# Perceptual Quality Assessment of Internet Videos

Jiahua Xu*
Alibaba Group
xujiahua@mail.ustc.edu.cn

Jing Li†
Alibaba Group
jing.li.univ@gmail.com

Xingguang Zhou
Alibaba Group
zxg1120101037@gmail.com

Wei Zhou
University of Science and Technology of China
weichou@mail.ustc.edu.cn

Baichao Wang
Alibaba Group
baichao.wbc@alibaba-inc.com

Zhibo Chen
University of Science and Technology of China
chenzhibo@ustc.edu.cn

## ABSTRACT

With the fast proliferation of online video sites and social media platforms, user, professionally and occupationally generated content (UGC, PGC, OGC) videos are streamed and explosively shared over the Internet. Consequently, it is urgent to monitor the content quality of these Internet videos to guarantee the user experience. However, most existing modern video quality assessment (VQA) databases only include UGC videos and cannot meet the demands for other kinds of Internet videos with real-world distortions. To this end, we collect 1,072 videos from Youku, a leading Chinese video hosting service platform, to establish the Internet video quality assessment database (Youku-V1K). A special sampling method based on several quality indicators is adopted to maximize the content and distortion diversities within a limited database, and a probabilistic graphical model is applied to recover reliable labels from noisy crowdsourcing annotations. Based on the properties of Internet videos originated from Youku, we propose a spatio-temporal distortion-aware model (STDAM). First, the model works blindly which means the pristine video is unnecessary. Second, the model is familiar with diverse contents by pre-training on the large-scale image quality assessment databases. Third, to measure spatial and temporal distortions, we introduce the graph convolution and attention module to extract and enhance the features of the input video. Besides, we leverage the motion information and integrate the frame-level features into video-level features via a bi-directional long short-term memory network. Experimental results on the self-built database and the public VQA databases demonstrate that our model outperforms the state-of-the-art methods and exhibits promising generalization ability.

## CCS CONCEPTS

• **Information systems** → **Multimedia databases**; • **Computing methodologies** → **Image processing**.

## KEYWORDS

Internet videos; perceptual quality; database; model

**ACM Reference Format:**
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 (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475486

## 1 INTRODUCTION

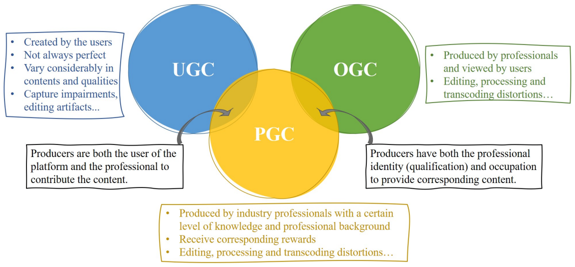
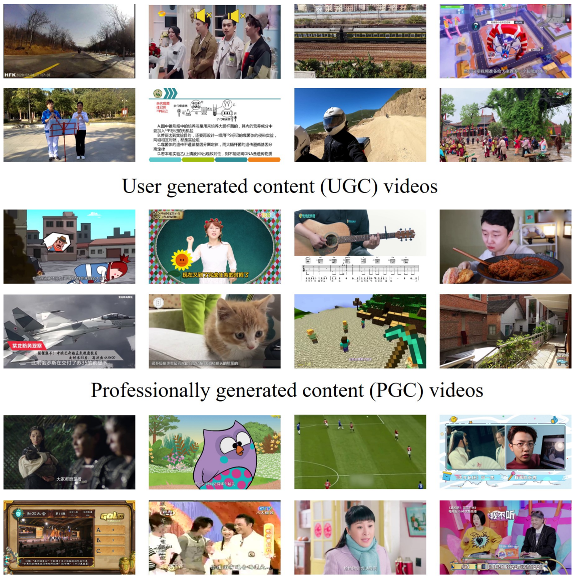

Created by the users
Not always perfect
Vary considerably in contents and qualities
Capture impairments, editing artifacts...
UGC
OGC
Produced by professionals and viewed by users
Editing, processing and transcoding distortions…
PGC
Producers are both the user of the platform and the professional to contribute the content.
Producers have both the professional identity (qualification) and occupation to provide corresponding content.
Produced by industry professionals with a certain level of knowledge and professional background
Receive corresponding rewards
Editing, processing and transcoding distortions…

**Figure 1: The descriptions of UGC, PGC and OGC videos.**

Videos account for the majority of Internet traffic in recent years [2]. Specifically, Internet videos can be classified into user generated content (UGC), professionally generated content (PGC), and occupationally generated content (OGC) as shown in Figure 1. During the processing chain, e.g. acquisition, compression, storage and transmission, multiple distortions will be introduced, leading to visual quality degradation [4, 8, 55, 57]. To guarantee the quality of experience (QoE) of end-users, image/video quality assessment (IQA/VQA) plays a significant role to guide the current image processing and video coding systems [58]. According to human engagement, VQA can be roughly divided into two categories, namely subjective quality assessment and objective quality assessment [40]. Subjective quality assessment requires human rating, thus is able to produce the most accurate quality labels for database construction [33]. However, it is labor-intensive and time-consuming, which is unsuitable for real-time applications. Therefore, to automatically predict the perceptual quality, objective quality assessment is developed and has been deeply researched in the past decades [5].

[*] Also with University of Science and Technology of China.
[†] Corresponding author.

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.
*MM '21, October 20–24, 2021, Virtual Event, China*
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475486



User generated content (UGC) videos

Professionally generated content (PGC) videos

Occupationally generated content (OGC) videos

**Figure 2: Sample videos in the Youku-V1K database.**

Traditional video quality assessment databases are usually constructed by selecting standard high-quality videos as the reference, and then adding synthetic corruptions to the reference. This is the common practice in previous years [40]. Later, with the development of online video sites and social media platforms, users can change their role into the producer to freely create contents, which is called UGC [44]. Different from conventional cases, the reference provided by the users are not always perfect and vary considerably in contents and qualities. For example, capture impairments including photography equipment limitations, bad shooting environments, improper camera parameters will deteriorate the visual quality.

Going a step further, PGC and OGC videos are also in great demands to attract more users and achieve larger business values. PGC videos are produced by professionals and viewed by users. OGC videos are produced by industry professionals with a certain level of knowledge and professional background, and these people will receive corresponding rewards [10]. Compared with UGC videos created by ordinary users, PGC and OGC videos are generally exhibited in high quality. However, camera language in PGC and OGC videos [7] will bring additional challenges, such as the case that blur is imposed on background or unimportant objects to highlight the main subject/object, which is considered as low-quality in conventional VQA problems. Besides, cross-platform sharing PGC and OGC videos will also introduce editing, processing and transcoding distortions. Therefore, the challenges of Internet VQA can be summarized as (a) no high-quality video for comparison, (b) diverse contents including meaningful scenarios (e.g. natural scenes, dramas, cartoons, screen content, etc.) as well as low content quality videos, (c) complex mixed distortions, namely capture distortions, compression, transmission error, transcoding artifacts, quality fluctuations and conflicts brought by camera language.

