


RESEARCH ARTICLE

Open Access



Mural classification model based on high- and low-level vision fusion

Jianfang Cao^{1,2*} , Hongyan Cui¹, Zibang Zhang¹ and Aidi Zhao¹

Abstract

The rapid classification of ancient murals is a pressing issue confronting scholars due to the rich content and information contained in images. Convolutional neural networks (CNNs) have been extensively applied in the field of computer vision because of their excellent classification performance. However, the network architecture of CNNs tends to be complex, which can lead to overfitting. To address the overfitting problem for CNNs, a classification model for ancient murals was developed in this study on the basis of a pretrained VGGNet model that integrates a depth migration model and simple low-level vision. First, we utilized a data enhancement algorithm to augment the original mural dataset. Then, transfer learning was applied to adapt a pretrained VGGNet model to the dataset, and this model was subsequently used to extract high-level visual features after readjustment. These extracted features were fused with the low-level features of the murals, such as color and texture, to form feature descriptors. Last, these descriptors were input into classifiers to obtain the final classification outcomes. The precision rate, recall rate and F1-score of the proposed model were found to be 80.64%, 78.06% and 78.63%, respectively, over the constructed mural dataset. Comparisons with AlexNet and a traditional backpropagation (BP) network illustrated the effectiveness of the proposed method for mural image classification. The generalization ability of the proposed method was proven through its application to different datasets. The algorithm proposed in this study comprehensively considers both the high- and low-level visual characteristics of murals, consistent with human vision.

Keywords: Vggnet model, Transfer learning, Mural classification, Feature fusion, Low-level features, SVM classifier

Introduction

Ancient Chinese murals have a long history and reflect the social and cultural characteristics of life at the time of their creation. Thus, they offer an important basis for understanding the development of human history, culture, art, science and technology and thus for further promoting the development of human civilization. Ancient Chinese murals constitute an indispensable class of ancient Chinese paintings [1]. Due to the numerous categories to which such murals may belong, the artificial classification of murals is onerous and time consuming, indirectly resulting in the slow development of mural research [2]. However, with the development and broad

application of digitization technology, an increasing number of ancient murals have gradually been digitized, which makes large-scale mural art analysis possible. On the one hand, the availability of a large number of digitized mural images provides researchers with abundant research data. On the other hand, these data present researchers with new questions regarding the effective use of such massive digital sources. Among these various questions, the question that is currently most important to resolve is how to use computers to analyze the elements contained in these mural images, based on which these images can then be effectively classified for further analysis, such as digital restoration, superresolution reconstruction or art value appraisal. Accordingly, such classification is expected to be of practical significance for art researchers conducting historical, anthropological and artistic investigations [3].

*Correspondence: caojianfangcn@163.com

¹ School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China

Full list of author information is available at the end of the article

Traditional computer-assisted mural classification methods utilize traditional classifiers based on low-level features extracted from murals [4]. Tang et al. [5] extracted contour features from mural images and used these features as similarity metrics between images. Later, Tang et al. [6] adopted scale-invariant feature transform (SIFT) features, combined with a support vector machine (SVM) classifier, to classify murals. Yang et al. [7] assessed the aesthetic visual style of murals in terms of composition, color and brightness attributes. Liu [8] extracted auspicious cloud elements, and Hao [9] obtained weighted characteristics such as colors, textures and contours of character images for mural classification. In the above methods, murals were classified using low-level features as proxies for high-level features for mural classification, and various approaches were adopted for mining the features of murals. Although these methods have some merits, they also have some limitations, as they are still unable to transcend the “semantic gap” and lack the ability to represent high-level semantics [10]. Nevertheless, on the basis of extracted low-level image features, the introduction of descriptors for the abstract representation of high-level features could enable the image features of murals to be comprehensively captured.

With the development of deep learning, deep neural networks have displayed a strong feature extraction ability, and end-to-end deep representations generated with such networks can effectively represent images at a more abstract level in addition to capturing their low-level information [11]. Among the available deep learning methods, convolutional neural networks (CNNs) have received widespread attention and are widely applied due to their good classification performance in the field of computer vision [12]. Shelhamer et al. [13] used AlexNet, the Visual Geometry Group network (VGG-Net) and GoogLeNet to construct fully convolutional networks for the semantic segmentation of images, thus greatly improving the semantic segmentation performance. To reduce economic losses caused by disease in the agricultural sector, Fuentes et al. [14] proposed a real-time deep-learning-based detector that could accurately and quickly detect tomato plant diseases and pest infestations. Ghazi et al. [15] used three popular deep learning architectures (i.e., GoogLeNet, AlexNet, and VGGNet) to identify plant species captured in photographs and assessed the different factors that influenced the performance of these networks. In the field of medicine, Lee et al. [16] developed an automatic deep feature classification (DFC) method for distinguishing benign angiomylipoma without visible fat (AMLwvf) from malignant clear cell renal cell carcinoma (ccRCC) using abdominal contrast-enhanced computed tomography (CECT) images, which further improved the quality of

the features used to distinguish AMLwvf and ccRCC in abdominal CECT images. However, in the mural domain, there is little literature on the application of CNNs for the classification of mural images, and related research is still in the nascent stage. Sun et al. [17] adopted four different algorithms (deep belief networks (DBNs), partial least squares regression (PLSR), principal component analysis with a support vector machine (PCA + SVM) and principal component analysis with an artificial neural network (PCA + ANN)) to classify the degree of flaking in the Mogao Grottoes based on the artificially labeled domains of interest. Later, Li et al. [18] proposed an unsupervised method of predicting the flaking degree for the Mogao Grottoes. However, due to the lack of supervision, the entire prediction process is complex. Wang et al. [19] adopted a two-layer CNN to extract the abstract features of murals and was able to classify ancient mural pigments. In a study by Caspari and Grespo [20], 3 convolution layers, 3 pooling layers and 2 fully connected (FC) layers were used to preliminarily analyze satellite images and promote research in the field of archeology. Li et al. [21] encoded the features of paintings by combining complex color shape descriptors with a 6-layer deep CNN for dynasty classification. Zou et al. [22] performed SVM classification by extracting SIFT and adjacent contour segment (kAS) feature descriptors to express shape features and adopted a depth confidence network to encode and refine the features. However, these classification efforts were still based on low-level CNNs. Therefore, the classification effects were not ideal. There are numerous ancient mural images; therefore, it is expensive to capture information from these images. On the other hand, the problem of overfitting is likely to occur when a deep CNN is adopted for the classification of only a limited number of digital murals. To address this challenge, transfer learning can be used for fine-tuning to adapt a pretrained model to different target domain, which greatly improves the classification efficiency of the resulting model. In recent years, transfer learning has seen a gradual increase in its scope of application and has become an effective means of possibly avoiding overfitting [23].

Therefore, in view of the current state of research as summarized above, in the study presented in this paper, we mainly investigated a means of comprehensively representing high- and low-level features of ancient mural images. First, we augmented the available samples of mural images and subjected a VGGNet model [24] to transfer learning. We extracted high-level mural features by fine-tuning the parameters of the model. Then, the extracted color histogram and local binary pattern (LBP) texture features were integrated to create feature descriptors capturing not only low-level color and

texture features but also the deep semantics of the mural images. Finally, these feature descriptors were input into an SVM classifier to automatically and efficiently classify the mural images.

The main contributions of this study include the following:

- (1) In this study, a pretrained classification model for natural images was adapted for application to ancient mural images through transfer learning. This approach not only makes use of the general characteristics of natural objects (because mural images are abstractly conceived depictions of natural objects) but also solves the problem of insufficient training data for mural image classification due to a small number of samples.
- (2) A feature extraction layer was designed based on feature fusion. Low-level features not only are used for the error analysis of high-level features but also are fused with the high-level features to form a feature descriptor that achieves high- and low-level visual fusion in a real sense to enrich the expression of mural features. This design is important to the classification task: it not only considers color and texture information that is of significance to the semantic expression of murals but also fully enables the extraction of semantic image features that are consistent with human vision.

