Aswan University Journal of Science and Technology Volume 4, issue 3

A Comparative Analysis of CNN Feature Extractors and Parameter Tuning with Ray Tune Search Algorithms for Image Quality Assessment

Hossam Mady, Adel Agamy, Abdelmageed Mohamed Aly, Mohamed Abdel-Nasser
Electrical Engineering Department, Faculty of Engineering, Aswan University, Aswan 81542, Egypt

Abstract
Image quality assessment (IQA) is crucial for the creation and assessment of visual intelligence systems to ensure end users receive high-quality visual content. Traditional IQA methods are frequently based on knowledge-driven, simplistic models. IQA has advanced significantly with the advent of deep learning, specifically convolutional neural networks (CNNs), which effectively model perceptual image distortions. This paper presents an extensive study on various CNN architectures as feature extractors in DISTS (Deep Image Structure and Texture Similarity) framework for IQA. Through the optimization of learnable parameters for various CNNs using various search algorithms and methods, we achieve substantial improvements in image quality assessment task. Our results show that optimized CNN-based metrics, particularly those built using VGG19 and SqueezeNet architectures, not only perform better but also outperform the CNN architectures used in the original DISTS model. These models closely match human perceptual judgments in their ability to capture and represent complex image features. This study opens the door for more accurate and user-aligned visual quality assessments by highlighting the potential of advanced deep learning techniques, especially when choosing the best CNN architecture and tuning method for particular task or application to improve the accuracy and reliability of IQA methods.

Keywords: Image Quality Assessment (IQA), DISTS Framework, Convolutional Neural Networks (CNNs), Hyperparameter Optimization.

1. Introduction

Nowadays, with the prevalence of visual intelligence products, images have become a crucial component of various applications in daily life. In our daily lives, image-based applications are everywhere. Image segmentation techniques can be utilized in medical imaging, object identification techniques can be employed in transportation hub monitoring, and image dehazing/deraining techniques can be significant in smart car autopilot [1].

Several kinds of distortions are incorporated into visual communication systems at nearly every stage, including acquisition, compression, transmission, and display. In this case, image quality assessment is required to guarantee and enhance the quality of the visual contents provided to end users [2]. Image quality assessment (IQA) can be applied to various areas like image acquisition [3], segmentation [4], image fusion [5], and medical imaging [6][7]. IQA methods fall into two categories - subjective and objective. While subjective evaluation, which involves human observers,

Corresponding author E-mail: dr.maisalem87@gmail.com

Received October 1, 2024, revised October 7, 2024, accepted October 9, 2024.

(ASWJST 2021/ printed ISSN: 2735-3087 and on-line ISSN: 2735-3095) https://journals.aswu.edu.eg/stjournal

is considered the most accurate way to judge image quality since people are ultimately the end users of most multimedia, it has some major drawbacks. Subjective tests are expensive and time-consuming, making them impractical for real-world use. They can also be affected by various external factors such as viewing distance, display type, lighting, an individual's vision and mood on a given day [8].

Therefore, there is a need to develop mathematical models that can predict the quality perception of the average human observer. These objective models would be able to replace subjective tests, which are not feasible options due to their cost and variability across test administrations and participants. The goal is to automate the assessment of image quality in a way that correlates well with human opinion [1][8].

For over 50 years, the area of full-reference IQA has been led by minimalistic, knowledge-based models containing few adjustable settings [9]. These knowledge-driven full-reference image quality assessment approaches have taken several computational models of the HVS from psychological vision science and made assumptions about the HVS's function in order to anticipate perceptual quality. However, it is challenging to guarantee the best performance by applying the HVS models to the real-world IQA problem because most of them are complex and were created in a limited and refined condition [10]. Some well-known instances involve techniques like mean squared error (MSE), structural similarity (SSIM) index [11], visual information fidelity (VIF) measure [12], most apparent distortion (MAD) calculation [13], and normalized Laplacian pyramid distance (NLPD) [14].

In recent years, the advent of deep learning has revolutionized IQA. Deep convolutional neural networks (CNNs) excel in extracting rich semantic information from high-dimensional data, making them highly effective for modeling perceptual image distortions [15][16][17]. Notably, pre-trained deep features from networks like VGG have proven valuable for perceptual quality measurement [16].

This paper presents a comparative analysis of various CNN architectures as feature extractors within the DISTS (Deep Image Structure and Texture Similarity) framework [18] for IQA. We employ different search algorithms to tune the learnable parameters, alpha and beta, for each CNN, comparing their performance to identify the best results. The optimized CNN-based metrics are evaluated on several state-of-the-art IQA databases, demonstrating the superior CNN architecture that outperforms even the original VGG16-based DISTS method in the image quality assessment task.

2. Related Work

Different types of IQA methods have been developed to evaluate the perceived quality of images. There are three classifications of objective image quality assessment: full-reference, reduced-reference, and no-reference image quality evaluation.

Full reference IQAs (FR-IQAs), which fall within the scope of this paper, evaluate the perceptual quality of a distorted image with respect to its reference image. These methods typically analyze pixel-level or feature-level differences between the reference and distorted images [17]. NR methods aim to assess image quality without requiring a reference image directly. Instead, they rely on intrinsic properties of the distorted image to estimate its quality. NR methods often leverage statistical analysis, machine learning algorithms, or image content analysis techniques to make quality predictions. Reduced-reference (RR) IQA methods operate with partial information from the reference image. These methods extract and compare specific features or characteristics from the reference and distorted images.

A straightforward and commonly used full-reference image quality metric is the mean square error

(MSE), which calculates the difference between a reference and distorted image. The Mean Squared Error (MSE) is calculated by taking the average of the squared differences between the original, undistorted image X and the tested "distorted" image Y. However, it has been observed that MSE does not align well with the perceived visual quality [19]. As a result, a diverse range of image quality metrics has been developed with the aim of better capturing the subjective assessment of image quality by humans. This multitude of metrics seeks to achieve a stronger correlation with the perceived visual quality and enhance the accuracy of image quality assessment [20].

