

Light-VQA: A Multi-Dimensional Quality Assessment Model for Low-Light Video Enhancement

Yunlong Dong

Shanghai Jiao Tong University
Shanghai, China
dongyangunlong@sjtu.edu.cn

Xiaohong Liu*

Shanghai Jiao Tong University
Shanghai, China
xiaohongliu@sjtu.edu.cn

Yixuan Gao

Shanghai Jiao Tong University
Shanghai, China
gaoyixuan@sjtu.edu.cn

Xunchu Zhou

Shanghai Jiao Tong University
Shanghai, China
zhou_xc@sjtu.edu.cn

Tao Tan

Macao Polytechnic University
Macao, China
taotanjs@gmail.com

Guangtao Zhai

Shanghai Jiao Tong University
Shanghai, China
zhaiguangtao@sjtu.edu.cn

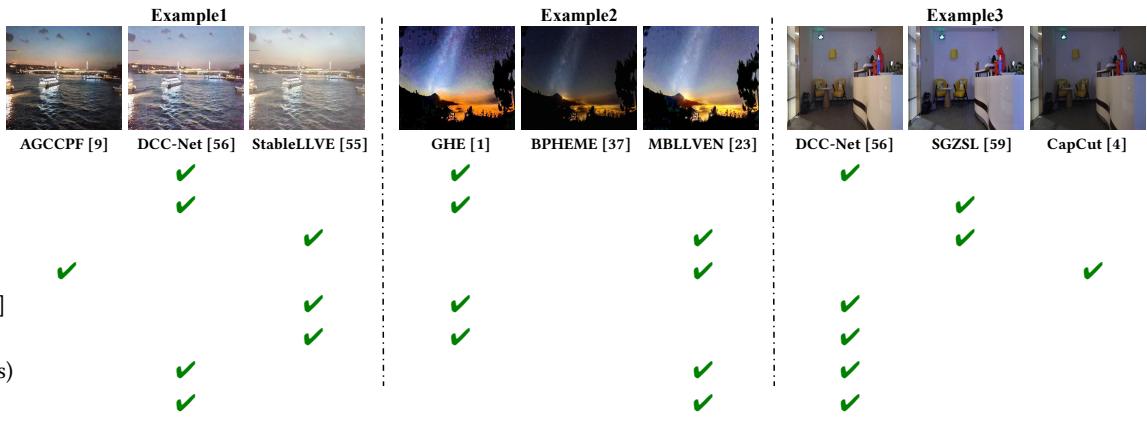


Figure 1: Which video has the best visual perceptual quality in each example listed? The above 9 figures are representative frames of sample enhanced videos obtained by applying different enhancement algorithms to corresponding original low-light videos. The concrete algorithms are listed below the figures. Then we use 6 state-of-the-art VQA models (V-BLIINDS [27], TLVQM [12], VIDEVAL [35], RAPIQUE [36], Simple-VQA [33], and FAST-VQA [42]) and the proposed Light-VQA to predict the quality of these enhanced videos. The check marks represent the enhanced video with the best perceptual quality predicted by each model. Mean Opinion Scores (MOSs), the ground-truth perceptual quality of enhanced videos, are obtained through a subjective experiment. It is evident that the prediction results of Light-VQA are highly consistent with human perception as compared to others. More detailed qualitative results can be found in Supplementary.

ABSTRACT

Recently, Users Generated Content (UGC) videos becomes ubiquitous in our daily lives. However, due to the limitations of photographic equipments and techniques, UGC videos often contain various degradations, in which one of the most visually unfavorable effects is the underexposure. Therefore, corresponding video enhancement algorithms such as Low-Light Video Enhancement (LLVE) have been proposed to deal with the specific degradation.

*Corresponding authors.

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 the author(s) 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 '23, October 29–November 3, 2023, Ottawa, ON, Canada.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0108-5/23/10...\$15.00

<https://doi.org/10.1145/3581783.3611923>

However, different from video enhancement algorithms, almost all existing Video Quality Assessment (VQA) models are built generally rather than specifically, which measure the quality of a video from a comprehensive perspective. To the best of our knowledge, there is no VQA model specially designed for videos enhanced by LLVE algorithms. To this end, we first construct a Low-Light Video Enhancement Quality Assessment (LLVE-QA) dataset in which 254 original low-light videos are collected and then enhanced by leveraging 8 LLVE algorithms to obtain 2,060 videos in total. Moreover, we propose a quality assessment model specialized in LLVE, named Light-VQA. More concretely, since the brightness and noise have the most impact on low-light enhanced VQA, we handcraft corresponding features and integrate them with deep-learning-based semantic features as the overall spatial information. As for temporal information, in addition to deep-learning-based motion features, we also investigate the handcrafted brightness consistency among video frames, and the overall temporal information is their concatenation. Subsequently, spatial and temporal information is

fused to obtain the quality-aware representation of a video. Extensive experimental results show that our Light-VQA achieves the best performance against the current State-Of-The-Art (SOTA) on LLVE-QA and public dataset. *Dataset and Codes can be found at <https://github.com/wenzhouyidu/Light-VQA>.*

CCS CONCEPTS

- Computing methodologies → Modeling and simulation.

KEYWORDS

Video quality assessment; low-light video enhancement; LLVE-QA dataset; spatial and temporal information fusion

ACM Reference Format:

Yunlong Dong, Xiaohong Liu*, Yixuan Gao, Xunchu Zhou, Tao Tan, and Guangtao Zhai. 2023. Light-VQA: A Multi-Dimensional Quality Assessment Model for Low-Light Video Enhancement. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3581783.3611923>

