Patch-VQ: 'Patching Up' the Video Quality Problem

Zhenqiang Ying[1*], Maniratnam Mandal[1*], Deepti Ghadiyaram[2†], Alan Bovik[1†]
[1]University of Texas at Austin, [2]Facebook AI
{zqying, mmandal}@utexas.edu, deeptigp@fb.com, bovik@ece.utexas.edu

Abstract

No-reference (NR) perceptual video quality assessment (VQA) is a complex, unsolved, and important problem for social and streaming media applications. Efficient and accurate video quality predictors are needed to monitor and guide the processing of billions of shared, often imperfect, user-generated content (UGC). Unfortunately, current NR models are limited in their prediction capabilities on real-world, "in-the-wild" UGC video data. To advance progress on this problem, we created the largest (by far) subjective video quality dataset, containing $38,811$ *real-world distorted videos and* $116,433$ *space-time localized video patches ('v-patches'), and* $5.5M$ *human perceptual quality annotations. Using this, we created two unique NR-VQA models: (a) a local-to-global region-based NR VQA architecture (called PVQ) that learns to predict global video quality and achieves state-of-the-art performance on 3 UGC datasets, and (b) a first-of-a-kind space-time video quality mapping engine (called PVQ Mapper) that helps localize and visualize perceptual distortions in space and time. The entire dataset and prediction models are freely available at* https://live.ece.utexas.edu/research.php.*.*

1. Introduction

User-generated content (UGC) and video streaming have exploded on social media platforms such as Facebook, Instagram, YouTube, and TikTok, each supporting millions and billions of users [63]. It has been estimated that each day, about 4 billion video views occur on Facebook [60] and 1 billion hours are viewed on YouTube [62]. Given the tremendous prevalence of Internet videos, it would be of great value to measure and control the quality of UGC videos, both on capture devices and at social media sites where they are uploaded, encoded, processed, and analyzed.

Full-reference (FR) video quality assessment (VQA) models perceptually compare quality against pristine videos, while no-reference (NR) models involve no such comparison. Thus, NR video quality monitoring could transform the processing and interpretation of videos on smartphones, social media, telemedicine, surveillance, and vision-guided robotics, in ways that FR models are unable to. Unfortunately, measuring video quality without a pristine reference is very hard. Hence, though FR models are successfully deployed at the largest scales [78], NR video quality prediction on UGC content remains largely unsolved, for several reasons.

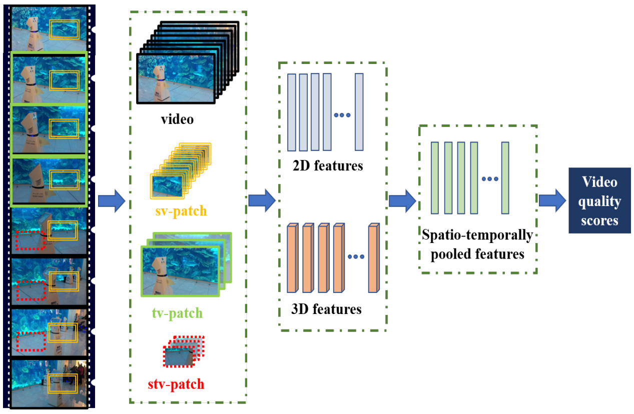


Fig. 1: **Modeling local to global perceptual quality:** From each video, we extract three spatio-temporal video patches (Sec. 3.1), which along with their subjective scores, are fed to the proposed video quality model. By integrating spatial (2D) and spatio-temporal (3D) quality-sensitive features, our model learns spatial and temporal distortions, and can robustly predict both global and local quality, a temporal quality series, as well as space-time quality maps (Sec. 5.2). Best viewed in color.

First, UGC video distortions arise from highly diverse capture conditions, unsteady hands of content creators, imperfect camera devices, processing and editing artifacts, frame rates, compression and transmission artifacts, and the way they are perceived by viewers. Inter-mixing of distortions is common, creating complex, composite distortions that are harder to model in videos. Moreover, it is well-known that the technical degree of distortion (e.g. amount of blur, blocking, or noise) does not correlate well with perceptual quality [75], because of neurophysiological processes that induce masking [47]. Indeed, equal amounts of distortions may very differently affect the quality of two different videos [52].

Second, most existing video quality resources are too

[*†] Equal contribution

small and unrepresentative of complex real-world distortions [12, 56, 30, 72, 73, 68, 69]. While three publicly available databases of authentically distorted UGC videos are available [23, 57, 74], they are far too small to train modern, data-hungry deep neural networks. What is needed are very large databases of videos corrupted by real-world distortions, subjectively rated by large numbers of human viewers. However, conducting large-scale psychometric studies is much harder and time-consuming (per video) than standard object/action classification tasks.

Finally, although a few NR algorithms achieve reasonable performance on small databases [42, 6, 28, 4, 35, 65, 37, 10], most of them fail to account for the complex space-time distortions common to UGC videos. UGC distortions are often transient (e.g., frame drops, focus changes, and transmission glitches) and yet may significantly impact the overall perceived quality of a video [55]. Most existing models are frame-based, or use sample frame differences, and cannot capture diverse temporal impairments.

We have made recent progress towards addressing these challenges, by learning to model the relationships that exist between local and global spatio-temporal distortions and perceptual quality. We built a large-scale public UGC video dataset of unprecedented size, comprising full videos and three kinds of spatio-temporal video patches (Fig. 1), and we conducted an online visual psychometric study to gather large numbers of human subjective quality scores on them. This unique data collection allowed us to successfully learn to exploit interactions between local and global video quality perception and to create algorithms that accurately predict video quality and space-time quality maps. We summarize our contributions below:

- **We built the largest video quality database in existence.** We sampled hundreds of thousands of open-source Internet UGC digital videos to match the feature distributions of social media UGC videos. Our final collection includes $38,811$ real-world videos of diverse sizes, contents, and distortions, 26 times larger than the most recent UGC dataset [74]. We also extracted three types of v-patches from each video, yielding $116,433$ space-time video patches ("v-patches") in total (Sec. 3.1).
- **We conducted the largest subjective video quality study to date**. Our final dataset consists of a total of 5.5M perceptual quality judgments on videos and v-patches from almost $6,300$ subjects, more than 9 times larger than any prior UGC video quality study (Sec. 3.2).
- **We created a state-of-the-art deep blind video quality predictor**, using a deep neural architecture that computes 2D video features using PaQ2PiQ [76], in parallel with 3D features using ResNet3D [20]. The 2D and 3D features feed a time series regressor [13] that learns to accurately predict both global video quality, as well as local space-time v-patch quality, by exploiting the relations between them. This new model, which we call Patch VQ (PVQ) achieves top performance on the new database as well as on smaller "in-the-wild" databases [57, 23], *without finetuning* (Secs. 4.1 and 5.3).
- **We also create another unique prediction model that predicts first-of-a-kind space-time maps** of video quality by learning global-to-local quality relationships. This second model, called the PVQ Mapper, helps localize, visualize, and act on video distortions (Sec. 5.2).

