DOI: 10.4208/JICS-2024-005
June 2024

Remote Sensing Image Scene Classification Based on Deep Learning Feature Fusion

Liqi Wang[1], Cheng Zhang[1], Yuchao Hou[2], Xiuhui Tan[1], Rong Cheng[1], Xiang Gao[1] and Yanping Bai[1,2,*]

[1] *School of Mathematics, North University of China, Taiyuan 030051, China*

[2] *School of Information and Communication Engineering, North University of China, Taiyuan 030051, China*

Abstract. In view that traditional manual feature extraction method cannot effectively extract the overall deep image information, a new method of scene classification based on deep learning feature fusion is proposed for remote sensing images. First, the Grey Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP) are used to extract the shallow information of texture features with relevant spatial characteristics and local texture features as well; second, the deep information of images is extracted by the AlexNet migration learning network, and a 256-dimensional fully connected layer is added as feature output while the last fully connected layer is removed; and the two features are adaptively integrated, then the remote sensing images are classified and identified by the Grid Search optimized Support Vector Machine (GS-SVM). The experimental results on 21 types of target data of the public dataset UC Merced and 7 types of target data of RSSCN7 produced average accuracy rates of 94.77% and 93.79%, respectively, showing that the proposed method can effectively improve the classification accuracy of remote sensing image scenes.

AMS subject classifications: 68U10, 68T05

Key words: Image classification, Convolutional Neural Network (CNN), Grey Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), migration learning, Support Vector Machine (SVM)

1 Introduction

Translated from *Journal of Nanjing University of Information Science & Technology*, 2023, 15(3): 346-356.

[*]Corresponding author. *Email addresses:* **1023252901@qq.com** (L. Wang), **baiyp666@163.com** (Y. Bai).

With the continuous development of remote sensing technology [1], remote sensing image classification has been widely applied in many fields such as land management, urban planning and traffic supervision [2]. However, at present, remote sensing scene images contain rich and complex information and structures, and there are still many challenges in how to make reasonable use of rich information in remote sensing images to obtain accurate and effective features [3].

Traditional manual feature extraction is commonly used in remote sensing image scene classification, including color histogram, texture feature, Global Image Descriptor (GIST), Scale Invariant Feature Transform (SIFT), etc. Li Fuyu et al. [4] pointed out that the remote sensing image registration technology based on SIFT has advantages in scale rotation invariance. Xu Junyi et al. [5] took Grey Level Co-occurrence Matrix (GLCM) as the first principal component of Principal Component Analysis (PCA), making full use of the robustness of GLCM in acquiring texture features. Although traditional manual features have good stability and the ability to express the overall shallow information, and are feasible to be directly applied to the scene classification task of low-resolution remote sensing images, traditional manual features are too dependent on manual design and cannot effectively extract the feature information of high-resolution remote sensing images, which makes them not widely applied in classification tasks.

In order to effectively solve the above problems and the lack of generalization ability and poor classification performance caused by a single feature, scholars have proposed a variety of feature fusion classification methods. Chen Xu et al. [6] proposed a texture classification algorithm based on the fusion features of GLCM and Tamura, which enhanced the robustness and classification performance of the algorithm by improving the rotation invariance of GLCM and reducing a large amount of redundant information. Zhang Qingchun et al. [7] used a multi-feature fusion algorithm to extract local entropy, texture features and other features to improve image classification performance. Wang Yu et al. [8] adopted a new spatial feature-the fusion of second-order moment feature and spectral feature-to achieve road refinement. Kang Jian et al. [9] used the RFB module to obtain high-level water body semantic information and multi-scale, integrated initial multi-scale features with original features in a deep level, completed the extraction of multi-scale features, and enhanced high-level water body semantic information features. These methods not only consider the global feature information, but also retain the shallow local information. Through the fusion of shallow information and global feature information, the generalization ability and classification performance of the algorithm are improved to a certain extent. However, this kind of feature fusion will increase the amount of computation, which leads to the increase of model complexity and the occurrence of overfitting. Therefore, PCA module is introduced in this algorithm, and appropriate principal component contribution rate is selected to remove redundant information and improve the classification performance of the model.

In recent years, due to the superiority of deep learning algorithm in image recognition, many scholars have introduced Convolutional Neural Network (CNN) into remote sensing image scene classification. Although CNN has achieved good results in the scene classification method, deep learning requires a large number of data labels, and it is difficult to obtain more reliable image feature information in the field of remote sensing image scene classification with fewer samples to learn. Transfer learning can obtain better results in scene classification under small sample condition by using ImageNet pre-training network. Han et al. [10] used the pre-trained AlexNet network combined with the spatial pyramid pooling method to improve the accuracy of scene classification. Some studies have shown that the effect of extracting CNN deep features for different ways of feature fusion and finally input the fusion features into SVM (Support Vector Machine) classification is better than that of CNN direct classification. Therefore, the Grid Search of SVM (GS-SVM) is selected as the final classifier in this paper. Li et al. [11] pointed out that the combination of pre-trained CNN features in scene classification showed better differentiation ability than the original CNN features. Lu et al. [12] adopted a Feature Aggregation Convolutional Neural Network (FACNN) applied to scene classification to learn image features by using pre-trained CNN as a feature extractor to explore semantic label information.

To sum up, traditional manual features rely on manual design, have weak feature expression ability, feature fusion increases model complexity and computational load, and deep learning has poor performance in the absence of a large number of data labels. Therefore, this paper proposes a remote sensing image scene classification method based on Guided Learning Local Binary Patterns Convolutional Neural Network (GL-LBP-CNN). Class 21 of UC Merced and 7 of RSSCN7 were classified. GLCM and LBP (Local Binary Patterns) were used to extract texture features with relevant spatial characteristics and shallow features of local texture features, and then AlexNet Convolutional Neural Network (AleCNN) was used to extract the depth features of the fully connected layer. After adaptive fusion of the extracted shallow features and deep features, redundant information was reduced through PCA. Finally, it was input into GS-SVM for classification.

