Google

Video Quality Dataset and Model in YouTube

Yilin Wang, Balu Adsumilli

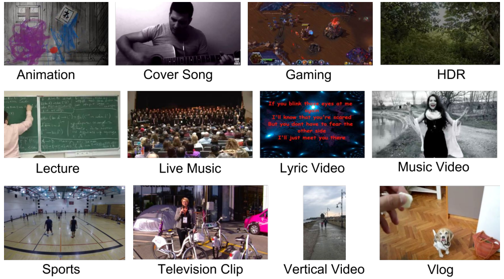
YouTube Media Algorithms Team

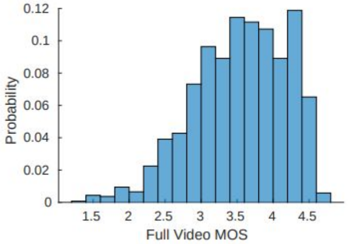
Part 1: Subjective Quality Datasets

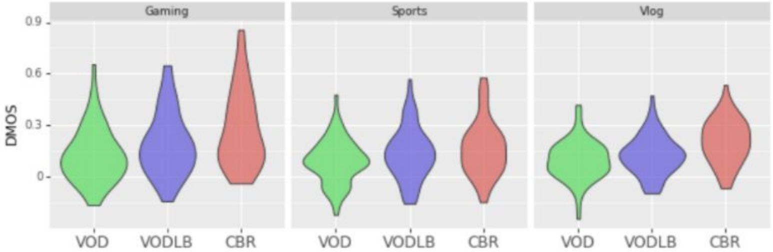
Google

YouTube UGC Dataset (media.withyoutube.com)

- Original videos (2019)
  - 1500 UGC (from 1.5 millions videos)
  - 15 categories (e.g. gaming, music, etc.)
  - multiple resolutions (from 360P to 4K)
- MOS for all originals (2020)
- DMOS for popular categories (2021)
  - multiple VP9 transcoded variants
- Content labels (2021)
  - 600+ refined fine grained labels







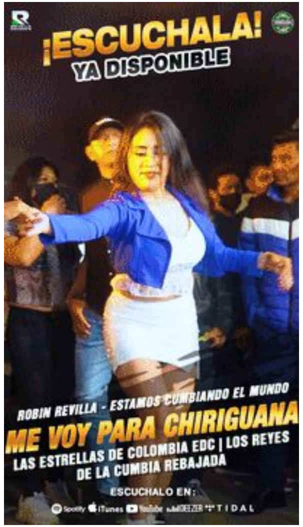
MOS: 4.55
Labels: Outdoor recreation(0.455), Game(0.455), Ball(0.455), Baseball bat(0.364), Cricket(0.182), Yo-yo(0.182), Walking(0.091), Mabinogi (video game)(0.091)

MOS: 4.33
Labels: Beach(0.917), Eating(0.500), Resort(0.417), Ibiza(0.333), Nail (anatomy)(0.333), Food(0.167), Swimming pool(0.083), Hotel(0.083),Bar(0.083)

Balu Adsumilli et al., "Launching a YouTube dataset of user-generated content", YouTube tech blog
Yilin Wang et al., "YouTube UGC Dataset for Video Compression Research", MMSP 2019
Joong Yim et al., "Subjective Quality Assessment for YouTube UGC Dataset", ICIP 2020
Yilin Wang et al., "Rich features for perceptual quality assessment of UGC videos", CVPR 2021

Google

Short Form Video (SFV): a new video form with billions of users

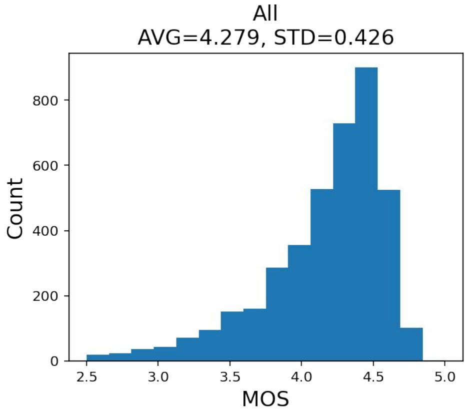


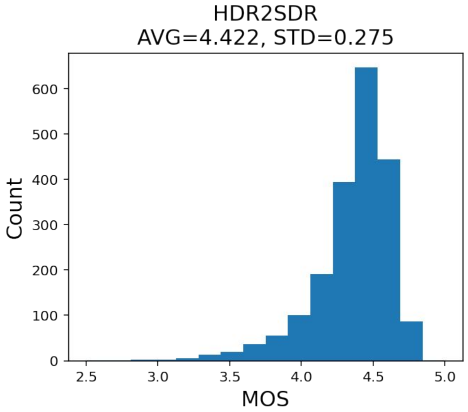
YouTube SFV-HDR Quality Dataset (2024)

