Journal of Business, Social and Technology
Volume 7, Issue 3, 1251-1269
e_ISSN: 2807-6362
https://bustechno.polteksci.ac.id/index.php/jbt/index
DOI: doi.org/10.59261/jbt.v7i3.737

Photo Aesthetic Assessment via Spatial and Frequency Domain Fusion Using a Vision Transformer Approach

***Reza Rachmadan**
Institut Teknologi Sepuluh Nopember, Indonesia

Shintami Chusnul Hidayati
Institut Teknologi Sepuluh Nopember, Indonesia

***Corresponding author:**
Reza Rachmadan, Institut Teknologi Sepuluh Nopember, Indonesia.
✉6025231073@student.its.ac.id

Article Info:
Article history:
Received: June 22, 2026
Revised: July 30, 2026
Accepted: August 11, 2026
Available Online: August 20, 2026

Keywords:
AADB Datasets; Adaptive Gate Fusion; Concatenation Fusion; Dual-branch; Fusion; Fast Fourier Transform; Image Aesthetic Assessment; Vision Transformer.

Abstract

Background: The rapid growth of digital media has increased the demand for automated image aesthetic assessment (IAA) in social media, creative industries, and e-commerce platforms. Although Vision Transformer (ViT)-based models have demonstrated promising performance, most existing approaches rely primarily on spatial representations while overlooking frequency-domain information, which captures complementary characteristics such as sharpness, texture, noise patterns, and bokeh effects.
Objective: This study proposes a Dual-Branch Late Fusion FFT-ViT architecture with concatenation-based fusion as an approach for integrating dual-domain representations to improve *photo aesthetic assessment* (PAA).
Methods: The proposed architecture consists of two parallel branches. The RGB branch employs a Vision Transformer (*ViT-Small*) to extract spatial and compositional features, while the FFT branch utilizes *ViT-Tiny* to capture frequency-domain characteristics associated with texture, sharpness, and image details. The extracted features from both branches are fused using a concatenation strategy before being passed to the regression layer.
Results: The proposed Dual-Branch FFT-ViT with concatenation fusion achieved the best performance, obtaining a PLCC of 0.7336, SRCC of 0.7347, MSE of 0.0193, MAE of 0.1117, and RMSE of 0.1388. Compared with the RGB-only ViT baseline, the proposed model improved the PLCC score by 0.0124, demonstrating the effectiveness of integrating spatial and frequency-domain features for aesthetic score prediction.
Conclusion: This study demonstrates that integrating spatial and frequency-domain representations through a dual-branch Vision Transformer architecture enhances *photo aesthetic assessment* performance.

To cite this article: Rachmadan, R., & Hidayati, S. C. (2026). Photo Aesthetic Assessment via Spatial and Frequency Domain Fusion Using a Vision Transformer Approach. *Journal of Business, Social and Technology*, 7(3), 1251-1269. https://doi.org/10.59261/jbt.v7i3.737

INTRODUCTION

Digital photos have become an indispensable part of modern life. The rapid growth of social media platforms and digital image-sharing services has significantly increased the production, circulation, and consumption of visual content. Images uploaded to platforms such as Instagram, Pinterest, Facebook, and other online media are not only used for communication and documentation but also for personal branding, marketing, entertainment, and creative expression. In this context, some photographs are perceived as more beautiful, attractive, or

visually appealing than others. Such aesthetic judgments are naturally performed by humans based on visual experience, including composition, lighting, color harmony, sharpness, subject arrangement, and overall emotional impression (Anwar et al., 2021; Zhou et al., 2022) .

Image Aesthetic Assessment (IAA) is a research field in artificial intelligence and computer vision that aims to enable computers to predict or replicate human aesthetic judgments of images. IAA has gained increasing attention because aesthetic quality is closely related to user experience, digital content recommendation, photo ranking, automatic image enhancement, and intelligent photography assistance. Recent studies have shown that deep learning-based IAA models can learn complex visual representations associated with photographic quality, artistic attributes, and human preferences more effectively than traditional handcrafted feature-based approaches (He et al., 2022; Jia et al., 2025; Ke et al., 2021). Therefore, the development of more accurate IAA models is essential for supporting real-world applications, including automated photo recommendation systems, creative content curation, artificial intelligence-based image quality enhancement, and instant feedback tools for beginner photographers.

Recent advances in IAA and image quality assessment have been strongly influenced by deep neural networks, particularly Transformer-based architectures. Vision Transformer-based models are increasingly adopted because they can capture long-range dependencies and global image structures that are important for understanding visual aesthetics. Studies such as MUSIQ demonstrate that Transformer-based models can process image quality information at different granularities and preserve global-local visual characteristics more effectively than conventional fixed-size processing approaches (Ke et al., 2021; Zhang et al., 2026). More recent studies also indicate that Vision Transformers can improve aesthetic prediction by preserving composition, aspect ratio, high-resolution details, and multi-scale visual information, all of which are highly relevant to aesthetic perception (Behrad et al., 2025). However, most current approaches still primarily focus on color and spatial-domain image information, while other potentially valuable representations remain underexplored.

One important representation that has not been fully utilized in IAA is frequency-domain information. Frequency-domain analysis using the Fast Fourier Transform (FFT) provides complementary visual information that cannot be completely captured from RGB images alone. FFT-based representations can describe texture patterns, edge structures, image sharpness, spectral energy distribution, and noise characteristics, which complement spatial-domain representations (Jiang et al., 2025). Previous image quality assessment studies have demonstrated that combining multiple visual perspectives, including aesthetic attributes, distortion characteristics, saliency information, and technical image features, can lead to more comprehensive quality prediction (Sun et al., 2024). This finding suggests that aesthetic prediction may benefit from a dual-representation approach that integrates spatial-domain features with frequency-domain characteristics.

Early Image Aesthetic Assessment (IAA) methods relied on handcrafted features, such as composition, color, symmetry, and texture. However, these features were insufficient to represent the complex and subjective nature of human aesthetic perception, motivating the adoption of deep learning approaches. The introduction of convolutional neural networks (CNNs) enabled automatic learning of aesthetic representations from large-scale datasets. Although CNN-based methods substantially improved prediction accuracy, they primarily captured local visual patterns and were less effective in modeling long-range relationships that influence image aesthetics, such as global composition, object placement, and scene structure (He et al., 2022; Ke et al., 2021).

Recent studies have increasingly adopted Transformer-based architectures because self-attention mechanisms effectively model global image relationships. Existing approaches have incorporated semantic attributes, photographic themes, saliency information, and technical quality features to improve aesthetic prediction performance (He et al., 2022; L. Li et al., 2023; Z. Li et al., 2025; J. Shi et al., 2024). Despite these advances, most existing methods still operate primarily on spatial-domain representations, leaving the integration of complementary frequency-domain information largely unexplored.

Vision Transformer (ViT) represents images as sequences of patches processed through self-attention mechanisms, enabling effective modeling of long-range spatial dependencies that

are important for Image Aesthetic Assessment. Unlike CNNs, ViT captures relationships between distant image regions, allowing more comprehensive representations of global composition and visual context.

Transformer-based models such as MUSIQ have demonstrated that global attention mechanisms improve image quality assessment and aesthetic prediction by preserving multi-scale visual information and image composition characteristics (Ke et al., 2021; Zhang et al., 2026). Subsequent studies further confirmed the effectiveness of ViT and multimodal Transformer architectures for aesthetic assessment through semantic-aware and attribute-aware representations (He et al., 2022; Z. Li et al., 2025; T. Shi et al., 2024). However, these approaches mainly learn representations from RGB image inputs and do not explicitly exploit frequency-domain information as an additional source of visual features.

Fast Fourier Transform (FFT) converts an image from the spatial domain into the frequency domain, enabling the analysis of image characteristics such as texture, edge information, sharpness, and noise patterns. These characteristics are closely related to technical image quality and can provide complementary information beyond pixel-level intensity representations. Recent studies have shown that frequency-domain representations improve image restoration and image quality assessment by preserving structural and texture-related information (Gao et al., 2025; Jiang et al., 2025; Sun et al., 2024). Nevertheless, the application of FFT-based representations as complementary features for Image Aesthetic Assessment remains limited, particularly within Transformer-based architectures.

Multi-branch architectures have been proposed to capture complementary visual information by processing different feature types in parallel before integrating them for prediction. Previous studies have incorporated semantic attributes, saliency information, and technical quality features, demonstrating that multiple visual perspectives can improve aesthetic assessment performance (He et al., 2022; L. Li et al., 2023; Sun et al., 2024).

Although these studies demonstrate the benefits of combining complementary information, most existing approaches still rely on multiple spatial-domain representations. The explicit fusion of RGB spatial features with FFT-based frequency representations within a Vision Transformer framework has received limited attention. This limitation motivates the proposed Dual-Branch FFT-ViT architecture, which integrates spatial and frequency-domain features to obtain richer image representations for photo aesthetic assessment.

