YouVQ: A new no-reference metric for UGC

Media Algorithms Team

Balu Adsumilli

Yilin Wang

VQEG, Dec. 2021

Agenda

01

What is UGC?

02

YouTube UGC Dataset

03

Introducing YouVQ

What is UGC?

What we see

YouTube video traffic

- 500 hours of video shared every minute
- >2B daily active users in 100+ countries in 80+ languages
- 70% of YouTube is watched on mobile devices
- ~1400 combinations of codecs, containers, resolutions, and formats
- most of the videos uploaded are user generated content

User Generated Content (UGC)

Content and emotion **>** narrative and quality

- Artifact-ridden: shaky cam, low light, portrait, overlays, heavily compressed
- Variability due to content creator, network, unusual viewing environment

Current Video Quality Metrics

Subjective

- Mean Opinion Score (MOS)
- Differential Mean Opinion Score (DMOS)

Objective

- Reference-based metrics: PSNR, SSIM, VMAF
  - Assumes a pristine original that the target should “get close to”
- No-reference metrics: banding, noise, NIQE
  - Does not depend on the original, pristine or otherwise

Are any of these good for UGC?

YouTube

Trouble with existing notions

High Pixel Difference ≠ Low Perceptual Quality

MSE=27.10 MSE=21.26

Left image: greater MSE. Right image: much lower spatial frequencies.
Human vision system has a stronger response to the lower spatial frequencies .

Need for accurate no-ref metric for UGC

Growing need for a reliable no-reference fidelity metric (not artifact)

- Original video is either not available or not a reference (not pristine)
    - same relative quality deltas map differently for varying original video quality

Original

Transcoded
PSNR= 43.77, SSIM=0.969, VMAF=89.34

Similar perceptual quality (DMOS~=0)

UGC Video Quality Assessment

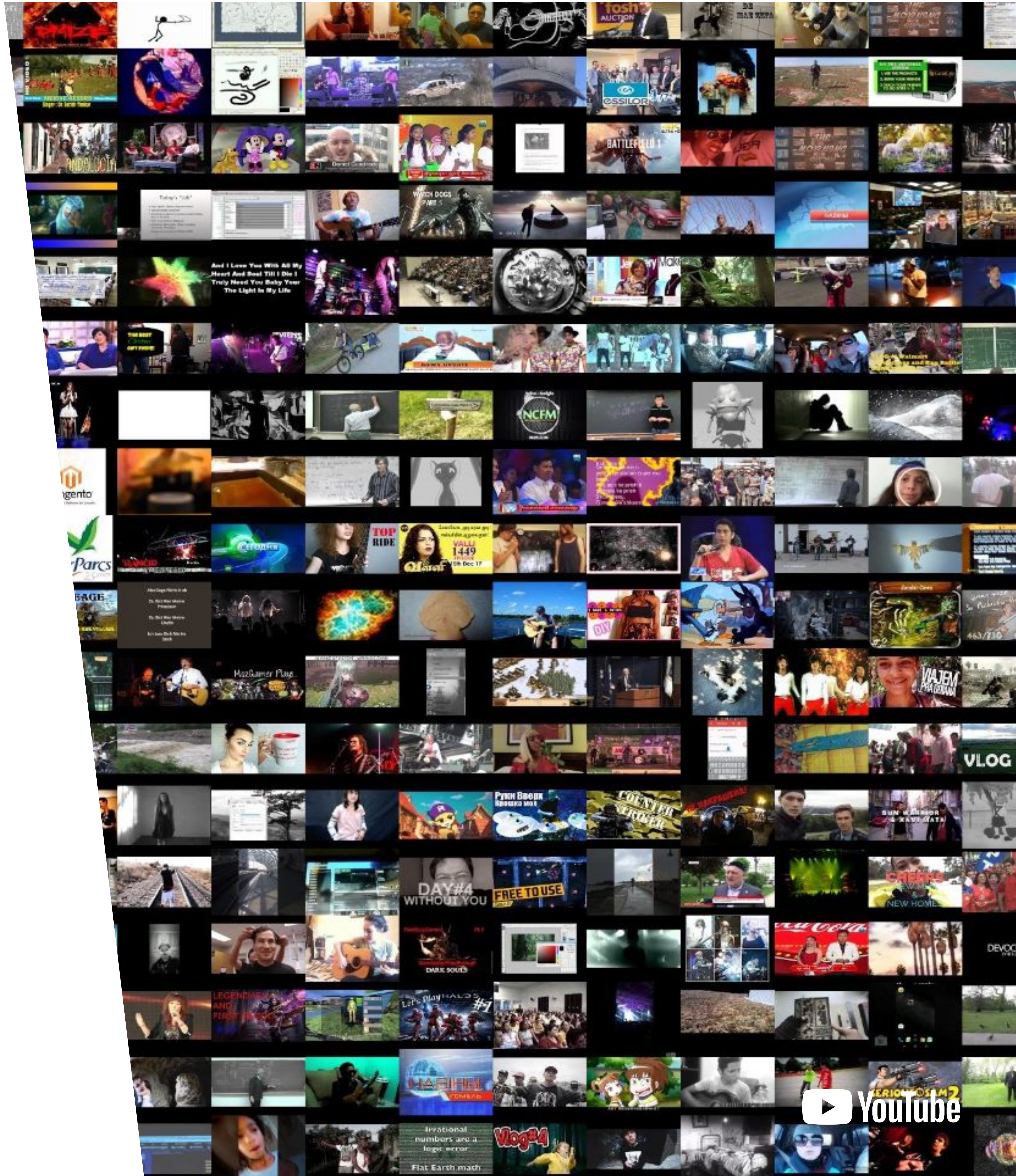
Foundational question: Will we need to rethink video quality metrics in the presence of non-pristine originals?

