# Deep Learning for No-Reference Image Quality Assessment

B. Padmaja, Habeeb Hussain Al Hamed, Aduri Jabili Reddy and Mannem Deepak Reddy
*Department of Computer Science & Engineering with Artificial Intelligence and Machine Learning,*
*Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana, India*

Keywords: Image Quality Assessment (IQA), KonIQ, CNN, ConvNeXt, PLCC, SRCC.

Abstract: Images are one of the fundamental modes of data storage and transfer. Images lose perceptual quality due to degradation from various sources like compression, corruption and noise with the degradation process is unknown. Modern deep convolutional neural networks (CNNs) are specialized for image processing tasks and can be used to restore an original image from its degraded copy in a process called image super resolution. Training such models needs huge amounts of multi-domain image data for better generalization. Evaluating these models in real world is even more difficult due to the lack of high resolution reference images. This project proposes and trains a CNN architecture based on ConvNeXt that assesses the quality of images by assigning a score to each image. The model achieves a score of 0.92 PLCC and 0.94 SRCC on the KonIQ test set on par the current SOTA convolutional models for IQA. The proposed model can in-turn be used to evaluate the real-world performance of deep learning models, that are trained to perform image super resolution, on images with no corresponding high-resolution reference images (blind).

## 1 INTRODUCTION

Our everyday lives generate a significant amount of image or video data in this era of rapidly advancing multimedia technology. Numerous conventional and learning-based lossy compression techniques have been proposed to lower the bandwidth and storage costs brought about by these data. However, it is challenging to quantify the distortion caused by these methods, and acquiring the Mean Opinion Score (MOS) manually is costly and presents challenges for receiving prompt feedback in a production setting. Therefore, we want to establish an accurate and efficient picture quality assessment metric that is near to subjective quality assessment and can be easily used in compression and other low-level vision activities in order to fulfill the increasing visual needs of the industry and the general public. The objective of the IQA task is to anticipate the subjective viewpoints of human viewers. The three primary kinds of IQA methods now in use are full-reference (FR), reduced-reference (RR), and no-reference (NR) methods. Although NR models, like NIQE (Gu et al., 2019), are highly adaptable in real-world settings, it is challenging to forecast raters' emotions in the absence of reference images. FR models, which are still often utilized in many visual reconstruction tasks, primarily focus on the changes in texture and structure between the distorted images and the reference image. The most well-known FR reference measures are Structural Similarity (SSIM) and Peak Signal-to-Noise Ratio (PSNR; Boose et al., 2018), which concentrate on the structural and pixel differences between two pictures, respectively. Because of their low computational complexity and outstanding performance on prior tasks, they are frequently chosen as optimization targets. Additionally, the evaluation of video quality has popularized fusion-based metrics like VMAF (Xu et al., 2021). But as deep learning technology advances—particularly with the use of GANs (Jiang et al., 2019) in image compression, restoration, and other domains—the reconstruction of images becomes more difficult to assess because it now includes noise that resembles real textures, sharper edges, and unrealistic generation artifacts. The quality of these photographs cannot be adequately evaluated by conventional metrics. In this regard, deep learning-based metrics for assessing perceptual quality (Zhao et al., 2021; Yamashita and Markov, 2020) perform better in the IQA challenge. Cheon et al. (Wang et al., 2022) suggested using a transformer (Ma et al., 2017) to deal with the bogus visuals because of the transformers' great expressive ability.

The KonIQ-10k (Gu et al., 2019) dataset's distinct features and adherence to real-world settings make it