1 INTRODUCTION

Compared to text and images also spreading widely [16, 57], videos are generally more entertaining and informative. However, due to the influence in photographic devices and skills, the quality of UGC videos often varies greatly. It is frustrating that the precious and memorable moment is degraded by photographic limitations (e.g., underexposure, low frame-rate, and low resolution). To address the problems mentioned above, specific video enhancement algorithms have been proposed, such as Low-Light Video Enhancement (LLVE) [9, 37, 56, 59], video frame interpolation [3, 25, 29, 30], and super-resolution reconstruction [19–21, 28, 51]. In this paper, we focus on the quality assessment of enhanced low-light videos. Low-light videos are often captured in the low- or back-lighting environments and suffer from significant degradations such as low visibility and noises [54]. Such degraded videos will challenge many computer vision downstream tasks [60] such as object detection, semantic segmentation, etc., which are usually resorted to videos with good quality. Therefore, many LLVE algorithms have been developed to improve the visual quality of low-light videos. To this end, one straightforward way is to split the video into frames and apply Low-Light Image Enhancement (LLIE) algorithms to enhance each frame of this video. Representative traditional LLIE algorithms include AGCCPF [9], GHE [1], and BPHEME [37]. There are also some deep-learning-based LLIE algorithms, such as MBLLEN [23], SGZSL [59], and DCC-Net [56]. However, applying LLIE algorithms directly to videos can lead to severe temporal instability. In order to fill the niche existing in LLIE, some LLVE algorithms that take temporal consistency into account are proposed, such as MBLL-VEN [23], SDSD [38], SMID [5], and StableLLVE [55].

Video Quality Assessment (VQA) is of great significance to facilitate the development of LLVE algorithms. Objective VQA can be divided into Full-Reference (FR) VQA [2], Reduced-Reference (RR) VQA [32], and No-Reference (NR) VQA [27] contingent on the amount of required pristine video information [33]. Due to the difficulty in obtaining reference videos, NR-VQA has attracted

a large number of researchers' attention. In the early development stages of NR-VQA, researchers often evaluate video quality based on handcrafted features [12, 13, 24, 27, 34–36], such as structure, texture, and statistical features. Recently, owing to the potential in practical applications, deep learning based NR-VQA models [15, 17, 33, 40, 42, 49, 52] have progressively dominated the VQA field. However, most existing VQA models are designed for general purpose. To the best of our knowledge, few models specifically evaluate the quality of videos enhanced by LLVE algorithms. One possible reason is the lack of corresponding datasets.

Therefore, in this paper, we elaborately build a Low-Light Video Enhancement Quality Assessment (LLVE-QA) dataset to facilitate the work on evaluating the performance of LLVE algorithms. Different from general datasets which commonly consist of original UGC videos with various degradations, LLVE-QA dataset contains 254 original low-light videos and 1,806 enhanced videos from representative enhancement algorithms, each with a corresponding MOS. Subsequently, we propose a quality assessment model specialized for low-light video enhancement, named Light-VQA. The framework of Light-VQA is illustrated in Figure 2. Considering that among low-level features, brightness and noise have the most impact on low-light enhanced VQA [54], in addition to semantic features and motion features extracted from deep neural network, we specially handcraft the brightness, brightness consistency, and noise features to improve the ability of the model to represent the quality-aware features of low-light enhanced videos. Extensive experiments validate the effectiveness of our network design.

The contributions of this paper are summarized as follows:

- (1) By leveraging representative LLVE algorithms on the collected videos with diverse content and various degrees of brightness, we conduct a subjective experiment to build a low-light video enhancement dataset, named LLVE-QA.
- (2) Benefiting from the built dataset, we propose a novel quality assessment model named Light-VQA *specifically designed* for low-light enhanced videos that integrates the luminance-sensitive handcrafted features into deep-learning-based features in both spatial and temporal information, which is then fused to obtain the quality-aware representation.
- (3) The proposed Light-VQA achieves the best performance as compared to 6 SOTA models on LLVE-QA and public dataset. We envision that the Light-VQA is promising to be a fundamental tool to assess the LLVE algorithms.

2 RELATED WORK

2.1 Low-Light Enhancement

To enhance the low-light videos, it is straightforward to split the low-light video into frames, so as to take advantage of existing LLIE algorithms. AGCCPF [9] enhances the brightness and contrast of low-light images using the gamma correction and weighted probability distribution of pixels. GHE [1] applies a transformation on image histogram to redistribute the pixel intensity, resulting in a more favorable visual result. BPHEME [37] enhances the low-light video by balancing the brightness preserving histogram with maximum entropy. In addition to the above traditional algorithms, low-light image enhancement algorithms based on deep learning [23, 56, 59]

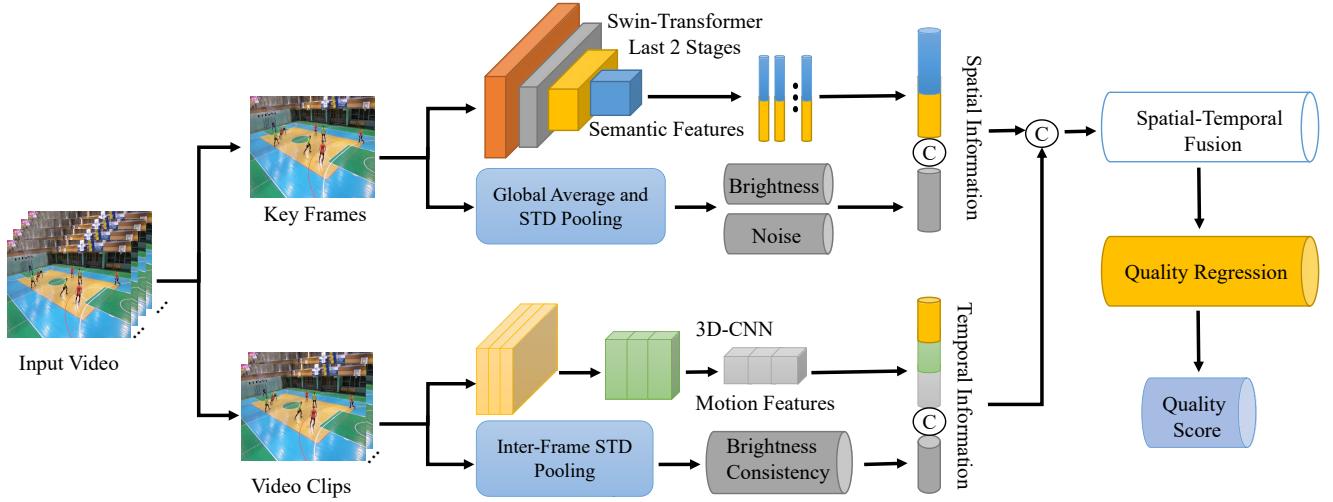


Figure 2: Framework of Light-VQA. The model contains the spatial and temporal information extraction module, the feature fusion module, and the quality regression module. Concretely, spatial information contains semantic features, brightness, and noise. Temporal information contains motion features and brightness consistency.

are rapidly emerging. Zhang *et al.* [56] propose a consistent network to improve illumination and preserve color consistency of low-light images. However, applying LLIE algorithms directly to videos can lead to temporal consistency problems such as motion artifacts and brightness consistency, which will ultimately reduce the quality of videos.

