CLIC 2022

Learned Image Compression and Perceptual Metric Challenge

Presented by Luca Versari,
Ross Cutler
and Nick Johnston

5th Workshop and Challenge on Learned Image Compression
New Orleans, LA

Welcome to CLIC 2022

- First Hybrid CLIC workshop
- First time in person since CVPR 2019 (Long Beach)
- Virtual through Zoom
  - https://www.eventscribe.net/2022/2022CVPR/ ----------------------->
  - Event will be recorded and available later this week
- 8:30 AM start, 5:45 PM end of poster session

Outline

- What is CLIC?
- Program
- Challenges / Tasks
  - Multiple Bitrate Image Compression Challenge
  - Video Compression Challenge
  - Perceptual Quality
- Future of CLIC

What is CLIC?

- **C**hallenge on **L**earned **I**mage **C**ompression (and beyond) and a **CVPR Workshop**
- It was started in 2018 by a team of researchers from ETH Zurich, Twitter and Google. Now organizers from Microsoft, Apple, Interdigital and Netflix have also joined the board!
- Our 2022 goals:
  - Define a benchmark and incentivise the development of learning-based compression methods for images and video (new since 2020)
  - Perceptual evaluation for images
  - Incentivize research in **learned compression of any kind,** and encourage development of new **perceptual quality metrics**

Organizers & Sponsors

CLIC

Organizers & Sponsors

CLIC

Google

CVL Computer Vision Lab

ETH zürich

NETFLIX

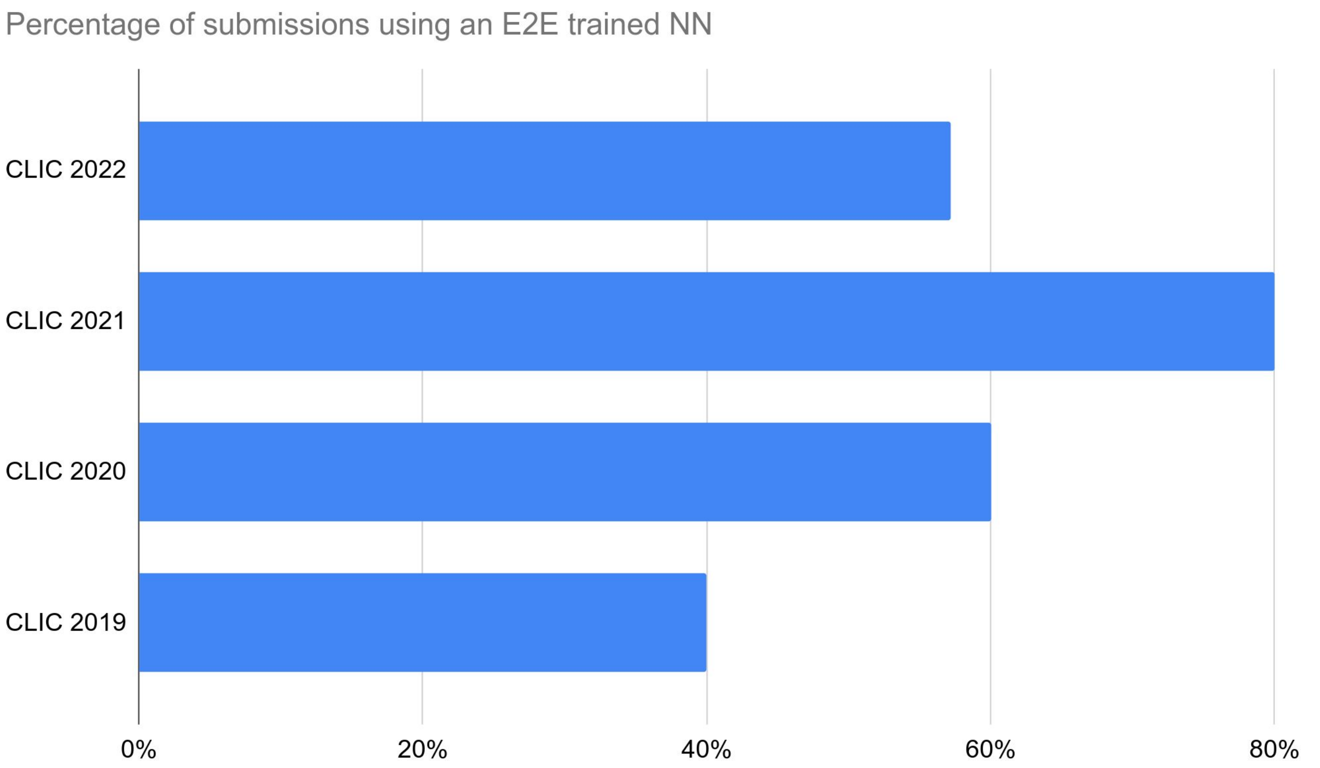
interdigital™

The Competition Tracks

- Multiple Bitrate Image Compression
  - Target an average of **0.075 bpp**, **0.15 bpp**, and **0.3 bpp**! (Started this three rate track last year).
  - High quality images from Unsplash.
- Video Compression
  - Target a fixed size (will get into detail later)
  - **0.1** mbps and **1.0** mbps
  - 30 10-second video from Pexels
- Perceptual Quality Evaluation
  - Request participants to submit metrics that are evaluated against the human ratings from the Multiple Bitrate Image Compression track
- Note:

  Full description & statistics are available at http://compression.cc/

Submission Trend



Challenge Format

- Development phase:
  - We release a new partitioned dataset (potentially to be used for training)
  - Participants develop new methods
  - Participants submit decoder (model + other tools) to evaluation server
  - The server evaluates the model, and updates the leaderboard
- Test Phase
  - Participants can no longer update the models/binaries
  - One week after the development phase ends we release the previously unseen test set
  - Participants upload compressed files, which we decompress with their previously submitted decoder
- Evaluation Phase
  - Human evaluation
  - Results released at this workshop

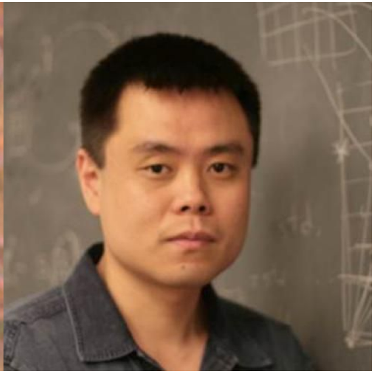
Workshop Program

Workshop Program

- Invited Speakers







Guo Lu

Beijing Institute of Technology

Debargha Mukherjee

Google

Auke Wiggers

Qualcomm

Tsachy Weissman

Stanford University

Zhou Wang

University of Waterloo

Workshop Program

- Talks by the winners of:
  - Image Compression Challenge
  - Video Compression Challenge
  - Perceptual Metric Challenge
- Panel Discussion
- Awards Ceremony
- Poster Session

Overview of the Day

Schedule (Preliminary)

Time (conference local time)	Talk/Activity	Speaker
08:30	Opening remarks	Nick Johnston (Google)
08:45	Invited speaker	Guo Lu (Beijing Institute of Technology)
09:15	Invited speaker	Debargha Mukherjee (Google, LLC)
09:45	Short break	
10:00	Dataset, Challenge, Rating Task	Luca Versari (Google) and Ross Cutler (Microsoft)
10:45	Invited speaker	Auke Wiggers (Qualcomm)
11:15	Image Track, 3rd place	
11:30	Image Track, 2nd place	
11:45	Image Track, 1st place	
12:00	Lunch break	
13:00	Invited speaker	Tsachy (Itschak) Weissman (Stanford University)
13:30	Video Track, 3rd place	
13:45	Video Track, 2nd place	
14:00	Video Track, 1st place	
14:15	Invited speaker	Zhou Wang (University of Waterloo)
14:45	Perceptual Quality Track, 3rd place	
15:00	Perceptual Quality Track, 2nd place	
15:15	Perceptual Quality Track, 1st place	
15:30	Short break	
15:45	Panel discussion	
16:45	Awards ceremony	
16:50	Poster session	
17:45	End of the workshop	

All times are local to conference venue.

Also on compression.cc

The Multiple Bitrate Image Compression Challenge

Why human evaluation?

Multiple Bitrate Image Compression: Human Evaluation

- Goal:
  - Use all images in the test set for the human evaluation (test set released after participants froze their models/code)
- Challenge:
  - Too many competitors, too many images, not enough rater time available to do all pairwise rating
- Solution:
  - Use the Pre-Selection Method from 2020 but include all participants and all images.

Designing the Test Set

- Fairness
  - Skin tone reproduction needs to be accounted for, so diversity is a must
  - Various scenery types need to be represented
  - Contents needs to be suitable for evaluation of compression methods
- Difficulty
  - How to find such a varied test?
  - How can we minimize human bias in this selection process?

