bitstream, such as quantization parameters, which helps weight distortions based on perceptual importance and improves correlation with MOS [21]. Although bitstream information has been extensively explored for general video quality prediction, the utilization of metadata information in STRA coding to construct a holistic VQA model remains under-explored. In this paper, we fill this gap by employing the metadata of subsampled videos to generate adaptive weights using a weight prediction module, thereby improving VQA performance through the incorporation of distinct information on video characteristics.

C. STRA Coding

Under a stringent bandwidth, video content providers proposed to optimize the bitrate ladder for input videos in terms of limited bitrate, spatial resolution, and frame rate. Netflix recommended a set of bitrate-resolution pairs for the H.264/AVC video encoder [7]. Some content adaptive STRA coding methods were further presented to obtain the optimal bitrate ladder [55]–[57]. Amirpour *et al*. took both spatial and temporal resolution into account, achieving double bitrate savings compared to spatial resolution optimization only [58]. The bitrate ladders were also adapted based on the VQA models [8], [9]. In this paper, the proposed perceptual-aware VQA model is optimized to provide a content- and resolution-adaptive bitrate ladder, demonstrating significant bitrate savings over existing objective quality models.

III. The Proposed Model

Our primary target is to develop a specific FR-VQA model for the video compressed by STRA coding. The framework of the proposed STRA-VQA model is shown in Fig. 2. We first employ a quality-aware CNN backbone to extract the spatial features of the reference and distorted video frames, respectively. Subsequently, the embeddings aggregated by distorted video features and the corresponding residual features are fed into a transformer encoder, the weights of which are determined by the metadata information (resolution and frame rate) of the subsampled video. Finally, we introduce the objective function to optimize the proposed network.

A. Frame-Wise Feature Transformation

1) Spatial Feature Extraction: Herein, we start to build a feature transformation module, which maps the source and reconstructed video frames to a more perceptually meaningful domain. To be specific, deep CNN backbones pre-trained by ImageNet [59] have exhibited strong capabilities in perceptual image processing [60], [61]. Thus, we take the pre-trained ResNet50 [62] as our feature extractor, which is further fine-tuned by the IQA benchmarks [22]. We denote the t-th frame of the source and reconstructed video to $\mathbf{X}^{(t)}$ and $\mathbf{Y}^{(t)}$, respectively. Then, the video frames are fed into the quality-aware feature extractor $f(\cdot)$, which equips with fixed fine-tuned weights $\mathbf{w}_f$ and can be expressed as follows,

$$\tilde{\mathbf{X}}^{(t)} = f(\mathbf{X}^{(t)}; \mathbf{w}_f)\,, \tilde{\mathbf{Y}}^{(t)} = f(\mathbf{Y}^{(t)}; \mathbf{w}_f)\,, \tag{1}$$

where $\tilde{\mathbf{X}}^{(t)} \in \mathbb{R}^{w\times h\times n}$ and $\tilde{\mathbf{Y}}^{(t)} \in \mathbb{R}^{w\times h\times n}$ are the perceptual feature maps of the source and reconstructed video, respectively. The notations w, h, and n stand for the width, height, and channel number of the output feature maps.

Subsequently, we perform the subtraction between source and reconstructed features to compute the corresponding residual features, which contain noticeable distortions in the perceptually meaningful feature space [30], [63]. This process can be expressed as follows,

$$\tilde{\mathbf{R}}^{(t)} = \tilde{\mathbf{X}}^{(t)} - \tilde{\mathbf{Y}}^{(t)}\,, \tag{2}$$

where $\tilde{\mathbf{R}}^{(t)} \in \mathbb{R}^{w\times h\times n}$ represents the residual features. Then, we apply the global mean pooling (GAP) to reduce the feature dimension and preserve the spatial content of each feature map. The feature variations are calculated by the global standard deviation pooling (GSP). It is worth mentioning that

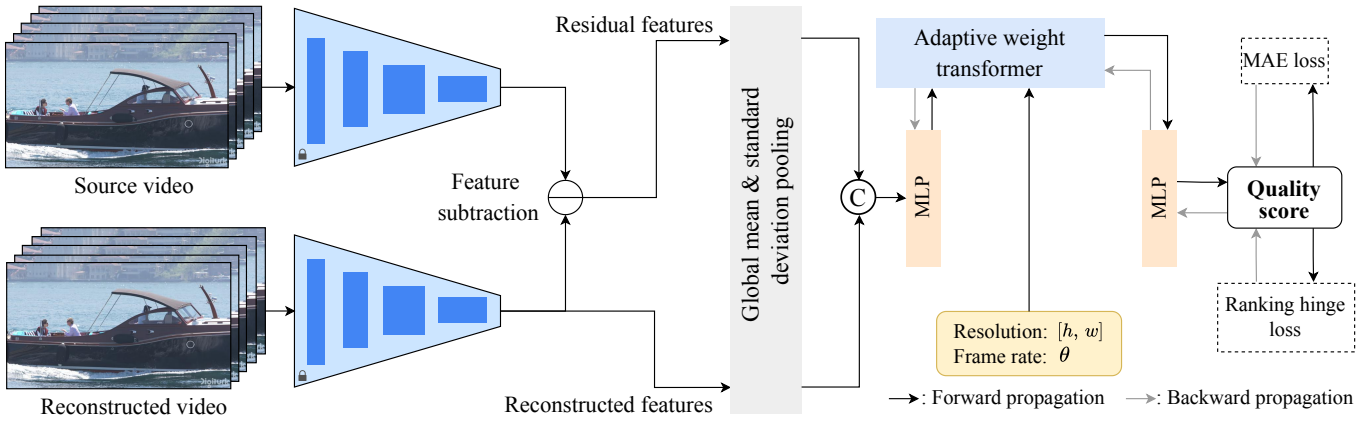

Fig. 2. Illustration of the architecture of proposed STRA-VQA model. We first extract the spatial features of the source and reconstructed videos. Then the aggregated features and the metadata information (resolution and frame rate of the downsampling video) are fed into a transformer, which generates adaptive weights to regress the overall video quality.

the GAP and GSP have been widely used in feature reduction in VQA since the obtained feature statistics are content-aware and distortion-sensitive [29], [32], [64]. As such, the GAP and GSP are carried out on both reconstructed and residual feature maps, respectively, and then the mean and standard deviation statistic features are concatenated along with the feature dimension as follows,

$$\tilde{\mathbf{y}}^{(t)} = \text{Concat}\left(\text{GAP}(\tilde{\mathbf{Y}}^{(t)}), \text{GSP}(\tilde{\mathbf{Y}}^{(t)})\right), \tag{3}$$

$$\tilde{\mathbf{r}}^{(t)} = \text{Concat}\left(\text{GAP}(\tilde{\mathbf{R}}^{(t)}), \text{GSP}(\tilde{\mathbf{R}}^{(t)})\right), \tag{4}$$

where $\tilde{\mathbf{y}}^{(t)} \in \mathbb{R}^{1\times 2n}$ and $\tilde{\mathbf{r}}^{(t)} \in \mathbb{R}^{1\times 2n}$ indicate the concatenated spatial feature vectors of reconstructed and residual feature maps, respectively.

2) Feature Aggregation: Given $\tilde{\mathbf{y}}^{(t)}$ and $\tilde{\mathbf{r}}^{(t)}$, we propose to combine them to obtain the final feature representations as follows,

$$\mathbf{Z}^{(t)} = \text{Concat}(\tilde{\mathbf{y}}^{(t)}, \tilde{\mathbf{r}}^{(t)}), \tag{5}$$

where $\mathbf{Z}^{(t)} \in \mathbb{R}^{1\times 4n}$ denotes the aggregated features of the t-th video frame. By utilizing one fully connected (FC) layer, we reduce the dimensions of aggregated features prior to feeding them to the long-range spatio-temporal interaction module, which can be expressed as follows,

$$\mathbf{E}^{(t)} = \mathbf{w_r} * \mathbf{Z}^{(t)} + \mathbf{b_r}, \tag{6}$$

where $\mathbf{w_r}$ and $\mathbf{b_r}$ are the learnable parameters of the FC layer. $\mathbf{E}^{(t)} \in \mathbb{R}^{1\times c}$ represents the compact feature containing both source and reconstructed information, and c is the feature dimension.

