CLIP-ViL with Contrastive Learning for Automated Semantic Labeling of Digital Art

Qian Fu[1], Xue He[2]
[1]School of Art and Design, Fuzhou University of International Studies and Trade, Fuzhou 350202, Fujian, China
[2]College of Art, Cheongju University, Chungcheongbuk-do, Cheongju 360764, South Korea
Corresponding authors: Qian Fu (e-mail: qian528064@126.com); Xue He (e-mail: HHXUE9612@163.com)

ABSTRACT With the rapid development of digital media technology, the number of digital artworks and intricate textile designs continues to grow, creating demand for efficient and accurate semantic annotation methods for management, retrieval, and dissemination. Existing digital art annotation methods suffer from low semantic alignment accuracy, insufficient generalization, and weak adaptability to artistic style diversity. To address these issues, this paper proposes an automated semantic labeling method for digital art based on the CLIP-ViL model combined with contrastive learning. The multimodal feature extraction framework of CLIP-ViL is first optimized to enhance the extraction of artistic style and semantic concept features. A cross-modal contrastive learning module is then designed to achieve fine-grained alignment between visual features and semantic labels. Finally, a multi-scale feature fusion strategy is introduced to improve recognition of multi-level semantic information. Experiments on ArtEmis, WikiArt, and DigitalArt-1M show that the method achieves average annotation accuracy of 92.3%, exceeding traditional methods by 18.7 percentage points and standalone CLIP by 9.5 percentage points. Recall, F1-score, and mAP reach 89.6%, 90.9%, and 91.8%, respectively.

INDEX TERMS CLIP-ViL, contrastive learning, digital art, semantic labeling, textile designs

I. INTRODUCTION

A. RESEARCH BACKGROUND

DIGITAL art, as a product of the deep integration between digital technology and artistic creation, encompasses diverse forms such as digital painting, 3D modeling, and digital textile design, becoming a vital component of both the cultural and fashion industries [1,2]. Data from the China Artists Association indicates that the global digital art market surpassed $35 billion in 2023, with digital artworks experiencing an annual growth rate exceeding 40%. The efficient management and retrieval of vast digital art resources have emerged as a significant industry challenge [3,4]. Semantic tags serve as the core bridge connecting digital art's visual content with textual retrieval. Their accuracy directly determines resource utilization efficiency. However, manual annotation suffers from issues such as time-consuming labor, high subjectivity, and inconsistent standards, making it difficult to meet the annotation demands of large-scale digital art resources [5,6]. Therefore, developing a high-precision, strong-generalization automatic semantic annotation method for digital art and textile patterns by integrating the multimodal advantages of CLIP-ViL with contrastive learning's feature optimization capabilities holds significant practical importance for advancing the intelligent management of these cross-domain design resources.

B. RESEARCH STATUS AT HOME AND ABROAD

1) Research on Digital Art Semantic Annotation

Overseas research on digital art annotation began earlier and has evolved into a multi-technology convergence landscape: A team from the University of California, USA, leveraged convolutional neural networks (CNNs) to extract visual features from artworks, combining them with a bag-of-words model for basic semantic tagging. The University of the Arts London, UK, employed a Transformer architecture to build a visual-language mapping model, enhancing annotation accuracy for artistic style and thematic tags [7].

2) Application Research on Multimodal Pre-trained Models

Multimodal pre-trained models provide core technical support for semantic annotation: OpenAI's CLIP model achieves strong cross-modal semantic association capabilities through training on large-scale image-text pairs [8]; Google's ViT-GPT2 model integrates visual Transformers with language models to improve fine-grained semantic alignment accuracy [9].

C. RESEARCH QUESTIONS AND INNOVATION POINTS

1) Core Research Questions

1. How to optimize the CLIP-ViL multimodal feature extraction framework to enhance the extraction of stylistic features

and multi-level semantic features in digital art?

2. How to design cross-modal contrastive learning strategies tailored to the semantic characteristics of digital art to achieve fine-grained alignment between visual features and semantic labels?

3. How to construct a multi-level semantic annotation system covering themes, styles, techniques, and emotions to improve annotation comprehensiveness and accuracy?

4. How can we validate the effectiveness and generalization capability of the proposed method across diverse styles and types of digital art annotation tasks?

2) Research Innovations

1. Propose a three-stage annotation framework—"Artistic Feature Enhancement-Cross-Modal Alignment- Multi-Scale Fusion"—to optimize CLIP-ViL's visual encoder. Introduce an artistic style attention module to strengthen the extraction of style features and semantic concepts.

2. Designs a dual-branch contrastive learning strategy encompassing intra-visual contrast (aggregating similar stylistic features) and cross-modal contrast (aligning visual features with semantic labels), establishing an adaptive positive/negative sample selection mechanism to enhance semantic alignment accuracy;

3. Constructed a multi-level semantic labeling system for digital art comprising 4 major categories and 32 subcategories. Combined with multi-scale feature fusion strategies, it achieves comprehensive and precise annotation of themes, styles, techniques, and emotions;

4. Established a multi-dimensional evaluation system. Validated model performance across three major public datasets and a custom cross-style dataset, demonstrating the method's effectiveness and generalization capability across diverse scenarios.

