Rich features for perceptual quality assessment of UGC videos

Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck,
Balu Adsumilli, Peyman Milanfar, Feng Yang
Google Inc.
[yilin, junjiek, htalebi, joonggonyim, birkbeck, badsumilli, milanfar, fengyang]@google.com

Abstract

Video quality assessment for User Generated Content (UGC) is an important topic in both industry and academia. Most existing methods only focus on one aspect of the perceptual quality assessment, such as technical quality or compression artifacts. In this paper, we create a large scale dataset to comprehensively investigate characteristics of generic UGC video quality. Besides the subjective ratings and content labels of the dataset, we also propose a DNN-based framework to thoroughly analyze importance of content, technical quality, and compression level in perceptual quality. Our model is able to provide quality scores as well as human-friendly quality indicators, to bridge the gap between low level video signals to human perceptual quality. Experimental results show that our model achieves state-of-the-art correlation with Mean Opinion Scores (MOS).

1. Introduction

Video streaming services currently consume the majority of today's internet traffic [23]. Service providers typically optimize and stream transcoded versions of the content that may have come from a professional (e.g., Netflix) or casual creator (e.g., social media). In the former case, the service provider has a pristine original and can rely on full-reference video quality assessment methods to optimize the quality / bitrate of transcodes sent to viewers. Instead, social media platforms often serve User Generated Content (UGC), where the non-pristine version shared by a user often has pre-existing distortions or compression artifacts.

Given the prevalence of UGC on social media sharing platforms, understanding the perceptual subjective video quality of such content (and compressed versions of it) are important to make informed quality of service trade-offs. Unlike early approaches at blind video and image quality assessment, where a set of pre-existing distortions is applied to pristine content [36, 16], the biggest challenge of UGC is its diversity due to several factors. First, the content could be a popular live show watched by millions of people, or a bunch of meaningless frames with no views. Second, the original quality could be created by a 4K HDR camera with professional post-processing, or captured by a low-end shaky camera. Finally, it is unclear how many additional operations have been applied on the video: some videos have been cropped, rescaled, or heavily compressed before being uploaded. The UGC video quality discussed in this paper is a generic concept, encompassing a mixture of content attractiveness, aesthetic quality [21, 29], and compression artifacts [39]. Each of these factors affects a viewers' expectation of the video quality, and may significantly influence their watching experience.

While an uploader may not be able to alter the subject matter of the video, feedback that quantifies the contributing perceptual quality factors may benefit the uploader. Also, understanding the quality of the original can be used by the service provider to optimize recommendation systems (if multiple videos are present at a single event) or to further compress low quality originals with little or no perceptual impact on the final result [35]. Taking advantage of such optimizations allows for better user experience at lower cost for the provider. Perceptual quality metrics are also becoming an integral part of image and video enhancement frameworks [13, 28, 40], and have shown promising results.

In this paper, we propose a framework to analyze video quality in a comprehensive way to allow for all the above applications. Unlike traditional video quality assessment metrics that work as black boxes outputting a single quality score, our model also provides human-friendly descriptors (as illustrated in Table 1) that decomposes the perceptual quality of the content into its constituent parts. Our contributions are as follows[1]:

- An enhanced dataset to explore distinct characteristics of UGC video quality, which contains subjective data for both original videos and corresponding transcoded versions. The collected ground truth data makes to it possible to understand the relationship between video content and perceptual quality, and improve content-aware video compression (Sec. 3).

[1] All data are available at https://media.withyoutube.com/ugc-dataset

UGC videos				
CoINVQ diagnosis report				
Compression level	0.924	0.022	0.015	0.039
Content labels	Dance, Musical ensemble, Outdoor recreation	Food, Recipe, Cooking	Vehicle, Car, Video game	Action-adventure game, Vehicle, Cartoon
Distortion types	Gaussian blur, Multiplicative noise, Color saturation	Color saturation, Denoise, Pixelate	Color shift, Quantization, Contrast change	Multiplicative noise, Gaussian blur, Color saturation
Quality predicted by single feature (CP, CT, DT)	(2.862, 3.621, 3.16)	(3.107, 3.172, 2.95)	(3.69, 3.376, 3.548)	(4.029, 3.89, 3.941)
Quality predicted by all features (CP+CT+DT)	2.955	3.03	3.448	3.971
MOS (from subjective tests)	2.754	2.881	3.29	3.795

Table 1. Understanding generic UGC video quality by Comprehensive Interpretation Network for Video Quality (CoINVQ), which provided an overall quality estimation as well as human friendly quality indicators, including compression level (0: low, 1: high), content labels (3800+ UGC entities [2]), distortion types (20+ artificial distortions [17]). CP, CT, DT: features from Compression, Content, and Distortion submodels. Besides a single quality score, CoINVQ report also reveals the rationale of quality assessment. For example, the first video has interesting contents (Dance and outdoor recreation), i.e. its content quality is good (CT=3.621). However, it is heavily compressed (CP=2.862) and has distortions like blur and noise (DT=3.16), which leads to a poor watching experience (CP+CT+DT=2.955).

