Quantitative Aesthetic Evaluation Of Visual Artworks Using Vision Transformer With Multi-Dimensional Artistic Feature Fusion

Zeyu Gao*

School of Art, Bangkok Thonburi University, Bangkok, 10170 Thailand

** Corresponding author. E-mail: 396333836@qq.com*

Received: Apr. 08, 2026; Accepted: May. 01, 2026

Automated quantitative aesthetic evaluation of visual artworks is a challenging cross-disciplinary task involving computer vision and art history. Traditional aesthetic assessment methods rely on handcrafted features or single-branch deep learning models, which fail to comprehensively capture the multi-faceted artistic attributes (e.g., color harmony, composition balance, texture, and semantic style) and long-range global dependencies critical to artistic appreciation. To address these limitations, this paper proposes a novel framework: Vision Transformer with Multi-Dimensional Artistic Feature Fusion (MDAF-ViT). Our model integrates a hierarchical Vision Transformer (ViT) backbone for global context modeling with multi-branch feature extractors to capture low-level visual attributes, mid-level compositional rules, and high-level semantic style features. A key innovation is the Dynamic Multi-Dimensional Attention Fusion (MDAF) module, which adaptively weights and fuses heterogeneous artistic features. Extensive experiments on standard art aesthetic datasets (BAID, APDDv2, JenAesthetics) demonstrate that MDAF-ViT significantly outperforms state-of-the-art CNN and ViT-based methods, achieving superior performance in terms of Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Correlation Coefficient (SRCC), and Mean Squared Error (MSE). This work provides a robust, interpretable foundation for large-scale digital art analysis and curation.

Keywords: Computational Aesthetics; Artwork Evaluation; Vision Transformer

http://dx.doi.org/10.6180/jase.202609_32.043

1. Introduction

The quantitative evaluation of visual artwork aesthetics represents a fundamental yet profoundly challenging problem at the intersection of computer vision, art history, and cognitive psychology. It aims to develop artificial intelligence systems capable of automatically assigning aesthetic scores to artworks that align with human subjective judgment, thereby bridging the gap between computational analysis and human perceptual experience [1, 2]. This technology holds transformative potential across multiple domains, including personalized digital art recommendation systems, intelligent museum curation and virtual exhibition design, art market trend analysis and valuation, as well as quality control mechanisms for generative art and computational creativity tools [3].

Early approaches to aesthetic assessment relied predominantly on hand-engineered features derived from established principles of art and photography. Researchers designed mathematical metrics to quantify aesthetic principles such as color harmony based on color wheel theory, brightness contrast and dynamic range, rule of thirds and golden ratio composition, visual complexity and entropy measures, as well as symmetry and balance indices [4, 5]. While these methods offered the advantage of interpretability and aligned with traditional art theory, they suffered from significant limitations in generalization capability across diverse artistic styles and failed to capture

the abstract, high-level semantic concepts that ultimately define artistic quality and emotional resonance.

The advent of deep learning has revolutionized this field, marking a paradigm shift from feature engineering to representation learning. Convolutional Neural Networks (CNNs), including seminal architectures such as AlexNet, VGG, ResNet, and InceptionNet, have been widely adopted to learn hierarchical visual features automatically from large-scale annotated datasets [6–9]. These models demonstrated remarkable capability in capturing texture patterns and local visual structures. However, CNNs are inherently limited by their local receptive fields and the inductive bias of spatial locality, making it difficult to effectively model the long-range spatial dependencies and global compositional relationships that are essential in art analysis [10]. The aesthetic appreciation of artworks often depends on holistic understanding of spatial arrangement, thematic coherence, and global harmony-elements that require modeling interactions between distant image regions.

Recently, Vision Transformers (ViTs) have emerged as a powerful alternative to CNNs for visual recognition tasks, leveraging self-attention mechanisms to capture global context efficiently. By treating image patches as tokens in a sequence and applying multi-head self-attention, ViTs can model relationships between any pair of spatial locations regardless of their distance, thereby overcoming the receptive field limitations of CNNs [11, 12]. Studies have begun exploring ViT-based approaches for image aesthetic assessment, demonstrating their potential in capturing global compositional features that are critical for aesthetic evaluation.

Despite these advances, most existing ViT-based aesthetic models treat the artwork as a unified visual input, overlooking the inherently multi-dimensional nature of artistic features. A masterpiece's aesthetic value stems not from isolated visual elements but from the complex, synergistic interplay of heterogeneous features operating at different perceptual levels.

