

MD-VQA: Multi-Dimensional Quality Assessment for UGC Live Videos

Zicheng Zhang^{1*}, Wei Wu^{2*}, Wei Sun^{1*}, Danyang Tu¹, Wei Lu¹,
Xiongkuo Min¹, Ying Chen², Guangtao Zhai^{1,3†}

¹Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University

²Alibaba Group

³MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

¹{zxc1998, sunguwei, danyangtu, SJTU-Luwei, minxiongkuo, zhaiguangtao}@sjtu.edu.cn,

²{guokui.wu, yingchen}@alibaba-inc.com

Abstract

User-generated content (UGC) live videos are often bothered by various distortions during capture procedures and thus exhibit diverse visual qualities. Such source videos are further compressed and transcoded by media server providers before being distributed to end-users. Because of the flourishing of UGC live videos, effective video quality assessment (VQA) tools are needed to monitor and perceptually optimize live streaming videos in the distributing process. In this paper, we address **UGC Live VQA** problems by constructing a first-of-a-kind subjective UGC Live VQA database and developing an effective evaluation tool. Concretely, 418 source UGC videos are collected in real live streaming scenarios and 3,762 compressed ones at different bit rates are generated for the subsequent subjective VQA experiments. Based on the built database, we develop a **Multi-Dimensional VQA (MD-VQA)** evaluator to measure the visual quality of UGC live videos from semantic, distortion, and motion aspects respectively. Extensive experimental results show that MD-VQA achieves state-of-the-art performance on both our UGC Live VQA database and existing compressed UGC VQA databases.

1. Introduction

With the rapid development of social media applications and the advancement of video shooting and processing technologies, more and more ordinary people are willing to tell their stories, share their experiences, and have their voice heard on social media or streaming media platforms such as Twitch, Tiktok, Taobao, etc. However, due to the lack of photography skills and professional equipment, the vi-

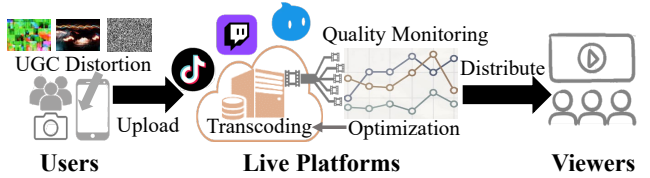


Figure 1. The distributing process of UGC live videos, where the users upload the videos degraded by the UGC distortions to the live platforms and the distorted videos are further compressed before being distributed to the viewers. The VQA models can monitor the quality changes of the compressed UGC live videos and adaptively optimize the transcoding setting.

sual quality of user-generated content (UGC) videos may be degraded by in-the-wild distortions [51]. What’s more, in common live platforms, live videos are encoded and distributed to end-users with very low delay, where compression algorithms have a significant influence on the visual quality of live videos because they can greatly reduce transmission bandwidth. As illustrated in Fig. 1, video quality assessment (VQA) tools play an important role in monitoring, optimizing, and further improving the Quality of Experience (QoE) of end-users in UGC live streaming systems.

Currently, many UGC VQA databases have been carried out [14, 21, 36, 45, 51] to address the impact of general in-the-wild distortions on video quality, while some compression VQA databases [22, 34, 40] are proposed to study the influence of compression artifacts. Then some compressed UGC VQA databases [1, 21, 46] are further constructed to solve the problem of assessing the quality of UGC videos with compression distortions. However, they are either small in scale or employ high-quality UGC videos as the sources, and all of the mentioned databases lack videos in live streaming scenes. Therefore, there is a lack of a proper **UGC Live VQA** database to develop and validate the video quality measurement tools for live streaming systems.

To address UGC Live VQA problems, we first con-

*These authors contributed equally to this work. The database is available at <https://tianchi.aliyun.com/dataset/148818?t=1679581936815>.

†Corresponding author.

Table 1. Review of common VQA databases, where 'UGC+Compression' refers to manually encoding the UGC videos with different compression settings.

Database	Year	Duration/s	Ref. Num.	Scale	Scope	Subjective Evaluating Format
CVD2014 [31]	2014	10-25	-	234	In-capture	In-lab
LIVE-Qualcomm [12]	2016	15	-	208	In-capture	In-lab
KoNViD-1k [14]	2017	8	-	1,200	In-the-wild	Crowdsourced
LIVE-VQC [36]	2018	10	-	585	In-the-wild	Crowdsourced
YouTube-UGC [45]	2019	20	-	1,500	In-the-wild	Crowdsourced
LSVQ [51]	2021	5-12	-	39,075	In-the-wild	Crowdsourced
UGC-VIDEO [21]	2019	>10	50	550	UGC + Compression	In-lab
LIVE-WC [53]	2020	10	55	275	UGC + Compression	In-lab
YT-UGC ⁺ (Subset) [46]	2021	20	189	567	UGC + Compression	In-lab
ICME2021 [1]	2021	-	1,000	8,000	UGC + Compression	In-lab
TaoLive(proposed)	2022	8	418	3,762	UGC + Compression	In-lab

struct a large-scale database named **TaoLive**, consisting 418 source UGC videos from the TaoBao [2] live streaming platform and the corresponding 3,762 compressed videos at various bit rates. Then we perform a subjective experiment in a well-controlled environment. Afterward, we propose a no-reference (NR) Multi-Dimensional VQA (**MD-VQA**) model to measure the visual quality of UGC live videos in terms of semantic, distortion, and motion aspects. The semantic features are extracted by pretrained convolutional neural network (CNN) model; the distortion features are extracted by specific handcrafted image distortion descriptors (i.e. blur, noise, block effect, exposure, and colorfulness); and the motion features are extracted from video clips by pretrained 3D-CNN models. Compared with existing UGC VQA algorithms, MD-VQA measures visual quality from multiple dimensions, and these dimensions correspond to key factors affecting live video quality, which thereby has better interpretability and performance. The contributions of this paper are summarized as below:

- **We build a large-scale UGC Live VQA database targeted at the compression artifacts on the UGC live videos.** We collect 418 raw UGC live videos that are diverse in content, distortion, and quality. Then 8 encoding settings are used, which provides 3,762 compressed UGC live videos in total.
- **We carry out a well-controlled in-lab subjective experiment.** 44 participants are invited to participate in the subjective experiment and a total of 165,528 subjective annotations are gathered.
- **A multi-dimensional NR-VQA model is proposed,** using pretrained 2D-CNN, handcrafted distortion descriptors, and pretrained 3D-CNN for the semantic, distortion, and motion features extraction respectively. The extracted features are then spatio-temporally fused to obtain the video-level quality score. The extensive experimental results validate the effectiveness of the proposed method.

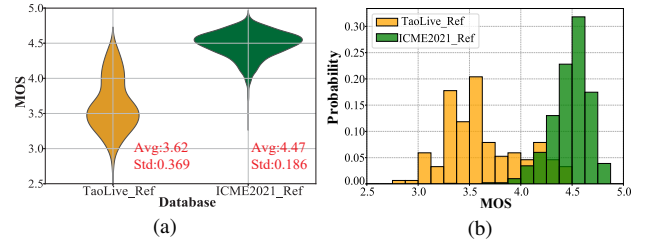


Figure 2. Comparison of the quality distribution of reference videos between the TaoLive and ICME2021 [1] databases. The source UGC videos in the ICME2021 database are centered on high-quality levels while the quality levels of the source UGC videos in the TaoLive database are more diverse.

2. Related Works

2.1. VQA Databases

During the last decade, many VQA databases [12, 14, 31, 36, 41, 42, 46, 51, 53] have been carried out to tackle the challenge of VQA problems and a review of the common VQA database is exhibited in Table 1. The early VQA databases [41, 42] usually collect limited numbers of source videos and manually introduce distortions such as compression and transmission error to generate distorted ones. Such databases are less diverse in content and distortions, which do not fit the scope of UGC videos. Then the CVD2014 [31], LIVE-Qualcomm [12], and LIVE-VQC [36] databases are formed with videos that are captured by real cameras. However, the scale of the mentioned in-capture databases is relatively small and the included distortions are much simpler. Later, UGC VQA databases such as KoNViD-1k [14], YouTube-UGC [45], and LSVQ [51] gather in-the-wild UGC videos from online platforms, which have more diverse content and distortions and have significantly promoted the development of UGC VQA tasks.

For UGC live videos, we can consider them as in-captured UGC videos followed by compressed distortions, where both in-the-wild distortions and compression distortions have a significant impact on the visual quality.

Although some works such as UGC-VIDEO [21], LIVE-WC [53], and YT-UGC+ [46] attempt to assess the quality of UGC videos caused by common compression algorithms, the relatively small size of these databases makes it difficult to support the mainstream of data-driven models, e.g. deep learning-based models in Section 2.2. The recent ICME2021 [1] database is large in scale. However, as shown in Fig. 2, source UGC videos in ICME2021 exhibit high visual quality, and thus can not reflect real source video quality distribution in live streaming systems. What's more, none of the mentioned databases includes source videos collected from practical live-streaming platforms.

2.2. UGC VQA models

Handcrafted-based: Handcrafted-based VQA models extract quality-aware features to model spatial and temporal distortions, such as natural scene statistics (NSS) features [23, 29, 33], artifacts [18, 27, 39], motion [10, 18, 39], etc. For example, VIIDEO [29] gives the intrinsic statistical regularities gathered from natural videos and assesses the video quality according to the regularities. V-BLIINDS [33] evaluates the video quality by using a spatio-temporal NSS model and a motion representation model. TLVQM [18] computes low complexity and high complexity quality-aware features in two steps to obtain the final video quality. VIDEVAL [39] carefully chooses representative quality-aware features among the mainstream NR-VQA models and regresses the representative features into quality values.

Deep learning-based: Considering the huge parameters of deep neural networks (DNN) and the relatively small scale of VQA databases, some VQA methods use pretrained DNN models for feature extraction. VSFA [20] extracts deep semantic features with a pre-trained DNN model and learns the quality-aware representation with gated recurrent units (GRUs). To enhance the understanding of video motion features, some studies [19, 37, 51] further attempt to extract motion features with 3D-CNN models pre-trained on the video action recognition databases to help detect video motion distortions and have yielded good performance. Later, some Transformer-based VQA methods are carried out. LSCT [52] first extracts frame-wise perceptual quality features and then feeds the features into a long short-term convolutional Transformer to predict the video quality. FAST/FASTER-VQA [48, 49] proposes Grid Minipatch Sampling and forms the video sampling results as fragments, which are put into a fragment-modified video swin transformer [24] for video quality representation.

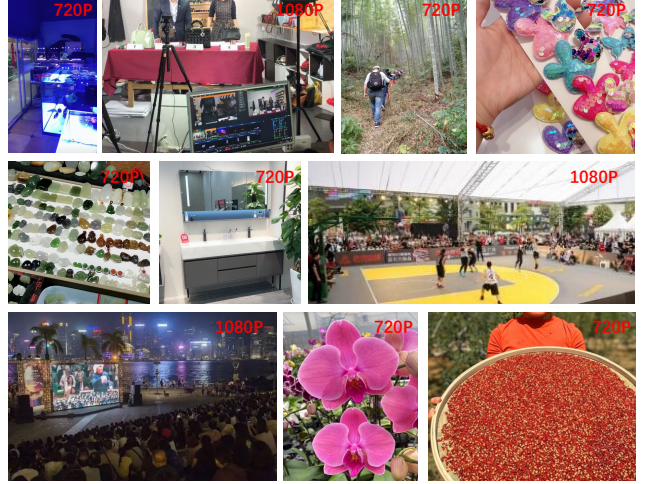


Figure 3. Sample frames of the videos in the proposed TaoLive Database, where the resolutions are marked on the top right. Additionally, the frame samples are cropped to delete some sensitive content such as human faces and watermarks for exhibition.