- We design a comprehensive framework to analyze UGC video quality from different aspects, such as semantic content, technical quality, and compression level, which brings new insights to interpret video perceptual quality as the interaction of complementary features (Sec. 4).

- The proposed model achieves state-of-the-art precision on UGC quality prediction, while also providing reliable indications for quality degradation caused by compression (Sec. 5).

2. Related Works

Perceptual quality for UGC videos is a broad concept. Besides compression artifacts, distortions introduced during the process of video production (like lens blur and camera shake) could also influence viewers' watching experience. Some large scale UGC image datasets have been released recently [12, 38, 6], but UGC video datasets are still very limited. Traditional public video quality datasets (e.g., LIVE datasets [24, 3, 9]) mainly focus on compression distortions introduced to pristine originals and contain limited UGC features. Some public UGC datasets, like YouTube-8M [2] and AVA [7] are designed for recognition and don't provide raw video data and corresponding MOS, making them less useful for quality assessment research. In contrast, a few large-scale UGC quality datasets [11, 26, 33] were released in the past two years that provide both raw videos and MOS. Within these datasets, YouTube's UGC dataset (YT-UGC) [33] is one of the most representative. It contains 1500 videos sampled from 1.5 million YouTube videos with the creative commons license. However, although one major goal of YT-UGC is to facilitate research on the practical applications of video compression and quality assessment, the current dataset doesn't contain any compressed versions of videos and corresponding differential-MOS (DMOS). Also videos within the provided coarse content categories show high quality diversity, making it difficult to establish a connection between content and quality.

Video quality assessment has been studied for decades and is still a challenging research topic. Reference quality metrics (e.g., PSNR, SSIM [36], and VMAF [16]) are designed for measuring relative quality changes from the reference (pristine original), and mainly focus on compression quality. Since traditional no-reference metrics [18, 20, 19, 34, 5] mainly rely on several manually designed features that are summarized from limited samples, they don't perform well on various UGC conditions. TLVQM [14] proposed 75 handcrafted features to handle various video distortions. Recent machine learning based metrics [15, 41] achieved significant improvements, benefiting from models pretrained on large scale datasets (e.g., ImageNet). However, these metrics are more or less biased on content related factors, and thus are less sensitive to small picture quality changes (e.g., may be caused by compression). How to build a metric for generic UGC video quality is still an open research topic.

3. YT-UGC$^+$: Content, Quality, Compression

To provide a comprehensive understanding of generic UGC video quality, we first explore and motivate the importance of UGC quality attributes (i.e., content labels, generic video quality, compression sensitivity) and connections among them. We reuse videos from the original YT-UGC dataset because of its broad diversity: 1500 20-second video clips, covering 15 categories (Animation, Cover Song, Gaming, HDR, How To, Lecture, Live Music, Lyric Video, Music Video, News Clip, Sports, Television Clip, Vertical Video, Vlog, and VR) and various resolutions (from 360P to 4K). We complement the original videos with content labels and compressed versions to enable thorough investigation on different aspects of UGC video quality. We call the enhanced dataset YT-UGC$^+$.

3.1. UGC Content Labels

Video content plays an important role in UGC perceptual quality and overall quality impression. Thus a good quality metric should have reasonable power of content recognition. Most UGC video datasets collect MOS [11, 26] or just provide content labels [2, 7]. The original YT-UGC dataset contains 15 high level content categories, which is too coarse to have significant descriptive power when predicting quality (explored further in Sec. 3.2). To enable further investigation on the connection between UGC content and perceptual quality, the first key feature of the YT-UGC$^+$ dataset is more fine-grained content labels.

We first used the public YT8M baseline model [2] to generate multiple labels, and selected top 12 confident labels as candidates. The advantage of using the YT8M model is that their labels (3862 coarse and fine-grained entities) have been refined for the UGC scenario, and the model was also trained on real YouTube clips, so we expect similar accuracy on the YT-UGC$^+$ dataset videos. Then we refined these candidate labels through a subjective test to get the final ground truth labels. Each video content was shown 4 times to the same subject, each time with 3 candidate labels as well as a "none of above" option. Each subject was asked to label 8 randomly selected videos, and finally every label on each video was voted by more than 10 subjects. We define the label confidence as its actual votes divided by its total shows. We set the minimum confidence at 0.2, then there are 610 individual labels from YT8M that appear on YT-UGC$^+$ videos. Besides high level content labels (e.g., Game and Musician), most labels are more concrete and fine-grained (e.g., Car, Tree, Dance, and Pet), which were more informative and descriptive of the content (Fig. 1).

3.2. From Content to Generic Video Quality

Diversity in generic video quality is an important characteristic of UGC. Fig. 2 shows MOS for the entire set as well as individual content categories. The majority of

Figure 1. Original videos with MOS and content labels.