Designing suitable Internet VQA metrics relies on an accurate labeling database. However, to our best of knowledge, there has been no existing work considering quality assessment for UGC, PGC and OGC videos. Therefore, we build the Internet video quality assessment database (Figure 2) with 1,072 videos collected from Youku (Youku-V1K). Representative videos are sampled according to their quality indicators [1] including spatial activity, temporal activity, predicted mean opinion score (MOS), etc. Afterwards, online subjective experiments are conducted to achieve human ratings. Finally, we apply a probabilistic graphical model [23] to recover the ground truth labels given the noisy and unreliable crowdsourcing annotations.

To cope with the challenges in Internet VQA, we propose a spatio-temporal distortion-aware model (STDAM). Firstly, this model takes only the video frames as input, which needs no reference video for comparison. Secondly, the frame-level model is pre-trained on an existing large-scale IQA database to be acquainted with diverse scenarios and image distortions. Thirdly, we apply several techniques to handle the complex distortions in Internet VQA, namely 1) *graph convolution module*, which aims to construct relations among long-distance pixels and cross-scale features, thus enlarge the receptive field of spatial distortions in Internet videos, 2) *attention module*, which enhances the feature representation for salient regions and important channels, and alleviate the conflicts brought by camera language mainly in PGC and OGC videos, 3) *optical flow module*, which corresponds to the quality degradation caused by camera motions especially in UGC videos, 4) *bi-directional long short-term memory (LSTM) module*, which deals with the quality fluctuation in Internet videos based on the assumption that the quality of current frame is influenced by its previous and later frames, and the importance of each frame to final quality prediction is different. The proposed model is verified on the self-built Youku-V1K database and several public VQA databases.

The main contributions can be listed as follows:

- With the carefully sampled videos to maintain content and distortion diversity, we establish the Youku-V1K database including UGC, PGC and OGC videos to provide a benchmark for designing and comparing VQA metrics.
- We conduct a subjective experiment on the self-built crowd rating system, and utilize a probabilistic graphical model to provide more reliable quality labels from noisy crowdsourcing annotations.
- We propose the STDAM to automatically predict the perceptual quality of Internet videos, which is validated on several databases.

## 2 RELATED WORK

### 2.1 Databases for VQA

Databases serve as the crucial benchmark for designing and evaluating algorithms in many computer vision tasks, e.g. classification, segmentation, detection, etc. Naturally, quality assessment of visual contents is also one of these tasks which rely on accurate labeling databases. Traditional VQA databases like IRCCyN/IVC 1080i [35], LIVE [40], CSIQ [45] only contain dozens of reference videos distorted with compression artifacts or transmission errors. Under such settings, the high-quality pristine videos are available and the distortions are all synthetic. However, it is not the case for Internet video quality assessment.

**Table 1: Database summary for video quality assessment.**

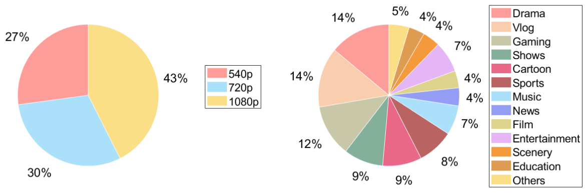
| Databases | | Source | # of videos(Ref/Dis) | Video length | Resolution | Distortion type | Subjective environment |
|---|---|---|---|---|---|---|---|
| With high-quality reference | IRCCyN/IVC 1080i [35] | High-quality reference | 24/168 | 9-12s | 1080p | synthetic | Laboratory |
| | LIVE [40] | High-quality reference | 10/150 | 8-10s | 768x432 | synthetic | Laboratory |
| | CSIQ [45] | High-quality reference | 12/216 | 10s | 832x480 | synthetic | Laboratory |
| Without high-quality reference | CVD2014 [34] | Captured | -/234 | 10-25s | 480p, 720p | authentic | Laboratory |
| | LIVE-Qualcomm [12] | Captured | -/208 | 15s | 1080p | authentic | Laboratory |
| | LIVE-VQC [42] | Captured | -/585 | 10s | 480p-1080p | authentic | Crowdsourcing |
| | KoNViD-1k [17] | Flicker | -/1200 | 8s | 540p | authentic (UGC) | Crowdsourcing |
| | YouTube-UGC [44] | YouTube | -/1380 | 20s | 360p-2160p | authentic (UGC) | Crowdsourcing |
| | Youku-V1K | Youku | -/1072 | 10s | 1080p | authentic (UGC+PGC+OGC) | Crowdsourcing |

Later, more databases containing no perfect-quality reference videos, e.g. CVD2014 [34], LIVE-Qualcomm [12], LIVE-VQC [42] are proposed focusing on the effect of cameras and in-capture distortions. In these databases, the perceptual quality is mainly affected by the inherent limitations, improper operations of cameras and unpredicted object motion and light condition. Besides, databases with videos uploaded by users and shared online [17, 44] also have no reference video for comparison. Apart from in-capture distortions [12], transcoding artifacts during the uploading and transmission process over limited bandwidth could be involved. These databases can be denoted as UGC VQA databases. However, the PGC and OGC videos account for a large part of online media too, which are not considered in the previous databases. Therefore, we build the Internet VQA database Youku-V1K with extremely diverse contents, distortions, and accurate quality labels. Moreover, a probabilistic graphic is applied to infer reliable quality labels from crowdsourcing results. The summary of existing VQA databases is listed in Table 1.

## 2.2 Objective VQA Metrics

To precisely evaluate the perceptual quality of videos, traditional methods mainly focus on the structural [47], gradient [29], motion [38], saliency [52] information and usually require reference videos for comparison. The hand-crafted features based on natural scene statistics and support vector regression (SVR) are frequently utilized to predict the quality [36] when the pristine videos are unavailable. Afterwards, with the development of deep learning technologies, we can assess the quality with convolutional neural networks (CNN) in an end-to-end manner. However, most of the above-mentioned methods are designed for synthetic distortions, which have poor generalization ability for the authentic distortions [28].

The existence of real-world VQA databases promotes the design of blind VQA (BVQA) model for videos suffering from various complex distortions. TLVQM [20] is proposed by considering low complexity features for each frame and high complexity features for representative frames. VIDEVAL [44] is another hand-crafted feature based model that derives from features of several famous blind IQA/VQA metrics. Feature selection in hand-crafted models is a knowledge-based process that relies on the rich experience and comprehensive understanding of the media contents and distortions. Later, some learning based methods are proposed, e.g., V-MEON [28] for compression artifacts, MLSP-VQA using multi-level spatially pooled features [13]. Zhang *et al.* [54] design a blind video quality assessment model in the 3D-DCT domain, and apply a resampling strategy from image to video. Considering the content-aware characteristics and temporal-memory effect, Li *et al.* develop an objective deep neural network VSFA [22] for quality assessment of videos in the wild. However, the spatio-temporal distortions and the conflicts brought by camera language cannot be well addressed in these methods, thus we propose a new model to effectively handle the complicated corruptions in Internet videos.



