Universität Konstanz

Methodologies for Large-Scale Crowdsourced Visual Quality Assessment

Oliver Wiedemann

Workshop on Large-Scale Subjective IQA, 14.12.2022

Project Goals and Overview

1. Understand visual quality
   - (Mostly) technical quality rather than aesthetics

2. Model and predict perception
   - What is the quality of a given image/video?

3. Enhance codecs and compression methods
   - Apply our insights to improve UX

Quality is a consensus-based property!

Crowdworkers Proven Useful

Crowd workers proven useful: A comparative study of subjective video quality assessment

Dietmar Saupe, Franz Hahn, Vlad Hosu
Igor Zingman, Masud Rana
Department of Computer and Information Science
University of Konstanz, Germany

Shujun Li
Department of Computer Science
Faculty of Engineering and Physical Sciences
University of Surrey, UK

Published at QoMEX 2016: [1]

Crowdworkers Proven Useful

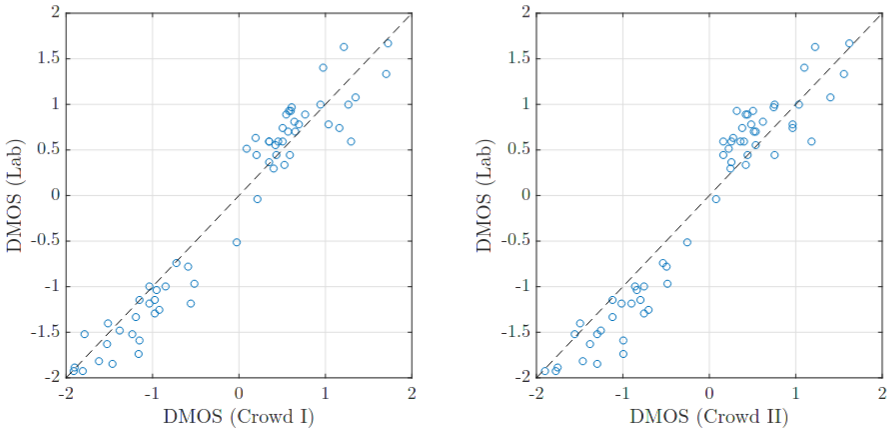


Fig. 1. Scatter plots comparing three assessments of DMOS values for 60 paired comparisons of video quality. Left: Crowd study 1 (strict quality control) versus DMOS derived from lab-based MOS values. Right: Crowd study 2 (mild quality control) versus lab. The Pearson correlation coefficients are 0.9687 (left), 0.9661 (right).

Selected Konstanz Image Quality Datasets

KonIQ-10k [2]
- 10,073 authentically distorted images
- 1.2 million ratings, 1459 crowd workers

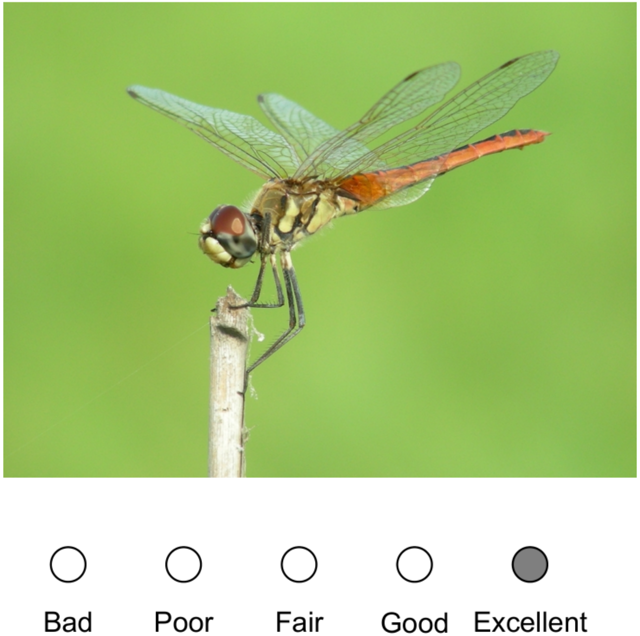
KADID-10k [3]
- 10,125 artificially distorted images
- 303,750 ratings by 2209 crowd workers

IQA-Experts-300 [4]
- 300 images, naturally + artificially dist.
- 70$ crowd ≈ 400$ pro freelancer ratings

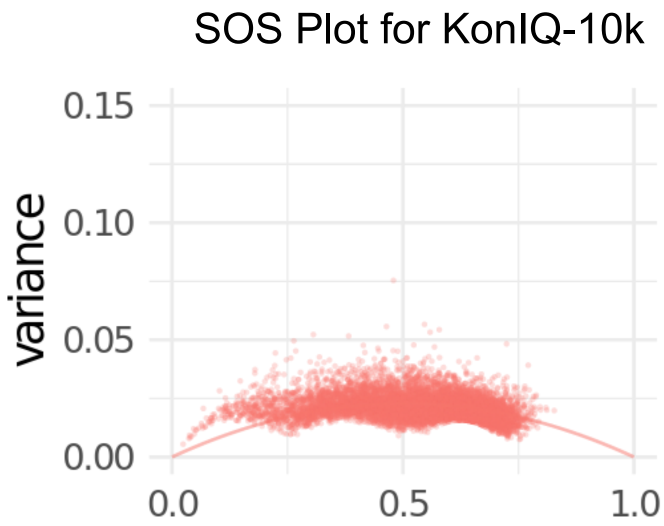
KonFIG-IQA [5]
- Artifact boosting
- Comparative study with score reconstruction