3. UGC Live Dataset and Human Study

3.1. Videos Collection

As illustrated in Fig. 1, users upload their live video streams to platforms, and platforms compress the video streaming at different bit rates and distribute one of them to viewers according to the Internet bandwidth or the viewers' choice. To reflect the real quality distribution of in-captured live video, we first collect large-scale uncompressed raw UGC videos from the Taolive [2] platform, a very popular live platform in China. Then, we manually select raw UGC videos that contain the scenes of technology, fashion, food, daily life, financial sales, etc, to ensure content diversity. In the last, we collect 418 raw UGC videos (110 videos have resolutions of 720P while 318 videos have resolutions of 1080P), and each raw UGC video is cropped into about 8s duration as source UGC live videos.

We use the open-source compression tools, *FFmpeg* [3], to compress source UGC live videos by 8 Constant Rate Factors (CRF) of H.265 including 16, 20, 24, 28, 32, 36, 40, and 44 to close to the distributing process of live platforms. Therefore, $3,344=418 \times 8$ compressed UGC videos are generated and a total of $3,762 = 3,344 + 418$ UGC videos are collected for evaluation, the samples of which are exhibited in Fig. 3.

3.2. Human Study

The human study is carried out in a well-controlled environment. 44 humans including 20 males and 24 females are invited to participate in the subjective experiment. The viewers are seated about 1.5 times the screen height (45cm) with normal indoor illumination and the videos are played on an iMac monitor which supports a resolution up to

4096×2304. Before the viewers start to evaluate the UGC live videos, a short introduction is given to get the viewers familiar with the equipment and quality assessment tasks. We split the experiment into 76 small sessions and each session contains 50 UGC live videos with no content overlap. The viewers participate in all the sessions individually and each session lasts about 30 minutes. There is at least 1-hour break between the sessions and each subject is allowed to attend no more than 2 sessions in a single day. During the sessions, each video is played only once and the viewers can rate the video quality from 1 to 5, with a minimum interval of 0.1. We make sure that each UGC live video is evaluated by the 44 invited viewers and 165,528=3,762×44 subjective ratings are collected in the end.

3.3. Subjective Data Analysis

According to the recommendation of ITU-R BT.500-13 [6], we compute the z-scores as the quality label of UGC live videos:

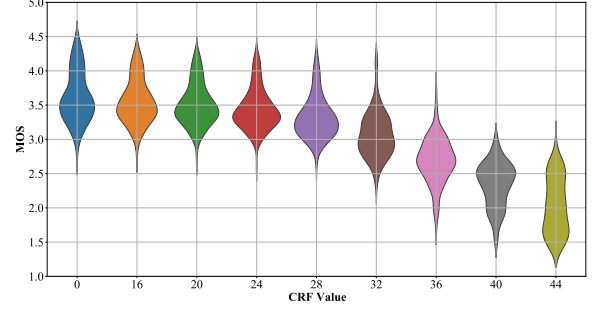
$$z_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i}, \quad (1)$$

where r_{ij} represents the quality rating given by the i -th subject on the j -th UGC live video, $\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} r_{ij}$, $\sigma_i = \sqrt{\frac{1}{N_i-1} \sum_{j=1}^{N_i} (r_{ij} - \mu_i)^2}$, and N_i is the number of UGC live videos evaluated by subject i . Then we remove the quality labels from unreliable subjects according to the recommended subject rejection procedure in [6]. Finally, the z-scores are linearly rescaled to [1, 5] and the mean opinion score (MOS) of the UGC video j is obtained by averaging the rescaled z-scores:

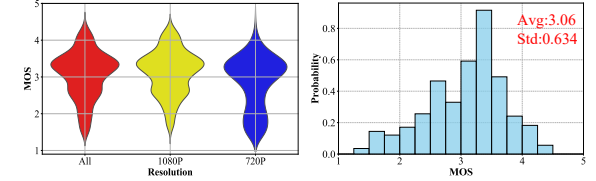
$$MOS_j = \frac{1}{M} \sum_{i=1}^M z'_{ij}, \quad (2)$$

where MOS_j represents the MOS for the j -th UGC video, M is the number of the valid subjects, and z'_{ij} are the rescaled z-scores.

We further plot the MOS distributions from the CRF and resolution perspectives. As shown in Fig. 4a, conservative CRF parameter selection (16~24) introduces slight perceptual quality loss to the UGC live videos. When CRF increases from 28 to 44, the downward trend of perceptual quality is more obvious. Moreover, when $CRF \geq 40$, nearly no compressed UGC video gains higher quality score than 3, which suggests that the 40+ CRF selection can result in a viewing experience below average. Such phenomena can provide useful guidelines for the compression strategy of live platforms. From Fig. 4b, we can find that the general quality of UGC videos with a resolution of 720P is lower than the UGC videos with a resolution of 1080P, which fits the common sense that lower resolutions lead to poorer visual quality levels.



(a) MOS distributions for different CRF values, where '0' indicates source videos.



(b) MOS distributions for different resolutions. (c) Detailed MOS distribution for the TaoLive database.

Figure 4. Illustration of the proposed TaoLive database's MOS distributions from different perspectives.

