Jurnal ELTIKOM :
Jurnal Teknik Elektro, Teknologi Informasi dan Komputer

Vol. 9, No. 1, June 2025, page. 23-34
ISSN 2598-3245 (Print), ISSN 2598-3288 (Online)
DOI: http://doi.org/10.31961/eltikom.v9i1.1398
Available online at http://eltikom.poliban.ac.id

NO-REFERENCE VIDEO QUALITY ASSESSMENT BASED ON THE DOVER FRAMEWORK USING A TRANSFER LEARNING METHOD

Ardhi Muda Ariska*, Tubagus Maulana Kusuma
Department of Technology and Engineering, Universitas Gunadarma, Indonesia
e-mail: ardhimudaariska@gmail.com, mkusuma@staff.gunadarma.ac.id

Received: 13 December 2024 – Revised: 3 March 2025 – Accepted: 4 March 2025

ABSTRACT

No-reference Video Quality Assessment (VQA) presents a critical challenge in digital multimedia. This study explores video quality measurement using the DOVER framework combined with a transfer learning method. While existing approaches often rely on end-to-end fine-tuning that requires substantial computational resources, this study introduces and validates a more efficient implementation. The model was built using Google Colab and Python, with the KoNViD-1k dataset as the training base. A head-only transfer learning approach was employed, using the DOVER framework as its foundation. This approach addresses a key research gap in resource-efficient no-reference VQA, as many state-of-the-art models remain impractical for real-world deployment due to high computational demands. The training process was conducted over 10 epochs with resource efficiency in mind. The head-only transfer learning technique allows for GPU memory optimization, showing minimal accuracy differences (1%–2%) compared to full end-to-end fine-tuning. Unlike previous studies that compromise performance for efficiency, this approach maintains competitive accuracy while significantly lowering computational costs. The results show that the proposed method delivers accurate and efficient video quality assessments, confirming the potential of the DOVER framework in no-reference VQA. This study highlights a practical balance between computational efficiency and assessment accuracy using transfer learning techniques.

Keywords*: accuracy, DOVER, efficiency, machine learning, transfer learning.*

I. INTRODUCTION

THE digital era is marked by an exponential increase in video content, largely driven by the growing popularity of User-Generated Content (UGC) [1]. Traditional video quality assessment methods, which primarily focused on technical aspects such as image sharpness and compression, are now considered inadequate. Researchers have highlighted the need for a more holistic approach—one that includes both technical parameters and aesthetic elements such as content and composition—to produce evaluations that better reflect user perception [2].

Recent developments in video quality assessment (VQA) have introduced innovative methodologies across various content domains. For instance, in AI-Generated Content (AIGC), benchmarking frameworks like AIGCBench have been designed to evaluate image-to-video generation tasks [3]. New approaches now incorporate multidimensional assessment criteria, including visual harmony, video-text consistency, and domain distribution gaps. Additionally, efficient deep learning methods such as FAST-VQA have been proposed to address computational challenges by introducing sampling strategies that preserve video quality information while reducing processing demands [4], [5].

No-Reference VQA remains a key challenge in digital multimedia, especially with the rapid expansion of content across streaming platforms, social media, and user-generated sources. Conventional no-reference VQA methods have relied mainly on technical visual features and often fall short in capturing the nuanced, subjective dimensions of video quality as perceived by users.

Recent progress in machine learning (ML) offers promising solutions to these limitations. By utilizing large-scale video datasets and advanced learning algorithms, ML-based models can capture the complex relationships between visual cues and perceived video quality. Transfer learning, in particular, has

emerged as a compelling technique, enabling models to adapt and generalize across various video types while maintaining computational efficiency.

This study aims to develop a VQA model using the DOVER (Decomposing Objective Video Evaluator) framework, combined with a novel transfer learning strategy. DOVER was chosen for its distinctive capability to integrate both technical and aesthetic evaluation components, resulting in a more comprehensive assessment aligned with human perception of video quality.

The proposed approach addresses key challenges in the field by overcoming the limitations of manual feature extraction, adapting to the dynamic nature of modern video content, and optimizing computational resources, particularly GPU memory. While existing approaches often rely on resource-intensive end-to-end fine-tuning, this study introduces a head-only transfer learning technique that enables efficient model development with only a 1–2% accuracy difference compared to full fine-tuning. This directly addresses a notable research gap in resource-efficient no-reference VQA, as many current state-of-the-art methods require excessive computational resources for practical use. By utilizing pre-trained models and innovative sampling strategies, this research aims to develop a more flexible and comprehensive video quality assessment method that maintains performance without compromising efficiency.

The KoNViD-1k dataset, specifically designed for real-world video analysis, is used to train the VQA model. This dataset was selected to improve the model's accuracy and generalizability across diverse video types. The implementation was conducted using Google Colab and Python, leveraging the platform's accessibility and machine learning capabilities.

The VQA method developed using the DOVER framework with transfer learning presents significant implications for various stakeholders in digital multimedia. For academic researchers, it advances more accurate and efficient VQA techniques, opening new avenues in computer vision, transfer learning, and video streaming research. For media institutions, it supports video streaming platforms in enhancing quality assessment, allowing for better content evaluation before publication. For content creators, it offers tools to improve video quality, potentially increasing viewer engagement and enhancing streaming experiences. For the broader public, it contributes to a more advanced understanding of video quality assessment technologies.