Methods

Extraction of low-level feature descriptors

Color histogram and LBP features

Pigments, which are essential components of murals, describe the color layout of a whole image. Color features intuitively reflect the color of an image and are the simplest features to extract. In particular, the statistical properties of the color histogram can be used to directly count the number of pixels of each color type.

$$C(m) = \sum_{i=0}^W \sum_{j=0}^H \delta(I[i,j] = m), 1 \leq m \leq M \quad (1)$$

where $C(m)$ represents the number of pixels in the m th-grade color space, i represents a color grade in the color histogram, and $\delta(\cdot)$ indicates whether the color value at position (i,j) in the image is equivalent to the m th color grade (if yes, this function returns a value of 1; otherwise, it returns 0). Equation (2) presents the normalization process, in which the number of pixels of each color grade is normalized and then divided by the total number of image pixels N to obtain the final characteristic vector *Hist*:

$$Hist = \left(\frac{c(1)}{N}, \dots, \frac{c(i)}{N}, \dots, \frac{c(m)}{N} \right), 0 \leq i \leq m \quad (2)$$

In addition, mural images are painted by humans, and each element contains a unique texture design, i.e., the linear grain present in the painting. The texture of a mural is also referred to as its grain. Therefore, texture features can also be used as low-level features of a mural. Among the texture features that have been developed to date, LBP descriptors are simple and effective local feature descriptors for images. Equation (3) gives the formula for calculating LBP descriptors.

$$L(x_c, y_c) = \sum_{r=1}^N s(p(r) - p(c)) * 2^r \quad (3)$$

where (x_c, y_c) is the central point of a domain and $p(c)$ represents the pixel value at that point. The circle with this point as its center and a radius of R is denoted by O . There are N points on the circle. Accordingly, r represents the r th pixel, $p(r)$ represents the pixel value of the r th pixel among the N points, and the definition of $s(\cdot)$ is given in Eq. (4). $s(\cdot)$ is a signifier function that sets the pixel value to 1 when the pixel value of a surrounding pixel is larger than that of the center pixel value and to 0 otherwise. Finally, an N -number signifier sequence containing only values of 0 and 1 is obtained.

$$s(x) = \begin{cases} 1, & |x| \geq 0, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

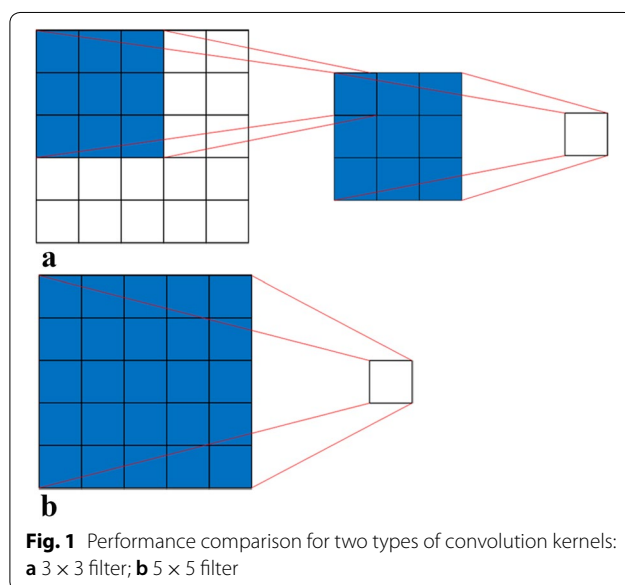
VGGNet model

VGGNet is an improvement over AlexNet; these models won first and second place, respectively, in the location and classification competitions of the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Of the VGGNet models, VGG-16 is the most commonly used network. The structure of VGG-16 is shown in Table 1, where the column titled 'Layer' indicates the network layer, the 'Size' column gives the size information for the image in each layer, the 'Filter' column specifies the convolution kernel, the 'Stride' column gives the step length, and the 'Dropout' column gives the probability of the random dropout of neurons. The number of categories is 1000.

As shown in Table 1, VGGNet is composed of 13 convolutional layers and 3 FC layers, among which the convolution kernels are all based on small 3×3 kernels. Compared with a large 5×5 convolution kernel, this design has the advantage that two stacked 3×3 convolutional layers having a field of vision equivalent to that of a 5×5 convolution kernel but fewer parameters. As shown in Fig. 1, under the assumption that the numbers of input

Table 1 VGGNet model

Layer	Size	Filter	Stride	Dropout
Input	224 × 224 × 3			
Conv1	224 × 224 × 64	3 × 3 × 64 × 2	1	
MaxPool1	112 × 112 × 64	2 × 2	2	
Conv2	112 × 112 × 128	3 × 3 × 128 × 2	1	
MaxPool2	56 × 56 × 128	2 × 2	2	
Conv3	56 × 56 × 256	3 × 3 × 256 × 3	1	
MaxPool3	28 × 28 × 256	2 × 2	2	
Conv4	28 × 28 × 512	3 × 3 × 512 × 3	1	
MaxPool4	14 × 14 × 512	2 × 2	2	
Conv5	14 × 14 × 512	3 × 3 × 512 × 3	1	
MaxPool5	7 × 7 × 512	2 × 2	2	
Fc6	4096			0.5
Fc7	4096			0.5
Softmax	1000			



and output channels are both 1, the size of the convolution kernel depicted in Fig. 1a is $3 \times 3 \times 2 = 18$, while that in Fig. 1b is $5 \times 5 = 25$. Thus, the merit of small 3×3 convolution kernels over a large 5×5 convolution kernel is evident. Increasing the number of network layers indirectly enriches the linear representation capabilities of the network, which is an implicit regular representation. The VGGNet model has good application prospects for transfer learning because of its features, i.e., deep layers and small convolution kernels.

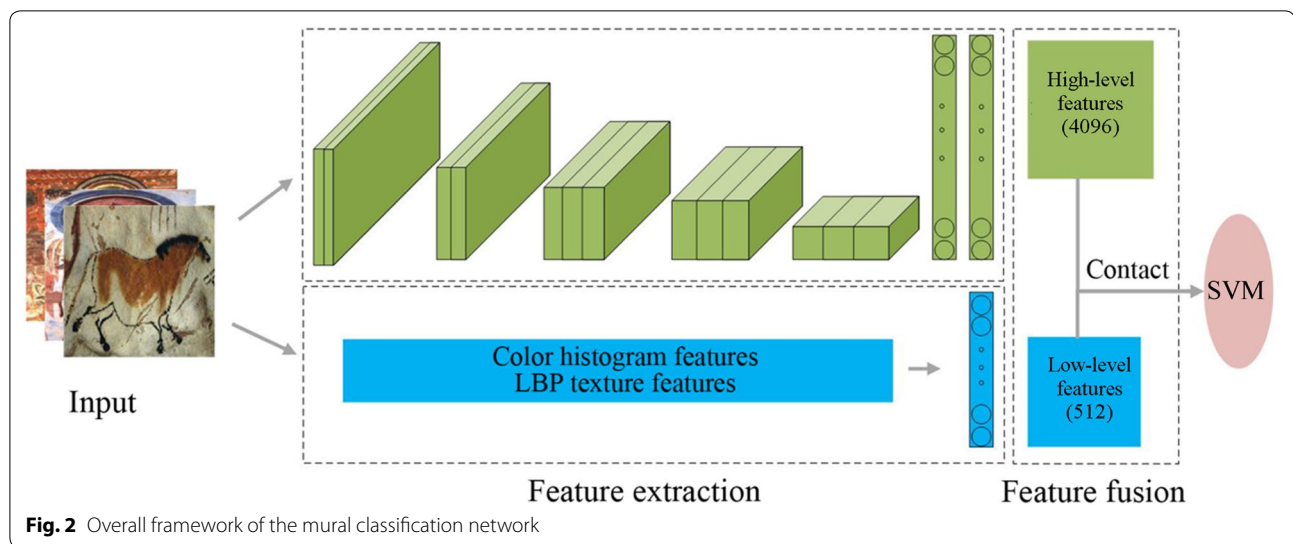
However, due to the large network depth of the VGGNet model, it is prone to overfitting when the sample size is small. Transfer learning can be implemented to

greatly mitigate this dilemma. In the transfer learning process, the parameters of a pretrained VGGNet model obtained through general training on a large-scale dataset, such as ImageNet, are fine-tuned on the knowledge acquired from the small target sample set. Finally, the fine-tuned model can be used to solve the classification problem for the application of interest. Based on this approach, in this study, we modified the VGGNet model through transfer learning to obtain high-level features of murals.

Mural classification based on the improved VGGNet model Network framework for mural classification

The quality of mural image extraction is a critical factor affecting the classification performance. The extraction of the low-level features of an image is simple and usually fast, yet these features are unable to provide a high-level semantic representation of the image. High-level features, defined in relation to low-level features, refer to features that can represent the image semantics to some extent. However, the gradient diffusion phenomenon becomes more evident in deeper network layers, which may cause the loss of some low-level features. Therefore, to integrate the advantages of human vision in each layer, we decided to combine low-level features and high-level features to better represent the information contained in mural images. First, the pretrained VGG-16 network was used as the basic network to extract high-level features. Next, color features and LBP texture features were extracted as low-level features. Then, these high-level and low-level features were fused to form the feature descriptor for mural classification. Finally, we developed a classification network for mural images based on an improved VGGNet model that integrates transfer learning and low-level features. The structure of the network is shown in Fig. 2.

