Learning Spatiotemporal Interactions for User-Generated Video Quality Assessment

Hanwei Zhu, Baoliang Chen, Lingyu Zhu, and Shiqi Wang, *Senior Member, IEEE*

***Abstract*—Distortions from spatial and temporal domains have been identified as the dominant factors that govern the visual quality. Though both have been studied independently in deep learning-based user-generated content (UGC) video quality assessment (VQA) by frame-wise distortion estimation and temporal quality aggregation, much less work has been dedicated to the integration of them with deep representations. In this paper, we propose a SpatioTemporal Interactive VQA (STI-VQA) model based upon the philosophy that video distortion can be inferred from the integration of both spatial characteristics and temporal motion, along with the flow of time. In particular, for each timestamp, both the spatial distortion explored by the feature statistics and local motion captured by feature difference are extracted and fed to a transformer network for the motion aware interaction learning. Meanwhile, the information flow of spatial distortion from the shallow layer to the deep layer is constructed adaptively during the temporal aggregation. The transformer network enjoys an advanced advantage for long-range dependencies modeling, leading to superior performance on UGC videos. Experimental results on five UGC video benchmarks demonstrate the effectiveness and efficiency of our STI-VQA model, and the source code will be available online at https://github.com/h4nwei/STI-VQA.**

***Index Terms*—No-reference video quality assessment, user-generated content, vision transformer.**

I. Introduction

THE services of user-generated content (UGC) videos on the social media have been growing explosively [1]. According to the report [2], internet video traffic from content delivery networks is estimated to exceed 82% of all bandwidth in 2022. The blossom of video data brings the exponential increase in the demand for high quality video services. However, the compression, unprofessional editing, and uncontrolled acquisition condition usually introduce visual distortions, leading to the unavoidable visual quality degradation. In this regard, there has been a strong desire of objective UGC video quality assessment (UGC-VQA) algorithms.

Generally speaking, UGC-VQA can be categorized into subjective and objective quality assessment [3]. Subjective VQA conducts the user studies in a controllable laboratory [4]–[7] or on crowdsourcing platforms [8]–[10], which provides reliable evaluation results since the human visual system (HVS) is the ultimate receiver of the UGC videos. However, subjective testing is time-consuming and expensive. By contrast, objective VQA offers an alternative way based upon computational algorithms to automatically predict video quality. For the UGC videos, the intrinsic unavailability of reference videos casts the UGC-VQA to a no-reference quality assessment task.

Although numerous no-reference VQA (NR-VQA) models demonstrate competitive performance on synthetic distortions [13]–[16], there is still a large domain gap between the synthetic and realistic distortions [17] for UGC videos, leading to unsatisfactory accuracy in predicting the quality. The major impediment to most NR-VQA models is the unpredictable temporal artifacts, which may be caused by fast-moving objects, sudden camera shaking, and rapid zoomed-in/out. In recent years, there are numerous learning-based NR-VQA models proposed to model the temporal characteristics explicitly and implicitly [3], [10], [18], [19], [19]–[22]. Regarding the explicit temporal modeling methods [3], [18], [22], the handcrafted methods show competitive performance on several small-scale benchmarks [4]–[6] using the statistical modeling in temporal domain. However, they are struggling on large-scale datasets [7], [10]. To jointly capture spatiotemporal features, there are still methods based upon 3D convolutional neural network (3D-CNN) or the motion knowledge transfer learned from action recognition task [10], [14], [17]. Despite that the performance has been boosted to some extent, those methods are limited by the scanty receptive field and substantial computational burden [23]. Moreover, recurrent neural networks (RNNs) and temporal pooling strategies have been adopted to implicitly model the motion information [19], [21]–[25], while those RNN-based methods suffer from gradient vanishing, especially for the long-range videos.

In this paper, we propose a SpatioTemporal Interactive VQA (STI-VQA) model and the design philosophy lies in two aspects: 1) the short-term memory persuades the visibility of distortions to be govern by the interactions of neighbouring components in spatiotemporal domains [26], [27]; 2) the long-term memory suggests that an exciting consideration is the possibility of involving the long dependencies based upon the aggregation of frame-wise quality [10], [28]. Herein, the image classification pre-trained backbones, which contain rich visual information and tend to be invariant to moderate geometric transformations (*i.e.*, background motion) [29], are adopted. The interaction in spatiotemporal space, which can be

This work is supported in part by the National Natural Science Foundation of China under 62022002, in part by Shenzhen Virtural University Park, The Science Technology and Innovation Committee of Shenzhen Municipality (Project No: 2021Szvup128), in part by the Hong Kong Research Grants Council General Research Fund (GRF) under Grant 11203220. (*Corresponding author: Shiqi Wang.)*

Hanwei Zhu, Baoliang Chen, and Lingyu Zhu are with Department of Computer Science, City University of Hong Kong, Hong Kong, China. (e-mail: {hanwei.zhu, blchen6-c, lingyzhu}@my.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).

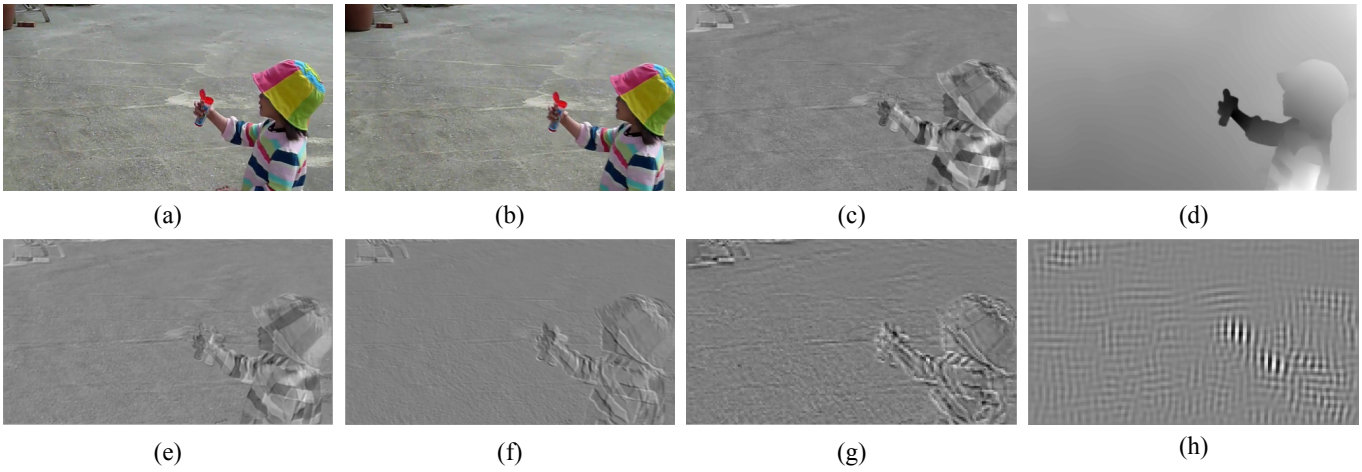

Fig. 1. Examples of the frame difference, optical flow, and motion-aware feature maps from four stages of the ResNeXt-101 [11]. (a)&(b) The consecutive frames of the "4265466447.mp4" video in KoNViD-1k [5]; (c) The frame-based motion map of (b), computed by subtraction of (b) and (a); (d) The flow map of (b), computed by RAFT [12]; (e)-(h) The selected motion-aware feature maps of (b), computed by subtraction the corresponding feature maps of (b) and (a) from the first to fourth stage with ResNeXt-101 [11].

achieved by the comparisons of features from the consecutive frames [28], [30]–[32], is specifically modeled based upon the difference between the consecutive frames in the feature domain. The design philosophy is intrinsically consistent with characteristics of HVS on perceiving the motion in videos. As shown in Fig. 1, we can observe the feature difference maps of each stage (*i.e.*, features at the same scales) highlight the moving objects, especially for the feature maps which contain more high-level semantic information. Finally, a hierarchical feature interaction transformer is designed to fully exploit the long-range dependencies of the spatiotemporal embeddings. Extensive experimental results validate that the proposed method outperforms the state-of-the-art models by a large margin on five UGC-VQA benchmarks. The main contributions of this paper can be summarized as follows,

- We propose a spatiotemporal interactive NR-VQA model, rooted in the view that HVS perceives the video quality in spatial and temporal domains interactively. The local motion introduced in frame-wise distortion estimation, and cross-scale interaction introduced in temporal aggregation lead to more discriminative feature learned for quality assessment.
- We propose a simple yet effective motion-aware feature extractor in capturing the distortion features in spatiotemporal domain. The interaction between local motion and spatial distortion is further constructed along with the flow of the frame.
- We develop a transformer network to model the long-dependency within a videos sequence. Moreover, the cross-scale feature communication is incorporated by a specific token introduced in the transformer, enabling the feature interactions at different scales.

