Taylor & Francis
Taylor & Francis Group

International Journal of Remote Sensing

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tres20

A lightweight convolution neural network based on joint features for Remote Sensing scene image classification

Cuiping Shi, Xinlei Zhang, Liguo Wang & Zhan Jin

To cite this article: Cuiping Shi, Xinlei Zhang, Liguo Wang & Zhan Jin (2023) A lightweight convolution neural network based on joint features for Remote Sensing scene image classification, International Journal of Remote Sensing, 44:21, 6615-6641, DOI: 10.1080/01431161.2023.2273246

To link to this article: https://doi.org/10.1080/01431161.2023.2273246

Published online: 09 Nov 2023.

INTERNATIONAL JOURNAL OF REMOTE SENSING
2023, VOL. 44, NO. 21, 6615–6641
https://doi.org/10.1080/01431161.2023.2273246

Taylor & Francis
Taylor & Francis Group

Check for updates

A lightweight convolution neural network based on joint features for Remote Sensing scene image classification

Cuiping Shi[a], Xinlei Zhang[b], Liguo Wang[c] and Zhan Jin[b]

[a]College of Information Engineering, Huzhou University, Huzhou, China; [b]College of Electronic and Communication Engineering, Qiqihar University, Qiqihar, China; [c]College of Information and Communication Engineering, Dalian Nationalities University, Dalian, China

ABSTRACT

Unlike natural images, remote sensing scene images usually contain one scene label and many object labels, and many object labels are arranged dispersedly, which brings great difficulties to feature extraction of scene label. To accurately identify scene labels from remote sensing scene images with multiple object labels, it is important to fully understand the global context of the image. In order to solve the challenges of multi-label scene images and improve the classification performance, a global context feature extraction module is proposed in this paper. The module combines the semantics information of different regions through a global pooling and three different scale sub-regions pooling, which makes the module have stronger ability of global feature representation. In addition, in order to fully understand the semantic content of remote sensing images, a three branch joint feature extraction module is constructed, which consists of the global context feature module, 3×3 convolution branch and identity branch are fused. Finally, a lightweight convolution neural network based on joint features (LCNN-JF) is constructed using traditional convolution, depthwise separable convolution, joint feature extraction module and classifier for remote sensing scene image classification. A series of experimental results on four datasets, UCM, AID, RSSCN and NWPU, demonstrate that the proposed method has better feature representation ability and can achieve better classification of remote sensing scene images.

ARTICLE HISTORY
Received 15 February 2023
Accepted 8 October 2023

KEYWORDS
global context feature; joint feature; lightweight convolutional neural network (LCNN); remote sensing scene image

1. Introduction

With the development of remote-sensing satellite technology, the resolution of remote sensing images that can be used for research has been significantly improved. At present, high-resolution remote sensing images have been used in various fields, such as ground target recognition and detection (Han et al. 2015), urban mapping (Hua, Mou, and Zhu 2020), natural disaster damage assessment (Liang et al. 2020), land use (Hua, Mou, and Zhu 2019; Wang, Xiong, and Ning 2019), and so on. However, high-resolution remote sensing scene images have more complex spatial structure, and usually a scene label contains multiple object labels. The scene label refers to the category of the whole remote

CONTACT Cuiping Shi shicuiping@qqhru.edu.cn College of Information Engineering, Huzhou University, Huzhou, China

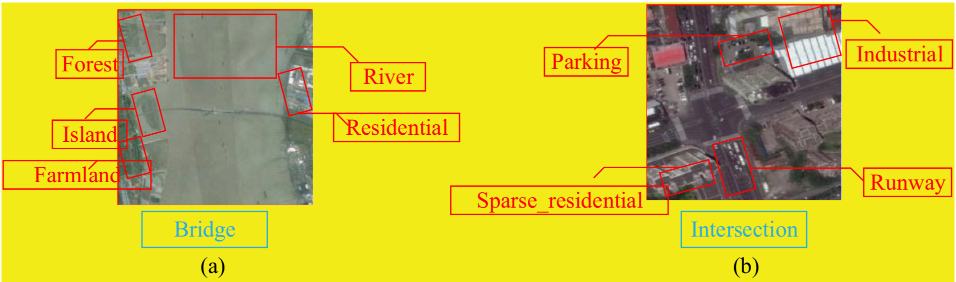


Figure 1. The complex spatial structure diagram of remote sensing scene image. Scene labels (shown in blue) and object labels (shown in red). (a) scene label: bridge, object label: forest, river, island, farmland, residential. (b) scene labels: intersection, object labels:parking,industrial, sparse_residential and runway.

sensing scene image. Object labels are labels for multiple objects contained in this scene image. It is difficult to understand the semantic content of scene images using local features alone. As shown in Figure 1(a), in addition to the bridge scene label, there are many object labels such as 'islands', 'rivers', and so on. In Figure 1(b), in addition to intersection scene labels, there are many object labels such as 'factory' and 'highway'. Multiple object labels in remote sensing scene images pose a challenge to correct recognition of scene labels.

In order to improve the performance of remote sensing scene image classification, many methods have been proposed by researchers. In the early days of image classification, feature extraction mainly relied on handmade feature descriptors. For example, Tian et al. (Tan et al. 2017) computed a low-rank constraint coefficient matrix using the low-rank representation in the image feature space, then uses the coefficient matrix to define the features and capture the global relationship between the images. Zhou et al. (Zhou et al. 2012) proposed a multi-label learning framework, in which each sample image is described by multiple labels, which can more fully represent complex scene images with multiple semantics. Wu et al. (Wu et al. 2019) proposed a framework for target detection of remote sensing images, which integrates feature extraction of channels, fast image pyramid matching and feature learning strategies. However, these hand-made feature representation methods have poor feature extraction ability, making it difficult to express the overall and advanced semantic features of remote sensing scene images. Using the handmade feature descriptor method to classify remote sensing scene images with rich spatial structure is not only a heavy task, but also a lack of adaptability and flexibility for different scenes, which results in insufficient classification performance and makes it difficult to meet the practical application of remote sensing scene classification.

In order to address the shortcomings of traditional feature extraction methods, machine learning-based feature extraction methods are gradually being widely used. In recent years, deep learning-based methods have become very popular, which are methods based on artificial neural networks for feature learning of input data. Among them, convolution neural network (CNN) is the most commonly used method in image classification tasks. In recent years, more convolution neural networks, such as VGG (Simonyan and Zisserman 2015), AlexNet (Krizhevsky, Sutskever, and Hinton 2017), ResNet (He et al. 2016), and MobleNet (Howard et al. 2017), which can learn more representative and distinctive features automatically, have been used in various fields of computer vision, such as target detection (Mahanand et al. 2021), semantic segmentation (Tasar, Tarabalka, and Alliez 2019), and scene classification (Wang et al. 2021). Due to the excellent performance of convolution neural network in computer vision, a large number of algorithms based on convolution neural network for remote sensing scene image classification have been proposed. For example, to explore semantic label information, Lu et al. (Lu, Sun, and Zheng 2019) proposed an end-to-end feature aggregation convolution neural network to learn the semantic information of remote sensing scene images using supervised convolution feature coding module and asymptotic aggregation strategy. This method integrates feature learning, feature aggregation and classifier into a unified end-to-end joint training framework, effectively improving the classification performance of convolution neural networks. He et al. (He et al. 2020) proposed a skip-layer connection covariance network based on convolution neural network. This method embeds the skip-layer connection module and the covariance pooling module into the traditional convolution neural network, effectively resolves the differences in the image data collection of remote sensing scenes, and can extract more representative features. To improve the performance of image classification for remote sensing scenes, Cheng et al. (Cheng et al. 2018) proposed a learning method of discriminant convolution neural network, which trains by optimizing a new discriminant objective function to make the image maps of different categories as far as possible and the image maps of the same scene category as close as possible. Cheng et al. (Cheng et al. 2018) applied convolution neural network to the classification of hyperspectral images and proposed a measurement-based learning framework to learn the spectral spatial characteristics of hyperspectral images. Jing et al. (Jing et al. 2020) proposed a method of image classification for remote sensing scene based on neural structure search. In addition, to improve the efficiency of neural search, an edge normalization technology was introduced into the algorithm. These methods are some creative improvement methods for remote sensing scene image classification. Most of them improve the classification performance of remote sensing scene images by only improving the network framework design and do not consider the characteristics of the image itself enough.

For remote sensing scene image classification, global context information is of great significance to improve the classification performance. In order to obtain global context features, most methods obtain global context features by using global average pooling or large convolution kernel. For example, the PoolNet (Liu et al. 2019) method used a modified pyramid pooling module to extract global context information. Wang et al. (Wang et al. 2018) used the convolution of three consecutive holes with different hole rates to enhance the receptive field of the network. Different from the existing methods, we use the four layer pyramid structure to extract the global features of remote sensing images at different scales. In addition, in order to improve the classification performance of remote sensing scene images, local features are added to the global feature extraction module, and identity connection is added to retain more shallow information. The shallow feature extraction module is composed of traditional convolution with convolution kernel size 3×3 and depthwise separable convolution with convolution kernel size 3×3. On the basis of the shallow feature extraction module, a joint feature extraction module is added to form the deep feature extraction module. Finally, a modular remote sensing scene image classification method is formed by combining the shallow feature extraction module and the deep feature extraction module.

In conclusion, the contributions of this paper can be summarized as the following three points.

