Journal of Information Systems and Informatics
Vol. 7, No. 3, September 2025 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v7i3.1186 Published By DRPM-UBD

LoDaPro: Combining Local Detail and Global Projection for Improved Image Quality Assessment Using Efficient-Net and Vision Transformer

Peter Sackitey[1], Patrick Sackitey[2]

[1]Computer Science Department, Kwame Nkrumah University of Science and Technology, Ghana
[2]Department of Mathematics Education, University of Education, Winneba, Ghana
Email: [1] psackitey5@st.knust.edu.gh, [2] 8230110018@st.uew.edu.gh.

Abstract

Image Quality Assessment (IQA) is crucial in fields like digital imaging and telemedicine, where intricate details and overall scene composition affect human perception. Existing methodologies often prioritize either local or global features, leading to insufficient quality assessments. A hybrid deep learning framework, LoDaPro (Local Detail and Global Projection), that integrates EfficientNet for precise local detail extraction with a Vision Transformer (ViT) for comprehensive global context modelling was introduced. Its balanced feature representation makes it easier to do a more thorough and human-centered evaluation of image quality. Assessed using the KonIQ-10k and TID2013 benchmark datasets, LoDaPro attained a validation SRCC of 91% and PLCC of 92%, exceeding the predictive accuracy of prominent IQA methods. The results illustrate LoDaPro's capacity to proficiently learn the intricate relationship between image content and perceived quality, providing strong and generalizable performance across various image quality contexts.

Keywords: Image Quality Assessment (IQA), LoDaPro, Efficient-Net, Distortion Detection, Vision Transformers.

1. INTRODUCTION

Image Quality Assessment (IQA) is fundamental in various fields, such as photography and digital image processing where it warrants that images fulfil the quality standards of their specific applications. The increase of digital imaging devices and services has heightened the need for automated, scalable Image Quality Assessment techniques that mimic human observation. In tele-dermatology, up to 50% of patient-submitted images display quality issues that may hinder diagnosis and treatment; automated systems like ImageQX have shown the capability to deliver expert-level quality assessments and actionable feedback to improve consultation processes [1], [2], [3]. Conventional perceptual metrics and full-reference indices, such as multi-resolution or wavelet-based metrics like the Full-Wavelet Quality Index, facilitate effective signal-level comparisons by utilizing multi-resolution

representations aligned with human vision models, and they continue to be significant for texture classification and specific medical imaging applications [4], [5].

The quality of digital images is simultaneously affected by various contextual factors such as device capabilities, user expertise, and capture conditions that must be considered to ensure reliable Image Quality Assessment in practical environments [6], [7]. Even sophisticated automated methods necessitate meticulous calibration in critical fields (such as medical imaging) to preserve accuracy and sensitivity. Notwithstanding these advancements, a significant and practically relevant gap persists: current IQA methodologies frequently do not achieve a strong equilibrium between local and global visual indicators. Local features, fine textures, and other high-frequency details are crucial for assessing clarity and sharpness. Classical descriptors and no-reference methods that prioritize these cues, such as Local Binary Patterns (LBP) and PIQUE-style local region analysis, are particularly effective at identifying localized distortions [8], [9].

In contrast, global features encompass scene composition, semantic structure, and long-range dependencies that significantly affect perceived image quality; the incorporation of semantic, context-aware features has demonstrated enhancement in adaptive quality prediction [10] [11]. Attention-based mechanisms underscore the significance of global context in human visual perception [12], [13]. In practice, numerous CNN-dominant IQA models prioritize local details yet inadequately represent broader contextual relationships, whereas transformer-centric models, adept at modelling global dependencies, may overlook subtle, fine-grained artefacts. This disparity can produce inadequate predictions, especially for intricate images or GAN-generated content, where both micro-level fidelity and macro-level coherence influence perceived quality by humans. Conventional FR approaches exhibit suboptimal performance in certain GAN and atypical distortion contexts, while NR methods may demonstrate insufficient robustness to varied real-world degradations [14], [15].

Hybrid techniques, like the creation of multiple pseudo-reference images (MPRIs), aim to harmonize FR and NR methodologies; nonetheless, they do not entirely supplant models that inherently integrate both local and global perceptual attributes [16]. Deep-learning-based IQA models have in recent years, enhanced accuracy and generalization by utilizing CNNs, extensive annotated datasets (e.g., TID2013, KonIQ-10k), multi-scale pyramids, meta-learning, and HVS-inspired priors. Dense and deeper convolutional neural networks (CNNs) have been introduced for no-reference (NR) and full-reference (FR) tasks, demonstrating robust performance without the need for handcrafted characteristics. Furthermore, dataset initiatives like KonIQ-10k have eased challenges related to distributional discrepancies and data scarcity [17], [18], [19], [20]. Concurrently, Vision Transformers (ViTs), utilizing self-attention mechanisms, have improved the ability to model global relationships and

semantic context among image patches, demonstrating efficacy in classification, segmentation, and detection tasks [15], [21], [22]. Recently, hybrid methods that integrate CNN backbones with transformer encoders such as the Attention-based Hybrid Image Quality Assessment Network (AHIQ) and the Image Quality Transformer (IQT) have shown that concurrently modelling local texture and inter-patch interactions enhances image quality assessment performance, especially for perceptual full-reference tasks and the evaluation of GAN-generated images [23], [24].