We start with a dataset

- Distributed across variations in content, complexity, resolutions, frame rates, formats
- Universal availability
- Ground truth subjective data

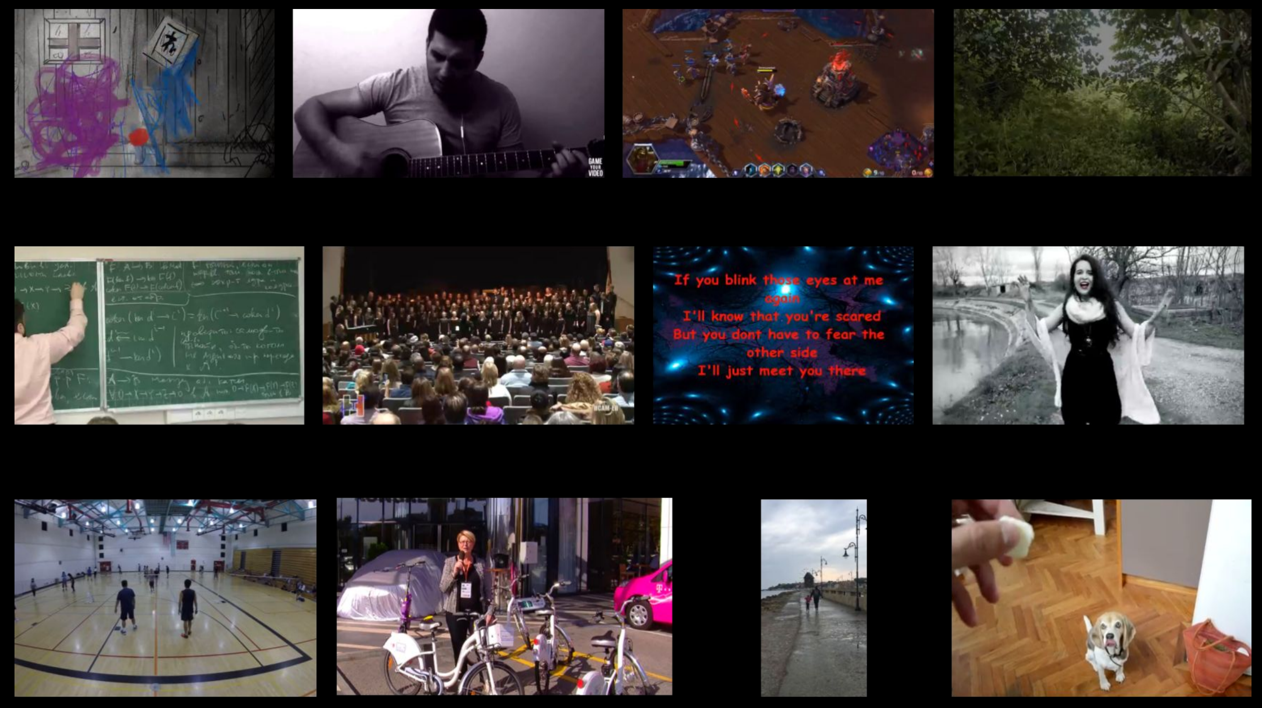
YouTube

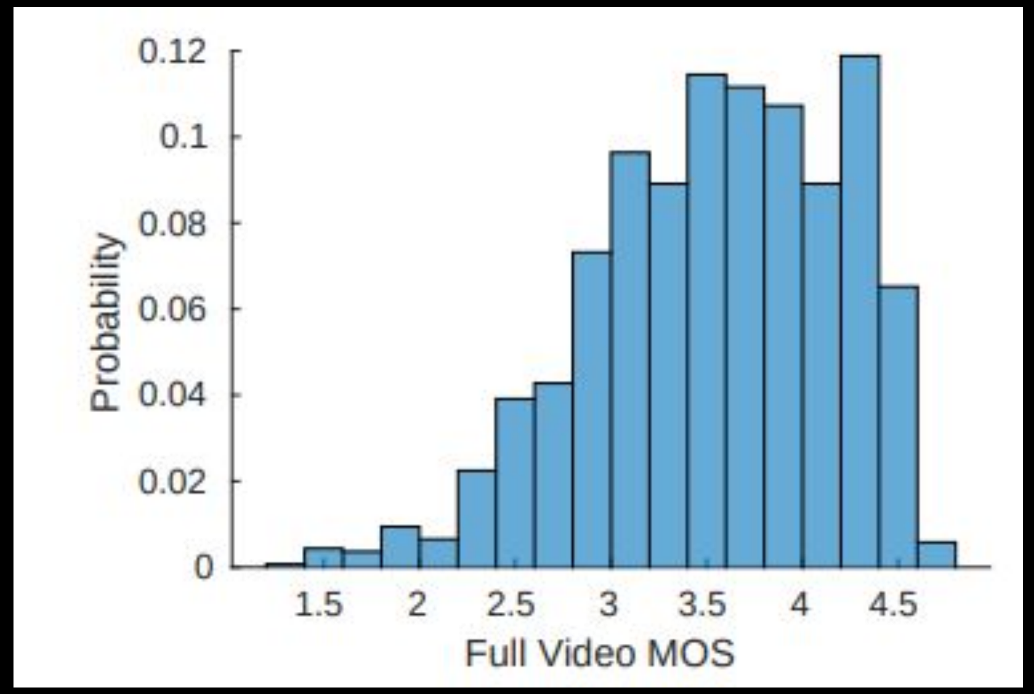
YouTube
UGC Dataset



YouTube UGC Dataset (YT-UGC): media.withyoutube.com

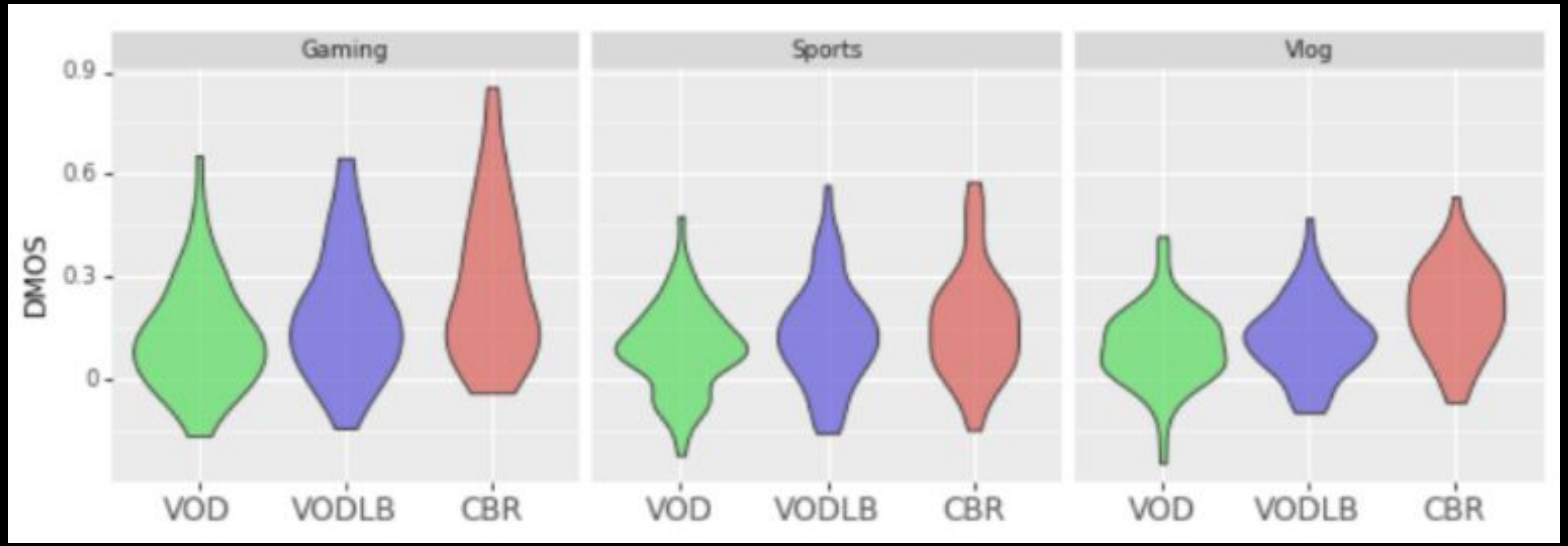
- 1500 Uploaded videos
  - Sourced from 1.5 million uploads
  - 15 content categories
  - Each video in multiple resolutions, fps





