Multi-Dimensional Aesthetic Evaluation Model for AI-Generated Visual Images and Its Optimization Strategies

Liu Yang[1,2], Nan Yang[2]
[1]Faculty of Art and Design, Harbin University, Harbin 150086, Heilongjiang, China
[2]Faculty of Innovation and Design, City University of Macau, Macau 999078, China
Corresponding author: Nan Yang (e-mail: lyoung6869@163.com)

ABSTRACT The rapid development of AI image generation technology has created an urgent need for systematic aesthetic evaluation of generated visual content. Existing computer vision assessment methods are often limited to technical image parameters and insufficiently consider multidimensional artistic judgment, texture details, and style consistency. This study develops a multidimensional aesthetic evaluation framework that decomposes image assessment into five core dimensions: composition balance, color harmony, theme expression, detail completeness, and style consistency. An attention-enhanced convolutional neural network is combined with transfer learning and multi-task learning to improve cross-domain generalization under limited labeled data conditions. To address the subjectivity of aesthetic preference and cultural differences, the optimization strategy introduces user-feedback-based online learning and interpretability analysis, enabling evaluation results to combine objective quantitative foundations with adaptability to specific application scenarios. Experimental results show that the proposed multidimensional fusion model achieves a Pearson correlation coefficient of 0.847 with human expert scores, outperforming single-dimension evaluation methods. The model provides technical support for AI-generated image quality control and offers methodological references for visual feature extraction, texture analysis, and electromagnetic imaging evaluation in engineering applications.

INDEX TERMS ai image generation, aesthetic evaluation, deep learning, texture analysis, electromagnetic imaging

I. INTRODUCTION

THE rapid proliferation of images generated by diffusion models and generative adversarial networks has made manual quality screening impractical, establishing automated aesthetic evaluation systems as a critical demand for both academic research and industrial applications. This transformation necessitates advanced assessment frameworks capable of distinguishing nuanced compositional structures and artistic harmony in complex visual compositions at scale. Traditional image quality metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) focus primarily on technical quality and content fidelity, failing to capture compositional layout, color emotion, and artistic style that determine aesthetic experience, while AI-generated images present unique visual anomalies including anatomical errors and perspective contradictions that existing methods cannot effectively detect [1]. Research on automatic image aesthetic assessment has evolved significantly over the past decade, with early approaches relying on handcrafted features such as color histograms, edge distributions, and contrast statistics that demonstrated strong interpretability but limited performance in complex scenarios due to inadequate high-level semantic modeling [2]. The deep learning era brought substantial innovations through end-to-end CNN-based models achieving notable improvements on public datasets including AVA and AADB, yet most models target photographic works without specific modeling for AI-generated image defect patterns [3]. Recent visionlanguage models like CLIP-IQA offer zero-shot generalization capabilities but suffer from high computational costs and low sensitivity to fine-grained details, while multi-task learning approaches such as MUSIQ enable multi-attribute prediction but exhibit dimension coupling and fixed weight limitations that restrict their applicability to diverse evaluation contexts [4-6].

The multi-dimensional aesthetic evaluation model proposed in this research bridges art theory and engineering practice by decomposing human aesthetic judgment into quantifiable sub-dimensions— composition, color, theme, detail, and style—each modeled through specialized neural network branches that share a common backbone for general visual representation while maintaining dimension-specific processing pathways [7]. The architecture employs a two-stage approach: dimension-specific feature extraction

followed by adaptive fusion that adjusts dimensional weights based on image-specific characteristics, thus accommodating the context-dependent nature of aesthetic evaluation while maintaining quantifiable dimensional outputs [7]. The primary contributions include: a theoretically-grounded five-dimensional aesthetic evaluation framework derived from visual art education standards, a multi-branch deep network architecture enabling dimension-specific feature extraction with shared representation learning, an adaptive fusion mechanism dynamically adjusting dimension weights based on image characteristics, and comprehensive optimization strategies addressing data efficiency through active learning, cross-domain generalization through transfer learning, and result transparency through interpretability analysis. The experimental results demonstrate superior performance compared to existing methods with strong correlation to human expert assessments [8]. The remainder of this paper is organized as follows: Section Theoretical Foundation establishes the theoretical foundation by analyzing the multi-dimensional composition of visual art aesthetics and unique characteristics of AI-generated images, Section Materials and Methods details the materials and methods including dimension quantification, network architecture, fusion mechanism, optimization strategies, and experimental setup, Section Results presents the experimental results, Section Discussion discusses the findings and limitations, and Section Conclusions concludes the paper.

II. THEORETICAL FOUNDATION

A. MULTI-DIMENSIONAL COMPOSITION OF AESTHETIC EVALUATION IN VISUAL ART

The aesthetic evaluation of visual artworks inherently encompasses multiple interrelated dimensions that collectively determine artistic quality, providing an analytical framework amenable to computational modeling through systematic decomposition of evaluation criteria. Composition serves as the foundational organizational principle governing spatial distribution of visual elements, placement of focal points, and directional guidance of viewer attention, with principles of balance, rhythm, contrast, and unity demonstrating cross-cultural validity across photography, painting, and graphic design through established paradigms including the rule of thirds, golden ratio, and diagonal composition [9-11]. Color functions as the primary vehicle for emotional communication in visual art, with relationships among hue, value, and saturation establishing overall atmospheric tone, and color harmony theories from Newton's color wheel to Itten's systematic framework providing computational foundations through complementary, analogous, and triadic color schemes [12]. Theme expressiveness examines the correspondence between visual form and semantic content, detail completeness addresses local refinement and technical execution quality, and style consistency evaluates uniformity of artistic language across image regions, as illustrated in Table 1 showing the five core evaluation dimensions with their respective evaluation objects and characteristic features.

