PeerJ Computer Science

Emotion classification of artistic images using domain adaptation and transfer learning

Jingwen Wang[1], Yisong Yang[2] and Kemal Polat[3]

[1] Faculty of Innovation and Design, City University of Macau, Macao, China
[2] Film and Television Academy, Yunnan Arts University, Kunming, China
[3] Faculty of Engineering, Department of Electrical and Electronics Engineering, Bolu Abant Izzet Baysal University, Bolu, Turkey

ABSTRACT

As society evolves, the appreciation and pursuit of art continue to grow. However, current technology struggles to intelligently interpret the emotional expressions conveyed in images. To enhance the understanding of emotions expressed in artistic images, we propose a novel emotion classification method that integrates domain adaptation and transfer learning. We first introduce an attention-based salient feature extraction technique designed to emphasize the primary artistic elements within an image and enhance the corresponding regions. Leveraging these salient features, we then develop a domain-adaptive image emotion classification model to capture semantic information and accurately recognize the emotional essence of artistic content. Experimental results validate the effectiveness of our approach, achieving a mean average precision (mAP) of 92.4% and an accuracy of 98.9%, demonstrating its capability to provide precise emotional interpretations of artworks. Our method offers a significant advancement in the intelligent analysis of artistic images, combining attention mechanisms, domain adaptation, and transfer learning to improve emotional understanding in visual art.

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Data Mining and Machine Learning, Social Computing, Neural Networks
Keywords Art image, Emotion classification, Domain adaptation, Transfer learning

Submitted 1 April 2025
Accepted 8 September 2025
Published 10 October 2025

Corresponding author
Yisong Yang, 290046@ynart.edu.cn

Academic editor
Siddhartha Bhattacharyya

Additional Information and Declarations can be found on page 18

DOI **10.7717/peerj-cs.3250**

INTRODUCTION

Artists express emotions through color, line, composition, and other techniques, making artistic images a vital research object for sentiment analysis. Sentiment classification of art images (*Zhang et al., 2024b*; *Silva, Cardoso & Rodrigues, 2025*; *Kulkarni & Dixit, 2025*; *Wu et al., 2025*) is not only helpful in excavating the intrinsic value of artworks but also provides strong support for art appreciation, education, psychotherapy and other fields. Artists express their emotions through methods such as color, line, and composition, which makes art images a vital research object for sentiment analysis. Emotion classification of art images is not only helpful in excavating the intrinsic value of artworks but also provides strong support for art appreciation, education, psychotherapy and other fields.

The emotional classification of art images presents multiple challenging difficulties. The first problem is that the emotional expressions contained in artistic images are more

complex and variable than those in everyday images. They convey a specific emotional atmosphere through various visual elements, such as color, line, shape, and composition (*Deng et al., 2024*; *Agung, Rifai & Wijayanto, 2024*). However, there are variants in the understanding of these elements among different cultural contexts, artistic schools and audience groups. Secondly, the existing models have limitations in generalization ability. When facing novel or unseen art images, it is difficult to accurately capture and parse the emotional features, which affects the accuracy of classification. Moreover, the emotional classification of art images often requires identifying a variety of emotional states, such as happiness, sadness, surprise, fear, anger, and disgust, in detail (*Khare et al., 2024*). However, most current emotion classification techniques can only achieve a relatively broad classification, which is difficult to meet the fine requirements of art image emotion classification (*Hosseini, Yamaghani & Poorzaker Arabani, 2024*; *Chen et al., 2024*). Finally, audiences from different cultures may have various emotional resonances for an art image. At the same time, different artistic schools and styles will also influence the audience's interpretation of the image's emotional impact. Therefore, ensuring that the emotion classification model can cross cultural and background boundaries and exhibit good adaptability and generalization ability is key in current research (*Zhang et al., 2024a*).

To address the problems above, numerous image sentiment classification algorithms have been proposed. *Yao et al. (2020)* developed a framework that simultaneously optimizes retrieval and classification tasks, introducing an adaptive similarity loss to adjust the disparities among various image pairs dynamically. *Tamil Priya & Divya Udayan (2020)* improved the accuracy of the image sentiment classification method by transfer learning technique. *Yadav & Vishwakarma (2020)* proposed a network that employs residual attention to focus on the key local regions rich in emotion in the image, thereby learning the spatial hierarchical structure of image features. *Xue et al. (2020)* proposed using attention to eliminate sample interference by decreasing the attention score during training. They introduced a weakly supervised learning method that focuses on essential positions within a sample. *Deng et al. (2021)* employed salient detection techniques to identify emotional regions in images, providing additional information for more accurate emotion recognition. *Ou et al. (2021)* proposed a pyramid network for context, utilizing ResNet-101 as the backbone to obtain multi-level emotion representations and extract context features through a multi-scale adaptive context module. Finally, it combines them for image emotion classification. *Zhang & Xu (2020)* introduced a method by a feature pyramid network, which combines features extracted at different depths. The results indicate that improved performance can be achieved by combining multi-scale features. *Wang et al. (2022)* employed a cross-channel Max pooling strategy to train a complete convolutional network for extracting discriminative feature maps. They then utilized an adversarial erasure operation to compel the network to identify distinct emotional discrimination regions. Finally, an adaptive feature fusion mechanism was employed to integrate discriminative and diverse emotional representations. *García-Domínguez et al. (2024)* used a style transfer network to change the memorability of images. *Hong, Zaheer & Wali (2024)* proposed a local color transfer method using a deep neural network. Given a source image and a target image corresponding to each object and background region in

the source image, the masks of each object and background region are predicted, and then local image color transfer is performed region-by-region using these masks. *Zhao et al. (2025)* introduced a color transfer method based on gradient updates. They utilized the structural similarity of the source image as a measure of content similarity. They performed multiple gradient updates on the output image until the color was sufficiently close to that of the reference image.

Although the methods above have achieved some success, specific approaches have not fully taken into account the potential impact of the background on the subject's emotions. For instance, some methods focus on analyzing local emotional regions. While they can capture significant local emotional information, irrelevant elements in the background under complex circumstances may interfere with the accurate judgment of the subject's emotions, leading to deviations in classification results. Other methods, although analyzing from a global perspective, may also introduce background noise during the multi-scale fusion process, which can affect the accurate extraction of local emotional correlations. Additionally, there remains a lack of sufficient model generalization ability and accuracy in emotional classification for new images. To address these challenges, we propose an artistic image sentiment classification scheme combining domain adaptation and transfer learning. The core of the scheme is to make full use of the ability to transfer cross-domain data and knowledge, thereby enhancing the recognition and understanding of art image emotions. Firstly, the advanced features process is used to effectively separate the image background from the subject, allowing the model to focus more on extracting the emotional features of the image subject. Then, given the diversity of artistic image styles and the complexity of emotional expression, we design a set of domain-adaptive frameworks to adapt to the sentiment classification task across different styles and artistic schools (*Sicilia, Zhao & Hwang, 2023*). This framework achieves effective knowledge transfer by minimizing the difference between the source and the target. At the same time, we integrate deep transfer learning technology. By pre-training a sentiment classification model that performs well on large-scale natural image datasets, we fine-tune it with the unique emotional features of artistic images. Considering the diversity of sentiment expression, we also explore the possibility of combining multimodal information, such as text descriptions and color analysis, to enrich the basis of sentiment classification. Through the multimodal fusion strategy, the model can capture the emotional features of images from multiple dimensions, thereby further improving the robustness and accuracy of classification.

