Bridging the Gap Between Saliency Prediction and Image Quality Assessment

Kirillov Alexey*† Andrey Moskalenko*‡§ Dmitriy Vatolin*§
*Lomonosov Moscow State University †Yandex
‡AIRI, Moscow, Russia §MSU Institute for Artificial Intelligence
{alexey.kirillov, andrey.moskalenko, dmitriy}@graphics.cs.msu.ru

***Abstract*—Over the past few years, deep neural models have made considerable advances in image quality assessment (IQA). However, the underlying reasons for their success remain unclear due to the complex nature of deep neural networks. IQA aims to describe how the human visual system (HVS) works and to create its efficient approximations. On the other hand, Saliency Prediction task aims to emulate HVS by determining areas of visual interest. Thus, we believe that saliency plays a crucial role in human perception.**

In this work, we conduct an empirical study that reveals the relation between IQA and Saliency Prediction tasks, demonstrating that the former incorporates knowledge of the latter. Moreover, we introduce a novel SACID dataset of saliency-aware compressed images and conduct a large-scale comparison of classic and neural-based IQA methods. Supplementary code and data will be available at https://huggingface.co/datasets/alexkkir/SACID.

***Index Terms*—Image Quality Assessment, Visual Saliency Prediction, Explainable AI**

I. Introduction

Image Quality Assessment (IQA) aims to measure image quality aligned with human visual perception. Improving IQA can greatly enhance user experience in tasks such as image compression, restoration, editing, and generation.

Despite achieving high correlations with human judgments, the internal workings of IQA models remain unclear. An important open question is whether IQA models implicitly adopt properties of human vision, particularly visual saliency—the human tendency to focus attention on certain image regions [1, 2]. Meanwhile, Saliency Prediction (SP) has advanced significantly, delivering accurate predictions of human attention.

In this work, we explore the connection of IQA and SP. Our main contributions are as follows:

- We propose a methodology to extract saliency maps from trained IQA models, which reveals that learning-based IQA methods incorporate saliency in their predictions and can even outperform saliency prediction baselines, such as center-prior.
- We present a method for parameter-free dual-task training strategy for IQA and Saliency Prediction, which reveals that these tasks are connected and can be solved simultaneously without quality drops.
- We conduct a subjective study with 1400+ assessors to evaluate the effectiveness of existing IQA metrics for non-uniform compressed content employing various saliency-aware coding strategies.

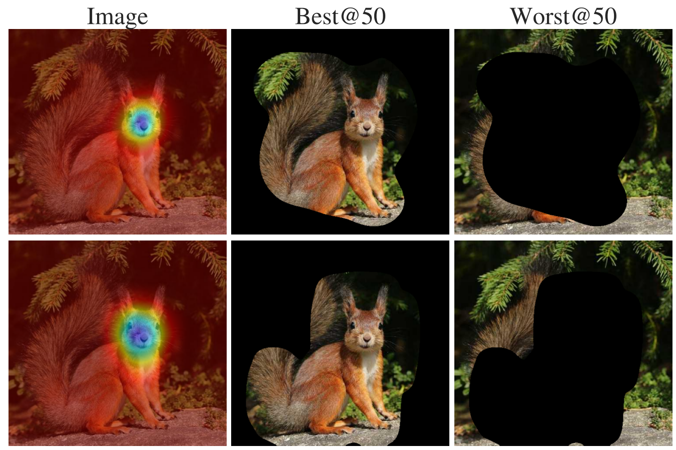

Fig. 1: Saliency and GradCAM identify important regions. Upper row: saliency maps predicted by SOTA model TranSalNet [3]. Bottom row: GradCAM extracted from our IQA model (Baseline-EfficientB0).

II. Related Work

A. Image Quality Assessment

IQA aims to measure the perceptual quality of an image by assessing image artifacts and distortions. Most IQA techniques can be categorized as Full-Reference (FR) or No-Reference (NR), with the latter more relevant for real-world applications [4], where a reference is not available.

Early approaches relied on natural scene statistics (NSS) [5, 6, 7, 8, 9] with hand-crafted features and regression models. As deep learning evolved, neural network-based approaches became popular, typically consisting of a convolutional or transformer backbone with a regression head. HyperIQA [10] proposed a multiscale feature, while MUSIQ [11] employed multiscale inputs and a vision transformer backbone.

