scientific reports

Check for updates

OPEN

Deep and shallow feature fusion framework for remote sensing open pit coal mine scene recognition

Yang Liu & Jin Zhang✉

Understanding land use and damage in open-pit coal mining areas is crucial for effective scientific oversight and management. Current recognition methods exhibit limitations: traditional approaches depend on manually designed features, which offer limited expressiveness, whereas deep learning techniques are heavily reliant on sample data. In order to overcome the aforementioned limitations, a three-branch feature extraction framework was proposed in the present study. The proposed framework effectively fuses deep features (DF) and shallow features (SF), and can accomplish scene recognition tasks with high accuracy and fewer samples. Deep features are enhanced through a neighbouring feature attention module and a Graph Convolutional Network (GCN) module, which capture both neighbouring features and the correlation between local scene information. Shallow features are extracted using the Gray-Level Co-occurrence Matrix (GLCM) and Gabor filters, which respectively capture local and overall texture variations. Evaluation results on the AID and RSSCN7 datasets demonstrate that the proposed deep feature extraction model achieved classification accuracies of 97.53% and 96.73%, respectively, indicating superior performance in deep feature extraction tasks. Finally, the two kinds of features were fused and input into the particle swarm algorithm optimised support vector machine (PSO-SVM) to classify the scenes of remote sensing images, and the classification accuracy reached 92.78%, outperforming four other classification methods.

Keywords Feature fusion, Graph convolutional network (GCN), Remote sensing (RS), Scene classification

Mineral resources are critical for human survival and development. Open-pit coal mining represents a distinct type of land use, characterised by significant disruption of the surface environment during the mining process. This disturbance leads to substantial alterations in the original landscape pattern[1]. Large-scale, prolonged, and unregulated open-pit coal mining causes serious damage and destruction to the land and ecological environment[2–4]. Assessing land use and damage in open-pit coal mining areas is crucial for effective scientific supervision and management[5]. High-resolution RS images encompass complex information and structures. Leveraging this detailed information to accurately extract and identify characteristics of open-pit coal mines can serve dual purposes: it provides a basis for decision-making in the event of sudden disasters and supports effective land management and reclamation planning of the mining area.

The methods of extracting features from RS scene images through supervised learning can roughly be divided into two categories: methods based on manual features and methods based on deep learning. The methods based on manual features[6–9] mainly focus on using a large number of engineering skills and domain expertise to design various ergonomic features, such as texture, spectrum, geometry, colour features or their combinations. These features represent the primary attributes of scene images and thus carry valuable information for scene classification[10]. A feature extraction method incorporating principal component analysis (PCA) and local binary pattern (LBP) was developed for hyperspectral images. R2FD2[11] combined a multi-channel autocorrelation strategy with log-Gabor wavelets in the feature detection stage, so as to detect points of interest with high repeatability and uniform distribution. LBP was used in previous research[12] to extract image local features such as edges and corners. These LBP features, along with global Gabor features and raw spectral features, were fused at both the feature level and decision level. While traditional manual features offer good stability and are effective in capturing overall shallow information, making them suitable for low-resolution RS image scene classification,

School of Mining Engineering, Taiyuan University of Technology, Shanxi, Taiyuan, China. ✉email: zjgps@163.com

they heavily depend on manual design. This reliance limits their effectiveness in extracting relevant feature information from high-resolution RS images, thus reducing their applicability in classification tasks.

Deep learning methods train a large amount of sample data through complex algorithms to learn the intrinsic laws and expression levels of the samples. This approach enables machines to develop analytical learning capabilities comparable to those of humans, leading to significant advancements in text, speech, and image recognition[13–15]. An increasing number of scholars have applied deep learning-based methods[16,17] to RS scene classification. As one of the representative algorithms of deep learning, CNN can mine the rich semantic information in RS scene images to obtain better performance[18–20]. CNN-based methods emphasise local perception but often overlook the long-range relationships between distant regions.

In recent years, various graph-based approaches have gained attention. GCNs leverage graph representation learning to integrate and analyse structural relationships and informational features between different nodes, enabling the understanding and reasoning of complex data. DAGCN[21] designed deep graph convolutional networks (DeepGCNs) to extract relationships between features within the HIS data. In addition, miniGCN[22] allows training of large-scale GCNs in small batches. Notably, miniGCNs can infer out-of-sample data without the need to retrain the network, thereby enhancing classification performance. In one study[23], a denoising matrix and an enhanced adjacency matrix (DREA) were designed to improve the model's focus on salient objects when extracting local features.

To address the limitations of existing methods, this paper introduces a framework called DS-NET for fusing deep features (DF) and shallow features (SF). Shallow features, including local and global texture characteristics, are extracted using the grey-level co-occurrence matrix (GLCM)[6] and Gabor filters[7]. Deep features are obtained through a key feature extraction module and a contextual feature extraction module. The key feature extraction module emphasises the relationships between neighbouring scale feature layers, focusing on small-scale features and integrating key regions of attention from these features into larger-scale features. Ultimately, all features are fused to enhance the overall performance. In the contextual feature extraction module, the GCN[24] model is used to mine the correlation of features within the scene. The deep and shallow features are fused and then subjected to PCA to reduce redundant information. The resulting reduced features are then input into a PSO-SVM[25,26] for classification, facilitating the recognition of mining scenes.

The main contributions of the present study can be summarised as follows:

(1) A three-branch feature extraction framework combining deep features and shallow features was designed;
(2) The combination of GLCM and Gabor was used to simultaneously extract the local variation features of the texture and the overall texture variation features;
(3) A two-branch deep feature extraction module was constructed, and the deep feature mainly consists of two modules: a key feature extraction module and a contextual feature extraction module. The key feature extraction module consists of multi-level feature extraction and an attention mechanism that highlights shallow information. The new attention mechanism captures the relationship between neighbouring features and adds the key information from the highlighted shallow features to the final feature layer. The contextual feature extraction module introduces the graph convolutional network (GCN) model to effectively reveal the correlation between local information in the scene to obtain finer features.

Related works

Gabor

Gabor filtering is a feature extraction method based on signal processing, where features that are not easily separable in the image space are analyzed at different frequencies by transforming the image space to the frequency domain, independently each feature. The expression of 2D Gabor wavelet kernel function is shown in (1):

$$\psi\ (\vec{x}) = \frac{\left\|\vec{k}\right\|^2}{\sigma^2} exp\left(\frac{-\left\|\vec{k}\right\|^2 \|\vec{x}\|^2}{2\sigma^2}\right)\left[exp\left(i\vec{k}\vec{x}\right) - exp\left(-\frac{\sigma^2}{2}\right)\right] \tag{1}$$

$$\overrightarrow{k} = \begin{pmatrix} k_x \\ k_y \end{pmatrix} = \begin{pmatrix} k_v \cos\phi_u \\ k_v \sin\phi_u \end{pmatrix}. \tag{2}$$

$\overrightarrow{x}$ represents the image coordinates at a given location, $\overrightarrow{k}$ is the center frequency of the filter, ϕ_u is the sampling direction, u is the direction of the gabor kernel, v is the scale of the gabor kernel, and σ denotes the radius of the Gaussian function. In Eq. (2) $k_v = \frac{k_{max}}{f^v} = 2^{-\frac{v+2}{2}}\pi$ and $\phi_u = u\pi/8$ denotes the Gaussian envelope function that constrains the plane.

Currently, Gabor can not only be used directly for feature extraction of images, but also combined with CNN to extract image feature information. The paper[27] proposed an effective image representation method based on the Gabor filtered completion local binary pattern (GCLBP) for land use scene classification. The paper[28] proposes an effective texture classification method that combines multi-resolution global and local Gabor features in pyramid space. F3DGF[29] explores phase-induced 3D Gabor based features to better utilize the two parts of Gabor features. This paper[30] proposes a three-dimensional (3D) Gabor convolutional network constructed from 3D Gabor filters that can be learned.

GLCM

GLCM is a feature extraction method that describes the spatial characteristics of image gray values. It summarizes the relationship between the gray values of the target image in different directions and angles, and determines the relevant matrix functions. The GLCM method has been verified by a large number of experiments to have a significant effect on image texture features. It can reflect the angle and step information of the input gray-scale two-dimensional image on the neighborhood spacing, direction, pixel mutation and other aspects. However, due to the large dimension of the gray-scale co-occurrence matrix of the original image, the information is complex. Therefore, a method is needed to simplify it, that is, to extract some of its feature quantities to describe the feature information of the original image, such as contrast, angular second moment, entropy and homogeneity. The formulas of the four descriptors are shown in (3)–(6).

(1) Contrast

$$Con = \sum_i \sum_j (i-j)^2 p(i,j). \tag{3}$$

(2) Angular Second Moment

$$Asm = \sum_i \sum_j p(i,j)^2. \tag{4}$$

(3) Entropy

$$Ent = -\sum_i \sum_j p(i,j)\, logp(i,j). \tag{5}$$

(4) Homogeneity

$$Hom = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j), \tag{6}$$

where $p(i,j)$ denotes the image is normalized after GLCM feature extraction.

Paper[31] establishes a GLCM-based texture orientation estimation method for remote sensing images. Paper[32] extracted GLCM-based features from base gray-scale images collected by UAVs to classify different types of crops. Paper[33] combines GLCM features and Rotation Invariant Uniform Local Binary Pattern (RIULBP) to achieve cloud detection.

Traditional manual features rely on specific objects, and these models are not generic and flexible enough.

CNN

There exist four key ideas for CNNs to exploit the properties of natural signals, namely local connectivity, shared weights, pooling, and multilayer usage. With the increase in computer performance, AlexNet[34] was proposed in 2012 and is considered to be the originator of deep learning image classification. After that more classical networks were proposed, such as VGGNet[35], GoogLeNet[36] and ResNet[37], which greatly improve the image classification accuracy. Paper[38] proposes a network that incorporates residual structures to extract depth details from images for detecting small multi-scale targets in complex scenes. Paper[39] constructed compact bilinear CNN models by improving the bilinear pooling method. This paper[40] introduces Li group machine learning into CNN modeling and proposes a new network model, Li group region influence network (LGRIN). Paper[41] proposed a multi-branch deep learning framework that effectively combines global contextual features with multi-scale features to recognize complex land scenes.

CNN-based methods have been better applied to RS scene categorization, but their poor ability to characterize the correlation between objects in a scene may lead to suboptimal performance in the classification task.