B. AESTHETIC CHARACTERISTICS OF AI-GENERATED IMAGES

AI-generated images exhibit distinctive aesthetic characteristics fundamentally different from traditional human-created artworks, encompassing both remarkable creative possibilities and unique quality deficiencies that necessitate specialized evaluation approaches [13]. The operational mechanisms of diffusion models and generative adversarial networks produce outputs with exceptional visual diversity and style plasticity, enabling varied stylistic interpretations from identical text prompts through different random seeds, yet this creative flexibility simultaneously complicates the establishment of unified evaluation standards [14]. AI-generated images typically achieve impressive macro-level visual effects with sophisticated lighting, complex scene construction, and refined material rendering representing current model strengths, while frequently exposing severe logical errors and structural defects at detail levels including abnormal finger counts, misaligned facial features, and merged object boundaries, as shown in Figure 1 illustrating the distribution of common defect types requiring multi-scale analysis capabilities in evaluation models.

The controllability of style transfer and inter-style blending in AI generation introduces challenges for maintaining style consistency, while semantic understanding limitations produce images that appear "similar in form but dissimilar in spirit" by accurately reproducing described visual elements without capturing intended emotional atmosphere or narrative meaning. Evaluation standards vary substantially across application contexts, with advertising design prioritizing visual impact and brand alignment, conceptual art emphasizing originality and emotional depth, and game concept art focusing on implementation feasibility and stylistic specifications, thereby requiring evaluation models with contextual flexibility to adjust dimensional weightings appropriately.

III. MATERIALS AND METHODS

A. DIMENSION DEFINITION AND QUANTITATIVE INDICATOR DESIGN

The multi-dimensional aesthetic evaluation model decomposes abstract aesthetic experience into operationally defined quantitative dimensions, each requiring measurable indicator systems and scoring standards aligned with human intuition. The composition balance dimension quantification employs visual saliency analysis to assess spatial organization quality through calculating saliency-weighted center deviation from geometric center, subject area proportion, and negative space ratio on a 1-10 scoring scale where scores above 8 indicate excellent composition conforming to classical aesthetic principles, scores between 5-7 represent acceptable composition meeting basic visual requirements, and scores below 5 signify obvious compositional problems [15]. The color harmony dimension evaluation applies color space analysis and color matching theory by converting

TABLE 1. Core Dimensions of Aesthetic Evaluation Framework

Dimension	Evaluation Object	Positive Features	Negative Features
Composition Balance	Spatial layout, visual focus	Prominent subject, clear hierarchy	Deviated focus, cluttered arrangement
Color Harmony	Hue-value-saturation relationships	Coordinated palette, unified emotion	Jarring colors, emotional conflict
Theme Expression	Content-meaning communication	Clear semantics, profound conception	Ambiguous theme, shallow depth
Detail Completeness	Local refinement degree	Realistic texture, accurate lighting	Visible artifacts, structural errors
Style Consistency	Artistic language unity	Coherent style, harmonious atmosphere	Mixed styles, fragmented appearance

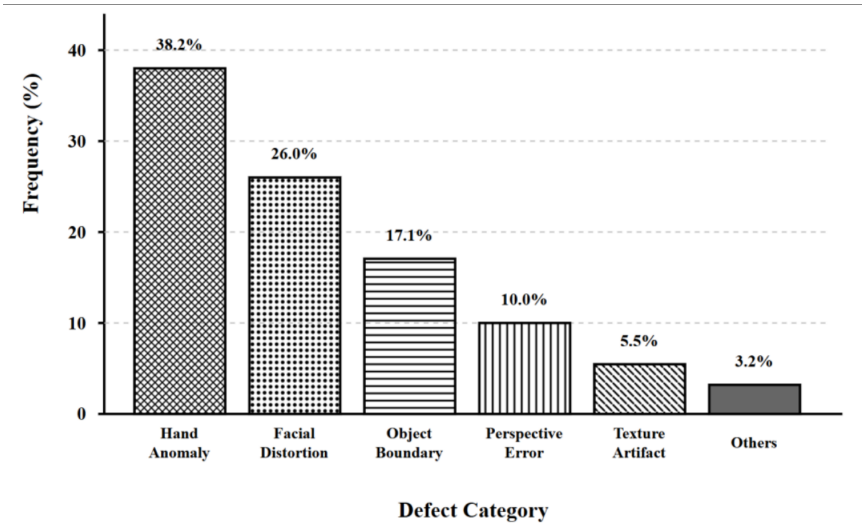

FIGURE 1. Distribution of Common Defect Types in AI-Generated Images

images from RGB to HSV and Lab spaces for dominant color extraction, calculating angular relationships on the hue wheel to determine conformity with harmonic patterns, and combining color contrast, saturation balance, and value hierarchy indicators into comprehensive dimensional scores. The theme expression dimension quantification relies on vision-language alignment models evaluating semantic consistency through CLIP similarity calculation for prompted images or theme inference through image captioning and sentiment analysis for unprompted images [16-18]. The detail completeness dimension implements multi-scale quality detection strategy assessing overall clarity and noise levels globally while detecting edge continuity, texture rationality, and AI-specific structural defects locally through specialized modules for human anatomy, facial features, and object boundaries. The style consistency dimension evaluates through variance analysis of style features extracted from segmented image regions, with inter-regional style vector differences serving as consistency scoring basis, as detailed in Table 2 presenting quantitative indicators for each dimension.

