KonIQ-10k-LT: Overcoming Score Priors in Blind Image Quality Assessment Under Imbalanced Distributions

Desen Yuan, Lei Wang
University of Electronic Science and Technology of China

***Abstract*—Blind Image Quality Assessment (BIQA) is essential in computational vision for predicting the visual quality of digital images without reference counterparts. Despite advancements through convolutional neural networks (CNNs), a significant challenge in BIQA remains the long-tail distribution of image quality scores, leading to biased training and reduced model generalization. To address this, we restructured the KonIQ-10k dataset to create an imbalanced version named KonIQ-10k-LT, manipulating the distribution of image quality scores to have opposing distributions in the training and validation sets. This restructuring increases the proportion of certain quality scores in the training set while decreasing them in the validation set. Experimental results show a significant performance decline of BIQA models on the KonIQ-10k-LT dataset compared to the original KonIQ-10k, highlighting the challenge posed by the long-tail distribution. To mitigate this issue, we propose a Proportion Weighted Balancing (PWB) method as a baseline, designed to enhance the robustness and generalization ability of BIQA models. Our findings demonstrate that the proposed WB method improves the performance and reliability of BIQA models under these challenging conditions.**

***Index Terms*—Blind image quality assessment, Long-tail learning**

I. Introduction

Blind Image Quality Assessment (BIQA) [1–7] is a crucial task in computational vision, aimed at predicting the visual quality of digital images without needing their pristine counterparts. The importance of robust and accurate Image Quality Assessment (IQA) methods has grown with the proliferation of digital images on social media platforms. Full-Reference (FR) IQA methods, requiring both the reference and distorted versions of an image, are limited in real-world scenarios where reference images are unavailable. Consequently, No-Reference (NR) IQA [8, 9] methods, which assess image quality without any reference image, have gained prominence.

Despite these advancements, a significant challenge in BIQA lies in the long-tail distribution of image quality scores within most datasets. This uneven distribution means that certain quality scores, or ground truths, are underrepresented, leading to biased training outcomes and reduced model generalization. As shown in Fig. 1, addressing this issue is crucial for developing robust BIQA models that perform well across the entire spectrum of image quality levels, ensuring accurate and reliable assessment even in real-world scenarios with diverse and complex distortions.

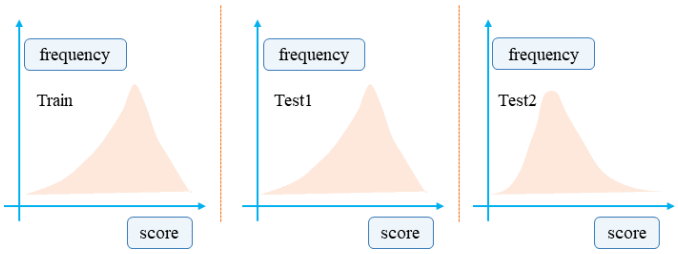

Figure 1: When the data distribution of the training set and the test set is consistent, the model often achieves good accuracy, as seen with Train and Test1. However, when the distribution of the test set significantly differs from that of the training set, it can lead to a decline in the model's performance, as observed with Test2.

To tackle the challenge of long-tail distribution in BIQA datasets, we restructured the widely used KonIQ-10k [10] dataset to create a new, imbalanced dataset named KonIQ-10k-LT. In our new dataset, we deliberately manipulated the distribution of image quality scores so that the training and validation sets have opposing distributions. Specifically, we increased the proportion of images with certain quality scores in the training set while decreasing their proportion in the validation set. Conversely, for quality scores that are underrepresented in the training set, we increased their proportion in the validation set. This approach ensures that the training set has more images with certain quality scores and fewer with others, whereas the validation set has the opposite distribution. This setup allows us to simulate real-world scenarios with distributional shifts, testing the robustness and generalization ability of our BIQA models across diverse and complex distortions.

We propose a reweighting approach as a baseline for long tail in BIQA. By leveraging the prior distribution of ground truth scores within the training set, we reweight the loss function to ensure uniform representation across different score ranges. Specifically, we divide the scores into 20 segments and assign weights based on the proportion of samples in each segment. This reweighting strategy balances the contribution of each score range during training, enhancing the model's ability to generalize across varying quality levels. Consequently, our approach improves the robustness and accuracy of BIQA

models, enabling them to effectively handle the diverse and complex distortions encountered in real-world applications.