Padmaja, B., Al Hamed, H. H., Reddy, A. J. and Reddy, M. D.
Deep Learning for No-Reference Image Quality Assessmentt.
DOI: 10.5220/0013597700004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 2*, pages 588-593
ISBN: 978-989-758-763-4

stand out as a top benchmark for assessing IQA algorithms. KonIQ-10k is an IQA dataset that includes 10,325 photos from the diversified YFCC100m (Xu et al., 2021) dataset. It differs from many other IQA datasets that frequently rely on artificially manufactured distortions to capture a wide range of content and photographic styles. Through a carefully managed crowdsourcing approach, these photographs were exposed to actual aberrations typically found in real-world circumstances, such as noise, blurring, and compression faults. By ensuring that the distortions accurately represented differences in image quality, this method improved the dataset's ecological validity.In this study, we explore the capabilities of the ConvNeXt (Boose et al., 2018) architecture for Image Quality Assessment (IQA), with a particular emphasis on applying it with the KonIQ-10k dataset. The ConvNext architecture is shown in Fig. 1. This dataset offers a strong standard for assessing IQA algorithms because of its extensive and varied collection of realistically warped photos matched with user-submitted quality ratings. We achieve this by carefully optimizing a ConvNeXt model that has already been trained for this dataset. This allows us to take use of the model's strong feature extraction capabilities and identify complex image quality patterns. We do a comprehensive assessment of our refined ConvNeXt model's predictive capacity for human perception of image quality using commonly used IQA measures, such as Spearman's Rank Correlation Coefficient (SRCC) and Linear Correlation Coefficient (LCC). We conduct a thorough comparative analysis against the most advanced IQA techniques in order to give a thorough grasp of our model's capabilities. We also examine the relative advantages and disadvantages of our ConvNeXt-based method in relation to the larger field of IQA research. Our goal is to advance the field of perception-aligned image quality evaluation by providing insightful information on the practicality and effectiveness of ConvNeXt for real-world IQA applications and encouraging more research into this intriguing architecture.his dataset offers a strong standard for assessing IQA algorithms because of its extensive and varied collection of realistically warped photos matched with user-submitted quality ratings.

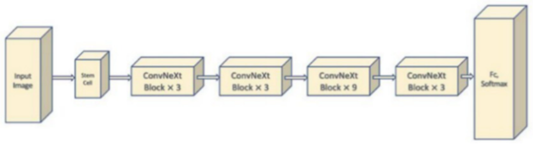

Figure 1: The overall architecture of the ConvNeXt Neural Network.

## 2 RELATED WORK

Deep-learned BIQA techniques (Gupta et al., 2020) have become a potent tool that acquires quality features directly and end-to-end from distorted images. These techniques, which use deep neural networks to automatically optimize quality forecasting models, outperform manually constructed BIQAs in terms of performance (Wang et al., 2022). There are two main categories of deep learning BIQAs: supervised learning-based and unsupervised learning-based techniques.

Annotated training data is necessary for supervised methods, but unsupervised learning-based techniques do not depend on high-quality labels during training. It is noteworthy that alternative learning modalities, like reinforcement learning (Yamashita and Markov, 2020), (Gu et al., 2021), are applicable to the deep-learned BIQA problem. In contrast to supervised and unsupervised BIQA techniques, these learning modalities are utilized less frequently. The primary aim of supervised learning-based BIQA (Harris et al., 2018) is to reduce the difference between the expected score and the human observers' subjective MOS value.

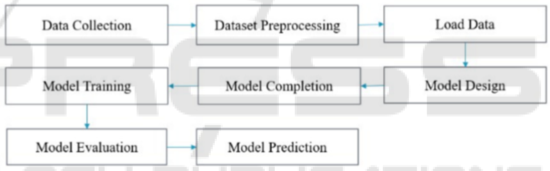

Figure 2: Process Flow of the Experiment.

