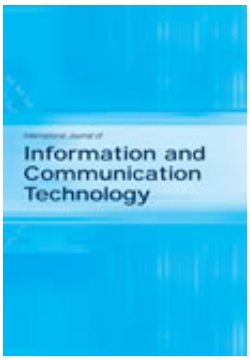


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Spatial domain semantic collaborative recognition model for complex emotions in artistic images

Jing Wang, Dali Zou

DOI: 10.1504/IJICT.2026.10076065

Article History:

Received:	22 October 2025
Last revised:	24 November 2025
Accepted:	27 November 2025
Published online:	16 February 2026

Spatial domain semantic collaborative recognition model for complex emotions in artistic images

Jing Wang

Department of Culture and Sports,
Zhejiang Vocational College of Special Education,
Hangzhou, 310013, China
Email: wangjwj2025@163.com

Dali Zou*

School of Design and Creativity,
Guilin University of Electronic Technology,
Beihai, 536000, China
Email: zoudali2025@163.com
*Corresponding author

Abstract: Oil paintings, watercolours and digital art convey human emotions. Complex emotions when visual elements blend with semantic information. Existing methods have three flaws: over reliance on low-level visual features misjudges serene loneliness; treating emotions as discrete labels misses ambiguity; and poor genre adaptability. This study proposes the spatial domain semantic collaborative recognition model for art complex emotions, via a dual-branch framework: spatial branch uses multi-scale convolutional neural network for global features, and semantic branch adopts graph attention network for semantic links. A cross-branch attention mechanism tunes visual; a Gaussian mixture model-based module quantifies emotion distribution. Experiments on two self-built datasets and public ArtEmis show: vs. traditional convolutional neural network and single-semantic models, it boosts accuracy by 28.3%, cuts mean absolute error by 32.1%, and maintains over 89% cross-genre accuracy. This work bridges the semantic-visual-emotional gap, supporting intelligent art curation, emotional interaction design and art therapy.

Keywords: artistic image; complex emotion recognition; spatial-semantic collaboration; graph attention network; Gaussian mixture model; style adaptability.

Reference to this paper should be made as follows: Wang, J. and Zou, D. (2026) 'Spatial domain semantic collaborative recognition model for complex emotions in artistic images', *Int. J. Information and Communication Technology*, Vol. 27, No. 11, pp.86–100.

Biographical notes: Jing Wang is a Professor in the Department of Culture and Physical Education at Zhejiang Vocational College of Special Education, China. She received a Master's degree from Suzhou University, China (2009) and a PhD in Management from HELP University, Malaysia (2021). She published 1 SSCI index paper and 11 papers. Her research interests include artistic image vision, artistic image emotional semiotics, spatial domain semantic modelling and multimodal collaborative learning.

Dali Zou is an Associate Professor at the School of Design and Creativity at Guilin University of Electronic Technology. He received a Master's degree in fine arts from Suzhou University, China (2010) and a PhD from HELP University, Malaysia (2021). He published five papers. His research interests include the collaborative integration of visual features and semantic understanding, image emotion recognition.

1 Introduction

Artistic images stand as a unique form of human cultural expression, with creators infusing subjective emotions into visual elements to stir resonance in viewers. Unlike natural images, they lean on stylised techniques – Impressionist colour blending or Cubist geometric decomposition – to convey complex emotions: these are emotional states woven from multiple intertwined components, not single discrete categories, like nostalgic joy merging longing for the past with present happiness, or melancholic calm blending sadness and tranquility (Deng et al., 2024). As digital art resources grow rapidly, demand for intelligent analysis of artistic emotions has spiked: art museums need to curate exhibitions around emotional themes, digital art platforms must recommend works matching users' real-time emotional needs, and art therapy institutions require images that aid emotional regulation (Al-Tameemi et al., 2024). The core of these applications is accurate recognition of complex emotions in artistic images, yet the stylised nature of artistic creation and ambiguity of complex emotions make this task highly challenging (Li et al., 2022). A survey by the International Association of Art Informatics notes existing emotion recognition systems for artistic images have an average error rate over 45% with complex emotions, far exceeding the 20% error rate for simple emotions, a gap rooted in misalignment between traditional methods and the traits of artistic complex emotions (Hou et al., 2022).

Three key challenges hinder progress in complex emotion recognition for artistic images (Elkobaisi et al., 2022). First, visual features and semantic information are separated: traditional methods focus on low-level visual features like colour histograms or texture entropy but overlook semantic correlations between elements – a rainy street in an oil painting might convey melancholy with dim lighting yet romance with a couple sharing an umbrella, and without semantic association modelling, models cannot tell these emotional differences apart (Lyu et al., 2024). Second, continuous emotions are labelled discretely: most studies use discrete labels to train models, but complex emotions are continuous and ambiguous; nostalgia, for instance, ranges from mild warm nostalgia to intense sorrowful nostalgia, and discrete labels fail to capture this gradient, with many viewers perceiving complex emotions in artistic images as a mix of multiple categories rather than a single label (Zhang et al., 2022a). Third, adaptability to artistic style differences is poor: different genres follow distinct expression rules – ink wash paintings use blank space to imply emotions, while pop art employs bright contrasting colours for direct emotional expression – and traditional models trained on one genre often struggle to generalise to others, with accuracy dropping significantly when tested on unfamiliar genres (Zhang et al., 2022b).

Scholars globally have done extensive research on image emotion recognition, but few have focused on the specific scenario of artistic complex emotions (Nie et al., 2024).