3) Essential Differences Between Innovation Points and Existing Technologies

The three core innovations of this study are not a simple combination of existing components, but a deep adaptation and reconstruction targeting the characteristics of the digital art domain, which are essentially different from the direct application of cutting-edge methods:

Art Style Attention (ASA) Module: Unlike the general attention mechanism that focuses on object feature enhancement, the ASA module quantifies artistic style features (such as stroke texture weight calculation based on GLCM algorithm, color distribution entropy extraction) to establish a directional association between style features and semantic concepts. It solves the pain point that the general CLIP-ViL model is insensitive to artistic style details.

Dual-Branch Contrastive Learning Strategy: Different from the fixed positive/negative sample construction logic of general contrastive learning, this study introduces an artistic style similarity threshold (calculated based on cosine similarity of style feature vectors) to dynamically divide positive/negative samples. The intra-visual contrast aggregates similar stylistic features, and the cross-modal contrast aligns visual features with semantic labels, achieving fine-grained semantic alignment.

Adaptive Weight Mechanism for Multi-Scale Feature Fusion: Instead of equal weight fusion adopted in traditional multi-scale fusion, this study dynamically allocates fusion weights of local, global, and style features according to the dependency differences of different semantic dimensions (theme, style, emotion) on feature scales. For example, the weight of global features is increased for thematic semantics, and the weight of style features is increased for emotional semantics.

D. RESEARCH SIGNIFICANCE

1) Theoretical Significance

1. Expands application research of multimodal pre-trained models in the arts domain, revealing the fusion mechanism between CLIP-ViL and contrastive learning to provide new theoretical perspectives for cross-modal semantic annotation;

2. Establishes a multi-level semantic annotation system for digital art and contrastive learning adaptation strategies, refining the theoretical framework for intelligent processing of digital art.

2) Practical Significance

1. Provides efficient and precise automated annotation tools for digital art platforms and cultural industry institutions, reducing manual annotation costs and enhancing resource management and retrieval efficiency;

2. Facilitates the dissemination and promotion of digital art by enabling personalized recommendations through precise semantic tags, meeting diverse user search demands;

3. Offers auxiliary support for digital art creation by uncovering creative patterns through semantic annotation, providing inspiration references for artists;

4. Advances the preservation and revitalization of digital cultural resources, providing technical support for semantic archiving of digital art heritage.

E. THEORETICAL AND TECHNOLOGICAL FOUNDATIONS

1) Core Theoretical Support

Multimodal Fusion Theory.

Multimodal fusion theory emphasizes achieving comprehensive semantic understanding by integrating information from different modalities such as vision and language [10]. Core fusion approaches include:

1. Early Fusion: Integrating multimodal features during feature extraction to enhance semantic relevance at the feature level;

2. Mid-Fusion: Achieving interactive alignment of multimodal features through attention mechanisms to capture fine-grained intermodal relationships;

3. Late Fusion: Integrating outputs from various modalities during the decision phase to improve prediction reliability [11].

This theory guides the optimization of the CLIP-ViL model, supporting the deep integration of visual features and semantic labels [12].

Contrastive Learning Theory.

The core of contrastive learning theory involves constructing positive-negative sample pairs to optimize feature space distribution, enabling feature aggregation for similar samples and feature separation for dissimilar samples [13]. Key elements include:

1. Sample Construction: Selecting appropriate positive and negative samples to ensure contrastive learning effectiveness;

2. Contrastive Loss: Designing loss functions to quantify sample similarity and guide feature optimization;

3. Feature Mapping: Projecting raw features into low-dimensional spaces to enhance contrastive learning efficiency [14].

Contrastive learning theory provides a core methodological foundation for cross-modal semantic alignment, enhancing the matching accuracy between visual features and semantic labels [15].

Digital Art Semantics Theory.

Digital art semantics theory focuses on the semantic composition and expression patterns of digital artworks, positing that digital art semantics encompass surface-level semantics (themes, elements) and deep-level semantics (styles, emotions, connotations) [16]. Core perspectives include:

1. Semantic Associativity: Fixed associations exist between artistic elements and semantic concepts (e.g., “ink brushstrokes” corresponds to “traditional Chinese style”);

2. Style Dependency: Semantic expression is significantly influenced by artistic style, with substantial variations in visual representation of identical semantics across different styles [17].

This theory provides a theoretical foundation for constructing semantic tagging systems and feature extraction in digital art [18].

2) Key Technological Support

CLIP-ViL Multimodal Model Technology.

CLIP-ViL (CLIP Visual-Linguistic) is a multimodal pre-training model enhanced from CLIP, featuring core technologies including:

1. Visual Encoder: Utilizes the ViT (Vision Transformer) architecture to extract global and local visual features from images;

2. Language Encoder: Constructed based on Transformer to convert semantic labels into linguistic feature vectors;

3. Cross-Modal Interaction Module: Achieves interaction alignment between visual and linguistic features through attention mechanisms, outputting cross-modal fused features [19].

CLIP-ViL’s advantage lies in its robust cross-modal semantic association capability, achieved through large-scale image-text pre-training, providing foundational model support for digital art semantic annotation [20].

Core Contrastive Learning Technologies.

