# Perception of video quality at a local spatio-temporal horizon

Andréas Pastor, Patrick Le Callet

# Research proposal: Perception of video quality at a local spatio-temporal horizon

Andréas Pastor
Patrick Le Callet
andreas.pastor@etu.univ-nantes.fr
patrick.le-callet@univ-nantes.fr
Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

## ABSTRACT

This paper contains the research proposal of Andréas Pastor that was presented at the MMSys 2022 doctoral symposium. Encoding video for streaming on Internet has become a major topic to reduce the consumption of bandwidth and latency. At the same time, the human perception of distortions has been explored in multiple research projects, especially for distortions generated by Coder-DECoder (CODEC) algorithms. These algorithms operate in a rate-distortion optimization paradigm to efficiently compress video content. This optimization can be driven by metrics that are most of the time not based on the human perception, and more importantly, not tuned to reflect the local perception of distortions by human eyes.

In this doctoral study, we proposed to work on the perception of localized distortion at a small temporal and spatial horizon. We present here the fundamental research questions and challenges in the domain with a focus on methods to collect perceptual judgments in subjective studies and metrics that can help us to derive an estimate of the perception of distortions by humans.

## CCS CONCEPTS

• **Human-centered computing**; • **Applied computing**;

## KEYWORDS

Perception, distortion, video quality, CODEC

**ACM Reference Format:**
Andréas Pastor and Patrick Le Callet. 2022. Research proposal: Perception of video quality at a local spatio-temporal horizon. In *13th ACM Multimedia Systems Conference (MMSys '22), June 14–17, 2022, Athlone, Ireland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3524273.3533931

## 1 MOTIVATION

Video encoding and research on optimization of encoding algorithms attempt to improve the compression performance and at the same time to keep the quality of content as high as possible to reduce the bandwidth consumption and latency of transmitted videos.


*MMSys '22, June 14–17, 2022, Athlone, Ireland*

ACM ISBN 978-1-4503-9283-9/22/06…$15.00
https://doi.org/10.1145/3524273.3533931

In the optimization phase of a CODEC encoder, different proposals are made for each coding unit (CU) and a selection is performed to reduce the cost (i.e. bit rate) while introducing the least possible distortions. This paradigm of being faithful to a reference block of pixels information can be measured by metrics like Sum of Squared Differences (SSD), and the Sum of Absolute Transformed Difference (SATD), which is an extension of the metric Sum of Absolute Difference (SAD).

Multiple research works have tried to replace these metrics with more perceptually based ones [5, 8, 9, 16, 23, 24, 28, 38, 46]. SSIM [39] and VMAF[21] are objective quality metrics that work well at a global video scale to predict the quality of video content. But their performances may vary and not be reliable at a localized scale, as needed in the decision process of CODEC. This research work will focus on creating a metric that can recognize locally the perceived distortion by human eyes and help to drive the encoding of videos.

This paper is organized as follows, we first introduce the research works related to this doctoral project in 2, then we discuss the different research questions in section 3. We present the work that we already performed in the section 4: how to adapt MLDS for inter-content scaling and a comparison study of different subjective methodologies for intra-content scale estimation. In section 5, we summarize the multiple other future works to conduct and finally a conclusion in section 6.

## 2 RELATED WORKS

In this part, we present the research works and domains that are interconnected with this doctoral project.

### 2.1 Subjective methodologies for perception evaluation

Perceptual studies through subjective experiments are important to collect ground truth information to help benchmark objective metrics, create datasets to train machine-learning or deep-learning models, and validate systems, like coding algorithms. Many subjective methodologies for video quality assessment VQA are standardized with detailed guidelines and instruction by ITU in the recommendations ITU-R Rec. BT.500 [10] for television and ITU-T Rec. P.910 [11] for multimedia applications.

During these subjective tests, most often collected data are used to locate stimuli on a perceptual scale, either by direct rating, e.g. using single stimulus methodologies like Absolute Category Rating (ACR) or double stimulus methodologies like Double Stimuli Impairment Scale (DSIS). Indirect rating is another type of method that relies on ranking or comparing stimuli: two-Alternative Forced Choice (2AFC) and pairwise comparison (PC) are some examples.

