https://doi.org/10.2352/EI.2026.38.10.HVEI-234

The influence of image semantic complexity on the performance of image quality metrics

***Peiyuan Zhang*[1], *Xinwei Liu*[1], *Marius Pedersen*[2] *and Sophie Triantaphillidou*[2]**
(1) Zhejiang Wanli University (ZWU), Ningbo, Zhejiang, China.
(2) Colourlab, Department of Computer Science, Norwegian University of Science and Technology, Gjøvik, Norway.

Abstract

Image quality assessment has been a longstanding area of research, with significant efforts dedicated to developing objective metrics that can reliably predict perceived image quality. While numerous image quality metrics have been proposed, ranging from traditional handcrafted approaches to modern machine learning-based models, their evaluation typically relies on statistical comparisons with subjective human ratings where correlation is the primary measure of performance. In this study, we explore an additional dimension in image quality evaluation: the impact of image semantic complexity on metric performance. Specifically, we hypothesize that the number of distinct semantic categories within an image influences the accuracy of image quality metrics. We evaluate 8 state-of-the-art image quality metrics across 2 benchmark datasets, analyzing performance variations with respect to image semantic complexity (category count), based on two vision-language models. Our findings reveal that for some image quality metrics there is an impact of semantic complexity and outliers. This suggests that content-aware evaluation could be crucial for future image quality research.

Introduction

The assessment of image quality has been an active research field for many years, with attention given to proposing objective image quality metrics to predict perceived image quality. Surveys show an impressive number of proposed image quality metrics [1], and nowadays machine learning based image quality metrics are frequently introduced [2, 3].

The evaluation of image quality metrics is carried out by comparing their prediction with the scores of subjective experiments. Common statistical measures include correlation, outlier ratio, root-mean-squared error, USTRESS [4]. These have been used extensively in the evaluation of image quality metrics [3, 5–9], and provide information on different aspects of the performance of the metrics.

In this work, we go beyond the normal assessment and investigate if the content of the images used for the evaluation influences the performance of image quality metrics. Our hypothesis is that the semantic complexity of images, here defined as the number of semantic categories in the image, influences performance. In view of this hypothesis, we analyze 8 state-of-the-art deep learning based image quality metrics on 2 datasets, in which category count has been based on two vision-language models. Olivia et al. [10] found that visual complexity is principally represented by the perceptual dimensions of quantity of objects, clutter, openness, symmetry, organization, and variety of colors. In our work, we focus on the first element, quantity of objects, represented by the number of semantic categories.

This paper is organized as follows. First, we introduce relevant background, then the methodology, before we present the experimental results. At last, we conclude and propose future work.

Background

Analysis of the performance of image quality metrics have been carried out by numerous researchers.

Samajdar and Quraishi [11] analyzed a set of image quality metrics and their performance for different distortions using correlation. Ponomarkenko et al. [12] evaluated image quality metrics on TID2013 for the full dataset and for subsets. Lin and Kup [13] did a similar analysis on image quality metrics on full datasets and their subsets. Pedersen and Hardeberg [1] also carried out a similar analysis of image quality metrics on various datasets, and subsets of the TID2008 dataset. Nouri et al. [14] also analyzed image quality metrics statistically in different datasets and image subsets. Hanhart et al. [15] benchmarked 35 image quality metrics for HDR content, and analyzed amongst other the impact of different color channels using correlation, outlier ratio, and RMSE.

Chetouani [16] analyzed image subsets and ranked image quality metrics based on their performance for each subset, and suggested that one might classify images to select the best performing image quality metric for a specific class, for example using the approach suggested by Charrier et al. [17].

Analysis has also been carried out with regard to single and multiple distortions, indicating differences in performance between image quality metrics in this aspect [18, 19].

Yang et al. [20] analyzed quality assessment of screen content images, which included analysis of an image with higher variation between the evaluated quality metrics.

Outlier ratio has been used to analyze performance, such as for SSIM [21], CBM [22], and by Sazzad et al. [23]. Group maximum differentiation was proposed by Ma et al. [24], which automatically selects subsets of image pairs from a given dataset where image quality metrics compete with each other.

The NIMA metric [25] calculates the distribution of quality scores, and performance evaluation of this metric was carried out using correlation and earth mover's distance.