The main contributions are as follows:

(1) By leveraging cross-domain data and knowledge transfer, we propose an artistic image emotion classification scheme that integrates domain adaptation and transfer learning, aiming to enhance the model's capability to recognize and understand the emotions conveyed in artistic images.

(2) To minimize the distribution disparity between the source domain and the target domain, we propose a domain adaptation-based image emotion classification method that adapts to emotion classification tasks across different styles and artistic genres.

RELATED WORKS

In image sentiment classification, numerous researchers have proposed various approaches (*Yadav & Vishwakarma, 2020*; *Wu et al., 2020*). *Yang et al. (2018a)* proposed a two-branch weakly supervised coupling framework, Weakly Supervised Coupled Network (WSCNet), which constructs a sentiment map through the object detection branch to generate the sentiment score of local semantic information, allowing for explicit perception of local regions in the image. *Rao, Li & Xu (2020)* designed Multi-level Deep Representation Network (MldrNet) to achieve effective sentiment classification of different image types by combining features at multiple levels, including semantics, aesthetics, and vision. The Multi-level Context Pyramid Network (MCPNet), designed by *Ou et al. (2021)*, enables multi-scale local sentiment correlation analysis and fusion based on a global view, recognizing objects with varying appearances and complexities, and achieves the best results in two-dimensional image sentiment analysis. *Lee, Ryu & Osanet (2022)* proposed the object semantic attention network, which extracts the subject target in the local branch and performs word embedding to represent the local target. *Yang et al. (2018b)* obtained the output of the convolutional neural network as a global sentiment feature map, coupled the original output through the sentiment heat map to form a local sentiment representation, and combined the global and local sentiment feature maps, thereby improving the complexity of the extracted features. *Xu et al. (2024)* combined three sub-analysis modules of salient object sentiment analysis, facial object sentiment analysis and image overall sentiment analysis, and input different scale sub-images of the image into three classification models with varying methods of training and purposes, and fused the output results of the three models to obtain more complex features. *Li et al. (2023)*, introduced a discriminant network strategy, considering that placing too much emphasis on locality can overlook global discriminant information. The sentiment graph and discriminant enhancement graph are applied to the features and summarized into a sentiment vector, serving as the basis for classification. It fully captured the overall information and local information of the emotional image.

MATERIALS AND METHODS

To address the interference of image background on subject emotion and the model's limited generalization ability, we propose an art image sentiment classification method based on domain adaptation and transfer learning (DATL). Firstly, we introduce the attention mechanism (AM) to further highlight the key emotional cues of the subject area by calculating the importance weights of different regions for the sentiment classification task, which can more accurately capture the emotional image features. Then, to overcome the challenges posed by the diversity of artistic image styles and the complexity of emotional expression, we design a domain-adaptive framework. This framework achieves effective knowledge transfer. We adopt an adversarial training strategy, where the domain discriminator and the feature extractor engage in an adversarial game, enabling the model to learn cross-domain common emotional features. We employ the pre-trained sentiment classification model and fine-tune it by combining the unique emotional features of art images. Through transfer learning, the model can leverage the general emotional features

learned from large-scale natural image datasets while learning the unique emotional expression patterns of artistic images. This strategy significantly enhances the model's adaptation ability and classification accuracy for new art images.

Data preprocessing steps

Image collection and labeling

We first collect a diverse dataset of artistic images, covering various forms such as paintings, digital art, and sketches. Each image in the dataset is assigned an emotion label, reflecting the emotion conveyed by the artwork (*e.g.*, happiness, sadness, anger, surprise). If a pre-labeled dataset is not available, we manually annotate the images or leverage crowdsourcing to gather emotion labels, ensuring the dataset includes a broad range of emotional expressions.

Image resizing

Artistic images come in various sizes, so we resize them to a consistent dimension (*e.g.*, 224 × 224 or 256 × 256 pixels) to maintain uniformity. To prevent distortion, we apply padding if necessary, preserving the aspect ratio of the images. This ensures that the visual composition of the artwork remains intact, which is crucial for emotional recognition.

Image normalization

For model efficiency, we normalize the pixel values of the images, scaling them to a range of 0 to 1. Additionally, to match the preprocessing standards used in transfer learning, we standardize the images by subtracting the mean pixel values and dividing by the standard deviation (*e.g.*, values used for ImageNet models). This enables the model to leverage pre-trained networks effectively, thereby improving convergence and accuracy.

Data augmentation

Given the variety in artistic styles, we apply data augmentation techniques to enhance the model's generalization ability. This includes:

- Random Flips: both horizontal and vertical flips are performed to introduce variability.
- Rotation: minor rotations (*e.g.*, −15 to +15 degrees) help simulate changes in image orientation.
- Color Jitter: adjustments to brightness, contrast, and saturation simulate different lighting and artistic styles.
- Random Cropping: this feature mimics the effect of zooming or focusing on different parts of the artwork.
- Scaling: random scaling is applied to introduce variability in image distances, enhancing the model's robustness.

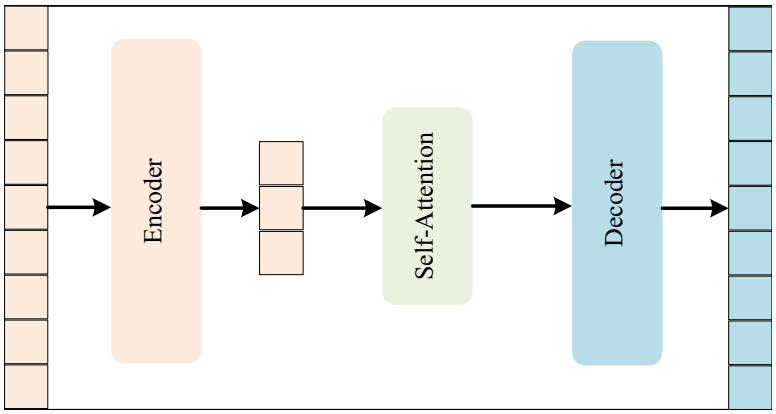

Figure 1 **The framework of AMSF.** Full-size DOI: 10.7717/peerj-cs.3250/fig-1

Saliency feature extraction of artistic images based on attention mechanism

To distinguish the main area of the art image and fully express the emotional subject of the art image, we propose an AM-based salient feature extraction method for art images (AMSF), as shown in Fig. 1. Firstly, we feed the features extracted by the encoder into the covariance pooling function to calculate a weight value for each channel. We then repeat this process for all channels to generate a matrix of feature weights. This weight matrix is then used to weigh the features, highlighting the feature information of interest. Finally, the weighted features are fed into the activation function for the AM weight value, located in the interval [0, 1].