The Baseline: ACR

- Requires participant training
- Fast response per user and image
- Requires many votes per image
- ITU-T Recommendation P.913
- KonIQ-10k: ACR via `crowdflower`

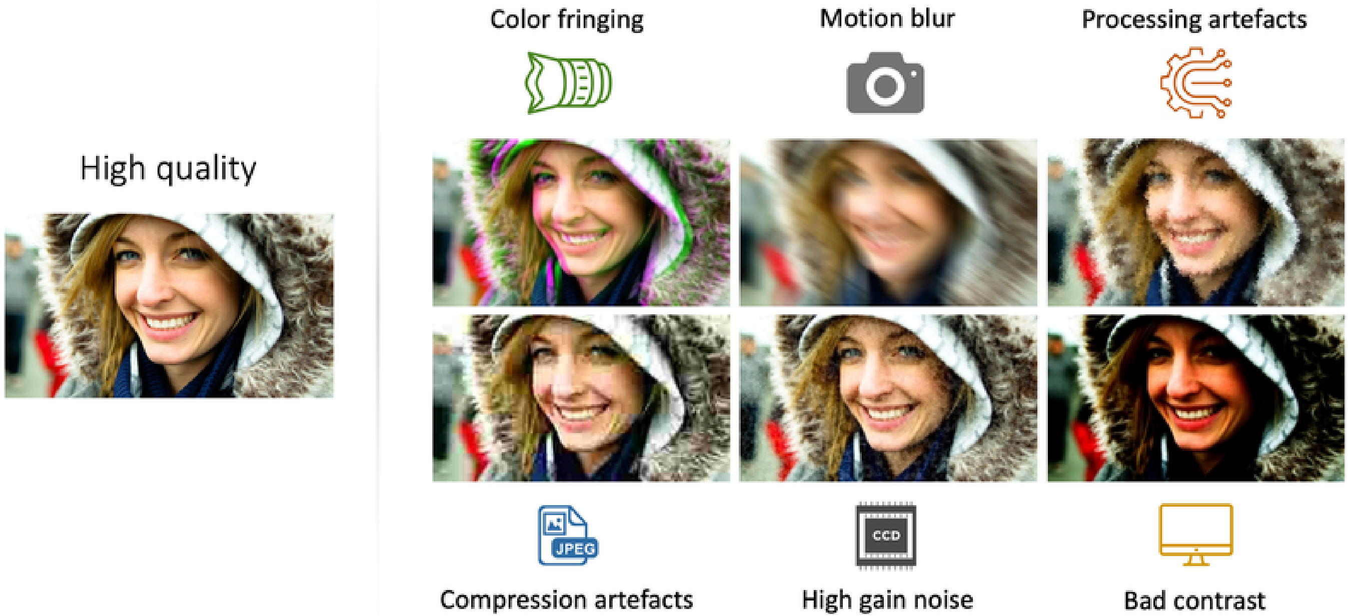


ACR: MOS vs Variance "Ripples"



note the lower edge of the scatterplot

Distortion Types and Magnitudes



Pairwise Comparisons

Unimpaired original.

JPEG compression ($qp = 10$).

Pairwise Comparisons

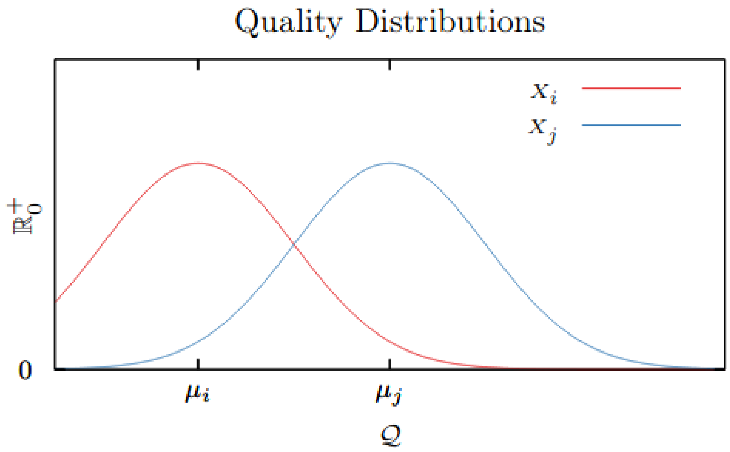
Raw experiment data will be a count matrix of preferences

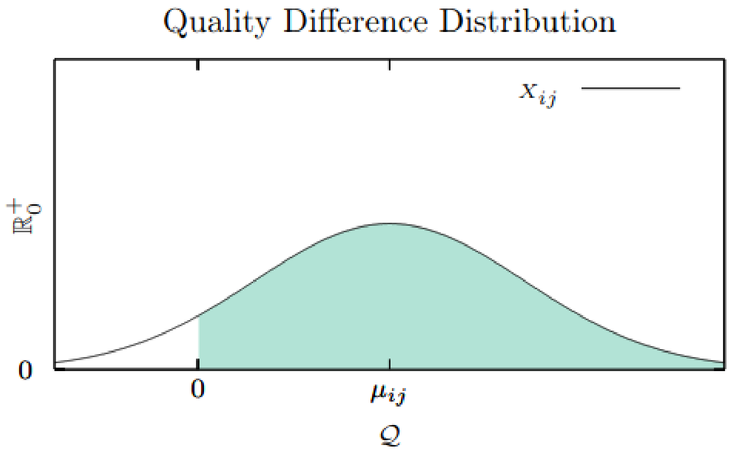
$$C_{i,j} = \begin{cases} \text{\# of votes preferring } i \text{ over } j \text{ for } i \neq j \\ 0 \text{ otherwise} \end{cases}$$

How do we get a reasonable scale from this?