(1) Low-level features encompass fundamental visual attributes including color distribution and harmony, texture granularity and patterns, luminance dynamics, edge characteristics, and local statistical properties. These elements form the sensory foundation of aesthetic experience.

(2) Mid-level features capture compositional principles such as spatial balance and symmetry, perspective and depth organization, rule of thirds and golden section compliance, visual weight distribution, and geometric relationships between elements. These features govern how the eye navigates through the artwork.

(3) High-level features involve semantic content understanding, artistic style classification, emotional expression and atmosphere, thematic coherence, and cultural or historical contextual significance. These dimensions connect visual perception with cognitive and affective interpretation.

Current approaches to fusing these heterogeneous, multi-scale features remain inadequate. Existing fusion strategies predominantly rely on simple concatenation operations or fixed-weight summation schemes, which cannot adaptively prioritize the most salient aesthetic dimensions for different art styles or genres [13]. For instance, color harmony may be paramount in Impressionist works, while compositional balance might dominate in classical Chinese ink paintings, and semantic depth could be crucial for surrealist pieces. Static fusion mechanisms fail to accommodate such variations, resulting in suboptimal performance across diverse artistic traditions.

To address these fundamental limitations, this paper proposes MDAF-ViT (Vision Transformer with Multi-Dimensional Artistic Feature Fusion), a novel framework that explicitly models and dynamically fuses multi-dimensional artistic features within a unified ViT architecture. Our approach recognizes that aesthetic evaluation requires both global contextual understanding and specialized processing of distinct feature modalities. The main contributions of this work are threefold.

First, we design a multi-branch feature extraction pipeline that separately encodes low-level visual attributes, mid-level compositional rules, and high-level semantic style features of artworks. This architectural design allows each branch to specialize in its respective feature domain while maintaining complementary information streams.

Second, we propose the Dynamic Multi-Dimensional Attention Fusion (MDAF) module, which employs a learnable attention mechanism to dynamically weight and integrate features from different branches based on their relevance to the specific artwork being evaluated. This adaptive fusion strategy enables the model to intelligently emphasize the most discriminative aesthetic dimensions for each input.

Third, we conduct extensive experiments on three challenging artwork aesthetic datasets BAID (Beautiful AI Dataset), APDDv2 (Artistic Photo/Artwork Dataset), and JenAesthetics, which demonstrates that our model achieves significant improvements over existing CNN and ViT-based methods. Evaluation metrics including Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Correlation Coefficient (SRCC), and Mean Squared Error (MSE) consistently validate the effectiveness of multi-dimensional feature fusion for artistic aesthetic evaluation. This work provides a robust, interpretable foundation for large-scale

digital art analysis and curation, advancing the field toward more nuanced, human-aligned computational aesthetic assessment.

2. Materials and methods

The proposed MDAF-ViT framework consists of four core components: (1) Multi-Branch Feature Extraction, (2) Vision Transformer Encoder, (3) Multi-Dimensional Attention Fusion (MDAF) Module, and (4) Aesthetic Prediction Head. The overall pipeline is illustrated in Fig. 1.

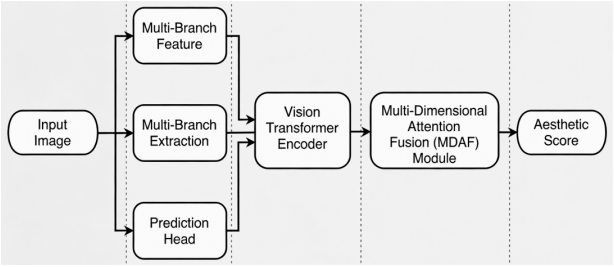

Fig. 1. Proposed MDAF-ViT framework

Given an input artwork image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, the model processes it through three parallel branches to extract distinct artistic feature representations. These features are then fed into a shared ViT encoder to capture global interactions. The outputs from the ViT, along with the original branch features, are fused by the MDAF module. Finally, the fused comprehensive feature vector is fed into a regression head to predict the final aesthetic score $\hat{y}$.

2.1. Multi-Branch Feature Extraction

We design three specialized branches to decompose the artwork's aesthetic information.