Some methods [12, 13] create paired models for both FR and NR scenarios, such as TOPIQ, which uses features from original, distorted, and difference images at each layer.

B. Saliency in Image Quality Assessment

Research indicates a connection between saliency and IQA [12, 14, 15, 16, 17]. Efforts to enhance IQA models by incorporating saliency include [15, 18, 19, 20]. Early studies [9] improved NSS-based metrics by reweighting error maps

TABLE I: Performance comparison on SALICON [30] dataset. GradCAM-extracted saliency maps from IQA models outperform center-prior baseline, but are surpassed by SOTA saliency prediction methods. The best results are **bold**, the second-best are <u>underlined</u>, and the third best are *italics*.

Type	Method	NSS ↑	SIM ↑	CC ↑	KLD ↓
Dummy	Center Prior	0,582	0,534	0,541	0,784
GradCAM from IQA	Baseline-EfficientNet	0,604	*0,584*	*0,637*	*0,700*
	Baseline-ResNet50	**0,621**	0,555	0,590	0,782
	TOPIQ [12]	0,597	0,552	0,572	0,759
	DBCNN [31]	0,579	0,566	0,590	0,707
	CLIP-IQA+ [32]	0,584	0,550	0,589	0,741
SOTA SP	MSINet [33]	*0,612*	<u>0,767</u>	<u>0,891</u>	**0,252**
	TranSalNet [3]	<u>0,613</u>	**0,776**	**0,903**	<u>0,258</u>

based on saliency. SGDNet [18] introduced a separate head to predict saliency and reweight features.

TranSLA [21] incorporates saliency as a query branch in cross-attention, while HVS-5M [19] uses a pre-trained SP model to reweight features. SCVC [22] aggregates patch scores using Gaussian functions and saliency maps (SM). LPIPS [23] performance can be enhanced by spatially reweighting feature maps [12]. However, some studies [14] report only slight improvement in VQA models from using saliency, and our research suggests models already incorporate saliency knowledge into their predictions.

C. AI Explainability

Neural networks, despite their high performance, remain a black box. Methods have been developed to provide insight into their internal operations. CAM [24] considers a simple case of a CNN with a head consisting of a single linear layer after global average pooling (GAP). GradCAM [25] extends CAM to models with arbitrary heads, using gradients w.r.t. the model's predictions. Subsequent works [26, 27, 28] propose heuristics for improving the method.

Some researchers have studied benchmarking explanation maps. In [26], they suggest masking important image areas and monitoring changes in model confidence. Work [29] introduced a metric, area over the perturbation curve, as a measure of explanation-map fidelity.

Several works address explainability in IQA. In [13], authors showed that IQA models only require half of an image to make accurate predictions. They divided each image into 12 square patches and examined predictions of the transformer-based IQA model as they masked various combinations. They found that using important regions preserves model quality, while using trivial regions decreases it. We simplify this approach and show that masking pixels based on saliency maps or GradCAM yields comparable results.

III. Experiments

A. Extracting Saliency from IQA models

We started our experiments by testing the hypothesis that ground-truth saliency (e.g. from eye-tracker) correlates with explanation GradCAM maps of IQA models, in other words, if it is possible to extract an approximation for saliency from a trained IQA model. In our experiments we used TOPIQ [12], DBCNN [31] and CLIP-IQA [32]. We also trained our simple baseline model, consisting of a backbone, GAP pooling, and MLP head. We tested two backbones: EfficientNet-B0 [34] and ResNet-50 [35]. To build GradCAMs we used HiResCAM [27] method, known to provably reflect the locations the model used for computation. Additionally, to improve the quality of maps and remove noise, we applied smoothing through augmentations and SVD decompositions of feature maps as recommended in [36]. Every image was passed six times through the model with small rotations. Then feature map from the last channel was channel-wise decomposed via SVD-decomposition and the first main component was taken.

We used straightforward saliency baselines – Center Prior, a Gaussian distribution that approximates the saliency averaged over the dataset – and two State-of-the-Art saliency prediction models [33, 3]. We evaluate the model quality based on common saliency metrics – NSS, SIM, CC, and KLD calculated on the validation split of SALICON dataset [30]. Before calculating metrics, we employed map transforms [37] as a post-processing transformation.