Thurstonian Reconstruction

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \qquad X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$$





Thurstonian Reconstruction

Thurstone proposed five versions with increasingly strict assumptions.

Case V: X_i and X_j are uncorrelated with equal variances.

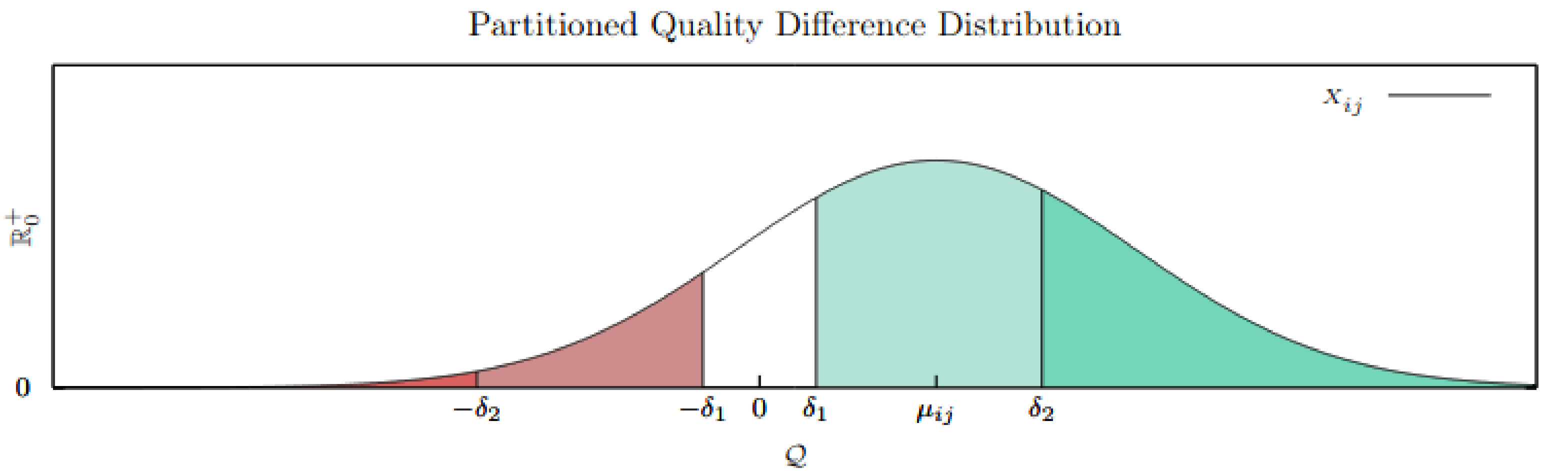
Setting $\sigma_i^2 = \sigma_j^2 = \frac{1}{2}$ leads to a simple closed form solution:

$$\hat{\mu}_{ij} = \Phi^{-1}\left(\frac{C_{i,j}}{C_{i,j} + C_{j,i}}\right)$$

To align multiple see [6], generally an optimization problem:

$$\arg\max_{\mu} L(\mu|C) \text{ sublect to } \sum_i \mu_i = 0$$

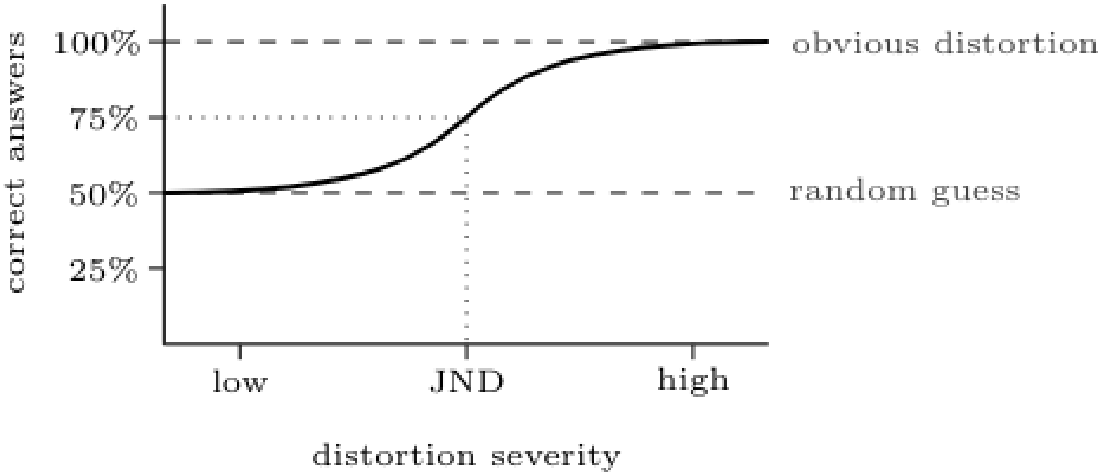
PCs with Multiple Options



ACR vs PC

	ACR	PC
1	Different understandings of quality, large variances for ratings	Independent of a nominal interpretation, faster responses
2	Saturation effect at range boundaries	No saturation effect by design
3	ACR scales are ordinal, difference in MOS between items does not translate well to perceptual difference	Reconstructed values are on an interval scale
4	Lack of meaningful units of measurement	Pair comparisons in units of JND

