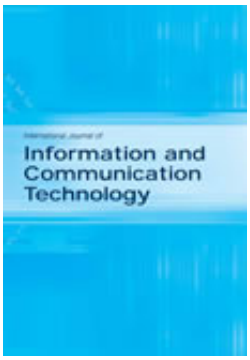
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

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

Emotion recognition in artistic images based on feature fusion and transfer learning

Laohui Liang

DOI: 10.1504/IJICT.2025.10074813

Article History:

Received:	07 September 2025
Last revised:	19 October 2025
Accepted:	20 October 2025
Published online:	12 December 2025

Emotion recognition in artistic images based on feature fusion and transfer learning

Laohui Liang

Faculty of Marxism,
Guangdong Mechanical and Electrical Polytechnic,
Guangzhou, 510000, China
Email: 2004010013@gdmec.edu.cn

Abstract: Currently, artistic images are scarce with limited sample sizes, and most sentiment analysis relies on low-level image features with low accuracy. To address this, this paper first extracts two-dimensional features from images in different colour spaces. It then employs multi-scale convolutional kernels to extract deep semantic information from images, fusing feature information from different dimensions to effectively preserve semantic features across scales. Finally, the transfer component analysis algorithm is employed to reduce dimensionality of features in source and target domains within original space. An improved joint subspace learning method is used to learn a feature transformation subspace, reducing the conditional probability distribution distance between source and target domains while balancing recognition accuracy across categories. Model optimisation is achieved through adversarial training. Experimental results demonstrate that the proposed model improves recognition accuracy by at least 3.82%, effectively enhancing the accuracy of emotional recognition in artistic images.

Keywords: artistic image emotion recognition; feature fusion; transfer learning; adversarial training; feature extraction.

Reference to this paper should be made as follows: Liang, L. (2025) 'Emotion recognition in artistic images based on feature fusion and transfer learning', *Int. J. Information and Communication Technology*, Vol. 26, No. 44, pp.1–17.

Biographical notes: Laohui Liang obtained her Master's in Information Engineering from Nanjing University in 2016. She is currently a Lecturer at Guangdong Polytechnic of Mechanical and Electrical Engineering. Her research interests include aesthetic education, music and artistic image emotion recognition.

1 Introduction

Art images, as important carriers of human emotional expression and cultural inheritance, touch people's hearts through their rich visual elements and profound connotations that transcend time and space boundaries. From ancient cave paintings to modern digital artworks, each image contains the creator's unique emotions, thoughts, and aesthetic concepts, while also being able to evoke diverse emotional resonance from viewers (Zabora et al., 2023). Art image emotion recognition aims to automatically analyze the emotional information conveyed in images through technical means. This not only helps

deepen understanding of the intrinsic value of art works but also brings development opportunities for multiple fields such as artistic creation and cultural dissemination (Tashu et al., 2021). Art images have high levels of abstraction, subjectivity, and diversity, with viewers from different cultural backgrounds, personal experiences, and aesthetic concepts possibly generating completely different emotional interpretations of the same work (Yang et al., 2023). At the same time, emotional expression in art images often combines multiple visual elements such as colour and texture. These elements are interrelated and influence each other, jointly forming complex and subtle emotional semantics (Gonzalez-Martin et al., 2024). Therefore, how to extract effective features from art images and accurately recognise their emotion categories has become a key issue urgently needing resolution in current research within this field.

Considering the characteristic that art images are entirely composed of low-level visual features, early works mostly focused on manually designing and extracting low-level image features to predict the emotions of artistic images. Gatys et al. (2017) listed artistic image features such as texture and shape based on image transformations, demonstrating that these features can be used to predict emotions. Yang et al. (2023) combined the Group Lasso model (Yang and Zou, 2015) with support vector machine (SVM) for art image emotion recognition, achieving an accuracy rate of only 75.24%. Liu et al. (2024) extracted four-dimensional texture representations from art images and analyzed their impact on emotional recognition using a saliency model. Ruan et al. (2021) designed and extracted artistic features from art images for emotion classification based on statistical analysis and artistic theory. Wang et al. (2023) suggested a multi-label transformation categorisation framework for joint learning in the domain of art picture emotion identification, integrating visual and textual information within the framework to address emotional classification problems for art images. Wang (2022) introduced nonlinear combination matrices (Xu et al., 2016) as transition classifiers for emotional feature representations, modelling relationships between different latent variables and conducting artistic image emotion recognition by simulating the process of inferring emotions.

As art image emotion recognition requires crossing deeper emotional semantic gaps, the above methods based on manually extracted features struggle to establish connections between visual elements contained in artistic images and human high-level emotional characteristics. Deep learning-based approaches for art image emotion identification automatically mine deep features from art images using neural network models, significantly enhancing identification efficiency. Gonzalez-Martin et al. (2024) put forward an art picture emotion recognition approach in light of a two-layer transfer convolutional neural network (CNN) model. Leveraging the hierarchical nature of CNN models, they used the ImageNet2012 dataset to extract universal low-level visual features from images and transferred these lower-level feature weights into networks with identical structures to extract deeper semantic features from images, achieving high recognition accuracy. Li et al. (2021) constructed a joint embedding network for visual and semantic features using an autoencoder, introduced attention mechanisms for feature fusion, and built an emotion classifier to achieve emotion classification. Zhang et al. (2021) obtained many candidate regions through object detection methods and combined emotional scores with object scores to select the top K ranked emotional regions, utilising a CNN model to capture characteristics from entire art images and emotional region features. Zhang et al. (2022) proposed extracting and combining semantic features at high

levels, aesthetic features at mid-levels, and visual features at low levels from CNN models for emotion recognition.