2. Related Work

Video Quality Datasets: Several public legacy video quality datasets [12, 56, 30, 72, 73, 68, 69] have been developed in the past decade. Each of these datasets comprises a small number of unique source videos (typically 10-15), which are manually distorted by one of a few synthetic impairments (e.g., Gaussian blur, compression, and transmission artifacts). Hence, these datasets are quite limited in terms of content diversity and distortion complexity, and do not capture the complex characteristics of UGC videos. Early "in-the-wild" datasets [46, 11] included fewer than 100 unique contents, while more recent ones such as KoNViD-1k [23], LIVE-VQC [57], and YouTube-UGC [74] contain relatively more videos (500-1500 per dataset), yet insufficient to train deep models. A more recent dataset, FlickrVid-150k [19] claims to contain a large number of videos, yet, has the following notable drawbacks: (a) Only 5 quality ratings were collected on each video which, given the complexity of the task, are insufficient to compute reliable ground truth quality scores (at least 15-18 is recommended [25]). (b) the database is not publicly available, hence limiting its use for any experiments or to validate its statistical integrity. (c) the videos are all drawn from Flickr, which is largely populated by professional and advanced amateur photographers, hence is not representative of social media UGC content.

Shallow NR VQA models: Early NR VQA models were distortion-specific [51, 14, 27, 26, 43, 8, 7] and focused mostly on transmission/compression-related artifacts. More recent and widely-used NR image quality prediction algorithms have been applied to frame difference statistics to create space-time video distortion models [42, 4, 38, 58, 49]. In all these models, handcrafted statistical features are used to train shallow regression models to predict perceptual video quality, achieving high performance on legacy datasets. Recently proposed models [35, 65, 49] use dozens or hundreds of such perceptually relevant features and achieve state-of-the-art performance on the leading UGC datasets, yet their predictive capability remains far below human performance.

Deep NR VQA models: There is more progress in the development of top-performing deep models for NR image quality prediction [15, 33, 5, 32, 59, 76, 41, 44, 10], but relatively fewer deep NR-VQA models exist. The au-

Table 1: **Summary of popular public-domain video quality datasets.** Legacy datasets contain singular synthetic distortions, whereas "in-the-wild" databases contain videos impaired by complex mixtures of diverse, real distortions.

Database	# Unique contents	# Video Duration (sec)	# Distortions	# Video contents	# V-Patch contents	Distortion type	Subjective study framework	# Annotators	# Annotations
MCL-JCV (2016) [72]	30	5	51	1,560	0	Compression	In-lab	150	78K
VideoSet (2017) [73]	220	5	51	45,760	0	Compression	In-lab	800	-
UGC-VIDEO (2019) [39]	50	> 10	10	550	0	Compression	In-lab	30	16.5K
CVD-2014 (2014) [46]	5	10-25	-	234	0	In-capture	In-lab	210	-
LIVE-Qualcomm (2016) [11]	54	15	6	208	0	In-capture	In-lab	39	8.1K
KoNViD-1k (2017) [23]	1,200	8	-	1,200	0	In-the-wild	Crowdsourced	642	$\approx$ 205K
LIVE-VQC (2018) [57]	585	10	-	585	0	In-the-wild	Crowdsourced	4,776	205K
YouTube-UGC (2019) [74]	1,500	20	-	1,500	4,500	In-the-wild	Crowdsourced	-	$\approx$ 600K
Proposed database (LSVQ)	$38,811$	5-12	-	$38,811$	$116,433$	In-the-wild	Crowdsourced	$6,284$	$5,545,594$

thors of [79] proposed a general-purpose NR VQA framework based on weakly supervised learning and a resampling strategy. The NR VSFA [37] model uses a CNN to extract frame-wise features followed by a gated recurrent unit to capture temporal features. These, and other attempts [70, 71, 79, 37] mostly perform well on legacy datasets [56, 72, 69] and struggle on in-the-wild UGC datasets [57, 74, 23]. MLSP-VQA [19] reports high performance on [23], but their code is not available, and we have been unable to reproduce their reported results.

3. Large-Scale Dataset and Human Study

Next, we present details of the newly constructed video quality dataset and the subjective quality study we conducted on it. The new database includes $38,811$ videos and $116,433$ "v-patches" extracted from them, on which we collected about 5.5M quality scores in total from around $6,300$ unique subjects. This new resource is significantly larger and more diverse than any legacy (synthetic distortion) databases [56, 12, 72, 73] or in-the-wild crowdsourced datasets [23, 57, 74] (26 times larger than [74]). We refer to the proposed dataset as the Large-Scale Social Video Quality (LSVQ) Database.

3.1. Building the Dataset

3.1.1 UGC-Like Data Collection and Sampling

We selected two large public UGC video repositories to source our data: the Internet Archive (IA) [61] and YFCC-100M [64], and collected a total of $400,000$ videos from them. Each video was randomly cropped to an average duration 7 seconds[1] using *ffmpeg* [2].

Sampling "UGC-like" videos: Our dataset distinguishes itself from other in-the-wild video datasets in several ways. First, unlike KoNViD-1k [23], we did not restrict the collected videos to have fixed resolutions or aspect ratios, making the proposed dataset much more representative of real-world content. Second, we did not apply scaling or further processing which could affect the quality of the content. Finally, to obtain "UGC-like" videos, we used a mixed integer programming method [67] to match a set of UGC feature histograms. Specifically, we computed the following 26 holistic spatial and temporal features on two video collections: (a) our aforementioned 400K video collection from IA and YFCC-100M and (b) 19K public, randomly selected videos from a social media website:

- *Absolute Luminance* $L = R + G + B$.
- *Colorfulness* using [21].
- *RMS Luminance Contrast* [48].
- Number of *detected faces* using [1].
- *Spatial Gaussian Derivative Filters* (3 scales, 2 orientations) from Leung-Malik filter bank [36].
- *Temporal Gaussian Derivatives* (3 scales) first averaged along temporal dimension, followed by computing the mean and standard deviation along the spatial dimension.