Dataset Overview

Sampling Pool	80,000 YouTube SFV with Creative Commons
Color Space	SDR, HDR
Resolution	1080 × 1920
Video length	5s
Content category	Animal, Cooking, Dance, Gameplay, Health, Hobby, Music, Society, Speech, Sports
Videos	SDR (2030), HDR2SDR (2000), HDR (2000)
Subjective scores (MOS in [1, 5])	SDR (2030), HDR2SDR (2000), HDR (300)

SDR MOS Analysis

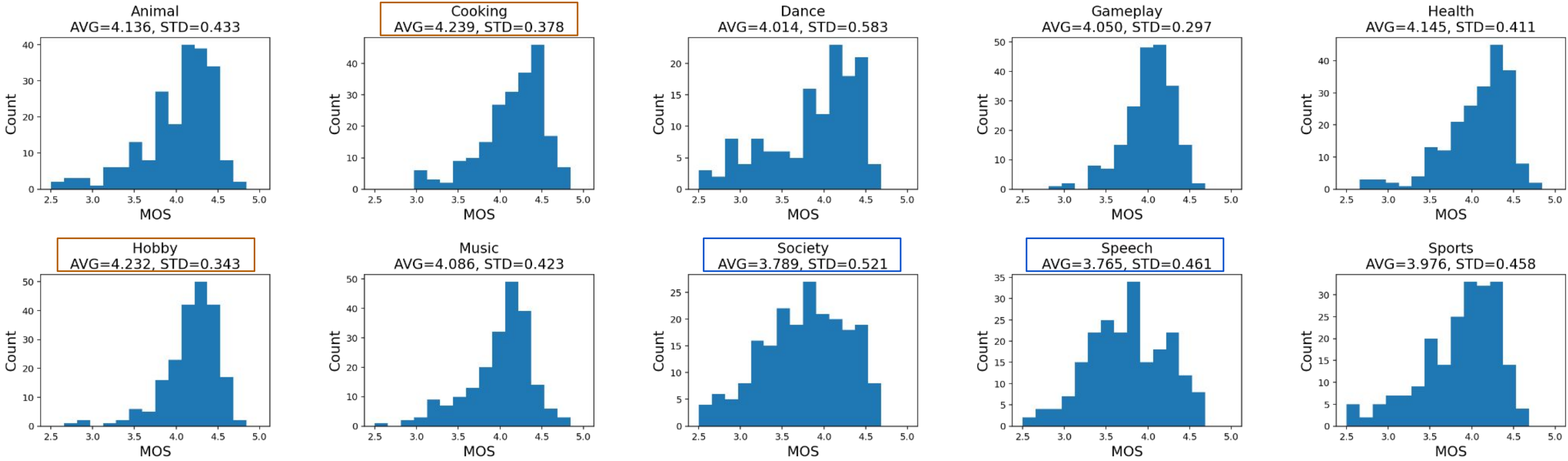




- Relatively narrow MOS distribution
  - 80% MOS values are within [3.8, 4.6]
- MOS of most HDR2SDR (90%) are higher than 4.0
  - potential reason 1: HDR videos are usually captured by high end devices (natively provides high picture quality)
  - potential reason 2: The color plays an important role in quality assessment

Google

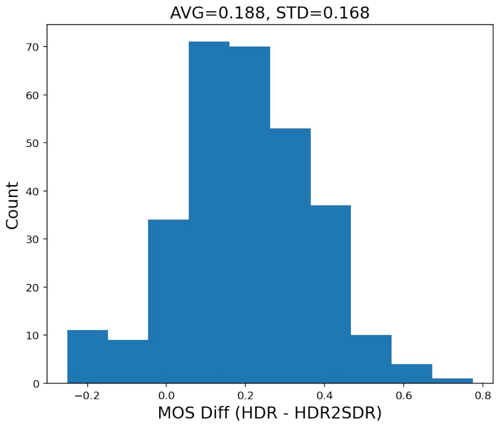
SDR MOS Analysis (per content category)