(a) (b)

**Figure 3: Resolution and category distribution of Youku-V1K database.**

# 3 YOUKU-V1K DATABASE

In this section, we first describe the video sampling process to construct the Youku-V1K database. Then, the detailed configurations of the subjective experiment are presented. Finally, given the noisy crowdsourcing ratings, we apply a probabilistic graphical to cleanse the data.

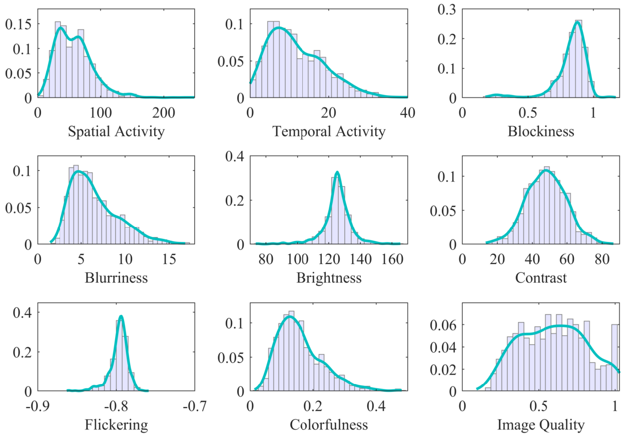


**Figure 4: Feature distributions of Youku-V1K database.**

### 3.1 Database Construction

10,000 video clips with diverse contents are collected from an online video site youku.com. Firstly, resolution and content are considered to conduct the pre-selection and there remain 3,000 videos for careful sampling. The video sampling strategy is similar to [46], while more quality indicators are considered in the sampling process including spatial activity, temporal activity, blockiness, blurriness, brightness, contrast, flickering, colorfulness and predicted image quality. The first eight factors are calculated with the image quality toolbox [1] and the last is computed by the VGG-16 network pretrained on image quality assessment databases. The sampling steps can be summarized as follows:

**(1)** Normalize the feature space for each video clip.
**(2)** Uniformly divide the normalized range into N(=3) bins for all features as done in [46].
**(3)** Change the order of bins randomly.
**(4)** For the current bin, select one video clip and add it to the database if the Euclidean distance of between this clip and clips in the database is greater than a threshold (0.3) according to [46].
**(5)** Switch to the next bin and repeat (4) until the database has enough samples.

After sampling, the distributions of different resolutions and content categories are shown in Figure 3. Besides, the feature distributions and uniformity of videos in each bin are shown in Figure 4 and Figure 5. As we can conclude from these figures, the content diversity and feature uniformity are promised. Note that all videos are resized to 1080p before the subjective experiment in accordance with the display resolution.

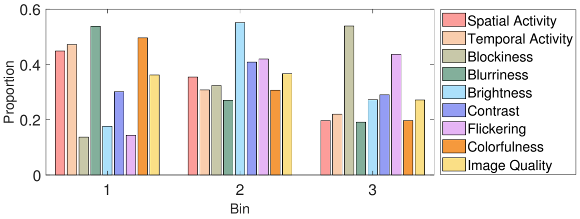
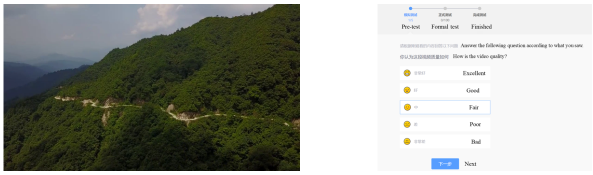


**Figure 5: Feature uniformity of Youku-V1K database.**

### 3.2 Subjective Testing

We adopt the Absolute Category Rating (ACR) [18] in our experiment. It is a single stimulus evaluation method and voting is performed after each viewing. The video quality is divided into five levels including 5-Excellent, 4-Good, 3-Fair, 2-Poor, 1-Bad. The user interface is shown in Figure 6, and the crowd rating system is developed by ourselves.


模拟测试 1/5 Pre-test
正式测试 0/100 Formal test
完成测试 Finished

请根据刚观看的内容回答以下问题 Answer the following question according to what you saw.

你认为这段视频质量如何 How is the video quality?

非常好 Excellent
好 Good
中 Fair
差 Poor
非常差 Bad

下一步 Next

(a) Viewing.

(b) Rating.

**Figure 6: User interface of subjective experiment.**

More than 22,000 crowdsourcing results are collected in our subjective experiment, and each video is rated by more than 15 observers which meets the requirements of Rec. ITU-R BT.500-9 [3]. The subjects are volunteers with payment aging from 18 to 49. 5 videos with different qualities are displayed in the pre-test to let the participants familiar with grading rules. In the formal test, each participant is asked to rate 100 unlabeled videos randomly. During the experiment, subjects can take breaks after rating a video to avoid eye fatigue.

### 3.3 Data Cleansing

In crowdsourcing tasks, noise always exists owing to the annotator's unreliability and task's difficulty. Thus, we adopt a probabilistic graphical model to recover the ground truth distribution of video rating.

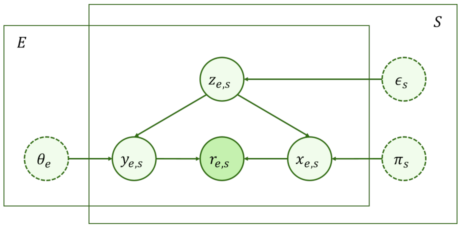
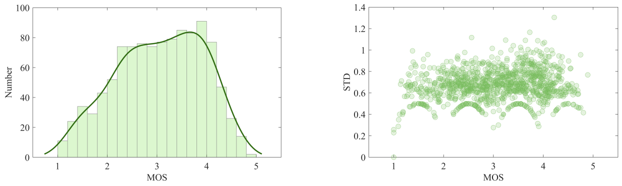


**Figure 7: The graphic model for recovering ground truth labeling. $r_{e,s}$ is given by the subjects, $y_{e,s}$, $x_{e,s}$ and $z_{e,s}$ are latent variables, $\theta_e$, $\epsilon_s$ and $\pi_s$ are parameters.**

The model is illustrated in Figure 7, where $S$ and $E$ are the total number of subjects and test videos, $r_{e,s}$ is the label given by subject $s$ to test video $e$, $y_{e,s}$, $x_{e,s}$ are the label given by subject $s$ for video $e$ according to the underlying ground truth distribution and subject's irregular behaviors. $z_{e,s}$ follows Bernoulli distribution determined by subject $s$. $\theta_{e,n}$ denotes the probability of acquiring score $n$ for video $e$, $\epsilon_s$ denotes how conscientious subject $s$ is, and $\pi_s$ denotes the irregular behavior of subject $s$. The conditional density of subjective rating [23] is given as:

$$
\begin{aligned}
p(R|\pi,\epsilon,\theta) &= \prod_{e,s\in A}\Big[\epsilon_s\Big(\prod_{n=1}^{N}\theta_{e,n}^{[r_{e,s}=n]}\Big) + (1-\epsilon_s)\Big(\prod_{n=1}^{N}\pi_{s,n}^{[r_{e,s}=n]}\Big)\Big], \\
s.t. \quad & 0 \le \theta_{e,n} \le 1, \sum_{n=1}^{N}\theta_{e,n} = 1, 0 \le \pi_{s,n} \le 1, \sum_{n=1}^{N}\pi_{s,n} = 1.
\end{aligned}
\tag{1}
$$

Accordingly, we apply Maximum Likelihood Estimates to infer the parameters, and the likelihood function is $\hat{\delta} = \arg \underset{\delta}{max} \log p(R|\delta)$, where $\delta = (\theta, \epsilon, \pi)$. Then, the ground truth quality can be expressed as the expectation of the estimated distribution $\sum_{n=1}^{N} n \cdot \theta_{e,n}$.