Due to the small sample size issue in datasets, few works directly train neural network models on art image emotion recognition datasets; researchers commonly adopt training strategies such as transfer learning. Integrating transfer learning algorithms represents one of the most effective and practical technical approaches currently available for enhancing the performance of artistic image emotion recognition models. This approach directly addresses the most critical data bottlenecks and model generalisation challenges in this field, providing a solid foundation for both research and practical applications. However, deeper challenges such as the subjectivity of emotions, cultural differences, and model interpretability still require collaborative solutions integrating multidisciplinary knowledge and more advanced algorithms. Lu and Wan (2022) proposed an art image emotion recognition method based on AlexNet using two stages of transfer learning, performing transfer learning first on a painting art classification dataset and then on an art image emotion dataset, ultimately obtaining a deep learning model capable of classifying emotions in art images. Jiang et al. (2024) adopted a transfer learning approach by transferring pre-trained model parameters to the target model and fine-tuning it, extracting semantic features from art images, performing linear fusion between them, and using this method for emotion recognition in art images. The results demonstrated that the method achieved high accuracy in predicting emotions of art images. Cheng et al. (2024) applied a progressive training strategy combined with domain transfer to fine-tune a P-CNN model on weakly labelled datasets, validating the effectiveness of CNNs and transfer learning techniques for emotion recognition in art images.

A comprehensive analysis of existing research reveals that current training approaches require substantial sample sizes and achieve excellent classification performance. However, abstract painting datasets feature limited samples. Direct application of deep learning methods in such scenarios can severely impact the network architecture, leading to overfitting and poor test results. To this end, this article proposes an emotion identification approach for artistic images based on feature fusion and transfer learning. First, contrast limited adaptive histogram equalisation (CLAHE) is employed to highlight the colour characteristics of artistic images. Two-dimensional characteristics are captured from the images in different colour spaces. Multi-scale convolutional kernels are used to capture characteristics from the deep semantic information of the model, compensating for the limitations of single-scale feature extraction. Second, feature information from different dimensions is concatenated and channel-wise mixed to further enhance information flow and expressive capability between feature map channels. Multi-scale information within the feature maps is fused to effectively preserve feature details across different scales. Finally, transfer component analysis is leveraged to perform feature dimensionality reduction for both the source and target domains in the original space, thereby reducing the distance among their marginal probability distributions. Subsequently, the reduced-dimensional source and target domains are processed. An enhanced joint subspace learning approach is utilised to derive a feature transformation subspace, aiming to minimise the discrepancy in conditional probability distributions between the source and target domains while ensuring balanced recognition performance across different classes. Through adversarial training, the model achieves performance optimisation for the task of emotion recognition in artistic images.

Experimental outcome implies that the suggested model achieves at least 3.82% and 5.55% improvements in accuracy and F1-score, respectively, compared to the baseline model, validating its effectiveness and superiority.

2 Relevant theory

2.1 Convolutional neural network

CNN is a deep learning architecture specialised in handling data with a grid-like topology., featuring strong local feature extraction capabilities as well as characteristics such as parameter sharing and hierarchical feature learning. In speech emotion recognition study, CNN is commonly adopted to capture depth emotion characteristics from spectrograms. A basic CNN typically includes convolutional levels, activation function levels, pooling levels, and fully linked levels (Kuo, 2016).

1 *Convolutional level:* By systematically sliding convolutional kernels across the input, it performs operations on all regions, enabling the effective capture of local features. Moreover, parameter sharing leads to a reduction in the total number of model parameters.

2 *Pooling level:* Placed after the convolutional level, it can decrease the amount of model parameters and lower the size of feature maps while reducing the correlation between features, thereby enhancing the robustness of the model (Vonder et al., 2023).

3 *Activation function:* A convolutional level processes feature maps only through complex linear transformations, whereas the application of an activation function introduces nonlinear transformation to CNNs, accelerating model convergence. Commonly used activation functions in CNNs include Sigmoid, Tanh, and ReLU.

4 *Fully linked level:* It is typically used to integrate and classify features extracted by the convolutional level and pooling level. Before being input into the fully connected layer, features usually pass through a flattening layer that processes them into one-dimensional vectors so that all information can be considered by the fully linked level. The final classification result of the network is then obtained through training and learning. Finally, a softmax function calculates the final classification probabilities, as expressed below.

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{c=1}^{C} e^{z_c}} \tag{1}$$

The primary advantage of CNNs over traditional neural network models (such as fully connected neural networks) lies in their highly efficient extraction of local features and parameter sharing mechanism. This endows them with significant strengths when processing data possessing spatial hierarchical structures.

2.2 Transfer learning algorithm

Transfer learning begins with data distribution adaptation and gradually develops into feature selection, subspace learning, manifold learning, and finally deep transfer learning. Data distribution adaptation exploits the discrepancies between the probability distributions of the source and target domains and mitigates them through a transformation process, it reduces the distributional divergence between the source and target domains. Subspace learning is divided into statistical feature alignment and manifold space alignment. Subspace learning is based on the assumption that source and target domain data share a similar distribution within a learned latent subspace. Classic algorithms for statistical feature alignment include subspace alignment (SA), subspace distribution alignment (SDA), second-order subspace alignment (CORAL) (Zhang et al., 2018). Manifold learning assumes current data are sampled from the same high-dimensional space, with classical algorithms including unsupervised domain adaptation in manifold space hypothesis (SGF), and unsupervised domain adaptation based on geodesic flow kernel (GFK) (Hosna et al., 2022).

Transfer component analysis (TCA) is a classic algorithm of transfer learning (Zheng et al., 2022). TCA assumes that the data distribution after feature mapping ϕ of original domain (X_t) and objective domain (X_s) are approximately equal $P(\phi(X_s)) \approx P(\phi(X_t))$. Assuming ϕ is known, to minimise the distance between the original field and target field, TCA uses a maximum mean discrepancy (MMD) matrix L shown in equation (2), calculates the difference among the means after mapping of the source domain and target domain $DISTANCE(X_s, X_t)$, as indicated in equation (3).

$$l_{ij} = \begin{cases} \frac{1}{n_1^2}, x_i, x_j \in X_s \\ \frac{1}{n_2^2}, x_i, x_j \in X_t \\ -\frac{1}{n_1 n_2}, others \end{cases} \tag{2}$$

$$DISTANCE\left(X_s, X_t\right) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi\left(x_i\right) - \frac{1}{n_2} \sum_{j=1}^{n_2} \phi\left(x_j\right) \right\|_H \tag{3}$$

where n_1 is the amount of instances from the original domain, and n_2 is the amount of instances from the target domain. By using properties of matrices, a kernel matrix K is introduced, as shown in equation (4), combining L transforms equation (3) into equation (5).

$$K = \begin{bmatrix} K_{s,s} & K_{s,t} \\ K_{t,s} & K_{t,t} \end{bmatrix} \tag{4}$$

$$tr(KL) - \lambda tr(K) \tag{5}$$

TCA is in light of MMD (Li et al., 2024), and utilising the kernel method to map data from both the source and target domains into a high-dimensional reproducing kernel Hilbert space. In this space, this approach achieves a balance between reducing domain

shift and the preservation of the data's inherent feature characteristics. Finally, a low-dimensional shared subspace is learnt in which the data distributions of the source and target domains are closely aligned.