Several full-reference IQA methods have been proposed to improve upon the limitations of the mean square error (MSE) metric, with the Structural Similarity (SSIM) index [11] emerging as a widely adopted standard in the field of image processing. The SSIM index takes into account the human visual system's sensitivity to structural information by considering three components: luminance similarity (comparing local mean luminance), contrast similarity (comparing local variances), and structural similarity (measured as local covariance). It has been recognized as a valuable metric that captures important perceptual aspects of image quality.

To combine image details at various resolutions and viewing conditions for IQA, multi-scale structural similarity index (MS-SSIM) [21] and information content weighted structural similarity index (IW-SSIM) [22] are proposed upon the foundation laid by the SSIM approach. By examining different angles of HVS the performance and speed of the FR-IQA algorithm are enhanced. However, these methods have several disadvantages. They can be computationally complex, have limitations in accurately assessing quality with specific types of distortions, lack adaptability to different content or viewing conditions, as they are often designed based on specific assumptions or models of visual perception, and may have a limited scope by focusing on specific aspects of image quality and potentially neglecting factors like color accuracy, texture preservation, or semantic relevance, which are important for comprehensive quality assessment.

Deep learning has revolutionized numerous fields of computer vision such as image segmentation [23][24], image classification [25], and object detection [26]. Following the success of CNNs in image classification tasks, deep learning has become prevalent across image processing domains. Liang et al. [27] introduced dual-path CNNs taking distorted and reference patches as input to predict quality scores. Kim and Lee [10] adopted a similar architecture comparing distorted patches to error maps. Recently, features from pre-trained CNNs have proven powerful for various vision tasks without need for retraining. Representations from networks like AlexNet [28] and GoogLeNet [29] pre-trained on ImageNet [30] achieved state-of-the-art results in classification, retrieval and other applications [31]. Motivated by this success, many FR-IQA methods now rely on deep features. Amirshahi et al. [32] measured similarity between AlexNet [28] activation maps on reference and distorted inputs. Bosse et al. [17] extracted features from VGG16 [33] patches, fusing distorted and reference vectors to predict quality. Richard Zhang et al. proposed the LPIPS method [16], which leverages the powerful representation capability of deep features to capture human perception. This method calculates a weighted mean squared error by comparing normalized activation maps of two images. LPIPS assesses image quality by directly comparing deep features extracted from a pre-trained VGG network at each local point. However, a limitation of these methods is that it does not adequately capture "visual texture," which encompasses repeated patterns that may vary in location, size, color, and orientation [34].

K. Ding et al. proposed DISTS [18], a method that leverages a modified version of the VGG16 network to comprehensively assess global structure and texture similarities, resembling the SSIM metric. It has been demonstrated through empirical evidence to be highly responsive to structural distortions and resilient against texture substitutions. However, it is worth noting that one potential

disadvantage of DISTS is that it may exhibit comparatively lower values of Spearman Rank Correlation Coefficient (SRCC) and Kendall's Rank Correlation Coefficient (KRCC) when evaluated on certain datasets, as compared to other IQA methods. While DISTS excels in capturing structural distortions and texture robustness, further research and exploration may be beneficial to enhance its performance in terms of correlation with human subjective ratings across a wider range of datasets.

Following this, K. Ding et al. also proposed a locally adaptive structure and texture similarity index for full-reference IQA (A-DISTS) [9]. However, A-DISTS may still demonstrate relatively lower values of SRCC and KRCC when assessed on specific datasets in comparison to other IQA methods. Additionally, the performance on global texture-related tasks may be slightly compromised.

3. Methodology

3.1. Unifying Structure and Texture Similarity (DISTS)

In 2020, K. Ding et al. proposed DISTS IQA method [18], which is the first full-reference IQA method insensitive to the resampling of visual textures. DISTS consists of two main components: a feature extraction backbone and the computation of image structure and texture similarity to determine the quality score. The feature extraction backbone, trained using VGG16, extracts feature maps at different layers such as conv1_2, conv2_2, conv3_3, conv4_3, and conv5_3. These feature maps capture various levels of information, including structure and texture details. The second component involves a weighted sum calculation.

DISTS is trained on the KADID-10k [35] dataset and exhibits improved performance in terms of PLCC and KRCC compared to conventional image quality datasets, acting as a baseline for SSIM [11].

DISTS demonstrates its ability to utilize a pre-trained and fixed VGG16 backbone to extract features from both the reference image and the distorted image, serving as input for IQA. The equation provided below presents the quality score for distorted images by combining the weighted summation of structural similarity and texture structure at various levels.

$$DISTS(x,y) = 1 - \sum_{i=0}^{m}\sum_{j=1}^{n_i}(\alpha_{ij} l(\check{x}_j^{(i)}, \check{y}_j^{(i)}) + \beta_{ij} s(\check{x}_j^{(i)}, \check{y}_j^{(i)})) \quad (1)$$

where $\{\alpha_{ij}, \beta_{ij}\}$ are positive learned values that are constrained such that,

$$\sum_{i=0}^{n}\sum_{j=1}^{ni}(\alpha_{ij} + \beta_{ij}) = 1 \quad (2)$$

These weights are tuned during training to align with human perception of picture quality.

The complete computational diagram of the original DISTS is illustrated in Figure 1. Six stages are included (raw pixels as the zeroth stage), and there are 3, 64, 128, 256, 512, and 512 feature maps in total at each level. At every step, measurements of the global texture and structural similarity are taken and mixed with a weighted summation.