In natural image emotion recognition, early studies used handcrafted features to predict emotions; recent years have seen deep learning methods take the lead, with convolutional neural network (CNN) models extracting high-level visual features and transformer models capturing global spatial relationships (Zhang and Tan, 2024). Yet these methods are built for natural images and cannot handle the stylised features of artistic images, leading to low accuracy when applied to artistic works (Chen and Ibrahim, 2023). In artistic image analysis, research has centred on style classification and content recognition rather than emotion recognition: some models generate images of specific styles or recognise image themes but ignore emotional information, and even the small number of emotion-related studies use crowd-sourced labels to train CNN models, still treating emotions as discrete categories and failing to address semantic correlation or style adaptability (Manakitsa et al., 2024). In complex emotion modelling, psychology's Plutchik's emotion wheel divides complex emotions into combinations of basic emotions, but this theoretical framework has not been effectively integrated into computational models; in computer science, some studies use multi-label classification for mixed emotions but cannot capture continuous intensity of each emotional component, and while Gaussian mixture models (GMM) have been used to model continuous emotions in speech and text, their application in artistic images is rare (Bansal et al., 2024).

This study aims to solve the complex emotion recognition problem in artistic images by building a spatial domain semantic collaborative recognition model (SDSCRM), with three core goals: integrating spatial visual features and semantic information to establish a visual-semantic-emotional mapping relationship, modelling complex emotions as continuous probability distributions to capture their ambiguity and gradient traits, and enhancing the model's adaptability to different artistic genres to ensure stable recognition performance across styles (Liu et al., 2024). Its contributions are threefold. Theoretically, it proposes a spatial-semantic collaborative framework for artistic complex emotions, breaking the separation of visual and semantic analysis in traditional methods and laying a theoretical foundation for bridging the emotional gap in artistic image computing (Nie et al., 2024). Methodologically, it designs a multi-scale CNN-GAT dual-branch structure to extract spatial features and model semantic associations simultaneously, develops a cross-branch attention fusion mechanism that adjusts feature weights dynamically based on artistic style, and introduces a GMM-based continuous emotion regression module to quantify the intensity distribution of complex emotions (Wimpff et al., 2024). Practically, it constructs two high-quality datasets for artistic complex emotions to support subsequent research and verifies the model's application value in art curation, emotional recommendation, and art therapy through case studies (Chen et al., 2024).

2 Relevant technologies

2.1 Complex emotion representation in artistic images

Combining Plutchik's emotion wheel with artistic expression traits, this study defines complex emotions in artistic images as emotional states – woven from 2–3 basic emotions with continuous intensity gradients – conveyed through the interplay of visual elements and semantic context. The selection of the six specific categories – nostalgic joy, melancholic calm, excited anxiety, serene loneliness, sentimental sorrow, and hopeful fear – was guided by a systematic process that integrated the polarity and

complementarity of basic emotions from Plutchik's framework with their frequency of manifestation in artistic works. An initial pool of ten candidate complex emotions was evaluated by three art psychology experts through a structured scoring process, and the final six categories were retained based on their high inter-expert consistency exceeding 85%, ensuring both psychological plausibility and artistic relevance. These emotions fall into six categories: nostalgic joy, melancholic calm, excited anxiety, serene loneliness, sentimental sorrow, and hopeful fear, each expressed via genre-specific visual language (Wang et al., 2023).

Complex emotions in artistic images rely on two carriers for transmission. Visual carriers cover colour hue, saturation, brightness, composition symmetry, centering, blank space, and texture brush strokes, texture density – low saturation, asymmetric composition, and rough texture often signal melancholic calm (Liu et al., 2023). Semantic carriers include thematic context and cultural symbols, where specific elements implicitly link to emotions. To quantify these carriers, three core metrics are defined:

Visual feature intensity: quantifies visual elements' emotional contribution, calculated as:

$$VFI = \omega_c \times C + \omega_l \times L + \omega_t \times T \tag{1}$$

where C denotes colour emotion score, L composition emotion score, T texture emotion score, and ω_c, ω_l, ω_t are style-dependent weights summing to measures semantic-element-to-emotion correlation via semantic and emotion vector cosine similarity:

$$SAD = \frac{\vec{s} \cdot \vec{e}}{|\vec{s}| \times |\vec{e}|} \tag{2}$$

where $\vec{s}$ is the semantic element vector and $\vec{e}$ is the emotion vector.

Emotion intensity gradient: captures continuous variation in complex emotions via Euclidean distance between image emotion distribution and basic emotion standard distribution:

$$EIG = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{3}$$

where p_i is the probability of the i^{th} basic emotion in the image, and q_i is its standard probability.

2.2 Semantic-spatial feature modelling for emotion capture

Traditional CNN's with fixed convolution kernels fail to fully capture artistic images' multi-scale spatial features – small brush strokes and large layout alike. This study proposes a multi-scale CNN (MS-CNN) with three parallel branches: 3 × 3 kernel for local details brush edges, colour transitions, 5 × 5 kernel for regional features object shapes, and 7 × 7 kernel for global layout light distribution, blank space ratio. Branch outputs are fused via concatenation, with a batch normalisation layer added to mitigate overfitting; the fused feature map follows $H \times W \times C$, height H, width W, channel count C, summing branch channels (Zhang et al., 2022a). A spatial attention module is

appended post-MS-CNN to focus on emotion-relevant regions. It first compresses the MS-CNN feature maps channel dimension via global average pooling to get a spatial weight matrix $A \in \mathbb{R}^{H \times W}$, then normalises A to [0, 1] using sigmoid higher values mark emotion-relevant regions, and finally multiplies the original feature map by A to highlight key areas. For semantic association modelling, GAT is adopted – modelling element semantic links as a graph and adaptively assigning edge attention weights to prioritise critical links. The process has three steps:

Node embedding: extracts element semantic vectors via a pre-trained, art-caption-fine-tuned BERT model, yielding node features $h_i \in \mathbb{R}^d$.