Pedersen and Cherepkova [26] analyzed outliers for a set of image quality metrics with regards to features; namely colorfulness, lightness, busyness, entropy and sharpness. They did not find a clear dependency between image properties and the prediction from the image quality, however, some common characteristics exist.

Triantaphillidou et al. [27] investigated the relationship between scene content and subjective results. Overall, they observed that quantification of scene features could match visual observations of the scenes, and that this could provide explanation on how these features impacted the perceptability of distortions.

To the best of our knowledge in depth analysis of outliers (points, either subjective human ratings or objective metric predictions, that significantly deviate from the expected or majority trend) have not been done. Therefore, in this paper, we analyze outliers with regard to the number of categories (semantic complexity), with the hypothesis that semantic complexity has an impact on the performance image quality metrics. Our hypothesis is based on findings that subjective image quality ratings are influenced by image busyness (or low-level image complexity) [27], which is usually present in images with many objects (images with high semantic complexity). Such complex images can result to outliers when it comes to correlating subjective results with image quality metric ratings.

Methodology

Image Quality Metrics

We have selected recent image quality metrics that are based on CNNs, Transformers, and CLIP, namely: ARNIQA [28] (regressor trained on KonIQ-10k), CNNIQA [29] (trained on KonIQ-10k), HyperIQA [30](trained on KonIQ-10k), NIMA [25] (trained on KonIQ-10k), MANIQA [31] (trained on KonIQ-10k), TReS [32] (trained on KonIQ-10k) CLIPIQA+ [33] (ViT-L/14, fine-tuned on KonIQ-10k), and QualiCLIP+ [34]. These image quality metrics represent recent methods, that have been shown to perform well on different datasets. They constitute a good basis for further analysis of their performance and if they are impacted by image content.

Datasets

We use the Koniq-10K [3] dataset for our analysis. This dataset has 10073 images, with camera distortions, which has been rated by 1459 crowd observers, with 120 ratings per image. The dataset has been used extensively for evaluation, but also for training deep learning image quality metrics. It contains different semantic content, and images were selected to span different attributes such as lightness, contrast, colorfulness, sharpness, content embeddings, which ensured further diversity of content.

In addition, we use the LIVE in the wild image quality Challenge dataset [35]. LIVE Challenge contains 1,162 images, captured using mobile devices, evaluated by over 8100 observers. The dataset has diverse camera distortions, and covers a wide range of semantic content.

Semantic complexity analysis

We use Qwen-VL-Max (Qwenvl) [36] and Doubao-Seed-1.6-vision (Doubao) [37], both large-scale vision-language models, to count the number of different categories in an image. A higher number of categories implies more objects in the image. Further, for each metric, we define outliers as the points furthest from the linear regression line fitted between its predicted Mean Opinion Scores (MOS) and subjective MOS. Our analysis is starting from a small percentage of outlier (5%) and increasing the number of outliers (up to 60%). We calculate the mean number of categories for the remaining images (inlier points). We perform this analysis for all the image quality metrics.

Figure 1 shows the semantic category count for the Koniq-10K dataset while Figure 2 shows the semantic category count for the LIVE Challenge dataset, processed with the two vision-language models. Some disagreement is expected, but in general their results correlate. We show the results for the Doubao model only, as the results are similar. We can notice that the range of categories go from a single category up to 20 categories in the image. Figure 3 shows examples of images from Koniq-10K that have few categories (top row) and images with the most categories (bottom row), while Figure 4 shows examples of images from LIVE Challenge that have few categories (top row) and images with the most categories.

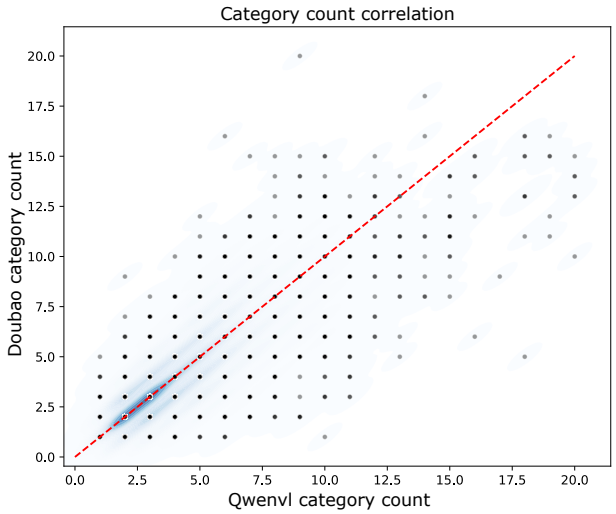


Figure 1: Correlation between category counts for Qwenvl and Doubao for the Koniq-10K dataset.