As shown in Fig. 2, first, features closely associated with the global information of the input mural image were obtained by fine-tuning the FC layers such that the 4096-dimensional output of the second FC layer would represent high-level features. Second, to ensure the fairness of the high-level features in expressing the features of the mural and avoid breaking the linear relation between the low- and high-level features, the high- and low-level features were simply concatenated to obtain the final feature descriptor of the mural. Because the low-level features were obtained as an eigenvector after normalization, they did not need to be normalized again during fusion. The network (referred to as TFNet) developed in this study is composed of three parts: the first performs high-level feature extraction, the second generates the low-level feature description, and the third applies the fusion process. The mural images input into



the model have dimensions of 224×224 . The different parts of the network are detailed as follows:

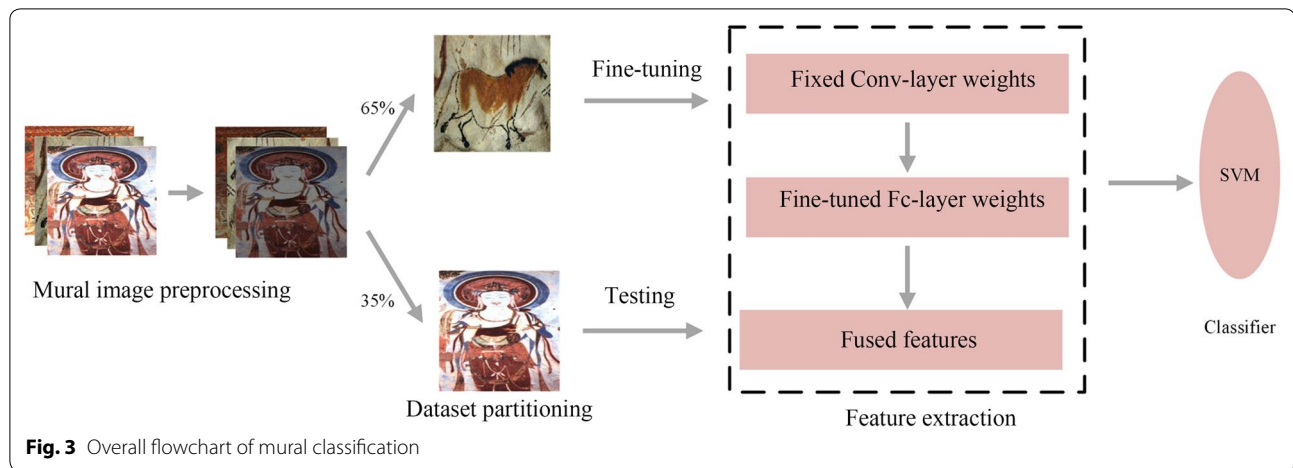
- (1) *High-level feature extraction.* The high-level features were extracted as the 4096-dimensional features obtained from the last FC layer of VGG-16 by inputting mural images into the pretrained network and fine-tuning the weights of the two FC layers while leaving the weights of the five convolution blocks unchanged (there are 13 convolutional layers in total, where the size of each convolution kernel is 3×3 ; the numbers of small convolution blocks contained in each main convolution block are 2, 2, 3, 3, and 3, in sequence, and the numbers of convolution kernels for each convolution are 64, 128, 256, 512, and 512, respectively).
- (2) *Low-level feature extraction.* The color histogram and LBP texture histogram were used to obtain 512-dimensional low-level features.
- (3) *Feature fusion.* The features at both levels were concatenated to form a 4608-dimensional feature descriptor for mural classification.
- (4) *Other details.* Based on references [15, 24, 25], three learning rates, i.e., 0.01, 0.001 and 0.0001, were set for use with the stochastic gradient descent (SGD) optimizer and the Adam optimizer on the training set (six experiments in total). Then, the models trained with each combination of a particular learning rate and a particular optimizer were tested on the test data, and the results were compared. Therefore, on the basis of the dataset used in this study, we concluded that the classification effect of the VGGNet model was optimal when the Adam optimizer was employed and the learning rate was set

to 0.001. The classifier was chosen to be an SVM classifier created using the LIBSVM toolkit with the default parameters. An SVM is a binary model. However, eight classes were considered in the developed classification model for mural images. Generally, there are three possible methods of constructing classifiers, i.e., one-to-one, one-to-multiple, and hierarchical. In this study, the one-to-one construction approach was selected in the LIBSVM toolkit, and in accordance with this method, a classifier was constructed between each pair of classes, resulting in a total of $k(k-1)/2$ classifiers. For the classification of unknown data, the data obtained based on these $k(k-1)/2$ classifiers were analyzed, and the results that appeared most frequently were taken as the final outcomes. Based on comparisons among a linear function kernel, a polynomial function kernel and a Gaussian radial basis function (RBF) kernel, the Gaussian kernel was chosen as the kernel function in this study. After the kernel function was determined, the gamma value was set to 0.01, 0.001 and $1/(\text{number of features})$ [26] to conduct experiments. The best effect was obtained when the gamma value was set to $1/(\text{number of features})$; thus, the kernel parameters were determined with this setting.

Description of the algorithms for mural classification

The flowchart for the classification of mural images using TFNet, as shown in Fig. 3, comprises the following algorithms.

Algorithm 1. Preprocessing.



Input: the sample set *Inputset*; the number of samples *N* with a division ratio of *r*;

Output: the training set *Trainset* and the test set *Testset*.

- (1) *DataSetEnhance(Inputset)*;
- (2) For each image contained in *Inputset*.
- (3) *Dataset* \leftarrow *getEnhancedSet(Inputset)*; /*add the image data in *Inputset* into *Dataset**/
- (4) End for.
- (5) *Trainset, Testset* \leftarrow *getPartitionedSet(Dataset, N, r)*; /*add $N \cdot r$ images in *Dataset* into *Trainset* and add the remaining $N \cdot (1-r)$ images into *Testset**/

Algorithm 2. Feature extraction.

Input: the sample set *Sample*;

Output: the image features *Fset*.

- (1) *FeatureExtraction(Sample)*;
- (2) For each image contained in *Sample*
- (3) *Lset* \leftarrow *getLowFeatures(Sample)*; /*extract the low-level features of the *Sample* images into the set *Lset**/
- (4) *Hset* \leftarrow *getHighFeatures(Sample)*; /*extract the high-level features of the *Sample* images into the set *Hset**/
- (5) *Fset* \leftarrow *getFeatures(Lset, Hset)*; /*fuse *Lset* with *Hset* and add the results into *Fset* as the feature descriptor for the current image*/
- (6) End for

Algorithm 3. Mural training.

Input: the training set *Trainset*;

Output: the classification model *Model*.

- (1) *ModelTrain(Trainset)*;

- (2) *TrainF* \leftarrow *FeatureExtraction(Trainset)*; /*add the image features in *Trainset* into *TrainF**/

- (3) *Model* \leftarrow *getFinetunedModel(Fset, Pre)*; /*add the fine-tuned pretrained model *Pre* into *Model**/

Algorithm 4. Mural testing.

Input: the test dataset *Testset*, the classification model *Model*;

Output: classification precision rate *Prec*.

- (1) *ModelTest(Testset)*;
- (2) *TestF* \leftarrow *FeatureExtraction(Testset)* /*add the image features in *Testset* into *TestF**/
- (3) *Prec* \leftarrow *Model(TestF)*; /*add the prediction outcomes of the model into *Prec**/

Experimental environment and experimental design

The experiments were performed on a PC running Windows 10 with an Intel Core i7-8750H CPU, a GTX 1070 GPU, 8 GB of memory and the Python-based TensorFlow deep learning framework.

The raw materials in this study were all obtained from artistic images scanned from the Tomb Murals of the Chinese Silk Road and the Complete Collection of China's Dunhuang Grotto Murals, spanning the Han, Tang, Sui, and Ming dynasties. These murals reflect a variety of subjects, such as stories, apparel and sutras, and exhibit diverse images and styles. Typical images of secular people, plants, bodhisattvas, animals, buildings, auspicious clouds, disciples of Buddha and Buddha himself were selected for the experiments. After severely damaged images were excluded, the numbers of images remaining in each of the above categories were 675, 227, 277, 263, 153, 122, 174 and 153, respectively. In total, 65% of the images were randomly selected from among the

various types of samples for use as the training set, and the remaining 35% of the images were used as the test set (see Table 2; the sample distribution shown in this table was used in all experiments in this study, except for sampling with replacement using the bootstrap method). However, the deep model adopted in this study was prone to overfitting when the sample size was small. Therefore, dataset augmentation was performed on the training set, including (1) changes to the image brightness; (2) transformation of the images into horizontal and vertical mirror images; and (3) background noise enhancement, mainly with Gaussian noise and salt-and-pepper noise. After these manipulations, the number of samples in the training set was increased by sixfold. Examples of the images after augmentation are shown in Fig. 4.