PC is considered a more reliable method since observers only need to indicate their preference on each pair of stimuli. Asking a judgment on a comparison is a more sensitive process than a direct annotation, which is important to improve the *discriminability* over stimuli.

At the same time, collecting these annotations can be time-consuming and expensive to perform. Due to the subjectivity of the data, we don't always have an agreement on the judgment of people and accurate estimation is important to reduce noise and uncertainty in collected data. In that case, to boost these subjective tests is essential to allocate annotation resources to the right stimuli.

Procedures such as Maximum Likelihood Difference Scaling (MLDS) [14, 25] relies on indirect rating and comparison of pairs of stimuli, and have shown some benefits in terms of discriminatory power, observers' cognitive load, and the number of trials required to achieve high reliability in the estimation of stimuli true perception. One of the disadvantages of MLDS is that it is a method for intra-content perception estimation, meaning that each of the stimuli estimated from one content cannot be directly related to the stimuli of another content since the estimations are made on separated perceptual scales.

Pairwise Comparison experiments are generally relying on a Pair Comparison Matrix (PCM) that contains the preferences of observers over pairs of stimuli and indicates how many times a stimulus has been preferred over another one. These PCM can be translated on a continuous scale, using models (e.g. Thurstone[35], Bradley and Terry [2], Tversky [36]).

A PCM can contain intra and inter-comparison of contents and estimate the location of stimuli on a unique perceptual scale. Unfortunately, the limitation is in the size of such matrix, since the number of possible comparisons is growing in $O(N^2)$, $N$ being the number of stimuli: introducing *efficiency* for a subjective protocol. A lot of previous works have focused on *active sampling* solutions [3, 7, 17, 18, 31, 42] and more recently with the work of [4, 19, 22, 26, 27, 34, 41]. Where the target is to select the most informative pairs and minimize *experimental effort* while maintaining accurate estimations and be *robustness* to bad annotator behavior (e.g. spammers), encountered in crowdsourcing environments.

## 2.2 Objective metrics for quality assessment

Objective metrics for quality evaluation are cost-efficient methods and avoid systematically running prohibitive resource-wise subjective experiments. Some metrics for VQA are VMAF[21], SSIM[39], MS-SSIM[40], PSNR, FSIM[44], VSI[43] and MAD[15]. VMAF is a machine learning-based approach that pools at a frame-level the features from other metrics VIF[33], DLM[20], computed at different resolutions in the video frames and information from the difference of consecutive frames as a motion feature. All of these metrics are performing at a global scale and not tuned to assess local quality in sub-parts of videos.

## 2.3 Deep-Learning features to estimate the human eye perception

With the advances in deep-learning computing, multiple models were developed to predict the quality of images and videos [1, 12, 13]. Other approaches [6, 32, 45] are focused only on local patches in the image. In the work on LPIPS[45], researchers focused on evaluating if features extracted from deep learning models, pre-trained on a variety of tasks not related to quality assessment, can be used as a proxy for the perceived quality in patches of size $64 \times 64$ pixels. The findings show that deep learning models can outperform some of the traditional objective metrics on specific types of distortions. A comment on these distortions types is, some of them come from various fields and applications of deep learning, such as super-resolution, generative models, colorization, and video deblurring, which are not the kind of distortions that we are trying to address in this doctoral project.

# 3 RESEARCH QUESTIONS

Based on the outcomes and literature in our domains, we selected to investigate the following research questions.

## 3.1 What is the best subjective methodology to estimate the perception of distortion?

Before creating an objective metric for local quality assessment, we need to collect data from small tubes: video localized at a small spatial and temporal horizon. Multiple subjective methodologies exist and need to be compared to find the one that is the most suited for our task. Our task can be split in two, with first an intra-content scaling where different stimuli from the same content are placed on their own grading scale. These scales, one per content, are important to understand the evolution of the perception with increasing levels of distortions, and since MLDS methodology uses a pre-ordering of stimuli this method can be the most suited.