B. Spatio-Temporal Quality Modeling

The spatial features lay the foundation for the proposed VQA model. Herein, we leverage the metadata information that contains explicit prior knowledge of the distortion patterns [16] for spatio-temporal quality modeling. More specifically, we build an adaptive weight transformer encoder, which applies the metadata information to predict the weights of the FC layers. The design philosophy brings two benefits:

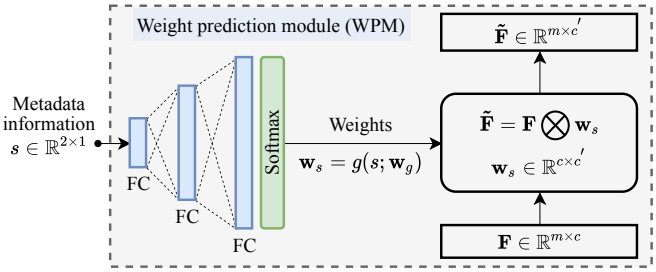

Fig. 3. Illustration of the proposed WPM module.

1) In contrast to fixed-weight deep learning models, our model generates adaptive weights to produce discriminative features that adapt to the distortions of the reconstructed video caused by STRA coding; 2) The transformer encoder enables the model to capture the long-range spatio-temporal quality dependencies.

1) Adaptive Weight Prediction: As shown in Fig. 3, we feed the metadata information into a weight prediction module (WPM), denoted as $g(\cdot; \mathbf{w}_g)$ and $\mathbf{w}_g$ is the parameters of network $g(\cdot)$. The WPM consists of three FC layers, and the softmax is used to activate and constrain the output weights to be $[0, 1]$. The output dimension of the FC layer increases to match the dimension of input spatial embedding. The metadata information s is obtained by the normalized spatial resolution and frame rate of the downsampling video. Formally, the weight generation procedure can be defined as

$$\mathbf{w}_s = g(s; \mathbf{w}_g), \tag{7}$$

where $\mathbf{w}_s \in \mathbb{R}^{c\times c'}$ is the predicted weight and c' represents the feature dimension. Then, the input features are reformulated by the generated weight as follows,

$$\tilde{\mathbf{F}} = \mathbf{F} \otimes \mathbf{w}_s, \tag{8}$$

where $\mathbf{F} \in \mathbb{R}^{m\times c}$ is the input features, $\otimes$ is the matrix multiplication, $\tilde{\mathbf{F}} \in \mathbb{R}^{m\times c'}$ is the feature obtained by the adaptive weights $\mathbf{w}_s$, and m is the number of features. It is worth noting that the weights are adaptive to the metadata

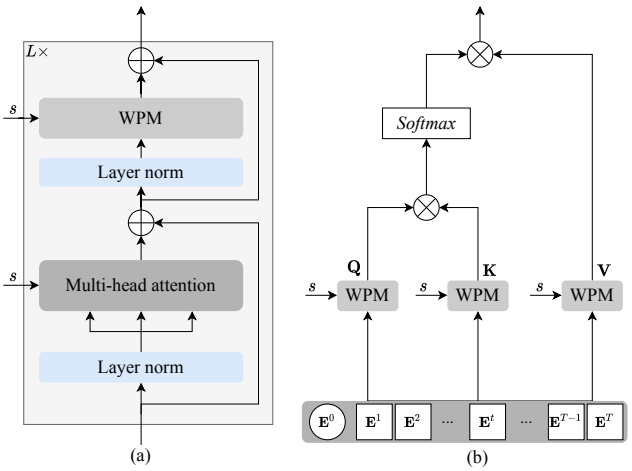

Fig. 4. (a) Illustration of the adaptive weight transformer encoder; (b) Flowchart of scaled dot-product attention.

information during the inference stage, which produces discriminative features.

2) Transformer Encoder: As shown in Fig. 2, we feed the compact feature $\mathbf{E}^{(t)}$ to a transformer encoder, which is deemed to be effective at capturing spatio-temporal dependencies over a long range. Meanwhile, we replace all the FC layers with the WPM such that we are able to learn the adaptive weight with respect to the metadata information. The structural detail of the proposed transformer is shown in Fig. 4 (a), which consists of three key components, including layer normalization (LN), multi-head self-attention (MSA), and WPM.

In particular, to efficiently communicate features in different frames along with the timeline, we design a quality token (denoted to $\mathbf{E}^{(0)}$) which is analogous to the classification task [65]. The token is randomly initialized at the first position of the embeddings. In addition, we employ sinusoidal positional encoding to record the order information of each frame, which not only enables the proposed transformer to be position-aware but also emphasizes the importance of temporal smoothness in the analysis of video quality. As such, the combined token can be formulated as follows,

$$\mathbf{U}^{(0)} = \left[\mathbf{E}^{(0)}, \ldots, \mathbf{E}^{(t)}, \ldots, \mathbf{E}^{(T)}\right] + \mathbf{E}^{(pos)}, \tag{9}$$

where $\mathbf{E}^{(t)} \in \mathbb{R}^{1\times c}$ and $\mathbf{E}^{(pos)} \in \mathbb{R}^{(1+T)\times c}$ represent the input embeddings and the position token, respectively. $\mathbf{U}^{(0)}$ is the combined embeddings for the MSA module, which is shown in Fig. 4 (b). Given a token embedding, the WPM is used to generate the mapping weights with metadata information s. Formally, the MSA can be formulated as follows,

$$\begin{aligned} \text{MSA}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i, s) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_H) \otimes \mathbf{w}_m, \\ \text{head}_i &= \textit{Softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^{\mathbf{T}}}{\sqrt{C}}\mathbf{V}_i\right), \\ \mathbf{w}_m &= g(s; \mathbf{w}_g), \end{aligned} \tag{10}$$

where $\mathbf{Q}_i$, $\mathbf{K}_i$, and $\mathbf{V}_i$ represent the query, key, and value embeddings for the i-th head. There are a total of H heads with the notation of head_i. $\mathbf{w}_m$ is the adaptive weight according to the WPM. Besides, the $\mathbf{Q}_i$, $\mathbf{K}_i$, and $\mathbf{V}_i$ are also obtained by WPM, which can be expressed as follows,

$$\begin{cases} \mathbf{Q}_i = \text{LN}\left(\mathbf{U}^{(0)}\right) \otimes \mathbf{w}_q \\ \mathbf{K}_i = \text{LN}\left(\mathbf{U}^{(0)}\right) \otimes \mathbf{w}_k \\ \mathbf{V}_i = \text{LN}\left(\mathbf{U}^{(0)}\right) \otimes \mathbf{w}_v \end{cases} \quad \begin{cases} \mathbf{w}_q = g(s; \mathbf{w}_g^q), \\ \mathbf{w}_k = g(s; \mathbf{w}_g^k), \\ \mathbf{w}_v = g(s; \mathbf{w}_g^v), \end{cases} \tag{11}$$

where $\mathbf{w}_q$, $\mathbf{w}_k$, and $\mathbf{w}_v$ are the adaptive weight generated by the WPM to map the input embedding to query, key, and value representations, respectively. Moreover, as shown in Fig. 4 (a), WPM is further used instead of multi-layer perceptrons. It is worth noting that the WPM modules are not weight-sharing. Hence, the WPM generates all the learnable parameters of the transformer encoder, thereby allowing the generation of discriminative features that are adaptive to the reconstructed video. The formulations of the transformer encoder can be expressed as follows,

$$\mathbf{V}^{(l)} = \mathbf{U}^{(l-1)} + \text{MSA}(\text{LN}(\mathbf{U}^{(l-1)}), s),\ l = 1 \cdots L, \tag{12}$$

$$\mathbf{U}^{(l)} = \mathbf{V}^{(l)} + \text{WPM}(\text{LN}(\mathbf{V}^{(l)}), s),\ l = 1 \cdots L, \tag{13}$$

where l represents the index of the transformer, and the maximum number is L. $\mathbf{V}^{(l)}$ is the intermediate embedding of l-th transformer encoder. Finally, the overall quality score $\tilde{Q}$ is regressed using the FC layer derived from the output quality token of the last transformer, which can be expressed as follows,

$$\tilde{Q} = \mathbf{w}_o * \mathbf{U}^{(L,0)} + \mathbf{b}_o, \tag{14}$$

where $\mathbf{w}_o$ and $\mathbf{b}_o$ are the learnable parameters of the FC layer, and $U^{(L,0)}$ stands for the quality token of the L-th transformer.