II. Research Method

Understanding the theoretical foundations and practical implementation of No-Reference VQA requires a thorough exploration of the methodological approach. This section outlines the key theoretical frameworks, computational strategies, and experimental design that support the proposed video quality evaluation method. By detailing the research methodology, we aim to provide a clear and rigorous explanation of the technical and theoretical decisions that shape this innovative VQA approach.

A. Theoretical Foundations

Transfer learning is a powerful machine learning technique that enables knowledge transfer from a source domain to a related target domain, particularly useful when data in the target domain is limited. In the context of VQA, this approach allows researchers to leverage the deep visual understanding embedded in pre-trained models, thereby improving performance and computational efficiency [6].

This study adopts a head-only transfer learning strategy, designed to optimize model adaptation with minimal computational cost [7]. By training only the fully-connected layers (the model's head) and retaining the pre-trained convolutional layers, this technique enables targeted model customization. It offers key benefits, including reduced computational demand, improved training stability, and preserved feature extraction capabilities—focusing the learning process on video quality assessment without sacrificing efficiency.

The proposed method strategically applies head-only transfer learning to develop an accurate no-reference VQA model using the KoNViD-1k dataset. By retraining only the model's head, the approach enables efficient learning of video quality features while minimizing architectural changes, thus achieving a strong balance between resource efficiency and prediction accuracy.

The DOVER (Decomposing Objective Video Evaluator) model introduces a novel approach to no-reference VQA, addressing the shortcomings of traditional methods that emphasize only technical aspects. Unlike conventional techniques, DOVER features a dual-branch architecture that simultaneously

assesses both aesthetic and technical components of video quality. This allows for a more holistic and perceptually aligned evaluation that reflects how users experience video content [8].

DOVER operates through two independent branches: the Aesthetic Branch and the Technical Branch. The Aesthetic Branch, trained on the AVA (Aesthetic Visual Analysis) dataset, evaluates visual elements such as composition, color, and lighting analyzing randomly sampled video frames. The Technical Branch extracts technical features including sharpness, distortion, and artifacts. These two sets of scores are then combined using empirically determined weights, enabling DOVER to deliver a holistic video quality evaluation that surpasses conventional single-dimensional methods.

This study focuses specifically on DOVER-Mobile, a lightweight variant optimized for resource-limited environments. With significantly fewer parameters (9.86M), lower computational demands (52.3 GFLOPs), and more efficient GPU memory usage, DOVER-Mobile offers a practical alternative to the standard DOVER model. It enhances efficiency while retaining the framework's strength in providing comprehensive, perception-aligned video quality assessment through its two-branch architecture.

The KoNViD-1k dataset is a benchmark collection for objective, no-reference VQA, known for its emphasis on real-world videos and detailed quality evaluations. Unlike traditional datasets, KoNViD-1k includes authentic videos from varied sources and features a broad range of distortions—such as blur, noise, compression, and color variations—supporting more robust and generalizable model training. Each video is annotated with subjective quality scores collected through multi-participant testing, serving as ground truth for model evaluation. The dataset also includes assessments of visual attributes like sharpness, color, contrast, and spatial detail, enabling a deeper analysis of video quality. Its use in this study is motivated by its strong relevance to real-world scenarios and its comprehensive coverage of perceptual quality dimensions [9].

Google Colab (Colaboratory) is a free, cloud-based development platform designed for executing Python code, particularly in machine learning, data science, and deep learning contexts. It offers interactive notebooks for writing, running, and visualizing Python code within a web browser. One of its main advantages is free access to GPU resources, which significantly accelerates model training. Colab is widely used by students, researchers, and practitioners for learning, prototyping, and collaboration [10].

Python is a high-level, versatile programming language known for its simple and readable syntax, making it accessible to beginners while remaining powerful for professionals. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming, and offers a rich ecosystem of libraries such as NumPy, Pandas, and TensorFlow. Python is widely used in various fields, including web development, data analysis, machine learning, and automation. Its key strengths include an intuitive, English-like syntax, a large and active user community, and flexibility in handling diverse tasks such as building web applications, analyzing data, developing machine learning models, and automating workflows [11].

Video quality metrics are quantitative measures used to assess the technical aspects of video quality through mathematical computations that yield numerical representations of specific characteristics. These metrics are essential for evaluating video performance and help professionals analyze the technical properties of video content. The main categories include error/distortion-based metrics, structural metrics, and basic technical metrics.

The most widely used metrics in these categories include Peak Signal-to-Noise Ratio (PSNR), which measures the ratio of maximum signal to noise and is expressed in decibels, and the Structural Similarity Index (SSIM), which evaluates structural similarities between images by considering luminance, contrast, and structure. Other important metrics include Mean Squared Error (MSE), Multi-Scale Structural Similarity (MS-SSIM), bitrate metrics (such as average and variable bitrate), and temporal metrics that assess frame rate consistency, frame dropping, and motion smoothness. These metrics offer comprehensive insights into video quality and support researchers and engineers in optimizing video processing and compression techniques [12], [13].

VQA is a comprehensive framework developed to evaluate video quality in response to the increasing demand for accurate and automated quality prediction. With video content accounting for 65% of internet traffic, the ability to measure and understand video quality has become critical. VQA approaches are generally divided into three types: Full-Reference (FR-VQA), which compares video against a high-quality reference; Reduced-Reference (RR-VQA), which uses partial reference data; and No-Reference (NR-VQA), which evaluates video quality without any reference [14], [15], [16].