Test Set Selection

- Addressing Fairness
    - We used unsplash as the source of images
    - Unsplash provides royalty free images, and allows searching by tags
    - We searched for location across all continents (i.e., for each continent we selected the same number of countries, and searched for their name)
    - From each search result, we took a random sample of images
- What we cannot account for:
    - Unsplash does have a photographer's bias in choice of subject
    - Many photographers like to photograph people, so many images in the test set have people
    - High end processing of photographs is most likely happening in the top results
    - Source is already compressed material. We downsample by a factor of 2 to compensate.

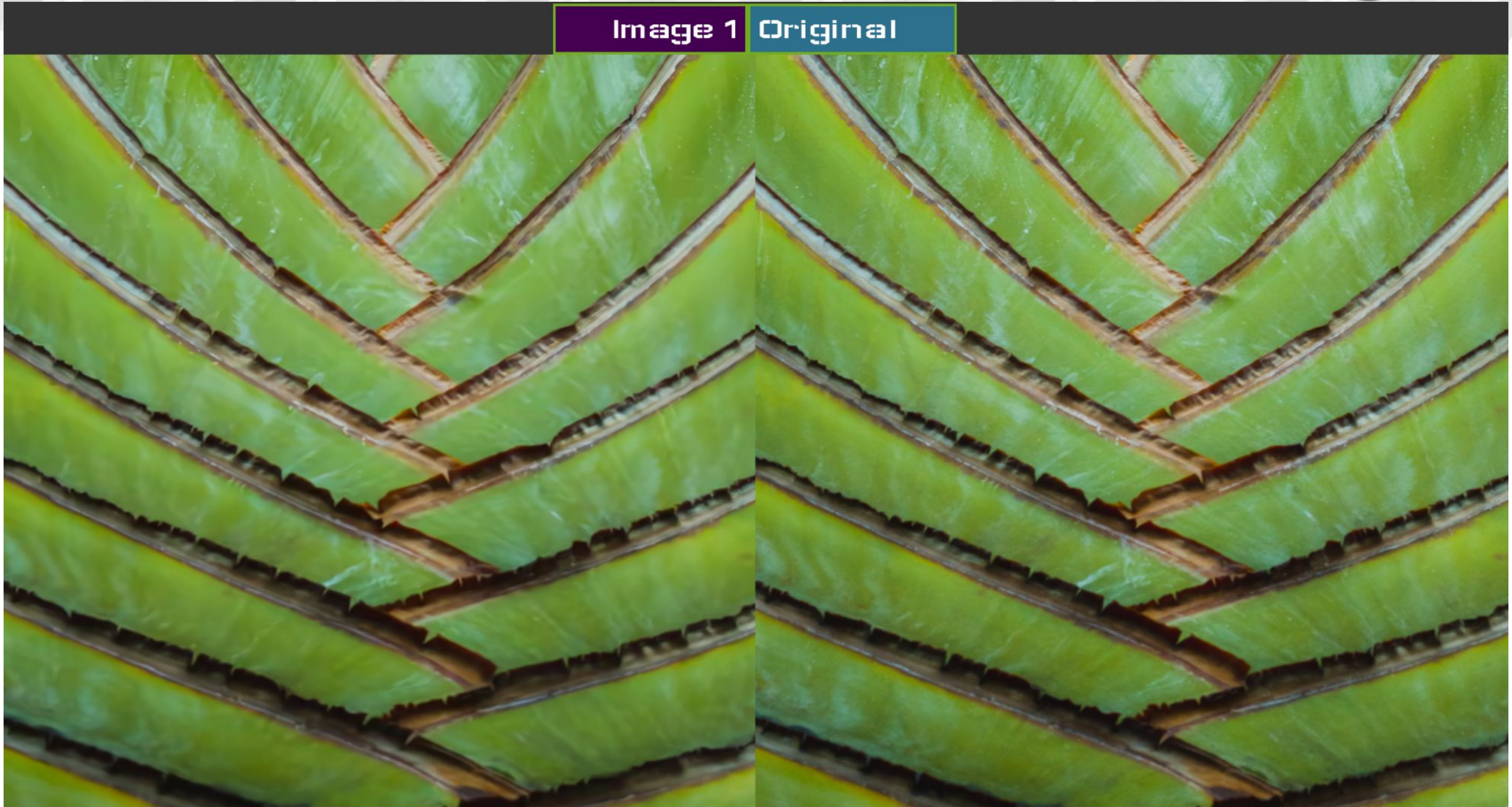
Test Set Selection

- Usefulness in compression evaluation
  - Fairness was addressed, and we believe we have one of the most diverse sets of images available for compression research
  - The set contains a wide range of processing styles for photographs, which should stress test methods which tend to enhance “normal” images to make them pop
- Possible Negative
  - Due to trying to avoid biases in these images, we don’t necessarily have “canonical” test images. No effort has been made to find such images

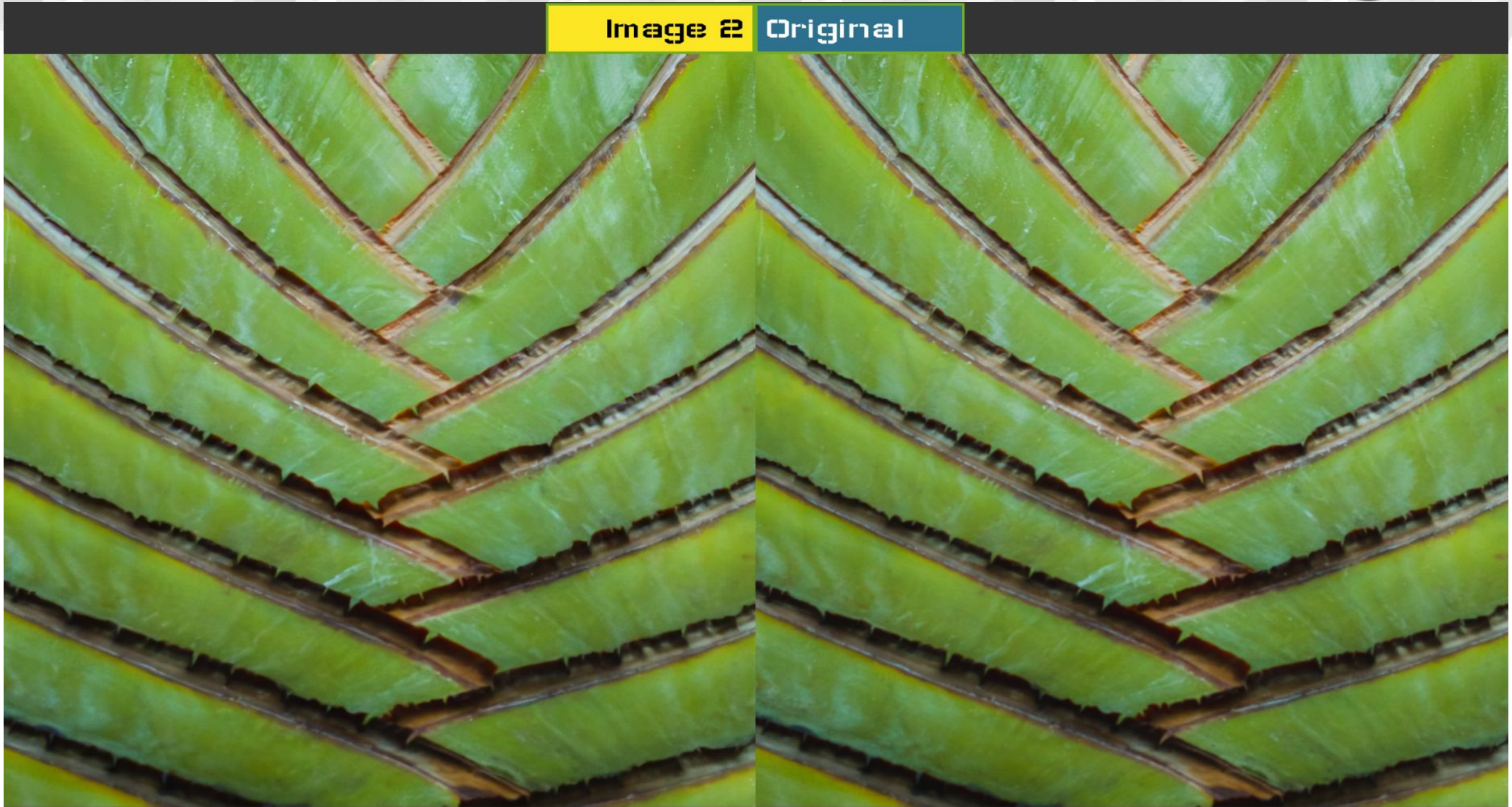
How to use an image?