The first five (spatial) features were computed on each frame, then the means and standard deviations of these features across all frames were obtained as the final features.

As mentioned, we sampled and matched feature histograms and in the end, arrived at about 39,000 videos, with roughly equal amounts from IA and YFCC-100M. Fig. 2 shows 16 randomly selected video frames from LSVQ, while Fig. 3 plots the diverse sizes, aspect ratios and durations of the final set of videos. It is evident that we obtained a diverse UGC video dataset that is representative of content, resolution, aspect ratios, and distortions.

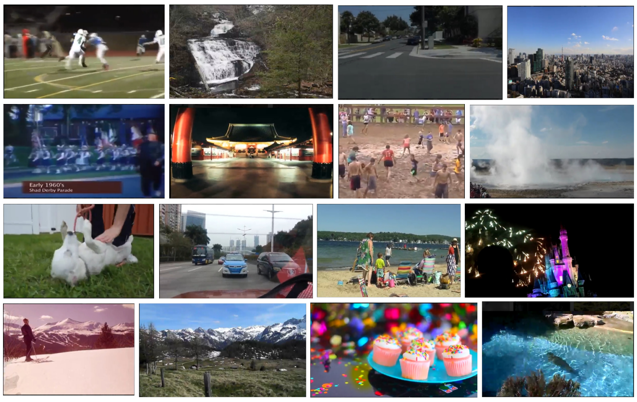


Fig. 2: **Sample video frames from the new database**, each resized to fit. The actual videos are of highly diverse sizes and resolutions.

[1] Cropping to a fixed duration was not possible, since a video must begin with a key frame to be decoded properly.

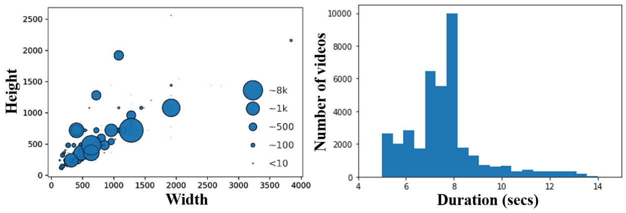


Fig. 3: **Left:** Scatter plot of video width versus video height with marker size indicating the number of videos having a given dimension in the new LSVQ database. **Right:** Histogram of the durations (in seconds) of the videos.

3.1.2 Cropping Video-Patches

To closely study and model the relationship between global and local spatio-temporal qualities, we randomly cropped three different kinds of video patches or "v-patches" from each video: a spatial v-patch (**sv-patch**), a temporal v-patch (**tv-patch**), and a spatio-temporal v-patch (**stv-patch**). All three patches are videos obtained by cropping an original video in space, time, or both space and time, respectively (Fig. 4). All v-patches have the same spatial aspect ratios as their source videos. Each sv-patch has the same temporal duration as their source videos, but cropped to 40% of spatial dimensions (16% of area). Each tv-patch has the same spatial size as its source, but clipped to 40% of temporal duration. Finally, each stv-patch was cropped to 40% along all three dimensions. Every v-patch is entirely contained within its source, but the volumetric overlap of each sv-patch and tv-patch with the same-source stv-patch did not exceed 25% (suppl. material).

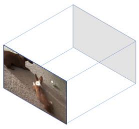
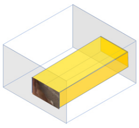
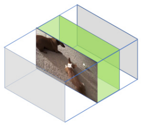
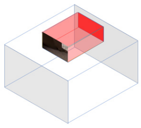
Full video	**Spatial Patch** sv-patch	**Temporal Patch** tv-patch	**Spatio-temporal Patch** stv-patch
			

Fig. 4: **Three kinds of video patches** (v-patches) cropped from random space-time volumes from each video in the dataset. All v-patches are videos.

3.2. Subjective Quality Study

Amazon Mechanical Turk (AMT) was used to collect human opinions on the videos and v-patches as in other studies [57, 74, 76, 16, 40]. We launched two separate AMT tasks - one for videos and the other for the three video patches. A total of $6,284$ subjects were allowed to participate on both tasks. On average, we collected 35 ratings on each video and v-patch. Subjects could participate in our study through desktops, laptops, or mobile devices.

3.2.1 AMT Study Design

The human intelligence task (HIT) pipeline is shown in Fig. 5. Each task began with general instructions, followed by a related quiz to check subjects' comprehension of the instructions, which they had to pass to proceed further. During training, each subject rated 5 videos to become familiar with the interface and the task. Then, they entered the testing phase, in which they rated 90 videos. Each video was played only once, following which the subject rated the video quality on a scale of 0-100 by sliding a cursor along the rating bar (suppl. material). Subjects could report a video as inappropriate (violent or pornographic), static or incorrectly oriented. We ensured that each video was downloaded before playback to avoid rebuffering and stalling. In the end, each subject answered several survey questions about the study conditions and their demographics.

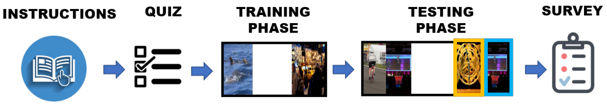


Fig. 5: **Study workflow** for both video and v-patch sessions.

3.2.2 Subject Rejection

Next, we summarize the several checks we employed at various stages of the AMT task to identify and eliminate unreliable subjects [57, 16] and participants with inadequate processing or network resources.

During Instructions: If a participant's browser window resolution, version, zoom, and the time taken to load videos did not meet our requirements (suppl. material), they were not allowed to proceed.

During Training: Although we ensured that each video was entirely downloaded prior to viewing, we also checked for any potential device-related video stalls. If the delay on any training video exceeded 2 seconds, or the total delay over the five training videos exceeded 5 seconds, the subject was not allowed to proceed (without prejudice). They were also stopped if a negative delay was detected (e.g., using plugins to speed up the video).

During Task: At the middle of each subject's task, we checked for instability of the internet connection, and if more than 50% of the videos viewed until then had suffered from hardware stalls, the subject was disqualified. We also checked whether the subject had been giving similar quality scores to all videos, or was nudging the slider only slightly, both indicative of insincere ratings.

Post task: In the test phase, of the 90 videos, 4, chosen at random, were repeated (seen twice at separate points), while another 4 were "golden" videos from KoNViD-1k [23], for which subjective ratings were available. After each task, we rejected a subject if their scores on the same repeated videos or on the gold standard videos were not similar enough.

Through all these careful checks, a total of 1,046 subjects were rejected over all sessions.