The remainder of this paper is organized as follows. In Section II, we provide a brief review on VQA datasets, NR-VQA models, and several closely relevant vision transformer architectures. Then we describe our proposed method in detail in Section III. The comprehensive experiments by comparing the proposed method with state-of-the-art NR-VQA methods are reported in Section IV. We draw the conclusions of this paper in Section V.

II. Related Works

In this section, we first provide a review of the existing VQA benchmarks. Then the traditional and deep learning-based NR-VQA models are introduced, followed by a brief summary of the vision transformer.

A. Databases for VQA

Researchers have dedicated numerous efforts to establishing high-quality VQA benchmarks since they provide reliable subjective data and greatly advance the development of VQA models. In particular, the VQA datasets can be divided into two categories according to the way how distorted videos are generated. Before 2014, the VQA datasets were dominated by synthetic distortions introduced from the video compression and transmission, constrained by limited scenes and distortion types [33]–[36]. Sehadrinathan *et al*. built the first synthetic distortions VQA dataset, containing 10 reference and 160 distorted videos [33]. Vu and Chandler applied more compression and transmission distortion types to construct a larger VQA dataset [36]. Waterloo IVC 4K VQA dataset is composed of 20 pristine 4K videos, and five video codecs with four distortion levels have been adopted to generate $1,200$ videos [37]. Wang *et al*. used the H. 264/AVC encoder to compress the UGC videos, building a large-scale VQA dataset with $7,200$ distorted videos [38]. These synthetically distorted VQA datasets have witnessed the success of many full-reference VQA (FR-VQA) models, such as MOVIE [28], FAST [39], VMAF [31], and DeepVQA [32]. Akamine *et al*. also proposed a adaptive spatial resolution reduction full-reference framework that saves the inference time without performance decrease [40]. However, those FR-VQA models are not applicable to the UGC video due to the unavailability of reference videos.

To address these issues, researchers have resorted to building the VQA datasets with realistic distortions. Nuutinen *et al*. used 78 video recording devices to capture five in-the-wild scenes, establishing the first realistic distortion VQA dataset [4]. Ghadiyaram *et al*. constructed a small-scale dataset that contains six typical in-capture distortions (artifacts, color, exposure, focus, sharpness, and stabilization) in smartphone cameras [6]. LIVE-VQC screened 585 videos from 1,000 videos using 101 devices, and a crowdsourcing subjective testing was carried out on Amazon Mechanical Turk (AMT) [8]. KoNViD-1k was the first VQA dataset that collects 1,200 UGC videos from a published large-scale video dataset [41] based on several strict selection strategies. Wang *et al*. constructed a YouTube UCG-VQA dataset, containing 1,380 AMT labeled videos [9], and recently more annotations (*i.e*., content labels and distortion types) are added to each video [42]. Ying *et al*. constructed so far the largest UGC VQA dataset that includes 38,811 source videos and 116,433 video patches [10]. In addition, Li *et al*. created the first professional UGC (PUGC) VQA dataset with comprehensive subjective user studies [7].

B. Objective NR-VQA Models

It is practical to design a NR-VQA model to evaluate video quality since the pristine counterpart is not always available. Herein, we briefly introduce the NR-VQA models, especially focusing on the temporal modeling due to its essentiality in the VQA model. The NR-VQA models share the pipeline of feature extraction, feature regression, and quality pooling. Extending the study of natural scene statistics (NSS) to the 3D domain have been widely exploited, aiming to jointly capture the spatial and temporal artifacts, such as 3D discrete cosine transformation (DCT) [43], 3D Gabor coefficients [44], and 3D mean subtracted and contrast normalized (MSCN) [45]. Manasa *et al*. applied the local and global optical flow statistics to model the motion activities [30]. Saad *et al*. proposed motion coherency models via frame difference to predict the video quality blindly [46]. Korhonen *et al*. classified the feature extraction into low complexity feature from every second frame and high complexity feature using keyframes, and the motion consistency and spatial features were averaged to obtain the quality score [18]. Moreover, Tu *et al*. conducted a comprehensive study on video temporal pooling [25], emphasizing the relative importance of each frame-level quality score. With the popularity of deep learning technologies, 3D CNN was used to model the spatiotemporal features [10], [14], [17] but they may not be appropriate for long-duration videos owing to the limited receptive field. RNN-based model attempted to build the long-range temporal effect with the content-aware features, including VSFA [19], MDTVSFA [21], RIRNet [47], and GSTVQA [24]. Alamgeer *et al*. employed a spatial and temporal selection method to train a CNN NR-VQA model, which largely decreases the running time and achieves competitive performance on synthetic distortion datasets [48]. In addition, Li *et al*. transferred the motion knowledge from the pre-trained video action recognition task, boosting the prediction accuracy by a large margin [23]. Despite existing works having shown remarkable progress, they still contain several intrinsic limitations as mentioned in Section I. In this paper, inspired by human short- and long-term memory mechanisms, a simple yet efficient motion-aware feature learning module is constructed, getting rid of the high computational consumption.

C. Vision Transformers

Motivated by the breakthrough achievement of the transformer on natural language processing (NLP) task [49]–[51], researchers extended the transformer to computer vision tasks. Dosovitskiy *et al*. first built a pure image classification Transformer architecture by splitting the image into multiple 16×16 patches [52]. Liu *et al*. utilized the shifted windows scheme to present a hierarchical vision transformer, showing superior results on several images and video tasks [53], [54]. ViViT involved four video transformer variants to explore the most efficient spatial and temporal encoder for video classification [55]. You *et al*. proposed a transformer-based NR-VQA method, which took the multi-scale quality-aware features as input, and then a transformer encoder was utilized to model the long-range dependencies of the concatenated features [56]. For more comprehensive surveys regarding vision transformer, please refer to [57], [58]. In this paper, we introduce a hierarchical feature aggregation transformer to model the long-range spatiotemporal interactions in UGC-VQA.

III. The Proposed Method

The overall framework of the proposed NR-VQA measurement is shown in Fig. 2, which is comprised of the frame-wise feature (*i.e*., spatial distortion and local motion) extractor and a long-range temporal aggregation module (*i.e*., transformer encoder). We adopt the pre-trained ResNeXt-101 to extract the hierarchical spatial distortion representations, which has been validated to deliver excellent performance on the quality assessment of in-the-wild images [59]. Subsequently, the local motion is built upon the feature difference of two consecutive frames. Then a transformer encoder learns the interaction between spatial distortion and local motion features, enjoying a strong capability of long-range dependencies capturing. A specific token is designed to enhance the scale communication of frame-wise distortion along the temporal dimension, aiming to establish the relationships between shallow and deep stages.