GCN

In 2017, Kipf et al. used GCN to solve the semi-supervised classification problem, and GCN delicately designed a method to extract features from graph data. So far, GCN has gained wide attention. In this paper[42], a feature fusion method based on deep residual GCN is proposed to study the relationship between HSI data. The adaptive cross-attention-driven spatial-spectral graph convolutional network (ACSS-GCN)[43] consists of a spatial GCN (Sa-GCN) subnetwork, a spectral GCN (Se-GCN) subnetwork, and a graph cross-attention fusion module (GCAFM). Spatial Sa-GCN and spectral features Se-GCN are fused together to reduce the noise effect. The novel two-branch deeper GCN (TBDGCN)[44] combines the advantages of hyperpixel-based GCN and pixel-based CNN to extract both hyperpixel-level and pixel-level features of HSI. A Differential Scale Restricted GCN (DSR-GCN)[45] for HSI classification is better able to model spatial structures with reliable and fine-grained maps and can capture more discriminative features in small sample learning (FSL) scenarios. The paper[46] proposes a CNN-GCN two-branch network for scene classification of high-resolution remote sensing images, which introduces contextual features that are easily neglected.

CNN essentially computes a weighted sum of local pixels to achieve spatial feature extraction without considering the structural relationships between image pixels, whereas GCN can make good use of the graph structural information to model the relationship between nodes through the adjacency matrix, and it is more specific in its feature extraction, which means that GCN can capture the global information of the graph, and has a great scope of application in the field of computer vision.

Proposed method

The proposed framework comprises two main components: the shallow feature extraction module and the deep feature extraction module. In the shallow feature extraction module, the GLCM is first used to measure the change amplitude and interconnections between pixel points, focusing on local texture variations. Subsequently, Gabor filters are applied to capture texture information across different frequencies, scales, and orientations, providing a robust description of overall texture changes. By integrating the local features obtained from GLCM with the global texture information extracted using Gabor filters, the framework constructs a comprehensive shallow texture feature that combines both local and global texture features. The deep feature extraction framework employs VGG-16 as the backbone model and is organised into two branches. In the first branch, a GCN is used to extract global features from the input image. The second branch focuses on extracting critical local information, complementing the global features obtained by the GCN. The contextual feature extraction module references the GCN model to explore the correlation between local information of the scene to obtain finer features. The key feature extraction module consists of multi-level feature extraction and an attention mechanism that highlights shallow information. The new spatial channel attention module captures the relationship between neighbouring features and adds the key information from the highlighted shallow features to the final feature layer. To facilitate feature mapping and to create connections between CNN and PSO-SVM algorithms, the abstract features obtained from the last layer are flattened into one-dimensional feature vectors. Finally, the extracted features from all modules are fused and the final recognition result is obtained by PSO-SVM. The overall framework of this network is shown in Fig. 1.

Two branch network architecture

Key feature extraction module

In the key feature extraction branch, VGG16 is utilised as the feature extractor. The VGG16 model comprises 13 convolutional layers and 3 fully connected layers, with approximately 138 million trainable parameters. Its core principle involves enhancing the network's depth by stacking multiple small-sized convolutional kernels and pooling layers, which improves the representation of image features. The model uses relatively small 3×3 convolutional kernels and 2×2 max pooling kernels, with each convolutional layer followed by a ReLU activation function. The structure of VGG16 is both straightforward and classical, making it a widely used benchmark model in deep learning. It excels in image classification tasks, effectively recognising and distinguishing between different object classes. Its simplicity and scalability make VGG16 a popular choice for transfer learning, where it serves as a foundational model that can be adapted for various computer vision applications.

VGG16 consists of five convolutional layers, three fully-connected layers, and SoftMax output layer, the layers are separated using max-pooling, and the activation units of all the hidden layers use the ReLU function. The five convolutional layers have 64 channels, 128 channels, 256 and two 512 channels. After the convolutional layers, the data is flattened into vectors using the Flatten function. These vectors then pass through three fully connected layers, each activated by ReLU, and the final predictions are produced using a SoftMax activation function.

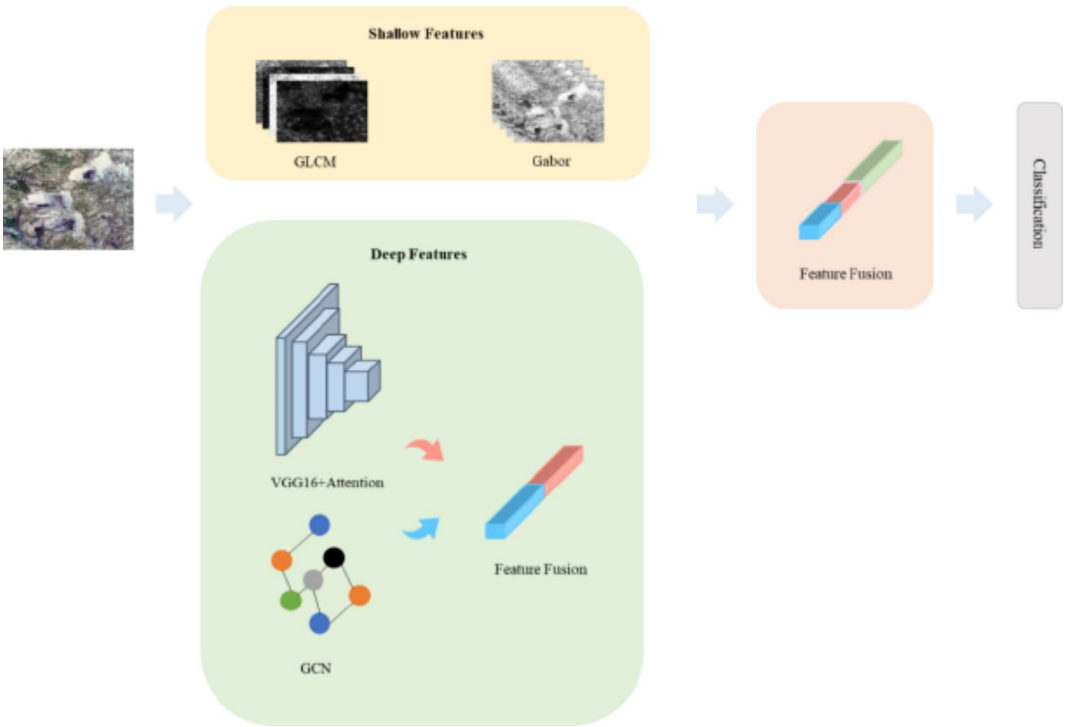

Fig. 1. DS-NET Overall Architecture.

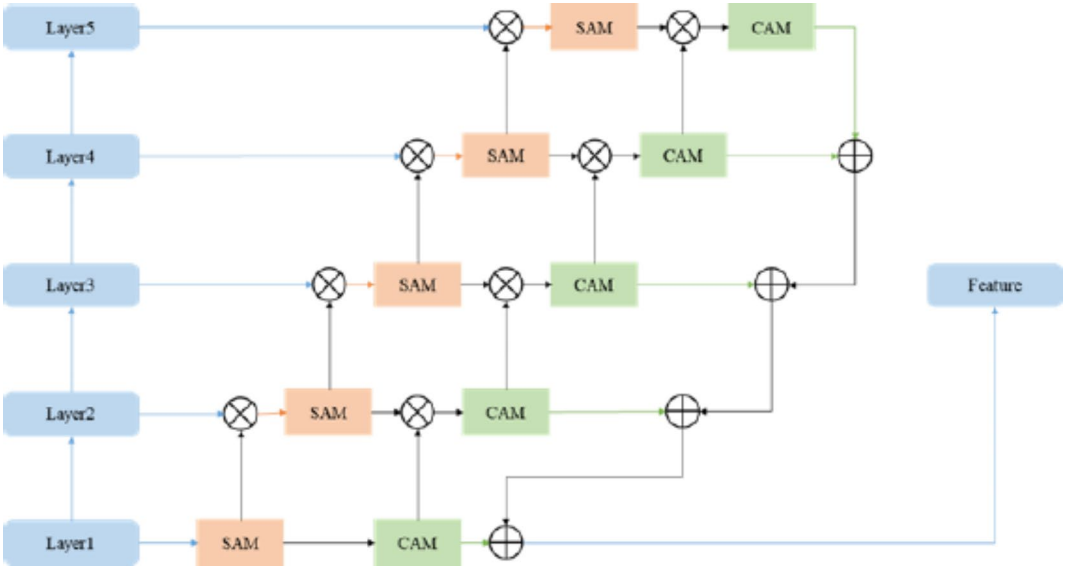

Fig. 2. Key feature extraction module architecture.

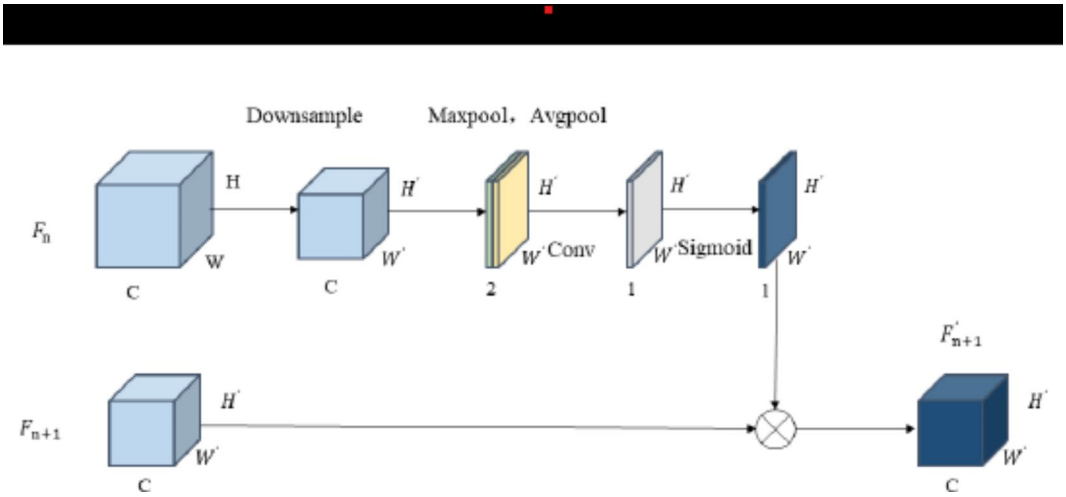

Fig. 3. Spatial Attention Module.

The feature extractor assigns the same weight to all channels when extracting features, which limits the classification performance of the algorithm. As such, improvements to the feature extractor are needed. In RS scene classification, global information is derived from the deep network, where deeper layers have larger receptive fields, resulting in feature maps that capture more extensive global context. Conversely, local information is extracted from the shallow network, where the receptive field of a single pixel is smaller, enabling the detection of finer details and small targets. This combination allows for a comprehensive understanding of both global and local features in scene classification tasks. The critical region in the classification task can be extracted by the spatial attention mechanism. Therefore, an attention module is incorporated between two neighbouring feature layers. This setup uses shallow features to guide the deep features and integrates the shallow local key features into the deep layer. The overall structure of this approach is illustrated in Fig. 2.