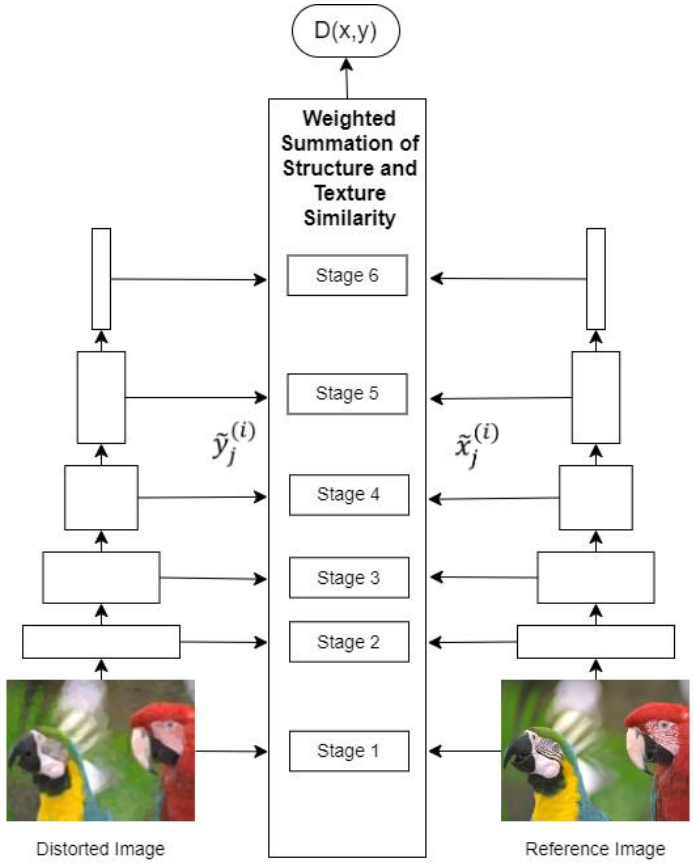

Figure 1. Perceptual representation for the original DISTS model.

3.2. Tune Search Algorithms

To optimize the performance of the models, we added a parameter tuning process. We used Ray Tune [[36], a potent framework for hyperparameter optimization that provides an extensive framework for hyperparameter tuning, to adjust the alpha and beta values in equation 1. Within the Ray Tune framework, we investigated various search algorithms, such as Random Search, Blend Search, Ax Search, Nevergrad Search, Bayesian Optimization (BO) Search, and Bayesian Opt/HyperBand (BOHB) search. We were able to locate the ideal values for alpha and beta by methodically and effectively searching the hyperparameter space with the help of these search algorithms.

3.2.1. Random Search

Random search [37] is a simple optimization algorithm that conducts a random exploration of the search space of a problem by sampling points and evaluating the objective function values. As such, it represents the 'default' and most basic way to do hyperparameter search, independent of gradients and any prior knowledge of the problem. The algorithm randomly generates a set of candidate solutions and then evaluates them in order to determine the performance. It repeats the process a number of times equal to a predefined number of iterations, or until a specified termination condition is reached. In this way, random search tries to achieve the optimum solution by luck.

3.2.2. Blend Search

Blend Search is a hyperparameter optimization algorithm developed by the flaml module [38] of the Ray Tune library. The essence of Blend Search is to combine several kinds of searches including Grid Search, Random Search, and even Adaptive Search. In this way, Blend Search tries to provide

a balance between exploration and exploitation so that fast convergence for the optimization results can be achieved.

3.2.3. Ax Search

Ax is a library for tuning hyperparameters: it performs Bayesian optimization. That enables the automation of the tuning of hyperparameters to get the best possible values and attain optimization for some objective function, such as model accuracy. In Ax, tunable hyperparameters are defined, and at the same time, their search space ranges are specified. It then proceeds to use Bayesian optimization in suggesting intelligent configurations to test, balancing exploration and exploitation. Ax intelligently explores the hyperparameter space by making suggestions from the result in previous iterations, techniques which are currently being developed using Gaussian processes and Bayesian statistics.

3.2.4. Bayesian Optimization (BO) search

The Bayesian Optimization (BO) search is a method used for hyperparameter tuning. It uses a probabilistic model of the objective function to guide the search towards promising regions of the hyperparameter space. By iteratively evaluating different hyperparameter configurations and updating the model based on the observed performance, BO gradually learns which configurations are likely to yield better results. It balances exploration and exploitation using an acquisition function, which determines the next configuration to evaluate.

3.2.5. Bayesian Opt/HyperBand (BOHB) search

The Ray Tune Bayesian Opt/HyperBand search algorithm [39] is actually a powerful search algorithm that combines the power of Bayesian Optimization with that of HyperBand. This links the two stages—exploration and then exploitation. During the exploration stage, it utilizes the HyperBand algorithm to evaluate the configurations of poor-performing sets of hyperparameters, discarding them while devoting more resources to the most promising ones. It then uses Bayesian Optimization in the exploitation phase, building a probabilistic model of the objective function that guides the search towards more promising areas.

3.2.6. Nevergrad Search

Nevergrad [40] is an optimization library from Facebook AI Research, offering various optimization algorithms. For hyperparameter tuning with Ray Tune, Nevergrad Search provides a search algorithm choice that allows you to leverage a wide variety of optimization algorithms offered by Nevergrad including but not limited to Differential Evolution and Particle Swarm Optimization to find the best set of hyperparameters for your machine learning model.

3.3 Machine Characteristics

A computer with an Intel(R) core(TM) i7-9700K processor (3.60 GHz) and 32 GB of RAM was used for this experiment. Using PyTorch version 2.2.0, we ran all computations on an NVIDIA GeForce GTX 1650 GPU, offloading all calculations to the CUDA-enabled GPU to improve performance and shorten runtime. CUDA version 12.1 was installed on our Windows 10 64-bit system in order to facilitate GPU acceleration.

4. Evaluation

4.1. IQA Datasets

Over the past 15 years, numerous IQA (Image Quality Assessment) databases have been developed, but there is currently no universally accepted standard dataset. These IQA datasets employ diverse subjective testing methodologies, varying numbers of images, and different types of distortions.

To facilitate our research and enable result comparisons, we chose the KADID-10k [35] database, which is highly regarded within the research community. Additionally, we selected TID2013 [41], LIVE [42], and CSIQ [13] datasets, as they were also chosen in the original DISTS paper.