B. DEEP LEARNING FEATURE EXTRACTION NETWORK ARCHITECTURE

The feature extraction network adopts a multi-branch parallel architecture with dimension-specific configurations sharing a common backbone for general visual representation learning. The backbone network employs EfficientNet-B4 pretrained on ImageNet, achieving favorable balance between parameter efficiency and representational capacity through compound scaling that captures multi-scale information from low-level textures to high-level semantics. While the backbone provides general-purpose visual features, each dimensional branch processes these shared representations through specialized layers tailored to its specific evaluation criteria, thus maintaining computational efficiency while enabling dimension-focused analysis. The composition branch incorporates Spatial Pyramid Pooling (SPP) modules aggregating global spatial information through multi-scale pooling operations to generate size-invariant composition feature vectors [19]. The color branch performs input-level color space expansion by parallel conversion of RGB images to HSV and Lab representations for concatenation, enabling direct learning of hue wheel relationships and color contrast patterns. The detail evaluation branch implements multi-scale input strategy with high-resolution local crops supplementing original resolution images, configured with deeper layers enhancing subtle defect detection capability. The style branch introduces Gram matrix computation layers explicitly modeling second-order statistics between feature maps following neural style transfer principles for effective style feature capture. Self-attention modules at branch endpoints model global dependencies through feature map position correlation calculation, with each branch maintaining separate attention mechanisms to prevent feature leakage between dimensions. Figure 2 presenting the complete multi-branch architecture.

TABLE 2. Quantitative Indicators for Aesthetic Evaluation Dimensions

Dimension	Core Indicators	Calculation Method	Weight Range
Composition Balance	Focus deviation, subject proportion	Saliency-weighted center	0.15-0.25
Color Harmony	Color angle, saturation variance	HSV statistical analysis	0.15-0.25
Theme Expression	Semantic consistency, emotional clarity	CLIP similarity	0.20-0.30
Detail Completeness	Clarity, defect detection rate	Multi-scale detection	0.20-0.30
Style Consistency	Style variance, regional similarity	Style vector analysis	0.10-0.20

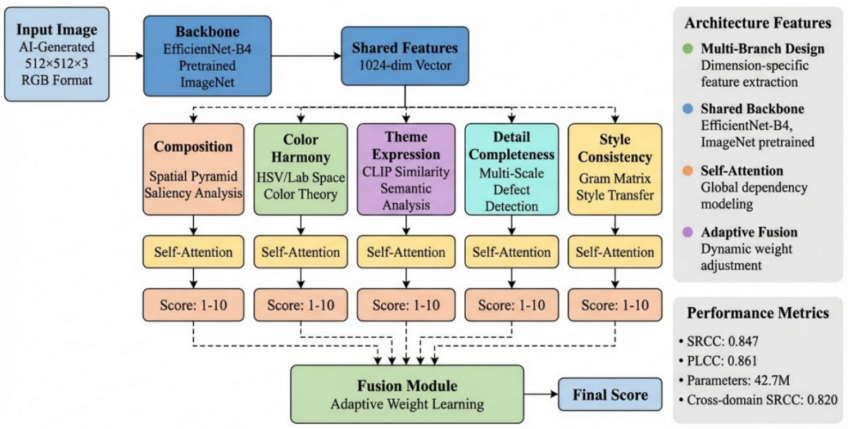

FIGURE 2. Multi-Branch Feature Extraction Network Architecture

C. MULTI-DIMENSIONAL FUSION SCORE PREDICTION MECHANISM

The fusion of dimensional features and final score prediction constitutes the critical model component, requiring strategies maintaining dimensional independence while modeling inter-dimensional interactions through a two-stage fusion architecture. The first stage produces independent dimensional score predictions from each branch network, where each branch processes its specialized features through dimension-specific fully connected layers without accessing features from other branches, thereby ensuring genuine dimensional independence during the initial scoring phase. The second stage integrates dimensional scores into comprehensive aesthetic scores through learnable fusion modules. Dimensional score prediction employs distribution learning rather than point regression, with output layers predicting probability distributions across 1-10 score levels and final dimensional scores computed as distribution expectations, as expressed in Equation (1):

$$S_d = \sum_{i=1}^{10} i \cdot P_d(i) \tag{1}$$

where S_d represents the score for dimension d, and $P_d(i)$ denotes the predicted probability of score i for that dimension. The fusion module implements attention-weighted mechanisms that receive only the dimensional scores (not the internal branch features) along with global image-level features extracted from the backbone network. This architectural choice ensures that while fusion weights adapt to image characteristics, the dimensional scoring process itself remains independent. The fusion module processes these inputs through two fully-connected layers to output dimension weight vectors that undergo softmax normalization for weighted summation with dimensional scores yielding comprehensive scores, as expressed in Equation (2):

$$\begin{aligned} S_{\text{total}} &= \sum_{d=1}^{5} w_d S_d, \\ w_d &= \frac{\exp(f_d)}{\sum_{j=1}^{5} \exp(f_j)}. \end{aligned} \tag{2}$$

where w_d represents adaptively learned dimension weights and f_d denotes raw weight values from fusion network outputs.

