Short-form UGC Video Quality Assessment Based on Multi-Level Video Fusion with Rank-Aware

Haoran Xu*
22251254@zju.edu.cn

Mengduo Yang*
yangmd2118@mails.jlu.edu.cn

Jie Zhou*
22251247@zju.edu.cn
Zhejiang University

Jiaze Li*
Jiaze_Li@zju.edu.cn
Zhejiang University

Abstract

Short-form UGC video platforms, such as Kwai and TikTok, have ushered in vigorous development. However, due to the variety of short video types and uneven quality, the workload of manual annotation is heavy. In this paper, video is decomposed into three levels (frame level, segment level, and video level) based on the view of data augmentation and multi-level fusion, and a new integrated framework is proposed to capture the spatial-temporal characteristics and relative rank information of different levels. It uses spatial-temporal data augmentation strategy, multi-level feature fusion, adaptive rank-aware loss, and redistributed model ensemble at all levels. These components allow our method not only to capture features at each level but also to mitigate the difficulty of identifying the relative rank of the two kinds of hard samples. Our framework achieves 5th place among all methods in the NTIRE 2024 Short-form UGC Video Quality Assessment Challenge. A large number of experiments show that our framework not only performs well on the KVQ dataset but also on other benchmark VQA datasets. It proves the generalization and superiority of our framework.

1. Introduction

In recent years, the number of users of short-form UGC video platforms has soared, and popular short videos often have more than 100 million plays [43]. Due to the passion of users to create, the forms of short videos are naturally diverse. Different from traditional long videos, short-form UGC videos are very short, usually only a few seconds [19]. This results in the quality of short videos changing faster and the difference is often greater than traditional videos. In addition, the short video platform itself often does some post-processing on the video [4], which affects the fluctuation of the video quality. Therefore, it is necessary to evaluate the quality of short-form UGC videos.

Traditional Video Quality Assessment (VQA) methods are generally divided into three types: full-reference, reduced reference, and no-reference, which are based on the reference information [20]. Almost all VQA methods typically follow a universal paradigm, that is, extracting visual features and designing regression or classification head for quality prediction from the extracted features [27].

Based on the above, two paths have emerged for VQA research: 1) extract stronger and more representative features, including using more popular and advanced backbones and modules, and pretraining on a larger range of datasets [42]; 2) use stronger mass fraction regression heads. We think about this problem from the perspective of spatial domain and temporal domain, from the macro to the micro. A typical video can be broken down into videos, segments, and frames. However, due to the lack of multi-level labels in existing VQA datasets, we designed a framework for three-level (frame-segment-video) supervised training based on the idea of data augmentation and labeled corresponding pseudo-labels. In addition, to guarantee the stability of training, we use the methods of redistribution and model integration in the three-level training framework, that is, training separately, inference once, distribution alignment, and label integration.

On the whole, to overcome the above difficulties, we propose a three-level integration framework for short-form UGC VQA. The contributions of this framework are summarized as follows:

- We proposal a multi-level framework. Globally, a three-level architecture is proposed to capture features at each

* These authors contributed equally to this work.

level, and locally, features on backbones from low level to high level are fused.
- Based on the view of data augmentation, data augmentation in spatial and temporal domain is employed on the three-level architecture respectively to improve the robustness of the model.
- In order to distinguish between two kinds of hard samples and relative rank information, we designed an adaptive relative rank loss.
- By utilizing the redistributed model integration strategy, the distribution of score labels is aligned and the training of the model is more stable.

The rest of this paper is organized as follows. In Sec. 2, we briefly review the existing Video Quality Assessment(VQA) methods. Our proposed framework is detailed in Sec. 3, and different experiments are presented in Sec. 4. Finally, Sec. 5 concludes this paper.

2. Related Work

2.1. Video Quality Assessment

The core mission of Video Quality Assessment (VQA) is to predict as accurately as possible the subjective quality as perceived by alignment human preferences. There are potential default metrics for human ratings of video quality. So a very natural idea is to evaluate video quality through handcrafted features [25, 32, 34]. Among these works, TLVQM [11] attempts to catch the spatial and temporal handcrafted features such as motion, jerkiness, blurriness, noise, etc. VIDEVAL [32] models diverse authentic distortions using different handcrafted features. However, handcrafted features often can not completely cover the video quality indicators, and the semantics of video content itself often affect the video quality evaluation [12, 34]. With the emergence and wide application of deep learning, the framework of end-to-end feature extraction and video quality assessment using regression head has gradually gained popularity [27]. The combination of manual features and end-to-end features is also a topic of exploration. TLVQM [10] and VIDEVAL [32] merge an extensive array of spatial-temporal features with traditional quality metrics in order to translate them into a numerical representation of video quality. Considering the utility of spatial and temporal domain aggregation under the perspective of aggregation [31], V-BLINDS [26] pioneers a model that harnesses spatio-temporal natural scene statistics (NSS) by evaluating the discrepancies in NSS attributes from frame to frame. VIIDEO [23] capitalizes on the intrinsic statistical patterns present in natural videos to tackle distortions that are specific to certain types of impairments. Finally, for a specific video, it is a feasible scheme to completely abandon the temporal domain and use the image quality assessment method to achieve a specific video quality assessment. This only requires the video to be drawn into frames according to certain rules [21, 22, 45]. Building upon this, subsequent investigations [12, 35] strive to fuse these tailored features with the semantic depth extracted from end to end models that have undergone preliminary training.