Just Noticeable Difference

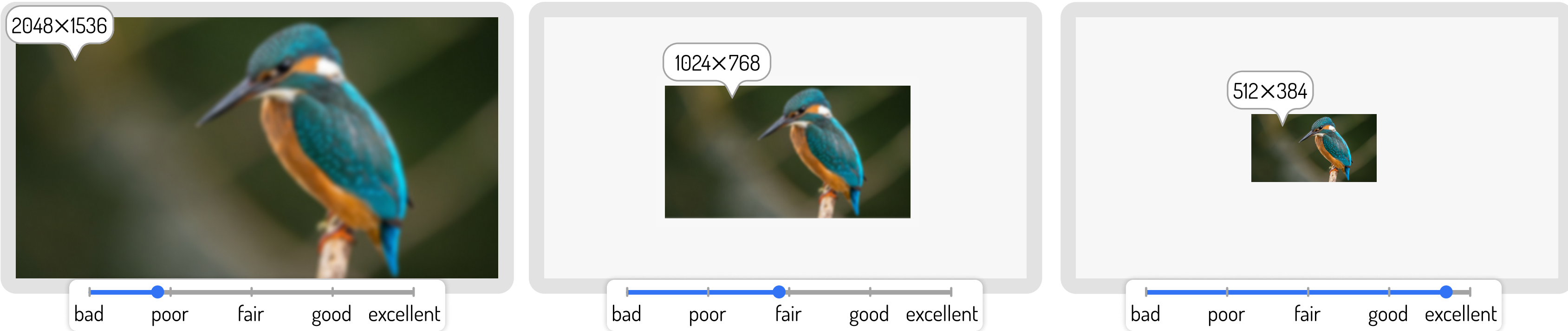


Artifact Boosting

Artifact Boosting

KonX: Cross-Resolution Assessment



Scaling affects subjective perception.