Video quality assessment involves various dimensions, including technical factors such as resolution, noise, color accuracy, and motion smoothness, as well as perceptual factors like visual appeal and user experience. To measure the performance of VQA models, researchers use three main metrics: the Spearman Rank Correlation Coefficient (SRCC), which evaluates how well the model ranks video quality; the Pearson Linear Correlation Coefficient (PLCC), which measures the strength of linear relationships between predicted and actual scores; and the Kendall Rank Correlation Coefficient (KRCC), which assesses scoring consistency. While subjective methods like the Mean Opinion Score (MOS) remain reliable, they are costly and time-consuming, making objective evaluation increasingly important in modern VQA research [17].

B. System Planning

This study aims to develop a novel VQA method using a transfer learning approach, focusing specifically on head-only fine-tuning with the KoNViD-1k dataset. Traditional manual video quality assessments are subjective, time-consuming, and often inconsistent, highlighting the need for more efficient and accurate evaluation techniques. Previous machine learning approaches have commonly relied on full fine-tuning, which demands significant computational resources and large training datasets, leaving opportunities for more optimized alternatives.

To address these limitations, the study employs Google Colab with a T4 GPU, providing a powerful and accessible computing environment. The NVIDIA Tesla T4 GPU, based on the Turing architecture, delivers up to 65 TFLOPS of mixed-precision performance using 2560 CUDA cores and Tensor Cores. With 16GB of GDDR6 memory and a 70W TDP, it efficiently accelerates various AI workloads [18]. The technical configuration includes Python 3.9, along with key libraries such as PyTorch, Torchvision, and OpenCV, and the KoNViD-1k dataset. By comparing head-only transfer learning with traditional full fine-tuning, the study seeks to optimize resource usage while improving the accuracy of video quality assessment. The goal is to offer new insights into digital video quality analysis and enable advanced applications in multimedia and video streaming technologies.

C. Preparing the Work Environment

Setting up the work environment involves creating a Python 3.9 virtual environment, installing the essential libraries required for the VQA model, and configuring the core algorithmic code. This initial step ensures that all necessary tools are in place for the effective development of the video quality assessment model.

D. Loading the KoNViD Dataset and DOVER Model

The process begins by loading the KoNViD dataset into Google Colab. This dataset consists of a collection of videos paired with corresponding quality labels. Next, the pre-trained DOVER-Mobile model is loaded. This model is a streamlined variant of the original DOVER, specifically designed for resource-constrained environments. A key feature of DOVER-Mobile is its use of the convnext_v2_femto (inflated) backbone across both branches, resulting in a significantly reduced model size.

DOVER-Mobile exemplifies efficient machine learning model design, containing only 9.86 million parameters—approximately 5.7 times fewer than the standard DOVER model. This reduction lowers computational complexity (52.3 GFLOPs) and memory usage (under 1.9 GB), enabling efficient execution on devices with limited resources. The model achieves notable performance, processing a single video in just 1.4 seconds, compared to 3.6 seconds for the standard model. These advantages make DOVER-Mobile well-suited for mobile and edge computing applications [8].

E. Adjusting the Model (Head-Only Transfer Learning)

Head-only transfer learning is a targeted technique that involves modifying only the head of the DOVER-Mobile model to adapt it to the specific characteristics of the KoNViD dataset. The head is responsible for generating the final output—in this case, video quality predictions. By adjusting this component, the model can more accurately map features extracted by the backbone to the quality attributes found in the KoNViD dataset.

The adaptation process involves retraining the head using the KoNViD dataset while preserving the pre-trained features of the backbone. Training is conducted over 10 epochs, each representing a full pass through the dataset. The key objectives of this approach are to improve prediction accuracy, reduce

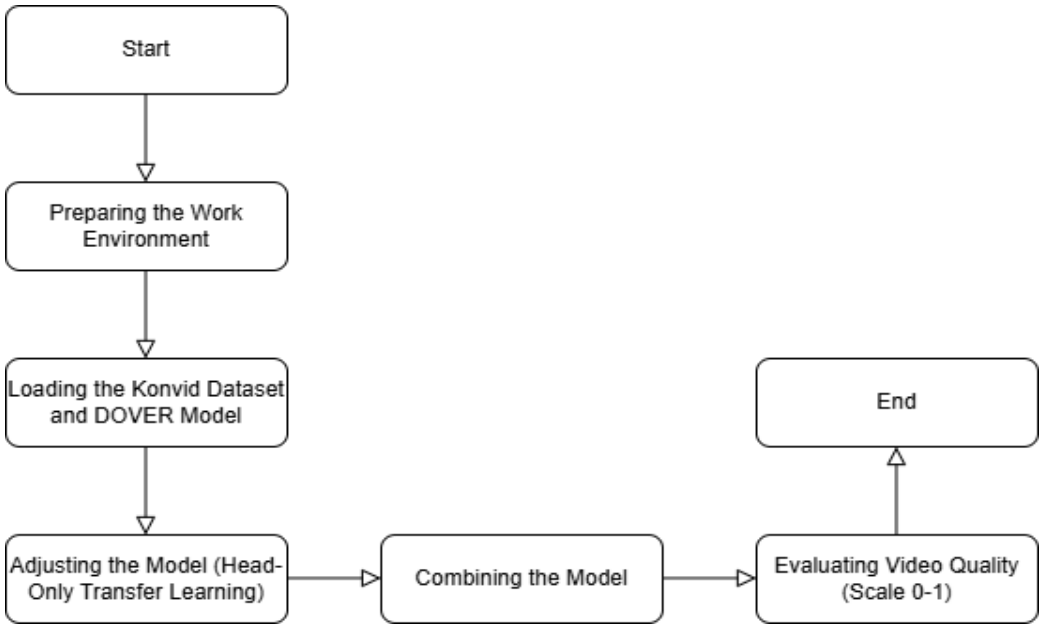

Figure 1. Flowchart of the Algorithm

training time, and optimize resource efficiency by leveraging the existing knowledge embedded in the DOVER-Mobile model. By limiting training to the head, the method creates a more efficient and adaptable model for video quality assessment that can better capture the unique patterns of the KoNViD dataset.