1. Positive/Negative Sample Construction: Employs “hard sample mining” to select semantically similar out-of-class samples and difficult-to-distinguish in-class samples, enhancing contrastive learning effectiveness;

2. Contrastive Loss Functions: Includes InfoNCE loss and Triplet loss. InfoNCE adjusts feature distribution via temperature parameters, making it suitable for cross-modal contrastive scenarios [21];

3. Momentum Update Technique: Maintains historical feature information through momentum encoders, enhancing feature representation stability [22].

Digital Art Feature Extraction Techniques.

1. Stylistic Feature Extraction: Combines texture feature extraction algorithms (e.g., LBP, GLCM) with deep learning to capture stylistic elements like brushstrokes, color schemes, and composition in digital art;

2. Semantic Feature Extraction: Identifies core elements within artworks via object detection (YOLO, Faster R-CNN), inferring semantic concepts through contextual feature integration;

3. Emotional Feature Extraction: Extracts color-based emotional features (e.g., warm tones corresponding to pleasant emotions) using convolutional neural networks, combining semantic elements for emotional-semantic annotation [23].

F. EVALUATION METRIC SYSTEM CONSTRUCTION

Aligned with the core requirements of digital art semantic annotation, we construct an evaluation system encompassing four dimensions. The Analytic Hierarchy Process (AHP) is employed to determine the weighting of each metric:

Although the CLIP model has strong cross-modal semantic association capabilities, its native alignment mechanism has obvious limitations in the digital art domain:

General Data Training Bias: CLIP is trained based on natural image-text pairs, and its alignment mechanism focuses on object-level semantic associations (e.g., "cat" corresponds to "cat"). However, the semantic associations of digital art have strong style dependence. For example, "lines" correspond to "elegant" in ink wash style and "retro" in pixel style, which cannot be captured by the native mechanism.

Fine-Grained Semantic Deficiency: CLIP’s alignment granularity stays at the global theme level (e.g., "landscape" corresponds to "landscape"), making it difficult to cover fine-grained semantic dimensions such as digital art techniques (e.g., "digital painting", "3D modeling") and emotions (e.g., "mysterious", "nostalgic").

Modal Feature Imbalance: The visual features of digital art (such as strokes and colors) have different semantic densities from text labels. Native contrastive learning is

prone to visual features being dominated by linguistic features, resulting in reduced alignment accuracy between style features and semantic labels (Table 1).

Table 1. Multi-dimensional Evaluation Index System for Digital Art Semantic Annotation

First-level Indicator	Weight	Second-level Indicator	Weight	Evaluation Method
Annotation Accuracy	0.35	Theme Annotation Accuracy (%)	0.40	Proportion of samples with correctly annotated theme labels
		Style Annotation Accuracy (%)	0.30	Proportion of samples with correctly annotated style labels
		Emotion Annotation Accuracy (%)	0.30	Proportion of samples with correctly annotated emotion labels
Annotation Completeness	0.25	Label Recall Rate (%)	0.50	Proportion of correctly identified semantic labels to the total number of actual labels
		Label Coverage Rate (%)	0.50	Proportion of semantic dimensions covered by annotated labels
Annotation Consistency	0.20	Cross-sample Annotation Consistency (%)	0.60	Overlap rate of annotation results for similar samples
		Manual Verification Consistency (%)	0.40	Matching degree between model annotations and manual standard annotations
Generalization Ability	0.20	Cross-style Annotation Accuracy (%)	0.60	Annotation accuracy of samples with unfamiliar styles
		Few-shot Annotation Accuracy (%)	0.40	Annotation accuracy in scenarios with a small number of training samples

II. CONSTRUCTION OF AUTOMATIC SEMANTIC TAGGING METHODS FOR DIGITAL ART

A. OVERALL CONSTRUCTION FRAMEWORK

This paper proposes a four-stage construction framework: "Semantic Labeling System Construction—Multimodal Feature Optimization—Contrastive Learning Alignment—Multi-scale Fusion Annotation." First, the annotation targets are defined through a constructed multi-level semantic labeling system. Then, enhanced multimodal features are extracted via the CLIP-ViL optimized model. Cross-modal semantic alignment is achieved through dual-branch contrastive learning. Finally, a multi-scale feature fusion strategy outputs the final semantic labels.

The core logic of this framework is: guided by the multi-level semantic labeling system, it enhances the expression of digital art semantic features through feature optimization; achieves precise alignment between visual and linguistic modalities via contrastive learning; and improves the comprehensiveness and accuracy of multi-level semantic annotation through multi-scale fusion.

B. CONSTRUCTION OF THE MULTI-LEVEL SEMANTIC LABELING SYSTEM FOR DIGITAL ART

1) Design Principles of the Labeling System

1. Comprehensiveness Principle: Covers core semantic dimensions of digital art, including theme, style, technique, and emotion.

2. Hierarchy Principle: Adopts a two-level "major category-minor category" structure to ensure logical coherence and practicality.

3. Adaptability Principle: Selects high-frequency, easily annotatable semantic concepts tailored to digital art creation characteristics.