Then the second task is inter-content scaling where stimuli of all the contents need to be positioned on a single perceptual scale. The goal is here to figure out which are the contents perceived as more or less impacted by the distortions generated by encoding algorithms. In [30], we proposed a modified version of the MLDS methodology to perform inter-content scaling. In [29], we applied the proposed method to start to collect data on small video patches.

## 3.2 How to generate interesting content to collect judgments?

From an existing dataset (e.g VideoSet [37]) that contains high-quality reference sources, it is possible to encode them with different parameters from a specific video CODEC, like AV1[1], and extract *tube-contents*: a set of tubes with a reference tube extracted at a specific location in the video source and multiple tubes extracted at the same location in the encoding/distorted version of the source. The size of the tubes is important. One tube will correspond to a Perceptual Unit (PU) corresponding to one degree of visual angle, around 64 pixels, and $200 - 300ms$ the fixation time of a gaze by eyes. These spatial and temporal horizons are important as they reflect the decision process made in the encoding algorithm at the level of a Coding Unit (CU).

[1] AV1 encoder v3.1.2, from AOM Alliance Open Media: https://aomedia.googlesource.com/aom/

**Figure 1: Example of 20 tube-contents selected for the first subjective experiment. Each column represents a tube-content and its distortion levels in increasing order from the top with the original tubes (not distorted) to bottom with the most distorted level.**

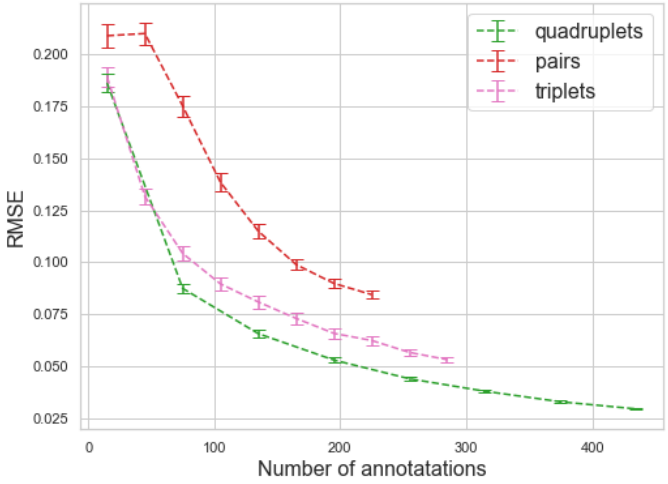


**Figure 2: Root Mean Squared Error RMSE evolution (averaged over the 8 contents) between ground truth estimate and partial estimate from randomly sub-set of annotators. A lower RMSE score indicates better performance from the methodology on the data.**

### 3.3 How to select distortions levels in each of the tube-content?

For each of the tube-contents that we extracted, we have multiple candidates to present to observers in a subjective experiment. A proxy of the perception of distortions can be used to sample efficiently on the perceptual continuum tube-candidates. VMAF has been widely adopted in the research community of VQA and can potentially be a good proxy to estimate the spacing between consecutive distortion levels in a tube-content. This proxy can also be corrected with the first estimations of a pilot subjective experiment. We already performed this step and corrected the bias in sensitivity to distortions of VMAF. Since this metric has been trained to assess the global quality of high-quality videos and not tuned for the small spatio-temporal horizon.

### 3.4 How Machine-Learning and Deep-Learning models help to predict the perception of distortions?

In a seminal paper [45], it has been found that the features learned by deep learning models pre-trained on various tasks, that are not directly related to quality assessment, can be used to extract information about the perception of distortions. However, some of these features are usually expensive to compute, requiring GPUs and large memories to run fast inferences. So, a study of the complexity and solution to extract only relevant features and simplify metrics will be part of this doctoral work.