Three indicators (i.e., the precision rate, recall rate and F1-score) that are commonly used in image classification were used for the comprehensive performance evaluation of the proposed algorithms. The specific definitions of each indicator are as follows.

$$Mural_P = \frac{TP}{TP + FP} \quad (5)$$

$$Mural_R = \frac{TP}{TP + FN} \quad (6)$$

$$Mural_F1 = \frac{2 \times Mural_P \times Mural_R}{Mural_P + Mural_R} \quad (7)$$

where TP represents the number of true positives (correct predictions), TN represents the number of true negatives (correct predictions), FP represents the number of false positives (incorrect predictions), FN represents the number of false negatives (incorrect predictions), and $Mural_P$, $Mural_R$ and $Mural_F1$ represent the precision rate, recall rate and F1-score, respectively, of mural classification.

Table 2 Numbers of murals selected for the experiments

Category	Training set	Test set
Buddha	99	54
Bodhisattva	180	97
Buddhist disciple	113	61
Secular person	438	237
Animal	170	93
Plant	147	80
Building	99	54
Auspicious cloud	78	44
Total	1325	720

Results and discussion

Comparison with different selected features

To validate the effectiveness of the low-level feature-fusion method proposed in this study, we performed comparisons in terms of the precision rate, recall rate and F1-score with a method involving only the 512-dimensional low-level features and a method involving only the 4096-dimensional high-level features, and the results are shown in Fig. 5.

Compared with the methods involving only the low-level features and only the high-level features, the fusion method proposed in this study yielded increases in the precision rate, recall rate and F1-score by 34.9, 34.74 and 34.03% and by 9.36, 8.72 and 8.33%, respectively. The reason for these improvements is that the fusion of low- and high-level features enables the extracted features not only to capture the main features of murals, i.e., color and texture, but also to mine the deep semantics implied in the images. Accordingly, our fused descriptor can more completely express the features of murals, thereby helping achieve better classification performance.

Comparison with traditional CNNs

For this study, the traditional AlexNet [27], VGGNet, GoogLeNet [28] and ResNet [29] transfer learning models were selected for comparison and to analyze the change in the precision rate in various experiments. In addition, the precision rate of the VGGNet-RCC model, which was obtained by replacing the color histogram used in this study with the state-of-the-art regional color co-occurrence (RCC) feature descriptor [30], was also considered. According to this model, an image is first segmented into nonintersecting regions. Then, a color co-occurrence matrix is constructed for each adjacent region, and finally, the constructed matrixes are summed and standardized to obtain features. In this experiment, the size of the codebook was set to 128. Table 3 compares the results in terms of the precision rate, recall rate, and F1-score, and Fig. 6 compares the precision rate curves showing the evolution with the number of iterations. The x-coordinate represents the number of iterations, and the y-coordinate is the precision rate. The red, blue, pink, green and navy lines represent the curves for the fine-tuned VGGNet, AlexNet, GoogLeNet, ResNet and VGGNet-RCC models, respectively.

As shown in Table 3, the model proposed in this study achieved the highest average precision rate, recall rate and F1-score of 80.64, 78.06 and 78.63%, respectively. The proposed model showed 5.46, 4, 6.99, 6.49 and 20.31% improvements in the precision rate compared with the AlexNet-F, VGGNet-F, GoogLeNet, ResNet-F, and VGGNet-RCC models, respectively. The main reason



Fig. 4 Examples of mural images

for this improvement lies in the fact that the high-level features of the VGGNet model are the most suitable for representing the mural images considered in this study. A greater or lesser depth is not conducive to full representation. As shown in Fig. 6, the model proposed in this study not only achieved the highest precision rate but also was the fastest to converge. Moreover, the proposed model was more stable than the VGGNet model, which had the highest precision rate among the transfer learning models considered for comparison, because it integrates both low- and high-level mural features to better simulate the feature representation of human vision. In addition, the use of an SVM classifier also contributed to

the increased stability of the model given the relatively small size of the image set used in this study. The overall precision rate of the VGGNet-RCC model was the lowest among the tested models. Although this model achieved the highest precision rate for murals in the Buddhist disciple and building categories, its precision rate for murals depicting secular people was the lowest. The reason for this result might be that the VGGNet-RCC model, which considers color spatiality, is suitable for extracting the features of single objects, such as those in the Buddhist disciple and building categories, but its adaptability for multiple-element categories, such as the secular person category, is poor.

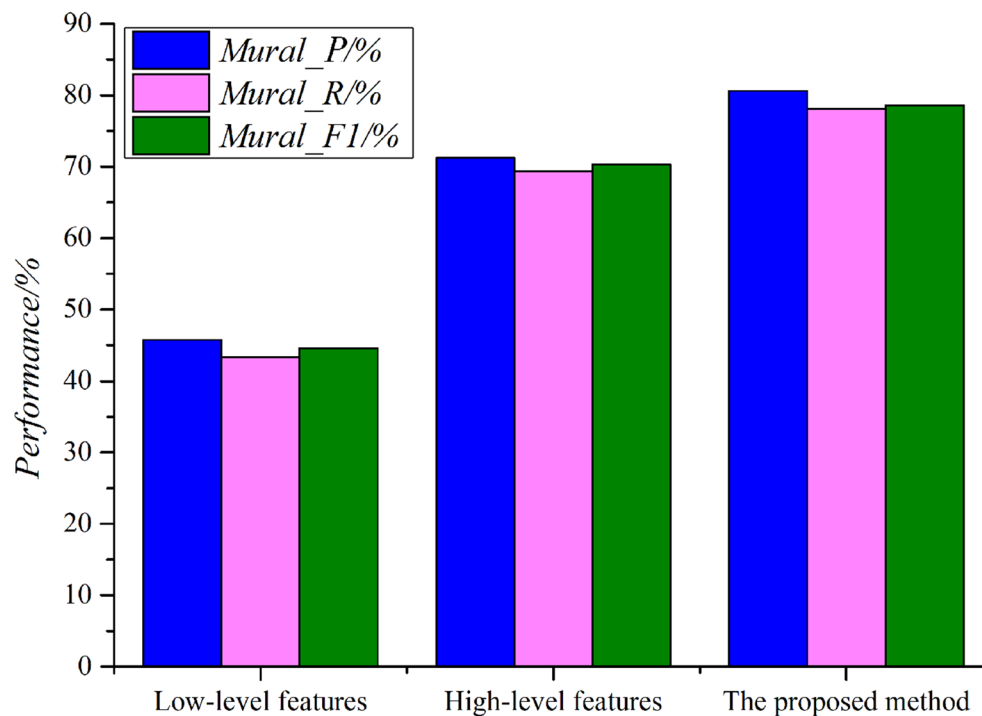


Fig. 5 Performance comparison among methods involving different selected features

Table 3 Comparisons with traditional algorithms in terms of the precision rate, recall rate and F1-score

Network	Precision rate/%	Recall rate/%	F1-score/%
TFNet	<i>80.64</i>	<i>78.06</i>	<i>78.63</i>
AlexNet-F	75.18	74.65	74.91
VGGNet-F	76.64	76.59	76.61
GoogLeNet-F	73.65	72.23	72.93
ResNet-F	74.51	74.42	74.46
VGGNet-RCC	60.33	58.71	59.51

The maximum value of precision rate, recall rate and F1-score in all models are in *italics*

Comparison with other improved CNNs

The proposed model was also compared with the models developed by Mehdipour [15], P. Cheng [25], Lee [16], G. Cheng [31] and Lin [32]. In the study by Mehdipour [15], both the training set and test set were augmented, after which fine-tuned VGGNet and GoogLeNet models were used as the optimal classifiers for plant classification. In the study by P. Cheng [25], the FC layers of CaffeNet and VGGNet were retrained on the training set after pretraining on the 2012 ILSVRC dataset. Lee et al. [16] proposed an automatic high-level feature classification method based on the extraction of 64-dimensional texture features, including histogram

and texture matrixes, and 7-dimensional shape features, including roundness and curvature; this method was used in combination with random forest classification. In the work of G. Cheng, the bag of convolutional features (BoCF) method was used to generate visual words from deep convolutional features, and off-the-shelf CNNs were used for classification [31]. Lin's model was developed based on a new deep neural network architecture in which color features were added to the first FC layer of a 5-layer CNN for multilabel image annotation [32]. Table 4 compares the results of these works in terms of the precision rates, recall rates, and F1-scores in various experiments, and Fig. 7 compares the precision rates, recall rates, and F1-scores for mural classification.