- Ground truth (MOS) for all videos

- Added DMOS for popular categories
- Added 600+ content labels

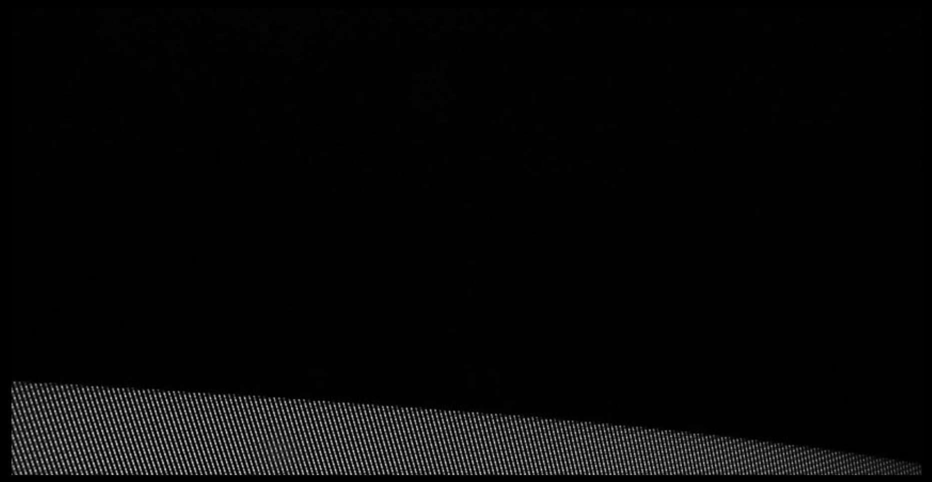
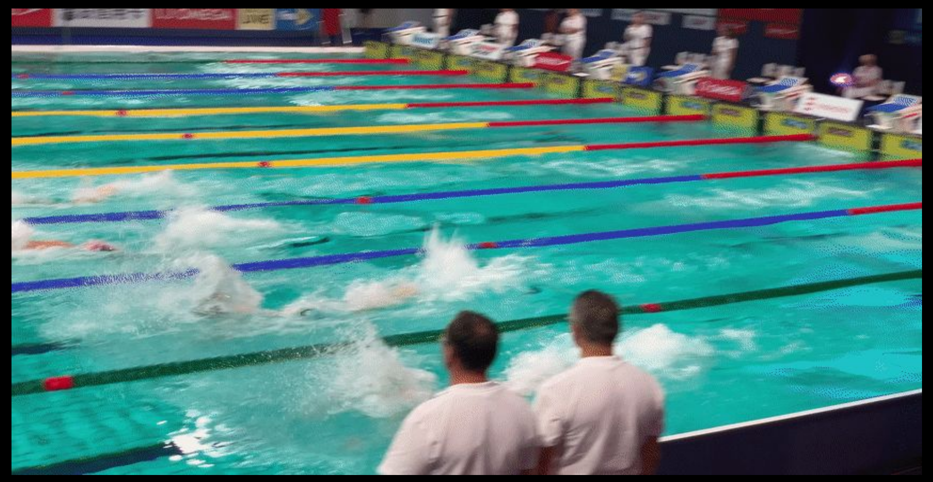


Balu Adsumilli et al., "[Launching a YouTube dataset of user-generated content]()", YouTube tech blog
Yilin Wang et al., "[YouTube UGC Dataset for Video Compression Research]()", MMSP 2019
Joong Yim et al., "[Subjective Quality Assessment for YouTube UGC Dataset]()", ICIP 2020
Yilin Wang et al., "[Rich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

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)

Perceptual Quality Assessment Aspects

	Low quality		High quality	
Video Content		MOS=2.052 (no meaningful content)		MOS=4.457 (intense games)
Distortions (Introduced during video production phase)		MOS=1.242 (heavy blur)		MOS=4.522 (high contrast, sharp edges)
Video Compression (introduced by compression or transmission)		MOS=2.372 (heavy compression)		MOS=4.646 (no compression artifacts)

UGC Video Quality Human Evaluation

Real-time strategy game
(interesting content)

Blurred texture

Heavily compressed text

Conclusion
Medium low quality
(MOS=2.761)

Explanation
Poor text and texture quality lead to bad game watching experience

YouTube

How do we scale UGC evaluation?

Auto-evaluation from multiple aspects:

- Content
- Distortion
- Compression

Report quality beyond a single score - folding in multiple high level interpretable indicators

Requirements for UGC metric

Comprehensively map to human evaluations accurately, folding in all the nuances of UGC