We embed channel attention into both the generator and the discriminator. The AM module in the discriminator aims to guide the generator to focus on those regions that are crucial for generating close to the real data distribution. In the generator, the core goal of the AM module is to highlight the parts within the region of interest that are different from the input. The generator pays special attention to the AM feature map regions of the training input samples to help judge the authenticity of the target image. The process of the channel AM is shown in Fig. 2. We adopt the covariance pooling method. Compared to traditional average pooling or Max pooling, covariance pooling utilizes second-order statistics to model the joint distribution of spectral and spatial information, thereby providing a richer feature representation. Assume that the image feature is I $\in$ RC $\times$ H $\times$ W. Then, we can represent the AM as follows.

$$I_1 = Var_{pool}(I) \tag{1}$$

$$I_2 = Conv(I_1) \tag{2}$$

$$I' = \sigma(FC(I_2)) \tag{3}$$

where I' denotes the feature vector output by the AM, σ(.) represents the activation function, FC(.) refers to the fully connected layer, Varpool(.) denotes covariance pooling.

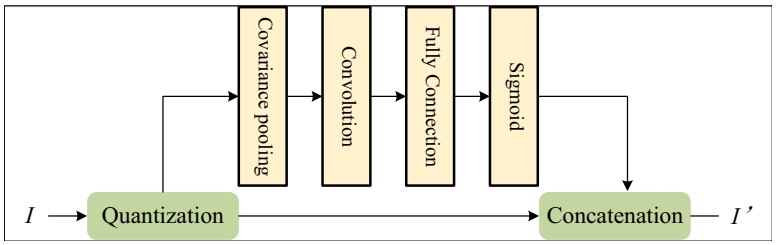

Figure 2 **Channel attention mechanism.** Full-size DOI: 10.7717/peerj-cs.3250/fig-2

After obtaining the weight vector output by all C channels of the AM, we weigh each channel of input I according to its corresponding attention weight.

Image sentiment classification method by domain adaptation

To enhance the generalization of the image sentiment classification model, we propose a domain-adaptive image sentiment classification method (DAIS).

We apply deep neural networks to extract high-order features of source and target images. To reduce the feature distribution gap between the two domains, we introduce a domain discriminator, which forms an adversarial relationship with the feature extractor. During training, the feature extractor gradually learns to generate features that are difficult for the domain discriminator to recognize and then aligns the source features with the target features. The primary task of the domain differentiator D is to distinguish the domain of the data, *i.e.*, to determine which domain the data originates from. D consists of five stacked convolutional structures. The convolution kernel in the first four layers has a size of 3 × 3, and the number of channels in each layer is 18, 36, 72, and 144, respectively. These convolutional layers are equipped with the LeakyReLU activation function and standardization layer BatchNorm. The last layer has a 1 × 1 kernel size and a single channel, followed by a sigmoid activation. The output of D is the confidence classification probability of the domain to which the image belongs.

On the other hand, the task of label classifier C is to classify the feature vector from the source domain and accurately predict its class label. C comprises a convolutional structure with four layers and two fully connected layers. Finally, C will output the probability distribution of the category to which the image belongs in the semantic space, as shown in Fig. 3.

The loss of the entire network system comprises two major components: the classifier loss and the discriminator loss. The trainable weights of the label classifier and the discriminator are denoted as δD and δC, respectively. For the network model, the optimal solution of its parameters can be expressed as:

$$\hat{\delta}_C = \arg\min_{\delta_C} L(C) \tag{4}$$

$$\hat{\delta}_D = \arg\max_{\delta_D} L(D). \tag{5}$$

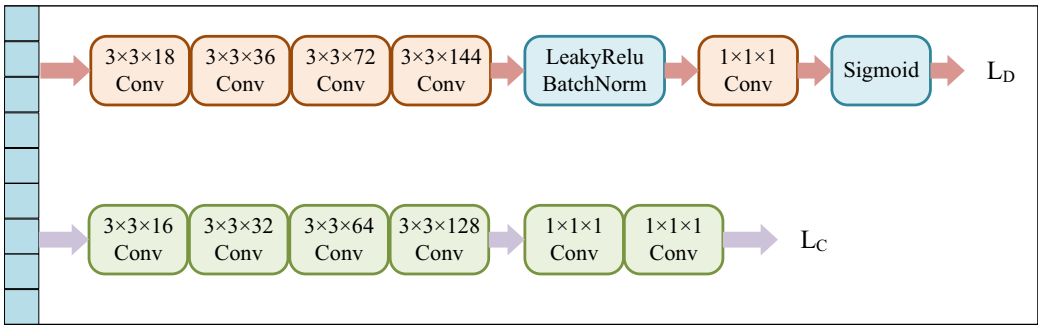

Figure 3 **Domain discriminator and label classifier.** Full-size DOI: 10.7717/peerj-cs.3250/fig-3

The classifier loss L(C) is determined by calculating the distance between the predicted classification probability y' of the source domain image and the true label y. Here, we adopt the cross-entropy LC as the loss metric of the classifier:

$$L(C) = \frac{1}{N}\sum_{i=1}^{N} L_C(y', y) \tag{6}$$

where N is the classes' number.

The domain discriminator loss L(D) is composed of the source domain discrimination loss LS(D) and the target domain loss LT(D). The source discriminative loss LS(D) is obtained by calculating the distance between the clean and the source labels yS = 1. The target loss LT (D) is calculated by the aloofness between the adversarial example and the target label yT = 0:

$$L(D) = L_S(D) + L_T(D) = \frac{1}{N}\sum_{i=1}^{N} L_C\left(y'_S, 1\right) + \frac{1}{M}\sum_{j=1}^{N} L_C\left(y'_T, 0\right). \tag{7}$$

Therefore, the loss L can be presented as:

$$L = L(D) + L(C). \tag{8}$$

EXPERIMENTS

Dataset and implementation settings

We use the ART500K Dataset (doi: https://doi.org/10.1145/3123266.3123405) for analyzing art image emotion classification results. ART500K is an extensive visual arts dataset employing over 500,000 images, each associated with more than 10 attribute labels. In addition to general labels such as artist, genre, and art movement, it also includes specialized labels, such as event, historical figure, and description. This dataset is versatile and can be utilized for various tasks, including visual arts classification, retrieval, and image captioning. Both the computer science and visual arts communities can derive significant benefits from it. During the training phase, we enhance computational efficiency by utilizing a Ryzen 9 9950X3D processor paired with four Nvidia A5000 GPUs. We selected Caffe as our framework and adjusted its settings according to the training

Table 1 Model settings.

Parameters	Value
Initial learning rate	4×10^{-3}
Epoch	56
Momentum	0.25

parameters specified in Table 1 for optimal performance. Meanwhile, we employ a series of data preprocessing techniques to enhance the artistic quality of images. Firstly, we adjust the image dimensions to meet the model's input requirements while removing irrelevant edge portions and normalizing the images. Secondly, we increase data diversity through geometric transformations, such as rotation, flipping, and scaling, and add Gaussian noise, salt-and-pepper noise, and other types of noise to the images to enhance the model's robustness against noise.

To valid our method, we select mean average precision (mAP) and accuracy (Acc) as the metrics for sentiment classification, with their respective formulas provided below:

$$\bar{P} = \frac{TP}{TP + FP} \tag{9}$$

$$\bar{R} = \frac{TP}{TP + FN} \tag{10}$$

$$AP = \bar{P} \times \sum_{n} (\bar{R}_n - \bar{R}_{n-1}) \tag{11}$$

$$mAP = \sum_{i=1}^{N} AP_i \tag{12}$$

where TP, FP and FN stand for true positive, false positive, and false negative examples, respectively. mAP combines precision and recall, making it suitable for multi-class classification tasks. In particular, emotional classification can provide a comprehensive reflection of the model's average performance across different categories. Additionally, Acc can intuitively represent the model's overall accuracy, making it applicable to tasks with balanced class distributions. It is a commonly used metric for quickly evaluating model performance. In addition, we also employ the precision (P), recall (R) and F1-score (F1) to evaluate the model. The entire metrics follows method (*Deng et al., 2024*).