(1) In order to obtain the global features of remote sensing images, we propose a four layer pyramid structure of global context feature extraction module. The module consists of four parts. The first part is to extract the global context information through average pooling and max pooling and to extract the global information of different scales by changing the size and step of the pooling kernel; the second part uses 1×1 convolution to preserve the weights of the global features extracted in the first part at different scales, while reducing the number of channels for the features; the third part is the upsampling layer, which restores the features of different scales to the same size as the original output through bilinear interpolation; the fourth part is the fusion layer, which fuses the global features of four different scales with the original input features to output the final global features.
(2) In order to further improve the classification performance, local features are added to the global feature extraction module to form a joint feature extraction module. This combination of local and global features can make the final prediction results more reliable. The joint feature extraction module consists of three branches, branch 2 is the global feature extraction module, and branch 3 is the local feature obtained by 3×3 convolution. To reduce the loss of shallow information during feature extraction, an identity branch is added.
(3) A modular remote sensing scene image classification method is presented. The method consists of shallow feature extraction module and deep feature extraction module. Shallow feature extraction module is composed of 3×3 traditional convolution and 3×3 depthwise separable convolution. The deep feature extraction module is composed of shallow module and joint feature extraction module. A series of experiments show that the proposed method can provide better classification performance.

The rest of this paper is as follows. In Section 2, the related work of global feature and multiscale feature networks is introduced. In Section 3, the global context feature extraction module, joint feature extraction module and lightweight convolutional neural network based on joint features (LCNN-JF) are introduced in detail. In Section 4, the proposed LCNN-JF method is compared with the advanced method. In Section 5, various visualization methods are used to discuss the performance of the proposed method. Section 6 is the conclusion of this paper.

2. Related works

2.1. Global features

To improve the perception field of the network, Liu et al. (Liu, Rabinovich, and Berg 2015) proposed ParseNet, which extracts the global context features of an image through global average pooling and fuses them with local features, which greatly improves the effectiveness of classification. Cao et al (Cao et al. 2019) overcomes the computational overhead of the traditional global attention module, presents a simplified global attention, and presents a GCNet by combining it with SEBlock. By capturing long-distance dependencies with simplified global self-attention, the method optimizes the ability of global context modelling while satisfying a relatively small amount of computations. The DeepLab network proposed by Chen et al. (Chen et al. 2017) uses four dilated convolution cascades to obtain global context information. Dilated convolution enlarges the field without increasing the number of parameters, and multiple dilated convolution cascades can achieve exponential growth of the field. To overcome the problem of effective weight reduction over long distances, a global average pooling layer is added after the last dilated convolution. Fu et al. (Fu et al. 2019) proposed a dual attention network, which uses dual attention to capture global context information to solve scene segmentation tasks. By introducing a self-attention mechanism to capture the spatial dependency between any two positions in the feature map, the global context information can be aggregated adaptively. In addition, channel attention is introduced to capture the channel dependency between any two feature maps.

2.2. Multi-scale feature network

Convolutional neural network extracts the features of the target through layer by layer abstraction. Although high-level features contain rich semantic information, it is difficult to accurately store the position information of objects due to low resolution. On the contrary, although the semantic information of low-level features is less, due to the high resolution, it can accurately contain the object position information. By fusing high-level semantic information and low-level semantic information to form a multi-scale feature network structure, the classification accuracy can be effectively improved. Multi-scale network structure is divided into multi-scale input, multi-scale feature fusion and multi-scale output.

2.2.1. Multi-scale input network

Multi-scale input network is to use images of multiple scales as input and then fuse the results. For example, in the Multi-task Cascaded Convolutional Networks (MTCNN) algorithm proposed by Zhang et al. (Zhang et al. 2016). In order to detect faces at the same scale, the original image is scaled to different scales before entering the network training, so as to enhance the robustness of the network to faces of different scales.

2.2.2. Multi-scale fusion network

The multi-scale fusion network consists of parallel multi-branch and serial hierarchical connections. Parallel structures can acquire features of different receptive fields at the same level, while serial structures can fuse features of different abstract levels. The inception module, proposed by Szegedy et al. in GoogleNet (Szegedy et al. 2015), is used by 1×1 convolution, 3×3 convolution, 5×5 convolution and 3×3 max pooling of four parallel branches, and finally fusing the four branches. The Feature Pyramid Network (FPN) algorithm proposed by Lin (Lin et al. 2017) et al. utilizes both low-level high-resolution features and high-level semantic features, and uses serial hierarchical connection method to fuse the features of different levels to achieve the prediction effect.

2.2.3. Multi-scale feature output network

Multi-scale feature output is to forecast in different feature scales and fuse the results. Liu et al. (Liu et al. 2016) predicted feature maps of different scales. The low-level feature map has a larger size and smaller field to detect small targets. The high-level feature map is small in size and has a large field to detect large targets. The Single Stage Headless Face Detector (SSHFD) algorithm proposed by Najibi et al. (Najibi et al. 2017) splits into branches starting from a larger resolution feature map, and each branch separately predicts targets of different scales.

3. Methodology

3.1. Global context feature extraction module

Convolution neural network can not perceive all the information of the original image because there is a local connection between the convolution layer and the pooling layer in the convolution structure. The larger the value of neural network receptive field is, the larger the range of the original image it can contact, which means that it contains more global information. The smaller the value is, the more local the features it contains. Zhou et al. (Zhou et al. 2014) proposed that the actual receptive field of convolutional neural networks is much smaller than the theoretical receptive field, resulting in many networks not fully integrating important global context information.

It is a simple method to obtain the global information of features by increasing the receptive field of convolution, but increasing the receptive field will bring a lot of computation, which will reduce the running speed of the network and is not conducive to optimization. Global pooling is a better method to extract global context information, but in remote sensing scene images, the feature map extracted by convolution contains rich spatial information. Directly using global pooling to fuse features into a single vector will lose the spatial information of features, and it is difficult to obtain good classification performance. To solve this problem, we propose a hierarchical global context information extraction module that includes the relationships between different scales and different sub regions. The hierarchical global context information extraction module combines the global context information and the local context information of sub regions, which is more helpful to distinguish various categories. As shown in Figure 2, the hierarchical global context information extraction module is composed of four parts. The first group is the global context features for extracting input features, which are composed of average pooling and maximum pooling. Assuming that the input feature is $U \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}\times\mathbb{C}}$, in group 1, the first level is the global context information generated by global pooling. Here, we use global pooling as global average pooling and global maximum pooling. Global average pooling takes the average value of all elements in the entire feature map and outputs it, so that more image background information can be retained. Maximum pooling is to take the maximum output of all elements in the entire feature map, discarding a large amount of redundant information in the network, and can retain more texture information of the image. Then, the output results of global average pooling and global maximum pooling are fused to obtain. This process can be expressed as

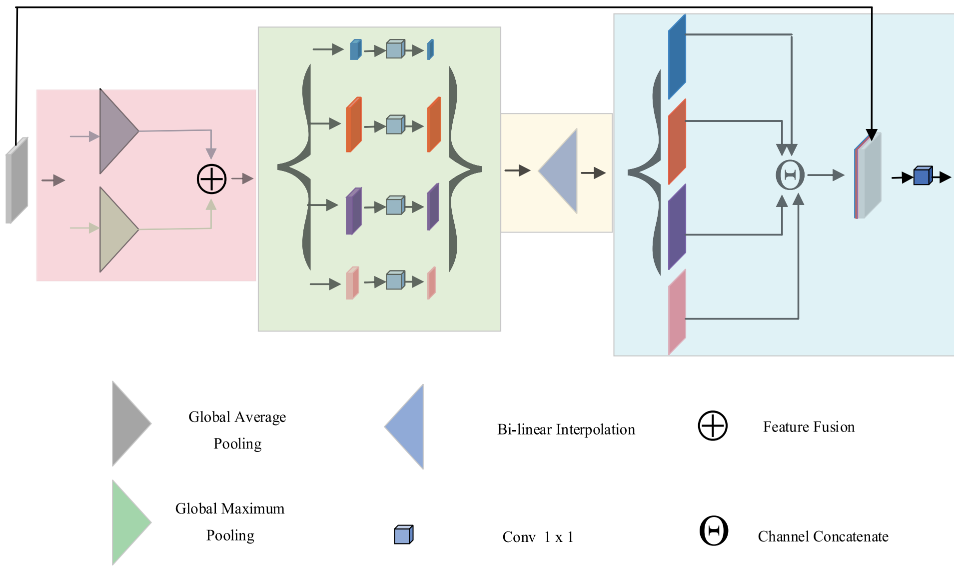


Figure 2. Global context feature module.

$$\mathbb{R}^{1\times1\times C} \leftarrow y_1 = \frac{1}{|\Re|}\sum_{(p,q)\in\Re} u_{pq} + \max_{(p,q)\in\Re} u_{pq} \tag{1}$$

In formula (1), $\frac{1}{|\Re|}\sum_{(p,q)\in\Re} u_{pq}$ represents the global average pooled output result, and $\max_{(p,q)\in\Re} u_{pq}$ represents the global max pooled output result. $|\Re|$ represents the number of all elements in the feature map, and u_{pq} represents the element at (p,q)in the rectangular area $\Re$.

The remaining three levels use pooled kernels with sizes of 2×2, 4×4 and 8×8 to extract the features of different sub regions, and the pooled outputs are$y_2 \in \mathbb{R}^{\frac{H}{2},\frac{W}{2},C}$, $y_3 \in \mathbb{R}^{\frac{H}{4},\frac{W}{4},C}$and $y_4 \in \mathbb{R}^{\frac{H}{8},\frac{W}{8},C}$, respectively. The local information$y_i$of each sub region can be expressed as

$$y_i = \frac{1}{|\Re_{mn}|}\sum_{(p,q)\in\Re_{mn}} u_{pq} + \max_{(p,q)\in\Re_{mn}} u_{pqi=2,3,4} \tag{2}$$

In formula (2), $\frac{1}{|\Re_{mn}|}\sum_{(p,q)\in\Re_{mn}} u_{pq}$ represents the output result of average pooling, and $\max_{(p,q)\in\Re_{mn}} u_{pq}$ represents the output result of max pooling. $\Re_{mn}$ refers to the rectangular area with the size of $m \times n$, which is 2×2, 4×4 and 8×8 respectively. u_{pq} represents the element at (p,q) in rectangular area $\Re_{mn}$, and $|\Re_{mn}|$ represents the number of elements in rectangular area $\Re_{mn}$.