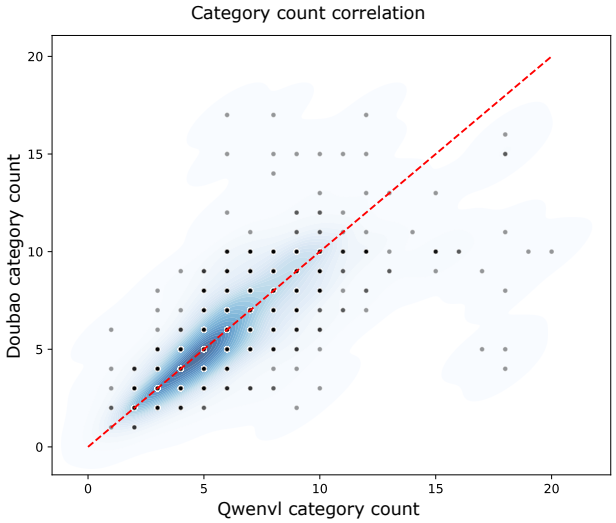


Figure 2: Correlation between category counts for Qwenvl and Doubao for the LIVE Challenge dataset.

Figure 3: Example of images from Koniq-10K with few categories (top row) and images with the most categories (bottom row).

Figure 4: Example of images from LIVE Challenge with few categories (top row) and images with the most categories (bottom row).

Results

We analyze the results as plots of the mean number of categories per image versus the outlier points (in percentage). We analyze the underpredicted outliers separately from the overpredicted outliers (i.e. points below and above the regression line). The results based on the category count using Doubao and the KONIQ-10K dataset are found in Figure 5. In general, the category count for outliers above the regression line (score overpredicted) are for most image quality metrics stable. TReS has a slight decline going from 5% outliers to 60% outliers. The exception is with MANIQA where the mean class count increases with the number of discarded outliers, with an indication that this metric overpredicts images with fewer classes. We can also notice that the outliers in MANIQA has a lower mean category count, at 10% outliers this is at 3 categories, while most of the other metrics have above 4 categories as the mean. In Figure 5 we analyze the outliers below the regression line, we notice a different behavior than those above the regression line. For ARNIQA, CNNIQA, TRES, and QUALICLIP+, we see that the category count increases with the number of outliers. This indicates that these metrics underpredict the quality of images with fewer categories. MANIQA has a different behaviour, where the most extreme outliers have more categories and the mean number of categories is reduced with increasing number of outliers. This indicates that MANIQA underpredicted images with more semantic complexity. For HyperIQA, NIMA and CLIPIQA+ we do not see an impact of category count. It is also interesting to notice that the mean category count for the outlier above and below the regression line is different. As an example, the 60% mean outlier category count is almost 4.4 for ARNIQA above the regression line but only 3.6 below the regression line. We see a similar tendency for CNNIQA, with 4.4 above the regression and 3.6 below. For MANIQA we see the opposite, with 3.4 as the mean category count at 60% outliers above and a bit above 4.8 mean category count in the points below the regression line.

When analyzing the results for Koniq-10K using Qwenvl we see the same trends as for Doubao in Figure 5. This is expected as Doubao and Qwenvl results are correlated, as shown in Figure 1.

Figure 6 shows the results for the LIVE Challenge dataset when using Doubao. On the left column showing the outlier above the regression line, we can notice that the curves have higher variability compared to Koniq-10k, which is expected as the size of the LIVE dataset is only approximately 10% of Koniq-10k. We notice that the CNNIQA curve has an increasing trend, for the first part, which indicates that the most extreme metric outliers originate from images with fewer categories. Looking at the outliers below the regression line in the right column, we notice that several image quality metrics (ARNIQA, CNNIQA, HyperIQA, NIMA, TReS, QualiCLIP+) all have curves that increase with the number of outliers, indicating that these metrics tend to underpredict the quality of images with fewer categories. Looking more closely at these, we can also notice that for many the category count is lower for the outlier below the regression line than those above. If we look at the 60% outlier mean, for ARNIQA this is 7 categories above and around 4,75 below, while for CNNIQA it is almost 6,5 above and 5,25 below, and a similar tendency for the other metrics. MANIQA and CLIPIQA does not have a clear trend.