As shown in Table 4, the model proposed in this study achieved 2.22, 6.78, 3, 8.61 and 4.33% improvements in the maximum precision rate over the Mehdipour, P. Cheng, Lee, G. Cheng and Lin models, respectively. Moreover, although the precision rate, recall rate and F1-score of the proposed model were low for the Buddhist disciple category, it achieved high precision rates, recall rates, and F1-scores for most categories. As seen from this table in combination with Fig. 7, the precision rate, recall rate and F1-score of the model proposed in this study for the classification of mural images generally reached the highest values. Collectively, the above data

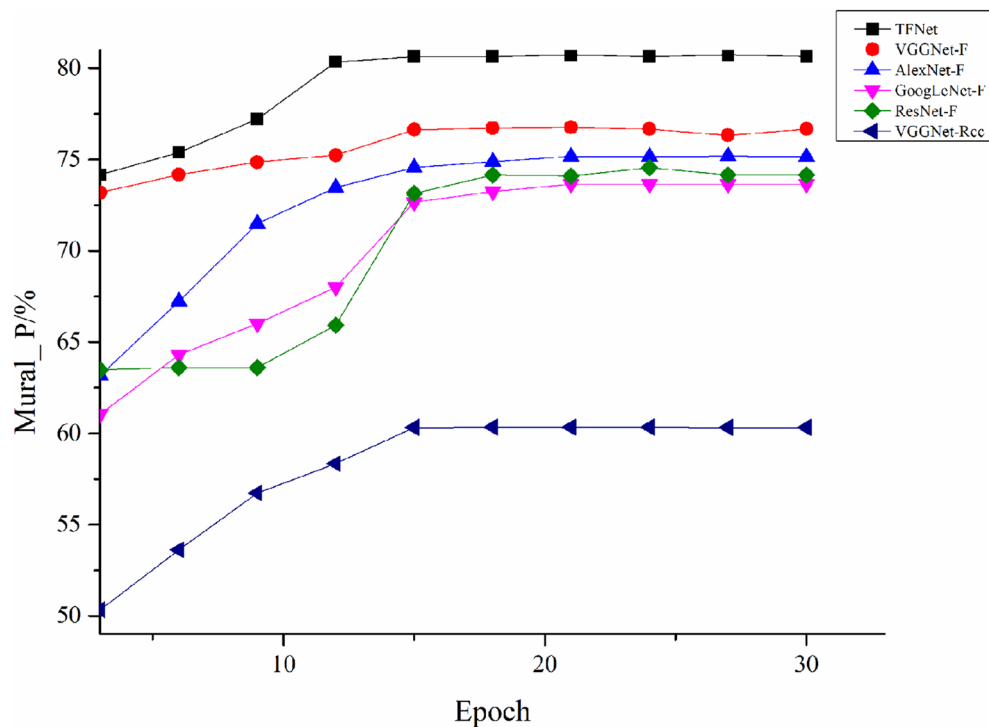


Fig. 6 Curves showing the evolution of the precision rate of each algorithm

sufficiently indicate the effectiveness of the proposed model for mural classification, mainly by virtue of the adoption of transfer learning to avoid overfitting when extracting high-level features. Moreover, the mural information can be fully represented by combining high- and low-level features, making the algorithm proposed in this study more stable and more robust than other models.

Comparison with traditional methods

For the mural dataset constructed in this study, the traditional SIFT feature descriptor, the histogram of ordered gradients (HOG) feature descriptor and a traditional backpropagation (BP) neural network were selected for comparative experiments. The color and texture features used in this study were also extracted using these traditional methods. A comparison of the corresponding results is summarized in Table 5, in which CL denotes the fused vector of both types of features. A bar graph comparing the classification performance of the different methods is shown in Fig. 8.

As shown in Table 5 and Fig. 8, although the size of the dataset used in this study was relatively small, the model developed in this study achieved a much higher identification rate than the traditional methods did, mainly because the features extracted using these traditional methods are low-level features extracted by human

beings, making it difficult for them to truly represent the image semantics and resulting in a low generalization ability. Furthermore, the overall classification precision rates of the traditional methods were low. The primary reasons for these results include the following: 1) the numbers of samples in the different categories in the test dataset were imbalanced, and 2) augmentation was applied to the dataset as a whole, rather than specifically for rare samples, which might present certain challenges for the traditional methods. Nevertheless, compared with the single features extracted via traditional methods, a CNN not only can directly and automatically learn features from the image pixels in both shallower and deeper layers but also can consider neighboring pixel values within the receptive field. Therefore, the classification performance of the model proposed in this study is much better than that of conventional methods. Furthermore, the proposed model also fuses high- and low-level mural features. As a result, the representation capability of our final descriptor is stronger than that of low-level features alone. The adoption of transfer learning also avoids the risk of overfitting arising from training the model directly and thus is beneficial for the classification of murals.

Table 4 Classification results of various models

Category	Index	Mehdipour	P. Cheng	Lee	G. Cheng	Lin	TFNet
Buddha	Precision	59.62	50.00	59.62	53.70	61.54	72.22
	Recall rate	57.39	47.62	58.17	51.95	59.22	70.18
	F1-score	58.48	48.78	58.89	52.81	60.36	71.19
Bodhisattva	Precision	78.43	74.23	73.53	73.20	74.51	81.44
	Recall rate	77.10	72.98	73.02	71.49	72.16	80.04
	F1-score	77.76	73.60	73.27	72.33	73.32	80.73
Disciple	Precision	51.67	47.54	50.00	44.26	50.00	45.90
	Recall rate	49.09	46.56	49.12	42.10	48.55	45.01
	F1-score	50.35	47.04	49.56	43.15	49.26	45.45
Secular person	Precision	90.57	86.08	89.06	87.76	87.55	91.14
	Recall rate	88.72	85.43	87.09	85.48	84.99	89.57
	F1-score	89.64	85.75	88.06	86.60	86.25	90.35
Animal	Precision	64.52	67.74	67.74	56.99	66.67	70.97
	Recall rate	63.06	64.28	66.53	56.04	64.30	69.15
	F1-score	63.78	65.96	67.13	56.51	65.46	70.05
Plant	Precision	72.15	51.25	65.82	57.50	67.09	61.25
	Recall rate	70.18	50.25	63.89	57.01	66.44	60.98
	F1-score	71.15	50.75	64.84	57.25	66.76	61.11
Building	Precision	69.09	74.07	72.73	68.52	69.09	83.33
	Recall rate	68.22	72.55	71.02	67.77	68.05	82.83
	F1-score	68.65	73.30	71.86	68.14	68.57	83.08
Auspicious cloud	Precision	84.78	84.09	93.48	72.73	84.78	90.91
	Recall rate	82.51	82.11	91.45	71.93	84.00	90.04
	F1-score	83.63	83.09	92.45	72.33	84.39	90.47
Average	Precision	78.42	73.86	77.64	72.03	76.31	80.64
	Recall rate	76.60	71.25	75.80	69.86	74.73	78.06
	F1-score	77.05	71.91	76.28	70.39	75.08	78.63

The maximum value of precision rate, recall rate and F1-score for each class of all models are in *italics*

Effectiveness of the proposed method

The use of different datasets is a persuasive approach for validating the adaptability of a model. We used two painting datasets, i.e., PT [33] and DH660 [34], to verify the effectiveness and adaptability of the proposed method.

PT consists of PT91 and PT13. The PT91 dataset contains 4266 images contributed by 91 artists [33], with 31–56 works per artist. Because these images belong to various categories and each category consists of a relatively small number of images, classification is difficult. In our experiment, 2275 images were used to constitute the training set, with the remaining 1991 images constituting the test set. Figure 9 shows some examples from the PT91 dataset. The PT13 dataset contains 2338 images associated with 13 artistic styles. In our experiment, 1250 images were used for training, with the remaining 1088 images used for testing. Some examples from this dataset are shown in Fig. 10.

The DH660 dataset contains 660 images of flying apsaras created during three different periods [34], with an average of 220 images from each period. In this study, half of the images were used for training, and the other half were used for testing. Figure 11 shows some examples of these images.