The spatial attention module (SAM) is shown in Fig. 3. Before extracting features, neighbouring features of the same size are made by means of F_n downsampling operation. Average pooling and maximum pooling operations are performed on F_n and the two pooling results are spliced according to the channels. Then, a convolution operation is performed to obtain the spatial attention weight matrix through the Sigmoid activation function. Finally, the weight matrix of F_n with respect to F_{n+1} is obtained.

The spatial attention is calculated as shown in Eq. (7).

$$F'_{n+1} = \sigma\left(f^5\left[F_n^{acg}; F_n^{max}\right]\right) \times F_{n+1}, \tag{7}$$

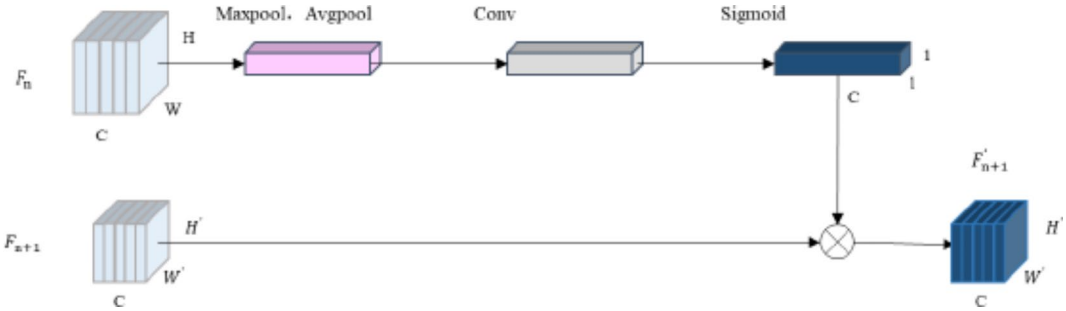

Fig. 4. Channel Attention Module.

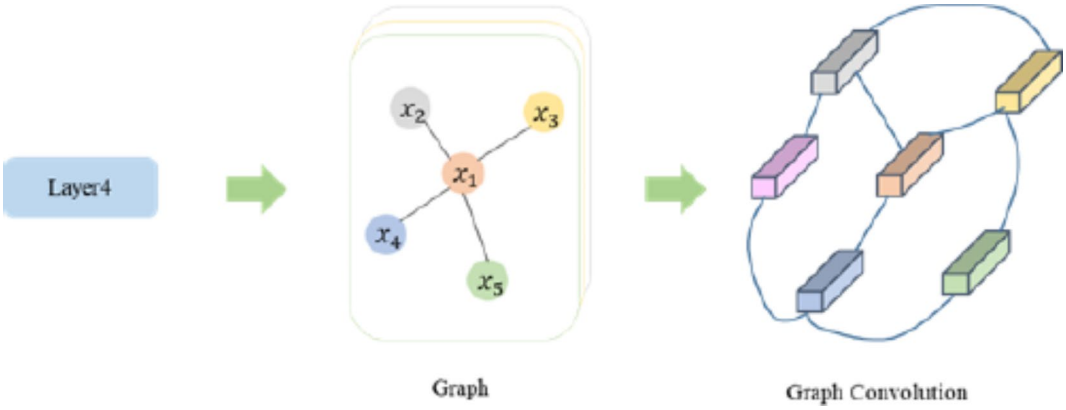

Fig. 5. Graph Convolutional Network.

where σ denotes the s-type function and f^5 represents a convolution operation with a convolution kernel size of 5×5.

To focus more effectively on key information in RS images and eliminate redundant data, feature learning is conducted along the channel dimension to assess the importance of each channel. In the CAM, shallow features are used to supervise deep features, as illustrated in Fig. 4. Since the number of feature channels may vary after applying the channel attention mechanism, 1×1 convolution is employed to standardise the number of channels before applying the channel attention.

The structure of the channel attention module is shown in Fig. 4. In order to obtain the supervision of F_n over F_{n+1}, the input feature map F_n is globally pooled and averagely pooled in the spatial dimension, and the features of the channel dimension, and the weights of each channel are learned based on the MLP. The pooling calculation Eq. (8) is as follows:

$$y_n = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_n(i,j), \tag{8}$$

where H and W denote the height and width of F_{n+1}, respectively, and σ denotes the s-type function.

The formula for channel attention is defined as denoted in Eq. (9):

$$F'_{n+1} = \sigma\ (C_1(y_n)) \times\ F_n, \tag{9}$$

where C_1 denotes a one-dimensional convolutional layer.

Contextual feature extraction module

The key feature extraction module, which utilises CNNs and attention mechanisms, refines features from R) images. Nonetheless, CNNs are limited to capturing short-range spatial information and cannot effectively model long-range dependencies within RS images. GCNs address this by leveraging their graph data structure to model long-distance spatial relationships. Traditional GCN algorithms, however, are computationally expensive due to the need to construct adjacency matrices for the entire dataset, which is particularly challenging for large-scale remote sensing applications. To mitigate this issue, we employ miniGCN, which enables the training of large-scale GCNs in smaller, manageable batches. As shown in Fig. 5, the feature map is obtained by using the fourth convolution layer of the VGG16 network in the key feature extraction module, and then a feature map with a size of 14×14×512 is obtained by down sampling through the pooling layer.

The relationship between feature nodes in the feature graph can be represented by $g = (\nu, \epsilon)$, where and denote the set of nodes and edges, respectively, and the set of edges E consists of the relationship between any two vertices V_i and V_j.

A is defined as the adjacency matrix of G, where A_{ij} represents the connection state between the i-th and j-th nodes. The adjacency matrix A is used to describe the relationship between vertices. The calculation of each element in A is shown in Eq. (10):

$$a_{i,j} = exp\left(-\frac{\|x_i - x_j\|}{\sigma^2}\right), \tag{10}$$

where x_i and x_j denote the feature vectors associated with vertices v_i and v_j, and σ is the width parameter of the function.

Assuming that the number of nodes in a graph set G is N, in order to reduce the huge computational cost of graph convolution, before each iteration of graph convolution, a random node sampler of size M (M ≪ N) is used to repetitively sample the vertices in the graph G until all the vertices have been sampled, generating a set of subgraphs shown in Eq. (11):

$$G = \left\{G_s = (V_s, E_s)\, 1,2,\cdots, \left\lceil N/M \right\rceil\right\}. \tag{11}$$

After the sampling is completed, the subgraph is input into GCN, which contains the graph convolution layer and the fully connected layer. Graph convolution realises the transmission of neighbourhood relations by aggregating the features between vertices v and all vertices $u \in V_g$. The GCN formula is shown in Eq. (12):

$$\tilde{H}_s^{(l+1)} = h\left(\tilde{D}_s^{-\frac{1}{2}} \tilde{A}_s \tilde{D}_s^{-\frac{1}{2}} \tilde{H}_s^{(l)} W^{(l)} + b_s^{(l)}\right), \tag{12}$$

where s denotes the sth subgraph and the sth batch of network training; w is the parameter matrix; h() is the activation function; and b is the bias parameter. The outputs of all subgraphs are cascaded to obtain the final output, as shown in Eq. (13):

$$H^{(l+1)} = \left[\tilde{H}_1^{(l+1)}, \cdots, \tilde{H}_s^{(l+1)}, \cdots \cdots, \tilde{H}_{\left\lceil N/M \right\rceil}^{(l+1)}\right]. \tag{13}$$

miniGCN reduces the GCN complexity while achieving better network local optimum results.

Feature fusion

The shallow feature F_1 has dimension m×1×1 and the deep feature F_2 has dimension n×1×1. The two sets of feature tensor are connected using the Concat operation which improves the performance of the model. Key features and GCN features are also fused using the Concat operation. The Concat operation connects the two features, where the dimensions of the features x and y are m and n, respectively. Consequently, the dimensions of the final output features become m + n. The feature fusion is denoted by Eq. (14):

$$F = [F_1; F_2]. \tag{14}$$

Scene recognition

In the present study, PSO-SVM was utilised as the final classification algorithm. By optimising the hyperparameters of the SVM through the PSO algorithm, the optimal set of hyperparameter configurations could be identified, thereby enhancing the SVM's performance in classification tasks. This approach automates the search of the hyperparameter space, reducing the need for manual adjustment and streamlining the optimisation process.

Experimental results and analysis

In the present study, three open-pit mines in Pingshuo were selected as the study area. The Pingshuo mining area is located at the border between Shuozhou city district and Pinglu district in Shanxi province, and the three open-pit mines in the Pingshuo mining area are Antaibao, Anjialing, and Dongdutian. The high-resolution RS images of the open-pit coal mine area used in the present study were sourced from historical images provided by Google Earth. The image of the study area is displayed in Fig. 6.

To evaluate the effectiveness of various features in open-pit coal mine scene recognition, different feature extraction methods—including GLCM, Gabor, and deep network features—were utilised. The texture features obtained from GLCM and Gabor were combined with deep features to construct classification features. By comparing the overall classification accuracy and scene recognition performance of these different feature combinations, the most effective classification features were identified. Initially, 1640 features were selected, including 20 texture features and 1620 depth features. Texture features included GLCM (8) and Gabor (12). The initial feature set had a large dimension and contained redundant data. Therefore, PCA was employed to reduce the feature set to 326 features.

Fig. 6. Remote sensing image of the study area.

Shallow feature extraction experiments

Dataset and experimental environment

The high-resolution satellite images of the surface coal mine area adopted in this section are from Google historical images, with an image resolution of about 2 m. Through visual interpretation and field investigation on the images, the feature types of the surface coal mine area were categorized into six categories, which are: coal mining area, pit slope, stripping area, agricultural land, buildings, and drainage field. The collected image data of the mining area were all cropped into 256 × 256 pixels sub-images, totaling 918 images. Each image was labeled by manual annotation.

The conventional feature extraction experiments were performed in the environment of MATLAB R2022b with the processor model AMD Ryzen5 3500X.

GLCM texture feature extraction results and analysis

Parameters such as texture statistics, neighbourhood window, texture direction and grey scale compression level during GLCM texture feature extraction affected the feature extraction effect. To simplify the calculations, the image grey levels were compressed to 64 levels, a 3 × 3 window was used, and a pixel distance of 1 was applied for calculating the GLCM texture features. The study analysed how different texture amounts and texture directions affected the recognition accuracy of various features in the study area.

