Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives

Haoning Wu*[1] Erli Zhang*[1] Liang Liao*[1] Chaofeng Chen[1]
Jingwen Hou[1] Annan Wang[1] Wenxiu Sun[2] Qiong Yan[2] Weisi Lin[1]
[1] S-Lab, Nanyang Technological University [2] Sensetime Research and Tetras AI

Abstract

The rapid increase in user-generated content (UGC) videos calls for the development of effective video quality assessment (VQA) algorithms. However, the objective of the UGC-VQA problem is still ambiguous and can be viewed from two perspectives: the ***technical perspective****, measuring the perception of distortions; and the* ***aesthetic perspective****, which relates to preference and recommendation on contents. To understand how these two perspectives affect overall subjective opinions in UGC-VQA, we conduct a large-scale subjective study to collect human quality opinions on the overall quality of videos as well as perceptions from aesthetic and technical perspectives. The collected Disentangled Video Quality Database (****DIVIDE-3k****) confirms that human quality opinions on UGC videos are universally and inevitably affected by both aesthetic and technical perspectives. In light of this, we propose the Disentangled Objective Video Quality Evaluator (****DOVER****) to learn the quality of UGC videos based on the two perspectives. The DOVER proves state-of-the-art performance in UGC-VQA under very high efficiency. With perspective opinions in DIVIDE-3k, we further propose* ***DOVER++****, the first approach to provide reliable clear-cut quality evaluations from a single aesthetic or technical perspective. Code at* `https://github.com/VQAssessment/DOVER`.

1. Introduction

Understanding and predicting human quality of experience (QoE) on diverse in-the-wild videos has been a long-existing and unsolved problem. Recent Video Quality Assessment (VQA) studies have gathered enormous human quality opinions [1–5] on in-the-wild user-generated contents (UGC) and attempted to use machine algorithms [6–8] to learn and predict these opinions, known as the **UGC-VQA problem** [9]. However, due to the diversity of contents in UGC videos and the lack of reference videos during subjective studies, these human-quality opinions are still ambiguous and may relate to different perspectives.

*The authors contribute equally to this paper.

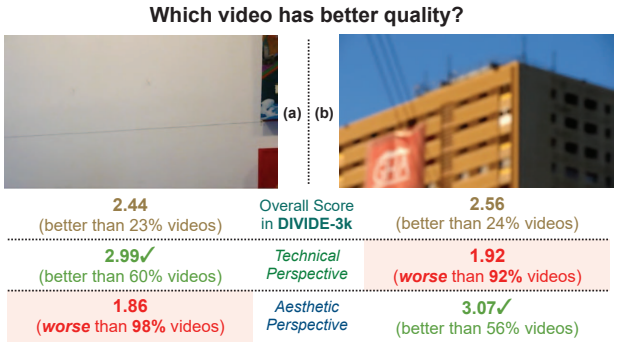


Figure 1. **Which video has better quality**: a clear video with meaningless contents **(a)** or a blurry video with meaningful contents **(b)**? Viewing from different perspectives (***aesthetic***/***technical***) may produce different judgments, motivating us to collect **DIVIDE-3k**, which is the first UGC-VQA dataset with opinions from multiple perspectives. More multi-perspective quality comparisons in our dataset are shown in *supplementary* ***Sec. A***.

Conventionally, VQA studies [9–13] are concerned with the ***technical perspective***, aiming at measuring distortions in videos (*e.g.*, *blurs, artifacts*) and their impact on quality, so as to compare and guide technical systems such as cameras [14, 15], restoration algorithms [16–18] and compression standards [19]. Under this perspective, the video with clear textures in Fig. 1(a) should have notably better quality than the blurry video in Fig. 1(b). On the other hand, several recent studies [2, 6, 7, 20, 21] notice that preferences on non-technical semantic factors (*e.g.*, *contents, composition*) also affect human quality assessment on UGC videos. Human experience on these factors is usually regarded as the ***aesthetic perspective*** [22–27] of quality evaluation, which considers the video in Fig. 1(b) as better quality due to its more meaningful contents and is preferred for content recommendation systems on platforms such as YouTube or TikTok. However, how aesthetic preference plays the impact on final human quality opinions of UGC videos is still debatable [1, 2] and requires further validation.

To investigate the impact of aesthetic and technical perspectives on human quality perception of UGC videos, we conduct the first comprehensive subjective study to collect opinions from both perspectives, as well as overall opinions on a large number of videos. We also conduct subjective

reasoning studies to explicitly gather information on how much each individual's overall quality opinion is influenced by aesthetic and technical perspectives. With overall 450K opinions on 3,590 diverse UGC videos, we construct the first Disentangled Video Quality Database (**DIVIDE-3k**). After calibrating our study on the DIVIDE-3k with existing UGC-VQA subjective studies, we observe that human quality perception on UGC videos is broadly and inevitably ***affected by both aesthetic and technical perspectives***. As a consequence, the overall subjective quality scores between the two videos in Fig. 1 with different qualities from either one of the two perspectives could be similar.

Motivated by the observation from our subjective study, we aim to develop an objective UGC-VQA method that accounts for both aesthetic and technical perspectives. To achieve this, we design the View Decomposition strategy, which divides and conquers aesthetic-related and technical-related information in videos, and propose the Disentangled Objective Video Quality Evaluator (**DOVER**). DOVER consists of two branches, each dedicated to focusing on the effects of one perspective. Specifically, based on the different characteristics of quality issues related to each perspective, we carefully design inductive biases for each branch, including *specific inputs, regularization strategies, and pre-training*. The two branches are supervised by the overall scores (affected by both perspectives) to adapt for existing UGC-VQA datasets [1, 3, 4, 28–30], and additionally supervised by aesthetic and technical opinions exclusively in the **DIVIDE-3k** (denoted as **DOVER++**). Finally, we obtain the overall quality prediction via a subjectively-inspired fusion of the predictions from the two perspectives. With the subjectively-inspired design, the proposed DOVER and DOVER++ not only reach better accuracy on the overall quality prediction but also provide more reliable quality prediction from aesthetic and technical perspectives, catering for practical scenarios.

Our contributions can be summarized as four-fold:

1) We collect the **DIVIDE-3k** (3,590 videos), the first UGC-VQA database that contains 450,000 subjective quality opinions from aesthetic and technical perspectives as well as their effects on overall quality scores.
2) By analyzing opinions, we observe that human quality perception is broadly affected by both aesthetic and technical perspectives in the UGC-VQA problem, better explaining the human perceptual mechanism on it.
3) We propose the **DOVER**, a subjectively-inspired video quality evaluator with two branches focusing on aesthetic and technical perspectives. The DOVER demonstrates state-of-the-arts on the **all** UGC-VQA datasets.
4) Our methods can provide quality predictions from a single perspective, which can be applied as metrics for camera systems (***technical***) or content recommendation (***aesthetic***), or for personalized VQA (Sec. 5.5).

2. Related Works

Databases and Subjective Studies on UGC-VQA. Unlike traditional VQA databases [28, 29, 31, 32], UGC-VQA databases [1, 3–5] directly collect from real-world videos from direct photography, YFCC-100M [33] database or YouTube [30] videos. With each video having unique content and being produced by either professional or non-professional users [7, 8], quality assessment of UGC videos can be more challenging and less clear-cut compared to traditional VQA tasks. Additionally, the subjective studies in UGC-VQA datasets are usually carried out by crowd-sourced users [34] with no reference videos. These factors may lead to the ambiguity of subjective quality opinions in UGC-VQA which can be affected by different perspectives.