In order to maintain the temporal consistency of videos better, specific LLVE algorithms [5, 23, 38, 55] are proposed. MBLL-VEN [23] processes low-light videos via 3D convolution to extract temporal information and preserve temporal consistency. Wang *et al.* [38] collect a new dataset that contains high-quality spatially-aligned video pairs in both low-light and normal-light conditions, and further design a self-supervised network to reduce noises and enhance the illumination based on the Retinex theory. Chen *et al.* [5] propose a siamese network and introduce a self-consistency loss to preserve color while suppressing spatial and temporal artifacts efficiently. StableLLVE [55] maintains the temporal consistency after enhancement by learning and inferring motion field (*i.e.*, optical flow) from synthesized short-range video sequences. In order to ensure the diversity of visual effects of the enhanced videos, we apply both LLIE and LLVE algorithms when constructing the LLVE-QA dataset.

2.2 VQA Datasets

In order to facilitate the development of VQA algorithms, many VQA datasets [7, 8, 11, 18, 31, 39, 52, 53] have been proposed. Videos in LIVE-Qualcomm [8] contain the following 6 distortion types: color, exposure, focus, artifacts, sharpness, and stabilization. LIVE-VQC [31] contains 585 videos, which are captured by various cameras with different resolutions. In addition to the common distortions, the visual quality of UGC videos is influenced by compression generated while they are uploaded to and downloaded from the Internet. UGC-VIDEO [18] and LIVE-WC [53] simulate the specific distortion by utilizing several video compression algorithms. KoNViD-1k [11], YouTube-UGC [39], and LSVQ [52] extensively

collect in-the-wild UGC videos from the Internet, greatly expand the scale of VQA datasets. Besides, VDPVE [7] is constructed to fill in the gaps of VQA datasets specially for video enhancement, which can further promote the refined development of VQA models. However, most of existing datasets only contain unprocessed UGC videos containing various distortions. While VDPVE takes videos after enhancement into account, it is still general and not targeted. Our LLVE-QA dataset focuses on original low-light videos and corresponding enhanced videos after LLVE, which lays a solid foundation for designing a specific LLVE quality assessment model.

2.3 NR-VQA Models

The traditional and naive NR-VQA models are based on handcrafted features [12, 13, 24, 27, 35]. These handcrafted features, including spatial features, temporal features, statistical features, and so on, can be extracted to learn the quality scores of videos. For example, V-BLIINDS [27] builds a Natural Scene Statistics (NSS) module to extract spatial-temporal features and a motion module to quantify motion coherency. The core of TLVQM [12] is to generate video features in two levels, in which low complexity features are extracted from the full sequence first, and then high complexity features are extracted in key frames which are selected by utilizing low complexity features. VIDEVAL [35] combines existing VQA methods together and proposes a feature selection strategy, which can choose appropriate features and then fuse them efficiently to predict the quality scores of videos.

With the rapid pace of technological advancements, VQA models based on deep learning [14, 15, 17, 33, 36, 40, 42–48, 50, 52, 58] have progressively emerged as the prevailing trend. For example, based on a pre-trained DNN model and Gated Recurrent Units (GRUs), VSFA [17] reflects the temporal connection between the semantic features of key frames well. BVQA [15] and Simple-VQA [33] further take the impact of motion features on videos into account and introduce motion features extracted by the pre-trained 3D CNN



Figure 3: Representative frames of two original videos and their corresponding enhanced videos.

models. Wang *et al.* [40] propose a DNN-based framework to measure the quality of UGC videos from three aspects: video content, technical quality, and compression level. FAST-VQA [42] creatively introduces a Grid Mini-patch Sampling to generate fragments, and utilizes a model with Swin Transformer [22] as the backbone to extract features efficiently from these fragments. RAPIQUE [36] leverages quality-aware statistical features and semantics-aware convolutional features, which first attempts to combine handcrafted and deep-learning-based features. While prior VQA models are designed for general UGC videos without exception, our Light-VQA model focuses on LLVE quality assessment by additionally introducing handcrafted brightness and noise features that significantly affect the quality of low-light videos and their corresponding enhanced results to improve the assessment accuracy.

3 DATASET PREPARATION

3.1 Video Collection

A high-quality dataset is a prerequisite for a well-performing model. To start with, we elaborately select 254 low-light videos from VDPVE [7], LIVE-VQC [31], YouTube-UGC [39], and SDSD [38] datasets. The low-light videos we choose contain diverse content and various degrees of brightness. Subsequently, we employ 7 low light enhancement algorithms (*i.e.*, AGCCPF [9], GHE [1], BPHEME [37], SGZSL [59], DCC-Net [56], MBLLVEN [23] and StableLLVE [55]) and one commercial software CapCut [4] respectively to obtain the enhanced videos. We further remove the videos with extremely poor visual quality due to the distortions generated in the process of enhancement. Eventually, 254 original low-light videos and 1,806 enhanced videos constitute our LLVE-QA dataset. To the best of our knowledge, this is the *first* dataset specifically designed for evaluating low-light video enhancement algorithms. Representative frames of two original videos and their corresponding enhanced videos are shown in Figure 3.

3.2 Subjective Experiment

We invite 22 subjects to participate in the subjective experiment. All of them are professional and experienced data labeling staff.

Subjects are required to score the quality of videos within the range of [0, 100]. The scoring criteria are that higher score corresponds to the better video quality. In the process of scoring a group of videos (including an original low-light video and corresponding enhanced videos), to make subjects not limited to the video content but pay more attention to the visual perceptual quality of the videos, we customize a scoring interface which is demonstrated in Supplementary. Subjects are supposed to score the original video first. When they score the enhanced videos, they can observe and compare them to the original video repeatedly. Compared to randomly shuffling the order of videos for scoring, the subjective quality scores obtained in this way can better reflect the visual perceptual difference caused by LLVE.