- Proposed idea:
  - Make raters choose a crop (768x768)
- Why crops?
  - Makes the rating task much more focused (fewer opportunities to have a more diverse set of artifacts that need to be disambiguated, and figured out which is more important)
- Why let raters *choose* which crop?
  - Choosing a random crop may yield completely uninformative regions of the image
  - Raters were able to choose “next crop” which would choose another random crop (and repeated this until a reasonable crop was found)

Rater interface



Rater interface



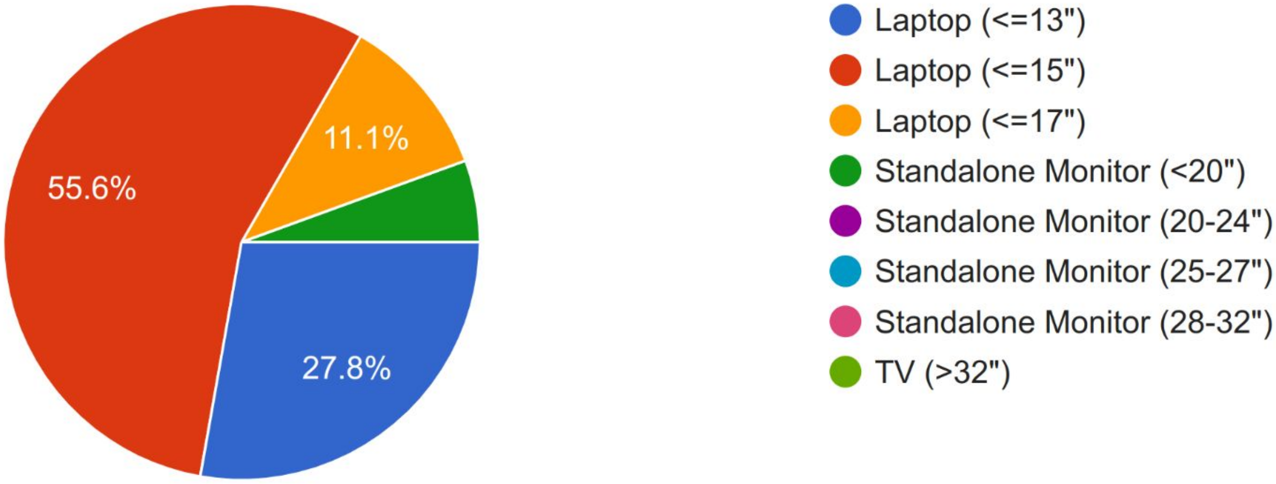
How to assign a score to each method?

- We employed the same methodology as CLIC 2020
- Multiple methods evaluated (each comparison is treated as a 2-player game):
  - **Monte Carlo Elo Ranking** (Developed for CLIC 2019)
  - New this year: single ranking for all bitrates
- Evaluating 3 bitrates means 3 Elo scores. How to get the global rank?
  - We used the harmonic mean of the Elo ranks (not scores) across all three bitrates

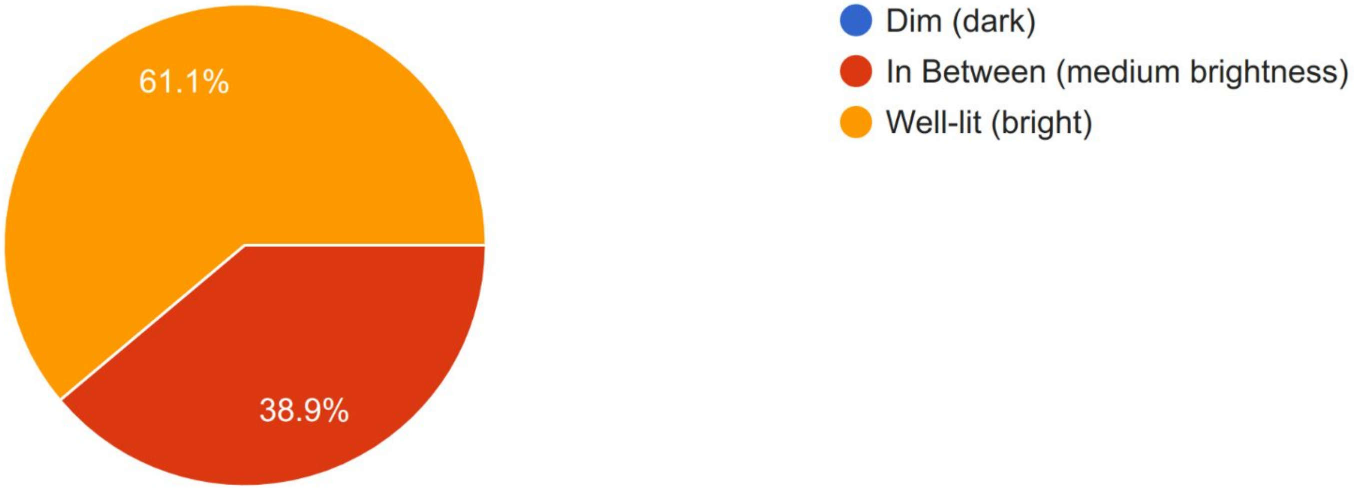
Data Quality

- We split the ratings into per-rater *sessions*
  - A 15+ minute break starts a new session
- We generated *gold questions* (10% of questions) which ask to compare A to B, with the original being identical to A.
- We excluded answers from sessions with <80% accuracy on gold questions.
- The rating UI forces the rater to
  - Spend at least 1 second before submitting an answer
  - Switch between images at least 3 times (A->B, B->A, A->B)