A. Frame-wise Quality Aware Feature Extraction

1) Spatial Distortion Estimation: To obtain the spatial distortion of each frame, a quality-aware feature extractor proposed in [59] is utilized. In particular, the extractor is based on the tailored ResNeXt-101 [11] network and fine-tuned for the quality prediction of images in the wild, demonstrating superior performance over most image quality assessment (IQA) models. The reasons we adopt this extractor lies in that: 1) compared with learning the extractor from the videos with insufficient labeled data, the model pre-trained on KonIQ-10k [60] database can highly reduce the risk of the over-fitting problem; 2) the quality-aware feature is learned in the fashion

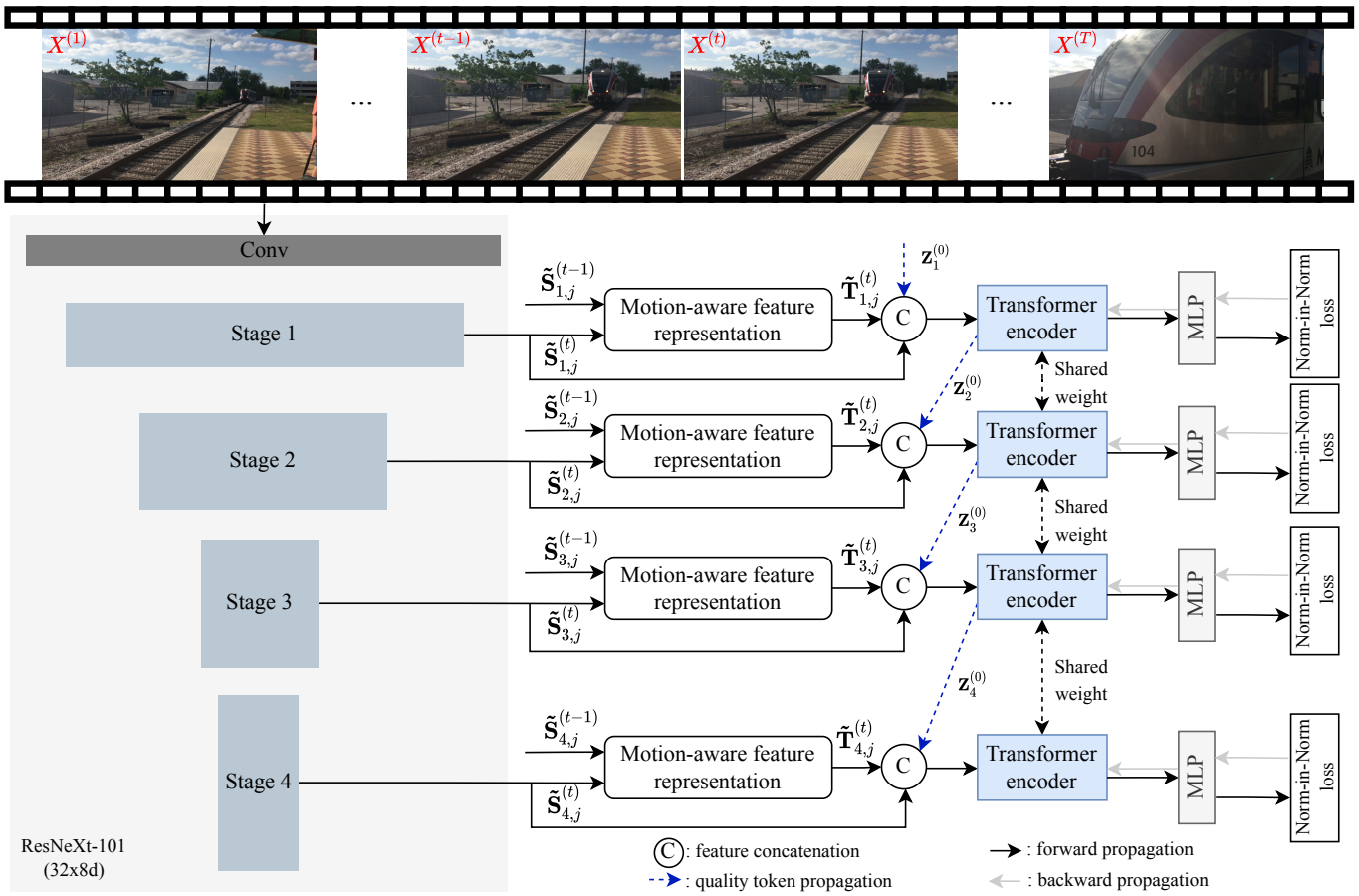

Fig. 2. Illustration of the overall framework of the STI-VQA model. It is consists of three main components: 1) the pre-trained spatial feature extractor for individual frame inputs; 2) the motion-aware feature representation module; 3) the transformer encoder and quality prediction.

of spatial statistics, *i.e.*, mean and standard deviation (std), which is independent of the frame size, further enhancing the generalization capability of our model. Herein, we use $\mathbf{X}^{(t)}$ to represent the t-th video frame, and $1 \leq t \leq T$. Subsequently, we feed $\mathbf{X}^{(t)}$ to the spatial feature extractor f with the pretrained weights w. The j-th feature map $\mathbf{S}_{i,j}^{(t)}$ at i-th stage thus can be acquired as follows,

$$\mathbf{S}_{i,j}^{(t)} = f_{i,j}(\mathbf{X}^{(t)}; w),\ i = 1 \cdots M, j = 1 \cdots N_i, \tag{1}$$

where M is the number of stages, and N_i is the number of feature maps at the i-th stage. In addition, the global mean and std pooling are used to convert the feature maps to feature statistics which can be expressed as follows,

$$\mu_{\mathbf{S}_{i,j}^{(t)}} = \frac{1}{\mathcal{D}} \sum_{k=1}^{\mathcal{D}} \mathbf{S}_{i,j}^{(t)}(k), \tag{2}$$

$$\sigma_{\mathbf{S}_{i,j}^{(t)}} = \sqrt{\frac{1}{\mathcal{D}} \sum_{k=1}^{\mathcal{D}} (\mathbf{S}_{i,j}^{(t)}(k) - \mu_{\mathbf{S}_{i,j}^{(t)}})^2}, \tag{3}$$

where $\mathcal{D}$ represents the total spatial size of j-th spatial feature map at i-th stage. k is the spatial index of each feature map, and $\mu_{\mathbf{S}_{i,j}^{(t)}}$ and $\sigma_{\mathbf{S}_{i,j}^{(t)}}$ indicate the mean and std values of the j-th spatial feature map at i-th stage, respectively. The feature statistics enjoy the advantages of both content-awareness and distortion-sensitivity. Then, we concatenate the mean and std statistic features as follows,

$$\tilde{\mathbf{S}}_{i,j}^{(t)} = \mu_{\mathbf{S}_{i,j}^{(t)}} \copyright \sigma_{\mathbf{S}_{i,j}^{(t)}}, \tag{4}$$

where $\tilde{\mathbf{S}}_{i,j}^{(t)}$ indicates the concatenated spatial features, and © represents the feature concatenation operation along with the feature dimension.

2) Local Motion Extraction: Motivated by the simplicity and effectiveness of frame difference in motion perception, we take one step forward to compute the difference in feature domain. As shown in Fig. 1, the feature difference maps exhibit almost the same visual appearance as the frame difference and optical flow map. Therefore, we calculate the feature difference of the consecutive frames to represent the local motion information as follows,

$$\mathbf{T}_{i,j}^{(t)} = \mathbf{S}_{i,j}^{(t)} - \mathbf{S}_{i,j}^{(t-1)}, \tag{5}$$

where $\mathbf{T}_{i,j}^{(t)}$ indicates the j-th motion-aware feature map at i-th stage. Analogously, the global mean and std pooling are utilized to compute the feature statistics. Meanwhile, we find the mean value of the difference map can be directly derived by the obtained spatial mean statistics as follows,

$$\mu_{\mathbf{T}_{i,j}^{(t)}} = \mu_{\mathbf{S}_{i,j}^{(t)} - \mathbf{S}_{i,j}^{(t-1)}} = \mu_{\mathbf{S}_{i,j}^{(t)}} - \mu_{\mathbf{S}_{i,j}^{(t-1)}}, \tag{6}$$

where $\mu_{\mathbf{T}_{i,j}^{(t)}}$ represents the mean values of the j-th motion-aware feature map at i-th stage. In addition, the std value of the difference map can also be calculated with the spatial std statistics as follows,

$$\begin{aligned}\sigma^2_{\mathbf{T}_{i,j}^{(t)}} &= \sigma^2_{\mathbf{S}_{i,j}^{(t)}-\mathbf{S}_{i,j}^{(t-1)}} \\ &= \sigma^2_{\mathbf{S}_{i,j}^{(t)}} + \sigma^2_{\mathbf{S}_{i,j}^{(t-1)}} + 2\sigma_{\mathbf{S}_{i,j}^{(t)}\mathbf{S}_{i,j}^{(t-1)}} \\ &= \sigma^2_{\mathbf{S}_{i,j}^{(t)}} + \sigma^2_{\mathbf{S}_{i,j}^{(t-1)}} + 2\sigma_{\mathbf{S}_{i,j}^{(t)}}\sigma_{\mathbf{S}_{i,j}^{(t-1)}}\rho_{\mathbf{S}_{i,j}^{(t)}\mathbf{S}_{i,j}^{(t-1)}}, \end{aligned} \tag{7}$$