This section presents a review of relevant literature and previous studies that serve as the primary references supporting the concepts. Furthermore, this review highlights the relevance of aesthetic annotation datasets, such as AADB and AVA, for evaluating the complexity of visual composition in landscape photography. A summary of the reviewed literature is presented in Table 1.

Table 1

Previous Research on CNNs and ViT for Photo Aesthetics

Researcher, Year	Model	Datasets	PLCC	SRCC
Kong dkk. (2016) ECCV	AADB-CNN	AADB (11K)	0.678	-
Talebi & Milanfar (2018) IEEE TIP and Tanawongsuwan (2026)	NIMA	AVA (255K)	0.612	0.592
Sheng dkk. (2018) and Zheng (2025) ACM MM	Attention-PAA	AVA & AADB	0.697	0.671
Ke dkk. (2021) ICCV	MUSIQ	AVA & AADB	0.728	0.726
Zeng dkk. (2019) IEEE TIP and Gonzalez-Naharro (2025)	Unified-PAA	AVA & AADB	0.689	0.659
Behrad dkk. (2025) CVPR	Dinov2-small (ViT Base)	AADB	0,695	0,682
Xu dkk. (2025) ICME	ViT-Tiny (NIMA)	AADB	0,703	0,712
He dkk. (2023)	ViT-Base	AVA	0,739	0,728

Table 1 summarizes eight previous studies in the field of photo aesthetic assessment that are relevant to this study, arranged in chronological order according to methodological development from 2016 to 2025. Significant technological advancements have been observed in the field of image aesthetic evaluation. In the early stages, most studies employed CNN-based models, such as AlexNet, InceptionV2, ResNet50, and VGG16. To enhance the models' ability to identify visually important regions within images, several studies incorporated attention mechanisms. Using these approaches, the PLCC scores ranged from 0.612 to 0.697. As deep learning architectures have evolved, more recent studies have adopted Vision Transformer (ViT), an architecture designed to capture long-range dependencies and model relationships among image patches more comprehensively. ViT is available in various configurations, ranging from lightweight versions, such as ViT-Tiny, to larger variants, such as ViT-Base. The results demonstrate improved performance compared with earlier CNN-based approaches, with PLCC values generally ranging from 0.695 to 0.739 (Behrad et al., 2025; He et al., 2023; Ke et al., 2021; Xu et al., 2025).

However, most existing image aesthetic assessment (IAA) approaches still primarily operate within the RGB spatial domain or combine RGB features with semantic, attribute-based, or saliency-based information. The explicit integration of spatial-domain ViT features with Fast Fourier Transform (FFT)-based frequency-domain features remains relatively unexplored, particularly within the context of the AADB dataset. This study addresses this research gap by proposing a Dual-Branch FFT-ViT architecture. The RGB branch employs ViT-Small to extract semantic, color, spatial, and compositional representations from image pixels, whereas the FFT branch utilizes ViT-Tiny to extract frequency-domain representations associated with image sharpness, texture characteristics, noise distribution, and spectral energy patterns.

The two representations are subsequently integrated through two fusion strategies: Concatenation Fusion and Adaptive Fusion. Concatenation Fusion is applied as a simple and interpretable fusion strategy to investigate whether spatial and frequency-domain representations provide complementary aesthetic information. In contrast, Adaptive Fusion employs a gating mechanism that enables the model to learn the relative contribution of each branch dynamically. By comparing these two fusion strategies, this study aims to evaluate whether dual-domain representation learning can improve aesthetic prediction performance while maintaining architectural interpretability and computational efficiency.

Collectively, previous studies demonstrate a growing trend toward improving image aesthetic assessment by learning richer visual representations through deep learning and Vision Transformer architectures. Although Transformer-based models consistently outperform traditional handcrafted features and conventional convolutional neural networks (CNNs) by capturing long-range spatial dependencies, most existing approaches still primarily rely on RGB images or incorporate additional semantic, saliency, or attribute-based information. These studies share the objective of improving aesthetic prediction through complementary visual cues; however, they differ in the types of information integrated into the learning framework. Although previous studies have demonstrated the effectiveness of Vision Transformers for image aesthetic assessment (Behrad et al., 2025; Ke et al., 2021) and explored complementary visual representations such as semantic attributes, saliency, and visual reasoning (L. Li et al., 2023; J. Shi et al., 2024), most existing approaches continue to depend mainly on RGB images and spatial-domain information (He et al., 2022; Ke et al., 2021). Consequently, the explicit integration of RGB-based spatial features with FFT-based frequency-domain information remains relatively unexplored in Image Aesthetic Assessment. Frequency-domain analysis provides complementary characteristics related to image sharpness, texture distribution, noise patterns, and spectral energy, which are difficult to capture using RGB information alone. Therefore, a research gap remains in developing a unified Vision Transformer framework capable of jointly exploiting spatial and frequency-domain representations for more accurate photo aesthetic assessment.

To address this gap, this study proposes a Dual-Branch FFT-ViT architecture that simultaneously processes two image representations through parallel branches. The first branch processes RGB images using the ViT-Small backbone to capture color information, semantic features, and spatial composition. The second branch processes the FFT-domain representation of the same image using the ViT-Tiny backbone to capture frequency-domain characteristics

related to texture, sharpness, and spectral energy distribution. The outputs of both branches are subsequently combined through two fusion strategies: simple concatenation-based fusion and adaptive fusion using a gating mechanism that learns the contribution of each branch. The research was conducted using the Aesthetic Attributes Database (AADB), which contains 9,958 photographs with human aesthetic score annotations and has been widely used in image aesthetic assessment research (He et al., 2022; Kong et al., 2016; Zhou et al., 2022).

The novelty of this study lies in the proposed Dual-Branch FFT-ViT architecture, which explicitly integrates spatial-domain and frequency-domain representations within a Vision Transformer framework for photo aesthetic assessment. Unlike previous studies that primarily rely on spatial information or semantic attributes, the proposed approach incorporates complementary frequency-domain characteristics through a parallel FFT branch and feature fusion mechanism. This integration provides richer image representations and contributes to improved aesthetic prediction performance on the AADB dataset.

The main contributions of this study are summarized as follows. First, a novel dual-branch architecture is proposed to integrate RGB spatial-domain and FFT-based frequency-domain representations within a Vision Transformer framework for Photo Aesthetic Assessment. Second, two feature fusion mechanisms, namely concatenation-based fusion and adaptive gate fusion, are systematically compared to investigate the effectiveness of different feature integration strategies. Third, comprehensive experiments conducted on the AADB dataset demonstrate that frequency-domain information provides complementary representations that improve aesthetic prediction performance compared with a single RGB-based branch.

METHOD

This study adopted an experimental model development approach to design, implement, and evaluate a Dual-Branch FFT-ViT architecture for photo aesthetic assessment. The overall research framework is illustrated in Figure 1, and each stage is described in the subsections that follow.

Figure 1 illustrates the overall research framework adopted in this study. The process began with the collection of images from the AADB dataset, followed by image preprocessing through resizing and bicubic interpolation. During the training process, random cropping was applied as a data augmentation technique to improve model generalization.