3 Feature extraction from artistic images and multi-scale feature fusion

3.1 Low-level feature extraction of artistic images

In order to address the issue that current research does not simultaneously consider both low-level and deep features of artistic images, this paper utilises CLAHE (Mondal et al., 2024) to highlight colour features of artistic images, separately extracts H-S two-dimensional features under the hue, saturation, value (HSV) colour space and chrominance red and chrominance blue (CrCb) two-dimensional features under the YCrCb colour space, sets different weights for the two colour features, adopts multi-scale convolution kernels to extract deep information from the image, concatenates and shuffles feature information of different dimensions to further enhance channel communication and expression capability among feature maps. Finally, it fuses the multi-scale information in the feature map so as to effectively retain characteristic information at different scales.

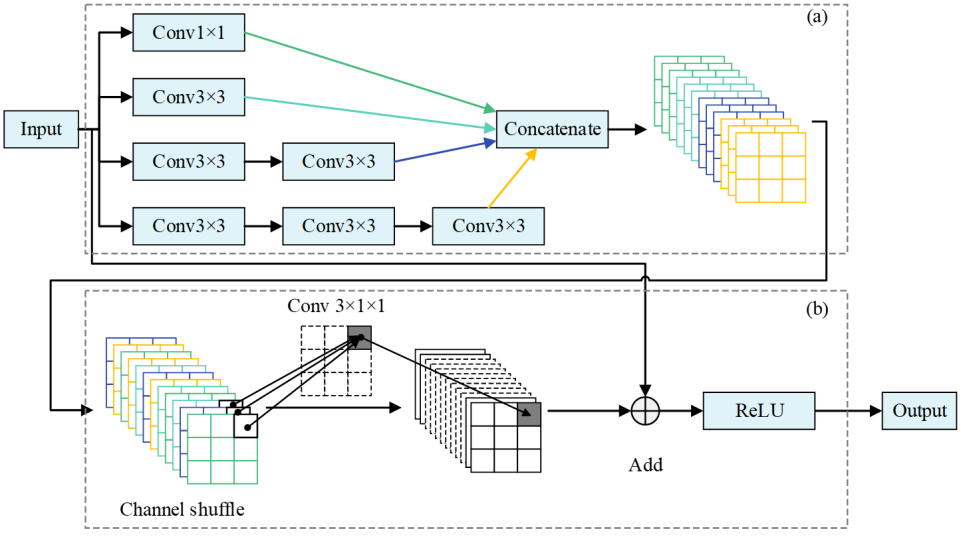
Colour features are the most basic low-level visual characteristics of images, containing the most critical and sensitive visual information. During research on low-level colour features of images, most scholars usually adopt adaptive histogram equalisation (AHE) (Song et al., 2025). The advantage of AHE lies in its capability to calculate histograms for each significant region of an image, re-adjust brightness values for uniform distribution, thus enhancing local contrast across different regions. However, the AHE method can easily introduce noise into images. Compared with AHE, CLAHE obtains corresponding transformation functions by limiting contrast in the neighbourhood of every pixel, effectively suppressing enhancement of image noise. Image colour feature extraction will be conducted based on the extracted image colour features after CLAHE processing. Considering that the HSV space of images is more consistent with human vision perception of colours, the picture is first converted from RGB format to HSV format. After CLAHE processing, hue-saturation two-dimensional histogram (H-S) method is used to extract significant H-S colour features in the image. In this two-dimensional histogram, the hue value (H) ranges between 0-180 and the saturation (S) ranges between 0–256.

In addition to common colour spaces such as HSV and RGB, there is another colour space that conforms to human vision characteristics called YCrCb. YCrCb is mainly used for optimising the transmission of colour video signals, where Y represents brightness; Cr represents hue, which indicates the red component value in colours; Cb represents saturation, which indicates the blue component value in colours. Based on the YCrCb colour space, the chrominance of images is primarily determined by the Cr and Cb channels. Therefore, only the two-dimensional spatial features of CrCb are considered when extracting colour features. The specific method is the same as mentioned above: convert the image from RGB format to YCrCb format, perform CLAHE processing and use the hue-saturation two-dimensional histogram method to extract significant CrCb colour features in the image. In this two-dimensional spatial feature, the hue value (Cr) ranges between 16–224 and saturation value (Cb) ranges between 16–224.

3.2 Multi-scale feature fusion

After obtaining the low-level features of artistic images, this article designs a multi-scale characteristic integration module that can simultaneously consider both the low-level and high-level characteristics of artistic images, as shown in Figure 1. First, multi-scale convolution kernels are adopted to extract semantic information from the network, making up for the shortcomings caused by single features. Second, feature information with different dimensions is concatenated and channel shuffled to further enhance the flow of information and expression ability between channels in the feature maps. Finally, multi-scale information in the feature maps is fused, thus effectively preserving feature information at different scales.