From Figs. 7, 8, 9 and 10, it is evident that the four GLCM texture statistics displayed varying distinguishing abilities for different features. For example, in Figs. 7a and b, 8a and b, 9a and b and 10a and b, the stripping area, coal mining area, and the discharge field exhibited stronger brightness, which indicates that their homogeneity and angular second-order moments were larger. Agricultural land and buildings were brighter in Figs. 7c and d, 8c and d, 9c and d and 10c and d, respectively, reflecting the larger entropy value of agricultural land and the larger contrast of buildings.

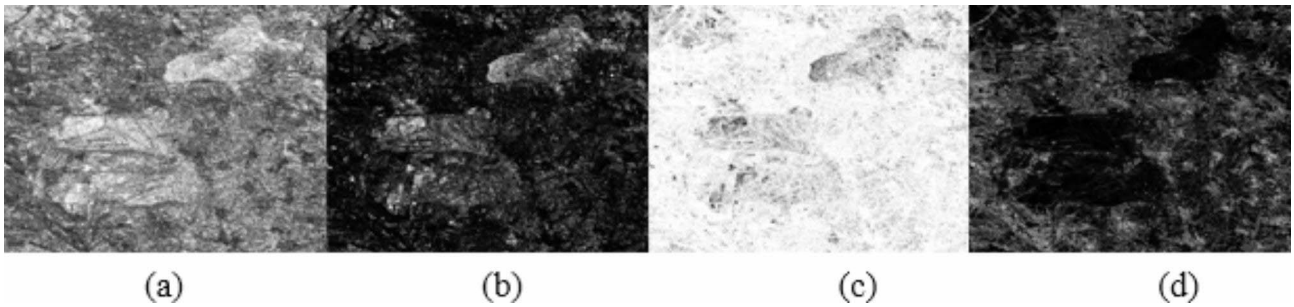

Fig. 7. GLCM 0-degree homogeneity, angular second moments, entropy, contrast.

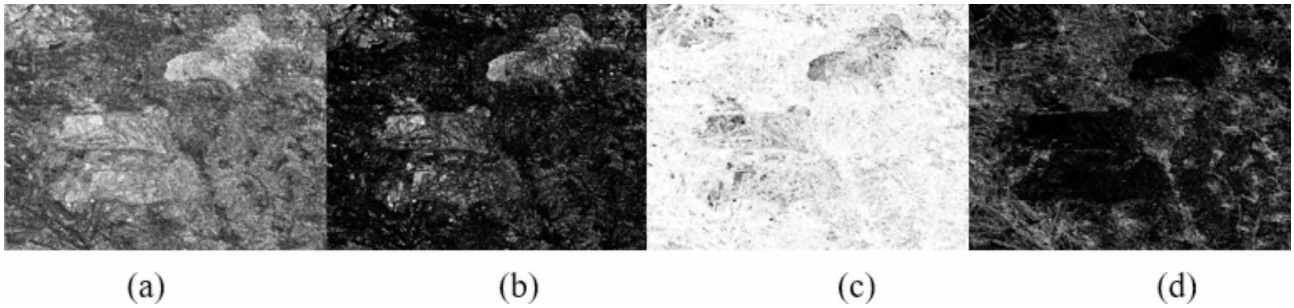

Fig. 8. GLCM 45-degree homogeneity, angular second moments, entropy, contrast.

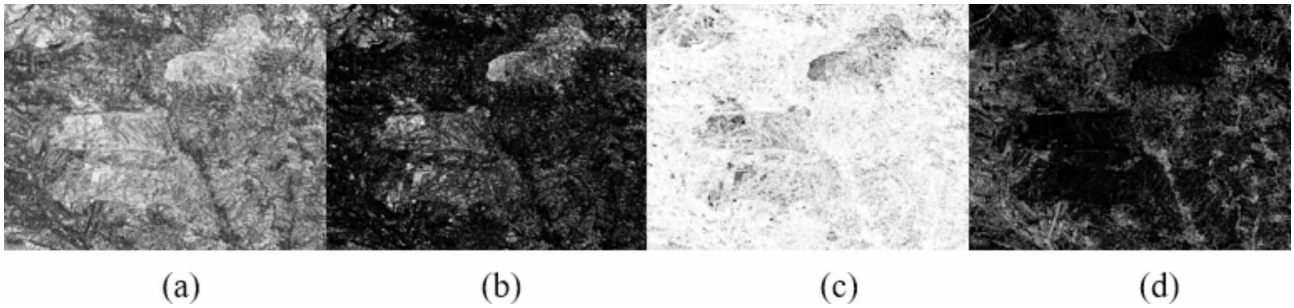

Fig. 9. GLCM 90-degree homogeneity, angular second moments, entropy, contrast.

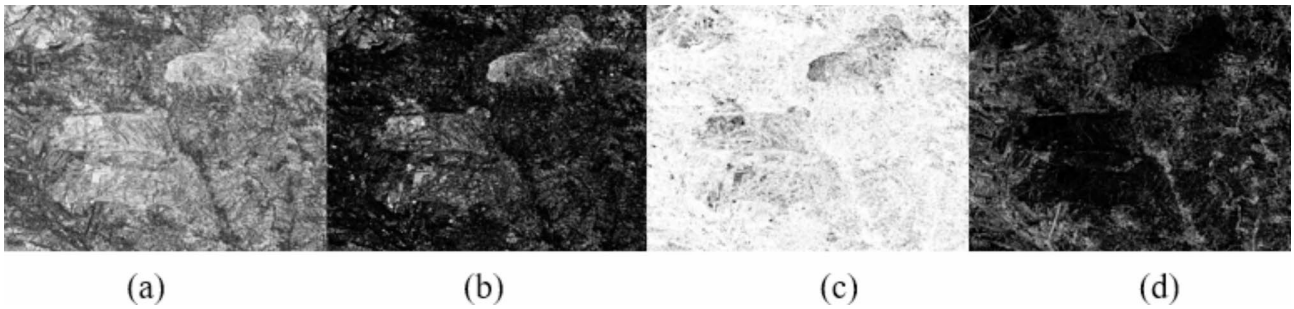

Fig. 10. GLCM 135-degree homogeneity, angular second moments, entropy, contrast.

Gabor texture feature extraction results and analysis

Parameters such as neighbourhood window, filter scale and direction during Gabor texture feature extraction affected the feature extraction results. The following figures show 24 Gabor features using a 21×21 neighbourhood window, a filter scale of 6 and a direction of 4.

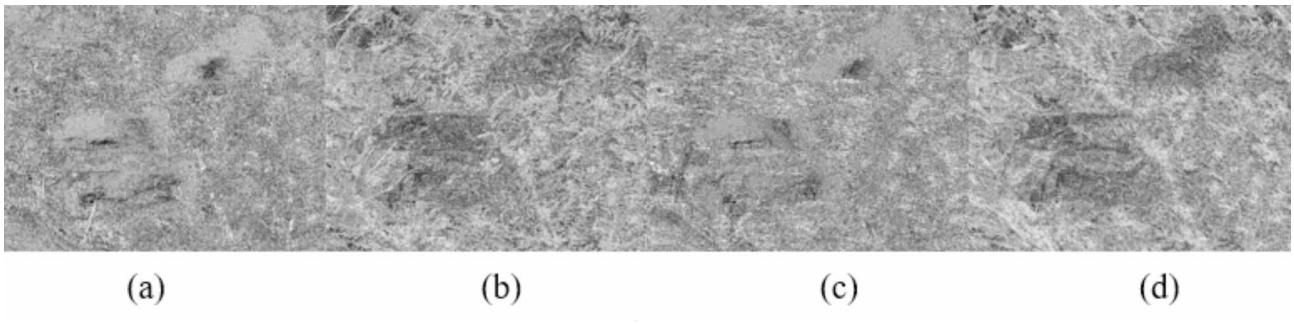

Fig. 11. Gabor scale I with angles 0, 45, 90, 135.

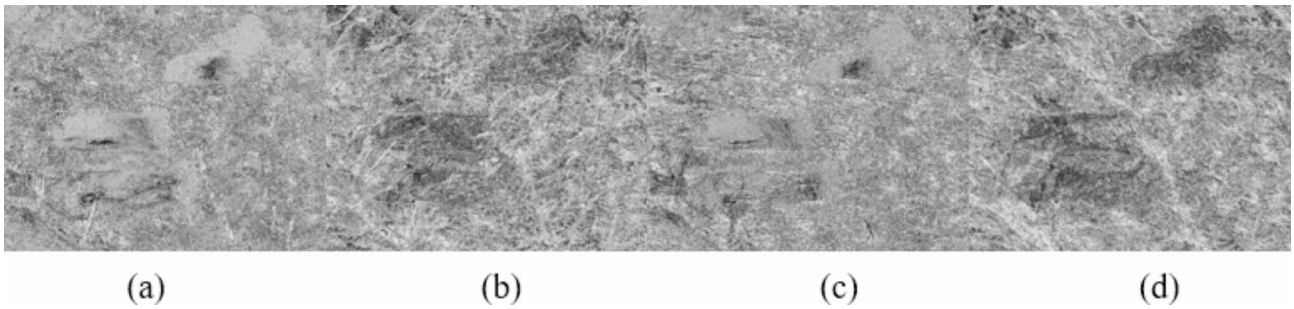

Fig. 12. Gabor scale II with angles 0, 45, 90, 135.

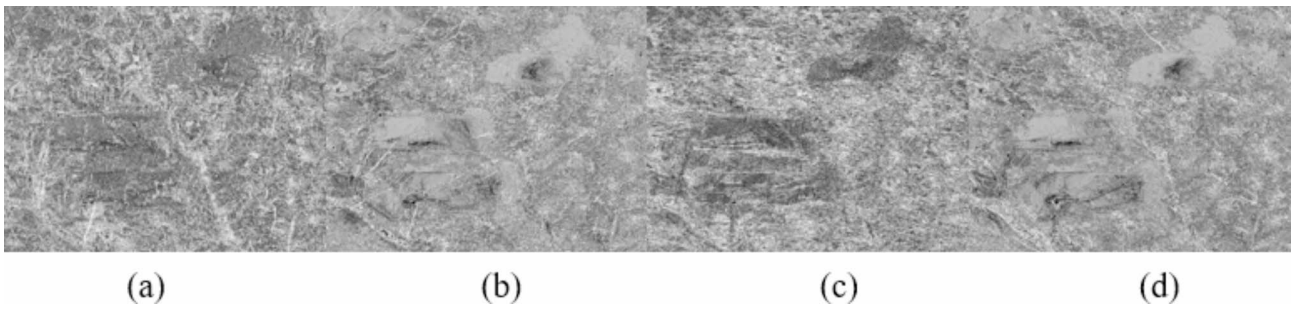

Fig. 13. Gabor scale III, with angles of 0, 45, 90, 135.