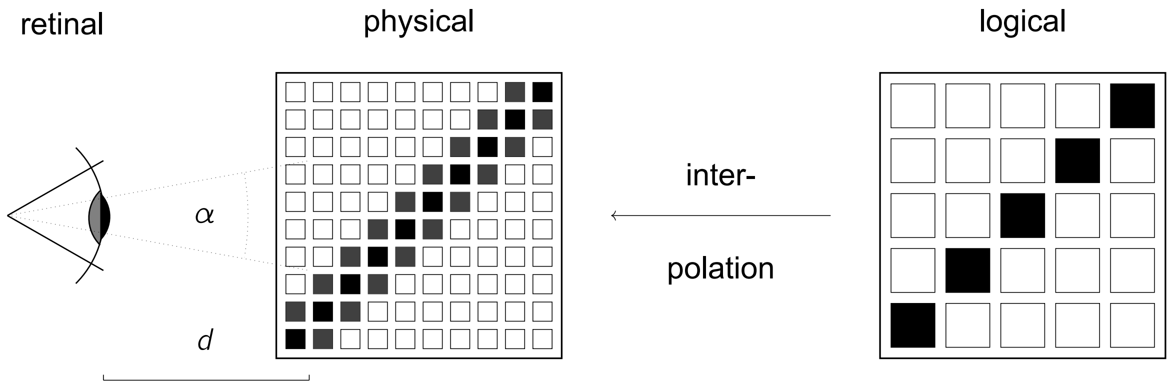
Image Scale vs CNNs

256×192px

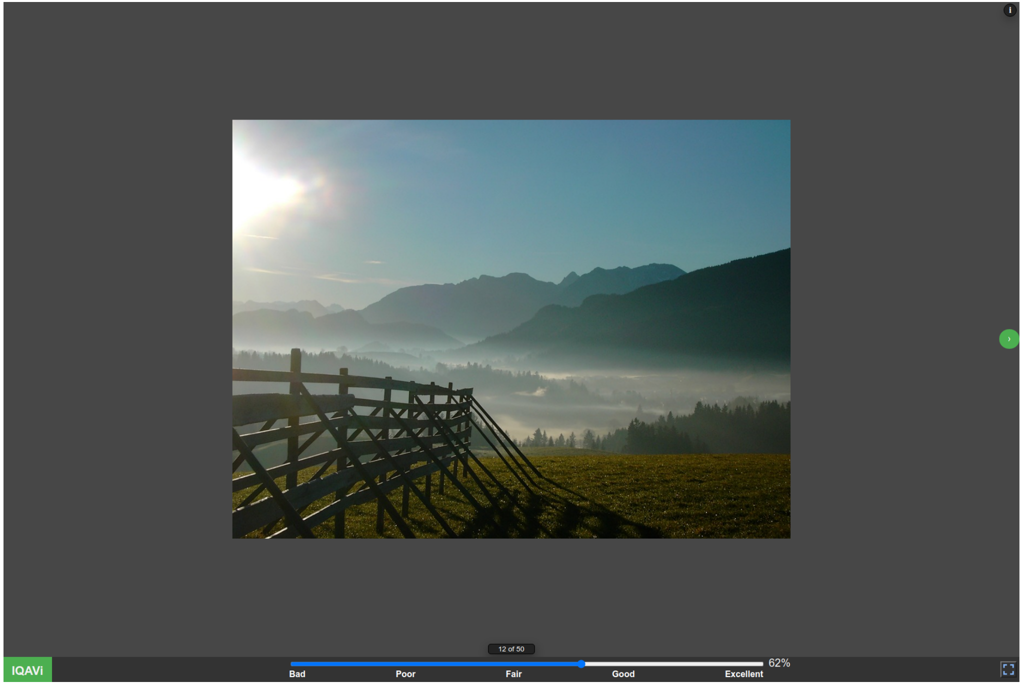
1024×768px

GradCAMs and predicted object classes change with scale.

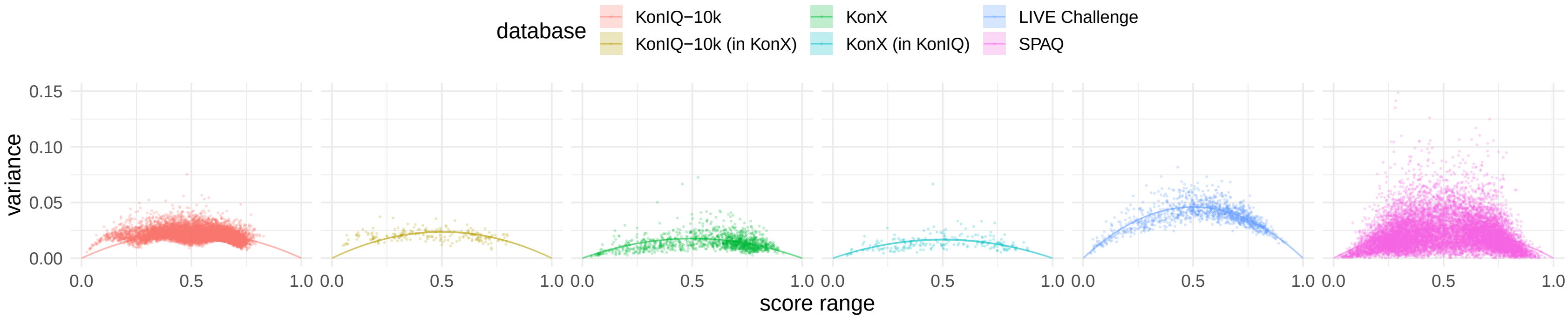
Careful with "Resolution"



The IQAVi Interface



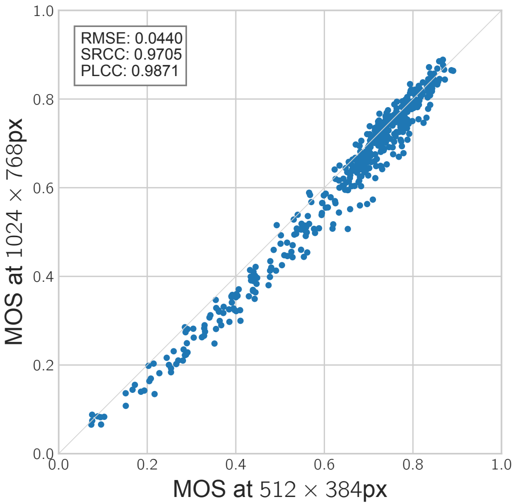
SOS Plots for Authentically Distorted DBs

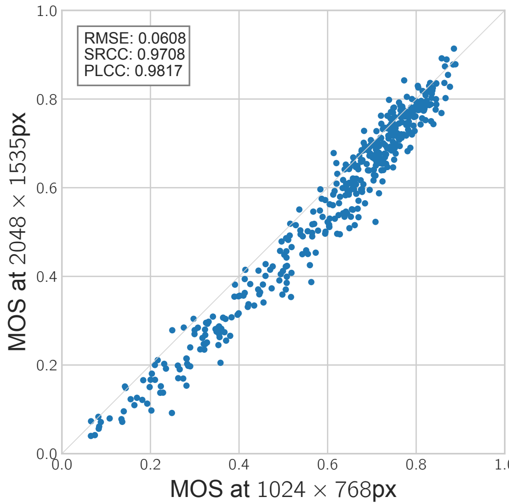


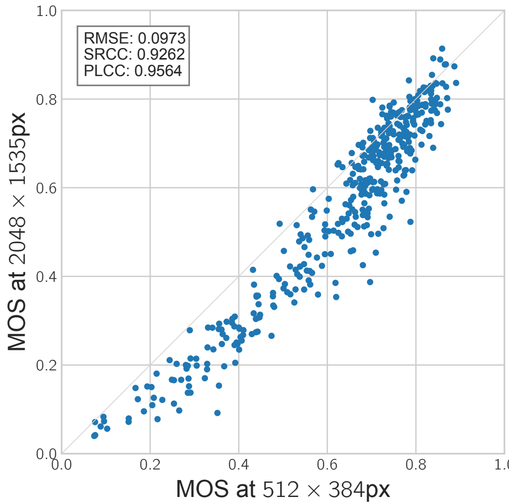
KonX: A Cross-Resolution IQA Benchmark

Sources	*Flickr* (KonIQ-10k) and *Pixabay*
#Images	210 from each source
Resolutions	2048×1535px, 1024×768px, 512×384px
Participants	19 in the full study
Annotations	2 per image at each resolution, 45360 in total