To further ensure a comprehensive and balanced evaluation of the classification model, we additionally computed Cohen's kappa and Matthews correlation coefficient (MCC).

The formulas are as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{13}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

where p_o is the observed agreement and p_e is the expected agreement by chance.

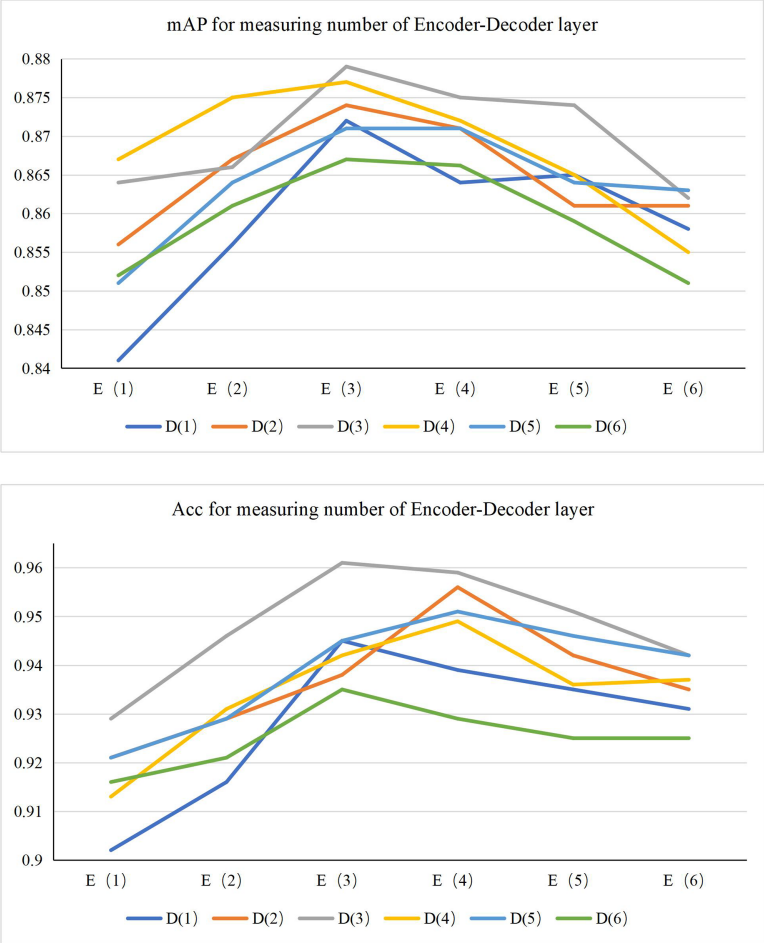

Figure 4 **Measuring experiments for the number of encoder-decoder layer.**
Full-size DOI: 10.7717/peerj-cs.3250/fig-4

Parameter experiments

Before conducting ablation experiments and comparison experiments, our first task is to optimize the number of encoder and decoder layers in the AMSF model. These two components are built based on the self-attention mechanism. The experiments are in Fig. 4. By observing the mAP performance index plot, we can see that when the encoder and decoder layers are set to 3, the model exhibits the best performance. At the same time, the Acc index plot also verifies this conclusion.

Furthermore, to gain a profound understanding of the rationale behind this optimal configuration, we delve into the specific impact of different layers on the model's performance. When the number of encoder and decoder layers is fewer than three, the model struggles to grasp the intricate features of the input data. This leads to inadequate information extraction, which subsequently affects the decoding process and ultimately compromises the prediction accuracy. This deficiency is evident in the performance decline observed in both the mAP and Acc metrics. Conversely, when the number of layers surpasses three, while the model is theoretically capable of learning deeper feature

Table 2 Ablation experiments.

AMSF	DAIS	mAP@1	mAP@5	Acc@1	Acc@5
Baseline		0.879	0.912	0.961	0.972
O		0.895	0.926	0.977	0.986
	O	0.903	0.921	0.979	0.984
O	O	0.924	0.951	0.989	0.994

representations, it may encounter practical challenges. Specifically, an excessive number of layers can introduce the risk of overfitting, where the model becomes overly tailored to the training data and performs poorly on unseen data. Additionally, it significantly escalates the computational complexity, leading to longer processing times. Such inefficiency is not conducive to deployment and can hinder the model's practical applicability.

Therefore, considering the performance of mAP and Acc metrics, as well as the model's complexity, training efficiency, and generalization ability, we determined the optimal configuration of three encoder-decoder layers for the AMSF model. This conclusion not only provides a solid baseline for subsequent ablation and comparison experiments but also indicates the direction for optimizing the model's performance in practical applications. In the following experiments, we will explore the model's adaptability in the dataset with this optimal configuration to obtain a more comprehensive and in-depth evaluation of model performance and optimization strategy.

Ablation experiments

We will thoroughly analyze the impact of the AMSF and DAIS modules on model performance. In this experiment, convolutional neural network (CNN) is the baseline. In Table 2, compared to the base model, the introduction of the AMSF module brings significant performance improvements: mAP@1 increases by 1.4%, mAP@5 increases by 1.4%, Acc@1 increases by 1.6%, and Acc@5 increases by 1.4%. The DAIS module also performs well, improving mAP@1 by 2.4%, mAP@5 by 0.9%, Acc@1 by 1.7%, and Acc@5 by 1.2%, respectively. It is particularly worth noting that when AMSF is combined with the DAIS module, the model's performance reaches its peak, with mAP@1 reaching 92.4%, mAP@5 reaching 95.1%, Accuracy@1 reaching 98.9%, and Accuracy@5 reaching 99.4%.

The AMSF module, as an attention-based salient feature extraction method for artistic images, dynamically adjusts the weights of different regions, allowing our model to concentrate on the most informative parts. This mechanism effectively enhances the accuracy and efficiency of feature extraction, resulting in significant improvements across multiple performance indicators. In particular, the performance of the AMSF module is particularly outstanding on mAP@1 and Acc@1, which reflects its advantages in accurately capturing key image features. The DAIS module is an image sentiment classification method based on domain adaptation. By minimizing the discrepancy between the source domain and the target domain, the model's capacity to generalize to unobserved data is enhanced. The improvement of this ability is particularly evident in the indicators mAP@1 and Acc@1, indicating that the DAIS module demonstrates a substantial impact in

Table 3 Comparison with article in terms of Acc.

Methods	Acc@1	Acc@5	Acc@10
LCDP (*Deng et al., 2024*)	0.877	0.914	0.923
MSVP (*Bisogni et al., 2024*)	0.894	0.923	0.942
EML (*Pan, Lu & Wang, 2024*)	0.923	0.965	0.988
POSER (*Govarthan et al., 2024*)	0.912	0.934	0.987
CNN (*Dixit & Satapathy, 2024*)	0.876	0.923	0.957
ESIF (*Kulkarni & Dixit, 2025*)	0.935	0.958	0.987
LSK (*Zhao et al., 2025*)	0.925	0.954	0.989
MTCN (*Zhang, Zhou & Qi, 2025*)	0.965	0.0978	0.994
HOLF (*Borgalli & Surve, 2025*)	0.935	0.973	0.986
OFEDA (*Akrout, 2025*)	0.956	0.974	0.989
Ours	0.989	0.994	0.997

Table 4 Comparison with article in terms of P, R F1, κ, MCC and ROC-AUC.