(a)

(b)

**Figure 8: (a) MOS distribution. (b) MOS versus STD distribution of Youku-V1K database.**

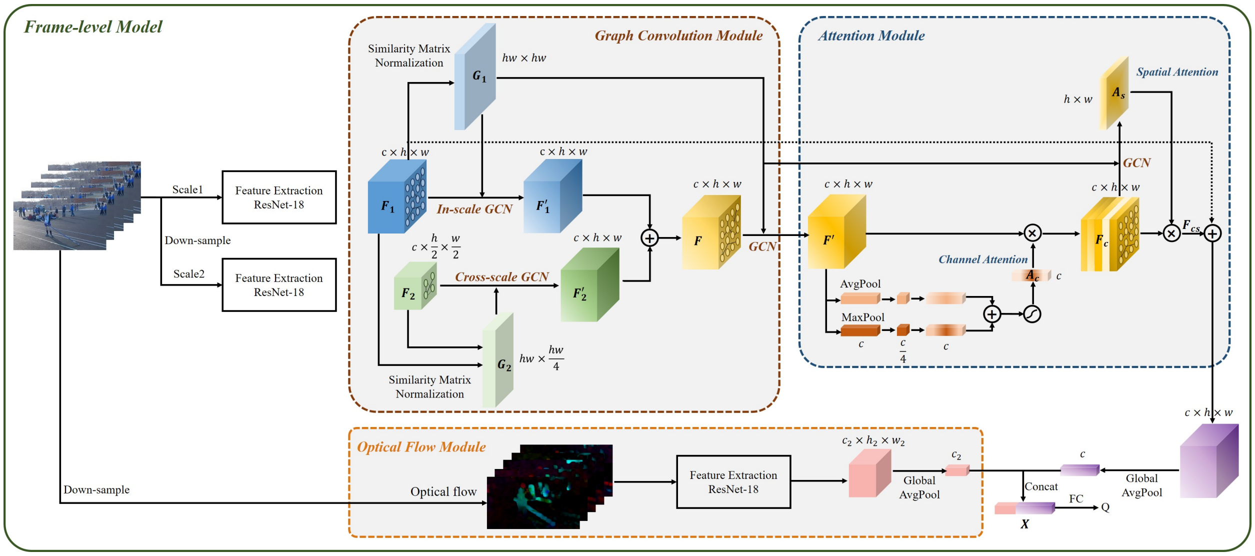


**Figure 9: The structure of the frame-level model including the graph convolution, attention and optical flow modules.**

After data cleansing, the MOS and MOS versus STD distribution are presented in Figure 8. As is shown, more high-quality videos are included and the standard deviations of the opinion scores parameter is 0.19, which falls in the range [0.11, 0.21] for standard VQA experiments [16]. Compared with previous UGC databases, Youku-V1K utilizes the probabilistic graphic model to guarantee the label accuracy with less intensive labor, which is verified in Section 5.4.

## 4 PROPOSED MODEL

Based on the proposed Youku-V1K database, we develop a no-reference spatio-temporal distortion-aware model (STDAM) to effectively evaluate the spatial, temporal distortions, and alleviate the conflicts brought by camera language. STDAM is illustrated in Figure 9 and Figure 10, and it can be decomposed into the frame-level and video-level model.

### 4.1 Frame-level Model

The frame-level model contains the graph convolution, attention and optical flow modules. The graph convolution module (GCM) and attention module (AM) are designed to handle spatial distortions and camera language conflicts in Internet videos. Specifically, GCM can capture long-distance [27] and cross-scale relations [24], thus ***enlarge the receptive field of spatial distortions***. AM is utilized to enhance the features of salient regions and discriminative channels for quality regression, and ***alleviate the conflicts brought by camera language mainly in PGC and OGC videos***. Moreover, the optical flow module (OFM) is utilized to measure the ***temporal distortions caused by the camera motions especially in UGC videos***.

**Graph Convolution Module:** The original frame $I_1$ and its down-sampled version $I_2$ are first fed into ResNet-18 [15] for different scales feature extraction, and the extracted features are represented as $F_1$ and $F_2$. To explore the long-distance relations and model visual dependency, we build the in-scale graph on $F_1$. The individual spatial locations are defined as the graph nodes $F_1 = [f_{11}, f_{12}, ..., f_{1N}] \in \mathbb{R}^{N \times c}$, where $N = h \times w$ indicates the number of graph nodes. Each location node is a $c$ dimensional vector, and $h$, $w$, $c$ denote the height, width and channels of extracted feature $F_1$. Then, the affinity between every two nodes are represented as the cosine similarity:

$$A_1(f_{1i}, f_{1j}) = \frac{f_{1i} \cdot f_{1j}}{\|f_{1i}\| \, \|f_{1j}\|}, \tag{2}$$

where $A_1$ is the affinity matrix for the in-scale graph. The normalized adjacency matrix $G_1$ is obtained using symmetric Laplacian normalization $G_1 = D^{-\frac{1}{2}} A_1 D^{-\frac{1}{2}}$, where $D$ denotes the diagonal matrix and $D_{ii} = \sum_j A_{1\,ij}$. The feature map $F'_1$ after in-scale graph convolution is $F'_1 = G_1 F_1 W_1$, where $W_1$ is the trainable weight matrix. Besides, we adopt cross-scale graph convolution for the down-sampled frame feature map $F_2$. The affinity of cross-scale graph nodes $A_2$ and the normalized adjacency matrix $G_2$ are defined as:

$$A_2(f_{1i}, f_{2j}) = \frac{f_{1i} \cdot f_{2j}}{\|f_{1i}\| \, \|f_{2j}\|}, \tag{3}$$

$$G_2^{ij} = \frac{\exp A_1(f_{1i}, f_{2j})}{\sum_{j=1}^{N} \exp A_1(f_{1i}, f_{2j})}. \tag{4}$$

Then, the cross-scale graph convolution is $F'_2 = G_2 F_2 W_2$, where $W_2$ is the trainable weight matrix. Afterwards, in-scale and cross-scale information are integrated through summation $F$ and one-layer graph convolution $F'$ as shown in Figure 9. By building the graphs, we can obtain more structure information and enlarge the number of propagation neighbors during graph convolution.