Group 2 uses 1×1 convolution to maintain the weight with global context characteristics obtained by group 1. Specifically, in each level, the different scale features y_i with global context information obtained after pooling and fusion are passed through 1×1 to maintain the weight and reduce the channel dimension of multi-level features. After 1×1

convolution, we use batch normalization to speed up network training and convergence, and then use*Rule*activation function to carry out network non-linearity and improve the expression ability of the network. The obtained results$t_1 \in \mathbb{R}^{1\times 1\times C/4}$,$t_2 \in \mathbb{R}^{H/2\times W/2\times C/4}$, $t_3 \in \mathbb{R}^{H/4\times W/4\times C/4}$, $t_4 \in \mathbb{R}^{H/8\times W/8\times C/4}$ can be expressed as:

$$t_i = \iota\{\kappa[w(y_i)]\} \tag{3}$$

In formula (3), $w \in \mathbb{R}^{\mathbb{1}\times\mathbb{1}\times\mathbb{C}/\mathbb{4}}$, κ represent batch normalization, and ι represents*Relu*activation function.

In group 3, the 1×1 convoluted sample of low-dimensional features is restored to the size of the original feature map. Specifically, bi-linear interpolation is used to up-sample the low-dimensional feature t_i generated by each level, and the features after sampling are r_i, $t_i \underset{upsamping}{\longrightarrow} r_i \in \mathbb{R}^{H\times W\times C}$, *upsamping* representing the up-sampling operation.

The fourth group is the feature fusion stage. The original feature $U \in \mathbb{R}^{H\times W\times C}$ and the features r_1, r_2, r_3, r_4 obtained at four levels are spliced to obtain $V = U\Theta r_1\Theta r_2\Theta r_3\Theta r_4$, where Θ represents the channel concatenate operation. Finally, all the context information obtained from the channel concatenate is integrated by 1×1 convolution to obtain the final global feature. Through different pyramid levels, different scale features are obtained, and then these features are aggregated. Therefore, different scale context information is aggregated.

3.2. Joint feature extraction module

The joint feature extraction module combines global context features and local features, which is more helpful to distinguish various categories. As shown in Figure 3, this module is composed of three branches, branch 1 is the identity branch, branch 2 is the global context feature module (GCFM) branch, and branch 3 is the local feature extraction branch. Suppose the input feature map is $U \in \mathbb{R}^{\mathbb{W}\times\mathbb{H}\times\mathbb{C}}$. Specifically, branch 2 uses the

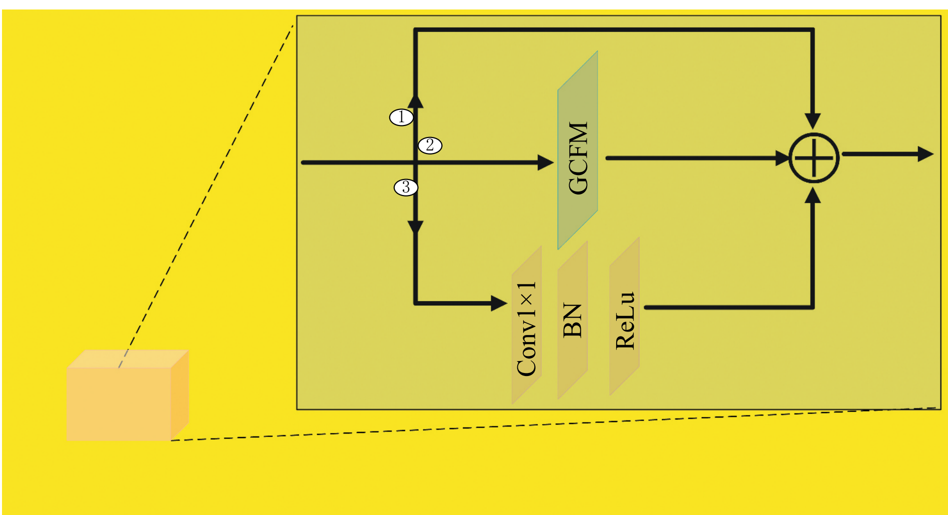


Figure 3. Joint feature extraction module.

pyramid pooling module with four levels to extract the extracted feature $V = U\Theta r_1\Theta r_2\Theta r_3\Theta r_4$, where r_1, r_2, r_3, r_4 are the hierarchical global context information extracted by the pyramid pooling module with four levels, which integrates the relationships between different scales and different sub regions, which is very important for understanding the semantic information of remote sensing scene images. Branch 3 is a local feature extracted using a convolution operation. Specifically, first, the input features are convoluted, and then, in order to speed up the convergence of the network, batch normalization is used after 3×3 convolution. Then, the modified linear unit (relu) activation function is used to perform nonlinear transformation on the features after batch normalization, so as to improve the representation ability of the extracted local features. The specific process is

$$\mathbb{R}^{W \times H \times C} \leftarrow T = \iota\{\kappa[w(U)]\} \tag{4}$$

In formula (4), T represents the local characteristics of the output of branch 3,$w \in \mathbb{R}^{1\times1\times C}$,$\kappa$represent batch normalization, andιrepresents the Relu activation function.

While extracting local and global features, branch 2 and branch 3 will inevitably lose some features. Here, branch 1 uses the identity branch to compensate the lost features, which also reduces the problem of performance degradation caused by network deepening. Finally, branch 1, branch 2 and branch 3 are fused to obtain the final joint feature $\mathbb{R}^{W \times H \times C} \leftarrow F = V \oplus T \oplus U$, where$\oplus$represents feature fusion.

3.3. Lightweight convolutional neural network based on joint features(LCNN-JF)

In order to find a lightweight convolutional neural network with balanced classification accuracy and running speed, we use a series of convolution operations, joint feature extraction module and classifier to form a lightweight convolutional neural network based on joint features. As shown in Figure 4, the overall structure of the proposed LCNN-JF method consists of six parts. Groups 1 to 5 extract the

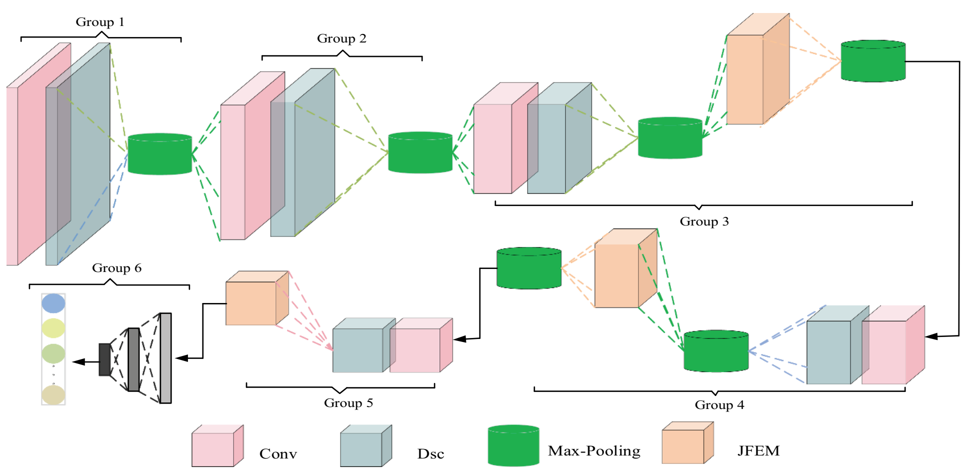


Figure 4. The overall structure diagram of the proposed LCNN-JF method.

features of remote sensing scene images. Groups 1 and 2 extract the shallow features of remote sensing scene images by using the shallow feature extraction module composed of traditional convolution (Conv) and depthwise separable convolution (Dssc). Groups 3, 4 and 5 add a joint feature extraction module (JFEM) to form a deep feature extraction module based on the shallow feature extraction module, which is used to extract the deeper complex features of remote sensing scene images. Further strengthen feature representation. Specifically, assuming that the input is $X \in \mathbb{R}^{\mathbb{W} \times \mathbb{H} \times \mathbb{C}}$, the feature $\tilde{F}_{\alpha}^{(2)}$ is obtained after the first and second groups of shallow feature extraction modules, and the calculation process of $\tilde{F}_{\alpha}^{(2)}$ is

$$\mathbb{R}^{\frac{H}{i} \times \frac{W}{i} \times iC'} \leftarrow F_{conv}^{(i)} = \iota\left\{\kappa\left[w_{conv}^{(i)}(\tilde{F}_{\alpha}^{(i-1)})\right]\right\}, i = 1, 2 \quad (5)$$

$$\mathbb{R}^{\frac{H}{i} \times \frac{W}{i} \times iC'} \leftarrow F_{dsc}^{(i)} = \iota\left\{\kappa\left[w_{dsc}^{(i)}(F_{conv}^{(i)})\right]\right\}, i = 1, 2 \quad (6)$$

$$\mathbb{R}^{\frac{H}{2i} \times \frac{W}{2i} \times iC'} \leftarrow \tilde{F}_{\alpha}^{(i)} = M_{\max pool}(F_{dsc}^{(i)}), i = 1, 2 \quad (7)$$