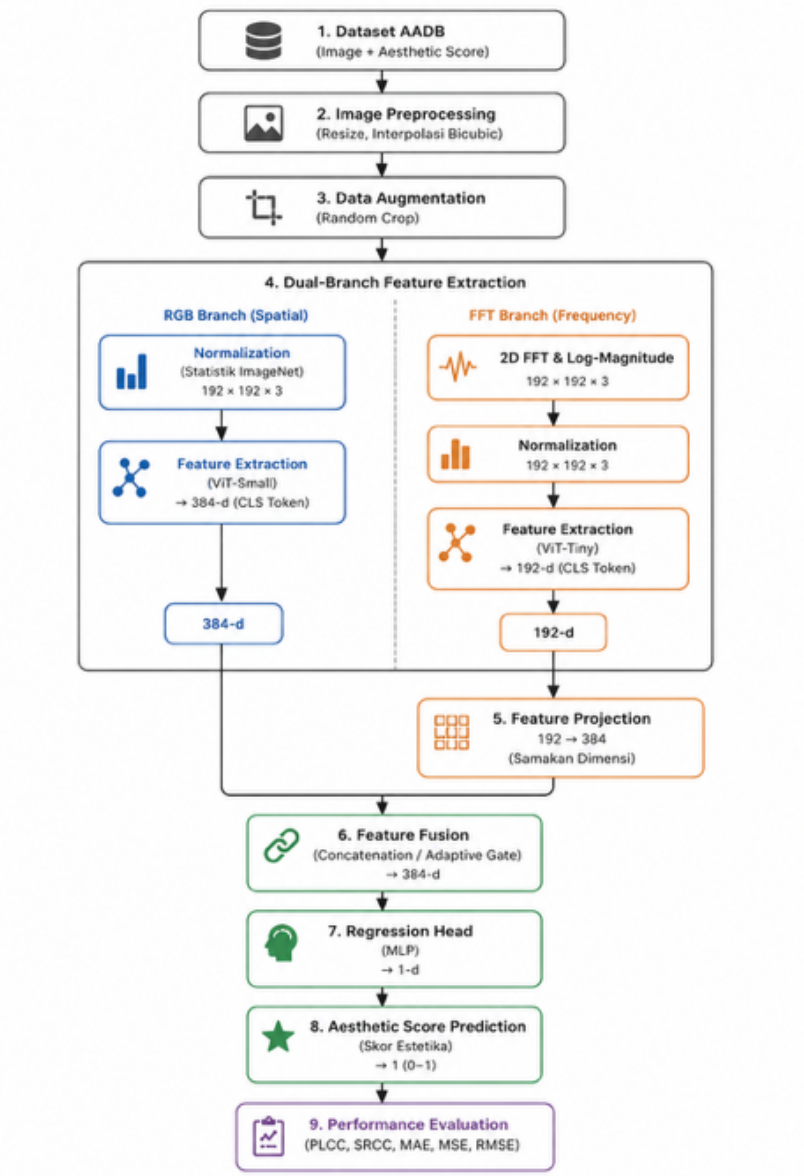

Figure 1. *Overall Research Framework of the Proposed Method*

The preprocessed images were processed in parallel through two complementary branches. The spatial branch normalized the RGB images using ImageNet statistics before extracting visual features with a ViT-Small backbone. Simultaneously, the frequency branch transformed the same images into the frequency domain using a two-dimensional Fast Fourier Transform (2D FFT) and logarithmic magnitude representation. The resulting frequency maps were normalized and processed using a ViT-Tiny backbone. Since the FFT branch generated a 192-dimensional feature vector, while the RGB branch generated a 384-dimensional feature vector, a feature projection layer was applied to align both representations within the same feature space.

Next, the projected FFT and RGB features were integrated using either Concatenation Fusion or Adaptive Gate Fusion to generate a unified 384-dimensional representation. Finally, the fused representation was passed through a regression head to predict the final aesthetic score. The effectiveness of the proposed framework was evaluated by comparing multiple experimental configurations using the PLCC, SRCC, MAE, MSE, and RMSE metrics.

Dataset AADB

This study used the Aesthetic Attributes Database (AADB) dataset developed by Kong et al. (2016) The dataset contains 9,958 photos collected from Flickr, representing a variety of

subjects, styles, and shooting conditions. Each photo is assigned an aesthetic score ranging from 0 to 1, obtained by aggregating ratings from multiple human judges through crowdsourcing platforms (Yang et al., 2025). Therefore, the assigned scores represent a collective human aesthetic consensus.

The dataset was divided into three partitions following the official split: 8,458 images (84.9%) for training, 500 images (5.0%) for validation, and 1,000 images (10.0%) for testing. The distribution of aesthetic scores across the three partitions was relatively balanced, with average scores of 0.511 for the training set, 0.517 for the validation set, and 0.525 for the testing set. This distribution ensured that no substantial bias existed among the dataset partitions.

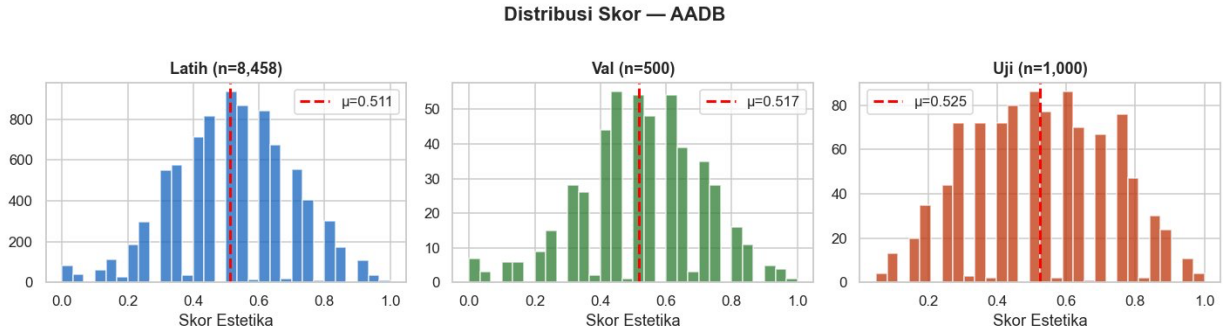

Figure 2. *Aesthetic Score Distribution of AADB Datasets on All Three Partitions (Train, Validate, Test)*

Figure 2 shows the distribution of aesthetic scores across all three data partitions. The shaped distribution resembled a bell curve, with peaks around values of 0.4 to 0.6, indicating that most photos had intermediate aesthetic qualities. Only a small percentage of photos received very low scores below 0.2 or very high scores above 0.8, which was consistent with the characteristics of the general aesthetic dataset.

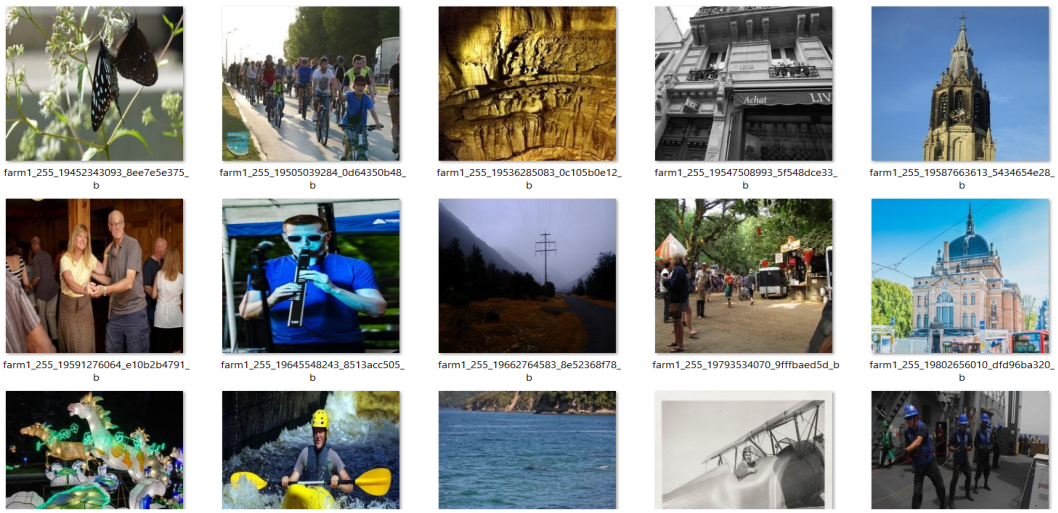

Figure 3. *Sample Photos from the AADB Datasets*

Data Pre-Proccessing

Each image underwent a series of pre-processing steps before being incorporated into the model. For the RGB branch, the images were resized to 256 × 256 pixels and then randomly cropped to 224 × 224 pixels during the training phase for data augmentation purposes. In the validation and testing phases, the images were directly resized to 224 × 224 pixels. Subsequently, normalization was applied using ImageNet statistics, with a mean (μ) of [0.485, 0.456, 0.406] and a standard deviation (σ) of [0.229, 0.224, 0.225].

Additional augmentation techniques were applied exclusively during the training phase to increase data diversity and reduce overfitting. These techniques included random horizontal

flipping with a probability of 0.5 and mild color variation using ColorJitter with an intensity of 0.15 for brightness, contrast, saturation, and hue adjustments. These augmentation techniques were not applied during validation and testing to ensure consistent and objective evaluation.

FFT Branch Feature Extraction

For the FFT branch, each image was processed through five stages of transformation. First, a 2D Fast Fourier Transform (FFT) was applied independently to each color channel (R, G, and B) using the torch.fft.fft2 function. Second, the resulting spectrum was shifted so that the low-frequency components were positioned at the center using the fftshift function. Third, the magnitude of the complex-valued spectrum was calculated. Fourth, a logarithmic transformation, defined as log(1 + |F|), was applied to reduce the excessive dynamic range of frequency magnitudes. Fifth, min–max normalization was performed to scale the values into the range of [0, 1]. The final output was a tensor with a size of (3, 224, 224), representing the distribution of frequency-domain energy across the three color channels.

Figure 4 presents examples of FFT feature extraction results from four images in the AADB dataset with different aesthetic scores. The results indicate that high-quality images with aesthetic scores above 0.75 tended to exhibit different spectral patterns compared with lower-quality images, particularly in the distribution of medium-frequency energy associated with image composition, texture characteristics, and object sharpness.