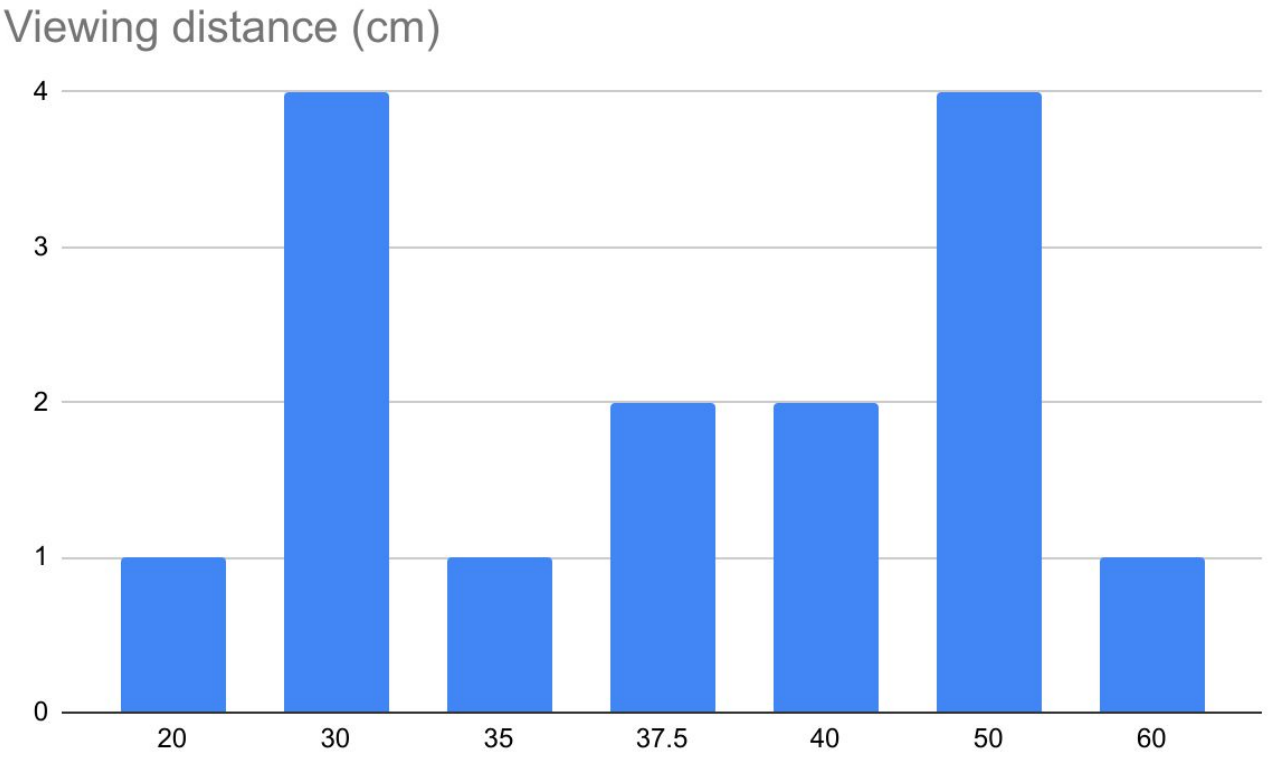
Rater Survey - Monitor size



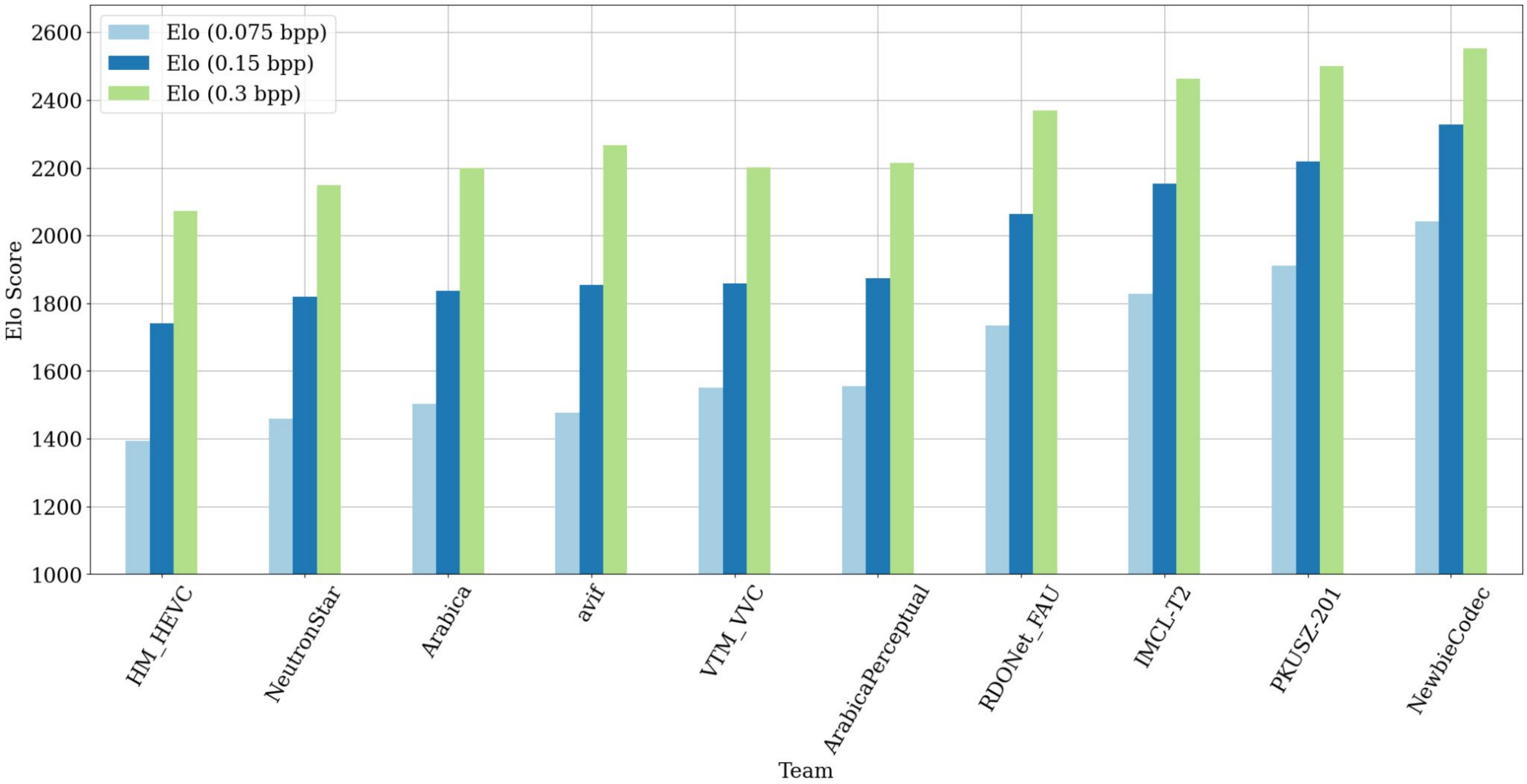
Rater Survey - Lighting environment



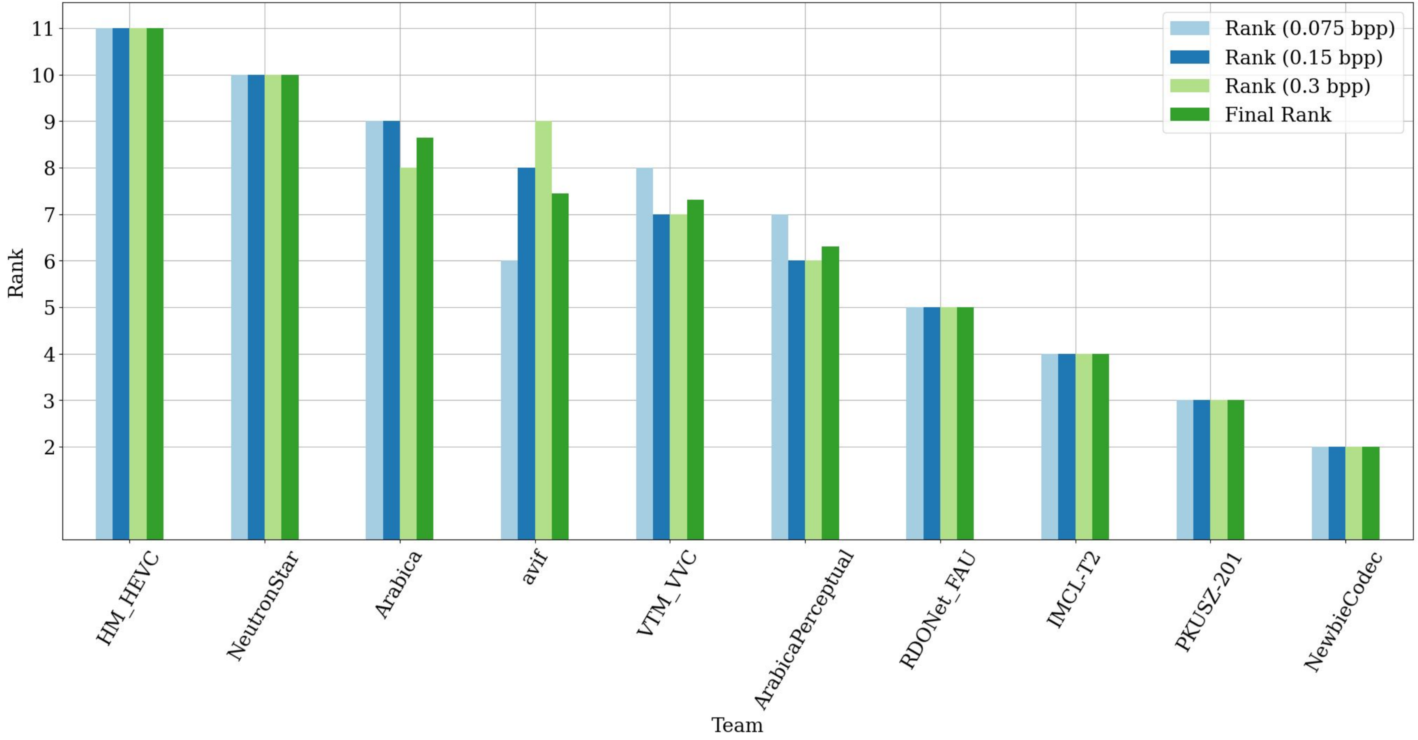
Rater Survey - viewing distances



Elo Scores (Higher=Better)



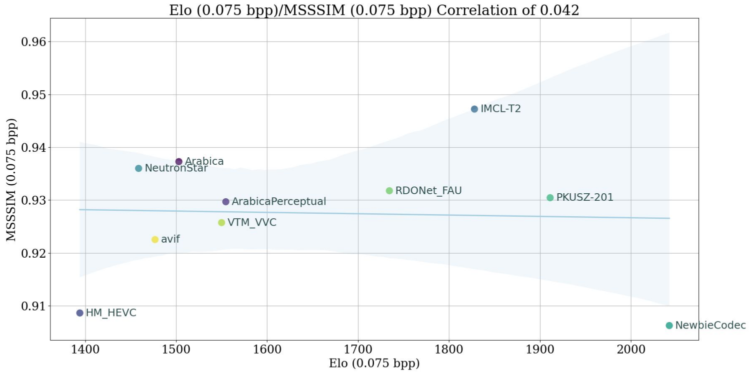
Final Rank = Harmonic Mean of Ranks
(Lowest Rank Wins)