Methods	P	R	F1	Cohen's Kappa (κ)	MCC	ROC-AUC
LCDP (*Deng et al., 2024*)	0.848	0.795	0.823	0.812	0.794	0.891
MSVP (*Pan, Lu & Wang, 2024*)	0.867	0.821	0.836	0.836	0.815	0.904
EML (*Govarthan et al., 2024*)	0.859	0.877	0.867	0.848	0.822	0.912
POSER (*Bisogni et al., 2024*)	0.869	0.857	0.862	0.864	0.837	0.917
CNN (*Dixit & Satapathy, 2024*)	0.864	0.838	0.845	0.778	0.754	0.871
ESIF (*Kulkarni & Dixit, 2025*)	0.891	0.908	0.893	0.843	0.817	0.910
LSK (*Zhao et al., 2025*)	0.884	0.859	0.872	0.826	0.804	0.895
MTCN (*Zhang, Zhou & Qi, 2025*)	0.936	0.891	0.926	0.857	0.832	0.918
HOLF (*Borgalli & Surve, 2025*)	0.874	0.914	0.896	0.832	0.810	0.906
OFEDA (*Akrout, 2025*)	0.926	0.901	0.917	0.849	0.823	0.913
Ours	0.958	0.941	0.947	0.942	0.936	0.972

accommodating diverse datasets and sentiment distributions. At the same time, although the improvement on mAP@5 and Acc@5 is relatively small, it still maintains a stable performance gain, which shows its broad application potential. When the AMSF is jointly applied with the DAIS module, their synergy is maximized. The accurate feature extraction provided by the AMSF module enables high-quality data input for the DAIS module, and the domain adaptation ability of the DAIS module further enhances the model's robustness across different contexts. This complementarity enables the joint model to achieve its peak in multiple performance metrics, particularly in mAP@1, mAP@5, Acc@1, and Acc@5, resulting in significant performance improvements.

Compare with others

We conduct an in-depth comparative evaluation of DATL's performance, utilizing article as a benchmark for comparison. As evidenced by the data presented in Tables 3, 4 and Fig. 5, our method excels across multiple core performance indicators, significantly

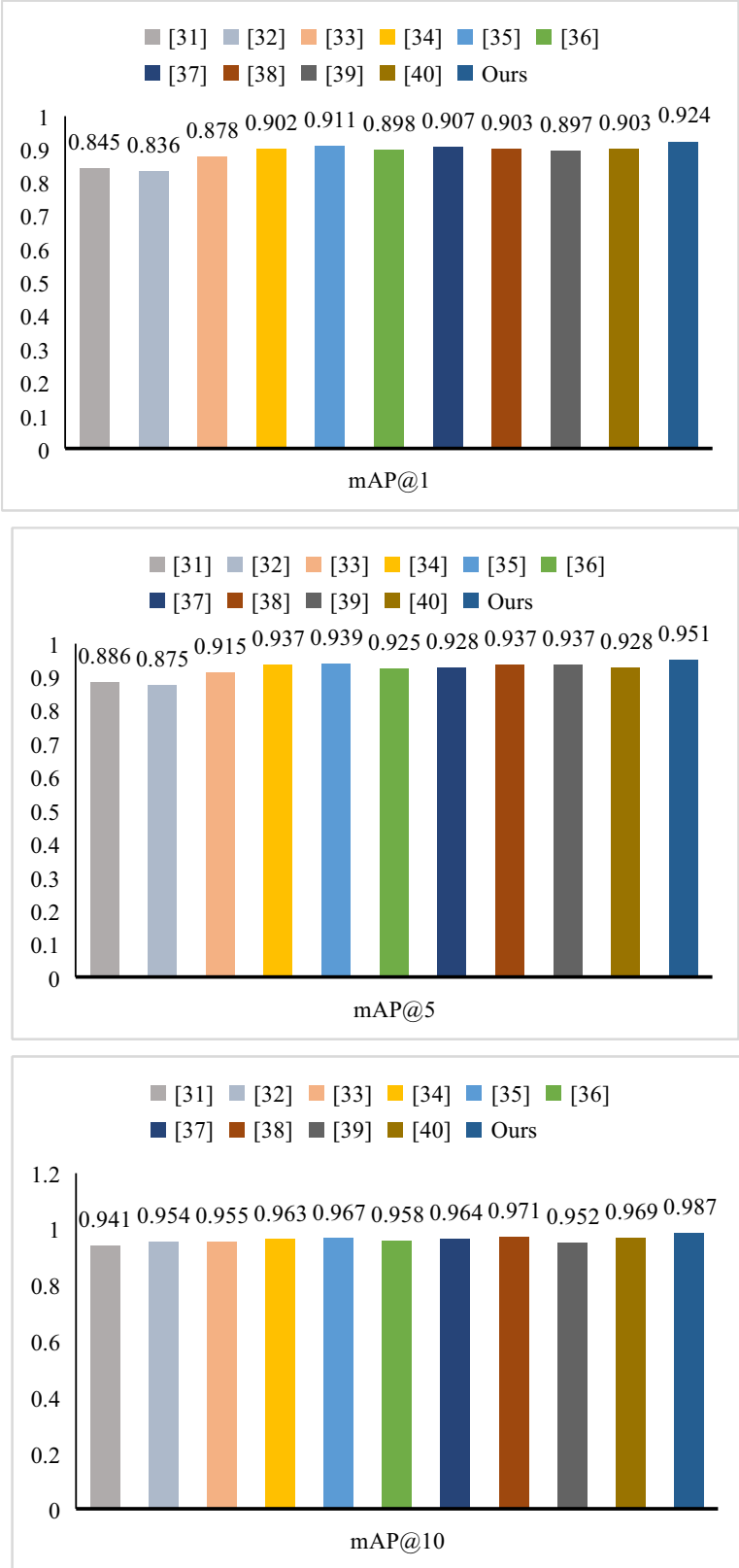

Figure 5 **Comparison of mAP with other methods.** Full-size DOI: 10.7717/peerj-cs.3250/fig-5

surpassing the selected comparative literature. In terms of mAP, our method attains scores of 0.924 for mAP@1, 0.951 for mAP@5, and 0.987 for mAP@10. Regarding Acc, our method achieves 0.989 at Acc@1, 0.994 at Acc@5, and 0.997 at Acc@10. When compared to articles *Li et al. (2023)* and *Deng et al. (2024)*, our method demonstrates a 7.9% improvement in mAP@1. Furthermore, in comparison to article (*Bisogni et al., 2024*), our method boasts a substantial advantage of 11.3% in Accuracy@1. In addition, with regard to P, R and F1, our method achieves a precision score of 0.958, a recall score of 0.941, and an F1-score of 0.947. Similarly, our method outperformed all the compared methods. These figures comprehensively underscore the superiority and advancement of DATL in terms of performance.