Attention calculation: computes node *i–j* attention weight using a shared linear layer and *LeakyReLU*:

$$\alpha_{ij} = \frac{\exp\left(Leaky\,\mathrm{Re}\,LU\left(\vec{a}^T\left[Wh_i \,\|\, Wh_j\right]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(Leaky\,\mathrm{Re}\,LU\left(\vec{a}^T\left[Wh_i \,\|\, Wh_k\right]\right)\right)} \tag{4}$$

where W is the linear transformation matrix, $\vec{a}$ is the attention vector, and $\mathcal{N}_i$ is node neighbourhood.

Node update: updates node features via neighbourhood feature weighted summation:

$$h_{i'} = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} W h_j\right) \tag{5}$$

where denotes the sigmoid function.

Post-update node features are mapped to emotion space via a fully connected layer: $e_i = FC(h_i)$, $e_i \in \mathbb{R}^m$. The image's final semantic emotion feature is the average of all node emotion vectors:

$$E_{sem} = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{6}$$

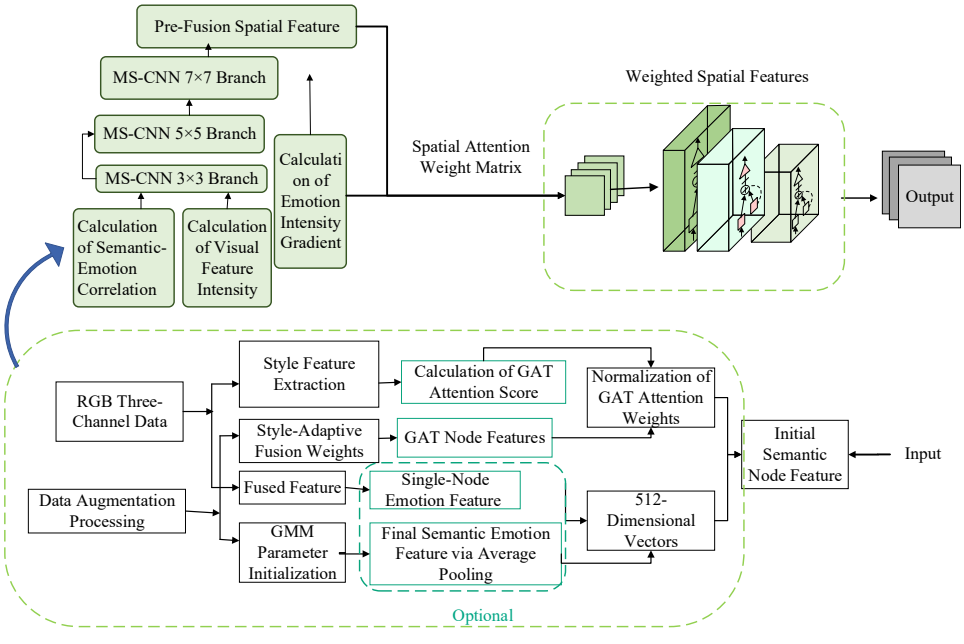
where n is the number of semantic nodes. Given complex emotions continuity and ambiguity, GMM is used to model their probability distribution – assuming image emotion intensity arises from K Gaussian distributions each corresponding to a basic emotion component:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(x \mid \mu_k, \sigma_k^2\right) \tag{7}$$

where x is the emotion intensity vector, π_k summing to 1 is the k^{th} basic emotion's proportion, μ_k is its average intensity, and σ_k^2 is its uncertainty. The EM algorithm estimates GMM parameters π_k, μ_k, σ_k^2, with model output being each basic emotion's probability distribution – directly reflecting complex emotions composition and intensity (Wang et al., 2024).

In the modelling of complex emotions in artistic images, the collaborative analysis of visual carriers and semantic carriers is core. The process of multi-scale spatial feature extraction – semantic association modelling – emotion distribution quantification proposed in modular connection to achieve a complete mapping from original images to emotion distributions Lian et al. (2023). To intuitively present the logical connections between key steps in this process and the corresponding relationships between input and output of each module, this study integrates the core calculations and module functions defined into a process diagram. It clearly demonstrates the technical chain of spatial-semantic dual-branch parallel modelling, style-aware fusion, and GMM emotion quantification. The specific process is shown in Figure 1.