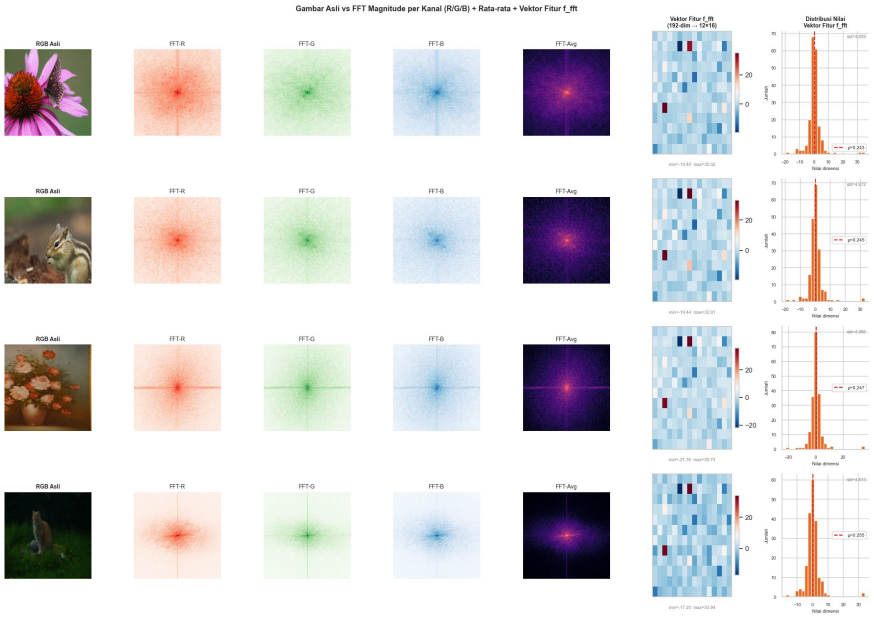

Figure 4. *Example of FFT Feature Extraction Results: Spectrum per Channel (R/G/B) and Feature Vector f_fft (192 dimensions)*

RGB Branch Feature Extraction

The ViT-Small backbone processed the normalized image by dividing it into 196 non-overlapping patches measuring 16 × 16 pixels, arranged in a 14 × 14 grid. Each patch was projected into a 384-dimensional embedding space and was subsequently processed together with a special classification token ([CLS]) through 12 Transformer blocks, each consisting of a multi-head self-attention mechanism and a feed-forward network. At the end of the processing stage, the [CLS] token was extracted as the global image representation, denoted as *image f_rgb ∈ R^384*.

Figure 5 presents an example of the RGB feature extraction pipeline for four images with different aesthetic scores. The feature vector f_{rgb} was visualized as a 16 × 24 heatmap representing 384 feature dimensions. The activation patterns within the feature vector exhibited consistent differences between the high-aesthetic images shown in the bottom row and the low-aesthetic images shown in the top row. This finding suggested that the ViT-Small backbone was capable of extracting representations relevant to aesthetic quality assessment..

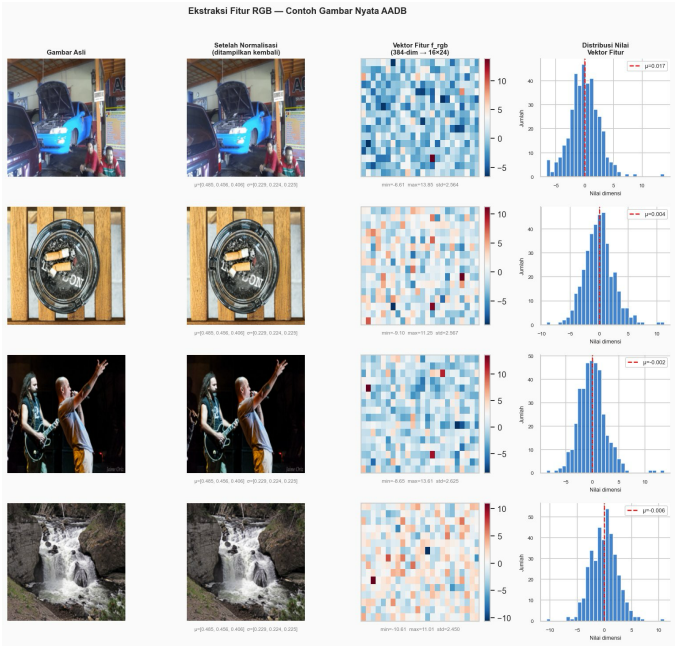

Figure 5. *RGB Feature Extraction Pipeline: Original Image, After Normalization, Feature Vector Heatmap f_rgb, and Its Value Distribution*

Proposed Model Architecture

The FFT-ViT Dual-Branch architecture consisted of four main blocks that operated sequentially. First, two backbone networks, namely ViT-Small for RGB inputs and ViT-Tiny for FFT inputs, processed the inputs in parallel and generated [CLS] tokens with dimensions of 384 and 192, respectively. Second, the LayerNorm and Linear projection layers transformed the 192-dimensional FFT feature representation into a 384-dimensional feature representation. The Linear and GELU layers were used to align the FFT feature dimensions with the RGB feature dimensions, enabling effective feature fusion. Third, the fusion module combined the two representations using one of the two fusion strategies evaluated in this study. Fourth, the regression head mapped the fused representation into aesthetic scores within the range of [0,1] through a progressive architecture consisting of LayerNorm(384), Linear(384 to 256), GELU, Dropout(0.3), Linear(256 to 128), GELU, Dropout(0.2), Linear(128 to 1), and a final Sigmoid activation function.

Two fusion strategies were evaluated in this study. First, Concatenation Fusion (DualNoGate) combined the RGB feature vector (384-dimensional) with the projected FFT feature vector (384-dimensional) into a single 768-dimensional vector. This combined representation was subsequently projected back to 384 dimensions through LayerNorm(768), Linear(768 to 384), and GELU activation. This approach assigned fixed and uniform contributions to both branches without learning adaptive weighting. Second, Adaptive Gate Fusion employed a gating mechanism that generated a dimension-wise weight vector of $\alpha \in (0,1)^{384}$ through a lightweight multilayer perceptron (MLP). The fused representation was calculated as:

$$f\_fused = \alpha \odot f\_rgb + (1-\alpha) \odot f\_fft$$

A value of α close to 1 indicated that the model relied more heavily on RGB features, whereas a value close to 0 indicated greater reliance on FFT features. A value around 0.5 suggested that both feature representations contributed relatively equally to the final prediction.

Training Protocol

The model was trained using Huber Loss with a parameter of $\delta = 0.1$, which provided greater robustness to outliers compared with the standard Mean Squared Error (MSE) loss. Optimization was performed using the AdamW optimizer with a weight decay coefficient of 1×10^{-4}. Three different learning rates were assigned to different model components: 1×10^{-5} for the RGB backbone to preserve pre-trained representations, 5×10^{-5} for the FFT backbone, and 3×10^{-4} for the feature fusion module and regression head. The learning rate was adjusted using Cosine Annealing, gradually decreasing to a minimum value of 1×10^{-7}.

During the first five training epochs, eight of the twelve RGB backbone blocks and six of the twelve FFT backbone blocks were frozen to allow the fusion module to stabilize before fine-tuning the entire network. After the warm-up stage, all network parameters were unfrozen and optimized jointly. Gradient clipping with a maximum norm of 1.0 and gradient accumulation with an effective batch size of 32 (16 samples per iteration with an accumulation step of 2) were applied to improve training stability. Early stopping with a patience value of 10 epochs based on the validation PLCC score was implemented to reduce the risk of overfitting.

Model performance was evaluated using the Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Correlation Coefficient (SRCC), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). PLCC and SRCC were selected because they are widely adopted evaluation metrics in Image Aesthetic Assessment (IAA), measuring prediction performance based on linear correlation and ranking consistency with human aesthetic judgments. MAE, MSE, and RMSE were additionally used to quantify absolute and squared prediction errors, providing complementary measurements of regression performance. These regression-based metrics were considered more appropriate than classification metrics, such as accuracy or F1-score, because the proposed model predicted continuous aesthetic scores rather than discrete class labels.

Trial Design

Six experiments were systematically designed and divided into two groups. The Basic Configuration Group (Experiments 1–4) evaluated four model configurations under identical training conditions. The Independent Replication Group (Experiments 5–6) repeated the two dual-branch architectures using different random initializations to assess the consistency and reliability of the proposed models. Experiments 1 (RGB baseline) and 4 (Adaptive Fusion) were trained for a maximum of 50 epochs with early stopping, whereas Experiments 2, 3, 5, and 6 were trained for 20 epochs under identical training settings to ensure fair comparisons. All models were evaluated using the same held-out test set, which was not used during the training process.