Table 1 quantifies the relation between outlier percentage and mean semantic category count on KonIQ and LIVE Challenge, using Spearman ρ (subtable (a)) and the linear slope (subtable (b)), for outliers above and below the regression line. We first notice a very clear asymmetry between the two sides. For the below-line (underpredicted) outliers, the relationship is consistently strong and positive on both datasets: on KonIQ, most metrics have ρ very close to 1 (typically ≈ 0.87–1.00) with clearly positive slopes (around 0.008–0.012 for several methods), and on LIVE Challenge the same pattern holds with uniformly high ρ (roughly ≈ 0.81–0.99) and even larger positive slopes (up to ≈ 0.016). This indicates that as we keep more extreme underpredicted outliers, their images tend to have higher semantic category counts. In contrast, for the above-line (overpredicted) outliers the behavior is much more method-dependent: some metrics still show positive association (e.g., ARNIQA and QualiCLIP+ on KonIQ; many methods on LIVE Challenge), while others show clear negative association (e.g., CNNIQA and TReS on KonIQ; HyperIQA and TReS on LIVE Challenge), with slopes changing sign accordingly. A particularly distinctive case is MANIQA on KonIQ, where the association flips direction across sides: it is strongly positive above the line ($\rho \approx 0.99$, slope ≈ 0.0064) but strongly negative below the line ($\rho \approx -0.99$, slope ≈ -0.0095), matching the plot-level impression that MANIQA's outlier behavior is tightly coupled to semantic category count but in opposite ways for over- vs under-prediction.

Table 2 summarizes the overall ranking performance of different NR-IQA models on the KonIQ-10k and LIVE Challenge datasets, measured in terms of SRCC and PLCC. These results