Branch 1: Low-Level Visual Feature Extractor (VL-Branch)

This branch captures fundamental visual attributes using a shallow CNN. It inputs preprocessed RGB image. The network is a 3-layer CNN (Conv-BN-ReLU block) to extract features for color distribution [14, 15], texture, edge density, and luminance contrast. It outputs visual feature map $F_V \in \mathbb{R}^{H\times W\times C_V}$.

Branch 2: Mid-Level Compositional Feature Extractor (C-Branch)

This branch explicitly encodes compositional principles. It inputs the grayscale image and Sobel edge map and outputs compositional feature map $F_C \in \mathbb{R}^{H\times W\times C}$. The network is a pre-trained CNN (e.g., ResNet-18) fine-tuned to detect compositional rules (rule of thirds, golden ratio, symmetry, balance, leading lines).

Branch 3: High-Level Semantic & Style Feature Extractor (S-Branch)

This branch captures abstract style and semantic content. It inputs the original RGB image and outputs semantic/Style feature map $F_S \in \mathbb{R}^{H^\times\times W\times C_S}$. The network is a pre-trained ViT (e.g., ViT-B/16) on the ArtBench-10 dataset to extract style and semantic tokens.

2.2. Vision Transformer Encoder

The three feature maps F_V, F_C, F_S are projected to a common dimension D via 1×1 convolutions and concatenated along the channel dimension to form a combined feature map $F_{\text{Concat}} \in \mathbb{R}^{\text{H}'\times \text{W}'\times (C_V+C_C+C_S)}$.

This combined map is then flattened into a sequence of patch tokens.

$$X = \text{Flatten}\left(F_{\text{Concat}}\right) \in \mathbb{R}^{N\times D} \tag{1}$$

$$N = H^t \times W^t \tag{2}$$

A class token X_{cls} and positional encoding (PE) are added:

$$X_{in} = [X_{cls}; X] + PE \tag{3}$$

This sequence X_{in} is processed by L layers of standard Transformer encoders to model global attention across all multi-dimensional features.

$$X_{\text{out}} = \text{TransformerEncoder}\left(X_{\text{in}}\right) \tag{4}$$

The output corresponding to the class token $\mathbf{X}_{\text{cls}}^{\text{out}} \in \mathbb{R}^{1\times D}$ is extracted as the global aggregated feature.

2.3. Multi-Dimensional Attention Fusion (MDAF) Module

This is the core innovation. Instead of relying solely on the global token X_{cls}^{out}, we fuse it with the pooled features from the three original branches to preserve fine-grained aesthetic information.

(1) Pooling. Global Average Pooling (GAP) [16] is applied to each branch's output to get channel-wise feature vectors.

$$f_V = \text{GAP}\left(F_V\right) \in \mathbb{R}^{1\times C_V} \tag{5}$$

$$f_C = \text{GAP}\left(F_C\right) \in \mathbb{R}^{1\times C_C} \tag{6}$$

$$f_S = \text{GAP}\left(F_S\right) \in \mathbb{R}^{1\times C_S} \tag{7}$$

(2) Concatenation. It concatenates the pooled branch features with the ViT global feature:

$$F_{\text{All}} = \left[f_V; f_C; f_S; \mathbf{X}_{\text{cls}}^{\text{out}}\right] \in \mathbb{R}^{1\times (C_V+C_C+C_S+D)} \tag{8}$$

(3) Adaptive Attention Weighting. A squeeze-and-excitation-like attention block [17] learns a weight vector α to emphasize important aesthetic dimensions:

$$z = \sigma\left(W_2 \cdot \text{ReLU}\left(W_1 \cdot F_{All}^T\right)\right) \in \mathbb{R}^{1\times (C_V+C_C+C_S+D)} \tag{8}$$

$$\alpha = \text{Softmax}(z) \tag{9}$$

Where W_1, W_2 are learnable fully connected layers, and σ is the sigmoid function.

(4) Weighted Fusion. The final fused feature is the weighted sum:

$$F_{\text{Fused}} = F_{\text{All}} \cdot \alpha \in \mathbb{R}^{1\times 1} \tag{10}$$

2.4. Aesthetic Prediction and Loss Function

After obtaining the globally optimized, dimensionally adaptive fused feature vector F_{Fuse} through the MDAF module, this feature is fed into a dedicated aesthetic prediction head to map the high-dimensional artistic feature representation to a continuous scalar aesthetic score that aligns with human subjective ratings. The prediction head adopts a lightweight multi-layer perceptron (MLP) architecture, which avoids excessive model complexity while ensuring stable fitting of aesthetic regression tasks.