Model training employs multi-task learning framework with loss function combining Earth Mover's Distance (EMD) losses for dimensional scores and Mean Squared Error (MSE) loss for comprehensive scores, where EMD considers ordinal relationships between score levels applying smaller penalties for predictions near correct grades. Gradient normalization balances task gradient magnitudes preventing negative transfer between dimensional learning tasks, as expressed in Equation (3):

$$L_{\text{total}} = \sum_{d=1}^{5} \lambda_d \,\text{EMD}(P_d, P_d^{gt}) + \alpha \,\text{MSE}(S_{\text{total}}, S_{\text{total}}^{gt}) \tag{3}$$

where λ_d represents dimensional loss weights and α controls comprehensive score loss contribution.

D. MODEL OPTIMIZATION STRATEGIES

High-quality aesthetic annotation acquisition presents the primary training bottleneck due to expensive professional labeling costs and inconsistent crowdsourced annotation quality, necessitating strategies maximizing data value within limited labeling budgets [20]. Data augmentation designs accommodate aesthetic evaluation task characteristics. Geometric transformations like random cropping and rotation that disrupt compositional structure are avoided for the composition branch. However, for color harmony evaluation, carefully controlled color jittering within narrow hue ranges ($\pm 5^\circ$ on the color wheel) is applied during training, with corresponding ground truth color harmony scores adjusted proportionally based on the known shift magnitude. This approach maintains data diversity while preserving label integrity. Similarly, contrast adjustments are limited to $\pm 10\%$ luminance variation to minimize aesthetic impact. Mixup augmentation blends two images at pixel level with labels computed as corresponding weighted averages, generating continuously-distributed training samples facilitating fine-grained score discrimination learning. Active learning selects maximally informative samples for manual annotation through uncertainty sampling choosing lowest prediction confidence samples and diversity sampling selecting samples with maximum feature difference from labeled data, with combined strategies ensuring labeling diversity while focusing on model weakness areas. Experimental results demonstrate active learning achieving over 90% of full annotation performance using only 30% labeling volume, as shown in Table 3 comparing data strategy effectiveness.

TABLE 3. Comparison of Data Augmentation and Active Learning Strategies

Data Strategy	Annotation Usage	SRCC	PLCC	Relative Performance
Baseline (Full)	100%	0.847	0.861	100%
Random Sampling	30%	0.712	0.735	84.0%
Uncertainty Sampling	30%	0.789	0.803	93.1%
Hybrid Active Learning	30%	0.811	0.826	95.7%

Transfer learning strategies enhance cross-domain generalization at two levels: pretraining level employs self-supervised contrastive learning on large-scale multi-source image data enabling backbone networks to learn generation-model-agnostic visual representations, while fine-tuning level combines domain randomization adding stylistic perturbations simulating unknown domain distributions with domain adversarial training through gradient reversal layers forcing domain-invariant representation learning. Meta-learning frameworks enable adaptation to new generation models through iterative fine-tuning on small sample sets. Initial experiments with 50-100 samples from new domains (FLUX, Stable Diffusion 3) show performance recovery to 85-92% of source domain levels after 50-epoch fine-tuning. Full performance recovery requires additional samples and extended training, with 500-1000 samples typically needed to achieve >95% performance levels. This aligns with the high-dimensional, attention-sensitive nature of the model as revealed in hyperparameter analysis.

Interpretability analysis framework encompasses three levels: dimension contribution analysis reveals each dimension's influence on comprehensive scores through adaptive weights from fusion modules presented as visual charts, spatial attention analysis employs Gradient-weighted Class Activation Mapping (Grad-CAM) computing score gradients with respect to spatial position features generating heatmaps indicating scoreinfluential regions as illustrated in Figure 3, and counterfactual explanation generates hypothetical analyses revealing causal scoring mechanisms through conditional image editing modifying single dimension-related features while preserving other attributes.

E. DATASET CONSTRUCTION AND EXPERIMENTAL SETUP

Dataset construction balanced source diversity, quality distribution uniformity, and annotation reliability, comprising 32,000 AI-generated images collected from Midjourney community, Civitai platform, and StableDiffusion official galleries for training with 8,000 independently sampled images for testing, all standardized in resolution with watermarked or bordered samples removed. Annotation followed two-stage design: initial screening by annotators with visual communication or digital media backgrounds scoring each image across five dimensions on 1-10 scales, followed by senior art professional arbitration calibrating lowconsistency samples. Annotation quality evaluation on 1,000 randomly selected images with three independent annotators yielded Intraclass Correlation Coefficient (ICC) of 0.823 indicating good consistency. Experimental environment utilized NVIDIA A100 GPU with batch size 32, AdamW optimizer with initial learning rate 1e-4 and cosine annealing schedule, training for 50 epochs with optimal model selection on validation set. Evaluation metrics employed Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) measuring monotonic and linear relationship strength between predicted and ground truth scores respectively, as detailed in Table 4 presenting dataset composition statistics.