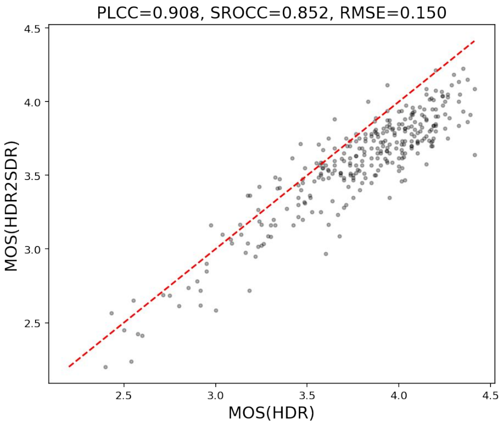


- Society and Speech have relatively uniform distributions (and lower average quality)
  - potential reason 1: many content were recorded in public spaces with restricted lighting and device control.
  - potential reason 2: viewers are not very interested in such contents and intuitively avoid giving very high scores.
- Cooking and Hobby have the highest average quality
  - potential reason 1: creators fully control the recording environments and are able to do sophisticated post-enhancements.
  - potential reason 2: contents are widely interesting

HDR MOS Analysis

- HDR experience is highly device-dependent
  - in-lab studies using Pixel 7 pro
- Most HDR MOS are significantly higher than corresponding HDR2SDR version
  - Viewer feedback: HDR videos are significantly brighter with more clear details than SDR versions





Google

Objective Metric Performance

- HDR2SDR has lower correlation than SDR videos
- PLCC for Gameplay is significantly lower than other categories

SDR_Gameplay_s0pc
MOS=4.26,
DOVER=3.26,
FAST-VQA=2.95,
FasterVQA=3.02

SDR_Gameplay_wcq2
MOS=4.15,
DOVER=2.31,
FAST-VQA=1.89,
FasterVQA=2.36

SDR_Gameplay_z5fz
MOS=4.22,
DOVER=3.20,
FAST-VQA=3.23,
FasterVQA=2.58

MOS correlations for all, native SDR, and HDR2SDR

	All		Native SDR		HDR2SDR	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
DOVER	0.781	0.702	0.793	0.750	0.618	0.496
FAST-VQA	0.797	0.752	0.789	0.789	0.664	0.543
FasterVQA	0.755	0.705	0.753	0.748	0.585	0.493

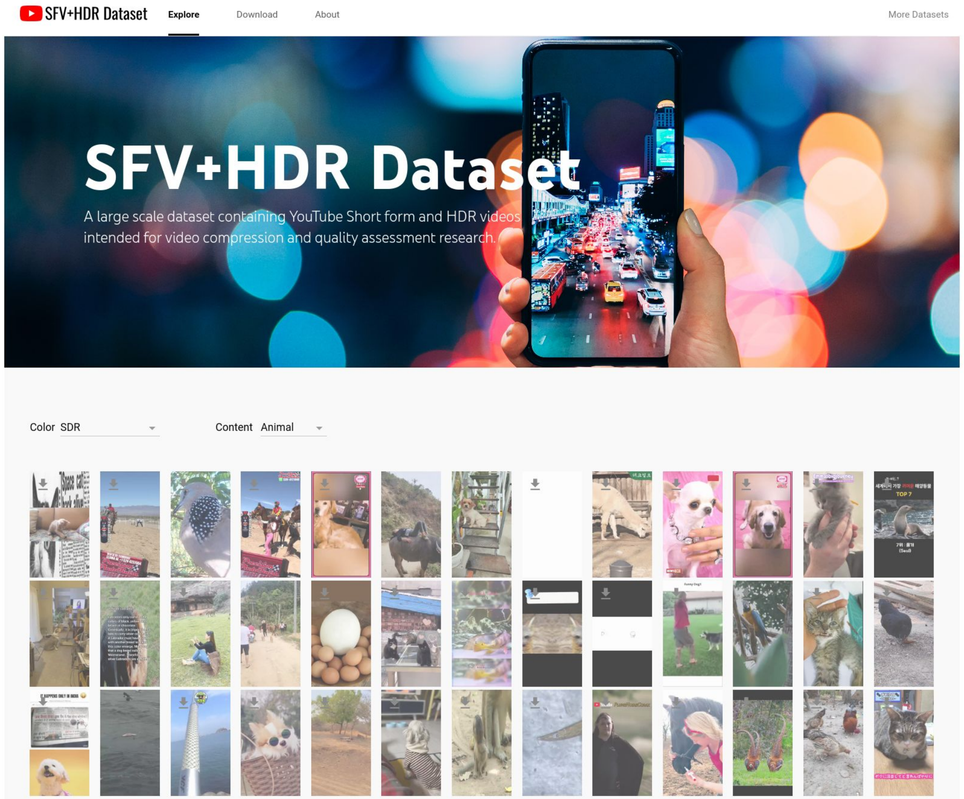
Per category MOS correlations for SDR videos