Target UGC centric no-reference, while still perform reliably with reference

Introducing YouVQ - a VQ metric for UGC

YouTube

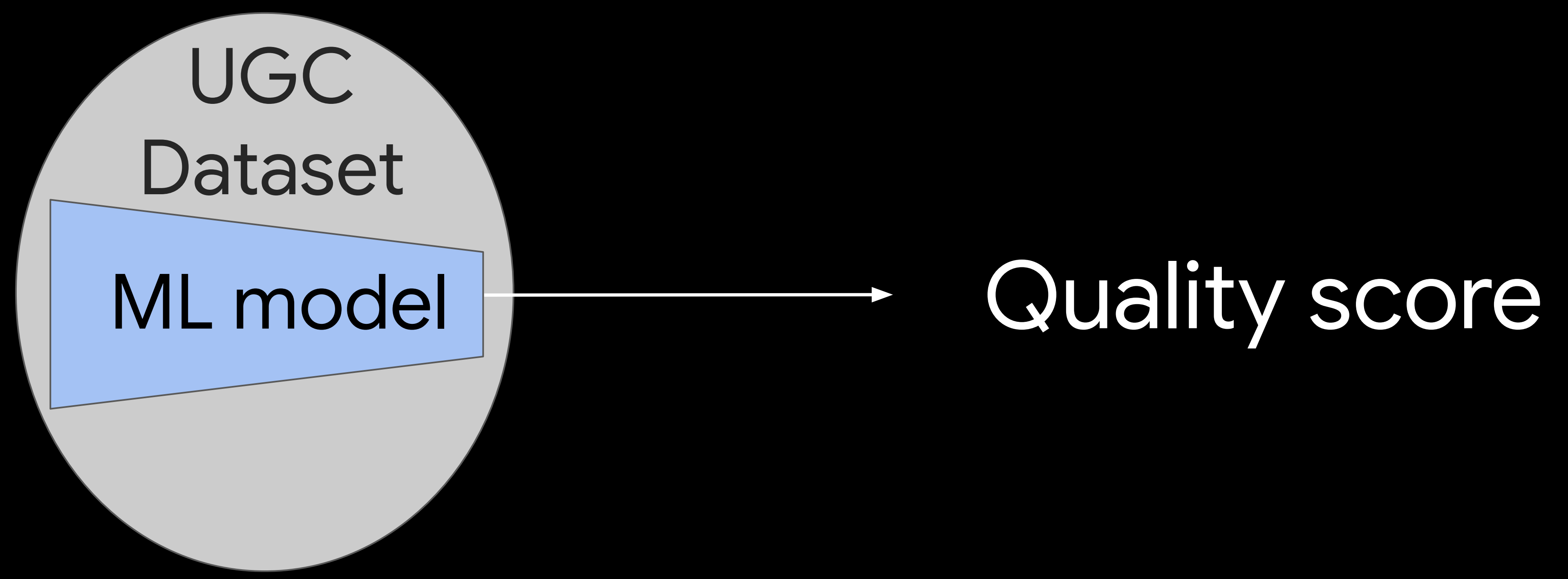
YouVQ Framework

UGC Video Quality Assessment (UGC-VQA)

- Existing handcrafted feature approaches (SSIM, VMAF, etc)
  - Difficult and time-consuming
  - Insufficient feature set (summarized from limited samples)

- Current Machine Learning approaches
  - Automatic feature learning
  - Suitable for large scale UGC data

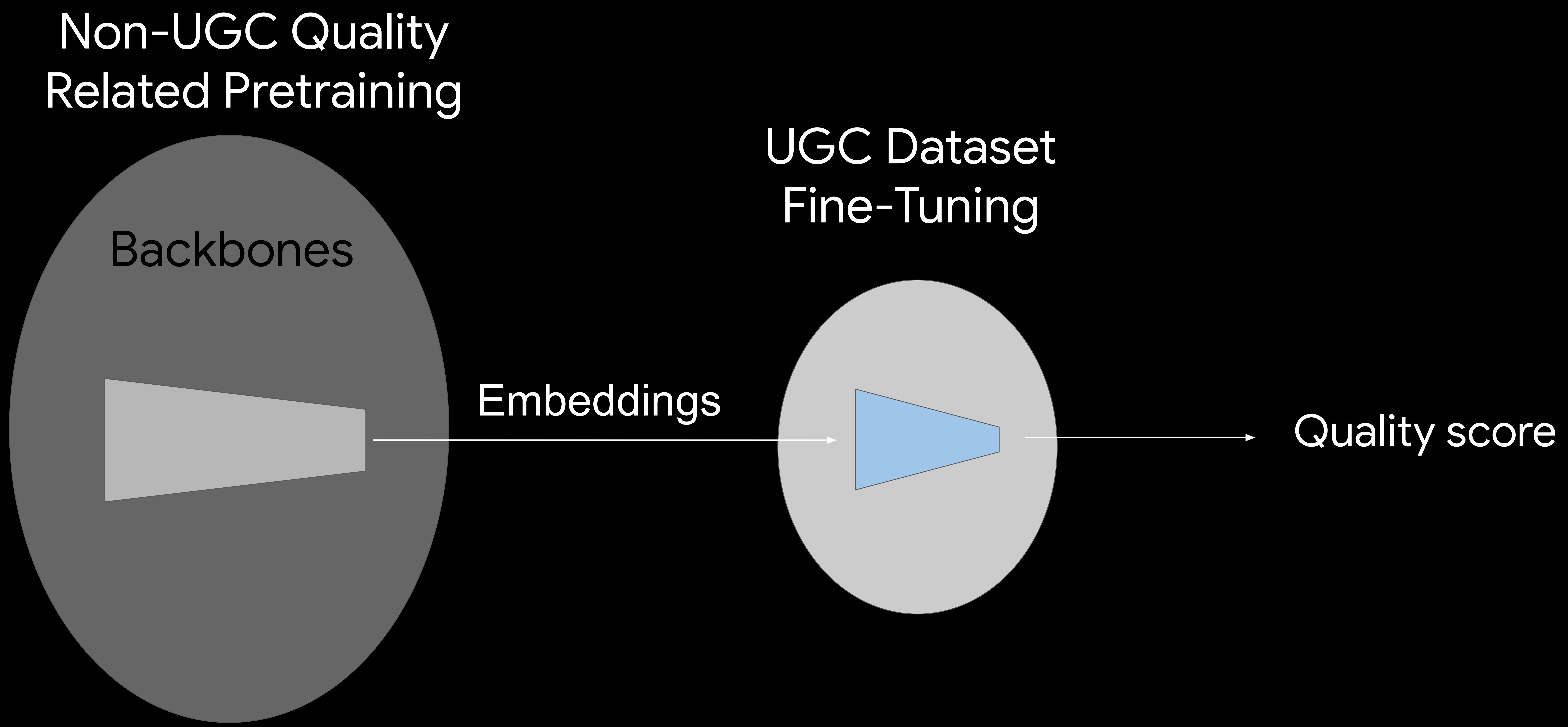
Direct training on UGC dataset:



Training data for UGC video quality assessment

- UGC datasets with quality labels
  - YT-UGC (1.5K), Patch-VQ (40K)
- Compare with non-quality datasets
  - Kinetics-600 (500K videos), YT8M (8M videos), ImageNet (14M images)
- Transfer Learning - preferred