The model is set to load pretrained weights, enabling it to begin training from an already established state. Training is configured for 10 epochs with a learning rate of 1e-3, and a 1e-1 multiplier is applied to the backbone network. A warmup period of 2.5 epochs is included to stabilize the early stages of training. To prevent overfitting, a weight decay of 0.05 is applied. The model The model also uses an Exponential Moving Average (EMA) of weights, which contributes to model stability by incorporating a history of weight updates rather than relying solely on the most recent ones. The best-performing model during training is saved based on validation performance. Data is processed in batches of 8, with 2 worker processes handling data loading. A fixed random seed of 42 is used to ensure reproducibility in data splitting.

A frame interval of 2 is applied during training to capture sufficient temporal information while avoiding redundancy by sampling every second frame. The model processes clips of 32 frames, enabling it to learn motion dynamics and temporal dependencies, a standard practice in deep learning for video analysis [2].

For the KoNViD dataset, which includes 1200 videos, two types of features are extracted: technical and aesthetic. Technical features are derived by segmenting each video into 7×7 fragments (32×32 each) and extracting three 32-frame clips at a frame interval of 2. Aesthetic features are extracted by resizing each frame to 224×224 and processing a single 32-frame clip using the same frame interval. The aesthetic fragments are also 32 frames long. In the loss calculation, the KoNViD dataset is assigned a weight of 0.540, influencing its contribution to the overall training objective.

The dataset is divided into 960 videos for training and 240 for validation. These configurations are designed to enhance the model's ability to extract meaningful technical and aesthetic features while maintaining computational efficiency.

F. Combining the Model

The model integration process is designed to enhance accuracy and reliability in video quality evaluation by replacing the original DOVER model's head with a custom-trained component tailored to the KoNViD dataset. This process involves a detailed examination of the model structure, which includes two key components: state_dict (containing model parameters, weights, and biases) and validation_results (providing performance metrics from training). The model includes layers such as technical_backbone.head, aesthetic_backbone.head, technical_head, and aesthetic_head, each holding weights and biases in tensor format.

Using PyTorch, the integration replaces these specific layers in the original DOVER model with weights from the head-only transfer learning model. By substituting the weights in technical_backbone.head, aesthetic_backbone.head, technical_head, and aesthetic_head, the model is recalibrated to better capture the characteristics of the KoNViD dataset. The resulting refined model is capable of evaluating video quality—particularly user-generated content—with greater precision. The objective is to develop a more adaptable, context-aware video quality assessment tool that delivers more nuanced and reliable evaluations across a variety of video types and sources.

G. Evaluating Video Quality (Scale 0-1)

The final integrated model predicts video quality by generating a numerical score between 0 and 1, with higher scores indicating better quality. Evaluation begins by segmenting the input video into smaller clips, which are then analyzed to produce two scores: a technical score assessing objective attributes such as resolution and frame rate, and an aesthetic score reflecting subjective aspects like composition and color.

These scores are weighted and combined to form a comprehensive quality assessment. This integrated evaluation approach supports applications such as video encoding optimization, content curation, and quality control across diverse multimedia platforms.

H. Flowchart of the Algorithm

The process of no-reference video quality assessment using transfer learning can be visualized through a flowchart, as shown in Figure 1, which consists of five main stages. The first stage is the preparation of the working environment, involving the setup of essential software and libraries such as Python and machine learning frameworks. The second stage includes loading the KoNViD dataset and the pre-trained DOVER-Mobile model. The KoNViD dataset supplies training data in the form of videos and corresponding quality labels, while the DOVER-Mobile model provides the base architecture for further development.

The third stage focuses on fine-tuning the DOVER-Mobile model using the head-only transfer learning approach. In this step, only the model's head—which is responsible for generating final predictions—is retrained using the KoNViD dataset to align more closely with the dataset's unique characteristics. In the fourth stage, the newly trained head is integrated into the original DOVER-Mobile model by replacing its existing head layers. This results in a combined model, specifically adapted for the target dataset.

The final stage is video quality evaluation. The integrated model processes input videos and outputs a quality score between 0 and 1, where higher values indicate better quality. This evaluation takes into account both technical aspects, such as resolution and frame rate, and aesthetic elements, including composition and color.

III. Result and Discussion

A. Evaluation Metrics

The evaluation of video quality prediction models relies on three key statistical metrics that collectively measure performance. Each metric focuses on a distinct aspect of correlation and ranking, offering a comprehensive view of the model's predictive accuracy. By analyzing the Spearman Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), and Kendall Rank Correlation Coefficient (KRCC), researchers can assess how well the model captures relationships between predicted and actual video quality.

B. SRCC Evaluation

The Spearman Rank Correlation Coefficient (SRCC) measures the model's ability to rank videos from highest to lowest quality. As shown in Figure 2, the SRCC graph indicates significant progress during the early stages of training. The SRCC value rose rapidly, reflecting the model's learning effectiveness and improvement in prediction ranking. After approximately 1,500 iterations, the SRCC stabilized at 0.86, suggesting that the model had effectively learned complex patterns in the data and achieved high ranking accuracy without signs of overfitting.