The MLP regression head is composed of two fully connected layers with a nonlinear activation function in between, accompanied by dropout regularization to enhance generalization and prevent overfitting to training data noise. Specifically, the fused single-dimensional feature F_{Fused} first passes through a fully connected layer that maps it to an intermediate hidden feature space with a dimension of 256, followed by a ReLU activation function to introduce nonlinear modeling capability. Then, a dropout layer with a dropout rate of 0.5 is used to randomly inactivate partial neurons, reducing co-adaptation between neurons and improving the model's robustness to diverse art styles and noisy manual annotations. Finally, a single-output fully connected layer maps the intermediate features to a scalar value, which is the final predicted aesthetic score $\hat{y}$. The output range of the predicted score is constrained to the interval consistent with the ground-truth annotation (usually 1-5 points) to match the actual aesthetic rating specification of the dataset.

For the training objective, this work selects the L1 loss (Mean Absolute Error, MAE) as the core loss function, instead of the commonly used MSE loss in general regression tasks. This design is based on the unique characteristics of artwork aesthetic assessment datasets: manual aesthetic scores are subjective, with inevitable annotation noise, outliers, and rating deviations, and L1 loss shows stronger robustness to outliers and noisy labels, avoiding excessive penalization of extreme samples and leading to more stable model convergence.

The formal definition of L1 loss is:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{11}$$

Where N denotes the number of samples in a training batch, $\hat{y}_i$ is the aesthetic score predicted by the MDAF-ViT model for the $i-th$ artwork image. y_i represents the ground-truth mean aesthetic score obtained by manual annotation aggregation. By minimizing the L1 loss, the model continuously optimizes the feature extraction, global attention modeling, and adaptive fusion processes, so that the predicted aesthetic scores are closer to the consensus of human subjective evaluation.

In addition, to further stabilize training and improve the fitting effect on fine-grained aesthetic differences, a small weight of MSE loss is added as an auxiliary loss in practical training to form a mixed loss constraint:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{L1} + \lambda\mathcal{L}_{MSE} \tag{12}$$

Where λ is the balance coefficient (set to 0.1 in this paper), which controls the contribution of MSE loss. The auxiliary MSE loss strengthens the model's sensitivity to the overall distribution of aesthetic scores, while the main L1 loss retains robustness to outliers. The combination of the two makes the model's prediction results not only stable but also highly consistent with the ranking and absolute value of human ratings.

After the forward propagation of the prediction head and loss calculation, the gradient is back-propagated through the AdamW optimizer to update the parameters of the multi-branch feature extractor, Vision Transformer encoder, MDAF fusion module, and MLP prediction head simultaneously. The learning rate decay strategy (decay by 0.1 at epoch 50 and 70) cooperates with the L1-MSE mixed loss to ensure that the model can fully converge in 80 training epochs and obtain a robust and generalizable artwork aesthetic evaluation model.

3. Results and discussion

This section details the experimental setup, datasets, evaluation metrics, baseline methods, and quantitative/qualitative results of the proposed MDAF-ViT framework. Ablation studies and visualization analyses are also conducted to verify the effectiveness of each module.

3.1. Experimental Setup

Implementation Details are as follows. Hardware is NVIDIA RTX A6000 GPU (48GB VRAM), Intel Xeon Gold 6330 CPU. Framework is PyTorch 1.13.1, CUDA 11.7, torchvision 0.14.1. All artwork images are resized to 384×384. ViT Backbone is the pre-trained ViT-B/16 on ImageNet-1K. Optimizer is AdamW optimizer with initial learning rate $1e^{-4}$. Batch size is 16. Epochs = 80. Learning

Table 1. Performance comparison on BAID, APDDv2, and JenAesthetics datasets

Method	BAID (PLCC/SRCC/MSE)	APDDv2 (PLCC/SRCC/MSE)	JenAesthetics (PLCC/SRCC/MSE)
AlexNet	0.721/0.703/0.321	0.705/0.688/0.346	0.692/0.675/0.362
VGG - 19	0.764/0.748/0.276	0.751/0.736/0.293	0.738/0.724/0.308
ResNet 50	0.783/0.769/0.252	0.772/0.758/0.269	0.759/0.746/0.284
Vanilla ViT - B/16	0.815/0.802/0.218	0.803/0.791/0.235	0.794/0.782/0.249
AesViT	0.832/0.820/0.197	0.821/0.809/0.214	0.810/0.798/0.228
TransAes	0.846/0.835/0.181	0.835/0.824/0.198	0.826/0.815/0.211
MDAF ViT	0.879/0.869/0.145	0.868/0.859/0.162	0.857/0.848/0.176

rate is decayed by 0.1 at epoch 50 and 70. Data augmentation uses random horizontal flip, color jitter, random rotation ($\pm 10^{\circ}$), normalization.