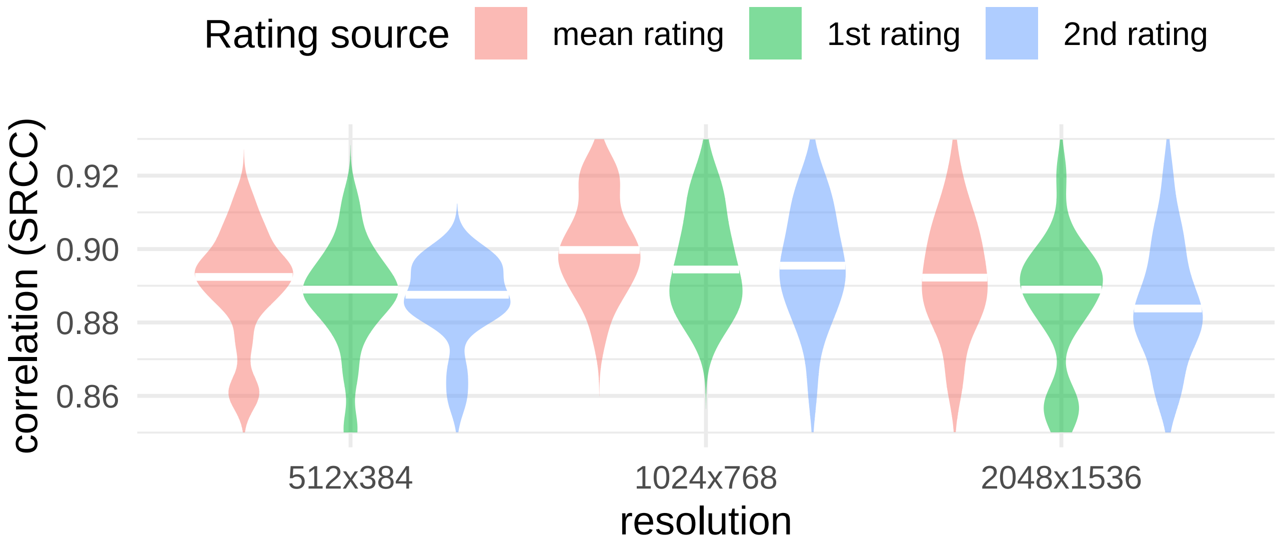
KonX MOS Scatterplots



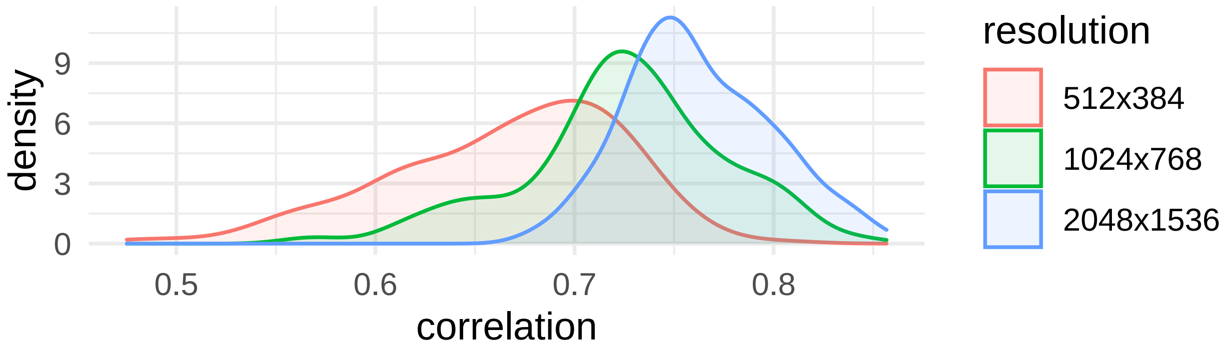




Correlations between KonX and KonIQ-10k

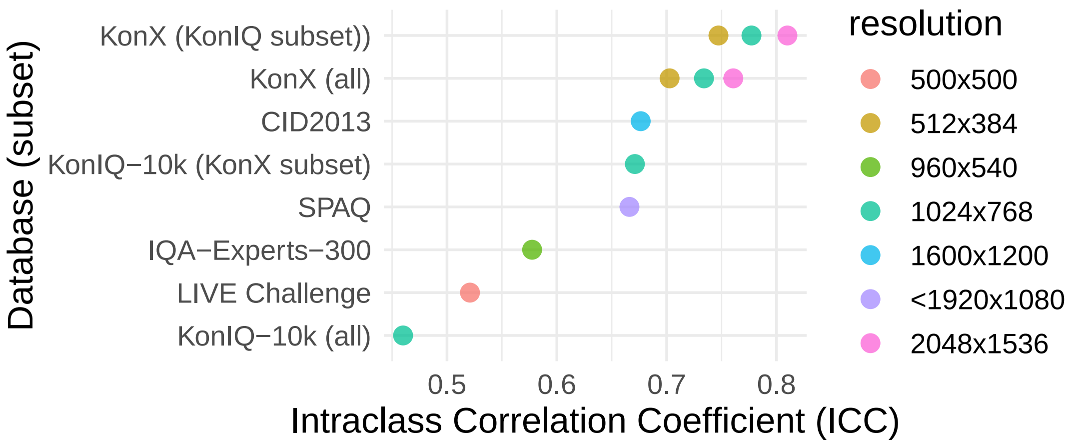


SRCCs Between KonX Participants by Resolution



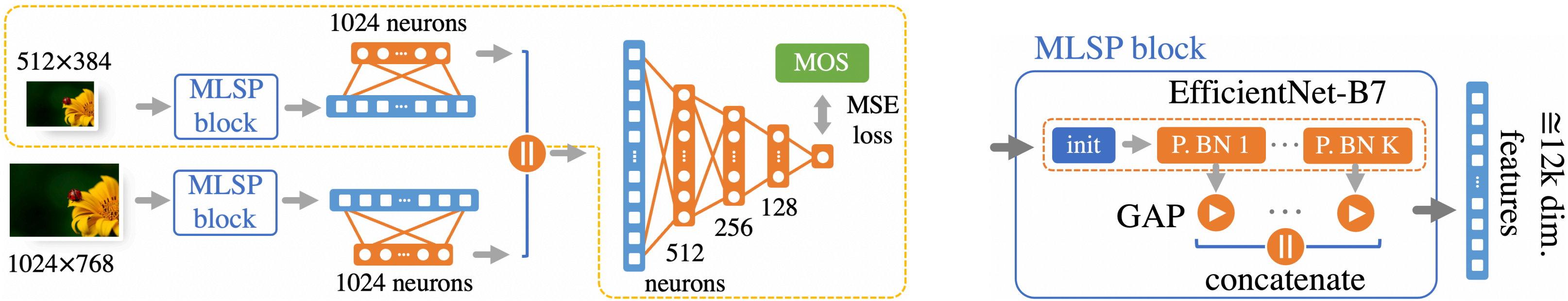
Larer images might be easier to assess.