Based on the above, deep learning-based approaches have risen to prominence, fueled by the advent of extensive VQA datasets [9, 37, 46]. Notably, the backbone with stronger feature extraction often needs to be pretrained on larger datasets, for example, the VSFA model [14] harnesses the pretrained ResNet-50 [8], which was trained on the ImageNet-1k [5] dataset, to capture spatial characteristics and subsequently utilizes Gated Recurrent Units (GRU) to model the temporal information. There are also some studies on the generalization of models, and for cross-dataset aspects, MDVSFA [15] delves into the mixture of datasets and it can alleviate overfitting challenges.

Concurrently, various methods [13, 40, 46] have begun to use video models, such as 3D-CNN [2, 7], which have been pretrained on action recognition datasets to capture temporal features. However, due to the increasing resolution of video, the computing resources required for end-to-end motion feature extraction are becoming more and more expensive. Therefore, these methods often only extract static motion features [13]. To address this computational bottleneck, FAST-VQA [38] and DOVER [41] introduce the grid mini-patch sampling (GMS) strategy, which involves sampling patches in spatial level at their native resolution, which is a sampling strategy to solve computing resource problems. DOVER [41] proposed the dual characteristics of the integration of aesthetic quality assessment and technical quality assessment.

2.2. Evolution of Visual Network

The ongoing revolution in visual network development is reshaping the landscape of the realm of deep learning. The architecture of these networks is bifurcated into two main streams: those designed for the intricate patterns of image and those tailored for the fluid nature of video sequences. A landmark in this progression is the inception of the Convolutional 3D (C3D) network [29], which masterfully adapts 3D-CNN to the temporal depth of video inputs. This breakthrough was swiftly followed by a suite of innovative models such as P3D [24], S3D [44], and R(2+1)D [30], each refining the aggregation between spatial and temporal dimensions to achieve a more efficient balance of computational resources and model accuracy.

On the other hand, there has been a discernible transition in the architecture of backbone networks from Convolutional Neural Networks (CNNs) to Transformers, particularly in the form of Vision Transformers (ViT) [6, 28]. The Swin Transformer [17], in particular, brings back stronger features associated with convolutions, such as local con-

nectivity, translation equivariance, and hierarchical structure, making it a versatile backbone suitable for a wide range of applications. The success of image Transformers has inspired further exploration of Transformer based video networks, i.e. from 2D to 3D, examples of which include ViViT [1], and Video Swin Transformer [18]. One of the key features of Transformers is their patch-wise operations. It divides and fuses patches through Patch Partition and Patch Merge modules. This makes them particularly suitable for processing inputs sampled using grid mini-patch sampling (GMS), reducing the consumption of computing resources, and better for capturing global features than CNN.

3. Method

Our UGC Video Quality Assessment (VQA) method is structured to address the complexity of UGC videos through a multi-faceted approach. We begin in Sec. 3.1 with an overview of our multi-scale feature extraction strategy, which is pivotal for capturing the rich and varied content present in UGC videos. We then proceed to Sec. 3.2, where we delve into the specifics of data augmentation across these scales. In Sec. 3.3, we introduce an adaptive rank-aware loss function designed to address the challenges posed by hard samples within the video data. This function is crucial for refining the model's ability to make accurate quality assessments, especially in scenarios with subtle quality distinctions. Finally, Sec. 3.4 is dedicated to discussing the integration of features extracted from different scales. Each part of this method is designed to work in harmony, providing a nuanced and robust framework for UGC VQA.

3.1. Multi-Level Feature Extraction

3.1.1 Video-Segment-Frame

In our quest to thoroughly evaluate video quality, we have developed a multi-faceted approach that captures the subtleties of multi-level feature extraction. As shown in Fig. 1, our network is designed to function across three key levels: Video, Segment, and Frame. Each level provides a unique contribution to the comprehensive assessment of video quality.

At the Video level, we employ the Swin Transformer, which is adept at capturing global features and understanding the overall context of the video content. This model operates on the entire video sequence, leveraging its hierarchical structure to extract features that are indicative of the video's quality as a whole.

Transitioning to the Segment level, we utilize a SlowFast network, which is specifically designed to capture both slow and fast motion features. This dual-rate approach allows the model to discern the nuances of motion within the video segments, providing a more detailed understanding of the dynamic aspects that contribute to video quality.

Finally, at the Frame level, we incorporate a series of Convolutional Network (ConvNet) blocks. These blocks are tasked with extracting local features from individual frames, focusing on the fine details and textures that are critical for a granular assessment of video quality.