- We propose a restructured dataset named KonIQ-10k-LT based on the widely used KonIQ-10k, designed to simulate extreme real-world distributional shifts by creating imbalanced training and validation sets with opposing quality score distributions.
- We introduce a reweighting approach for the loss function, leveraging the prior distribution of ground truth scores to balance the representation across different score ranges, thereby enhancing the robustness and generalization ability of BIQA models.

II. Related Work

Recent advances in Blind Image Quality Assessment (BIQA) and No Reference Image Quality Assessment (NR-IQA) have shifted towards new paradigms and deep learning methods. Initially, BIQA focused on Natural Scene Statistics (NSS) [11–13] and shallow learning techniques using codebooks. The advent of deep learning has revolutionized BIQA, enabling end-to-end optimization by integrating feature extraction with quality prediction. This shift has led to the exploration of adaptive convolution, self-attention mechanisms, and new objective functions to enhance metric correlation and convergence speed.

Prior to deep learning, NR-IQA methods relied on NSS to detect distortions or used machine learning for mapping features to quality scores. These traditional methods struggled with real-world distortions. The introduction of deep learning brought about methods using CNN features [1, 14–16]. However, the robustness of an image quality assessment model is crucial for its practical application. It is essential to maintain model stability when the test distribution differs from the training distribution.

III. KONIQ-LT:Dataset Creation and Analysis

The data split method of KonIQ-10k-LT results in a different overall MOS distribution in the test data compared to the training data. This split is achieved by reorganizing the KonIQ-10k dataset. To achieve the different overall MOS distribution in the KonIQ-10k-LT dataset, we reorganized the KonIQ-10k dataset through a series of systematic steps. Initially, we grouped the data into 20 bins based on MOS scores, using a step size of 5. The distribution proportions for training and testing sets within each bin were predefined. For example, bins with scores (hundred mark system) from 0-5 had a training proportion of 0.1 and a testing proportion of 0.9, while bins with scores from 95-100 had a training proportion of 0.9 and a testing proportion of 0.1.

Figure 2 shows the training and testing split results of the KonIQ-10k-LT dataset obtained through our processing. It is evident that there is a significant difference between the distributions of the training and testing sets in this dataset. This setup allows for evaluating model performance under more extreme conditions.

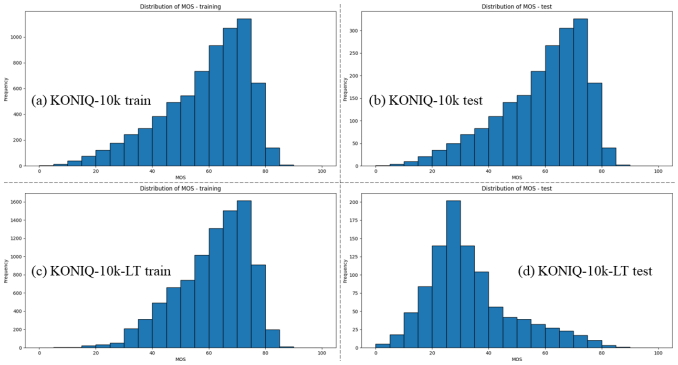

Figure 2: The visualization of the training and test set distributions for KonIQ-10k and KonIQ-10k-LT shows that the training and test sets for KonIQ-10k-LT have become more inconsistent, posing a greater challenge for the model.

IV. Benchmarking BIQA Models on KONIQ-LT

To demonstrate the challenge posed by our KonIQ-10k-LT dataset splits, we report the performance results of several mainstream BIQA models trained on the KonIQ-10k and KonIQ-10k-LT training sets and evaluated on the corresponding test sets. Additionally, we introduce a Gap metric to assess the performance differences between the different dataset splits. The results are presented in Table 1.

Table I: We compared the performance of existing IQA models on the KonIQ-10k-LT test split with their performance on the KonIQ-10k split.

Dataset	KonIQ-10k		KonIQ-10k-LT		Gap	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
CNNIQA [1]	0.793	0.752	0.708	0.659	0.085	0.091
VGG-16 [17]	0.890	0.860	0.835	0.777	0.055	0.083
ResNet-50 [18]	0.895	0.871	0.831	0.771	0.064	0.1
DBCNN [19]	0.911	0.892	0.853	0.801	0.058	0.091
NIMA [20]	0.912	0.891	0.851	0.800	0.061	0.091
TRes [21]	0.915	0.889	0.857	0.801	0.058	0.088
TOPIQ [22]	0.920	0.893	0.895	0.853	0.025	0.04
MANIQA [23]	0.936	0.920	0.896	0.845	0.04	0.075

In Table 1, we present a comparative analysis of several Image Quality Assessment (IQA) models trained and evaluated on the KonIQ-10k and KonIQ-10k-LT dataset splits. The models evaluated include CNNIQA, VGG-16, ResNet-50, DBCNN, NIMA, TRes, TOPIQ, and MANIQA.