Intraclass Correlation Coefficients

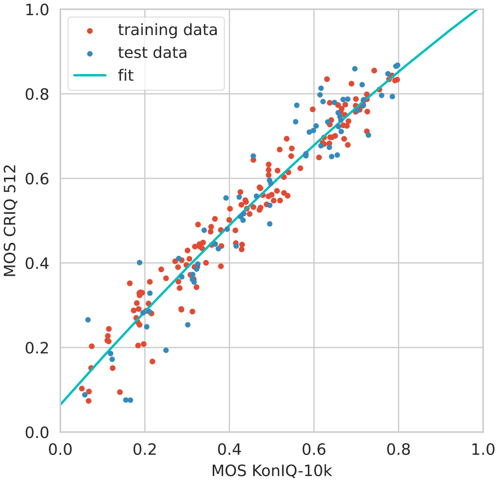


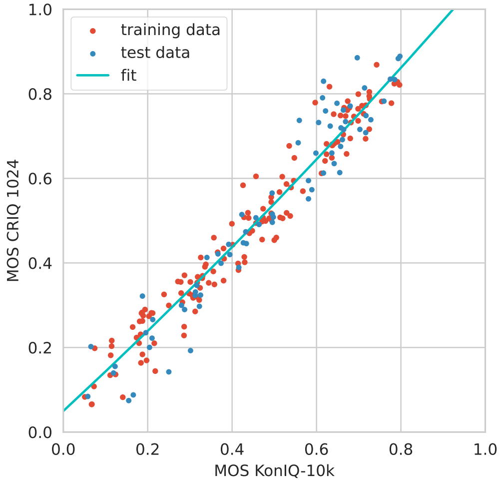
Agreement of individual scores per images is high in KonX.

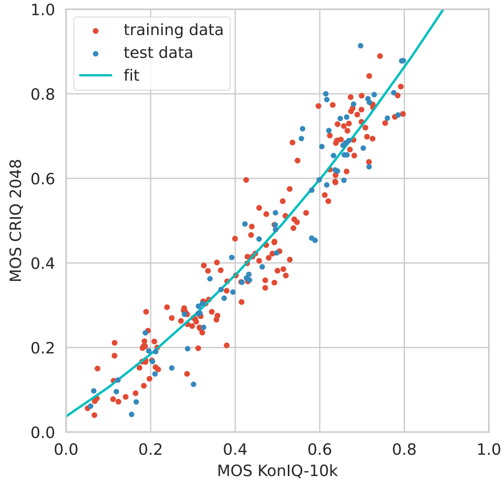
Effnet-2C-MLSP



Training on Remapped KonIQ-10k







Reduces MAE by 12.8%.