4.2. Evaluation Criteria

To assess the objective performance of IQA (Image Quality Assessment), the correlation between Mean Opinion Scores (MOS) and the objective IQA quality score is widely used. This correlation is typically measured using three metrics: PLCC, SRCC, and KRCC.

SRCC and KRCC evaluate the monotonicity of the predicted quality score, while PLCC measures the linearity of the predicted quality score. Higher values for these correlation metrics indicate better IQA prediction performance.

The Pearson linear correlation coefficient (PLCC) specifically requires the produced scores to exhibit linearity in relation to subjective ratings. Considering the non-linear relationship between IQM scores and human assessors' scores [43][44], we have decided not to utilize the Pearson linear correlation coefficient (PLCC) for our evaluation.

4.3. Results

In our study, we conducted separate experiments using different Convolutional Neural Networks (CNNs) as the feature extraction backbone. Specifically, we utilized popular models such as VGG19 [33], ResNet50 [45], Resnet101 [45], AlexNet [28], SqueezeNet [46], and Xception [47]. These CNN architectures are known for their ability to capture intricate visual patterns and hierarchies of features from input images.

Specifically, we divide the convolutional sections of each network into five parts to obtain feature maps. In addition, we use the input image itself as another feature map, resulting in a total of six feature maps per image. The inclusion of the input image as a feature map was motivated by the desired property of the transformation, which should be injective [18]. To compare the feature maps of the reference and distorted image, we globally applied two similarity functions: the texture similarity function $l(\cdot)$ and the structure similarity function $s(\cdot)$. By leveraging $l(\cdot)$ and $s(\cdot)$, we were able to assess the perceptual quality of the images based on their texture and structural characteristics. Finally, we aggregated the results from the similarity functions to calculate a final score. l and s represent texture similarity and structure similarity respectively and are equal:

$$l = \frac{2\mu_{\tilde{x}j}^{(i)}\mu_{\tilde{y}j}^{(i)} + c_1}{\left(\mu_{\tilde{x}j}^{(i)}\right)^2 + \left(\mu_{\tilde{y}j}^{(i)}\right)^2 + c_1} \tag{3}$$

$$s = \frac{2\sigma_{\tilde{x}j}^{(i)}\sigma_{\tilde{y}j}^{(i)} + c_2}{\left(\sigma_{\tilde{x}j}^{(i)}\right)^2 + \left(\sigma_{\tilde{y}j}^{(i)}\right)^2 + c_2} \tag{4}$$

Where μ_x and σ_x are the mean and standard deviation of the image x respectively and σ_{xy} is the covariance of the x and y images.

We have done extensive hyperparameter optimization using various datasets such as TID2013, KADID10K, LIVE, and CSIQ. Several search algorithms have been run to widely explore the hyperparameter space and retain the best configurations for every dataset-CNN combination. Various algorithms tested were: Random Search, Blend Search, Bayesian Optimization, HyperBand (BOHB), Nevergrad Search, and Ax Search. Ray tune was used to test about 50 hyperparameter samples for each optimization run. We tried to find the optimal values of alpha and beta weight factors and best settings for individual CNN architectures on various image quality

datasets by comprehensively evaluating a large number of parameter settings.
Tables 1-6 report the Spearman Rank Correlation Cofficient (SRCC) and Kendall Rank Correlation Cofficient (KRCC) values achieved using optimized hyperparameters obtained with an extensive search using various algorithms. In Table 1, we list the best achieved SRCC and KRCC for each combination of CNN with dataset for the Random Search. Similarly, for Ax Search in Table 2 and in Table 3 for Blend Search and in Table 4 for BayesOptSearch. In Table 5, the results are reported that are obtained also by using Bayesian Optimization together with HyperBand, and this is known as TuneBOHB. Finally, Table 6 shows the results for Nevergrad Search in each case.

Table 1: SRCC and KRCC using Random Search

CNN Model	LIVE [42]		CSIQ [13]		TID2013 [41]		KADID10K [35]	
	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
VGG19 [33]	0.954	0.813	0.947	0.801	0.856	0.666	0.899	0.723
SqueezeNet [46]	0.947	0.804	0.917	0.742	0.866	0.680	0.889	0.708
AlexNet [28]	0.944	0.796	0.898	0.713	0.834	0.647	0.874	0.686
Xception [47]	0.662	0.488	0.463	0.316	0.426	0.294	0.259	0.171
ResNet101 [45]	0.913	0.753	0.860	0.648	0.742	0.549	0.772	0.572

Table 2: SRCC and KRCC using Ax Search

CNN Model	LIVE [42]		CSIQ [13]		TID2013 [41]		KADID10K [35]	
	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
VGG19 [33]	0.955	0.814	0.946	0.800	0.856	0.666	0.898	0.723
SqueezeNet [46]	0.947	0.804	0.917	0.742	0.868	0.690	0.885	0.698
AlexNet [28]	0.946	0.797	0.898	0.713	0.836	0.648	0.876	0.687
ResNet50 [45]	0.883	0.719	0.793	0.612	0.759	0.565	0.779	0.580
Xception [47]	0.671	0.496	0.532	0.372	0.48	0.337	0.239	0.158
ResNet101 [45]	0.929	0.766	0.869	0.660	0.745	0.551	0.773	0.572

Table 3: SRCC and KRCC using Blend Search

CNN Model	LIVE [42]		CSIQ [13]		TID2013 [41]		KADID10K [35]	
	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
VGG19 [33]	0.955	0.815	0.947	0.801	0.856	0.666	0.899	0.723
SqueezeNet [46]	0.949	0.808	0.919	0.745	0.866	0.681	0.898	0.718
AlexNet [28]	0.944	0.795	0.899	0.714	0.835	0.647	0.876	0.687
Xception [47]	**0.853**	**0.664**	**0.677**	**0.493**	0.428	0.296	**0.392**	**0.293**
ResNet101 [45]	0.915	0.758	0.862	0.649	**0.771**	**0.575**	**0.776**	**0.579**