where $\sigma^2_{\mathbf{T}_{i,j}^{(t)}}, \sigma^2_{\mathbf{S}_{i,j}^{(t)}}$ indicate the variance values of motion-aware and spatial feature maps, $\sigma_{\mathbf{S}_{i,j}^{(t)}\mathbf{S}_{i,j}^{(t-1)}}$ is the covariance value of the spatial features at the t and $t-1$ timestamp, and $\rho_{\mathbf{S}_{i,j}^{(t)}\mathbf{S}_{i,j}^{(t-1)}}$ is the linear correlation coefficient of the spatial feature $\mathbf{S}_{i,j}^{(t)}$ and $\mathbf{S}_{i,j}^{(t-1)}$. Generally speaking, the consecutive frames are highly correlated that we assume $\rho_{\mathbf{S}_{i,j}^{(t)}\mathbf{S}_{i,j}^{(t-1)}} = 1$, and the Eqn. (7) can be derived as,

$$\sigma^2_{\mathbf{T}_{i,j}^{(t)}} = \sigma^2_{\mathbf{S}_{i,j}^{(t)}} + \sigma^2_{\mathbf{S}_{i,j}^{(t-1)}} + 2\sigma_{\mathbf{S}_{i,j}^{(t)}}\sigma_{\mathbf{S}_{i,j}^{(t-1)}}, \tag{8}$$

$$\begin{aligned}\sigma_{\mathbf{T}_{i,j}^{(t)}} &= \sqrt{(\sigma_{\mathbf{S}_{i,j}^{(t)}} + \sigma_{\mathbf{S}_{i,j}^{(t-1)}})^2}, \\ \sigma_{\mathbf{T}_{i,j}^{(t)}} &= \sigma_{\mathbf{S}_{i,j}^{(t)}} + \sigma_{\mathbf{S}_{i,j}^{(t-1)}}. \end{aligned} \tag{9}$$

By concatenating the mean and std statistic values with the feature dimension, the motion-aware features $\tilde{\mathbf{T}}_{i,j}^{(t)}$ is given by,

$$\tilde{\mathbf{T}}_{i,j}^{(t)} = \mu_{\mathbf{T}_{i,j}^{(t)}} \copyright \sigma_{\mathbf{T}_{i,j}^{(t)}}. \tag{10}$$

3) Feature Aggregation: The extracted hierarchical spatial and temporal features are also aggregated by the feature concatenation operation as follows,

$$\mathbf{E}_{i,j}^{(t)} = \tilde{\mathbf{S}}_{i,j}^{(t)} \copyright \tilde{\mathbf{T}}_{i,j}^{(t)}, \tag{11}$$

where we interpret $\mathbf{E}_{i,j}^{(t)}$ as the j-th spatiotemporal features at i-th stage. Before feeding the spatiotemporal features to the transformer, we reduce the dimension of the feature from different stages to a fixed length C via fully connected (FC) layers, which can be expressed as

$$\mathbf{Z}_i^{(t)} = \mathbf{FC}_i(\mathbf{E}_i^{(t)}), \tag{12}$$

where $\mathbf{FC}_i$ indicates the i-th FC layer for the features from i-th stage. Herein, it is worth noting that we omit the subscript j in the spatiotemporal features because the FC layers are designed to reduce the feature dimension from N_i to C.

B. Temporal Quality Aggregation

As shown in Fig. 2, after obtaining the multi-scale spatiotemporal features, we introduce a transformer encoder to model the long-range dependencies. Due to the computational complexity and scale ambiguity, instead of feeding all the spatiotemporal features from different stages to the transformer simultaneously, we design a hierarchical transformer encoder that takes the spatiotemporal features from the same stage as input. The features at different layers contain distinct context information, facilitating the transformer to capture the long-range spatiotemporal interaction from various aspects.

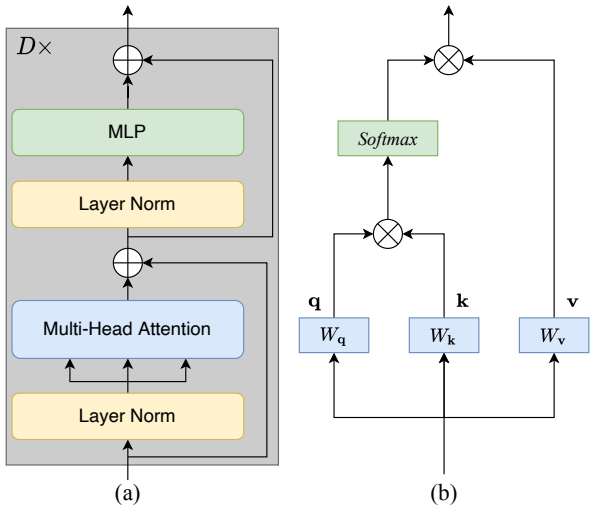

Fig. 3. (a) Illusion of the transformer encoder by courtesy of [52]; (b) Self-attention module for the spatiotemporal embeddings; $\bigoplus$: element-wise summation. $\bigotimes$: matrix multiplication.

The multi-scale processing scheme also matches the HVS mechanism [24], [61], [62], providing the more stable and accurate quality modeling.

In analogous with the transformer used in the classification task [52], we design a quality token $\mathbf{Z}_i^{(0)}$ to build the communication of features in different stages along with the time flow. The quality token is randomly initialized at the first stage, and then progressively delivered from the shallowest to the deepest stage. Besides the scale information, the time order of each frame also plays an important role in video quality regression. To equip the position-aware capability of the transformer encoder, sinusoidal positional encoding is further adopted, treated as the position recorder of each frame. Thus, the combined tokens of the i-th stage ($\mathbf{X}_i^{(0)}$) is formulated as follows,

$$\mathbf{X}_i^{(0)} = \left[\mathbf{Z}_i^{(0)}, \ldots, \mathbf{Z}_i^{(t)}, \ldots, \mathbf{Z}_i^{(T)}\right] + \mathbf{Z}_i^{(pos)}, \tag{13}$$

where $\mathbf{Z}_i^{(t)} \in \mathbb{R}^{1\times C}$ and $\mathbf{Z}_i^{(pos)} \in \mathbb{R}^{(1+T)\times C}$ denote the input tokens and the position embeddings of the i-th stage, respectively. $[\cdot]$ is the token concatenation. The detail of the transformer encoder is shown in Fig. 3, which is composed of the layer normalization (LN), multiheaded self-attention (MSA), and MLP. The transformer encoder can be formulated as follows,

$$\mathbf{Y}_i^{(d)} = \mathbf{X}_i^{(d-1)} + \mathrm{MSA}(\mathrm{LN}(\mathbf{X}_i^{(d-1)})),\ d = 1 \cdots D, \tag{14}$$

$$\mathbf{X}_i^{(d)} = \mathbf{Y}_i^{(d)} + \mathrm{MLP}(\mathrm{LN}(\mathbf{Y}_i^{(d)})),\ d = 1 \cdots D, \tag{15}$$

where d is the index of the transformer and the total number is D. Moreover, $\mathbf{Y}_i^{(d)}$ is the intermediate tokens of d-th tokens at i-th stage. After obtaining the output embeddings from different stages, we use the MLP to regress the quality tokens to compute the final quality scores,

$$\tilde{Q}_i = \mathrm{MLP}_i(\mathbf{X}_i^{(D)}(0)),\ i = 1 \cdots M, \tag{16}$$

where $\tilde{Q}_i$ indicates the quality score of i-th stage, and $\mathbf{X}_i^{(D)}(0)$ is the quality token of the D-th transformer at i-th stage. Each

MPL contains one normalization layer followed with one FC layer. It is worth noting that the output of the deepest stage is recognized as the overall quality score.