C. Objective Loss Function

In this subsection, the learnable parameters in the WPM and FC layers are optimized by the difference between the predicted and ground-truth MOS values. We design a joint loss function that ensures the linearity and monotonicity of the predicted video score simultaneously. Specifically, for the linearity term, we minimize the mean absolute error (MAE) among model predictions and MOS values, which can be expressed as follows,

$$\ell_l(\tilde{Q}_i, Q_i) = \frac{1}{B}\sum_{i=1}^{B}\left|\tilde{Q}_i - Q_i\right|, \tag{15}$$

where B is the batch size, $\tilde{Q}_i$ and Q_i represent the predicted quality score and normalized MOS, respectively. For the monotonicity term, we compute the ranking hinge loss according to the following formulation,

$$\ell_m(\tilde{Q}_i, Q_i) = \sum_{i,j}\max\left((\tilde{Q}_i - \tilde{Q}_j)\cdot \text{sgn}(Q_i - Q_j), 0\right), \tag{16}$$

where $\text{sgn}(\cdot)$ is the sign function. Finally, we combine the linearity and monotonicity loss to obtain the joint objective loss function as follows,

$$\ell(\tilde{Q}_i, Q_i) = \ell_l(\tilde{Q}_i, Q_i) + \beta \cdot \ell_m(\tilde{Q}_i, Q_i), \tag{17}$$

where β is the balance parameter between two terms.

IV. VALIDATIONS

In this section, we first introduce the experimental setup, which contains VQA datasets, the implementation details of the proposed method, and the evaluation methodology. Subsequently, we compare the proposed method with the advanced quality assessment methods on the subsampling and compression VQA datasets. Finally, the proposed model is validated with extensive ablation studies.

A. Experimental Setup

1) VQA Datasets: We employ four subsampling and compression VQA datasets to validate the performance of the proposed VQA model, including AVT-VQDB-UHD-1 [15], ETRI-LIVE STSVQ [16], LIVE-YT-HFR [14], and MCL-V [10]. We summarize the details of each dataset as follows:

- AVT-VQDB-UHD-1 (test #4): There are five source videos with 4K resolution and 60 fps frame rate. The source videos are spatially downsampled to five different resolutions (1440/1080/720/480/360p) and temporally downsampled to three frame rates (30/24/15 fps). Then the H.264 is used to compress the subsampled videos with eight pre-defined bitrates. The spatial resolution is modified by the bicubic interpolation. Frame dropping is adopted to reduce the frame rate. The number of reconstructed videos is 120.
- ETRI-LIVE STSVQ: A total of 15 sources are selected, each with a spatial resolution of 4K and a frame rate of 120/60 fps. The Lanczos interpolation is used to downsample the source video to three resolutions (1080/720/520p). By using the frame drop, the source frame rates are reduced by half (60/30 fps). Linear frame interpolation is used to recover the missing frames. Subsampled videos are compressed using HEVC with five adaptive bitrates, resulting in 437 reconstructed videos.
- LIVE-YT-HFR: The dataset solely analyzes the video quality with respect to frame rate variations and compression levels. The frame dropping decreases the frame rates of the six source video to six chosen frame rates (120/92/82/60/30/20 fps). Through the VP9 encoder, the subsampled videos are compressed to yield 480 videos with five compression levels.
- MCL-V: The number of the source video is 12, and the spatial resolution is 1080p and frame rates are 24/30 fps. Lanczos algorithm is applied to downscale the video spatially, and it has a resolution of 720p. The H.264 is used to compress the source and subsampled videos with four adaptive bitrates. The bilinear algorithm super-resolved the compressed video to the original resolution. There are 96 reconstructed videos in total.

It is worth noting that the spatial resolutions and frame rates of the reconstructed videos may not be equal to the source videos in AVT-VQDB-UHD-1 [15] and LIVE-YT-HFR [14]. As such, we apply the bicubic interpolation and frame duplication to upsample the spatial resolution and frame rate, which is the same as the settings in [17]. In order to reduce the bias resulting from uncertainty in randomly training-testing set splitting, the experiment is repeated ten times, and each dataset is divided with the ratio around 80% and 20% for training and testing [16], respectively. Video content is strictly enforced independently between training and testing sets. After ten repeated experiments, we present the average and standard deviation of Spearman's rank-order correlation coefficient (SRCC) as well as the Pearson linear correlation coefficient (PLCC) values.

2) Implementation Details: We adopt the IQA finetuned ResNet50 [62] backbone to extract the quality-aware spatial features of source and reconstructed videos, respectively. The configuration of the spatial feature extractor [22], [62] can be found in the Appendix. The backbone is tailored at the last residual block, containing 2048 feature maps. Notably, the input videos are maintained to the original spatial resolutions and frame rates. We set the batch size to 16. The compact feature dimension C equals 128. There are five transformer encoders in Eqn. (12), and the hidden layer number is 64 on each of its six heads. We train the model by minimizing the joint objective loss (Eqn. (17)) with the Adam optimizer [68], and β is equal to 1. The initial learning rate is set to 10^{-3}, and a weight decay 0.8 is multiplied for every 5 epoch. The total number of training epochs is 100.

3) Evaluation Methodology: The linearity and monotonicity between subjective quality scores and model predictions are calculated by using two standard performance criteria suggested by the video quality experts group (VQEG) [69], *i.e.*, SRCC and PLCC. Models with larger SRCC and PLCC values have better performance. Prior to computing the PLCC value, the predicted quality scores are mapped to the same scale as MOSs using the five-parameter nonlinear logistic function [69], which can be expressed as follows,

$$\hat{Q}_i = \eta_1 \left(\frac{1}{2} - \frac{1}{\exp(\eta_2(\tilde{Q}_i - \eta_3))} \right) + \eta_4 \tilde{Q}_i + \eta_5, \qquad (18)$$

where $\eta_1 \sim \eta_5$ represent the fitting parameters for logistic regression.

B. Experimental Results

Herein, we compare the proposed model with a variety of full-reference IQA (FR-IQA) and FR-VQA models. Video quality scores are calculated for FR-IQA models by averaging the frame-wise image scores, including PSNR, SSIM [23], VIF [66], LPIPS [63], and DISTS [67]. In addition, six state-of-the-art FR-VQA models are selected for comparison: ST-RRED [42], SpEED [43], FRQM [28], C3DVQA [30], VMAF [24], VSTR [17], VFIPS [45], and RankDVQA [46]. ST-RRED [42] and SpEED [43] are FR-VQA models based on natural video statistics. VMAF [24] is a widely used machine learning-based model. The FRQM [28] is a wavelet domain FR-VQA method. C3DVQA [30] VFIPS [45], and RankDVQA [46] are FR-VQA models using advanced deep learning techniques. VSTR [17] is a specifically designed method to measure the quality of the STRA coding videos. In order to make a fair comparison, we use the same training-testing splits to run the FR-IQA and FR-VQA models on each dataset. For the learning-based model, we directly utilize the given models to evaluate the result. For VSTR, we report the results from [17].

TABLE I
PERFORMANCE COMPARISON OF THE PROPOSED MODEL AGAINST OBJECTIVE QUALITY MODELS ON FOUR SPATIAL AND TEMPORAL SUBSAMPLED VQA DATASETS [10], [14]–[16]. THE AVERAGE SRCC AND PLCC RESULTS AMONG TEN SPLITS ALONG WITH STANDARD DEVIATION VALUE IN THE BRACKET. THE BEST TWO RESULTS ARE HIGHLIGHTED IN BOLDFACE AND UNDERLINE, RESPECTIVELY.