Category	DOVER PLCC/SRCC	FAST-VQA PLCC/SRCC	FasterVQA PLCC/SRCC
Animal	0.848/0.775	0.829/0.793	0.786/0.735
Cooking	0.753/0.731	0.733/0.775	0.646/0.664
Dance	0.883/0.851	0.882/0.866	0.860/0.833
Gameplay	0.639/0.545	0.634/0.557	0.640/0.558
Health	0.784/0.691	0.810/0.768	0.745/0.712
Hobby	0.596/0.568	0.708/0.693	0.606/0.617
Music	0.772/0.724	0.738/0.721	0.745/0.728
Society	0.842/0.843	0.770/0.796	0.759/0.798
Speech	0.843/0.841	0.827/0.826	0.805/0.810
Sports	0.826/0.781	0.789/0.778	0.749/0.729

Now available at: media.withyoutube.com/sfv-hdr

Including

- all raw SDR, HDR, and HDR2SDR videos
- MOS (crowdsourcing) for all SDR, and HDR2SDR videos
- MOS (in-lab) for selected HDR and corresponding HDR2SDR version



Yilin Wang, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, "Youtube SFV+HDR Quality Dataset", ICIP 2024

Part 2: Universal Video Quality (UVQ) model

Universal Video Quality (UVQ) model

UVQ Quality Report:

Overall quality score in [1, 5]

- Interpretation of UVQ scores
  - [1, 3.5): relatively low
  - [3.5, 4.1]: medium/fair
  - [4.1, 5]: relatively high
- Noticeable diff: 0.05~0.1 UVQ DMOS
- Score for this example: 3.15 (low quality)

Quality labels

- From high level (semantic) to low level (pixel difference)
- Labels for this example
  - Content: strategy video game,
  - Distortion: gaussian blur, pixelate
  - Compression: medium high

YouTube Media Algorithms & Google Research

Yilin Wang et al., "Rich features for perceptual quality assessment of UGC videos", CVPR 2021