Table 4: SRCC and KRCC using BayesOptSearch

CNN Model	LIVE [42]		CSIQ [13]		TID2013 [41]		KADID10K [35]	
	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
VGG19 [33]	0.954	0.813	0.947	0.802	**0.857**	**0.667**	0.898	0.723
SqueezeNet [46]	0.951	0.806	**0.929**	**0.763**	**0.883**	**0.699**	**0.907**	**0.732**
AlexNet [28]	0.945	0.795	**0.909**	**0.727**	**0.843**	**0.652**	**0.880**	**0.692**
ResNet50 [45]	**0.933**	**0.776**	**0.881**	**0.706**	**0.784**	**0.587**	**0.794**	**0.598**
Xception [47]	0.698	0.521	0.527	0.368	**0.566**	**0.407**	0.273	0.182
ResNet101 [45]	**0.935**	**0.779**	**0.873**	**0.669**	0.747	0.553	0.771	0.571

Table 5: SRCC and KRCC using Bayesian Optimization and HyperBand

CNN Model	LIVE [42]		CSIQ [13]		TID2013 [41]		KADID10K [35]	
	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
VGG19 [33]	0.955	0.814	0.946	0.801	0.856	0.666	0.898	0.723
SqueezeNet [46]	**0.955**	**0.816**	0.922	0.751	0.870	0.685	0.89	0.706
AlexNet [28]	**0.948**	**0.800**	0.902	0.719	0.839	0.649	0.875	0.685
ResNet50 [45]	0.915	0.758	0.803	0.619	0.760	0.567	0.779	0.58
Xception [47]	0.651	0.478	0.476	0.327	0.464	0.321	0.273	0.182
ResNet101 [45]	0.915	0.758	0.862	0.650	0.744	0.551	0.769	0.570

Table 6: SRCC and KRCC using Nevergrad Search

CNN Model	LIVE [42]		CSIQ [13]		TID2013 [41]		KADID10K [35]	
	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
VGG19 [33]	**0.956**	**0.819**	**0.948**	**0.802**	**0.857**	**0.667**	**0.900**	**0.725**
SqueezeNet [46]	0.950	0.805	0.917	0.742	0.868	0.690	0.886	0.699
AlexNet [28]	0.946	0.796	0.898	0.713	0.837	0.645	0.875	0.686
ResNet50 [45]	0.880	0.717	0.871	0.692	0.760	0.567	0.78	0.581
Xception [47]	0.656	0.482	0.493	0.34	0.44	0.304	0.273	0.182
ResNet101 [45]	0.920	0.761	0.860	0.649	0.746	0.552	0.771	0.571

Table 1 to 6 outlines that, among various search algorithms applied, the best results in hyperparameters optimization search for the VGG19 model has been outlined by the NevergradSearch algorithm. This table shows NevergradSearch is particularly good for very

complex architectures like VGG19. On the other hand, the BayesOptSearch algorithm proved to be the most suitable for adjusting the alpha and beta parameters in most models across different datasets, indicating its versatility and robustness in hyperparameter optimization.

As shown in Figure 2, the SqueezeNet-based model, despite being the most lightweight, demonstrates outstanding performance regarding the TID2013 and KADID-10k-related datasets, surpassing the other models. Moreover, on other datasets such as LIVE and CSIQ, it shows strong competitiveness with the VGG19-based model. SqueezeNet performs exceptionally well with larger datasets, as evidenced by the fact that the SqueezeNet-based model's performance values (SRCC and KRCC) are the best with the KADID-10K and TID2013 datasets. The VGG19-based model performs the best when evaluated with CSIQ and LIVE testing datasets and yields good competitiveness on the others. This is anticipated as its deeper architecture and higher complexity enable it to represent richer, more complex features of an image.

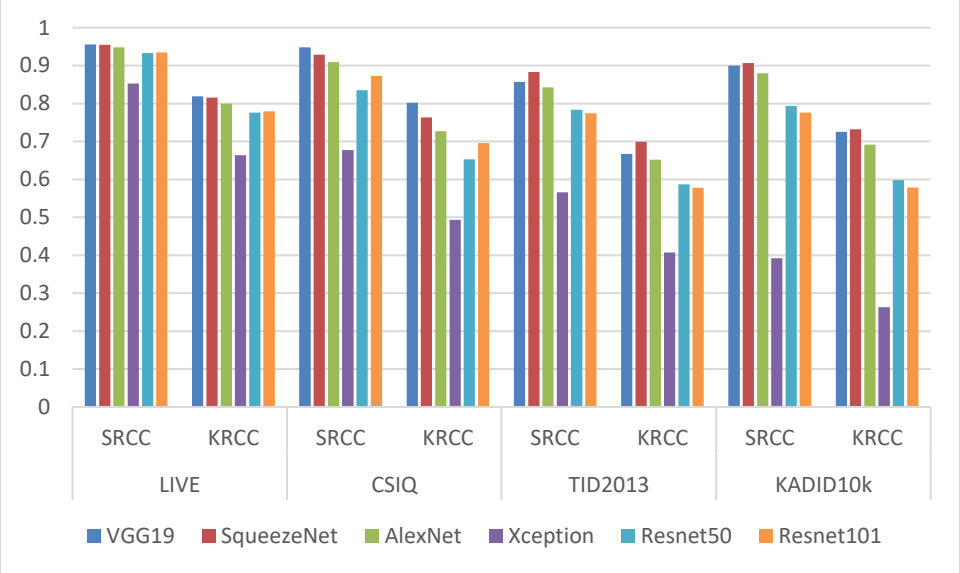

Figure 2. Comparative Analysis of CNN Models: Chart illustrating the SRCC and KRCC values across diverse datasets, highlighting the performance variations in a crucial metric.