C. Objective Loss Function

Herein, we apply the "Norm-in-Norm" loss function to optimize the transformer encoder and MLP, which performs better than mean absolute error (MAE) and mean squared error (MSE) on IQA in terms of convergence speed and prediction accuracy [59]. Specifically, The "Norm-in-Norm" loss first calculates the mean and L^q norm statistics, and then the scores are normalized by the obtained statistics. The last step is to compute the difference between the normalized predictions and the normalized subjective scores. We denote the predicted quality scores to $\left\{\tilde{Q}_i^{(l)}\right\}_{l=1}^{B}$ and the subjective quality scores to $\left\{Q^{(l)}\right\}_{l=1}^{B}$, where l is the index of the batch size B. Therefore, the loss is mathematically formulated as follows,

$$\ell(\tilde{r}_i^{(l)}, r^{(l)}) = \frac{1}{\epsilon}\sum_{l=1}^{B}\left\|\tilde{r}_i^{(l)} - r^{(l)}\right\|^p, \quad (17)$$

$$\tilde{r}_i^{(l)} = \frac{\tilde{Q}_i^{(l)} - \mu_{\tilde{Q}_i^{(l)}}}{\left\|\tilde{Q}_i^{(l)} - \mu_{\tilde{Q}_i^{(l)}}\right\|^q}, \; r^{(l)} = \frac{Q^{(l)} - \mu_{Q^{(l)}}}{\left\|Q^{(l)} - \mu_{Q^{(l)}}\right\|^q}, \quad (18)$$

where ϵ is a normalization hyper-parameter, $\tilde{r}_i^{(l)}$ and $r^{(l)}$ indicate the normalized predicted quality scores from i-th stage and subjective quality scores respectively, $\mu_{\tilde{Q}_i^{(l)}}$ and $\mu_{Q^{(l)}}$ are the mean values of predicted quality scores at i-th stage and subjective quality scores respectively, and $\|\cdot\|$ is the norm operation.

The overall loss is computed by the summation of M individual loss from different stages,

$$\ell = \sum_{i=1}^{M}\ell(\tilde{Q}_i, Q). \quad (19)$$

IV. Experiments

In this section, we first present our experimental setup, including the implementation details of the proposed model, benchmark datasets, and evaluation criteria. We then compare the proposed method with the state-of-the-art quality assessment models in intra-dataset and cross-dataset settings. Finally, we perform extensive ablation studies to isolate the contributions of the proposed approach.

A. Experimental Setup

1) Implementation Details: We feed all the video frames with full resolution to the pre-trained ResNeXt-101 [11] to extract the spatial distortion features. The motion-aware features of the first frame are set to zeros. The stage number M is 4, and the feature maps number $N_i = \{512, 1024, 2048, 4096\}$. The reduced feature dimension C is equal to 128. We set number of transformer encoder to $D = 5$ in Eqn. (14), and each of them has 6 heads with the hidden layer number 64. Training is conducted by minimizing Eqn. (19) with the default Adam stochastic optimization package [63] using batch size of $B = 18$. The initial learning rate is set to 10^{-3} and a decay factor of 0.8 was multiplied for every 2 epochs. The total epoch number is 20. In the "Norm-in-Norm" loss, we inherit the best parameters from [59] and set $p = 1$ and $q = 2$. In addition, ϵ is equal to $2^p B^{1-\frac{p}{q}}$ when p is less than q [59].

2) Benchmarks: We use five UGC-VQA benchmarks to verify the effectiveness of the proposed model, including KoNViD-1k [5], LIVE-VQC [8], YouTube-UGC [9], LSVQ [10], and PUGC [7]. The details regarding the corresponding benchmarks are summarized in Table I. It should be noted that we use the $38,811$ full resolution videos in LSVQ to train the methods with the given split. We provide the testing results on two video subsets, which are partitioned by whether the video resolution is larger than or equal to 1080p [10]. For the YouTube-UGC dataset, we report the results on the available $1,228$ videos. For intra-database experiments, we randomly divide each dataset into 60% for training, 20% for validation, and 20% for testing and perform the experiment ten times to eliminate the bias caused by unpredictability in training-validation-testing set splitting. We report the median and std values of Spearman's rank-order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC) after ten repeated experiments. For the cross-dataset experiments, we utilize one of the UGC-VQA datasets for training, and the other datasets were adopted for testing individually.

3) Evaluation Criteria: Two widely used performance criteria SRCC and PLCC are used to assess prediction monotonicity and accuracy between the predicted and subjective quality scores. Both of them reflect better performance with larger values. As suggested by the Video Quality Experts Group (VQEG) [71], we apply a five-parameter nonlinear logistic function to map the objective scores to the same scale as subjective scores before computing the PLCC value,

$$\hat{Q}_i = \beta_1\left(\frac{1}{2} - \frac{1}{\exp(\beta_2(\tilde{Q}_i - \beta_3))}\right) + \beta_4\tilde{Q}_i + \beta_5, \quad (20)$$

where $\beta_1 \sim \beta_5$ are logistic regression parameters that need to be fitted.

B. Performance Comparisons

1) Performance on Intra-Dataset: Herein, we compare the proposed NR-VQA model with several representative no-reference IQA (NR-IQA) and NR-VQA models. We treat the NR-IQA models as baseline, and the corresponding quality can be obtained by averaging the frame-level quality scores including opinion-unaware models - NIQE [64] and CORNIA [67], and opinion-aware models - BRISQUE [65], FRIQUEE [66] and ResNet-50 [68]. In addition, the competing UGC-VQA models are selected with various design methodologies: VIIDEO [69], V-BLIINDS [46], TLVQM [18], and VIDEVAL [3] are natural video statistical models; VSFA [19] and GSTVQA [24] are RNN-based model; PVQ is designed by 3D-CNN [10]; CNN-TLVQM [70]; RAPIQUE [22] are hybrid methods based on tradition and deep learning; LSCT-PHIQNet is a transformer-based model [56]. We use the implementations from the original authors to retrain and test VSFA [19],

TABLE I
COMPARISON OF THE EXISTING UGC-VQA DATABASES. N/A: NOT AVAILABLE.

Database	# scenes	# videos	Resolutions	Framerates	Duration	Formats	Data	Year
KoNViD-1k [5]	$1,200$	$1,200$	540p	$24-30$	8	MP4	MOS+σ	2017
LIVE-VQC [8]	585	585	240p-1080p	$19-30$	10	MP4	MOS	2018
YouTube-UGC [9]	$1,380$	$1,380$	360p-4K	$15-60$	20	MKV	MOS+σ	2019
LSVQ [10]	$38,811$	$38,811$	$\leq$ 4K	N/A	$5-12$	MP4	MOS+σ	2021
PUGC [7]	$10,000$	$10,000$	$\leq$ 1080p	$15-30$	5	MP4	MOS+σ	2021

TABLE II
PERFORMANCE COMPARISON ON KONVID-1K [5], LIVE-VQC [8], AND YOUTUBE-UGC [9] WITH INTRA-DATASET SETTING. THE MEDIAN SRCC AND PLCC RESULTS AMONG TEN SPLITS ALONG WITH STD VALUE IN THE BRACKET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLDFACE.