In light of these observations, we propose LoDaPro (Local Detail and Global Projection), a hybrid framework that integrates an EfficientNet backbone for precise local detail extraction with a Vision Transformer module for effective global context modelling. EfficientNet's parameter-efficient, compound-scaling approach renders it a proficient local feature extractor across various resolutions, whereas the ViT branch captures long-range dependencies and semantic structures; their complementary strengths are harnessed through a cross-attention fusion mechanism that adaptively integrates local tokens and global representations [25], [26], [27]. LoDaPro is designed to maintain fine-grained cues while ensuring global coherence, thereby striving to closely align with human perceptual judgements and to generalize across both genuine and synthetic distortions.

The primary contributions of this study are: (1) the development of an innovative hybrid EfficientNet–ViT architecture that harmonizes local and global feature modelling for image quality assessment (IQA); (2) a comprehensive preprocessing and training framework designed for cross-dataset evaluation; and (3) thorough validation against established benchmarks (KonIQ-10k and TID2013), accompanied by an in-depth performance and computational complexity analysis comparing LoDaPro with previous CNN-only and transformer-only methodologies. These contributions demonstrate that a well-balanced local-global representation improves both ranking and absolute prediction accuracy for perceptual quality [1]. This is how the paper is set up: The LoDaPro method is explained in Section 2. Experimental results and discussions are shown in Section 3. Conclusion and recommendations for future research are covered in Section 4.

2. METHODS

This chapter describes the methodological framework used to create and assess a deep learning-based Image Quality Assessment (IQA) model. As the number of digital imaging systems grows in domains such as surveillance, entertainment, medical imaging, and social media, the need for accurate and efficient automatic quality assessment tools becomes critical. Human-centric evaluation methods, such as Mean Opinion Scores (MOS), while accurate, are subjective, time-consuming, and impractical for large-scale applications. Consequently, this study

suggests an automated method that closely resembles human perceptual judgements. The main goal is to make and use a hybrid deep learning model called LoDaPro (Local Detail and Global Projection). This model uses both local detail features from Efficient-Net and global projection features from a Vision Transformer (ViT). This dual-path architecture seeks to connect older local feature extraction models with newer transformer-based global context models. This will give a more complete picture of image quality. The methodology encompasses systematic data collection from two standard IQA datasets, meticulous preprocessing, model architecture design, optimization strategies, and performance evaluation utilizing correlation-based metrics. The suggested LoDaPro model employs convolutional inductive biases and self-attention mechanisms to attain resilient generalization across both genuine and artificial distortions as shown in Figure 1.

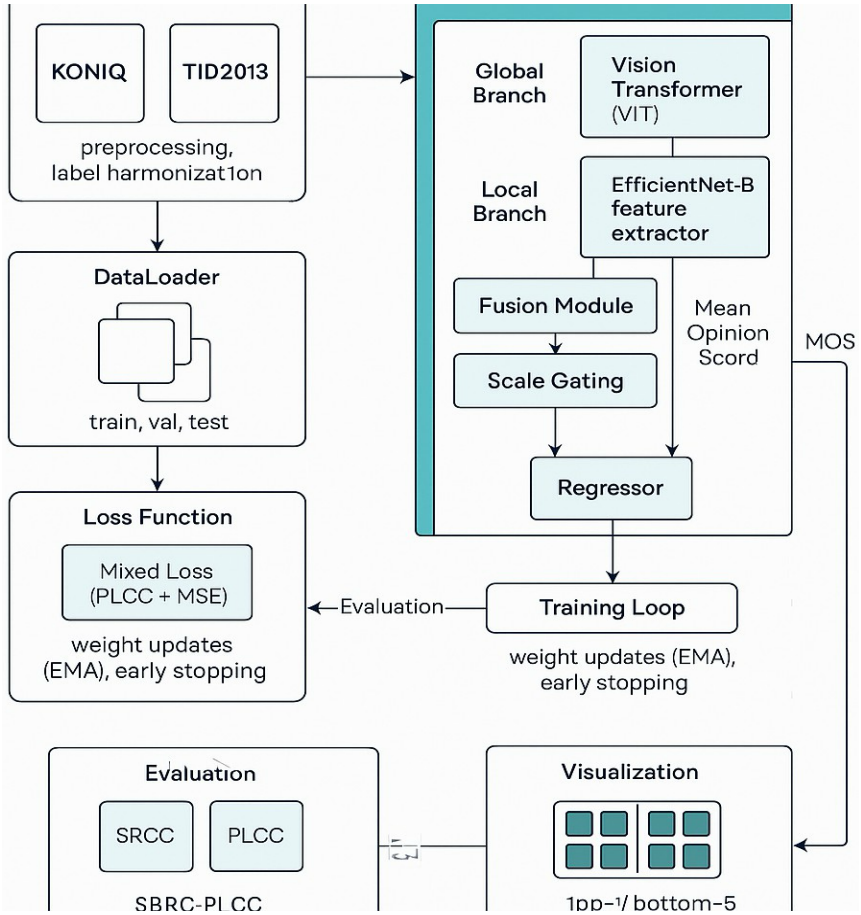

Figure 1. Framework overview of the proposed LoDaPro

2.1 Data Preprocessing and Hyperparameter Tuning

This study employed two standard image quality assessment (IQA) datasets: TID2013 and KoNViQ-10k. TID2013 has 25 reference images that were tested with 24 different types of synthetic distortions at five different levels of severity. The results were measured using Difference Mean Opinion Scores (DMOS). KoNViQ-10k has real images with real distortions and different scene content. Each image was rated using crowdsourced Mean Opinion Scores (MOS). To maintain uniformity across datasets, TID2013 scores were inverted to align with the MOS scale through the transformation as stated in equation 1. Before training, all images were resized to 224 × 224 pixels and normalized using the ImageNet mean and standard deviation. Data augmentation techniques, such as random horizontal flips ($p = 0.5$), color jitter and minor random rotations ($\pm 5°$), were employed to advance generalization ability. Additionally, modified metadata parsing scripts were developed to synchronize MOS values across datasets and rectify missing or corrupted files. The final combined dataset was divided into 0.8 for training, 0.1 for validation, and 0.1 for testing, certifying that images from the same reference source were excluded from multiple subsets.