(a) Reference image	(b) Gaussian noise	(c) Quantization noise	(d) Image denoising
MOS ↑	3.865	2.194	1.541
↓	0.241	0.326	0.493
↓	0.255	0.312	0.496
↓	0.244	0.308	0.372

VGG19

AlexNet

SqueezeNet

(a) Reference image

DMOS ↓

VGG19 ↓

AlexNet ↓

SqueezeNet

(b) Gaussian blurring

0.777

0.496

0.511

0.517

(c) Contrast decrements

0.458

0.211

0.194

0.205

(d) JPEG compression

0.816

0.462

0.471

0.466

(a) Reference image 1

DMOS ↓

VGG19 ↓

AlexNet ↓

SqueezeNet

(b) White noise

28.45

0.071

0.062

0.066

(a) Reference image 2

DMOS ↓

VGG19 ↓

AlexNet ↓

SqueezeNet

(d) Fast fading

52.171

0.285

0.277

0.270

Figure 3. Sample images from TID2013, LIVE, and CSIQ with diverse distortion types, accompanied by MOS | DMOS scores, and quality assessments from each model.

Tables 7, and 8 provide the SRCC and KRCC for different distortions seen in CSIQ and LIVE datasets. It is clear that VGG19 and SqueezeNet outperform the rest of the CNN models in various types of distortions.

Table 7: SRCC and KRCC values for different distortions seen in the LIVE dataset. In each column the highest correlation is shown by red, the second highest by a blue.

MODEL	Wn		Jp2k		Jpeg		Gblur		Fastfading	
	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
VGG19 [32]	0.978	0.867	0.964	0.832	0.977	0.867	0.973	0.863	0.940	0.796
SqueezeNet [46]	0.985	0.893	0.981	0.879	0.979	0.869	0.962	0.835	0.920	0.774
AlexNet [27]	0.980	0.874	0.955	0.814	0.973	0.854	0.969	0.843	0.915	0.761
ResNet50 [45]	0.962	0.828	0.946	0.790	0.967	0.840	0.928	0.776	0.934	0.788
Xception [47]	0.957	0.819	0.916	0.736	0.947	0.793	0.967	0.846	0.836	0.658
ResNet101 [45]	0.962	0.828	0.945	0.788	0.965	0.832	0.928	0.777	0.934	0.789

Table 8: SRCC and KRCC values for different distortions seen in the CSIQ dataset. In each column the highest correlation is shown by red, the second highest by a blue.

MODEL	AWGN		Jp2k		Fnoise		Blur		Contrast	
	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
VGG19 [32]	0.933	0.769	0.968	0.856	0.960	0.822	0.968	0.849	0.942	0.793
SqueezeNet [46]	0.950	0.795	0.972	0.862	0.952	0.802	0.972	0.857	0.935	0.776
AlexNet [27]	0.944	0.786	0.967	0.845	0.949	0.796	0.969	0.847	0.945	0.798
ResNet50 [45]	0.936	0.772	0.957	0.827	0.947	0.794	0.811	0.640	0.642	0.438
Xception [47]	0.859	0.667	0.952	0.812	0.907	0.723	0.629	0.464	0.676	0.451
ResNet101 [45]	0.922	0.750	0.951	0.815	0.947	0.791	0.791	0.618	0.675	0.463

In addition to accuracy, the computational complexity and inference speed of Convolutional Neural Networks (CNNs) play a critical role in determining their suitability for real-time applications. To compare the efficiency of each model, we evaluate the frames per second (FPS) performance of the CNNs employed in this study, including VGG19, SqueezeNet, AlexNet, ResNet50, ResNet101, and Xception. FPS is measured under identical hardware conditions using the same image quality assessment task. This metric provides an indication of how many images a model can process per second, giving an empirical assessment of computational load.

The results in Table 9 indicate significant variations in FPS across the different architectures. SqueezeNet, despite being the best model in this study, also demonstrated the highest FPS at 142, making it a strong candidate for real-time applications where both speed and accuracy are critical. VGG19, another high-performing model in terms of accuracy, achieved an FPS of 28, which reflects its more complex architecture and subsequently slower processing speed. AlexNet, ranked third in accuracy, performed efficiently in terms of speed with 120 FPS, offering a good balance between accuracy and real-time capability. On the other hand, ResNet50 and ResNet101, which performed poorly in terms of accuracy in this task, exhibited moderate to low FPS scores at 46 and 27, respectively, underscoring their inefficiency in both accuracy and speed for this particular application. Finally, Xception, which performed the worst in terms of accuracy, achieved a high FPS of 139.

Table 9: Comparison of Convolutional Neural Networks (CNNs) in terms of frames per second (FPS) performance during an image quality assessment task, reflecting the computational complexity of each model.

MODEL	Frames Per Second (FPS)
VGG19 [33]	28
SqueezeNet [46]	142
AlexNet [28]	120
ResNet50 [45]	46
Xception [47]	139
ResNet101 [45]	27

4.4. Comparison to the State of the Art

In this section, we compare the results of our top two models with state-of-the-art metrics in Image Quality Assessment (IQA). The aim is to evaluate the performance of our models and ascertain their effectiveness in the field. Table 10 displays the Spearman Rank Order Correlation Coefficient (SRCC) and Kendall Rank Correlation Coefficient (KRCC) values obtained by our models, along with those of other existing models, including the original DISTS which is based on VGG16.

Table 10: Performance comparison of our top two models against five IQA models on four standard IQA databases. Larger SRCC, and KRCC numbers represent better performance. Bold indicates the top result, and underlining signifies the second-best performance.