Objective Methods for UGC-VQA. Classical VQA methods [9–12, 35–42] employ handcrafted features to evaluate video quality. However, they do not take the effects of semantics into consideration, resulting in reduced accuracy on UGC videos. Noticing that UGC-VQA is deeply affected by semantics, deep VQA methods [2, 6, 13, 43–50] are becoming predominant in this problem. For instance, VSFA [6] conducts subjective studies to demonstrate videos with attractive content receive higher subjective scores. Therefore, it uses the semantic-pretrained ResNet-50 [51] features instead of handcrafted features, followed by plenty of recent works [1, 2, 13, 21, 52–54] that improve the performance for UGC-VQA. However, these methods, which are directly driven by ambiguous subjective opinions, can hardly explain what factors are considered in their quality predictions, hindering them from providing reliable and explainable quality evaluations on real-world scenarios (*e.g.*, distortion metrics and recommendations).

3. The DIVIDE-3k Database

In this section, we introduce the proposed Disentangled Video Quality Database (**DIVIDE-3k**, Fig. 2), along with the multi-perspective subjective study. The database includes 3,590 UGC videos, on which we collected 450,000 human opinions. Different from other UGC-VQA databases [1–3], the subjective study is conducted in-lab to reduce the ambiguity of perspective opinions.

3.1. Collection of Videos

Sources of Videos. The 3,590-video database is mainly collected from two sources: 1) the YFCC-100M [33] social media database; 2) the Kinetics-400 [55] video recognition database, collected from YouTube, which has in total 400,000 videos. Voices are removed from all videos.

Getting the subset for annotation. Similar to existing studies [1, 3], we would like the sampled video database able to represent the overall quality of the original larger database. Therefore, we first histogram all 400,000 videos

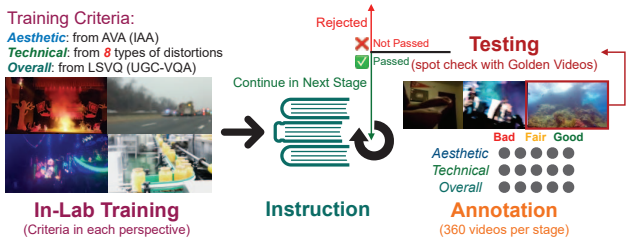


Figure 2. The in-lab subjective study on videos in **DIVIDE-3k**, including Training, Instruction, Annotation and Testing, discussed in Sec. 3.2.

with spatial [11], temporal [12], and semantic indices [56]. Then, we randomly select a subset of 3,270 videos from the 400,000 videos that match the histogram from the three dimensions [57] as in [1, 3]. Several examples from DIVIDE-3k are provided in the supplementary. We also select 320 videos from the LSVQ [1], the most recent UGC-VQA database, to examine the calibration between DIVIDE-3k and existing UGC-VQA subjective studies (see in Tab. 2).

3.2. In-lab Subjective Study on Videos

To ensure a clear understanding of the two perspectives, we conduct in-lab subjective experiments instead of crowdsourced, with 35 trained annotators (including 19 male and 16 female) participating in the full annotation process of Training, Testing and Annotation. All videos are downloaded to local computers before annotation to avoid transmission errors. The main process of the subjective study is illustrated in Fig. 2, discussed step-by-step as follows. *Extended details about the study are in supplementary* ***Sec. A***.

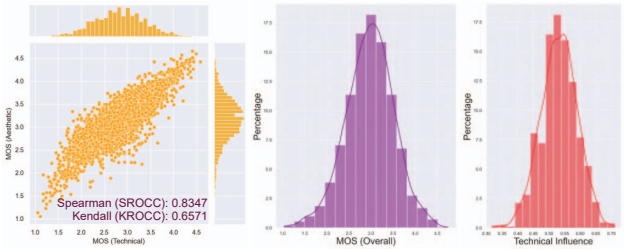
Training. Before annotation, we provide clear criteria with abundant examples of the three quality ratings to train the annotators. For ***aesthetic rating***, we select example images with *good*, *fair* and *bad* aesthetic quality from the aesthetic assessment database AVA [22], each for 20 images, as calibration for aesthetic evaluation. For ***technical rating***, we instruct subjects to rate purely based on technical distortions and provide 5 examples for each of the following eight common distortions: *1) noises; 2) artifacts; 3) low sharpness; 4) out-of-focus; 5) motion blur; 6) stall; 7) jitter; 8) over/under-exposure.* For overall quality rating, we select 20 videos each with *good*, *fair* and *bad* quality as examples, from the UGC-VQA dataset LSVQ [1].

During Experiment: Instruction and Annotation. We divide the subjective experiments into 40 videos per group, and 9 groups per stage. Before each stage, we instruct the subjects on how to label each specific perspective:

- **Aesthetic Score**: Please rate the video's quality based on aesthetic perspective (*e.g.*, semantic preference).
- **Technical Score**: Please rate the video's quality with only consideration of technical distortions.
- **Overall Score**: Please rate the quality of the video.
- **Subjective Reasoning**: Please rate how **your** overall score is impacted by aesthetic or technical perspective.

Specifically, for the subjective reasoning, subjects need to rate the proportion of ***technical*** impact in the overall score for each video among $[0, 0.25, 0.5, 0.75, 1]$, while rest proportion is considered as ***aesthetic*** impact.

Testing with Golden Videos. For testing, we randomly insert 10 **golden videos** in each stage as a spot check to ensure the quality of annotation, and the subject will be rejected and not join the next stage if the annotations on the golden videos severely deviate from the standards.



(a) Correlation between ***Aesthetic*** and ***Technical*** scores **(b)** Distribution of **Overall** scores **(c)** Distribution of **Subjective Reasonings**

Figure 3. **Statistics in DIVIDE-3k**: **(a)** The correlations between aesthetic and technical perspectives, and distributions **(b)** of overall quality (MOS) & **(c)** subject-rated proportion of technical impact on overall quality.

Table 1. **Effects of Perspectives**: The correlations between different perspectives and overall quality (MOS) for all 3,590 videos in **DIVIDE-3k**.

Correlation to MOS	MOS_A	MOS_T	$\text{MOS}_\text{A} + \text{MOS}_\text{T}$	$0.428\text{MOS}_\text{A} + 0.572\text{MOS}_\text{T}$
Spearman (SROCC↑)	0.9350	0.9642	0.9827	**0.9834**
Kendall (KROCC↑)	0.7894	0.8455	0.8909	**0.8933**

Table 2. **Calibration with Existing**: The correlations of between different ratings in **DIVIDE-3k** and existing scores in LSVQ [1] ($\text{MOS}_\text{existing}$).

Correlation to $\text{MOS}_\text{existing}$	MOS_A	MOS_T	$\text{MOS}_\text{A} + \text{MOS}_\text{T}$	MOS
Spearman (SROCC↑)	0.6956	0.7374	0.7632	**0.7680**
Kendall (KROCC↑)	0.5073	0.5469	0.5797	**0.5822**

3.3. Observations

Effects of Two Perspectives. To validate the effects of two perspectives, we first quantitatively assess the correlation between the two perspectives and overall quality. Denote the mean aesthetic opinion as MOS_A, mean technical opinion as MOS_T, mean overall opinion as MOS, the Spearman and Kendall correlation between different perspectives are listed in Tab. 1. From Tab. 1, we notice that the weighted sum of both perspectives is a better approximation of overall quality than either single perspective. Consequently, methods [1, 6, 58] that naively regress from overall MOS might not provide pure technical quality predictions due to the inevitable effect of aesthetics. The best/worst videos (Fig. 4) in each dimension also support this observation.