These datasets are different in scale and therefore required different training times. To maintain consistency in this study, we performed 30 rounds of training for each dataset. A comparison of the results is summarized in Table 6.

As shown in Table 6, the precision rates of the method proposed in this study on PT91, PT13 and DH660 were 60.01, 66.93, and 96.56%, respectively, being 6.91, 4.73 and 5.32% higher than those reported in the literature (53.10% and 62.2% for PT91 and PT13, respectively [30], and 91.24% for DH660 [34]). As seen from this table in combination with the results shown in Fig. 12, the proposed model exhibited excellent

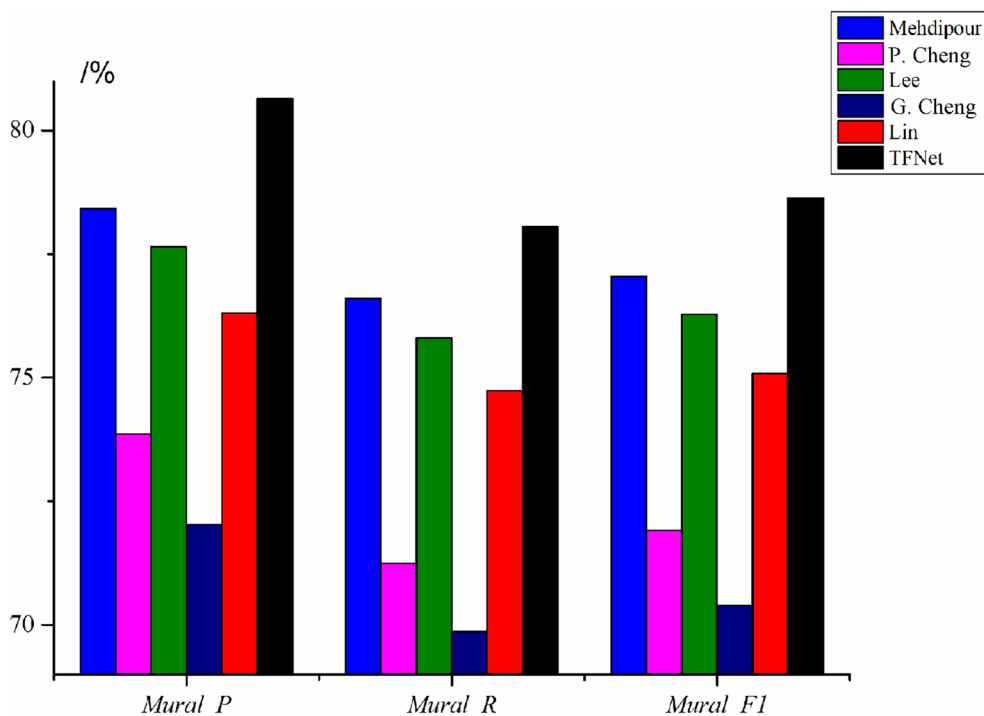


Fig. 7 Performance comparison of various classification models

classification performance on these different datasets. On the one hand, the fusion of the low- and high-level features is suitable for expressing the characteristics of the mural dataset used in this study (noticeably, the proposed model outperformed the models presented in [33] and [34]); on the other hand, the application of the low-level features in the error propagation for the high-level features in this study enables high integration of high- and low-level visual characteristics to fully simulate the human visual system. Accordingly, these results validate the effectiveness and further prove the adaptability of our model.

Robustness of the proposed model

To verify the robustness of the proposed model, we used the bootstrap method to perform idealized sampling with replacement. Usually, this method is used to evaluate the generalization errors of a model. Traditional bootstrapping requires a large number of sampling instances and targets traditional machine learning with a small sample size. By contrast, deep learning methods often

use large volumes of data, and the models are complex. Furthermore, deep learning requires extensive time and resources; specifically, the time required for one iteration is approximately 1 h. Therefore, a large number of sampling instances is not feasible for deep learning. In this study, the number of sampling instances was set to 10, considering the limitations of both resources and time. Equation (9) is the bootstrapping formula for calculating the precision rate of this model under an infinite number of sampling instances:

$$FinalP = \frac{1}{b} \sum_{i=1}^b (0.632 * TrueP + 0.368 * TotalP) \quad (9)$$

where $TrueP$ represents the precision rate on the test set, $TotalP$ represents the precision rate on all samples, and b is the number of sampling instances (equal to 10 here). Table 7 shows the outcomes based on 10 sampling instances (the training set was not subjected to data augmentation in this experiment).

As shown in Table 7, at a confidence level of 0.95, the bootstrap confidence interval is (82.02, 82.73), which

Table 5 Classification performance of the model proposed in this study compared with traditional algorithms

Category	Index	BP	HOG	SIFT	LBP	COLOR	CL	TFNet
Buddha	Precision	56.44	39.66	47.14	51.61	42.65	62.00	72.22
	Recall rate	55.23	37.19	46.52	48.49	40.92	61.06	70.18
	F1-score	55.83	38.39	46.83	50.00	41.77	61.53	71.19
Bodhisattva	Precision	56.52	37.61	43.87	49.78	39.37	37.25	81.44
	Recall rate	55.01	35.62	41.99	48.03	37.67	35.08	80.04
	F1-score	55.75	36.59	42.91	48.89	38.50	36.13	80.73
Disciple	Precision	50.69	27.63	31.93	42.98	40.57	38.20	45.90
	Recall rate	48.69	27.00	30.46	40.91	37.68	37.01	45.01
	F1-score	49.67	27.31	31.18	41.92	39.07	37.60	45.45
Secular person	Precision	63.16	25.45	42.90	25.97	39.80	50.90	91.14
	Recall rate	61.54	23.69	42.11	23.92	37.56	49.09	89.57
	F1-score	62.34	24.54	42.50	24.90	38.65	49.98	90.35
Animal	Precision	49.31	47.35	26.36	45.63	37.08	40.49	70.97
	Recall rate	48.92	45.66	25.10	43.99	35.72	38.88	69.15
	F1-score	49.11	46.49	25.71	44.79	36.39	39.67	70.05
Plant	Precision	60.57	34.68	30.88	47.01	29.25	40.23	61.25
	Recall rate	58.46	32.17	28.55	46.03	28.11	38.93	60.98
	F1-score	59.50	33.38	29.67	46.51	28.67	39.57	61.11
Building	Precision	56.31	11.65	35.81	42.78	31.80	71.11	83.33
	Recall rate	55.22	10.01	33.56	39.86	29.53	67.99	82.83
	F1-score	55.76	10.77	34.65	41.27	30.62	69.52	83.08
Auspicious cloud	Precision	41.41	43.10	29.17	47.26	27.81	30.25	90.91
	Recall rate	40.03	42.14	27.60	45.62	26.59	28.96	90.04
	F1-score	40.71	42.61	28.36	46.43	27.19	29.59	90.47
Average	Precision	54.61	33.68	35.96	43.66	36.04	45.74	80.64
	Recall rate	53.29	31.84	34.06	41.59	35.13	43.52	78.06
	F1-score	53.94	32.73	34.98	42.60	35.58	44.60	78.63

The maximum value of precision rate, recall rate and F1-score for each class of all models are in *italics*

CL: fused vector of color and texture features

indicates that the results obtained based on the method proposed in this study deviate from the precision rate in the ideal state. This gap is primarily due to the imbalance in the dataset used in this study. The volume of data in the Buddha category represented approximately 1/3 of the total volume of eight categories, which led to insufficient learning for the other categories, thereby decreasing the precision rate. Methods of addressing such data imbalance to further improve the performance of the model may be an important direction for future studies.