Results are presented in Tab. I. GradCAM better predicts saliency than center-prior, indicating that IQA models understand saliency distribution and allow extracting such proxy SM in a zero-shot mode.

B. Saliency Masking

Inspired by [13], we investigate the significance of individual image regions on image quality by masking them. The primary aim was to determine which interpretation maps best represent important areas of the image: saliency or GradCAM. For each image, the corresponding explanation map was used, with high values indicating important areas. Subsequently, all pixels in the image whose values did not exceed the threshold were masked.

Masking of different image portions was performed using two strategies: Most Relevant First (MoRF) and Least Relevant First (LeRF), as proposed in [29]. In MoRF, pixels with values above the threshold were masked, preserving trivial regions. Conversely, in LeRF, pixels with values below the threshold were masked, preserving important regions. Perturbations, such as filling with black color or ImageNet mean values, were applied. Before thresholding the explanation maps, Gaussian blur with a large kernel (approximately 101 pixels) and a small sigma was applied to ensure the calculation of thresholds corresponding to all quantiles. Fig. 1 shows examples of images whose regions are filled with black. For the experiment, images from the KonIQ-10k dataset were used. The performance of IQA models on masked images was calculated in terms of Spearman's Rank Order Correlation Coefficient (SROCC).

Fig. 2 presents the results. A comparison was also made to the method from [13], referred to as "HalfImage".

Masking important areas (MoRF) decreases image quality more than masking trivial areas (LeRF). The same observation

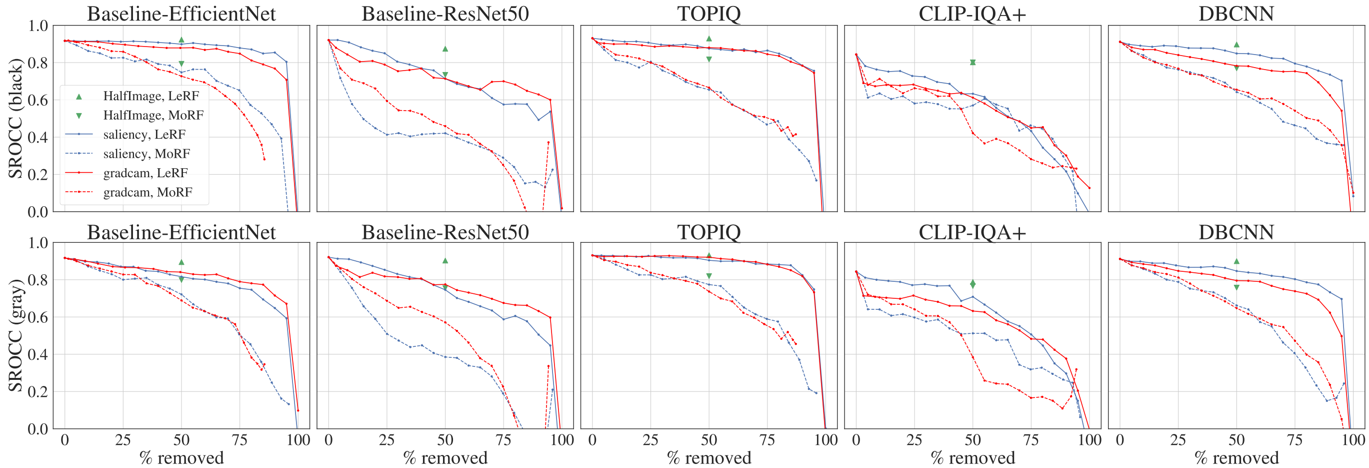

Fig. 2: Masking input images according to saliency and GradCAM maps with black (upper row) and ImageNet mean (bottom row). We note a strong relation between the behavior of correlations when masking with saliency and GradCAM maps.

holds for GradCAM. Interestingly, masking with saliency results in a larger performance gap between LeRF and MoRF regions compared to HalfImage, even though the latter employs knowledge of pretrained IQA models. This finding suggests that saliency maps effectively highlight regions that are crucial for IQA models. Furthermore, it is noted that the behavior of models when masked with GradCAM maps and saliency exhibits significant similarity. This similarity can be interpreted as an existing relationship between the attention of IQA models and human visual attention.