3.2.3 Data Cleaning

Following the subject rejection, we conducted extensive data cleaning: **(1)** We excluded all scores provided by the subjects who were blocked, or for whom $>$ 50% of the

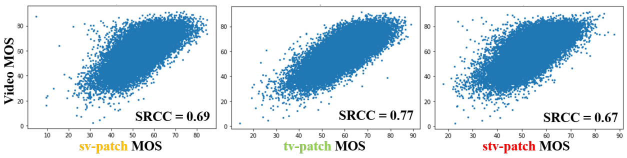


Fig. 6: **Scatter plots of patch-video MOS correlations** Video MOS vs sv-patch (left), tv-patch (middle) and stv-patch (right) MOS cropped from the same video.

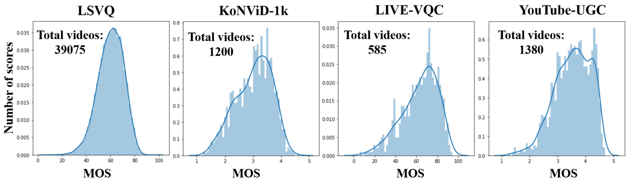


Fig. 7: **Ground Truth MOS histograms of four "in-the-wild" databases**. Starting from left, proposed LSVQ dataset, KoNViD-1k [23], LIVE-VQC [57], and YouTube-UGC [74].

videos stalled during a session. **(2)** We removed ratings given by people who did not wear their prescribed lenses during the study (1.13%), as uncorrected vision could affect perceived quality. **(3)** We applied ITU-R BT.500-14 [25] (Annex 1, Sec 2.3) standard rejection to screen the remaining subjects. This resulted in 301 subjects being rejected (about 2.6%). **(4)** To detect (and reject) outliers, we first calculated the kurtosis coefficient [3] of each score distribution, to determine normality. We then applied the Z-score method in [24] if the distribution deemed Gaussian-like, and the Tukey IQR method [66] otherwise (suppl. material). The total number of ratings collected after cleaning was around 5.5M (1.4M on videos and 4.1M on v-patches).

3.2.4 Data Analysis

Inter-subject consistency: On the cleaned data, we conducted an inter-subject consistency test [76, 57]. Specifically, we randomly divided the subjects into two equal and disjoint sets and computed the Spearman Rank Correlation Coefficient (**SRCC**) [31] between the two sets of MOS over 50 such random splits. We achieved an average SRCC of **0.86** on full videos, and **0.71**, **0.71** and **0.67** for sv-patches, tv-patches, and stv-patches, respectively. This indicates a high degree of agreement between the human subjects, implying a successful screening process (suppl. material).

Intra-subject consistency: We computed the Linear Correlation Coefficient (**LCC**) [50] between subjective MOS against the original scores on the "golden" videos, obtaining a median PCC of **0.96** on full videos, and **0.946**, **0.95**, and **0.937** for sv-patches, tv-patches, and stv-patches, respectively. These high correlations further validate the efficacy of our data collection process.

Relationship between patch and video quality: Fig 6 shows scatter plots of the video MOS against each type of v-patch MOS. The calculated SRCC between the video MOS and the sv-patch, tv-patch and stv-patch MOS was **0.69**, **0.77**, and **0.67** respectively, indicating strong relationships between global and local quality, even though the v-patches are relatively small volumes of the original video data.

MOS Distributions: Fig. 7 plots the MOS distribution of the videos in the new dataset as compared to other popular "in-the-wild" video quality databases [23, 57, 74]. The new dataset has a narrower distribution than the others, which again, matches actual social media data. Such a narrow distribution makes it more challenging to create predictive models that can parse finely differing levels of quality.

4. Modeling a Blind Video Quality Predictor

Taking advantage of the unique potential of the new dataset (Sec. 3), we created a deep video quality prediction model, which we refer to as Patch-VQ (PVQ), and a spatio-temporal quality mapper called PVQ-Mapper, both of which we describe next.

4.1. Overview

Contrary to the way most deep image networks are trained, we did not crop, subsample, or otherwise process the input videos. Any such operation would introduce additional spatial and/or temporal artifacts, which can greatly affect video quality. Processing input videos of diverse aspect ratios, resolutions, and durations, however, makes training an end-to-end deep network impractical. To address this challenge, PVQ extracts spatial and temporal features on unprocessed original videos, and uses them to learn the local to global spatio-temporal quality relationships. As illustrated in Fig 8, PVQ involves three sequential steps: feature extraction, feature pooling, and quality regression. First, we extract features from both the 2D and 3D network streams, thereby capturing the spatial and temporal information from the whole video. Three kinds of v-patch features are also extracted from the output of both networks, using spatial and temporal pooling layers to capture local quality information. Finally, the pooled features from the video and the v-patches are processed by a time series network that effectively captures perceptual quality changes over time and predicts a single quality score per video. We provide more details of each step below.

4.2. Feature Extraction

To capture the spatial aspects of both perceptual video quality and frame content, we extracted per frame (2D) spatial features using the PaQ-2-PiQ backbone pre-trained on the LIVE-FB Dataset [76]. To capture temporal distortions, such as flicker, stutter, and focus changes, we extracted spatio-temporal (3D) features using a 3D ResNet-18 [20] backbone, pre-trained on the Kinetics dataset [29].

4.3. Feature Pooling

Spatial and temporal pooling is applied in stages to extract features from the specified spatio-temporal regions of

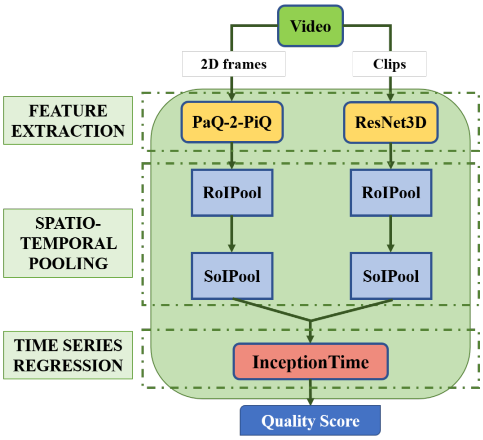


Fig. 8: **Illustrating the proposed PVQ model** which involves 3 sequential steps: feature extraction, spatio-temporal pooling, and temporal regression (Sec. 4.1).

interest (v-patches), allowing us to model local-to-global space-time quality relationships.

Spatial Pooling: The extracted 2D and 3D features are independently passed through a spatial RoIPool (region-of-interest pooling) layer [17, 18], with regions specified by the 3D v-patch coordinates. RoIPool helps compute a feature map with a fixed spatial extent of 2 × 2. The RoIPool layer generates 4 feature vectors of size 2048 per frame and video clip, for all three v-patches and the full video.