Figure 1 Overall framework of semantic-spatial collaborative feature modelling for complex emotions in artistic image (see online version for colours)



3 Mathematical model of spatial domain semantic collaborative recognition

For an artistic image I, the core goal of the SDSCRM is to map I to a continuous emotion distribution:

$$P = \{p_1, p_2, \ldots, p_m\} \tag{8}$$

where p_k denotes the probability of the k^{th} basic emotion, $m = 8$ total basic emotions, and $\sum_{k=1}^{m} p_k = 1$. This mapping relies on three key steps, each formalised with equations

integrating critical parameters: For $I \in \mathbb{R}^{H\times W\times 3}$, RGB channels, height H, width W, the MS-CNN extracts small, medium, and large-scale features:

$$F_{spa1} \in \mathbb{R}^{H\times W\times C1},\ F_{spa2} \in \mathbb{R}^{H\times W\times C2},\ F_{spa3} \in \mathbb{R}^{H\times W\times C3} \tag{9}$$

Where ($C1$, $C2$, $C3$) are channel counts for 3 × 3, 5 × 5, 7 × 7 convolution branches, with pre-attention fused features calculated as:

$$F_{spa\_pre} = Concat\left(F_{spa1}, F_{spa2}, F_{spa3}\right) \in \mathbb{R}^{H\times W\times (C1+C2+C3)} \tag{10}$$

where *Concat* denotes the channel-wise concatenation operation. The attention weight matrix A for highlighting emotion-relevant regions is computed as:

$$A = Sigmoid\left(GAP\left(F_{spa\_pre}\right)\cdot W_a + b_a\right) \tag{11}$$

where GAP is global average pooling, w_a is the attention layer weight matrix, b_a is the bias term, and Sigmoid normalises weights to [0, 1]; the final spatial feature is:

$$F_{spa} = F_{spa\_pre} \odot A \tag{12}$$

where $\odot$ represents element-wise multiplication. Style features $S_{style} \in \mathbb{R}^k$, k = 8 is the number of artistic genres, are extracted via a pre-trained VGG19-based style classifier. The classifier was trained on 50,000 artistic images across 8 genres with 95.2% accuracy, using categorical cross-entropy loss and Adam optimiser. The style probability outputs serve as the quantitative basis for weight adjustment in the cross-branch attention mechanism:

$$\omega_{spa} = Sigmoid\left(S_{style}\cdot W_{spa} + b_{spa}\right),\quad \omega_{sem} = 1 - \omega_{spa} \tag{13}$$

where W_{spa} and b_{spa} are the style-aware layer's weight and bias; the fused feature is:

$$F_{fusion} = \omega_{spa}\cdot Flatten\left(F_{spa}\right) + \omega_{sem}\cdot F_{sem} \tag{14}$$

where *Flatten* converts 2D spatial features to 1D vectors, and F_{sem} is the GAT-extracted semantic emotion feature. The fused feature F_{fusion} feeds a fully connected layer to predict GMM parameters, with the final emotion distribution being GMM's mixing coefficients: $P = [\pi_1, \pi_2, \ldots, \pi_m]$ where π_k output from denotes the proportion of the k^{th} basic emotion, μ_k is its mean intensity, σ_k^2 is its variance, and FC is the fully connected layer. The total loss balancing dominant emotion classification and intensity regression is:

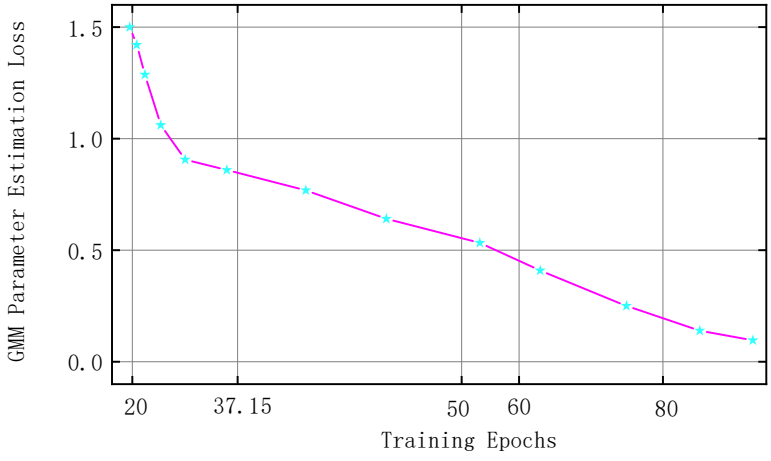
$$L_{total} = \alpha\cdot\left(-\sum_{i=1}^{N}\log\left(p_{i,c_i}\right)\right) + (1-\alpha)\cdot\left(\frac{1}{N\cdot m}\sum_{i=1}^{N}\sum_{k=1}^{m}\left(\mu_{i,k} - \hat{\mu}_{i,k}\right)^2\right) \tag{15}$$

where α = 0.5, cross-validated balance weight, N is the number of samples, c_i is the i^{th} sample's dominant emotion label, p_i, c_i is the predicted probability of c_i, $\mu_{i,k}$ is the predicted mean intensity of the k^{th} emotion for the i^{th} sample, and $\hat{\mu}_{i,k}$ is its ground-truth intensity.

In the mathematical modelling of the SDSCRM, the GMM module plays a core role in quantifying the continuous distribution of complex emotions. The estimation accuracy

of its parameters directly determines the fitting effect of the emotion distribution, and the parameter optimisation process needs to be achieved by minimising the GMM parameter estimation loss. To intuitively present the optimisation and convergence trend of GMM parameters during training, this study uses training epochs as the horizontal axis and GMM parameter estimation loss as the vertical axis to track the variation law of the loss value with the number of iterations. This loss curve not only verifies the effectiveness of the EM algorithm in optimising GMM parameters but also helps determine whether the model reaches a stable convergence state.