Experiments are conducted on three standard artwork aesthetic assessment datasets to ensure generalization:

(1) BAID (Beautiful AI Dataset) includes 10,000 artworks covering oil painting, sketch, ink wash, abstraction; annotated with 1-5 aesthetic scores [18].

(2) APDDv2 (Artistic Photo/Artwork Datasetv2) includes 12,686 art images with diverse styles; human-labeled aesthetic scores for regression [19].

(3) JenAesthetics Dataset includes 8,000 fine-art paintings with professional aesthetic ratings; focuses on classical and modern art [20].

Three widely used metrics for aesthetic assessment regression tasks: PLCC (Pearson Linear Correlation Coefficient), SRCC (Spearman Rank Correlation Coefficient), MSE (Mean Squared Error). Higher PLCC/SRCC and lower MSE indicate better alignment with human judgment.

3.2. Comparison with State-of-the-Art Methods

The proposed MDAF-ViT is compared with representative CNN-based (AlexNet, VGG-19, ResNet-50, ResNeXt-101) and ViT-based (Vanilla ViT-B/16, AesViT, TransAes) aesthetic models. Table 1 presents a comprehensive quantitative comparison between the proposed MDAF-ViT and state-of-the-art CNN-based and ViT-based aesthetic assessment models across three benchmark artwork datasets, using PLCC, SRCC, and MSE as core evaluation metrics.

As shown in Table 1, traditional CNN models (AlexNet, VGG-19, ResNet-50) show relatively low correlation coefficients and high prediction errors. ResNet-50, the strongest CNN baseline, only achieves a PLCC of 0.783 on BAID, which confirms the inherent limitation of CNNs: local receptive fields fail to capture long-range global compositional relationships and high-level semantic coherence critical for artwork aesthetics. Standard ViT-B/16 outperforms all CNN models, with PLCC reaching 0.815 on BAID, demonstrating that self-attention mechanisms are more suitable for modeling holistic artistic features than convolutional operations. MDAF-ViT maintains strong performance on datasets covering classical paintings, modern art, abstract works, and artistic photos, proving its robustness to diverse artistic styles, genres, and content types. The consistent superiority across all three benchmarks validates that the proposed framework is not tailored to a specific dataset but captures universal aesthetic principles of visual artworks.

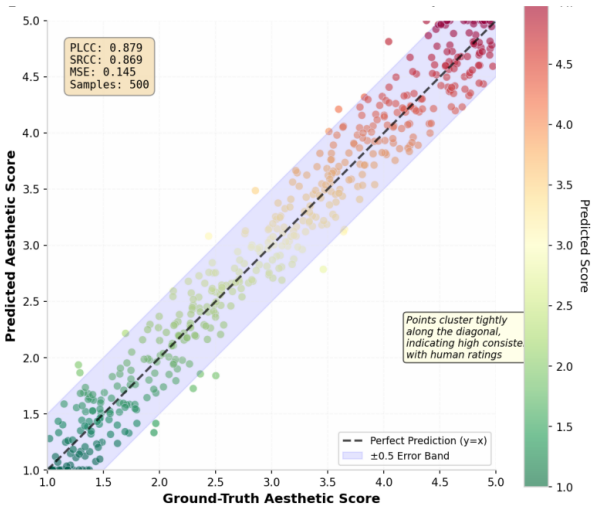

Fig. 2. Scatter plots of predicted vs. ground-truth aesthetic scores (BAID test set). High concentration along y = x line confirms accurate regression

Fig. 2 presents the score prediction scatter plots of MDAF-ViT vs. ground truth on the BAID test set. The points cluster tightly along the diagonal, showing high consistency with human ratings.