Figure 1 Multi-scale feature fusion module (see online version for colours)



This paper combines multi-dimensional convolution kernels with depthwise separable convolutions to build an Xception structure that improves the feature extraction capability of deep neural networks while reducing network computation costs. Depthwise separable convolutions split traditional convolutions into two operations (Jang et al., 2023). First, 2D convolutions are applied on every channel of the input characteristic pictures, a process referred to as depth convolution. Then, a 1 × 1 convolution kernel is used to combine channel features, an operation known as point convolution. By these two steps, the amount of network parameters can be reduced, making the network more lightweight while maintaining its feature detection efficiency. Suppose the size of the input characteristic picture is $d_f \times d_f$, the amount of input channels is m, the convolution kernel size is $d_k \times d_k$, the convolution stride is 1, and the number of output channels is n. Subsequently, the parameter count params and computational cost floating point operations (FLOPs) for depthwise separable convolutions are given by equations (6) and (7), respectively.

$$params = d_k \times d_k \times m + m \times n \tag{6}$$

$$FLOPs = d_k \times d_k \times m \times d_f \times d_f + m \times n \times d_f \times d_f \tag{7}$$

From the above formulas, it can be seen that the amount of parameters and computational complexity of depthwise separable convolutions are closely related to the convolution kernel size $d_k \times d_k$ and the number of output channels n. In equation (6), the number of point convolutions is $m \times n$, which still accounts for most of the total algorithm parameters. To reduce computational complexity, this paper replaces the point convolutions in depthwise separable convolutions with channel-wise convolutions. By sparsifying the connections between inputs and outputs, the 1 × 1 convolution kernel slides along the channel dimension to further compress and accelerate the model. The parameter of the channel-wise convolution is $d_c \times d_k \times d_k$, usually set to $d_c \leq m$, where d_c is the number of input channels sampled in a single operation. The improved depthwise separable convolutions have parameter count params and computational cost FLOPs as follows.

$$params = d_k \times d_k \times m + d_c \tag{8}$$

$$FLOPs = d_f \times d_f \times m \times d_k \times d_k + d_c \times n \times d_f \times d_f \tag{9}$$

This model constructs a multi-scale feature fusion module by employing convolutional kernels of varying sizes. Its core approach involves capturing features at different scales within an image using kernels of distinct dimensions. These multi-scale features are then integrated through feature concatenation, weighted fusion, or attention mechanisms to form a more discriminative feature representation. This paper adopts convolution kernels of 1 × 1, 3 × 3, 5 × 5, and 7 × 7 to construct a multi-scale characteristic extraction module as shown in Figure 1. In practical applications, larger convolution kernels for example 5 × 5 and 7 × 7 can be replaced by cascading multiple 3 × 3 convolution kernels. This ensures the same receptive field while significantly reducing the number of parameters. In addition, this paper introduces channel-shuffle (Li et al., 2021), which randomly shuffles the concatenated feature maps along the channel dimension to enhance inter-channel feature interaction of the feature maps and further improve the model's generalisation ability.

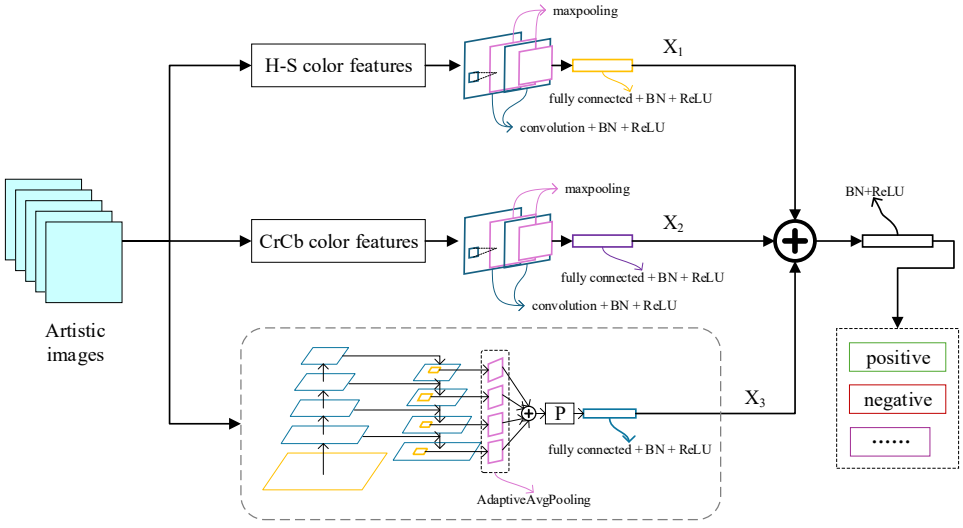
4 Emotion recognition in artistic images based on feature fusion and transfer learning

4.1 Feature transformation subspace learning

To address the issue of low emotional recognition accuracy and large deviations between categories in current research, this paper proposes an artistic image emotion identification model in light of feature fusion and transfer learning. The suggested artistic picture emotion recognition model is shown in Figure 2. First, TCA is adopted to transform the fused characteristics, and for the original domain and target domain after TCA transformation, an improved joint subspace learning (ICSL) approach is suggested to learn a characteristic transformation subspace that simultaneously considers source domain labels and pseudo-labels generated iteratively from both the original and target domains. To prevent this characteristic transformation subspace from overfitting, $L_{2,1}$

norms and distance metric matrices are added to preserve certain distance constraints of the original space.

Figure 2 The suggested artistic image emotion recognition model (see online version for colours)



The marginal distribution distances between transformed source domain sample features X_s and target domain sample features X_t are reduced. To minimise the conditional distribution divergence between the source and target domains, conventional transfer learning approaches address solely the marginal distribution shift between domains, overlooking the divergence in their conditional distributions. Thus, this article suggests an ICSL appraoch to find a feature transformation subspace M that minimises the conditional distribution distances among the original and target domains. Traditional transfer learning methods only learn the features themselves of the original domain and target domain, thus ignoring the conditional distribution distance between the original domain and target domain. ICSL considers iteratively generated pseudo-labels on top of traditional transfer learning methods, adding a feature transformation subspace learning for same-class pseudo-labels as shown in equation (10).

$$\min_M \left(\left\| X_s^T M - f_s \right\|_F^2 + \left\| X_t^T M - f_t \right\|_F^2 + \left\| X_c^T M - f_c \right\|_F^2 \right) \tag{10}$$

where f_s is the feature representation of source domain labels under M, initialising f_s as category label features. f_t represents the target domain's feature expression under M, initialise f_t randomly and randomly assign category labels to the target domain in accordance with the style of source domain category label features. X_c is a collection of original domain and target domain instances recognised as belonging to the same class, while f_c represents data instances from both the source domain and target domain identified as class c under M. Regarding determination of class c, traditional SVM classification algorithms can be used.