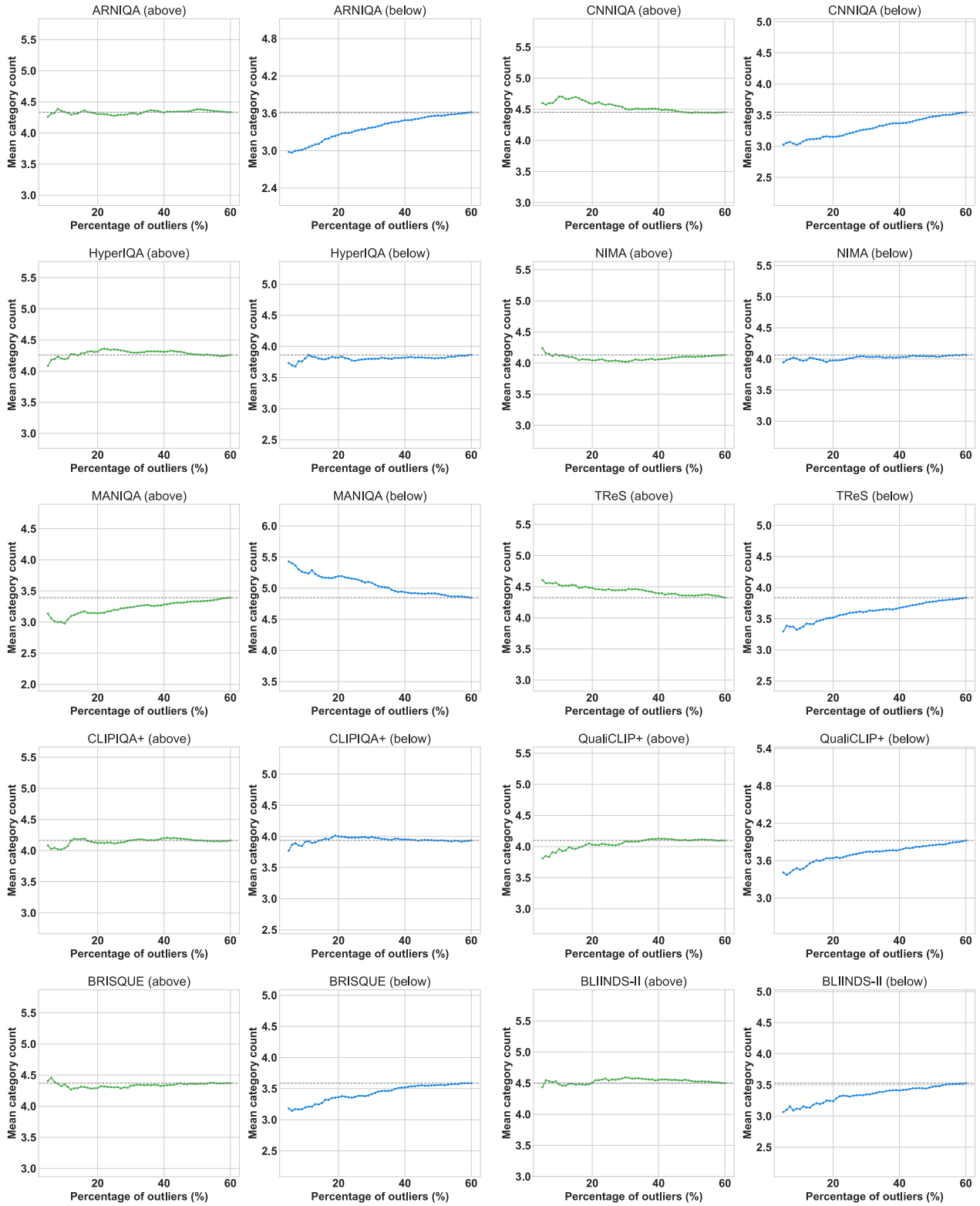


Figure 5: Mean category count plotted against percentage of outliers for the Koniq-10K dataset and using Doubao. The dotted line shows the average object count after removing 60% of the largest-error outliers. Please note that the mean category count axis is different for each plot.