The area of BIQAs has made great progress since the advent of supervised learning. By utilizing certain techniques, current supervised learning-based BIQAs address the issue of sparse training data. The primary foundation of sample-based BIQAs is increasing the capacity of training samples, which typically use patchlevel quality characteristics to forecast an image-level score. All patches inside a specific picture are immediately shared with the same image-level annotations via allocation-based BIQAs (Wang et al., 2022). The association between patch-level characteristics and the overall image-level quality scores has been originally demonstrated using these approaches. Weighted features (Bulat et al., 2018), weighted judgments (Zhang et al., 2021), and other enhancements have been made with the goal of improving this enhancements have been made with the goal of improving this connection. Weighted decisions (Zhang et al., 2021), voting decisions (Jiang et al., 2019), and weighted features (Bulat et al., 2018) have all been included in an effort to further refine this correlation. In more recent attempts, a generic feature representation has been learnt and the input picture has been expanded into multi-scale patches (Dosovitskiy et al., 2021). Nevertheless, these allocation-based techniques have difficulties capturing highly non-stationary feature representations due to the inherent uncertainty in picture content (Jiang et al., 2019). Allocation-based BIQAs can perform better and produce more accurate quality predictions by creating more resilient feature representations that can adjust to local fluctuations and complex relationships between the content and distortion. generational. Patch-level scores are typically used by generation-based BIQAs to develop a supervised model. The relationship between various modalities may be used in the future to investigate the training data volume problem. The resilience of the forecasting performance is enhanced when a deep-learned model's low-level embedding characteristics are enriched from various angles through the expansion of data modalities (Zhao et al., 2021).

# 3 METHADOLOGY

The goal of this research is to forecast picture quality using the ConvNeXt architecture without requiring an ideal reference image. No-Reference Image Quality Assessment (NR-IQA) is a crucial task in practical scenarios. Fig. 2 shows the overall process flow of this research. Attempting to evaluate pictures taken with a phone camera or ones that have been Photoshopped is difficult as there is rarely a perfect original to compare them to. Modern deep learning models such as ConvNeXt have demonstrated exceptional ability to comprehend pictures, doing well on tasks like object recognition and scene classification. Its power is in the way it is able to extract relevant information from photographs. The goal of this research is to use this power for NR-IQA. Training a ConvNeXt model to identify favorable picture characteristics in the absence of a perfect example is our aim.

## 3.1 ConvNeXt Architecture

The core of ConvNeXt's architecture is a hierarchical structure made up of several steps, each of which is in charge of extracting features at various sizes. As information moves across the layers, this hierarchical representation enables the network to gradually learn ever-more-complex patterns. ConvNeXt is mostly composed of specialized blocks that were modeled by the design of the Swin Transformer. These blocks make use of depthwise convolutions, which individually process each channel to lower computing costs while maintaining spatial information. Most importantly, ConvNeXt blocks have inverted bottleneck layers, which provide effective feature map reduction and extension, hence improving the network's capacity to collect fine features. Fig. 3 shows the ConvNeXt blocks along with the inverted bottleneck.

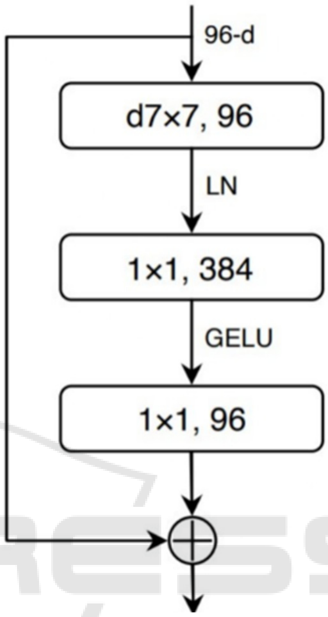

Figure 3: The ConvNeXt block built using an inverted.

ConvNeXt uses Layer Normalization, in contrast to standard convolutional networks, which frequently utilize Batch Normalization. This decision enhances training stability, especially when utilizing varied batch sizes or training data. ConvNeXt uses Global Average Pooling to minimize the spatial dimensions of the feature maps after processing the input picture via the hierarchical stages. This produces a succinct, global representation of the image information. For the ultimate prediction, this representation is subsequently given to the network's classification head—in our example, the regression head. ConvNeXt offers an effective framework for image analysis by fusing the advantages of transformer-based and convolutional architectures.

We predict that its effective architecture and capacity to extract both local and global picture characteristics will allow our model to global picture characteristics will allow our model to efficiently learn the little patterns and aberrations suggestive of high-quality images.

