

# A Deep Learning based No-reference Quality Assessment Model for UGC Videos

Wei Sun  
Shanghai Jiao Tong University  
Shanghai, China  
sunguwei@sjtu.edu.cn

Wei Lu  
Shanghai Jiao Tong University  
Shanghai, China  
SJTU-Luwei@sjtu.edu.cn

Xiongkuo Min  
Shanghai Jiao Tong University  
Shanghai, China  
minxiongkuo@sjtu.edu.cn

Guangtao Zhai\*  
Shanghai Jiao Tong University  
Shanghai, China  
zhaiguangtao@sjtu.edu.cn

## ABSTRACT

Quality assessment for User Generated Content (UGC) videos plays an important role in ensuring the viewing experience of end-users. Previous UGC video quality assessment (VQA) studies either use the image recognition model or the image quality assessment (IQA) models to extract frame-level features of UGC videos for quality regression, which are regarded as the sub-optimal solutions because of the domain shifts between these tasks and the UGC VQA task. In this paper, we propose a very simple but effective UGC VQA model, which tries to address this problem by training an end-to-end spatial feature extraction network to directly learn the quality-aware spatial feature representation from raw pixels of the video frames. We also extract the motion features to measure the temporal-related distortions that the spatial features cannot model. The proposed model utilizes very sparse frames to extract spatial features and dense frames (i.e. the video chunk) with a very low spatial resolution to extract motion features, which thereby has low computational complexity. With the better quality-aware features, we only use the simple multilayer perception layer (MLP) network to regress them into the chunk-level quality scores, and then the temporal average pooling strategy is adopted to obtain the video-level quality score. We further introduce a multi-scale quality fusion strategy to solve the problem of VQA across different spatial resolutions, where the multi-scale weights are obtained from the contrast sensitivity function of the human visual system. The experimental results show that the proposed model achieves the best performance on five popular UGC VQA databases, which demonstrates the effectiveness of the proposed model. The code is available at <https://github.com/sunwei925/SimpleVQA>.

## CCS CONCEPTS

• **Computing methodologies** → **Modeling methodologies**.

\*Corresponding author: Guangtao Zhai.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '22, October 10–14, 2022, Lisboa, Portugal

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9203-7/22/10...\$15.00

<https://doi.org/10.1145/3503161.3548329>

## KEYWORDS

video quality assessment, UGC videos, deep learning, feature fusion

### ACM Reference Format:

Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai\*. 2022. A Deep Learning based No-reference Quality Assessment Model for UGC Videos. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3503161.3548329>

## 1 INTRODUCTION

With the proliferation of mobile devices and wireless networks in recent years, User Generated Content (UGC) videos have exploded over the Internet. It has become a popular daily activity for the general public to create, view, and share UGC videos through various social media applications such as YouTube, TikTok, etc. However, UGC videos are captured by a wide variety of consumers, ranging from professional photographers to amateur users, which makes the visual quality of UGC videos vary greatly. In order to ensure the Quality of Experience (QoE) of end-users, the service providers need to monitor the quality of UGC videos in the entire streaming media link, including but not limited to video uploading, compressing, post-processing, transmitting, etc. Therefore, with billions of video viewing and millions of newly uploaded UGC videos every day, an effective and efficient video quality assessment (VQA) model is needed to measure the perceptual quality of UGC videos.

Objective VQA can be divided into full-reference (FR), reduced-reference (RR), and no-reference (NR) according to the amount of pristine video information needed. Since there is no reference video for in-the-wild UGC videos, only NR VQA models are qualified for evaluating their quality. Although NR VQA algorithms [21, 23, 26] have been studied for many years, most of them were developed for Professionally Generated Content (PGC) videos with synthetic distortions, where the pristine PGC videos are shot by photographers using professional devices and are normally of high quality, and the distorted PGC videos are then degraded by specific video processing algorithms such as video compression, transmission, etc. So, previous VQA studies mainly focus on modeling several types of distortions caused by specific algorithms, which makes them less effective for UGC videos with in-the-wild distortions. To be more specific, the emerging UGC videos pose the following challenges to the existing VQA algorithms for PGC videos:

First, the distortion types of UGC videos are diverse. A mass of UGC videos are captured by amateur users, which may suffer various distortion types such as under/over exposure, low visibility, jitter, noise, color shift, etc. These authentic distortions are introduced in the shooting processing and cannot be modeled by the single distortion type, which thereby requires that the VQA models have a more strong feature representation ability to qualify the authentic distortions. Second, the content and forms of UGC videos are extremely rich. UGC videos can be natural scenes, animation [35], games [46, 47], screen content, etc. Note that the statistics characteristics of different video content vary greatly. For example, the natural scenes statistics (NSS) features [22–24, 26] are commonly used in the previous VQA studies to measure the distortions of natural scene content, but they may be ineffective for computer-generated content like animation or games. In addition, live videos, videoconferencing, etc. are also ubiquitous for UGC videos nowadays, whose quality is severely affected by the network bandwidth. Third, due to the advancement of shooting devices, more high resolution [18] and high frame rate [19, 52, 53] videos have emerged on the Internet. The various kinds of resolutions and frame rates are also important factors for video quality. What's more, users can view the UGC videos through mobile devices anywhere and at any time, so the display [25] and the viewing environment such as ambient luminance [29], etc. also affect the perceptual quality of UGC videos to a certain extent. However, these factors are rarely considered by previous studies.

The recently released large-scale UGC VQA databases such as KoNViD-1k [8], YouTube UGC [36], LSVQ [44], etc. have greatly promoted the development of UGC VQA. Several deep learning based NR VQA models [14, 15, 37, 40, 44] have been proposed to solve some challenges mentioned above and achieve pretty good performance. However, there are still some problems that need to be addressed. First, the previous studies either use the image recognition model [15][44] or the pretrained image quality assessment (IQA) models [37][40][14] to extract frame-level features, which lacks an end-to-end learning method to learn the quality-aware spatial feature representation from raw pixels of video frames. Second, previous studies usually extract the features from all video frames and have a very high computational complexity, making them difficult to apply to real-world scenarios. Since there is much redundancy spatial information between the adjacent frames, we argue that there is not necessary to extract the features from all frames. Third, the spatial resolution and frame rate of UGC videos as well as other factors such as the display, viewing environment, etc. are still rarely considered by these studies. However, these factors are very important for the perceptual quality of UGC videos since the contrast sensitivity of the human visual system (HVS) is affected by them.