4. Proposed Method

The framework of the proposed MD-VQA model is illustrated in Fig. 5, which includes the feature extraction module, feature fusion module, and feature regression module. Specifically, quality-aware features are extracted from multiple dimensions including the semantic, distortion, and motion aspects. What's more, the feature error between adjacent frames is employed to reflect the temporal quality fluctuation. Then the obtained multi-dimensional features are fused in spatio-temporal manners and mapped to quality scores via the quality regression module.

4.1. Feature Extraction

Given a video whose number of frames and frame rate is n and r , we split the video into $\frac{n}{r}$ clips for feature extraction and each clip lasts for 1s. For each clip C_i (i represents the index of the clip), $2L$ frames are uniformly sampled for semantic and distortion feature extraction while the whole clip is employed for motion feature extraction.

4.1.1 Semantic Feature Extraction

Different semantic contents shall have diverse impacts on humans' tolerance for different distortions [20]. For example, humans are more able to endure blur distortions on flat and texture-less objects such as clear sky and smooth walls [55, 58]. However, the blur distortions can be unacceptable on objects that are rich in texture such as rough rocks and complex plants. It is also believed that semantic information

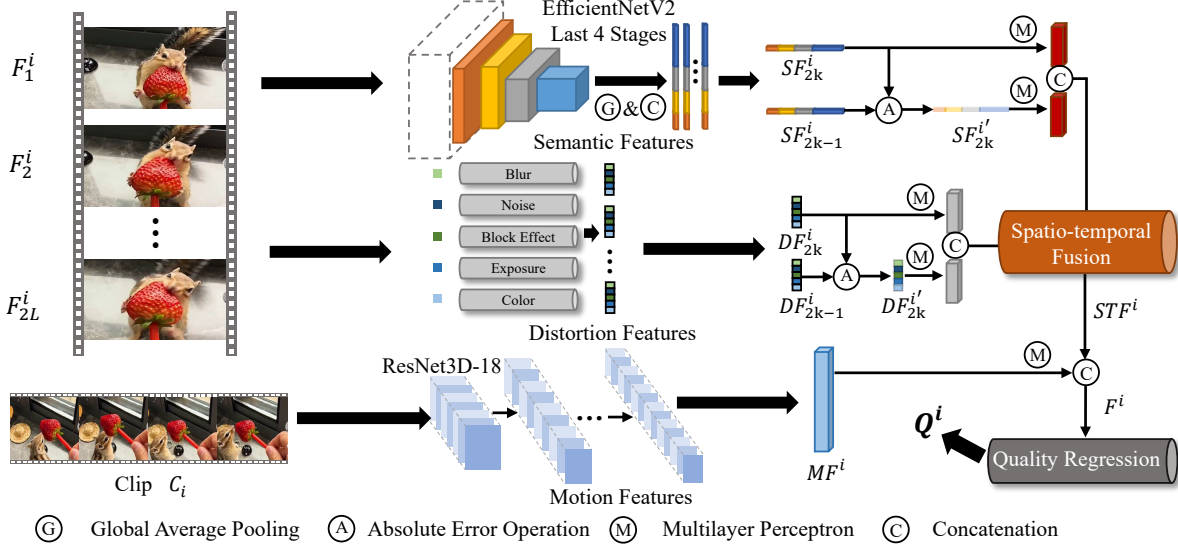


Figure 5. The framework of the proposed method, where the semantic, distortion, and motion features are extracted by the pretrained EfficientNetV2 [38], handcrafted distortion descriptors, and pretrained ResNet3D-18 [13] respectively. The absolute error between the adjacent frames' semantic and distortion features is used to reflect the temporal quality fluctuations. Finally, the multi-dimensional features are spatio-temporally fused and regressed into quality values.

can help identify the existence and extent of the perceived distortions [9]. Moreover, it has been proven that human perception is highly affected by compression [15, 50]. Thus it is quite reasonable to incorporate the semantic information into the compressed UGC video quality assessment.

Considering that visual perception is a hierarchical structure, that is, input visual information is perceived hierarchically from low-level features to high-level [26, 43, 44], we propose to utilize multi-scale features extracted from the last 4 stages of the pretrained EfficientNetV2 [38] as the frame-level semantic information:

$$SF_l^i = \alpha_1 \oplus \alpha_2 \oplus \alpha_3 \oplus \alpha_4, l \in \{1, \dots, 2L\},$$

$$\alpha_j = \text{GAP}(L_j(F_l^i)), j \in \{1, 2, 3, 4\}, \quad (3)$$

where SF_l^i indicates the extracted semantic features from the l -th sampled frame F_l^i of clip C_i , $\oplus(\cdot)$ stands for the concatenation operation, $\text{GAP}(\cdot)$ represents the global average pooling operation, $L_j(F_l^i)$ stands for the feature maps obtained from j -th last layer of EfficientNetV2, and α_j denotes the average pooled features from $L_j(F_l^i)$.

4.1.2 Distortion Feature Extraction

Various types of distortions exist in UGC videos and only utilizing semantic information is insufficient to model the distortion perception of UGC videos. What's more, the original UGC distortions can exhibit dissimilar quality representations under different levels of compression. For example, the blur is less sensitive to compression [56, 57] since compression usually wipes out high-frequency information,

and noise can be eased or even disappear when higher compression levels are applied [5]. Therefore, to better improve the quality representation of the proposed method, some handcrafted distortion descriptors are further employed for quality prediction, which include blur [54], noise [7], block effect [47], exposure [18], and colorfulness [32]. Then the frame-level distortion features can be derived as:

$$DF_l^i = \Psi(F_l^i), l \in \{1, \dots, 2L\}, \quad (4)$$

where DF_l^i represents the extracted distortion features from the l -th sampled frame F_l^i of clip C_i and $\Psi(\cdot)$ stands for the distortion feature extraction process.