The brightness of a feature in the image correlates with its energy value. As shown in Figs. 11, 12, 13, 14, 15 and 16, the brightness of buildings decreased with increasing scale and was higher at 0 and 90 degrees for the same scale. Conversely, as the scale increased, the brightness of the earth removal site increased while the brightness of the coal mining area decreased, suggesting a clearer distinction between these features and others. Overall, larger scales improved the extraction capability for larger texture patterns but reduced the ability to capture smaller textures.

Statistical characterization of GLCM texture for different features

To quantitatively analyse the differences in GLCM texture features across different types in the study area, GLCM was calculated for the mean values of each feature at four directions: 0°, 45°, 90°, and 135°. The mean and standard deviation of GLCM were computed for each type of feature sample across various texture features (homogeneity, angular second moments, entropy, and contrast) in the images of the study area. The results are presented in Table 1. Overall, the standard deviations of angular second moments and homogeneity were relatively small compared to those of entropy and contrast, indicating that these two texture features are more advantageous for feature extraction.

Comparing the standard deviation of different features of the same feature in Tables 1 and 2, an observation can be made that the standard deviation of coal mining area was smaller in angular second moments and homogeneity features but larger in contrast features. The standard deviation of buildings was smaller in angular

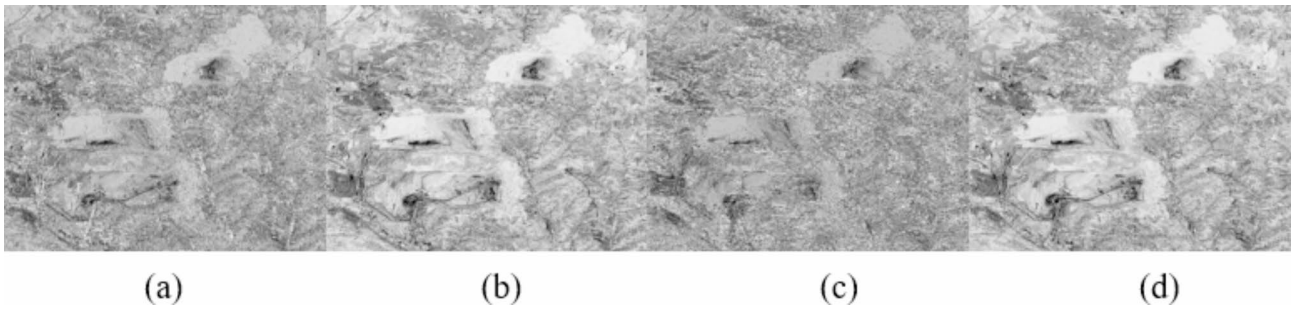

Fig. 14. Gabor scale IV with angles 0, 45, 90, 135.

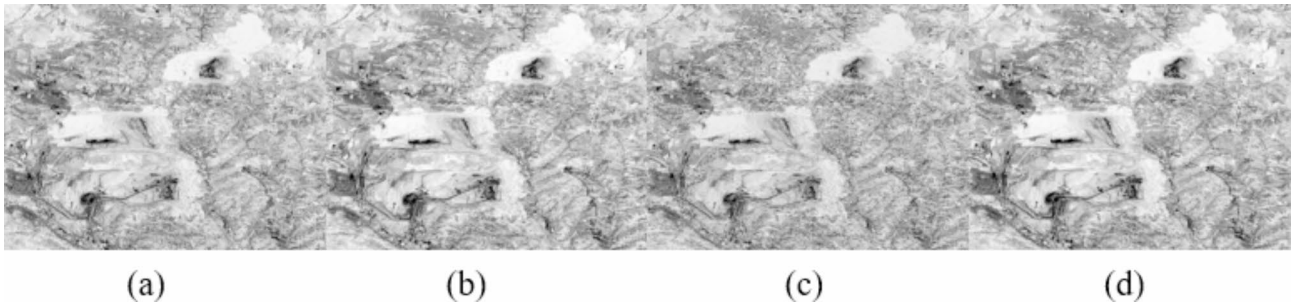

Fig. 15. Gabor scale V with angles 0, 45, 90, 135.

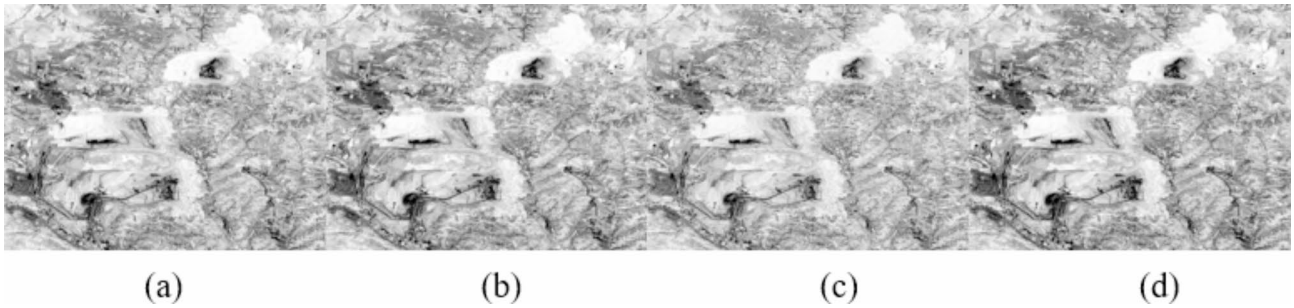

Fig. 16. Gabor scale VI with angles 0, 45, 90, 135.

Object	Entropy		Angular second moments	
	Mean value	Standard deviation	Mean value	Standard deviation
Stripping area	3.068505336	0.116398727	0.093197482	0.010045371
Pit slope	3.591584814	0.187953289	0.047202999	0.008125491
Coal mining area	3.121770516	0.128776815	0.070978562	0.008220239
Buildings	4.47736324	0.136819877	0.017369565	0.002448701
Agricultural land	3.483054469	0.151328326	0.063318376	0.008048864
Dump site	1.895596015	0.084789267	0.248515761	0.020585119

Table 1. Characteristic statistics of entropy and angular second moments for different features in the study area.

second moments features but larger in contrast features, indicating that different GLCM texture features had varying abilities to distinguish the same feature.

By comparing the mean values of the same feature across different feature types in Tables 1 and 2, and using the angular second moments feature as an example, it was observed that the values for stripping areas, coal

Object	Contrast		Homogeneity	
	Mean value	Standard deviation	Mean value	Standard deviation
Stripping area	0.909397778	0.260668931	0.761521525	0.036315922
Pit slope	1.324195341	0.486754122	0.733341037	0.054549724
Coal mining area	0.584402118	0.165328178	0.80586437	0.037015652
Buildings	3.841928198	1.143338804	0.566066677	0.043366824
Agricultural land	1.224498663	0.392730883	0.730960737	0.045699904
Dump site	0.387519207	0.094584882	0.847960242	0.027751769

Table 2. Characterization statistics of contrast and homogeneity of different features in the study area.

mining areas, and soil disposal sites were higher compared to other features, while the value for buildings was the smallest. This indicates that the angular second moments feature enhanced the distinguishability between stripping areas, coal mining areas, and soil disposal sites compared to other features, with a particularly strong distinction from buildings. These findings reflect the varying effectiveness of the same feature in distinguishing different feature types. The quantitative results align with the qualitative analysis results.

Statistical characterization of Gabor texture for different features

In order to quantitatively evaluate the ability of six scales and four directions of Gabor filtering to extract features for each feature in the study area, the mean and variance of different features on each wavelet texture feature were calculated according to the methods described in the previous section. The coefficient of variation, derived from these mean and variance values, was then computed. Both the coefficient of variation and the standard deviation indicate the degree of feature discretisation. A lower coefficient of variation signifies less feature discretisation, making it easier to distinguish the feature. The results are presented in Table 3.

Comparing the coefficients of variation for the same scale across different directions in Table 3, it was observed that the coefficients for directions two and four were larger than those for directions one and three. This indicates that directions one and three were more effective for feature extraction across the various objects in the study area. Additionally, when comparing coefficients of variation for different scales within the same direction, scales three, four, five, and six exhibited much smaller coefficients than scales one and two. This suggests that feature extraction capability improved with increasing scale. Therefore, considering the comprehensive image features, scales four, five, and six were selected as the final classification features.

Comparing the coefficients of variation of the same feature for different features in Table 3, the coefficient of variation for the stripped area at the second direction of the sixth scale was the smallest compared to other features, while the coefficient for buildings was the largest. This finding aligns with the results of the feature analysis. Further, examining the different texture features in Table 2 reveals that the coefficients of variation for the stripping area, agricultural land, and soil discharge field were relatively small across each texture feature. In contrast, the coefficients of variation for pit slopes, coal mining areas, and buildings were larger. This indicates that Gabor filtering was more effective for extracting features from the stripping area, agricultural land, and soil discharge field compared to these three categories.

Deep feature extraction experiments

Datasets

To evaluate the performance of the DS module, experiments were conducted on two public remote sensing scene datasets, the AID dataset[47] and the RSSCN7 dataset[48]. The image sizes in both datasets were uniformly scaled to 256 × 256 pixels to meet the model input requirements.

The AID dataset consists of 30 scene categories, each with 200 to 400 high-resolution remote sensing images. The size of the images in the AID dataset is 600 × 600.

The RSSCN7 dataset contains 2800 remotely sensed images with 7 scene categories which are grassland, forest, farmland, parking lot, residential area, industrial area, and river and lake, each of which consists of 400 images sampled based on 4 different scales. The pixel size of each image in this dataset is 400*400, which makes it more challenging due to the diversity of the scene images, which are derived from different seasons and weather variations and sampled at different scales.

Experimental setup and evaluation metrics

The experiments were implemented using TensorFlow framework and executed on windows operating system. To speed up the training process, we utilize GeForce RTX 3070Ti with 16G RAM. Adam's algorithm is selected as the optimizer, batch size is set to 16 and learning rate is initialized to 0.0001. 50% of the images from each scene category are randomly selected as the training set.

In remote sensing scene classification, Overall Accuracy (OA) is commonly used as an accuracy evaluation metric. OA is the number of correctly classified samples divided by the number of all samples.

Experiments were conducted to compare the model proposed in this paper with a variety of excellent scene recognition methods (GoogLeNet, VGG16 + SVM, CaffeNet, MCNN[49], SPP-NET[50], FACNN[51], ADFF[52], WSPM-CRC[53], GBNet+[54], EfficientNetB3 -Basic[15], Coutourlet CNN[55], SE-MDPMNet[56], Wavelet CNN[57], and TEX-Nets-LF[58]) are compared to validate the models. The results are shown in Tables 4 and 5.