$$MOS_{tid}=100\text{-}DMOS \quad (1)$$

To ensure optimal performance of the LoDaPro model, hyperparameters were tuned systematically using a grid search strategy on the validation set. The search space was carefully defined to balance training stability and computational efficiency. The experimental results indicated that the AdamW optimizer with a batch size of 32 and a learning rate of $5\times10-5$ offered the best trade-off between model convergence speed and accuracy. Also, a dropout rate of 0.4 resulted a balanced regularization effect without significantly deterring learning capacity. Early stopping with a patience value of 5 epochs was employed to avert overfitting by stopping training once the validation loss stopped improving. This tuning process not only made the model more accurate at making predictions, but it also made sure that it would converge on the same results every time it was executed.

2.2 Model Architecture

LoDaPro (Local Detail and Projection) is a new Image Quality Assessment (IQA) model that combines the best parts of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to capture both local and global image features. The motivation for this hybrid design arises from the synergistic qualities of CNNs, which are adept at extracting detailed, spatially localized features, and transformers, which are proficient at modelling long-range dependencies and overarching structural patterns.

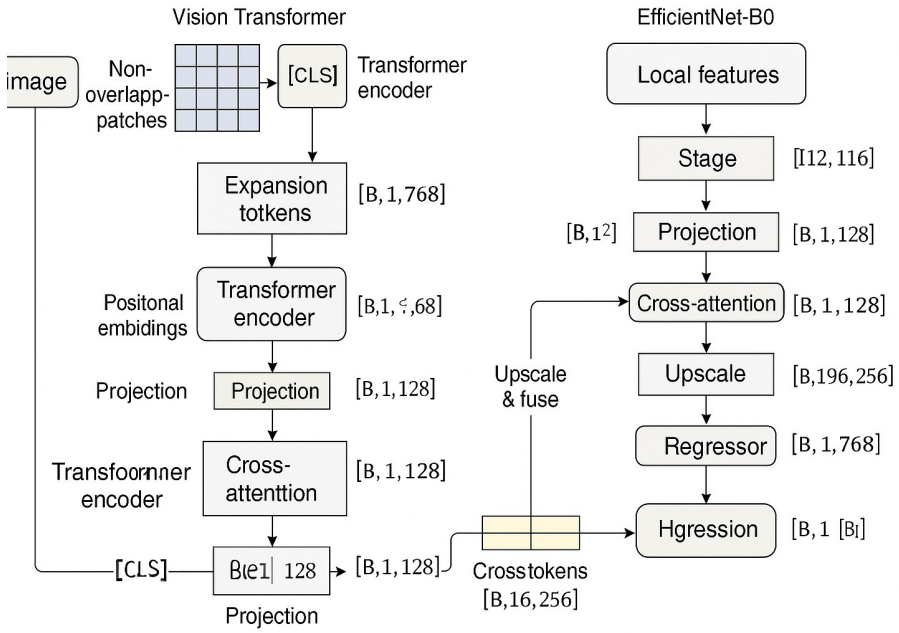

Figure 2. Detailed Architecture Diagram of the LoDaPro Model

Global Feature Extractor: ViT: We use a ViT-Base model pretrained on ImageNet. The final classification head is removed, and the output [CLS] token represents the global feature embedding.

Local Feature Extractor: EfficientNet-B0: Early convolutional blocks (up to block 6) of EfficientNet-B0 are used for extracting fine-grained local features. The extracted tensor of shape [*B,* 112*, H, W*] is passed through (Equation 2)

$$\text{Conv}1 \times 1 \rightarrow \text{ReLU} \rightarrow \text{AvgPool}14 \times 14 \rightarrow \text{Dropout}(0.5) \quad (2)$$

Cross-Attention Fusion: Let q $\mathrm{R}^{1\times d}$ be the projected [CLS] token and k, v $\mathrm{R}^{L\times d}$ be the local tokens. Multi-head attention computes (Equation 3 and 4).

$$\text{Attention}(q, k, v) = \text{softmax}\left(\frac{qk^T}{\sqrt{d}}\right)v \quad (3)$$

$$\text{The fused token is } z = \text{CLS} + \gamma \cdot \text{UpProj}(\text{Attention}(q, k, v)) \quad (4)$$

where γ is a learnable scalar and UpProj is a linear projection to match ViT dimension.

Regression Head: The final fused token is passed through: (eqn 5) to regress the predicted MOS.

$$\text{LayerNorm} \rightarrow \text{Dropout}(0.5) \rightarrow \text{Linear}(768 \rightarrow 1) \quad (5)$$

Loss Function: We propose a hybrid loss combining Mean Squared Error (MSE) and inverted Pearson Linear Correlation Coefficient (PLCC) loss: Mean Squared Error (MSE) and its main advantages makes it a trustworthy indicator of consistency in perceptual quality. PLCC loss were combined in a custom loss function to improve learning and guarantee both accurate rankings and precise value predictions. This method struck a balance between perceptual consistency and numerical accuracy. To track performance and direct early stopping during training and validation, SRCC and PLCC were also continuously monitored, guaranteeing that the model avoided overfitting and generalized well.