2 Feature Extraction

High-resolution remote sensing image scene classification mainly uses image spatial information and a small amount of spectral information to identify the scene category of remote sensing image [13]. In this paper, when classifying remote sensing images, texture features with relevant spatial characteristics are extracted through Grey Level Co-occurrence Matrix. However, it is limited to focus only on the overall texture features. Local Binary Patterns can effectively extract local texture features of remote sensing images, and finally build a shallow texture feature that integrates the overall feature and local feature. Then the transfer learning module based on AlexNet network is added to extract the deep features of the remote sensing image, and finally the shallow features and deep features are fused to obtain the final features of the remote sensing image.

2.1 Gray co-occurrence matrix

Grey Level Co-occurrence Matrix (GLCM) reflects texture information such as direction, adjacent interval, and change amplitude in gray image [14]. Generally, the four most commonly used features are used to extract image texture features. The second moment reflects the uniformity of image gray distribution and the thickness of texture. Contrast reflects the depth of texture; Autocorrelation reflects the similarity of rows or columns of elements in a matrix $\boldsymbol{G}$; Entropy reflects the amount of information in an image.
The second moment:

$$M = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} G^2(i,j). \tag{1}$$

contrast ratio:

$$t = \sum_{n=0}^{L-1} n^2 \left\{ \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} G(i,j) \right\}. \tag{2}$$

autocorrelation:

$$c = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{ijG(i,j) - \mu_1\mu_2}{\sigma_1^2 \sigma_2^2}, \tag{3}$$

among them:

$$\mu_1 = \sum_{i=0}^{L-1} i \sum_{j=0}^{L-1} G(i,j)\,,$$

$$\mu_2 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} G(i,j)\,,$$

$$\sigma_1^2 = \sum_{i=0}^{L-1} (i-\mu_1)^2 \sum_{j=0}^{L-1} G(i,j)\,,$$

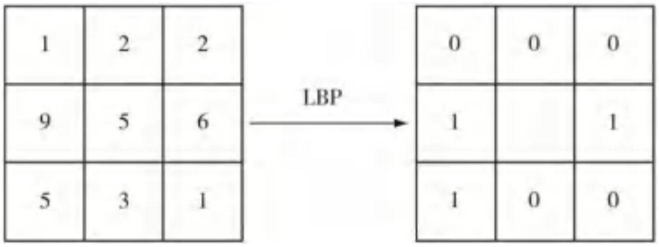
$$\sigma_2^2 = \sum_{i=0}^{L-1} (i-\mu_2)^2 \sum_{j=0}^{L-1} G(i,j).$$

entropy:

$$e = -\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} G(i,j) \log G\,(i,j). \tag{4}$$

2.2 Local Binary Patterns

The LBP operator is defined as taking the central pixel of each block as the threshold in a 3 × 3 block and comparing the gray value of the surrounding eight pixels with it. If the surrounding pixel value is greater than the central pixel value, the position of the pixel is marked as 1; otherwise, it is 0. In this way, the comparison of 8 points in the 3 × 3 neighborhood can produce 8 bits of binary number (usually converted to decimal LBP code), that is, the LBP value of the center pixel of the region is obtained, and this value is used to reflect the texture information of the region. As shown in Figure 1.



1	2	2
9	5	6
5	3	1

0	0	0
1		1
1	0	0

Figure 1: LBP diagram

DOI https://doi.org/10.4208/JICS-2024-005 | Generated on 2025-04-02 16:21:02

The original window can be represented as binary: 00010011; Decimal: 19. Expressed by the formula as

$$\mathrm{LBP}(x_c, y_c) = \sum_{p=0}^{p-1} 2^p s(i_p - i_c), \tag{5}$$

where, P represents the p pixel point in the 3×3 window except the central pixel point, i_p represents the gray value of the p pixel point in the field, and i_c represents the gray value of the central pixel point.

The formula for $s(x)$ in formula (5) is as

$$s(x) = \begin{cases} 1, x \geq 0, \\ 0, x < 0, \end{cases} \tag{6}$$

where x is the value of $i_p - i_c$.

2.3 Transfer Learning

The object classification and recognition methods of deep learning under natural images have become more and more mature, and the classification model under natural images is applied to the feature extraction of remote sensing images, which can solve the problem of difficult training due to the lack of scene classification training data of remote sensing images to a certain extent [15].

Compared with other CNN architectures with complex structure and depth (such as GoogLeNet, VGG-16, etc.), AlexNet is a CNN architecture with simple structure, which is easy to train and optimize [10]. In consideration of the balance between model computation and accuracy, the AlexNet network is selected for the experiment [16]. The AlexNet network model has five convolutional layers and three fully connected layers, among which the last fully connected layer is 1 000 dimensional, and a 1×256 fully connected layer is added when the fully connected layer is removed. Finally, the output 256-dimensional vector is used as the extracted remote sensing image feature. The original AlexNet network model is shown in Figure 2 [10].

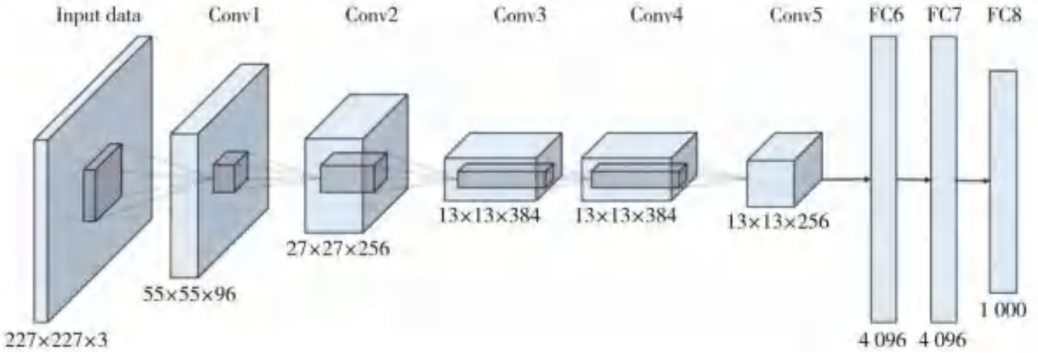


Figure 2: AleNet network framework

2 Methods of this paper

The GL-LBP-CNN method is proposed by using GS-SVM (Grid Search Support Vector Machine) as a classifier.

DOI https://doi.org/10.4208/JICS-2024-005 | Generated on 2025-04-02 16:21:02
OPEN ACCESS

2.1 GS-SVM model

The mesh search algorithm is applied to the parameter optimization of SVM. It optimizes the parameters of the estimation function by cross-validation method to obtain the optimal learning algorithm [17].