4.1.3 Motion Feature Extraction

UGC live videos are often bothered with motion distortions resulting from the unstable shooting environment as well as the restricted bit rates. However, these motion distortions such as the video shaking and the motion blur are difficult to recognize from the spatial features alone. Furthermore, video compression deeply depends on motion estimation [11, 25], which indicates that motion distortions can influence the quality of video compression. Therefore, to help the model better understand the motion information, we propose to use the pretrained 3D-CNN backbone, ResNet3D-18 [13], to capture clip-level motion distortions:

$$MF^i = \Gamma(C_i), \quad (5)$$

where MF^i denotes the motion features extracted from clip C_i and $\Gamma(\cdot)$ represents the motion feature extraction operation.

To sum up, given the i -th clip C_i of the video, we can obtain the clip-level semantic features $SF^i \in \mathbb{R}^{2L \times N_S}$, the distortion features $DF^i \in \mathbb{R}^{2L \times N_D}$, and the motion features $MF^i \in \mathbb{R}^{1 \times N_M}$, where N_S , N_D , and N_M represent the number of channels for the semantic, distortion, and motion features respectively.

4.2. Feature Fusion

It has been proven in [30] that videos with better quality tend to have smaller quality fluctuations while videos with lower quality tend to have larger quality fluctuations. Therefore, to quantify the fluctuations that are highly correlated with human perception, we propose to employ the absolute error between adjacent semantic and distortion features for temporal quality fluctuations reflection:

$$\begin{aligned} SF_{2k}^{i'} &= |SF_{2k}^i - SF_{2k-1}^i|, k \in \{1, \dots, L\}, \\ DF_{2k}^{i'} &= |DF_{2k}^i - DF_{2k-1}^i|, k \in \{1, \dots, L\}, \end{aligned} \quad (6)$$

where $SF_{2k}^{i'}$ and $DF_{2k}^{i'}$ represent the absolute error between adjacent semantic and distortion features. Then the spatio-temporal fusion can be derived as:

$$\begin{aligned} SD_{2k}^i &= \omega(SF_{2k}^i) \oplus \omega(DF_{2k}^i) \oplus \omega(SF_{2k}^{i'}) \oplus \omega(DF_{2k}^{i'}), \\ STF^i &= W_L^1 (SD^{i^T}), \end{aligned} \quad (7)$$

where $\oplus(\cdot)$ stands for the concatenation operation, $\omega(\cdot)$ represents the learnable Multilayer Perceptron (MLP), $SD_{2k}^i \in \mathbb{R}^{1 \times N_{SD}}$ indicates the frame-level spatial-fused features obtained from semantic and distortion features, $SD^{i^T} \in \mathbb{R}^{N_{SD} \times L}$ is the transposition result of the clip-level semantic and distortion features $SD^i \in \mathbb{R}^{L \times N_{SD}}$, W_L^1 is a learnable linear mapping operation to fuse the SD^{i^T} in the temporal domain, and we finally obtain the spatio-temporal fused features $STF^i \in \mathbb{R}^{N_{SD} \times 1}$. To further introduce the quality-aware motion features, we concatenate the spatio-temporal features with the motion features:

$$F^i = STF^{i^T} \oplus \omega(MF^i), \quad (8)$$

where $STF^{i^T} \in \mathbb{R}^{1 \times N_{SD}}$, the final clip-level quality-aware representation $F^i \in \mathbb{R}^{1 \times (N_{SD} + N'_M)}$ and N'_M is adjusted number of channels for the motion features after MLP operation.

4.3. Feature Regression

After the feature extraction process described above, we use the three-stage fully-connected layers to regress the clip-level quality-aware representation F^i into quality values:

$$Q^i = \mathbf{FC}(F^i), \quad (9)$$

where $\mathbf{FC}(\cdot)$ indicates the fully-connected layers and Q_i stands for the quality value of clip C_i . Consequently, the

overall UGC live video quality can be obtained via average pooling:

$$Q = 1/n \sum_r^n Q^i, \quad (10)$$

where Q is the video quality value and $\frac{n}{r}$ represents the number of clips. We simply use the Mean Squared Error (MSE) as the loss function:

$$Loss = \frac{1}{n} \sum_{m=1}^n (Q'_m - Q_m)^2 \quad (11)$$

where n indicates the number of videos in a mini-batch, Q'_m and Q_m are the subjective quality labels and predicted quality levels respectively.

5. Experiment

In this section, we first give the details of the experimental setup. Then we validate the proposed MD-VQA model with other mainstream VQA models on the proposed TaoLive database and the other two UGC compression VQA models. The ablation study and cross database validation are conducted to investigate the contributions of different groups of features and the generalization ability of the VQA models. Finally, we test the proposed MD-VQA model on two in-the-wild UGC VQA databases.

5.1. Benchmark Databases

The proposed TaoLive database and two compressed UGC VQA databases including the LIVE-WC [53] and YT-UGC+ [46] databases are selected as the benchmark databases. For all the databases, we follow the common practice and split the databases with an 80%-20% train-test ratio. Additionally, all the databases are validated separately. To fully evaluate the stabilization and performance of the VQA models, the split is randomly conducted 30 times and the average results are recorded as the final performance.

5.2. Implementation Details

The EfficientNetV2 [38] backbone is fixed with the EfficientNetV2-S weights pretrained on the ImageNet database [8] for semantic feature extraction while the ResNet3D-18 [13] is fixed with the weights pretrained on the Kinetics-400 [16] database. All the frames are maintained with the original resolution for the semantic, distortion, and motion feature extraction. The Adam optimizer [17] is employed with the initial learning rate set as 0.001. If the training loss has not decreased for 5 epochs, the learning rate will be reduced to half. The default number of epochs is set as 50. The parameter L described in Section 4.1 is set as 8, which means 16 frames are uniformly sampled for the semantic and distortion feature extraction for a single clip.