Conclusions

This study proposes the idea of applying transfer learning and feature fusion in the classification of mural images to solve the overfitting problem that can easily occur when training deep models on small samples. First, due to the small size of the mural dataset used in this study, data augmentation was performed to expand the dataset. Next, high- and low-level image features were extracted to construct a joint feature descriptor for each mural, thereby enriching the representation of the mural features. Then, transfer learning was adopted to fine-tune a pretrained VGGNet model on the mural dataset, making the model more suitable for extracting mural features. Moreover, we effectively solved the overfitting problem

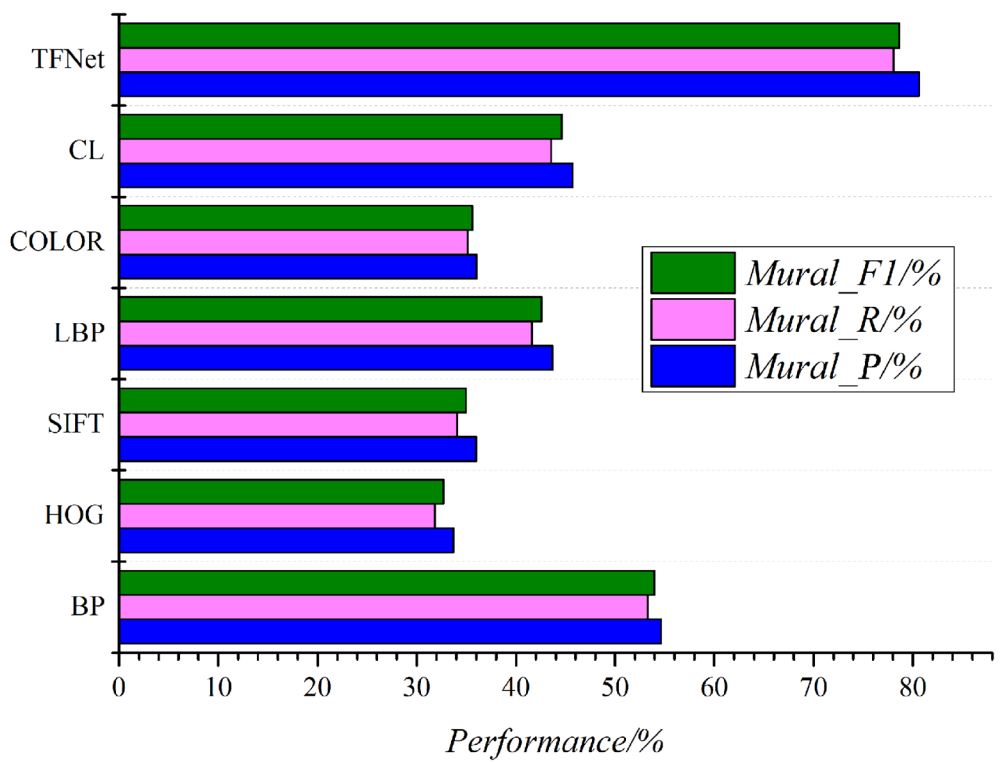


Fig. 8 Classification performance of the model developed in this study compared with traditional algorithms

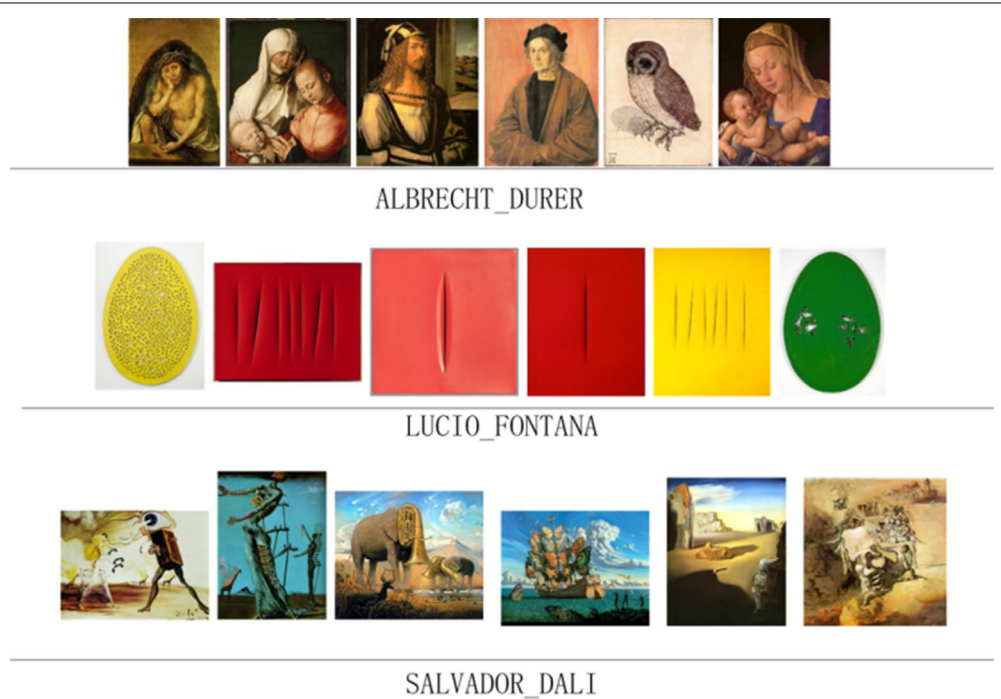


Fig. 9 Some example images from the PT91 dataset



Fig. 10 Some example images from the PT13 dataset



Fig. 11 Some example images from the DH660 dataset

that is prone to occur in deep models trained on small samples to improve the generalization ability of the model. Finally, the precision rate, recall rate and F1-score of the model proposed in this study reached 80.64, 78.06 and 78.63%, respectively, through suitable adjustment of the parameters. The classification model proposed in

this study achieved a higher recognition rate than other classification models and significantly improved classification performance, thus validating its effectiveness and motivating follow-up research.

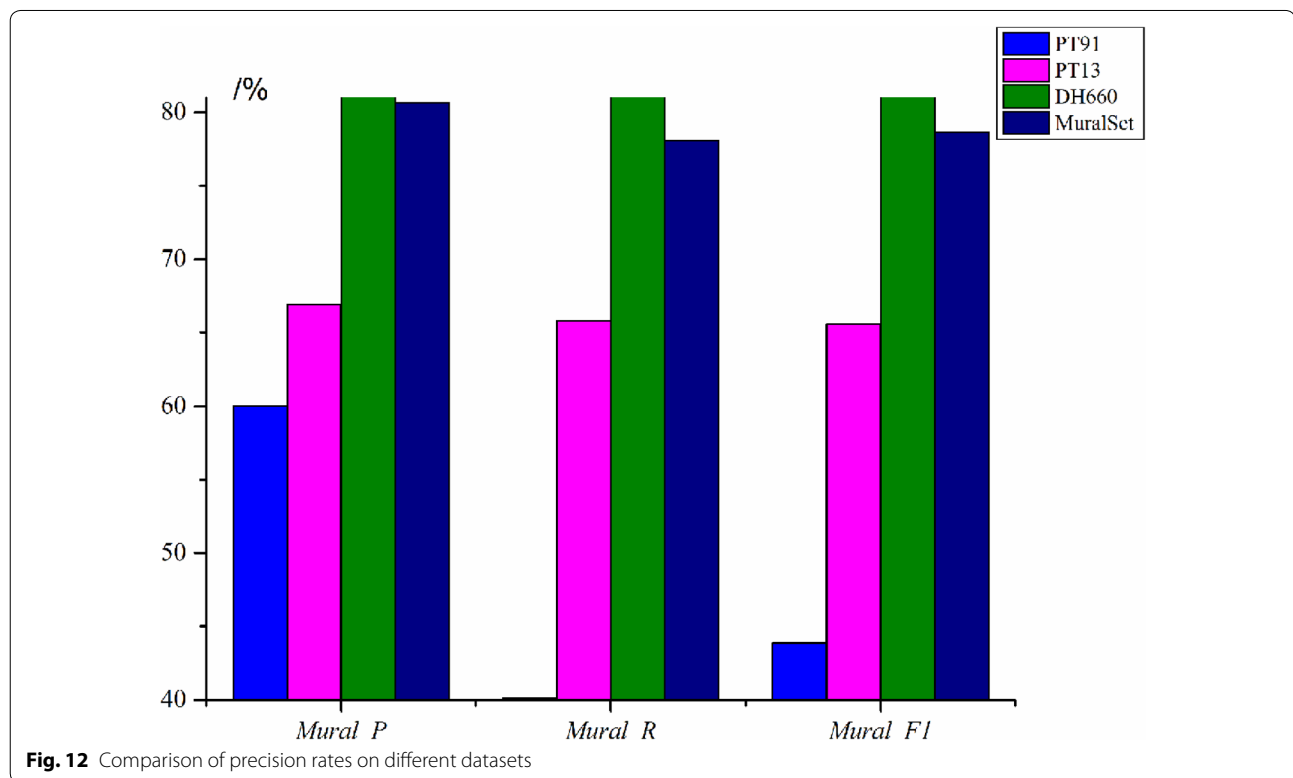
However, there are still some limitations of this study. Due to the large number of parameters of the

Table 6 Precision rates, recall rates and F1-scores of the proposed method on different datasets

Dataset	Precision	Recall	F1-score
PT91	60.01	40.12	43.88
PT13	66.93	65.81	65.57
DH660	96.56	96.56	96.55
MuralSet	80.64	78.06	78.63

VGGNet model, the computation is onerous. Therefore, in future studies, first, the network structure can be optimized by reducing the number of model parameters to find a rapid and effective classification model for mural images. Second, during transfer learning, if the transfer learning technique is incorrectly applied or a proper validation strategy is not enacted, the resulting model cannot achieve a satisfactory effect;

accordingly, multiple challenges remain in the application of transfer learning. The negative transfer phenomenon can reduce the recognition rate of a model rather than increasing its capacity. For instance, in the cross-field transfer learning performed in this study, the size of the target dataset and that of the source dataset differed greatly. Errors in the transfer learning strategy might still result in overfitting, resulting in a decreased precision rate. Therefore, determining how to better combine transfer learning with the model proposed in this study remains a challenge. Third, the numbers of dimensions of the low- and high-level mural image features were large, which influenced the training time of the proposed network. Therefore, effective dimension compression methods should be considered in the future.