Direct Transfer Learning

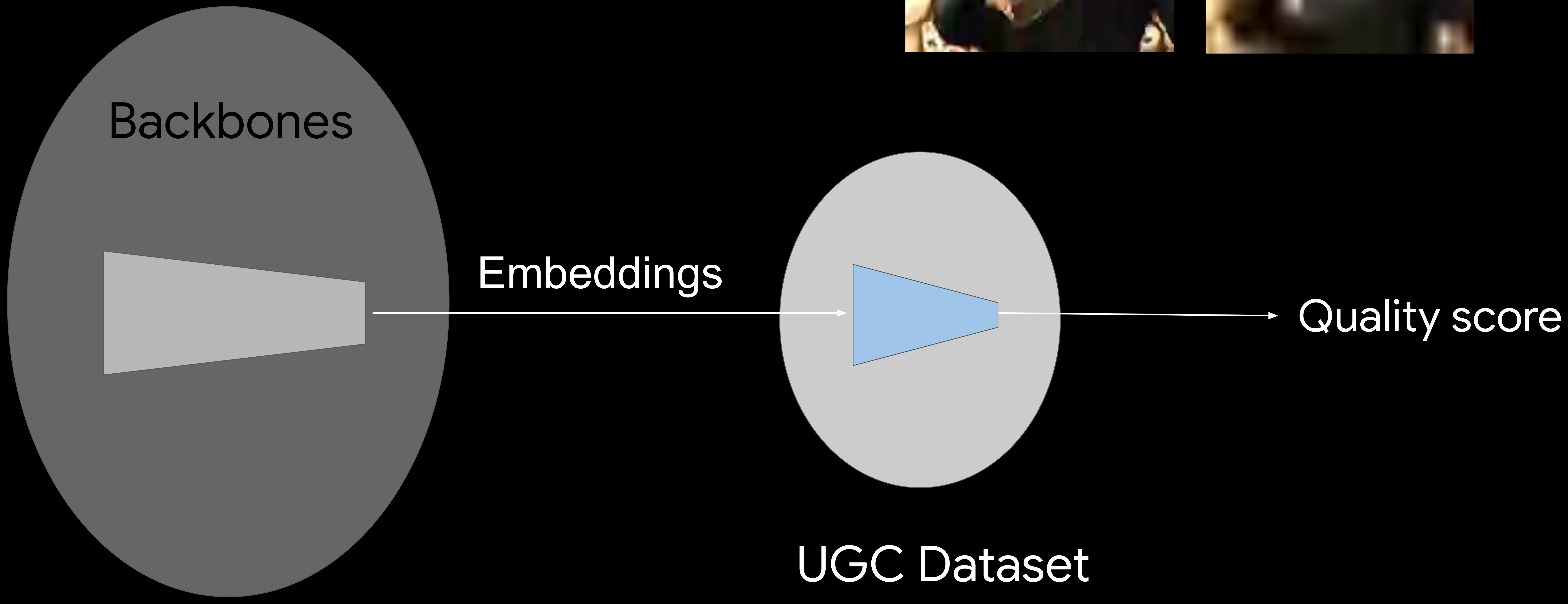


Direct Transfer Learning

For recognition: similar

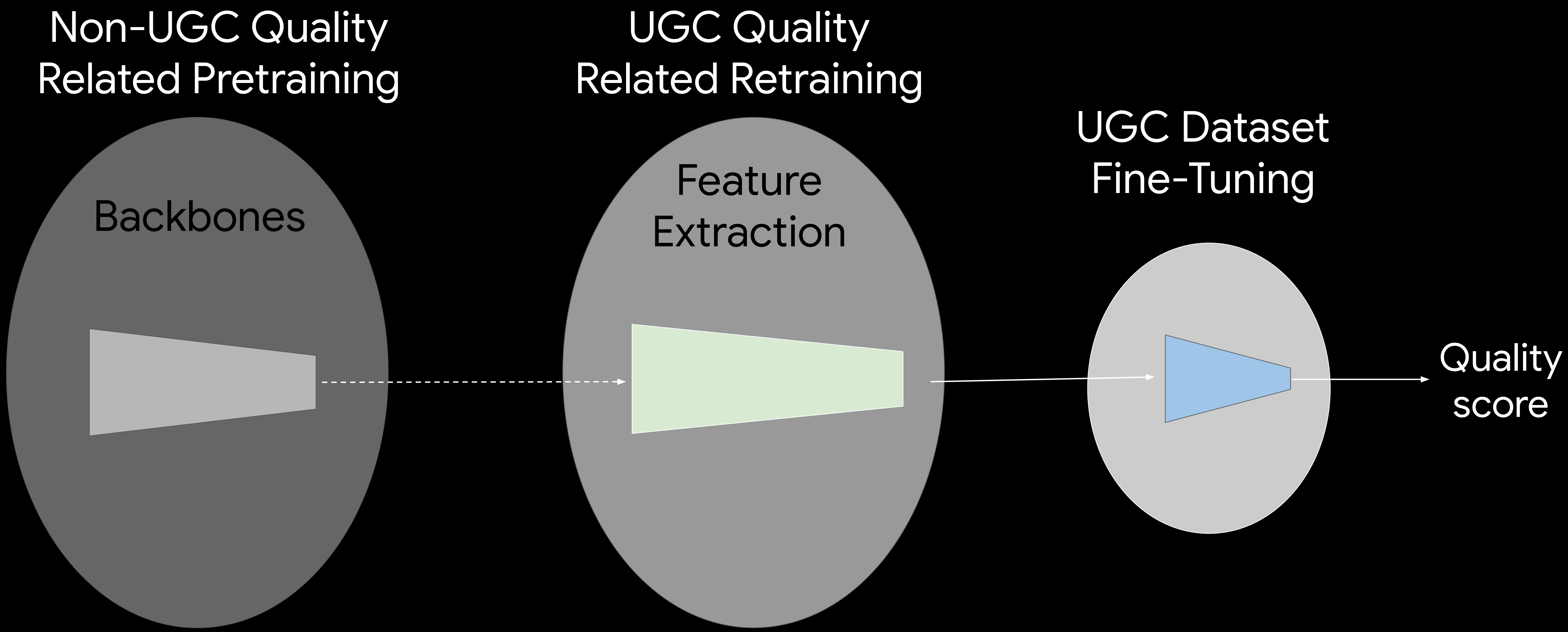
For video quality: very different

Non-UGC Quality Related Pretraining



UGC Dataset Fine-Tuning

Retraining on quality related data



Effectiveness of UGC quality related retraining

Evaluated on YT-UGC MOS

Backbone (EfficientNet-b0)	*PLCC*	*SRCC*
Raw (ImageNet, frozen weights)	0.624	0.612
Raw (ImageNet, trainable weights)	0.671	0.690
Retrained (KADIS-700K, frozen weights)	**0.732**	**0.735**
Retrained (KADIS-700K, trainable weights)	**0.732**	**0.738**