Method	ETRI-LIVE STSVQ		AVT-VQDB-UHD-1		LIVE-YT-HFR		MCL-V	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
PSNR	0.490 (0.112)	0.487 (0.127)	0.565 (0.147)	0.605 (0.126)	0.712 (0.082)	0.727 (0.088)	0.541 (0.096)	0.543 (0.103)
SSIM [23]	0.543 (0.119)	0.531 (0.124)	0.709 (0.156)	0.680 (0.123)	0.654 (0.184)	0.635 (0.120)	0.710 (0.122)	0.709 (0.157)
VIF [66]	0.612 (0.098)	0.631 (0.125)	0.691 (0.133)	0.699 (0.173)	0.681 (0.143)	0.702 (0.112)	0.743 (0.078)	0.747 (0.138)
LPIPS [63]	0.641 (0.129)	0.642 (0.153)	0.634 (0.176)	0.620 (0.154)	0.692 (0.075)	0.705 (0.096)	0.761 (0.088)	0.752 (0.121)
DISTS [67]	0.621 (0.137)	0.632 (0.167)	0.713 (0.149)	0.721 (0.132)	0.721 (0.127)	0.729 (0.105)	0.796 (0.105)	0.782 (0.093)
ST-RRED [42]	0.558 (0.133)	0.561 (0.161)	0.733 (0.069)	0.742 (0.109)	0.620 (0.082)	0.623 (0.119)	0.789 (0.132)	0.802 (0.098)
SpEED [43]	0.467 (0.128)	0.478 (0.113)	0.741 (0.113)	0.735 (0.105)	0.550 (0.124)	0.564 (0.136)	0.841 (0.089)	0.843 (0.091)
FRQM [28]	0.347 (0.217)	0.309 (0.169)	0.445 (0.174)	0.466 (0.171)	0.654 (0.131)	0.668 (0.097)	-	-
C3DVQA [30]	0.563 (0.143)	0.578 (0.145)	0.639 (0.118)	0.620 (0.132)	0.730 (0.098)	0.741 (0.137)	0.785 (0.108)	0.792 (0.118)
VMAF [24]	0.665 (0.153)	0.670 (0.180)	0.869 (0.056)	0.875 (0.049)	0.786 (0.121)	0.783 (0.098)	0.828 (0.082)	0.830 (0.087)
VSTR [17]	0.771 (0.114)	0.778 (0.121)	0.800 (0.113)	0.818 (0.134)	0.782 (0.104)	0.765 (0.114)	0.852 (0.070)	0.854 (0.075)
VFIPS [45]	0.644 (0.016)	0.657 (0.108)	0.777 (0.187)	0.789 (0.329)	0.719 (0.245)	0.706 (0.152)	0.757 (0.109)	0.764 (0.096)
RankDVQA [46]	0.755 (0.171)	0.760 (0.103)	0.856 (0.123)	0.860 (0.109)	0.722 (0.145)	0.731 (0.158)	0.832 (0.054)	0.841 (0.048)
STRA-VQA	**0.845 (0.159)**	**0.837 (0.168)**	**0.957 (0.036)**	**0.960 (0.029)**	**0.799 (0.048)**	**0.806 (0.052)**	**0.857 (0.099)**	**0.864 (0.116)**

1) Results on Intra-Dataset: The results of four subsampling and compression VQA datasets are summarized in Table I, from which we are able to have the following observations. First, the STRA-VQA model performs the best compared to both FR-IQA and FR-VQA models on each VQA dataset, indicating the proposed adaptive weights and long-range temporal dependency are effective in evaluating the video quality. Second, the advanced models (*i.e.*, VMAF, VSTR, and STRA-VQA) are VQA-based and equipped with reliable temporal modules, demonstrating the importance of capturing temporal artifacts in video quality prediction. Third, since FRQM is designed for videos with various frame rates, it works better on LIVE-YT-HFR but shows subpar performance on others because the hybrid spatial and temporal distortions are more complicated. Fourth, the transformer-based methods (VFIPS and RankDVQA) demonstrate good performance across four benchmarks, highlighting the significance of long-range dependencies within videos for effective video quality prediction. Last but not least, the specifically designed spatial and temporal subsampled FR-VQA model VSTR achieves competitive results but still underperforms the proposed VQA model.

2) Results on Individual Subsampling Type: Herein, we further investigate the performance of different subsampled spatial resolution and frame rate levels on the ETRI-LIVE STSVQ individually. We adopt the same competing algorithms and experimental settings to conduct the experiments. The SRCC and PLCC results on half and full frame rates are reported in Table II, from which we find an interesting trend that the comparison methods achieve better performance on the video with the full frame rates. It may be due to the fact that the full frame rate videos do not have to interpolate the missing frames, getting rid of a number of temporal distortions, such as jerkiness and flickering. Consequently, a good temporal model is crucial in predicting video quality across different frame rates. In addition, FRQM is specifically designed for quality assessment when the frame rate varies, such that it would be difficult to handle compression distortions.

TABLE II
PERFORMANCE COMPARISON OF THE PROPOSED MODEL AGAINST OBJECTIVE QUALITY MODELS ON ETRI-LIVE STSVQ [16] ACROSS DIFFERENT SUBSAMPLING SPATIAL RESOLUTIONS. THE BEST TWO RESULTS ARE HIGHLIGHTED IN BOLDFACE AND UNDERLINE, RESPECTIVELY.

Method	Half frame rate		Full frame rate		Overall	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
PSNR	0.463	0.472	0.645	0.636	0.490	0.487
SSIM [23]	0.487	0.389	0.805	0.784	0.543	0.531
VIF [66]	0.534	0.549	0.764	0.785	0.612	0.631
LPIPS [63]	0.589	0.623	0.781	0.775	0.641	0.642
DISTS [67]	0.563	0.583	0.765	0.753	0.621	0.632
ST-RRED [42]	0.512	0.623	0.828	0.830	0.558	0.561
SpEED [43]	0.501	0.539	0.850	0.845	0.467	0.478
FRQM [28]	0.185	0.193	-	-	0.347	0.309
C3DVQA [30]	0.432	0.441	0.601	0.612	0.563	0.578
VMAF [24]	0.613	0.624	0.737	0.735	0.665	0.670
VSTR [17]	0.640	0.665	**0.888**	**0.887**	0.771	0.778
VFIPS [45]	0.621	0.650	0.751	0.734	0.644	0.657
RankDVQA [46]	0.633	0.645	0.795	0.798	0.755	0.760
STRA-VQA	**0.816**	**0.790**	0.865	0.852	**0.845**	**0.837**

Based on Table III, we can observe that both IQA and VQA models perform better on high spatial resolutions (2160p). It is reasonable because the upsampling methods inevitably introduce additional spatial distortions, which brings more difficulties in quality prediction. The VSTR also demonstrates competitive performance from low to high spatial resolutions, in addition to the proposed model which achieves the best performance. In summary, we believe the performance gain of

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED MODEL AGAINST OBJECTIVE QUALITY MODELS ON ETRI-LIVE STSVQ [16] ACROSS DIFFERENT SUBSAMPLING FRAME RATES. THE BEST TWO RESULTS ARE HIGHLIGHTED IN BOLDFACE AND UNDERLINE, RESPECTIVELY.