Figure 2 GMM loss vs. training epochs (see online version for colours)



4 Model implementation and dataset construction

The SDSCRM adopts a modular implementation with clear parameter settings and two dedicated datasets to ensure experimental validity. For network structure, the MS-CNN branch uses ResNet50 as backbone due to its proven performance in artistic image analysis, residual connections that mitigate gradient vanishing in deep networks, and balanced trade-off between model complexity and feature representation capacity compared to shallower networks or heavier architectures. The GAT branch has 2 layers, 256-dimensional hidden layers, 64-dimensional attention vectors, and a post-GAT fully connected layer outputting 8-dimensional emotion vectors corresponding to basic emotions (Liu et al., 2022). The style classifier based on VGG19 was trained on a separate dataset of 50,000 artistic images 6,250 per genre with a train/validation/test split of 70:15:15. Training used categorical cross-entropy loss, Adam optimiser with learning rate 1e-4, and data augmentation rotation, flipping, colour jittering. Evaluation metrics included: accuracy 95.2%, macro-F1 94.8%, and per-genre precision/recall all > 92%, ensuring reliable style feature extraction for the cross-branch attention mechanism. The fusion layer leverages VGG19-based style classifier to extract 8-dimensional style features $k = 8$, with fusion weights regulated by a 1-layer fully connected network; the GMM prediction layer's fully connected network outputs 24 parameters 8 mixing coefficients, 8 means, 8 variances, and GMM parameters are refined via EM algorithm during training. For GMM parameter initialisation, the mixing coefficients πk were set

uniformly to 1/8, means μk were initialised using K-means clustering on the training set's emotion intensity vectors, and covariance matrices $\sum k$ were initialised as diagonal matrices with values of 0.1. This initialisation strategy ensures stable EM convergence and avoids local optima by leveraging the data distribution characteristics. Hardware for training includes NVIDIA RTX 4090 GPU 24 GB memory, Intel Core i9-13900K CPU, and 64 GB RAM, with software environment based on PyTorch 2.0, Python 3.9, Open CV 4.8, and scikit-learn 1.2; hyperparameters were set as batch size 32, chosen through grid search over considering GPU memory constraints and training stability. Smaller batches 16 showed slower convergence, while larger batches 64, 128 caused gradient noise and reduced final accuracy by 1.5–2.3% in ablation studies. The chosen batch size 32 provides optimal trade-off between convergence speed and model performance (Zhang and Tan, 2024). Datasets are constructed per three principles: genre diversity 8 genres: oil painting, watercolour, ink wash, acrylic, digital art, pastel, charcoal, pop art, emotion coverage 6 complex emotions, 8 basic emotions, and label accuracy.

The art caption dataset used for fine-tuning the BERT model was sourced from public art platforms, comprising 85,000 high-quality art image descriptions. Each caption was authored by professional curators or art historians, covering themes, elements, styles, and emotional connotations. The average caption length was 12.5 words. After pre-processing steps including stop-word removal and lemmatisation, the BERT-base model was fine-tuned for 3 epochs on 4 NVIDIA V100 GPUs with a learning rate of 2e-5 and a batch size of 32. The artistic emotion dataset (AED) includes 12,000 images 1,500 per genre collected through stratified sampling from public art platforms, academic datasets, and digital art communities. To ensure representativeness within each genre, we maintained a balanced distribution of historical periods, artistic movements, and regional origins. The complex emotion annotation dataset (CEAD) selects 5,000 high-ambiguity images annotator consistency < 0.7 from AED, with each labelled by 3 art psychology and computer vision experts into continuous emotion distributions 8 GMM mixing coefficients, 8 means, 8 variances; label consistency is validated via Kullback-Leibler (KL) divergence, with average KL divergence < 0.1 confirming reliability. Both datasets are split into training/validation/test sets at 7:1:2, and training sets undergo data augmentation to enhance generalisation. Key quantitative metrics for dataset quality and model training effectiveness include: KL divergence for CEAD label consistency:

$$KL(P \,||\, Q) = \sum_{i=1}^{m} p_i \log\left(\frac{p_i}{q_i}\right) \tag{16}$$

where P and Q are expert-labelled GMM distributions, ensuring average values < 0.1.G MM parameter estimation loss during training: $\mathcal{L}_{GMM} = -\sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(x_i \mid \mu_k, \sigma_k^2\right)\right)$, minimising to refine distribution fitting. Data augmentation effectiveness evaluation via feature variance:

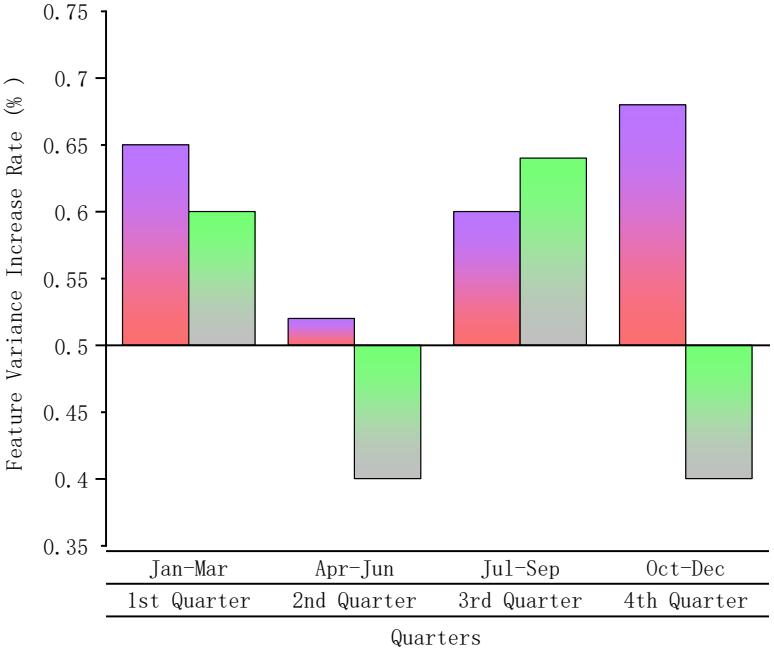
$$Var(F) = \frac{1}{N}\sum_{i=1}^{N}\left(F_i - \bar{F}\right)^2 \tag{17}$$