Let the training set $T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_i, y_i)\} \in (X \times Y)^l$, where $\boldsymbol{x}_i$ is the eigenvector and y_i is the label corresponding to $\boldsymbol{x}_i$.

Selecting appropriate kernel function $\kappa(\boldsymbol{x}, \boldsymbol{x}')$ and parameter C, Lagrange multiplier α is introduced to solve the optimization problem:

$$\underset{\alpha}{\mathrm{Max}} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \kappa(\boldsymbol{x}, \boldsymbol{x}'),$$
$$\text{s.t. } \sum_{i=1}^{l} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1,2, \cdots, l. \tag{7}$$

the optimal solution is obtained: $\boldsymbol{\alpha}^* = (\alpha_1^*, \cdots, \alpha_l^*)^{\mathrm{T}}$.
choose any component of $\boldsymbol{\alpha}^*$ and calculate the threshold b^*:

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* \kappa(\boldsymbol{x}, \boldsymbol{x}_i). \tag{8}$$

RBF kernel function is used to improve SVM and map samples to high-dimensional space, which can better handle nonlinear data and improve the accuracy of SVM classification. The RBF kernel function $\kappa(\boldsymbol{x}, \boldsymbol{x}_i)$ is

$$\kappa(\boldsymbol{x}, \boldsymbol{x}_i) = \exp(-\mathrm{g}\|(\boldsymbol{x} - \boldsymbol{x}_i)^2\|), \tag{9}$$

where g is the radius of the kernel function.

Construct the decision function:

$$f(\boldsymbol{x}) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i^* y_i \kappa(\boldsymbol{x}, \boldsymbol{x}_i) + \boldsymbol{b}^*\right). \tag{10}$$

When SVM performs image classification prediction, it is usually necessary to select appropriate relevant parameters C and g to obtain better classification accuracy. The method adopted in this paper aims to obtain the optimal model parameters by optimizing the parameters.

2.2 Methods and Procedures

This paper combines GL-LBP-CNN method and GS-SVM algorithm based on feature fusion, and the specific steps are as follows.

1) Global texture features and local texture features with spatial information were extracted by GLCM and LBP respectively;

2) Deep features of remote sensing images were extracted by AleCNN;

3) The extracted GLCM features, LBP features and deep features are adaptive fusion;

4) Input fusion features into GS-SVM algorithm to classify remote sensing image scenes.

The method flow of this paper is shown in Figure 3.

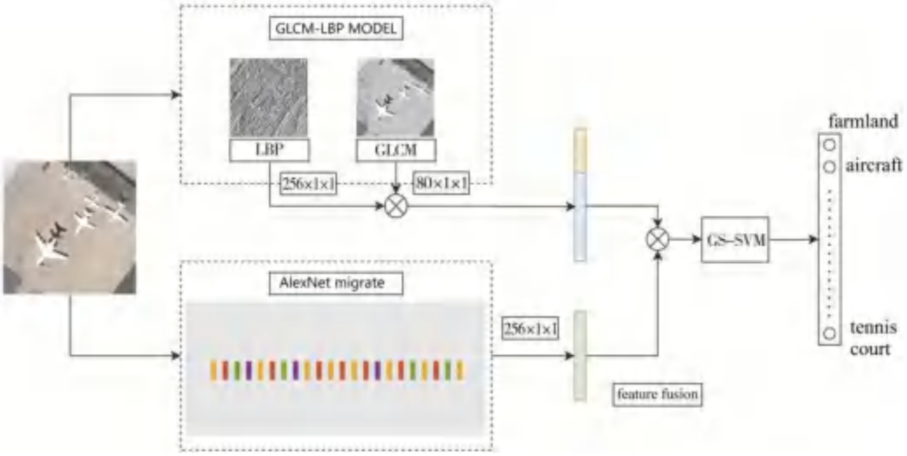


Figure 3: Flowchart of GL-LBP-CNN approach

3 Experimental results and analysis

3.1 Data Set

The UC Merced dataset comes from the United States Geological Survey, including 21 categories such as aircraft and golf courses, each category contains 100 images, the image size is 256 × 256, and the image resolution is 0.3 m. RSSCN7 dataset contains 2 800 remote sensing images, which are from seven typical scene categories. They are meadows, fields, industries, rivers and lakes, forests, residential areas and parking lots, each of which contains 400 images sampled on four different scales. The sample size of the above two datasets is much smaller than that of imageNet, open images and other datasets commonly used in deep learning. Examples of images for each class of the two data sets are shown in Figures 4 and 5, respectively.

UC Merced data set link: http: / / weegee.vision.ucmerced.edu / datasets / landuse. html.

RSSCN7 data set link: https: / / hyper.ai / datasets / 5440.

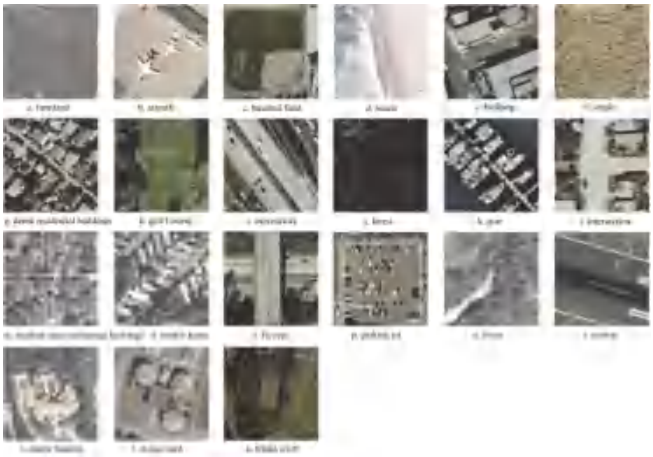


Figure 4: Samples of UCM dataset

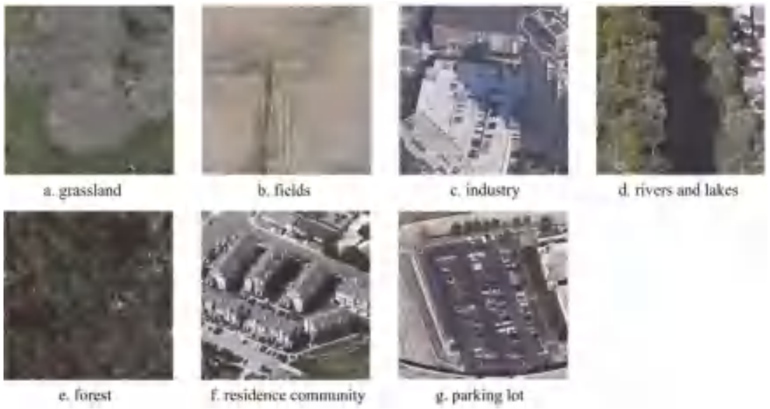

a. grassland
b. fields
c. industry
d. rivers and lakes
e. forest
f. residence community
g. parking lot