Method		540p		720p		1080p		2160p		Overall	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
FR-IQA	PSNR	0.393	0.405	0.421	0.433	0.457	0.468	0.563	0.564	0.490	0.487
	SSIM [23]	0.535	0.521	0.513	0.509	0.510	0.513	0.582	0.570	0.543	0.531
	VIF [66]	0.581	0.590	0.572	0.580	0.568	0.573	0.634	0.641	0.612	0.631
	LPIPS [63]	0.603	0.611	0.616	0.621	0.634	0.629	0.658	0.661	0.641	0.642
	DISTS [67]	0.596	0.604	0.589	0.593	0.581	0.592	0.678	0.683	0.621	0.632
FR-VQA	ST-RRED [42]	0.534	0.551	0.525	0.532	0.502	0.508	0.631	0.603	0.558	0.561
	SpEED [43]	0.345	0.386	0.435	0.441	0.464	0.473	0.529	0.538	0.467	0.478
	FRQM [28]	0.279	0.283	0.311	0.318	0.316	0.321	0.426	0.449	0.347	0.309
	C3DVQA [30]	0.456	0.463	0.503	0.512	0.537	0.541	0.676	0.684	0.563	0.578
	VMAF [24]	0.521	0.542	0.627	0.635	0.646	0.636	0.740	0.752	0.665	0.670
	VSTR [17]	0.720	0.750	0.770	0.770	0.750	0.740	0.850	0.840	0.771	0.778
	VFIPS [45]	0.638	0.649	0.748	0.760	0.679	0.686	0.759	0.765	0.644	0.657
	RankDVQA [46]	0.644	0.657	0.740	0.761	0.698	0.713	0.767	0.773	0.755	0.760
	STRA-VQA	**0.730**	**0.754**	**0.827**	**0.815**	**0.848**	**0.853**	**0.890**	**0.883**	**0.845**	**0.837**

	PSNR	SSIM	VIF	LPIPS	DISTS	ST-RRED	SpEED	FRQM	C3DVQA	VMAF	VSTR	VFIPS	RankDVQA	STRA-VQA
PSNR	-	-	0	0	0	0	-	1	0	0	0	0	0	0
SSIM	1	-	-	0	0	-	1	1	-	0	0	0	0	0
VIF	1	1	-	-	-	1	1	1	1	0	0	0	0	0
LPIPS	1	1	1	-	-	1	1	1	1	-	0	-	0	0
DISTS	1	1	-	-	-	-	1	1	1	0	0	-	0	0
ST-RRED	1	-	0	0	0	-	1	1	-	0	0	0	0	0
SpEED	-	-	0	0	0	0	-	1	0	0	0	0	0	0
FRQM	0	0	0	0	0	0	0	-	0	0	0	0	0	0
C3DVQA	1	1	-	0	0	-	1	1	-	0	0	0	0	0
VMAF	1	1	1	-	1	1	1	1	0	-	0	-	0	0
VSTR	1	1	1	1	1	1	1	1	1	1	-	1	1	0
VFIPS	1	1	1	-	-	1	1	1	1	0	0	-	0	0
RankDVQA	1	1	1	1	1	1	1	1	1	1	0	1	-	0
STRA-VQA	1	1	1	1	1	1	1	1	1	1	1	1	1	-

(a) ETRI-LIVE STSVQ

	PSNR	SSIM	VIF	LPIPS	DISTS	ST-RRED	SpEED	FRQM	C3DVQA	VMAF	VSTR	VFIPS	RankDVQA	STRA-VQA
PSNR	-	0	0	0	0	0	0	1	0	0	0	0	0	0
SSIM	1	-	-	1	-	0	0	1	1	0	0	0	0	0
VIF	1	-	-	1	-	0	0	1	1	0	0	0	0	0
LPIPS	1	0	0	-	0	0	0	1	-	0	0	0	0	0
DISTS	1	1	1	1	-	-	0	1	1	0	0	0	0	0
ST-RRED	1	1	1	1	-	-	-	1	1	0	0	0	0	0
SpEED	1	1	1	1	1	-	-	1	1	0	0	0	0	0
FRQM	-	0	0	0	0	0	0	-	0	0	0	0	0	0
C3DVQA	1	0	0	-	0	0	0	1	-	0	0	0	0	0
VMAF	1	1	1	1	1	1	1	1	1	-	1	1	-	0
VSTR	1	1	1	1	1	1	1	1	1	0	-	-	0	0
VFIPS	1	1	1	1	1	1	1	1	1	0	-	-	0	0
RankDVQA	1	1	1	1	1	1	1	1	1	0	1	1	-	0
STRA-VQA	1	1	1	1	1	1	1	1	1	1	1	1	1	-

(b) AVT-VQDB-UHD-1

Fig. 5. The matrix represents statistical significance based on F-test derived from (a) ETRI-LIVE STSVQ [16] and (b) AVT-VQDB-UHD-1 [15]. In these matrices, the symbol "1" indicates that the performance of the model represented by the row is statistically superior to that of the model represented by the column. The "0" symbolizes that the row model's performance is statistically inferior. The "-" symbol denotes that the performances of the row and column models are statistically indistinguishable.

the proposed model arises from the following two aspects: 1) the metadata information enables the WPM model to generate adaptive weights, leading to discriminative features; 2) the transformer encoder provides the strong capability in long-range temporal modeling, facilitating the model to capture the complicated spatial and temporal distortions.

C. Statistical Significance Analysis

In this subsection, we conduct the F-test statistical validation [70] on ETRI-LIVE STSVQ [16] and AVT-VQDB-UHD-1 [15]. The results of the F-test between the two models are shown in Fig. 5, where a symbol "1" indicates that the model in the row significantly outperforms the model in the column, while a symbol "0" indicates that the model in the row significantly underperforms the model in the column. A "-" symbol denotes that the performances of the row and column models are statistically indistinguishable. Based on the results, we can find that the deep learning-based VQA models generally perform better than traditional models. In addition, STRA-VQA significantly outperforms all the state-of-the-art objective quality assessment models on ETRI-LIVE STSVQ and AVT-VQDB-UHD-1, indicating the effectiveness of the proposed model for videos that undergo initial subsampling followed by compression.

TABLE IV
ABLATION STUDIES OF THE PROPOSED MODEL ON ETRI-LIVE STSVQ [16] AND AVT-VQB-UHD-1 [15] DATASETS.

Experiment ID	Model Variations	ETRI-LIVE STSVQ		AVT-VQDB-UHD-1	
		SRCC	PLCC	SRCC	PLCC
1	STRA-VQA w concatenate operation	0.843	0.835	0.946	0.953
2	STRA-VQA w/o residual feature	0.817	0.820	0.941	0.943
3	STRA-VQA w/o WPM	0.835	0.829	0.944	0.947
4	STRA-VQA w resolution	0.840	0.833	0.952	0.954
5	STRA-VQA w frame rate	0.839	0.833	0.949	0.950
6	STRA-VQA w/ average token	0.842	0.835	0.952	0.955
7	STRA-VQA w/o MAE loss	**0.847**	**0.840**	0.955	0.957
8	STRA-VQA w/o ranking hinge loss	0.833	0.833	0.939	0.942
9	STRA-VQA w/o IQA finetune	0.823	0.828	0.936	0.940
10	STRA-VQA (Default)	0.845	0.837	**0.957**	**0.960**

D. Ablation Studies

A series of ablation experiments are conducted on the ETRI-LIVE STSVQ [16] and AVT-VQB-UHD-1 [15] datasets to reveal the functionality of each component to the proposed method.

1) Impact of Residual Features: From Eqn. (2) to Eqn. (5), we compute the residual feature statistics (*i.e.*, mean and standard deviation) of the source and reconstructed videos and concatenate them together to obtain the final feature representations. First, We compare the proposed model with the model that directly concatenates the spatial feature of the source and reconstructed videos. The results are shown in Experiment 1 in Table IV, from which we observe the direct feature concatenation presents competitive experimental results but still underperforms the proposed model. To further verify whether the source videos are beneficial, we solely feed the reconstructed videos into the quality regression model, reducing it to a no-reference video quality model. As shown in Experiment 2 of Table IV, we find that adding the source videos to the proposed model results in better results owing to the guidance of the reference information.

2) Impact of WPM: One of the major contributions of this work is the WPM module that takes the metadata information as input and predicts adaptive weights. We remove the WPM module, and the weights of the network are retrained by the input videos. The test results are listed in Experiment 3 of Table IV. It is apparent that the WPM improves the model performance on both VQA datasets. In addition, we have conducted additional experiments to evaluate the effectiveness of each component of the metadata information, including the resolution and frame rate. We systematically remove each component and explore the impact on the overall performance. As shown in Table IV, Experiments 4 and 5 demonstrate that both the resolution and frame rate significantly contribute to the performance of our model and perform better than the model (Experiment 3) without the metadata information. When either of these components is removed, there is a noticeable decrease in performance, confirming their importance in the WPM. We may draw the conclusion that the proposed WPM module is of benefit to the quality prediction for the videos with subsampling and compression distortions.