The classification accuracy on the AID dataset reaches 97.52%. Because the proposed method constructs a two-branch model and fully learns global and contextual information, it is more capable of expressing the

Stripping area	Scale I			
	16.63688009	493.5740727	11.61809629	445.5510048
	Scale II			
	1.396448852	9.695541082	0.964914364	8.110308734
	Scale III			
	0.624533267	4.304443783	0.504857421	3.372372757
	Scale IV			
	0.809280868	1.927258499	0.677685859	1.528821539
	Scale V			
	0.560130776	2.506741448	0.499295886	2.122431262
	Scale VI			
	1.032509763	0.473908829	0.860003291	0.453405497
Pit slope	Scale I			
	10.4301172	351.9863398	9.706117411	294.30449
	Scale II			
	0.850810284	7.243367585	0.798855709	5.756146499
	Scale III			
	0.395193606	3.149535104	0.37661261	2.48682803
	Scale IV			
	0.575931704	1.391488122	0.561019775	1.200979215
	Scale V			
	0.407568507	1.849787228	0.40346919	1.637700233
	Scale VI			
	0.748664671	0.332513851	0.74348945	0.323075415
Coal mining area	Scale I			
	16.63688009	493.5740727	11.61809629	445.5510048
	Scale II			
	1.129865104	10.18999934	1.043842007	10.05762736
	Scale III			
	0.582422468	4.359978033	0.55313722	4.393484575
	Scale IV			
	0.805363555	1.973466197	0.782500652	2.041487041
	Scale V			
	0.592957158	2.882527497	0.581963795	3.077638366
	Scale VI			
	1.122815536	0.553012331	1.107286694	0.581410942
Buildings	Scale I			
	21.58054436	1323.433079	16.54555972	1368.909228
	Scale II			
	1.921129932	28.0160313	1.418425954	28.59024324
	Scale III			
	0.868521994	10.91005449	0.72536459	11.34243031
	Scale IV			
	1.108554304	4.217898992	0.986862527	4.359926419
	Scale V			
	0.729026593	5.470301158	0.678299066	5.689181749
	Scale VI			
	1.426916514	0.701727827	1.327099597	0.734870947

Continued

Agricultural land	Scale I			
	12.98974101	402.0191377	10.82124099	467.6800789
	Scale II			
	1.057151152	7.89178637	0.919656032	9.472930703
	Scale III			
	0.512238995	3.398838395	0.439191823	4.05857784
	Scale IV			
	0.73077492	1.523620323	0.626119329	1.745290525
	Scale V			
	0.51039492	2.147476437	0.459856926	2.390167866
	Scale VI			
	0.933146126	0.414534806	0.862567681	0.426186829
Dump site	Scale I			
	9.309919583	171.501162	8.989684225	166.4415943
	Scale II			
	0.745495931	3.432253683	0.726807258	3.307083557
	Scale III			
	0.330164002	1.60089376	0.322818709	1.50073381
	Scale IV			
	0.500408483	0.812162392	0.492858145	0.783798574
	Scale V			
	0.355324068	1.006842572	0.350599351	0.961750973
	Scale VI			
	0.656073467	0.253124323	0.64879848	0.250270915

Table 3. Differences in statistical characteristics of Gabor texture for different features in the study area.

Method	50%
GoogLeNet	86.39
CaffeNet	89.53
VGG16 + SVM	89.64
MCNN	91.8
SPP-NET	91.45
FACNN	95.45
ADFF	94.75
WSPM-CRC	95.11
GBNet+	95.48
Proposed	97.52

Table 4. Overall accuracy of the classification methods on the AID dataset.

high-level semantics of the scene.The RSSCN7 dataset has multi-scale and multi-angle attributes for each scene category, which makes it challenging to classify, and the proposed algorithm achieves the highest classification accuracy of 96.73% with 50% of the training samples in the RSSCN7 dataset, which is better than the above advanced scene classification algorithms. The proposed algorithm achieves the highest classification accuracy of 96.73% with 50% of the training samples of RSSCN7 dataset, which is better than the above advanced scene classification algorithms.

Feature combination experiments (using svm)

In order to quantitatively compare the classification results of different texture feature combinations, the following three classification features were constructed: GLCM texture feature (GLCM), Gabor texture feature (Gabor), depth feature (deep), Gabor texture feature combined with GLCM texture feature (Gabor + GLCM), Gabor texture feature combined with depth feature (Gabor + deep), GLCM texture feature combined with depth feature (GLCM + deep) and three types of feature combination (GLCM + Gabor + deep). SVM was used to classify the described feature combinations, and the results are shown in Fig. 17.

As shown in Fig. 17, the classification accuracy of single features both traditional shallow features and deep features was lower than that of combined features, and the accuracy of a single traditional feature combined with deep features was higher than the classification accuracy of the combination of two traditional features. As such,

Method	50%
GoogLeNet	85.84
VGG16 + SVM	87.18
CaffeNet	88.25
WSPM-CRC	93.9
SE-MDPMNet	94.71
EfficientNetB3-Basic	94.39
Wavelet CNN	94.89
TEX-Nets-LF	94
Coutourlet CNN	95.54
Proposed	96.73

Table 5. Overall accuracy of classification methods on RSSCN7 dataset.

the three features were ultimately combined. This combination led to varying degrees of improvement in both classification accuracy and the kappa coefficient, indicating enhanced performance in scene recognition.

The results indicate that shallow features extracted through traditional methods could represent image information more effectively when fused with other features, compared to using a single feature alone. The integration of two different shallow features provided complementary information, thereby enhancing the model's generalisation ability and classification performance. However, neither single traditional features nor their combinations outperformed the deep features in classification performance. This suggests that combining traditional features with deep features was more effective for extracting higher-level semantic information from images, thereby improving the classification accuracy of high-resolution remote sensing scene images.

Scene recognition based on particle swarm optimization

Accuracy analysis of different classification methods

The radial basis kernel function requires a penalty coefficient c and a kernel parameter γ. If these parameters are too large, the detection accuracy of the training sample set will be high, while the detection accuracy of the test sample set will be low. On the contrary, if they are disproportionately small, the detection accuracy will be so low as to be unsatisfactory, rendering the model useless. In the present study, the SVM was optimised using the particle swarm algorithm.

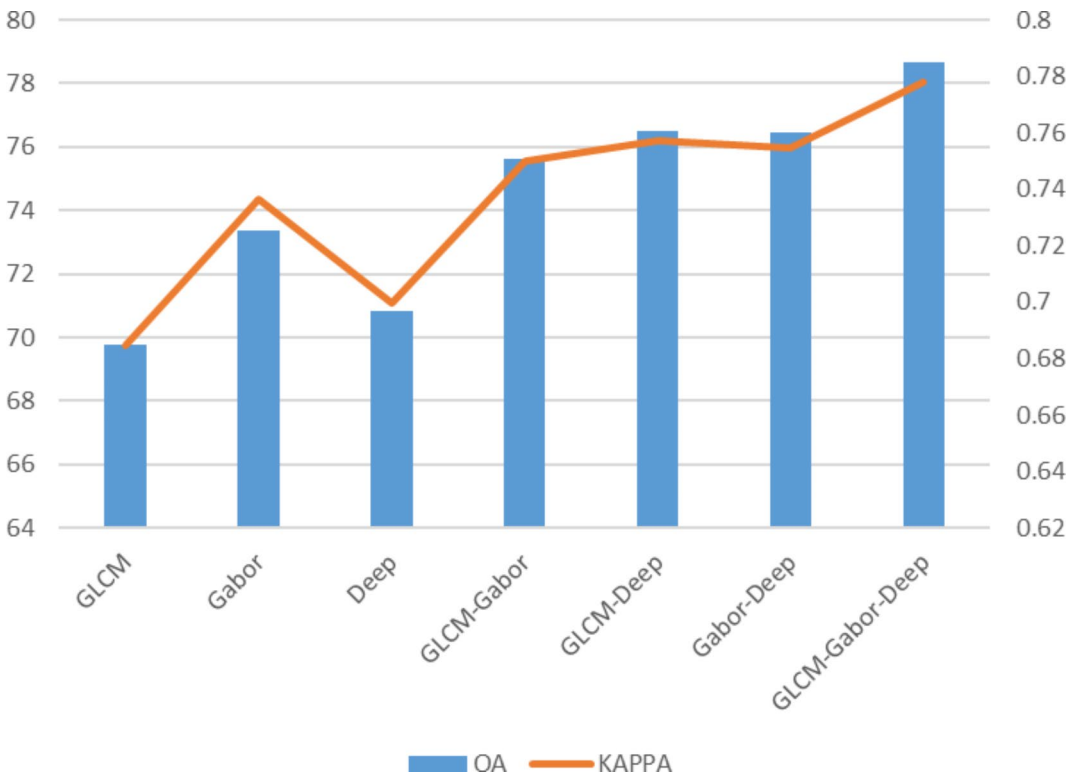

Fig. 17. Overall accuracy for different feature combinations.

Comparing PSO-SVM with Random Forest, Decision Tree and XGBoost, Fig. 18 shows the overall accuracy of classification with Kappa coefficient. From the table, an observation can be made that the overall classification accuracy and Kappa coefficient achieved by the method optimised using the particle swarm algorithm were higher compared to other methods. This indicates that the PSO algorithm is more effective for classifying high-resolution remote sensing image scenes.

Principal component analysis dimensionality reduction

A certain correlation was identified between the features, which suggested the presence of some invalid and redundant information that could affect classification results and reduce accuracy. To address this issue, in the present study, the PCA method was used to analyse and reduce the dimensionality of the combined feature vectors, and the accuracy of the reduced dimensionality was improved by 10.02%. In addition, the PCA dimensionality reduction effectively increased the accuracy of classification.

As shown in Fig. 19, among the machine learning models SVM, RF, DT, and XGBoost trained and tested with the dimensionality-reduced 1,920 features, PSO-SVM outperformed all other models, achieving an accuracy of 92.78%.

Table 6 presents the evaluation metrics for each class in the mining dataset using the PSO-SVM classification method. The results indicate that four classes of scenes achieved recognition accuracies of 90% or above. However, the stripping areas and dump sites did not reach 90% accuracy, primarily due to their similar textures, which increased the probability of misclassification between these two classes. Overall, the recognition accuracy exceeded 92%, demonstrating that the proposed scene recognition method, which combined deep and shallow features, effectively recognised scenes in mining areas.

Comparison of scene recognition results in open pit coal mining areas

Table 7 presents the results of scene recognition for remote sensing images of the study area using various methods. The comparison shows that the optimal accuracy of the CNN-only model was 76.52%. In contrast, the combined GCN method proposed in the present study achieved an accuracy of 92.31%, marking an improvement of 15.79% points over the best-performing ResNet101 model. This significant increase demonstrates that the proposed method has a clear advantage in recognising scenes in open-pit coal mining areas, showcasing a stronger recognition capability.