Calibration with Existing Study. To validate whether the observation can be extended for existing UGC-VQA subjective studies, we select 320 videos from LSVQ [1] to compare quality opinions from multi-perspectives with existing scores of these videos. As shown in Tab. 2, the overall quality score is more correlated with the existing score than scores from either perspective, further suggesting considering human quality opinion as a fusion of both perspectives might be a better approximation in the UGC-VQA problem.

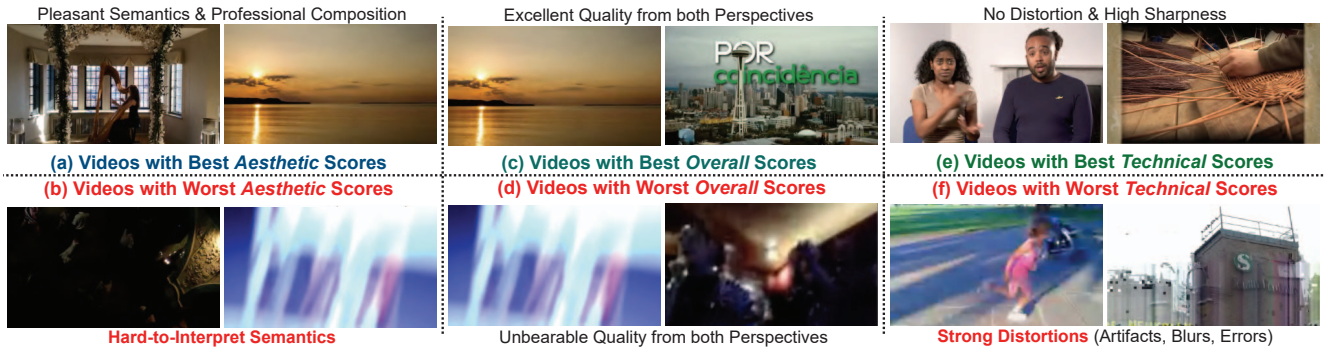


Figure 4. Videos with best and worst scores in aesthetic perspective, technical perspective and overall quality perception in the **DIVIDE-3k**. The aesthetic perspective is more concerned with semantics or composition of videos, while the technical perspective is more related to low-level textures and distortions.

Subjective Reasoning. In the DIVIDE-3k, we conducted the first subjective reasoning study during the human quality assessment. Fig. 3(c) illustrates the mean technical impact for each video, ranging among $[0.364, 0.698]$. The results of reasoning further explicitly validate our aforementioned observation, that human quality assessment is affected by opinions from both aesthetic and technical perspectives.

4. The Approaches: DOVER and DOVER++

Observing that overall quality opinions are affected by both aesthetic and technical perspectives from subjective studies in **DIVIDE-3k**, we propose to distinguish and investigate the aesthetic and technical effects in a UGC-VQA model based on the View Decomposition strategy (Sec. 4.1). The proposed Disentangled Objective Video Quality Evaluator (**DOVER**) is built up with an aesthetic branch (Sec. 4.2) and a technical branch (Sec. 4.3). The two branches are separately supervised, either both by overall scores (denoted as **DOVER**) or by respective aesthetic and technical opinions (denoted as **DOVER++**), discussed in Sec. 4.4. Finally, we discuss the subjectively-inspired fusion (Sec. 4.5) to predict the overall quality from DOVER.

4.1. Methodology: Separate the Perceptual Factors

From DIVIDE-3k, we notice that aesthetic and technical perspectives in UGC-VQA are usually associated with different perceptual factors. Specifically, as illustrated in (Fig. 4**(a)&(b)**), aesthetic opinions are mostly related to *semantics, composition* of objects [24, 27, 59], which are typically high-level visual perceptions. In contrast, the technical quality is largely affected by low-level visual distortions, *e.g.*, *blurs, noises, artifacts* [1, 13, 21, 60, 61] (Fig. 4**(e)&(f)**).

The observation inspires the View Decomposition strategy that separates the video into two views: the **Aesthetic View** (S_{A}) that focus on aesthetic perception, and **Technical View** (S_{T}) for vice versa. With the decomposed views as inputs, two separate aesthetic ($\mathbf{M}_{\mathrm{A}}$) and technical branches ($\mathbf{M}_{\mathrm{T}}$) evaluate different perspectives separately:

$$Q_{\mathrm{pred,A}} = \mathbf{M}_{\mathrm{A}}(S_{\mathrm{A}}); Q_{\mathrm{pred,T}} = \mathbf{M}_{\mathrm{T}}(S_{\mathrm{T}}) \quad (1)$$

Despite that most perception related to the two perspectives can be separated, a small proportion of perceptual factors are related to both perspectives, such as **brightness** related to both *exposure (technical)* [29] and *lighting (aesthetic)* [26], or **motion blurs** (which is occasionally considered as *good aesthetics* but typically regarded as *bad technical quality* [62]). Thus, we don't separate these factors and keep them in both branches. Instead, we employ inductive biases (*pre-training, regularization*) and specific supervision in the DIVIDE-3k to further drive the two branches' focus on corresponding perspectives, introduced as follows.

4.2. The Aesthetic Branch

To help the aesthetic branch focus on the aesthetic perspective, we first pre-train the branch with Image Aethetic Assessment database AVA [22]. We then elaborate the Aesthetic View (S_{A}) and additional regularization objectives.

The Aesthetic View. *Semantics* and *Composition* are two key factors deciding the aesthetics of a video [24, 59, 63]. Thus, we obtain Aesthetic View (see Fig. 5(b)) through *spatial downsampling* [64] and *temporal sparse frame sampling* [65] which preserves the semantics and composition of original videos. The downsampling strategies are widely applied in many existing state-of-the-art aesthetic assessment methods [24, 25, 27, 66–68], further proving that they are able to preserve aesthetic information in visual contents. Moreover, the two strategies significantly reduce the sensitivity [9–12] on technical distortions such as *blurs, noises, artifacts* (via spatial downsampling), *shaking, flicker* (via temporal sparse sampling), so as to focus on aesthetics.