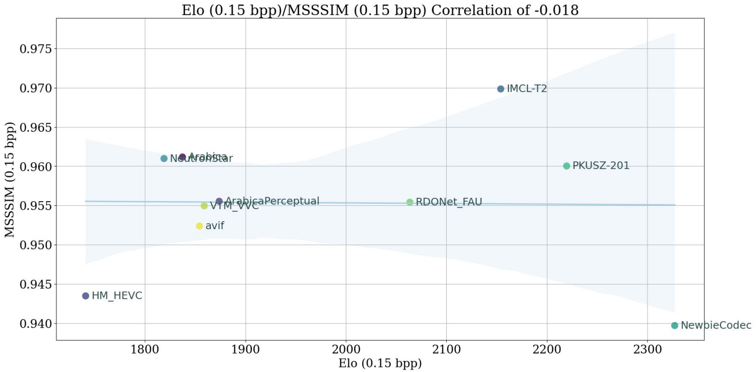


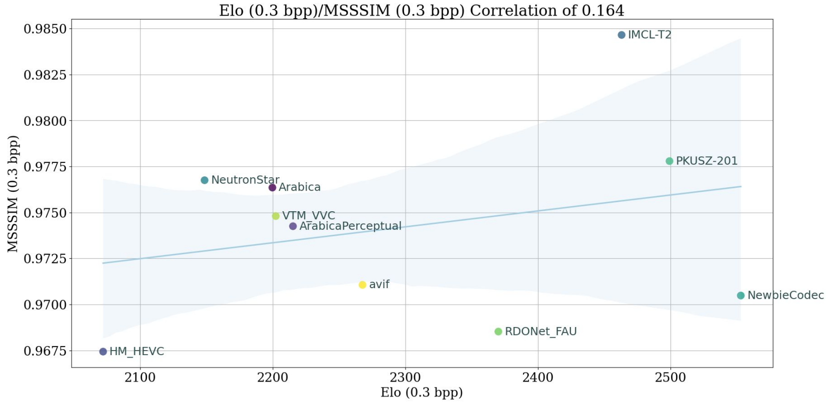
Originals have a rank of 1.

MS-SSIM vs. Elo Score

There should be a positive correlation

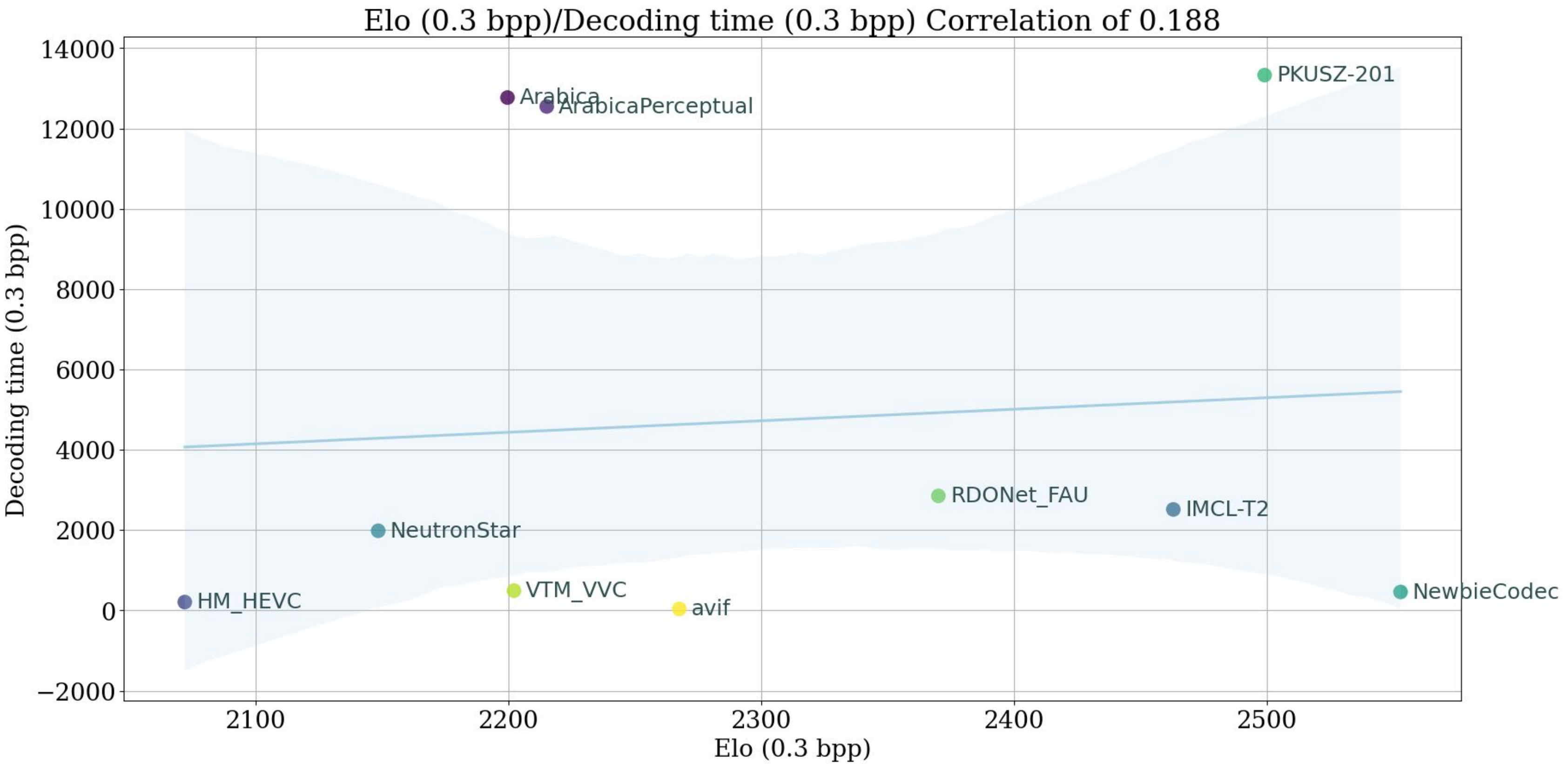






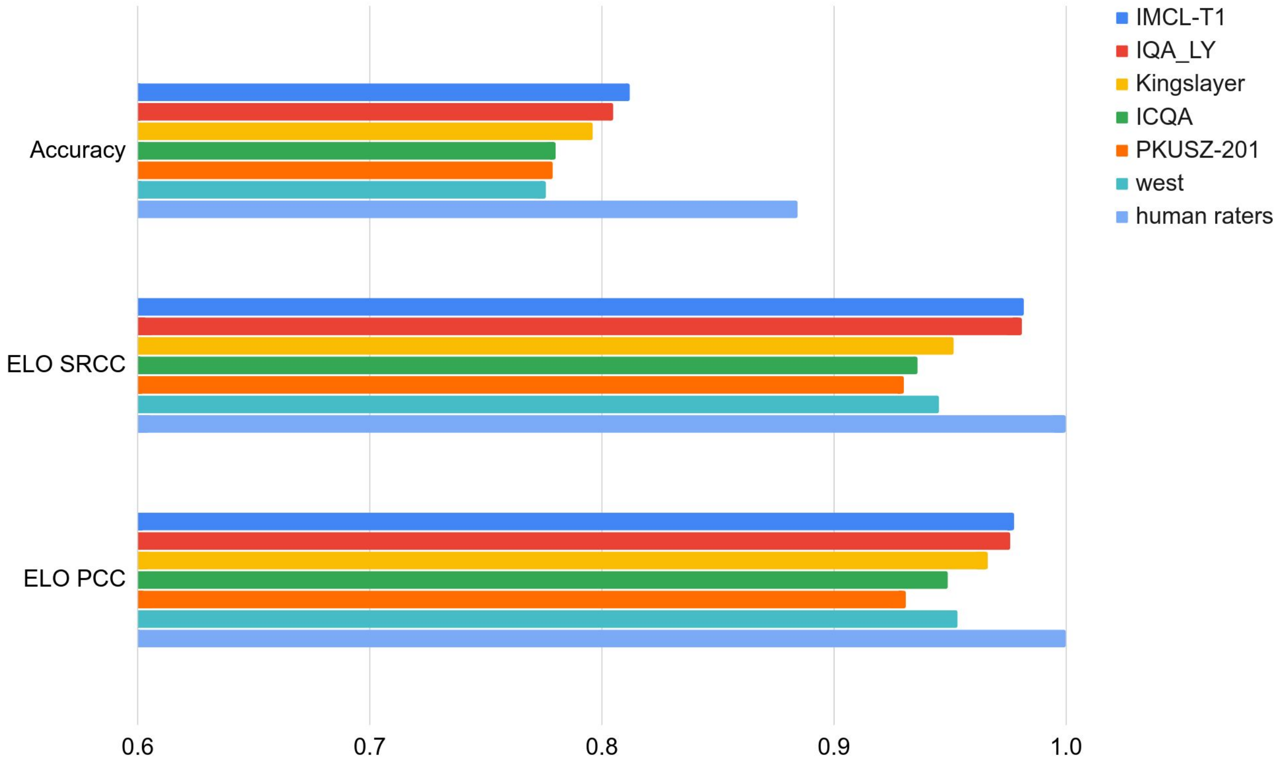
Runtime vs. Elo Score @ 0.3bpp

We expect a positive correlation



Perceptual Metric Evaluation

Accuracy and Correlation



CVPR CLIC 2022 Video Track Video Quality Assessment

Ross Cutler
Microsoft Corp.