C. Dual-Task Training

After discovering a close connection between IQA and saliency prediction (SP), we decided to train a model on both tasks simultaneously and determine whether we could do so without decreasing performance. We implemented two approaches to achieve this goal.

In the first approach, termed Baseline-Sal-Loss, we added a small decoder to our baseline model. It consisted of a single convolutional layer with a 1×1 kernel, followed by a sigmoid activation function to integrate saliency loss into the model.

The second approach, called Baseline-GradCAM-Loss, utilizes the GradCAM method as a secondary output, which we incorporated into the model's loss function to enhance saliency prediction. The GradCAM calculation is as follows:

$$L^c_{\text{GradCAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right), \quad \alpha^c_k = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A^k_{ij}}$$

where y^c is model prediction and A^k is k-th channel of features from last layer. As IQA model has only one output, c equals 1. To ensure differentiability, we detached α^c_l from the computational graph, calculating gradients exclusively over the channel maps. This modification is integral to the Baseline-GradCAM-Loss model.

We used 80% of the KonIQ-10k dataset for training and the remaining 20% for testing. Additionally, we evaluated our model on the CLIVE, SALICON, and CAT2000 datasets. Since KonIQ-10k does not provide saliency maps, we generated proxy SM using the MSINet model. Each experiment was repeated 10 times to ensure consistency, with the averaged results presented in Table 1.

We compared the dual-task models with the baseline model trained solely on IQA using the saliency extraction technique, Center Prior, and other State-of-the-Art saliency models. Our findings reveal that both dual-task approaches allow for effective training on IQA and SP tasks concurrently. These approaches maintain the performance of IQA models while significantly improving the results for SP.

D. Non-uniform-Compression Dataset

We hypothesized that the existing IQA datasets lack sufficient complexity (due to almost-uniform compression and degradation), hindering the ability to detect emerging capabilities in IQA models, utilizing saliency. Consequently, we constructed a new dataset of high-quality images compressed with nonuniform codecs called Saliency-Aware Compressed Images Dataset (SACID). Specifically, we employed the custom codec [41] and sourced 50 images from the CLIC-2021 dataset.

We generated saliency maps for each image using the TranSalNet model and applied compression with four presets: one without saliency consideration and three that incorporate saliency. For each preset, we selected three bitrates, yielding bit-per-pixel (bpp) values of 0.08, 0.12, and 0.16 – similar to those in the CLIC-2021 challenge. Our saliency-aware compression used settings of $(\texttt{saliency\_s0}, \texttt{saliency\_bitrate}) \in [(75, 80), (60, 80), (60, 40)]$, creating a diverse set of images. We generated a total of 720 nonuniformly compressed images for evaluation.

To obtain subjective scores, we conducted pairwise comparisons using over 1,400 assessors from the crowdsourcing platform `Subjectify.us`. Assessors were presented with

TABLE II: Simultaneous learning of two tasks.

Method	IQA				Saliency Prediction							
	KonIQ-10k [38]		CLIVE [39]		SALICON [30]				CAT2000 [40]			
	SROCC ↑	PLCC ↑	SROCC ↑	PLCC ↑	NSS ↑	SIM ↑	CC ↑	KLD ↓	NSS ↑	SIM ↑	CC ↑	KLD ↓
Center Prior	-	-	-	-	0.578	0.536	0.544	0.779	0.303	0.599	0.771	0.672
Baseline	0.913	0.931	**0.862**	**0.848**	0.604	0.584	0.637	0.700	0.308	0.613	0.781	0.585
Baseline-Sal-Loss	**0.914**	**0.930**	*0.845*	*0.839*	**0.618**	0.688	0.803	0.429	0.372	*0.665*	0.818	0.433
Baseline-GradCAM-Loss	*0.912*	0.931	0.852	0.840	0.615	*0.703*	*0.825*	*0.396*	**0.382**	**0.676**	0.826	**0.403**
MSINet [33]	-	-	-	-	0.612	0.767	0.891	**0.252**	*0.370*	0.664	*0.820*	*0.421*
TranSalNet [3]	-	-	-	-	*0.613*	**0.776**	**0.903**	0.258	0.369	0.670	**0.829**	0.407

3 verification and 25 random pairs of images with different picture qualities and were asked the following question: "You will be shown sequential pairs of images with different picture quality. For each pair, select the image that has the most acceptable quality for viewing, or note that the quality in this pair is almost the same".