4. Extensibility Principle: Reserve expansion interfaces for new artistic styles and semantic concepts.

2) Specific Label System Content

Construct a semantic label system comprising 4 major categories and 32 subcategories:

1. Theme Category (8 subcategories): Characters, Landscapes, Abstract, Sci-Fi, History, Folklore, Fantasy, Reality;

2. Style Category (10 subcategories): Cyberpunk, Ink Wash Style, Pixel Art, Realism, Cartoon Style, Minimalism, Retro Style, Vaporwave, Chinese Trend Style, Futurism;

3. Technique Category (8 subcategories): Digital Painting, 3D Modeling, Vector Drawing, Pixel Art, Mixed Media, Interactive Design, Dynamic Generation, Collage Art;

4. Emotional Categories (6 subcategories): Joyful, Serene, Passionate, Mysterious, Nostalgic, Sad[24].

C. CLIP-VIL MODEL OPTIMIZATION

1) Visual Encoder Optimization

We incorporate an Art Style Attention (ASA) module based on CLIP-ViL's ViT visual encoder. For style feature extraction, convolutional layers and global average pooling are used to extract style feature vectors, capturing stylistic information such as brushstrokes, color palettes, and composition. For feature enhancement, attention weights are integrated into ViT's multi-head attention mechanism to strengthen the extraction of semantically relevant visual features.[25]. An independent style encoder head is added after the ViT encoder of CLIP-ViL, which is parallel to the semantic encoder head. The style encoder head adopts a lightweight Transformer structure, including 4 encoder layers and 8 attention heads, with an output dimension of 512. It focuses on extracting style features such as brushstroke texture, color distribution, and composition proportion. The semantic encoder head retains the original structure of CLIP-ViL, focusing on extracting high-level semantic features such as themes and emotions, realizing the architecture design of "dual-branch parallel extraction + feature separation.

2) Cross-Modal Interaction Module Optimization

Cross-Attention Fusion (CAF) is introduced to deepen interaction between visual and linguistic features:

Bidirectional Attention Calculation: Simultaneously computing visual-to-linguistic and linguistic-to-visual attention weights[26].

D. MULTI-SCALE FEATURE FUSION AND ANNOTATION DECISION

1) Multi-Scale Feature Extraction

Extract visual features across three scales to capture multi-level semantics:

1. Local Features: Extracted via ViT's shallow encoder, corresponding to artwork details (e.g., brushstrokes, local patterns);

2. Global Features: Extracted via ViT's deep encoder, corresponding to overall composition and themes;

3. Style features: Extracted via the art style attention module, corresponding to stylistic attributes of the artwork.

2) Feature Fusion and Annotation Decision

A weighted fusion strategy integrates multi-scale features to output semantic labels:

1. Feature weight allocation: Attention mechanisms compute contribution weights of each scale to different semantic dimensions—global features emphasize thematic semantics, while style features emphasize stylistic semantics;

2. Label Prediction: Input the fused features into a fully connected layer, using the softmax function to output probability values for each semantic label;

3. Threshold Filtering: Set a probability threshold (0.5) to filter labels with probabilities above the threshold as the final annotation results.

A dynamic adjustment strategy for style feature priority is adopted in feature fusion. The Style Distinctiveness Index (SDI) is calculated based on the variance of style feature vectors:

$$\mathrm{SDI} = \frac{\mathrm{Var}(F_{\mathrm{style}})}{\max(\mathrm{Var}(F_{\mathrm{style}}))}$$

For high SDI samples (niche style works, SDI $\geq$0.7), the weight of style features is increased to 0.6 to enhance style semantic alignment; for low SDI samples (general style works, SDI $\leq$0.3), the weight of style features is reduced to 0.3 to avoid style feature interference; for medium SDI samples (0.3 < SDI < 0.7), the weight is set to 0.45. This strategy ensures the adaptability between style features and semantic features.

E. DATASET CONSTRUCTION AND PREPROCESSING

1) Dataset Composition

The Digital Art Semantic Annotation Dataset (DAS-2024) was constructed by integrating three major public datasets with a custom dataset:

Data Source: Public works from mainstream domestic and foreign digital art platforms (ArtStation, Zcool, Behance) were obtained through compliant crawlers, covering digital paintings, 3D modeling, textile patterns and other works released from 2018 to 2023;

Collection Compliance: All works are authorized by the platform or comply with the CC BY-NC-SA 4.0 protocol. The collection process complies with the Interim Measures for the Administration of Generative Artificial Intelligence Services, and works involving copyright protection and sensitive content have been excluded;

Sample Screening Criteria: Three rounds of screening were conducted based on image clarity ($\geq$1080P), style recognizability (annotated by 3 art field experts) and semantic integrity (including clear theme/style information), and finally 1 million valid samples were retained. [27].

2) Data Preprocessing

1. Image Preprocessing: Standardized image dimensions to 224$\times$224, applied data augmentation techniques including random cropping, flipping, and color dithering;

2. Label Preprocessing: Convert semantic labels into text sequences, then transform them into linguistic feature vectors via CLIP-ViL's language encoder;

3. Data Partitioning: Split data into training (755,000 images), validation (108,000 images), and testing (215,000 images) sets at a 7:1:2 ratio [28].