Conclusion

A high-resolution remote sensing image open pit coal mining scene recognition method was proposed that combines shallow and deep features. The method employs GLCM and Gabor to extract shallow features and CNN and GCN to extract deep features. These features are then fused to obtain a comprehensive set of fusion

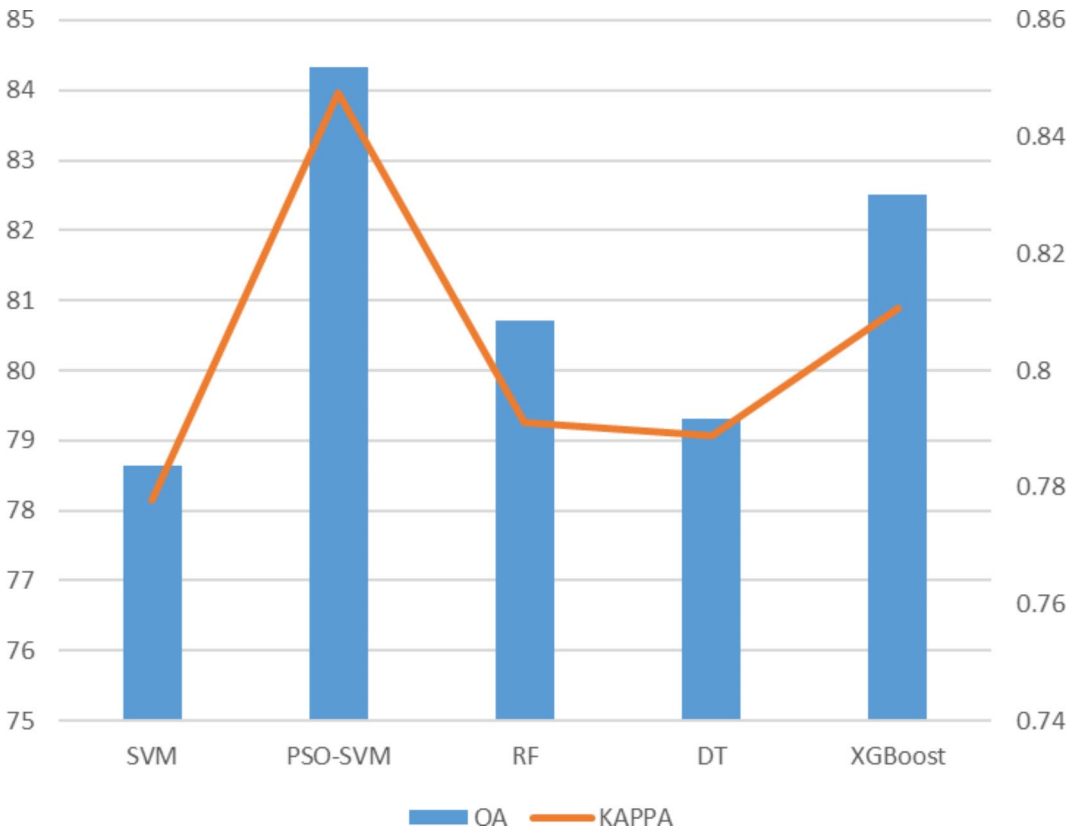

Fig. 18. Comparison of different classification methods.

Class	OA	Precision	Recall	F1
Stripping area	89.15	89.25	90.6	90.38
Pit slope	90.08	91.61	90.37	91.99
Coal mining area	91.52	92.65	92.41	92.53
Buildings	91.33	91.35	90.6	90.97
Agricultural land	92.44	93.16	91.19	92.65
Dump site	90.77	89.19	90.28	89.73

Table 6. Accuracy, precision, recall, and F1 for each category in the dataset.

Method	Precision	Recall	F1
VGG16	63.84	79.23	74.66
VGG19	70.67	81.36	73.15
ResNet50	74.87	75.42	79.04
ResNet101	76.52	77.87	81.73
ResNet152	75.63	75.31	79.63
Proposed	92.31	87.19	89.78

Table 7. Precision, Recall, and F1 of various methods for Scene Recognition in Open Coal Mining areas.

features. Classification accuracy varies with different feature combinations, so each feature was tested separately, in pairs, and as a full set to determine the optimal feature combination. The optimal kernel function for SVM was selected by evaluating four different kernel functions on the best feature set. Given the large number of features and potential redundancy, PCA was used for dimensionality reduction on the optimal feature set. Finally, the reduced texture features were input into the PSO-SVM classifier, achieving effective recognition of the mining scenes. In order to evaluate the performance of the methods, extensive experiments were conducted on several classifiers, including Random Forest, Decision Tree, and XGBoost. The comparison focused on overall accuracy and the Kappa coefficient of the algorithms. The experimental results demonstrate that the feature extraction model proposed in this study exhibits strong extraction capabilities for mining areas, proving to be effective for mining scene recognition.

First, due to the manual combination of deep and shallow features before inputting them into the classifier, the traditional modular processing approach increases the time required for scene recognition. To address this, the plan is to optimise the model further to enable end-to-end processing. Second, the framework encounters challenges in distinguishing between similar categories, such as stripping areas and discharge sites, which are often confused. Future work will focus on enhancing the model's ability to differentiate between these similar areas.

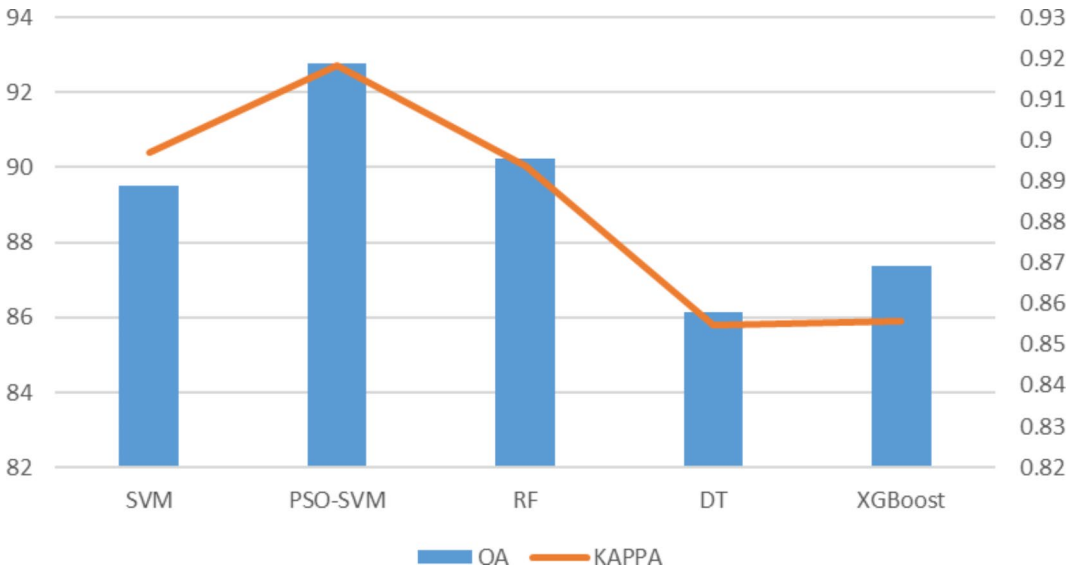

Fig. 19. Comparison of different classification methods after dimensionality reduction.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 3 April 2024; Accepted: 11 September 2024
Published online: 15 October 2024