In formulas (5) - (7), $\tilde{F}_{\alpha}^{0} = X$, $w_{conv}^{(i)} \in \mathbb{R}^{H' \times W' \times iC'}$ represent the traditional convolution of Group i, where H' and W' represent the height and width of the convolution kernel respectively, iC'represents the number of channels of the convolution kernel, from group 1 to group 2, the number of channels of the output feature increases twice, $\frac{H}{i}$ and $\frac{W}{i}$ represent the height and width of the output feature respectively, from group 1 to group 2, the spatial size of the output feature is reduced to half of the original, $w_{conv}^{(i)} \in \mathbb{R}^{H \times W \times iC'}$ represents the depthwise separable convolution of group i, and $M_{\max pool}(\cdot)$ represents the max pool operation with pool kernel and pool step of 2. The depth feature extraction process from group 3 to group 5 is

$$F_{conv}^{(i)} = \iota\left\{\kappa\left[w_{conv}^{(i)}(\tilde{F}_{\alpha}^{(i-1)})\right]\right\}, i = 3, 4, 5 \quad (8)$$

$$F_{dsc}^{(i)} = \iota\left\{\kappa\left[w_{dsc}^{(i)}(F_{conv}^{(i)})\right]\right\}, i = 3, 4, 5 \quad (9)$$

$$\tilde{F}_{\alpha}^{(i)} = M_{\max pool}(F_{dsc}^{(i)}), i = 3, 4, 5 \quad (10)$$

$$\hat{F}^{(i)} = \Phi(F_{dsc}^{(i)}), i = 3, 4, 5 \quad (11)$$

$$\tilde{F}_{\beta}^{(i)} = M_{\max pool}(\hat{F}^{(i)}), i = 3, 4 \quad (12)$$

In formula (8) and formula (9), $F_{conv}^{(3)} = F_{dsc}^{(3)} \in \mathbb{R}^{\frac{W}{4} \times \frac{H}{4} \times 4C'}$, $F_{conv}^{(4)} = F_{dsc}^{(4)} \in \mathbb{R}^{\frac{W}{16} \times \frac{H}{16} \times 8C'}$, $F_{conv}^{(5)} = F_{dsc}^{(5)} \in \mathbb{R}^{\frac{W}{64} \times \frac{H}{64} \times 16C'}$; In formula (10), $\tilde{F}_{\alpha}^{(3)} \in \mathbb{R}^{\frac{W}{8} \times \frac{H}{8} \times 4C'}$, $\tilde{F}_{\alpha}^{(4)} \in \mathbb{R}^{\frac{W}{32} \times \frac{H}{32} \times 8C'}$, $\tilde{F}_{\alpha}^{(5)} \in \mathbb{R}^{\frac{W}{64} \times \frac{H}{64} \times 16C'}$; In formula (11), $\hat{F}^{(3)} \in \mathbb{R}^{\frac{W}{8} \times \frac{H}{8} \times 4C'}$, $\hat{F}^{(4)} \in \mathbb{R}^{\frac{W}{32} \times \frac{H}{32} \times 8C'}$, $\hat{F}^{(5)} \in \mathbb{R}^{\frac{W}{64} \times \frac{H}{64} \times 16C'}$; In formula (12), $\tilde{F}_{\beta}^{(3)} \in \mathbb{R}^{\frac{W}{16} \times \frac{H}{16} \times 4C'}$, $\tilde{F}_{\beta}^{(4)} \in \mathbb{R}^{\frac{W}{32} \times \frac{H}{32} \times 8C'}$. Compared with the shallow feature extraction module, the deep feature extraction module adds a joint feature extraction module and a maximum pooling layer, as shown in formula (11) and formula (12), where $\Phi(\cdot)$ represents the joint feature extraction module.

The sixth group is the classifier module, which is composed of global average pooling, full connection layer and softmax classifier. Suppose that the characteristic of group 5 output is $G = [g_1; g_2; \ldots; g_i; \ldots; g_N] \in \mathbb{R}^{N\times W\times H\times C}$, where $[;;\ldots;]$ represents the cascade operation along the batch dimension, N represents the batch size of input data, and W, H, C represents the width, height and number of channels of input data respectively. The output result of the global average pooling layer is $O = [o_1; o_2; \ldots; o_i; \ldots; o_N] \in \mathbb{R}^{N\times C}$, so the processing process of $g_i \in \mathbb{R}^{H\times W\times C}$ by the global average pooling layer can be expressed as

$$o_i = \frac{\sum_{h=1}^{H}\sum_{w=1}^{W} g_i}{H \times W} \tag{13}$$

Using global average pooling can reduce the risk of over fitting in the process of model training. In addition, using the global average pooling layer before the full connection layer can reduce the destruction of feature space information. Then input the global average pooled output result $o_i \in O$ to the full connection layer with the number of categories Z, and the output result is $J \leftarrow [j_1, j_2, \ldots, j_i, \ldots j_Z] \equiv FC(o_i)$. Finally, input the output result J of the full connection layer into the softmax function to obtain the output result $Y = [y_1, y_2, \ldots, y_i, \ldots, y_Z]$, then the output result y_i of the softmax classifier can be expressed as

$$y_i = \frac{e^{J[i]}}{\sum_{k=1}^{Z} e^{J[k]}} \tag{14}$$

In formula (14), $J[i]$ represents the i-th element in J (index starts from 1).

In this paper, cross-loss entropy is used as the loss function. Assuming that $K = [k_1, k_2, \ldots, k_i, \ldots, k_Z]$ represents the one-hot coding result of the input data, the loss function *loss* can be expressed as

$$loss = -\sum_{i=1}^{Z} k_i \log(y_i) \tag{15}$$

In formula (15), Z represents the number of scene categories, and y_i represents the output result of softmax classifier.

4. Experimental results and analysis

4.1. Dataset settings

The SIRI-WHU remote-sensing image dataset (Zhao et al. 2016) was published by Wuhan University in 2016. The dataset contains 12 scene categories, with 200 images per scene category, 200×200 pixels per image, and 2 metres of spatial resolution. SIRI-WHU remote sensing image dataset resources come from Google Earth and cover mainly urban areas in China. The WHU-RS19 (Xia et al. 2018) dataset is a remote sensing image dataset obtained from Google's satellite imagery. The data, published by Wuhan University in 2011, contain 19 categories, each with 50 images. AID Remote Sensing Image Dataset (Xia et al. 2017) is a challenging dataset that covers different seasonal scene images from China, Germany and the United States. The dataset consists of 30 categories of scene images, each of which contains 220 to 420 images with pixels of 600×600 and spatial resolution ranging from 0.5 m to 0.8 m. The NWPU (Cheng, Han, and Lu 2017) remote sensing image dataset consists of images obtained by Google Satellite in different seasons, illumination and angles. The dataset contains 45 scene categories, including airplanes, airports, baseball courts, and basketball courts. Each scene category has 700 images, with each image having 256×256 pixels and a spatial resolution ranging from 0.2 m to 30 m. The UC (Yang and Newsam 2010) dataset is a land-use image remote sensing dataset with 21 scene categories. The dataset contains 100 images in each scene category, each with a pixel size of 256×256 and a spatial resolution of 1 foot.

4.2. Setting of the Experiments

4.2.1. Dataset settings

Based on previous studies, the five datasets are divided into the following for a more effective evaluation of the proposed LCNN-JF method:

SIRI-WHU dataset:

(1) 50% of the images were randomly selected as training sets and the remaining 50% as test sets;

(2) 80% of the images were randomly selected as training sets and the remaining 20% as test sets;

WHU-RS19 dataset:

(1) 40% of the images are randomly selected as training sets and the remaining 60% as test sets;

(2) 60% of the images were randomly selected as training sets and the remaining 40% as test sets;

UC dataset:

80% of the images were randomly selected as training sets and the remaining 20% as test sets.

AID dataset:

(1) 20% of the images were randomly selected as training sets and the remaining 80% as test sets;

(2) 50% of the images were randomly selected as training sets and the remaining 50% as test sets;

NWPU dataset:

(1) 10% of the images are randomly selected as training sets and the remaining 90% as test sets;

(2) 20% of the images were randomly selected as training sets and the remaining 80% as test sets;

4.2.2. Parameter settings

During the experiment, the initial learning rate was set to 0.01, the network was optimized using the momentum gradient descent function, and the momentum coefficient was set to 0.9, and the batch size was 16. In addition, the computer parameters used in the experiment are as follows:

CPU:AMD Ryzen 7 4800 H with Radeon Rraphics@2.90 GHz; RAM: 16 G; GPU: RTX2060; Solid state hard disk: 1T; Operating system: Window 11.

4.3. Experimental result

4.3.1. Overall performance of proposed methods

To validate the performance of the proposed method, the overall accuracy (OA), Kappa, F1, and confusion matrix were used as evaluation indicators. OA is the most commonly used indicator in classification tasks and represents the ratio between predicting the correct number of samples on all test sets and the total number of samples. Confusion matrix is a visual matrix used to represent the performance of the algorithm. Each column in the matrix represents the predicted value, each row represents the actual category, the diagonal value represents the probability that the current class is correctly classified, and the value outside the diagonal represents the probability that the corresponding class is incorrectly classified. The Kappa coefficient is calculated based on the confusion matrix and is an indicator for consistency testing, which is to check the consistency of the predicted and actual results of the model. The F1 score, also known as the balanced F score, is an indicator of model accuracy, which takes into account both the accuracy and recall of classification models. The experimental results are shown in Table 1. From Table 1, it can be seen that the proposed LCNN-JF method performs well in a variety of training scales for five datasets. Especially in SIRI-WHU dataset with 80% training ratio, WHU-RS19 with 60% training ratio and WHU-RS19 with 40% training ratio and UC dataset with 80% training ratio, the OA value of the proposed method is more than 99%. In addition, it has better performance on AID and NWPU datasets which are more difficult to train.

4.3.2. Experimental results on SIRI-WHU dataset

On the SIRI-WHU dataset, performance comparisons with the most advanced methods are shown in Table 2. From Table 2, it can be seen that the proposed LCNN-JF method achieves the best performance in both training scales. When the training proportion is

Table 1. Experimental results of the proposed LCNN-JF method on five datasets.

