Table 1. Comparing the **DIVIDE-3k** with recent non-reference VQA databases.

Database	UGC?	Subjects	#Content	#Video	Perspective	Labels
CVD2014 (2014) [1]	✗	**In-lab**	6	234	In-capture Distortions	MOS + σ
LIVE-Qualcomm (2017) [2]	✗	**In-lab**	42	208	In-capture Distortions	MOS
KoNViD-1k (2018) [3]	✓	Crowdsource	1200	1200	Subjective Quality	MOS + σ
LIVE-VQC (2019) [4]	✓	Crowdsource	585	585	Subjective Quality	MOS
Youtube-UGC (2020) [5]	✓	Crowdsource	1380	1380	Subjective Quality	MOS
LSVQ (2020) [6]	✓	Crowdsource	39,076	39,076	Subjective Quality	MOS
MSU-VQB (2022) [7]	✗	Crowdsource	36	2,486	Compression Distortions	MOS
DIVIDE-3k (Ours)	✓	**In-lab**	3,590	3,590	**Technical Distortions + Aesthetic Preference + Subjective Quality**	MOS + σ

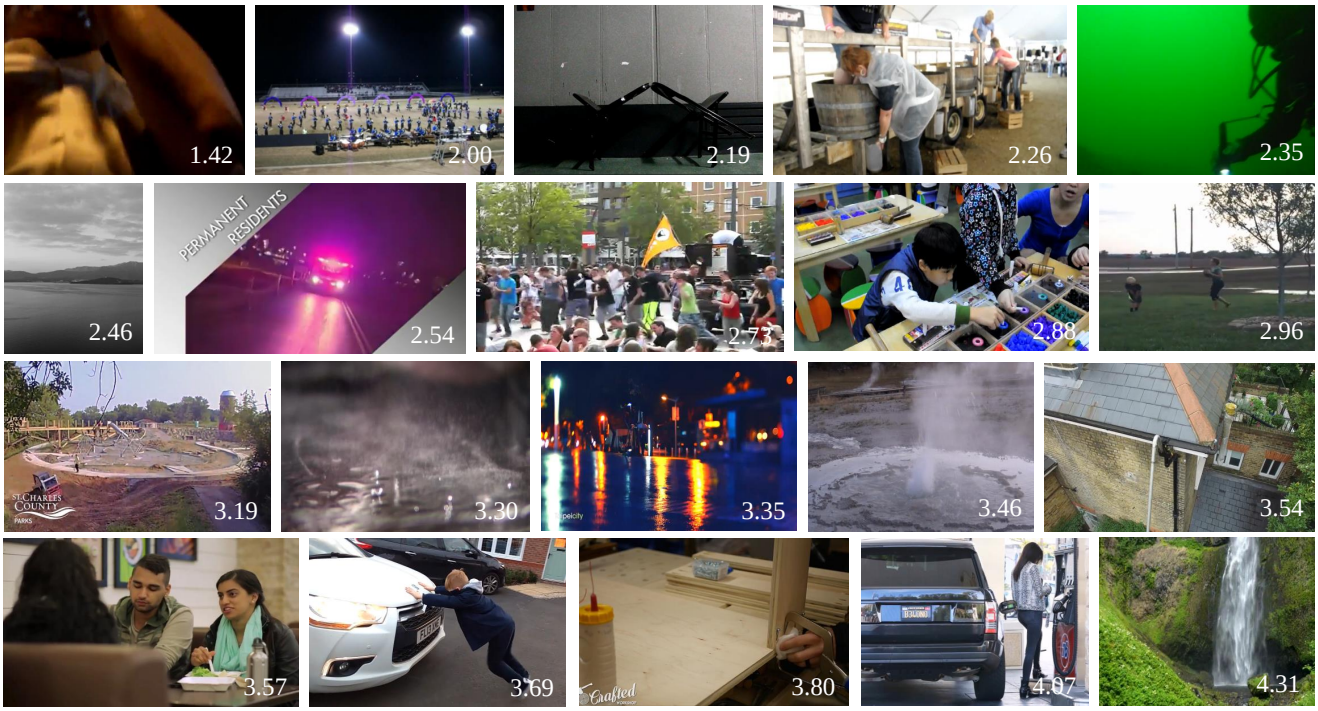

Figure 1. 20 examples of videos in the **DIVIDE-3k**, together with their overall quality scores. All videos are reshaped to fit in the figure and sorted with ascending overall quality from *upper-left* to *lower-right*. Please zoom in to better view technical details.

A. Extended Details on the DIVIDE-3k

A.1. Overview

Compared with existing UGC-VQA databases, the proposed DIVIDE-3k database has several specific features:

1. It contains diverse in-the-wild contents, with each video corresponding to its unique content and directly sampled from real-world videos. The videos are also sampled to match the quality distribution of large multimedia databases [8], so as to better represent human quality opinions on videos in the real world.

2. The subjective study is conducted with a well-controlled **in-lab** protocol. All subjects complete the subjective study on their personal laptops with the pre-downloaded annotation package. Moreover, all subjects have participated in Training before starting the annotation. The in-lab study process could better reduce the ambiguity of the quality scores.

3. Besides analysis on correlation between multiple perspective, the subjective study has the first-of-its-kind explicit **Subjective Reasoning** study to rate the impact of two perspectives (***technical***, ***aesthetic***) on the final subjective quality scores. It further explicitly supports the observations that the subjective quality perception on UGC videos is based on both perspectives.

Examples with their final mean opinion scores in the DIVIDE-3k are shown in Fig. 1. Among these videos, it could be noticed that aesthetic and technical perceptions are both related to the quality scores. Details as follows.

A.2. Statistics of Videos

In the DIVIDE-3k database, the videos are with multiple resolutions ranging from 240P to 1080P. The video duration

in the database ranges from 2s to 13s, with an average of around 10s (**more than 97% of videos are longer than 5s**). Videos with resolution more than 1080P are reshaped to 1080P; similarly, videos with duration longer than 12s are trimmed to 12s via *ffmpeg*. All the videos are in *.mp4* format, with frame rate ranging from 24 to 30.