## 3.2 Training

We use a transfer learning strategy to efficiently train our ConvNeXt-based NR-IQA model by utilizing the insights from pre-training on the large ImageNet (Wang et al., 2022) dataset. This approach greatly speeds up the training process on the KonIQ-10K dataset and improves generalization capabilities by starting our model using weights that have already been trained to extract rich and relevant features from a wide range of pictures. The train-test split was selected to be 70% (KonIQ$^{train}$)-30% (KonIQ$^{test}$) respectively. We make use of the well-liked and effective optimization method known as the Adam (Ma et al., 2019) optimizer, which is ideal for deep learning model training. When compared to more conventional optimization techniques like stochastic gradient descent (SGD), Adam's approach allows for faster convergence and perhaps improved performance by adjusting the learning rate for each parameter during training. Early stopping (Gu et al., 2021) is one of the strategies we use to avoid overfitting. This method keeps track of the model's training results on a held-out validation set. To avoid the model from remembering the training data at the price of generalizing to unknown cases, the training process is stopped if the validation loss does not improve after a predefined period of epochs (patience). We carefully adjusted the Adam optimizer's hyperparameters, including the epsilon value and learning rate, to guarantee effective and stable training. In addition, we tested with various batch sizes to determine, considering our hardware limitations, the ideal trade-off between training time and memory usage.

The machine used throughout the whole training procedure has a single NVIDIA T4 GPU with 15GB of VRAM. We were able to investigate a larger variety of hyperparameters and model topologies in a fair amount of time because to this hardware acceleration, which dramatically shortened training times. The training data, preprocessed as stated previously, was supplied to the model in batches, and the weights of the ConvNeXt backbone and the regression head were simultaneously updated to minimize the mean squared error (MSE) between the predicted and the mean squared error (MSE) between the predicted and ground-truth picture quality scores. Throughout the training phase, we kept a close eye on both the training loss and validation loss to evaluate the model's convergence and generalizability to new data.

# 4 RESULTS AND DISCUSSION

The results of our studies are shown in this part, offering some insight into the efficacy of the suggested ConvNeXt-based NR-IQA model. Analyzing the convergence and generalization capacity of the model, we first examine the training dynamics. Next, we visually examine the learnt convolutional kernels with the goal of comprehending the intrinsic representations of picture quality within the model. Lastly, we present sample inference findings that illustrate the advantages and disadvantages of the model by contrasting its predictions with ground-truth quality ratings.

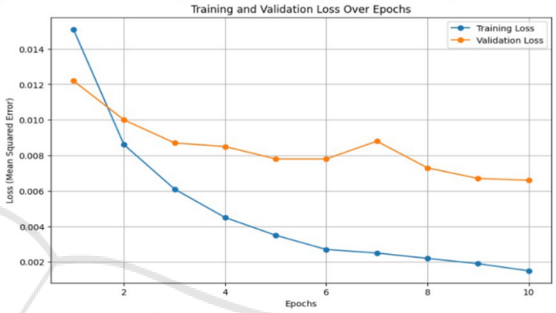

Figure 4: The training loss and validation loss curves.

## 4.1 Training Results

Fig. 4 shows the training loss and validation loss curves. The model achieved 0.0066 Mean Absolute Error loss in the prediction of image quality scores on the test-set towards the end of the training. Early stopping was used to stop at this value just before the model was beginning to over-fit. The loss on the training set reached as low as 0.0015 which is also the standard deviation of the image quality scores as predicted by the model. The Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank Correlation Coefficient (SRCC) between the test set and the model's prediction were 0.92 and 0.94 respectively.

## 4.2 Comparison with the State of Art

The results of our suggested strategy on the KonIQ-10k dataset are shown in Table 1, which shows how well it performs in comparison to state-of-the-art (SOTA) methods. Interestingly, the BIQA technique (Xu et al., 2023) uses a multi-crop ensemble approach for both training and testing, which has its own set of problems despite attempting to address the shortcomings of fixed-shape CNNs. Sampling 25 crops each picture increases computing cost during inference and introduces randomness because of the crop selection procedure, even though it may mitigate the fixed-shape restriction. Because each crop only shows a small section of the image, it may miss important global quality indicators.

Our method performs well on assessing the quality of images.