Figure 5: Samples of RSSCN7 dataset

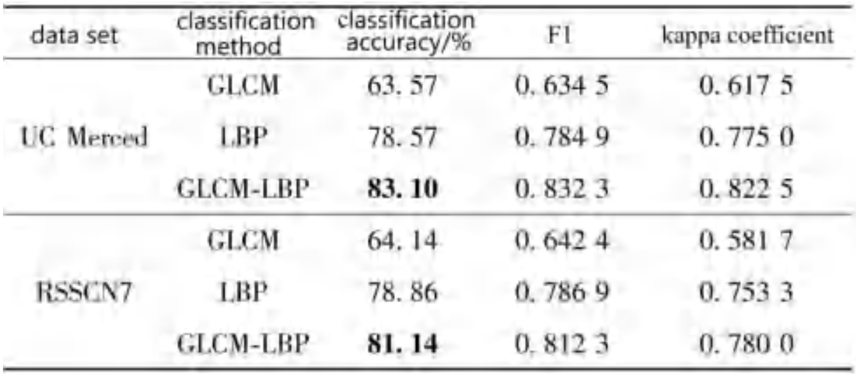
3.2 Experimental Setup

The experiments were carried out in MATLAB R2019b with AMD Ryzen5 3500X processor model. In the model experiment of this paper, the training set and test set on UC Merced data set are 80% and 20% of each class respectively, and the training set and test set on RSSCN7 data set are 50% of each class. AlexNet pre-training network is based on ImageNet data set, the batch size is 10 in the experiment, and the initial learning rate is 0.0001. The number of iterations is ten. In the experiment, classification accuracy, time cost, confusion matrix, F1 and kappa coefficient were used as evaluation indexes.

3.3 Comparative experiment

To compare the performance of the manual feature module compared with a single manual feature, the classification accuracy, F1 and kappa coefficients of the two benchmark datasets were compared by GLCM-LBP, GLCM and LBP manual features. The results are shown in Table 1.

Table 1: Comparison of classification accuracy, F1, and kappa coefficient between manual feature modules



data set	classification method	classification accuracy/%	F1	kappa coefficient
UC Merced	GLCM	63. 57	0. 634 5	0. 617 5
	LBP	78. 57	0. 784 9	0. 775 0
	GLCM-LBP	**83. 10**	0. 832 3	0. 822 5
RSSCN7	GLCM	64. 14	0. 642 4	0. 581 7
	LBP	78. 86	0. 786 9	0. 753 3
	GLCM-LBP	**81. 14**	0. 812 3	0. 780 0

Note: 95% of the principal components of GLCM, LBP and GLCM-LBP are selected, with bold text indicating the highest classification accuracy.

DOI https://doi.org/10.4208/JICS-2024-005 | Generated on 2025-04-02 16:21:02

As can be seen from Table 1, compared with the single feature of GLCM and LBP in UC Merced dataset, the classification accuracy of GLCM-LBP module is increased by 19.53 percentage points and 4.53 percentage points respectively, and the F1 score and kappa coefficient are improved to different degrees. In the RSSCN7 dataset, the classification accuracy of GLCM-LBP module is improved by 17 percentage points and 2.28 percentage points respectively compared with the single feature of GLCM and LBP, and also has better performance in F1 score and kappa coefficient. The results show that the ability of the traditional manual feature extraction shallow features to express image information after fusion is better than that of single features, and the two different shallow features can complement each other, thus enhancing the generalization ability and classification performance of the model. However, no matter GLCM-LBP module or the single manual feature such as GLCM and LBP, the classification performance is not good on the two benchmark data sets, indicating that the manual feature cannot effectively extract the higher-level semantic information of the image and effectively solve the high-resolution remote sensing scene image classification ability.

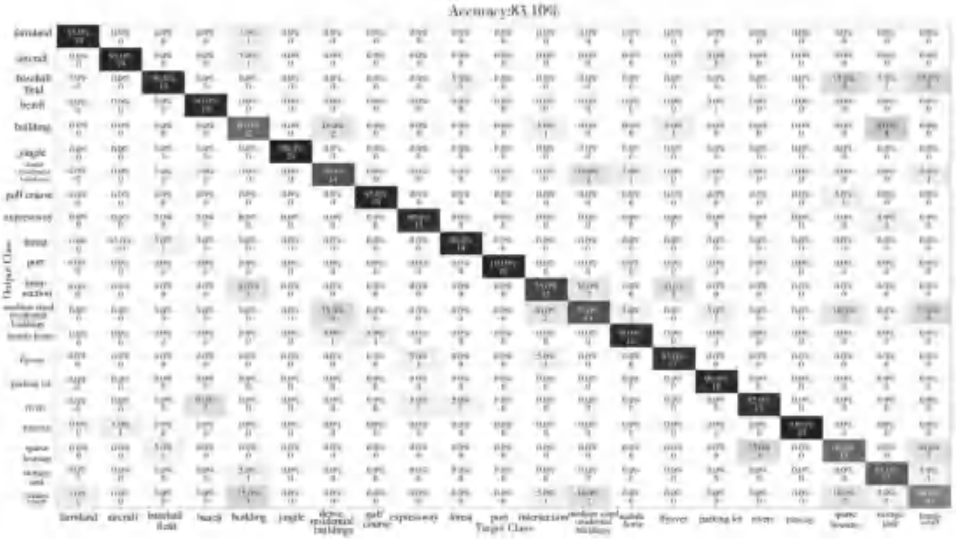
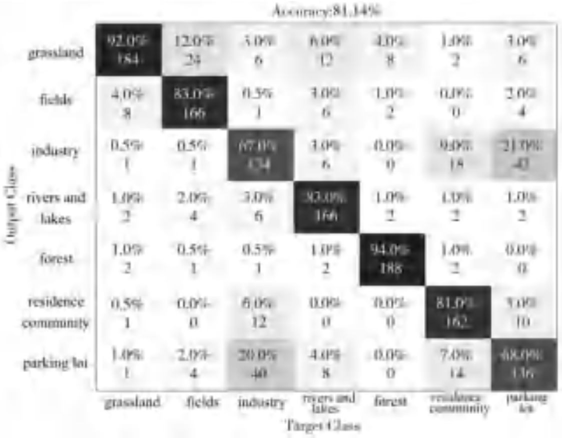