Through this hierarchical model architecture, we extract a rich set of features at each level, which are then combined through a Redistribution and Ensemble process to form a comprehensive video quality score. This multi-level integration enhances the accuracy of our assessments and provides a robust framework capable of handling the complexity and variability of UGC video content.

To maintain the stability of the model's convergence and to further refine its performance, we employ distinct loss functions at each level. The use of different losses is strategically designed to address the unique challenges and characteristics of each level. By doing so, we optimize the learning process at each stage, ensuring that the model can effectively capture and generalize from the diverse features present in UGC videos. This methodological choice not only improves the accuracy of our video quality assessments but also fortifies the model against the intricacies of real-world video content.

In the subsequent content, we will explore the specific loss functions used at each level and explain how they contribute to the model's overall performance, highlighting the importance of this tailored approach in achieving reliable and nuanced video quality evaluations.

$$\mathcal{L}_{\text{Segment}} = \mathcal{L}_1(\mathbf{y}, \hat{\mathbf{y}}) + \lambda_1 \mathcal{L}_{\text{Rank}}(\mathbf{y}, \hat{\mathbf{y}}) \tag{1}$$

$$\mathcal{L}_{\text{Frame}} = \mathcal{L}_{\text{SmoothL1}}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda_2 \mathcal{L}_{\text{Rank}}(\mathbf{y}, \hat{\mathbf{y}}) \tag{2}$$

$$\mathcal{L}_{\text{Video}} = \mathcal{L}_{\text{PLCC}}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda_3 \mathcal{L}_{\text{Rank}}(\mathbf{y}, \hat{\mathbf{y}}) \tag{3}$$

To enhance the precision of the predictions of the model, we have tailored specific loss functions for each level.Meanwhile, we use a uniform Rank loss, the details of which are provided in Sec. 3.3,. This approach takes advantage of the unique characteristics of each level, ensuring that our model can effectively capture the diverse nuances present in various types of video content.

$$\text{MSE}(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2, \tag{4}$$

$$\mathcal{L}_{\text{PLCC}}(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \left(\frac{\text{MSE}(\mathbf{x}, \mathbf{y})}{4} + \frac{\text{MSE}(\rho \cdot \mathbf{x}, \mathbf{y})}{4} \right), \tag{5}$$

$$\text{with } \rho = \frac{1}{N} \sum_{i=1}^{N} x_i \cdot y_i, \tag{6}$$

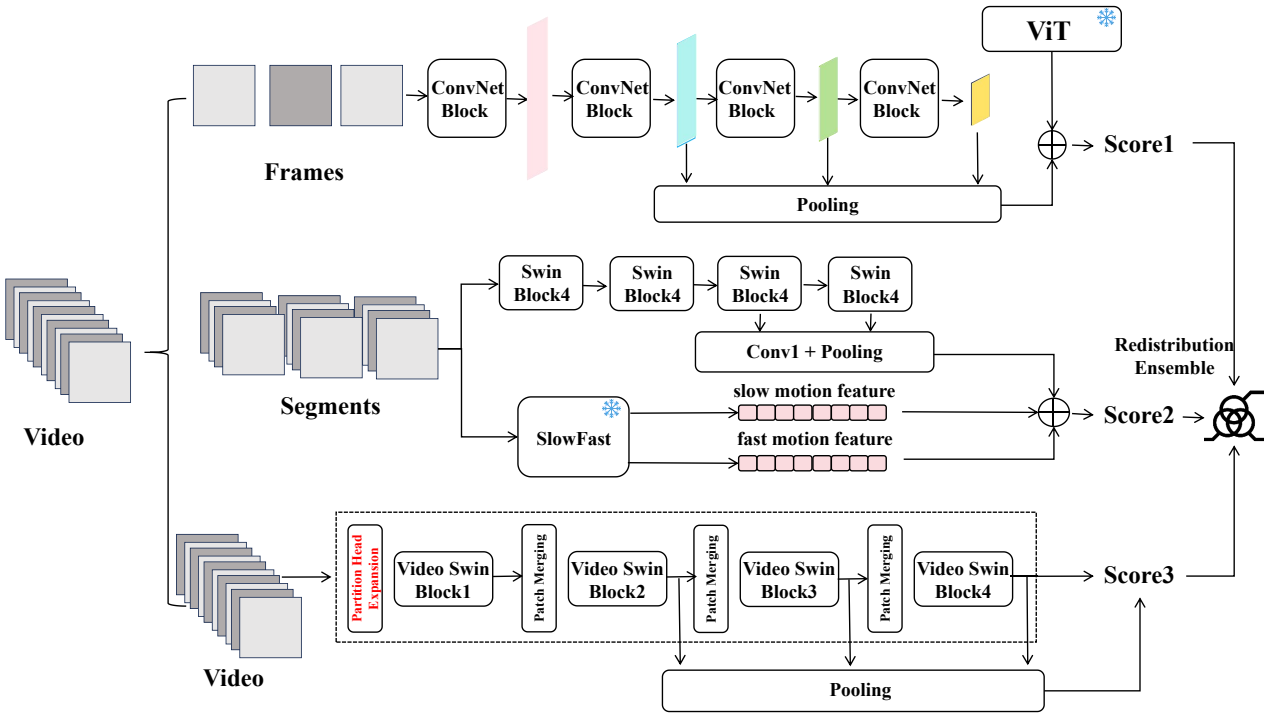

Figure 1. An overview of our proposed VQA framework, which is divided into frame, segment, and video, the three main components. The final results from these three levels are ensembled to provide a comprehensive assessment of video quality.