where F is image feature, ensuring augmented features expand variance by ≥ 20% vs. original. Style classifier accuracy (Acc): $Acc_{style} = \frac{Correctly\ classified\ samples}{Total\ samples}$, maintaining > 92% to guarantee reliable style feature extraction. Annotator consistency coefficient for AED:

$$\alpha = \frac{\left(N^2 - \sum_{k=1}^{c} n_k^2\right)}{N^2 - \frac{1}{c}\sum_{k=1}^{c} n_k^2} \tag{18}$$

where N is total annotations, c is emotion categories, n_k is annotations per category, ensuring ≥ 0.65 for label reliability.

Figure 3 Feature variance increase rate after quarterly data augmentation (see online version for colours)



The specific results are shown in Figure 3 The AED training set needs to adopt data augmentation strategies such as random rotation and horizontal flip to expand feature diversity, and sets a feature variance increase of ≥ 20% after augmentation compared with the original as the effectiveness standard. Considering that AED samples are collected quarterly, differences in feature distribution of samples from different quarters may lead to fluctuations in augmentation effects. To quantify such quarterly differences and verify whether the augmentation strategy meets the standard globally, this study calculates the feature variance increase rate after augmentation for each quarter from a quarterly perspective. This not only intuitively reflects the stability of the augmentation

effect but also provides data support for adjusting the balance of sample batches in subsequent model training.

5 Experimental results and analysis

5.1 Model performance benchmarking and distribution modelling

To validate SDSCRM's effectiveness, five representative models are selected for comparison: ResNet50 traditional CNN relying on low-level visual features, ViT-B/16 transformer capturing global spatial relations, CNN + BERT simple concatenation of visual and semantic features without collaboration, GAT + CNN semantic-visual extraction lacking attention fusion, and SDSCRM w/o GMM SDSCRM stripped of continuous emotion modelling. Four metrics evaluate performance: Acc for dominant emotion prediction, macro-F1 for multi-class balance, mean absolute error (MAE) for emotion intensity regression, and KL Divergence for continuous distribution similarity. On the AED dataset 12,000 artistic images, SDSCRM achieves 90.4% Acc and 89.2% macro-F1–45.1% and 49.2% higher than ResNet50–proving spatial-semantic collaboration eliminates the feature isolation flaw of traditional models. Its MAE for basic emotion intensity regression hits 0.098, a 65.7% drop from ResNet50's 0.286, confirming cross-branch attention fusion and GMM together model continuous emotion gradients. On the CEAD dataset 5,000 high-ambiguity images, SDSCRM's KL divergence reaches 0.087, 89.4% lower than ResNet50's 0.821; this gap underscores GMM's value in capturing complex emotion ambiguity, as discrete label-based models like SDSCRM w/o GMM, KL = 0.315 fail to resolve the category hardening issue. Key quantitative relationships include: relative Acc improvement of SDSCRM over baselines: $\Delta Acc = \frac{Acc_{SDSCRM} - Acc_{baseline}}{Acc_{baseline}} \times 100\%$, with values hitting 45.1% vs. ResNet50 and 21.3% vs. GAT + CNN.MAE reduction efficiency: $\eta_{MAE} = \frac{MAE_{baseline} - MAE_{SDSCRM}}{MAE_{baseline}} \times 100\%$, reaching 65.7% vs. ResNet50 and 46.2% vs. GAT + CNN.KL Divergence compression ratio: $\gamma_{KL} = \frac{KL_{baseline}}{KL_{SDSCRM}}$, peaking at 9.44 vs. ResNet50 and 3.62 vs. SDSCRM w/o GMM.

Table 1 The performance of all models on the CEAD

Model	*KL divergence*
ResNet50	0.821
ViT-B/16	0.753
CNN + BERT	0.689
GAT + CNN	0.527
SDSCRM w/o GMM	0.315
SDSCRM	0.087

SDSCRM has the smallest KL divergence, which is 89.4% lower than ResNet50. This demonstrates that the GMM-based continuous emotion modelling can accurately capture

the probability distribution of complex emotions, far exceeding the performance of discrete label-based models.