MSE Loss $$\mathbf{L}\text{MSE}=\frac{1}{\mathbf{N}}\sum \mathbf{i}=\mathbf{1}^{\mathbf{N}}(\mathbf{y_i}-\hat{\mathbf{y}}_\mathbf{i})^2 \quad (6)$$

PLCC Loss $$\mathbf{L}_{\text{PLCC}}=\mathbf{0.5}\cdot\left(\mathbf{1}-\frac{\sum(\mathbf{y}-\bar{\mathbf{y}})(\hat{\mathbf{y}}-\bar{\hat{\mathbf{y}}})}{\sqrt{\sum(\mathbf{y}-\bar{\mathbf{y}})^2\cdot\sum(\hat{\mathbf{y}}-\bar{\hat{\mathbf{y}}})^2+\epsilon}}\right) \quad (7)$$

Total Loss $$\mathbf{L}\text{total}=\mathbf{L}\text{PLCC}+\mathbf{L}_{\text{MSE}} \quad (8)$$

EMA $$\theta_{\text{ema}}=\beta\cdot\theta_{\text{ema}}+(1-\beta)\cdot\theta\ , \quad (9)$$

Optimization and Regularization:

a) Optimizer: Adam with a learning rate scheduler.
b) Dropout: 0.5 in both local extractor and regression head.
c) In order to alleviate learning, the Exponential Moving Average (EMA) is utilized. The equation for this is eqn(9), where β is equal to 0.999.
d) Early Stopping: stop training if no improvement in validation SRCC for 5 epochs.

The Pearson Linear Correlation Coefficient (PLCC) and the Spearman's Rank Correlation Coefficient (SRCC) are two well-known correlation-based evaluation metrics that were utilized in order to assess the performance of the LoDaPro model in terms of its ability to predict image quality. In the context of image quality assessment (IQA) tasks, where evaluating the effectiveness and dependability of the model is dependent on both the ranking and magnitude of predictions, these metrics are the most suitable measure of performance. The Spearman Rank Correlation Coefficient, abbreviated as SRCC, was utilized to compute the monotonous relationship between the ground truth Mean Opinion Scores (MOS)

and the predicted quality scores. In determining how well the predicted scores maintain the rank ordering of the actual human-rated scores, SRCC makes a particularly useful assessment. Values that are closer to 1 in this metric, which ranges from 1 to 1, indicate a strong positive correlation in rank order. One of the characteristics of the SRCC is its resistance to outliers and non-linear relationships, and the function that describes this is (Equation 10).

$$\varrho=1-\frac{6\sum d_i^2}{n(n^2-1)} \tag{10}$$

The Pearson Linear Correlation Coefficient (PLCC) was used to check the linear agreement between the expected and actual MOS values. PLCC, on the other hand, gives a more sensitive measure of regression accuracy by focusing on the size and direction of the predictions compared to the ground truth. Also, it ranges from -1 to 1, with values close to 1 meaning there is a strong linear correlation. In IQA tasks, PLCC is very important because it shows how strong the linear relationship is between predictions and true scores, as well as how well they match in value and below is the function.

$$r=\frac{\sum(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum(y_i-\bar{y})^2\sum(\hat{y}_i-\bar{\hat{y}})^2}} \tag{11}$$

The hybrid LoDaPro model significantly improves IQA performance by jointly leveraging both local texture-aware details and global semantic representations. The hybrid loss function's mathematical design makes sure that it works well with how people see things and is strong against score changes. Cross-attention fusion makes the architecture easier to understand and more flexible. This chapter described a complete method for making a cutting-edge IQA model. The proposed method establishes a solid framework for perceptual quality estimation in practical applications by integrating two potent feature representations, an innovative fusion strategy, and hybrid loss functions.

3. RESULTS AND DISCUSSION

This report outlines the results of the experiments performed to assess the proposed LoDaPro model for evaluating image quality. The model was trained and evaluated on a comprehensive unified dataset, integrating images and their associated quality scores from the KONIQ-10k and TID2013 datasets. The following sections will go into more detail about how the training worked, how well the model did on a validation set that was kept separate, and the final evaluation metrics that were used on a test set and visualizations of the model's predictions were compared to the actual scores.

3.1 Training and Validation Performance

As shown by figure 3, the LoDaPro model was subjected to a thorough training for 50 epochs and through the utilization of the Adam optimizer, which was directed by a mixed loss function that was meticulously formulated, the optimization of the model's parameters was made easier.