Database		KoNViD-1k [5]		LIVE-VQC [8]		YouTube-UGC [9]	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
NR-IQA	NIQE [64]	0.542 (0.034)	0.553 (0.034)	0.596 (0.057)	0.629 (0.051)	0.238 (0.049)	0.278 (0.043)
	BRISQUE [65]	0.657 (0.035)	0.656 (0.034)	0.593 (0.068)	0.638 (0.063)	0.382 (0.052)	0.395 (0.049)
	FRIQUEE [66]	0.747 (0.026)	0.748 (0.026)	0.658 (0.054)	0.700 (0.059)	0.765 (0.030)	0.757 (0.032)
	CORNIA [67]	0.717 (0.025)	0.714 (0.024)	0.672 (0.047)	0.718 (0.042)	0.597 (0.041)	0.605 (0.040)
	ResNet-50 [68]	0.802 (0.026)	0.810 (0.023)	0.663 (0.051)	0.721 (0.043)	0.718 (0.028)	0.710 (0.028)
NR-VQA	VIIDEO [69]	0.299 (0.056)	0.300 (0.054)	0.033 (0.086)	0.215 (0.090)	0.058 (0.054)	0.153 (0.050)
	V-BLIINDS [46]	0.710 (0.031)	0.704 (0.030)	0.694 (0.050)	0.718 (0.050)	0.559 (0.050)	0.555 (0.047)
	TLVQM [18]	0.773 (0.024)	0.769 (0.024)	0.799 (0.037)	0.803 (0.036)	0.669 (0.031)	0.659 (0.030)
	VIDEVAL [3]	0.783 (0.022)	0.780 (0.022)	0.752 (0.039)	0.751 (0.042)	0.778 (0.025)	0.773 (0.026)
	VSFA [19]	0.784 (0.023)	0.794 (0.021)	0.715 (0.035)	0.742 (0.029)	0.754 (0.019)	0.763 (0.018)
	CNN-TLVQM [70]	0.825 (0.015)	0.822 (0.020)	0.819 (0.019)	0.824 (0.018)	0.692 (0.026)	0.691 (0.027)
	RAPIQUE [22]	0.793 (0.020)	0.803 (0.021)	0.743 (0.025)	0.751 (0.028)	0.753 (0.020)	0.766 (0.023)
	GSTVQA [24]	0.810 (0.016)	0.814 (0.019)	0.759 (0.026)	0.799 (0.025)	0.751 (0.027)	0.768 (0.025)
	LSCT-PHIQNet [56]	0.838 (0.013)	0.839 (0.016)	0.789 (0.043)	0.810 (0.033)	0.825 (0.027)	0.833 (0.030)
	STI-VQA	**0.856 (0.012)**	**0.866 (0.027)**	**0.831 (0.030)**	**0.836 (0.030)**	**0.853 (0.025)**	**0.844 (0.030)**

TABLE III
PERFORMANCE COMPARISON ON LSVQ [10], LSVQ-1080P [10], AND PUGC [7] WITH INTRA-DATASET SETTING. THE MEDIAN SRCC AND PLCC RESULTS AMONG TEN SPLITS ARE LISTED. THE BEST RESULTS ARE HIGHLIGHTED IN BOLDFACE.

Database	LSVQ [10]		LSVQ-1080p [10]		PUGC [7]	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE [65]	0.579	0.576	0.497	0.531	0.138	0.137
TLVQM [18]	0.772	0.774	0.589	0.616	0.853	0.865
VIDEVAL [3]	0.794	0.783	0.545	0.554	0.864	0.872
VSFA [19]	0.801	0.796	0.675	0.704	0.875	0.888
PVQ [10]	0.827	0.828	0.711	0.739	-	-
STI-VQA	**0.849**	**0.853**	**0.773**	**0.781**	**0.925**	**0.927**

TABLE IV
CROSS-DATASET COMPARISON WHEN THE MODELS ARE TRAINED ON LSVQ [10], AND TESTED ON KONVID-1K [5] AND LIVE-VQC [8]. THE BEST RESULTS ARE HIGHLIGHTED IN BOLDFACE.

Training	LSVQ [10]			
Testing	KoNViD-1k [5]		LIVE-VQC [8]	
	SRCC	PLCC	SRCC	PLCC
BRISQUE [65]	0.646	0.647	0.524	0.536
TLVQM [18]	0.732	0.724	0.670	0.691
VIDEVAL [3]	0.751	0.741	0.630	0.640
VSFA [19]	0.784	0.794	0.734	0.772
PVQ [10]	0.791	0.795	0.770	0.807
STI-VQA	**0.829**	**0.831**	**0.791**	**0.816**

GSTVQA [24], CNN-TLVQM [70], RAPIQUE [22], and LSCT-PHIQNet [56] under the same experimental settings. The results of the remaining comparison models are obtained from the UGC-VQA benchmark[1] based upon experiments with random 100 train-test (80%-20%) splits, and the median and std values of SRCC and PLCC are reported.

We first present the results on three medium-scale UCG-VQA datasets in Table II, including KoNViD-1k [5], LIVE-VQC [8], and YouTube-UGC [9]. We can observe that the proposed model outperforms the comparison NR-IQA and NR-VQA models on all three datasets in terms of two criteria. In addition, NR-IQA models (*e.g.*, FRIQUEE [66] and ResNet-50 [68]) demonstrate competitive performance on UGC-VQA datasets, which reflects the handcrafted and pre-trained spatial features are promising for capturing the spatial distortion in UGC video. Subsequently, VIIDEO [69] is subpar in the UCG video datasets since it only works for synthetic distortions. In addition, the explicit motion modeling methods (*e.g.*, CNN-TLVQM [70], RAPIQUE [22] and STI-VQA) outperform the remaining methods, which validates the essential of the motion information modeling in UCG-VQA. Moreover, the transformer-based method - LSCT-PHIQNet [56] achieves a competitive performance on three benchmarks, verifying the long-range dependencies within the video play an important role in video quality prediction.

As shown in Table III, we compare the proposed method against five quality measures on two large-scale UGC datasets (*i.e.*, LSVQ [10] and PUGC [7]). The proposed model outperforms the comparison methods by a significant margin, which

[1]https://github.com/vztu/NR-VQA_Benchmark

TABLE V
PERFORMANCE COMPARISON ON KONVID-1K [5], LIVE-VQC [8], AND YOUTUBE-UGC [9] WITH CROSS-DATASET SETTING. THE MEDIAN SRCC AND PLCC RESULTS OF TEN SPLITS ARE LISTED. THE BEST RESULTS ARE HIGHLIGHTED IN BOLDFACE.

Training	KoNViD-1k [5]				LIVE-VQC [8]				YouTube-UGC [9]			
Testing	LIVE-VQC [8]		YouTube-UGC [9]		KoNViD-1k [5]		YouTube-UGC [9]		KoNViD-1k [5]		LIVE-VQC [8]	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE [65]	0.581	0.579	0.271	0.293	0.653	0.623	0.201	0.181	0.541	0.556	0.317	0.333
CORNIA [67]	0.689	0.723	0.326	0.362	0.711	0.702	0.223	0.234	0.576	0.580	0.461	0.466
V-BLIINDS [46]	0.610	0.629	0.150	0.171	0.631	0.615	0.057	0.032	0.300	0.281	0.163	0.166
VSFA [19]	0.687	0.694	0.352	0.361	0.683	0.688	0.322	0.348	0.635	0.642	0.578	0.597
TLVQM [18]	0.706	0.739	0.342	0.359	0.632	0.645	0.217	0.229	0.523	0.531	0.352	0.361
VIDEVAL [3]	0.641	0.657	0.081	0.063	0.654	0.629	0.258	0.239	0.656	0.641	0.386	0.400
RAPIQUE [22]	0.613	0.654	0.301	0.324	0.667	0.659	0.370	0.397	0.624	0.643	0.601	0.615
GSTVQA [24]	0.711	0.721	0.424	0.430	0.692	0.694	0.443	0.468	0.647	0.643	0.604	0.613
STI-VQA	**0.725**	**0.733**	**0.441**	**0.450**	**0.762**	**0.745**	**0.514**	**0.528**	**0.781**	**0.786**	**0.645**	**0.653**

TABLE VI
ABLATION STUDIES ON KONVID-1K [5], LIVE-VQC [8], AND YOUTUBE-UGC [9] USING THE IQA FINE-TUNED RESNEXT-101 [59]. THE BEST RESULTS ARE HIGHLIGHTED IN BOLDFACE.

Spatial statistics	Motion statistics	Single stage	Multiple stages	KoNViD-1k [5]		LIVE-VQC [8]		YouTube-UGC [9]	
				SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
✓		✓		0.812	0.805	0.801	0.802	0.813	0.806
✓			✓	0.821	0.818	0.813	0.813	0.825	0.823
	✓	✓		0.804	0.800	0.795	0.799	0.805	0.800
	✓		✓	0.811	0.808	0.806	0.808	0.817	0.815
✓	✓	✓		0.832	0.829	0.820	0.821	0.831	0.829
✓	✓		✓	**0.856**	**0.866**	**0.831**	**0.836**	**0.853**	**0.844**

TABLE VII
ABLATION STUDIES ON KONVID-1K [5], LIVE-VQC [8], AND YOUTUBE-UGC [9] USING THE IMAGENET PRE-TRAINED RESNEXT-101 [11]. THE BEST RESULTS ARE HIGHLIGHTED IN BOLDFACE.