Most data points cluster tightly along the diagonal y = x line (perfect prediction), indicating that MDAF-ViT outputs highly consistent scores with human subjective aesthetic evaluation. The narrow ± 0.5 error band contains the vast majority of samples, meaning prediction deviations are

small and stable. The model performs well on both low-score (1.0-2.5) and high-score (3.5-5.0) artworks, without obvious bias toward over-estimation or under-estimation. No extreme outliers appear, showing that the L1 loss used in training effectively suppresses disturbance from noisy human annotations. The scatter distribution directly supports the numerical metrics in Table 1, proving that the high PLCC/SRCC and low MSE stem from genuine alignment with human aesthetics, not accidental metric optimization.

Fig. 3 shows case-level aesthetic score prediction by MDAF-ViT on three representative art styles: Impressionist Painting, Classical Chinese Ink Painting, and Abstract Art, along with key aesthetic contributors.

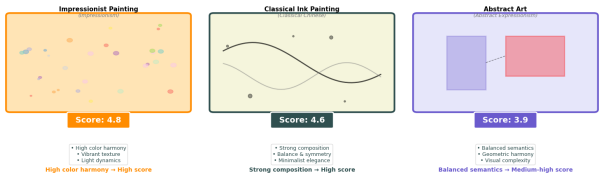

Fig. 3. Qualitative Aesthetic Score Prediction on Sample Artworks

Impressionist Painting receives a high score (4.8) mainly due to excellent color harmony and texture, which aligns with the aesthetic essence of Impressionism. Classical Ink Painting scores 4.6, dominated by strict composition, balance, symmetry, and minimalist elegance, consistent with traditional oriental art criteria. Abstract Art scores 3.9, relying on balanced semantic expression and geometric harmony, reflecting the model's ability to evaluate non-representational art reasonably. The score explanation reveals that MDAF-ViT does not make black-box predictions; it weights low-, mid-, and high-level features according to stylistic characteristics, consistent with real-world art appreciation logic. The model correctly identifies the core aesthetic dimensions of Western oil painting, Chinese ink wash, and abstract art, demonstrating its capacity to generalize across cultural and stylistic boundaries.

3.3. Ablation Study

Ablation experiments verify the contribution of multi-branch extraction and the MDAF fusion module. Base: Vanilla ViT-B/16; w/MB: Multi-branch feature extraction (VL/C/S branches); w/MDAF: Dynamic Multi-Dimensional Attention Fusion; MDAF-ViT: Full model. Table 2 validates the contribution of core components via progressive ablation: Vanilla ViT-B/16 → +Multi-branch (MB) → +MB+Concat Fusion → Full MDAF-ViT (MB + MDAF).

Adding multi-branch extraction boosts PLCC from 0.815 to 0.848 (+3.3% absolute improvement), confirming that separating low-, mid-, and high-level artistic features significantly enhances representation. Replacing simple concatenation with the MDAF module further increases PLCC from 0.859 to 0.879 and reduces MSE from 0.166 to 0.145. This proves that adaptive feature weighting is superior to fixed fusion strategies, as it emphasizes the most salient aesthetic dimensions for each input. The full MDAF-ViT achieves the best performance, showing that multi-branch decomposition and dynamic attention fusion are complementary and together form a high-performance aesthetic evaluation system. Multi-branch extraction significantly boosts performance (PLCC+3.3%).

Table 2. Ablation results on BAID dataset

Setting	PLCC	SRCC	MSE
Vanilla ViT-B/16	0.815	0.802	0.218
+ Multi-branch (MB)	0.848	0.836	0.179
+MB+ Concate Fusion	0.859	0.847	0.166
+MB+MDAF (Ours)	0.879	0.869	0.145

Dynamic attention fusion (MDAF) outperforms simple concatenation. Full MDAF-ViT achieves the best results.

Table 3 evaluates the importance of each feature branch by removing one branch at a time from the full MDAF-ViT.

Table 3. Branch ablation on BAID

Setting	PLCC	SRCC
Full MDAF-ViT	0.879	0.869
w/o Low-Level (VL) Branch	0.851	0.840
w/o Mid-Level (C) Branch	0.843	0.831
w/o High-Level (S) Branch	0.847	0.835

Removing any single branch causes clear performance drops (PLCC decreases by at least 0.028), meaning low-level visual, mid-level compositional, and high-level semantic/style features are all essential for accurate aesthetic assessment. Removing the mid-level composition branch leads to the largest PLCC drop ($0.879 \rightarrow 0.843$), suggesting that compositional rules are the most critical factor in artwork aesthetics for the tested dataset. Removing the low-level visual branch or high-level semantic branch also causes noticeable declines, confirming that no single dimension can fully represent artistic quality. The results support the paper's core motivation: artwork aesthetics rely on the synergistic integration of multi-level features, not isolated visual elements.