Figure 6: GLCM-LBP obfuscation matrix in the UC Merced dataset

Figures 6 and 7 show the confusion matrix of GLCM-LBP module in UC Merced data set and RSSCN7 data set respectively. According to the confusion matrix of UC Merced data set, the classification accuracy of seven categories, namely buildings, dense residences, intersections, medium residences, sparse residences, oil storage tanks and tennis courts, did not exceed 83.1%, and these categories have complex image features. Moreover, buildings, dense houses, medium houses and sparse houses have the characteristics of high inter-class similarity and low intra-class similarity, which greatly interferes with the performance of model classification. Most of the misclassified images in the dense housing category were misclassified into the other three categories. According to the confusion matrix of RSSCN7 data set, the classification accuracy of the three categories of industry, residential area and parking area did not exceed 81.14%.

DOI https://doi.org/10.4208/JICS-2024-005 | Generated on 2025-04-02 16:21:02

Among them, 78.79% of the images with wrong classification of industrial area were wrongly classified into residential area and parking area, and 84.21% of the images with wrong classification of residential area were wrongly classified into industrial area and parking area. 81.25% of the images misclassified by parking categories were misclassified to industrial and residential areas. Therefore, traditional manual features cannot fully express the information of complex feature images, and cannot effectively solve the problem of high inter-class similarity and low intra-class similarity.

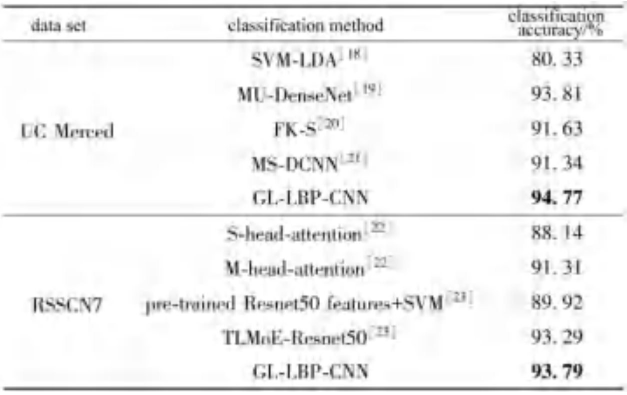


	grassland	fields	industry	rivers and lakes	forest	residence community	parking lot
grassland	92.0% 184	12.0% 24	3.0% 6	6.0% 12	4.0% 8	1.0% 2	3.0% 6
fields	4.0% 8	83.0% 166	0.5% 1	3.0% 6	1.0% 2	0.0% 0	2.0% 4
industry	0.5% 1	0.5% 1	67.0% 134	3.0% 6	0.0% 0	9.0% 18	21.0% 42
rivers and lakes	1.0% 2	2.0% 4	3.0% 6	83.0% 166	1.0% 2	1.0% 2	1.0% 2
forest	1.0% 2	0.5% 1	0.5% 1	1.0% 2	94.0% 188	1.0% 2	0.0% 0
residence community	0.5% 1	0.0% 0	6.0% 12	0.0% 0	0.0% 0	81.0% 162	5.0% 10
parking lot	1.0% 1	2.0% 4	20.0% 40	4.0% 8	0.0% 0	7.0% 14	68.0% 136

Figure 7: GLCM-LBP obfuscation matrix in the RSSCN7 dataset

The GL-LBP-CNN method was compared with SVM-LDA [18], MU-DenseNet [19], FK-S [20] and MS-DCNN [21] on the UC Merced dataset. Compared with S-head-attention [22], M-head-attention [22], pre-trained Resnet50 features+SVM [23] and TLMoE-Resnet50 [23] on RSSCN7 dataset, The results are shown in Table 2.

Table 2: Comparison of classification accuracy between different methods



data set	classification method	classification accuracy/%
UC Merced	SVM-LDA[18]	80.33
	MU-DenseNet[19]	93.81
	FK-S[20]	91.63
	MS-DCNN[21]	91.34
	GL-LBP-CNN	**94.77**
RSSCN7	S-head-attention[22]	88.14
	M-head-attention[22]	91.31
	pre-trained Resnet50 features+SVM[23]	89.92
	TLMoE-Resnet50[23]	93.29
	GL-LBP-CNN	**93.79**

Note: The principal component contribution rates of GL-LBP-CNN in UC Merced data set and RSSCN7 data set were 95% and 90% respectively, with bold text indicating the highest classification accuracy.

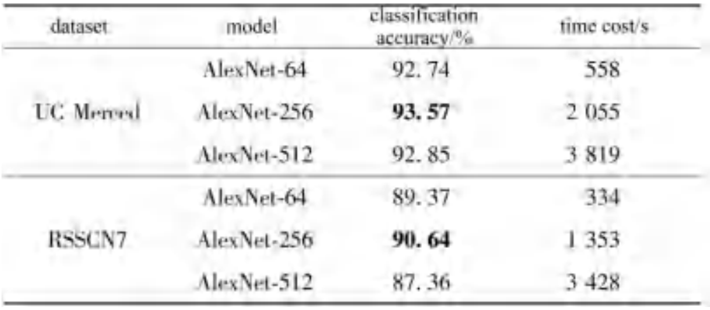
As can be seen from Table 2, the classification accuracy of GL-LBP-CNN on UC Merced and RSSCN7 data sets is 94.77% and 93.79% respectively, both of which are better than the comparison method. This method not only has the superiority in classification accuracy, but also has strong generalization ability, and can be applied to various scene classification data sets.