MOS ranges from 3 to 4 (on a scale of 1-5), while HDR has the highest average quality (4.02) and CoverSong has the lowest average quality (3.25). The quality of HowTo, LyricVideo, NewsClip, and TelevisionClip are relatively uniformly distributed, while Gaming, Sports, and VerticalVideo have a bias on the high quality range. All content categories have large standard deviations on MOS implying it is difficult to map the high level content labels to perceptual quality at video level.

We also explored the correlation between MOS and the collected content labels from Sec. 3.1. To simplify the problem, we divided the MOS range into 3 quality levels (low, medium, and high) with thresholds 3.0 and 3.8 as low and high quality bars respectively. Armed with more fine-grained content labels, we start to find more interesting correlations between content and quality. For example, 52 video clips contain label "Strategy video game", and 65% of them are in the high quality range, which is much higher than the high quality ratio of label "Video game" (50%). In contrast, the label "Forest" appears on 7 videos, and 5 of them belong to low quality. The combination of multiple fine-grained content labels and deep content features can be even more indicative of the video quality (discussed in Sec. 4.1). The connection between content and visual quality is still an open question, and the content labels provided in YT-UGC$^+$ can be a benchmark to evaluate the power of a quality metric from content aspects.

3.3. UGC Compression Sensitivity

UGC compression has received more and more research interest recently [35]. However, most existing UGC datasets only contain original videos and their MOS. To enable more future research, we conducted another subjective experiment to collect MOS for compressed UGC videos. We selected all 720P and 1080P videos (189 originals) from three popular content categories (Gaming, Sports, and

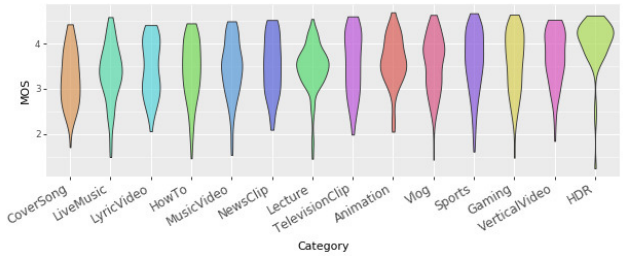

Figure 2. MOS distributions for original videos by content type.

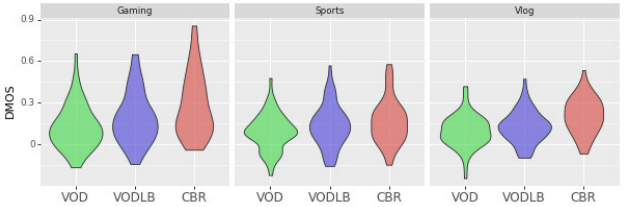

Figure 3. DMOS distributions for transcoded variants.

Figure 4. Low (top) & high (bottom) compression sensitivity

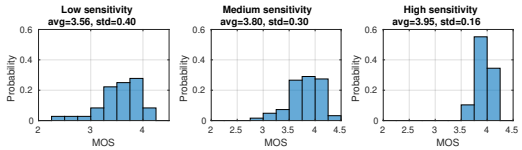

Figure 5. MOS distributions for original versions in different compression sensitivity levels.

Vlog). Each original video (the same one used in Sec. 3.2) was then transcoded by VP9 into three variants: Video-On-Demand (VOD), Video-On-Demand with Lower Bitrate (VODLB), and Constant Bit Rate (CBR), using the recommended VP9 settings and target bitrates [1]. VOD and VODLB are two-pass transcoding at the native resolution, where VODLB used the recommended target bitrate from the lower resolution (i.e., using 720P bitrate to compress 1080P video, and 480P bitrate for 720P video). CBR is one pass transcoding, which is widely used in live streaming. We restricted the device display height to be within 700P and 800P as that was the most popular resolution reported in the YT-UGC crowd-sourcing platform [37]. Videos were played in full screen mode, and subjects were asked to score quality between 1 and 5 for all 4 versions of the same content (playing in random order to reduce the influence of personal preferences). Finally, each video clip was rated by more than 30 subjects.

Fig. 3 shows the distribution of DMOS (= MOS(orig) - MOS(v)) for 3 variants v (VOD, VODLB, and CBR). In general, the VOD version has better quality than the VODLB version due to the higher target bitrate; and the CBR version tends to have the lowest quality since 1 pass transcoding is less optimized than 2 passes. The DMOS of the compressed variants is an important complement of the original MOS. A good quality metric should also be sensitive to the difference among those variants, i.e., it should have reasonable correlation with DMOS.

As pointed out in [35], the recommended settings may not be optimal for low quality UGC inputs, since the default settings have some bias on high quality inputs to avoid decreasing quality too much. To further investigate the impact of compression on UGC perceptual quality, we classified the videos into 3 compression sensitivity levels based on DMOS: low, medium, and high. For low sensitivity videos, all compressed variants' DMOS are less than T^{sc}_{low} (=0.1 in this paper). For high sensitivity, all DMOS are greater than T^{sc}_{high} (=0.2 in this paper). Other videos belong to the medium level. In general, low sensitivity means there are no significant quality differences between original and transcoded version. Those videos' bitrates could be further reduced to save users' bandwidth. For high sensitivity videos, the recommended VOD settings still causes noticeable quality degradation and should be improved.