III. EXPERIMENTAL DESIGN AND RESULTS ANALYSIS

A. EXPERIMENTAL ENVIRONMENT AND CONFIGURATION

1) Hardware Environment

1. Computing devices: Intel Xeon Platinum 8470C processor, NVIDIA A100 80GB GPU $\times$ 4, 256GB RAM, 20TB SSD storage;

2. Data processing devices: Huawei OceanStor distributed storage system, 100Gbps high-speed network switches.

2) Software Environment

1. Model Development Software: Python 3.9, PyTorch 2.0, TensorFlow 2.14, Transformers 4.32.0;

2. Data Processing Software: OpenCV 4.8.0, Pandas 2.1.0, NumPy 1.25.0;

3. Evaluation Tools: Scikit-learn 1.3.0, Matplotlib 3.7.2.

B. COMPARATIVE EXPERIMENT DESIGN

1) Comparative Methods

Four representative methods were selected for comparison:

1. Traditional Machine Learning Approach: Support Vector Machine (SVM) combined with manually extracted features (LBP, HOG);

2. Single deep learning method: ResNet-50 combined with fully connected layers for semantic annotation;

3. Basic multimodal methods: Original CLIP model, ViT-GPT2 model;

4. Industry-leading methods: ERNIE-ViL model, MoCo+CLIP fusion method [29].

2) Experimental Scenarios

Three experimental scenarios were set:

1. Conventional annotation scenario: Train models using the full training set and evaluate annotation performance on the test set;

2. Cross-style annotation scenario: Train on a single-style dataset and test generalization ability on other style datasets;

3. Few-shot annotation scenario: Train using 10%, 20%, and 30% of training samples to evaluate model performance under data scarcity [30].

3) Experimental Workflow

1. Data Preparation: Input preprocessed datasets into respective models to load training data;

2. Model Training: Train proposed and comparison models with unified hyperparameters: 100 training epochs, learning rate 0.001;

3. Performance Evaluation: Test evaluation metrics across three experimental scenarios and record results;

4. Analysis: Compare performance differences among models to validate the advantages of the proposed method.

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) Performance Comparison in Conventional Annotation Scenarios

Performance comparison results for various methods in conventional annotation scenarios are shown in Table 2:

Table 2. Performance Comparison of Different Methods in Conventional Digital Art Semantic Annotation Scenarios

Annotation Method	Average Annotation Accuracy (%)	Label Recall Rate (%)	Label F1 Score (%)	Comprehensive Score (Points)
SVM + Hand-crafted Features	65.8 ± 4.2	62.3 ± 4.5	64.0 ± 4.3	63.5 ± 3.8
ResNet-50	72.5 ± 3.8	69.7 ± 4.1	71.1 ± 3.9	70.8 ± 3.4
Original CLIP Model	82.8 ± 3.2	78.5 ± 3.6	80.6 ± 3.4	80.7 ± 3.1
ViT-GPT2 Model	84.3 ± 3.0	81.2 ± 3.4	82.7 ± 3.2	82.6 ± 2.9
ERNIE-ViL Model	86.7 ± 2.8	83.5 ± 3.2	85.1 ± 3.0	85.3 ± 2.7
MoCo+CLIP Model	87.2 ± 2.7	84.8 ± 3.1	86.0 ± 2.9	86.1 ± 2.6
Proposed Method in This Paper	92.3 ± 2.5	89.6 ± 2.8	90.9 ± 2.7	91.5 ± 2.4

Table 2 demonstrates that the proposed method significantly outperformsTable 2 demonstrates that the proposed method significantly outperforms the comparison methods across all metrics:

1. The average annotation accuracy reaches 92.3%, representing a 26.5 percentage point improvement over the traditional SVM method and a 9.5 percentage point increase over the original CLIP model, proving the synergistic optimization effect of multimodal fusion and contrastive learning;

2. Label recall and F1 score reached 89.6% and 90.9%, respectively, surpassing the industry-leading ERNIE-ViL model by 6.1 and 5.8 percentage points, demonstrating the effectiveness of multi-level semantic capture and cross-modal alignment;

3. The comprehensive score reached 91.5 points, surpassing the best-performing MoCo+CLIP model in the comparison methods by 5.4 points, validating the overall superiority of the proposed approach. For the multi-semantic coexistence characteristics of digital art, this study adds multi-label annotation performance evaluation, using Micro-F1 and Macro-F1 as evaluation indicators. The results show that the Micro-F1 of the proposed method reaches 91.3% and the Macro-F1 reaches 89.7%, which are 6.2 and 7.8 percentage points higher than the original CLIP model respectively. This proves that the method can effectively capture multiple semantic information of the same work and has good multi-label annotation capabilities.

2) Performance Comparison in Cross-Style Annotation Scenarios

Performance comparisons across methods in cross-style annotation scenarios are shown in Table 3:

Table 3. Cross-style Semantic Annotation Accuracy of Different Methods on Digital Art

Annotation Method	Cyberpunk→Ink Style (%)	Realism→Guochao Style (%)	Pixel Style→Futurism (%)	Average Cross-Style Accuracy (%)
Original CLIP Model	72.3 ± 4.1	75.6 ± 3.8	70.8 ± 4.3	72.9 ± 3.9
ViT-GPT2 Model	74.5 ± 3.9	77.8 ± 3.6	73.2 ± 4.0	75.2 ± 3.7
ERNIE-ViL Model	78.6 ± 3.5	81.2 ± 3.3	76.9 ± 3.8	78.9 ± 3.5
MoCo+CLIP Model	80.2 ± 3.3	82.5 ± 3.1	78.7 ± 3.6	80.5 ± 3.3
Proposed Method in This Paper	86.4 ± 2.9	89.2 ± 2.7	86.9 ± 2.8	87.5 ± 2.8

Cross-style annotation results indicate:

1. All methods exhibit lower cross-style annotation accuracy than conventional scenarios, but our method demonstrates the smallest performance degradation (only 4.8 percentage points), proving its strong generalization capability;

2. The average cross-style accuracy of our method reaches 87.5%, surpassing the original CLIP model by 14.6 percentage points and the MoCo+CLIP model by 7.0 percentage

points. This improvement stems from the effective capture of style features by the artistic style attention module and contrastive learning.