A.3. Eligibility of Subjects

According to the recommendation of ITU-R BT.500 [9], we collect opinions of 35 subjects, including 19 males and 16 females to complete the whole study. The age of the subjects ranges from 20 to 26. During the study, all subjects have participated the **Golden Test** in each stage, that they need to correctly label (*pre-set label* ± 1) more than 7 out of the 10 golden videos (spot check out of 360 videos in the stage) to pass the test. Otherwise, the subject will be rejected for the next stage, and a complimentary subject will be trained to replace the subject in the next stage.

A.4. Training Materials for the Subjective Study

Before annotation, all subjects have passed through training on all three perspectives (for aesthetic, technical, and overall UGC-VQA quality perception). The details for the training process are introduced as follows.

Training for the Aesthetic Perspective. Following the ITU recommendations [9] and existing studies [10–12], during the aesthetic labelling, the subjects are generally asked to score the aesthetics on the video: **good** refers to '*absolutely preferred*' (with score 5), **fair** refers to '*neither preferred not disliked*' (with score 3), **bad** refers to '*absolutely disliked*' (with score 1). Considering the subjectiveness of aesthetic assessment [10, 13], we do not provide a strict criteria for the aesthetic perspective. Instead, we provide three hint-level criterions for aesthetic study:

1. Do the video contents have clear, appealing or meaningful semantics?
2. Do the video has good composition, i.e. do the target objects occur in good positions of the video?
3. Do you have positive feeling about the content of the video?

More importantly, during the training for the aesthetic perspective, all subjects are shown with 60 images, all from the Image Aesthetic Assessment dataset AVA [10]: 20 with **good aesthetics** (score: 8-10), 20 with **fair aesthetics** (score: 4-6), 20 with **bad aesthetics** (score: 1-3). The training images will be published with the dataset.

Training for the Technical Perspective. Unlike the aesthetic perspective, the definitions of technical distortions are more concrete and clearly-defined in many prior arts [1, 2, 4, 14]. After going through these arts, we ask subjects to **mainly** focus on the eight common distortions:

1. Noises. Usually resulted by **Video Capturing**, noises are grain-like fake textures that does not related to original objects in images or videos.
2. Artifacts. Related to **Video Compression**, they look like mini-blocks with rectangular edges.
3. Low Sharpness. Also known as general blurs (not caused by focus or motion), low sharpness can be caused by many reaaons.
4. Out-of-focus Blur. Resulted by **Video Capturing**. It means the **main object** is not in-focus.
5. Motion Blur. Resulted by **Video Capturing**, motion blur means apparent streaking of moving objects.
6. Stall. Usually resulted by **Video Transmission**, stall happens when some frames are lost, and one frame directly jump to a non-adjacent frame.
7. Jitter. Resulted by **Video Capturing**, jitter denotes the video *shaking* between adjacent frames.
8. Over/Under-exposure. It happens when the object is too dark (or too bright) to be recognizable.

For each of the eight types of distortions, we provide five examples (in total 40 examples for these distortions) to assist the subjects to understand the cases that are related to these distortions. Moreover, besides these common distortions, all other distortions related to video capturing, compression and restoration **are also instructed to be taken into consideration** during the technical perspective rating.

Training for the Overall Quality. During the training for overall quality assessment, we would like the subjects to better retrieve the original UGC-VQA problem [15], or, express their original opinions on their quality of experience (QoE) for the videos. Therefore, we remind the subjects they may consider the abovementioned factors or perspectives, yet their final judgements should be based on their overall experience on ***the quality of the videos***. Specifically, to better align their opinions with existing UGC-VQA studies, we choose 60 videos from the largest-ever UGC-VQA dataset LSVQ, including 20 with **good quality** (original score ≥ 70), 20 with **fair quality** (original score among $[40, 70)$), 20 with **bad quality** (original score < 40). The training videos will be published with the dataset.

A.5. Interface of Subjective Study

The interface of the subjective study in the DIVIDE-3k is shown in Fig. 3. For each video, subjects need to rate the ***Aesthetic Score***, ***Technical Score***, ***Overall Score*** of the video. Moreover, in the **Subjective Reasoning**, the subjects are instructed to rank the impact of the aesthetic and technical perspectives. Subjects can proceed to the next video only when all the four dimensions are scored. Returning to previous videos to modify labels is allowed.

A.6. Post-processing of Opinions

Opinion Cleaning. According to the recommendations of ITU-R BT.500 [9], we remove the unreliable opinions by