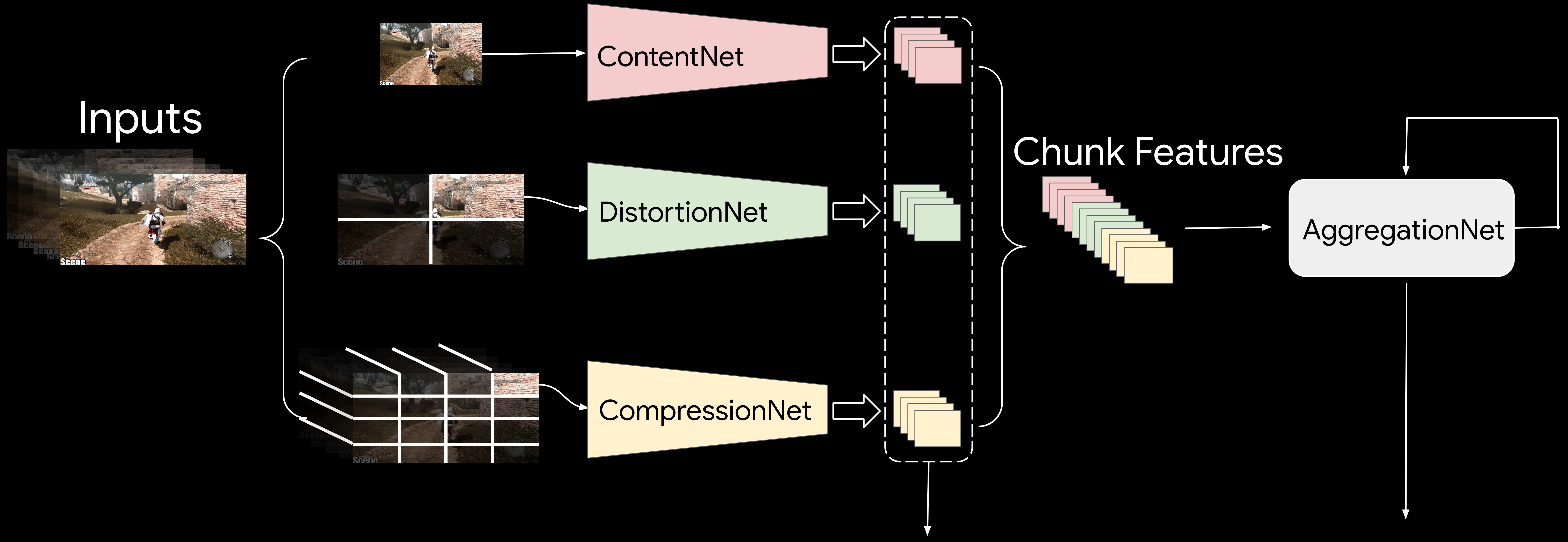
Direct transfer learning

With quality related retraining

PLCC, SRCC: correlation coefficients in [0, 1], the higher the better.