Spatial statistics	Motion statistics	Single stage	Multiple stages	KoNViD-1k [5]		LIVE-VQC [8]		YouTube-UGC [9]	
				SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
✓		✓		0.769	0.771	0.769	0.772	0.786	0.782
✓			✓	0.778	0.782	0.773	0.780	0.797	0.790
	✓	✓		0.761	0.762	0.758	0.761	0.777	0.774
	✓		✓	0.782	0.786	0.783	0.787	0.803	0.801
✓	✓	✓		0.800	0.802	0.791	0.795	0.814	0.808
✓	✓		✓	**0.824**	**0.830**	**0.810**	**0.813**	**0.825**	**0.819**

TABLE VIII
SRCC AND PLCC COMPARISON RESULTS OF THE ABLATION STUDY ON QUALITY TOKEN REGRESSION.

Dataset	KoNViD-1k [5]		LIVE-VQC [8]		YouTube-UGC [9]	
STI-VQA variants	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Independent quality token at each stage	0.825	0.823	0.803	0.818	0.814	0.807
Average pooling for the tokens at each stage	0.806	0.812	0.789	0.790	0.792	0.784
STI-VQA (Progressively propagate quality token)	**0.856**	**0.866**	**0.831**	**0.836**	**0.853**	**0.844**

demonstrates that the proposed transformer-based model is able to achieve better performance with more training data. Besides, we believe the performance gain arises for the following reasons: 1) the fine-tuning strategy performed on IQA task enables the spatial feature extractor to be distortion-sensitive; 2) the proposed effective motion-aware feature extraction module explicitly models the content movements within consecutive frames; 3) the transformer encoder facilitates learning the long-range spatiotemporal interactions of UGC videos.

2) Performance on Cross-Dataset: In real-world applications, the proposed quality measure often encounters the unexpected contents and distortions, which requires the model to possess strong generalization capability. Thus, researchers adopt cross-dataset experiments to verify the generalization capability. In this subsection, we select eight NR-VQA models and four UGC datasets to conduct the cross-dataset validations. The SRCC and PLCC results are summarized in Table IV and Table V, from which we can draw the following observations. In particular, all the quality models perform worse compared to the intra-dataset results, demonstrating the domain shift among each dataset, unprecedentedly challenging the existing methods. For example, when the methods are trained on

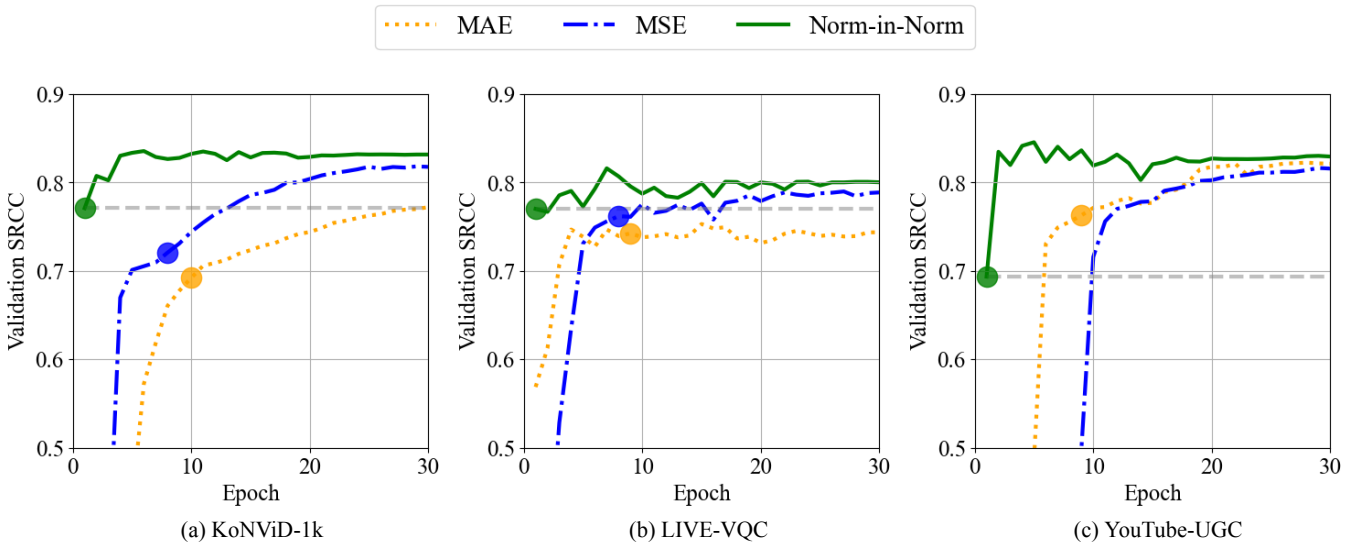

Fig. 4. The SRCC results of STI-VQA were optimized by MAE, MSE, and "Norm-in-Norm" on KoNViD-1k [5], LIVE-VQC [8], and YouTube-UGC [9], respectively.

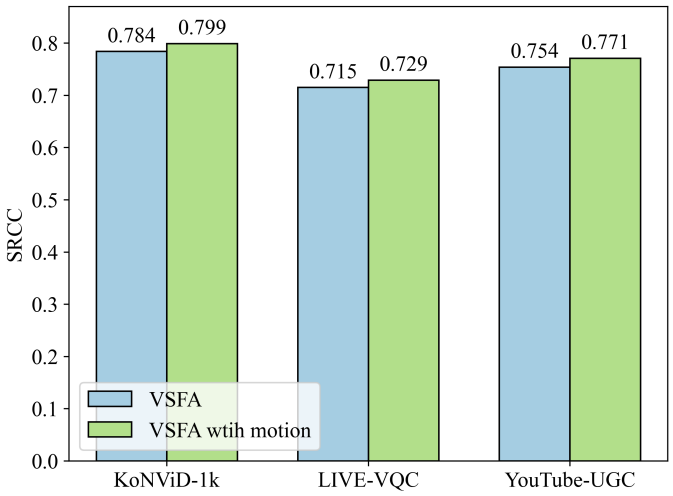

Fig. 5. SRCC results of the VSFA and VSFA with the proposed motion-aware features on KoNViD-1k [5], LIVE-VQC [8], and YouTube-UGC [9].

KoNViD-1k, the best testing result SRCC is 0.441 on the YouTube-UGC, which is far from satisfaction. Therefore, it is highly desirable to improve the generalization capability. In addition, it is worth noting that GSTVQA [24] performs better than other RNN-based NR-VQA models since they impose the feature representation to the Gaussian distribution, which improves the generalization capability. Moreover, the proposed method still achieves the best performance under the cross-dataset setting for both medium-scale datasets [5], [8], [9] and the large-scale dataset [10].

C. Ablation Studies

In order to investigate the contribution of each component in the proposed method, we carry out comprehensive ablation experiments on three UGC datasets [5], [8], [9] using the following configurations.

1) Spatial Feature Extractor: We first compare spatial feature extractors which are pre-trained by ImageNet [72] or fine-tuned by the realistically distorted IQA dataset [60]. The results are shown in Table VI and Table VII, respectively. We can find that the ResNeXt-101 [11] fine-tuned by the realistically distorted images provides better quality prediction in terms of both SRCC and PLCC values. It is reasonable that UGC-VQA suffers from similar spatial distortions with the images in KoNIQ-10k [60], enabling the features to be sensitive to various degradation.

2) Motion-aware Feature Representation: We further verify the effectiveness of the proposed motion-aware feature representation module, which is utilized to explicitly model the motion within a local region. As shown in Table VI and Table VII, the ablation of the motion-aware features results in a significant performance drop with respect to SRCC and PLCC values. Thus, we may draw the conclusion that the proposed motion-aware feature representation module facilitates modeling the temporal distortions, complementing the spatial features in UGC-VQA prediction.

3) Hierarchical Features: In this subsection, we replace the features from the multiple stages with the deepest stage, which has been regarded as content-aware features and achieved compromising performance in UGC-VQA [19], [23]. The comparison results are listed in Table VI and Table VII. We find that although the features from the single stage show competitive results with the state-of-the-art methods (see Table II), the hierarchical features still perform better. This observation is consistent with previous IQA or VQA models [24], [47], [62] that multi-scale processing scheme benefits the quality prediction.