We found that there are 36 (19%) videos at low sensitivity and 29 (15%) at high sensitivity, which means a significant amount of UGC videos could be further optimized (either reducing bitrate or improving quality). Two examples from low and high levels are shown in Fig. 4. Fig. 5 shows the MOS distribution for original versions with different compression sensitivity levels, where we can see the average MOS in low sensitivity is significantly lower than the other levels'. It matches the conclusion in [35] that people are less sensitive to quality changes in low quality videos than in high quality ones. Although videos at the high sensitivity level have the highest average MOS, some videos with high MOS (> 4) fall in (or below) the medium sensitivity group. This implies that in addition to input quality, compression sensitivity may be affected by other factors (e.g., video content). Compression sensitivity is a core characteristic of UGC, and we hope our data inspires more advanced optimizations for UGC compression and transcoding.

4. CoINVQ Framework

To further explore intrinsic properties of UGC perceptual quality, we propose a framework called Comprehensive Interpretation Network for Video Quality (CoINVQ) to (1) extract quality related features in various aspects to facili-

tate deeper understanding of the video, enabling customized treatment for improving video quality, and (2) predict the quality of the video with comprehensive features that allow for more quantitative analysis. Specifically, CoINVQ captures video quality effects from multiple aspects:

- **Content**: the meaningfulness and attractiveness of the video content inevitably affects viewers' attention as well as their quality sensitivity.
- **Distortion (technical quality)**: distortions could be introduced during the video producing stage. Some distortions are intended (e.g., proper sharpness filters or fine-tuning color saturation) and may have positive impact on perceptual quality. Unintended distortions (e.g., motion blur or jitter) have negative impact.
- **Compression level**: many UGC videos are compressed before being shared publicly due to bandwidth restrictions. Unlike technical quality, which is an intrinsic property of the video, compression artifacts are usually introduced by a third-party (e.g., upload app) and is adjustable. The impact of compression highly depends on the spatial/temporal complexity of content, and applying the same settings to different videos may cause completely different artifacts. Due to the complexity of video compression, we treat compression level as an individual aspect of the overall video quality.
- **Temporal aggregation**: a video is a sequence of small chunks (or frames), and those chunks could have different perceptual quality. Is the average chunk quality score a good representative of the overall video quality, or is a more complicated temporal aggregation strategy needed?

In our framework (Fig. 6), we decompose UGC quality understanding into these 4 sub-problems and feed input frames into 3 subnets (ContentNet, DistortionNet, and CompressionNet) to extract corresponding 2D deep features as well as high level quality indicators: content label, distortion types, and compression level (as shown in Table 1). The features are then concatenated together and aggregated through an AggregationNet to obtain the overall video quality estimation. Note that our framework is different from [8], where the baseline model was shared across different quality assessment tasks. As discussed in Sec. 5.2, using separate networks to learn different features performs better than learning all features through a single network.

Due to use cases and model training restrictions, the subnet inputs are slightly different from each other. ContentNet requires the entire frame to run overall semantic classification, and the result should not be affected by the input resolution, so we sampled frames at 1 frame per second (fps, common in video recognition applications) and resize them into a small resolution. DistortionNet and CompressionNet work on the native resolution to avoid rescaling artifacts. Frames are split into disjoint patches for feature extraction,

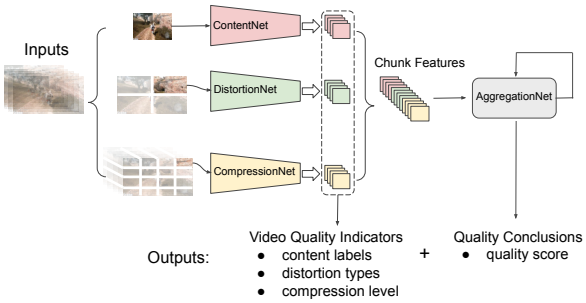

Figure 6. Overview of CoINVQ framework. Frames and chunks extracted from the input video are fed into three sub DNN models to extract corresponding features and quality indicators which are then aggregated to obtain the final quality score.

and patch features are then stitched together to obtain features for the entire frame. DistortionNet only takes a single frame (1 fps) while CompressionNet requires multiple frames (5 fps) to capture both spatial and temporal artifacts.

4.1. ContentNet

ContentNet is our video classification model that provides semantic-level embeddings for the UGC quality assessment task. It is a multi-label classification model trained on single video frames. The outputs of ContentNet are content-sensitive embedding features and the predicted content labels (defined in [2]). As can be seen in Fig. 7, the predicted content labels represent the frame semantics.