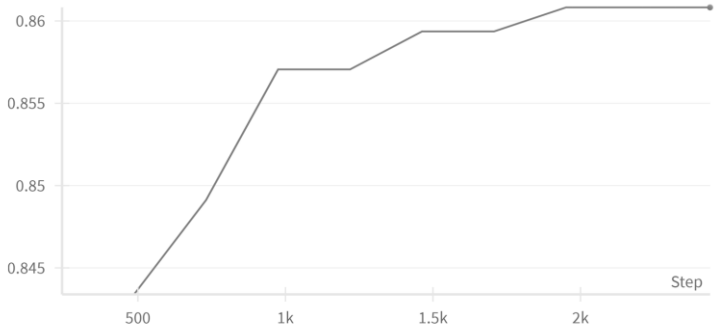

Figure 2. SRCC Graph

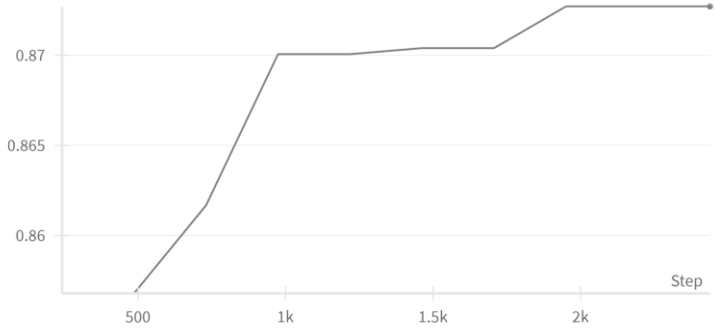

Figure 3. PLCC Graph

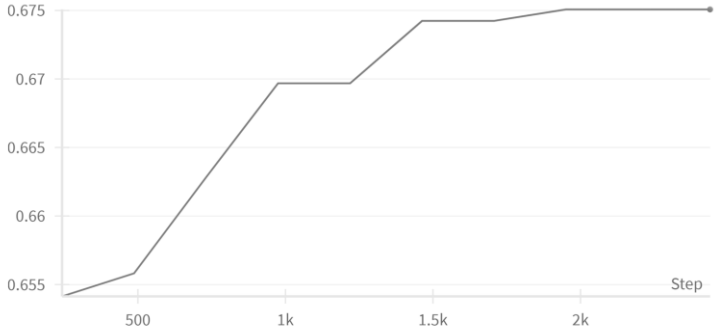

Figure 4. KRCC Graph

C. PLCC Evaluation

The Pearson Linear Correlation Coefficient (PLCC) assesses the strength of the linear relationship between the predicted and actual quality scores. As shown in Figure 3, the PLCC graph exhibits a similar trend to SRCC, with marked improvement during initial training. The model ultimately reached a peak PLCC value of 0.87, indicating a strong alignment between predictions and actual video quality scores. The plateau after 1,500 iterations confirms the model's stability and consistent predictive performance.

D. KRCC Evaluation

The Kendall Rank Correlation Coefficient (KRCC) evaluates the consistency of the model in assigning higher scores to higher-quality videos. As depicted in Figure 4, the KRCC graph shows a pattern of rapid initial improvement followed by stabilization. The model achieved a maximum KRCC value of 0.675, indicating a strong correlation between predicted and actual rankings. The consistent performance after reaching the plateau reflects the model's ability to learn ranking relationships effectively without overfitting.

E. Trade-offs in Epoch Selection for Transfer Learning

The decision to train only the head layers for 10 epochs aligns with the core principles of Deep Transfer Learning (DTL). DTL addresses limitations of traditional Deep Learning (DL) models—such as

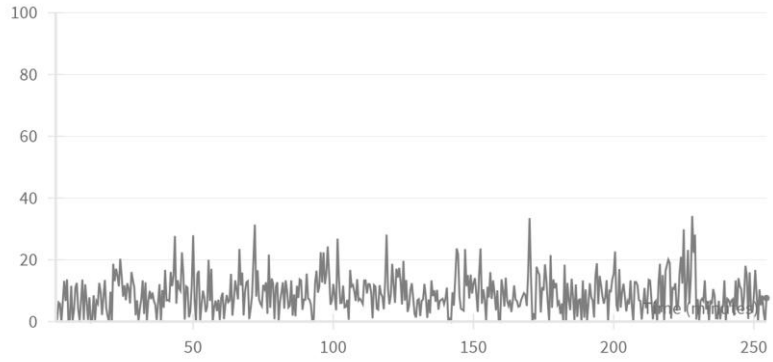

Figure 5. GPU Utilization Graph

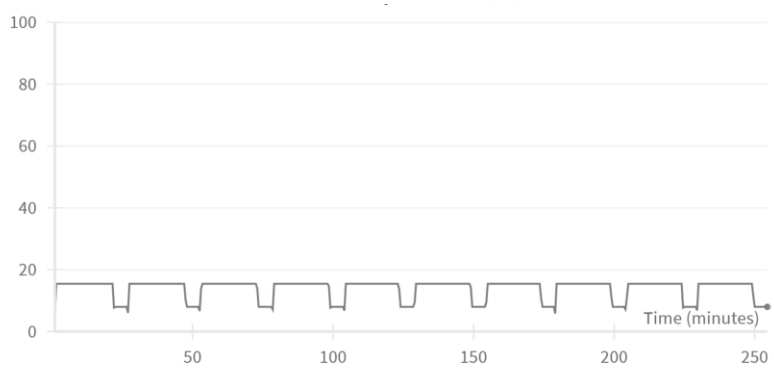

Figure 6. GPU Memory Allocated Graph

reliance on large labeled datasets and high training costs—by leveraging pre-trained models to accelerate learning on new target data [6]. This approach reduces computational demands, making it suitable for resource-constrained environments.