3.3 Data Analysis

In order to measure the visual perceptual difference between original and enhanced videos, we calculate three video attributes [7]: brightness, contrast, and colorfulness, which are normalized and shown in Figure 4. Colorfulness is not significantly changed before and after video enhancement, while the brightness and contrast have undergone major changes, which is in line with visual perception. Since there is a large amount of redundant information between adjacent frames, we only select a subset of all video frames for processing. The concrete calculation process are listed as follows [11]:

(1) *Brightness*: Given a video frame, we convert it to grayscale and compute the average of pixel values. Then the brightness result of a video is obtained by averaging the brightness of all selected frames.

(2) *Contrast*: For a video frame, its contrast is obtained simply by computing standard deviation of pixel grayscale intensities [26]. Then we average the contrast results of all selected frames to get the contrast of a video.

(3) *Colorfulness*: We utilize Hasler and Suesstrunk's metric [10] to calculate this attribute. Specifically, given a video frame in RGB format, we compute $rg = R - G$ and $yb = \frac{1}{2}(R + G) - B$ first, and then the colorfulness is calculated by $\sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + \frac{3}{10}\sqrt{\mu_{rg}^2 + \mu_{yb}^2}$.

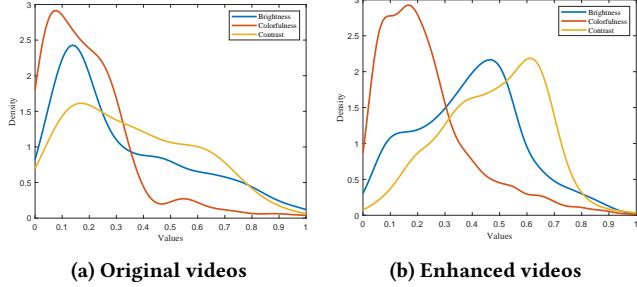


Figure 4: Distributions of brightness, contrast, and colorfulness over the original and enhanced videos in our LLVE-QA dataset.

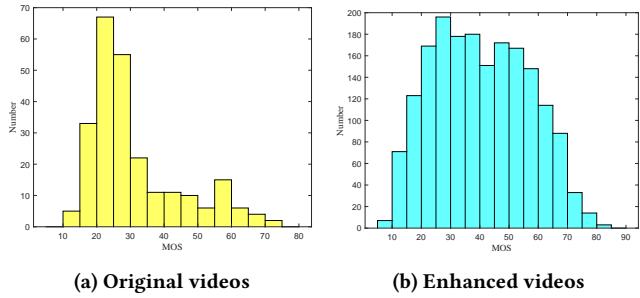


Figure 5: Detailed MOS distributions of the original and enhanced videos in our LLVE-QA dataset.

where σ^2 and μ represent the variance and mean value respectively. Finally, we average the colorfulness values of all selected frames to obtain the colorfulness of a video.

After the subjective experiment, we collect 45,320 (*i.e.*, $22 \times 2,060$) scores in total. Considering the standard deviation can well reflect the distribution of data, we calculate the standard deviation of 2,060 scores generated by each subject and reject two invalid subjects whose standard deviations of ratings are significantly lower than others. Finally, we obtain 20 valid subjects and MOSs for all videos in LLVE-QA dataset. In order to provide the insights in the difference between the MOSs of original videos and enhanced videos, we draw the MOS distributions in Figure 5. The relatively uniform MOS distribution in Figure 5b reflects the diversity of visual quality of enhanced videos obtained by our selected LLVE algorithms.

4 PROPOSED METHOD

Benefiting from the built dataset, we propose a multi-dimensional quality assessment model named Light-VQA for low-light video enhancement. This model consists of the spatial and temporal information extraction module, the feature fusion module, and the quality regression module as shown in Figure 2. Specifically, spatial information extracted from key frames contains deep-learning-based semantic features, handcrafted brightness, and noise features. Temporal information extracted from video clips contains deep-learning-based motion features and handcrafted brightness consistency features. Then they are fused to obtain the quality-aware representation. Finally, we utilize two Fully Connected (FC) layers to regress fused features into the video quality score.

4.1 Spatial Information

Since the adjacent frames of a video contain plenty of redundant contents, spatial information shows the extreme sensitivity to the video resolution and is not quite relevant to the video frame rate. Therefore, in order to reduce the computational complexity, we uniformly select k key frames from the video to extract spatial information. In Light-VQA, we design two branches to simultaneously extract features in a video. Concretely, one is for deep-learning-based features, which contain rich semantic information, the other is for handcrafted features, which contain brightness and noise specifically designed for evaluating the quality of low-light and corresponding enhanced videos.

Swin Transformer [22] has achieved more excellent performance than traditional CNNs in computer vision tasks such as image classification, object detection, and segmentation. For deep-learning-based features, we utilize the semantic information extracted from the last two stages of the pre-trained Swin Transformer:

$$\begin{aligned} SF_i &= \alpha_1 \oplus \alpha_2, i \in \{1, \dots, k\}, \\ \alpha_j &= GAP(F_i^j), j \in \{1, 2\}, \end{aligned} \quad (1)$$

where SF_i indicates the extracted semantic features of the i -th sampled key frame of a video, \oplus represents the concatenation operation, $GAP(\cdot)$ stands for the global average pooling operation, F_i^j indicates the feature maps in the i -th key frame generated from the j -th last stage of Swin Transformer, and α_j denotes the features after average pooling. For handcrafted features, we extract brightness and noise features which influence the quality of low-light and corresponding enhanced videos greatly [54] to better improve the quality-aware representation of Light-VQA:

$$\begin{aligned} BF_i &= \Theta(F_i), \\ NF_i &= \Psi(F_i), \end{aligned} \quad (2)$$

where BF_i and NF_i indicate the extracted brightness and noise features from the i -th sampled key frame, respectively. $\Theta(\cdot)$ and $\Psi(\cdot)$ represent the extraction process of brightness and noise features, respectively.