References

1. Wu, Z. H., Lei, S. G., Lu, Q. Q., Bian, Z. F. & Ge, S. J. Spatial distribution of the impact of surface mining on the landscape ecological health of semi-arid grasslands. *Ecol. Ind.* **111**, 105996. https://doi.org/10.1016/j.ecolind.2019.105996 (2020).
2. Zhang, P. et al. Establishment of landslide early-warning indicator using the combination of numerical simulations and case matching method in Wushan open-pit mine. *Front. Earth Sci.* **10**, 960831. https://doi.org/10.3389/feart.2022.960831 (2022).
3. Jia, L., Wang, J. & Gao, S. Landslide risk evaluation method of open-pit mine based on numerical simulation of large deformation of landslide. *Sci. Rep.* **13**, 15410. https://doi.org/10.1038/s41598-023-42736-4 (2023).
4. Tzampoglou, P. & Loupasakis, C. Hydrogeological hazards in open pit Coal Mines–investigating triggering mechanisms by validating the European Ground Motion Service Product with Ground Truth Data. *Water* **15** (8), 1474. https://doi.org/10.3390/w15081474 (2023).
5. Madasa, A., Orimoloye, I. R. & Ololade, O. O. Application of geospatial indices for mapping landcover use change detection in a mining area. *J. Afr. Earth Sc.* **175**, 104108. https://doi.org/10.1016/j.jafrearsci.2021.104108 (2021).
6. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Trans. Syst. Man. Cybernetics* **3**(6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314 (1973).
7. JALSR, N. Object detection using gabor filters. *Pattern Recogn.* **30**(2), 295–309. https://doi.org/10.1016/S0031-3203(96)00068-4 (1997).
8. Lowe, D. G. Distinctive image features from Scale-Invariant keypoints. *Int. J. Comput. Vision* **60**(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94 (2004).
9. Dalal, N. & Triggs, B. Histograms of oriented gradients for human detection. In *Proceedings of International Conference on Computer Vision and Pattern Recognition* San Diego, USA: 886–893 https://doi.org/10.1109/CVPR.2005.177 (2005).
10. Chen, H., Miao, F., Chen, Y., Xiong, Y. & Chen, T. A. Hyperspectral image classification method using multifeature vectors and optimized KELM. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **14**, 2781–2795. https://doi.org/10.1109/JSTARS.2021.3059451 (2021).
11. Zhu, B. & R_2FD_2 Fast and robust matching of Multimodal Remote sensing images via repeatable feature detector and rotation-invariant feature descriptor. *IEEE Trans. Geosci. Remote Sens.* **61**, 5606115. https://doi.org/10.1109/TGRS.2023.3264610 (2023).
12. Li, W., Chen, C., Su, H. & Du, Q. Local binary patterns and Extreme Learning Machine for Hyperspectral Imagery classification. *IEEE Trans. Geosci. Remote Sens.* **53**(7), 3681–3693. https://doi.org/10.1109/TGRS.2014.2381602 (2015).
13. Preethi, P. & Mamatha, H. R. Region-based convolutional neural network for segmenting text in epigraphical images. *Artif. Intell. Appl.* **1**(2), 119–127. https://doi.org/10.47852/bonviewAIA2202293 (2022).
14. Bhosle, K. & Musande, V. Evaluation of Deep Learning CNN Model for Recognition of Devanagari Digit. *Artif. Intell. Appl.* **1**(2), 114–118. https://doi.org/10.47852/bonviewAIA3202441 (2023).
15. Chen, H. et al. M^3FuNet: An Unsupervised Multivariate Feature Fusion Network for Hyperspectral Image Classification. In IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–15, Art no. 5513015. https://doi.org/10.1109/TGRS.2024.3380087 (2024).
16. Cao, R., Fang, L., Lu, T. & He, N. Self-attention-based Deep Feature Fusion for Remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **18**(1), 43–47. https://doi.org/10.1109/LGRS.2020.2968550 (2021).
17. Deng, P., Xu, K. & Huang, H. When CNNs Meet Vision Transformer: a Joint Framework for Remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **19**, 8020305. https://doi.org/10.1109/LGRS.2021.3109061 (2022).
18. Alhichri, H., Alswayed, A. S., Bazi, Y., Ammour, N. & Alajlan, N. A. Classification of Remote sensing images using EfficientNet-B3 CNN Model with attention. *IEEE Access*. **9**, 14078–14094. https://doi.org/10.1109/ACCESS.2021.3051085 (2021).
19. Lv, P. et al. A spatial-Channel feature preserving Vision Transformer for Remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **60**, 4409512. https://doi.org/10.1109/TGRS.2022.3157671 (2022).
20. Wang, W., Chen, Y., Ghamisi, P. & Transferring, C. N. N. With adaptive learning for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **60**, 5533918. https://doi.org/10.1109/TGRS.2022.3190934 (2022).
21. Bai, J. et al. Hyperspectral image classification based on deep attention graph Convolutional Network. *IEEE Trans. Geosci. Remote Sens.* **60**, 5504316. https://doi.org/10.1109/TGRS.2021 (2022).
22. Hong, D. et al. Graph Convolutional Networks for Hyperspectral Image classification. *IEEE Trans. Geosci. Remote Sens.* **59**(7), 5966–5978. https://doi.org/10.1109/TGRS.2020.3015157 (2021).
23. Yuan, Z. Q. et al. Remote sensing Cross-modal text-image Retrieval based on global and local information. *IEEE Trans. Geosci. Remote Sens.* **60**, 5620616. https://doi.org/10.1109/TGRS.2022.3163706 (2022).
24. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of International Conference on Learning Representations* San Juan, Puerto Rico. https://doi.org/10.48550/arXiv.1609.02907 (2016).
25. Kennedy, J. & Eberhart, R. Particle Swarm Optimization. In *Proceedings of International Conference on Neural Networks* Perth, Western Australia. https://doi.org/10.1109/ICNN.1995.488968 (1995).
26. Cortes, C., Vapnik, V., Support-Vector & Networks *Mach. Learn.* **20**(3), 273–297 https://doi.org/10.1023/A:1022627411411. (1995).
27. Chen, C. et al. Gabor-Filtering-Based Completed Local Binary Patterns for Land-Use Scene Classification. In *Proceedings of the IEEE International Conference on Multimedia Big Data* Beijing, China, 324–329. https://doi.org/10.1109/BigMM.2015.23 (2015).
28. Wang, J., Fan, Y. Y., Li, Z. H. & Lei, T. Texture classification using multi-resolution global and local Gabor features in pyramid space. *Signal. Image Video Process.* **13**, 163–170. https://doi.org/10.1007/s11760-018-1341-6 (2018).
29. Cai, R. L. & Shang, G. W. Flexible 3-D Gabor features fusion for hyperspectral imagery classification. *J. Appl. Remote Sens.* **15**(3), 036508. https://doi.org/10.1117/1.JRS.15.036508 (2021).
30. Pan, H. Z., Liu, M. Q., Ge, H. M. & Yuan, Q. Learnable three-dimensional Gabor convolutional network with global affinity attention for hyperspectral image classification. *Chin. Phys. B* **31**(12), 120701–120701. https://doi.org/10.1088/1674-1056/ac8cd7 (2022).
31. Zheng, G. et al. Development of a Gray-Level Co-occurrence Matrix-based texture orientation estimation method and its application in sea surface wind direction Retrieval from SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **56**(9), 5244–5260. https://doi.org/10.1109/TGRS.2018.2812778 (2018).
32. Iqbal, N., Mumtaz, R., Shafi, U. & Zaidi, S. M. H. Gray level co-occurrence matrix (GLCM) texture based crop classification using low altitude remote sensing platforms. *PeerJ Comput. Sci.* **7**, e536. https://doi.org/10.7717/peerj-cs.536 (2021).
33. Sun, H., Li, S., Zheng, X. & Lu, X. Remote sensing scene classification by gated bidirectional network. *IEEE Trans. Geosci. Remote Sens.* **58**(1), 82–96. https://doi.org/10.1109/TGRS.2019.2931801 (2022).

34. Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90. https://doi.org/10.1145/3065386 (2012).
35. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Boston, MA. https://doi.org/10.48550/arXiv.1409.1556 (2015).
36. Szegedy, C. et al. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Boston, USA: 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
37. Bae, W., Yoo, J. & Ye, J. Beyond Deep Residual Learning for Image Restoration: Persistent Homology-Guided Manifold Simplification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Workshops. Honolulu, HI, USA: 1141–1149. https://doi.org/10.1109/CVPRW.2017.152 (2017).
38. Xie, Q., Zhou, D., Tang, R., Feng, H. A. & Deep CNN-Based detection method for Multi-scale Fine-Grained objects in Remote sensing images. *IEEE Access.* **12**, 15622–15630. https://doi.org/10.1109/ACCESS.2024.3356716 (2024).
39. Li, E., Samat, A., Du, P., Liu, W. & Hu, J. Improved Bilinear CNN Model for Remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **19**, 8004305. https://doi.org/10.1109/LGRS.2020.3040153 (2022).
40. Xu, C., Zhu, G., Shu, J. A. & Lightweight Robust lie group-convolutional neural networks joint representation for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **60**, 5501415. https://doi.org/10.1109/TGRS.2020.3048024 (2022).
41. Khan, S. D. & Basalamah, S. Multi-branch Deep Learning Framework for Land Scene classification in Satellite Imagery. *Remote Sens.* **15**(13), 3408. https://doi.org/10.3390/rs15133408 (2023).
42. Chen, R., Li, G., Dai, C. L. & H. & Feature Fusion via Deep residual graph Convolutional Network for Hyperspectral Image classification. *IEEE Geosci. Remote Sens. Lett.* **19**, 6011805. https://doi.org/10.1109/LGRS.2022.3192832 (2022).
43. Yang, J. Y., Li, H. C., Hu, W. S., Pan, L. & Du, Q. Adaptive cross-attention-driven spatial–spectral graph Convolutional Network for Hyperspectral Image classification. *IEEE Geosci. Remote Sens. Lett.* **19**, 6004705. https://doi.org/10.1109/LGRS.2021.3131615 (2022).
44. Yu, L., Peng, J., Chen, N., Sun, W. & Du, Q. Two-branch deeper graph Convolutional Network for Hyperspectral Image classification. *IEEE Trans. Geosci. Remote Sens.* **61**, 5506514. https://doi.org/10.1109/TGRS.2023.3257369 (2023).
45. Xue, Z., Liu, Z. & Zhang, M. D. S. R. G. C. N. Differentiated-scale restricted Graph Convolutional Network for few-shot hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **61**, 5504918. https://doi.org/10.1109/TGRS.2023.3253248 (2023).
46. Deng, P. F., Xu, K. J. & Huang, H. CNN-GCN-based dual-stream network for scene classification of remote sensing images. *Natl. Remote Sens. Bull.* **25**(11), 2270–2282. https://doi.org/10.11834/jrs.20210587 (2021).
47. Xia, G. S. et al. AID: a Benchmark Data Set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **55** (7), 3965–3981. https://doi.org/10.1109/TGRS.2017.2685945 (2017).
48. Zou, Q., Ni, L., Zhang, T. & Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **12**(11), 2321–2325. https://doi.org/10.1109/LGRS.2015.2475299 (2015).
49. Liu, Y., Zhong, Y. & Qin, Q. Scene classification based on Multiscale Convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **56**(12), 7109–7121. https://doi.org/10.1109/TGRS.2018.2848473 (2018).
50. Han, X., Zhong, Y., Cao, L., Zhang, L. & Pre-Trained AlexNet Architecture with pyramid pooling and Supervision for high spatial resolution remote sensing image scene classification. *Remote Sens.* **9**(8), 848. https://doi.org/10.3390/rs9080848 (2017).
51. Lu, X., Sun, H., Zheng, X. A. & Feature Aggregation Convolutional Neural Network for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **57**(10), 7894–7906. https://doi.org/10.1109/TGRS.2019.2917161 (2019).
52. He, N., Fang, L., Li, S., Plaza, J. & Plaza, A. Skip-Connected Covariance Network for Remote Sensing Scene Classification. I*EEE Transactions on Neural Networks and Learning Systems.* 31(5), 1461–1474. https://doi.org/10.1109/TNNLS.2019.2920374 (2020).
53. Liu, B. D. et al. Weighted spatial pyramid matching collaborative representation for remote-sensing-image scene classification. *Remote Sens.* **11**(5), 518. https://doi.org/10.3390/rs11050518 (2019).
54. Sun, H., Li, S., Zheng, X. & Lu, X. Remote sensing scene classification by gated bidirectional network. *IEEE Trans. Geosci. Remote Sens.* **58**(1), 82–96. https://doi.org/10.1109/TGRS.2019.2931801 (2020).
55. Liu, M. et al. C-CNN: Contourlet Convolutional neural networks. *IEEE Trans. Neural Networks Learn. Syst.* **32**(6), 2636–2649. https://doi.org/10.1109/TNNLS.2020.3007412 (2021).
56. Zhang, B., Zhang, Y., Wang, S. A. & Lightweight Discriminative model for remote sensing scene classification with Multidilation Pooling Module. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **12**(8), 2636–2653. https://doi.org/10.1109/JSTARS.2019.2919317 (2019).
57. Fujieda, S., Takayama, K. & Hachisuka, T. Wavelet convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Salt Lake City, USA. http://arxiv.org/abs/1805.08620 (2018).
58. Anwer, R. M., Khan, F. S., Weijer, J. V. D. & Laaksonen, J. TEX-Nets: Binary Patterns Encoded Convolutional Neural Networks for Texture Recognition. In *Proceedings of the ACM on International Conference on Multimedia Retrieval* Bucharest, Romania: 125–132. https://doi.org/10.1145/3078971.3079001 (2017).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Project No. 42171424).

Author contributions

Conceptualization, methodology, and validation were carried out by Y.L.; original draft preparation by Y.L.; review and editing were carried out by J.Z. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

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