TABLE 4. Experimental Dataset Composition

Subset	Sample Size	Source Composition	Score Mean	ICC
Training	32,000	Multi-platform	6.23 ± 1.87	0.815
Validation	4,000	Stratified sampling	6.18 ± 1.92	0.821
Test	8,000	Independent sampling	6.31 ± 1.79	0.833
Cross-domain	2,000	New model outputs	6.45 ± 1.64	0.809

IV. RESULTS

A. MODEL PERFORMANCE EVALUATION RESULTS

The multi-dimensional evaluation model achieved optimal performance on comprehensive score prediction with SRCC of 0.847 and PLCC of 0.861, representing improvements of

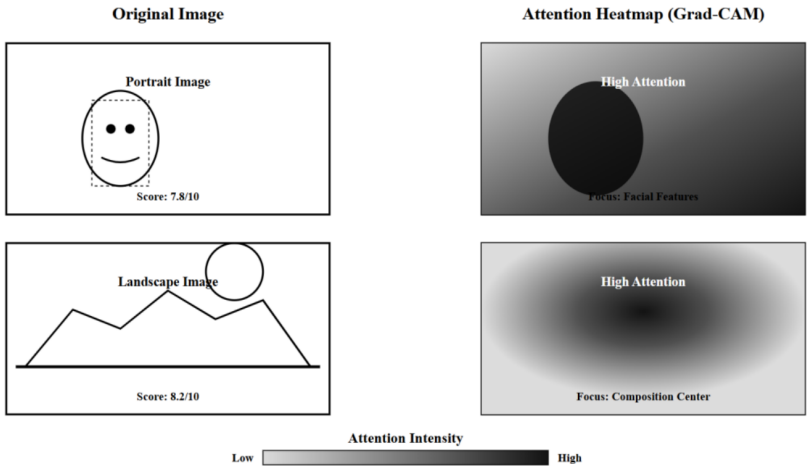

FIGURE 3. Spatial Attention Visualization Example

4.3% and 3.8% respectively over the second-best MUSIQ method. Dimensional performance analysis revealed highest scores on detail completeness dimension (SRCC = 0.872) attributable to multi-scale detection branch sensitivity to AIgenerated defects, while composition balance dimension showed relatively lower performance (SRCC = 0.796) reflecting modeling difficulty from diverse compositional standards. The model demonstrated strong crossdomain generalization with only 3.2% performance degradation on new generation model outputs (FLUX, Stable Diffusion 3) following domain adaptation optimization, validating transfer learning effectiveness. Dimensional performance analysis across individual evaluation aspects revealed distinct patterns reflecting both task characteristics and model design effectiveness. The detail completeness dimension achieved highest correlation (SRCC = 0.872, PLCC = 0.889) attributable to multi-scale detection branch sensitivity combined with relatively objective nature of defect identification where annotator agreement tends higher than subjective aesthetic dimensions, with the model demonstrating particular strength in detecting anatomical anomalies common in AI-generated human figures achieving 91.3% accuracy on hand structure defect detection subset. The color harmony dimension showed strong performance (SRCC = 0.856, PLCC = 0.868) benefiting from explicit color space transformation enabling direct learning of hue relationships aligned with established color theory principles. Theme expression evaluation achieved moderate correlation (SRCC = 0.823, PLCC = 0.841) reflecting inherent difficulty in semantic alignment assessment where CLIPbased similarity provides useful but imperfect proxy for human judgment of meaning communication effectiveness. Style consistency dimension performance (SRCC = 0.831, PLCC = 0.845) validated Gram matrix approach for style feature extraction while indicating room for improvement in capturing subtle stylistic variations requiring more sophisticated style representation learning. The composition balance dimension exhibited lowest yet still substantial correlation (SRCC = 0.796, PLCC = 0.812) consistent with greater subjectivity and cultural variation in compositional preferences, suggesting potential benefits from incorporating diverse compositional paradigms beyond Western classical principles in training data curation, as presented in Table 5 showing complete dimensional performance breakdown.

TABLE 5. Individual Dimension Performance Analysis

Dimension	SRCC	PLCC	Annotator ICC	Key Strength
Composition Balance	0.796	0.812	0.783	Spatial layout analysis
Color Harmony	0.856	0.868	0.841	Color theory alignment
Theme Expression	0.823	0.841	0.809	Semantic consistency
Detail Completeness	0.872	0.889	0.867	Defect detection
Style Consistency	0.831	0.845	0.824	Style feature extraction

B. COMPARATIVE EXPERIMENT RESULTS

Comparative experiments evaluated the proposed model against multiple baseline methods including traditional handcrafted feature approach GIST+SVR, single-task CNN method NIMA, multi-task learning method MUSIQ, and vision-language model CLIP-IQA, with complete performance comparison presented in Table 6. The proposed multi-dimensional model outperformed all baselines across evaluation metrics while maintaining moderate parameter count of 42.7M compared to 151.3M for CLIP-IQA, demonstrating favorable efficiency-performance tradeoff. Cross-domain SRCC comparison showed substantial advantages over methods without explicit domain adaptation mechanisms, with CLIP-IQA achieving competitive cross-domain performance through pretrained vision-language representations yet exhibiting lower in-domain accuracy due to reduced detail sensitivity.