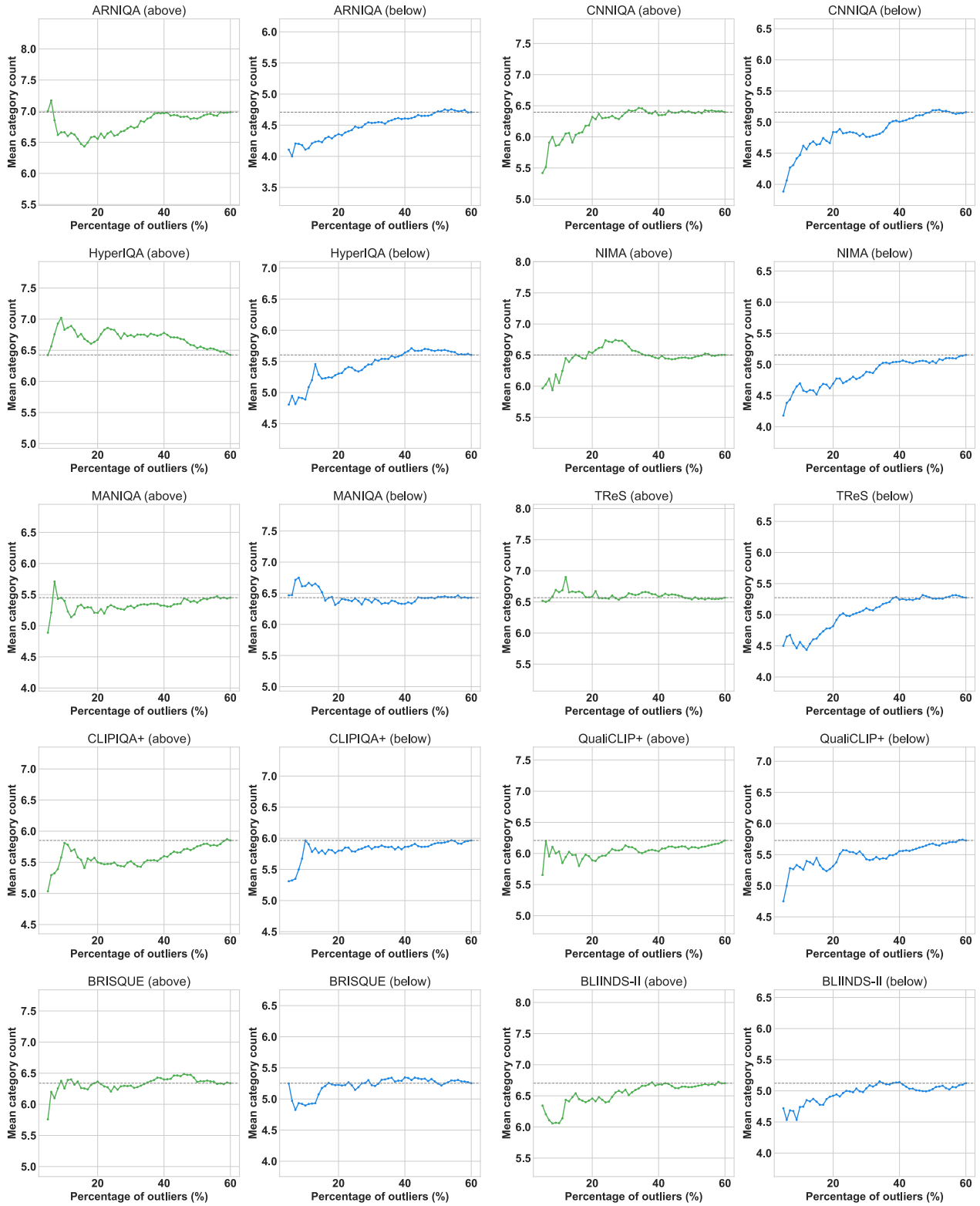


Figure 6: Mean category count plotted against percentage of outliers for the LIVE Challenge dataset and using Doubao. The dotted line shows the average object count after removing 60% of the largest-error outliers. Please note that the mean category count axis is different for each plot.

provide a global performance reference for the models analyzed in Table 1, enabling the subsequent investigation of how semantic category complexity influences model outlier behavior relative to their overall ranking accuracy.

The results for Qwenvl on the LIVE dataset gives similar results to that of Doubao. This is expected as Doubao and Qwenvl correlated in Figure 2.

Although higher semantic category counts are associated with increased outlier rates (Figure 5 and Figure 6), many high-semantic-complexity images still lie close to the regression line, implying that semantic richness alone is not sufficient to produce large errors. Based on our observations, high-complexity images that become outliers often co-occur with additional challenging visual factors (e.g., difficult illumination such as backlighting/low-light, strong local artifacts, or atypical structural degradations), whereas high-complexity images with more balanced illumination and less dominant artifacts tend to remain near the regression. The broadly shared trend across multiple IQA models suggests a common sensitivity to such "semantic complexity $\times$ distortion" interactions rather than a single architecture-specific artifact; nevertheless, the model-dependent variations (e.g., MANIQA) indicate that architecture/training strategy modulates this sensitivity. Since outliers are defined relative to MOS, we interpret the divergence primarily as a model–MOS mismatch under semantically rich content, rather than MOS being directly driven by semantic category diversity.

Conclusion and Future Work

We have analyzed how image semantic complexity, here by number of categories, in an image influence the performance of image quality metrics. For the Koniq-10k and LIVE Challenge datasets we have counted the number of categories, and investigated if outliers have more categories than inliers. Our analysis shows indications that for most image quality metrics, outliers have more categories than inliers. This indicates that image quality metrics struggle more with predicting the quality of complex images containing more categories.

Future work can include more semantic analysis of the categories, and how this impacts the performance.

Acknowledgments

Marius Pedersen and Sophie Triantaphillidou were supported by the project "Quality and Content: understanding the influence of content on subjective and objective image quality assessment" (grant 324663) from the Research Council of Norway.

References

[1] Marius Pedersen, Jon Yngve Hardeberg, et al. Full-reference image quality metrics: Classification and evaluation. *Foundations and Trends® in Computer Graphics and Vision*, 7(1):1–80, 2012.

[2] Jie Yang, Mengjin Lyu, Zhiquan Qi, and Yong Shi. Deep learning based image quality assessment: A survey. *Procedia Computer Science*, 221:1000–1005, 2023.

[3] 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.

[4] Pedro Latorre-Carmona, Rafael Huertas, Marius Pedersen, and Samuel Morillas. Proposal of a new fidelity measure between computed image quality and observers quality scores accounting for scores variability. *Journal of Visual Communication and Image Representation*, 90:103704, 2023.

[5] Seyed Ali Amirshahi, Marius Pedersen, and X Yu Stella. Image quality assessment by comparing cnn features between images. *Journal of Imaging Science and Technology*, 60:1–10, 2016.