Introduction

- New crowdsourcing platform for VQA
- Validation of the platform
- Results of CLIC video compression track

Video quality assessment

- Lab studies (e.g., ITU-T P.910) are the gold standard, but they are expensive, slow, not practical in a pandemic
- Crowdsourcing
  - Unknown participants
  - Working at own environment
  - Using own devices
  - No moderator
- We introduce an open-source framework with participant eligibility tests, environment and setup tests and reliability checks

Related work

Tool	Measures	Rater qual.	Viewing cond.	HW	Network	Accur.	Repro.
QualityCrowd [15, 16]	ACR, DSCQS	N	N	N	N	Y	N
WESP [17]	ACR, ACR-HR, DCR, PC	N	N	N	N	N	N
avrateNG [19]	ACR	N	N	N	N	Y	N
Ours	ACR, ACR-HR, DCR	Y	Y	Y	Y	Y	Y

Table 1: Open-source crowdsourcing video quality assessment systems

Framework

- Multiple scripts to automate the process
- Test methods
  - Absolute Category Rating (ACR)
  - ACR – Hidden reference
  - Degradation Category Rating (DCR)
  - Comparison Category Rating (CCR)
- Scales
  - 5 and 9 point Likert scale
- Can be used with any crowdsourcing platform or dedicated panel

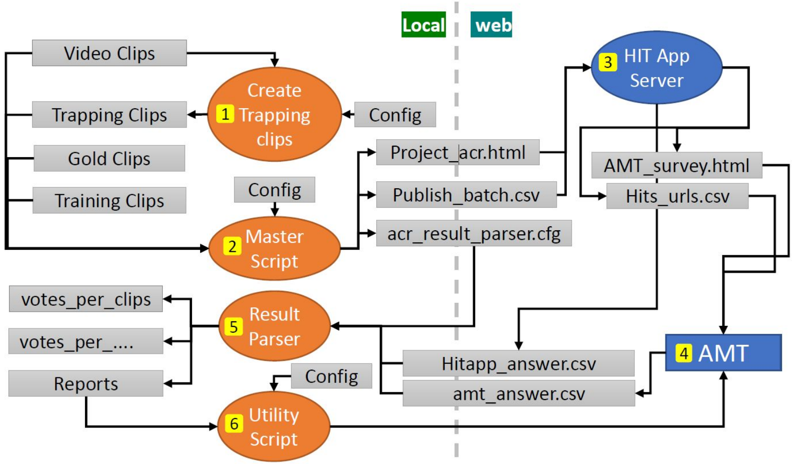
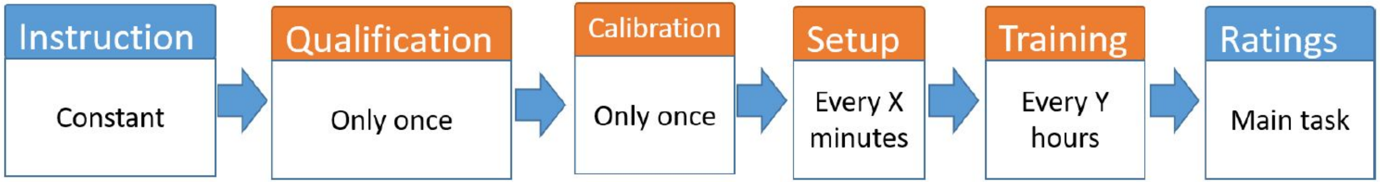


Fig. 1: Data Flow Diagram.

Test components



- The test is designed in different sections from participants perspective
- Rating sections: 10-12 clips to be rated
- Background hardware/network checks:
  - Resolution
  - Screen refresh rate
  - PC or Mobile
  - Network test

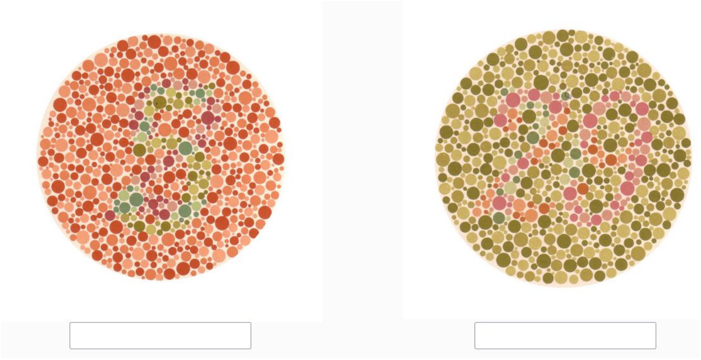
- Video playback component:
  - Full-screen (with/-out scaling)
  - Record playback duration
  - Force to watch until the end

Qualification

- Normal color vision Test
  - 2 plates from Ishihara test instead of 15

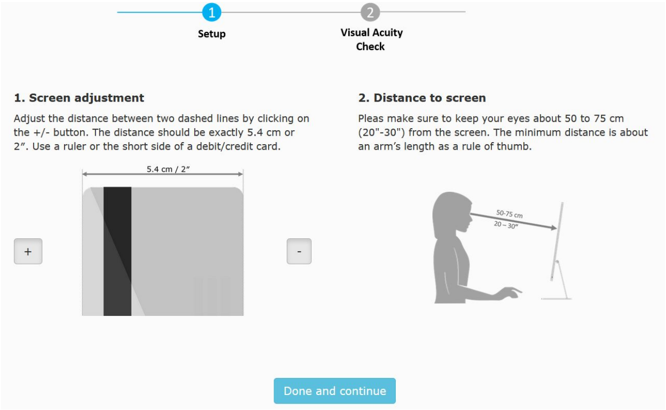
Pretest:

- 300 AMT and 191 from online color-blind communities
- Decision tree: 98% accuracy (sensitivity 0.996, specificity 0.95)



- Normal or corrected-to-normal Visual Acuity
- P.910: No error on the 20/30 line of a standard chart

- 5 Landolt ring optotypes



Setup I

- Ask to perform Resolution, Color and Brightness Calibration
  - For Windows/Mac devices raters are asked

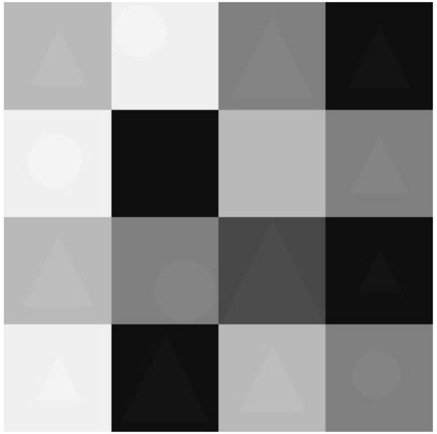
Q1. How many **circles** and **triangles** do you see in the image?

Circles:

Triangles:

Check my answer

You can try it up to **4** times.



4 circles
10
triangles

Setup II

- Viewing distance test
- 3 image pairs
- Blur effect, detected if
  - Too close
  - In proper distance
  - Even if too far
- Rater asked to adjust their distance if failed

3. Which image has a better quality compared to the other one? **Pictures may be blurry.**

Image A

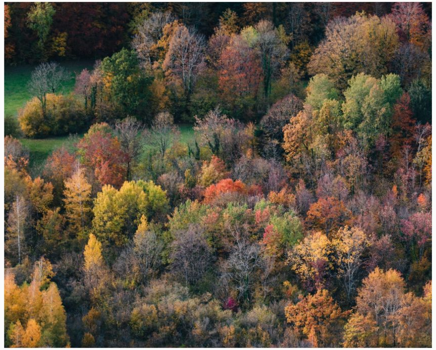
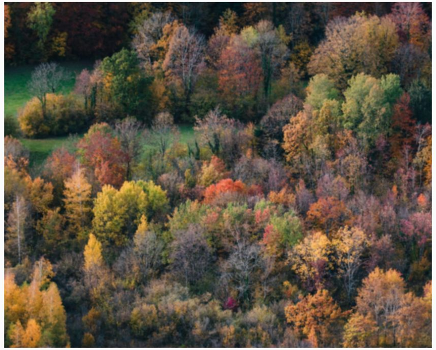


Image B



- Quality of **Image A** is better.
- Difference is **not detectable**.
- Quality of **Image B** is better.