In this paper, to address the challenges mentioned above, we propose a very simple but effective deep learning based VQA model for UGC videos. The proposed framework is illustrated in Figure 1, which consists of the feature extraction module, the quality regression module, and the quality pooling module. For the feature extraction module, we extract quality-aware features from the spatial domain and the spatial-temporal domain to respectively measure the spatial distortions and motion distortions. Instead of using the

pretrained model to extract the spatial features in the previous studies, we propose to train an end-to-end spatial feature extraction network to learn quality-aware feature representation in the spatial domain, which thereby makes full use of various video content and distortion types in current UGC VQA databases. We then utilize the action recognition network to extract the motion features, which can make up the temporal-related distortions that the spatial features cannot model. Considering that the spatial features are sensitive to the resolution while the motion features are sensitive to the frame rate, we first split the video into continuous chunks and then extract the spatial features and motion features by using a key frame of each chunk and all frames of each chunk but at a low spatial resolution respectively. So, the computational complexity of the proposed model can be greatly reduced.

For the quality regression module, we use the multilayer perception (MLP) network to map the quality-aware features into the chunk-level quality scores, and the temporal average pooling strategy is adopted to obtain the final video quality. In order to solve the problem of quality assessment across different resolutions, we introduce a multi-scale quality fusion strategy to fuse the quality scores of the videos with different resolutions, where the multi-scale weights are obtained from the contrast sensitivity function (CSF) of HVS by considering the viewing environment information. The proposed models are validated on five popular UGC VQA databases and the experimental results show that the proposed model outperforms other state-of-the-art VQA models by a large margin. What's more, the proposed model trained on a large-scale database such as LSVQ [44] achieves remarkable performance when tested on the other databases without any fine-tuning, which further demonstrates the effectiveness and generalizability of the proposed model.

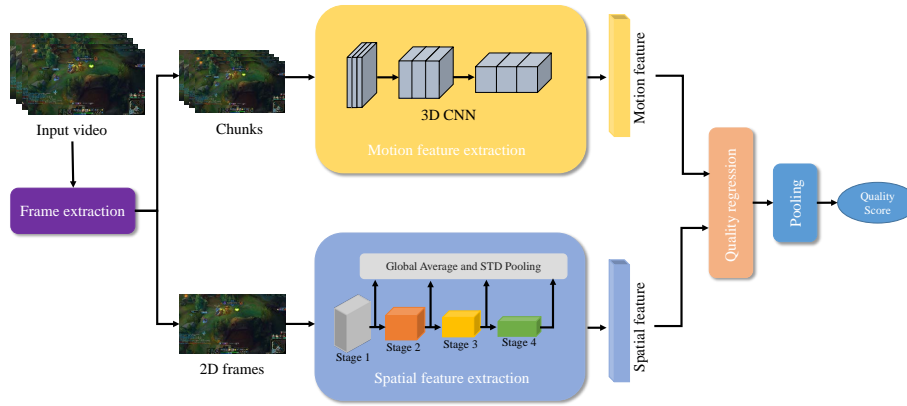
In summary, this paper makes the following contributions:

- (1) We propose an effective and efficient deep learning based model for UGC VQA, which includes the feature extraction module, the quality regression module, and the quality pooling module. The proposed model not only achieves remarkable performance on the five popular UGC VQA databases but also has a low computational complexity, which makes it very suitable for practical applications.
- (2) The feature extraction module extracts two kinds of quality-aware features, the spatial features for spatial distortions and the spatial-temporal features for motion distortions, where the spatial features are learned from raw pixels of video frames via an end-to-end manner and the spatial-temporal features are extracted by a pretrained action recognition network.
- (3) We introduce a multi-scale quality fusion strategy to solve the problem of quality assessment across different resolutions, where the multi-scale weights are obtained from the contrast sensitivity function of the human visual system by considering the viewing environment information.

## 2 RELATED WORK

### 2.1 Handcrafted feature based NR VQA Models

A naive NR VQA method is to compute the quality of each frame via popular NR IQA methods such as NIQE [24], BRISQUE [22],



**Figure 1: The network architecture of the proposed model. The proposed model contains the feature extraction module, the quality regression module, and the quality pooling module. The feature extraction module extracts two kinds of features, the spatial features and the motion features.**

CORNIA [42] etc., and then pool them into the video quality score. A comparative study of various temporal pooling strategies on popular NR IQA methods can refer to [32]. The temporal information is very important for VQA. V-BLIINDS [26] is a spatio-temporal natural scene statistics (NSS) model for videos by quantifying the NSS feature of frame-differences and motion coherency characteristics. Mittal *et al.* [23] propose a training-free blind VQA model named VIIDEO that exploits intrinsic statistics regularities of natural videos to quantify disturbances introduced due to distortions. TLVQM [12] extracts abundant spatio-temporal features such as motion, jerkiness, blurriness, noise, blockiness, color, etc. at two levels of high and low complexity. VIDEVAL [33] further combines the selected features from typical NR I/VQA methods to train a SVR model to regress them into the video quality. Since video content also affects its quality, especially for UGC videos, understanding the video content is beneficial to NR VQA. Previous handcrafted feature based methods are difficult to understand semantic information. Hence, some studies [13, 34] try to combine the handcrafted features with the semantic-level features extracted by the pretrained CNN model to improve the performance of NR VQA models. For example, CNN-TLVQM [13] combines the handcrafted statistical temporal features from TLVQM and spatial features extracted by 2D-CNN model trained for IQA. RAPIQUE [34] utilizes the quality-aware scene statistics features and semantics-aware deep CNN features to achieve a rapid and accurate VQA model for UGC videos.