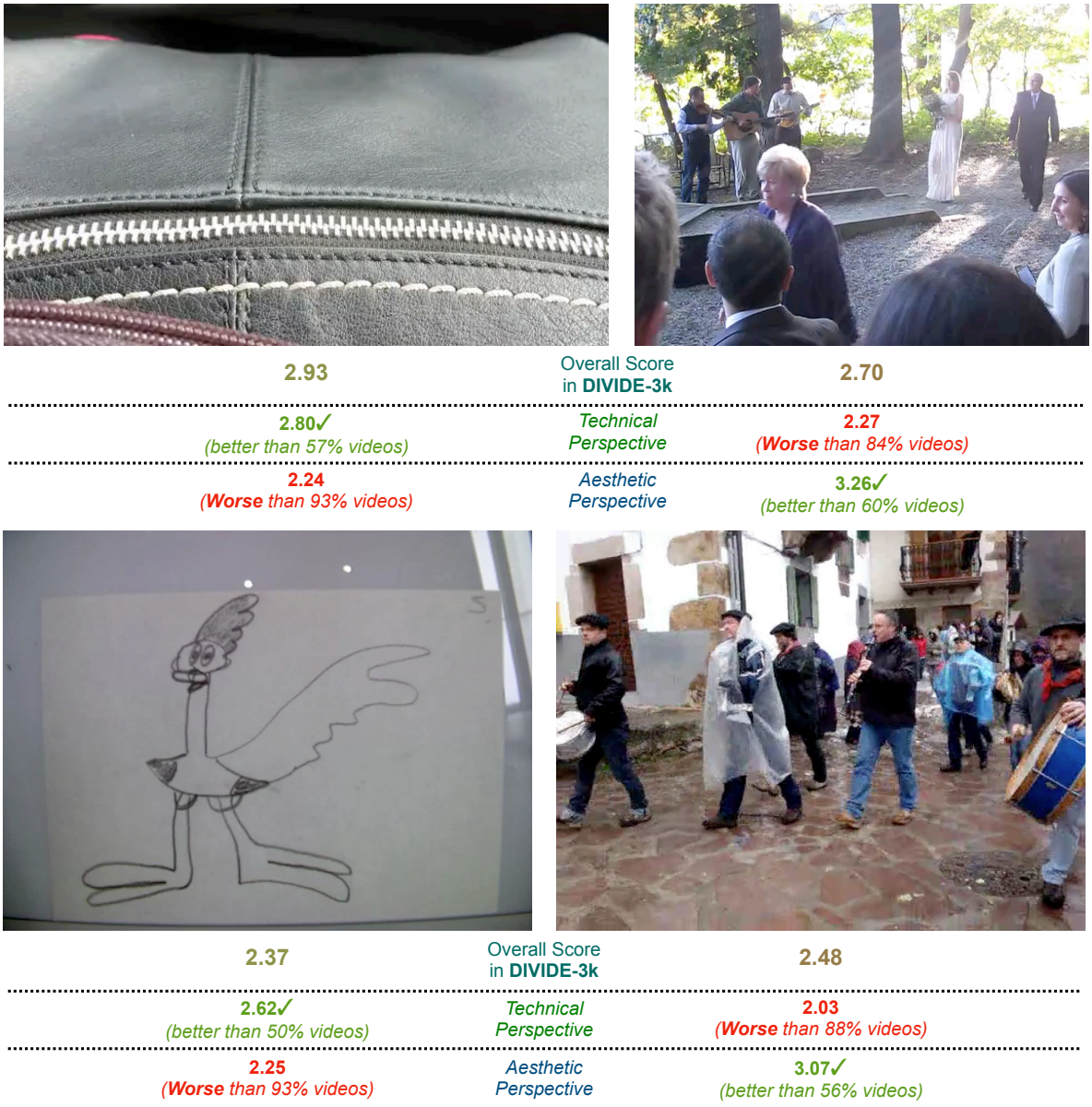

Figure 2. Two extended pairs of examples in the **DIVIDE-3k** that ***aesthetic*** and ***technical*** perspectives **view "quality" differently**. The left video in each pair has better technical quality than the right one, yet worse aesthetics due to relative meaningless or unappealing contents.

leaving out the opinions provided by the most deviated subjects in each stage (whose opinions show least correlation to the primary mean opinion in the batch). After the opinion cleaning, each video has an average of 31 annotations and at least 28 annotations.

Mean Opinion Scores. The mean opinion scores for each branch is directly obtained from an average of cleaned opinions in the respective perspectives. Therefore, the final score range is among $[1, 5]$. Specifically, for the aesthetic perspective, the $\min$ and $\max$ values of the $\mathrm{MOS_A}$ are 1.06 and 4.66, respectively. The $\min$ and $\max$ for the $\mathrm{MOS_T}$ (technical perspective) are 1.06 and 4.57, and 1.06 and 4.74 for the overall MOS.

A.7. Subjective Divergence between Perspectives

In this part, we discuss more examples than the **Fig. 1** in the main paper, where the subjects have apparently different opinions from different two perspectives. In Fig. 2, we further show two additional pairs of videos in the DIVIDE-3k where the left video has worse ***aesthetic quality*** but the right video has notably worse ***technical quality***. It is also easy to notice that the final quality scores are not only decided by one perspective in the two cases, while such rivalry (one has worse contents, one has stronger distortions) can

Figure 3. **The Interface for the Subjective Study** in the **DIVIDE-3k** database. Subjects are required to rate four scores: the ***Aesthetic*** score, the ***Technical*** score, the overall quality score, and the **Subjective Reasoning Study** on impact of perspectives during scoring.

result in either the left one or the right one to have better subjective quality score.

B. *User Studies* on Existing UGC-VQA Dataset

B.1. Before Experiments: Data Preparation

Data Source. The videos used for the *User Studies* are from the two test sets of LSVQ [6] dataset: $\text{LSVQ}_{\text{test}}$ and $\text{LSVQ}_{1080\text{p}}$ (the largest test sets), and the proposed DOVER is trained on the corresponding training set of LSVQ. To make sure that the resolution of videos does not affect subjective ratings, we only select pairs where both videos are in $\text{LSVQ}_{\text{test}}$ or $\text{LSVQ}_{1080\text{p}}$ (100 pairs for each case). We also remove the `ia_batch*` subsets in LSVQ as some videos in these subsets contain very severe decoding errors.

Why Evaluate on Diverged Pairs? In a proportion of videos, the professionalism of photographers is associated with what technical equipment they would use to record or generate a video. The existence of these videos could lead to the relatively good overall performance on the biased evaluators (aesthetic branch and technical branch): they are focusing on different aspects of the video quality but they happen to be similar. These videos cannot reflect the ability of the aesthetic branch and the technical branch on separation the perceptions between the two issues. The videos with diverged aesthetic branch and technical branch predictions, though, could be potentially the other proportion of videos with different aesthetic and technical quality, and henceforth selected for subjective studies to evaluate the disentanglement ability of DOVER. Following conclusions of several recent studies [16–18], we evaluate on the results of pairwise rank comparisons instead of direct quality score. The selection process for pairs is discussed as follows.