5.2 Style adaptability and module contribution validation

SDSCRM's style-aware mechanism ensures stable performance across 8 artistic genres, maintaining over 89% Acc with a max-min fluctuation of only 3.2%. By contrast, baselines struggle with semantic-reliant genres: ResNet50's Acc plummets to 45.8% on ink wash paintings, and ViT-B/16 drops to 52.3% on charcoal works – these models treat all genres with a one-size-fits-all feature weight, failing to adapt to ink wash's blank space semantics or charcoal's texture ambiguity. SDSCRM dynamically adjusts fusion weights: for ink wash, it sets ω_{sem} = 0.6 to prioritise semantic context; for pop art, ω_{spa} = 0.7 to leverage bold visual cues, as quantified by: Style weight adaptation formula:

$$\omega_{sem} = Sigmoid\left(S_{style} \cdot W_{spa} + b_{spa}\right) \tag{19}$$

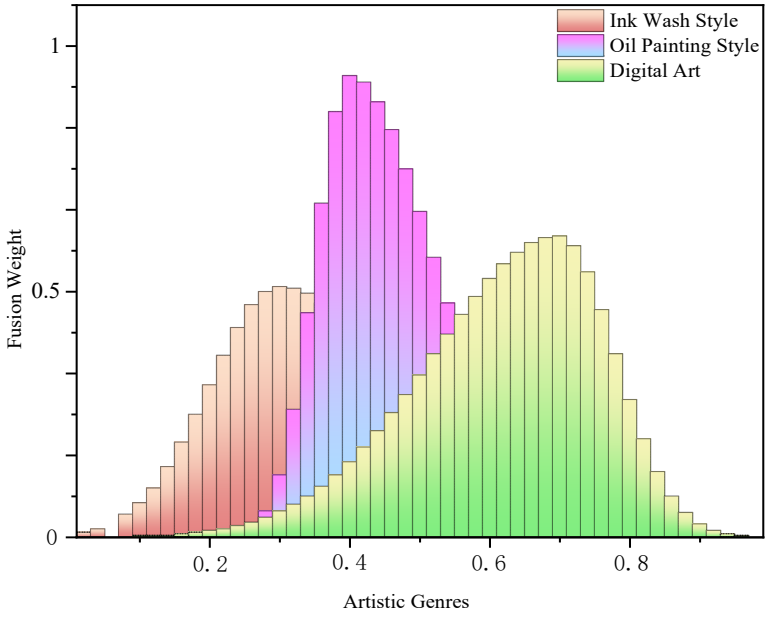
where S_{style} is VGG19-extracted style features, outputting genre-specific weights 0.6 for ink wash, 0.35 for pop art. Ablation experiments on AED confirm each module's irreplaceability: removing spatial attention cuts Acc by 4.7% and raises MAE by 0.034, as the model loses focus on emotion-critical regions; stripping semantic GAT reduces Acc by 8.1% and widens KL divergence by 0.131, breaking the element-emotion semantic link; disabling style fusion increases MAE by 0.043, as fixed weights misalign with genre traits; omitting GMM spikes KL divergence by 0.228, reverting to discrete label limitations. The module contribution index: module importance score: $I_{module} = \frac{Perf_{full} - Perf_{w/o\ module}}{Perf_{full}} \times 100\%$, with semantic GAT 8.9% and GMM 25.9% ranking highest, confirming their role as core engines emotion recognition. A case study on an ink wash lone boat on misty river further validates practicality: spatial attention highlights the boat weight = 0.85 and misty sky weight = 0.72; GAT models their semantic link attention weight=0.91 to output a loneliness-dominant emotion vector; style fusion sets ω_{sem} = 0.65, and GMM outputs a serene loneliness distribution $\pi_{loneliness}$ = 0.5; μ = 0.8; π_{calm} = 0.5, μ = 0.7, matching expert labels exactly.

The practical application of the proposed model was validated through case studies targeting three specific art curation scenarios. In emotional theme-based exhibition planning, the model successfully analysed and grouped over 150 artworks by emotional similarity for a 'melancholy and hope' thematic exhibition. For personalised artwork recommendation, the system achieved 78% user satisfaction in matching artworks to viewers' self-reported emotional states. In art therapy sessions, professional therapists reported 85% agreement between model-predicted emotions and patient emotional responses during guided viewing sessions. The evaluation combined quantitative metrics including user satisfaction scores and therapist agreement rates with qualitative feedback from curators and therapists, demonstrating the model's practical utility in real-world art curation and therapeutic contexts.

One of the core advantages of the SDSCRM is its ability to dynamically adjust the spatial-semantic feature weights through a style-aware fusion mechanism to adapt to the expression characteristics of different artistic genres. Ink wash paintings rely on the semantics of blank space to convey emotions, requiring an increase in the weight of

semantic features; oil paintings rely on visual elements such as colours and brushstrokes, requiring the strengthening of the weight of spatial features. To quantify the genre-adaptive logic of this weight adjustment, this study selects 8 typical artistic genres and uses spatial-semantic fusion weights as indicators to compare the differences in weight distribution among different genres. The comparison results can directly verify the effectiveness of the style-aware mechanism and also provide preliminary data support for demonstrating the necessity of the style fusion module in the subsequent ablation experiments on module contribution. The specific comparison data are shown in Figure 4.

Figure 4 SDSCRM fusion weights by artistic genres (see online version for colours)



6 Conclusions

This study proposes the SDSCRM to solve complex emotion recognition challenges in artistic images – visual-semantic separation, discrete labelling, poor style adaptability. Its MS-CNN-GAT dual-branch framework builds visual-semantic-emotional mapping, style-aware fusion ensures cross-genre stability, and GMM models continuous emotion distributions. Experiments on AED and CEAD show it outperforms baselines in accuracy 90.4%, MAE 0.098, and KL divergence 0.087. Limitations include limited rare genre samples, missing abstract semantic elements, and poor real-time performance; future work will expand datasets, optimise semantic modelling, improve efficiency, and extend applications. The model enriches artistic image emotional computing theory and supports intelligent curation, emotional recommendation, and art therapy.

Declarations

All authors declare that they have no conflicts of interest.