## 2.2 Deep learning based NR VQA Models

With the release of several large-scale VQA databases [8, 36, 44], deep learning based NR VQA models [2, 11, 14, 15, 31, 37, 40, 43, 44] attract many researchers' attention. Liu *et al.* [17] propose a multi-task BVQA model V-MEON by jointly optimizing the 3D-CNN for quality assessment and compression distortion classification. VSFA [15] first extracts the semantic features from a pre-trained CNN model and then uses a gated recurrent unit (GRU) network to model the temporal relationship between the semantic features of video frames. The authors of VSFA further propose MDVSFA

[16], which trains the VSFA model on the multiple VQA databases to improve its performance and generalization. RIRNet [4] exploits the effect of motion information extracted from the multi-scale temporal frequencies for video quality assessment. Ying *et al.* [44] propose a local-to-global region-based NR VQA model that combines the spatial features extracted from a 2D-CNN model and the spatial-temporal features from a 3D-CNN network. Wang *et al.* [37] propose a feature-rich VQA model for UGC videos, which measures the quality from three aspects, compression level, video content, and distortion type and each aspect is evaluated by an individual neural network. Xu *et al.* [40] first extract the spatial feature of the video frame from a pre-trained IQA model and use the graph convolution to extract and enhance these features, then extract motion information from the optical flow domain, and finally integrated the spatial feature and motion information via a bidirectional long short-term memory network. Li *et al.* [14] also utilize the IQA model pre-trained on multiple databases to extract quality-aware spatial features and the action recognition model to extract temporal features, and then a GRU network is used to model spatial and temporal features and regress them into the quality score. Wen and Wang [39] propose a baseline I/VQA model for UGC videos, which calculates the video quality by averaging the scores of each frame and frame-level quality scores are obtained by a simple CNN network.

## 3 PROPOSED MODEL

The framework of the proposed NR VQA model is shown in Figure 1, which consists of the feature extraction module, the quality regression module, and the quality pooling module. First, we extract the quality-aware features from the spatial domain and the spatial-temporal domain via the feature extraction module, which are utilized to evaluate the spatial distortions and motion distortions respectively. Then, the quality regression module is used to map the quality-aware features into chunk-level quality scores. Finally, we perform the quality pooling module to obtain the video quality score.

### 3.1 Feature Extraction Module

In this section, we expect to extract the quality-aware features that can represent the impact of various distortion types and content on visual quality. The types of video distortion can be roughly divided into two categories: the spatial distortions and the motion distortions. The spatial distortions refer to the artifacts introduced in the video frames, such as noise, blur, compression, low visibility, etc. The motion distortions refer to the jitter, lagging due, etc., which are mainly caused by unstable shooting equipment, fast-moving objects, the low network bandwidth, etc. Therefore, we need to extract the quality-aware features from these two aspects.

Note that the characteristics of the spatial features and motion features are quite different. The spatial features are sensitive to the video resolution but insensitive to the video frame rate since the adjacent frames of the video contain lots of redundancy spatial information and higher resolution can represent more abundant high-frequency information, while motion features are the opposite because the motion distortions are reflected on the temporal dimension and these features are usually consistent for local regions of the frames.

Therefore, considering these characteristics, given a video  $V$ , whose number of frames and frame rate are  $l$  and  $r$  respectively, we first split the video  $V$  into  $N_c$  continuous chunks  $c = \{c_i\}_{i=1}^{N_c}$  at an time interval  $\tau$ , where  $N_c = l/(r * \tau)$ , and there are  $N_f = r * \tau$  frames in each chunk  $c_i$ , which is denoted as  $c_i = \{x_{i,j}\}_{j=1}^{N_f}$ . Then we only choose a key frame  $x_{i, key}$  in each chunk to extract the spatial features and the motion features of each chunk are extracted using all frames in  $c_i$  but at a very low spatial resolution. As a result, we can greatly reduce the computation complexity of the VQA model with little performance degradation.

**3.1.1 Spatial Feature Extraction Module.** Given a frame  $x$ , we denote  $f_w(x)$  as the output of the CNN model  $f$  with trainable parameters  $w = \{w_k\}$  applied on the frame  $x$ . Assume that there are  $N_s$  stages in the CNN model, and  $f_w^k(x)$  is the output feature maps extracted from the  $k$ -th stage, where  $f_w^k(x) \in \mathbb{R}^{H_k \times W_k \times C_k}$ , and  $H_k$ ,  $W_k$ , and  $C_k$  are the height, width, and the number of channels of the feature maps  $f_w^k(x)$  respectively. In the following, we use the  $f_w^k$  to replace the  $f_w^k(x)$  for simplicity.

It is well known that the features extracted by the deep layers of the CNN model contain rich semantic information, and are suitable for representing content-aware features for UGC VQA. Moreover, previous studies indicate that the features extracted by the shallow layers of the CNN models contain low-level information [28, 48], which responds to low-level features such as edges, corners, textures, etc. The low-level information is easily affected by the distortion and is therefore distortion-aware. Hence, we extract the quality-aware features via calculating the global mean and stand deviation of feature maps extracted from all stages of the CNN model. Then, we apply global average and stand deviation pooling operations on the feature maps  $f_w^k$ :

$$\begin{aligned} \mu_{f_w^k} &= \text{GP}_{\text{avg}}(f_w^k), \\ \sigma_{f_w^k} &= \text{GP}_{\text{std}}(f_w^k), \end{aligned} \quad (1)$$

where  $\mu_{f_w^k}$  and  $\sigma_{f_w^k}$  are the global means and stand deviation of feature maps  $f_w^k$  respectively. Finally, we concatenate the  $\mu_{f_w^k}$  and  $\sigma_{f_w^k}$  to derive the spital feature representation of our NR VQA model:

$$\begin{aligned} F_s^k &= \text{cat}([\mu_{f_w^k}, \sigma_{f_w^k}]), \\ F_s &= \text{cat}(\{F_s^k\}_{k=1}^{N_s}). \end{aligned} \quad (2)$$

**3.1.2 Motion Feature Extraction Module.** We extract the motion features as the complementary quality-aware features since UGC videos are commonly degraded by the motion distortions caused by the unstable shooting equipment or low bit rates in the living streaming or videoconferencing. The spatial features are difficult to handle these distortions because they are extracted by the intra-frames while motion distortions occur in the interframes. Therefore, the motion features are also necessary for evaluating the quality of UGC videos. Here, we utilize the pretrained action recognition model as the motion feature extractor to obtain the motion features of each video chunk. The action recognition model is designed to detect different kinds of action classes, so the feature representation of the action recognition network can reflect the motion information of the video to a certain extent. Therefore, given the video chunk  $c$  and the action recognition network MOTION, we can obtain the motion features:

$$F_m = \text{MOTION}(c) \quad (3)$$

where  $F_m$  represents the motion features extract by the action recognition network.