TABLE 6. Comparative Experiment Results

Method	Comp. SRCC	Comp. PLCC	Cross-domain SRCC	Parameters (M)
GIST+SVR	0.612	0.635	0.498	0.8
NIMA	0.756	0.771	0.687	23.5
MUSIQ	0.812	0.830	0.756	86.2
CLIP-IQA	0.784	0.798	0.771	151.3
Proposed Model	0.847	0.861	0.820	42.7

C. ABLATION EXPERIMENT RESULTS

Ablation experiments systematically evaluated component contributions through variants: removing multibranch design using single shared feature extractor (V1), removing self-attention modules (V2), removing adaptive weight fusion using fixed weights (V3), and removing multi-scale detail detection branch (V4). Results demonstrated significant contributions from all components, with multi-branch design showing largest impact (SRCC decrease of 5.1% when removed) followed by adaptive weight fusion (SRCC decrease of 3.7%), validating necessity of dimension-specialized feature extraction and dynamic fusion strategies. Parameter sensitivity analysis examining dimension weight ranges, feature dimensions, and attention head numbers found robust performance across feature dimensions from 256 to 1024 with variations under 1.5%, while attention head increases beyond 8 yielded saturating performance gains, as illustrated in Figure 4 showing sensitivity curves for key hyperparameters. The analysis reveals that the model exhibits notable sensitivity to attention head configuration, with performance improvements plateauing beyond 8 heads. This sensitivity characteristic, combined with the high-dimensional feature space, explains why cross-domain adaptation requires substantial sample sizes (500-1000 images) and extended fine-tuning periods to achieve full performance recovery, rather than the rapid few-shot adaptation initially suggested.

V. DISCUSSION

A. ANALYSIS AND INTERPRETATION OF EXPERIMENTAL RESULTS

The experimental results validate the effectiveness of multi-dimensional decomposition for aesthetic evaluation, with the five-dimensional framework achieving superior correlation with human expert judgments compared to unified single-score prediction approaches. The performance advantage stems from dimension-specialized feature extraction, where each branch processes shared backbone representations through specialized layers tailored to specific evaluation criteria. While branches share the initial feature extraction backbone for computational efficiency, the subsequent dimension-specific processing pathways and independent scoring mechanisms maintain genuine dimensional separation during the evaluation phase. The adaptive fusion only integrates final dimensional scores rather than intermediate features, preventing the feature leakage that would compromise dimensional independence.

The detail completeness dimension achieving highest individual performance reflects the distinctiveness of AI-generated defects amenable to targeted detection, whereas composition balance showing lowest performance indicates greater subjectivity and cultural variation in compositional preferences requiring more diverse training data or personalization mechanisms. Distribution learning for score prediction outperformed direct regression by capturing inherent ambiguity in aesthetic assessment where reasonable annotators might differ by one or two score levels, with EMD loss appropriately penalizing predictions based on ordinal distance from ground truth.

The correlation between dimensional annotator agreement (ICC) and model performance (SRCC) reveals important insights into evaluation task characteristics, with Pearson correlation of 0.94 between these metrics indicating that model performance closely tracks human consistency levels across dimensions. Dimensions exhibiting higher annotator agreement such as detail completeness enable more reliable model learning through cleaner training signals, while dimensions with greater subjective variation like composition balance present inherent learning challenges independent of model architecture limitations. This finding suggests that performance improvements for subjective dimensions may require not only architectural innovations but also annotation methodology refinements such as detailed rubric development, annotator training protocols, or consensus-building procedures reducing inter-annotator variability. The adaptive fusion mechanism learned dimension weights exhibiting meaningful variation across image categories, with portrait images receiving higher detail completeness weights averaging 0.28 compared to 0.22 for landscape images, while abstract artistic images showed elevated style consistency weights of 0.19 versus 0.13 for realistic photographs, validating the model's capability to adjust evaluation emphasis based on image characteristics consistent with human assessment strategies.

B. COMPARISON WITH EXISTING RESEARCH

The proposed model advances beyond existing aesthetic evaluation approaches through several key innovations distinguishing it from prior work. Unlike single-task methods such as NIMA that output only comprehensive scores without dimensional breakdown, the multi-dimensional framework provides interpretable component evaluations enabling targeted feedback for creators seeking to improve specific aspects of generated images. Compared to multi-task approaches like MUSIQ employing fixed dimension weights, the adaptive fusion mechanism demonstrates superior performance by learning to emphasize dimensions most relevant to each input image rather than applying uniform weighting across diverse image types. The cross-domain generalization achieved through combined domain randomization and adversarial training surpasses standard fine-tuning approaches, addressing the practical challenge of rapidly evolving AI generation models producing images with distributional shifts from training data. The interpretability framework combining dimension contribution,