Temporal Pooling: The RoIPool layer is followed by an SoIPool (segment-of-interest pooling) layer [9] that helps compute a feature map with a fixed temporal extent. Specifically, an SoIPool layer with a fixed temporal extent of 16 is applied on both 2D and 3D features of each v-patch and the full video. The SoIPool layer yields 4 feature vectors of size 16 × 2048 per all three v-patches and the full video.

4.4. Temporal Regression

The resulting space-time quality features are fed to InceptionTime [13], a state-of-the-art deep model for Time Series Classification (TSC). InceptionTime consists of a series of inception modules (with intermittent residual connections) followed by a global average pooling and a fully connected layer. The inception modules learn changes in the quality features over time, which is crucial to accurately predict global video quality. Although RNNs have been used to model temporal video quality [37, 77], we have found that InceptionTime [13] is much faster and easier to train compared to RNN, does not suffer from vanishing gradients, and gives better performance.

5. Experiments

Train and test splits: The entire dataset of videos, v-patches, and human annotations was divided into a training and two test sets. We first selected those videos having both of their spatial dimensions greater than 720, and reserved it for use as a secondary testing set (about 9% of the LSVQ dataset: 3.5K videos and 10.5K v-patches). About 93.2% of the videos in the reserved set have resolutions 1080p or higher, hence we will refer to it as “Test-1080p”. On the remaining videos, we applied a typical 80-20 split, yielding about 28.1K videos (and 84.3K v-patches) for training, and 7.1K videos (and 21.3K v-patches) for testing.

Table 2: **Performance on full-size videos in the LSVQ dataset.** Higher values indicate better performance. Picture based model is *italicized.*

	Test		Test-1080p	
Model	**SRCC**	**LCC**	**SRCC**	**LCC**
BRISQUE [45]	0.579	0.576	0.497	0.531
TLVQM [35]	0.772	0.774	0.589	0.616
VIDEVAL [65]	0.794	0.783	0.545	0.554
VSFA [37]	0.801	0.796	0.675	0.704
PVQ (w/o v-patch)	0.814	0.816	0.686	0.708
PVQ (w/ v-patch)	**0.827**	**0.828**	**0.711**	**0.739**

Input processing and training: Each video was divided into 40 clips of 16 continuous frames. For feature extraction, we used a batch size of 8 for 3D ResNet-18 and 128 for PaQ-2-PiQ. For spatial and temporal pooling, we provide sets of spatio-temporal coordinates $(x_1, x_2, y_1, y_2, t_1, t_2)$ of each v-patch. When training InceptionTime, we used a batch size of 128 and L1 loss to predict the output quality scores (details in suppl. material).

Baselines and metrics: The model comparisons were done on both videos and v-patches. We compared with a popular image model BRISQUE [45], by extracting frame-level features and training an SVR and two other shallow NR VQA models, TLVQM [35] and VIDEVAL [65], that perform very well on existing UGC video databases. We also trained the VSFA [37], which extracts frame-level ResNet-50 [22] features followed by a GRU layer to predict video quality. To study the efficacy of our local-to-global model, we trained two versions of our PVQ model, one with, and the other without the spatio-temporal v-patches. All models were trained and evaluated on the same train/test splits. Following the common practice in the field of video quality assessment, we report the performance using the correlation metrics SRCC and LCC.

5.1. Predicting global video quality

The quality prediction performance of the compared models on the new LSVQ dataset is summarized in Table 2. As is evident, the shallow learner using traditional features (BRISQUE [45]) did not perform well on our dataset. TLVQM [35], VSFA [37], and VIDEVAL [65] performed better, indicating that they are capable of learning complex distortions. While both PVQ models (with and without patches) outperformed other models, including the v-patch data resulted in a performance boost on both test sets. Particularly on higher resolution test videos (Test-1080p), the proposed PVQ model (trained with v-patches) outperforms the strongest baseline by **3.6%** on SRCC.

Table 3: **Results on the three v-patches** in the LSVQ dataset. Picture based model is *italicized*.

	sv-patch		tv-patch		stv-patch	
Model	**SRCC**	**LCC**	**SRCC**	**LCC**	**SRCC**	**LCC**
BRISQUE [45]	0.469	0.417	0.465	0.485	0.476	0.462
TLVQM [35]	0.575	0.543	0.523	0.536	0.561	0.563
VIDEVAL [65]	0.596	0.570	0.633	0.634	0.662	0.636
VSFA [37]	0.654	0.609	0.688	0.681	0.685	0.670
PVQ (w/o v-patch)	0.723	0.717	0.696	0.701	0.651	0.643
PVQ (w/ v-patch)	**0.737**	**0.720**	**0.701**	**0.700**	**0.711**	**0.707**

Table 4: **Ablation studies** conducted on the Test split of the LSVQ dataset. Higher values indicate better performance.

Model	SRCC	LCC	# parameters
PVQ$_{2D}$ (w/ v-patch)	0.774	0.774	16.3 M
PVQ$_{3D}$ (w/ v-patch)	0.805	0.805	38.3 M
PVQ (w/ sv-patch)	0.815	0.815	54.2 M
PVQ (w/ tv-patch)	0.817	0.818	54.2 M
PVQ (w/ stv-patch)	0.824	0.826	54.2 M
PVQ$_{\text{Mobile}}$ (w/ v-patch)	0.774	0.779	10.9 M

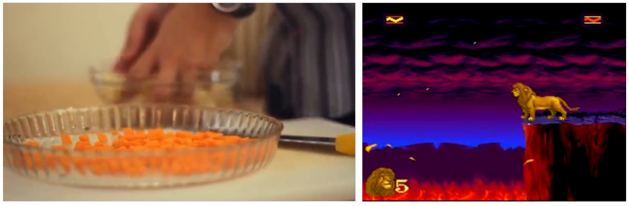
Performance on each v-patch: Table 3 sheds light on the capability of the compared models in predicting local quality. The two PVQ models delivered the best performance on all three types of v-patches, with the PVQ model trained on v-patches outperforming all baselines. From Tables 2 and 3, we may conclude that PVQ effectively captures global and different forms of local spatio-temporal video quality.

Contribution of 2D and 3D streams: We also studied the contribution of the 2D and 3D features towards the performance of PVQ by training separate models on 2D (PVQ$_{2D}$) and 3D (PVQ$_{3D}$) features alone (Table 4). As can be observed, PVQ$_{3D}$ achieves higher performance than PVQ$_{2D}$ on both test sets. This further asserts that 3D features are more capable of capturing complex spatio-temporal distortions, and thus more favorable for VQA.