Therefore, given the video chunk  $c$ , we first select a key frame in the chunk to calculate the spatial features  $F_s$ . Then, we calculate the motion features  $F_m$  using the whole frames but at a low spatial resolution in the video chunk. Finally, we obtain the quality-aware features for the video chunk  $c$  by concatenating the spatial features and motion features:

$$F = \text{cat}([F_s, F_m]), \quad (4)$$

### 3.2 Quality Regression Module

After extracting quality-aware feature representation by the feature extraction module, we need to map these features to the quality scores via a regression model. In this paper, we use the multi-layer perception (MLP) as the regression model to obtain the chunk-level quality due to its simplicity and effectiveness. The MLP consists of two fully connected layers and there are 128 and 1 neuron in each layer respectively. Therefore, we can obtain the chunk-level quality score via

$$q = f_{\text{wFC}}(F), \quad (5)$$

where  $f_{\text{wFC}}$  denotes the function of the two FC layers and  $q$  is the quality of the video chunk.

### 3.3 Quality Pooling Module

As stated in Section 3.1, we split the video  $V$  into  $N_c$  continuous chunks  $\{c_i\}_{i=1}^{N_c}$ . For the chunk  $c_i$ , we can obtain its chunk-level quality score  $q_i$  via the feature extraction module and the quality regression module. Then, it is necessary to pool the chunk-level scores into the video level. Though many temporal pooling methods have been proposed in literature [32][15], we find that the temporal

**Table 1: Summary of the benchmark UGC VQA databases. Time duration: Seconds.**

Database	Videos	Scenes	Resolution	Time Duration	Format	Distortion Type	DATA	Environment
KoNViD-1k [8]	1,200	1,200	540p	8	MP4	Authentic	MOS + $\sigma$	Crowd
YouTube-UGC [36]	1500	1500	360p-4K	20	YUV, MP4	Authentic	MOS + $\sigma$	Crowd
LSVQ [44]	38,811	38,811	99p-4K	5-12	MP4	Authentic	MOS + $\sigma$	Crowd
LBVD [3]	1,013	1,013	240p-540p	10	MP4	Authentic, Transmission	MOS + $\sigma$	In-lab
LIVE-YT-Gaming [45]	600	600	360p-1080p	8-9	MP4	Authentic	MOS	Crowd

averaging pooling achieves the best performance from Section 4.3.2. Therefore, the video-level quality is calculated as:

$$Q = \frac{1}{N_c} \sum_{i=1}^{N_c} q_i, \quad (6)$$

where  $q_i$  is the quality of the  $i$ -th chunk and  $Q$  is the video quality evaluated by the proposed model.

### 3.4 Loss Function

The loss function used to optimize the proposed models consists of two parts: the mean absolute error (MAE) loss and rank loss [39]. The MAE loss is used to make the evaluated quality scores close to the ground truth, which is defined as:

$$L_{MAE} = \frac{1}{N} \sum_{i=1}^N |Q_i - \hat{Q}_i|, \quad (7)$$

where the  $\hat{Q}_i$  is the ground truth quality score of the  $i$ -th video in a mini-batch and  $N$  is the number of videos in the mini-batch.

The rank loss is further introduced to make the model distinguish the relative quality of videos better, which is very useful for the model to evaluate the videos with similar quality. Since the rank value between two video quality is non-differentiable, we use the following formula to approximate the rank value:

$$L_{rank}^{ij} = \max(0, |\hat{Q}_i - \hat{Q}_j| - e(\hat{Q}_i, \hat{Q}_j) \cdot (Q_i - Q_j)), \quad (8)$$

where  $i$  and  $j$  are two video indexes in a mini-batch, and  $e(\hat{Q}_i, \hat{Q}_j)$  is formulated as:

$$e(\hat{Q}_i, \hat{Q}_j) = \begin{cases} 1, & \hat{Q}_i \geq \hat{Q}_j, \\ -1, & \hat{Q}_i < \hat{Q}_j, \end{cases} \quad (9)$$

Then,  $L_{rank}$  is calculated via:

$$L_{rank} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N L_{rank}^{ij} \quad (10)$$

Finally, the loss function can be obtained by:

$$L = L_{MAE} + \lambda \cdot L_{rank}, \quad (11)$$

where  $\lambda$  is a hyper-parameter to balance the MAE loss and the rank loss.

### 3.5 Multi-scale Quality Fusion Strategy

Previous studies evaluate the video quality either using the original spatial resolution or a fixed resized spatial resolution, which ignore that videos are naturally multi-scale [53]. Some existing work [38][25][20] shows that considering the multi-scale characteristics can improve the performance of image quality assessment. So, we

propose a multi-scale quality fusion strategy to further improve the evaluation accuracy of the VQA model and this strategy is very useful to compare the quality of videos with different spatial resolutions.

**3.5.1 Multi-scale Video Quality Scores.** We first resize the resolution of the video into three fixed spatial scales, which are 540p, 720p, and 1080p, respectively. We do not downscale the video from the original scale to several lower resolution scales, which is a more common practice in previous studies. That is because when users watch videos in an application, the resolution of videos is actually adapted to the resolution of the playback device, and the modern display resolution is normally larger than 1080p. So, the perceptual quality of the low-resolution videos is also affected by the up-sampling artifacts, which also need to be considered by VQA models. Therefore, given a VQA model, we can derive three quality of videos at three scales, which are denoted as  $Q_1$ ,  $Q_2$ , and  $Q_3$  respectively.

**3.5.2 Adaptive Multi-scale Weights.** The weight of each scale is obtained by considering the human psychological behaviors and the visual sensitivity characteristics. It is noted that the contrast perception ability of the HVS depends on the spatial frequency of the visual signal, which is modeled by the contrast sensitivity function (CSF). Specifically, we first define a viewing resolution factor  $\xi$  as:

$$\xi = \frac{\pi \cdot d \cdot n}{180 \cdot h_s \cdot 2}, \quad (12)$$

where the unit of  $\xi$  is cycles per degree of visual angle (cpd),  $d$  is the viewing distance (inch),  $h_s$  is the height of the screen (inch), and  $n$  denotes the number of pixels in the vertical direction of the screen. For the above three spatial scales of video, we can obtain the corresponding  $\xi$ , which are denoted as  $\xi_1$ ,  $\xi_2$ , and  $\xi_3$  respectively. We use  $\xi$  to divide the spatial frequency range of the corresponding scale, which covers one section of the CSF formulated by:

$$S(u) = \frac{5200e^{(-0.0016u^2(1+100/L)^{0.08})}}{\sqrt{(1 + \frac{144}{X_0^2} + 0.64u^2)(\frac{63}{L^{0.83}} + \frac{1}{1-e^{(-0.02u^2)}})}} \quad (13)$$

where  $u$ ,  $L$ , and  $X_0^2$  indicate spatial frequency (cpd), luminance (cd/m<sup>2</sup>), and angular object area (squared degrees), respectively.