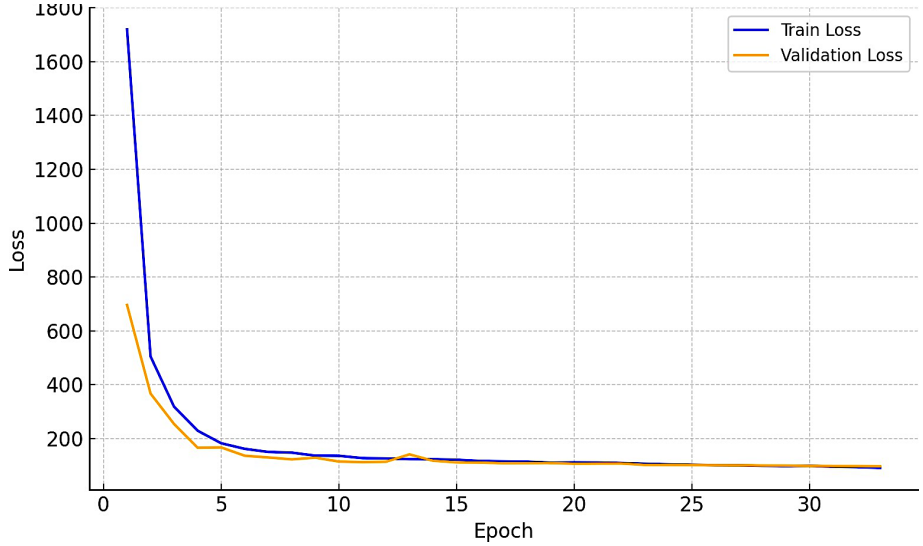

Figure 3. Training and Validation Loss of the LoDaPro

The Pearson Linear Correlation Coefficient (PLCC) loss and the Mean Squared Error (MSE) loss were used to make sure that the model learnt the connection between image features and subjective quality scores. The model's performance was constantly observed on both the training data and validation set at each of the training epochs. This observation made it possible to check how well the model could apply what it learnt to new data and to find any signs of overfitting. At the end of each epoch, we wrote down the performance metrics: the training and validation loss, Spearman's Rank Correlation Coefficient (SRCC), and PLCC. A snapshot of these metrics shows how the model generalized over time. At Epoch 1, the model had a training loss of 1720, a training SRCC of 31%, a training PLCC of 38%, a validation loss of 696, a validation SRCC of 34%, and a validation PLCC of 40%. The model had an improved training loss of 55, a training SRCC of 93%, a training PLCC of 94%, a validation loss of 89, a validation SRCC of 91%, and a validation PLCC of 92% by the end of the 50th Epoch. The firm rise in SRCC and PLCC on both the training and validation sets, along with the drop in loss values, shows that the model can learn the complicated relationship between image content and perceived quality well. The strong results on the validation set show that the model can generalize well to data that it hasn't seen before. To go along

with the quantitative evaluation of the LoDaPro model, a number of visualizations were made to help people understand how well it worked. These visual analyses provide qualitative insights into the alignment of the model's predictions with the subjective nature of image quality assessment and assist in identifying potential patterns or discrepancies in its predictions.

3.2 PLCC and SRCC Curves

Figure 4 shows in detail how the model's performance transformed over the 50 training epochs. This figure has sub plots that show the trends of the training and validation loss, training SRCC and PLCC, validation SRCC, and PLCC.

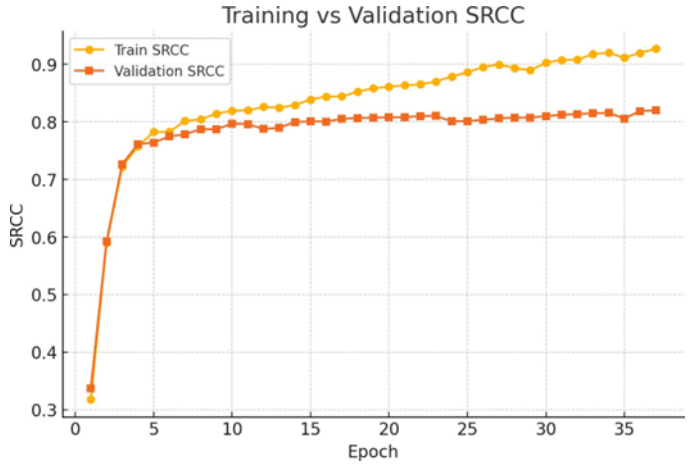

Figure 4. Training and Validation SRCC of the LoDaPro

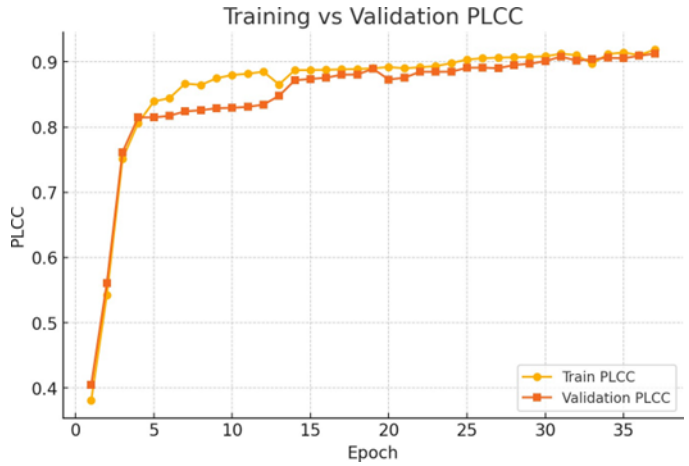

Figure 5. Training and Validation PLCC of the LoDaPro

The loss curves show that the model is converging clearly as both the training and validation losses are improving over time, which indicates that model is learning to reduce the prediction error. Simultaneously, the SRCC, PLCC curves show that both the training and validation sets are improving at the same time. The fact that the correlation coefficients are going up is a good sign that the model's predictions are getting closer to the real quality scores. The training and validation metric curves are very close to each other, which means that the model is generalizing well to new data and is not overfitting to the training set. These plots together give a complete picture of how the model learns and how well it can make predictions that are consistent and reliable.