[6] Rameez Wajid, Atif Bin Mansoor, and Marius Pedersen. A human perception based performance evaluation of image quality metrics. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Ryan McMahan, Jason Jerald, Hui Zhang, Steven M. Drucker, Chandra Kambhamettu, Maha El Choubassi, Zhigang Deng, and Mark Carlson, editors, *Advances in Visual Computing*, pages 303–312, Cham, 2014. Springer International Publishing.

[7] Seyed Ali Amirshahi, Marius Pedersen, and Azeddine Beghdadi. Reviving traditional image quality metrics using cnns. In *Color and Imaging Conference*, volume 26, pages 241–246. Society for Imaging Science and Technology, 2018.

[8] Marius Pedersen. Evaluation of 60 full-reference image quality metrics on the CID:IQ. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1588–1592. IEEE, 2015.

[9] Aladine Chetouani and Marius Pedersen. Image quality assessment without reference by combining deep learning-based features and viewing distance. *Applied Sciences*, 11(10):4661, 2021.

[10] A. Olivia, M. L. Mack, Shrestha M., and Peeper A. Identifying perceptual dimensions of visual complexity of scenes. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2004.

[11] Tina Samajdar and Md. Iqbal Quraishi. Analysis and evaluation of image quality metrics. In J. K. Mandal, Suresh Chandra Satapathy, Manas Kumar Sanyal, Partha Pratim Sarkar, and Anirban Mukhopadhyay, editors, *Information Systems Design and Intelligent Applications*, pages 369–378, New Delhi, 2015. Springer India.

[12] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015.

[13] Weisi Lin and C.-C. Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011.

[14] Anass Nouri, Christophe Charrier, Abdelhakim Saadane, and Christine Fernandez-Maloigne. Statistical comparison of no-reference images quality assessment algorithms. In *2013 Colour and Visual Computing Symposium (CVCS)*, pages 1–5. IEEE, 2013.

[15] Philippe Hanhart, Marco V Bernardo, Manuela Pereira, António M G. Pinheiro, and Touradj Ebrahimi. Benchmarking of objective quality metrics for hdr image quality assessment. *EURASIP Journal on Image and Video Processing*, 2015(1):39, 2015.

Model	KonIQ (Above)	KonIQ (Below)	LIVE (Above)	LIVE (Below)
ARNIQA	0.505	1.000	0.655	0.985
CNNIQA	-0.939	0.998	0.809	0.945
HyperIQA	-0.020	0.560	-0.609	0.902
NIMA	0.124	0.873	0.190	0.968
MANIQA	0.989	-0.993	0.628	-0.206
TReS	-0.963	0.997	-0.351	0.959
CLIPIQA+	0.425	-0.064	0.642	0.809
QualiCLIP+	0.841	0.998	0.720	0.912
BRISQUE	0.527	0.995	0.551	0.678
BLIINDS-II	0.207	0.996	0.893	0.754

(a) Spearman correlation (ρ) between outlier percentage and mean semantic category count.

Model	KonIQ (Above)	KonIQ (Below)	LIVE (Above)	LIVE (Below)
ARNIQA	0.00087	0.01186	0.00688	0.01227
CNNIQA	-0.00456	0.00985	0.01096	0.01613
HyperIQA	0.00055	0.00125	-0.00502	0.01380
NIMA	-0.00001	0.00157	0.00401	0.01323
MANIQA	0.00639	-0.00950	0.00334	-0.00281
TReS	-0.00408	0.00909	-0.00129	0.01614
CLIPIQA+	0.00164	0.00042	0.00675	0.00597
QualiCLIP+	0.00420	0.00851	0.00419	0.00982
BRISQUE	0.00082	0.00816	0.00363	0.00574
BLIINDS-II	0.00071	0.00801	0.00951	0.00731

(b) Linear regression slope between outlier percentage and mean semantic category count.

Table 1: Relationship between outlier percentage and semantic category count across different NR-IQA models on KonIQ-10k and LIVE datasets.

Model	KonIQ SRCC	KonIQ PLCC	LIVE SRCC	LIVE PLCC
ARNIQA	0.798	0.836	0.676	0.733
CNNIQA	0.776	0.820	0.609	0.640
HyperIQA	0.947	0.956	0.750	0.793
NIMA	0.951	0.964	0.793	0.833
MANIQA	0.905	0.933	0.832	0.849
TReS	0.935	0.948	0.771	0.807
CLIPIQA+	0.873	0.889	0.773	0.779
QualiCLIP+	0.872	0.890	0.805	0.831
BRISQUE	0.764	0.767	0.164	0.193
BLIINDS-II	0.651	0.657	0.247	0.254

Table 2: Overall performance (SRCC and PLCC) of different NR-IQA models on KonIQ-10k and LIVE Challenge datasets. PLCC is computed after standard non-linear mapping.