The weight of each scale is calculated as the area under the CSF within the corresponding frequency covering range:

$$w_i = \frac{1}{Z} \int_{\xi_{i-1}}^{\xi_i} S(u) du, \quad i \in \{1, 2, 3\}, \quad (14)$$

where  $i$  from 1 to 3 corresponds the finest to coarsest scale respectively, and  $\xi_0$  corresponds the viewing resolution factor of 0.  $Z$  is a normalization factor such that  $\sum_i w_i = 1$ .

**Table 2: Performance of the SOTA models and the proposed model on the KoNViD-1k, YouTube-UGC, LBVD, and LIVE-YT-Gaming databases. W.A. means the weight average results. The best performing model is highlighted in each column.**

Type	Database Criterion	KoNViD-1k		YouTube-UGC		LBVD		LIVE-YT-Gaming		W.A.	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
IQA	NIQE	0.542	0.553	0.238	0.278	0.327	0.387	0.280	0.304	0.359	0.393
	BRISQUE	0.657	0.658	0.382	0.395	0.435	0.446	0.604	0.638	0.513	0.525
	GM-LOG	0.658	0.664	0.368	0.392	0.314	0.304	0.312	0.317	0.433	0.440
	VGG19	0.774	0.785	0.703	0.700	0.676	0.673	0.678	0.658	0.714	0.712
	ResNet50	0.802	0.810	0.718	0.710	0.715	0.717	0.729	0.768	0.744	0.751
	KonCept512	0.735	0.749	0.587	0.594	0.626	0.636	0.643	0.649	0.650	0.660
VQA	V-BLIINDS	0.710	0.704	0.559	0.555	0.527	0.558	0.357	0.403	0.566	0.578
	TLVQM	0.773	0.769	0.669	0.659	0.614	0.590	0.748	0.756	0.699	0.689
	VIDEVAL	0.783	0.780	0.779	0.773	0.707	0.697	0.807	0.812	0.766	0.762
	RAPIQUE	0.803	0.818	0.759	0.768	0.712	0.725	0.803	0.825	0.767	0.781
	VSFA	0.773	0.775	0.724	0.743	0.622	0.642	0.776	0.801	0.721	0.736
	Li <i>et al.</i>	0.836	0.834	0.831	0.819	-	-	-	-	-	-
	Pro.	<b>0.856</b>	<b>0.860</b>	<b>0.847</b>	<b>0.856</b>	<b>0.844</b>	<b>0.846</b>	<b>0.861</b>	<b>0.866</b>	<b>0.851</b>	<b>0.856</b>

**Table 3: Performance of the SOTA models and the proposed models on the LSVQ database. Pro. M.S. refers to the proposed model implemented by the multi-scale quality fusion strategy. W.A. means the weighted average results. The best performing model is highlighted in each column.**

Database Criterion	Test		Test-1080p		W.A.	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
TLVQM	0.772	0.774	0.589	0.616	0.712	0.722
VIDEVAL	0.794	0.783	0.545	0.554	0.712	0.707
VSFA	0.801	0.796	0.675	0.704	0.759	0.766
PVQ	0.827	0.828	0.711	0.739	0.789	0.799
Li <i>et al.</i>	0.852	0.854	<b>0.772</b>	0.788	0.825	0.832
Pro.	0.864	0.861	0.756	0.801	0.829	0.841
Pro. M.S.	<b>0.867</b>	<b>0.861</b>	0.764	<b>0.803</b>	<b>0.833</b>	<b>0.842</b>

Therefore, the multi-scale fusion quality score  $Q_m$  is calculated as:

$$Q_m = \prod_{i=1}^3 Q_i^{w_i}, \quad (15)$$

## 4 EXPERIMENTAL VALIDATION

### 4.1 Experimental Protocol

**4.1.1 Test Databases.** We test the proposed model on the five UGC VQA database: KoNViD-1k [8], YouTube-UGC [36], LSVQ [44], LBVD [3], and LIVE-YT-Gaming [45]. We summarize the main information of the databases in Table 1. The LSVQ database is the largest UGC VQA database so far, and there are 15 video categories such as animation, gaming, HDR, live music, sports, etc. in the YouTube-UGC database, which is more diverse than other databases. The LBVD database focuses on the live broadcasting videos, of which the videos are degraded by the authentic transmission distortions. The LIVE-YT-Gaming database consists of streamed gaming videos, where the video content is generated by computer graphics.

**4.1.2 Implementation Details.** We use the ResNet50 [7] as the backbone of the spatial feature extraction module and the SlowFast R50 [6] as the motion feature extraction model for the whole experiments. The weights of the ResNet50 are initialized by training on the ImageNet dataset [5], the weights of the SlowFast R50 are fixed by training on the Kinetics 400 dataset [10], and other weights are randomly initialized. For the spatial feature extraction module, we resize the resolution of the minimum dimension of key frames as 520 while maintaining their aspect ratios. In the training stage, the input frames are randomly cropped with the resolution of 448×448. If we do not use the multi-scale quality fusion strategy, we crop the center patch with the same resolutions of 448×448 in the testing stage. Note that we only validate the multi-scale quality fusion strategy on the model trained by the LSVQ database since there are enough videos with various spatial resolutions in it. For the motion feature extraction module, the resolution of the videos is resized to 224×224 for both the training and testing stages. We use PyTorch to implement the proposed models. The Adam optimizer with the initial learning rate 0.00001 and batch size 8 are used for training the proposed model on a server with NVIDIA V100. The hyper-parameter  $\lambda$  is set as 1. For simplicity, we select the first frame in each chunk as the key frame. For the multi-scale quality fusion strategy, there are  $d = 35$ ,  $n = 1080$ ,  $h = 11.3$ ,  $L = 200$ , and  $X_0^2 = 606$ , and the final multi-scale weights for UGC videos are  $w_1 = 0.8317$ ,  $w_2 = 0.0939$ , and  $w_3 = 0.0745$ .