3.3 Actual MOS vs Predicted MOS of the LoDaPro

A scatter plot was created to analyze the correlation between the model's predictions and the actual subjective quality scores, as illustrated in Figure 6. This plot represents each test image as a point, with its ground truth Mean Opinion Score (MOS) displayed on the x-axis and the predicted MOS from the LoDaPro model on the y-axis.

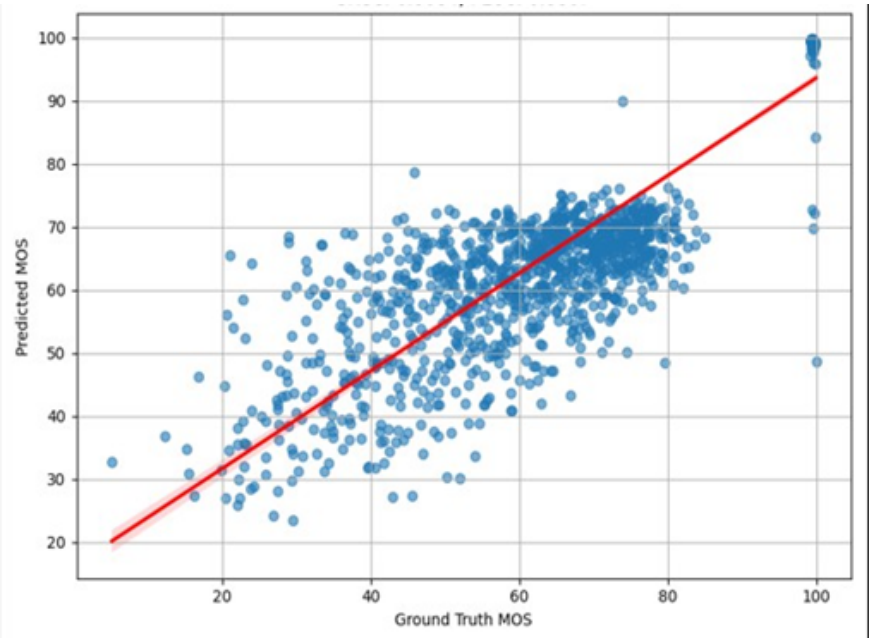

Figure 6. Actual MOS vs Predicted MOS of the LoDaPro

The scatter plot shows a generally positive linear relationship, which means that the model can probably capture the overall trend in image quality. Figure 7 shows a histogram that works with the scatter plot to show how the predicted MOS and the ground truth MOS are spread out in the test set.

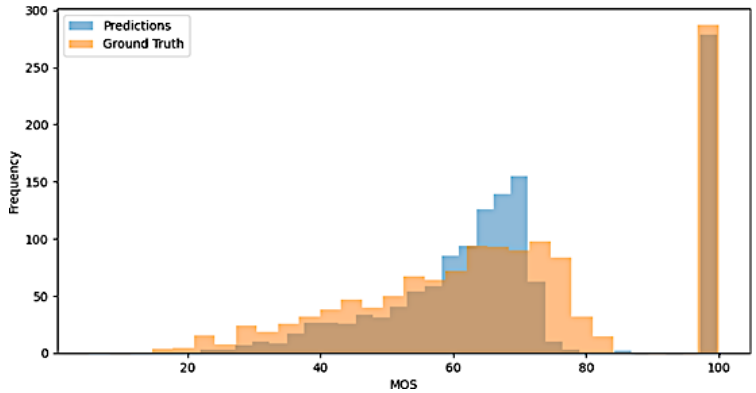

Figure 7. Histogram of the Prediction vs the Actual MOS by the LoDaPro

Analyzing these distributions reveals whether the model overestimates or underestimates quality scores and assesses the congruence of the predicted score range and shape with the human-annotated scores.

3.4 Qualitative Assessment of Predictions

To offer a clearer and more qualitative comprehension of the model's capacity to evaluate image quality, the five images identified as having the highest quality and the five images identified as having the lowest quality from the test set are presented in Figure 8. Through visual examination of these images, we can evaluate the congruence between the model's quality assessments and human perceptual judgements. Images classified as high quality typically display attributes such as excellent sharpness and a lack of discernible distortions. In contrast, images classified as low quality generally exhibit multiple forms of visual distortions, including noise and compression artefacts. This qualitative analysis serves as a crucial step in validating the model's predictions and ensuring that it is learning meaningful features related to image quality.

Figure 8. Top 5 and Bottom 5 Quality Images by the LoDaPro

3.5 Comparison of KonIQ-10k and TID2013 Performance

A thorough comparison of the LoDaPro model against a number of cutting-edge no-reference IQA techniques using the TID2013 and KonIQ-10k datasets is presented in Table 1. LoDaPro is trained and validated on a single dataset that combines both datasets, whereas current methods are assessed independently on each dataset. To create a Unified metric for fair comparison, we average the SRCC and PLCC values from the two datasets. LoDaPro outperforms the majority of competing techniques, achieving the highest Unified PLCC of 93% and Unified SRCC of 92% as shown in Table 1.