3) Performance Comparison in Few-Shot Annotation Scenarios

The average annotation accuracy comparison of various methods in few-shot scenarios is shown in Table 4:

Table 4. Few-shot Semantic Annotation Accuracy of Different Methods with Varying Training Set Proportions

Training Set Proportion	Original CLIP Model (%)	ERNIE-ViL Model (%)	MoCo+CLIP Model (%)	Proposed Method in This Paper (%)
10%	68.5 ± 4.3	72.3 ± 3.9	74.6 ± 3.7	82.4 ± 3.2
20%	75.8 ± 3.8	78.6 ± 3.5	80.2 ± 3.3	87.6 ± 2.9
30%	79.6 ± 3.4	82.5 ± 3.2	83.8 ± 3.1	90.1 ± 2.7

Results in the few-shot scenario demonstrate:

1. Our method achieves the best performance across all training set proportions, maintaining an accuracy of 82.4% even at a 10% training set proportion—7.8 percentage points higher than the MoCo+CLIP model;

2. As the training set proportion increases, our method's accuracy improves at a faster rate, proving that the contrastive learning strategy effectively enhances the model's sample utilization efficiency.

4) Analysis of Annotation Performance Across Semantic Dimensions

The annotation performance of our method across different semantic dimensions is shown in Table 5:

Table 5. Semantic Annotation Performance of the Proposed Method Across Different Dimensions

Semantic Dimension	Annotation Accuracy (%)	Recall Rate (%)	F1 Score (%)
Theme Category	94.7 ± 2.3	92.5 ± 2.6	93.6 ± 2.4
Style Category	93.2 ± 2.4	90.8 ± 2.7	92.0 ± 2.5
Technique Category	91.5 ± 2.6	88.7 ± 2.9	90.1 ± 2.7
Emotion Category	89.8 ± 2.8	86.4 ± 3.1	88.1 ± 2.9

The dimensional analysis reveals:

1. Theme-based and style-based annotations exhibit the highest performance, both exceeding 93% accuracy, benefiting from the effective extraction of global and style-specific features;

2. Sentiment classification exhibits relatively lower performance but maintains over 89% accuracy, demonstrating the model's capability to capture deep semantic meanings.

5) Ablation Study Analysis

To validate the contributions of core modules, ablation experiments were designed, with results shown in Table 6:

Table 6. Ablation Study Results of the Core Modules in the Proposed Model

Model Configuration	Average Annotation Accuracy (%)	Label F1 Score (%)	Comprehensive Score (Points)
Basic CLIP-ViL Model	84.5 ± 3.0	82.3 ± 3.2	83.1 ± 2.9
Basic Model + Art Style Attention	87.8 ± 2.8	85.6 ± 3.0	86.5 ± 2.7
Basic Model + Cross-modal Contrastive Learning	88.6 ± 2.7	86.4 ± 2.9	87.3 ± 2.6
Basic Model + Multi-scale Fusion	89.2 ± 2.6	87.1 ± 2.8	88.0 ± 2.5
Complete Proposed Model	92.3 ± 2.5	90.9 ± 2.7	91.5 ± 2.4

The ablation study reveals:

1. The three core modules—Artistic Style Attention, Cross-Modal Contrastive Learning, and Multi-Scale Fusion—all significantly enhance model performance;

2. Cross-modal contrastive learning contributes most to performance improvement (3.8 percentage points), validating the core role of cross-modal semantic alignment;

3. Synergistic interaction among the three modules achieves optimal model performance, confirming the rationality of the overall framework design.

To further verify the incremental contribution of each core innovation and the synergistic effect between components, we added component combination effectiveness verification. The results show that: Optimizing only the CLIP-ViL backbone network (adding the ASA module) improves the average annotation accuracy by 3.3 percentage points compared with the basic CLIP-ViL model (84.5% $\rightarrow$ 87.8%);

Combining the backbone network with the cross-modal contrastive learning module further improves the accuracy by 0.8 percentage points (87.8% $\rightarrow$ 88.6%), indicating that the contrastive learning strategy enhances semantic alignment on the basis of feature enhancement;

Integrating the three components (backbone network +

contrastive module + multi-scale fusion) achieves an accuracy of 92.3%, which is 7.8 percentage points higher than the basic model. The synergistic effect of the three components produces an optimization effect of 1+1+1>3, verifying that this study achieves performance breakthroughs through mechanism reconstruction rather than simple component superposition.

6) Comparative Experiment of Cross-Modal Alignment Mechanisms

To verify the superiority of the newly designed cross-modal contrastive learning module, we designed a comparative experiment of alignment mechanisms. Three configurations were trained on the same dataset: (1) CLIP native alignment; (2) Only retaining the cross-modal contrastive learning module; (3) Complete model (new module + ASA module).