Inspired by recent success in deep CNN classifiers, we opt to fine-tune some existing pre-trained models on our data. We experimented with pre-trained ResNet-V2-50 [10] and EfficientNet [30] on the ImageNet dataset [22]. To adapt the classification CNN to our data, the following changes are applied to the baseline CNN model: 1) A fully connected layer with output feature maps of size (16, 16, 100) is added before the last layer, and 2) The last layer (head) is changed to output 3862 logits, which corresponds to the total number of video classes. We resize the input frames to 496×496, and use a cross-entropy loss to fine-tune the baseline CNN in a multi-label mode. Note that input frames may have up to 20 labels.

Our classification results for various baseline CNNs are shown in Table 2. In most cases both EfficientNet models outperform the ResNet model. Also, the difference between the two EfficientNet models tested on the YT8M data is not significant, but the gap seems to get larger on the UGC data. Note that our ultimate goal is to deploy the ContentNet embeddings into our perceptual quality assessment model. Hence, in our supplementary results we compare the impact of each model on the overall quality prediction problem. We found that retraining on YT8M achieved significantly higher correlations than directly using embeddings from ImageNet. These observations justify our efforts on

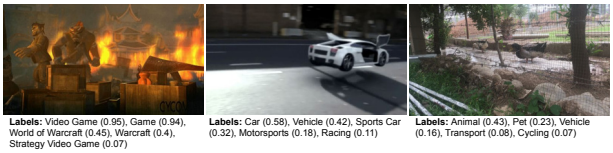

Figure 7. Examples of predicted content labels. The top-5 class predictions and probabilities are reported for each frame.

Model	YT8M			UGC-VQ		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
ResNet-V2-50	0.325	0.554	0.659	0.234	0.425	0.517
EfficientNet-b0	0.463	0.721	0.792	0.196	0.426	0.531
EfficientNet-b7	0.460	0.723	0.788	0.249	0.455	0.605

Table 2. Accuracy of the multi-label ContentNet model on the YT8M and the UGC datasets which have a total of 3862 classes.

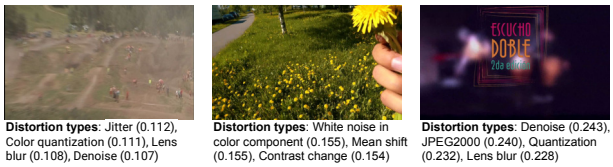

Figure 8. Examples of predicted distortion types.

fine-tuning CNNs specifically on UGC data. Finally, given the performance and the computational complexity of each CNN, we decide to use EfficientNet-b0 as our ContentNet.

4.2. DistortionNet

User generated content naturally contains various distortions (e.g., contrast change or denoise), which are orthogonal to video content features, and could have good or bad impact on perceptual quality. To learn quality related features in the distortion domain, we build our second sub-model called DistortionNet. The outputs of DistortionNet are distortion-sensitive embedding features and the detected distortion types (as shown in Fig. 8).

To achieve this goal, we train the network on synthetically distorted images from KADIS-700K and KADID-10K [17]. The dataset provides pristine original images and 25 distortion filters, like High sharpen, Denoise, and Gaussian blur. Each filter can generate distortions in 5 different levels, so each original has 125 distorted variants. We used EfficientNet-b0 (pre-trained on ImageNet) as the backbone network. The training loss contains three parts. The first one is cross-entry loss, L_T^{DT}, for multi-label (distortion type) classification. The second one is the pairwise hinge loss, L_P^{DT}, between the two randomly selected variants with the same distortion type. We use $L_T^{DT}+L_P^{DT}$ as loss to train the initial DistortionNet on KADIS-700K. Then we fine-tune the model on KADID-10K dataset. Since KADID-10K also provides ground truth MOS, we use a separate MLP head to predict the MOS scores and train with a L2 distance loss (L_M^{DT}), and the total loss for training on KADID-10K is $L^{DT} = L_T^{DT} + L_P^{DT} + L_M^{DT}$. The final model achieves an accuracy of 0.97 on distortion type classification, and 0.74 MOS correlation on KADID-10K dataset.

Figure 9. Examples of predicted compression levels.

4.3. CompressionNet

Most video sharing platforms transcode the original video into different bitrates/resolutions to meet device and network requirements. Commonly used video compression strategies are lossy, causing noticeable quality degradation. Such compression artifacts could heavily influence the watching experience, so we built an isolated sub-model to learn compression related features. The outputs of CompressionNet are compression-sensitive embedding features as well as a continuous compression level score in the range of 0 (no compression) to 1 (heavy compression) (see Fig. 9).

Due to the limited public UGC video sets with associated ground truth MOS, we used self-supervised learning to train the model. The input original videos are transcoded by VP9 with two different compression strengths: VOD (2 passes) with recommended bitrate (75kbps) and CBR (1 pass) with a low bitrate (20kbps). The underlying quality order is: Orig $\approx$ VOD $>$ CBR. Five frames were uniformly sampled from the original and transcoded clips, and fed into a shared $D3D$ model [27] to get predicted compression level. A $(1, 4, 4, 100)$ feature layer, inserted before the fully connected layer, is used to extract compression features.