Cross-scale Regularization. To better reduce technical impact in this branch, we obtain the over-downsampled view ($S_{\mathrm{A}\downarrow}$) during training by over-downsampling the videos with up to $11.3\times$ downscaling ratio. We then observe that the $S_{\mathrm{A}\downarrow}$ can barely keep any technical distortions but still remains similar aesthetics with S_{A} and even the original videos (see Fig. 5(b) *upper-right*). Furthermore, conclusions from existing study [69] suggest that feature dissimilarity among different scales (*e.g.*, $S_{\mathrm{A}\downarrow}$ and S_{A}) is related

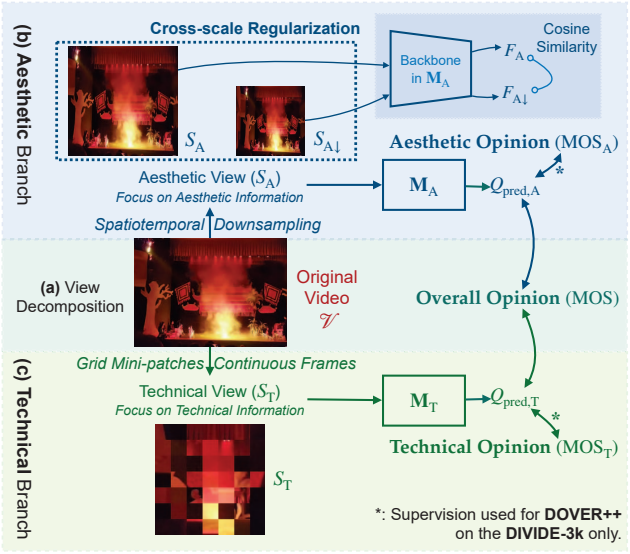


Figure 5. The proposed Disentangled Objective Video Quality Evaluator (**DOVER**) and **DOVER++** via **(a)** View Decomposition (Sec. 4.1), with the **(b)** Aesthetic Branch (Sec. 4.2) and the **(c)** Technical Branch (Sec. 4.3). The equations to obtain the two views are in Supplementary **Sec. E**.

to technical distortions. Henceforth, we impose the Cross-scale Restraint ($\mathcal{L}_{\text{CR}}$, Fig. 5(e)) as a regularization to further reduce the technical influences in the aesthetic prediction by encouraging the feature similarity between $S_{\text{A}\downarrow}$ and S_{A}:

$$\mathcal{L}_{\text{CR}} = 1 - \frac{F_{\text{A}} \cdot F_{\text{A}\downarrow}}{\|F_{\text{A}}\|\|F_{\text{A}\downarrow}\|} \tag{2}$$

where F_{A} and $F_{\text{A}\downarrow}$ are output features for S_{A} and $S_{\text{A}\downarrow}$.

4.3. The Technical Branch

In the technical branch, we would like to keep the technical distortions but obfuscate the aesthetics of the videos. Thus, we design the Technical View (S_{T}) as follows.

The Technical View. We introduce *fragments* [13] (as in Fig. 5(c)) as Technical View (S_{T}) for the technical branch. The *fragments* are composed of randomly cropped patches stitched together to retain the technical distortions. Moreover, it discarded most content and disrupted the compositional relations of the remaining, therefore severely corrupting aesthetics in videos. Temporally, we apply *continuous frame sampling* for S_{T} to retain temporal distortions.

Weak Global Semantics as Background. Many studies [13, 60, 70] suggest that technical quality perception should consider global semantics to better assess distortion levels. Though most content is discarded in S_{T}, the technical branch can still reach 68.6% accuracy for Kinetics-400 [55] video classification, indicating it can preserve weak global semantics as background information to distinguish textures (*e.g.*, sands) from distortions (*e.g.*, noises).

4.4. Learning Objectives

Weak Supervision with Overall Opinions. With the observation in Sec. 3.3, the overall MOS can be approximated as a weighted sum of MOS_{A} and MOS_{T}. Moreover, the subjectively-inspired inductive biases in each branch can reduce the perception of another perspective. The two observations suggest that if we use overall opinions to separately supervise the two branches, the prediction of each branch could be majorly decided by its corresponding perspective. Henceforth, we propose the Limited View Biased Supervisions ($\mathcal{L}_{\text{LVBS}}$), which minimize the relative loss* between predictions in each branch with the overall opinion MOS, as the objective of DOVER, applicable on all databases:

$$\begin{aligned}\mathcal{L}_{\text{LVBS}} = \mathcal{L}_{\text{Rel}}(Q_{\text{pred,A}}, \text{MOS}) \\ + \mathcal{L}_{\text{Rel}}(Q_{\text{pred,T}}, \text{MOS}) + \lambda_{\text{CR}}\mathcal{L}_{\text{CR}}\end{aligned} \tag{3}$$

Supervision with Opinions from Perspectives. With the DIVIDE-3k database, we further improve the accuracy for disentanglement with the Direct Supervisions ($\mathcal{L}_{\text{DS}}$) on corresponding perspective opinions for both branches:

$$\mathcal{L}_{\text{DS}} = \mathcal{L}_{\text{Rel}}(Q_{\text{pred,A}}, \text{MOS}_{\text{A}}) + \mathcal{L}_{\text{Rel}}(Q_{\text{pred,T}}, \text{MOS}_{\text{T}}) \tag{4}$$

and the proposed **DOVER++** is driven by a fusion of the two objectives to jointly learn more accurate overall quality as well as perspective quality predictions for each branch:

$$\mathcal{L}_{\text{DOVER++}} = \mathcal{L}_{\text{DS}} + \lambda_{\text{LVBS}}\mathcal{L}_{\text{LVBS}} \tag{5}$$

4.5. Subjectively-inspired Fusion Strategy

From the subjective studies, we observe that the MOS can be well-approximated as $0.428\text{MOS}_{\text{A}} + 0.572\text{MOS}_{\text{T}}$. Henceforth, we propose to similarly obtain the final overall quality prediction (Q_{pred}) from two perspectives: $Q_{\text{pred}} = 0.428Q_{\text{pred,A}} + 0.572Q_{\text{pred,T}}$ via a simple weighted fusion. With better accuracy on all datasets (Tab. 9), the strategy by side validates the subjective observations in Sec. 3.3.

5. Experimental Evaluation

In this section, we answer two important questions:

- Can the aesthetic and technical branches better learn the effects of corresponding perspectives (Sec. 5.2)?
- Can the fused model more accurately predict overall quality in UGC-VQA problem (Sec. 5.3)?

Moreover, we include ablation studies (Sec. 5.4) and an outlook for personalized quality evaluation (Sec. 5.5).

5.1. Experimental Setups

Implementation Details. In the aesthetic branch, we use S_{A} with size 224×224 during inference and over-downsampled $S_{\text{A}\downarrow}$ size 128×128 to better exclude technical

*A criterion [71] based on the linear and rank correlation between predictions and labels. Details provided in supplementary **Sec. E**.

Table 3. **Quantitative Evaluation on Perspectives** of DOVER (weakly-supervised) and DOVER++ (fully-supervised) in the DIVIDE-3k, by evaluating the correlation across different predictions and subjective opinions. *w/o* Decomposition denotes both branches with original videos as inputs.

Method	SROCC/PLCC	MOS_A	MOS_T
w/o Decomposition ($w/$ MOS_A&MOS_T)	$Q_{pred,A}$	0.7482/0.7576	0.7941/0.8039
	$Q_{pred,T}$	0.7234/0.7430	0.8190/0.8233
DOVER *w/o* MOS_A&MOS_T	$Q_{pred,A}$	0.7489/0.7607	0.7877/0.8044
	$Q_{pred,T}$	0.7153/0.7382	0.8213/0.8295
DOVER++ *w/* MOS_A&MOS_T	$Q_{pred,A}$	**0.7683/0.7779**	0.7584/0.7708
	$Q_{pred,T}$	0.7015/0.7230	**0.8376/0.8443**

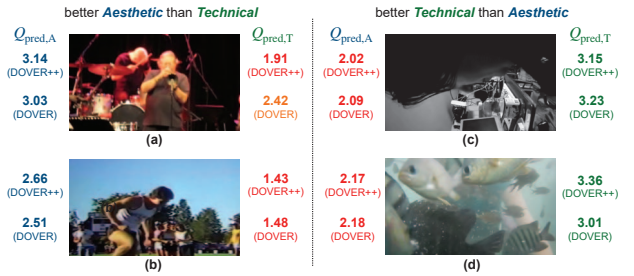


Figure 6. **Qualitative Studies on Perspectives** of DOVER/DOVER++: Visualizations of videos in the **DIVIDE-3k** where aesthetic and technical predictions are diverged. *More visualizations in supplement.* ***Sec. D***.

quality issues. $N = 32$ frames are sampled uniformly from each video and the backbone is inflated-ConvNext [72] Tiny pre-trained with AVA [22]. In the technical branch, we crop single patches at size $S_f = 32$ from 7×7 spatial grids and sample a clip of 32 continuous frames during training, and three clips during inference. The backbone of the technical branch is Video Swin Transformer [73] Tiny with GRPB [13]. λ_{CR} is set as 0.3, and λ_{LVBS} is set as 0.5.