3) Impact of Quality Token: Herein, we compare the model that leverages a quality token ($\mathbf{E}^{(0)}$) to obtain the final quality score to the mean pooling scheme. To be specific, we compute the average of all the feature embeddings to obtain the final feature representation that is regressed by one FC layer (referring to Eqn. (14)). The comparison results are shown in Table IV (Experiment 6), from which we can see that the average pooling scheme underperforms the proposed model with the quality token.

4) Impact of Loss Function: We further study the model trained with a single loss function, *i.e.*, MAE or ranking hinge loss, to verify the functionalities of the proposed joint loss function, as shown in Eqn. (17). As shown in Experiment 7 and 8 of Table IV, we can find that two variants demonstrate competitive performance on two datasets. In addition, it is worth noting that though the model individually trained with rank hinge loss (Experiment 8) outperforms the joint loss function, the rank loss emphasizes monotonicity and cannot ensure the same range between the predicted score and MOS. As a result, we use the joint loss function where the MAE is complementary to the ranking hinge loss, obtaining a more meaningful quality score when compared with a single ranking hinge loss function.

5) Impact of Spatial Feature Extractor: Herein, we validate the utility of the IQA finetuned feature extractor (ResNet50 [62]). We ablate the IQA finetuned weights in the backbone [22] and directly apply the weights pre-trained from ImageNet [59] to train the model. The results are shown in Table IV (Experiment 9), where significant performance degradation can be observed in two datasets in terms of SRCC and PLCC values. As such, the performance improvement of the fine-tuned IQA backbone is attributed to the shrinkage of the domain gap between image recognition and quality evaluation, allowing the feature extractor to compensate for spatial distortions.

V. APPLICATION OF STRA-VQA: BITRATE LADDER OPTIMIZATION

Apart from objectively assessing the quality of videos compressed with STRA coding, an effective FR-VQA model should be capable of optimizing the bitrate ladder to provide

TABLE V
QUANTITATIVE COMPARISON ON FIVE VIDEO SEQUENCES OF THE PROPOSED MODEL AGAINST SSIM [23] AND VMAF [24] WITH RESPECT TO THE ACCURACY AND BD-RATE SAVING OF THE PREDICTED CONVEX HULLS.

Video detail				Accuracy			BD-rate (%)		
Name	Duration	Resolution	Frame rate	SSIM	VMAF	STRA-VQA	SSIM	VMAF	STRA-VQA
Daydreamer	8s	2160p	60 fps	0.333	0.167	0.333	4.180	-2.901	-3.711
Fr-041 debris	8s	2160p	60 fps	0.333	0.667	1.000	1.985	-7.417	-13.230
Giftmord	8s	2160p	60 fps	0.500	0.833	0.667	-7.207	-7.221	-0.612
Sparks cut 13	8s	2160p	60 fps	0.500	0.833	1.000	-8.973	-20.049	-19.994
Sparks cut 15	8s	2160p	60 fps	0.167	0.500	0.500	10.256	16.799	1.693
Average				0.367	0.600	**0.700**	0.048	-4.158	**-7.171**

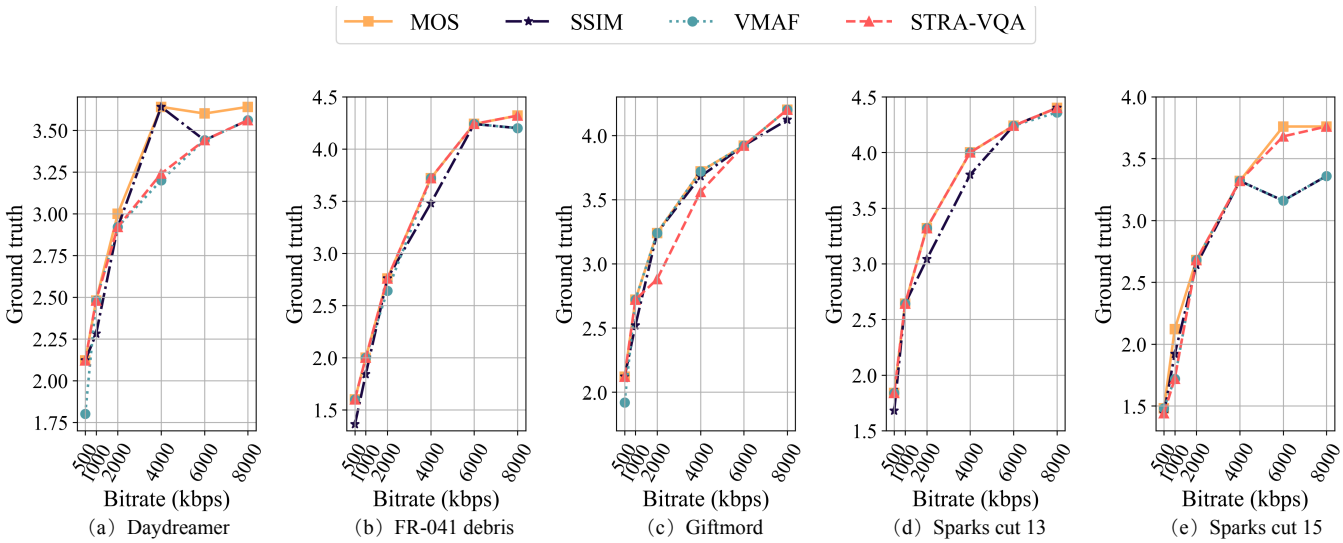

Fig. 6. Illustration of the rate-distortion curves in terms of ground truth value based on the convex hull of SSIM [23], VMAF [24], and the proposed model (STRA-VQA), respectively.

optimal content distribution services. As such, we compare the proposed model with two classical image and video quality measures, *i.e.*, SSIM [23] and VMAF [24], on the AVT-VQDB-UHD-1 (test #4) dataset [15]. As mentioned in Sec. IV-A1, AVT-VQDB-UHD-1 (test #4) contains five available source videos with 4K spatial resolution and 60 fps frame rates as listed in Table V. Five spatial resolutions and three frame rates are used to generate the subsampled videos. These subsampled videos are further compressed with eight pre-defined bitrates. Within the eight pre-defined bitrates, we select the videos compressed by six bitrates, mimicking a wide range of real-world bandwidth conditions and subsampling strategies. The combination of the chosen bitrates, spatial resolutions, and frame rates can be found in the Appendix.

For a given bitrate, the objective is to find the optimal subsampling strategy according to an objective quality assessment model. As such, for each quality assessment model, we can generate a bitrate ladder that takes both the temporal and spatial resolutions into consideration. The optimal combination that is selected by MOS is regarded as the ground-truth. By comparing the predicted one from the objective quality assessment model with the one from MOS, the accuracy can be obtained accordingly. As shown in Table V, the proposed method outperforms competing objective quality models, demonstrating that the proposed model is better aligned with human visual perception. Additionally, based on the formation of the convex hull, we present the rate-distortion curves between the bitrates and the corresponding MOS values in Fig. 6. It is worth noting that the MOS value is determined by the video with the optimal quality score based on a specific quality assessment model. As shown in Fig. 6, it is reasonable that the upper bounder is the bitrate ladder optimized by MOS. The curve generated by STRA-VQA is approaching the results of MOS and performs better than SSIM and VMAF in most test videos, such as *Fr-041 debris*, *Sparks cut 13*, and *Sparks cut 15*. Moreover, we calculate the Bjontegaard delta rates (BD-rate) [71] for the convex hull results and compare them to a randomly selected solution for each bitrate setting. We repeat this process 100 times to reduce uncertainty bias and summarize the average BD-rate savings in Table V. Our findings reveal several interesting observations. The bitrate ladder optimized by SSIM shows an apparent loss for some sequences, indicating that the mixture of subsampling and compression distortions prevent the IQA model from performing well. In addition, the proposed STRA-VQA model achieves the highest average bitrate savings for the five videos, which demonstrates its superiority in predicting bitrate ladders in STRA coding.