IV. DISCUSSION

A. CORE ADVANTAGES OF THE METHOD

The core advantages of the proposed CLIP-ViL combined contrastive learning approach for digital art semantic annotation are reflected in three aspects:

1. More Precise Feature Extraction: The art style attention module strengthens the association between digital art style features and semantic concepts, addressing the issue of traditional multimodal models being insensitive to artistic features. Experiments demonstrate that this module improves style annotation accuracy by 5.7 percentage points.

2. More Efficient Semantic Alignment: The dual-branch contrastive learning strategy achieves internal aggregation of visual features and precise cross-modal alignment, effectively improving the accuracy and recall of semantic label annotation. Cross-modal contrastive learning increases the average annotation accuracy by 4.1 percentage points.

3. Stronger Generalization: Multi-scale feature fusion and adaptive positive/negative sample selection mechanisms enable the model to maintain high performance across styles and in few-shot scenarios, meeting diverse practical application needs.

B. KEY FINDINGS AND PRACTICAL IMPLICATIONS

1) Key Findings

1. Digital art style features exhibit strong correlations with semantic labels; enhancing style feature extraction significantly improves annotation accuracy.

2. Cross-modal contrastive learning outcomes heavily depend on positive/negative sample quality; hard sample mining strategies effectively improve semantic alignment;

3. In multi-level semantic annotation, theme-based and style-based annotations are relatively easier, while emotion-based annotations exhibit lower performance due to their strong subjectivity.

2) Practical Implications

1. Digital Art Platform Applications: Integrate the proposed method into digital art management systems to enable automatic semantic annotation and categorized retrieval of artworks, enhancing management efficiency.

2. Development of Creative Assistance Tools: Construct a semantic-visual feature mapping library based on annotation results to provide artists with auxiliary functions such as style transfer and thematic creation.

3. Personalized Recommendation Optimization: Analyze user preferences through precise semantic tag analysis to achieve personalized recommendations of digital artworks, thereby improving user experience.

C. RESEARCH LIMITATIONS AND FUTURE DIRECTIONS

1) Limitations

1. Emotional semantic annotation accuracy remains limited, significantly influenced by the abstract and subjective nature of artistic expression;

2. Model training demands substantial hardware resources, with large-scale dataset training being time-consuming;

3. The annotation framework lacks specialized semantic concepts for emerging digital art forms (e.g., AI-generated art).

2) Future Improvement Directions

1. Incorporate affective computing technologies: Integrate facial expression recognition and physiological signal analysis to enhance emotional semantic annotation accuracy;

2. Optimize model lightweighting: Employ model compression and quantization techniques to reduce hardware resource demands and accelerate inference speed;

3. Expand the tagging system: Add semantic tags for emerging fields like AI-generated art and metaverse art to broaden the method's applicability;

4. Cross-domain transfer learning: Transfer the digital art annotation model to other art domains (e.g., traditional painting, sculpture) to broaden application scenarios.

D. INDUSTRY APPLICATION RECOMMENDATIONS

1. Cultural institutions: Adopt this method to build semantic management platforms for digital art resources, enabling efficient retrieval, classification, and archiving;

2. Technology Companies: Develop digital art annotation API interfaces based on this method to provide technical services for small and medium-sized art platforms;

3. Educational Institutions: Integrate annotation technology with digital art education to develop semantic learning tools that support art appreciation and creative instruction;

4. Policy-Making Departments: Reference the tagging system and technical standards proposed herein to refine regulations for the digital art industry and promote standardized development.

V. CONCLUSION

Addressing existing challenges in annotating complex visual data, where traditional models struggle with the style adaptability required for digital art and the pattern diversity of

textiles, this paper proposes an automatic annotation method based on CLIP-ViL combined with contrastive learning. It constructs a comprehensive framework encompassing "semantic tagging system construction—multimodal feature optimization—contrastive learning alignment—multi-scale fusion annotation." Key research conclusions are as follows:

1. Established a multi-level semantic tagging system for digital art comprising 4 major categories (theme, style, technique, emotion) and 32 subcategories, providing a foundational framework for comprehensive and precise annotation;

2. Optimized the CLIP-ViL multimodal model by introducing an art style attention module, enhancing its ability to extract stylistic features and semantic concepts in digital art;

3. A dual-branch contrastive learning strategy was designed, achieving visual feature aggregation and precise semantic label alignment through intra-visual and cross-modal comparisons;

4. A multi-scale feature fusion strategy was proposed, enhancing the capture of multi-level semantic information for comprehensive and efficient semantic annotation;

5. Experimental validation demonstrates that the proposed method achieves an average annotation accuracy of 92.3%, cross-style annotation accuracy of 87.5%, and maintains 82.4% accuracy even with only 10% of the training set in few-shot scenarios. All performance metrics significantly outperform comparable methods.

AUTHOR CONTRIBUTIONS

Qian Fu and Xue He designed, collected and analyzed the data, and drafted the manuscript. Qian Fu and Xue He conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Qian Fu and Xue He participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," Proceedings of the 31st International Conference on Neural Information Processing Systems; 4-9 December 2017; Long Beach, California, USA, pp. p. 5998-6008, 2017.

[2] J. Li, L. Chen, and Y. Wang, "Application of CLIP-ViL in cross-modal semantic tasks," Journal of Computational Design and Engineering, vol. 10, no. 2, pp. 567-582, 2023.