Google

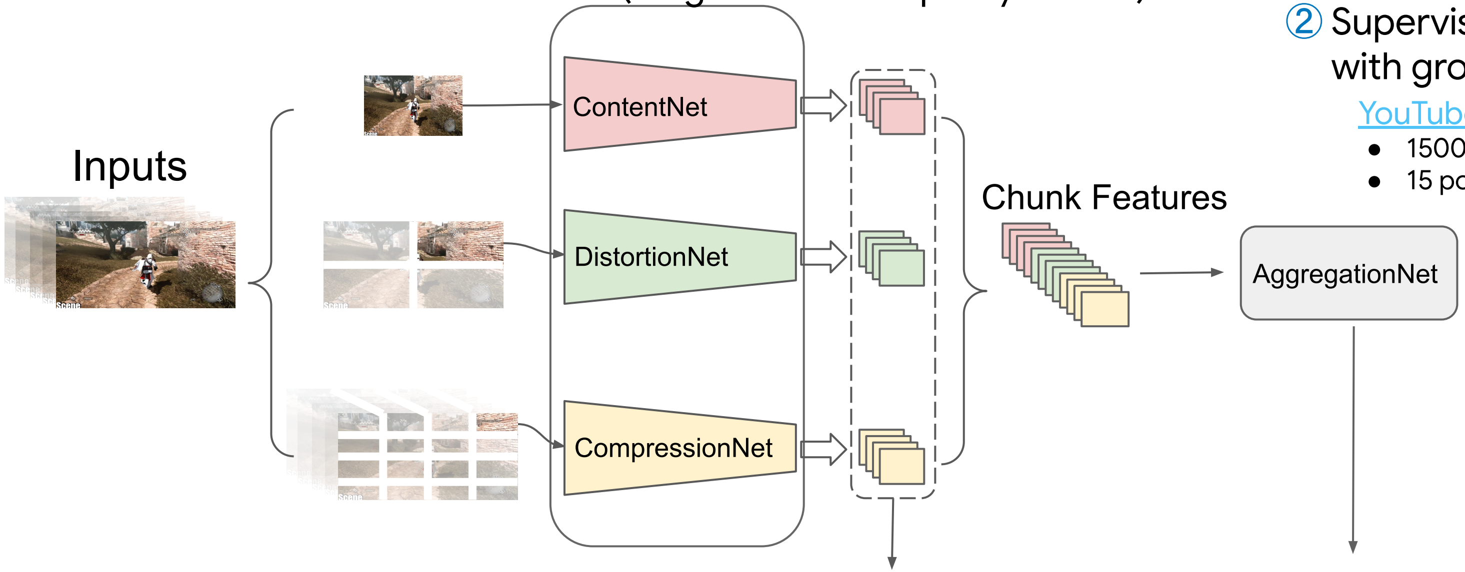
UVQ Framework

① Self-supervised Learning with **Millions** of training videos (no ground truth quality scores)

② Supervised Learning with ground truth data

[YouTube UGC Dataset]

- 1500 sampled from 1.5M videos
- 15 popular content categories



Outputs:

Video Quality Indicators
- content labels
- distortion types
- compression level

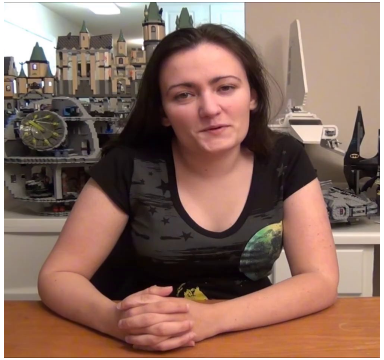
\+

Quality Conclusions
- quality score

Public link: github.com/google/uvq

A common reliability issue for no-ref quality metrics

Compression

	Original	Compressed (47.9kBps)	Compressed (25.8kBps)
			
Raw model	4.25	4.19	4.25
Refined model	4.46	4.34	4.23

Enhancement

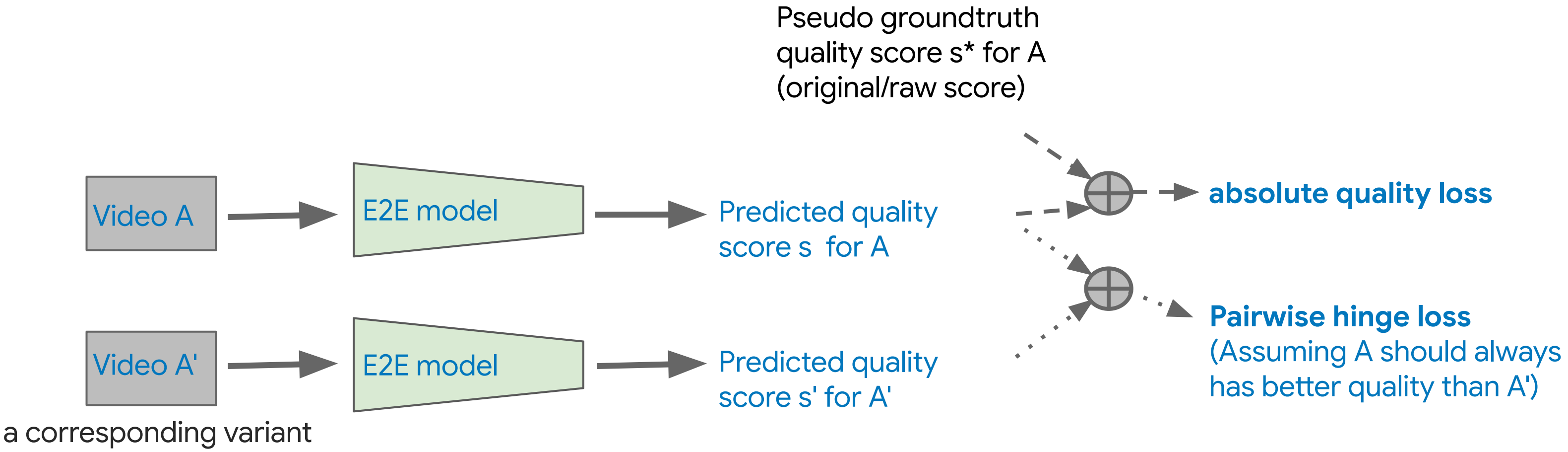
	Original	Sharpened	Over-sharpened
			
Raw model	3.06	3.49	3.19
Refined model	3.81	4.07	3.20

Solution: Retraining the model with both originals and variants

- Goal
  - To make the model more reliable/sensitive to small changes
- Original videos
  - ~30k videos from YT8M, 80% training, 20% validation
- Compression variants (33 variants)
  - AV1 / VP9 / H264: 100 kbps, 250 kbps, ..., 2500 kbps, 3000 kbps (11 bitrates)
  - Predefined quality order
    - For a given codec, higher bitrate has higher (or equal) quality.
    - For a given bitrate, newer codec has higher (or equal) quality (AV1 >= VP9 >= H264)
- Enhancement variants (6 variants)
  - Ffmpeg unsharp filter (s1, s2, s3 properly sharpened variants, s4, s5, s6 over-sharpened variants)
  - Predefined quality order
    - s6 <= s5 <= s4 <= orig <= s1 <= s2 <=s3
- Label: original/raw predicted quality scores