Dataset	OA	Kappa	F1
SIRI-WHU(50%)	97.28±0.25	98.06±0.26	98.28±0.72
SIRI-WHU(80%)	99.05±0.16	99.34±0.15	99.45±0.26
WHU-RS19(40%)	98.50±0.52	98.72±0.43	99.05±0.34
WHU-RS19(60%)	99.00±0.15	99.23±0.52	99.46±0.28
UC(80%)	99.52±0.25	98.96±0.02	99.50±0.39
AID(20%)	93.05±0.46	93.89±0.75	94.05±0.42
AID(50%)	96.65±0.15	97.12±0.28	96.92±0.49
NWPU(10%)	91.36±0.29	92.20±0.16	92.16±0.52
NWPU(20%)	93.25±0.16	93.69±0.34	94.05±0.45

Table 2. Comparisons of LCNN-JF and advanced methods on 50% and 80% SIRI-WHU datasets.

Method	OA(50%)	OA(80%)	Parameter
DMTM	91.52	-	-
Siamese ResNet50	95.75	97.50	-
Siamese AlexNet	83.25	88.96	-
Siamese VGG-16	94.50	97.30	-
Fine-tune MobileNetV2	95.77±0.16	96.21±0.31	3.5M
SE-MDPMNet	96.96±0.19	98.77±0.19	5.17M
LPCNN	-	89.88	-
SICNN	-	93.00	-
Pre-trained-AlexNet-SPP-SS	-	95.07±1.09	-
SRSCNN	93.44	94.76	-
Proposed	97.28±0.25	99.05±0.16	5M

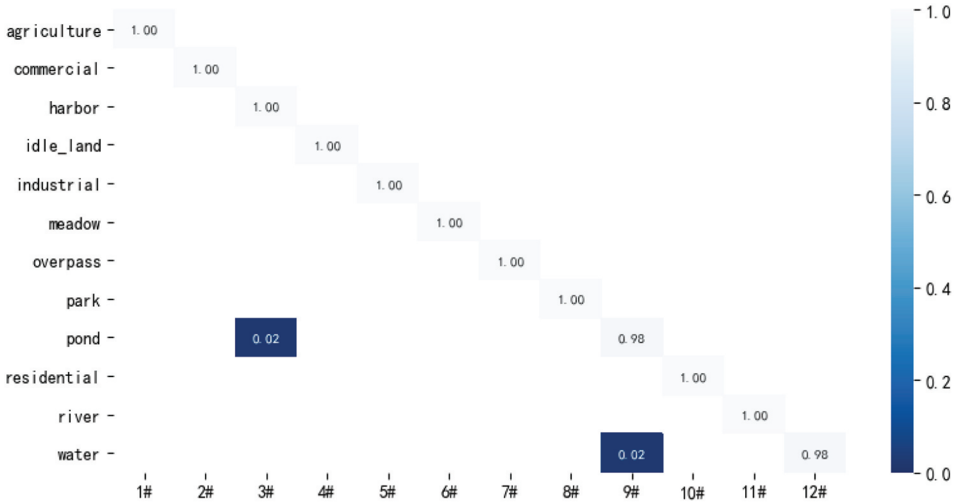


Figure 5. Confusion matrix of the proposed LCNN-JF method on a training scale of 80% SIRI-WHU dataset.

50%, the OA value of the proposed method is 97.28%, 1.32% higher than that of SE-MDPMNet (Zhang, Tang, and Zhao 2019). Under the training proportion of 80%, the classification accuracy reaches 99.05%, and 0.73% higher than that of SE-MDPMNet (Zhang, Zhang, and Wang 2019). Compared with the lightweight Fine-tune MobileNetV2 (Zhang, Tang, and Zhao 2019) method, the classification accuracy under two training scales is 2.51% and 3.29% higher than that of Fine-tune MobileNetV2 (Zhang, Zhang, and Wang 2019) method, although the number of parameters is slightly higher.

Figure 5 shows the confusion matrix of the proposed method on the SIRI-WHU dataset with a training scale of 80%. From Figure 5, it can be seen that, in addition to the two confusing scenarios of 'pond' and 'water', the proposed method can completely correct the classification of other scenario experiments in the dataset. Because 'pond' scenes and 'water' scenes contain the same elements, it is easy to confuse when classifying. Nevertheless, the proposed method still performs well in classification.

Table 3. Comparisons of the proposed LCNN-JF and advanced methods on 40% and 60% WHU-RS19 datasets.

Method	OA(40%)	OA(60%)	Parameter
CaffeNet	95.11±1.20	96.24±0.56	60.97M
VGG-VD-16	95.44±0.60	96.05±0.91	138.36M
GoogLeNet	93.12±0.82	94.71±1.33	7M
Fine-tune MobileNetV2	96.82±0.35	98.14±0.33	3.5M
SE-MDPMNet	98.46±0.21	98.97±0.24	5.17M
DCA by addition	-	98.70±0.22	-
Two-Stream Deep Fusion Framework	98.23±0.56	98.92±0.52	-
TEX-Net-LF	98.48±0.37	98.88±0.49	-
Proposed	98.50±0.52	99.01±0.15	5M

4.3.3. Experimental results on WHU-RS19 dataset

The experimental comparison results of LCNN-JF and advanced methods proposed on WHU-RS19 dataset are shown in Table 3. From Table 3, we can see that the proposed method has the best performance advantage among all the comparison methods. When the training proportion was 40%, the classification accuracy was 5.93%, 0.59%, and 2.23% higher than the lightweight methods GoogLeNet (Xia et al. 2017), SE-MDPMNet (Zhang, Tang, and Zhao 2019) and Fine-tune MobileNetV2 (Zhang, Zhang, and Wang 2019, respectively. When the training proportion was 60%, the classification accuracy of the proposed method reached 98.52%, which was 4.8%, 0.54% and 1.37% higher than that of GoogLeNet; Xia et al. 2017), SE-MDPMNet (Zhang, Tang, and Zhao 2019) and Fine-tune MobileNetV2 (Zhang, Zhang, and Wang 2019), respectively. In particular, compared with the Fine-tune MobileNetV2 (Zhang, Tang, and Zhao 2019) method, the proposed method achieves a good trade-off between classification accuracy and model complexity.

The confusion matrix of the proposed LCNN-JF method on the WHU-RS19 dataset with 60% training ratio is shown in Figure 6. From Figure 6, you can see that similar to the confusion results on the SIRI-WHU dataset, the classification errors are caused by the same spatial layout of the 'port' and 'river' scenes (such as water and forests in both the 'port' and 'river' scenes). However, the proposed method still performs well.

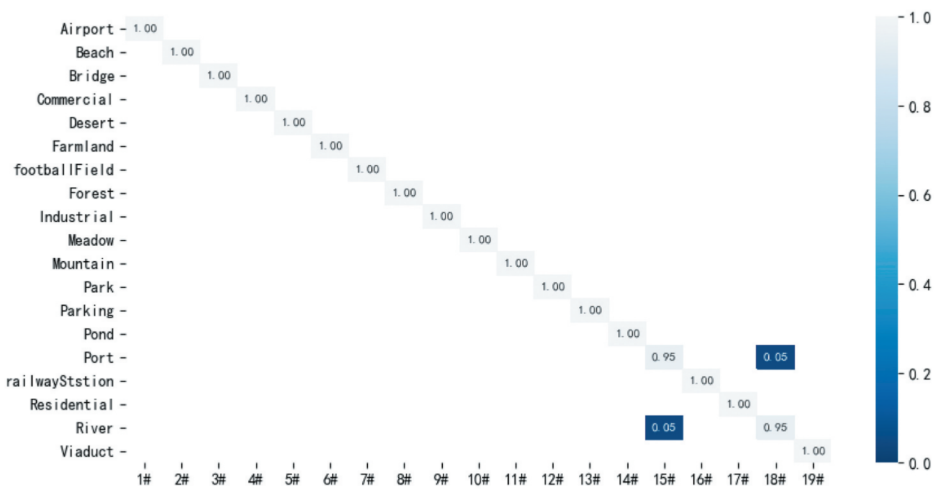


Figure 6. Confusion matrix of the proposed method on a training scale of 60% WHU-RS19 dataset.

Table 4. Comparisons of LCNN-JF and advanced methods on UCM datasets with 80% training ratio.

Method	OA	Parameter
ResNet+WSPM-CRC	97.95	23M
ADFF	98.81±0.51	23M
LCNN-BFF Method	99.29±0.24	6.2M
VGG16 with MSCP	98.36±0.58	-
Gated Bidirectional+global feature Method	98.57±0.48	138M
Feature Aggregation CNN	98.81±0.24	130M
Skip-Connected CNN	98.04±0.23	6M
Discriminative CNN	98.93±0.10	130M
ABM-CNN	99.50±0.23	5.6M
VGG16-DF	98.97	130M
Scale-Free CNN	99.05±0.27	130M
VGG16+CapsNet	99.05±0.24	22M
Semi-Supervised Representation Learning	94.05±1.2	210M
Siamese CNN	94.29	-
Siamese ResNet50 with R.D	94.76	-
Bidirectional Adaptive Feature Fusion Method	95.48	130M
Multiscale CNN	96.66±0.90	60M
SAFF	97.02±0.78	15M
proposed	99.50±0.25	5M

4.3.4. Experimental results on UCM dataset

The experimental results of the proposed LCNN-JF method and advanced method in the UCM dataset with a training ratio of 80% are shown in Table 4. From Table 4, it can be seen that the proposed method has a parameter of 5 M and a classification accuracy of 99.50. A good trade-off between classification accuracy and model complexity is achieved. Compared with the lightweight methods LCNN-BFF Method (Shi, Wang, and Wang 2020), ABM-CNN (Shi, Zhao, and Wang 2021) and Skip-Connected CNN (He et al. 2020), the classification accuracy is improved by 0.23%, 0.02% and 1.48, respectively.