## 4.3 Visualization

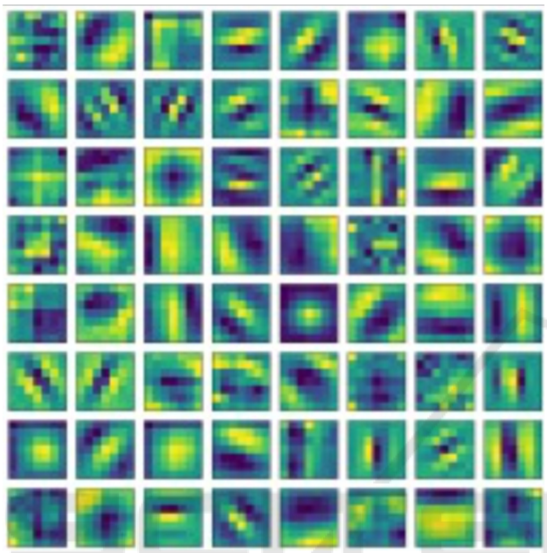

Figure 5: The visualizations of learnt kernels in the first convolutional layer of the network.

We show the convolutional kernels from the first layer of the network to have a better grasp of the features learnt by our ConvNeXt-based NR-IQA model. Some of these learnt kernels are shown in Fig. 5. It's interesting to note that a large number of these kernels have patterns that resemble texture analyzers, edge detectors, and noise-sensitive filters. This implies that the model is capable of recognizing subtle aspects of images, such sharpness, smoothness, and the existence of artifacts, that are frequently linked to perceptual quality.

Visually evaluate the model's performance. A variety of photos, including ones with different levels of blur, noise, and compression errors, are displayed in this figure. We show the associated ground-truth score and the model's projected quality score for each image. The findings show that, for most distortions, our model accurately represents the relative quality differences over a range of distortions and generally agrees well with human perception. Furthermore, we show example inference results in Fig 6.

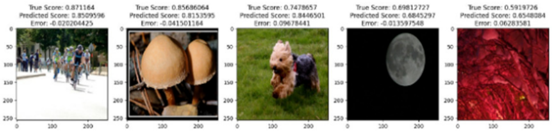

Figure 6: Inferencing visualization of the trained model.

Table 1: Comparision of SOTA methods on KoniqIQ-10k.

| APPROACH | PLCC | SRCC |
|---|---|---|
| WaDIQaM (Ke et al., 2022) | 0.794 | 0.799 |
| BIECON (Ledig et al., 2017) | 0.620 | 0.681 |
| SFA (Yamashita and Markov, 2020) | 0.859 | 0.847 |
| BRISQUE (Ma et al., 2019) | 0.668 | 0.701 |
| BIQA (Xu et al., 2023) | 0.902 | 0.918 |
| HOSA (Dosovitskiy et al., 2021) | 0.674 | 0.697 |
| PQR (Wang et al., 2015) | 0.873 | 0.881 |
| ILNIQE (Gu et al., 2021) | 0.501 | 0.524 |
| DBCNN (Xu et al., 2021) | 0.832 | 0.887 |
| MetaliQA (Jiang et al., 2019) | 0.847 | 0.888 |
| *ConvNeXt (Ours)* | ***0.920*** | ***0.940*** |

## 5 CONCLUSION AND FUTURE WORK

The usefulness of the ConvNeXt architecture for No-Reference Image Quality Assessment (NR-IQA) was investigated in this study. Through the utilization of the KonIQ-10k dataset, we were able to refine the pre-trained ConvNeXt model and create an NR-IQA model that can reliably predict image quality scores without the need for reference pictures. Our findings show that the ConvNeXt-based model outperforms current state-of-the-art techniques in Spearman's Rank Correlation Coefficient (SRCC) and Linear Correlation Coefficient (LCC). The model was shown to be able to effectively capture low-level picture information linked to perceptual quality through the visualization of the learnt kernels. While more research is necessary to handle a variety of visual distortions and investigate other ConvNeXt versions, this work lays a solid basis for utilizing ConvNeXt's capabilities for precise and effective NR-IQA, opening the door for its use in a variety of practical image processing and computer vision applications.