Experiment 5 replicated the Concatenation Fusion model from Experiment 3 without modifying the network architecture, backbone selection, hyperparameters, learning rates, or training protocol. The only difference was the random initialization of model parameters at the beginning of training. Similarly, Experiment 6 replicated the Adaptive Fusion model under the same controlled conditions. These replication experiments were designed to examine whether the performance of the proposed dual-branch architecture remained consistent across different optimization trajectories, thereby providing additional evidence regarding model robustness and reliability rather than relying solely on the outcome of a single training run.

Variations between repeated experiments were expected because deep neural network training is inherently stochastic. Random weight initialization, mini-batch shuffling, and dropout introduce different optimization trajectories that may lead to slightly different local optima. Nevertheless, consistent performance across independent replications indicates that the proposed architecture was stable and did not depend on a specific initialization condition. Therefore, observed performance variations reflected the stochastic characteristics of the optimization process rather than instability within the proposed Dual-Branch FFT-ViT architecture. Although future research should include multi-seed evaluations with mean and standard deviation reporting, the replication experiments conducted in this study provided additional evidence supporting the reliability and reproducibility of the proposed approach.

RESULTS AND DISCUSSION

Results of Each Trial

Table 2 presents the complete results of all experiments conducted on the test dataset. The primary evaluation metrics used are the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Correlation Coefficient (SRCC), which measure the strength of the correlation between model predictions and human aesthetic scores; values closer to 1 indicate stronger agreement and better performance. In addition, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are used to quantify prediction errors, where lower values indicate better performance, with values closer to 0 representing more accurate

predictions.

Table 2
Comparison Table of Experimental Results

No.	Configuration	PLCC ↑	SRCC ↑	MSE ↓	MAE ↓	RMSE ↓
1	RGB-ViT (Kontrol)	0,7212	0,7193	0,0199	0,1123	0,1410
2	FFT-only	0,5288	0,5124	0,0300	0,1403	0,1732
3	Concatenation Fusion	0,7004	0,7010	0,0211	0,1168	0,1453
4	Adaptif Gate Fusion	0,7183	0,7108	0,0200	0,1120	0,1416
5	Concatenation Fusion. (Ablation) ★	0,7336	0,7347	0,0193	0,1117	0,1388
6	Adaptif Gate Fusion (Ablation)	0,7230	0,7164	0,0198	0,1122	0,1406

From Table 2, the RGB-only control model in the first experiment achieved a PLCC value of 0.7212, which serves as the baseline performance. The FFT-only model in Experiment 2 obtained the lowest performance (PLCC = 0.5288), indicating that frequency-domain information alone is insufficient for reliable image aesthetic assessment because it lacks semantic understanding and compositional context. The Concatenation Fusion approach in Experiment 3 achieved a PLCC value of 0.7004, whereas the Adaptive Fusion approach in Experiment 4 improved the PLCC value to 0.7183. These findings demonstrate that integrating spatial-domain and frequency-domain features is more effective than relying on either domain independently.

These results indicate that RGB and FFT features provide complementary information for image aesthetic assessment. RGB features primarily capture semantic content, object arrangement, and color harmony, whereas FFT features emphasize structural characteristics, including texture patterns, edge frequency distributions, and fine-grained visual details. By preserving feature representations from both branches before integration, Concatenation Fusion generates a richer feature space that allows the subsequent fusion and regression layers to learn more discriminative aesthetic representations. In contrast, Adaptive Fusion integrates the two branches through learned weighting mechanisms, which may reduce the contribution of certain complementary features during the optimization process. Since image aesthetics are influenced by multiple visual attributes simultaneously, maintaining comprehensive information from both spatial and frequency domains enhances the representational capacity of the model and contributes to improved aesthetic prediction performance.

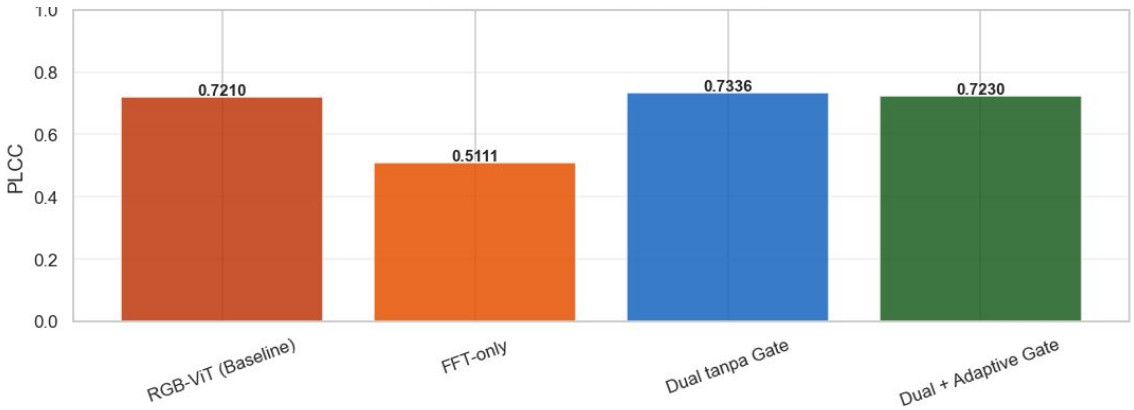

Figure 6. *Comparison of PLCC Four Main Configurations and Two Replication Configurations*

Figure 6 visualizes the comparison of PLCC values across configurations. Experiments 5 and 6 were not designed to introduce new architectures but rather to evaluate the reproducibility and optimization potential of the proposed dual-branch models under different stochastic initialization conditions. Since all architectural components and hyperparameters remained

unchanged, any performance variations can be attributed to the stochastic nature of deep neural network optimization. The proposed model surpassed the RGB-only model, which achieved a PLCC value of 0.7212, with an improvement of 0.0124. Beyond PLCC, the proposed model also demonstrated superior performance across all other evaluation metrics, achieving an SRCC of 0.7347, an MAE of 0.1117, and an RMSE of 0.1388, representing the best overall results. Furthermore, both dual-branch models outperformed the RGB-only baseline: the Concatenation Fusion model achieved a PLCC of 0.7336, which was 0.0124 higher than the RGB-only model, while the Adaptive Fusion model achieved a PLCC of 0.7230, representing an improvement of 0.0018 over the baseline. These findings demonstrate that dual-branch architectures can outperform single-branch approaches when parameter initialization conditions facilitate better convergence during training.

Training Curve Analysis

Figure 7 presents the training curves of the RGB-ViT control model from the first experiment. The model achieved its highest validation PLCC value at approximately the 20th epoch and subsequently stabilized before training was automatically terminated at the 22nd epoch using an early stopping mechanism. The training PLCC curve continued to increase, indicating a reasonable performance gap between the training and validation datasets, which is a common phenomenon in deep neural network models. This behavior suggests the presence of mild overfitting; however, the early stopping strategy effectively prevented further degradation of generalization performance.

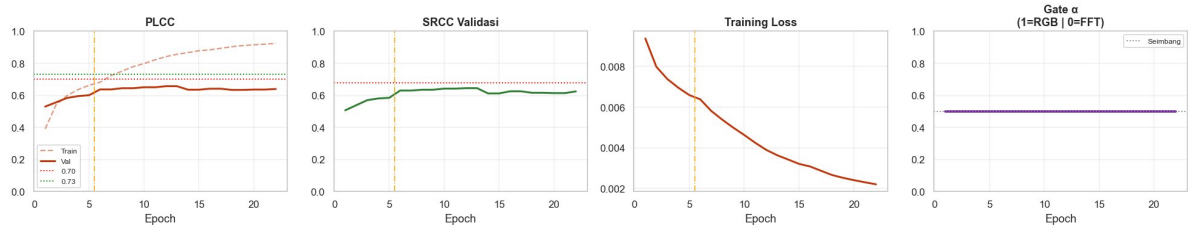

Figure 7. *Training Curve for Model RGB-ViT*

Figure 8 shows the FFT-ViT Dual-Branch training curve with Concatenation Fusion in the fifth experiment. The most notable aspect is the fourth panel, which displays the gate values (α) throughout the training process. In contrast to the RGB-only model, which does not include a gate mechanism and is displayed as a straight line, the α values in the Adaptive Fusion model fluctuate around 0.486 with minor variations, indicating that the active gate mechanism dynamically adjusts the contributions of the two branches. An α value lower than 0.5 suggests a slight tendency for the model to rely more on the FFT branch.