VI. Conclusions

We propose an FR-VQA model STRA-VQA that effectively evaluates the quality of videos that are corrupted with subsampling and compression. The novelty of our model lies in the integration of downsampling metadata information into a transformer encoder, adapting the weights of the FC layers to the down-sampled video. This enables the model to generate discriminative features and capture long-range spatio-temporal quality dependencies. Experimental results demonstrate that STRA-VQA outperforms state-of-the-art methods in terms of SRCC and PLCC on four VQA datasets. Additionally, our model has been applied to bitrate ladder optimization, yielding promising results in bitrate savings at the same quality levels.

Appendix

In this section, we provide more details of the configuration of the spatial feature extractor [22], [62] in Fig. VI. Furthermore, the chosen bitrates, spatial resolutions, and frame rates are shown in Table VII.

TABLE VI
THE CONFIGURATION OF THE SPATIAL FEATURE EXTRACTOR – RESNET-50 [22], [62]. WE OMIT THE NONLINEARITY, NORMALIZATION, AND SKIP CONNECTION FOR SIMPLICITY.

Layer Name	Network Parameter
Convolution	7×7, 64, stride 2
Max Pooling	3×3, stride 2
Residual Block 1	[1×1, 64, stride 1 3×3, 64, stride 1 1×1, 256, stride 1] ×3
Residual Block 2	[1×1, 128, stride 1 3×3, 128, stride 2 1×1, 512, stride 1] ×1 [1×1, 128, stride 1 3×3, 128, stride 1 1×1, 512, stride 1] ×3
Residual Block 3	[1×1, 256, stride 1 3×3, 256, stride 2 1×1, 1024, stride 1] ×1 [1×1, 256, stride 1 3×3, 256, stride 1 1×1, 1024, stride 1] ×5
Residual Block 4	[1×1, 512, stride 1 3×3, 512, stride 2 1×1, 2048, stride 1] ×1 [1×1, 512, stride 1 3×3, 512, stride 1 1×1, 2048, stride 1] ×2

TABLE VII
THE EXPERIMENTAL SETTINGS OF DIFFERENT BITRATES, RESOLUTIONS, AND FRAME RATES.

Bitrate (kbps)	Subsampling strategy (resolution, frame rate)
500	(360p, 15 fps), (360p, 24 fps), (480p, 15 fps)
1000	(360p, 24 fps), (480p, 15 fps), (480p, 24 fps), (720p, 24 fps)
2000	(480p, 24 fps), (720p, 24 fps), (720p, 30 fps), (1080p, 24 fps)
4000	(720p, 30 fps), (1080p, 24 fps), (1080p, 30 fps), (1440p, 30 fps)
6000	(1080p, 30 fps), (1440p, 30 fps), (1440p, 60 fps), (2160p, 30 fps)
8000	(1440p, 60 fps), (2160p, 30 fps), (2160p, 60 fps)

References

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] S. Ma, L. Zhang, S. Wang, C. Jia, S. Wang, T. Huang, F. Wu, and W. Gao, "Evolution of AVS video coding standards: Twenty years of innovation and development," *Science China Information Sciences*, vol. 65, no. 9, pp. 1–24, 2022.

[4] J. Chen, Y. Ye, and S. Kim, "Algorithm description for versatile video coding and test model 6 (VTM 6)," in *Joint Video Exploration Team*, 2019, pp. 1–92.

[5] D. Mukherjee, J. Han, J. Bankoski, R. Bultje, A. Grange, J. Koleszar, P. Wilkins, and Y. Xu, "A technical overview of VP9—the latest open-source video codec," *SMPTE Motion Imaging Journal*, vol. 124, no. 1, pp. 44–54, 2015.

[6] Y. Chen, D. Murherjee, J. Han, A. Grange, Y. Xu, Z. Liu, S. Parker, C. Chen, H. Su, U. Joshi *et al.*, "An overview of core coding tools in the AV1 video codec," in *Picture Coding Symposium*, 2018, pp. 41–45.

[7] A. Anne, L. Zhi, M. Megha, C. J. De, and R. David, "Per-title encode optimization," Dec. 2015. [Online]. Available: https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2

[8] M. Afonso, F. Zhang, and D. R. Bull, "Video compression based on spatio-temporal resolution adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 275–280, 2018.

[9] V. V. Menon, H. Amirpour, M. Ghanbari, and C. Timmerer, "Perceptually-aware per-title encoding for adaptive video streaming," in *IEEE International Conference on Multimedia and Expo*, 2022, pp. 1–6.

[10] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-C. J. Kuo, "MCL-V: A streaming video quality assessment database," *Journal of Visual Communication and Image Representation*, vol. 30, pp. 1–9, 2015.

[11] M. Cheon and J.-S. Lee, "Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 7, pp. 1467–1480, 2017.

[12] A. Mackin, F. Zhang, and D. R. Bull, "A study of high frame rate video formats," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1499–1512, 2018.

[13] Q. Huang, S. Y. Jeong, S. Yang, D. Zhang, S. Hu, H. Y. Kim, J. S. Choi, and C.-C. J. Kuo, "Perceptual quality driven frame-rate selection (PQD-FRS) for high-frame-rate video," *IEEE Transactions on Broadcasting*, vol. 62, no. 3, pp. 640–653, 2016.

[14] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Access*, vol. 9, pp. 108 069–108 082, 2021.

[15] R. R. R. Rao, S. Göring, W. Robitza, B. Feiten, and A. Raake, "AVT-VQDB-UHD-1: A large scale video quality database for UHD-1," in *IEEE International Symposium on Multimedia*, 2019, pp. 170–177.

[16] D. Y. Lee, S. Paul, C. G. Bampis, H. Ko, J. Kim, S. Y. Jeong, B. Homan, and A. C. Bovik, "A subjective and objective study of space-time subsampled video quality," *IEEE Transactions on Image Processing*, vol. 31, pp. 934–948, 2021.

[17] D. Y. Lee, J. Kim, H. Ko, and A. C. Bovik, "Video quality model of compression, resolution and frame rate adaptation based on space-time regularities," *IEEE Transactions on Image Processing*, to appear 2022.

[18] A. Raake, M.-N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P. 1203.1," in *IEEE International Conference on Quality of Multimedia Experience*, 2017, pp. 1–6.

[19] R. R. R. Rao, S. Göring, P. List, W. Robitza, B. Feiten, U. Wüstenhagen, and A. Raake, "Bitstream-based model standard for 4K/UHD: ITU-T P. 1204.3—model details, evaluation, analysis and open source implementation," in *International Conference on Quality of Multimedia Experience*, 2020, pp. 1–6.