Datasets. Despite evaluating DOVER and DOVER++ on the proposed **DIVIDE-3k** (3,590 videos) database, we also evaluate DOVER with the large-scale UGC-VQA dataset, LSVQ [1] (39,072 videos), and on three smaller UGC-VQA datasets, KoNViD-1k [3] (1,200 videos), LIVE-VQC [4] (585 videos), and YouTube-UGC [5] (1,380 videos).

5.2. Evaluation on Two Perspectives

In this section, we quantitatively and qualitatively evaluate the perspective prediction ability of proposed methods in the **DIVIDE-3k** (Sec. 5.2.1). The divergence map and pairwise user studies further prove that the two branches in DOVER better align with human opinions on corresponding perspectives on existing UGC-VQA databases (Sec. 5.2.2).

5.2.1 Evaluation on the DIVIDE-3k

Quantitative Studies. In Tab. 3, we evaluate the cross-correlation between the aesthetic and technical predictions in DOVER or DOVER++ and human opinions from the two perspectives in the DIVIDE-3k, compared with baseline (*with respective labels as supervision, but without View Decomposition*). DOVER shows a stronger perspective preference than the baseline even without using the respective labels, proving the effectiveness of the decomposition strategy. DOVER++ more effectively disentangle the two perspectives with each branch around **7%** more correlated with respective opinions than opinions from another perspective.

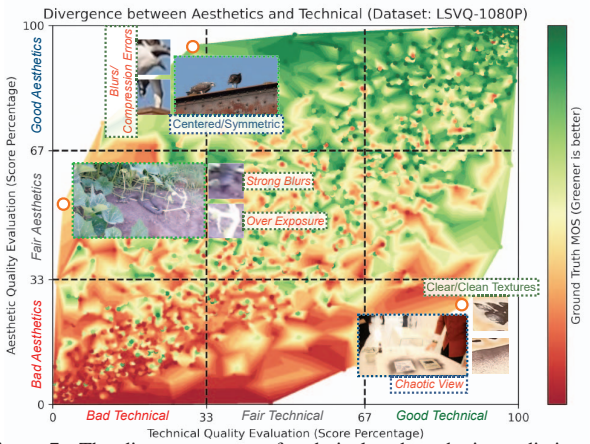


Figure 7. The divergence map of technical and aesthetic predictions of DOVER in LSVQ [1] dataset. Similar as Fig. 6, the videos with diverged scores also align with human opinions of aesthetic and technical quality.

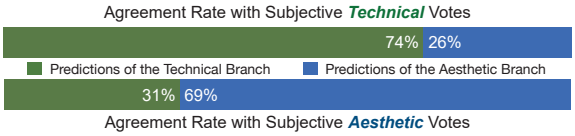


Figure 8. **User Studies on Diverged Pairs** when technical and aesthetic branches in DOVER predict differently, demonstrating that predictions of each branch are more aligned with corresponding subjective opinions.

Qualitative Studies. In Fig. 6, we visualize several videos with diverged predicted aesthetic and technical scores. The two videos with better aesthetic scores (Fig. 6**(a)&(b)**) have clear semantics yet suffer from blurs and artifacts; on the contrary, the two with better technical scores (Fig. 6**(c)&(d)**) are sharp but with chaotic composition and unclear semantics. These examples align with human perception of the two perspectives, proving that both variants can effectively provide disentangled quality predictions.

5.2.2 Evaluation on Existing UGC-VQA Datasets

The Divergence Map. In Fig. 7, we visualize the divergence map between predictions in two branches (trained and tested on LSVQ [1]) and examine the videos where two branches score most differently, noted in ***orange circles***. Among these videos, the aesthetic branch can distinguish between bad (chaotic scene, Fig. 7 *downright*) and good (symmetric view, Fig. 7 *upleft*) aesthetics, while the technical branch can detect technical quality issues (*blurs, over-exposure, compression errors* at Fig. 7 *upleft*).

Pairwise User Studies. We further conduct ***user studies*** to measure whether the two evaluators can distinguish the two perspectives on these diverged cases. Specifically, we evaluate on diverged pairs $\{\mathcal{V}_1, \mathcal{V}_2\}$ where aesthetic branch predicts $\mathcal{V}_1$ is obviously better (*at least one score higher when scores are in the range* $[1, 5]$) yet technical branch predicts $\mathcal{V}_2$ is obviously better. After random sampling 200 pairs

Table 4. Benchmark on official splits on the large-scale UGC-VQA dataset LSVQ [1]. First, second and third bests are labelled in **red**, **blue** and **boldface**.

Training Set: $\text{LSVQ}_{\text{train}}$ [1]	Inference Computational Cost			Intra-dataset Evaluations				Generalization Evaluations			
Testing Set/	*on a 1080P, 10-second video*			**$\text{LSVQ}_{\text{test}}$**		**$\text{LSVQ}_{\text{1080p}}$**		**KoNViD-1k**		**LIVE-VQC**	
Methods	GFLOPs	CPU Time	GPU Time	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
Classical Approaches (based on handcraft features):											
TLVQM (TIP, 2019) [10]	NA	248s	NA	0.772	0.774	0.589	0.616	0.732	0.724	0.670	0.691
VIDEVAL (TIP, 2021) [9]	NA	895s	NA	0.795	0.783	0.545	0.554	0.751	0.741	0.630	0.640
Deep Approaches (based on deep neural network features):											
VSFA (ACM MM, 2019) [6]	40919	466s	11.1s	0.801	0.796	0.675	0.704	0.784	0.795	0.734	0.772
⋆ Patch-$\text{VQ}_{\text{w/o patch}}$ (CVPR, 2021) [1]	58501	539s	13.8s	0.814	0.816	0.686	0.708	0.781	0.781	0.747	0.776
⋆ Patch-$\text{VQ}_{\text{w/ patch}}$ (CVPR, 2021) [1]	– – *same as above* – –			0.827	0.828	0.711	0.739	0.791	0.795	0.770	0.807
⋆ Li *et al.* (TCSVT, 2022) [58]	112537	1567s	27.6s	0.852	0.855	0.771	0.782	0.834	0.837	0.816	0.824
FAST-VQA (ECCV, 2022) [13]	**279.1**	**8.8s**	**45ms**	**0.876**	**0.877**	**0.779**	**0.814**	**0.859**	**0.855**	**0.823**	**0.844**
DOVER (Ours)	**282.3**	**9.7s**	**47ms**	**0.888**	**0.889**	**0.795**	**0.830**	**0.884**	**0.883**	**0.832**	**0.855**
Improvement to existing best	–	–	–	+1.3%	+1.3%	+2.0%	+2.0%	+2.9%	+3.3%	+1.0%	+1.3%

Table 5. Performance benchmark on existing smaller UGC-VQA datasets. All experiments are conducted under 10 train-test splits.