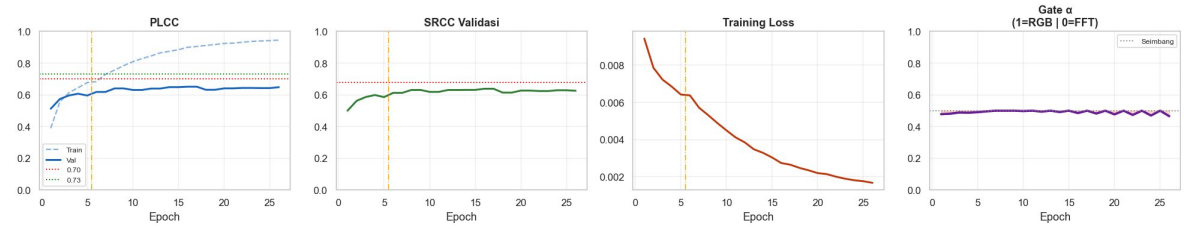

Figure 8. *Dual-Branch FFT-ViT Concatenation Fusion Training Curve*

Comparative Dashboard Analysis

Figure 9 presents a summary of the comparison between the best-performing model in this study, namely the Dual-Branch Vision Transformer FFT with Concatenation Fusion from the fifth experiment, and the RGB-ViT comparison model through several types of analysis. Overall, both models demonstrate similar prediction patterns; however, the Dual-Branch model achieves a slightly higher correlation, with a PLCC score of 0.7336 compared with 0.7212 for the RGB-ViT model, indicating that its predictions are closer to the reference scores. This finding suggests that incorporating frequency-domain information using FFT can improve performance, although the improvement is relatively modest. The distribution of the α values also indicates that the model

tends to utilize information from both the RGB and FFT branches relatively equally, without excessively depending on either branch. In terms of prediction errors, the Dual-Branch model achieves a slightly lower MAE value than RGB-ViT, indicating that the average difference between the model predictions and human aesthetic judgments is smaller.

Analysis based on the α value groups shows that the model performance remains stable across various levels of balance between RGB and FFT information. Furthermore, the PLCC curve throughout the training process demonstrates that both models experience similar performance improvements before eventually reaching convergence; however, the Dual-Branch model achieves better generalization performance when evaluated on completely unseen data. Overall, the results presented in this figure demonstrate that combining spatial-domain information (RGB) and frequency-domain information (FFT) can enhance the model's capability to assess image aesthetics more accurately compared with using RGB information alone.

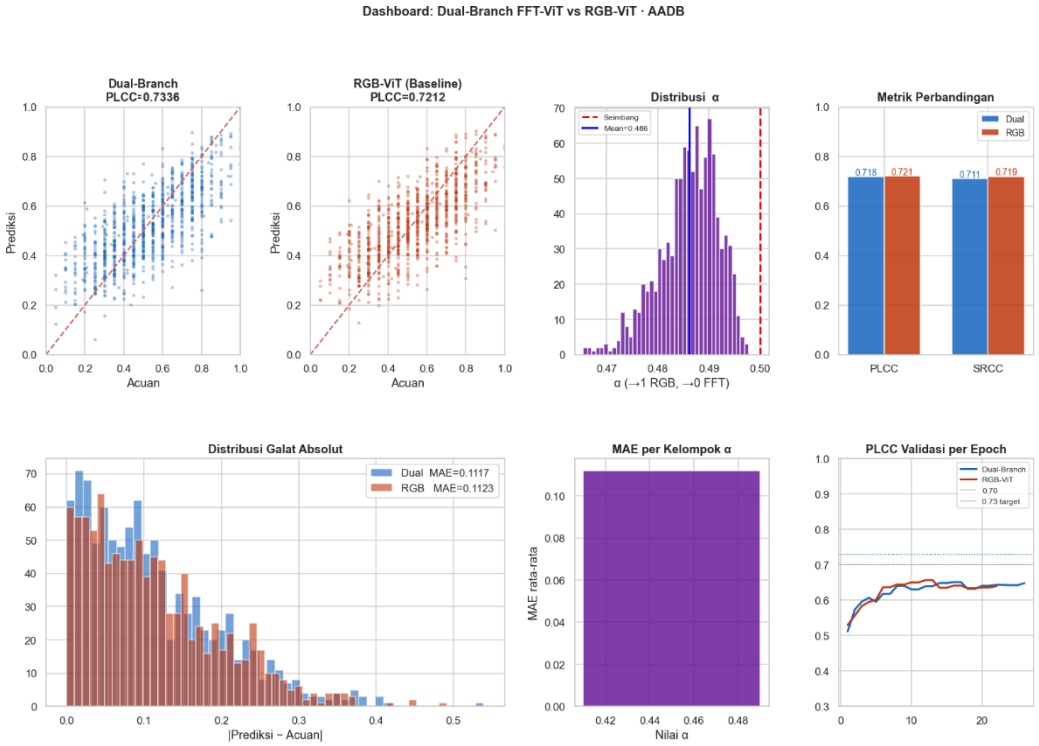

Figure 9. *Full Comparison Dashboard: Scatter Plot, α Distribution, MAE, and PLCC Curve per Epoch*

Comparison of Experimental Results with Previous Research

After obtaining the experimental results from the eight test scenarios, the best-performing results achieved in this study need to be compared with those reported in previous research in the field of Photo Aesthetic Assessment (PAA). This comparison aims to evaluate the relative performance of the proposed model against current state-of-the-art approaches while also validating whether the developed RGB-FFT fusion approach can achieve competitive performance compared with previous methods, including both CNN-based and Vision Transformer (ViT)-based methods. Based on the experimental results presented in the table, the Random Seed Difference-Initiated Concatenation Fusion model achieved the best performance, with a PLCC score of 0.7336 and an SRCC score of 0.7347. This performance surpassed that of the RGB-ViT baseline, which obtained a PLCC score of 0.7212 and an SRCC score of 0.7193, as well as the results of all other fusion scenarios. These scores were subsequently used as the primary benchmark for comparison with eight previous studies.

Table 3
Comparison of Experimental Results with Previous Research

No.	Researcher & Year	Model	Datasets	PLCC	SRCC
1	Kong dkk. (2016) and Yang (2025) ECCV	AADB-CNN	AADB (11K)	0.678	-

No.	Researcher & Year	Model	Datasets	PLCC	SRCC
3	Sheng dkk. (2018) ACM MM and Zheng (2025)	Attention-PAA	AVA & AADB	0.697	0.671
4	Ke dkk. (2021) and Zhang (2026) ICCV	MUSIQ	AVA & AADB	0.728	0.726
5	Zeng dkk. (2019) IEEE TIP Gonzalez-Naharro (2025)	Unified-PAA	AVA & AADB	0.689	0.659
6	Behrad dkk. (2025) CVPR	Dinov2-small (ViT Base)	AADB	0.695	0,682
7	Xu dkk. (2025) ICME	ViT-Tiny (NIMA)	AADB	0.703	0.712
9	This Research	RGB-ViT & FFT-ViT Fusion	AADB	0.733	0.734

Table 3 shows that the proposed New Initiation Concatenation Fusion model in this study demonstrates superior or comparable performance to most previous studies. Compared with CNN-based approaches such as AADB-CNN (Kong et al., 2016; Yang et al., 2025), NIMA (Talebi & Milanfar, 2018; Tanawongsuwan & Phongsuphap, 2026), Attention-PAA (Sheng et al., 2018; Zheng et al., 2025), and Unified-PAA (Gonzalez-Naharro et al., 2025; Zeng et al., 2019), the proposed model achieves significant improvements in PLCC values of 0.0556, 0.1216, 0.0366, and 0.0446, respectively (Talebi & Milanfar, 2021; Tanawongsuwan & Phongsuphap, 2026). These results indicate that the fusion approach combining the spatial domain (RGB) and frequency domain using FFT provides a more effective feature representation than conventional CNN-based approaches that rely solely on pixel-domain information, as theoretically predicted by Kong et al. (2016). This finding is also consistent with previous research by Y. Zhang et al. (2024), which emphasized the importance of frequency-related technical features, such as image sharpness and blur, in aesthetic quality assessment (Yang et al., 2025).

When compared with the most relevant ViT-based studies, the proposed model also demonstrates competitive performance. Compared with the RGB-only baseline model on the AADB dataset reported by Behrad et al. (2025) and Xu et al. (2025), the proposed model achieves higher performance, with PLCC improvements of 0.0386 and 0.0216, respectively. This suggests that the integration of an FFT branch through a concatenation-based fusion scheme, combined with the proposed weight initialization strategy, enhances the capability of the fundamental ViT architecture on the same dataset. Meanwhile, compared with MUSIQ (Ke et al., 2021; Zhang et al., 2026), which employs a multi-scale ViT architecture, the proposed model achieves a marginally lower PLCC value by 0.0056. Compared with ViT-Base (He et al., 2023) evaluated on the AVA dataset, the proposed model shows a slightly lower performance difference of 0.0054, despite using a substantially smaller dataset. Specifically, the AVA dataset contains approximately 255,000 images, whereas the AADB dataset used in this study contains approximately 11,000 images. These findings suggest that the proposed fusion strategy can achieve competitive aesthetic assessment performance despite being trained on a considerably smaller dataset.