### 3.5 How can these subjective perceptual judgments be integrated into a modern CODEC like AV1?

Finally, the objective metric created from perceptual judgments needs to be integrated into the coding recipes of the CODEC. Taking into account the different modes of a CODEC and being efficient to not overwhelm the already existing complexity of video encoding algorithms.

## 4 FIRST INVESTIGATIONS AND RESULTS

The first task to address the research questions is to study the different subjective methodologies available, especially the MLDS methodology, with its two versions: one based on quadruplets and the second one on triplets and compare them to pairwise comparison (PC). We targeted to study how these methods can estimate the perception of distortions by observers on a perceptual scale.

The methodologies need to be compared for intra and inter-content scaling. First, in intra-content scaling, we compared the methodologies and their efficiency in the number of annotators required to estimate precisely a scale for each content.

In figure 1, we present the tube-contents we selected to perform this first subjective study. Reference and distorted versions of a tube are extracted from video encoding using AV1 and varying the Quantization Parameter QP, to act on the level of distortion.

We used the software provided with MLDS to estimate a perceptual curve per content, and we compared which methodology produced the best results. Only 8 out of the 20 tube contents were used in 3 subjective tests. A first subjective test where annotators are presented with pair of intra-content tubes following the PC methodology design. In the second one, annotators rated triplets, and in the last one, quadruplets. We avoid as much as possible to recruit annotators to participate in more than one experiment, to prevent any bias.

In figure 2, we present the results when comparing the different subjective annotation methodologies. First, a reference ground truth is obtained by using all the data collected in the 3 experiments. Then an increasing number of judgments from different annotators are used from each experiment to check how efficient is a methodology when only X annotators are available to perform the task. We can see in the figure, that with only 80 annotations quadruplet methodology can have the same accuracy as triplets with 110 annotations or pairwise comparison with 200 annotations. These findings are in favor of quadruplets to collect data in intra-content comparison.

For inter-content comparison, the MLDS methodology doesn't have a procedure to estimate the scale between the stimuli of different contents. In the paper [30], we proposed a solution to perform inter-content scaling using quadruplets. We still need to perform a comparative study against pairwise comparison to validate that our proposed solution is more efficient.

## 5 FUTURE WORKS

The objective is next to create an annotated large-scale dataset of localized spatio-temporal tubes and a method to select these tubes. Since it is costly financially and in resources to collect large-scale annotated datasets, we need to define which are the interesting tube-contents and their distortion levels to then collect subjective judgments from crowdsourcing annotators. This task can be performed using already existing objective quality metrics as a proxy of local quality and select tube-contents where there is a large disagreement between the metrics. More traditional solutions exist using Spatial Information SI and Temporal Information TI statistics.

Since the existing metrics are not tuned for the local perception of distortion or not based on the perception of the Human Visual System, we need to identify potential good features from ML-trained metrics or Deep Learning based models which can benefit the prediction of local quality. This work will lead to the creation of a metric that can be then used in video CODEC. The selection from the different candidates of a block in the encoding process can be solved using this newly created metric.

## 6 CONCLUSION

In this paper, we presented the research questions on this doctoral project toward a perceptually oriented lightweight localized video quality metric. We introduce the motivations and the challenges induced by this topic, as well as a short overview of the related research works. We also stated the methods that we will use to achieve that goal and discussed the ongoing, partially published, and future works of the Ph.D. thesis.

## ACKNOWLEDGMENTS

This work is supervised by University Prof. Patrick Le Callet from Nantes Université, France, and founded by Netflix, Inc. I would like to thanks also people from Netflix, for joining our regular meetings and providing valuable feedback: Dr. Lukáš Krasula, Dr. Xiaoqing Zhu, and Dr. Zhi Li.

## REFERENCES

[1] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing* 27, 1 (2017), 206–219.

[2] Ralph A. Bradley and Milton E. Terry. 1952. The Rank Analysis of Incomplete Block Designs — I. The Method of Paired Comparisons. *Biometrika* 39 (1952), 324–345.

[3] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 193–202.