**4.1.3 Comparing Algorithms.** We compare the proposed method with the following no-reference models:

- IQA models: NIQE [24], BRISQUE [22], GM-LOG [41], VGG19 [27], ResNet50 [7], and KonCept512 [9].
- VQA models: V-BLIINDS [26], TLVQM [12], VIDEAL [33], RAPIQUE [34], VSFA [15], PVQ [44], and Li *et al.* [14].

Since the number of videos in the LSVQ database is relatively large, we only compare some representative VQA models on the LSVQ database and omit the methods which perform poorly on the other four UGC databases.

**4.1.4 Evaluation Criteria.** We adopt two criteria to evaluate the performance of VQA models, which are Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC). PLCC reflects the prediction linearity of the VQA algorithm and SRCC indicates the prediction monotonicity. An excellent VQA model should obtain the value of SRCC and PLCC close to 1. Before calculating the PLCC, we follow the same procedure in [1] to map the objective score to the subject score using a four-parameter logistic function.

For KoNViD-1k, YouTube-UGC, LBVD, and LIVE-YT-Gaming databases, we randomly split these databases into the training set with 80% videos and the test set with 20% videos for 10 times, and report the median values of SRCC and PLCC. For the LSVQ database, we follow the same training and test split suggested by [44] and report the performance on the test and test-1080p subsets.

## 4.2 Performance Comparison with the SOTA Models

The performance results of the VQA models on the KoNViD-1k, YouTube-UGC, LBVD, and LIVE-YT-Gaming databases are listed in Table 2, and on the LSVQ database are listed in Table 3. From Table 2 and Table 3, we observe that the proposed model achieves the best performance on all five UGC VQA databases and leads by a large margin, which demonstrates that the proposed model does have a strong ability to measure the perceptual quality of various kinds of UGC videos. For the test-1080p subset of the LSVQ database, the proposed model is inferior to Li *et al.*, which may be because the spatial resolution of most videos in the test-1080p subset is larger than 1080p while the proposed model resizes the spatial resolution of test videos into 448×448, so the proposed model has a relatively poor ability to represent the characteristics of high-resolution videos. Through the multi-scale quality weighting fusion strategy, the proposed model can significantly improve the performance on the test-1080p subset.

Then, most handcrafted feature based IQA models perform poorly on these UGC VQA databases especially for the LBVD and LIVE-YT-Gaming databases since they are designed for natural scene images with synthetic distortions and are difficult to handle the complex in-the-wild distortions and other video types such as gaming, videoliving, etc. It is worth noting that through fine-tuning the deep CNN baseline i.e. ResNet50 on the VQA databases, it can achieve a pretty good performance, which also indicates that spatial features are very important for VQA tasks. For the NR VQA methods, the hand-crafted feature based NR VQA methods such as TLVQM and VIDEVAL achieve pretty well performance by incorporating the rich spatial and temporal quality features, such as NSS features, motion features, etc., but they are inferior to the deep learning based NR VQA methods due to the strong feature representation ability of CNN. VSFA extracts the spatial features from the pretrained image recognition model, which are not quality-aware, and achieves relatively poor performance when compared with other deep learning based methods. PVQ and Li *et al.* methods both utilize the pretrained IQA model and pretrained action recognition model to extract spatial and motion features respectively, and they perform better than other compared NR I/VQA methods but are inferior to the proposed model. Through training an end-to-end

**Table 4: The results of ablation studies on the LSVQ database. S and M means the spatial features and motion features respectively, and S\* means that the spatial features are extracted by the pretrained image classification network.**

	Database Criterion	Test		Test-1080p	
		SRCC	PLCC	SRCC	PLCC
Feature	S*+M	0.847	0.841	0.732	0.774
	S	0.827	0.829	0.702	0.757
	M	0.660	0.669	0.569	0.621
Regression	GRU	0.858	0.855	0.735	0.788
	Transformer	0.860	0.861	0.753	0.799
Pooling	Method in [15]	0.860	0.858	0.733	0.786
	1D CNN based	0.864	0.862	0.739	0.790

**Table 5: The SRCC results of cross-database evaluation. The model is trained on the LSVQ database.**

Database	KoNViD-1k	YouTube-UGC	LBVD	LIVE-YT-Gaming
Pro.	<b>0.860</b>	0.789	0.689	0.642
Pro. M.S.	0.859	<b>0.822</b>	<b>0.711</b>	<b>0.683</b>

spatial feature extractor, the proposed model can take advantage of various video content and distortion types in the UGC databases and learn a better quality-aware feature representation. As a result, the proposed model achieves the best performance on all five UGC VQA databases.

## 4.3 Ablation Studies

In this section, we conduct several ablation studies to investigate the effectiveness of each module in the proposed model, including the feature extraction module, and the quality regression module. All the experiments are tested on the LSVQ database since it is the largest UGC VQA model and is more representative.

**4.3.1 Feature Extraction Module.** The proposed model consists of the spatial feature extractor that learns the end-to-end spatial quality-aware features and the motion feature extractor that utilizes a pretrained action recognition model to represent motion information. Therefore, we first do not train the spatial feature extractor and directly use the weights trained on the ImageNet database to study the effect of the end-to-end training strategy for the spatial feature extractor. Then, we only use the end-to-end trained spatial features or the pretrained motion features to evaluate the quality of UGC videos to investigate the effect of these two kinds of features. The results are listed in Table 4. First, it is observed that the model using the motion features is inferior to the model using the spatial features and both of them are inferior to the proposed model, which indicates that both spatial and motion features are beneficial to the UGC VQA task and the spatial features are more important. Then, we find that end-to-end training for the spatial feature extractor can significantly improve the evaluation performance, which demonstrates that end-to-end trained spatial features represent better than that extracted by the pretrained image classification model.