In head-only transfer learning, freezing the early layers of a pre-trained model helps prevent catastrophic forgetting, where extensive weight updates may erase previously acquired knowledge. Instead, training focuses on the fully connected head layers, allowing the model to adapt to the target dataset while retaining general feature representations learned in earlier layers [6]. Since earlier layers typically extract generic features and later layers handle task-specific learning [6], fine-tuning the head ensures a focused and efficient adaptation.

The SRCC, PLCC, and KRCC results in this study support the adequacy of 10 training epochs. Rapid performance gains during early training and stabilization around 1,500 iterations (approximately 10 epochs) indicate effective learning and convergence. Extending training beyond this point would likely increase computational costs without meaningful performance improvements, and may risk overfitting.

Progressive learning techniques—which extend pre-trained models by adding new layers for task-specific purposes—also underscore the importance of preserving early layer knowledge while allowing for specialized adaptation [19]. The head-only transfer learning approach reflects this principle by promoting efficiency, stability, and adaptability within a controlled training duration.

In summary, the use of 10 epochs achieves a practical balance between computational efficiency and model accuracy. It leverages the strengths of DTL by maintaining foundational knowledge and refining only the task-specific components, minimizing the risks of catastrophic forgetting and overfitting.

F. Training Efficiency and Resource Utilization

The training process, which lasted 4 hours, 14 minutes, and 42 seconds, provides valuable insights into resource utilization and performance. As shown in Figure 5, GPU utilization consistently remained below 30%, indicating that the system did not fully leverage the available computational capacity. However, this observation warrants a nuanced interpretation.

On the one hand, the low GPU utilization and minimal memory allocation, as shown in Figures 5 and 6, suggest the training process had computational headroom—potentially allowing for further optimization. On the other hand, this could also reflect the model's efficiency, where the training objectives were

TABLE 1
COMPARISON OF MODEL PERFORMANCE

	SRCC	PLCC	KRCC
Model after transfer learning integration	0.861	0.873	0.675
NR DOVER	0.9169	0.9208	0.7663
NR FAST-VQA	0.8694	0.8785	0.6977
NR FasterVQA	0.8219	0.8432	0.6402
NR VIDEVAL	0.7267	0.7703	0.5396

TABLE 2
COMPARISON OF MODEL PERFORMANCE **ON THE** VSC2022 DATASET

Video ID	DOVER Model	Model after transfer learning integration
Video 1	0.602	0.560
Video 2	0.614	0.572
Video 3	0.536	0.502
Video 4	0.602	0.550
Video 5	0.561	0.522
Video 6	0.513	0.497
Video 7	0.489	0.461
Video 8	0.602	0.531
Video 9	0.508	0.476
Video 10	0.614	0.564

Figure 7. Sample Frame of the Video

achieved without overburdening the GPU. Factors contributing to this may include algorithmic efficiency, optimized data loading, or a lightweight model architecture that demands less computational power.

While increasing GPU utilization might improve training speed, it is important to recognize that the current usage level could represent an effective balance between performance and efficiency. A more in-depth analysis should examine factors influencing GPU performance, such as batch size, data preprocessing methods, and code-level optimizations, to determine whether any adjustments are warranted. The goal is to identify the optimal configuration that minimizes training time without incurring unnecessary computational costs.

G. Comparison of Model Performance

The performance comparison of video quality prediction models reveals strong results across multiple evaluation metrics. This study assessed the ability of different machine learning models to predict and rank video quality using three key statistical measures: Spearman Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), and Kendall Rank Correlation Coefficient (KRCC). The transfer learning model achieved an SRCC of 0.861, a PLCC of 0.873, and a KRCC of 0.675.

These metrics reflect the model's predictive capabilities. The high SRCC indicates a strong ability to rank videos accurately from highest to lowest quality. The PLCC value of 0.873 demonstrates a strong linear correlation between predicted scores and actual quality ratings. Meanwhile, the KRCC value of 0.675 confirms the model's consistency in assigning higher scores to higher-quality videos.

When compared with other models—such as NR DOVER, NR FAST-VQA, NR FasterVQA, and NR VIDEVAL—as shown in Table 1, the transfer learning model demonstrates competitive performance

[20]. These results suggest that transfer learning effectively leverages prior knowledge from large datasets to enhance video quality prediction. Its performance can be attributed to key strengths such as computational efficiency, high accuracy, and flexibility through fine-tuning of the model's final layers.

H. Evaluating Video Quality

The final model, developed through the integrated transfer learning approach, predicts video quality using a numerical score between 0 and 1, with higher values indicating better quality. The evaluation process involves dividing each video into smaller segments, which are analyzed to generate two separate scores: a technical quality score that captures objective factors such as resolution and frame rate, and an aesthetic quality score that reflects subjective aspects like composition and color. These scores are weighted and combined to produce an overall quality rating. This final score can support various applications, including video encoding optimization and quality-based content selection.