Fig. 4 visualizes the adaptive weighting of low-, mid-, and high-level features by the MDAF module across Impressionism, Classical Painting, and Surrealism.

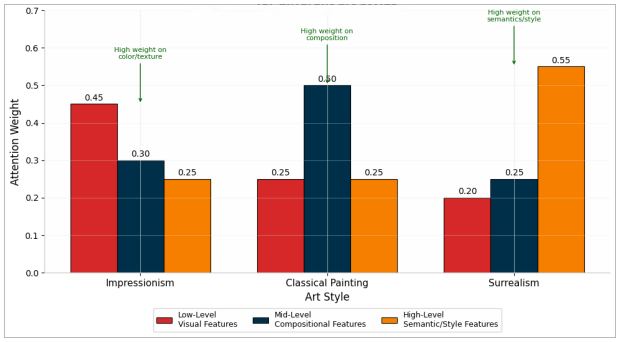

Fig. 4. Attention weight distribution of the MDAF module for different art styles

For Impressionism, the model assigns high weight to low-level color/texture features, matching the style's focus on light and color. For Classical Painting, the model emphasizes mid-level compositional features, consistent with classical art's emphasis on balance, perspective, and structural harmony. For Surrealism, the model prioritizes high-level semantic/style features, reflecting the importance of thematic meaning and conceptual expression.

The dynamic attention mechanism automatically identifies and strengthens the dominant aesthetic dimension of each style, which static fusion methods cannot achieve. This adaptive behavior explains why MDAF-ViT outperforms fixed-fusion baselines in Tables 1 and 2. The weight distribution provides transparent reasoning for aesthetic scores, making the model applicable in interpretable art analysis, intelligent curation, and art education tools. It confirms that the model learns human-like aesthetic evaluation logic rather than purely statistical patterns.

4. Conclusions

This study proposes the MDAF-ViT framework to address the limitations of existing methods in automated quantitative aesthetic evaluation of visual artworks. By integrating a hierarchical Vision Transformer backbone with a three-branch feature extraction pipeline, the model effectively captures low-level visual attributes, mid-level compositional rules, and high-level semantic style features. The core Dynamic Multi-Dimensional Attention Fusion module adaptively weights and integrates heterogeneous artistic features, overcoming the drawbacks of static fusion strategies. Extensive experiments on BAID, APDDv2, and JenAesthetics datasets demonstrate that MDAF-ViT surpasses state-of-the-art CNN and ViT-based models, with higher PLCC and SRCC values and lower MSE, showing stronger alignment with human aesthetic judgment. Ablation studies verify the efficacy of multi-branch extraction and dynamic attention fusion. This work provides a robust, interpretable approach for large-scale digital art analysis, intelligent curation, and generative art quality assessment, offering a reliable technical basis for the intersection of computational vision and art research.

References

[1] D. Jonauskaite, N. Dael, L. Baboulaz, L. Chèvre, I. Cierny, N. Ducimetière, A. Fekete, P. Gabioud, H. Leder, M. Vetterli, et al., (2024) *"Interactive digital engagement with visual artworks and cultural artefacts enhances user aesthetic experiences in the laboratory and museum"* **International Journal of Human–Computer Interaction *40***(6): 1369–1382. DOI: 10.1080/10447318.2022.2143767.

[2] E. Stamkou, D. Keltner, R. Corona, E. Aksoy, and A. S. Cowen, (2024) *"Emotional palette: A computational mapping of aesthetic experiences evoked by visual art"* **Scientific Reports *14***(1): 19932. DOI: 10.1038/s41598-024-69686-9.

[3] E. A. Vessel and H. Ovadia, (2025) *"The role of the default mode network in aesthetic appeal"* **Current Opinion in Behavioral Sciences *66***: 101608. DOI: 10.1016/j.cobeha.2025.101608.

[4] T. Shi, C. Chen, X. Li, and A. Hao, (2024) *"Semantic and style based multiple reference learning for artistic and general image aesthetic assessment"* **Neurocomputing *582***: 127434. DOI: 10.1016/j.neucom.2024.127434.

[5] X. Zhang, Y. Xiao, J. Peng, X. Gao, and B. Hu, (2024) *"Confidence-based dynamic cross-modal memory network for image aesthetic assessment"* **Pattern Recognition *149***: 110227. DOI: 10.1016/j.patcog.2023.110227.