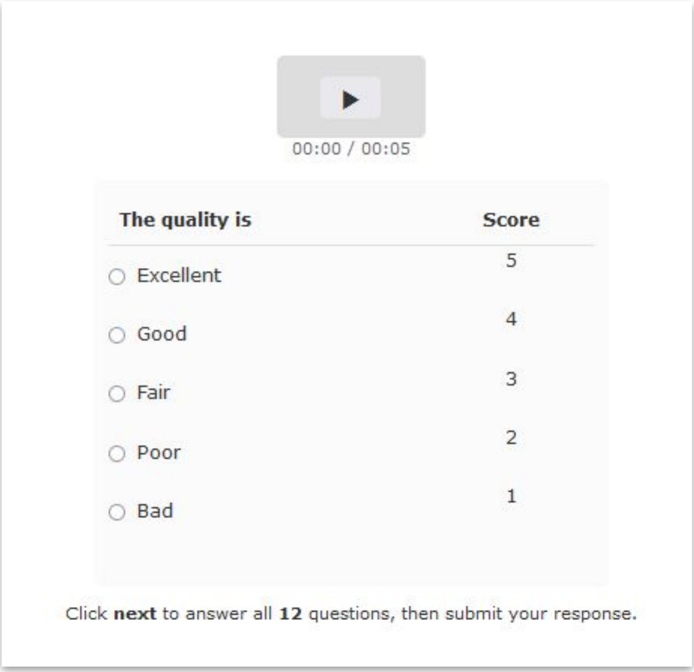
Training + Rating

Training

- Every 60 minutes
- Anchoring
- One trapping question with feedback

Ratings

- 10 clips + 1 gold question + 1 trapping


00:00 / 00:05

The quality is	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

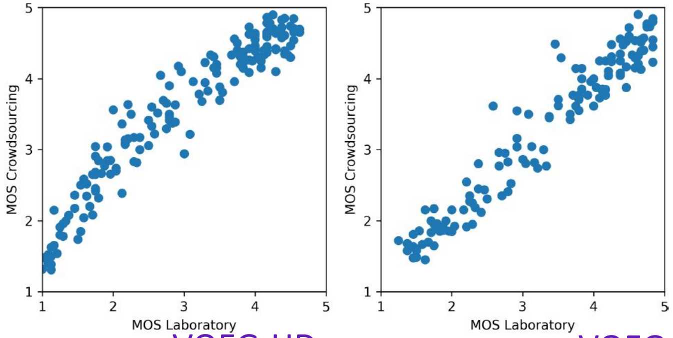
Click **next** to answer all **12** questions, then submit your response.

Validation

- VQEG HD3 and VQEG HD5
  - 168 sequences
  - Ratings per clip:
- Videos re-encoded using x264, CRF 17
- On average PCC 0.952
- Shows platform is accurate compared to lab study

Table 2: Comparison between laboratory and crowdsourcing experiments.

Dataset	MOS			DMOS		
	PCC	SPCC	RMSE FOM	PCC	SPCC	RMSE FOM
VQEG HDTV3 -run1	0.956	0.949	0.333	0.948	0.949	0.362
VQEG HDTV3 -run2	0.964	0.951	0.302	0.946	0.939	0.370
VQEG HDTV3 -run3	0.959	0.949	0.323	0.940	0.942	0.389
VQEG HDTV3 -run4	0.917	0.913	0.455	0.904	0.922	0.489
VQEG HDTV3 -run5	0.947	0.923	0.367	0.932	0.909	0.415
VQEG HDTV5	0.970	0.957	0.278	0.965	0.958	0.299



(a) VQEG-HD 3 (b) VQEG-HD 5

Reproducibility

- 5 repetitions on different days with different raters
- Shows system is highly repeatable

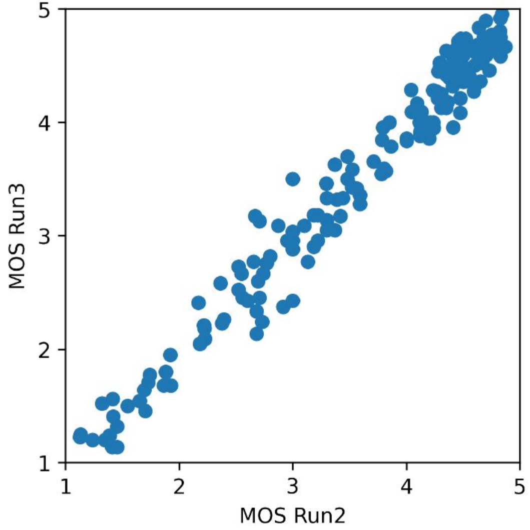


Table 3: Correlation coefficients between five runs of the VQEGHD3 dataset. Pearson correlation coefficient on upper triangle and Spearman's rank correlation coefficient on lower triangle.

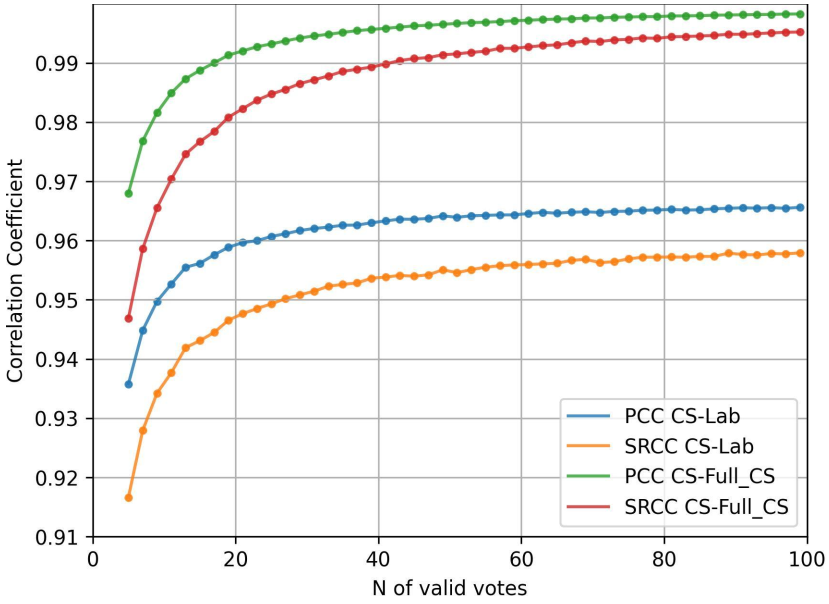
	Run 1	Run 2	Run 3	Run 4	Run 5
Run 1		0.984	0.987	0.957	0.977
Run 2	0.959		0.985	0.957	0.977
Run 3	0.974	0.969		0.952	0.972
Run 4	0.943	0.941	0.942		0.956
Run 5	0.954	0.947	0.942	0.933	

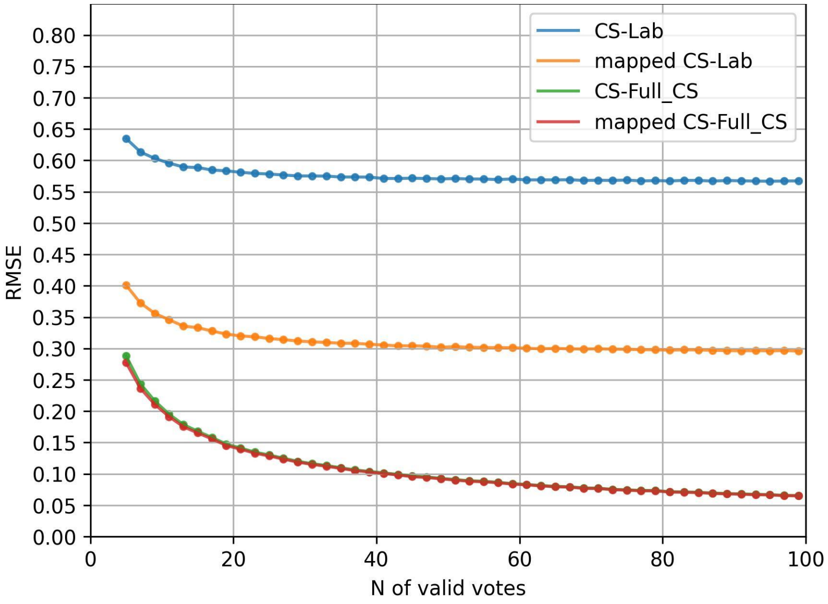
Ablation study

Case	PCC	SPCC	RMSE	RMSE after mapping
All passed	**0.96**	**0.96**	**0.62**	**0.31**
Gold clips failed	0.57	0.53	1.02	0.93
Play back duration failed	0.62	0.57	1.12	0.89
Brightness check failed	0.89	0.88	0.84	0.52
Straight liners	0.29	0.30	1.53	1.09
Viewing distance - passed	**0.93**	**0.92**	**0.76**	**0.41**
Viewing distance - failed	0.83	0.78	0.95	0.62
VAT Passed & All criteria passed	**0.91**	**0.90**	**0.88**	**0.47**
VAT Failed & All criteria passed	0.87	0.89	0.96	0.59
Complete test -all passed	**0.95**	**0.95**	**0.72**	**0.34**
No calibration	0.91	0.89	0.80	0.34
No Trapping clip	0.92	0.92	0.88	0.46