CNN Model	LIVE [42]		CSIQ [13]		TID2013 [41]		KADID10K [35]	
	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
DISTS [18]	0.954	0.811	0.929	0.767	0.830	0.639	0.887	0.709
PieAPP [48]	0.919	0.750	0.892	0.715	<u>0.876</u>	<u>0.683</u>	0.836	0.647
LPIPS [16]	0.932	0.765	0.876	0.689	0.670	0.497	0.843	0.653
A-DISTS [9]	0.955	0.812	<u>0.942</u>	<u>0.796</u>	0.836	0.642	0.890	0.715
SWLGV [49]	-	-	0.922	0.755	0.804	0.637	0.840	0.655
DeepDC [50]	0.940	0.781	0.937	0.774	0.844	0.651	<u>0.905</u>	<u>0.733</u>
VGG19-based (ours)	**0.956**	**0.819**	**0.948**	**0.802**	0.857	0.667	0.900	0.725
SqueezeNet (ours)	<u>0.955</u>	<u>0.816</u>	0.929	0.763	**0.883**	**0.699**	**0.907**	**0.732**

Remarkably, our two models, based on VGG19 and SqueezeNet, consistently emerged as either the best, second-best, or highly competitive performers across the evaluated metrics. The VGG19-based model showed superior performance in the CSIQ, LIVE, and KADID-10k datasets, demonstrating its ability to capture intricate image features and align closely with human perceptual judgments. The SqueezeNet-based model, despite being the most lightweight, performed exceptionally well, particularly on the TID2013 and kadid10k datasets, and showcased strong competitiveness on other datasets such as LIVE and CSIQ. This highlights SqueezeNet's efficiency and robustness in various IQA scenarios.

Overall, our comparative analysis demonstrates that the proposed models are highly effective for IQA tasks, often surpassing or matching the performance of existing state-of-the-art methods. This underscores the potential of leveraging optimized CNN architectures for developing robust and accurate image quality assessment tools.

5. Conclusion and Future Work

In this paper, we conducted a comprehensive analysis of several CNNs as feature extractors in a DISTS framework for IQA task. We illustrated that with proper optimization of learnable parameters, the proposed CNNs gained tremendous improvements in IQA performance. Our results very well support the effectiveness of deep learning methods in precisely characterizing and predicting perceptual image quality, thus outperforming classic IQA methods.

These experiments turned out quite impressive for models like VGG19 and SqueezeNet across different image quality datasets. More interestingly, SqueezeNet

happened to be the lightest among all those models; it matched the performance or even excelled in some cases, proving its efficiency and strength. These models are very good at learning fine patterns of visuals and feature hierarchies and staying close to human perceptual judgments. These models clearly show better performance, thus underlining the necessity of choosing appropriate CNN architectures and their optimum parameters for a certain IQA task.

Some promising results are obtained, but some future works remain open. One perspective is to study more CNN architectures and advanced deep learning techniques, including transformer models, in order to enhance IQA performance. Other perspectives are new search algorithms, such as metaheuristic algorithms, to optimize model parameters more efficiently, and taking into account texture similarity and tolerance of texture variation to make IQA models more robust.

It is obvious from our results that deep learning–based approaches bring a significant contribution to the field of image quality assessment. Further research in techniques and refinement will be done with a view to building more accurate and reliable models, all contributing toward higher quality in the visual content delivered to the end-user.

References

[1] L. Wang, "A survey on IQA," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2109.00347

[2] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," Nov. 01, 2020, *Science in China Press*. doi: 10.1007/s11432-019-2757-1.

[3] F. Xiao, J. E. Farrell, and B. A. Wandell, "Psychophysical thresholds and digital camera sensitivity: the thousand-photon limit," in *Digital Photography*, SPIE, Feb. 2005, p. 75. doi: 10.1117/12.587468.

[4] J. Chen, T. N. Pappas, A. Mojsilovic, and B. E. Rogowitz, "PERCEPTUALLY-TUNED MULTISCALE COLOR-TEXTURE SEGMENTATION." [Online]. Available: http://peacock.ece.utk.edu/FeatureTest/.

[5] G. Piella' and H. Heijmans, "A NEW QUALITY METRIC FOR IMAGE FUSION," in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, Barcelona, Spain, 2003. doi: 10.1109/ICIP.2003.1247209.

[6] H. H. Barrett, "Objective assessment of image quality: effects of quantum noise and object variability," *J. Opt. Soc. Am. A*, vol. 7, no. 7, 1990.

[7] H. H. Barrett, J. L. Denny, R. F. Wagner, and K. J. Myers, "Objective assessment of image quality. II. Fisher information, Fourier crosstalk, and figures of merit for task performance," *J. Opt. Soc. Am. A*, vol. 12, no. 5, 1995.

[8] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and Objective Quality Assessment of Image: A Survey," Jun. 2014, [Online]. Available: https://arxiv.org/abs/1406.7799

[9] K. DIng, Y. Liu, X. Zou, S. Wang, and K. Ma, "Locally Adaptive Structure and Texture Similarity

for Image Quality Assessment," in *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2021, pp. 2483–2491. doi: 10.1145/3474085.3475419.

[10] J. Kim and S. Lee, "Deep Learning of Human Visual Sensitivity in Image Quality Assessment Framework," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[12] H. R. Sheikh and A. C. Bovik, "A VISUAL INFORMATION FIDELITY APPROACH TO VIDEO QUALITY ASSESSMENT".

[13] D. M. Chandler and E. C. Larson, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J Electron Imaging*, vol. 19, no. 1, pp. 1–21, Jan. 2010, doi: 10.1117/1.3267105.

[14] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized Laplacian pyramid," *Electronic Imaging*, vol. 2016, no. 16, pp. 1–6, 2016.

[15] Q. Sang, Z. Shu, L. Liu, C. Hu, and Q. Wu, "Image quality assessment based on self-supervised learning and knowledge distillation," *J Vis Commun Image Represent*, vol. 90, p. 103708, Feb. 2023, doi: 10.1016/J.JVCIR.2022.103708.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric." [Online]. Available: https://www.github.com/richzhang/PerceptualSimilarity.

[17] S. Bosse, D. Maniry, K. R. Müller, T. Wiegand, and W. Samek, "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, Jan. 2018, doi: 10.1109/TIP.2017.2760518.

[18] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image Quality Assessment: Unifying Structure and Texture Similarity," Apr. 2020, doi: 10.1109/TPAMI.2020.3045810.

[19] B. Girod, "What's wrong with mean-squared error," *Digital Images and Human Vision*, pp. 207–220, 1993.