At the Video level, we incorporate a PLCC (Pearson-Linear Correlation Coefficient) loss, which measures the linear correlation between the predictions of the model and the ground truth.

$$\mathcal{L}_1(\mathbf{y}, \hat{\mathbf{y}}) = \sum_i |y_i - \hat{y}_i|, \tag{7}$$

At the Segment level, we combine a L1 loss, which is adept at handling pixel-wise errors and preserving image details

$$\mathcal{L}_{\text{SmoothL1}}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_i l(y_i, \hat{y}_i) \tag{8}$$

$$\text{with } l(y_i, \hat{y}_i) = \begin{cases} \frac{1}{2\gamma}(y_i - \hat{y}_i)^2, & |y_i - \hat{y}_i| < \gamma, \\ |y_i - \hat{y}_i| - \frac{\gamma}{2}, & \text{otherwise.} \end{cases} \tag{9}$$

Finally, at the Frame level, we employ a Smooth L1 loss, denoted as $\mathcal{L}_{\text{SmoothL1}}(\mathbf{y}, \hat{\mathbf{y}})$, with y_i and $\hat{y}_i$ representing the true and predicted values for the i-th sample, respectively. The element-wise loss function $l(y_i, \hat{y}_i)$ providing a robust measure against outliers and ensuring smoothness in the gradient of the loss function.

The integration of results from the three different levels will be described in Sec. 3.4. By integrating these multi-level features into our network, we ensure that our model is capable of capturing the complex interplay between global video patterns and local quality indicators, thereby providing a robust and nuanced assessment of User Generated Content (UGC) video quality.

3.1.2 Muti-Level inner Model

Within the Swin Transformer model, which is a key component at the Video and Segment levels, we focus on extracting and aggregating features from the latter three layers of the model.We posit that each layer within the Swin Transformer is attuned to capturing features at different levels of abstraction.The lower layers are typically more sensitive to fine-grained, textural details, while the higher layers synthesize this information to form a more comprehensive understanding of the video content.By aggregating features from the latter three layers (which we will refer to as the "aggregated feature set"), we enable the model to leverage the rich, hierarchical representations that Swin Transformer provides.This aggregated feature set is then used to inform the quality assessment, ensuring that the model's predictions are grounded in a nuanced understanding of the video's visual elements.

$$\mathbf{F}_{\text{concat}} = \text{Concat}(\mathbf{F}_{n-2}, \mathbf{F}_{n-1}, \mathbf{F}_n), \tag{10}$$

$$\mathbf{F}_{\text{agg}} = \text{ReLU}(\mathbf{F}_{\text{concat}}), \tag{11}$$

The features $\mathbf{F}_i$ from the last three layers of the Swin Transformer, where $i = n-2, n-1, n$, are concatenated along the channel dimension to form $\mathbf{F}_{\text{concat}}$. Then, the ReLU activation function is applied to $\mathbf{F}_{\text{concat}}$ to produce the aggregated feature set $\mathbf{F}_{\text{agg}}$. This process captures both low-level and high-level information from the video frames.

Through this dual-pronged approach to multi-level feature extraction, we ensure that our model is well-equipped to handle the complexity and variability of UGC videos.

3.2. Data Augmentation

3.2.1 Frame-Level and Data Augmentation Feature Fusion

At the Frame level, our methodology centers on enhancing the granularity of video quality assessment by leveraging data augmentation techniques. We initiate this process by extracting individual frames from the video, which serves as the fundamental unit for quality evaluation at this level. In a pivotal move that amplifies our dataset, we broadcast the video's quality score to corresponding frames, thereby endowing each frame with the same quality metric as its source video. This strategy not only significantly inflates the volume of training data but also refines the model's comprehension of subtle detail features that are critical for accurate VQA.

To capture both local and global characteristics inherent in video frames, we employ a dual-model approach: the ConvNet for local feature extraction and the Vision Transformer (ViT) for global feature comprehension. The ConvNet, with its deep convolutional architecture, is adept at identifying localized features and textures that may affect perceived video quality. In contrast, the ViT, known for its transformative ability to process spatial relationships, aids in distilling a holistic understanding of the frame's quality.

Through this comprehensive Frame-Level Data Augmentation and Feature Fusion strategy, our model is equipped to dissect the complex tapestry of video quality with precision and finesse. By amalgamating the strengths of ResNet and ViT with the adaptiveness of our loss function, we forge a pathway to a more nuanced and insightful VQA framework.

3.2.2 Segment-Level Data Sampling and Augmentation

In order to capture the dynamic essence of videos and to fortify the robustness of our model, we have adopted a segmented approach to video analysis. By dividing the video into distinct segments, we enable the model to focus on shorter, more manageable sequences that encapsulate the temporal evolution of visual content.