4.2 Distance-based regularisation

To prevent overfitting caused by minimisation of conditional distribution distance between the original domain and target domain, constraints are imposed on M. The $L_{2,1}$ norm of M is added to equation (10) as shown in equation (11), where V is a diagonal matrix with each element being ε.

$$\|M\|_{2,1} = \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{n} M_{ij}^2} = 2tr(M^T V M) \tag{11}$$

In traditional $L_{2,1}$ norms, the definition of v_{ii} is given in equation (12), and ε is an additional introduced parameter. Furthermore, the value of $\|m^i\|^2$ approaches 0, while $L_{2,1}$ norm is unpredictable in the original space. Therefore, when ε approaches 0.

$$v_{ii} = \frac{1}{2\sqrt{\|m^i\|^2 + \varepsilon}} \tag{12}$$

According to the assumption of manifold learning, x_i and x_j that are close in the original space remain close in the subspace after feature transformation. Thus, data samples from the original domain and target domain, denoted as x_1, …, x_n, generate a distance metric matrix G by computing the K nearest neighbours for each instance and marking those nearest neighbours' positions as 1 while setting others to 0. Finally, the result of the distance metric regularisation between the original domain and target domain is shown in equation (13).

$$\begin{aligned} distance1(M) &= \frac{1}{2}\sum_{i,j=1}^{n} \left\| x_i' - x_j' \right\|_F^2 g_{ij} \\ &= \sum_{i=1}^{n} \left(x_i^T M\right)^T \left(x_i^T M\right) d_{ii} - \sum_{i,j=1}^{n} \left(x_i^T M\right)^T \left(x_j^T M\right) g_{ij} \\ &= tr(M^T X J X^T M) \end{aligned} \tag{13}$$

where $J = D - G$ is a Laplacian matrix, D is a diagonal matrix where each element is defined as per equation (14), d_{ii} denotes summing over the columns of G.

$$d_{ii} = \sum_{j} g_{ij} \tag{14}$$

Considering the original space distance metric regularisation after TCA between the source domain and target domain to decrease their marginal distribution distances and conditional distribution distances. On this basis, add a distance metric regularisation for the class label space of the source domain B, as shown in Equation (15), which ensures data samples from the same class within the source domain remain close in this space, where $J' = D' - G'$.

$$\begin{aligned} distance2(M) &= tr(M^T B D' B^T M) - tr(M^T B G' B^T M) \\ &= tr(M^T B J' B^T M) \end{aligned} \tag{15}$$

Finally, the objective function of the distance metric matrix regularisation is shown in equation (16).

$$distance(M) = tr\left(M^T XJX^T M\right) + tr\left(M^T BJ^T B^T M\right) \tag{16}$$

4.3 Model iterative optimisation

To improve the model's training efficiency, this paper proposes an adversarial training-based domain adaptation learning (ATDA) model. The objective is to address performance degradation caused by distribution differences between domains when applying a model from the original domain to another target domain. ATDA is an adversarial training-based domain adaptation learning model consisting mainly of three components: a characteristic fuser, a label classifier, and a domain discriminator. Among these, the domain discriminator acts as the core unit in DANN for achieving domain adaptation, with its role being to determine the domain provenance of features, framing it as a binary classification task. However, at the same time, the goal of the feature fuser is to generate features that are challenging to differentiate among the source and target domains. Therefore, the model's training process can be viewed as adversarial training between the characteristic fuser and the domain discriminator, leading to a reduction in the divergence between the feature distributions of the source and target domains, ultimately achieving effective emotion transfer between them and improving the model's performance on the target domain.

The gradient reversal layer is a key component in ATDA, positioned between the feature fuser and domain discriminator, achieving adversarial training of features across domains. The training objective of the domain adaptation network is to minimise the artistic image emotion label classification loss L_c while maximising the domain categorisation loss L_d. The label categorisation loss is generated based on source domain label information and used to measure the accuracy of label classification, with its loss function expressed as follows, where θ_f and θ_c represent trainable parameters in the feature extractor and label classifier respectively, $G_f(\cdot)$ is the feature extraction function for generating output after data samples pass through the characteristic extractor, $G_y(\cdot)$ is the label forecasting function for generating artistic image emotion classification labels of instances, and n denotes the amount of labelled training samples in the source domain.

$$L_c\left(y;\theta_f,\theta_c\right) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(G_y\left[G_f\left(x_i\right)\right]_{y_i}\right) \tag{17}$$

The loss function of artistic image emotion classification used in domain adaptation training can be expressed as follows, where θ_d represents trainable parameters in the domain discriminator, $G_d(\cdot)$ is the domain discrimination function for generating domain classification results, and m denotes the amount of unlabeled training sinstances in the target domain.

$$\begin{aligned} L_d\left(d;\theta_f,\theta_d\right) &= -\frac{1}{n}\sum_{i=1}^{n}\log\left(G_d\left[G_f\left(x_i\right)\right]_{d_i}\right) \\ &= -\frac{1}{m}\sum_{j=1}^{m}\log\left(G_d\left[G_f\left(x_j\right)\right]_{d_j}\right) \end{aligned} \tag{18}$$

Therefore, the final objective function is shown in equation (19), where λ is used to control the weight between different parts of the loss. By jointly optimising these two parts of the loss function, ATDA can learn a feature representation shared between the original domain and target domain, achieving generalisation of the artistic image emotion recognition model on the target domain. It can be observed that this is a maximin problem, which attempts to find a saddle point using parameters θ_f, θ_c and θ_d, as shown in equations (20) and (21). At the saddle point, the artistic image emotion classification loss for labels reaches its minimum, while the artistic image emotion classification loss reaches its maximum.

$$L\left(y, d; \theta_f, \theta_c, \theta_d\right) = L_c\left(y; \theta_f, \theta_c\right) - \lambda L_d\left(d; \theta_f, \theta_d\right) \tag{19}$$

$$\left(\hat{\theta}_f, \hat{\theta}_c\right) = \arg\min_{\theta_f, \theta_c} L\left(y, d; \theta_f, \theta_c, \hat{\theta}_d\right) \tag{20}$$

$$\hat{\theta}_d = \arg\max_{\theta_d} L\left(y, d; \hat{\theta}_f, \hat{\theta}_c, \theta_d\right) \tag{21}$$