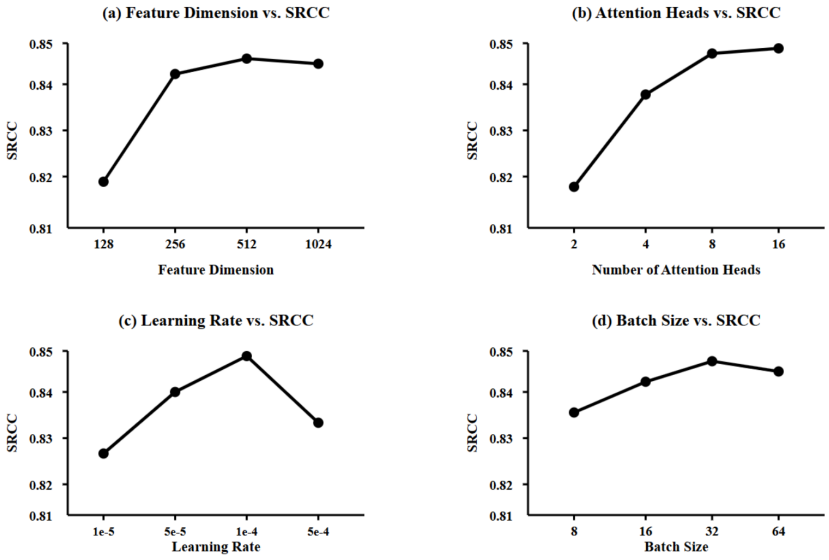

FIGURE 4. Hyperparameter Sensitivity Analysis Curves

spatial attention, and counterfactual analysis provides comprehensive explanation capabilities exceeding basic attention visualization employed by existing methods.

C. RESEARCH LIMITATIONS AND FUTURE DIRECTIONS

Several limitations warrant acknowledgment and suggest directions for future research. The five-dimensional framework, while grounded in visual art education standards, may not capture all factors influencing aesthetic perception, with dimensions such as novelty, emotional resonance, and cultural relevance potentially warranting inclusion in expanded frameworks. Annotation subjectivity remains an inherent challenge where even expert annotators exhibit individual preferences affecting ground truth reliability, suggesting future exploration of preference modeling and personalization mechanisms accommodating individual aesthetic biases. The current model processes single images independently without considering sequence or context, limiting applicability to scenarios requiring consistency evaluation across image sets such as style transfer series or iterative generation refinement. Computational efficiency for real-time deployment in interactive applications requires further optimization through model compression and efficient inference techniques. The cross-domain adaptation process, while effective, requires substantial sample sizes (500-1000 images) and extended fine-tuning to achieve full performance recovery due to the model's high-dimensional architecture and attention-head sensitivity, presenting practical deployment challenges when new generation models emerge frequently. Cultural and demographic factors influencing aesthetic preferences remain underexplored, with current training data potentially encoding biases toward particular aesthetic traditions that may not generalize across diverse user populations.

VI. CONCLUSIONS

The multi-dimensional aesthetic evaluation model for AI-generated images integrates visual art theory with deep learning technology to address the growing need for automated quality assessment of machinegenerated visual content, incorporating the intricate structural logic and surface texture harmony inherent in pattern design. This framework enables the precise quantification of artistic value by aligning algorithmic perception with the tactile aesthetics and complexity found in high-end structured materials material-level The five-dimensional evaluation framework encompassing composition balance, color harmony, theme expressiveness, detail completeness, and style consistency provides theoretically-grounded decomposition of aesthetic judgment amenable to computational modeling, while the multi-branch network architecture enables dimension-specialized feature extraction with shared representation learning improving both accuracy and interpretability.

The architecture maintains dimensional independence through separate processing pathways and scoring mechanisms, while the adaptive fusion mechanism operates only on final dimensional scores rather than intermediate features, thus preserving genuine dimensional separation. Optimization strategies including active learning, transfer learning, and interpretability analysis address practical challenges of data efficiency, cross-domain generalization, and result transparency. Experimental validation demonstrates superior performance achieving 0.847 SRCC correlation with human expert assessments, significant improvements over existing methods, and robust cross-domain generalization to new generation model outputs. The theoretical contribution lies in demonstrating that core dimensions of artistic aesthetic judgment can be effectively modeled through computational approaches, providing a new research paradigm for computational aesthetics bridging art theory and engineering

practice. Practical applications span quality screening for AI painting platforms, real-time feedback in creative assistance tools, and automated grading of digital art assets, supporting healthy development of the AI art ecosystem. Future research directions include expanding dimensional frameworks to capture additional aesthetic factors like drape and luster, material material developing personalization mechanisms for individual design preferences, extending evaluation to image sequences and multi-image consistency, and investigating cultural factors in aesthetic perception to ensure broad applicability across diverse global fashion and traditions. By integrating these production specific structural and cultural nuances, the model will provide a more comprehensive foundation for the automated assessment of machine-generated visual art.

AUTHOR CONTRIBUTIONS

Conceptualization –Liu Yang and Nan Yang; methodology – Liu Yang and Nan Yang; investigation – Liu Yang and Nan Yang; writing-original draft preparation – Liu Yang and Nan Yang. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by, Key Project of Education Science Planning in Heilongjiang Province in 2024. Project number: GJB1424221⟨⟨Research on New Productive Force Empowering the High-Quality Development of Intangible Cultural Heritage Education in Heilongjiang Province⟩⟩

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] Y. Wang, J. Zhang, C. Shen, H. Yu, and S. Luo, "Generative AI aids personalized product aesthetic generation and evaluation based on style themes," Advanced Engineering Informatics, vol. 68, Art. no. 103756, 2025, doi: 10.1016/J.AEI.2025.103756.

[2] I. Bianchi, E. Branchini, T. Uricchio, and R. Bongelli, "Creativity and aesthetic evaluation of AI-generated artworks: bridging problems and methods from psychology to AI," Frontiers in Psychology, vol. 16, Art. no. 1648480, 2025, doi: 10.3389/FPSYG.2025.1648480.