Fusion Model Behavior Analysis

The behavior of the concatenation fusion mechanism was further analyzed based on the learned α values. The top row presents images with the lowest α values, ranging from approximately 0.465 to 0.468, indicating that these images rely slightly more on the FFT branch. These images generally exhibit poor illumination conditions, lower aesthetic scores ranging from 0.188 to 0.400, and less visually appealing characteristics. In contrast, the bottom row presents images with higher α values, ranging from approximately 0.497 to 0.498, indicating that the model achieves a more balanced contribution between the RGB and FFT branches. These images show more diverse aesthetic scores, ranging from 0.150 to 0.750, and contain more varied visual content.

An important finding from this analysis is that all learned α values fall within a very narrow range between 0.465 and 0.498, with no values approaching 0, which would indicate

complete FFT dominance, or 1, which would indicate complete RGB dominance. This result confirms that the model has learned to maintain a nearly balanced contribution between both branches, with a slight tendency to rely more on frequency-domain information for lower-quality images. This adaptive balance suggests that the proposed fusion mechanism effectively integrates complementary spatial and frequency representations for image aesthetic quality assessment.

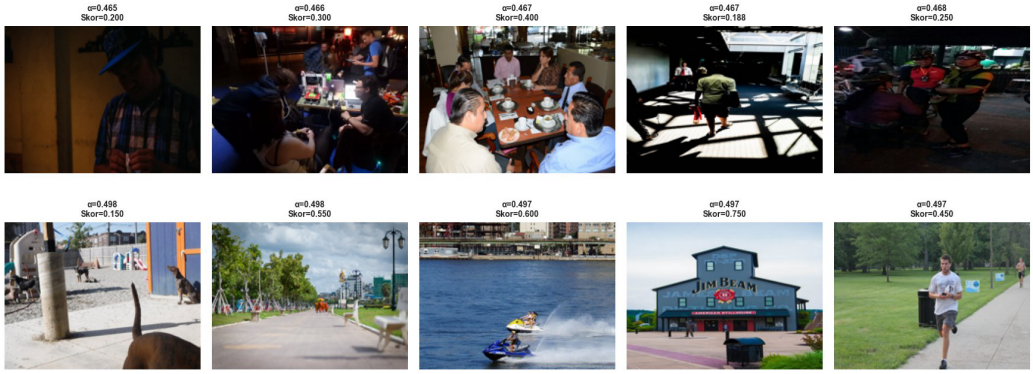

Figure 10. *Gate α Analysis: Photos with Low α Values (Top Row) vs Higher α (Bottom Row)*

Examples of Best Model Predictions

Figure 11 shows an example of the best model prediction from the fifth experiment, in which concatenation fusion with random-state initialization achieved a PLCC score of 0.7336. The 18 test images were divided into three aesthetic score categories: low scores (<0.30), medium scores (0.40–0.60), and high scores (≥0.70). Each image is accompanied by a human ground truth (GT) score and a model-predicted score, along with a comparative bar chart illustrating the difference between the two values. The green image border indicates a highly accurate prediction with an error of less than 0.05, orange indicates moderate accuracy with an error between 0.05 and 0.12, and red indicates a large prediction error exceeding 0.12.

The model demonstrates strong predictive capability for medium- and high-scoring photos, particularly nature and macro photographs that exhibit distinctive frequency-domain characteristics. For low-scoring photos, the model tends to predict values higher than the ground truth (overestimation), which is a common behavior of regression models applied to aesthetic datasets with imbalanced score distributions. This phenomenon is associated with the subjective nature of human aesthetic judgments, which are challenging to fully replicate using computational models.

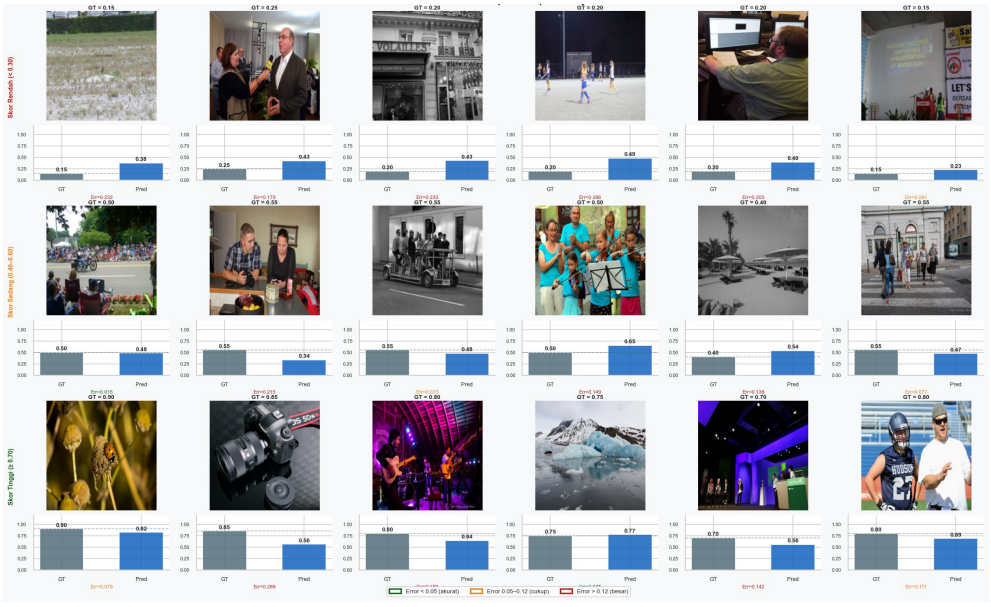

Figure 11. *Best Model Prediction Example (Trial 5) on 18 Images from AADB Test Data*

Discussion

Initialization Variance and Its Implications

One of the important findings of this study was the noticeable performance variation between experiments with identical model configurations but different random initializations. Concatenation Fusion in Experiment 3 achieved a PLCC of 0.7004, whereas the replicated configuration in Experiment 5 achieved a PLCC of 0.7336, representing an improvement of 0.0332. Since both experiments used the same architecture, hyperparameters, and training protocol, this variation can be attributed to differences in the stochastic optimization process rather than changes in the proposed model architecture.

This phenomenon occurs because the initialization of network weights, the order of shuffled mini-batches, and the dropout patterns are determined by the random number generator at the beginning of training. When the number of training epochs is relatively limited (20 epochs in this study), the optimization process may converge to different local optima, making the final performance more sensitive to the initial random state.

This observation is consistent with previous deep learning studies, which have reported that stochastic optimization algorithms may produce different solutions even under identical experimental settings due to random initialization and data shuffling. Therefore, evaluating a model using multiple random seeds has become a widely recommended practice for improving the reliability and reproducibility of experimental results. In the context of image aesthetic assessment, these findings indicate that the performance improvement of the proposed Dual-Branch FFT-ViT is not solely attributable to a favorable initialization but reflects the robustness of the proposed architecture across different optimization trajectories.

These findings motivate the adoption of multi-seed evaluation, in which each experiment is repeated using several random seeds and the results are reported as the mean and standard deviation. Such an evaluation protocol would provide a more reliable estimate of model performance and improve the reproducibility of future studies.

Limitations and Future Work

Despite the promising performance of the proposed Dual-Branch FFT-ViT architecture, several limitations should be acknowledged. First, employing two Vision Transformer backbones increases computational complexity and memory consumption compared with a conventional single-branch RGB model, resulting in longer training and inference times. Although the FFT transformation itself introduces only a small computational overhead, the additional feature extraction branch substantially increases the overall number of trainable parameters. Consequently, the proposed architecture is more suitable for offline aesthetic assessment scenarios than for real-time applications on resource-constrained devices.

Second, all experiments were conducted exclusively on the AADB dataset. Although AADB is a widely used benchmark for image aesthetic assessment, its image distribution may not fully represent other aesthetic datasets, such as AVA or Flickr-AES. Therefore, the generalization capability of the proposed model across different aesthetic datasets remains to be investigated. Future work should evaluate the proposed architecture on multiple public datasets and perform cross-dataset validation to better assess its robustness and generalization ability.

Finally, the replication experiments revealed that model performance is influenced by stochastic optimization, particularly when the number of training epochs is relatively limited. Although the proposed architecture consistently demonstrated competitive performance, future studies should adopt multi-seed evaluation and report the mean and standard deviation of performance metrics to provide a more reliable assessment of model stability and reproducibility.

CONCLUSION

This study successfully proposes and evaluates the Dual-Branch FFT-ViT architecture for photo aesthetic assessment by combining spatial-domain (RGB) and frequency-domain (FFT) representations. Through six systematic trials conducted on the AADB dataset, several key conclusions can be drawn.