4) Quality Token: In the proposed method, the quality token is designed for cross-scale feature communication. Herein, we compare the progressively propagated quality token with the independent quality tokens and mean pooling strategy. The independent quality tokens represent that we randomly initialize

the quality token at each stage, and mean pooling strategy indicates we simply compute the mean value of all frame tokens except for the quality token. The comparison result are given in Table VIII, from which we observe that the designed progressively propagate quality token outperforms both the independent quality token and average pooling strategy. Thus, the cross-scale feature interaction improves the video quality prediction.

5) Objective Loss: MAE and MSE have been frequently used to optimize networks for IQA and VQA tasks because of their simplicity and reliability. However, experimental results exhibit that MAE and MSE suffer from slow convergence problem [59]. Alternatively, the proposed architecture is optimized by the "Norm-in-Norm" loss. Here we summarize the comparison results of the MAE, MSE, and "Norm-in-Norm" in Fig 4. It is worth mentioning that the results for setting the total epoch to 30 since the MAE and MSE need more epochs to reach convergence. It is obvious that the "Norm-in-Norm" loss not only outperforms the other losses in terms of prediction performance among the three UGC benchmarks [5], [8], [9], but also converges faster.

D. Improving Existing NR-VQA Model

The designed motion-aware feature extraction module is model-agnostic *i.e.*, it can be adopted to complement spatial features for more accurate quality prediction. To verify its effectiveness, we incorporate the motion-aware feature extractor into a general deep learning-based NR-VQA model VSFA [19], which indicates the spatial feature statistics are concatenated with motion-aware feature statistics. Fig. 5 illustrates the results on three UGC datasets [5], [8], [9]. We can observe that the modified VSFA model achieves 1.5%, 1.4%, and 1.7% SRCC gains on KoNViD-1k, LIVE-VQC, and YouTube-UGC, respectively. As a result, the experimental results show that the proposed motion-aware feature representation module consistently enhances the quality of prediction results while imposing no additional computational burdens.

V. Conclusions

In this paper, we have proposed a spatiotemporal feature interaction NR-VQA model by exploring the short-term memory and long-term dependencies. More specifically, we explicitly extract the motion-aware feature representations using the local temporal difference module. Furthermore, we propose a hierarchical feature aggregation transformer to fully exploit the long-range dependencies of the spatial and temporal embeddings. A randomly initialized regression token is progressively fed from the shallowest to the deepest layer, facilitating the cross-scale feature interaction. Experimental results show that the proposed NR-VQA model significantly improves the performance over the state-of-the-art NR-VQA models on five UGC video benchmarks. Moreover, the superior performance achieved by transferring the motion-aware feature extractor to the general VQA model further demonstrates the effectiveness and generalization capability of the proposed method.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments, and Zhaoqian Li for insightful suggestions on model implementation.

References

[1] W. Geyser, "Social media marketing benchmark report 2022," Nov. 2021. [Online]. Available: https://influencermarketinghub.com/social-media-marketing-benchmark-report

[2] Cisco Mobile VNI, "Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021 white paper," Sept. 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

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

[4] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014 – A database for evaluating no-reference video quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3073–3086, Jul. 2016.

[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 *International Conference on Quality of Multimedia Experience*, 2017, pp. 1–6.

[6] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2061–2077, Sept. 2018.

[7] G. Li, B. Chen, L. Zhu, Q. He, H. Fan, and S. Wang, "PUGCQ: A large scale dataset for quality assessment of professional user-generated content," in *ACM International Conference on Multimedia*, 2021, pp. 3728—3736.

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

[9] Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC dataset for video compression research," in *IEEE International Workshop on Multimedia Signal Processing*, 2019, pp. 1–5.

[10] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ: 'Patching Up' the video quality problem," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 019–14 029.

[11] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5987–5995.

[12] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*, 2020, pp. 402–419.

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

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

[15] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, "No-reference pixel video quality monitoring of channel-induced distortion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 605–618, Apr. 2012.

[16] J. Søgaard, S. Forchhammer, and J. Korhonen, "No-reference video quality assessment using codec analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 10, pp. 1637–1650, Oct. 2015.

[17] Y. Liu, J. Wu, L. Li, W. Dong, J. Zhang, and G. Shi, "Spatiotemporal representation learning for blind video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, to appear, 2021.

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

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

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

[21] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-wild videos with mixed datasets training," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1238–1257, Jan. 2021.

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

[23] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *CoRR*, vol. abs/2108.08505, 2021.

[24] 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*, to appear, 2021.

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

[26] J. Fischer and D. Whitney, "Serial dependence in visual perception," *Nature neuroscience*, vol. 17, no. 5, pp. 738–743, Mar. 2014.

[27] 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*, to appear 2021.

[28] 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.

[29] J. Kim, W. Jung, H. Kim, and J. Lee, "CyCNN: A rotation invariant CNN using polar mapping and cylindrical convolution layers," *CoRR*, vol. abs/2007.10588, 2020.

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

[31] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," Jun. 2016. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[32] 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.

[33] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[34] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan, "IVP subjective quality video database," Nov. 2011. [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/

[35] Y. Pitrey, M. Barkowsky, R. Pépion, P. Le Callet, and H. Hlavacs, "Influence of the source content and encoding configuration on the perceived quality for scalable video coding," *Human Vision and Electronic Imaging XVII*, vol. 8291, pp. 460–467, Feb. 2012.

[36] 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.

[37] Z. Li, Z. Duanmu, W. Liu, and Z. Wang, "AVC, HEVC, VP9, AVS2 or AV1?—A comparative study of state-of-the-art video encoders on 4K videos," in *International Conference on Image Analysis and Recognition*, 2019, pp. 162–173.

[38] H. Wang, G. Li, S. Liu, and C.-C. J. Kuo, "Challenge on quality assessment of compressed UGC videos," Jun. 2021. [Online]. Available: http://ugcvqa.com/

[39] 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.

[40] W. Y. Akamine, P. G. Freitas, and M. C. Farias, "A framework for computationally efficient video quality assessment," *Signal Processing: Image Communication*, vol. 70, pp. 57–67, Sept. 2019.

[41] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, Feb. 2016.

[42] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 435–13 444.

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

[44] K. Zhu, C. Li, V. Asari, and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 4, pp. 533–546, Apr. 2014.

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

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

[47] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi, "RIRNet: Recurrent-in-recurrent network for video quality assessment," in *ACM International Conference on Multimedia*, 2020, pp. 834–842.

[48] S. Alamgeer, M. Irshad, and M. C. Farias, "Cnn-based no-reference video quality assessment method using a spatiotemporal saliency patch selection procedure," *Journal of Electronic Imaging*, vol. 30, no. 6, p. 063001, Nov. 2021.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Annual Conference on Neural Information Processing Systems*, 2017, pp. 5998–6008.

[50] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[51] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, 2020, pp. 1877–1901.

[52] 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.

[53] 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 *IEEE International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[54] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," *CoRR*, vol. abs/2106.13230, 2021.

[55] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in *IEEE International Conference on Computer Vision*, 2021, pp. 6836–6846.

[56] J. You, "Long short-term convolutional transformer for no-reference video quality assessment," in *ACM International Conference on Multimedia*, 2021, pp. 2112–2120.

[57] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on visual transformer," *CoRR*, vol. abs/2012.12556, 2020.

[58] S. H. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *CoRR*, vol. abs/2101.01169, 2021.

[59] D. Li, T. Jiang, and M. Jiang, "Norm-in-norm loss with faster convergence and better performance for image quality assessment," in *ACM International Conference on Multimedia*, 2020, pp. 789–797.

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

[61] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *IEEE International Conference on Image Processing*, 1995, pp. 444–447.

[62] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conference on Signals, Systems & Computers*, 2003, pp. 1398–1402.

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[64] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

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

[66] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 32, 1–25, Jan. 2017.
[67] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.
[68] 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.
[69] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, Jan. 2016.
[70] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *ACM International Conference on Multimedia*, 2020, pp. 3311–3319.
[71] 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
[72] 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.

Hanwei Zhu received the B.E 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 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.S. degree from the Wuhan University of Technology in 2018 and the master's 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.

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 Assistant 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 serves as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.