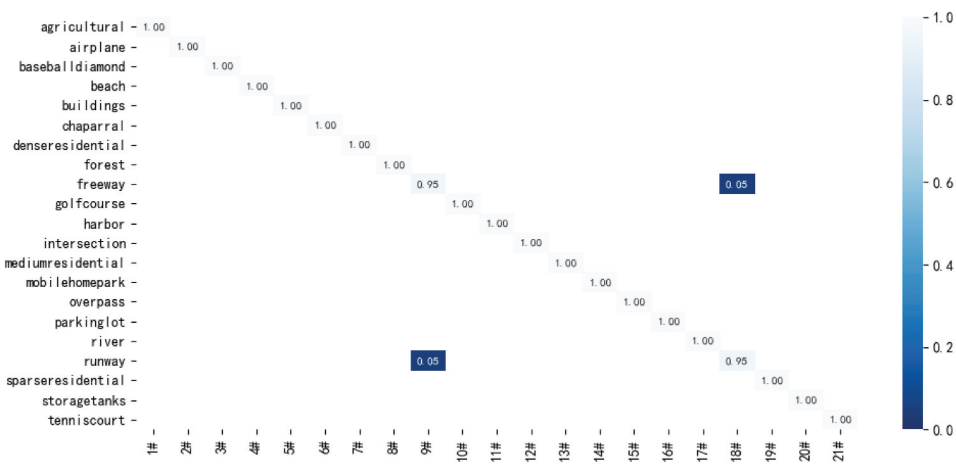


Figure 7. Confusion matrix of the proposed method on a training scale of 80% UC dataset.

Table 5. Comparisons of LCNN-JF and advanced methods on 20% and 50% AID datasets.

Method	OA(20%)	OA(50%)	Parameter
Bidirectional Adaptive Feature Fusion Method	-	93.56	130M
SAFF	90.25±0.29	93.83±0.28	15M
Skip-Connected CNN	91.10±0.15	93.30±0.13	6M
Gated Bidirectional Method	90.16±0.24	93.72±0.34	18M
Gated Bidirectional+global feature Method	92.20±0.23	95.48±0.12	138M
Feature Aggregation CNN	-	95.45±0.11	130M
AlexNet with MSCP	88.99±0.38	92.36±0.21	-
VGG16 with MSCP	91.52±0.21	94.42±0.17	-
Discriminative CNN	85.62±0.10	94.47±0.12	60M
TSDFF	-	91.8	-
LCNN-BFF Method	91.66±0.48	94.64±0.16	6.2M
ABM-CNN	93.27±0.22	95.54±0.13	5.6M
ResNet50	92.39±0.15	94.69±0.19	25.61M
DDRL-AM method	92.36±0.10	96.25±0.05	-
Fine-tuning Method	86.59±0.29	89.64±0.36	130M
Proposed	93.05±0.46	96.65±0.15	5M

The confusion matrix for the proposed method on a UCM dataset with 80% training ratio is shown in Figure 7. As we can see from Figure 7, the proposed method achieves complete recognition of almost all scenarios in a UCM dataset. The performance advantages of the proposed method are further verified.

4.3.5. Experimental results on WHU-RS19 dataset

The AID dataset is a more challenging dataset than the SIRI-WHU, WHU-RS19, and UC datasets. Referring to the classification of the dataset, we divided the proportion of samples used for training into 20% and 50%. The comparison results are shown in Table 5. From Table 5, it can be seen that the classification performance of the proposed method is better than that of the comparison method under both training scales. With a training proportion of 20%, the classification accuracy of the proposed method was 93.05%, which exceeded all the comparison methods, being 1.69%, 1.66% and 0.78% higher than DDRL-AM method (Li et al. 2020), ResNet50 (Li et al. 2020) and ABM-CNN (Shi, Zhao, and Wang 2021), respectively. When the training proportion was 50%, the classification accuracy of the proposed method was 96.65%, which was 4.35%, 3.01%, 1.4% and 2.11% higher than Skip-Connected CNN (He et al. 2020), LCNN-BFF Method (Shi, Wang, and Wang 2020), DDRL-AM Method (Li et al. 2020) and ABM-CNN (Shi, Zhao, and Wang 2021), respectively. The performance advantages of the proposed method over the comparison method are further verified.

The confusion matrix of the proposed method on the AID dataset with a training proportion of 50% is shown in Figure 8. From Figure 8, it can be seen that the proposed method achieves a classification accuracy of more than 90% for all scenarios in AID dataset, and 100% for ‘viaduct’ scenarios. Of all the scenarios in this dataset, ’school’ and ‘industrial’ scenarios are the most likely to be confused. Similar building shapes and spatial structures in both scenarios result in lower classification accuracy for schools and factories, 91% and 94%, respectively.

Proposed	93.05±0.46	96.65±0.15	5M

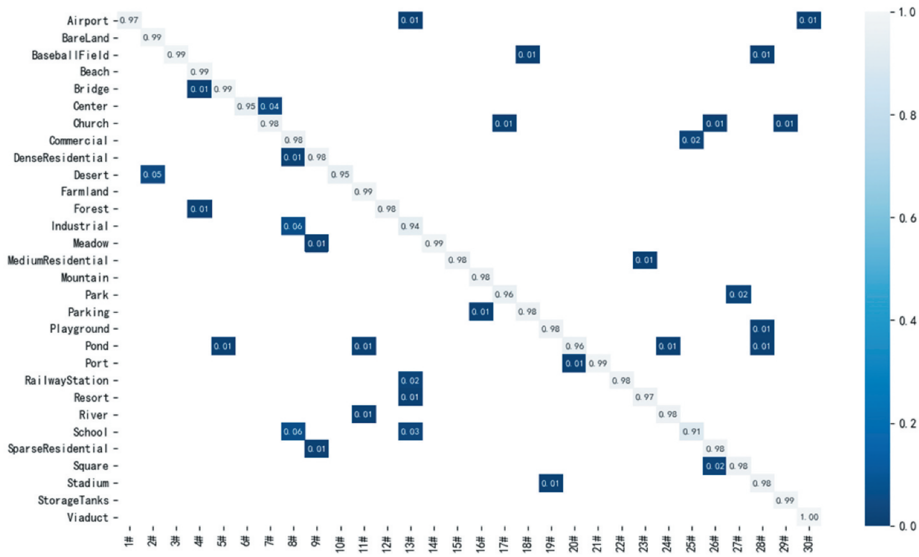


Figure 8. Confusion matrix of the proposed method on a training scale of 50% AID dataset.

Table 6. Comparisons of LCNN-JF and advanced methods on 10% and 20% NWPU datasets.

Method	OA(10%)	OA(20%)	Parameter
SAFF	84.38±0.19	87.86±0.14	15M
Skip-Connected CNN	84.33±0.19	87.30±0.23	6M
Discriminative with AlexNet	85.56±0.20	87.24±0.12	130M
Discriminative with VGG16	89.22±0.50	91.89±0.22	130M
VGG16+CapsNet	85.05±0.13	89.18±0.14	130M
LCNN-BFF Method	86.53±0.15	91.73±0.17	6.2M
ABM-CNN	88.99±0.14	92.42±0.14	5.6M
Contourlet CNN	85.93±0.51	89.57±0.45	12.6M
ResNet50	86.23±0.41	88.93±0.12	25.61M
InceptionV3	85.46±0.33	87.75±0.43	45.37M
Fine-tuning Method	87.15±0.45	90.36±0.18	130M
AlexNet with MSCP	81.70±0.23	85.58±0.16	-
VGG16 with MSCP	85.33±0.17	88.93±0.14	-
Proposed	91.36±0.29	93.25±0.16	5M

4.3.6. Experimental results on NWPU dataset

In the NWPU dataset, the experimental results of the proposed and advanced methods with 10% and 20% training ratio are shown in Table 6. From Table 6, it can be seen that the proposed method achieves remarkable performance in both training scales. When the training proportion is 10%, the proposed method is 3.14% higher than Discriminative with VGG16 (Cheng et al. 2018), 3.37% higher than ABM-CNN (Shi, Zhao, and Wang 2021), and 5.21% higher than Fine-tuning Method (Xia et al. 2017). When the training proportion is 20%, the classification accuracy of the proposed method is 93.25%, which has better performance advantages than all the comparison methods. The classification accuracy of the proposed

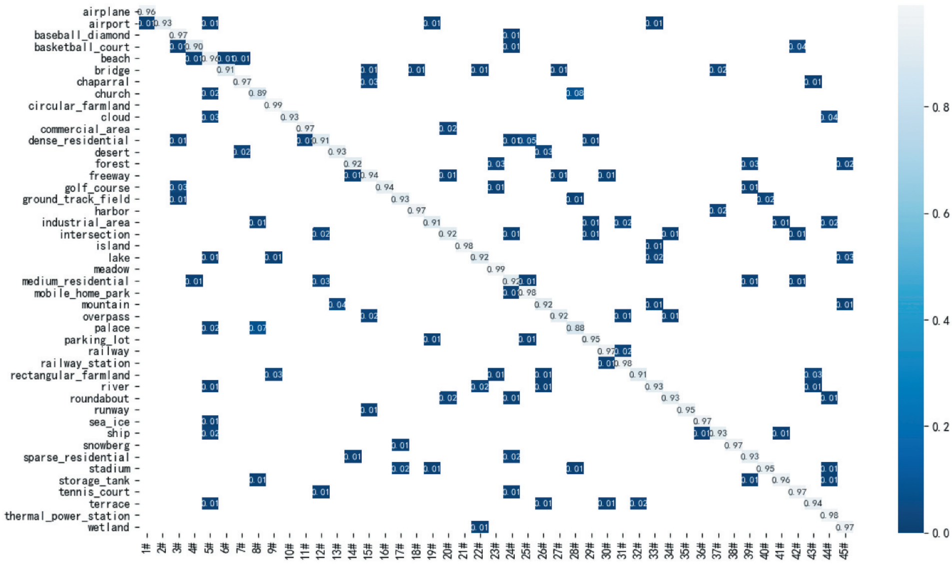


Figure 9. Confusion matrix of the proposed method on a training scale of 50% AID dataset.

methods is 1.83%, 6.95% and 2.52% higher than that of the lightweight methods ABM-CNN (Shi, Zhao, and Wang 2021), Skip-Connected CNN (He et al. 2020) and LCNN-BFF Method (Shi, Wang, and Wang 2020), respectively.