First, the best-performing model across all trials was the Concatenation Fusion strategy in the independent replication experiment (Experiment 5), achieving a PLCC score of 0.7336 and an

SRCC score of 0.7347. This performance exceeded that of the RGB baseline by 0.0124 points in terms of PLCC. Second, the FFT-only representation was insufficient as a standalone feature representation, achieving a PLCC score of only 0.5288; however, when combined with RGB features within a dual-branch framework, it provided meaningful complementary information. Third, Adaptive Fusion consistently outperformed static Concatenation Fusion under comparable experimental conditions, demonstrating the advantages of adaptive weighting mechanisms for integrating spatial and frequency-domain features. Fourth, variation caused by different random initializations (approximately ±0.010–0.033 PLCC) represents an inherent characteristic of stochastic training processes that should be considered in experimental design and evaluation. Fifth, the fusion gate parameter ($\alpha \approx 0.486$) indicates that the model utilizes both branches in a relatively balanced manner, with image-dependent variations reflecting meaningful adaptive behavior.

For future research, several development directions are suggested. First, multi-seed evaluation using a minimum of 3–5 different random seeds should be conducted to obtain a more stable performance estimate and enable statistical confidence interval reporting. Second, domain-specific normalization for FFT features, where the mean (μ_FFT) and standard deviation (σ_FFT) are calculated from the training dataset rather than adopting ImageNet normalization statistics, has the potential to improve the consistency and quality of frequency-domain representations across different initialization conditions.

Third, a cross-attention-based fusion mechanism between patch tokens from the two branches, rather than relying only on the [CLS] tokens, may capture richer local spatial-frequency interactions. Fourth, validation on additional aesthetic datasets, such as AVA and Flickr-AES, is necessary to confirm the generalizability of the proposed approach beyond the AADB dataset. Fifth, exploring alternative frequency-domain representations, such as the Discrete Wavelet Transform (DWT), which simultaneously preserves spatial and frequency information, may provide a promising alternative to the FFT magnitude spectrum.

ACKNOWLEDGEMENT

The author would like to thank the supervisors at the Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember, for their guidance and support during this research. The AADB dataset used in this study is a publicly available dataset and managed by its creator.

AUTHOR CONTRIBUTION STATEMENT

Reza Rachmadan contributed to the conceptualization, research design, methodology, data collection, model development, data analysis, and preparation of the original manuscript. Shintami Chusnul Hidayati contributed to research supervision, validation, technical evaluation, interpretation of results, and critical review and editing of the manuscript. Both authors collaboratively contributed to the refinement of the research methodology, discussion of the findings, and finalization of the manuscript.

REFERENCES

Anwar, A., Kanwal, S., Tahir, M., Saqib, M., Uzair, M., Rahmani, M. K. I., & Ullah, H. (2021). A Survey on Image Aesthetic Assessment. *ArXiv*. https://doi.org/10.48550/arXiv.2103.11616

Behrad, F., Tuytelaars, T., & Wagemans, J. (2025). Charm: The Missing Piece in ViT Fine-Tuning for Image Aesthetic Assessment. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7815–7824. https://doi.org/10.1109/CVPR52734.2025.00732

Gao, M., Song, C., Zhang, Q., Zhang, X., Li, Y., & Yuan, F. (2025). Research Progress on Color Image Quality Assessment. *Journal of Imaging*, *11*(9), 307. https://doi.org/10.3390/jimaging11090307

Gonzalez-Naharro, L., Flores, M. J., Martínez-Gómez, J., & Puerta, J. M. (2025). Analysis of image aesthetics assessment as a positive-unlabeled problem. *Signal Processing: Image Communication*, 117441.

He, S., Ming, A., Zheng, S., Zhong, H., & Ma, H. (2023). EAT: An Enhancer for Aesthetics-Oriented

Transformers. *Proceedings of the 31st ACM International Conference on Multimedia*, 1023–1032. https://doi.org/10.1145/3581783.3611881

He, S., Zhang, Y., Xie, R., Jiang, D., & Ming, A. (2022). Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, 942–948. https://doi.org/10.24963/ijcai.2022/132

Jia, M., Xing, X., He, J., & Xu, X. (2025). Self-Supervised Image Aesthetic Assessment Based on Transformers. *International Journal of Semantic Computing*, *18*(4), 519–536. https://doi.org/10.1142/S1469026824500299

Jiang, X., Zhang, X., Gao, N., & Deng, Y. (2025). When Fast Fourier Transform Meets Transformer for Image Restoration. *Computer Vision – ECCV 2024*, *15103*, 381–402. https://doi.org/10.1007/978-3-031-72995-9_22

Ke, J., Wang, Q., Wang, Y., Milanfar, P., & Yang, F. (2021). MUSIQ: Multi-Scale Image Quality Transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5128–5137. https://doi.org/10.1109/ICCV48922.2021.00510

Kong, S., Shen, X., Lin, Z., Mech, R., & Fowlkes, C. (2016). Photo Aesthetics Ranking Network with Attributes and Content Adaptation. *Computer Vision – ECCV 2016*, *9905*, 662–679. https://doi.org/10.1007/978-3-319-46448-0_40

Li, L., Huang, Y., Wu, J., Yang, Y., Li, Y., Guo, Y., & Shi, G. (2023). Theme-Aware Visual Attribute Reasoning for Image Aesthetics Assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(9), 4798–4811. https://doi.org/10.1109/TCSVT.2023.3249185

Li, Z., Yan, X., Wei, X., & Shao, F. (2025). IAACLIP: Image Aesthetics Assessment via CLIP. *Electronics*, *14*(7), 1425. https://doi.org/10.3390/electronics14071425

Sheng, K., Dong, W., Ma, C., Mei, X., Huang, F., & Hu, B.-G. (2018). Attention-Based Multi-Patch Aggregation for Image Aesthetic Assessment. *Proceedings of the 26th ACM International Conference on Multimedia*, 879–886. https://doi.org/10.1145/3240508.3240554

Shi, J., Gao, P., & Smolic, A. (2024). Blind Image Quality Assessment via Transformer Predicted Error Map and Perceptual Quality Token. *IEEE Transactions on Multimedia*, *26*, 4641–4651. https://doi.org/10.1109/TMM.2023.3325719

Shi, T., Chen, C., Wu, Z., Hao, A., & Fang, Y. (2024). Improving Image Aesthetic Assessment via Multiple Image Joint Learning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *20*(11). https://doi.org/10.1145/3687128

Sun, W., Zhang, W., Cao, Y., Cao, L., Jia, J., Chen, Z., Zhang, Z., Min, X., & Zhai, G. (2024). Assessing UHD Image Quality from Aesthetics, Distortions, and Saliency. *ArXiv*. https://doi.org/10.48550/arXiv.2409.00749

Talebi, H., & Milanfar, P. (2018). NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing*, *27*(8), 3998–4011. https://doi.org/10.1109/TIP.2018.2831899

Talebi, H., & Milanfar, P. (2021). Learning to Resize Images for Computer Vision Tasks. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 487–496. https://doi.org/10.1109/ICCV48922.2021.00055

Tanawongsuwan, R., & Phongsuphap, S. (2026). Flip-Robust Neural Image Assessment (FR-NIMA) for Spatially Consistent IQA. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, *20*(2), 303–318.

Xu, L., Xu, J., Yang, Y., Wang, X., Huang, Y.-J., & Li, Y. (2025). CLIP Brings Better Features to Visual Aesthetics Learners. *2025 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. https://doi.org/10.1109/ICME59968.2025.11210032

Yang, H., Li, Y., Jin, X., Zhou, X., Shi, P., & Liu, Y. (2025). Aesthetic multi-attributes network for image captioning. *Computers and Electrical Engineering*, *123*, 110103.

Zeng, H., Cao, Z., Zhang, L., & Bovik, A. C. (2019). A Unified Probabilistic Formulation of Image Aesthetic Assessment. *IEEE Transactions on Image Processing*, *29*, 1548–1561. https://doi.org/10.1109/TIP.2019.2941778

Zhang, R., Dong, W., Lu, L., Zhou, Z., Zhang, T., & Niu, W. (2026). Multi-scale wavelet vision transformer with HDR-aware attention for high dynamic range image quality assessment. *The Journal of Supercomputing*, *82*(9), 509.

Zheng, Q., Shi, Y., & Wu, Y. (2025). Image Aesthetic Assessment Based on Multi-level Hierarchical

Adaptive Fusion. *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 354–368.

Zhou, X., Jin, X., Lv, J., Huang, H., Mao, M., & Cui, S. (2022). Aesthetic Attributes Assessment of Images with AMANv2 and DPC-CaptionsV2. *ArXiv Preprint ArXiv:2208.04522*.