Selection for Diverged Pairs. To select the pairs where the two evaluators (the aesthetic branch and the technical branch) predict differently with high confidences, we first normalize the predictions of each evaluator, so that the scores of both evaluators are rescaled to $[1, 5]$ (scale of the DIVIDE-3k) as follows:

$$\hat{\text{Q}}_{\text{pred,A}} = \frac{1}{1+e^{-\frac{\text{Q}_{\text{pred,A}}-\overline{\text{Q}_{\text{pred,A}}}}{\sigma(\text{Q}_{\text{pred,A}})}}} \times 4 + 1 \quad (1)$$

$$\hat{\text{Q}}_{\text{pred,T}} = \frac{1}{1+e^{-\frac{\text{Q}_{\text{pred,T}}-\overline{\text{Q}_{\text{pred,T}}}}{\sigma(\text{Q}_{\text{pred,T}})}}} \times 4 + 1 \quad (2)$$

After normalizing predicted scores, to select the diverged video pair $(\mathcal{V}_1, \mathcal{V}_2)$, we constrain that $\hat{\text{Q}}^{\mathcal{V}_1}_{\text{pred,T}} - \hat{\text{Q}}^{\mathcal{V}_2}_{\text{pred,T}} > 1$ and $\hat{\text{Q}}^{\mathcal{V}_1}_{\text{pred,A}} - \hat{\text{Q}}^{\mathcal{V}_2}_{\text{pred,A}} < -1$, or vice versa. These diverged pairs are the video pairs where the proposed DOVER recognizes that one video has *notably* better aesthetic quality, but another has *notably* better technical quality.

After pair selection, we get around 38,000 feasible video pairs following the rules above. Then, 200 random pairs (with seed 42) are sampled from all feasible pairs and used for subjective studies. Code for pair selection is appended.

Information on Subjects. Following recommendations from ITU [9], we select 15 subjects with age 19 to 25 in two different countries, where each subject is instructed to provide 400 binary opinions (200 for aesthetic, 200 for technical). Each subject is not allowed to view judges from other subjects to avoid influences from one another.

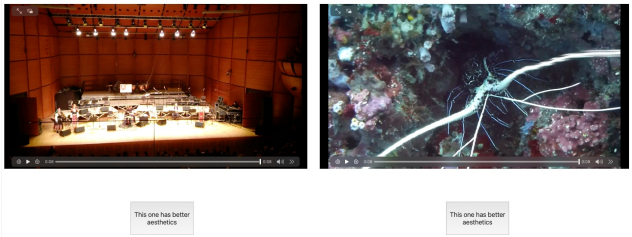

Figure 4. **User Study Interface** for aesthetic quality comparison. The subjective expert is instructed to select *which one has better aesthetics* between the two videos. Similar for technical quality comparison (the *aesthetics* in the button is changed to *technical quality* respectively).

B.2. During User Studies

Annotation Interface. The annotation interface is shown in Fig. 4. To avoid Internet-transmission-based stalls that may change the quality of original videos [1, 19], we require the annotators to download all videos into local directories first and annotate through a local browser.

Instructions for aesthetic quality comparison.

In this task, you are instructed to assess which one has better aesthetic quality between two videos, specifically based on the following aspects:

1. Do the video contents have clear, appealing or meaningful semantics?
2. Do the video has good composition, i.e. do the target objects occur in good positions of the video?
3. Do you have positive feeling about the content of the video?

Please be at most subjective on judging which video has overall better aesthetics by your preference, ***WITHOUT*** considering the following aspects:

1. The textures
2. The artifacts, and noises
3. The picture clearness (whether it is blurry or not)
4. Other technical-related issues

We advise you to view the videos without zooming and view only once to have a overall subjective judgement on the aesthetics of this video.

Instructions for technical quality comparison.

In this task, you are instructed to assess which one has better technical quality between two videos, only based on the following aspects:

1. The artifacts, and noises (stronger is worse)
2. The temporal quality: does the video have very strong flicker? (stronger is worse)
3. The picture clearness (whether it is blurry or not)
4. Other technical-related issues

In this part, we advise you to zoom each video into full screen; for better judgement, you may stop at middle of the video to see more clearly.

Be sure ***not to consider*** the contents / composition in the videos during the stage of technical evaluation.

Training videos. Besides randomly selecting 200 pairs of videos from all video pairs that follow our data preparation requirements for the blind annotations, we also select 10 pairs as ***training videos*** with gold subjective labels (*i.e.* certificated ground truth binary opinions on both aesthetic and technical quality comparison between the videos the pairs) from the research team to train the subjective evaluators prior to their subjective studies. The training video pairs are also randomly mixed in the video pairs that are needed to be annotated and the opinions of the annotator is valid only when he/she correctly labels more than 70% (7/10) of training video pairs. All 15 subjects have passed the validity test.

B.3. Results

Success Cases. We visualized several successful cases (*i.e.* when the subjective annotators agree with both aesthetic and technical comparison of DOVER between videos in the diverged pair) for the aesthetic-technical disentanglement via the proposed DOVER in Fig. 5, Fig. 6 and Fig. 7. Specifically, we can notice that the technical branch is very sensitive to textures especially clearness of videos, and is also sensitive to global technical quality factors such as under-exposure (Fig. 5, left) and over-exposure (Fig. 7, left). On the contrary, the aesthetic branch not only very sensitive on the chaotic compositions of the three examples in the right of the figures, but also able to recognize the very commonly-agreed good aesthetics for photography as in (Fig. 7, left). These cases further demonstrates the effectiveness of the DOVER on disentangling both effects in UGC-VQA. *We also show the video version of the demos (*`Fine-graind Subjective Studies Result Demo.mp4`*) in the supplementary package.*

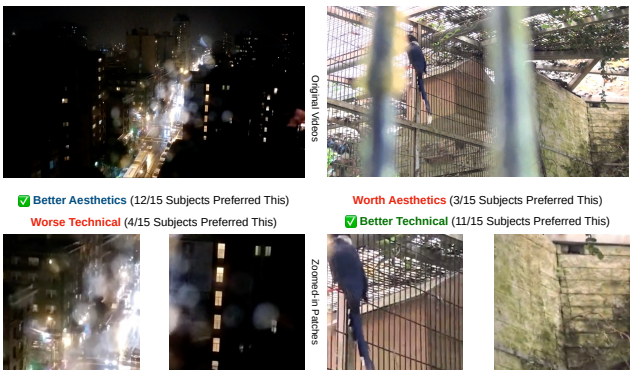

Figure 5. **Success case (I)** when the proposed DOVER can disentangle the aesthetic and technical quality of videos. The video in the left has apparently better aesthetics (good composition) but worse technical quality due to the distortions (blurs, noises, color errors, under-exposures).