Table 2. Experimental performance on the compressed UGC VQA databases. ‘Hand’ indicates using handcrafted-based features while ‘Deep’ indicates using deep learning-based features. Best in **red** and second in **blue**.

Method	Hand	Deep	LIVE-WC		YT-UGC ⁺		TaoLive	
			SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow
BRISQUE (TIP, 2012) [28]	✓	×	0.787	0.788	0.303	0.309	0.771	0.777
TLVQM (TIP, 2019) [18]	✓	×	0.838	0.830	0.672	0.697	0.862	0.869
VIDEVAL (TIP, 2021) [39]	✓	×	0.812	0.825	0.660	0.662	0.914	0.910
VSFA (ACM MM, 2019) [20]	×	✓	0.856	0.857	0.784	0.783	0.920	0.917
PVQ (CVPR, 2021) [51]	×	✓	0.901	0.909	0.775	0.776	0.916	0.919
BVQA (TCSVT, 2022) [19]	×	✓	0.912	0.916	0.777	0.781	0.926	0.922
SimpleVQA (ACM MM, 2022) [37]	×	✓	0.927	0.920	0.789	0.784	0.932	0.926
MD-VQA(Ours)	✓	✓	0.931	0.937	0.822	0.828	0.942	0.945

Table 3. Experimental performance of the ablation study, where SF, DF, and MF indicate the semantic features, distortion features, and motion features respectively.

Feature	LIVE-WC		YT-UGC ⁺		TaoLive	
	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow
SF	0.911	0.910	0.785	0.787	0.931	0.934
DF	0.640	0.658	0.277	0.381	0.603	0.638
MF	0.841	0.858	0.537	0.561	0.909	0.912
SF+DF	0.925	0.922	0.805	0.824	0.935	0.936
SF+MF	0.921	0.924	0.792	0.789	0.940	0.941
DF+MF	0.857	0.861	0.573	0.631	0.925	0.926
All	0.931	0.937	0.822	0.828	0.942	0.945

Table 4. Ablation study results for absolute error (ABS) and feature fusion module (FFM), where ABS is replaced with error and FFM is replaced with concatenation.

Model	LIVE-WC		YT-UGC ⁺		TaoLive	
	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow
w/o ABS	0.912	0.916	0.814	0.814	0.925	0.920
w/o FFM	0.913	0.915	0.811	0.817	0.921	0.928
All	0.931	0.937	0.822	0.828	0.942	0.945

5.3. Competitors & Criteria

To fully evaluate the performance of the proposed method, we select several popular quality assessment models for comparison, which include BRISQUE [28], VSFA [20], TLVQM [18], VIDEVAL [39], PVQ [51], BVQA [19] and SimpleVQA [37]. It’s worth mentioning that BRISQUE belongs to NR-IQA models and we obtain the video quality features by averaging the features extracted from each frame with BRISQUE. The other VQA models are trained with the default parameter setting defined by their authors.

Two criteria are adopted to evaluate the performance of the quality assessment models, which include the Spearman Rank Order Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC). Before calculating the criteria values, a four-parameter logistic regression function [35] is utilized to fit the predicted scores to

the scale of MOSs. The value range for SRCC and PLCC is [0,1] and better models should yield higher SRCC and PLCC values.

5.4. Performance Discussion

The experimental performance on the three compressed UGC VQA databases is shown in Table 2, from which we can draw several conclusions. (a) The proposed MD-VQA achieves first place and surpasses the second place (SimpleVQA [37]) by about 0.004, 0.033, and 0.010 in terms of SRCC values on the LIVE-WC, YT-UGC⁺, and TaoLive databases respectively, which demonstrates its effectiveness of predicting the quality levels of compressed UGC videos. (b) The handcrafted-based methods (BRISQUE, TLVQM, and VIDEVAL) are significantly inferior to the deep learning-based methods (VSFA, PVQ, BVQA, SimpleVQA, and MD-VQA). It can be explained that the handcrafted based methods hold the prior experience of NSS, which comes from the pristine videos. However, the characteristics of compressed UGC videos are far more complicated and do not suit the prior knowledge of natural regularities. (c) All the VQA methods experience performance drops on the YT-UGC⁺ database compared with the other two databases. The YT-UGC⁺ database uses the recommended VP9 settings and target bit rates [4] for compression while the LIVE-WC and TaoLive databases control the compression by varying the CRF parameters of H.264 and H.265 respectively. It might be because the recommended VP9 compression settings do not monotonically reduce the video bit rates, thus being more challenging for quality prediction.

5.5. Ablation Study

To investigate the contributions of different features employed in MD-VQA, we conduct the ablation study in this section. The experimental results for employing different types of features are shown in Table 3. Combining features yield better performance than using a single group of features and employing all features leads to the best perfor-

mance among the combinations of different features, which confirms the contributions of the semantic, distortion, and motion features. Additionally, by comparing the performance of SF, DF, and MF models, we can see that the SF model achieves first place on all the databases, which indicates that the semantic features make the most devotion to the final performance. What’s more, the distortion and motion features achieve only 0.277 and 0.537 in terms of SRCC values on the YT-UGC⁺ database. This is because 1/3 of the YT-UGC⁺ database’s compressed UGC videos are obtained from game videos. The game videos’ distortion and motion characteristics differ greatly from the natural videos. To illustrate, noise usually does not exist in game videos and the motion blur effect is manually introduced if applied. Such phenomena result in the relatively low performance of the distortion and motion features. We also conduct the ablation study for using absolute error (ABS) and feature fusion module (FFM) and the results are listed in Tab 4, from which we can find that both ABS and FFM make contributions to the final results.