YouTube

YouVQ: YouTube Video Quality Assessment Framework



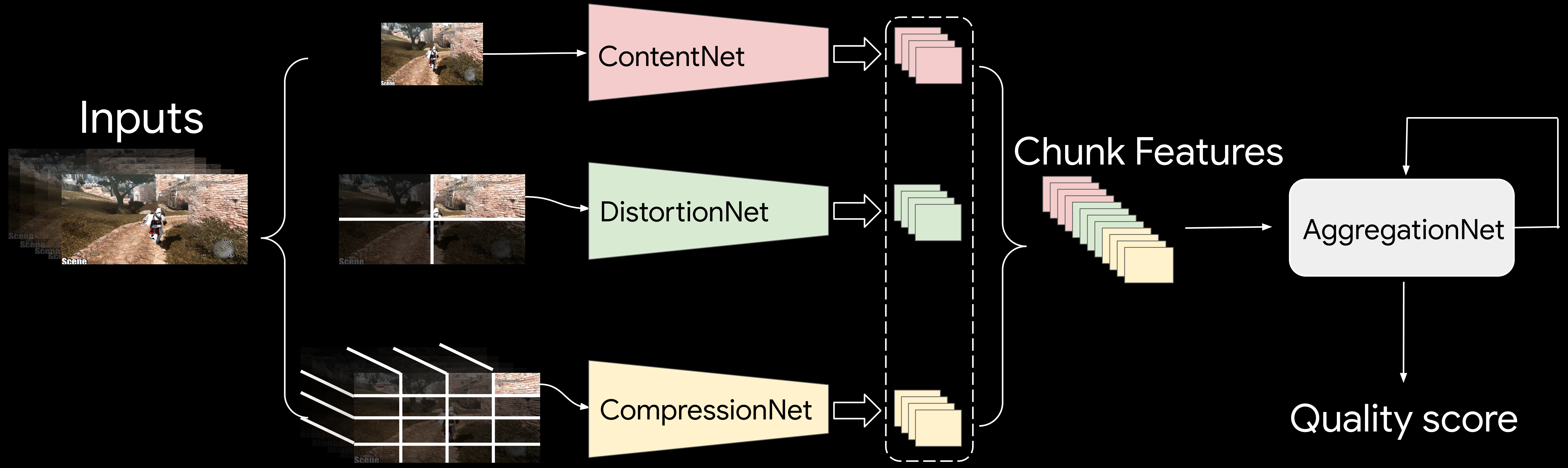
Outputs: Video Quality Indicators

- content labels
- distortion types
- compression level

\+ Quality score

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

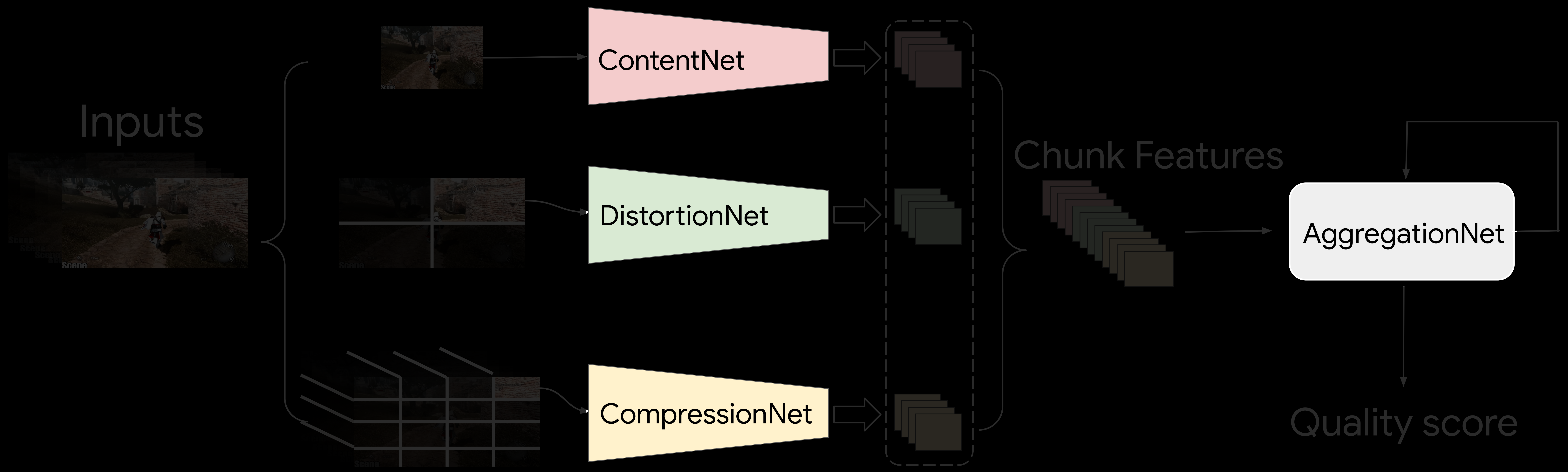
YouVQ: YouTube Video Quality Assessment Framework



Benefits of YouVQ framework:

- Self-supervised learning on raw UGC videos, no longer restricted by labeled MOS.
- Complementary features learned from different quality related aspects.
- Works on native resolutions, and sensitive to local details.

YouVQ: YouTube Video Quality Assessment Framework

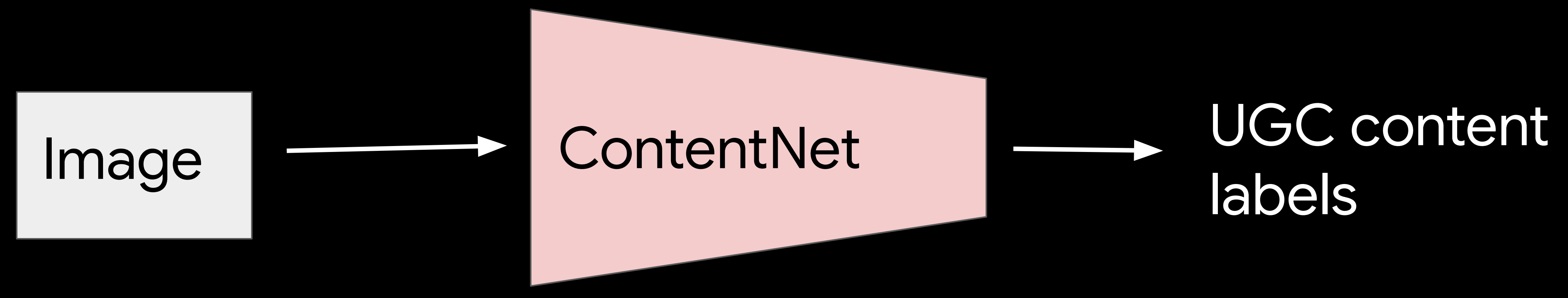


Benefits of YouVQ framework:

- Self-supervised learning on raw UGC videos, no longer restricted by labeled MOS.
- Complementary features learned from different quality related aspects.
- Works on native resolutions, and sensitive to local details.

YouVQ Features: ContentNet (CT)

- Multi-label classification
- Model trained on 100k YT8M videos
  - Inputs: single image
  - Outputs: 3862 UGC content labels
  - Loss: cross-entropy
- Backbone: EfficientNet-b0 (pre-trained on ImageNet)



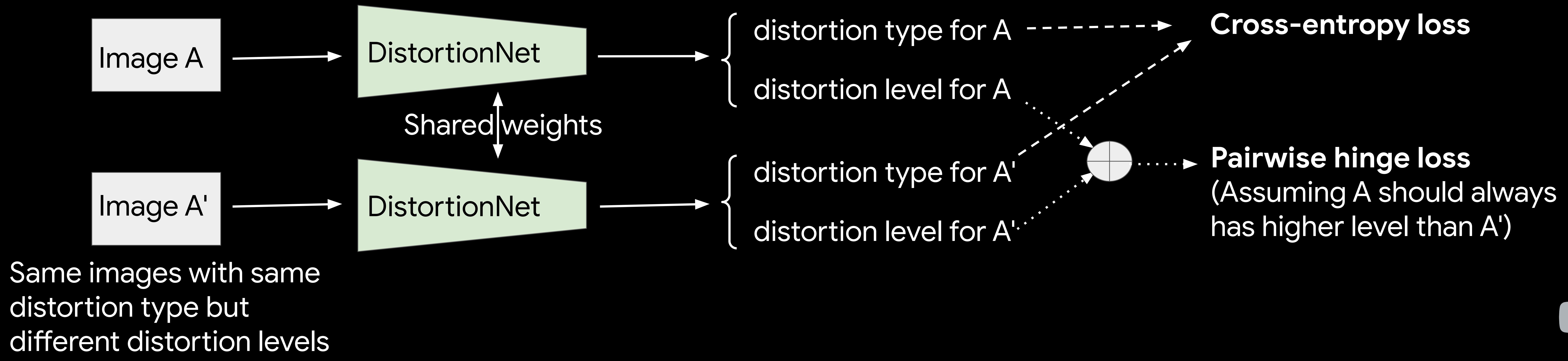
Backbone model	*#Params*	*#FLOPS*	*YT8M Classification Accuracy*		
			Top-1	Top-5	Top-10
ResNet-50	23.5M	3.8B	0.325	0.554	0.659
EfficientNet-b0	5.3M	0.39B	0.463	0.721	0.792
EfficientNet-b7	66M	37B	0.460	0.723	0.788

Correlation on YT-UGC quality scores	*MOS*		*DMOS*	
	b0	b7	b0	b7
Content features only	0.628	0.615	0.584	0.357
Content+Compression	0.787	0.774	0.672	0.652
Content+Distortion	0.750	0.752	0.390	0.334
All three features	0.802	0.796	0.539	0.497

No gain when using EfficientNet-b7 feature for quality assessment.