Target (Fine-tuning) Quality Dataset		**LIVE-VQC** (585)		**KoNViD-1k** (1200)		**YouTube-UGC** (1380)		*Weighted Average*	
Methods	**Source (Pre-training) Quality Dataset**	(240P - **1080P**)		(540P)		(360P - **2160P(4K)**)			
		SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
TLVQM (TIP, 2019) [10]	NA (*pure handcraft*)	0.799	0.803	0.773	0.768	0.669	0.659	0.732	0.726
VIDEVAL (TIP, 2021) [9]	NA (*pure handcraft*)	0.752	0.751	0.783	0.780	0.779	0.773	0.772	0.772
RAPIQUE (OJSP, 2021) [74]	*handcraft* + KoNiQ [75]	0.755	0.786	0.803	0.817	0.759	0.768	0.774	0.790
CNN+TLVQM (ACMMM, 2020) [53]	*handcraft* + KoNiQ [75]	0.825	0.834	0.816	0.818	0.809	0.802	0.815	0.814
CNN+VIDEVAL (TIP, 2021) [9]	*handcraft* + KoNiQ [75]	0.785	0.810	0.815	0.817	0.808	0.803	0.806	0.810
VSFA (ACMMM, 2019) [6]	*None*	0.773	0.795	0.773	0.775	0.724	0.743	0.752	0.765
Patch-VQ (CVPR, 2021) [1]	PaQ-2-PiQ [61]	0.827	0.837	0.791	0.786	NA	NA	NA	NA
CoINVQ (CVPR, 2021) [7]	*self-collected*	NA	NA	0.767	0.764	0.816	0.802	NA	NA
Li *et al.* (TCSVT, 2022) [58]	*fused* ([15, 75–77])	0.834	0.842	0.834	0.836	0.818	0.826	0.823	0.833
FAST-VQA (ECCV, 2022) [13]	LSVQ [1]	**0.849**	**0.862**	**0.891**	**0.892**	**0.855**	**0.852**	**0.868**	**0.869**
DOVER (ours)	LSVQ [1]	**0.860**	**0.875**	**0.909**	**0.906**	**0.890**	**0.891**	**0.891**	**0.891**
– *improvement to existing best*		+1.6%	+1.4%	+2.0%	+1.6%	+3.9%	+3.8%	+2.6%	+2.5%

Table 6. Performance benchmark on the **DIVIDE-3k**. All experiments are conducted under 10 train-test splits with random seed 42.

Training/Testing on		**DIVIDE-3k** (3590)		
Methods	**Pre-training Dataset**	SROCC↑	PLCC↑	KROCC↑
TLVQM (2019) [10]	NA (*pure handcraft*)	0.6461	0.6807	0.4699
VIDEVAL (2021) [9]	NA (*pure handcraft*)	0.7056	0.7162	0.5233
RAPIQUE (2021) [74]	*handcraft* + KoNiQ [75]	0.7341	0.7547	0.5498
VSFA (2019) [6]	NA	0.7254	0.7386	0.5395
MDTVSFA (2021) [50]	NA	0.7522	0.7409	0.5647
UNIQUE (2021) [78]	*fused* ([15, 75–77])	0.7529	0.7637	0.5634
Li *et al.* (2022) [58]	*fused* ([15, 75–77])	0.7967	0.8125	0.6138
FAST-VQA (2022) [13]	LSVQ [1]	**0.8184**	**0.8288**	**0.6285**
DOVER (Ours)	LSVQ [1]	**0.8331**	**0.8438**	**0.6480**
DOVER++ (Ours)	LSVQ [1]	**0.8442**	**0.8537**	**0.6603**

Table 7. Zero-shot or cross-dataset evaluations on the **DIVIDE-3k**. None of the listed methods has been trained on the DIVIDE-3k.

Evaluating on		**DIVIDE-3k** (3590)		
Zero-shot (Opinion-Unaware) VQA Approaches:				
Methods	**Training on**	SROCC↑	PLCC↑	KROCC↑
NIQE (2013) [11]	None	0.3524	0.3839	0.2634
TPQI (2022) [12]	None	0.4407	0.4432	0.3045
CLIP-IQA (2022) [56]	CLIP [79]	0.5882	0.5910	0.4067
BVQI (2023) [80]	CLIP [79]	0.6678	0.6802	0.4842
Cross-dataset Evaluation (training on LSVQ):				
Patch-VQ (2021) [1]	LSVQ [1]	0.6454	0.6713	0.4489
Li et al. (2022) [58]		0.7318	0.7524	0.5395
DOVER (Ours)		**0.7727**	**0.7806**	**0.5799**

in this way, we ask 15 subjects to choose **which one has better aesthetic (or technical) quality in the pair**. After post-processing the subject choices with popular votes, we calculate the agreement rates between subjective votes and predictions (in Fig. 8). Each subjective perspective is notably more agreed with corresponding branch predictions, demonstrating that even without the respective labels, the DOVER can still learn to primarily disentangle the two perspectives. *More details are in supplementary (**Sec. B**).*

5.3. Evaluation on Overall Quality Prediction

5.3.1 Results on Existing UGC-VQA Datasets

Results on LSVQ. In Tab. 4, we train the DOVER on the large-scale UGC-VQA dataset, LSVQ [1], and test it on five different existing UGC-VQA datasets. The proposed DOVER outperforms state-of-the-arts for intra-dataset evaluations by improving up to **2.0%** PLCC. When testing on datasets other than LSVQ as generalization evaluation, the DOVER has shown more competitive performance. It improves PLCC on FAST-VQA by **3.3%** on KoNViD-1k, the UGC-VQA dataset with more diverse contents, further suggesting the importance of modelling from the aesthetic perspective in quality assessment on videos of diverse contents.

Results on Smaller UGC-VQA Datasets. Following [13], we pre-train the proposed DOVER on LSVQ instead of IQA datasets [61, 75, 76] and then fine-tune the proposed method on three smaller UGC-VQA datasets and list the results in Tab. 5. DOVER has reached unprecedented performance on all three datasets (mean PLCC > 0.89), and outperformed FAST-VQA with an average of **2.6%** improvement under exactly the same training process. The results further prove the effectiveness of considering aesthetic and technical perspectives separately and explicitly in UGC-VQA.

Table 8. **Ablation Study of DOVER (I):** the View Decomposition scheme.

Testing Set/ Variants/Metric	**LSVQ$_{test}$** SROCC/PLCC	**LSVQ$_{1080p}$** SROCC/PLCC	**KoNViD-1k** SROCC/PLCC	**LIVE-VQC** SROCC/PLCC
w/o Decomposition	0.859/0.858	0.752/0.798	0.851/0.850	0.816/0.834
Feature Aggregation	0.873/0.874	0.776/0.811	0.863/0.864	0.813/0.839
DOVER (Ours)	**0.888/0.889**	**0.795/0.830**	**0.884/0.883**	**0.832/0.855**

Table 9. **Ablation Study of DOVER (II):** Accuracy of single branch predictions and the effect of subjectively-inspired fusion (denoted as *SIF*).