The loss function contains two parts: pairwise loss (L_P^{CP}) and contrastive loss (L_C^{CP}). We set $L_P^{CP} = sigmoid((orig - cbr) * K)$ (where $K = 4$) to evaluate the compression level difference between the Original and CBR versions. To compute the contrastive loss, features were projected into a 1d space (1×1600) by two dense layers (to form a nonlinear mapping), and the similarity ($sim(x, y)$) between two features is defined by their feature distance [32]: $L_C^{CP} = sim(orig, vod)/(sim(orig, vod) + sim(orig, cbr) + sim(vod, cbr))$. The final loss is defined as $L^{CP} = L_P^{CP} + L_C^{CP}$.

4.4. AggregationNet

A common way to estimate video quality is to use the average of frame quality scores, and it performs well on most public video datasets [31]. It is unclear whether such basic pooling methods still work well in UGC scenario, or whether more elaborate aggregation strategies could achieve better accuracy. To investigate the impact of temporal pooling strategies, we compared 3 aggregation mod-

els, AvgPool, LSTM, and ConvLSTM, on YT-UGC original MOS. LSTM and ConvLSTM are classical temporal models. In AvgPool model, each chunk feature is filtered by a 1×1 Conv2D layer (256 units) to refine the feature space. Those refined features are then sent through a shared 2D head (formed by BatchNormalization, Activation(relu), GlobalMaxPool2D, Dropout, and Dense layers) separately to get per chunk scores, whose average is used as the final quality score. The experimental results show that AvgPool has better performance than LSTM and ConvLSTM on our dataset, which suggests that the majority of UGC videos still have relatively consistent quality. This matches the observation in [35] where the average chunk MOS has 0.976 correlation with the entire video MOS for original videos in YT-UGC dataset. Detailed comparison can be found in supplementary material.

We use the absolute difference between ground truth MOS and predicted score as the loss to train these 3 aggregation models. The difference of predicted MOS between target and reference videos gives the predicted DMOS. In this way, no-reference CoINVQ scores can be used to measure quality degradation caused by video compression.

5. Experimental Results

5.1. Implementation Details

All sub-networks of the CoINVQ framework were trained separately. ContentNet and DistortionNet are frame based and use EfficientNet-b0 [30] (pretrained on ImageNet) as the backbone network. CompressionNet is trained by D3D model [27] (pretrained on Kinetics-600 [4]), to learn both spatial and temporal video features.

Three datasets were used to retrain these sub models. We randomly selected 1-second chunks from 100k 1080P video from YT8M dataset, downscaled into 180P (320×180) to remove noticeable compression artifacts, and used them as the original version for training CompressionNet. We then selected another 100k frames from YT8M 1080P videos, using YT8M baseline model to get top 20 predicted labels, then joined with the ground truth video labels to obtain the refined labels for that particular frame to train the ContentNet. For DistortionNet, the input images are cropped and resized to 360x640. We pool the final convolution layer to the size of $(8, 8, 100)$ as our deep distortion features. For each MLP head, we use a single fully connected layer with 512 units. The model was first trained on the KADIS-700K (excluding type 13 and 23 due to the license issue), then fine-tuned on the KADID-10K dataset [17].

ContentNet and DistortionNet were trained on 30 V100 GPU with a batch size of 4 using RMSProp optimizer. The learning rate starts from 0.001 is decayed by a factor of 0.99 every 15k steps, and training converged around 2.5 million steps. CompressionNet was trained on a TPU v2 with a batch size 64. Learning rate is 0.0001 with a cosine decay, and the training converged after 10k steps.

To align with the display resolution of the subjective data, all YT-UGC original videos are first rescaled to 720P, and divided into 4 disjoint 360P patches for DistortionNet and 16 180P patches for CompressionNet. Deep embedding features from 3 pre-trained submodels: Compression (CP), Content (CT), and Distortion (DT), are used to train AggregationNet on YT-UGC original MOS data. The training was on a TPU v2 (batch size 256, learning rate 0.0001 with a cosine decay), and converged around 20k steps.

5.2. Evaluation on Original MOS

In the following experiments, we evaluate the model performance against the YT-UGC original MOS data. We use 5-fold cross-validation with consistent splits for all tests and report average results over the test folds. All compared metrics are evaluated at the original video resolutions with their default parameters, and output scores are rescaled into $[1, 5]$ using a nonlinear logistic function [25].

Table 3 compared CoINVQ with popular no-reference metrics. We can see non-learning based metrics (BRISQUE [18], NIQE [20], and VIIDEO [19]) didn't perform well on UGC cases, probably because UGC patches don't always follow some traditional characteristics found in pristine videos, like Natural Scene Statistics.

All machine learning based metrics were fine-tuned on the YT-UGC dataset with the same 5 splits. We first evaluated two recent video quality metrics (TLVQM [14] and VSFA [15]). TLVQM is based on 75 hand-crafted features and then fine-tuned with Support Vector Regression (SVR) and Random Forest Regression (RFR). VSFA is a state-of-the-art deep learning based video quality metric. It performs better than metrics based on hand-crafted features, achieving similar correlation as CoINVQ(CT+DT).