Our method achieves the highest values in both metrics (kappa = 0.942, MCC = 0.936), significantly outperforming existing approaches such as POsed *vs* Spontaneous Emotion Recognition (POSER) (*Bisogni et al., 2024*) and Multimodal Temporal Context Network (MTCN) (*Zhang, Zhou & Qi, 2025*). Cohen's kappa accounts for agreement beyond chance, making it particularly suitable for evaluating emotion classification tasks where class distributions may be skewed. MCC, as a correlation coefficient between observed and predicted classifications, provides a robust summary of binary and multi-class prediction quality. These results further validate the robustness, reliability, and generalization ability of the proposed DATL method across diverse emotional categories.

As a feature extraction method, the core value of AMSF lies in accurately identifying and extracting the salient features in an image. Within the framework of DATL, this ability of AMSF is particularly important because it directly affects the effectiveness of transfer learning. By simulating the human visual attention mechanism, AMSF can highlight the most noticeable regions or objects in an image, thereby extracting more representative features. These features can better express the commonality, which helps improve the model's generalization ability. AMSF can mitigate the impact of such differences on the transfer learning effect to some extent by extracting salient features. Although AMSF itself is not directly used for sentiment classification, the salient features extracted by AMSF can be used as essential inputs for sentiment classification models. In DATL, these features can accurately represent the image content, thereby improving the accuracy of sentiment classification.

DAIS utilizes domain adaptation technology to transfer sentiment classification knowledge, thereby adapting to the image data distribution in different domains. By utilizing much-labeled data in the source domain, DAIS can effectively alleviate this issue and improve the sentiment classification model in the target domain. DAIS can transfer sentiment classification knowledge, allowing the model to maintain stable performance across different image data distributions. Through domain adaptation technology, DAIS can enhance the model's generalization ability. This enables the model to better adapt to various image sentiment classification tasks and enhance its overall classification performance.

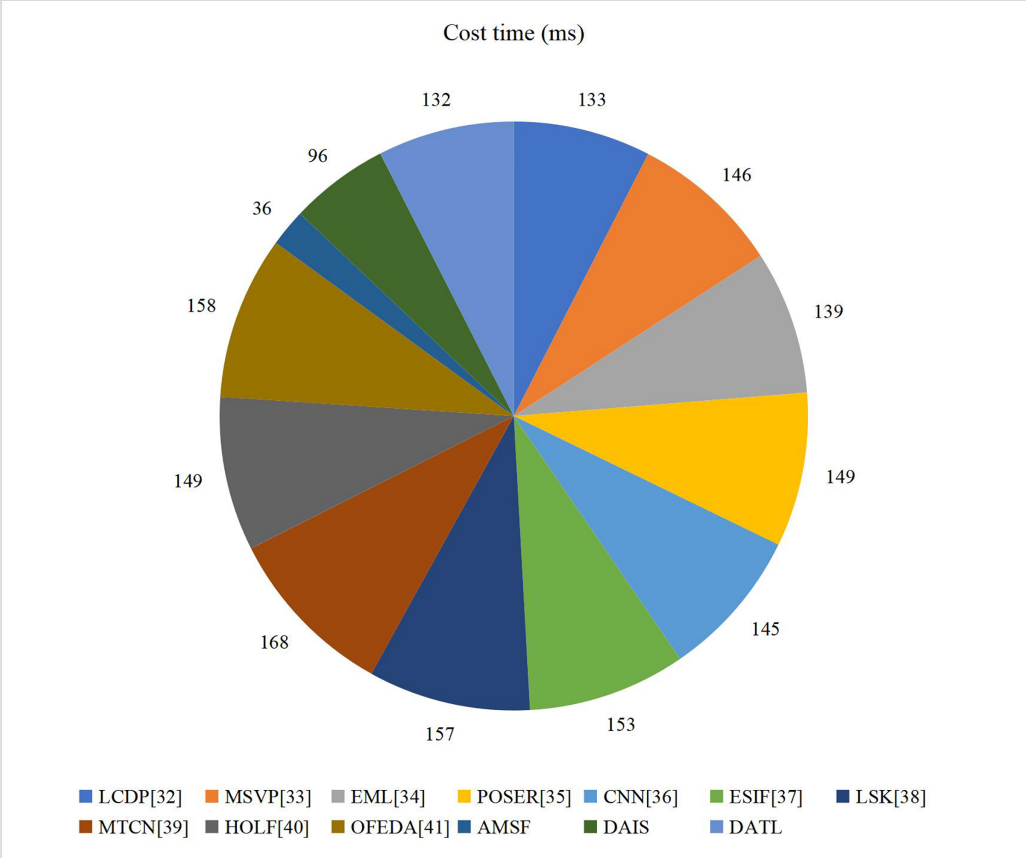

Figure 6 **Implement time of AMSF, DAIS and DATL.** Full-size DOI: 10.7717/peerj-cs.3250/fig-6

Application testing

Firstly, we conduct a comprehensive and in-depth evaluation of the running efficiency of DATL, aiming to verify its real-time performance and efficiency in processing image sentiment classification tasks. To more intuitively illustrate the running efficiency of DATL, we compare its running time with that of various current mainstream and advanced image sentiment classification methods. The results are presented in Fig. 6 as charts. Through a careful review of Fig. 6, we can find that under the same test environment and conditions, AMSF only takes 36 milliseconds to run, which fully demonstrates its ability to process image feature extraction efficiently. Although DAIS is relatively complex, it can still complete the sentiment classification task in a remarkably short time, taking only 96 milliseconds. By combining the running times of AMSF and DAIS, the total processing time of DATL for the entire image sentiment classification process is only 132 milliseconds, which is the best among similar methods. DATL outperforms all other methods in the comparison by a significant margin. This not only means that DATL can complete the image sentiment classification task in a shorter time, but also provides strong support for its wide deployment and real-time processing in practical applications. Therefore, we can conclude that the excellent performance of DATL