5.6. Cross Database Performance

Since UGC videos are diverse in contents and distortions, We carry out the cross database validation to test the generalization ability of the VQA models in this section. The VQA models are trained on the TaoLive database and tested on the other two compressed UGC VQA databases. The experimental results are listed in Table 5. The proposed MD-VQA model has surpassed all the compared VQA models on both LIVE-WC and YT-UGC⁺ databases, which proves its strong generalization ability. The handcrafted-based VQA models (BRISQUE, TLVQM, and VIDEVAL) perform badly on the YT-UGC⁺ database and the deep learning-based methods also undergo significant performance drops from the LIVE-WC database to the YT-UGC⁺ database. The reason is that the YT-UGC⁺ database employs VP9 [4] while the TaoLive database utilizes H.265 for compression respectively. The different compression standards can bring unignorable gaps between the data distributions. Therefore the quality representation learned from the TaoLive database is less effective on the YT-UGC⁺ database but works well on the H.264 compressing LIVE-WC database.

5.7. In-the-wild Performance

Although the proposed MD-VQA model focuses on the compressed UGC VQA issues, we also test its performance on some mainstream in-the-wild UGC VQA databases, which includes the KoNViD-1k [14] and LIVE-VQC [36] databases. These databases are not focused on compression and contain a wide range of distortions. Validation on these databases can reflect the VQA models’ ability to handle general UGC VQA issues. The experimental results are

Table 5. Experimental performance of cross databases, where the VQA models are all pretrained on the TaoLive database.

Method	LIVE-WC		YT-UGC ⁺	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑
BRISQUE [28]	0.708	0.709	0.026	0.059
TLVQM [18]	0.562	0.583	0.155	0.184
VIDEVAL [39]	0.557	0.583	0.077	0.132
VSFA [20]	0.701	0.698	0.357	0.399
SimpleVQA [37]	0.711	0.723	0.388	0.394
MD-VQA	0.742	0.728	0.440	0.448

Table 6. Experimental performance of in-the-wild UGC databases.

Method	KoNViD-1k		LIVE-VQC	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑
BRISQUE [28]	0.657	0.658	0.593	0.638
TLVQM [18]	0.773	0.769	0.799	0.803
VIDEVAL [39]	0.783	0.780	0.752	0.751
VSFA [20]	0.785	0.797	0.716	0.775
SimpleVQA [37]	0.856	0.860	0.811	0.815
MD-VQA	0.851	0.853	0.814	0.839

exhibited in Table 6, from which we can find that the proposed MD-VQA outperforms the compared VQA models on the LIVE-VQC databases and gets the second-ranking on the KoNViD-1k database. This implies that the proposed MD-VQA remains competitive not only for compression-specific VQA issues and can be taken as a strong baseline for in-the-wild UGC VQA tasks as well.

6. Conclusion

In this paper, we focus on the compressed UGC VQA issues. To meet the practical needs of live platforms, we carry out a large-scale compressed UGC VQA database called **TaoLive**. Unlike the common compression VQA databases that employ high-quality videos as source videos, the TaoLive database collects the 418 source UGC videos that cover a wide quality range and generate the compressed videos by varying the CRF parameters with H.265. A well-controlled subjective experiment is conducted to gather the quality labels for the compressed UGC videos. Further, we propose a VQA model (MD-VQA) to assess the quality of the compressed UGC videos from the semantic, distortion, and motion dimensions. The extensive experimental results confirm the effectiveness of the proposed method.

7. Acknowledgement

This work was supported in part by NSFC (No.62225112, No.61831015), the Fundamental Research Funds for the Central Universities, National Key R&D Program of China 2021YFE0206700, and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

- [1] Wang, Haiqiang and Li, Gary and Liu, Shan and Kuo, C.-C. Jay, “ICME 2021 UGC-VQA Challenge.”, [Online] Available: <http://ugcvqa.com/>. 1, 2, 3
- [2] Taobao Alibaba, Inc., “TaoLive.” [Online] Available: <https://taolive.taobao.com>. 2, 3
- [3] FFmpeg Group, “FFmpeg” [Online] Available: <https://ffmpeg.org/>. 3
- [4] Google, Inc., “VP9 encoding recommendations.” [Online] Available: <https://developers.google.com/media/vp9>. 7, 8
- [5] Osama K Al-Shaykh and Russell M Mersereau. Lossy compression of noisy images. *IEEE TIP*, 7(12):1641–1652, 1998. 5
- [6] RECOMMENDATION ITU-R BT. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union*, 2002. 4
- [7] Guangyong Chen, Fengyuan Zhu, and Pheng Ann Heng. An efficient statistical method for image noise level estimation. In *IEEE/CVF CVPR*, pages 477–485, 2015. 5
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF CVPR*, pages 248–255, 2009. 6
- [9] Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *IEEE QoMEX*, pages 1–6, 2016. 5
- [10] Joshua Peter Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C Bovik. Chipqa: No-reference video quality prediction via space-time chips. *IEEE TIP*, 30:8059–8074, 2021. 3
- [11] Borko Furht, Joshua Greenberg, and Raymond Westwater. *Motion estimation algorithms for video compression*, volume 379. Springer Science & Business Media, 2012. 5
- [12] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE TCSVT*, 28(9):2061–2077, 2017. 2
- [13] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *IEEE/CVF CVPR*, pages 6546–6555, 2018. 5, 6
- [14] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *IEEE QoMEX*, pages 1–6, 2017. 1, 2, 8
- [15] Nikil Jayant, James Johnston, and Robert Safranek. Signal compression based on models of human perception. *Proceedings of the IEEE*, 81(10):1385–1422, 1993. 5
- [16] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6
- [17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6
- [18] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE TIP*, 28(12):5923–5938, 2019. 3, 5, 7, 8
- [19] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE TCSVT*, 2022. 3, 7
- [20] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM MM*, pages 2351–2359, 2019. 3, 4, 7, 8
- [21] Yang Li, Shengbin Meng, Xinfeng Zhang, Shiqi Wang, Yue Wang, and Siwei Ma. Ugc-video: perceptual quality assessment of user-generated videos. In *IEEE MIPR*, pages 35–38, 2020. 1, 2, 3
- [22] Zhuoran Li, Zhengfang Duanmu, Wentao Liu, and Zhou Wang. Avc, hevc, vp9, avs2 or av1?—a comparative study of state-of-the-art video encoders on 4k videos. In *International Conference on Image Analysis and Recognition*, pages 162–173, 2019. 1
- [23] Liang Liao, Kangmin Xu, Haoning Wu, Chaofeng Chen, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring the effectiveness of video perceptual representation in blind video quality assessment. In *ACM MM*, pages 837–846, 2022. 3
- [24] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE/CVF CVPR*, pages 3202–3211, 2022. 3
- [25] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *IEEE/CVF CVPR*, pages 11006–11015, 2019. 5
- [26] Wei Lu, Wei Sun, Wenhan Zhu, Xiongkuo Min, Zicheng Zhang, Tao Wang, and Guangtao Zhai. A cnn-based quality assessment method for pseudo 4k contents. In *IFTC*, pages 164–176. Springer, 2022. 5
- [27] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. St-greed: Space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE TIP*, 30:7446–7457, 2021. 3
- [28] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12):4695–4708, 2012. 7, 8
- [29] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE TIP*, 25(1):289–300, 2015. 3
- [30] Manish Narwaria, Weisi Lin, and Anmin Liu. Low-complexity video quality assessment using temporal quality variations. *IEEE TMM*, 14(3):525–535, 2012. 6
- [31] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. Cvd2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE TIP*, 25(7):3073–3086, 2016. 2
- [32] Karen Panetta, Chen Gao, and Sos Agaian. No reference color image contrast and quality measures. *IEEE transactions on Consumer Electronics*, 59(3):643–651, 2013. 5
- [33] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE TIP*, 23(3):1352–1365, 2014. 3