5 Experimental results and performance analysis

The MART dataset (Liu and Agaian, 2024) is selected as the experimental dataset. This dataset contains 1,293 artistic image works from the Trento and Rovereto Modern Art Museums. Most of these painters are Italian, with some also coming from Europe and America. The dataset includes six emotion labels: happiness, sadness, anger, fear, surprise, and disgust. To demonstrate the effectiveness of this method for artistic image emotion recognition tasks, during training, a five-fold cross-validation approach is used with a default of 20 epochs, learning rate set to 0.005, batch size set to 64, using the cross-entropy loss function, and selecting SGD optimiser with momentum set at 0.9.The experimental processor is Intel i5-8279U, with a clock speed of 2.40 GHz. The experiment runs based on the deep learning framework PyTorch under Python 3.7, using Linux operating system for development in PyCharm and GPU training with NVIDIA Tesla V100-SXM2-16 GB GPU having a total VRAM of 16160 MiB.

In this article's experiments, SCEP (Li et al., 2021), CNNAR (Zhang et al., 2021), IAlexNet (Lu and Wan, 2022), P-CNN (Cheng et al., 2024), and the proposed model FFTL are selected for comparison. The recognition accuracy rates of different models for six categories of artistic image emotions are shown in Table 1. Five models achieve lower recognition accuracy for sadness and anger compared to the other four emotions. The average recognition accuracies across six artistic image emotions for SCEP, CNNAR, IAlexNet, P-CNN, and FFTL are 77.94%, 82.9%, 86.8%, 90.4%, and 94.35%, respectively, demonstrating relatively excellent performance in artistic image emotion recognition.

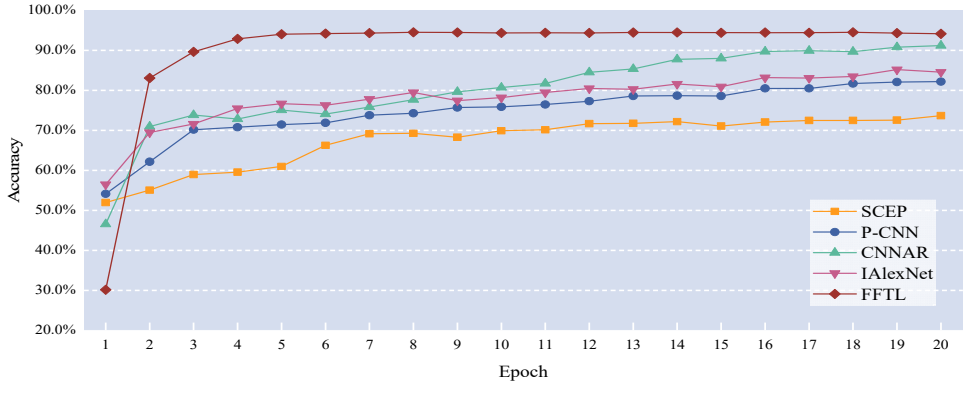
In the five-fold cross-validation of each model, one fold with the best prediction results on the training set is selected and the accuracy and loss value changes throughout the entire training process are plotted for that fold. Figure 3(a) shows a polyline comparison chart of the accuracy of different models as iterations change. Figure 3(b) demonstrates the loss performance of different models as iterations vary. It can be seen that the FFTL model converges very quickly compared to other models. When epoch = 4,

the loss value of the FFTL model drops from 4 to below 0.5; when epoch = 6, the loss value has already approached zero and its accuracy reaches 95%. This indicates that low-level features in art images also play an important role in emotion recognition. From an intuitive perspective, compared with texture features, colour features contain richer emotional semantic information. More importantly, fusing image low-level features with deep features can effectively identify the image's emotional semantic information and enhance the training performance of models.

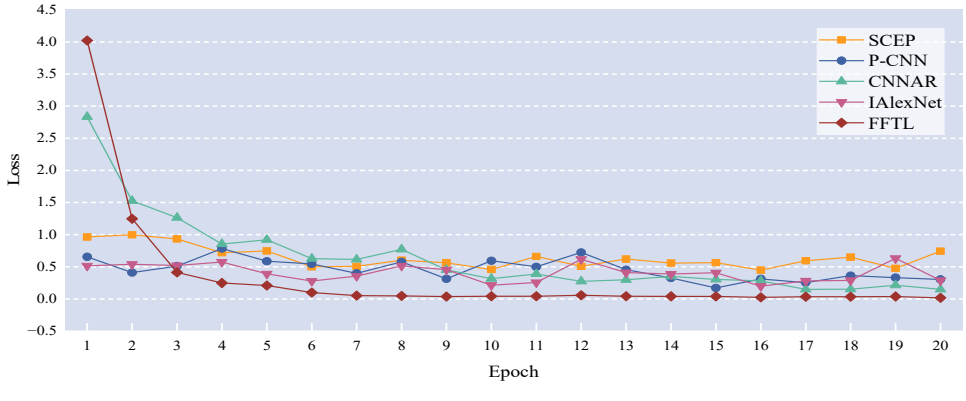
Table 1 Accuracy of different models in recognising emotions in six categories of artistic images (%)

Emotional categories	*Joy*	*Sadness*	*Anger*	*Fear*	*Surprise*	*Disgust*
SCEP	78.51	80.64	76.49	80.96	74.22	76.84
CNNAR	84.25	81.62	80.59	83.91	83.16	83.87
IAlexNet	88.52	86.37	84.52	86.94	87.51	86.92
P-CNN	91.08	87.51	88.45	92.41	90.02	92.95
FFTL	95.29	91.58	93.32	95.33	93.75	96.82

Figure 3 Accuracy and loss performance of each model, (a) accuracy rate change line chart (b) loss change line chart (see online version for colours)