## REFERENCES

Ahmad, W., et al. “A new generative adversarial network for medical images super-resolution." Scientific Reports, 2022.

Bosse, S., et al. “Deep neural networks for no-reference and full-reference image quality assessment." IEEE TIP, 2018.

Bulat, A., et al. “To learn image super-resolution, use a GAN to learn how to do image degradation first." ECCV, 2018.

Dosovitskiy, A., et al. “An image is worth 16x16 words: Transformers for image recognition at scale."ICLR, 2021.

Egger, J., et al. “Deep learning—a first meta-survey of selected reviews across scientific disciplines." PeerJ Computer Science, 2021.

Gu, J., et al. “Blind image quality assessment using self-attention mechanisms." ICCV Workshops, 2021.

Gu, J., et al. “Blind super-resolution with iterative kernel correction." CVPR, 2019.

Gu, J., et al. “NTIRE 2022 challenge on perceptual image quality assessment." CVPR Workshops, 2022.

Gupta, R., et al. “Super-resolution using GANs for medical imaging." Procedia Computer Science, 2020.

Haris, M., et al. “Deep back-projection networks for super-resolution." CVPR, 2018.

Huang, Y., et al. “Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding." CVPR, 2017.

Jiang, K., et al. “Deep distillation recursive network for remote sensing imagery super-resolution." Remote Sensing, 2018.

Jiang, K., et al. “Edge-enhanced GAN for remote sensing image super-resolution." IEEE TGRS, 2019.

Ke, J., et al. “Vision transformers in image quality assessment." CVPR, 2022.

Ledig, C., et al. “Photo-realistic single image super-resolution using a generative adversarial network." CVPR, 2017.

Liu, P., et al. “A comprehensive review on no-reference image quality assessment." Information Fusion, 2021.

Ma, K., et al. “Blind image quality assessment: From scene statistics to deep learning." IEEE Signal Processing Magazine, 2017.

Ma, K., et al. “Perceptual image quality assessment with deep neural networks: A review." IEEE TIP, 2019.

Mittal, A., et al. “Making a 'completely blind' image quality analyzer." IEEE Signal Processing Letters, 2013.

Mittal, A., et al. “NIQE: A no-reference image quality evaluation metric based on natural scene statistics." IEEE Signal Processing Letters, 2013.

Wang, K., et al. “A transformer-based network for blind image quality prediction." IEEE TIP, 2022.

Wang, P., et al. “A comprehensive review on deep learning-based remote sensing image super-resolution methods." Earth-Science Reviews, 2022.

Wang, R., et al. “Deep learning for no-reference image quality assessment." IJCV, 2022.

Wang, X., et al. “ESRGAN: Enhanced super-resolution generative adversarial networks." ECCV Workshops, 2018.

Wang, Z., Chen, J., & Hoi, S. C. H." Deep learning for image super-resolution: A survey."IEEE TPAMI, 2021.

Wang, Z., et al. “A feature-enriched completely blind image quality evaluator." IEEE Signal Processing Letters, 2015.

Xu, Y., et al. “TE-SAGAN: An improved generative adversarial network for remote sensing super-resolution images." Remote Sensing, 2022.

Xu, Z., et al. “A survey on model evaluation metrics for no-reference image quality assessment." ACM Computing Surveys, 2021.

Xu, Z., et al. “Few-shot remote sensing image scene classification based on metric learning and local descriptors." Remote Sensing, 2023.

Yamashita, K., & Markov, K. "Medical image enhancement using super-resolution methods." ICCS, 2020.

Yue, Z., et al. “Blind image super-resolution with elaborate degradation modeling on noise and kernel." CVPR, 2022.

Zhang, J., et al. “Single-image super-resolution of remote sensing images with real-world degradation modeling" Remote Sensing, 2022.

Zhang, W., et al. “Deep learning-based no-reference image quality assessment: A survey." Neurocomputing, 2021.

Zhao, Q., et al. “A self-attention-based transformer network for no-reference image quality prediction." ICCV, 2021.