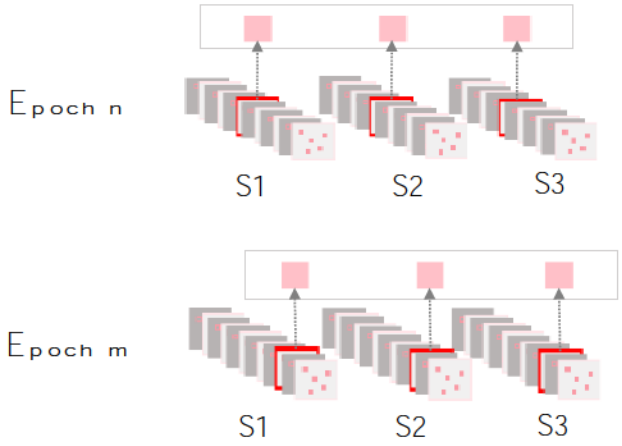

Figure 2. A schematic of the sampling strategy employed at the Segment Level of our framework.

As depicted in Fig. 2, within the video stream, frames are extracted in discrete segments, which may overlap or be separated by intervening frames. To maintain consistency in the relative positioning of frames, a key frame is selected from each segment during the same epoch, ensuring that the key frames across segments are aligned. These key frames are then used to extract patches, which serve as the input for the Segments level model. It is important to note that the positioning of key frames is randomized across different epochs, introducing variability to enhance the model's generalization capabilities. This approach ensures that the model is trained on a diverse set of video content, capturing both temporal and spatial features crucial for effective video quality assessment.

Additionally,to enhance the variability and challenge faced by the model, we incorporate randomness in the selection of segment starting points. This approach prevents the model from developing path dependencies, encouraging it to adapt to various contexts and conditions present within the video content.

Through these segment-level data sampling and augmentation techniques, we equip our model with the capability to analyze and comprehend the intricate details and temporal nuances of video content. This approach not only enriches the dataset but also refines the ability of the model to deliver precise and reliable video quality assessments.

3.3. Adaptive Rank-Aware Loss

Based on the official description of the competition-provided dataset, we became aware of two special types of data that pose unique challenges for video quality assessment.

The first type consists of non-homogeneous video pairs, where distinct video content receives the same Mean Opin-

ion Score (MOS), as exemplified in Figure 3a and 3b, both receiving an MOS of 2.861. This scenario emphasizes the need for our model to discern quality beyond mere content differences.

The second type is characterized by homogeneous video pairs, which have the same overall content but exhibit significant differences in MOS after undergoing adaptive enhancement and preprocessing. As shown in Figure 3c and 3d, the MOS scores for these pairs are 0.214 and 2.861, respectively, highlighting the influence of processing on perceived quality.

(a) (b) (c) (d)

Figure 3. The illustrate of two types of hard samples: (a) and (b) non-homogeneous video pairs, and (c) and (d) homogeneous video pairs.

In our method, we propose an Adaptive Rank-Aware Loss function to effectively handle the challenges posed by the coexistence of homogenous and heterogeneous data within Kwai UGC video dataset [19]. This loss function is designed to differentiate between hard samples and enhance the model's ability to make fine-grained distinctions in video quality.

The loss function is formulated as follows:

$$e(y_i^{gt}, y_j^{gt}) = \begin{cases} 1, & y_i^{gt} \geq y_j^{gt} \\ -1, & y_i^{gt} < y_j^{gt} \end{cases} \tag{12}$$

$$\mathcal{L}_{\text{r}} = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \left[\max(0, -e(y_i^{gt}, y_j^{gt})(y_i - y_j)) \right]^2 \tag{13}$$

$$\mathcal{L}_{\text{Rank}} = M\mathcal{L}_{\text{r}} + \lambda(1-M)M_\alpha\mathcal{L}_{\text{r}} \tag{14}$$

In this equation, M_α denotes a distance indicator function that utilizes the threshold of the ground truth score of video pair $(y_i^{gt}, y_j^{gt}$, where M_α is 1 if $\left|y_i^{gt} - y_j^{gt}\right| < c$ else is 0 if$\left|y_i^{gt} - y_j^{gt}\right| \geq c$. m is the number of all pairs in the same batch, and $M = [y_i^{class} = y_j^{class}]$ is an indicator that is 1 if the videos in the pair have sample class label (homogeneous videos), and 0 otherwise. The margin parameter λ is introduced to control the trade-off between homogeneous loss and non-homogeneous pairs loss.

The first term in the loss function penalizes homogeneous pairs that have wrong relative rank, encouraging the model to learn to distinguish the first kind of hard sample, that is, the fine-grained feature difference of homogeneous data. The second term in the loss function penalizes non-homogeneous pairs that have wrong relative rank and the score label gap is within a certain threshold. Because we found that when the score label gap of non-homogeneous data is too large, the distance between their features will be also very large, and then the penalty relative relationship will affect the final convergence result of the model. Thus the second term encourages the model to learn the second kind of hard sample, punishing pairs with too similar score labels.