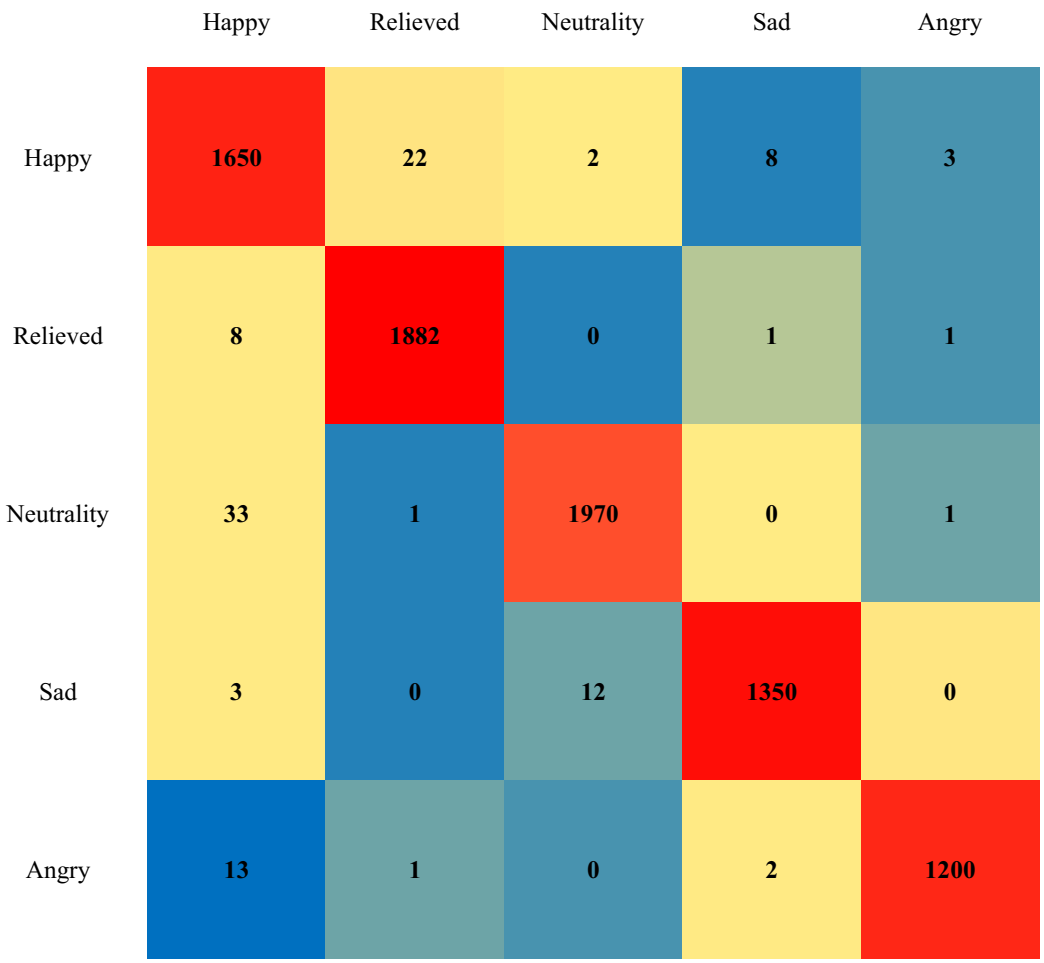

Figure 7 **Confusion matrix of DATL classification results.**
Full-size DOI: 10.7717/peerj-cs.3250/fig-7

in terms of running efficiency has been fully verified, which can not only quickly process the image sentiment classification task but also ensure the accuracy and reliability of the classification results while maintaining high efficiency.

Then, to conduct a more thorough assessment of DATL's performance on the image emotion classification task, we select several representative emotion categories, including happy, relieved, neutral, sad, and angry. These emotion categories encompass a wide range of human emotion dimensions, which is helpful for comprehensively examining the classification ability of DATL. To intuitively illustrate the classification effect of DATL, we utilize a confusion matrix, a powerful visualization tool, to present the classification results graphically. The specific results are shown in Fig. 7. By observing the confusion matrix, we can see that DATL exhibits extremely high accuracy in distinguishing these emotion categories, and the classification confusion between each emotion category is relatively low, indicating that DATL has excellent stability and consistency in distinguishing different emotional states.

Discussions

After the above series of carefully designed experiments, we have conducted a comprehensive and in-depth validation of the two core modules, AMSF and DAIS. The experimental results show that our method not only successfully classifies the sentiment of artistic images but also achieves a significant classification effect, fully demonstrating the effectiveness of DATL in practical applications.

In our experiments, we observe that the AMSF module can efficiently extract the salient features in art images, which are crucial for subsequent sentiment classification. By simulating the attention mechanism of human vision, AMSF can focus on the most emotionally expressive parts of the image, thereby improving the accuracy and pertinence of feature extraction. At the same time, the DAIS module utilizes domain adaptation technology to address the issue of data distribution differences effectively. Through transfer learning, DAIS can transfer sentiment classification knowledge from the source to the target, allowing the model to maintain stable performance across different artistic image datasets. This feature makes our method more flexible and general, allowing it to adapt to a variety of complex sentiment classification tasks. In addition, we also evaluate the overall performance of DATL, including aspects of running efficiency, robustness and scalability. The experiments conclude that DATL demonstrates strong performance in handling large-scale art image datasets, can complete the sentiment classification task in a short time, and exhibits good adaptability for different types of art images.

In summary, our method has achieved remarkable results in the task of art image sentiment classification, which benefits from the effective combination of two core modules of AMSF and DAIS, as well as the flexibility and versatility of the DATL framework. In the future, we will continue to optimize and improve this method, explore more application scenarios, and make greater contributions to the development of artificial intelligence and image processing technology.

CONCLUSION AND LIMITATION

To gain a deep understanding of the emotions conveyed in art images, we propose an artistic image sentiment classification method that combines domain adaptation and transfer learning. The core of this method lies in the construction of two modules. The first is an attention-based salient feature extraction method for art images, which accurately highlights the main body of artistic expression and effectively captures the key emotional elements in the image. The second is the image emotional classification upon domain adaptation, which utilizes the salient image content to achieve accurate emotional recognition of art image content. The experiments conclude that our method achieves a 92.4% mAP score and 98.9% accuracy in realistic and figurative artistic image sentiment classification, which fully demonstrates its effectiveness and reliability in practical applications. This result provides an accurate emotional interpretation of artworks in the fields of art appreciation and sentiment analysis.

The model may struggle to generalize across diverse artistic styles, particularly in abstract or unconventional art, and its predictions may not always align with human

emotional perceptions due to the subjectivity inherent in art. Its performance is also highly dependent on the training dataset, which may introduce bias if not diverse enough. Additionally, the method is computationally complex, making real-time processing challenging, and it may overfit specific datasets, reducing its effectiveness on unseen artworks. Furthermore, the model lacks interpretability, making it difficult to understand why a particular emotion is assigned to an image. Future work should focus on expanding dataset diversity, improving model interpretability, and optimizing computational efficiency for broader and more reliable applications.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Jingwen Wang conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Yisong Yang performed the experiments, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Kemal Polat analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The ART500K Dataset is available at: https://deepart.hkust.edu.hk/ART500K/art500k.html.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3250#supplemental-information.

REFERENCES

Agung ES, Rifai AP, Wijayanto T. 2024. Image-based facial emotion recognition using convolutional neural network on emognition dataset[J]. *Scientific reports* **14(1)**:14429 DOI 10.1038/s41598-024-65276-x.

Akrout B. 2025. Deep facial emotion recognition model using optimal feature extraction and dual-attention residual U-Net classifier. *Expert Systems* **42(1)**:e13314 DOI 10.1111/exsy.13314.

Bisogni C, Cascone L, Nappi M, Pero C. 2024. POSER: POsed vs spontaneous emotion recognition using fractal encoding. *Image and Vision Computing* **144(10)**:104952 DOI 10.1016/j.imavis.2024.104952.

Borgalli RA, Surve S. 2025. A hybrid optimized learning framework for compound facial emotion recognition. In: *International Conference on Cognitive Computing and Cyber Physical Systems*. Singapore: Springer, 447–459.

Chen H, Shao F, Mu B, Jiang Q. 2024. Image aesthetics assessment with emotion-aware multibranch network. *IEEE Transactions on Instrumentation and Measurement* **73**:1–15 DOI 10.1109/TIM.2024.3365174.

Deng S, Wu L, Shi G, Xing L, Jian M, Xiang Ye, Dong R. 2024. Learning to compose diversified prompts for image emotion classification. *Computational Visual Media* **10(6)**:1169–1183 DOI 10.1007/s41095-023-0389-6.