[6] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, (2015) *"Rating image aesthetics using deep learning"* **IEEE Transactions on Multimedia *17***(11): 2021–2034. DOI: 10.1109/TMM.2015.2477040.

[7] H. Jang and J.-S. Lee, (2021) *"Analysis of deep features for image aesthetic assessment"* **IEEE Access *9***: 29850–29861. DOI: 10.1109/ACCESS.2021.3060171.

[8] J. McCormack and A. Lomas. "Understanding aesthetic evaluation using deep learning". In: *International conference on computational intelligence in music, sound, art and design (part of EvoStar)*. Springer. 2020, 118–133. DOI: 10.1007/978-3-030-43859-3_9.

[9] X. Tian, Z. Dong, K. Yang, and T. Mei, (2015) *"Query-dependent aesthetic model with deep learning for photo quality assessment"* **IEEE Transactions on Multimedia *17***(11): 2035–2048. DOI: 10.1109/TMM.2015.2479916.

[10] Y. Deng, C. C. Loy, and X. Tang, (2017) *"Image aesthetic assessment: An experimental survey"* **IEEE Signal Processing Magazine *34***(4): 80–106. DOI: 10.1109/MSP.2017.2696576.

[11] Y. Ke, Y. Wang, K. Wang, F. Qin, J. Guo, and S. Yang, (2023) *"Image aesthetics assessment using composite features from transformer and CNN"* **Multimedia Systems *29***(5): 2483–2494. DOI: 10.1007/s00530-023-01141-7.

[12] M. Carrasco, C. González-Martín, J. Aranda, and L. Oliveros, (2026) *"Vision Transformer attention alignment with human visual perception in aesthetic object evaluation"* **Plos one *21***(4): e0344006. DOI: 10.1371/journal.pone.0344006.

[13] S. Li, H. Liang, M. Xie, and X. He. "Multi-scale and multi-patch aggregation network based on dual-column vision fusion for image aesthetics assessment". In: *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2024, 1–6. DOI: 10.1109/ICME57554.2024.10687850.

[14] H. Takimoto, F. Omori, and A. Kanagawa, (2021) *"Image aesthetics assessment based on multi-stream CNN architecture and saliency features"* **Applied Artificial Intelligence *35***(1): 25–40. DOI: 10.1080/08839514.2020.1839197.

[15] D. Soydaner and J. Wagemans, (2024) *"Multi-task convolutional neural network for image aesthetic assessment"* **Ieee Access *12***: 4716–4729. DOI: 10.1109/ACCESS.2024.3349961.

[16] C.-H. Lee, J.-L. Shih, W.-L. Su, C.-C. Lien, and C.-C. Han. "Combination of Global Maximum Pooling and Local Average Pooling for Unsupervised Fine-Grained Image Retrieval". In: *Proceedings of the 2024 7th Artificial Intelligence and Cloud Computing Conference*. 2024, 299–307. DOI: 10.1145/3719384.3719427.

[17] J. Yao, J. Yao, Y. Yang, and C. Huang. "Hierarchical Adaptive Position Encoding-Based Transformer for Point Cloud Analysis". In: *International Conference on Neural Information Processing*. Springer. 2024, 197–210. DOI: 10.1007/978-981-96-6576-1_14.

[18] R. Egele, J. Junior, J. CS, J. N. van Rijn, I. Guyon, X. Baró, A. Clapés, P. Balaprakash, S. Escalera, T. Moeslund, et al., (2024) *"Ai competitions and benchmarks: Dataset development"* **arXiv preprint arXiv:2404.09703**: DOI: 10.48550/arXiv.2404.09703.

[19] U. Lee, Y. Son, J. Shin, G. Byun, Y. Lee, J. Koh, M. Jeon, and H. Kim. "LLaVA-Docent-V2: Improving Data Quality and Pedagogical Data Generation to Train Large Multimodal Models for Art Appreciation Education". In: *International Conference on Intelligent Tutoring Systems*. Springer. 2025, 213–228. DOI: 10.1007/978-3-031-98284-2_17.

[20] S. A. Amirshahi, G. U. Hayn-Leichsenring, J. Denzler, and C. Redies. "Jenaesthetics subjective dataset: analyzing paintings by subjective scores". In: *European Conference on Computer Vision*. Springer. 2014, 3–19. DOI: 10.1007/978-3-319-16178-5_1.