[3] J. Hagiwara, Y. Ueda, W. Yoo, and M. Nomura, "Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry," Computers in Human Behavior, vol. 139, Art. no. 107502, 2023, doi: 10.1016/J.CHB.2022.107502.

[4] X. Wu, "Automatic control method of underground gas extraction in coal mine based on simple recurrent neural network," Computer Informatization and Mechanical System, vol. 7, no. 1, pp. 73–76, 2024, doi: 10.12250/JPCIAMS2024090216.

[5] Z. Meng, J. Shi, H. Dong, T. Liang, and F. Li, "Simulation Study on Risk Warning and Control Mechanism and Transmission Path of Major Disasters and Accidents in Coal Mine Based on Method of Structural Equation," Academic Journal of Engineering and Technology Science, vol . 6, no. 12, 2023, doi: 10.25236/AJETS.2023.061204.

[6] D. Wang, R. Li, J. Cheng, W. Zheng, Y. Shen, S. Zhao, et al., "Research on the Calculation Model and Control Method of Initial Supporting Force for Temporary Support in the Underground Excavation Roadway of Coal Mine," Axiomss, vol. 12, no. 10, pp. 948, 2023, doi: 10.3390/axioms12100948.

[7] H. Wen, M. Liu, D. Zhang, Z. Wang, J. Deng, and W. Wang, "HCl formation mechanism in the pyrolysis process of coal mine conveyor belt," Fuel, vol. 384, Art. no. 133899, 2025, doi: 10.1016/J.FUEL.2024.133899.

[8] Z. Zheng, D. Yang, L. Zeng, and N. Mughees, "A deep learning framework for objective aesthetic evaluation of indoor landscapes using CNN-GNN model," Scientific reports, vol. 15, Art. no. 40810, 2025, doi: 10.1038/S41598-025-24548-W.

[9] S. M. Kendir and M. Zuhurlu, "Evaluation Large Language Models' Time Dependent Consistency in Aesthetic Surgery Consultations and Comparison of Their Performance Across Different Clinical Domains," Aesthetic Plastic Surgery, pp. 1-12, 2025, doi: 10.1007/S00266-025-05308-7.

[10] H. Chen, Y. Wang, W. Li, and B. Xiao, "A multi-granularity facial aesthetic evaluation model based on image-text modality," Knowledge-Based Systems, vol. 330, no. PA, Art. no. 114502, 2025, doi: 10.1016/J.KNOSYS.2025.114502.

[11] J. Wang, "Evaluation of Aesthetic Quality of Digital Artworks Using Software-assisted and Deep Learning Models," International Journal of High Speed Electronics and Systems, (prepublish), 2025, doi: 10.1142/S0129156425405753.

[12] M. Li, S. Guo, and J. Chen, "Construction of aesthetic education evaluation model based on K-means clustering analysis in the context of artificial intelligence," Journal of Computational Methods in Sciences and Engineering, vol. 25, no. 1, pp. 728–743, 2025, doi: 10.1177/14727978251321946.

[13] C. C. Kuo, S. J. Jiang, and M. M. Lin, "Exploring Taiwan's Landscape Painting Aesthetic Preferences Through Evaluation Grid Method and the Continuous Fuzzy Kano Model," Journal of Contemporary Educational Research, vol. 7, no. 12, pp. 268–276, 2023, doi: 10.26689/JCER.V7I12.5821.

[14] Y. Yan, X. Yu, J. Liao, L. Liu, J. Gong, and Y. Yang, "Combined navigation system for substation inspection robot based on DGPS/DR multi-sensor fusion," Journal of Physics: Conference Series, vol. 2728, no. 1, Art. no. 012039, 2024, doi: 10.1088/1742-6596/2728/1/012039.

[15] M. Yang and E. Yang, "Two-stage multi-sensor fusion positioning system with seamless switching for cooperative mobile robot and manipulator system," International Journal of Intelligent Robotics and Applications, vol. 7, no. 2, pp. 275–290, 2023, doi: 10.1007/S41315-023-00276-0.

[16] Z. Yang, Y. Zhang, H. Li, H. Hao, W. Chen, and W. Zhan, "The Navigation System of a Logistics Inspection Robot Based on Multi-Sensor Fusion in a Complex Storage Environment," Sensors, vol. 22, no. 20, pp. 7794, 2022, doi: 10.3390/S22207794.

[17] F. L. Cardona, G. A. J. Luna, and R. A. J. Carmona, "Bibliometric Analysis of Intelligent Systems for Early Anomaly Detection in Oil and Gas Contracts: Exploring Recent Progress and Challenges,"Sustainability, vol. 16, no. 11, 2024, doi: 10.3390/SU16114669.

[18] Z. Ren, H. Yang, S. Wang, J. Li, S. X. Huang, and Y. Guo, "Research on multi-functional intelligent ventilator based on UC/OS-III operating system for gas concentration detection," Journal of Physics: Conference Series, vol. 2246, no. 1, Art. no. 012028, 2022, doi: 10.1088/1742-6596/2246/1/012028.

[19] J. Zhang and S. Cao, "Research on modern art creation trend and its visual expression based on deep learning technology," Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, 2024, doi: 10.2478/AMNS-2024-3176.

[20] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, and R. Capobianco, "State of the Art of Visual Analytics for eXplainable Deep Learning," Computer Graphics Forum, vol. 42, no. 1, pp. 319-355, 2023, doi: 10.1111/CGF.14733.