Therefore, given a video, we first uniformly select k key frames, and then extract deep-learning-based and handcrafted features through two branches respectively. Finally, quality-aware spatial information is obtained by concatenating semantic features, brightness and noise features together:

$$SI_i = SF_i \oplus BF_i \oplus NF_i, \quad (3)$$

where SI_i indicates the ultimate spatial information of the i -th sampled key frame.

4.2 Temporal Information

Different from spatial information, temporal information is extremely susceptible to video frame-rate variations but not sensitive to resolution [33]. Therefore, in order to preserve adequate temporal information while reducing computational complexity, we uniformly split the video into k clips with lower resolution for temporal information extraction. Concretely, similar to the extraction of spatial information, we design two branches to obtain deep-learning-based and handcrafted features respectively. One is for

Table 1: Experimental performance on our constructed LLVE-QA dataset and subset of KoNViD-1k. Our proposed Light-VQA achieves the best performance. ‘Handcrafted’ and ‘Deep Learning’ denote two types of leveraged features. The handcrafted models are inferior to deep-learning-based models. Best in red and second in blue.

VQA Model	Handcrafted	Deep Learning	LLVE-QA			Subset of KoNViD-1k		
			SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓
V-BLIINDS (TIP, 2014) [27]	✓		0.7123	0.7130	11.6185	0.5927	0.6157	13.2098
TLVQM (TIP, 2019) [12]	✓		0.7321	0.7401	9.0957	0.4260	0.4671	12.4164
VIDEVAL (TIP, 2021) [35]	✓		0.5294	0.5233	13.9555	0.4963	0.4052	12.6448
RAPIQUE (OJSP, 2021) [36]	✓	✓	0.5890	0.5922	13.2555	0.3861	0.4751	14.5661
Simple-VQA (ACM MM, 2022) [33]		✓	0.8984	0.8983	7.2287	0.6978	0.7101	10.0307
FAST-VQA (ECCV, 2022) [42]		✓	0.9156	0.9159	6.3528	0.7064	0.7156	10.7450
Light-VQA	✓	✓	0.9374	0.9393	5.6523	0.7975	0.7860	8.8070

motion, the other is for brightness consistency which is a significant feature in low-light videos.

For deep-learning-based features, we utilize a pre-trained Slow-Fast network [6] to extract motion features for each video clip:

$$MF_i = \Phi(V_i), \quad (4)$$

where V_i indicates the i -th video clip, $\Phi(\cdot)$ denotes the extraction operation of motion features, and MF_i stands for the extracted motion features from the i -th video clip. For handcrafted features, we extract brightness consistency features:

$$CF_i = \Gamma(V_i), \quad (5)$$

where Γ indicates the extraction operation of brightness consistency features, and CF_i denotes the extracted brightness consistency features from the i -th video clip.

To sum up, given a video, we split it into k clips with lower resolution uniformly, and then extract deep-learning-based and handcrafted features through two branches respectively. Finally, temporal information is obtained by concatenating motion features and brightness consistency features together:

$$TI_i = MF_i \oplus CF_i, \quad (6)$$

where TI_i indicates the temporal information of the i -th video clip.

4.3 Spatial-Temporal Fusion

After obtaining the both spatial and temporal information, it is essential to fuse them to get a more comprehensive feature expression. In this paper, we utilize Multi-Layer Perception (MLP) as the fusion module to integrate spatial with temporal information due to its simplicity and effectiveness. Various fusion strategies based on attention mechanisms can be included, but they are beyond the scope of this paper. Specifically, given spatial information SI_i extracted from the i -th key frame and temporal information TI_i extracted from the i -th video clip, we concatenate them first and then pass them through a MLP:

$$FF_i = \mathcal{F}(SI_i \oplus TI_i), \quad (7)$$

where $\mathcal{F}(\cdot)$ indicates the learnable feature fusion that contains one FC layer with 1,024 neurons and one REctified Linear Unit (RELU), and FF_i represents features after fusion of the i -th video clip (the i -th key frame can be regarded as one frame in the i -th video clip

but with original resolution). All elements in FF_i are calculated jointly by SI_i and TI_i .

4.4 Quality Regression

Subsequently, we utilize another two FC layers to regress quality-aware representation FF_i into the video quality score:

$$Q_i = FC(FF_i), \quad (8)$$

where Q_i indicates the quality score of the i -th video clip. Finally, the overall score of the entire video is obtained by averaging the quality scores of all k video clips:

$$Q = \frac{1}{k} \sum_{i=1}^k Q_i, \quad (9)$$

where Q is the quality score of the video and k indicates the number of video clips.

Our loss function for training is composed of two parts [33]: Mean Absolute Error (MAE) loss and rank loss [41]. MAE loss is widely used in various deep learning tasks, and in this paper it is defined as:

$$L_{MAE} = \frac{1}{N} \sum_{m=1}^N |Q_m - \hat{Q}_m|, \quad (10)$$

where \hat{Q}_m is the predicted MOS for the m -th video in a batch and N is the batch size. Rank loss can help the network to learn the relative quality of different videos, which exactly coincides with our need to compare the quality of different LLVE algorithms. Specifically, the rank loss is defined as follows:

$$L_{rank} = \frac{1}{N^2} \sum_{m=1}^N \sum_{n=1}^N \max(0, |\hat{Q}_m - \hat{Q}_n| - e(\hat{Q}_m, \hat{Q}_n) \cdot (Q_m - Q_n)), \quad (11)$$

where m and n are two different videos in one training batch, and $e(\hat{Q}_m, \hat{Q}_n)$ is formulated as:

$$e(\hat{Q}_m, \hat{Q}_n) = \begin{cases} 1, & \hat{Q}_m \geq \hat{Q}_n, \\ -1, & \hat{Q}_m < \hat{Q}_n, \end{cases} \quad (12)$$

Then, the optimization objective can be obtained by:

$$L = L_{MAE} + \beta \cdot L_{rank}, \quad (13)$$

where β is a hyper-parameter for balancing the MAE loss and the rank loss.