**Attention Module:** It is introduced to enhance the features for discriminative channels and salient regions. At first, the spatial information is aggregated by average and max pooling operations. A shared multi-layer perceptron is followed to explore the inter-channel relations [48] and infer better channel-wise attention. We merge both features by element-wise summation, and the total

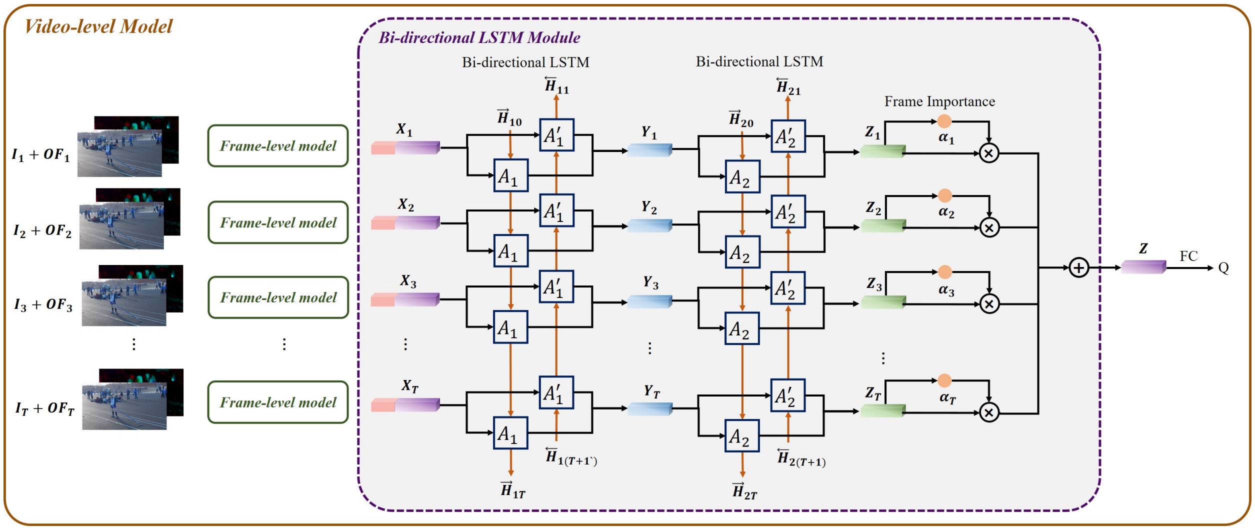


**Figure 10: The structure of the video-level model including the frame-level model and the bi-directional LSTM module.**

channel attention can be expressed as:

$$A_c = \sigma(\text{MLP}(AvgPool(F')) + \text{MLP}(MaxPool(F'))), \tag{5}$$

$$F_c = A_c \otimes F', \tag{6}$$

where $F_c$ is the channel-refined feature, $\otimes$ denotes element-wise multiplication, and $\sigma$ is the Sigmoid activation function. Then, we utilize graph convolution to generate the spatial attention map by capturing the relations among adjacent and long-distance regions. During the graph convolution, the channel dimension is reduced to one for highlighting the most informative regions. The spatial attention map are fused with feature $F_c$, which is complementary to channel attention:

$$A_s = G_1(G_1 F_c W_{s1}) W_{s2}, \tag{7}$$

$$F_{cs} = A_s \otimes F_c, \tag{8}$$

where $F_{cs}$ denotes the spatial and channel refined feature, $W_{s1}$ and $W_{s2}$ are the trainable weight matrix to reduce channel dimension. Finally, we apply a residual connection by adding $F_1$ and $F_{cs}$, and this feature can be utilized for global average pooling.

**Optical Flow Module:** Camera shake and movement will bring visual quality degradation. Thus, to effectively evaluate such distortion, optical flow is computed for motion information representation. Since we mainly focus on the global motion of cameras and the computation complexity of optical flow is high, we down-sample the video frames before optical flow generation (default in OpenCV) [9]. ResNet-18 is followed to extract motion-aware features from the optical flow maps. The motion-aware features are then concatenated with channel and spatial refined features as the frame-level feature vectors $X$ for quality estimation.

## 4.2 Video-level Model

As is shown in Figure 10, we feed the features extracted from the frame-level model into the bi-directional LSTM [37] module, which is aimed at evaluating ***the influence of quality fluctuations in Internet videos***.

**Bi-directional LSTM Module:** The frame-level features $[X_1, ..., X_T]$ are sent to the bi-directional LSTM $(A/A')$, since the current frame quality is influenced by previous and next frames according to [39]. The hidden states of the first LSTM layer are initialized as $\vec{H}_{10}, \overleftarrow{H}_{1(T+1)}$. The current hidden state $\vec{H}_{1t}, \overleftarrow{H}_{1t}$ are calculated from the current input $X_t$, and previous/next hidden state $\vec{H}_{1(t-1)}/\overleftarrow{H}_{1(t+1)}$ as follows:

$$\vec{H}_{1t} = A_1(X_t, \vec{H}_{1(t-1)}), \tag{9}$$

$$\overleftarrow{H}_{1t} = A_1'(X_t, \overleftarrow{H}_{1(t+1)}), \tag{10}$$

where $A_1$, $A_1'$ are the LSTM units and $t$ denotes the current frame index. The first layer output vector $Y_t$ for current frame is obtained by concatenating $\vec{H}_{1t}$ and $\overleftarrow{H}_{1t}$. There are two bi-directional LSTM layers in our framework, and the final feature representation for each frame is $Z_t$. One fully-connected layer is adopted to reduce the feature dimension of $Z_t$ to 1, which indicates the frame importance. After that, Softmax normalization is utilized to guarantee the sum of the frame importance equaling to 1. The video-level feature $Z$ is then denoted as:

$$Z = \sum_{t=1}^{T} \alpha_t Z_t, \tag{11}$$

where $\alpha_t$ indicates the weight for the $t$-frame, and $T$ frames are contained in total. We apply two fully-connected (FC) layers for quality regression to estimate the final perceptual quality of an Internet video.

## 4.3 Implementation Details

The experiments are conducted on NVIDIA 1080Ti GPUs and the model is implemented with PyTorch. The ResNet-18 feature extraction network is initialized with the weights pre-trained on ImageNet [6]. Firstly, we train the frame-level model with graph convolution and attention module on the KonIQ-10k database [25], which is a IQA database with over 10,000 images. Therefore, it can be familiar with diverse contents and distortions. Secondly, the total frame-level model (including optical flow module) is trained with frames of Internet videos. $L_2$ loss and Adam optimizer [19] with an initial learning rate 0.0001 is adopted in this step. The learning rate is scaled by 0.25 every 5 epochs and 20 epochs are required for training the frame-level model. Finally, we uniformly sample 10 frames from each video to train the video-level model as done in [56]. Loss function and optimizer are the same as the previous step, but we only need to train the video-level model for 10 epochs with the weight of frame-level model fixed. The mini-batch size is set to 16 during training and the MOS values are scaled to [0, 1] in our experiment.

# 5 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we compare our proposed STDAM with state-of-the-art (SOTA) metrics on the self-built Youku-V1K and other public VQA databases. Ablation study and cross database tests are also conducted to verify the effectiveness and robustness of STDAM.