Contribution of each v-patch: To study the relative contributions of the three types of v-patches in PVQ, we trained three separate models utilizing each patch separately (Table 4). Among the three, we observe that the highest performance is achieved when trained on stv-patches. Though stv-patches have relatively the least volume (Fig. 4), they contain the most localized information on video quality distortions, which could explain its better performance.

Mobile-friendly version: We also implemented an efficient version of PVQ for mobile and embedded vision applications (PVQ$_{\text{Mobile}}$), using the 2D and 3D versions of MobileNetV2 [54, 34] for the two branches, and by reducing the RoIPool output size to 1×1. Though there is a 6% decrease in performance as compared to PVQ (w/ v-patch), our mobile model requires only $1/5$ as many parameters (Table 4) compared to PVQ (w/ v-patch) and $1/2$ as many parameters compared to VSFA [37] (24M parameters).

Failure cases: The video in Fig 9 (a) was rated with a high score (MOS = 75.7) by human subjects, but was underrated by PVQ (predicted MOS = 47.4). We believe that an aesthetic "bokeh" blur effect was interpreted as high-quality content by subjects but such high levels of blur caused the model to predict low quality. The video in Fig 9 (b) was overrated by PVQ (predicted MOS = 54.7), considerably higher than the subject rating (MOS = 21). The video is of a computer-generated game and does not appear very distorted. Yet, the subjects may have expected higher resolution content for modern video games. These cases illustrate the challenges of creating models that closely align with human perception, while also highlighting the content diversity in the proposed dataset.



(a) Predicted: 47.6 Ground Truth MOS: 75.7

(b) Predicted: 54.7 Ground Truth MOS: 21

Fig. 9: **Failure cases:** Frames from video examples where predictions differed the most from the human quality judgements.

5.2. Predicting perceptual quality maps

We adapted the PVQ model (Sec. 4) to compute spatial and temporal quality maps on videos. Because of its flexible network architecture, PVQ is capable of predicting quality on any number (and sizes) of local spatio-temporal patches of an input video. We exploited this to create a temporal quality series and a first-of- its-kind video quality map predictor, dubbed PVQ Mapper.

Temporal quality series: A video is uniformly divided into 16 small temporal clips of 16 (continuous) frames each[2] and a single quality score per clip is computed, thus capturing a temporal series of perceptual qualities across a video.

Space-time quality maps: For space-time quality maps, we further divide each frame of each temporal clip defined above into a grid of 16×16 non-overlapping spatial blocks of the same aspect ratio as the frame and compute a local space-time video clip quality. Bi-linear interpolation was applied to spatially re-scale the spatio-temporal quality predictions to match the input dimensions.

Fig. 10 depicts the temporal quality series and magma color space-time quality maps that were α-blended ($\alpha = 0.8$) with original frames picked from the center of each clip. The series shows the video quality evolving over time. As may be observed, PVQ Mapper was able to accurately capture local quality loss, distinguishing blurred and under-exposed areas from high-quality regions, and high-quality stationary backgrounds from fast-moving, streaky objects.

Do v-patches matter for quality maps? Fig. 11 shows spatial quality maps on two sample videos generated by

[2] By changing the number of frames in each clip, the quality predictions can be made less or more dense.

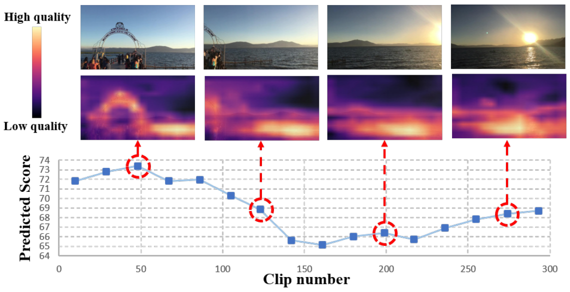


Fig. 10: **Space-time quality maps:** Space-time quality maps generated on a video using the PVQ Mapper (Sec. 5.2), and sampled in time for display. Four video frames are shown at top, with spatial quality maps (blended with the original frames using magma color) immediately under, while the bottom plots show the evolving quality of the video. Best viewed in color.

PVQ Mapper, trained with and without using v-patches. In Fig. 11 (a), the object in the foreground is focus blurred, whereas in Fig. 11 (b), the dog is motion blurred and the desk is underexposed. These local quality distortions are not captured with PVQ Mapper (w/o v-patch) as indicated in the middle row, but are distinctly evident in the output of PVQ Mapper (w/ v-patch) as indicated in the bottom row. This indicates that PVQ Mapper that uses v-patches is able to better learn from both global and local video quality features and human judgments of them, and hence predict more accurate quality maps.

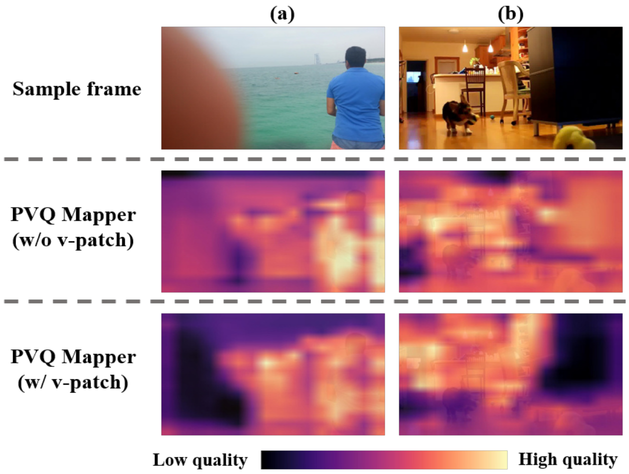


Fig. 11: **Improvement in quality maps when PVQ-Mapper is trained with patches** illustrating that learning from both local space-time and global video quality yields more accurate predictions. Best viewed in color.

5.3. Cross-database comparisons

To emphasize the validity and generalizability of the PVQ model, we also tested it on the two popular, yet much smaller "in-the-wild" video databases: KoNViD-1k [23] and LIVE-VQC [57] (Table 1). We compared the performance of PVQ against other popular models when each model was separately trained and tested on both datasets. As shown in Table 5, PVQ competes very well with other models on KoNViD-1k, while improves the SRCC on LIVE-VQC by **2.8%** compared to the strongest baseline. To further study the generalizability of PVQ, we also compared the performance of all models when trained on the proposed dataset (LSVQ) but tested on the two aforementioned datasets. From Table 6, it may be seen that PVQ transferred very well to both datasets. Specifically, our model outperforms the strongest baseline by **0.7%** and **3.6%** boost in SRCC on KoNViD-1k and LIVE-VQC respectively. This degree of database independence, both highlights the representativeness of the new LSVQ dataset and the general efficacy of the proposed PVQ model.