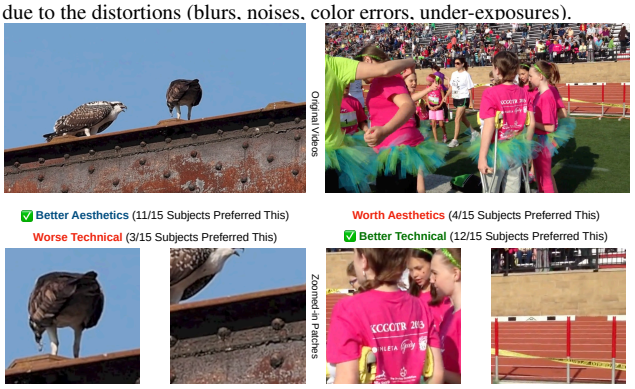

Figure 6. **Success case (II)** when the proposed DOVER can disentangle the aesthetic and technical quality of videos. The video in the left has apparently better aesthetics (symmetry composition, clear semantics) but relative worse technical quality due to the distortions (blurs).

Failure Cases. Without respective supervision, the proposed DOVER is not perfect on decoupling aesthetic and technical effects with still around 30% error predictions (26% for technical, 31% for aesthetics). When we look at the failure cases, we notice that most of them are pretty "finer-grained" cases where there are different aesthetic and technical concerns in the two videos of the pair, as illustrated in Fig. 8. For the video in the *left*, there are some typical light spots which are usually preferred in photography, yet the content is relatively meaningless compared with the *right* one. For technical considerations, the *left* one is much more blurry but the other is with unacceptable artifacts from compression. This cases suggest that though DOVER could be able to consider different issues for technical branch and aesthetic branch, it is still not so well in better reasoning their effects in aesthetic and technical perception of video quality, especially in the finer-grained cases. To achieve finer-grained predictions, the optimal way would be intro-

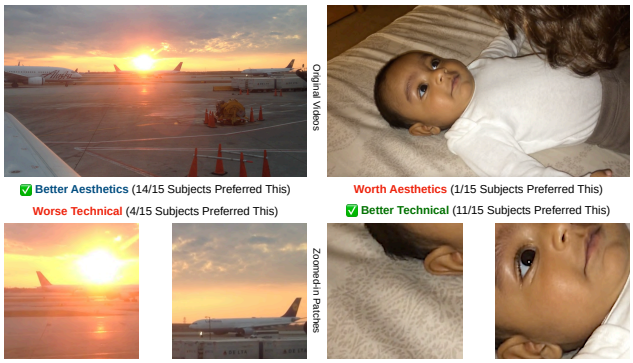

Figure 7. **Success case (III)** when the proposed DOVER can distinguish the aesthetic and technical quality of videos. The video in the left has apparently better aesthetics with almost every subjects agreed (with very good content for photography: sunset) but worse technical quality due to the distortions from the technical perspective (over-exposure, blurs).

Figure 8. A typical failure case of DOVER on comparing aesthetic and technical quality. The video in the *right* has very strong compression artifacts where the one in the *left* is very blurry: the technical branch prefers the *right* while subjective opinions slightly prefer the *left*.

ducing separate supervisions for both qualities, yet this is currently unavailable in existing UGC-VQA datasets.

C. Extended Qualitative Results

C.1. Feature Dissimilarity Curves

We visualize the Feature Dissimilarity Curves In Fig. 9 to further demonstrate the effectiveness of the Cross-scale Regularization. Without the regularization, the dissimilarity of features between different scales (S_{A} and $S_{\mathrm{A}}\downarrow$) can be reduced but not fully removed, which are closely related to the remaining perception on low-level technical issues. With the Cross-scale Restraint, the dissimilarity could be totally removed, so as to better help the aesthetic branch focus on non-technical issues.

C.2. More Divergence Maps

Fig. 10 and Fig. 11 show the divergence maps on $\mathrm{LSVQ}_{\mathrm{test}}$, LIVE-VQC, KoNViD-1k and YouTube-UGC respectively. The DOVER on all these datasets show divergent predictions on the two evaluators, especially the YouTube-UGC where the two evaluators only have 0.793 SRCC, 0.810 PLCC, 0.603 KRCC and 80.1% concordance. In LIVE-VQC, two evaluators are relatively similar (0.906 SRCC, 86.7% concordance), which might be due to the limited diversity of input contents (all videos are shot by smartphones on common events or objects, without post-production). We also notice that though in general the technical branch predictions have better correlation with MOS labels, in the edge cases the aesthetic branch is usually more accurate, which might be suggesting that aesthetic quality is usually not so much deviated in UGC videos but with bad aesthetics can significantly degrade the overall quality of videos. **This conclusion also aligns with the subjective observations as in main paper Sec. 3.3, that the aesthetic perspective has more impact during the extreme cases.**

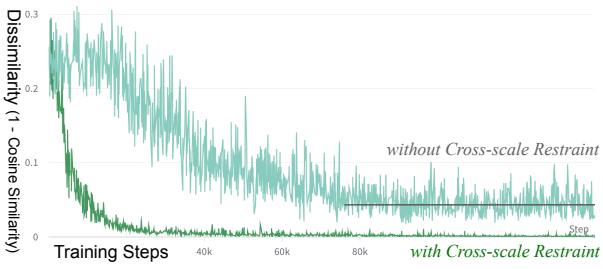

Figure 9. Dissimilarities between different S_{A} downsampled to 128×128 (denoted as $S_{\mathrm{A}\downarrow}$) and 224×224. With the Cross-scale Restraint, the aesthetic branch can extract consistent representations across scales.