The table shows a significant decline in performance for all models when transitioning from the KonIQ-10k to the KonIQ-10k-LT dataset, indicating the increased difficulty introduced by the KonIQ-10k-LT split. For example, CNNIQA's PLCC drops from 0.793 to 0.708, and SRCC from 0.752 to 0.659. Advanced models like ResNet-50 and DBCNN also exhibit performance drops, with ResNet-50's PLCC reducing from 0.895 to 0.831 and SRCC from 0.871 to 0.771, while DBCNN's PLCC decreases from 0.911 to 0.853 and SRCC from 0.892 to 0.801. TOPIQ shows the smallest decline, with

PLCC dropping from 0.920 to 0.895 and SRCC from 0.893 to 0.853, while MANIQA, despite having the highest scores, also sees a decline with PLCC decreasing from 0.936 to 0.896 and SRCC from 0.920 to 0.845.

These results underscore the KonIQ-10k-LT dataset's ability to create a more challenging evaluation scenario, testing the robustness and generalization of BIQA models. The Gap metric further quantifies the performance differences, highlighting the need for improved methodologies to address long-tail distribution challenges in real-world applications.

V. Methodology

We propose a weighted loss function for image quality assessment named Proportion Weighted Balancing (PWB) to address the long-tail distribution problem in the training data. Given a batch of model predictions $\mathbf{s}$ and ground truth Mean Opinion Scores (MOS) $\mathbf{t}$, the following steps are taken:

A. Proportion Calculation and Weighting

Let $\mathbf{p}$ be the vector of proportions for each score segment in the training dataset. To mitigate the effect of imbalanced data, we compute the weights as the inverse of these proportions:

$$\mathbf{w} = \frac{1}{\mathbf{p} + \epsilon} \tag{1}$$

where ϵ is a small constant added to avoid division by zero.

B. Score Segmentation

The score segments are defined by equally spaced bin edges. Let $\mathbf{b}$ be the vector of bin edges for the score segments, with K bins covering the range of possible scores:

$$\mathbf{b} = [b_0, b_1, \ldots, b_K] \tag{2}$$

C. Weighted Mean Squared Error (MSE) Loss

For each ground truth score t_i in the batch, we determine the corresponding bin index k_i such that:

$$k_i = \text{bucketize}(t_i, \mathbf{b}) \tag{3}$$

where bucketize is a function that maps the score t_i to its respective bin index k_i. The weight w_{k_i} for each sample is then retrieved based on its bin index.

The unweighted MSE loss for the predictions $\mathbf{s}$ and ground truth $\mathbf{t}$ is computed as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (s_i - t_i)^2 \tag{4}$$

where N is the batch size. The weighted loss is then given by:

$$\mathcal{L}_{\text{weighted}} = \frac{1}{N} \sum_{i=1}^{N} w_{k_i} (s_i - t_i)^2 \tag{5}$$

Finally, the overall loss is the mean of the weighted loss values:

$$\mathcal{L} = \mathbb{E}[\mathcal{L}_{\text{weighted}}] \tag{6}$$

This approach ensures that the loss function accounts for the imbalance in the training data, giving more emphasis to underrepresented score segments.

VI. Experimental Setup and Results

A. Experimental Settings

We evaluated various NR-IQA models and the PWB method on KonIQ-10k and KonIQ-10k-LT datasets. The performance of the NR-IQA models was assessed using two standard evaluation metrics: Spearman Rank Order Correlation Coefficient (SRCC) [24] and Pearson Linear Correlation Coefficient (PLCC) [25]. To ensure consistency, we ran multiple experiments on the models and reported the average results. For the model training setup, we used a batch size of 16 and applied a cosine decay schedule for optimization, utilizing either SGD or Adam. The learning rate was maintained as specified in the original publications of each model.

B. Performance of PWB

Table II provides a detailed performance comparison of several networks, both with and without the integration of PWB, evaluated on the KonIQ-10k and KonIQ-10k-LT datasets. The Gap metric is introduced to quantify the difference in performance between the KonIQ-10k and KonIQ-10k-LT datasets, thus highlighting the models' ability to generalize to a more challenging distribution.

ResNet-50's performance declines from 0.895 to 0.831 in PLCC and from 0.871 to 0.771 in SRCC when transitioning to KonIQ-10k-LT. With PWB (ResNet-50-PWB), the scores improve to 0.897 and 0.875 on KonIQ-10k, and 0.875 and 0.823 on KonIQ-10k-LT, showing a reduced Gap.