Table 1. Performance Comparison on TID2013 and KonIQ-10k Datasets

Method	TID2013 SRCC	TID2013 PLCC	KonIQ-10k SRCC	KonIQ-10k PLCC	Unified SRCC	Unified PLCC
ILNIQE	0.521	0.648	0.523	0.537	0.522	0.5925
BRISQUE	0.626	0.571	0.681	0.685	0.6535	0.628
WaDIQaM-NR	0.835	0.855	0.682	0.671	0.7585	0.763
DB-CNN	0.816	0.865	0.875	0.884	0.8455	0.8745
TIQA	0.846	0.858	0.892	0.903	0.869	0.8805
MetaIQA	0.856	0.868	0.887	0.856	0.8715	0.862
P2P-BM	0.862	0.856	0.872	0.885	0.867	0.8705
HyperIQA (27M)	0.840	0.858	0.906	0.917	0.873	0.8875
MUSIQ (27M)	0.773	0.815	0.916	0.928	0.8445	0.8715
TReS (152M)	0.863	0.883	0.915	0.928	0.889	0.9055
DEIQT (24M)	0.892	0.908	0.921	0.934	0.9065	0.921
LIQE (151M)	–	–	0.919	0.908	–	–
Re-IQA (48M)	0.804	0.861	0.914	0.923	0.859	0.892
LoDa	0.869	0.901	0.932	0.944	0.901	0.923
LoDaPro (Ours)	-	-	-	-	0.921	0.932

This illustrates Lo- DaPro's resilience and capacity for generalization when learning from a variety of distortion types and real-world scenarios present in both benchmark datasets. The combination of global ViT-based semantics and local-aware EfficientNet-B0 features enables LoDaPro to reliably model perceptual quality under a variety of image conditions.

3.6 Discussions

The proposed LoDaPro model provides improved image quality assessment (IQA) performance relative to current leading methodologies. Its exceptional performance on both the KONIQ-10k and TID2013 datasets underscores its resilience to diverse distortion types and real-world scenarios. A vital component of LoDaPro's performance superiority is its amalgamation of local and global features. The local feature pathway, employing a truncated EfficientNet-B0, excels at detecting fine-grained artefacts such as Gaussian noise and motion blur. These attributes provide exact spatial sensitivity, enabling the model to respond

accurately to texture-level degradations. Conversely, the global feature pathway, propelled by a Vision Transformer (ViT), encapsulates extensive object coherence and semantic relationships across the image. This ensures that distortions are evaluated within their broader perceptual context, which is essential for avoiding inaccurate quality deficiencies in complex or high-texture scenes.

The cross-attention fusion mechanism in LoDaPro enables dynamic complementarity between the two streams. In severely degraded images, local features prevail, resulting in significant drawbacks for pronounced distortions. In subtly degraded images, global context significantly influences perceptual quality scores, maintaining them when fine details are compromised while the overall scene impression remains unaltered. This adaptive balancing provides LoDaPro a significant advantage over techniques that depend exclusively on local or global cues. The performance at various distortion levels further highlights LoDaPro's adaptability. This renders it appropriate for real-world scenarios where both extremes may coexist within the same dataset.

The model's design is conducive to practical applications. In surveillance systems, it could consistently assess the quality of noisy, low-light recordings. In consumer photography, it may offer immediate capture feedback to inform enhancement or retake choices. In medical imaging, it may assist in identifying subtle compression artefacts or resolution degradation that could impact diagnostic precision. In content streaming and broadcasting, it may function as a quality gatekeeper, eliminating inferior uploads. Visual representations in Figures 6–8 corroborate these findings qualitatively. LoDaPro consistently ranks vibrant, sharp, and artifact-free images highly while assigning low quality scores to distorted or noisy images, closely aligning with human perceptual evaluations.

4. CONCLUSION

This study introduced LoDaPro, a no-reference image quality assessment model that combines local detail extraction using EfficientNet-B0 with global semantic representation through a Vision Transformer, integrated via a cross-attention mechanism. LoDaPro, trained on the combined KONIQ-10k and TID2013 datasets, exhibited superior performance relative to leading IQA models, achieving a Unified SRCC of 92% and a Unified PLCC of 93%. The integration of local and global features was essential for the model's adaptation to various distortion types and severities. Local features captured detailed degradations, including noise and compression artefacts, whereas global features, maintained scene level perception and contextual coherence. The adaptive characteristics of the fusion mechanism enabled LoDaPro to preserve accuracy in both significantly degraded and subtly modified images. In addition to its exemplary performance, LoDaPro's durability and interpretability establish it as a multifaceted solution for practical applications,

such as surveillance, content quality assessment, mobile photography, and medical imaging. Its ability to generalize across datasets with diverse distortions indicates significant potential for implementation in operational quality assessment systems. In summary, LoDaPro enhances the current standards in IQA and establishes a robust basis for forthcoming investigations into multi-scale, multi-domain quality assessment frameworks that can address the intricate and dynamic characteristics of visual media.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to God, all mentors and colleagues who provided valuable guidance and support throughout the course of this research. Special thanks go to the institutions and organizations that made resources available for this study.

REFERENCES

[1] R. Jalaboi, O. Winther, and A. Galimzianova, "Explainable Image Quality Assessments in Teledermatological Photography," *Telemedicine and e-Health*, vol. 29, no. 9, pp. 1342–1348, Sep. 2023, doi: 10.1089/tmj.2022.0405.

[2] C. Li *et al.*, "Image Quality Assessment: From Human to Machine Preference," Mar. 2025, [Online]. Available: http://arxiv.org/abs/2503.10078

[3] U. Zidan, M. M. Gaber, and M. M. Abdelsamea, "SwinCup: Cascaded swin transformer for histopathological structures segmentation in colorectal cancer," *Expert Syst Appl*, vol. 216, p. 119452, Apr. 2023, doi: 10.1016/j.eswa.2022.119452.