The confusion matrix for the proposed method on a 20% NWPU dataset is shown in Figure 9. As you can see from Figure 9, 'palace' and 'church' have the lowest classification accuracy, 88% and 89%, respectively. Except for the two scenes of 'palace' and 'church', all the other scenes are classified with more than 90% accuracy.

4.4. Model complexity analysis

To further prove the efficiency of the proposed method, seven advanced methods such as LCNN-BFF (Shi, Wang, and Wang 2020), GoogLeNet (Xia et al. 2017), CaffeNet (Xia et al. 2017), VGG-VD-16 (Xia et al. 2017), Fine-tune MobileNetV2 (Zhang, Tang, and Zhao 2019), SE-MDPMNet (Zhang, Zhang, and Wang 2019) and Contourlet CNN (Liu et al. 2018) were selected for model complexity comparison. A comparative experiment was conducted on a 50% training scale AID dataset. Parameters and Floating point operations (FLOPs) were selected to measure the complexity of the model. The experimental results are shown in Table 7. Table 7 shows that the proposed method has the highest classification accuracy on AID datasets. Although the number of parameters is slightly higher than Fine-tune MobileNetV2 (Zhang, Zhang, and Wang 2019), the FLOPs value of the proposed method is 313.2 M lower and the classification accuracy is 2.94% higher than that of the method. The proposed method achieves a good trade-off between classification accuracy and model complexity.

Table 7. Comparing the complexity of LCNN-JF and advanced methods on 50% AID dataset.

Method	OA	Parameter	FLOPs
LCNN-BFF	94.64	6.1M	24.6M
GoogLeNet	85.84	7M	1.5G
CaffeNet	88.25	60.97M	715M
VGG-VD-16	87.18	138M	15.5G
Fine-tune MobileNetV2	94.71	3.5M	334M
SE-MDPMNet	92.64	5.17M	3.27G
Contourlet CNN	95.54	12.6M	2.1G
Proposed	97.65	5M	20.8M

Table 8. ATT comparison of the proposed model on UC dataset with advanced methods.

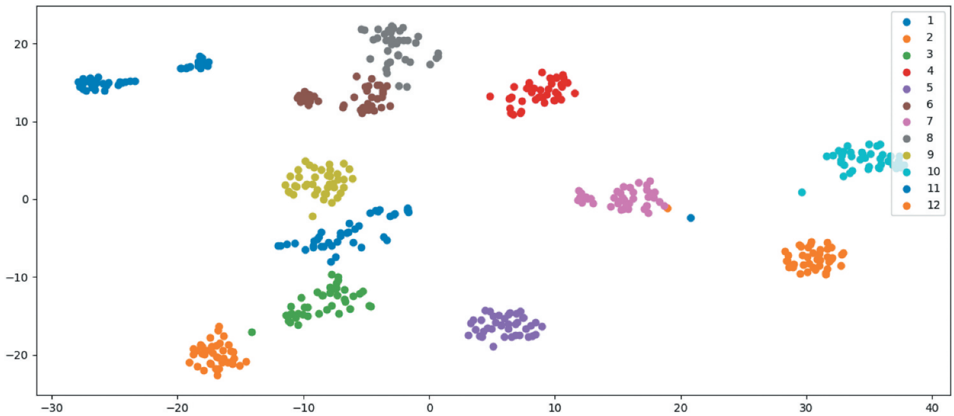
Method	ATT(s)
Siamese ResNet_50	0.053
Siamese AlexNet	0.028
Siamese VGG-16	0.039
LCNN-BFF	0.029
Gated Bidirectional+global feature Method	0.052
Gated Bidirectional Method	0.048
Proposed	0.015

4.5. Model speed comparison

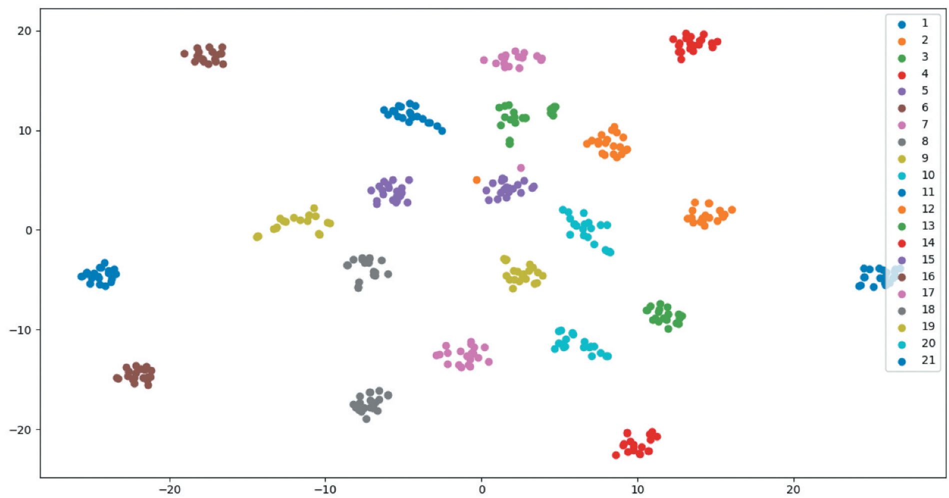
Usually, Average training time (ATT) is used to measure the average time required for a model to train a picture. Select Siamese ResNet_50 (Liu et al. 2019), Siamese AlexNet (Li et al. 2019), Siamese VGG-16 (Liu et al. 2019), LCNN-BFF (Shi, Wang, and Wang 2020), Gated Bidirectional+global feature method (Sun et al. 2020) and Gated Bidirectional Method (Sun et al. 2020) are compared experimentally. The results of the comparison are shown in Table 8. As you can see from Table 8, the ATT value of the proposed method is 0.015 s, 0.014 s less than LCNN-BFF (Shi, Wang, and Wang 2020), and 0.024 s less than the Siamese VGG-16 (Liu et al. 2019) method. The efficiency of the proposed method is further verified.

5. Discussions

In order to further discuss the performance of the proposed method, in this section, various visualization results of the proposed method are provided. Firstly, the t-distributed stochastic neighbour embedding (T-SNE) (Maaten et al. 2008) method is adopted, which is a visualization method for dimension reduction proposed by Laurens van der Maaten and Geoffrey Hinton in 2008. The T-SNE visualization method considers both the local and global relationships of the data, which can give the validity of a method from the perspective of visualization. The visualization results of the proposed methods using T-SNE on the SIRI-WHU and UC datasets are shown in Figure 10. From Figure 10, we can see that on both datasets, this method reduces the distance within the same semantic



(a)



(b)

Figure 10. T-SNE visualization results diagram. (a). T-SNE visualization results on the SIRI-WHU dataset. (b). T-SNE visualization results on UC datasets.

clusters, reduces confusion between different semantic clusters, and effectively improves the classification accuracy on the two datasets.

Then, gradient weighted class activation map (GradCAM) is utilized to visually discuss some scenes of UC dataset. Grad-CAM is avisualization method proposed by Selvaraju et al. (Selvaraju et al. 2017),which is used to locate the category-related areas in an image, and to show thelevel of interest in the related areas by the colour depth. The visualizationresults are shown in Figure 11. From Figure 11, we can see that the proposed method pays much more attention to thescene labels of the input image than to the object labels, which improves the classification performance of remote sensingscene images.

Finally, the UC dataset is randomly predicted by theproposed method, and the results are shown in Figure 12. From Figure 12, it canbe seen that the confidence of the proposed

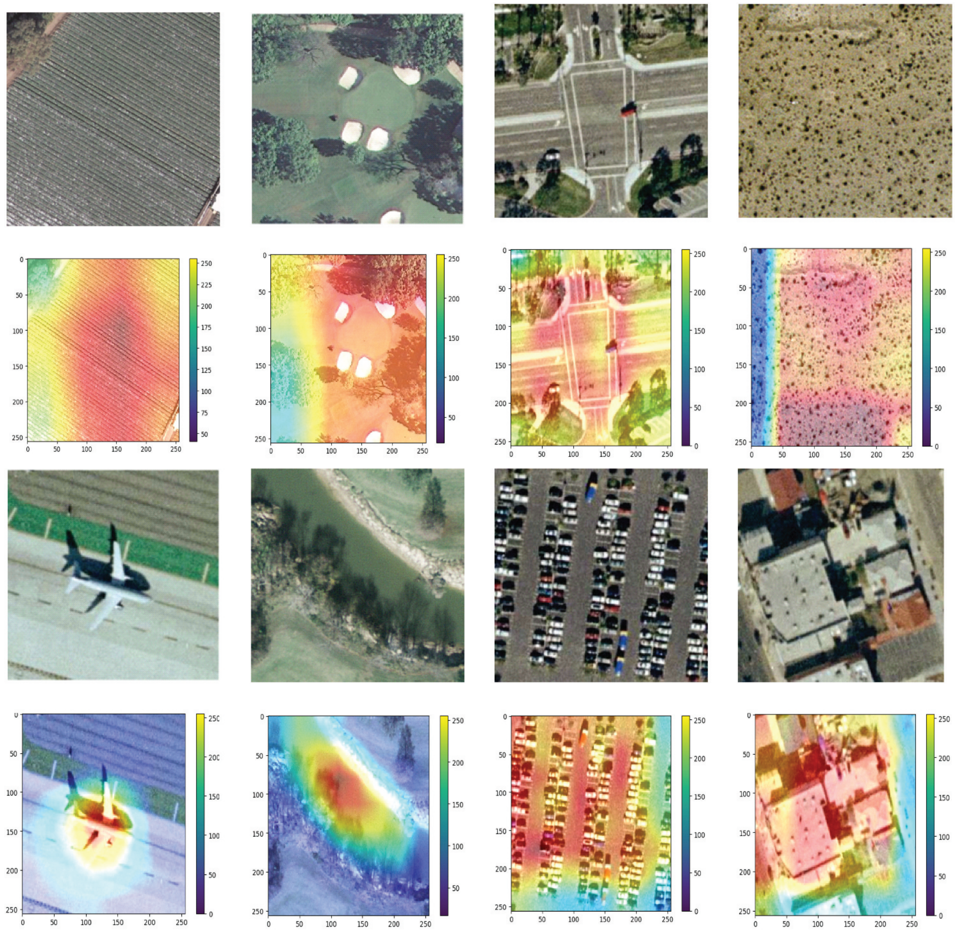


Figure 11. The visualization results on UC dataset by Grad CAM.