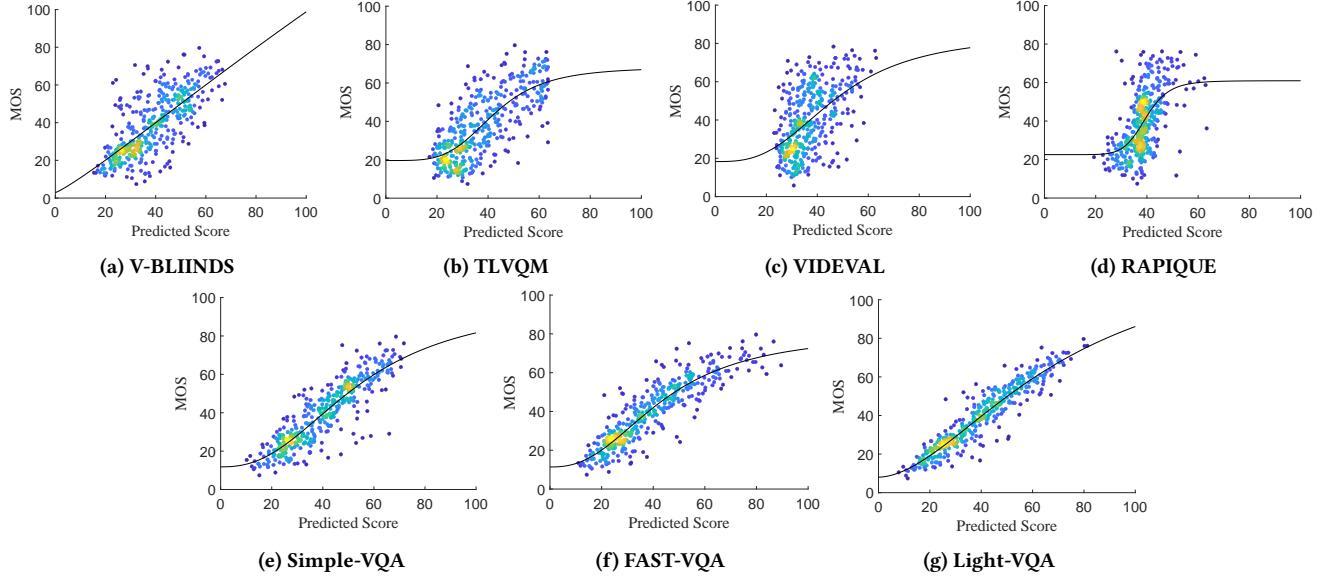


Figure 6: The scatter plots of the predicted scores versus the MOSs. The curves are obtained by a four-order polynomial nonlinear fitting. It is evident that the predicted scores of our proposed VQA bear the closest resemblance to the MOSs.

5 EXPERIMENT

5.1 Performance Comparisons with the SOTA VQA Models

To validate the effectiveness of Light-VQA on LLVE-QA dataset, we compare it with 6 state-of-the-art VQA models including V-BLIINDS [27], TLVQM [12], VIDEVAL [35], RAPIQUE [36], Simple-VQA [33], and FAST-VQA [42]. We utilize the same training strategy to train all models on the LLVE-QA dataset and ensure their convergence. Then we test them on the testing set. The numbers of videos in training set, validation set, testing set are 1260, 400, and 400 respectively. The overall experimental results are shown in Table 1. Figure 6 shows the scatter plots of the predicted MOSs versus the ground-truth MOSs on LLVE-QA dataset for 7 VQA models listed in Table 1. The curves shown in Figure 6 are obtained by a four-order polynomial nonlinear fitting. According to Table 1, Light-VQA achieves the best performance in all 7 models and leads the second place (*i.e.*, FAST-VQA) by a relatively large margin, which demonstrates its effectiveness for the perceptual quality assessment of low-light video enhancement.

5.2 Cross Dataset Performance

To examine the cross-dataset performance of the model, we conduct experiments on the subset of low light videos in KoNViD-1k [11]. The distributions of three attributes (*i.e.*, brightness, colorfulness and contrast) and MOS of the subset are shown in Figure 7. We directly leverage the models pre-trained on LLVE-QA dataset to perform testing on the newly built subset with ease and efficiency. The overall experimental results on subset of KoNViD-1k are shown in Table 1. Since LLVE-QA dataset includes both low-light videos and their corresponding enhanced versions, whereas KoNViD-1k exclusively consists of low-light videos, the quality-aware representation learned from LLVE-QA dataset is less effective on KoNViD-1k. All

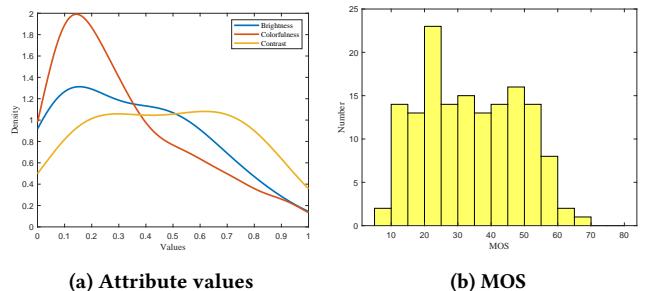


Figure 7: Detailed attribute and MOS distribution for low light video subset of KoNViD-1k [11].

methods have experienced the decline of performance. However, our proposed Light-VQA still surpasses the other 6 VQA methods by a large margin, which demonstrates its good generalization ability in terms of the quality assessment of low light videos.

5.3 Ablation Studies

In this subsection, a series of ablation experiments are conducted to analyze the contributions of different modules in Light-VQA. Table 2 shows the experimental results of ablation studies. *Model 1* only utilizes Semantic Features (SF) extracted by Swin Transformer [22]. *Model 2* only utilizes Motion Features (MF) extracted by SlowFast R50 [6]. *Model 3* utilizes both SF and MF, and obtains the results after passing them through the Feature Fusion (FF) module. Based on *Model 3*, *Model 4* adds handcrafted Brightness and Noise Features (BF + NF) that belong to spatial information together with SF. Based on *Model 3*, *Model 5* adds handcrafted Brightness Consistency Features (CF) that belong to temporal information coupled with MF. *Model 6* utilizes all the spatial information and temporal information, but instead of performing feature fusion, Multiple Linear Regression (MLR) is used as a replacement. *Model 7* is the complete

Table 2: Experimental results of ablation studies on LLVE-QA dataset. Best in red and second in blue. [Key: SF: Semantic Features, BF: Brightness Features, NF: Noise Features, MF: Motion Features, CF: Brightness Consistency Features, FF: Feature Fusion, MLR: Multiple Linear Regression]