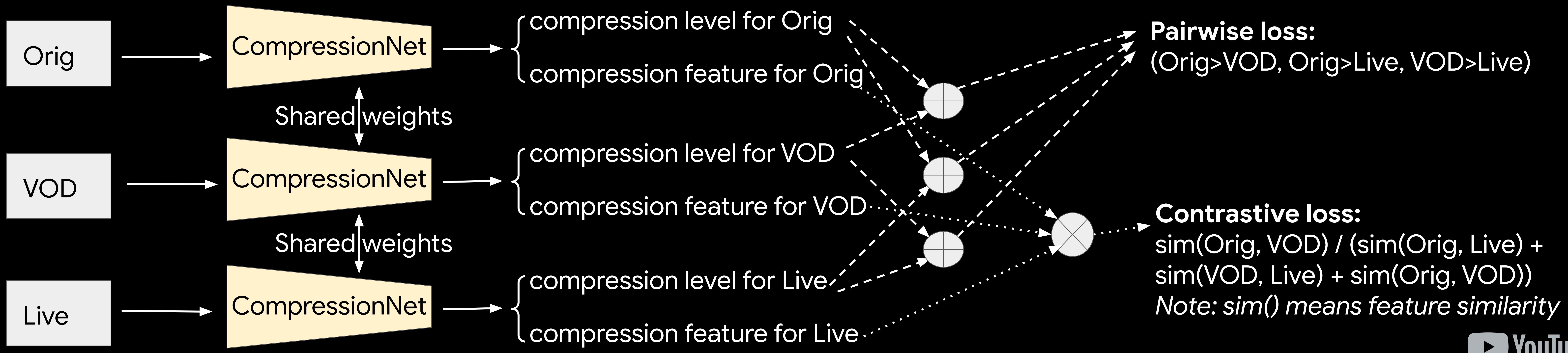
YouVQ Features: DistortionNet (DT)

- Synthetic distortions
  - 23 types, e.g. Gaussian noise and motion blur
  - distorted variants in 5 levels per type
- Model trained on KADIS-700K images
  - Inputs: two images with the same distortion type
  - Outputs: distortion type and level
  - Loss: cross-entropy + pairwise hinge loss
- Backbone: EfficientNet-b0 (pre-trained on ImageNet)



YouVQ Features: CompressionNet (CP)

- Self-supervised learning
- Compressing original videos with recommended VP9 settings for VOD and Live
- Model trained on YT8M 1080p videos
  - Inputs: original and its VOD and Live versions
  - Outputs: compression level in [0, 1] + compression feature (last layer outputs)
  - Loss: pairwise loss + contrastive loss
- Backbone: D3D (pre-trained on Kinetics-600)



YouVQ feature aggregation

- AggregationNet
  - Training with YouVQ features on YT-UGC original MOS
  - Three candidate aggregation models
    - AvgPool, LSTM, ConvLSTM
  - AvgPool performs best
    - most UGC videos have relatively consistent quality among frames

Feature	***AvgPool***			***LSTM***			***ConvLSTM***		
	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
CP+CT+DT	0.802	0.816	0.382	0.767	0.771	0.411	0.760	0.764	0.418

Correlations on YT-UGC MOS

YouVQ Performance

Accuracy

Correlations with YT-UGC MOS

YouVQ Features	***PLCC***	***SRCC***
CP (Compression)	0.770	0.785
CT (Content)	0.628	0.628
DT (Distortion)	0.726	0.744
CP+CT	0.787	0.801
CP+DT	0.790	0.802
CT+DT	0.750	0.767
CP+CT+DT	**0.802**	**0.816**

Increasing

Generalizability on MOS Prediction for UGC

Model fine-tuned on YT-UGC MOS		Directly predicting on KoNViD-1k MOS
PLCC (YouVQ)	0.802	0.670
PLCC (best of other evaluated metrics)	0.761 (from VSFA)	0.602 (from VSFA)
Metrics compared	BRISQUE, NIQE, VIIDEO, TLVQM(SVR), TLVQM(RFR), VSFA	TLVQM(SVR), TLVQM(RFR), VSFA

Generalizability on DMOS Prediction for UGC

Evaluated on YT-UGC DMOS (not re-trained)

- 189 originals + three VP9 variants

Pred DMOS = YouVQ(ref) - YouVQ(target)

- Sensitive to compression
- Good correlations without retraining

Metric	*PLCC*
PSNR	0.402
SSIM	0.493
VMAF	0.401
LPIPS	0.524
TLVQM	0.276
VSFA	0.403
YouVQ	**0.660**

Comprehensive Quality Indicators

ContentNet

- top-10 label accuracy
  - 0.792 on YT8M
  - 0.53 on YT-UGC

Content labels: Car (0.58), Vehicle (0.42), Sports Car (0.32), Motorsports (0.18), Racing (0.11)

DistortionNet

- evaluated on KADID-10K
- distortion classification accuracy 0.97

Distortion types: Jitter (0.112), Color quantization (0.111), Lens blur (0.108), Denoise (0.107)

CompressionNet

- self-supervised learning
- high accuracy on predicting pairwise order of compression level

Compression level: 0.892 (high)

YouTube

Locating Local Quality Issues

YouVQ provides patchwise quality assessment

Patch at time t = 1
compression level = 0.000

Patch at time t = 2
compression level = 0.904

How YouVQ works in practice

YouVQ diagnosis report

From ContentNet (CT)

Video game, Strategy video game, World of Warcraft, etc

From DistortionNet (DT)

Multiplicative noise, Gaussian blur, Color saturation, Pixelate, etc

From CompressionNet (CP)

0.559 (medium high compression)

Predicted quality score in [1, 5]

(CP, CT, DT): (3.151, 3.901, 3.216)

(CP+CT+DT): 3.149 (medium low quality)

Summary

We introduced YouVQ for UGC video quality assessment

- It is a comprehensive framework to analyze UGC video quality and makes the VQ score more interpretable
- Maps very well to ground truth human evaluations
- Performs consistently and reliably for no-reference, works equally well when reference is present (pristine or otherwise)

Videos and subjective data are available on media.withyoutube.com

Thank you!