**Table 6: Comparison of computational complexity for the six VQA models and two proposed models. Time: Second.**

Methods	V-BLIINDS	TLVQM	VIDEVAL	VSFA	RAPIQUE	Li <i>et al.</i>	Pro.	Pro. M.S.
Time	61.982	219.992	561.408	56.424	38.126	61.971	<b>6.929</b>	8.448

**4.3.2 Quality Regression Module.** In this paper, we use the MLP as the regression model to derive the chunk-level quality scores. However, in previous studies, some sequential models such as GRU [15], Transformer [14], etc. are also adopted to further consider the influence of the features extracted from adjacent frames. Here, we also adopt these methods as a comparison to investigate whether sequential models can improve the performance of the proposed models. Specifically, we replace the MLP module with the GRU and Transformer and keep other experimental setups the same. The results are listed in Table 4. We observe that models using GRU and Transformer are both inferior to the proposed model, which means that the MLP module is enough to regress the quality-aware features to quality scores though it is very simple. This conclusion is also consistent with [37]. The reason is that the proposed model and the model in [37] calculate the chunk-level quality score and the effect of adjacent frames are considered in the quality-aware features (i.e. motion features), while other VQA models [15] [14] calculate the frame-level quality scores, which may need to consider the effect of adjacent frames in the quality regression module.

**4.3.3 Quality Pooling Module.** The proposed model uses the temporal average pooling method to fuse the chunk-level quality scores into the video level. It is noted that previous studies also propose several temporal pooling methods for VQA. In this section, we test two temporal pooling methods, which are the subjectively-inspired method introduced in [15] and a learning based temporal pooling method using the 1D CNN. The results are listed in Table 4. From Table 4, we observe that the average pooling strategy achieves similar performance to the learning based pooling method, and both of them are superior to the subjectively-inspired methods. Since the average pooling strategy is simpler and does not increase the extra parameters, we use the temporal average pooling method in this paper.

#### 4.4 Cross-Database Evaluation

UGC videos may contain various kinds of distortions and content, most of which may not exist in the training set. Hence, the generalization ability of the UGC VQA model is very important. In this section, we use the cross-database evaluation to test the generalization ability of the proposed model. Specifically, we train the proposed model on the LSVQ database and test the trained model on the other four UGC VQA databases. We list the results in Table 5. It is observed that the proposed model achieves excellent performance in cross-database evaluation. The SRCC results on the KoNViD-1k and YouTube-UGC databases both exceed 0.8, which have surpassed most VQA models trained on the corresponding database. We find that the multi-scale quality fusion strategy can significantly improve the performance on the databases containing videos with different spatial resolutions (YouTube-UGC, LBVD, and LIVE-YT-Gaming), which further demonstrates its effectiveness.

It is also observed that the performance on the LBVD and LIVE-YT-Gaming databases is not good as the other two databases. The reason is that the LBVD and LIVE-YT-Gaming databases contain live broadcasting and gaming videos respectively, which may rarely exist in the LSVQ database. Since the single database can not cover all kinds of video types and distortions, we may further improve the generalization ability of the proposed model via the multiple database training strategy [30] [51] or the continual learning manner [49] [50].

#### 4.5 Computational Complexity

The computational complexity is a very important factor that needs to be considered in practical applications. Hence, we test the computational complexity in this section. All models are tested on a computer with i7-6920HQ CPU, 16G RAM, and NVIDIA Quadro P400. The deep learning based models and the handcrafted based models are tested using the GPU and CPU respectively. We report the running time for a video with the resolution of 1920×1080 and time duration of eight seconds in Table 6. It is seen that the proposed model has a considerably low running time compared with other VQA models. The reason is that we use very sparse frames to calculate the spatial features while other deep learning based methods need dense frames. Moreover, we extract the motion features at a very low resolution, which only adds little computational complexity to the proposed model. The very low computational complexity makes the proposed model suitable for practical applications.

### 5 CONCLUSION

In this paper, we propose an effective and efficient NR VQA model for UGC videos. The proposed model extracts the quality-aware features from the spatial domain and the spatial-temporal domain to measure the spatial distortions and motion distortions respectively. We train the spatial feature extractor in an end-to-end training manner, so the proposed model can make full use of the various spatial distortions and content in the current VQA database. Then, the quality-aware features are regressed into the quality scores by the MLP network, and the temporal average pooling is used to obtain the video-level quality scores. We further introduce the multi-scale quality fusion strategy to address the problem of quality assessment across different spatial resolutions. The experimental results show that the proposed model can effectively measure the quality of UGC videos.

### ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (61831015, 61901260) and the National Key R&D Program of China 2021YFE0206700.