[20] W. Lin and C. C. Jay Kuo, "Perceptual visual quality metrics: A survey," *J Vis Commun Image Represent*, vol. 22, no. 4, pp. 297–312, May 2011, doi: 10.1016/J.JVCIR.2011.01.005.

[21] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "MULTI-SCALE STRUCTURAL SIMILARITY FOR IMAGE QUALITY ASSESSMENT," in *IEEE Asilomar Conference on Signals, System and Computers*, 2003, pp. 1398–1402.

[22] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, May 2011, doi: 10.1109/TIP.2010.2092435.

[23] W. M. Abd-Elhafiez, A. H. Abu Bakr, E. A. Zanaty, and M. E. Hussein, "Medical Image Segmentation Using Deep Learning: Review," *Aswan University Journal of Sciences and Technology*, vol. 3, no. 1, 2023, [Online]. Available: https://journals.aswu.edu.eg/stjournal

[24] M. E. Rayed, S. M. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. F. Mridha, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Inform Med Unlocked*, vol. 47, p. 101504, Jan. 2024, doi: 10.1016/J.IMU.2024.101504.

[25] A. H. Mohamed, M. Refaat, and A. M. Hemeida, "Image classification based deep learning: A Review," *Aswan University Journal of Sciences and Technology*, vol. 2, no. 1, pp. 11–35, Jun. 2022, doi: 10.21608/AUJST.2022.259887.

[26] K. Vaishnavi, G. P. Reddy, T. B. Reddy, N. Ch. S. Iyengar, and S. Shaik, "Real-time Object

Detection Using Deep Learning," *Journal of Advances in Mathematics and Computer Science*, vol. 38, no. 8, pp. 24–32, Jun. 2023, doi: 10.9734/JAMCS/2023/V38I81787.

[27] Y. Liang, J. Wang, X. Wan, Y. Gong, and N. Zheng, "Image quality assessment using similar scene as reference," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9909 LNCS, pp. 3–18, 2016, doi: 10.1007/978-3-319-46454-1_1/FIGURES/5.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks." [Online]. Available: http://code.google.com/p/cuda-convnet/

[29] C. Szegedy *et al.*, "Going Deeper with Convolutions," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 1–9, Sep. 2014, doi: 10.1109/CVPR.2015.7298594.

[30] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 248–255, 2009, doi: 10.1109/CVPR.2009.5206848.

[31] D. Varga, "A Combined Full-Reference Image Quality Assessment Method Based on Convolutional Activation Maps," *Algorithms 2020, Vol. 13, Page 313*, vol. 13, no. 12, p. 313, Nov. 2020, doi: 10.3390/A13120313.

[32] S. A. Amirshahi, M. Pedersen, and S. X. Yu, "Image quality assessment by comparing CNN features between images," *Journal of Imaging Science and Technology*, vol. 60, no. 6, Nov. 2016, doi: 10.2352/J.ImagingSci.Technol.2016.60.6.060410.

[33] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[34] J. Portilla and E. P. Simoncelli, "Parametric texture model based on joint statistics of complex wavelet coefficients," *Int J Comput Vis*, vol. 40, no. 1, pp. 49–71, 2000, doi: 10.1023/A:1026553619983/METRICS.

[35] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A Large-scale Artificially Distorted IQA Database," Mar. 2018, doi: 10.1109/TIP.2020.2967829.

[36] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A Research Platform for Distributed Model Selection and Training," Jul. 2018, [Online]. Available: http://arxiv.org/abs/1807.05118

[37] J. Bergstra, J. B. Ca, and Y. B. Ca, "Random Search for Hyper-Parameter Optimization Yoshua Bengio," 2012. [Online]. Available: http://scikit-learn.sourceforge.net.

[38] C. Wang, Q. Wu, M. Weimer, and E. Zhu, "FLAML: A Fast and Lightweight AutoML Library," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.04706

[39] S. Falkner, A. Klein, and F. Hutter, "BOHB: Robust and Efficient Hyperparameter Optimization at Scale," Jul. 2018, [Online]. Available: http://arxiv.org/abs/1807.01774

[40] J. 'Rapin and O. 'Teytaud, "Nevergrad - A gradient-free optimization platform,." [Online]. Available: https://GitHub.com/FacebookResearch/Nevergrad

[41] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process Image Commun*, vol. 30, pp. 57–77, Jan. 2015, doi: 10.1016/j.image.2014.10.009.

[42] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "Image and video quality assessment research at LIVE." [Online]. Available: http://live.ece.utexas.edu/research/quality/.

[43] S. Kastryulin, J. Zakirov, N. Pezzotti, and D. V. Dylov, "Image Quality Assessment for Magnetic Resonance Imaging," Mar. 2022, doi: 10.1109/ACCESS.2023.3243466.

[44] A. Mason *et al.*, "Comparison of Objective Image Quality Metrics to Expert Radiologists' Scoring of Diagnostic Quality of MR Images," *IEEE Trans Med Imaging*, vol. 39, no. 4, pp. 1064–1072, Apr. 2020, doi: 10.1109/TMI.2019.2930338.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[46] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," Feb. 2016, [Online]. Available: http://arxiv.org/abs/1602.07360

[47] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," Oct. 2016, [Online]. Available: http://arxiv.org/abs/1610.02357

[48] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual Image-Error Assessment through Pairwise Preference," Jun. 2018, [Online]. Available: http://arxiv.org/abs/1806.02067

[49] D. Varga, "Full-Reference Image Quality Assessment Based on Grünwald–Letnikov Derivative, Image Gradients, and Visual Saliency," *Electronics (Switzerland)*, vol. 11, no. 4, Feb. 2022, doi: 10.3390/electronics11040559.

[50] H. Zhu, B. Chen, L. Zhu, S. Wang, and W. Lin, "DeepDC: Deep Distance Correlation as a Perceptual Image Quality Evaluator," Nov. 2022, [Online]. Available: http://arxiv.org/abs/2211.04927.