DOI https://doi.org/10.4208/JICS-2024-005 | Generated on 2025-04-02 16:21:02

3.4 Ablation experiment

In order to further explore the ability of pre-trained AlexNet's fully connected layer to express image information in different dimensions, the experiment compared the output features of 64-dimensional, 256-dimensional and 512-dimensional fully connected layers, and the results were shown in Table 3.

Table 3: Comparison of classification accuracy and time overhead of AlexNet for different fully connected layers

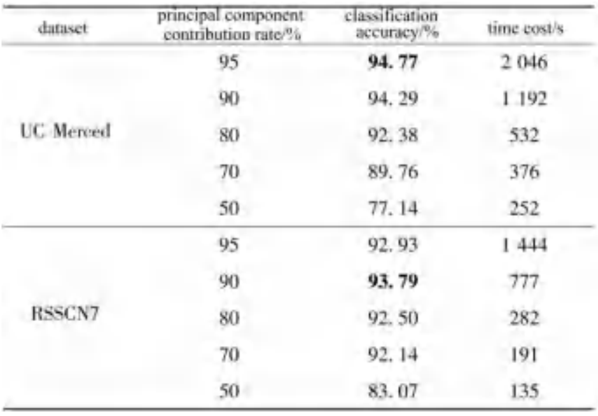


dataset	model	classification accuracy/%	time cost/s
UC Merced	AlexNet-64	92. 74	558
	AlexNet-256	**93. 57**	2 055
	AlexNet-512	92. 85	3 819
RSSCN7	AlexNet-64	89. 37	334
	AlexNet-256	**90. 64**	1 353
	AlexNet-512	87. 36	3 428

Note: 95% of the principal components are selected, and bold text indicates the highest classification accuracy.

As can be seen from Table 3, in UC Merced data set and RSSCN7 data set, the dimension of the full connection layer of pre-trained AlexNet is proportional to the time cost, and the multiples are basically consistent. In both datasets, the classification accuracy of AlexNet-256 is better than that of AlexNet-64 and AlexNet-512. The results show that the output feature dimension of the fully connected layer is too high, it will produce some redundant information, and the effect in SVM classifier is not ideal. After feature fusion, the output feature dimension of scene image increases, and the complexity and computation amount of the model are increased, which greatly increases the time cost of the model. In this paper, the effects of different principal component contribution rates on GL-LBP-CNN classification accuracy and time cost are compared to obtain a better parameter. The results are shown in Table 4.

Table 4: Comparison of classification accuracy and time overhead under different principal component contributions



dataset	principal component contribution rate/%	classification accuracy/%	time cost/s
UC Merced	95	**94. 77**	2 046
	90	94. 29	1 192
	80	92. 38	532
	70	89. 76	376
	50	77. 14	252
RSSCN7	95	92. 93	1 444
	90	**93. 79**	777
	80	92. 50	282
	70	92. 14	191
	50	83. 07	135

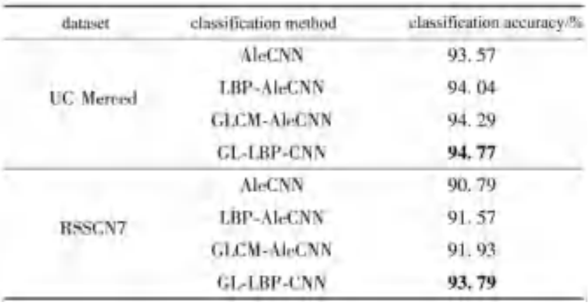
As can be seen from Table 4, on the UC Merced dataset, the classification accuracy reached 94.77% when the principal component contribution rate was 95%; when the principal component contribution rate was 50%, the classification accuracy was only 77.14%, indicating that too much effective information was lost in the process of PCA dimension reduction. In the process of weighing time overhead and classification accuracy, classification accuracy is crucial, so 95% principal component contribution rate is selected.

On RSSCN7 data set, when the principal component contribution rate is 90%, the classification accuracy is the highest (93.79%); when the principal component contribution rate is 50%, the classification accuracy is only 83.07%. Considering the classification accuracy and time cost, 90% principal component contribution rate is finally selected.

The above experimental results show that feature fusion will generate redundant information while increasing the complexity and computation of the model, which will reduce the classification accuracy of the model to some extent. The PCA method introduced in this paper can effectively reduce the redundant information and the calculation amount of the model, and reduce the time cost while ensuring the classification accuracy.

In order to explore the effectiveness of shallow layer information in the GL-LBP-CNN method on improving the scene classification performance of the pre-trained network, the single manual features of GLCM and LBP are merged with AleCNN respectively, and the results are shown in Tables 5 and 6.

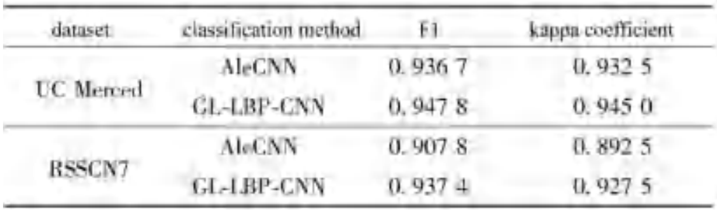
Table 5: Comparison of classification accuracy between different feature fusion methods



dataset	classification method	classification accuracy/%
UC Merced	AleCNN	93. 57
	LBP-AleCNN	94. 04
	GLCM-AleCNN	94. 29
	GL-LBP-CNN	**94. 77**
RSSCN7	AleCNN	90. 79
	LBP-AleCNN	91. 57
	GLCM-AleCNN	91. 93
	GL-LBP-CNN	**93. 79**

Note: 95% of the principal components of UC Merced data set and 90% of the principal components of RSSCN7 data set were selected, with bold text indicating the highest classification accuracy.

Table 6: Comparison of F1 and kappa coefficients between pre-trained AleCNN and GL-LBP-CNN



dataset	classification method	F1	kappa coefficient
UC Merced	AleCNN	0. 936 7	0. 932 5
	GL-LBP-CNN	0. 947 8	0. 945 0
RSSCN7	AleCNN	0. 907 8	0. 892 5
	GL-LBP-CNN	0. 937 4	0. 927 5

Note: The principal component of UC Merced data set was 95%, and that of RSSCN7 data set was 90%.