Model	Spatial Information		Temporal Information		Fusion Method		LLVE-QA		
	SF	BF + NF	MF	CF	FF	MLR	SRCC↑	PLCC↑	RMSE↓
1	✓						0.9120	0.9143	6.5377
2				✓			0.8446	0.8438	8.8438
3	✓			✓		✓	0.9223	0.9245	6.6829
4	✓		✓	✓		✓	0.9324	0.9354	5.8278
5	✓			✓		✓	0.9299	0.9310	5.9764
6	✓	✓	✓	✓		✓	0.9231	0.9243	6.7132
7	✓	✓	✓	✓	✓		0.9374	0.9393	5.6523

Table 3: The average scores predicted by Light-VQA on original low-light videos, results of StableLLVE w/o Light-VQA, and results of StableLLVE w/ Light-VQA.

Dataset	Origin	w/o Light-VQA	w/ Light-VQA
Average score	39.0933	59.4926	86.2688

model we propose, in which we fuse all the spatial and temporal information, and obtain the best results.

Feature Extraction Module. For Light-VQA, both spatial and temporal information is composed of deep-learning-based and hand-crafted features. First, *Model 1* and *Model 2* are designed to verify the contribution of deep-learning-based features in spatial information and temporal information, respectively. It can be observed from the results that semantic features in spatial information are significantly superior to motion features in temporal information. When we fuse them in *Model 3*, the performance of the model is further improved. Second, based on *Model 3*, *Model 4*, and *Model 5* are designed to prove the effectiveness of handcrafted features in spatial information and temporal information respectively. It is evident that both of them obtain better results compared to *Model 3*. When we add them all in *Model 7*, the final model Light-VQA exhibits the best performance.

Feature Fusion Module. In this paper, we utilize MLP as the feature fusion module to integrate spatial-temporal information. To verify its effectiveness, we train two models separately, one of which only contains temporal information and the other only contains spatial information, and then we utilize Multiple Linear Regression (MLR) to get the predicted score:

$$Q_i^m = a \cdot Q_i^s + b \cdot Q_i^t + c, \quad (14)$$

where Q_i^s indicates the score obtained by spatial information module, Q_i^t indicates the score obtained by temporal information module, and Q_i^m denotes the score after MLR. a , b , and c are parameters to be fitted in MLR. By comparing the results of *Model 6* and *Model 7* in Table 2, it is evident that our feature fusion module plays a role in improving the prediction performance.

5.4 Refinement for Training LLVE Algorithms

To demonstrate that Light-VQA can be utilized as a metric to facilitate the development of LLVE algorithms by approaching the

human visual system, we use Light-VQA as a loss function to train a recent low-light video enhancement algorithm named StableLLVE [55]. Experimental results show that training with Light-VQA as a loss function yields videos with better perceptual quality compared to the original training method. The dataset we use for experiments is from SDSD [38]. The average scores predicted by Light-VQA on original low-light videos, results of StableLLVE w/o Light-VQA, and results of StableLLVE w/ Light-VQA are shown in Table 3. The detailed qualitative comparisons of are shown in Figure 8. It is evident that the results of StableLLVE training with Light-VQA have better perceptual quality.



Figure 8: The detailed qualitative comparisons of original low-light videos, results of StableLLVE w/o Light-VQA, and results of StableLLVE w/ Light-VQA.

6 CONCLUSION

In this paper, we focus on the issue of evaluating the quality of LLVE algorithms. To facilitate our work, we construct a LLVE-QA dataset containing 2,060 videos. Concretely, we collect 254 original low-light videos that contain various scenes and generate the remaining videos by utilizing different LLVE algorithms. Further, we propose an effective VQA model named Light-VQA specially for low-light video enhancement. Concretely, we integrates the luminance-sensitive handcrafted features into deep-learning-based features in both spatial and temporal information extractions. Then we fuse them to obtain the overall quality-aware representation. Extensive experimental results have validated the effectiveness of our Light-VQA. For future work, we will enable the Light-VQA to evaluate the recovery performance of overexposed videos.

ACKNOWLEDGMENTS

This work was supported in part by the Shanghai Pujiang Program under Grant 22PJ1406800, in part by the National Natural Science Foundation of China under Grant 62225112 and Grant 61831015.