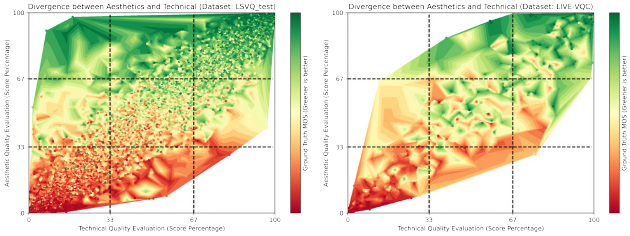

Figure 10. The divergence between aesthetic and technical evaluation by the proposed DOVER model in LIVE-VQC [4] and $\mathrm{LSVQ}_{\mathrm{test}}$ [6].

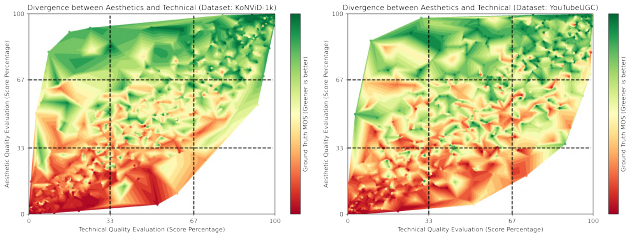

Figure 11. The divergence between aesthetic and technical evaluation by the proposed DOVER model in KoNViD-1k [3] and YouYube-UGC [5].

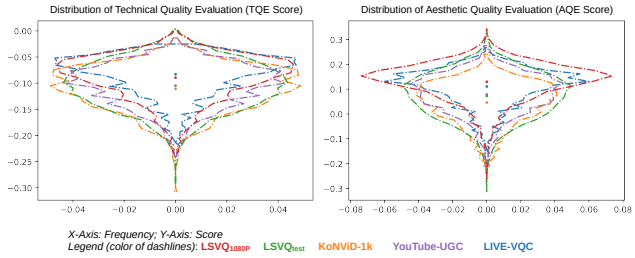

Figure 12. Distributions of aesthetic quality evaluation (aesthetic branch score) and technical quality evaluation (technical branch score) for different datasets.

C.3. Statistics on Different Datasets

We visualize the distributions of predicted aesthetic and technical quality of videos in different datasets by DOVER. As all the scores are predicted via the same model (and weights) and reach very good relative correlation with every single dataset, we can utilize the statistical information about the predicted scores as a reference for the quality distributions of different datasets. As illustrated in Fig. 12, LIVE-VQC has the best technical quality (*while compression processes are excluded in this dataset*) among all sets, yet $\text{LSVQ}_{1080\text{P}}$ has better overall aesthetic quality (*while a proportion of videos are shot by professional users*). $\text{LSVQ}_{\text{test}}$ contains more old videos (which were created decades ago) from the Internet Archive (IA) [21] (48%) than $\text{LSVQ}_{1080\text{P}}$ (23%), causing the difference between their aesthetic quality distributions ($\text{LSVQ}_{1080\text{P}}$ has sharper distributions). The KoNViD-1k is the worst for both quality evaluation among all datasets and is also the earliest one among them, suggesting the overall quality of UGC videos are notably improving during recent years.

D. Generalization on *Real-World* Videos

Quality assessment datasets might have different quality distribution from in-the-wild videos due to data pre-filtering processes [3, 4]. Thus, to examine the generalization ability of the proposed DOVER on its both aesthetic and technical evaluators, we direct test it on a randomly sampled 3000-video subset of Kinetics-400 [20], an action recognition dataset directly collected from UGC videos on YouTube platform. As illustrated in Fig. 13, the aesthetic branch and technical branch in the proposed DOVER can effectively identify aesthetic or technical quality on the random in-the-wild subset. The video with worse aesthetic quality has multiple negative aesthetic issues: *chaotic composition, upside down view, unappealing content*, where the one with best aesthetic quality has *shallow depth-of-field and rule-of-thirds composition*, proving very good aesthetic experience. The technically worse or best videos are also with *most unacceptable artifacts* or *very sharp and clean textures* respectively. The demo video for the in-the-wild comparison is appended as `In-the-wild Demos.mp4`.

E. Detailed Structures of the DOVER

E.1. the Aesthetic Branch

Equation for the Aesthetic View. Given a video $\mathcal{V} = \{V_i | i = 0, 1, \ldots, T-1\}$ with T total frames, the aesthetic view S_{A} with N sparse frames and spatial size s is formulated as:

$$\mathcal{F} = \{V_{\mathcal{U}(\frac{j\times N}{T}, \frac{(j+1)\times N}{T})} |_{j=0}^{T-1}\} = \{F_i|_{i=0}^{N-1}\} \quad (3)$$

$$S_{\text{A},i} = \text{downsample}(\mathbf{b} \otimes F_i, s) \quad (4)$$

where $\mathcal{F}$ is the remained frames after sparse sampling (F_i is the i-th frame in it), $\mathcal{U}(a, b)$ denotes uniform index sampling between (a, b), $\mathbf{b}$ is the blur kernel, $\otimes$ denotes element-wise multiplication, and $\text{downsample}(\cdot, s)$ denotes downsampling a frame into size $s \times s$. Moreover, the over-downsampled views ($S_{\text{A},\downarrow}$) is defined as follows:

$$S_{\text{A}\downarrow,i} = \text{downsample}(\text{downsample}(\mathbf{b} \otimes F_i, s), s^-) \quad (5)$$

where s^- is the size for the over-downsampled views.

Examples of the aesthetic view is illustrated in Fig. 14(a).

Cross-scale Friendly Feature Extractor. Following existing practices [22], we include the ImageNet-1k [23] pre-trained backbone as the content extractor for the aesthetic branch. We also choose a traditional convolution-based backbone ConvNeXt [24] to be friendly to multi-scale learning. As the temporal content relations are also noticed to be influential in the UGC-VQA problem [25, 26], we inflate the 2D ConvNext backbone with strategies as in [27] with the "1, 1, 3" inflation strategy to better consider both spatial and temporal aesthetic information in UGC videos.