Table 5: **Cross-database comparison 1:** Performance when all models are separately trained and tested on KoNViD-1k [23] and LIVE-VQC [57].

	KoNViD-1k [23]		**LIVE-VQC** [57]	
Model	**SRCC**	**LCC**	**SRCC**	**LCC**
BRISQUE [45]	0.657	0.658	0.592	0.638
NRVQA-NSTSS [49]	0.642	0.653	-	-
V-BLIINDS [53]	0.710	0.704	0.694	0.718
VSFA [37]	0.773	0.775	0.773	0.795
TLVQM [35]	0.773	0.769	0.799	0.803
VIDEVAL [65]	0.783	0.780	0.752	0.751
RIRNet [10]	0.775	0.7812	-	-
PVQ (w/o v-patch) (Sec. 4)	**0.791**	**0.786**	**0.827**	**0.837**

Table 6: **Cross-database comparison 2:** Performance when all models are separately trained on the new LSVQ database, then evaluated on KoNViD-1k [23] and LIVE-VQC [57] **without fine-tuning.**

	KoNViD-1k [23]		**LIVE-VQC** [57]	
Model	**SRCC**	**LCC**	**SRCC**	**LCC**
BRISQUE [45]	0.646	0.647	0.524	0.536
TLVQM [35]	0.732	0.724	0.670	0.691
VIDEVAL [65]	0.751	0.741	0.630	0.640
VSFA [37]	0.784	0.794	0.734	0.772
PVQ (w/o v-patch) (Sec. 4)	0.782	0.781	0.747	0.776
PVQ (w/ v-patch) (Sec. 4)	**0.791**	**0.795**	**0.770**	**0.807**

6. Concluding Remarks

Predicting perceptual video quality is a long-standing problem in vision science, and more recently, deep learning. In recent years, it has dramatically increased in importance along with tremendous advances in video capture, sharing, and streaming. Accurate and efficient video quality prediction demands the tools of large-scale data collection, visual psychometrics, and deep learning. To progress towards that goal, we built a new video quality database, which is substantially larger and diverse than previous ones. The database contains patch-level annotations that enable us (and others) to make global-to-local and local-to-global quality inferences, culminating in the accurate and generalizable PVQ model. We also created a space-time video quality mapping model, called PVQ Mapper, which utilizes learned patch quality attributes to accurately infer local space-time video quality, and is able to generate accurate spatio-temporal quality maps. We believe that the new LSVQ dataset, the PVQ model, and PVQ Mapper, can significantly advance progress on the UGC VQA problem, and enable quality-based monitoring, ingestion, and control of billions of videos streamed on social media platforms.

References

[1] Face detection using haar cascades. *OpenCV-Python Tutorials*, [Online] Available: https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_objdetect/py_face_detection/py_face_detection.html.

[2] FFmpeg. [Online] Available: https://ffmpeg.org/.

[3] *Measure of Kurtosis*, pages 343–343. Springer New York, New York, NY, 2008.

[4] M. A. Saad A. Mittal and A. C. Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2015.

[5] S. Bosse, D. Maniry, T. Wiegand, and W. Samek. A deep neural network for image quality assessment. In *2016 IEEE Int'l Conf. Image Process. (ICIP)*, pages 3773–3777, Sep. 2016.

[6] M. Klimpke C. Keimel, J. Habigt and K. Diepold. Design of no-reference video quality metrics with multiway partial least squares regression. *2011 Third International Workshop on Quality of Multimedia Experience*, pp. 49-54, 2011.

[7] T. Oelbaum C. Keimel and K. Diepold. No-reference video quality evaluation for high-definition video. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 1145–1148, 2009.

[8] J. E. Caviedes and F. Oberti. No-reference quality metric for degraded and enhanced video. *Digit. Video Image Qual. Perceptual Coding*, pp. 305–324, 2017.

[9] Y. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018.

[10] Pengfei Chen, Leida Li, Lei Ma, Jinjian Wu, and Guangming Shi. *RIRNet: Recurrent-In-Recurrent Network for Video Quality Assessment*, page 834–842. Association for Computing Machinery, New York, NY, USA, 2020.

[11] A. C. Bovik A. K. Moorthy P. Panda D. Ghadiyaram, J. Pan and K. C. Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Trans. Circ. and Syst. for Video Tech.*, 2017. [Online] LIVE-Qualcomm Database: http://live.ece.utexas.edu/research/incaptureDatabase/index.html.

[12] M. Naccari S. Tubaro F. De Simone, M. Tagliasacchi and T. Ebrahimi. A h.264/avc video database for the evaluation of quality metrics. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 2430–2433, 2010.

[13] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean. Inceptiontime: Finding alexnet for time series classification, 2019.

[14] M. Tagliasacchi G. Valenzise, S. Magni and S. Tubaro. No-reference pixel video quality monitoring of channel-induced distortion. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 605–618, 2011.

[15] D. Ghadiyaram and A. C. Bovik. Blind image quality assessment on real distorted images using deep belief nets. In *IEEE Global Conference on Signal and Information processing*, volume pp. 946–950, pages 946–950, Atlanta, GA, 2014.

[16] 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, Jan 2016.

[17] R. Girshick. Fast R-CNN. In *IEEE Int'l Conf. on Comput. Vision (ICCV)*, page 1040–1049, 2015.

[18] R. Girshick. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Adv. Neural Info Process Syst (NIPS)*, 2015.

[19] F. G. Hahn, V. Hosu, H. Lin, and D. Saupe. No-reference video quality assessment using multi-level spatially pooled features, 2019.

[20] K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[21] D. Hasler and S. E. Suesstrunk. Measuring colorfulness in natural images. In *SPIE Conf. on Human Vision and Electronic Imaging VIII*, 2003.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vision and Pattern Recogn.*, pages 770–778, 2016.

[23] 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 *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2017. [Online] Database: http://database.mmsp-kn.de/konvid-1k-database.html.

[24] B. Iglewicz and D. C. Hoaglin. Volume 16: How to Detect and Handle Outliers. *The ASQC Basic References in Quality Control: Statistical Techniques*, 1993.

[25] International Telecommunication Union. ITU-R BT.500-14, methodologies for the subjective assessment of the quality of television images. [Online] Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-14-201910-I!!PDF-E.pdf.