[20] R. R. Ramachandra Rao, S. Göring, and A. Raake, "AVQBits-Adaptive

video quality model based on bitstream information for various video applications," *IEEE Access*, vol. 10, pp. 80 321–80 351, 2022.
[21] R. R. Ramachandra Rao, S. Göring, and A. Raake, "Enhancement of pixel-based video quality models using meta-data," *Electronic Imaging*, vol. 2021, no. 9, pp. 264–271, 2021.
[22] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, Mar. 2021.
[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[24] L. Zhi, A. Anne, K. Ioannis, M. Anush, and M. Megha, "Toward a practical perceptual video quality metric," Jun. 2016. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652
[25] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, Feb. 2010.
[26] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao, "HodgeRank on random graphs for subjective video quality assessment," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 844–857, 2012.
[27] Q. Xu, J. Xiong, Q. Huang, and Y. Yao, "Robust evaluation for quality of experience in crowdsourcing," in *ACM International Conference Multimedia*, 2013, pp. 43–52.
[28] F. Zhang, A. Mackin, and D. R. Bull, "A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling," in *IEEE International Conference on Image Processing*, 2017, pp. 300–304.
[29] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *European Conference on Computer Vision*, 2018, pp. 219–234.
[30] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, "C3DVQA: Full-reference video quality assessment with 3D convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 4447–4451.
[31] 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 Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1903–1916, 2021.
[32] H. Zhu, B. Chen, L. Zhu, and S. Wang, "Learning spatiotemporal interactions for user-generated video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1031–1042, 2023.
[33] Q. Xu, J. Xiong, X. Cao, and Y. Yao, "Parsimonious mixed-effects HodgeRank for crowdsourced preference aggregation," in *ACM International Conference on Multimedia*, 2016, pp. 841–850.
[34] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 1–21, 2010.
[35] Q. Wu, H. Li, F. Meng, K. N. Ngan, B. Luo, C. Huang, and B. Zeng, "Blind image quality assessment based on multichannel feature fusion and label transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 425–440, 2015.
[36] Q. Wu, H. Li, K. N. Ngan, and K. Ma, "Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2078–2089, 2017.
[37] B. Chen, H. Li, H. Fan, and S. Wang, "No-reference screen content image quality assessment with unsupervised domain adaptation," *IEEE Transactions on Image Processing*, vol. 30, pp. 5463–5476, 2021.
[38] R. Ma, Q. Wu, K. N. Ngan, H. Li, F. Meng, and L. Xu, "Forgetting to remember: A scalable incremental learning framework for cross-task blind image quality assessment," *IEEE Transactions on Multimedia*, to appear 2023.
[39] P. V. Vu and D. M. Chandler, "ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *Journal of Electronic Imaging*, vol. 23, no. 1, pp. 1 – 25, 2014.
[40] J. Wu, Y. Liu, W. Dong, G. Shi, and W. Lin, "Quality assessment for video with degradation along salient trajectories," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2738–2749, Nov. 2019.
[41] K. Zeng and Z. Wang, "3D-SSIM for video quality assessment," in *IEEE International Conference on Image Processing*, 2012, pp. 621–624.
[42] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684–694, 2012.
[43] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1333–1337, 2017.
[44] Y. Liu, J. Wu, A. Li, L. Li, W. Dong, G. Shi, and W. Lin, "Video quality assessment with serial dependence modeling," *IEEE Transactions on Multimedia*, vol. 24, pp. 3754–3768, 2021.
[45] Q. Hou, A. Ghildyal, and F. Liu, "A perceptual quality metric for video frame interpolation," in *European Conference on Computer Vision*. Springer, 2022, pp. 234–253.
[46] C. Feng, D. Danier, F. Zhang, and D. Bull, "RankDVQA: Deep VQA based on ranking-inspired hybrid training," in *IEEE Winter Conference on Applications of Computer Vision*, 2024, pp. 1648–1658.
[47] R. M. Nasiri, Z. Duanmu, and Z. Wang, "Temporal motion smoothness and the impact of frame rate variation on video quality," in *IEEE International Conference on Image Processing*, 2018, pp. 1418–1422.
[48] F. Yang and S. Wan, "Bitstream-based quality assessment for networked video: A review," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 203–209, 2012.
[49] ITU, "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport—quality integration module," 2021. [Online]. Available: https://www.itu.int/rec/T-REC-P.1203.3/en
[50] C. Keimel, J. Habigt, M. Klimpke, and K. Diepold, "Design of no-reference video quality metrics with multiway partial least squares regression," in *International Workshop on Quality of Multimedia Experience*, 2011, pp. 49–54.
[51] M.-N. Garcia, D. Dytko, and A. Raake, "Quality impact due to initial loading, stalling, and video bitrate in progressive download video services," in *International Workshop on Quality of Multimedia Experience*, 2014, pp. 129–134.
[52] N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, and P. Demeester, "Constructing a no-reference H. 264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 8, pp. 1322–1333, 2013.
[53] D. C. Mocanu, J. Pokhrel, J. P. Garella, J. Seppänen, E. Liotou, and M. Narwaria, "No-reference video quality measurement: added value of machine learning," *Journal of Electronic Imaging*, vol. 24, no. 6, pp. 061 208–061 208, 2015.
[54] A. Raake, S. Borer, S. M. Satti, J. Gustafsson, R. R. R. Rao, S. Medagli, P. List, S. Göring, D. Lindero, W. Robitza *et al.*, "Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P. 1204," *IEEE Access*, vol. 8, pp. 193 020–193 049, 2020.
[55] L. Toni, R. Aparicio-Pardo, K. Pires, G. Simon, A. Blanc, and P. Frossard, "Optimal selection of adaptive streaming representations," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 11, no. 2s, pp. 1–26, 2015.
[56] J. De Cock, Z. Li, M. Manohara, and A. Aaron, "Complexity-based consistent-quality encoding in the cloud," in *IEEE International Conference on Image Processing*, 2016, pp. 1484–1488.
[57] Z. Duanmu, W. Liu, Z. Li, and Z. Wang, "Modeling generalized rate-distortion functions," *IEEE Transactions on Image Processing*, vol. 29, pp. 7331–7344, 2020.
[58] H. Amirpour, C. Timmerer, and M. Ghanbari, "PSTR: Per-title encoding using spatio-temporal resolutions," in *IEEE International Conference on Multimedia and Expo*, 2021, pp. 1–6.
[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[60] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016, pp. 694–711.
[61] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of full-reference image quality models for optimization of image processing systems," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1258–1281, 2021.
[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
[64] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *ACM International Conference on Multimedia*, 2019, pp. 2351–2359.

[65] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021, pp. 1–22.
[66] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.
[67] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.
[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
[69] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," 2000. [Online]. Available: http://www.vqeg.org
[70] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[71] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *ITU SG16 Doc. VCEG-M33*, 2001.

Hanwei Zhu received the B.S. and M.S. degrees from the Jiangxi University of Finance and Economics, Nanchang, China, in 2017 and 2020, respectively. He is currently pursuing a Ph.D. degree in the Department of Computer Science, City University of Hong Kong. His research interest includes perceptual image processing and computational photography.

Baoliang Chen (Member, IEEE) received his B.S. degree in Electronic Information Science and Technology from Hefei University of Technology, Hefei, China, in 2015, his M.S. degree in Intelligent Information Processing from Xidian University, Xian, China, in 2018, and his Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2022. He is currently a postdoctoral researcher with the Department of Computer Science, City University of Hong Kong. His research interests include image/video quality assessment and transfer learning.

Lingyu Zhu received the B.Eng. degree from the Wuhan University of Technology in 2018 and MA.Eng degree from Hong Kong University of Science and Technology in 2019. He is currently pursuing a Ph.D. degree at the City University of Hong Kong. His research interests include image/video quality assessment, image/video processing, and deep learning.

Peilin Chen received the B.S. degree in Software Engineering from Sun Yat-sen University, Guangzhou, China, in 2018 and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong SAR, China, in 2023. He is currently a Postdoctoral Researcher with the Department of Computer Science, City University of Hong Kong. His current research interests include visual data processing and semantic communication.

Linqi Song (Senior Member, IEEE) received the B.S. and M.S. degrees from Tsinghua University, China, and the Ph.D. degree in electrical engineering from the University of California at Los Angeles (UCLA), USA. He was a Post-Doctoral Scholar with the Department of Electrical and Computer Engineering, UCLA. He is currently an Assistant Professor with the Department of Computer Science, City University of Hong Kong, and a Research Scientist with the City University of Hong Kong Shenzhen Research Institute. His research interests include artificial intelligence, information theory, machine learning, and big data. He has received the Hong Kong RGC Early Career Scheme in 2019 and the Best Paper Awards from IEEE MIPR 2020 and China Communications 2023.

Shiqi Wang (Senior Member, IEEE) received the B.S. degree in computer science from the Harbin Institute of Technology in 2008 and the Ph.D. degree in computer application technology from Peking University in 2014. From 2014 to 2016, he was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. From 2016 to 2017, he was a Research Fellow with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore. He is currently an Associate Professor with the Department of Computer Science, City University of Hong Kong. He has proposed more than 50 technical proposals to ISO/MPEG, ITU-T, and AVS standards, and authored or coauthored more than 200 refereed journal articles/conference papers. His research interests include video compression, image/video quality assessment, and image/video search and analysis. He received the Best Paper Award from IEEE VCIP 2019, ICME 2019, IEEE Multimedia 2018, and PCM 2017. His coauthored article received the Best Student Paper Award in the IEEE ICIP 2018. He was a recipient of the 2021 IEEE Multimedia Rising Star Award in ICME 2021. He served or serves as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON CYBERNETICS. He is also the technical program co-chair of IEEE ICME 2024.