A comparison between the original DOVER model and the enhanced model using transfer learning reveals subtle differences in scoring. For the same video, the transfer learning model produced a score of 0.587, while the original DOVER model yielded 0.643—a variation of 0.056. This suggests that the transfer learning process may have led the model to adopt a more conservative evaluation approach, possibly applying stricter criteria to aspects such as resolution, frame rate, composition, and color. Further research is needed to identify which specific evaluation parameters became more sensitive following transfer learning.

I. Evaluating Video Quality from Other Datasets

This study also evaluated video quality using 10 sample videos from the VSC2022 dataset, selected for its diverse visual characteristics and rigorous curation process. The dataset includes videos modified through resizing, cropping, and filtering, and adheres to strict inclusion criteria—each video must have a minimum resolution of 320×320 pixels and a duration of at least 5 seconds. To ensure a focus solely on visual quality, audio was removed, and adaptive blur was applied to regions depicting humans [21].

The evaluation compared the original DOVER model with the model enhanced through transfer learning. As shown in Table 2, the transfer learning model consistently produced lower quality scores across all 10 videos, with an average difference of 0.04 points. Despite this, both models preserved the same relative ranking of video quality: videos 2 and 10 received the highest scores (0.614 from the original model), while video 7 received the lowest (0.489). The findings suggest that the transfer learning model became more conservative and sensitive to subtle visual variations. This is particularly evident in videos 1, 4, and 8, which received identical scores from the original model but different scores from the adapted model. These differences point to the development of more refined and nuanced evaluation criteria in the transfer learning-based assessment model.

This indicates that videos 2 and 10, featuring a grassy plain and a person on a beach, respectively (sample frames shown in Figure 7), consistently received the highest evaluation scores. In contrast, video 7, a collage of four unrelated clips placed in the corners of the frame with minimal visual coherence, received the lowest score (also shown in Figure 7).

While the VSC2022 dataset provided a controlled environment for evaluation, the generalizability of the findings to more diverse video datasets remains an important consideration. The lower scores produced by the transfer learning model suggest increased sensitivity to dataset-specific visual characteristics. Future research should explore the model's performance on datasets with varying content types, compression levels, and resolutions, and investigate how preprocessing choices such as audio removal and adaptive blurring affect outcomes. This will help determine whether the model's heightened sensitivity translates into accurate video quality assessment across a broader range of real-world scenarios.

J. Future Directions and Related Work

Future research should explore the seamless integration of the DOVER framework with head-only transfer learning into real-time video streaming workflows. This would enable continuous, accurate quality assessment with minimal computational overhead. The development of multi-dimensional VQA models, as described in [22], which use pre-trained CNNs and distortion-specific metrics to evaluate content meaning, visual artifacts, and motion, should also be considered. However, such methods often require high computational resources, particularly in real-time applications. The optimized head-only transfer learning model proposed in this study presents a promising alternative for incorporating comprehensive quality evaluations with reduced resource demands.

Given the growing importance of Quality of Experience (QoE) in video streaming, as highlighted in [23], continued research into efficient and accurate VQA methods remains essential. Building on server-side QoE estimation frameworks, such as those proposed in [24], which combine network parameters and visual features using models like PatchVQ, future work should investigate integrating the DOVER framework into live streaming environments. This would support real-time video quality adjustments, enhancing the viewer's experience.

Subsequent research should prioritize the development of adaptive streaming systems that use VQA scores from the DOVER model to dynamically adjust bitrate and resolution to optimize QoE. Integrating additional features, similar to those used in [22], could improve the model's accuracy and robustness while preserving computational efficiency. Large-scale user studies correlating VQA scores with subjective QoE ratings will be essential to validate the model's ability to reflect user perception accurately. Furthermore, deploying the VQA model on edge computing platforms should be explored to reduce latency and enhance scalability, enabling real-time quality evaluation closer to end users [23]. Expanding the dataset to include a broader range of distortions and video content will also enhance the model's generalizability and reliability.

IV. CONCLUSION

This study successfully developed a no-reference video quality assessment method using the DOVER framework combined with a transfer learning approach. The resulting model strong performance in video quality prediction, achieving an SRCC of 0.861, PLCC of 0.873, and KRCC of 0.675. These metrics indicate a high correlation between the model's predictions and actual video quality scores. By applying head-only transfer learning, the model effectively captured complex patterns in video data and delivered accurate assessments of both technical and aesthetic quality. In tested samples, it produced a quality score of 0.587, closely aligning with perceived visual quality. These findings highlight the potential of the DOVER framework and transfer learning in building a reliable video quality assessment model applicable to digital multimedia and video streaming environments.

Based on the findings, several directions for further research are recommended. First, incorporating additional features may improve the model's predictive accuracy. Second, expanding the dataset to include more diverse content and detailed annotations could enhance generalizability. Third, the model holds promise for practical applications such as video compression optimization, development of video editing tools, and integration with streaming platforms. Lastly, investigating factors like content variation, lighting, and noise can offer deeper insights into the model's performance boundaries and areas for refinement.

REFERENCES

[1] M. L. B. dos Santos, "The 'so-called' UGC: an updated definition of user-generated content in the age of social media," *Online Information Review*, vol. 46, no. 1, pp. 95–113, 2022.

[2] H. Wu *et al.*, "Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives," in *International Conference on Computer Vision (ICCV)*, 2023.