## 5.1 Databases and Performance Measures

Four databases are utilized in our experiment, namely self-built Youku-V1K, KoNViD-1k [17], LIVE-VQC [42] and YouTube-UGC [44]. The latter three databases are described in Section 2.1, and all the databases for performance comparison contain no reference videos, thus full-reference VQA metrics cannot be evaluated. Spearman's rank order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC) are adopted to measure the prediction monotonicity and accuracy. Their values are in the range of [0,1] and the higher value means the better performance. Note that before calculating PLCC, a five-parameter logistic regression is adopted as suggested by Video Quality Experts Group [14].

## 5.2 Performance Evaluation

To evaluate the performance of the proposed STDAM, we divide each database into 80% training and 20% testing videos [43]. According to [26], the experiments are conducted 10 times with random train-test splitting operation to avoid content bias, and we report the median value and standard deviation of SROCC and PLCC for performance comparison. The compared "completely blind" opinion unaware metrics which do not require training include NIQE [32], ILNIQE [53] for IQA and VIIDEO [31] for VQA. The opinion aware metrics which require training are BRISQUE [30], GM-LOG [50], HIGRADE [21], FRIQUEE [11], CORNIA [51], HOSA [49], pre-trained VGG-19 [41], pre-trained ResNet-50 [15] for IQA and V-BLIINDS [36], TLVQM [20], VIDEVAL [44] for VQA. The experimental results are listed in Table 2, and we can observe that our proposed STDAM achieves competitive performance on all four databases. For Youku-V1K, KoNViD-1k and YouTube-UGC databases, STDAM can achieve 3%-6% performance improvement compared with SOTA methods. Besides, the VQA models including V-BLIINDS, TLVQM and VIDEVAL perform much better than other IQA models especially on LIVE-VQC database, since this database contains more temporal distortions caused by large camera motions. Therefore, the VQA metrics considering motion-related features usually achieve higher performance. Moreover, the standard deviations of SROCC and PLCC for Youku-V1K database are generally smaller than other databases, indicating that the quality labels for the videos in Youku-V1K database are more accurate and reliable.

**Table 2: SROCC and PLCC performance comparison on four VQA databases. The best and second-best performing results are marked in boldface and underlined.**

| SROCC | Youku-V1K | KoNViD-1k | LIVE-VQC | YouTube-UGC |
|---|---|---|---|---|
| NIQE | 0.5782(±0.0112) | 0.5417(±0.0347) | 0.5957(±0.0571) | 0.2379(±0.0487) |
| ILNIQE | 0.4427(±0.0121) | 0.5264(±0.0294) | 0.5037(±0.0712) | 0.2918(±0.0502) |
| VIIDEO | 0.4210(±0.0124) | 0.2988(±0.0561) | 0.0332(±0.0856) | 0.0580(±0.0536) |
| BRISQUE | 0.7804(±0.0268) | 0.6567(±0.0351) | 0.5929(±0.0681) | 0.3820(±0.0519) |
| GM-LOG | 0.7930(±0.0241) | 0.6578(±0.0324) | 0.5881(±0.0683) | 0.3678(±0.0589) |
| HIGRADE | 0.8486(±0.0170) | 0.7206(±0.0302) | 0.6103(±0.0680) | 0.7376(±0.0338) |
| FRIQUEE | 0.8512(±0.0182) | 0.7472(±0.0263) | 0.6579(±0.0536) | 0.7652(±0.0301) |
| CORINA | 0.8464(±0.0176) | 0.7169(±0.0245) | 0.6719(±0.0473) | 0.5972(±0.0413) |
| HOSA | 0.8480(±0.0144) | 0.7654(±0.0224) | 0.6873(±0.0462) | 0.6025(±0.0344) |
| VGG-19 | 0.8647(±0.0180) | 0.7741(±0.0288) | 0.6568(±0.0536) | 0.7025(±0.0281) |
| ResNet-50 | 0.8791(±0.0157) | 0.8018(±0.0255) | 0.6636(±0.0511) | 0.7183(±0.0281) |
| V-BLIINDS | 0.7822(±0.0245) | 0.7101(±0.0314) | 0.6939(±0.0502) | 0.5590(±0.0496) |
| TLVQM | 0.7832(±0.0237) | 0.7729(±0.0242) | **0.7988(±0.0365)** | 0.6693(±0.0306) |
| VIDEVAL | 0.8294(±0.0183) | 0.7832(±0.0216) | 0.7522(±0.0390) | 0.7787(±0.0254) |
| STDAM | **0.9141(±0.0089)** | **0.8448(±0.0189)** | 0.7931(±0.0340) | **0.8341(±0.0306)** |

| PLCC | Youku-V1K | KoNViD-1k | LIVE-VQC | YouTube-UGC |
|---|---|---|---|---|
| NIQE | 0.6046(±0.0097) | 0.5530(±0.0337) | 0.6286(±0.0512) | 0.2776(±0.0431) |
| ILNIQE | 0.4685(±0.0110) | 0.5400(±0.0337) | 0.5437(±0.0717) | 0.3302(±0.0579) |
| VIIDEO | 0.4148(±0.0128) | 0.3002(±0.0539) | 0.2146(±0.0903) | 0.1534(±0.0498) |
| BRISQUE | 0.7801(±0.0278) | 0.6576(±0.0342) | 0.6380(±0.0632) | 0.3952(±0.0486) |
| GM-LOG | 0.7958(±0.0545) | 0.6636(±0.0315) | 0.6212(±0.0636) | 0.3920(±0.0594) |
| HIGRADE | 0.8507(±0.0166) | 0.7269(±0.0287) | 0.6332(±0.0652) | 0.7216(±0.0334) |
| FRIQUEE | 0.8508(±0.0185) | 0.7482(±0.0257) | 0.7000(±0.0587) | 07571(±0.0324) |
| CORINA | 0.8479(±0.0188) | 0.7135(±0.0236) | 0.7183(±0.0420) | 0.6057(±0.0399) |
| HOSA | 0.8485(±0.0144) | 0.7664(±0.0207) | 0.7414(±0.0410) | 0.6047(±0.0347) |
| VGG-19 | 0.8704(±0.0156) | 0.7845(±0.0246) | 0.7160(±0.0481) | 0.6997(±0.0281) |
| ResNet-50 | 0.8821(±0.0154) | 0.8104(±0.0229) | 0.7205(±0.0434) | 0.7097(±0.0276) |
| V-BLIINDS | 0.7844(±0.0249) | 0.7037(±0.0301) | 0.7178(±0.0500) | 0.5551(±0.0465) |
| TLVQM | 0.7849(±0.0243) | 0.7688(±0.0238) | 0.8025(±0.0360) | 0.6590(±0.0302) |
| VIDEVAL | 0.8304(±0.0187) | 0.7803(±0.0233) | 0.7514(±0.0420) | 0.7733(±0.0257) |
| STDAM | **0.9120(±0.0074)** | **0.8415(±0.0173)** | **0.8204(±0.0342)** | **0.8297(±0.0279)** |

## 5.3 Ablation Study

We conduct the ablation experiments to verify the effectiveness of each key component in our proposed STDAM. The modified ResNet-18 with fully-connected network for regression is regarded as the baseline model (BL), and we analyze the validity of graph convolution, attention, optical flow, and bi-directional LSTM modules as well as the pre-training stage on IQA databases.