[26] S. Forchhammer J. Søgaard and J. Korhonen. No-reference video quality assessment using codec analysis. *Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 10, pp. 1637–1650, 2015.

[27] L. P. Kondi K. Pandremmenou, M. Shahid and B. Lövström. A noreference bitstream-based perceptual model for video quality estimation of videos affected by coding artifacts and packet losses. *Human Vision and Electronic Imaging XX*, vol. 9394, pp.93941F, 2015.

[28] V. Asari K. Zhu, C. Li and D. Saupe. No-reference video quality assessment based on artifact measurement and statistical analysis. *Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 4, pp. 533–546, 2014.

[29] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[30] C. Keimel, A. Redl, and K. Dieopold. The TUM high definition video datasets. volume pp. 97-102, 07 2012.

[31] Maurice George Kendall. Rank correlation methods. 1948.

[32] J. Kim and S. Lee. Fully deep blind image quality predictor. *IEEE J. of Selected Topics in Signal Process.*, vol. 11, no. 1, pp. 206-220, Feb 2017.

[33] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and

A. C. Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130-141, Nov 2017.

[34] Okan Kopuklu, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll. Resource efficient 3d convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

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

[36] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, vol. 43, no. 1, pp. 29-44, 2001. [Online] Filter Bank: https://www.robots.ox.ac.uk/~vgg/research/texclass/filters.html.

[37] D. Li, T. Jiang, and M. Jiang. Quality assessment of in-the-wild videos. 2019.

[38] X. Li, Q. Guo, and X. Lu. Spatiotemporal statistics for video quality assessment. *IEEE Transactions on Image Processing*, 25(7):3329–3342, 2016.

[39] Y. Li, S. Meng, X. Zhang, S. Wang, Y. Wang, and S. Ma. UGC-VIDEO: perceptual quality assessment of user-generated videos, 2019.

[40] H. Lin, V. Hosu, and D. Saupe. Koniq-10K: Towards an ecologically valid and large-scale IQA database. *arXiv preprint arXiv:1803.08489*, March 2018.

[41] X. Liu, J. van de Weijer, and A. D. Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *IEEE Int'l Conf. on Comput. Vision (ICCV)*, page 1040–1049, 2017.

[42] A. C. Bovik M. A. Saad and C. Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.

[43] S. Stavrou M. T. Vega, D. C. Mocanu and A. Liotta. Predictive no-reference assessment of video quality. *Signal Processing: Image Communication*, vol. 52, pp. 20–32, 2017.

[44] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202-1213, March 2018.

[45] 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.

[46] 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*, 25(7):3073–3086, 2016.

[47] J. Park, S. Lee, and A.C. Bovik. VQpooling: Video quality pooling adaptive to perceptual distortion severity. *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 610-620, Feb. 2013.

[48] E. Peli. Contrast in complex images. *J. Opt. Soc. Am. A*, vol. 7, no. 10, pp. 2032–2040, Oct 1990.

[49] S. V. Reddy Dendi and S. S. Channappayya. No-reference video quality assessment using natural spatiotemporal scene statistics. *IEEE Transactions on Image Processing*, 29:5612–5624, 2020.

[50] Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[51] M.-N. Garcia S. Argyropoulos, A. Raake and P. List. No-reference video quality assessment for sd and hd h. 264/avc sequences based on continuous estimates of packet loss visibility. *2011 Third International Workshop on Quality of Multimedia Experience*, pp. 31-36, 2011.

[52] E. Allen S. Triantaphillidou and R. Jacobson. Image quality comparison between JPEG and JPEG2000. I. Psychophysical investigation. *Journal of Imaging Science and Technology*, vol. 51, no. 3, pp. 248-258, 2007.

[53] 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*, 21(8):3339–3352, 2012.

[54] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[55] K. Seshadrinathan and A. C. Bovik. Temporal hysteresis model of time varying subjective video quality. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1153–1156. IEEE, 2011.

[56] 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*, 19(6):1427–1441, 2010.

[57] 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, Feb. 2019. [Online] LIVE VQC Database: http://live.ece.utexas.edu/research/LIVEVQC/index.html.

[58] Z. Sinno and A. C. Bovik. Spatio-temporal measures of naturalness. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1750–1754, 2019.

[59] H. Talebi and P. Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998-4011, Aug 2018.

[60] *99Firms*. Facebook Video Statistics. [Online] Available: https://99firms.com/blog/facebook-video-statistics/.

[61] *Internet Archive*. Moving Image Archive. [Online] Available: https://archive.org/details/movies.

[62] *Maryam Mohsin, Oberlo*. 10 Youtube Statistics Every Marketer Should Know in 2020. [Online] Available: https://www.oberlo.com/blog/youtube-statistics.

[63] *Omnicore*. TikTok by the Numbers. [Online] Available: https://www.omnicoreagency.com/tiktok-statistics/.

[64] 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. 2015. [Online] Dataset Browser: http://projects.dfki.uni-kl.de/yfcc100m/.

[65] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik. UGC-VQA: Benchmarking blind video quality assessment for user generated content, 2020.

[66] J. Tukey. *Exploratory data analysis*. Addison-Wesley Pub. Co, Reading, Mass, 1977.

[67] 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, Nov 2017.

[68] Video Quality Experts Group (VQEG). VQEG HDTV phase I database. [Online] Available: https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx.

[69] P. V. Vu and D. M. Chandler. Vis3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *J. Electron. Imag.*, vol. 23, no. 1, p. 013016, Feb. 2014.

[70] S. Ahn J. Kim W. Kim, J. Kim and S. Lee. Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 219–234, 2018.

[71] Z. Duanmu W. Liu and Z. Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. *Proc. ACM Multimedia Conf. (MM)*, pp. 546–554, 2018.

[72] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C. . J. Kuo. Mcl-jcv: A JND-based h.264/avc video quality assessment dataset. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1509–1513, 2016.

[73] H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, Xin Zhou, M-O. Pun, X. Jin, R. Wang, X. Wang, Y. Zhang, J. Huang, S. Kwong, and C. C. Jay Kuo. Videoset: A large-scale compressed video quality dataset based on JND measurement, 2017.

[74] Y Wang, S Inguva, and B Adsumilli. Youtube UGC dataset for video compression research. 2019.

[75] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98-117, Jan 2009.

[76] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. C. Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3572–3582, 2020.

[77] J. You and J. Korhonen. Deep neural networks for no-reference video quality assessment. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2349–2353, 2019.

[78] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, April 2004.

[79] 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*, 29(8):2244–2255, 2019.