Model Refinement

- Goal
  - To be more sensitive and reliable to small variances
- Pairwise training framework



Improved Model Reliability

- Evaluated on selected challenging cases (1000 videos)
- Metrics
  - MOS correlation (PLCC)
  - Flip rate: the ratio of variant pairs counter to the predefined order

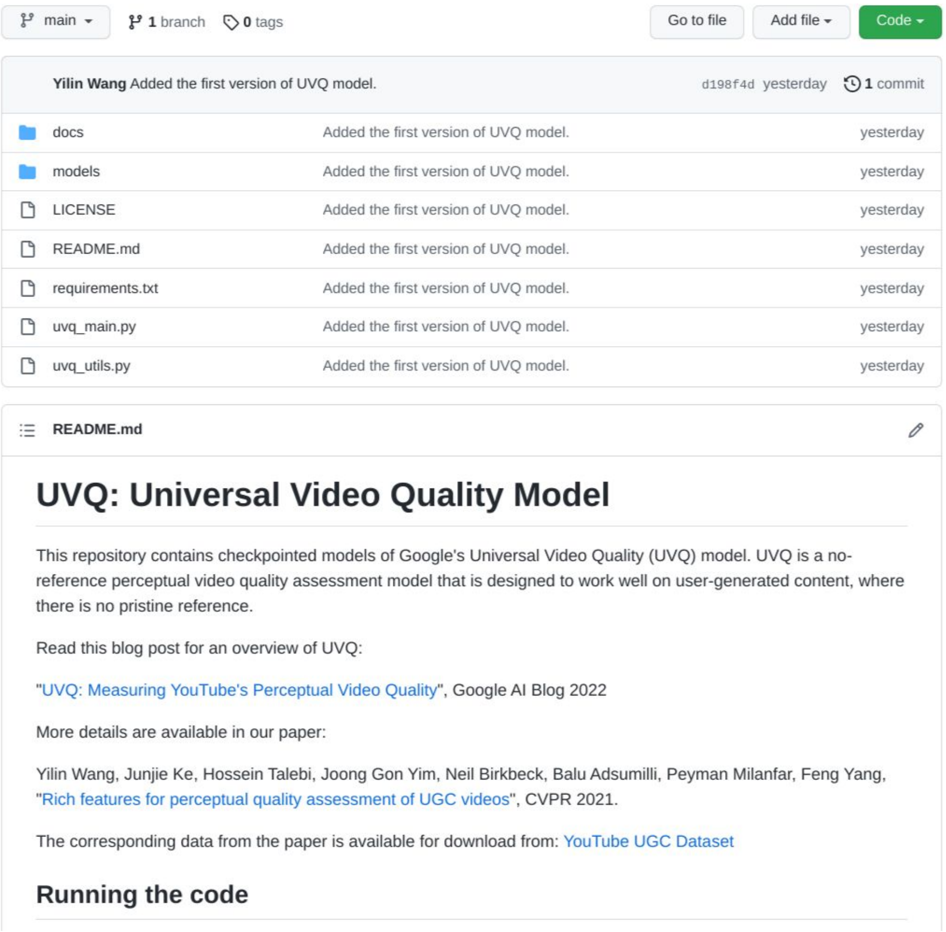
	PLCC↑	Flip rate (Compression)↓	Flip rate (Enhancement)↓
Original/raw UVQ	0.8200	0.1802	0.3325
Refined UVQ	0.8324	0.0339	0.0204

Google

UVQ public version

- Public link: github.com/google/uvq
- In the folder
  - UVQ models + runnable scripts
- Input
  - "video_id,length,filepath"
- Outputs
  - **overall scores + labels + raw features**

The pytorch version and the new robust version will be available next year.



Thanks!

Google

Join us at the AOMedia reception
Google and Meta lounges - ICC 3
13:45-14:30

Thank you!