**Graph Convolution Module:** We first evaluate the graph convolution module by adding this module to the baseline model (BL+GCM). The experiments are conducted on the four databases to better illustrate the influence of spatial distortion. As is shown in Figure 11, BL+GCM can achieve 1%-3% improvement on the databases.

**Attention Module:** By adding attention module to BL+GCM (BL+GCM+AM), SROCC performance of our model further improves 0.01-0.02. The attention maps are visualized in Figure 12 to validate that human attention is focused on salient objects when assessing the visual quality.

**Pre-training Stage:** The ability to handle spatial distortions in Internet videos can be raised by pre-training on IQA databases with diverse contents and distortions (BL+GCM+AM+PS). We can observe from Figure 11 that BL+GCM+AM+PS brings up to 5% improvement on YouTube-UGC database.

**Optical Flow Module:** It is introduced to better deal with the temporal distortion caused by large camera motions. The entire frame model is denoted as BL+GCM+AM+PS+OFM, and it can further achieve much improvements on LIVE-VQC databse.

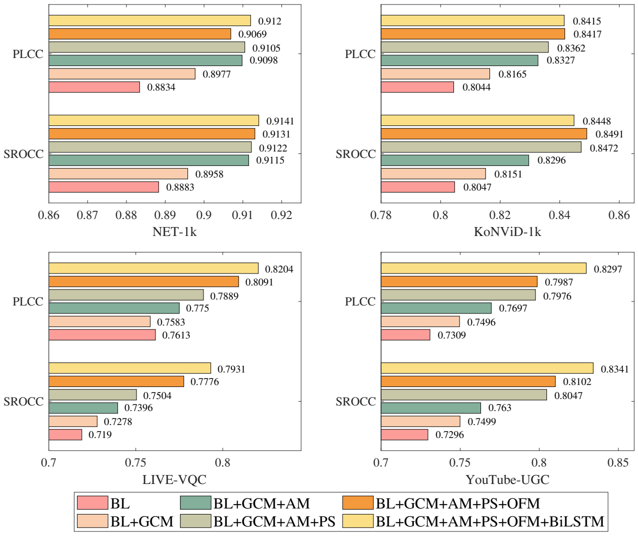


**Figure 11: Ablation study on four VQA databases.**

**Bi-directional LSTM Module:** We also exhibit the improvements brought by the bi-directional LSTM module (BL+GCM+AM+PS+OFM+BiLSTM). The effects of previous and next frame to current frame quality are considered in this module, and we can obtain the highest SROCC value for majority of the databases.

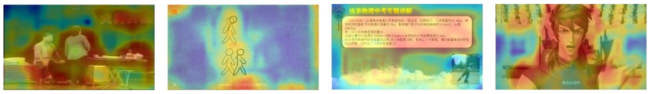


**Figure 12: Spatial attention maps for some video frames in Youku-V1K database.**

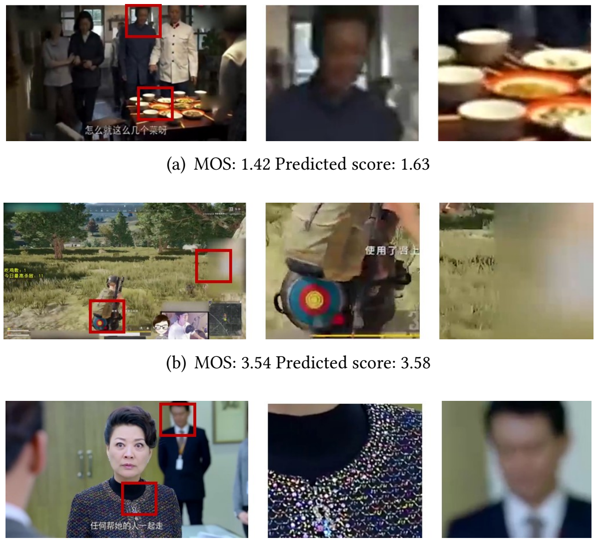
We exhibit the samples of video frames with their associate MOS values and predicted qualities in Figure 13. The predicted scores are very close to the ground truth labels, which means the proposed STDAM is able to handle local distortions (e.g. Figure 13 (b)) and camera language conflicts (e.g. Figure 13 (c)).

## 5.4 Cross Database Tests

To evaluate the generalization ability of STDAM and applicability of Youku-V1K database, cross database experiments are conducted. We train STDAM on one database and test it on another. In Table 3, we report the cross database performance of STDAM. Compared with the SROCC and PLCC presented in [44], our proposed STDAM demonstrates better performance, representing robust generalization capacities. Moreover, the applicability of Youku-V1K database is verified. The model pre-trained on Youku-V1K shows better generalization ability on YouTube-UGC compared to the other two databases, which demonstrates the effectiveness of Youku-V1K.

## 6 CONCLUSION

In this paper, we build the Youku-V1K database for perceptual quality assessment of Internet videos (e.g. UGC, PGC and OGC videos), which provides a benchmark for designing and comparing VQA metrics. Besides, we conduct a subjective experiment with the minimal human effort by applying a probabilistic graphical model to recover ground truth labels from noisy crowdsourcing ratings. Finally, we propose a spatio-temporal distortion-aware model called STDAM for blind Internet VQA. It is composed of the graph convolution, attention, optical flow, and bi-directional LSTM modules to handle diverse contents, complex distortions and camera language conflicts. Experimental results on the self-built and public VQA databases demonstrate that our model has the capability to distinguish complicated distortions and make human-like quality estimations. We believe that the Youku-V1K database as well as the STDAM model can provide useful guidance to the screening, optimization, distribution steps of current online video sites and social media platforms.



(a) MOS: 1.42 Predicted score: 1.63

(b) MOS: 3.54 Predicted score: 3.58

(c) MOS: 4.75 Predicted score: 4.43

**Figure 13: The samples of video frames with their associate MOS values and predicted quality scores.**

**Table 3: SROCC and PLCC performance of cross database tests.**

| Train/Test | KoNViD-1k | | LIVE-VQC | | YouTube-UGC | | Youku-V1K | |
|---|---|---|---|---|---|---|---|---|
| | SROCC | PLCC | SROCC | PLCC | SROCC | PLCC | SROCC | PLCC |
| KoNViD-1k | - | - | 0.7316 | 0.7699 | 0.6750 | 0.7033 | 0.8579 | 0.8486 |
| LIVE-VQC | 0.7086 | 0.7139 | - | - | 0.6840 | 0.6691 | 0.8351 | 0.8312 |
| YouTube-UGC | 0.7476 | 0.7380 | 0.6718 | 0.7052 | - | - | 0.8625 | 0.8625 |
| Youku-V1K | 0.7193 | 0.7255 | 0.6624 | 0.6808 | 0.7358 | 0.7314 | - | - |

## ACKNOWLEDGMENTS

This work was supported in part by NSFC under Grant U1908209, 61632001, 62021001 and the National Key Research and Development Program of China 2018AAA0101400.