We evaluated all models from the PYIQA toolbox [42], saliency-aware versions of PSNR and SSIM (EW-PSNR and EW-SSIM [9]), our model variants (Baseline, Baseline-Sal-Loss, and Baseline-GradCAM-Loss), and salient deep-learning models (SGDNet [18], etc.). Models were compared in terms of SROCC, PLCC and fraction of concordant pairs (FracCP). To calculate FracCP, groups corresponding to different images were considered, and in each group, the fraction of ordered pairs was counted. The resulting numbers were averaged across all groups.

Results are listed in Tab. III. Notably, salient versions of PSNR and SSIM remarkably outperformed the originals. But Baseline, Baseline-Sal-Loss, and Baseline-GradCAM-Loss models demonstrated similar performance, implying that saliency fails to enhance deep-learning metrics significantly. We attribute this to the reliance of methods on conventional datasets, where most distortions are uniform, despite being trained with saliency. Thus, IQA models are limited by current compression standards and may show lower correlations in the non-uniform compression domain.

IV. Discussion

Our experiments reveal a clear link between saliency prediction and IQA. Using GradCAM to extract saliency maps from IQA models shows they implicitly capture human attention better than simple baselines like center-prior. Our dual-task training approach demonstrates that explicitly using saliency maintains or slightly improves IQA performance without adding model complexity.

Masking experiments confirm the significance of salient image regions, as masking them significantly reduces IQA scores. However, directly incorporating saliency into neural IQA methods showed limited gains on our non-uniform compression dataset (SACID). This limitation likely stems from existing IQA datasets being dominated by uniform distortions, restricting models' ability to leverage saliency cues effectively. Therefore, building datasets with diverse, saliency-driven, and non-uniform distortions is important for further research.

TABLE III: Quantitative results on SACID.

Type	Method	SROCC ↑	PLCC ↑	FracCP ↑
NSS based	MS-SSIM [6]	0.807	0.849	0.507
	BRISQUE [8]	0.817	0.833	*0.572*
	PSNR	0.822	0.858	0.505
	SSIM [5]	0.835	*0.880*	0.502
	VIF [7]	*0.848*	0.891	0.526
	EW-SSIM [9]	0.850	0.867	0.636
	EW-PSNR [9]	**0.875**	**0.893**	**0.678**
FR	AHIQ [43]	0.789	0.817	*0.546*
	LPIPS [23]	0.818	*0.854*	0.512
	PieAPP [44]	*0.833*	0.869	0.535
	TOPIQ-FR [12]	0.836	0.848	**0.629**
	DISTS [45]	**0.871**	**0.904**	0.589
NR	MANIQA [46]	0.752	0.773	0.564
	TReS [47]	0.765	0.800	0.559
	HyperIQA [10]	0.784	0.802	0.562
	SGDNet-None [18]	0.820	0.853	0.651
	PaQ2PiQ [48]	0.848	0.858	0.619
	SGDNet-Output [18]	0.854	0.879	**0.687**
	Baseline-GradCAM-Loss	0.863	0.865	0.624
	CLIP-IQA+ [32]	0.869	*0.896*	0.598
	Baseline	0.871	0.881	0.609
	TOPIQ-NR [12]	0.876	0.886	0.629
	Baseline-Sal-Loss	*0.880*	0.886	0.630
	MUSIQ [11]	0.888	0.899	*0.662*
	DBCNN [31]	**0.901**	**0.917**	0.666

V. Conclusion

This study examined the relationship between saliency prediction and image-quality assessment. We propose a technique for the extraction of saliency maps from IQA models and empirically demonstrate their reliance on salient regions. Additionally, we propose a parameter-free dual-task learning approach without sacrificing quality. Finally, we curated a dataset of non-uniformly compressed images and performed a large-scale comparison. We conclude that saliency can remarkably improve NSS-based methods, while learning-based methods face slight improvement since they already take saliency priors into account during conventional training.

Acknowledgments

This research was supported by Russian Science Foundation under grant 24-21-00172, https://rscf.ru/en/project/24-21-00172/ and was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

References

[1] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience*, 1995.

[2] X. Gao, W. Lu, D. Tao, and X. Li, "Image quality assessment and human visual system," in *Visual Communications and Image Processing 2010*, 2010.