The DBCNN model also shows performance gains with PWB integration. Without PWB, DBCNN's scores are 0.911 and 0.892 on KonIQ-10k, and 0.853 and 0.801 on KonIQ-10k-LT. With PWB (DBCNN-PWB), the scores rise to 0.915 and 0.891 on KonIQ-10k, and 0.876 and 0.836 on KonIQ-10k-LT, with a smaller Gap.

The MANIQA model, which already performs well, improves further with PWB. Without PWB, MANIQA scores 0.936 and 0.920 on KonIQ-10k, and 0.896 and 0.845 on KonIQ-10k-LT. With PWB (MANIQA-PWB), the scores increase to 0.944 and 0.930 on KonIQ-10k, and 0.910 and 0.871 on KonIQ-10k-LT. These results clearly demonstrate that the integration of PWB significantly enhances the robustness and generalization capabilities of BIQA models, particularly in handling the challenges posed by long-tail distributions in real-world applications.

C. Ablation Study

Table III presents an ablation study on the effect of different batch sizes on the performance of the DBCNN-PWB model. The study evaluates the model on both the KonIQ-10k and KonIQ-10k-LT datasets using batch sizes of 8, 16, 24, 32, 40, and 48. Performance metrics include the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-order Correlation Coefficient (SRCC).

The results indicate that the DBCNN-PWB model maintains robust performance across various batch sizes. For instance, the highest PLCC and SRCC scores on the KonIQ-10k dataset

Table II: Performance using PWB integration across various networks.

Dataset	KonIQ-10k		KonIQ-10k-LT		Gap	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
VGG-16	0.890	0.860	0.835	0.777	0.055	0.083
VGG-16-PWB	**0.895**	**0.861**	**0.852**	**0.813**	**0.043**	**0.048**
ResNet-50	0.895	0.871	0.831	0.771	0.064	0.1
ResNet-50-PWB	**0.897**	**0.875**	**0.875**	**0.823**	**0.022**	**0.052**
DBCNN	0.911	0.892	0.853	0.801	0.058	0.091
DBCNN-PWB	**0.915**	**0.891**	**0.876**	**0.836**	**0.039**	**0.055**
MANIQA	0.936	0.920	0.896	0.845	0.04	0.075
MANIQA-PWB	**0.944**	**0.930**	**0.910**	**0.871**	**0.034**	**0.059**

are 0.915 and 0.891, respectively, achieved with a batch size of 16. On the KonIQ-10k-LT dataset, the best scores of 0.876 (PLCC) and 0.836 (SRCC) are also obtained with a batch size of 16. Although slight variations in performance are observed with different batch sizes, the overall results demonstrate that the DBCNN-PWB model is robust and maintains consistent performance across a range of batch sizes.

Table III: Ablation study on batch sizes.

Dataset	KonIQ-10k		KonIQ-10k-LT	
Method	PLCC	SRCC	PLCC	SRCC
DBCNN-PWB 8	0.911	0.883	0.872	0.832
DBCNN-PWB 16	**0.915**	**0.891**	**0.876**	**0.836**
DBCNN-PWB 24	0.913	0.888	0.873	0.831
DBCNN-PWB 32	0.911	0.887	0.871	0.828
DBCNN-PWB 40	0.911	0.886	0.867	0.828
DBCNN-PWB 48	0.909	0.885	0.866	0.827

D. Under different loss scenarios

Table IV provides a comparison of different loss function methods in the Image Quality Assessment (IQA) field, highlighting the effectiveness and versatility of the PWB. The methods evaluated include MSE, NIN [26], and PLCC losses, with and without PWB integration, using the ResNet50 model on both the KonIQ-10k and KonIQ-10k-LT datasets. Performance metrics reported are the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-order Correlation Coefficient (SRCC).

The results demonstrate that the integration of PWB consistently improves performance across all loss functions tested. For the ResNet50 model with MSE loss, the PLCC and SRCC scores improve from 0.895 and 0.871 on KonIQ-10k to 0.897 and 0.875, and from 0.831 and 0.771 on KonIQ-10k-LT to 0.875 and 0.823 with PWB. Similarly, for the NIN loss, PWB integration enhances the scores from 0.909 and 0.886 to 0.918 and 0.893 on KonIQ-10k, and maintains the scores on KonIQ-10k-LT with a slight improvement.