[3] F. Fan, C. Luo, W. Gao, and J. Zhan, "AIGCBench: Comprehensive Evaluation of Image-to-Video Content Generated by AI." 2024. [Online]. Available: https://arxiv.org/abs/2401.01651

[4] B. Qu, X. Liang, S. Sun, and W. Gao, "Exploring AIGC Video Quality: A Focus on Visual Harmony, Video-Text Consistency and Domain Distribution Gap." 2024. [Online]. Available: https://arxiv.org/abs/2404.13573

[5] H. Wu *et al.*, "FAST-VQA: Efficient End-to-end Video Quality Assessment with Fragment Sampling," in *Proceedings of European Conference of Computer Vision (ECCV)*, 2022.

[6] M. Iman, H. R. Arabnia, and K. Rasheed, "A Review of Deep Transfer Learning and Recent Advancements," *Technologies*, vol. 11, no. 2, 2023, doi: 10.3390/technologies11020040.

[7] U. Evci, V. Dumoulin, H. Larochelle, and M. C. Mozer, "Head2toe: Utilizing intermediate representations for better transfer learning," in *International Conference on Machine Learning*, PMLR, 2022, pp. 6009–6033.

[8] H. Wu, "Open Source Deep End-to-End Video Quality Assessment Toolbox." 2022. [Online]. Available: http://github.com/timothyhtimothy/fast-vqa

[9] V. Hosu *et al.*, "The Konstanz natural video database (KoNViD-1k)," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–6. doi: 10.1109/QoMEX.2017.7965673.

[10] M. Canesche, L. Bragança, O. P. V. Neto, J. A. Nacif, and R. Ferreira, "Google Colab CAD4U: Hands-On Cloud Laboratories for Digital Design," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5. doi: 10.1109/ISCAS51556.2021.9401151.

[11] A. Rawat, "A Review on Python Programming," *International Journal of Research in Engineering, Science and Management*, vol. 3, no. 12, pp. 8–11, Dec. 2020.

[12] O. Keleş, M. A. Yılmaz, A. M. Tekalp, C. Korkmaz, and Z. Doğan, "On the Computation of PSNR for a Set of Images or Video," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5. doi: 10.1109/PCS50896.2021.9477470.

[13] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Structural similarity index (SSIM) revisited: A data-driven approach," *Expert Systems with Applications*, vol. 189, p. 116087, 2022, doi: https://doi.org/10.1016/j.eswa.2021.116087.

[14] X. Min, H. Duan, W. Sun, Y. Zhu, and G. Zhai, "Perceptual video quality assessment: a survey," *Science China Information Sciences*, vol. 67, no. 11, p. 211301, Oct. 2024, doi: 10.1007/s11432-024-4133-3.
[15] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, "Learning Generalized Spatial-Temporal Deep Feature Representation for No-Reference Video Quality Assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1903–1916, 2022, doi: 10.1109/TCSVT.2021.3088505.
[16] S. Dost, F. Saud, M. Shabbir, M. G. Khan, M. Shahid, and B. Lovstrom, "Reduced reference image and video quality assessments: review of methods," *EURASIP Journal on Image and Video Processing*, vol. 2022, no. 1, p. 1, Jan. 2022, doi: 10.1186/s13640-021-00578-y.
[17] D. Li, T. Jiang, and M. Jiang, "Unified Quality Assessment of in-the-Wild Videos with Mixed Datasets Training," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1238–1257, Apr. 2021, doi: 10.1007/s11263-020-01408-w.
[18] NVIDIA Corporation, "NVIDIA T4 Tensor Core GPU." [Online]. Available: https://www.nvidia.com/en-us/data-center/tesla-t4/
[19] M. Iman, H. R. Arabnia, and R. M. Branchinst, "Pathways to Artificial General Intelligence: A Brief Overview of Developments and Ethical Issues via Artificial Intelligence, Machine Learning, Deep Learning, and Data Science," in *Advances in Artificial Intelligence and Applied Cognitive Computing*, H. R. Arabnia, K. Ferens, D. De La Fuente, E. B. Kozerenko, J. A. Olivas Varela, and F. G. Tinetti, Eds., in Transactions on Computational Science and Computational Intelligence, Cham: Springer International Publishing, 2021, pp. 73–87. doi: 10.1007/978-3-030-70296-0_6.
[20] A. Antsiferova, S. Lavrushkin, M. Smirnov, A. Gushchin, D. Vatolin, and D. Kulikov, "Video compression dataset and benchmark of learning-based video-quality metrics," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Curran Associates, Inc., 2022, pp. 13814–13825. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/59ac9f01ea2f701310f3d42037546e4a-Paper-Datasets_and_Benchmarks.pdf
[21] E. Pizzi *et al.*, "The 2023 video similarity dataset and challenge," *Computer Vision and Image Understanding*, vol. 243, p. 103997, 2024, doi: https://doi.org/10.1016/j.cviu.2024.103997.
[22] Z. Zhang *et al.*, "MD-VQA: Multi-dimensional quality assessment for UGC live videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1746–1755.
[23] K. Bouraqia, E. Sabir, M. Sadik, and L. Ladid, "Quality of Experience for Streaming Services: Measurements, Challenges and Insights," *IEEE Access*, vol. 8, pp. 13341–13361, 2020, doi: 10.1109/ACCESS.2020.2965099.
[24] G. Margetis, G. Tsagkatakis, S. Stamou, and C. Stephanidis, "Integrating Visual and Network Data with Deep Learning for Streaming Video Quality Assessment," *Sensors*, vol. 23, no. 8, p. 3998, Apr. 2023, doi: 10.3390/s23083998.