Deng Z, Zhu Q, He P, Zhang D, Luo Y. 2021. A saliency detection and gram matrix transform-based convolutional neural network for image emotion classification. *Security and Communication Networks* **2021(1)**:6854586 DOI 10.1155/2021/6854586.

Dixit C, Satapathy SM. 2024. Deep CNN with late fusion for real time multimodal emotion recognition. *Expert Systems with Applications* **240(1)**:122579 DOI 10.1016/j.eswa.2023.122579.

García-Domínguez M, Domínguez Cé, Heras Jó, Mata E, Pascual V. 2024. Deep style transfer to deal with the domain shift problem on spheroid segmentation. *Neurocomputing* **569(11)**:127105 DOI 10.1016/j.neucom.2023.127105.

Govarthan PK, Peddapalli SK, Ganapathy N, Ronickom JFA. 2024. Emotion classification using electrocardiogram and machine learning: a study on the effect of windowing techniques. *Expert Systems with Applications* **254(15)**:124371 DOI 10.1016/j.eswa.2024.124371.

Hong H, Zaheer W, Wali A. 2024. Visual sentiment analysis using data-augmented deep transfer learning techniques. *Multimedia Systems* **30(2)**:103 DOI 10.1007/s00530-024-01308-w.

Hosseini SS, Yamaghani MR, Poorzaker Arabani S. 2024. Multimodal modelling of human emotion using sound, image and text fusion. *Signal, Image and Video Processing* **18(1)**:71–79 DOI 10.1007/s11760-023-02707-8.

Khare SK, Blanes-Vidal V, Nadimi ES, Acharya UR. 2024. Emotion recognition and artificial intelligence: a systematic review (2014-2023) and research recommendations. *Information Fusion* **102(3)**:102019 DOI 10.1016/j.inffus.2023.102019.

Kulkarni D, Dixit VV. 2025. Hybrid classification model for emotion detection using electroencephalogram signal with improved feature set. *Biomedical Signal Processing and Control* **100(6)**:106893 DOI 10.1016/j.bspc.2024.106893.

Lee SE, Ryu C, Osanet PE. 2022. Object semantic attention network for visual sentiment analysis. *IEEE Transactions on Multimedia* **25**:7139–7148 DOI 10.1109/tmm.2022.3217414.

Li Z, Lu H, Zhao C, Feng L, Gu G, Chen W. 2023. Weakly supervised discriminate enhancement network for visual sentiment analysis. *Artificial Intelligence Review* **56(2)**:1763–1785 DOI 10.1007/s10462-022-10212-6.

Ou H, Qing C, Xu X, Jin J. 2021. Multi-level context pyramid network for visual sentiment analysis. *Sensors* **21(6)**:2136 DOI 10.3390/s21062136.

Pan J, Lu J, Wang S. 2024. A multi-stage visual perception approach for image emotion analysis. *IEEE Transactions on Affective Computing* **15(3)**:1786–1799 DOI 10.1109/taffc.2024.3372090.

Rao T, Li X, Xu M. 2020. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters* **51(3)**:2043–2061 DOI 10.1007/s11063-019-10033-9.

Sicilia A, Zhao X, Hwang SJ. 2023. Domain adversarial neural networks for domain generalization: when it works and how to improve. *Machine Learning* **112(7)**:2685–2721 DOI 10.1007/s10994-023-06324-x.

Silva N, Cardoso PJS, Rodrigues JMF. 2025. Sentiment classification model for landscapes. In: *International Conference on Human-Computer Interaction*. Cham: Springer, 375–393.

Tamil Priya D, Divya Udayan J. 2020. Transfer learning techniques for emotion classification on visual features of images in the deep learning network. *International Journal of Speech Technology* **23(2)**:361–372 DOI 10.1007/s10772-020-09707-w.

Wang X, Liu C, Hu M, Yao Y, Ren F. 2022. Feature discrimination and diversification for image emotion recognition. *Journal of Electronic Imaging* **31(2)**:023002 DOI 10.1117/1.JEI.31.2.023002.

Wu L, Qi M, Jian M, Zhang H. 2020. Visual sentiment analysis by combining global and local information. *Neural Processing Letters* **51(3)**:2063–2075 DOI 10.1007/s11063-019-10027-7.

Wu H, Zhou D, Guo Z, Song Z, Li Yu, Wei X, Zhou Q. 2025. A video-based cognitive emotion recognition method using an active learning algorithm based on complexity and uncertainty. *Applied Sciences* **15(1)**:462 DOI 10.3390/app15010462.

Xu Q, Wei Y, Yuan S, Wu J, Wang L, Wu C. 2024. Learning emotional prompt features with multiple views for visual emotion analysis. *Information Fusion* **108(1)**:102366 DOI 10.1016/j.inffus.2024.102366.

Xue L-Y, Mao Q-R, Huang X-H, Chen J. 2020. NLWSNet: a weakly supervised network for visual sentiment analysis in mislabeled web images. *Frontiers of Information Technology & Electronic Engineering* **21(9)**:1321–1333 DOI 10.1631/FITEE.1900618.

Yadav A, Vishwakarma DK. 2020. A deep learning architecture of RA-DLNet for visual sentiment analysis. *Multimedia Systems* **26(4)**:431–451 DOI 10.1007/s00530-020-00656-7.

Yang J, She D, Lai Y-K, Rosin PL, Yang M-H. 2018a. Weakly supervised coupled networks for visual sentiment analysis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7584–7592.

Yang J, She D, Sun M, Cheng M-M, Rosin PL, Wang L. 2018b. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia* **20(9)**:2513–2525 DOI 10.1109/TMM.2018.2803520.

Yao X, She D, Zhang H, Yang J, Cheng M-M, Wang L. 2020. Adaptive deep metric learning for affective image retrieval and classification. *IEEE Transactions on Multimedia* **23**:1640–1653 DOI 10.1109/tmm.2020.3001527.

Zhang H, Xu M. 2020. Weakly supervised emotion intensity prediction for recognition of emotions in images. *IEEE Transactions on Multimedia* **23**:2033–2044 DOI 10.1109/tmm.2020.3007352.

Zhang J, Qin Q, Liu X, Ye Qi, Du W. 2024a. Emotion-wise feature interaction analysis-based visual emotion distribution learning. *The Visual Computer* **40(3)**:1359–1368 DOI 10.1007/s00371-023-02854-6.

Zhang J, Zheng L, Wang M, Guo D. 2024b. Training a small emotional vision language model for visual art comprehension. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 397–413.

Zhang X, Zhou J, Qi G. 2025. Multimodal temporal context network for tracking dynamic changes in emotion. *The Journal of Supercomputing* **81(1)**:71 DOI 10.1007/s11227-024-06484-0.

Zhao X, Li X, Jiang R, Tang B. 2025. Decoupled cross-attribute correlation network for multimodal sentiment analysis. *Information Fusion* **117**:102897 DOI 10.1016/j.inffus.2024.102897.

Zhao Q, Xia Y, Long Y, Xu Ge, Wang J. 2025. Leveraging sensory knowledge into text-to-text transfer transformer for enhanced emotion analysis. *Information Processing & Management* **62(1)**:103876 DOI 10.1016/j.ipm.2024.103876.