[3] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu, "TranSalNet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, 2022.

[4] L. Wang, "A survey on iqa," *arXiv preprint arXiv:2109.00347*, 2021.

[5] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.

[6] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2003.

[7] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *The first international workshop on video processing and quality metrics for consumer electronics*, 2005.

[8] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, 2012.

[9] H. Liu and I. Heynderickx, "Studying the added value of visual attention in objective image quality metrics based on eye movement data," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009.

[10] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. 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*, 2020.

[11] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale Image Quality Transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[12] 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," 2023.

[13] J. You, Y. Lin, and J. Korhonen, "Half of an image is enough for quality assessment," 2023.

[14] X. Wang, A. Katsenou, and D. Bull, "UGC Quality Assessment: Exploring the Impact of Saliency in Deep Feature-Based Quality Assessment," 2023.

[15] H. Liu, U. Engelke, J. Wang, P. Le Callet, and I. Heynderickx, "How does image content affect the added value of visual attention in objective image quality assessment?" *IEEE Signal Processing Letters*, 2013.

[16] Hantao Liu and I. Heynderickx, "Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data," *IEEE Transactions on Circuits and Systems for Video Technology*, 2011.

[17] L. Lin, Y. Zheng, W. Chen, C. Lan, and T. Zhao, "Saliency-Aware Spatio-Temporal Artifact Detection for Compressed Video Quality Assessment," 2023.

[18] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "SGDNet: An End-to-End Saliency-Guided Deep Neural Network for No-Reference Image Quality Assessment," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[19] A.-X. Zhang, Y.-G. Wang, W. Tang, L. Li, and S. Kwong, "HVS Revisited: A Comprehensive Video Quality Assessment Framework," 2022.

[20] R. Cai and M. Fang, "Blind image quality assessment by simulating the visual cortex," *The Visual Computer*, 2023.

[21] M. Zhu, G. Hou, X. Chen, J. Xie, H. Lu, and J. Che, "Saliency-Guided Transformer Network combined with Local Embedding for No-Reference Image Quality Assessment," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.

[22] J. Ji, K. Xiang, and X. Wang, "SCVS: blind image quality assessment based on spatial correlation and visual saliency," *The Visual Computer*, 2023.

[23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference CVPR*, 2018.

[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," 2015.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[26] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*, 2018.

[27] R. L. Draelos and L. Carin, "Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks," 2021.

[28] S.-A. Rebuffi, R. Fong, X. Ji, H. Bilen, and A. Vedaldi, "NormGrad: Finding the Pixels that Matter for Training," 2019.

[29] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, "Sanity Checks for Saliency Metrics," 2019.

[30] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[31] 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*, 2018.

[32] J. Wang, K. C. K. Chan, and C. C. Loy, "Exploring CLIP for Assessing the Look and Feel of Images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[33] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual Encoder-Decoder Network for Visual Saliency Prediction," *Neural Networks*, 2020.

[34] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019.

[35] 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.

[36] J. Gildenblat and contributors, "Pytorch library for cam methods," https://github.com/jacobgil/pytorch-grad-cam, 2021.

[37] V. Lyudvichenko, M. Erofeev, Y. Gitman, and D. Vatolin, "A semi-automatic saliency model and its application to video compression," in *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2017, pp. 403–410.

[38] 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*, 2020.

[39] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, 2015.

[40] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research. arxiv 2015," *arXiv preprint arXiv:1505.03581*, 2019.

[41] V. Lyudvichenko, M. Erofeev, Y. Gitman, and D. Vatolin, "A semi-automatic saliency model and its application to video compression," in *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2017.

[42] C. Chen, "IQA PyTorch," https://github.com/chaofengc/IQA-PyTorch, 2021.

[43] S. Lao, Y. Gong, S. Shi, S. Yang, T. Wu, J. Wang, W. Xia, and Y. Yang, "Attentions help cnns see better: Attention-based hybrid image quality assessment network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[44] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "Pieapp: Perceptual image-error assessment through pairwise preference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[45] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *arXiv:2004.07728*, 2020.

[46] 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," 2022.

[47] S. A. Golestaneh, S. Dadsetan, and K. 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*, 2022.

[48] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. C. Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *2020 IEEE/CVF Conference CVPR*, 2020.