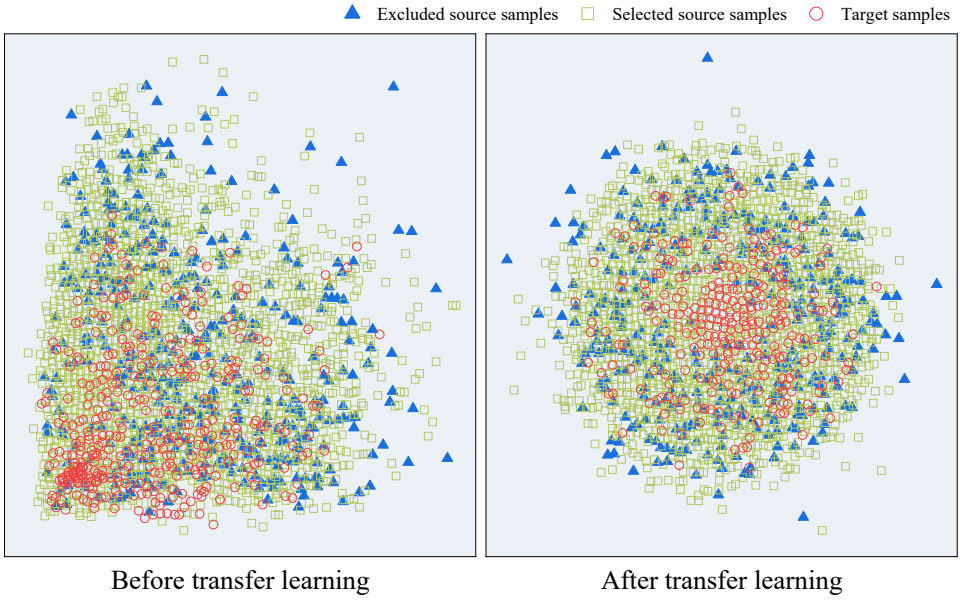
(a)



(b)

At the same time, this paper also analyzes the impact of transfer learning methods on the distribution of art image data. As shown in Figure 4, green triangles represent source domain samples selected based on deep style features, blue triangles are excluded source domain samples, and red circles are target domain training samples. Both visualisation results with and without TCA transformation indicate that selecting a transfer learning method can prioritise choosing source domain samples located near the centre of the target domain distribution for network training while eliminating some source domain samples far from the centre, thereby alleviating negative transfer issues. To improve the prediction accuracy of difficult-to-train samples, a small number of source domain samples away from the distribution centre are also used for training networks to prevent model overfitting.

Figure 4 The impact of transfer learning methods on the distribution of artistic image data (see online version for colours)



This paper further compares the recognition performance of different models using emotional identification metrics for example accuracy, precision, and F1, as shown in Table 2. The FFTL model achieves an accuracy of 95.48% and an F1-score of 96.08%, which represent increases of at least 3.82% and 5.55%, respectively, compared to the baseline models. SCEP realises visual feature and semantic feature fusion through an autoencoder; however, emotional recognition in art images requires a large amount of labelled data to guide model learning for emotion-related features. When labels are lacking, the model may only learn low-level visual features while ignoring high-level semantic information closely related to emotions. CNNAR realises art image emotion recognition through CNN, but CNN only learns low-level characteristics via shallow networks and high-level characteristics via deep networks. However, art emotions may span multiple levels, leading to poor performance in art image emotion recognition. IAlexNet implements art image transfer using AlexNet, but the intermediate features lack intuitive interpretation, making it difficult to locate reasons for model misclassification.

P-CNN utilises traditional transfer learning for art image emotion recognition. This model usually fixes the low-level parameters of a pre-trained CNN and only fine-tunes the top-layer classifier. However, the low-level features may not match the target emotional task, resulting in poor feature adaptability. FFTL significantly improves the effectiveness of art image emotion recognition by designing a multi-scale characteristic integration module to achieve deep integration of low-level characteristics and semantic characteristics of art images and through transfer learning to reduce bias among classification accuracy.

Table 2 Comparison of recognition performance across different models

Model	*Accuracy (%)*	*Precision (%)*	*F1 (%)*
SCEP	79.01	77.58	80.63
CNNAR	82.14	84.52	85.61
IAlexNet	87.52	89.05	88.58
P-CNN	91.97	90.51	91.03
FFTL	95.48	93.61	96.08

6 Conclusions

Aesthetic image emotion recognition is currently a research hotspot in affective computing. Aiming at the problem that there are few sources of aesthetic images, small sample sizes, and most emotional analysis uses low-level features of the image with poor accuracy, this article proposes an aesthetic image emotion recognition model based on fused feature extraction and transfer learning. First, CLAHE is used to highlight the colour characteristics of artistic images, extracting H-S 2D features under HSV colour space and CrCb 2D features under YCrCb colour space separately. Multi-scale convolutional kernels are adopted to extract deep semantic information from the network, compensating for the shortcomings brought about by single-scale feature extraction. Secondly, different dimensional feature information is concatenated and channel shuffled further to enhance information flow between feature map channels and improve expression capacity of features. Multi-scale information in the characteristic maps is fused so that the characteristic information at different scales can be effectively preserved. Finally, transfer component analysis algorithm is used for dimensionality reduction on source domain and target domain in original space, making their marginal probability distribution distance smaller. After processing the reduced dimensional source domain and target domain, an improved joint subspace learning method was used to study a characteristic transformation subspace that reduces the conditional probability distribution distance between the source domain and the target domain as well as balances recognition accuracy among categories. At the same time, to prevent overfitting of the feature transformation subspace, the norm of the subspace and distance metric regularisation between source domain and target domain are added. In addition, to enhance the distinguishability of the source domain, distance metric regularisation for source domain category label features is also introduced. Experimental outcome indicates that the accuracy and F1 of the proposed model are 95.48% and 96.08%, respectively,

outperforming the comparison models and effectively achieving the artistic image emotion recognition task.

Declarations

All authors declare that they have no conflicts of interest.

References

Cheng, J., Yang, L. and Tong, S. (2024) 'Painting style and sentiment recognition using multi-feature fusion and style migration techniques', *Informatica*, Vol. 48, No. 21, pp.127–138.