➡ This shows each check in the platform gets us closer to the lab study

Number of votes

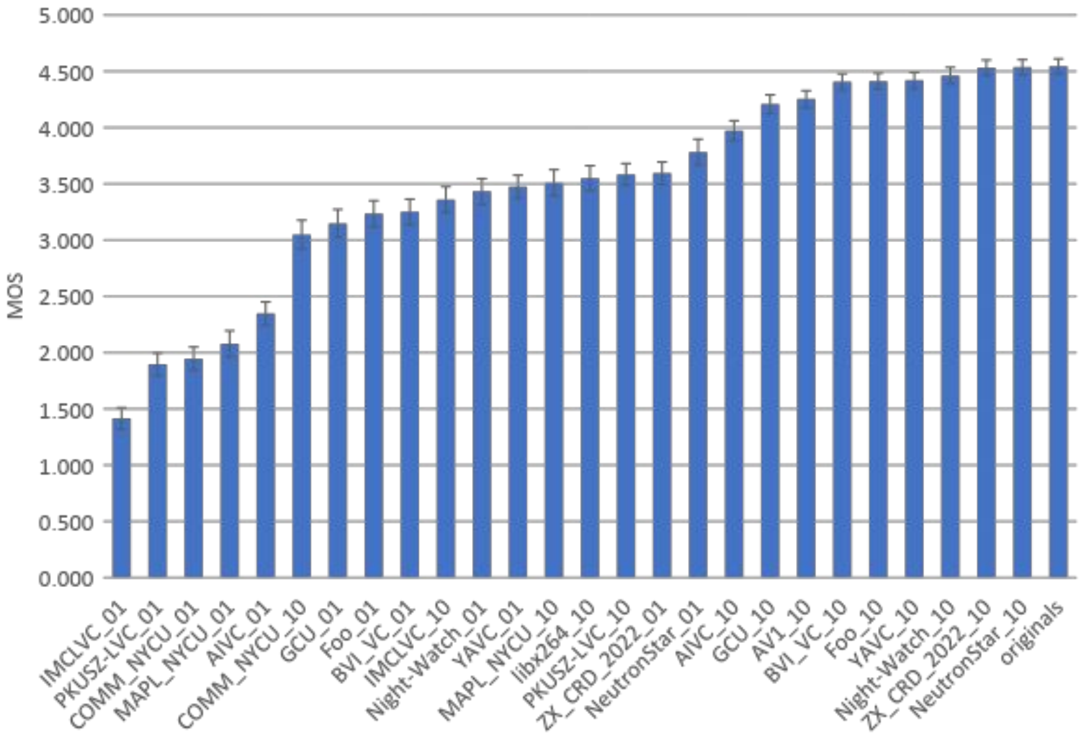




- **CS-Lab:** statistic between subset of CS and Lab
- **CS-Full_CS:** statistic between subset of CS and Full CS
- With N ~ 20 ratings we get close to the max CS-Lab and CS-CS PCC

Results: Round 1

- ACR
- 7 ratings per clip
- $1 per HIT
- Team_01: 0.1 Mbps
- Team_10: 1.0 Mbps



Results: Round 2

- Top 6 teams in each track
- 14 ratings per clip

Team name	MOS	95% CI
NeutronStar_10	4.450	0.05
ZX_CRD_2022_10	4.431	0.05
YAVC_10	4.410	0.05
Night-Watch_10	4.346	0.05
Foo_10	4.327	0.05
BVI_VC_10	4.306	0.05
NeutronStar_01	3.214	0.08
ZX_CRD_2022_01	3.084	0.07
YAVC_01	2.979	0.07
Night-Watch_01	2.793	0.08
BVI_VC_01	2.673	0.08
Foo_01	2.551	0.08

ANOVA (p-values)

	NeutronStar_01	ZX_CRD_2022_01	YAVC_01	Night-Watch_01	BVI_VC_01
NeutronStar_01					
ZX_CRD_2022_01	0.000				
YAVC_01	0.000	0.006			
Night-Watch_01	0.000	0.000	0.000		
BVI_VC_01	0.000	0.000	0.000	0.005	
Foo_01	0.000	0.000	0.000	0.000	0.009

NeutronStar_01, ZX_CRD_2022_01, YAVC_01 are significantly difference ($p < 0.01$)

	NeutronStar_10	ZX_CRD_2022_10	YAVC_10	Night-Watch_10	Foo_10
NeutronStar_10					
ZX_CRD_2022_10	0.426				
YAVC_10	0.259	0.731			
Night-Watch_10	0.008	0.058	0.124		
Foo_10	0.004	0.032	0.073	0.789	
BVI_VC_10	0.000	0.006	0.016	0.381	0.548

NeutronStar_01, ZX_CRD_2022_01, YAVC_01 are significantly tied, separated from Night-Watch_10 ($p < 0.06$)

ACR comparison to existing objective metrics

	PCC	SRCC
PSNR	0.69	0.67
MS-SSIM	0.75	0.79
VMAF	0.89	0.86

➡ These existing objective metrics are insufficient to evaluate / stack rank ML codecs

Comparison to DCR

- DCR reduces the content bias
- DCR gives similar results to ACR
  - PCC: 0.976
  - SRCC: 0.994
- The top 3 for 0.1 and 1.0 Mbps tracks don't change
  - There are some differences
- Note there are no public DCR lab studies to compare with
- DCR takes 2X longer to rate after qualification

Conclusion

- Platform in process of being standardized in ITU-T
- Platform available at: http://github.com/microsoft/P.910
- Paper: A crowdsourced implementation of ITU-T P. 910
  - Babak Naderi, Ross Cutler
- Next steps:
  - Create an objective full reference VQA model with PCC > 0.95 and SRCC > 0.95
  - Release this FRVQA and dataset to promote ML codec development

Lunch Break from 12:20 pm to 13:20pm CDT

In Person Poster Session in Evening: Hall D/E 225a-253a

Break from 15:45 pm to 15:55pm CDT

In Person Poster Session in Evening: Hall D/E 225a-253a

Potential Changes for 2023

Potential Changes for 2023

- Realism in image compression - topics for the Panel Discussion
    - Impose a much tighter runtime limit when using a GPU (e.g., 1x the time it takes VVC to decode on CPU)?
    - Create a track specific to "realistic" codecs (i.e., "1000 FLOPs/pixel")?
- Year-round evaluation server
    - Fixed validation set to track progress over time.
    - Test set released / decoder fix released before next workshop (as we currently do).

Potential Changes for 2023

- Video perceptual metrics
  - Have a similar track as our image perceptual metric, except on video
- Community raters
  - Training and getting time for expert raters is expensive.
  - Involving more raters from the compression community would be beneficial to a year-round evaluation setup.

Awards Ceremony

Prize Structure

- Top 3 on the leaderboard allotted for a monetary prize.
  - Limited to academic submissions.
- *New* Best Student Paper Award (for the paper only track).
- After conference, contact me (nickj at google.com) and ETH Zurich will disperse prize money (all listed awards in USD).

Perceptual Metric Track

1. IMCL-T1 ($600)
2. IQA_LY (Prize ineligible)
3. Kingslayer ($600)

Image Compression Track

1. NewbieCodec (Prize ineligible)
2. PKUSZ-201 ($600)
3. IMCL-T2 ($600)

Video Compression Track

1. NeutronStar (Prize ineligible)
2. ZX_CRD_2022 ($600)
3. YAVC ($600)

Best Student Paper Award

- Encourage more student participation (student as first author)
- Challenge tracks are very important (and also very competitive)

"Neural Face video Compression using Multiple Views" by Anna Volokitin et al.

$400 prize.

Poster Sessions: Hall D/E

225a-253a

CLIC 2022

Thank you for Attending the
5th Challenge on Learned Image Compression

Poster Session Now

See you in 2023!