# REFERENCES

[1] AGH University of Science and Technology. [n. d.]. Video Quality Indicators. http://vq.kt.agh.edu.pl/metrics.html.

[2] Christos G Bampis, Zhi Li, and Alan C Bovik. 2018. Spatiotemporal feature integration and model fusion for full reference video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 8 (2018), 2256–2270.

[3] BT, RECOMMENDATION ITU-R. 2002. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union* (2002).

[4] Zhibo Chen, Wei Zhou, and Weiping Li. 2017. Blind stereoscopic video quality assessment: From depth perception to overall experience. *IEEE Transactions on Image Processing* 27, 2 (2017), 721–734.

[5] Sathya Veera Reddy Dendi and Sumohana S Channappayya. 2020. No-Reference Video Quality Assessment Using Natural Spatiotemporal Scene Statistics. *IEEE Transactions on Image Processing* 29 (2020), 5612–5624.

[6] 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.

[7] D.G.Eugene. [n. d.]. Understanding the language of the camera. https://www.thehindu.com/in-school/signpost/understanding-the-language-of-the-camera/article8580792.ece.

[8] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. 2020. Perceptual Quality Assessment of Smartphone Photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3677–3686.

[9] Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*. Springer, 363–370.

[10] Ganlu(Z5130000). [n. d.]. UGC, PGC and OGC? https://z5130000.wordpress.com/2018/06/01/ugc-pgc-and-ogc/.

[11] Deepti Ghadiyaram and Alan C Bovik. 2017. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of vision* 17, 1 (2017), 32–32.

[12] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. 2017. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 9 (2017), 2061–2077.

[13] Franz Götz-Hahn, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. 2019. No-reference video quality assessment using multi-level spatially pooled features. *arXiv preprint arXiv:1912.07966* (2019).

[14] Video Quality Experts Group et al. 2000. Final report from the video quality experts group on the validation of objective models of video quality assessment. In *VQEG meeting, Ottawa, Canada, March, 2000*.

[15] 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.

[16] Tobias Hoβfeld, Raimund Schatz, and Sebastian Egger. 2011. SOS: The MOS is not enough!. In *2011 third international workshop on quality of multimedia experience*. IEEE, 131–136.

[17] 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.

[18] P ITU-T RECOMMENDATION. 1999. Subjective video quality assessment methods for multimedia applications. *International telecommunication union* (1999).

[19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[20] Jari Korhonen. 2019. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing* 28, 12 (2019), 5923–5938.

[21] Debarati Kundu, Deepti Ghadiyaram, Alan C Bovik, and Brian L Evans. 2017. No-reference quality assessment of tone-mapped HDR pictures. *IEEE Transactions on Image Processing* 26, 6 (2017), 2957–2971.

[22] 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.

[23] Jing Li, Suiyi Ling, Junle Wang, Zhi Li, and Patrick Le Callet. 2020. A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing. In *Proceedings of the 28th ACM International Conference on Multimedia*.

[24] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. 2020. Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 214–223.

[25] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 2018. KonIQ-10K: Towards an ecologically valid and large-scale IQA database. *arXiv preprint arXiv:1803.08489* (2018).

[26] Kwan-Yee Lin and Guanxiang Wang. 2018. Hallucinated-IQA: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 732–741.

[27] Dong Liu, Rohit Puri, Nagendra Kamath, and Subhabrata Bhattacharya. 2020. Composition-Aware Image Aesthetics Assessment. In *The IEEE Winter Conference on Applications of Computer Vision*. 3569–3578.

[28] 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.

[29] Wen Lu, Ran He, Jiachen Yang, Changcheng Jia, and Xinbo Gao. 2019. A spatiotemporal model of video quality assessment via 3D gradient differencing. *Information Sciences* 478 (2019), 141–151.

[30] 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.

[31] 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.

[32] 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.

[33] Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo De Veciana. 2012. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE Journal of Selected Topics in Signal Processing* 6, 6 (2012), 652–671.

[34] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. 2016. CVD2014—A database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing* 25, 7 (2016), 3073–3086.

[35] Stéphane Péchard, Romuald Pépion, and Patrick Le Callet. 2008. Suitable methodology in subjective video quality assessment: a resolution dependent paradigm.

[36] 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.

[37] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[38] Kalpana Seshadrinathan and Alan Conrad Bovik. 2009. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE transactions on image processing* 19, 2 (2009), 335–350.

[39] K. Seshadrinathan and A. C. Bovik. 2011. Temporal hysteresis model of time varying subjective video quality. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1153–1156.

[40] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. 2010. Study of subjective and objective quality assessment of video. *IEEE transactions on Image Processing* 19, 6 (2010), 1427–1441.

[41] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[42] Zeina Sinno and Alan Conrad Bovik. 2018. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing* 28, 2 (2018), 612–627.

[43] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 2020. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3667–3676.

[44] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. 2020. UGC-VQA: Benchmarking Blind Video Quality Assessment for User Generated Content. *arXiv preprint arXiv:2005.14354* (2020).

[45] Phong V Vu and Damon M Chandler. 2014. ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging* 23, 1 (2014), 013016.

[46] 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.

[47] Zhou Wang, Ligang Lu, and Alan C Bovik. 2004. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication* 19, 2 (2004), 121–132.

[48] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[49] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. 2016. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing* 25, 9 (2016), 4444–4457.

[50] 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.

[51] 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.

[52] Junyong You, Touradj Ebrahimi, and Andrew Perkis. 2013. Attention driven foveated video quality assessment. *IEEE Transactions on Image Processing* 23, 1 (2013), 200–213.

[53] Lin Zhang, Lei Zhang, and Alan C Bovik. 2015. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing* 24, 8 (2015),

2579–2591.

[54] Yu Zhang, Xinbo Gao, Lihuo He, Wen Lu, and Ran He. 2018. Blind video quality assessment with weakly supervised learning and resampling strategy. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 8 (2018), 2244–2255.

[55] Wei Zhou and Zhibo Chen. 2020. Deep Local and Global Spatiotemporal Feature Aggregation for Blind Video Quality Assessment. *arXiv preprint arXiv:2009.03411* (2020).

[56] Wei Zhou, Zhibo Chen, and Weiping Li. 2018. Stereoscopic video quality prediction based on end-to-end dual stream deep neural networks. In *Pacific Rim Conference on Multimedia*. Springer, 482–492.

[57] Wei Zhou, Qiuping Jiang, Yuwang Wang, Zhibo Chen, and Weiping Li. 2020. Blind quality assessment for image superresolution using deep two-stream convolutional networks. *Information Sciences* (2020).

[58] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. 2020. MetaIQA: Deep Meta-learning for No-Reference Image Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14143–14152.