## REFERENCES

- [1] Jochen Antkowiak, T Jamal Baina, France Vittorio Baroncini, Noel Chateau, France FranceTelecom, Antonio Claudio França Pessoa, F Stephanie Colonese, Italy Laura Contin, Jorge Caviedes, and France Philips. 2000. Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000. (2000).
- [2] Yuqin Cao, Xiongkuo Min, Wei Sun, and Guangtao Zhai. 2021. Deep Neural Networks For Full-Reference And No-Reference Audio-Visual Quality Assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1429–1433.
- [3] Pengfei Chen, Leida Li, Yipo Huang, Fengfeng Tan, and Wenjun Chen. 2019. QoE evaluation for live broadcasting video. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 454–458.
- [4] Pengfei Chen, Leida Li, Lei Ma, Jinjian Wu, and Guangming Shi. 2020. RIRNet: Recurrent-in-recurrent network for video quality assessment. In *Proceedings of the 28th ACM International Conference on Multimedia*. 834–842.
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li-Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slow-fast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6202–6211.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [8] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 2017. The Konstanz natural video database (KoNViD-1k). In *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE, 1–6.
- [9] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* 29 (2020), 4041–4056.
- [10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).
- [11] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. 2018. Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 219–234.
- [12] Jari Korhonen. 2019. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing* 28, 12 (2019), 5923–5938.
- [13] Jari Korhonen, Yicheng Su, and Junyong You. 2020. Blind natural video quality prediction via statistical temporal features and deep spatial features. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3311–3319.
- [14] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. 2021. Blindly Assess Quality of In-the-Wild Videos via Quality-aware Pre-training and Motion Perception. *arXiv preprint arXiv:2108.08505* (2021).
- [15] Dingquan Li, Tingting Jiang, and Ming Jiang. 2019. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2351–2359.
- [16] Dingquan Li, Tingting Jiang, and Ming Jiang. 2021. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision* 129, 4 (2021), 1238–1257.
- [17] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. 2018. End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks.. In *ACM Multimedia*. 546–554.
- [18] Wei Lu, Wei Sun, Xiongkuo Min, Wenhan Zhu, Quan Zhou, Jun He, Qiyuan Wang, Zicheng Zhang, Tao Wang, and Guangtao Zhai. 2022. Deep Neural Network for Blind Visual Quality Assessment of 4K Content. *arXiv preprint arXiv:2206.04363* (2022).
- [19] Pavan C Madhusudana, Xiangxu Yu, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. 2021. Subjective and objective quality assessment of high frame rate videos. *IEEE Access* 9 (2021), 108069–108082.
- [20] Xiongkuo Min, Kede Ma, Ke Gu, Guangtao Zhai, Zhou Wang, and Weisi Lin. 2017. Unified blind quality assessment of compressed natural, graphic, and screen content images. *IEEE Transactions on Image Processing* 26, 11 (2017), 5462–5474.
- [21] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Mylene CQ Farias, and Alan Conrad Bovik. 2020. Study of subjective and objective quality assessment of audio-visual signals. *IEEE Transactions on Image Processing* 29 (2020), 6054–6068.
- [22] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* 21, 12 (2012), 4695–4708.
- [23] Anish Mittal, Michele A Saad, and Alan C Bovik. 2015. A completely blind video integrity oracle. *IEEE Transactions on Image Processing* 25, 1 (2015), 289–300.
- [24] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* 20, 3 (2012), 209–212.
- [25] Abdul Rehman, Kai Zeng, and Zhou Wang. 2015. Display device-adapted video quality-of-experience assessment. In *Human Vision and Electronic Imaging XX*, Vol. 9394. International Society for Optics and Photonics, 939406.
- [26] Michele A Saad, Alan C Bovik, and Christophe Charrier. 2014. Blind prediction of natural video quality. *IEEE Transactions on Image Processing* 23, 3 (2014), 1352–1365.
- [27] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [28] Wei Sun, Xiongkuo Min, Guangtao Zhai, Ke Gu, Huiyu Duan, and Siwei Ma. 2019. MC360QA: a multi-channel CNN for blind 360-degree image quality assessment. *IEEE Journal of Selected Topics in Signal Processing* 14, 1 (2019), 64–77.
- [29] Wei Sun, Xiongkuo Min, Guangtao Zhai, Ke Gu, Siwei Ma, and Xiaokang Yang. 2020. Dynamic backlight scaling considering ambient luminance for mobile videos on lcd displays. *IEEE Transactions on Mobile Computing* (2020).
- [30] Wei Sun, Xiongkuo Min, Guangtao Zhai, and Siwei Ma. 2021. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *arXiv preprint arXiv:2105.14550* (2021).
- [31] Wei Sun, Tao Wang, Xiongkuo Min, Fuwang Yi, and Guangtao Zhai. 2021. Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–6.
- [32] Zhengzhong Tu, Chia-Ju Chen, Li-Heng Chen, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. 2020. A comparative evaluation of temporal pooling methods for blind video quality assessment. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 141–145.
- [33] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. 2021. UGC-VQA: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing* (2021).
- [34] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. 2021. Rapique: Rapid and accurate video quality prediction of user generated content. *arXiv preprint arXiv:2101.10955* (2021).
- [35] Tao Wang, Zicheng Zhang, Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. 2022. Subjective Quality Assessment for Images Generated by Computer Graphics. *arXiv preprint arXiv:2206.05008* (2022).
- [36] Yilin Wang, Sasi Inguva, and Balu Adsumilli. 2019. YouTube UGC dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSp)*. IEEE, 1–5.
- [37] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. 2021. Rich features for perceptual quality assessment of UGC videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13435–13444.
- [38] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. Ieee, 1398–1402.
- [39] Shaoguo Wen and Junle Wang. 2021. A strong baseline for image and video quality assessment. *arXiv preprint arXiv:2111.07104* (2021).
- [40] Jiahua Xu, Jing Li, Xingguang Zhou, Wei Zhou, Baichao Wang, and Zhibo Chen. 2021. Perceptual Quality Assessment of Internet Videos. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1248–1257.
- [41] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C Bovik, and Xiangchu Feng. 2014. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. *IEEE Transactions on Image Processing* 23, 11 (2014), 4850–4862.
- [42] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 1098–1105.
- [43] Fuwang Yi, Mianyi Chen, Wei Sun, Xiongkuo Min, Yuan Tian, and Guangtao Zhai. 2021. Attention Based Network For No-Reference UGC Video Quality Assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1414–1418.
- [44] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. 2021. Patch-VQ: Patching Up the Video Quality Problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14019–14029.
- [45] Xiangxu Yu, Zhenqiang Ying, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. 2022. Subjective and Objective Analysis of Streamed Gaming Videos. *arXiv preprint arXiv:2203.12824* (2022).
- [46] Saman Zadtootaghaj, Nabajeet Barman, Steven Schmidt, Maria G Martini, and Sebastian Möller. 2018. NR-GVQM: A no reference gaming video quality metric. In *2018 IEEE International Symposium on Multimedia (ISM)*. IEEE, 131–134.
- [47] Saman Zadtootaghaj, Steven Schmidt, Saeed Shafiee Sabet, Sebastian Möller, and Carsten Griwodz. 2020. Quality estimation models for gaming video streaming services using perceptual video quality dimensions. In *Proceedings of the 11th ACM Multimedia Systems Conference*. 213–224.
- [48] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.
- [49] Weixia Zhang, Dingquan Li, Chao Ma, Guangtao Zhai, Xiaokang Yang, and Kede Ma. 2021. Continual learning for blind image quality assessment. *arXiv preprint arXiv:2102.09717* (2021).
- [50] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. 2021. Task-specific normalization for continual learning of blind image quality models. *arXiv preprint*

- arXiv:2107.13429* (2021).
- [51] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. 2021. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing* 30 (2021), 3474–3486.
- [52] Qi Zheng, Zhengzhong Tu, Yibo Fan, Xiaoyang Zeng, and Alan C Bovik. 2022. No-Reference Quality Assessment of Variable Frame-Rate Videos Using Temporal Bandpass Statistics. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1795–1799.
- [53] Qi Zheng, Zhengzhong Tu, Pavan C Madhusudana, Xiaoyang Zeng, Alan C Bovik, and Yibo Fan. 2022. FAVER: Blind Quality Prediction of Variable Frame Rate Videos. *arXiv preprint arXiv:2201.01492* (2022).