By incorporating this Adaptive Rank-Aware Loss, our model is better equipped to handle the complexity of UGC videos, where the quality differences can be subtle and the data distribution is highly varied. This loss function plays a crucial role in improving the overall performance of our VQA framework.

3.4. Training Stability and Model Ensemble

We enhance the robustness and stability of our model training by employing an ensemble of predictions from the Video, Segment, and Frame level models. This ensemble is achieved through a weighted sum of normalized and activated predictions.

$$\mathbf{P}_{\text{scaled}} = \text{Scale}(\sigma(\frac{\mathbf{x} - \bar{x}}{s})) \tag{15}$$

$$\mathbf{P}_{\text{ensemble}} = \sum_{i=1}^{3} w_i \cdot \mathbf{P}_{\text{scaled},i} \tag{16}$$

In this notation, $\mathbf{x}_{\text{norm}}$ represents the normalized prediction vector from the i-th level model, where $\bar{x}$ is the mean and s is the standard deviation of the feature vector $\mathbf{x}$. The sigmoid function σ is applied to $\mathbf{x}_{\text{norm}}$ to obtain $\mathbf{P}_{\text{sigmoid}}$, which maps the normalized predictions into the range (0, 1).

The scaling function Scale then adjusts the sigmoid-activated predictions $\mathbf{P}_{\text{sigmoid}}$ to the desired range, which is from 0 to 5 in this case. The scaled predictions $\mathbf{P}_{\text{scaled},i}$ are combined with their respective weights w_i, which are determined based on the model's validation performance.

The final ensemble prediction $\mathbf{P}_{\text{ensemble}}$ is computed as a weighted sum of the scaled predictions from the Video, Segment, and Frame levels. This ensemble approach leverages the strengths of each level to produce a more accurate and robust video quality assessment.

4. Experiments

4.1. Training Strategies

In the training phase, we tailored our approach for each level of the model. At the Video Level, we set the batch size to 4 and assigned a weight of 0.3 to the rank loss. We employed the AdamW optimizer with an initial learning rate of 1×10^{-3}. Following a warm-up period spanning 3 epochs, the learning rate was modulated using a cosine decay schedule. The weight decay for the optimizer was configured at 1×10^{-2}, and the model underwent training for a total of 30 epochs.

For the Segments Level, training was conducted for 10 epochs with a batch size of 20. The learning rate was initialized at 1×10^{-5} and decremented by a factor of 0.95 every 2 epochs.

At the Frame Level, we utilized a batch size of 30 and initialized the learning rate at 4×10^{-4}. The AdamW optimizer was set with a weight decay of 0.01, and the training encompassed 30 epochs in total.

At the Video and Segment levels, we harness the power of the Swin Transformer. Initially, the Swin Transformer undergoes pre-training for 20 epochs on the Large-scale Structural Video Quality (LSVQ) dataset, necessitated by the limited data volume provided by the competition. This pre-training phase equips the model with generalizable features essential for Video Quality Assessment (VQA). We then proceed to fine-tune the Swin Transformer on our proprietary dataset, a critical step that refines its feature representation capabilities. This fine-tuning process ensures that the model is finely attuned to the specific characteristics of our content, aligning its features with the unique attributes of our dataset.

Additionally, To capture the rich dynamic content within videos more effectively, we have employed the SlowFast model at the segment level. This model samples the video at different temporal densities, generating two complementary feature streams: a fast feature stream that captures rapid motion and transient changes at a high sampling rate, and a slow feature stream that provides broader spatial context information at a lower sampling rate. This dual-stream approach allows us to comprehensively capture and analyze motion features within user-generated content (UGC) videos, thereby enabling a more accurate assessment of video quality.

Prior research has highlighted the advantages of increasing fragment sizes for more effectively capturing local information. In line with this, we have adopted a strategy similar to that used in the ZOOM-VQA framework[48], which involves patch head expansion. Specifically, we have set the patch size to 6 and applied zero-padding around the existing convolutional kernels to accommodate the enlarged patches. This enhancement allows our model to more accurately capture subtle local features present in the UGC video content.

During the training phase, we implement a rigorous five-fold cross-validation procedure to ensure the robustness and generalizability of our models. This methodological choice facilitates a thorough exploration of the model's performance across different subsets of the data, thereby guarding against overfitting and enhancing the reliability of our results.

The implementation details, including the programming language and hardware specifications, are as follows:

- Platform: PyTorch 2.2.1
- Language: Python 3.8.8
- CUDA Version: 12.2
- Hardware: 32G V100

4.2. Experiment Results

In this section, we present the results of our experiments and compare the performance of various methods using different metrics. Two common evaluation metrics for performance comparison, Spearman Rank Order Correlation Coefficient (SROCC) and Pearson Linear Correlation Coefficient (PLCC) are calculated as follows:

$$SROCC = 1 - \frac{6\sum_{i=1}^{N}(d_i^2)}{(N(N^2-1))}, \tag{17}$$

$$PLCC = \frac{\sum_{i=1}^{N}(s_i - \bar{s})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2}\sqrt{\sum_{i=1}^{N}(p_i - \bar{p})^2}}, \tag{18}$$

where d_i is the difference in ranks, s_i and p_i are the scores for the predicted and ground truth, respectively, and $\bar{s}$ and $\bar{p}$ are the mean scores.