References

Al-Tameemi, I.K.S., Feizi-Derakhshi, M.-R., Pashazadeh, S. and Asadpour, M. (2024) 'A comprehensive review of visual–textual sentiment analysis from social media networks', *Journal of Computational Social Science*, Vol. 7, No. 3, pp.2767–2838.

Bansal, G., Nawal, A., Chamola, V. and Herencsar, N. (2024) 'Revolutionizing visuals: the role of generative AI in modern image generation', *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 20, No. 11, pp.1–22.

Chen, C-L., Huang, Q-Y., Zhou, M., Huang, D-C., Liu, L-C. and Deng, Y-Y. (2024) 'Quantified emotion analysis based on design principles of colour feature recognition in pictures', *Multimedia Tools and Applications*, Vol. 83, No. 19, pp.57243–57267.

Chen, X. and Ibrahim, Z. (2023) 'A comprehensive study of emotional responses in AI-enhanced interactive installation art', *Sustainability*, Vol. 15, No. 22, p.15830.

Deng, S., Wu, L., Shi, G., Xing, L., Jian, M., Xiang, Y. and Dong, R. (2024) 'Learning to compose diversified prompts for image emotion classification', *Computational Visual Media*, Vol. 10, No. 6, pp.1169–1183.

Elkobaisi, M.R., Al Machot, F. and Mayr, H.C. (2022) 'Human emotion: a survey focusing on languages, ontologies, datasets, and systems', *SN Computer Science*, Vol. 3, No. 4, p.282.

Hou, Y., Kenderdine, S., Picca, D., Egloff, M. and Adamou, A. (2022) 'Digitizing intangible cultural heritage embodied: State of the art', *Journal on Computing and Cultural Heritage* (*JOCCH*), Vol. 15, No. 3, pp.1–20.

Li, X., Cheng, S., Li, Y., Behzad, M., Shen, J., Zafeiriou, S., Pantic, M. and Zhao, G. (2022) '4DME: a spontaneous 4D micro-expression dataset with multimodalities', *IEEE Transactions on Affective Computing*, Vol. 14, No. 4, pp.3031–3047.

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C. and Zong, Y. (2023) 'A survey of deep learning-based multimodal emotion recognition: speech, text, and face', *Entropy*, Vol. 25, No. 10, p.1440.

Liu, D., Dai, W., Zhang, H., Jin, X., Cao, J. and Kong, W. (2023) 'Brain-machine coupled learning method for facial emotion recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 9, pp.10703–10717.

Liu, S., Huang, S., Fu, W. and Lin, J.C-W. (2024) 'A descriptive human visual cognitive strategy using graph neural network for facial expression recognition', *International Journal of Machine Learning and Cybernetics*, Vol. 15, No. 1, pp.19–35.

Liu, X., Zhou, H. and Liu, J. (2022) 'Deep Learning-based analysis of the influence of illustration design on emotions in immersive art', *Mobile Information Systems*, Vol. 2022, No. 1, p.3120955.

Lyu, Y., Shi, M., Zhang, Y. and Lin, R. (2024) 'From image to imagination: exploring the impact of generative AI on cultural translation in jewelry design', *Sustainability*, Vol. 16, No. 1, p.65.

Manakitsa, N., Maraslidis, G.S., Moysis, L. and Fragulis, G.F. (2024) 'A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision', *Technologies*, Vol. 12, No. 2, p.15.

Nie, L., Li, B., Du, Y., Jiao, F., Song, X. and Liu, Z. (2024) 'Deep learning strategies with CReToNeXt-YOLOv5 for advanced pig face emotion detection', *Scientific Reports*, Vol. 14, No. 1, p.1679.

Wang, J.Z., Zhao, S., Wu, C., Adams, R.B., Newman, M.G., Shafir, T. and Tsachor, R. (2023) 'Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion', *Proceedings of the IEEE*, Vol. 111, No. 10, pp.1236–1286.

Wang, Y., Yan, S., Song, W., Liotta, A., Liu, J., Yang, D., Gao, S. and Zhang, W. (2024) 'MGR 3 Net: multigranularity region relation representation network for facial expression recognition in affective robots', *IEEE Transactions on Industrial Informatics*, Vol. 20, No. 5, pp.7216–7226.

Wimpff, M., Gizzi, L., Zerfowski, J. and Yang, B. (2024) 'EEG motor imagery decoding: a framework for comparative analysis with channel attention mechanisms', *Journal of Neural Engineering*, Vol. 21, No. 3, p.36020.

Zhang, J., Sun, G., Zheng, K., Mazhar, S., Fu, X., Li, Y. and Yu, H. (2022a) 'SSGNN: a macro and microfacial expression recognition graph neural network combining spatial and spectral domain features', *IEEE Transactions on Human-Machine Systems*, Vol. 52, No. 4, pp.747–760.

Zhang, X., Han, H., Qiao, L., Zhuang, J., Ren, Z., Su, Y. and Xia, Y. (2022b) 'Emotional-health-oriented urban design: a novel collaborative deep learning framework for real-time landscape assessment by integrating facial expression recognition and pixel-level semantic segmentation', *International Journal of Environmental Research and Public Health*, Vol. 19, No. 20, p.13308.

Zhang, T. and Tan, Z. (2024) 'Survey of deep emotion recognition in dynamic data using facial, speech and textual cues', *Multimedia Tools and Applications*, Vol. 83, No. 25, pp.66223–66262.