[4] Marc Demers, Pascal E Fortin, Antoine Weill, Yongjae Yoo, Jeremy R Cooperstock, et al. 2021. Active Sampling for Efficient Subjective Evaluation of Tactons at Scale. In *2021 IEEE World Haptics Conference (WHC)*. IEEE, 1–6.

[5] Sai Deng, Jingning Han, and Yaowu Xu. 2020. VMAF Based Rate-Distortion Optimization for Video Coding. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–6.

[6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. 2020. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728* (2020).

[7] Mark E Glickman and Shane T Jensen. 2005. Adaptive paired comparison design. *Journal of statistical planning and inference* 127, 1-2 (2005), 279–293.

[8] Yi-Hsin Huang, Tao-Sheng Ou, Po-Yen Su, and Homer H Chen. 2010. Perceptual rate-distortion optimization using structural similarity index as quality metric. *IEEE Transactions on Circuits and Systems for Video Technology* 20, 11 (2010), 1614–1624.

[9] Yi-Hsin Huang, Tao-Sheng Ou, Po-Yen Su, and Homer H. Chen. 2010. Perceptual Rate-Distortion Optimization Using Structural Similarity Index as Quality Metric. *IEEE Transactions on Circuits and Systems for Video Technology* 20, 11 (2010), 1614–1624. https://doi.org/10.1109/TCSVT.2010.2087472

[10] ITU Recommendation BT.500-14. 2019. Methodologies for the Subjective Assessment of the Quality of Television Images.

[11] ITU-T Recommendation P.910. 2008. Subjective video quality assessment methods for multimedia applications.

[12] Jongyoo Kim and Sanghoon Lee. 2017. Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1676–1684.

[13] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. 2018. Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 219–234.

[14] Kenneth Knoblauch, Laurence T Maloney, et al. 2008. MLDS: Maximum likelihood difference scaling in R. *Journal of Statistical Software* 25, 2 (2008), 1–26.

[15] Eric Cooper Larson and Damon Michael Chandler. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging* 19, 1 (2010), 011006.

[16] Bin Li, Houqiang Li, Li Li, and Jinlei Zhang. 2014. Lambda domain rate control algorithm for High Efficiency Video Coding. *IEEE transactions on Image Processing* 23, 9 (2014), 3841–3854.

[17] Jing Li, Marcus Barkowsky, and Patrick Le Callet. 2012. Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment. In *2012 19th IEEE International Conference on Image Processing*. IEEE, 629–632.

[18] Jing Li, Marcus Barkowsky, and Patrick Le Callet. 2013. Boosting paired comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs. In *Stereoscopic Displays and Applications XXIV*, Vol. 8648. International Society for Optics and Photonics, 86481V.

[19] Jing Li, Rafal K. Mantiuk, Junle Wang, Suiyi Ling, and Patrick Le Callet. 2018. Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation. *CoRR* abs/1810.08851 (2018). arXiv:1810.08851 http://arxiv.org/abs/1810.08851

[20] Songnan Li, Fan Zhang, Lin Ma, and King Ngi Ngan. 2011. Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Transactions on Multimedia* 13, 5 (2011), 935–949.

[21] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. 2016. Toward a practical perceptual video quality metric. *The Netflix Tech Blog* 6, 2 (2016).

[22] Suiyi Ling, Jing Li, Anne-Flore Perrin, Zhi Li, Lukás Krasula, and Patrick Le Callet. 2020. Strategy for Boosting Pair Comparison and Improving Quality Assessment Accuracy. *CoRR* abs/2010.00370 (2020). arXiv:2010.00370 https://arxiv.org/abs/2010.00370

[23] Zhengyi Luo, Chen Zhu, Yan Huang, Rong Xie, Li Song, and C-C Jay Kuo. 2021. VMAF Oriented Perceptual Coding Based on Piecewise Metric Coupling. *IEEE Transactions on Image Processing* 30 (2021), 5109–5121.