In the KVQ dataset, two additional ranking metrics are added due to the presence of homogeneous and non-homogeneous data. Rank1(ranking score in homogeneous data) and Rank2(ranking score in non-homogeneous data) use the following formulas to calculate:

$$\text{Rank1} = \frac{\sum_{i=1}^{S_1}\mathbb{1}(r_{y_{\text{pre}}} = r_{y_{\text{gt}}})}{\sum_{i=1}^{S_1}(\mathbb{1}(r_{y_{\text{pre}}} = r_{y_{\text{gt}}}) + \mathbb{1}(r_{y_{\text{pre}}} \neq r_{y_{\text{gt}}}))}, \tag{19}$$

$$\text{Rank2} = \frac{\sum_{i=1}^{S_2}\mathbb{1}(r_{y_{\text{pre}}} = r_{y_{\text{gt}}})}{\sum_{i=1}^{S_2}(\mathbb{1}(r_{y_{\text{pre}}} = r_{y_{\text{gt}}}) + \mathbb{1}(r_{y_{\text{pre}}} \neq r_{y_{\text{gt}}}))}, \tag{20}$$

where $r_{y_{\text{pre}}}$ and $r_{y_{\text{gt}}}$ are the ranks of the predicted and ground truth, respectively, and $\mathbb{1}$ is the indicator function. S_1 contains 250 homogeneous video pairs of selected, while S_2 contains 250 non-homogeneous video pairs for ranking labeling. Rank1 and Rank2 represent the scores of two sample scenarios(homogeneous pairs and non-homogeneous

pairs but the difference of score is less than 0.5) where the relative rank is difficult to distinguish.

The final score is computed as a weighted sum of these metrics:

$$\text{Final\_Score} = 0.45 \cdot \text{PLCC} + 0.45 \cdot \text{SROCC} + 0.05 \cdot \text{Rank1} + 0.05 \cdot \text{Rank2}. \quad (21)$$

In other datasets, the Rank1 and Rank2 score is not calculated.

Based on the above metrics, our framework achieves 5th place in the NTIRE 2024 Short-form UGC Video Quality Assessment Challenge[16]. We report the results about the NTIRE 2024 Short-form UGC Video Quality Assessment Challenge in Table 1.

Table 1. The comparison of test accuracy of different methods. The best results are **bolded**.The organizers of the competition did not provide the actual ranking on the Val set, so we found the corresponding score on the Val set ranking based on the username on the test set.

Team Name	Val Score	Test Score	Final Ranking
SJTU MMLab	0.9087	0.9228	1
IH-VQA	0.9088	0.9145	2
TVQE	0.8115	0.9120	3
BDVQAGroup	0.9090	0.9116	4
Ours	**0.9054**	**0.8932**	**5**
MC2 Lab	0.8857	0.8855	6
Padding	0.8651	0.8689	7
ysy0129	0.8620	0.8655	8
lizhibo	0.8616	0.8641	9
YongWu	0.8304	0.8555	10
we are a team JH_Chen	0.8239	0.8242	11
dulan	0.8139	0.8098	12
D-H dzx	0.8447	0.7677	13

We also compare with current other SOTA VQA methods on the three UGC VQA databases are shown in Table 2. We can see that our approach works best on KVQ, KoNViD-1K and YouTube-VQC datasets. This proves the superiority and generalization of our framework.

Table 2. The comparison of test accuracy of different methods. The "N/A" means missing corresponding results in the original paper.The best results are **bolded**.

Method	KVQ		KoNViD-1k		YouTube-VQC	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
VIQE[49]	0.221	0.397	0.628	0.638	0.513	0.476
TLVQM[10]	0.490	0.509	0.773	0.768	0.669	0.659
RAPIQUE[36]	0.740	0.717	0.803	0.817	0.759	0.768
VIDEVAL[33]	0.369	0.639	0.773	0.768	0.669	0.659
VSFA[13]	0.762	0.765	0.773	0.775	0.724	0.743
GSTVQA[3]	0.786	0.781	0.814	0.825	N/A	N/A
PVQ[47]	0.794	0.801	0.791	0.786	N/A	N/A
SimpleVQA[27]	0.840	0.847	0.856	0.860	0.847	0.856
FastVQA[39]	0.832	0.834	0.891	0.892	0.855	0.852
KSVQ[19]	0.867	0.869	0.922	0.921	0.900	0.912
Ours	**0.903**	**0.907**	**0.934**	**0.931**	**0.911**	**0.931**

4.3. Ablation Studies

In this section, we analyze the effectiveness of the proposed framework by conducting ablation studies on the KVQ dataset. We evaluate six major components: frame branch, segment branch, video branch, redistribution ensemble, adaptive rank loss, and data augmentation.The results are shown in the Table 3. Model 1 (M1) only uses frame branch to gain quality score. Model 2 (M2) uses frame branch and segment branch, while model 3 (M3) also utilizes video branch beside the above ones. M2 and M3 use direct model ensemble, while M4 uses redistribution model ensemble method. Compared to M4, Model 5 (M5) adds adaptive rank loss. Finally, M6 is the method we proposed.