**Fig. 12** Comparison of precision rates on different datasets**Table 7 Classification precision results for the proposed model based on 10 sampling instances according to the bootstrap method**

	1	2	3	4	5	6	7	8	9	10	Average
FinalP	82.18	81.18	82.49	82.99	82.54	80.77	82.34	81.79	83.26	82.12	82.14

Abbreviations

SIFT: Scale-invariant feature transform; SVM: Support vector machine; VGGNet: Visual Geometry Group network; DFC: Deep feature classification; CECT: Contrast-enhanced computed tomography.

Acknowledgements

None.

Authors' contributions

All authors contributed to the current work. JFC devised the study plan, led the writing of the article and supervised the entire process. HYC and ZBZ conducted the experiments and collected the data, and ADZ performed the analyses. All authors read and approved the final manuscript.

Funding

This work was supported by the Natural Science Foundation of Shanxi Province [grant number 201701D21059], a Project of Key Basic Research in Humanities and Social Sciences of Shanxi Colleges and Universities [grant number 20190130], an Art and Science Planning Project of Shanxi Province [grant number 2017F06], the Platform and Personnel Specialty of Xinzhou [grant number 20180601], the Teaching Reform Innovation Project of Xinzhou Teachers University [grant number JGZD202004], a Teaching Reform Innovation Project of Colleges and Universities in Shanxi Province [grant number J2019168], and an Education Science Planning Project of the 13th 5 year Plan of the Key Discipline Project of Shanxi Province [grant number GH-17059].

Availability of data and materials

All data used for analysis in this study are included within the article.

Competing interests

The authors declare that they have no competing interests.

Author details

¹ School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China. ² Department of Computer Science and Technology, Xinzhou Teachers University, No. 10 Peace West Street, Xinzhou 034000, China.

Received: 22 July 2020 Accepted: 6 November 2020

Published online: 24 November 2020

References

- Jiang SQ, Huang QM, Ye QX, Gao W. An effective method to detect and categorize digitized traditional Chinese paintings. *Pattern Recogn Lett*. 2006;27:734–46.
- Sun MJ, Zhang D, Wang Z, et al. Monte carlo convex hull model for classification of traditional Chinese paintings. *Neurocomputing*. 2016;171:788–97.
- Li XY, Zhuang YT, Pan YH. The technique and system of content-based image retrieval. *J Comput Res Dev*. 2001;38:344–54.
- Huang KQ, Ren WQ, Tan TN. A review on image object classification and detection. *Chin J Comput*. 2014;36:1225–40.
- Tang DW, Lu DM, Yang B, Xu DQ. Similarity metrics between mural images with constraints of the overall structures of contours. *J Image Graph*. 2013;18:968–75.
- Tang DW, Lu DM, Xu DQ, Yang B. Clustered multiple instance learning for mural image classification. *J Image Graph*. 2014;19:708–15.
- Yang B, Xu RQ, Tang DW, Yang X, Zhao L. Aesthetic visual style assessment on Dunhuang Murals. *J Shanghai Jiaotong Uni (Sci)*. 2014;19:28–34.
- Liu XJ. Research on feature extraction and evolution patterns of auspicious cloud in the Dunhuang Grotto Murals. Dissertation. Wuhan University of Technology, Wuhan, 2014.
- Hao YB. Research and implementation on classification algorithm with people of ancient Chinese murals based on style characteristics. Dissertation. Tianjin University, Tianjin, 2016.
- Lejbølle AR, Nasrollahi K, Moeslund TB. Enhancing person re-identification by late fusion of low-, mid- and high-level features. *IET Biomet*. 2018;7:125–35.
- Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–44.
- Rawat W, Wang ZH. Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput*. 2017;29:2352–449.
- Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2016. <https://doi.org/10.1109/TPAMI.2016.2572683>.
- Fuentes A, Yoon S, Kim SC, Park DS. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*. 2017;17:1–21.
- Mehdipour Ghazi M, Yanikoglu B, Aptoula E. Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing*. 2017;235:228–35.
- Lee H, Hong H, Kim J, Jung DC. Deep feature classification of angiomyolipoma without visible fat and renal cell carcinoma in abdominal contrast-enhanced CT images with texture image patches and hand-crafted feature concatenation. *Med Phys*. 2018;45:1550–61.
- Sun MJ, Zhang D, Wang Z, Ren JC, Chai BL, Sun JZ. What's wrong with murals at Mogao Grottoes: a near-infrared hyperspectral image method. *Sci Rep*. 2015;5:14371.
- Li P, Sun MJ, Wang Z, Chai BL. OPTICS-based unsupervised method for flaking degree evaluation on the murals in Mogao Grottoes. *Sci Rep*. 2018;8:15954.
- Wang YN, Zhu DN, Wang HQ, Wang K. Multi-spectral image classification of mural pigments based on CNN [J/OL]. *Laser & Optoelectronics Progress*, 1–16. <http://kns.cnki.net/kcms/detail/31.1690.TN.20190521.1045.008.html>. Accessed 23 Aug 2019.
- Caspari G, Grespo P. Convolutional neural networks for archaeological site detection-Finding "princely" tombs. *J Archaeol Sci*. 2019;104998(1–104998):9.
- Li QQ, Zou Q, Ma D, Wang Q, Wang S. Dating ancient paintings of Mogao Grottoes using deeply learnt visual codes. *Sci China Inf Sci*. 2018;61:092105.
- Zou Q, Cao Y, Li QQ, Huang CH, Wang S. Chronological classification of ancient paintings using appearance and shape features. *Pattern Recogn Lett*. 2014;49:146–54.
- Zhuang FZ, Luo P, He Q, Shi ZZ. Survey on transfer learning research. *J Softw*. 2015;26:26–39.
- Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[EB/OL]. (2015-04-10). <https://arxiv.org/pdf/1409.1556.pdf>. Accessed 23 Aug 2019.
- Cheng PM, Malhi HS. Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. *J Digit Imaging*. 2016;30:234–43.
- Chen T, Ju S, Ren F, Fan M, Gu Y. EEG emotion recognition model based on the LIBSVM classifier. *Measurement*. 2020. <https://doi.org/10.1016/j.measurement.2020.108047>.
- Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*. New York, NY: Curran Associates; 2012. p. 1097–105. <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>.
- Szegedy C, Vanhoucke V, Ioffe S, Shelens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE; 2016. p. 2818–26. [arXiv:1512.00567](https://arxiv.org/abs/1512.00567).
- He KM, Zhang XY, Ren SQ, Sun J. Deep Residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE; 2016. p. 770–8. [arXiv:1512.03385](https://arxiv.org/abs/1512.03385).
- Zou Q, Qi X, Li Q, Wang S. Discriminative regional color co-occurrence descriptor. In: *Proceeding of the IEEE international conference on image processing*. Piscataway: IEEE; 2015. p. 696–700. <https://cse.sc.edu/~songwang/document/icip15a.pdf>.
- Cheng G, Li Z, Yao X, Guo L, Li KM. Remote sensing image scene classification using bag of convolutional features. *IEEE Geosci Remote Sens Lett*. 2017;14:1735–9.
- Lin Y, Zhang HG. Automatic image annotation via combining low-level colour feature with features learned from convolutional neural networks. *NeuroQuantology*. 2018;16:679–85.
- Khan SF, Beigpour S, Weijer J, Felsberg M. Painting-91: a large scale database for computational painting categorization. *Mach Vis Appl*. 2014;25:1385–97.

34. Zou Q, Ni LH, Hu ZW, Li QQ, Wang S. Local pattern collocations using regional co-occurrence factorization. *IEEE Trans Multimedia*. 2017;19:492–505.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