[24] Chengyue Ma, Karam Naser, Vincent Ricordel, Patrick Le Callet, and Chunmei Qing. 2016. An adaptive Lagrange multiplier determination method for dynamic texture in HEVC. In *2016 IEEE International Conference on Consumer Electronics-China (ICCE-China)*. IEEE, 1–4.

[25] Laurence T Maloney and Joong Nam Yang. 2003. Maximum likelihood difference scaling. *Journal of Vision* 3, 8 (2003), 5–5.

[26] Hui Men, Hanhe Lin, Mohsen Jenadeleh, and Dietmar Saupe. 2021. Subjective Image Quality Assessment With Boosted Triplet Comparisons. *IEEE Access* 9 (2021), 138939–138975.

[27] Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafał K Mantiuk. 2021. Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2559–2566.

[28] Karam Naser, Vincent Ricordel, and Patrick Le Callet. 2016. Modeling the perceptual distortion of dynamic textures and its application in HEVC. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3787–3791.

[29] Andréas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, and Patrick Le Callet. 2022. On the Accuracy of Open Video Quality Metrics for Local Decision in AV1 Video Codec. In *2022 IEEE International Conference on Image Processing (ICIP)*. 4013–4017. https://doi.org/10.1109/ICIP46576.2022.9897469

[30] Andréas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, and Patrick Le Callet. 2022. Improving Maximum Likelihood Difference Scaling Method To Measure Inter Content Scale. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2045–2049. https://doi.org/10.1109/ICASSP43922.2022.9746681

[31] Thomas Pfeiffer, Xi Alice Gao, Yiling Chen, Andrew Mao, and David G Rand. 2012. Adaptive polling for information aggregation. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[32] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. 2018. PieAPP: Perceptual Image-Error Assessment through Pairwise Preference. *CoRR* abs/1806.02067 (2018). arXiv:1806.02067 http://arxiv.org/abs/1806.02067

[33] Hamid R Sheikh and Alan C Bovik. 2006. Image information and visual quality. *IEEE Transactions on image processing* 15, 2 (2006), 430–444.

[34] Edwin Simpson and Iryna Gurevych. 2020. Scalable Bayesian preference learning for crowds. *Machine Learning* (2020), 1–30.

[35] Louis Leon Thurstone. 1927. A Law of Comparative Judgement. *Psychological Review* 34 (1927), 278–286.

[36] Amos Tversky. 1972. Elimination by aspects: A theory of choice. *Psychological review* 79, 4 (1972), 281.

[37] Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeong-Hoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, Yun Zhang, Jiwu Huang, Sam Kwong, and C.-C. Jay Kuo. 2017. VideoSet: A Large-Scale Compressed Video Quality Dataset Based on JND Measurement. *CoRR* abs/1701.01500 (2017). arXiv:1701.01500 http://arxiv.org/abs/1701.01500

[38] Shiqi Wang, Abdul Rehman, Zhou Wang, Siwei Ma, and Wen Gao. 2011. SSIM-motivated rate-distortion optimization for video coding. *IEEE Transactions on Circuits and Systems for Video Technology* 22, 4 (2011), 516–529.

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[40] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. Ieee, 1398–1402.

[41] Qianqian Xu, Jiechao Xiong, Xi Chen, Qingming Huang, and Yuan Yao. 2018. Hodgerank with information maximization for crowdsourced pairwise ranking aggregation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[42] Peng Ye and David Doermann. 2014. Active sampling for subjective image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4249–4256.

[43] Lin Zhang, Ying Shen, and Hongyu Li. 2014. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing* 23, 10 (2014), 4270–4281.

[44] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing* 20, 8 (2011), 2378–2386.

[45] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *CoRR* abs/1801.03924 (2018). arXiv:1801.03924 http://arxiv.org/abs/1801.03924

[46] Chen Zhu, Yan Huang, Rong Xie, and Li Song. 2021. HEVC VMAF-oriented Perceptual Rate Distortion Optimization using CNN. In *2021 Picture Coding Symposium (PCS)*. 1–5. https://doi.org/10.1109/PCS50896.2021.9477459