[16] Aladine Chetouani. Full reference image quality assessment: Limitation. In *2014 22nd International Conference on Pattern Recognition*, pages 833–837. IEEE, 2014.

[17] Christophe Charrier, Olivier Lézoray, and Gilles Lebrun. A machine learning regression scheme to design a fr-image quality assessment algorithm. In *European Conference on Colour in Graphics, Imaging, and Vision*, pages 35–42, 2012.

[18] Marius Pedersen and Jon Yngve Hardeberg. A new spatial filtering based image difference metric based on hue angle weighting. *Journal of Imaging Science and Technology*, 56(5):50501–1, 2012.

[19] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Hybrid no-reference quality metric for singly and multiply distorted images. *IEEE Transactions on Broadcasting*, 60(3):555–567, 2014.

[20] Huan Yang, Yuming Fang, and Weisi Lin. Perceptual quality assessment of screen content images. *IEEE Transactions on Image Processing*, 24(11):4408–4421, 2015.

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

[22] Xinbo Gao, Tao Wang, and Jie Li. A content-based image quality metric. In Dominik Ślezak, JingTao Yao, James F. Peters, Wojciech Ziarko, and Xiaohua Hu, editors, *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, pages 231–240, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[23] ZM Parvez Sazzad, Yoshikazu Kawayoke, and Yuukou Horita. No reference image quality assessment for jpeg2000 based on spatial features. *Signal Processing: Image Communication*, 23(4):257–268, 2008.

[24] Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group mad competition-a new methodology to compare objective image quality models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2016.

[25] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

[26] Marius Pedersen and Olga Cherepkova. Content-based image quality assessment. *International Journal of Imaging and Robotics*, 20(3), 2020.

[27] Sophie Triantaphillidou, Elizabeth Allen, and R Jacobson. Image quality comparison between jpeg and jpeg2000. ii. scene dependency, scene analysis, and classification. *Journal of Imaging Science and Technology*, 51(3):259–270, 2007.

[28] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 189–198, 2024.

[29] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014.

[30] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3667–3676, 2020.

[31] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022.

[32] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1220–1230, 2022.

[33] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2555–2563, 2023.

[34] Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. Quality-aware image-text alignment for opinion-unaware image quality assessment. *arXiv preprint arXiv:2403.11176*, 2024.

[35] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE transactions on image processing*, 25(1):372–387, 2015.

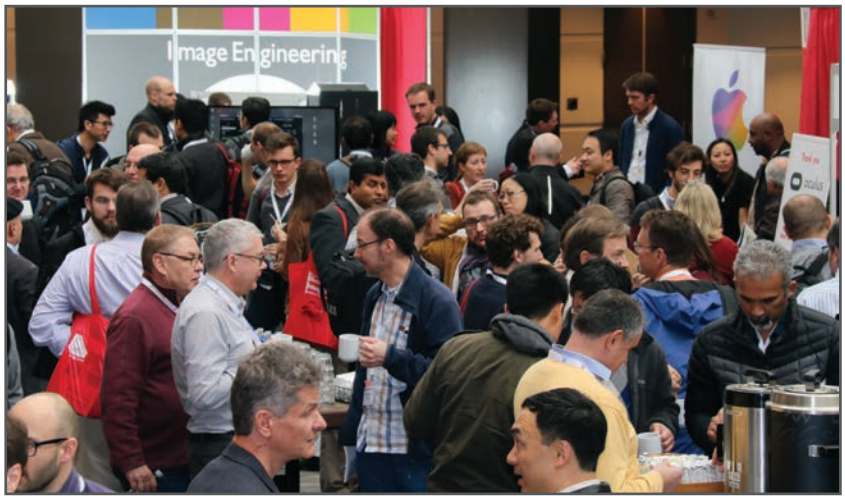
[36] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[37] ByteDance Seed Team. The doubao seed1.6. `https://seed.bytedance.com/en`, 2025. Accessed: 2025-09-12.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org