As can be seen from Table 5, different manual features complement the image information extracted by AleCNN to different degrees. On UC Merced dataset, the classification accuracy of LBP-AleCNN and GLCM-AleCNN is better than that of AleCNN. Although the classification accuracy of GLCM is lower than that of LBP, it is slightly better than that of LBP in the feature information supplement of AleCNN, indicating that the two texture features have different information supplement when they act on deep semantic features. On RSSCN7 dataset, the classification accuracy of LBP-AleCNN and GLCM-AleCNN is also better than that of AleCNN. The experimental results show that the shallow feature is effective for improving the AleCNN scene classification ability. As can be seen from Table 6, compared with pre-trained AleCNN, GL-LBP-CNN achieved better results on F1 and kappa coefficients on the two data sets, indicating that the GL-LBP-CNN method achieved higher accuracy in the accuracy and recall rate of each class, and the predicted results were in good agreement with the actual classification results.

Figures 8 and 9 show the confusion matrix of GL-LBP-CNN method in UC Merced data set and RSSCN7 data set respectively. By comparing Figure 8 with Figure 6, it can be seen that the classification accuracy of seven categories, namely buildings, dense houses, intersections, medium-sized houses, sparse houses, oil storage tanks and tennis courts, has improved to varying degrees, and the average classification accuracy of the seven categories has increased from 64.29% to 87.14%. By comparing Figure 9 with Figure 7, it can be seen that the classification performance of the three categories of industry, residential area and parking area has been greatly improved. Among them, the classification error rate of residential area is only 0.5%, and it is not wrongly classified into industry and parking area; the classification error rate of parking area is only 8%, and the images wrongly classified into industry and residential area have been greatly reduced. It is shown that the GL-LBP-CNN method greatly increases the ability of the model to express complex image information, and effectively solves the problem of high inter-class similarity and low intra-class similarity.

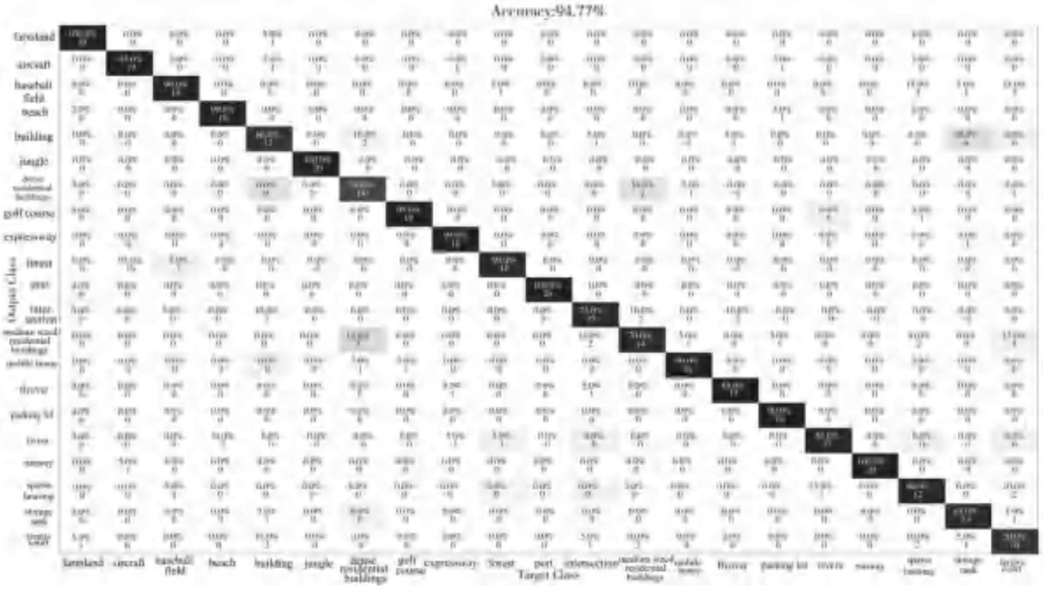
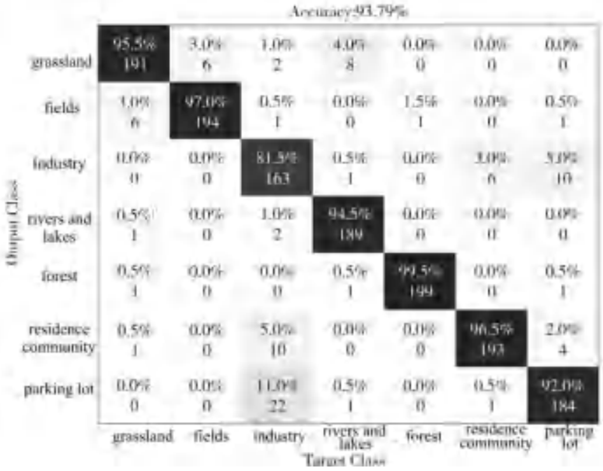


Figure 8: GL-LBP-CNN obfuscation matrix in the UC Merced dataset

DOI https://doi.org/10.4208/JICS-2024-005 | Generated on 2025-04-02 16:21:02



Accuracy93.79%

Output Class \ Target Class	grassland	fields	industry	rivers and lakes	forest	residence community	parking lot
grassland	95.5% 191	3.0% 6	1.0% 2	4.0% 8	0.0% 0	0.0% 0	0.0% 0
fields	3.0% 6	97.0% 194	0.5% 1	0.0% 0	1.5% 1	0.0% 0	0.5% 1
industry	0.0% 0	0.0% 0	81.5% 163	0.5% 1	0.0% 0	3.0% 6	5.0% 10
rivers and lakes	0.5% 1	0.0% 0	1.0% 2	94.5% 189	0.0% 0	0.0% 0	0.0% 0
forest	0.5% 1	0.0% 0	0.0% 0	0.5% 1	99.5% 199	0.0% 0	0.5% 1
residence community	0.5% 1	0.0% 0	5.0% 10	0.0% 0	0.0% 0	96.5% 193	2.0% 4
parking lot	0.0% 0	0.0% 0	11.0% 22	0.5% 1	0.0% 0	0.5% 1	92.0% 184