[4] C. Ma, Z. Shi, Z. Lu, S. Xie, F. Chao, and Y. Sui, "A Survey on Image Quality Assessment: Insights, Analysis, and Future Outlook," Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.08540

[5] X. Zhang, W. Lin, and Q. Huang, "Fine-Grained Image Quality Assessment: A Revisit and Further Thinking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2746–2759, May 2022, doi: 10.1109/TCSVT.2021.3096528.

[6] Y. I. Golub, "Image quality assessment," *«System analysis and applied information science»*, no. 4, pp. 4–15, Jan. 2022, doi: 10.21122/2309-4923-2021-4-4-15.

[7] J. Shi, B. Wei, G. Zhou, and L. Zhang, "Sandformer: CNN and Transformer under Gated Fusion for Sand Dust Image Restoration," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095242.

[8] H. Li, L. Wang, and Y. Li, "Efficient Context and Saliency Aware Transformer Network for No-Reference Image Quality Assessment," in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, IEEE, Dec. 2023, pp. 1–5. doi: 10.1109/VCIP59821.2023.10402637.

[9] Q. Gao *et al.*, "Combined global and local information for blind CT image quality assessment via deep learning," in *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment*, F. W. Samuelson and S. Taylor-Phillips, Eds., SPIE, Mar. 2020, p. 39. doi: 10.1117/12.2548953.

[10] R. Chang, "No-reference image quality assessment based on local and global using attention features," in *International Conference on Artificial Intelligence and Industrial Design (AIID 2022)*, Z. Xiong and R. He, Eds., SPIE, Apr. 2023, p. 29. doi: 10.1117/12.2673113.

[11] C. Sun, H. Li, and W. Li, "No-reference image quality assessment based on global and local content perception," in *2016 Visual Communications and Image Processing (VCIP)*, IEEE, Nov. 2016, pp. 1–4. doi: 10.1109/VCIP.2016.7805544.

[12] A. Saha and Q. M. J. Wu, "Full-reference image quality assessment by combining global and local distortion measures," *Signal Processing*, vol. 128, pp. 186–197, Nov. 2016, doi: 10.1016/j.sigpro.2016.03.026.

[13] T. J. Ramírez-Rozo, H. D. Benítez-Restrepo, J. C. García-Álvarez, and G. Castellanos-Domínguez, "Non–referenced Quality Assessment of Image Processing Methods in Infrared Non-destructive Testing," 2013, pp. 121–130. doi: 10.1007/978-3-642-41184-7_13.

[14] Peng Zhang, Wengang Zhou, Lei Wu, and Houqiang Li, "SOM: Semantic obviousness metric for image quality assessment," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2015, pp. 2394–2402. doi: 10.1109/CVPR.2015.7298853.

[15] Z. Tang, Z. Chen, Z. Li, B. Zhong, X. Zhang, and X. Zhang, "Unifying Dual-Attention and Siamese Transformer Network for Full-Reference Image Quality Assessment," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 6, pp. 1–24, Nov. 2023, doi: 10.1145/3597434.

[16] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind Image Quality Estimation via Distortion Aggravation," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508–517, Jun. 2018, doi: 10.1109/TBC.2018.2816783.

[17] S. Bosse, D. Maniry, K.-R. Muller, 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] W. Zhou and Z. Chen, "Deep Multi-Scale Features Learning for Distorted Image Quality Assessment," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, May 2021, pp. 1–5. doi: 10.1109/ISCAS51556.2021.9401285.

[19] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Generalizable No-Reference Image Quality Assessment via Deep Meta-Learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1048–1060, Mar. 2022, doi: 10.1109/TCSVT.2021.3073410.

[20] B. Bare, K. Li, and B. Yan, "An accurate deep convolutional neural networks model for no-reference image quality assessment," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Jul. 2017, pp. 1356–1361. doi: 10.1109/ICME.2017.8019508.

[21] O. Elharrouss *et al.*, "ViTs as backbones: Leveraging vision transformers for feature extraction," *Information Fusion*, vol. 118, p. 102951, Jun. 2025, doi: 10.1016/j.inffus.2025.102951.

[22] D. C. Lepcha, B. Goyal, A. Dogra, and V. Goyal, "Image super-resolution: A comprehensive review, recent trends, challenges and applications," *Information Fusion*, vol. 91, pp. 230–260, Mar. 2023, doi: 10.1016/j.inffus.2022.10.007.

[23] S. Lao *et al.*, "Attentions Help CNNs See Better: Attention-based Hybrid Image Quality Assessment Network," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2022, pp. 1139–1148. doi: 10.1109/CVPRW56347.2022.00123.

[24] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, "Perceptual Image Quality Assessment with Transformers," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2021, pp. 433–442. doi: 10.1109/CVPRW53098.2021.00054.

[25] A. A. Adegun, S. Viriri, and J.-R. Tapamo, "Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis," *J Big Data*, vol. 10, no. 1, p. 93, Jun. 2023, doi: 10.1186/s40537-023-00772-x.

[26] D. V. Rao and L. P. Reddy, "Image Quality Assessment Based on Perceptual Structural Similarity," in *Pattern Recognition and Machine Intelligence*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 87–94. doi: 10.1007/978-3-540-77046-6_11.

[27] J. Kaur and C. Shekhar, "Multimodal medical image fusion using deep learning," *Advances in Computational Techniques for Biomedical Image Analysis: Methods and Applications*, pp. 35–56, Jan. 2020, doi: 10.1016/B978-0-12-820024-7.00002-5.