Effectiveness of Multi-Level. Comparing M1, M2 and M3, it can be observed that multi-level models fusion and ensemble enable the framework to perceive video quality more effectively.

Effectiveness of Redistribution ensemble. Comparing M3 with M4, we can find that Redistribution ensemble makes sense. This indicates that the distribution between multi-level models may be different, and therefore distribution alignment is required for uniform aggregation.

Effectiveness of Adaptive rank loss. Comparing M4 with M6, adaptive rank loss can capture the relative rank relationship between two kind of hard cases.

Effectiveness of Data Augmentation. Comparing M4 with M6, Data augmentation can improve the performance and robustness of the framework, both in terms of spatial domain and temporal domain.

Table 3. The comparison of test accuracy of different methods. The best results are **bolded**.

Method	Frame	Segment	Video	Redistribution	Adaptive rank loss	Data Augmentation	KVQ	
							SROCC	PLCC
M1	✓					✓	0.851	0.856
M2	✓	✓				✓	0.866	0.875
M3	✓	✓	✓			✓	0.879	0.868
M4	✓	✓	✓	✓		✓	0.883	0.872
M5	✓	✓	✓	✓	✓		0.892	0.894
M6	✓	✓	✓	✓	✓	✓	0.903	0.907

5. Conclusion

In this paper, we introduce a novel three-level integration framework for short-form UGC video quality assessment, addressing the dynamic quality variations and diverse content nature. Our approach integrates global and local features across video, segment, and frame levels, enhanced by an adaptive relative rank loss function for fine-grained distinctions between hard samples. Achieving 5th place in the NTIRE 2024 Challenge, our contributions are expected to propel advancements in video quality assessment models.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[3] Baoliang Chen, Lingyu Zhu, Guo Li, Fangbo Lu, Hongfei Fan, and Shiqi Wang. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1903–1916, 2021. 8

[4] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. Perceptual image quality assessment with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 433–442, 2021. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[7] 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. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[9] 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 *2017 Ninth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2017. 2

[10] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019. 2, 8

[11] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019. 2

[12] 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. 2

[13] 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. 2, 8

[14] 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. 2

[15] 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. 2

[16] Xin Li, Kun Yuan, Yajing Pei, Yiting Lu, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. Ntire 2024 challenge on short-form ugc video quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 8

[17] 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. 2

[18] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022. 3

[19] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos, 2024. 1, 6, 8

[20] Kede Ma and Yuming Fang. Image quality assessment in the modern age. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 5664–5666, New York, NY, USA, 2021. Association for Computing Machinery. 1

[21] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708, 2012. 2

[22] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 2

[23] 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. 2

[24] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 2

[25] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, 2012. 2

[26] 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. 2

[27] 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. 1, 2, 8

[28] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2

[29] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[30] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[31] 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 *2020 IEEE International Conference on Image Processing (ICIP)*, pages 141–145. IEEE, 2020. 2

[32] 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. 2

[33] 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. 8

[34] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan Bovik. Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, PP:1–1, 2021. 2

[35] Zhengzhong Tu, Xiangxu Yu, 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. 2

[36] Zhengzhong Tu, Xiangxu Yu, 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. 8

[37] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019. 2

[38] 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 *European conference on computer vision*, pages 538–554. Springer, 2022. 2

[39] 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 *European conference on computer vision*, pages 538–554. Springer, 2022. 8

[40] 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*, 2023. 2

[41] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives, 2023. 2

[42] Wei Wu, Shuming Hu, Pengxiang Xiao, Sibin Deng, Yilin Li, Ying Chen, and Kai Li. Video quality assessment based on swin transformer with spatio-temporal feature fusion and data augmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1846–1854, 2023. 1

[43] Lei Xiao, Anton Kaplanyan, Alexander Fix, Matthew Chapman, and Douglas Lanman. Deepfocus: learned image synthesis for computational displays. *ACM Trans. Graph.*, 37 (6), 2018. 1

[44] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 2

[45] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1098–1105. IEEE, 2012. 2

[46] Zhenqiang Ying, Maniratnam Mandal, Deepti 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. 2

[47] Zhenqiang Ying, Maniratnam Mandal, Deepti 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. 8

[48] Kai Zhao, Kun Yuan, Ming Sun, and Xing Wen. Zoom-vqa: Patches, frames and clips integration for video quality assessment, 2023. 7

[49] Qi Zheng, Zhengzhong Tu, Xiaoyang Zeng, Alan C Bovik, and Yibo Fan. A completely blind video quality evaluator. *IEEE Signal Processing Letters*, 29:2228–2232, 2022. 8