Testing Set/			**LSVQ$_{test}$**	**LSVQ$_{1080p}$**	**KoNViD-1k**	**LIVE-VQC**
$Q_{\text{pred,A}}$	$Q_{\text{pred,T}}$	*SIF*	SROCC/PLCC	SROCC/PLCC	SROCC/PLCC	SROCC/PLCC
✓			0.855/0.856	0.738/0.782	0.844/0.853	0.792/0.826
	✓		0.877/0.878	0.778/0.812	0.861/0.855	0.825/0.844
✓	✓		0.885/0.886	0.792/0.826	0.880/0.880	0.829/0.849
✓	✓	✓	**0.888/0.889**	**0.795/0.830**	**0.884/0.883**	**0.832/0.855**

5.3.2 Results on the DIVIDE-3k

Training and Testing on DIVIDE-3k. We first benchmark recent state-of-the-arts by conducting training and testing in the DIVIDE-3k. As shown in Tab. 6, the two semantic-unaware classical methods [9, 10] are performing notably worse and DOVER again achieves state-of-the-art. It is also noteworthy that with aesthetic and technical scores as auxiliary labels, DOVER++ further improves the performance for overall quality prediction. This further suggests that better modeling of the two perspectives can finally benefit overall quality assessment in the UGC-VQA problem.

Zero-shot and Cross-dataset Evaluations. We also benchmark the opinion-unaware (*i.e.* zero-shot) VQA approaches on the **DIVIDE-3k**. Among them, the recent BVQI [80] reaches the best performance by considering both technical and semantic (aesthetic-related) criteria. Moreover, we benchmark the best approaches in Tab. 4 on the cross-dataset generalization from LSVQ to the DIVIDE-3k, where the proposed DOVER again outperforms other methods, suggesting the alignment between the proposed objective approach and subjective database.

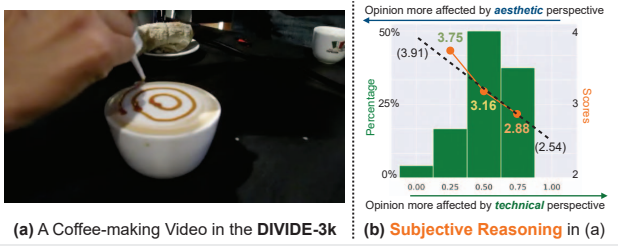
5.4. Ablation Studies

Effects of View Decomposition. In Tab. 8, we compare the proposed View Decomposition strategy with common strategies in UGC-VQA by keeping other parts the same. First of all, it is much better than the variant *w/o Decomposition* that directly takes the original videos as inputs of both branches, showing the effectiveness of decomposition. Moreover, with backbone and input kept the same, DOVER with separate supervisions is also notably better than *Feature Aggregation*, which first concatenates features from two branches together and then regress them to the quality scores, as applied by several existing approaches [1, 49, 58].

Effects of Subjectively-Inspired Fusion. We discuss the fusion strategy in Tab. 9. As shown in the table, only considering one branch will bring a notable performance decrease, and directly obtaining the fused quality as $Q_{\text{pred,A}}+Q_{\text{pred,T}}$ without weights is also less accurate than subjectively-inspired fusion. These results further validate the subjective observations found in the DIVIDE-3k.

Table 10. **Ablation Study of DOVER++**: Effects of different objectives.

Loss Objectives			**DIVIDE-3k** (3590)		
	$\mathcal{L}_{\text{LVBS}}$	$\mathcal{L}_{\text{DS}}$	SROCC↑	PLCC↑	KROCC↑
w/o MOS_{A}&MOS_{T}	✓		0.8331	0.8438	0.6480
w/ MOS_{A}&MOS_{T}		✓	**0.8357**	**0.8455**	**0.6521**
	✓	✓	**0.8442**	**0.8537**	**0.6603**



(a) A Coffee-making Video in the **DIVIDE-3k** **(b)** **Subjective Reasoning** in (a)

Legend ●—● : Mean Opinion Score in Different Groups - - - - - : DOVER++ Prediction *w.r.t.* weight of $Q_{\text{pred,T}}$

Figure 9. For the video **(a)**, the impact of aesthetic and technical perspectives on the final quality rating **(b) varies among individuals**. By adjusting fusion weights, DOVER++ can align with opinions from different groups.

Ablation Studies of DOVER++. In Tab. 10, we further discuss whether the extra objective ($\mathcal{L}_{\text{DS}}$) can improve accuracy of overall quality prediction. By combining $\mathcal{L}_{\text{DS}}$ with $\mathcal{L}_{\text{LVBS}}$, it contributes to around 1% performance gain. It is also noteworthy that even without direct MOS labels for supervision, the $\mathcal{L}_{\text{DS}}$ only can still outperform $\mathcal{L}_{\text{LVBS}}$. All these results suggest that explicitly considering "quality" in UGC-VQA into a sum of two perspectives is a good approximation to the human perceptual mechanism.

5.5. Outlook: Personalized Quality Evaluation

During the subjective reasoning study, we further find out that the effect of each perspective varies among different individuals. For instance, the video in Fig. 9(a) has better aesthetics and worse technical quality (*blurry, under-exposed*), and different individuals consider the technical impact differently while rating the overall opinion (Fig. 9(b)). Moreover, with more consideration of the technical perspective, subjects tend to rate lower scores on the video. With DOVER++, if we adaptively fuse between $Q_{\text{pred,A}}$ and $Q_{\text{pred,T}}$, we find that the differently-fused results can better predict the quality perception of individual subject groups, suggesting its primary capability to provide quality evaluation catering for personalized requirements.

6. Conclusion

In this paper, we present the DIVIDE-3k database and the first subjective study aimed at exploring the impact of aesthetic and technical perspectives on UGC-VQA, which reveals that both perspectives impact human quality opinions. In light of this observation, we propose the objective quality evaluators, DOVER and DOVER++, that achieve two objectives: **1)** significantly improving overall UGC-VQA performance; **2)** decoupling effects of two perspectives, so as to be applicable to specific real-world scenarios where pure technical or aesthetic quality metrics are needed.

7. Acknowledgement

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-vq: 'patching up' the video quality problem," in *CVPR*, June 2021, pp. 14 019–14 029. 1, 2, 3, 4, 6, 7, 8

[2] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, "Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild," in *IEEE Access 9*. IEEE, 2021, pp. 72 139–72 160. 1, 2

[3] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (konvid-1k)," in *QoMEX*, 2017, pp. 1–6. 1, 2, 3, 6

[4] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612–627, 2019. 1, 2, 6

[5] Y. Wang, S. Inguva, and B. Adsumilli, "Youtube ugc dataset for video compression research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 2019, pp. 1–5. 1, 2, 6

[6] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *ACM MM*, 2019, p. 2351–2359. 1, 2, 3, 7

[7] Y. Wang, J. Ke, H. Talebi, J. G. Yim, N. Birkbeck, B. Adsumilli, P. Milanfar, and F. Yang, "Rich features for perceptual quality assessment of ugc videos," in *CVPR*, June 2021, pp. 13 435–13 444. 1, 2, 7

[8] J. Xu, J. Li, X. Zhou, W. Zhou, B. Wang, and Z. Chen, "Perceptual quality assessment of internet videos," in *ACM MM*, 2021. 1, 2

[9] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Ugc-vqa: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021. 1, 2, 4, 7, 8

[10] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019. 1, 2, 4, 7, 8

[11] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013. 1, 2, 3, 4, 7

[12] L. Liao, K. Xu, H. Wu, C. Chen, W. Sun, Q. Yan, and W. Lin, "Exploring the effectiveness of video perceptual representation in blind video quality assessment," in *ACM MM*, 2022. 1, 2, 3, 4, 7