Gatys, L.A., Ecker, A.S. and Bethge, M. (2017) 'Texture and art with deep neural networks', *Current Opinion in Neurobiology*, Vol. 46, pp.178–186.

Gonzalez-Martin, C., Carrasco, M. and Wachter Wielandt, T.G. (2024) 'Detection of emotions in artworks using a convolutional neural network trained on non-artistic images: a methodology to reduce the cross-depiction problem', *Empirical Studies of the Arts*, Vol. 42, No. 1, pp.38–64.

Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z. and Azim, M.A. (2022) 'Transfer learning: a friendly introduction', *Journal of Big Data*, Vol. 9, No. 1, pp.102–113.

Jang, J.-G., Quan, C., Lee, H.D. and Kang, U. (2023) 'Falcon: lightweight and accurate convolution based on depthwise separable convolution', *Knowledge and Information Systems*, Vol. 65, No. 5, pp.2225–2249.

Jiang, Z., Zaheer, W., Wali, A. and Gilani, S. (2024) 'Visual sentiment analysis using data-augmented deep transfer learning techniques', *Multimedia Tools and Applications*, Vol. 83, No. 6, pp.17233–17249.

Kuo, C-C.J. (2016) 'Understanding convolutional neural networks with a mathematical model', *Journal of Visual Communication and Image Representation*, Vol. 41, pp.406–413.

Li, B., Ren, H., Jiang, X., Miao, F., Feng, F. and Jin, L. (2021) 'SCEP – a new image dimensional emotion recognition model based on spatial and channel-wise attention mechanisms', *IEEE Access*, Vol. 9, pp.25278–25290.

Li, J., Ye, Z., Gao, J., Meng, Z., Tong, K. and Yu, S. (2024) 'Fault transfer diagnosis of rolling bearings across different devices via multi-domain information fusion and multi-kernel maximum mean discrepancy', *Applied Soft Computing*, Vol. 159, pp.11–20.

Li, W., Li, J., Li, J., Huang, Z. and Zhou, D. (2021) 'A lightweight multi-scale channel attention network for image super-resolution', *Neurocomputing*, Vol. 456, pp.327–337.

Liu, S. and Agaian, S.S. (2024) '3DEmo: for portrait emotion recognition with new dataset', *ACM Journal on Computing and Cultural Heritage*, Vol. 17, No. 2, pp.1–26.

Liu, S., Agaian, S. and Grigoryan, A. (2024) 'PortraitEmotion3D: a novel dataset and 3D emotion estimation method for artistic portraiture analysis', *Applied Sciences*, Vol. 14, No. 23, pp.23–35.

Lu, J. and Wan, X. (2022) 'Image recognition algorithm based on improved AlexNet and shared parameter transfer learning', *Academic Journal of Computing & Information Science*, Vol. 5, No. 12, pp.6–14.

Mondal, K., Rabidas, R. and Dasgupta, R. (2024) 'Single image haze removal using contrast limited adaptive histogram equalization based multiscale fusion technique', *Multimedia Tools and Applications*, Vol. 83, No. 5, pp.15413–15438.

Ruan, S., Zhang, K., Wu, L., Xu, T., Liu, Q. and Chen, E. (2021) 'Color enhanced cross correlation net for image sentiment analysis', *IEEE Transactions on Multimedia*, Vol. 26, pp.4097–4109.

Song, H., Wang, Z., Cao, W., Zhang, Y. and Leng, X. (2025) 'Infrared image enhancement based on guided filtering and adaptive algorithm and its FPGA implementation', *Microwave and Optical Technology Letters*, Vol. 67, No. 1, pp.70–85.

Tashu, T.M., Hajiyeva, S. and Horvath, T. (2021) 'Multimodal emotion recognition from art using sequential co-attention', *Journal of Imaging*, Vol. 7, No. 8, pp.15–27.

Vonder, L., Elvira, T. and Ochoa, O. (2023) 'An analysis of explainability methods for convolutional neural networks', *Engineering Applications of Artificial Intelligence*, Vol. 117, pp.56–62.

Wang, D. (2022) 'Research on the art value and application of art creation based on the emotion analysis of art', *Wireless Communications and Mobile Computing*, Vol. 10, No. 3, pp.24–36.

Wang, M., Zhao, Y., Wang, Y., Xu, T. and Sun, Y. (2023) 'Image emotion multi-label classification based on multi-graph learning', *Expert Systems with Applications*, Vol. 231, pp.12–20.

Xu, X., He, L., Lu, H., Shimada, A. and Taniguchi, R-I. (2016) 'Non-linear matrix completion for social image tagging', *IEEE Access*, Vol. 5, pp.6688–6696.

Yang, H., Fan, Y., Lv, G., Liu, S. and Guo, Z. (2023) 'Exploiting emotional concepts for image emotion recognition', *The Visual Computer*, Vol. 39, No. 5, pp.2177–2190.

Yang, Y. and Zou, H. (2015) 'A fast unified algorithm for solving group-lasso penalize learning problems', *Statistics and Computing*, Vol. 25, No. 6, pp.1129–1141.

Zabora, V., Kasianenko, K., Pashukova, S., Alforova, Z. and Shmehelska, Y. (2023) 'Digital art in designing an artistic image', *Amazonia Investiga*, Vol. 12, No. 64, pp.300–305.

Zhang, J., Duan, Y. and Gu, X. (2021) 'Research on emotion analysis of Chinese literati painting images based on deep learning', *Frontiers in Psychology*, Vol. 12, pp.72–85.

Zhang, J., Liu, X., Chen, M., Ye, Q. and Wang, Z. (2022) 'Image sentiment classification via multi-level sentiment region correlation analysis', *Neurocomputing*, Vol. 469, pp.221–233.

Zhang, J., Yu, J. and Tao, D. (2018) 'Local deep-feature alignment for unsupervised dimension reduction', *IEEE Transactions on Image Processing*, Vol. 27, No. 5 pp.420–2432.

Zheng, Z., Zhao, W., Hable, B., Gong, Y., Wang, X., Shannon, R.W. and Liu, K. (2022) 'Transfer learning-based independent component analysis', *IEEE Transactions on Automation Science and Engineering*, Vol. 21, No. 1, pp.783–798.