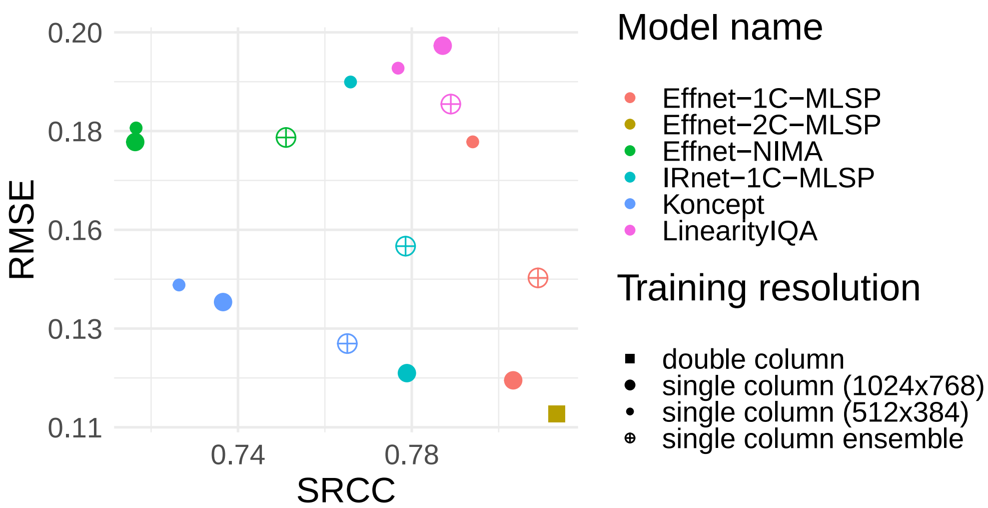
Cross-Database Model Performance

Models	KonIQ-10k		Live Challenge		SPAQ	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
LinearityIQA	0.9299	0.9415	0.8114	0.8404	0.8442	0.8422
Effnet-NIMA	0.7635	0.7788	0.6886	0.7269	0.7896	0.7936
IRN-1C-MLSP	0.8601	0.8932	0.8005	0.8310	0.8523	0.8553
Effnet-2C-MLSP	**0.9490**	**0.9596**	**0.8327**	**0.8595**	**0.8641**	**0.8641**

Results on CRIQ Splits

Model	Training Resolution	SRCC						PLCC					
		512×384px		1024×768		2048×1536		512×384px		1024×768		2048×1536	
		KonIQ	Pixabay	KonIQ	Pixabay	Koniq	Pixabay	Koniq	Pixabay	Koniq	Pixabay	Koniq	Pixabay
KonCept	512	0.8807	0.3047	0.8264	0.2703	0.6821	0.3112	0.8535	0.3049	0.7522	0.2670	0.6016	0.2690
	1024	0.8251	0.2658	0.8888	0.4175	0.8165	0.4518	0.6968	0.2658	0.8845	0.4201	0.8420	0.4926
Effnet-NIMA	512	0.8506	0.3101	0.7648	0.3739	0.5505	0.4010	0.8357	0.3682	0.7664	0.4118	0.5928	0.3972
	1024	0.8568	0.2506	0.8840	0.3184	0.8185	0.3925	0.8449	0.3105	0.8849	0.3895	0.8423	0.4503
LinearityIQA	512	**0.9436**	0.3818	0.9111	0.3994	0.7611	0.4485	**0.9416**	0.4681	0.9068	0.4670	0.7933	0.4859
	1024	0.9141	0.3849	**0.9452**	0.4519	0.9023	0.4935	0.9087	0.4311	0.9435	0.4813	0.9115	0.5291
IRN-1C-MLSP	512	0.9279	0.3197	0.9093	0.3490	0.8072	0.4501	0.9274	0.4155	0.9046	0.4355	0.8326	0.4967
	1024	0.8949	0.3117	0.9320	0.4190	0.9076	0.5037	0.8992	0.4003	0.9313	0.4876	0.9160	**0.5596**
Effnet-2C-MLSP	512	0.9273	0.3955	0.9056	0.4457	0.7900	0.5149	0.9248	0.4689	0.9035	0.5063	0.8252	0.5391
	1024	0.8918	0.3762	0.9358	0.4844	0.9105	**0.5415**	0.8957	0.4443	0.9361	**0.5422**	0.9228	0.5857
	both	0.9234	**0.4058**	0.9426	**0.4715**	**0.9276**	0.5132	0.9251	**0.4783**	**0.9437**	0.5220	**0.9325**	**0.5596**

RMSE vs SROCC on KonX



References

[1] Dietmar Saupe, Franz Hahn, Vlad Hosu, Igor Zingman, Masud Rana, and Shujun Li. Crowd workers proven useful: A comparative study of subjective video quality assessment. In QoMEX 2016: 8th International Conference on Quality of Multimedia Experience, 2016.

[2] 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.

[3] 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.

[4] Vlad Hosu, Hanhe Lin, and Dietmar Saupe. Expertise screening in crowdsourcing image quality. In QoMEX 2018: Tenth International Conference on Quality of Multimedia Experience, 2018.

[5] Hui Men, Hanhe Lin, Mohsen Jenadeleh, and Dietmar Saupe. Subjective Image Quality Assessment with Boosted Triplet Comparisons. arXiv e-prints, page arXiv:2108.00201, July 2021.

[6] Kristi Tsukida and Maya R Gupta. How to analyze paired comparison data. Technical report, Washington University, Seattle, Dept. of Electrical Engineering, 2011.