- [34] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE TIP*, 19(6):1427–1441, 2010. 1
- [35] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE TIP*, 15(11):3440–3451, 2006. 7
- [36] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE TIP*, 28(2):612–627, 2018. 1, 2, 8
- [37] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *ACM MM*, 2022. 3, 7, 8
- [38] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021. 5, 6
- [39] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE TIP*, 30:4449–4464, 2021. 3, 7, 8
- [40] Phong V Vu and Damon M Chandler. Vis3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging*, 23(1):013016, 2014. 1
- [41] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset. In *IEEE ICIP*, pages 1509–1513, 2016. 2
- [42] Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeonghoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, et al. Videoset: A large-scale compressed video quality dataset based on jnd measurement. *Journal of Visual Communication and Image Representation*, 46:292–302, 2017. 2
- [43] Tao Wang, Wei Sun, Xiongkuo Min, Wei Lu, Zicheng Zhang, and Guangtao Zhai. A multi-dimensional aesthetic quality assessment model for mobile game images. In *IEEE VCIP*, pages 1–5. IEEE, 2021. 5
- [44] Tao Wang, Zicheng Zhang, Wei Sun, Xiongkuo Min, Wei Luand, and Guangtao Zhai. Subjective quality assessment for images generated by computer graphics. In *IEEE MMSP*, pages 1–5. IEEE, 2022. 5
- [45] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *IEEE International Workshop on Multimedia Signal Processing*, pages 1–5, 2019. 1, 2
- [46] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of ugc videos. In *IEEE/CVF CVPR*, pages 13435–13444, 2021. 1, 2, 3, 6
- [47] Zhou Wang, Hamid R Sheikh, and Alan C Bovik. No-reference perceptual quality assessment of jpeg compressed images. In *IEEE ICIP*, 2002. 5
- [48] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. 2022. 3
- [49] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *arXiv preprint arXiv:2210.05357*, 2022. 3
- [50] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):5–26, 2020. 5
- [51] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq: patching up the video quality problem. In *IEEE/CVF CVPR*, pages 14019–14029, 2021. 1, 2, 3, 7
- [52] Junyong You. Long short-term convolutional transformer for no-reference video quality assessment. In *ACM MM*, pages 2112–2120, 2021. 3
- [53] Xiangxu Yu, Neil Birkbeck, Yilin Wang, Christos G Bampis, Balu Adsumilli, and Alan C Bovik. Predicting the quality of compressed videos with pre-existing distortions. *IEEE TIP*, 30:7511–7526, 2021. 2, 3, 6
- [54] Yibing Zhan and Rong Zhang. No-reference image sharpness assessment based on maximum gradient and variability of gradients. *IEEE TMM*, 20(7):1796–1808, 2017. 5
- [55] Zicheng Zhang, Wei Sun, Xiongkuo Min, Tao Wang, Wei Lu, and Guangtao Zhai. A full-reference quality assessment metric for fine-grained compressed images. In *IEEE VCIP*, pages 1–4. IEEE, 2021. 4
- [56] Zicheng Zhang, Wei Sun, Xiongkuo Min, Wenhan Zhu, Tao Wang, Wei Lu, and Guangtao Zhai. A no-reference evaluation metric for low-light image enhancement. In *IEEE ICME*, pages 1–6. IEEE, 2021. 5
- [57] Zicheng Zhang, Wei Sun, Xiongkuo Min, Wenhan Zhu, Tao Wang, Wei Lu, and Guangtao Zhai. A no-reference deep learning quality assessment method for super-resolution images based on frequency maps. In *IEEE ISCAS*, pages 3170–3174. IEEE, 2022. 5
- [58] Zicheng Zhang, Wei Wu, Ying Cheng, Xiongkuo Min, Guangtao Zhai, et al. Perceptual quality assessment for fine-grained compressed images. *VCIR*, 90:103696, 2023. 4