We then evaluated several frame-based models. The first two models were retrained EfficientNet-b0 (pretrained on ImageNet) with frozen and trainable weights on frames extracted from YT-UGC originals (100 frames per video, assuming all frames have same MOS as the video's). The result of EfficientNet-b0 (frozen) roughly matched CoINVQ(CT), since both of them learned content related features with a trainable head. We also tested DistortionNet (retrained on KADIS-700K and KADIS-10K with MOS), and their correlations are around 0.73, close to CoINVQ(DT), but worse than CoINVQ(CT+DT). This implies many pre-learned content features were overwhelmed by new learned distortion features in the DistortionNet, but much better preserved when content and compression features are learnt separately and joined as in CoINVQ.

For CoINVQ models, combining features performs better than using a single feature, and CP+CT+DT has the highest correlation on all feature combinations. Compression and distortion features seem to have higher impact on visual quality than content features, but content features

Model	PLCC	SROCC	RMSE
BRISQUE [18]	0.112	0.121	0.639
NIQE [20]	0.105	0.236	0.640
VIIDEO [19]	0.146	0.130	0.637
TLVQM(SVR) [14]	0.697	0.722	0.479
TLVQM(RFR) [14]	0.719	0.730	0.448
VSFA [15]	0.761	0.761	0.431
EfficientNet-b0(frozen)	0.624	0.612	0.509
EfficientNet-b0(finetune)	0.671	0.690	0.474
DistortionNet(frozen)	0.732	0.735	0.443
DistortionNet(finetune)	0.732	0.738	0.435
CoINVQ (CP)	0.770	0.785	0.408
CoINVQ (CT)	0.628	0.628	0.495
CoINVQ (DT)	0.726	0.744	0.434
CoINVQ (CP+CT)	0.787	0.801	0.395
CoINVQ (CP+DT)	0.790	0.802	0.391
CoINVQ (CT+DT)	0.750	0.767	0.421
CoINVQ (CP+CT+DT)	**0.802**	**0.816**	**0.382**

Table 3. No-reference metrics on YT-UGC original MOS.

Model	Trained on KoNViD-1k			Trained on YT-UGC		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
TLVQM(SVR)	0.758	0.763	0.421	0.482	0.484	0.599
TLVQM(RFR)	0.723	0.723	0.445	0.515	0.521	0.628
VSFA	**0.792**	**0.782**	**0.398**	0.602	0.599	0.425
CoINVQ						
(CP)	0.725	0.727	0.441	0.517	0.521	0.550
(CT)	0.674	0.667	0.469	0.525	0.535	0.542
(DT)	0.737	0.742	0.436	0.602	0.614	0.511
(CP+CT)	0.755	0.755	0.417	0.636	0.647	0.493
(CP+DT)	0.747	0.749	0.427	0.628	0.643	0.498
(CT+DT)	0.757	0.760	0.415	0.630	0.637	0.497
(CP+CT+DT)	0.764	0.767	0.413	**0.670**	**0.685**	**0.480**

Table 4. Performance on KoNViD-1k original MOS.

also bring in additional gain when combining with other features. Thus the 3 extracted features all inform the perceptual quality. Adding a new feature gives 1 to 2% increase in correlation, which implies those features cover complementary aspects of the perceptual quality.

We also evaluate CoINVQ on another UGC dataset KoNViD-1k [11] by retraining AggregationNet with the original MOS. Again we used 5-fold cross-validation and report average results on test folds (Table 4). The combined features (CP+CT+DT) still achieve the highest correlations among CoINVQ models, and adding additional features has positive impact. CoINVQ(DT) gets better correlations than CoINVQ(CP), probably due to fewer videos in KoNViD-1k having compression issues than in YT-UGC dataset. The retrained CoINVQ(CP+CT+DT) performs slightly better than TLVQM, but worse than VSFA, which may also be relevant to the lower importance of compression features on KoNViD-1k dataset. Using models trained on YT-UGC MOS to predict on KoNViD-1k video quality, then our CoINVQ models outperforms both VSFA and TLVQM.

Feature	PLCC	SROCC	RMSE
PSNR	0.402	0.389	0.099
SSIM [36]	0.493	0.479	0.093
VMAF [16]	0.401	0.399	0.143
LPIPS [41]	0.524	0.507	0.095
TLVQM(SVR) [14]	0.180	0.123	0.207
TLVQM(FRF) [14]	0.276	0.246	0.149
VSFA [15]	0.403	0.384	0.151
CoINVQ (CP)	0.640	0.590	0.196
CoINVQ (CT)	0.570	0.511	0.106
CoINVQ (DT)	0.315	0.325	0.235
CoINVQ (CP+CT)	**0.660**	**0.594**	**0.192**
CoINVQ (CP+DT)	0.476	0.459	0.232
CoINVQ (CT+DT)	0.312	0.335	0.201
CoINVQ (CP+CT+DT)	0.500	0.490	0.203

Table 5. Results on YT-UGC$^+$ compressed DMOS (not retrained).