E.2. the Technical Branch

Equation for the Technical View. The Technical View (S_{T}) [28] is formulated as:

$$S_{\text{T},i,[u\times S_f:(u+1)\times S_f, v\times S_f:(v+1)\times S_f]} \quad (6)$$

$$= P_{i,u,v} \quad (7)$$

$$= \text{RCrop}(\mathcal{V}_{i,[\frac{u\times H}{G_f}:\frac{(u+1)\times H}{G_f}, \frac{v\times W}{G_f}:\frac{(v+1)\times W}{G_f}]}, s_f) \quad (8)$$

where $P_{i,u,v}$ is the patch at the i-th frame, u-th horizontal grid, v-th vertical grid. $G_f \times G_f$ is the number of grids where patches are cropped, and $\text{RCrop}(\cdot, s_f)$ denotes randomly cropping a patch sized $S_f \times s_f$. Several examples for the technical view is shown in Fig. 14(b).

Patch-based Feature Extractor. To adapt to the characteristics of S_{T}, we choose the patch-based Swin-GRPB backbone as proposed in [28]. As discussed in [30], the structure

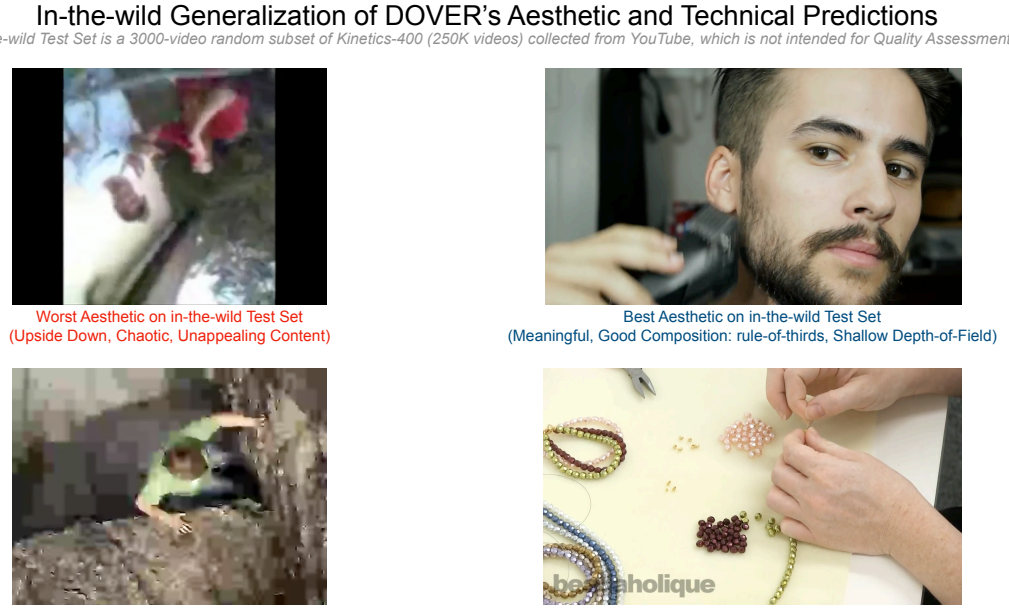

Figure 13. Videos with worst and best ***aesthetic*** and ***technical*** quality on the in-the-wild test set sub-sampled from Kinetics-400 [20], effectively reflecting human perception on aesthetic and technical quality of videos. The corresponding demo video is in `In-the-wild Demos.mp4`.

Table 2. Ablation study on the inductive biases in the Aesthetic Branch.

	Overall Accuracy of the DOVER				Accuracy of the Aesthetic Branch Only			
Testing Set/	**$\text{LSVQ}_{\text{test}}$**	**$\text{LSVQ}_{\text{1080p}}$**	**KoNViD-1k**	**LIVE-VQC**	**$\text{LSVQ}_{\text{test}}$**	**$\text{LSVQ}_{\text{1080p}}$**	**KoNViD-1k**	**LIVE-VQC**
Variants/Metric	SRCC/PLCC	SRCC/PLCC	SRCC/PLCC	SRCC/PLCC	SRCC/PLCC	SRCC/PLCC	SRCC/PLCC	SRCC/PLCC
Group 1: Variants for the Aesthetic View:								
cropping instead of downsampling	0.878/0.878	0.770/0.809	0.858/0.854	0.823/0.842	0.808/0.814	0.638/0.675	0.733/0.778	0.740/0.775
keeping spatial aspect ratio	0.887/0.887	0.793/0.828	0.883/0.883	0.831/0.854	0.857/0.858	0.740/0.786	0.846/0.855	0.792/0.825
temporal continuous frames	0.880/0.881	0.780/0.819	0.863/0.859	0.828/0.847	0.832/0.834	0.716/0.765	0.827/0.829	0.758/0.798
temporal global random frames	0.883/0.884	0.788/0.824	0.868/0.867	0.830/0.849	0.843/0.845	0.726/0.777	0.833/0.842	0.778/0.813
Group 2: Variants for Pre-training Settings:								
w/o AVA [29] pre-training	0.886/0.887	0.792/0.826	0.882/0.880	0.828/0.840	0.851/0.853	0.736/0.779	0.836/0.838	0.788/0.817
Group 3: Variants for Regularization Strategies:								
w/o Cross-scale Regularization	0.884/0.885	0.787/0.823	0.876/0.875	0.830/0.851	0.855/0.853	0.743/0.787	0.842/0.851	0.781/0.814
Accu. for technical branch only	0.877/0.878	0.778/0.812	0.861/0.855	0.825/0.844	NA	NA	NA	NA
DOVER (Ours)	**0.888/0.889**	**0.795/0.830**	**0.884/0.883**	**0.832/0.855**	0.855/0.856	0.738/0.782	0.843/0.852	0.792/0.826

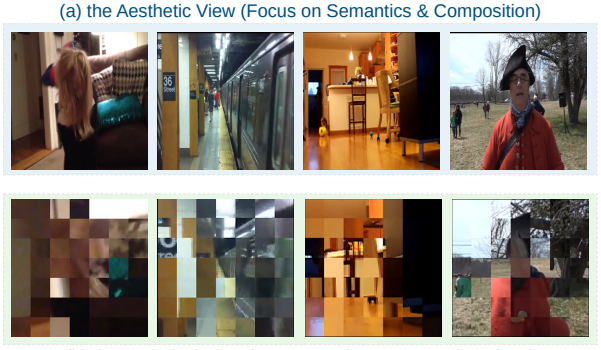

Figure 14. Examples for (a) the Aesthetic View (S_{A}) and (b) the Technical View (S_{T}) under *View Decomposition*, focusing on different perspectives.

could sufficiently avoid misunderstanding between ***banding artifacts*** and ***edges among patches*** (for instance, in the *lower-left* of the Fig. 13, the video with strong banding artifacts can be recognized as worst technical quality). The feature extractor is pre-trained with Kinetics-400 [27], so as to also be able to understand weak global semantics as background for distortion assessment.