REFERENCES

- [1] Ahmed S Abutaleb. Automatic thresholding of gray-level pictures using two-dimensional entropy. *Computer Vision, Graphics, and Image Processing*, 47(1):22–32, 1989.
- [2] Christos G Bampis, Zhi Li, and Alan C Bovik. Spatiotemporal feature integration and model fusion for full reference video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2256–2270, 2018.
- [3] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019.
- [4] ByteDance. Capcut. <https://www.capcut.cn/>, 2017.
- [5] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3185–3194, 2019.
- [6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.
- [7] Yixuan Gao, Yuqin Cao, Tengchuan Kou, Wei Sun, Yunlong Dong, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. VDPVE: VQA dataset for perceptual video enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop(CVPRW)*, 2023.
- [8] Deepali 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 Transactions on Circuits and Systems for Video Technology*, 28(9):2061–2077, 2017.
- [9] Bhupendra Gupta and Mayank Tiwari. Minimum mean brightness error contrast enhancement of color images using adaptive gamma correction with color preserving framework. *International Journal for Light and Electron Optics*, 127(4):1671–1676, 2016.
- [10] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pages 87–95, 2003.
- [11] 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 *Proceedings of the 9th International Conference on Quality of Multimedia Experience*, pages 1–6, 2017.
- [12] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019.
- [13] Jari Korhonen, Yicheng Su, and Junyong You. Blind natural video quality prediction via statistical temporal features and deep spatial features. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3311–3319, 2020.
- [14] Tengchuan Kou, Xiaohong Liu, Jun Jia, Wei Sun, Guangtao Zhai, and Ning Liu. Stablevqa: A deep no-reference quality assessment model for video stability. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.
- [15] 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 Transactions on Circuits and Systems for Video Technology*, 32(9):5944–5958, 2022.
- [16] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *arXiv preprint arXiv:2306.04717*, 2023.
- [17] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2351–2359, 2019.
- [18] Yang Li, Shengbin Meng, Xinfeng Zhang, Shiqi Wang, Yue Wang, and Siwei Ma. Ugc-video: perceptual quality assessment of user-generated videos. In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval*, pages 35–38, 2020.
- [19] Xiaohong Liu, Lei Chen, Wenyi Wang, and Jiying Zhao. Robust multi-frame super-resolution based on spatially weighted half-quadratic estimation and adaptive btv regularization. *IEEE Transactions on Image Processing*, 27:4971–4986, 2018.
- [20] Xiaohong Liu, Lingshi Kong, Yang Zhou, Jiying Zhao, and Jun Chen. End-to-end trainable video super-resolution based on a new mechanism for implicit motion estimation and compensation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2416–2425, 2020.
- [21] Xiaohong Liu, Kangdi Shi, Zhe Wang, and Jun Chen. Exploit camera raw data for video super-resolution via hidden markov model inference. *IEEE Transactions on Image Processing*, 30:2127–2140, 2021.
- [22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.
- [23] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *Proceedings of the British Machine Vision Conference*, volume 220, page 4, 2018.
- [24] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2015.
- [25] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017.
- [26] Eli Peli. Contrast in complex images. *JOSA A*, 7(10):2032–2040, 1990.
- [27] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, 23(3):1352–1365, 2014.
- [28] Zhihao Shi, Xiaohong Liu, Chengqi Li, Linhui Dai, Jun Chen, Timothy N Davidson, and Jiyang Zhao. Learning for unconstrained space-time video super-resolution. *IEEE Transactions on Broadcasting*, 68:345–358, 2021.
- [29] Zhihao Shi, Xiaohong Liu, Kangdi Shi, Linhui Dai, and Jun Chen. Video frame interpolation via generalized deformable convolution. *IEEE transactions on multimedia*, 24:426–439, 2021.
- [30] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang. Video frame interpolation transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17482–17491, 2022.
- [31] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2018.
- [32] Rajiv Soundararajan and Alan C Bovik. Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):684–694, 2012.
- [33] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 856–865, 2022.
- [34] Zhengzhong Tu, Chia-Ju Chen, Li-Heng Chen, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. A comparative evaluation of temporal pooling methods for blind video quality assessment. In *Proceedings of the IEEE International Conference on Image Processing*, pages 141–145, 2020.
- [35] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021.
- [36] Zhengzhong Tu, Xiangyu Xu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021.
- [37] Chao Wang and Zhongfu Ye. Brightness preserving histogram equalization with maximum entropy: a variational perspective. *IEEE Transactions on Consumer Electronics*, 51(4):1326–1334, 2005.
- [38] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9700–9709, 2021.
- [39] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *Proceedings of the 21st IEEE International Workshop on Multimedia Signal Processing*, pages 1–5, 2019.
- [40] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13435–13444, 2021.
- [41] Shaoguo Wen and Junle Wang. A strong baseline for image and video quality assessment. *arXiv preprint arXiv:2111.07104*, 2021.
- [42] 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. In *Proceedings of the European Conference on Computer Vision*, pages 538–554, 2022.
- [43] 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, 2022.
- [44] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. Discovqa: Temporal distortion-content transformers for video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023.
- [45] Haoning Wu, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Erli Zhang, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring opinion-unaware video quality assessment with semantic affinity criterion. In *Proceedings of International Conference on Multimedia and Expo (ICME)*, 2023.
- [46] Haoning Wu, Liang Liao, Annan Wang, Chaofeng Chen, Jingwen Hou Hou, Erli Zhang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Towards robust text-prompted semantic criterion for in-the-wild video quality assessment, 2023.
- [47] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [48] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Towards explainable video quality assessment: A database and a language-prompted approach. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 2023.
- [49] Jiahua Xu, Jing Li, Xingguang Zhou, Wei Zhou, Baichao Wang, and Zhibo Chen. Perceptual quality assessment of internet videos. In *Proceedings of the 29th ACM*

- International Conference on Multimedia*, pages 1248–1257, 2021.
- [50] Fuwang Yi, Mianyi Chen, Wei Sun, Xiongkuo Min, Yuan Tian, and Guangtao Zhai. Attention based network for no-reference ugc video quality assessment. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1414–1418, 2021.
- [51] Guanghao Yin, Zefan Qu, Xinyang Jiang, Shan Jiang, Zhenhua Han, Ningxin Zheng, Xiaohong Liu, Huan Yang, Yuqing Yang, Dongsheng Li, and Lili Qiu. Online video streaming super-resolution with adaptive look-up table fusion. *arXiv preprint arXiv:2303.00334*, 2023.
- [52] Zhenqiang Ying, Maniratnam Mandal, Deepthi Ghadiyaram, and Alan Bovik. Patch-vq: patching up the video quality problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14019–14029, 2021.
- [53] Xiangyu Yu, Neil Birkbeck, Yilin Wang, Christos G Bampis, Balaji Adsumilli, and Alan C Bovik. Predicting the quality of compressed videos with pre-existing distortions. *IEEE Transactions on Image Processing*, 30:7511–7526, 2021.
- [54] Guangtao Zhai, Wei Sun, Xiongkuo Min, and Jiantao Zhou. Perceptual quality assessment of low-light image enhancement. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(4):1–24, 2021.
- [55] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4967–4976, 2021.
- [56] Zhao Zhang, Huan Zheng, Richang Hong, Mingliang Xu, Shuicheng Yan, and Meng Wang. Deep color consistent network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1899–1908, 2022.
- [57] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A perceptual quality assessment exploration for aigc images. *arXiv preprint arXiv:2303.12618*, 2023.
- [58] Zicheng Zhang, Wei Sun, Yingjie Zhou, Haoning Wu, Chunyi Li, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Advancing zero-shot digital human quality assessment through text-prompted evaluation. *arXiv preprint arXiv:2307.02808*, 2023.
- [59] Shen Zheng and Gaurav Gupta. Semantic-guided zero-shot learning for low-light image/video enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 581–590, 2022.
- [60] Yinqiang Zheng, Mingfang Zhang, and Feng Lu. Optical flow in the dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6749–6757, 2020.