[13] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling," in *ECCV*, 2022. 1, 2, 4, 5, 6, 7

[14] DxOMark, "Dxomark photography benchmark." [Online]. Available: http://dxomark.com/ 1

[15] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *CVPR*, 2020, pp. 3677–3686. 1, 7

[16] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Basicvsr: The search for essential components in video super-resolution and beyond," in *CVPR*, 2021. 1

[17] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *ICCV Workshops*, 2021. 1

[18] Y. Wang, Y. Lu, Y. Gao, L. Wang, Z. Zhong, Y. Zheng, and A. Yamashita, "Efficient video deblurring guided by motion magnitude," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1

[19] T. Wiegand, "Draft itu-t recommendation and final draft international standard of joint video specification," 2003. 1

[20] D. Li, T. Jiang, W. Lin, and M. Jiang, "Which has better visual quality: The clear blue sky or a blurry animal?" *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1221–1234, 2019. 1

[21] H. Wu, C. Chen, L. Liao, J. Hou, W. Sun, Q. Yan, and W. Lin, "Discovqa: Temporal distortion-content transformers for video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 2, 4

[22] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *CVPR*, 2012, pp. 2408–2415. 1, 3, 4, 6

[23] V. Hosu, B. Goldlücke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *CVPR*, 2019, pp. 9367–9375. 1

[24] J. Hou, H. Ding, W. Lin, W. Liu, and Y. Fang, "Distilling knowledge from object classification to aesthetics assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1, 4

[25] X. Zhang, X. Gao, W. Lu, L. He, and J. Li, "Beyond vision: A multimodal recurrent attention convolutional neural network for unified image aesthetic prediction tasks," *IEEE Transactions on Multimedia*, vol. 23, pp. 611–623, 2021. 1, 4

[26] Y. Yang, L. Xu, L. Li, N. Qie, Y. Li, P. Zhang, and Y. Guo, "Personalized image aesthetics assessment with rich attributes," in *CVPR*, 2022, pp. 19 861–19 869. 1, 4

[27] J. Hou, S. Yang, and W. Lin, "Object-level attention for aesthetic rating distribution prediction," in *ACM MM*, 2020, p. 816–824. 1, 4

[28] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "Cvd2014—a database for evaluating no-reference video quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3073–3086, 2016. 2

[29] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2061–2077, 2018. 2, 4

[30] J. G. Yim, Y. Wang, N. Birkbeck, and B. Adsumilli, "Subjective quality assessment for youtube ugc dataset," in *ICIP*, 2020, pp. 131–135. 2

[31] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010. 2

[32] P. V. Vu and D. M. Chandler, "Vis3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *Journal of Electronic Imaging*, vol. 23, 2014. 2

[33] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, p. 64–73, 2016. 2

[34] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016. 2

[35] ——, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, 2017. 2

[36] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, pp. 684–694, 2013. 2

[37] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, pp. 3350–3364, 2011. 2

[38] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction," *IEEE Trans. Image Process.*, 2021. 2

[39] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, no. 2, 2016. 2

[40] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012. 2

[41] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2016. 2

[42] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012. 2

[43] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Blind video quality assessment with weakly supervised learning and resampling strategy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 2244–2255, 2019. 2

[44] J. You and J. Korhonen, "Deep neural networks for no-reference video quality assessment," in *ICIP*, 2019, pp. 2349–2353. 2

[45] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *ECCV*, 2018. 2

[46] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, "Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 2

[47] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi, "Rirnet: Recurrent-in-recurrent network for video quality assessment," *ACM MM*, 2020. 2

[48] Y. Liu, X. Zhou, H. Yin, H. Wang, and C. C. Yan, "Efficient video quality assessment with deeper spatiotemporal feature extraction and integration," *Journal of Electronic Imaging*, vol. 30, pp. 063 034 – 063 034, 2021. 2

[49] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for ugc videos," *arXiv preprint arXiv:2204.14047*, 2022. 2, 8

[50] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-wild videos with mixed datasets training," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1238–1257, 2021. 2, 7

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. 2

[52] J. You, "Long short-term convolutional transformer for no-reference video quality assessment," in *ACM MM*, 2021, p. 2112–2120. 2

[53] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *ACM MM*, 2020, p. 3311–3319. 2, 7

[54] H. Wu, L. Liao, A. Wang, C. Chen, J. H. Hou, E. Zhang, W. S. Sun, Q. Yan, and W. Lin, "Towards robust text-prompted semantic criterion for in-the-wild video quality assessment," 2023. 2

[55] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, A. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *ArXiv*, vol. abs/1705.06950, 2017. 2, 5

[56] J. Wang, K. C. K. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," 2022. 3, 7

[57] V. Vonikakis, R. Subramanian, J. Arnfred, and S. Winkler, "A probabilistic approach to people-centric photo selection and sequencing," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2609–2624, 2017. 3

[58] B. Li, W. Zhang, M. Tian, G. Zhai, and X. 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*, 2022. 3, 7, 8

[59] B. Zhang, L. Niu, and L. Zhang, "Image composition assessment with saliency-augmented multi-pattern pooling," *arXiv preprint arXiv:2104.03133*, 2021. 4

[60] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020. 4, 5

[61] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *CVPR*, 2020. 4, 7

[62] *ATQAM/MAST'20: Joint Workshop on Aesthetic and Technical Quality Assessment of Multimedia and Media Analytics for Societal Trends*. New York, NY, USA: Association for Computing Machinery, 2020. 4

[63] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *ECCV*, 2016. 4

[64] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981. 4

[65] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2740–2755, 2019. 4

[66] J. Hou, W. Lin, Y. Fang, H. Wu, C. Chen, L. Liao, and W. Liu, "Towards transparent deep image aesthetics assessment with tag-based content descriptors," *IEEE Transactions on Image Processing*, 2023. 4

[67] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018. 4

[68] X. Zhang, X. Gao, W. Lu, and L. He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Transactions on Multimedia*, vol. PP, pp. 1–1, 04 2019. 4

[69] V. Bhateja, A. Kalsi, and A. Srivastava, "Reduced reference iqa based on structural dissimilarity," in *2014 International Conference on Signal Processing and Integrated Networks (SPIN)*, 2014, pp. 63–68. 4

[70] H. Wu, C. Chen, L. Liao, J. Hou, W. Sun, Q. Yan, J. Gu, and W. Lin, "Neighbourhood representative sampling for efficient end-to-end video quality assessment," *arXiv preprint arXiv:2210.05357*, 2022. 5

[71] D. Li, T. Jiang, and M. Jiang, "Norm-in-norm loss with faster convergence and better performance for image quality assessment," in *ACM MM*, 2020, p. 789–797. 5

[72] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *CVPR*, 2022, pp. 11 976–11 986. 6

[73] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *CVPR*, 2022. 6

[74] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Rapique: Rapid and accurate video quality prediction of user generated content," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 425–440, 2021. 7

[75] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020. 7

[76] A. Ciancio, A. L. N. T. Targino da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 64–75, 2011. 7

[77] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2015. 7

[78] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, Mar. 2021. 7

[79] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. 7

[80] H. Wu, L. Liao, J. Hou, C. Chen, E. Zhang, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring opinion-unaware video quality assessment with semantic affinity criterion," *arXiv preprint arXiv:2302.13269*, 2023. 7, 8