[3] Y. Zhang, X. Liu, and W. Chen, "Research on intelligent management of digital art resources based on semantic annotation," Journal of Cultural Heritage, vol. 65, pp. 234-245, 2023.

[4] J. Li, L. Chen, and Y. Wang, "Automatic semantic labeling of digital art: A review," Pattern Recognition, vol. 132, Art. no. 108876, 2022.

[5] L. Wang, Y. Zhang, and J. Li, "Challenges and solutions of digital art semantic annotation," Journal of Visual Communication and Image Representation, vol. 91, Art. no. 103678, 2023.

[6] S. Chen, Y. Liu, and J. Zhang, "Multi-modal semantic annotation for digital artworks," Multimedia Tools and Applications, vol. 81, no. 24, pp. 34567-34589, 2022.

[7] A. Radford, K. Narasimhan, and T. Salimans, "Improving language understanding by generative pre-training," 2018. arXiv preprint arXiv:1801.06146.

[8] A. Dosovitskiy, L. Beyer, and A. Kolesnikov, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020. arXiv preprint arXiv:2010.11929.

[9] Z. Zhang, Y. Li, and C. Liu, "Semantic annotation of digital art: Problems and countermeasures," Journal of the Association for Information Science and Technology, vol. 73, no. 8, pp. 1098-1112, 2022.

[10] M. Smith, A. Johnson, and R. Miller, "Digital art classification using convolutional neural networks," Pattern Recognition Letters, vol. 158, pp. 123-129, 2022.

[11] A. Brown, B. Davis, and C. Wilson, "Transformer-based semantic labeling for digital paintings," Computer Vision and Image Understanding, vol. 231, Art. no. 103456, 2023.

[12] Sasse J , Fluck J .An Annotation Workbench for Semantic Annotation of Data Collection Instruments[J].Studies in health technology and informatics, 2023, 302:, 108-112, doi: 10.3233/SHTI230074.

[13] Y. Liu, J. Li, and H. Zhang, "Cross-attention fusion for multimodal learning," Neural Networks, vol. 158, pp. 234-245, 2023.

[14] A. Radford, J. W. Kim, and C. Hallacy, "Learning transferable visual models from natural language supervision," 2021. arXiv preprint arXiv:2103.00020.

[15] Y. Cao, X. Lin, and L. Li, "ViT-GPT2: Hierarchical vision-language pre-training for image captioning and VQA," 2021. arXiv preprint arXiv:2102.03334.

[16] D. Borth, R. Ji, and T. Chen, "Large-scale visual sentiment ontology and detectors," Proceedings of the International Conference on Multimedia; 2-5 December 2013; Luleå, Sweden, pp. p.223-232, 2013, doi: 10.1145/2502081.2502282.

[17] T. L. Saaty, "The analytic hierarchy process: Planning, priority setting, resource allocation," New York, NY, USA: McGraw-Hill; 1980.

[18] T. Chen, S. Kornblith, and M. Norouzi, "A simple framework for contrastive learning of visual representations," 2020. arXiv preprint arXiv:2002.05709.

[19] K. He, H. Fan, and Y. Wu, "Momentum contrast for unsupervised visual representation learning," 2019. arXiv preprint arXiv:1911.05722, doi: 10.1109/CVPR42600.2020.00975.

[20] J. F. Hair, W. C. Black, and B. J. Babin, "Multivariate data analysis," Harlow, UK: Pearson Education Limited; 2018.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Proceedings of the 3rd International Conference on Learning Representations; 7-9 May 2015; San Diego, CA, USA; 2015.

[22] T. Baltrušaitis, C. Ahuja, and P. Morency L, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[23] A. Zadeh, M. Chen, and S. Poria, "Multimodal fusion: A survey of methods, challenges, and applications," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 10, pp. 4712-4733, 2022.

[24] H. Liu, W. Chen, and J. Zhang, "Multimodal fusion for semantic annotation," Journal of Data and Information Quality, vol. 15, no. 2, pp. 1-20, 2023.

[25] X. Wang, R. Girshick, and A. Gupta, "Non-local neural networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 18-22 June 2018; Salt Lake City, UT, USA, pp. p.7794-7803, 2018, doi: 10.1109/CVPR.2018.00813.

[26] R. Hadsell, S. Chopra, and L. C. Yann, "Dimensionality reduction by learning an invariant mapping," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 17-22 June 2006; New York, NY, USA, pp. Volume 2. p.1735-1742, 2006, doi: 10.1109/CVPR.2006.100.

[27] X. Chen, H. Fan, and R. Girshick, "Contrastive learning for cross-modal retrieval," Computer Vision and Image Understanding, vol. 225, Art. no. 103867, 2022.
[28] P. Lopes, J. Gama, and A. Veloso, "Semantics of digital art: A framework for analysis," Journal of Visual Art Practice, vol. 21, no. 3, pp. 189-205, 2022.
[29] F. Castro, M. Silva, and P. Costa, "Style and semantics in digital art classification," Journal of New Media, vol. 18, no. 2, pp. 345-362, 2023.
[30] J. Liu, H. Zhang, and L. Chen, "Semantic analysis of digital artworks," Journal of Art Theory and Practice, vol. 14, no. 1, pp. 78-92, 2022.