The most notable improvement is observed with the PLCC loss, where the ResNet50-PLCC model's scores increase from 0.910 and 0.887 to 0.921 and 0.898 on KonIQ-10k, and from 0.868 and 0.815 to 0.881 and 0.825 on KonIQ-10k-LT with PWB integration. These consistent improvements across different loss functions underline the general applicability and robustness of the PWB method in enhancing the performance of IQA models.

Table IV: Comparison with other loss function methods in the IQA field.

Dataset	KonIQ-10k		KonIQ-10k-LT	
Method	PLCC	SRCC	PLCC	SRCC
ResNet50-MSE	0.895	0.871	0.831	0.771
ResNet50-MSE-PWB	**0.897**	**0.875**	**0.875**	**0.823**
ResNet50-NIN	0.909	0.886	0.871	0.816
ResNet50-NIN-PWB	**0.918**	**0.893**	**0.875**	**0.820**
ResNet50-PLCC	0.910	0.887	0.868	0.815
ResNet50-PLCC-PWB	**0.921**	**0.898**	**0.881**	**0.825**

E. PWB improves generalization ability

To showcase the generalization capability of the proposed PWB method, we conducted a cross-dataset evaluation. We trained the ResNet50 and TOPIQ models on the KonIQ-10k dataset and then tested them on the CSIQ, LIVEC, and KADID-10k datasets. As shown in Table V, PWB effectively enhances the generalization performance of NR-IQA models.

Table V: Comparison of cross-dataset performance on public benchmarks.

Train dataset	KonIQ-10k					
Test dataset	CSIQ		LIVEC		KADID	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
ResNet50	0.744	0.716	0.790	0.756	0.553	0.511
TOPIQ	0.775	0.734	0.845	0.819	0.555	0.523
ResNet50-PWB	**0.755**	**0.736**	**0.798**	**0.763**	**0.543**	**0.562**
TOPIQ-PWB	**0.779**	**0.738**	**0.850**	**0.821**	**559**	**0.527**

VII. Conclusion

In this study, we addressed the significant challenge of long-tail distribution in Blind Image Quality Assessment (BIQA) by creating an imbalanced dataset, KonIQ-10k-LT, which features opposing distributions in the training and validation sets. Our experiments demonstrated a notable decline in BIQA model performance on KonIQ-10k-LT compared to the original KonIQ-10k, highlighting the difficulty posed by the long-tail distribution. To mitigate this issue, we proposed a Weight Balancing (WB) method, which effectively improved the robustness and generalization of BIQA models. Our findings underscore the importance of addressing distributional biases to enhance the accuracy and reliability of BIQA methods in real-world scenarios.

REFERENCES

[1] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[3] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 143–14 152.

[4] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, no. 11, pp. 1–52, 2020.

[5] D. Yuan and L. Wang, "Dual-criterion quality loss for blind image quality assessment," in *ACM Multimedia 2024*, 2024.

[6] D. Yuan, "Balancing easy and hard distortions: A multi-rate knowledge distillation strategy for blind image quality assessment," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8265–8269.

[7] L. Wang, Q. Wu, D. Yuan, K. N. Ngan, H. Li, F. Meng, and L. Xu, "Learning with noisy low-cost mos for image quality assessment via dual-bias calibration," *arXiv preprint arXiv:2311.15846*, 2023.

[8] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2017.

[9] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.

[10] 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*, vol. 29, pp. 4041–4056, 2020.

[11] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[12] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE transactions on Image Processing*, vol. 21, no. 8, 2012.

[13] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[14] H. Zeng, L. Zhang, and A. C. Bovik, "A probabilistic quality representation approach to deep blind image quality prediction," *arXiv preprint arXiv:1708.08190*, 2017.

[15] W. Wu, T. Dai, X. Huang, F. Ma, and J. Xiao, "Image augmentation with controlled diffusion for weakly-supervised semantic segmentation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6175–6179.

[16] ——, "Top-k pooling with patch contrastive learning for weakly-supervised semantic segmentation," *arXiv preprint arXiv:2310.09828*, 2023.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.

[20] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE transactions on image processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[21] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1220–1230.

[22] C. Chen, J. Mo, J. Hou, H. Wu, L. Liao, W. Sun, Q. Yan, and W. Lin, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE Transactions on Image Processing*, 2024.

[23] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. 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*, 2022, pp. 1191–1200.

[24] J. H. Zar, "Spearman rank correlation," *Encyclopedia of biostatistics*, vol. 7, 2005.

[25] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[26] D. Li, T. Jiang, and M. Jiang, "Norm-in-norm loss with faster convergence and better performance for image quality assessment," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 789–797.