F. More Implementation Details

F.1. Training Objective $\mathcal{L}_{\text{Sup}}$

In this section, we discuss the concrete design of the relative loss function $\mathcal{L}_{\text{Rel}}$ (the relative loss). Inspired by several studies [31–33], we restrain the monotonicity between predicted scores and MOS ($\mathcal{L}_{\text{mono}}$) and the linearity between them ($\mathcal{L}_{\text{lin}}$). The fusion loss described as follows is the same as existing state-of-the-arts [28, 34]:

$$\mathcal{L}_{\text{mono}} = \sum_{i,j} \max((Q^i_{\text{pred}} - Q^j_{\text{pred}}) \operatorname{sgn}(\text{MOS}^j - \text{MOS}^i), 0) \tag{9}$$

$$\mathcal{L}_{\text{lin}} = (1 - \frac{(Q_{\text{pred}} - \overline{Q_{\text{pred}}}) \cdot (\text{MOS} - \overline{\text{MOS}})}{\|Q_{\text{pred}} - \overline{Q_{\text{pred}}}\|_2 \|\text{MOS} - \overline{\text{MOS}}\|_2})/2 \quad (10)$$

$$\mathcal{L}_{\text{rel}} = \mathcal{L}_{\text{lin}} + 0.1\mathcal{L}_{\text{mono}} \quad (11)$$

where $\text{sgn}(\cdot)$ denotes the sign function, $a \cdot b$ denotes the inner product of a and b, and Q_{pred} and MOS are vectors that refer to predictions and ground truth labels in a batch.

F.2. Evaluation Metrics

We introduce the Pearson linear correlation coefficient (PLCC) and the Spearman's rank-order correlation coefficient (SRCC) as evaluation metrics. PLCC computes the linear correlation between a series of predicted scores Q_{pred} and ground truth scores MOS, while SRCC assesses the rank correlation. They are formulated as below:

$$\text{PLCC} = \frac{(Q_{\text{pred}} - \overline{Q_{\text{pred}}}) \cdot (\text{MOS} - \overline{\text{MOS}})}{\|Q_{\text{pred}} - \overline{Q_{\text{pred}}}\|_2 \|\text{MOS} - \overline{\text{MOS}}\|_2}) \quad (12)$$

$$\text{SRCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \quad (13)$$

where $\mu(\cdot)$ is the mean value, d_i is the distance of rank orders between predictions and ground truth of video i.

We also include the concordance C (a metric for agreement between two binary rank evaluators, higher is better) as the evaluation metric for the results of ***User Studies***, which is calculated as follows:

$$\text{C} = \frac{\text{concordant pairs}}{\text{concordant pairs} + \text{discordant pairs}} \quad (14)$$

Specifically, for our subjective studies, a concordance pair means that at least 8 (among 15) subjective annotators agree with the corresponding objective prediction by DOVER, which others are considered as discordant pairs.

G. Extended Quantitative Results

G.1. Ablation Studies on the Inductive Biases

We discuss the design of the inductive biases in the aesthetic branch in Tab. 2. All three types of inductive biases, including *inputs, pre-training and regularization strategies* have contributed to more accurate final quality prediction in the DOVER. The design of the aesthetic view is specifically dicussed, where the proposed way has outperformed several variants and proved most contribution to the overall accuracy of the DOVER.

References

[1] 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. 1, 2, 4

[2] 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. 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, 6, 7

[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, 7

[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, 6

[6] 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, 4, 6

[7] A. Antsiferova, S. Lavrushkin, M. Smirnov, A. Gushchin, D. S. Vatolin, and D. Kulikov, "Video compression dataset and benchmark of learning-based video-quality metrics," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [Online]. Available: https://openreview.net/forum?id=My5AI9aM49R 1

[8] 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. 1

[9] "Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures," ITU-R Rec. BT.500, 2000. 2, 4

[10] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *CVPR*, June 2018. 2

[11] B. Zhang, L. Niu, and L. Zhang, "Image composition assessment with saliency-augmented multi-pattern pooling," *arXiv preprint arXiv:2104.03133*, 2021. 2

[12] 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. 2

[13] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *ECCV*, 2016. 2

[14] 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

[15] 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. 2

[16] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "Pieapp: Perceptual image-error assessment through pairwise preference," in *CVPR*, June 2018. 4

[17] J. Gu, H. Cai, H. C. Chen, X. Ye, J. S. Ren, and C. Dong, "Pipal: a large-scale image quality assessment dataset for perceptual image restoration," in *ECCV*, 2020, pp. 633–651. 4

[18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595. 4

[19] 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. 4

[20] 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. 7, 8

[21] I. Archive, "Moving image archive." [Online]. Available: https://archive.org/details/movies 7

[22] 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

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255. 7

[24] 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. 7

[25] J. You, "Long short-term convolutional transformer for no-reference video quality assessment," in *ACM MM*, 2021, p. 2112–2120. 7

[26] 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. 7

[27] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, July 2017. 7, 8

[28] 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. 7, 8

[29] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *CVPR*, 2012, pp. 2408–2415. 8

[30] 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. 7

[31] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Exploiting unlabeled data in cnns by self-supervised learning to rank," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019. 8

[32] 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. 8

[33] ——, "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. 8

[34] 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. 8