5.3. Evaluation on Compressed Video DMOS

The 5 trained models from the previous section were used to predict the quality scores on compressed videos (i.e., no retraining was done for this section). For no-reference metrics (CoINVQ, TLVQM, VSFA), we compute DMOS as the MOS difference between the original version and the transcoded variant.

Table 5 shows the correlation of different features on DMOS prediction, as well as some popular reference quality metrics. The full-reference metrics, PSNR, SSIM, and VMAF, highly depend on pixel level difference or handcrafted features, so their performance is worse than deep-learned metric like LPIPS [41]. LPIPS mainly relies on features extracted from ImageNet and is slightly less correlated to UGC compression than our content features directly extracted from UGC videos. CoINVQ(CP) had better correlation than CoINVQ(CT), as expected, and combining CP and CT gives the best performance. An interesting observation is that distortion features seem not as useful for compression quality prediction, and adding them in the model may give worse results. It suggests that we should carefully choose features when dealing with different quality prediction tasks. Naively combining as many features as possible may not always return the best results.

6. Conclusion

In this paper, we discussed the main challenges in UGC quality assessment. We complement YT-UGC dataset with content labels and compressed videos with corresponding DMOS to get a new dataset YT-UGC$^+$. A comprehensive framework for UGC quality assessment is proposed, and we demonstrated that combining features learned from different quality aspects can achieve better performance than single features. There are still many open questions for UGC quality assessment, and we hope our work inspires more advanced research in this area.

References

[1] VP9 encoding recommendations. https://developers.google.com/media/vp9. 4

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2, 3, 5

[3] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik. Study of temporal effects on subjective video quality of experience. *IEEE Transactions on Image Processing*, 26(11), 2017. 2

[4] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *CoRR*, abs/1808.01340, 2018. 7

[5] Chao Chen, Mohammad Izadi, , and Anil Kokaram. A no-reference perceptual quality metric for videos distorted by spatially correlated noise. *ACM Multimedia*, 2016. 2

[6] T. Y. Chiu, Y. Zhao, and D. Gurari. Assessing image quality issues for real-world problems. In *CVPR*, 2020. 2

[7] Chunhui Gu, Chen Sun, et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 2, 3

[8] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5

[9] D. Ghadiyaram, J. Pan, and A.C. Bovik. A subjective and objective study of stalling events in mobile streaming videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(1), 2017. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 5

[11] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2017. 2, 3, 8

[12] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 2

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711, 2016. 1

[14] J. Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 2019. 2, 7, 8

[15] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019. 2, 7, 8

[16] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *Blog, Netflix Technology*, 2016. 1, 2, 8

[17] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 2, 6, 7

[18] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 2012. 2, 7, 8

[19] A. Mittal, M. A. Saad, and A. C. Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 2016. 2, 7, 8

[20] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 22(3), 2013. 2, 7, 8

[21] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 1

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[23] Sandvine. The global internet phenomena report. https://www.sandvine.com/phenomena, 2019. 1

[24] 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), 2010. 2

[25] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, Nov 2006. 7

[26] Z. Sinno and A. C. Bovik. Large scale subjective video quality study. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 276–280, Oct 2018. 2, 3

[27] J. C. Stroud, D. A. Ross, C. Sun, J. Deng, and R. Sukthankar. D3d: Distilled 3d networks for video action recognition. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 6, 7

[28] Hossein Talebi and Peyman Milanfar. Learned perceptual image enhancement. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–13. IEEE, 2018. 1

[29] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 2018. 1

[30] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning, 2019*. 5, 7

[31] Zhengzhong Tu, Chia-Ju Chen, Li-Heng Chen, Neil Birkbeck, Balu Adsumilli, and Alan Bovik. A comparative evaluation of temporal pooling methods for blind video quality assessment. In *ICIP*, 2020. 6

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[33] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube UGC dataset for video compression research. In *2019 IEEE 21st*

International Workshop on Multimedia Signal Processing (MMSP), 2019. 2

[34] Yilin Wang, Sang-Uok Kum, Chao Chen, and Anil Kokaram. A perceptual visibility metric for banding artifacts. *IEEE International Conference on Image Processing*, 2016. 2

[35] Yilin Wang, Hossein Talebi, Feng Yang, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, and Peyman Milanfar. Video transcoding optimization based on input perceptual quality. In *Proc. SPIE 11510, Applications of Digital Image Processing XLIII*, 2020. 1, 3, 4, 7

[36] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 2004. 1, 2, 8

[37] J. G. Yim, Y. Wang, N. Birkbeck, and B. Adsumilli. Subjective quality assessment for youtube UGC dataset. In *ICIP*, 2020. 4

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

[39] Kai Zeng, Tiesong Zhao, Abdul Rehman, and Zhou Wang. Characterizing perceptual artifacts in compressed video streams. In *Proc. SPIE 9014, Human Vision and Electronic Imaging XIX*, 2014. 1

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2, 8