Figure 9: GL-LBP-CNN obfuscation matrix in the RSSCN7 dataset

4 Conclusion

The GL-LBP-CNN method proposed in this paper can not only extract the shallow features combined with the whole image and the local image, but also integrate the depth features extracted by AleCNN. With PCA dimensionality reduction, problems such as the increase of computation amount after the expansion of feature dimension are effectively solved, and the classification efficiency of the algorithm is improved without affecting the classification accuracy. SVM is optimized by grid search to improve the classification performance of the classifier. The experimental results on UC-Merced dataset and RSSCN7 dataset show that the proposed model is superior to the comparison method, with the average classification accuracy of 94.77% and 93.79%, respectively. In the future, we will further optimize the structure of the pre-trained network and design more lightweight and efficient methods from the perspective of channel attention, multi-scale feature fusion and optimization classifier.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 61774137), the Shanxi Basic Research Program (Grant No. 202103021224195, 202103021224212, 202103021223189, 20210302123019) and the Research Project of Returned overseas Students in Shanxi Province (Grant No. 2020-104, 2021-108).

References

[1] B. Zhao, Y. F. Zhong, G. S. Xia, et al., Dirichlet-derived multiple topic scene classification model

DOI https://doi.org/10.4208/JICS-2024-005 | Generated on 2025-04-02 16:21:02

for high spatial resolution remote sensing imagery, *IEEE Trans. Geosci. Remote Sens.*, 2016, 54(4): 2108-2123.

[2] S. Z. Chen, and Y. L. Tian, Pyramid of spatial relations for scene-level land use classification, *IEEE Trans. Geosci. Remote Sens.*, 2015, 53(4): 1947-1957.

[3] D. H. Yu, H. T. Guo, Q. Xu, et al., Hierarchical attention and bilinear fusion for remote sensing image scene classification, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2020, 13: 6372-6383.

[4] F. Y. Li, and F. M. Ye, Summarization of SIFT-based remote sensing image registration techniques, *Remote Sens. Land Resour.*, 2016, 28(2): 14-20.

[5] J. Y. Xu, F. B. Xu, Y. Q. Zhang, et al., Monitoring suspected pollution on unutilized land using gray-level co-occurrence matrices, *J. Beijing Univ. Technol.*, 2018, 44(11): 1423-1433.

[6] X. Chen, Y. Z. Gao, S. J. Chen, et al., Classification of textured materials based on T-GLCM and Tamura fusion features, *J. Nanjing Univ. Inf. Sci. Technol. (Nat. Sci. Ed.)*, 2022, 1-11.

[7] Q. C. Zhang, G. F. Tong, Y. Li, et al., River detection in remote sensing images based on multi-feature fusion and soft voting, *Acta Opt. Sinica*, 2018, 38(6): 320-326.

[8] Y. Wang, H. Y. He, W. Tan, et al., High resolution remote sensing image road extraction algorithm based on multi-feature fusion, *Remote Sens. Inf.*, 2019, 34(1): 111-116.

[9] J. Kang, H. Y. Guan, Y. T. Yu, et al., RFA-LinkNet: a novel deep learning network for water body extraction from high-resolution remote sensing images, *J. Nanjing Univ. Inf. Sci. Technol. (Nat. Sci. Ed.)*, 2023, 15(2): 160-168.

[10] X. B. Han, Y. F. Zhong, L. Q. Cao, et al., Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification, *Remote Sens.*, 2017, 9(8): 848.

[11] E. Z. Li, J. S. Xia, P. J. Du, et al., Integrating multilayer features of convolutional neural networks for remote sensing scene classification, *IEEE Trans. Geosci. Remote Sens.*, 2017, 55(10): 5653-5665.

[12] X. Q. Lu, H. Sun, and X. T. Zheng, A feature aggregation convolutional neural network for remote sensing scene classification, *IEEE Trans. Geosci. Remote Sens.*, 2019, 57(10): 7894-7906.

[13] X. L. Qian, J. Li, G. Cheng, et al., Evaluation of the effect of feature extraction strategy on the performance of high-resolution remote sensing image scene classification, *J. Remote Sens.*, 2018, 22(5): 758-776.

[14] G. T. Nie, and H. Huang, A survey of object detection in optical remote sensing images, *Acta Autom. Sinica*, 2021, 47(8): 1749-1768.

[15] D. H. Yu, B. M. Zhang, C. Zhao, et al., Scene classification of remote sensing image using ensemble convolutional neural network, *J. Remote Sens.*, 2020, 24(6): 717-727.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Commun. ACM*, 2017, 60(6): 84-90.

[17] C. Zhang, Y. C. Hou, Y. Q. Jiao, et al., Research on concrete image classification based on three-channel separation feature fusion and support vector machine, *J. Graphics*, 2021, 42(6): 917-923.

[18] F. Zhang, B. Du, and L. P. Zhang, Saliency-guided unsupervised feature learning for scene classification, *IEEE Trans. Geosci. Remote Sens.*, 2015, 53(4): 2175-2184.

[19] Y. Y. Zhang, B. H. Zhang, Y. F. Zhao, et al., Remote sensing image classification based on dual-channel deep dense feature fusion, *Laser Technol.*, 2021, 45(1): 73-79.

[20] B. Zhao, Y. F. Zhong, L. P. Zhang, et al., The fisher kernel coding framework for high spatial resolution scene classification, *Remote Sens.*, 2016, 8(2): 157.

[21] S. H. Xu, X. D. Mu, P. Zhao, et al., Scene classification of remote sensing image based on multi-

scale feature and deep neural network, *Acta Geod. Cartogr. Sinica*, 2016, 45(7): 834-840.

[22] Y. F. Li, X. J. Fan, X. B. Yang, et al., Remote sensing image classification framework based on self-attention convolutional neural network, *J. Beijing Forestry Univ.*, 2021, 43(10): 81-88.

[23] X. Gong, Z. L. Chen, L. Wu, et al., Transfer learning based mixture of experts classification model for high-resolution remote sensing scene classification, *Acta Opt. Sinica*, 2021, 41(23): 2301003.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Global Science Press and / or the editor(s). Global Science Press and / or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

DOI https://doi.org/10.4208/JICS-2024-005 | Generated on 2025-04-02 16:21:02