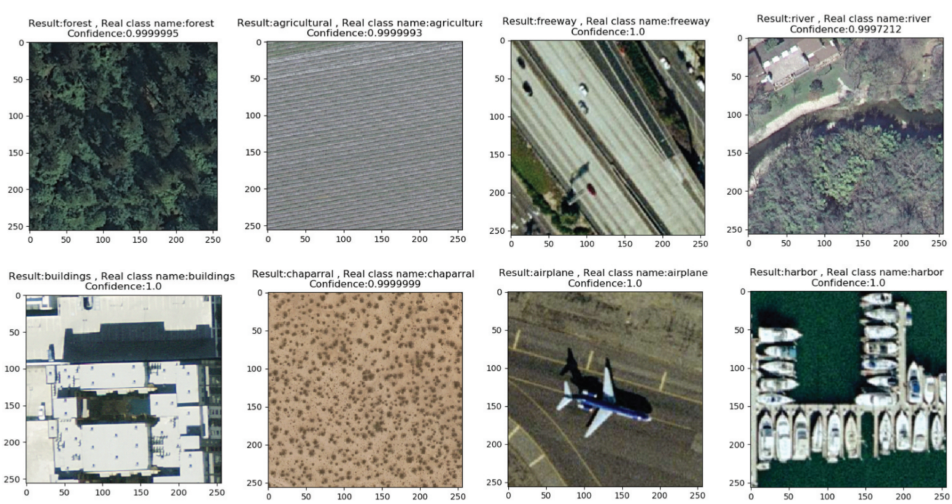


Figure 12. Randomly predicted results on the UC dataset.

method for random prediction of allscenarios is more than 99%, and some scenarios are even 100%, which furtherproves the validity of the proposed method.

6. Conclusions

This paper presents a multi-scale global feature extraction module, which combines global features of multiple scales through global pooling and sub-region pooling operations, effectively improving the characterization ability of features. In addition, based on the global feature extraction module, a joint feature extraction module is proposed. The module consists of three branches. Branch 1 being 3×3 convolution is utilized to extract local features. Branch 2 is a global feature extraction module. To reduce the loss of information during the feature extraction process, an identity branch is adopted to compensate for the features, and the three branches are fused. Then, a lightweight modular convolution neural network is constructed using the joint feature extraction module for remote sensing scene image classification, and a series of experiments prove the superiority of this method. Although the proposed method has achieved good classification results, some scenes that difficult to be distinguished in the dataset (such as 'palace' and 'church' scenes in NWPU dataset) should be further recognized. The next step is to find a method to identify these scenes more accurately and further improve the classification performance of remote sensing scene images.

Acknowledgements

The authors would like to thank the editors and the reviewers for their help and suggestion.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was funded in part by the National Natural Science Foundation of China (42271409), in part by the Heilongjiang Science Foundation Project of China under Grant LH2021D022, in part by the Leading Talents Project of the State Ethnic Affairs Commission, and in part by the Fundamental Research Funds in Heilongjiang Provincial Universities of China under Grant 145209149.

Data availability statement

Data associated with this research are available online. The UCM Merced dataset is available for download at http://weegee.vision.ucmerced.edu/datasets/landuse.html. NWPU dataset is available for download at http://www.escience.cn/people/JunweiHan/NWPURESISC45.html. AID dataset is available for download at https://captain-whu.github.io/AID/. SIRI-WHU dataset is available for download at http://www.lmars.whu.edu.cn/prof_web/zhongyanfei/ecode.html. WHU-RS19 dataset is available for download at https://paperswithcode.com/dataset/whu-rs19.

References

Cao, Y., J. Xu, S. Lin, F. Wei, and H. Hu. 2019. "GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond." 2019 IEEE/cvf International Conference On Computer Vision Workshops (iccvw), 1971–1980. https://doi.org/10.1109/ICCVW.2019.00246.

Cheng, G., J. Han, and X. Lu. 2017. "Remote Sensing Image Scene Classification: Benchmark and State of the Art." *Processing IEEE* 105 (10): 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998.

Cheng, G., Z. Li, J. Han, X. Yao, and L. Guo. 2018. "Exploring Hierarchical Convolutional Features for Hyperspectral Image Classification." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 56 (11): 6712–6722. https://doi.org/10.1109/TGRS.2018.2841823.

Cheng, G., C. Yang, X. Yao, L. Guo, and J. Han. 2018. "When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 56 (5): 2811–2821. https://doi.org/10.1109/TGRS.2017.2783902.

Chen, L. C., G. Papandreou, F. Schroff, and H. Adam. 2017. "Rethinking Atrous Convolution for Semantic Image Segmentation." *arXiv: 1706,05587v3*. https://doi.org/10.48550/arXiv.1706.05587.

Fu, J., J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. 2019. "Dual Attention Network for Scene Segmentation." *arXiv: 1809,02983v4*. https://doi.org/10.48550/arXiv.1809.02983.

Han, J., D. Zhang, G. Cheng, L. Guo, and J. Ren. 2015. "Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 53 (6): 3325–3337. https://doi.org/10.1109/TGRS.2014.2374218.

He, N., L. Fang, S. Li, J. Plaza, and A. Plaza. 2020. "Skip-Connected Covariance Network for Remote Sensing Scene Classification." *IEEE Transactions on Neural Networks and Learning Systems* 31 (5): 1461–1474. https://doi.org/10.1109/TNNLS.2019.2920374.

He, K., X. Zhang, S. Ren, and J. Sun 2016 "Deep Residual Learning for Image Recognition." Proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR), SC USA, 770–778.

Howard, A. G., M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. 2017. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." *arXiv* 1704.04861. [Online].https://arxiv.org/abs/1704.04861.

Hua, Y., L. Mou, and X. X. Zhu. 2019. "Recurrently Exploring Class-Wise Attention in a Hybrid Convolutional and Bidirectional LSTM Network for Multi-Label Aerial Image Classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 149:188–199. https://doi.org/10.1016/j.isprsjprs.2019.01.015.

Hua, Y., L. Mou, and X. X. Zhu. 2020. "Relation Network for Multilabel Aerial Image Classification." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 58 (7): 4558–4572. https://doi.org/10.1109/TGRS.2019.2963364.

Jing, W., Q. Ren, J. Zhou, and H. A. Song. 2020. "AutoRsisc: Automatic Design of Neural Architecture for Remote Sensing Image Scene Classification." *Pattern Recognition Letters* 140:186–192. https://doi.org/10.1016/j.patrec.2020.09.034.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2017. "ImageNet Classification with Deep Convolutional Neural Networks." *Communications of the ACM* 60 (6): 84–90. https://doi.org/10.1145/3065386.

Liang, L., W. Zhao, X. Hao, Y. Yang, K. Yang, L. Liang, and Q. Yang. 2020. "Image Registration Using Two-Layer Cascade Reciprocal Pipeline and Context-Aware Dissimilarity Measure." *Neurocomputing* 371:1–14. https://doi.org/10.1016/j.neucom.2019.06.101.

Li, J., D. Lin, Y. Wang, G. Xu, Y. Zhang, C. Ding, and Y. Zhou. 2020. "Deep Discriminative Representation Learning with Attention Map for Scene Classification." *Remote Sensing of Environment* 12:1366. https://doi.org/10.3390/rs12091366.

Lin, T. Y., P. Dollar, R. Grishick, K. He, B. Hariharan, and S. Belongie. 2017. "Feature Pyramid Networks for Object Detection." *arXiv: 1612,03144v2*. https://doi.org/10.48550/arXiv.1612.03144.

Li, B., W. Su, H. Wu, R. Li, W. Zhang, W. Qin, and S. Zhang. 2019. "Aggregated Deep Fisher Feature for VHR Remote Sensing Scene Classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12 (9): 3508–3523. https://doi.org/10.1109/JSTARS.2019.2934165.

Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, Y. C. Fu, and A. C. Berg. 2016. "SSD: Single Shot MultiBox Detector." *arXiv* 1512.02325v5. https://doi.org/10.1007/978-3-319-46448-0_2.

Liu, J. J., Q. Hou, M. M. Cheng, J. Feng, and J. Jiang 2019. "A Simple Pooling-Based Design for Real-Time Salient Object Detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, USA, 3917–3926.

Liu, B. D., J. Meng, W. Y. Xie, S. Shao, Y. Li, and Y. Wang. 2019. "Weighted Spatial Pyramid Matching Collaborative Representation for Remote-Sensing-Image Scene Classification." *Remote Sensing of Environment* 11 (5): 518. Art. no. https://doi.org/10.3390/rs11050518.

Liu, W., A. Rabinovich, and A. C. Berg. 2015. "ParseNet: Looking Wider to See Better." *arXiv: 1506,04579v2*. https://doi.org/10.48550/arXiv.1506.04579.

Liu, Y., Y. Zhong, F. Fei, Q. Zhu, and Q. Qin. 2018. "Scene Classification Based on a Deep Random-Scale Stretched Convolutional Neural Network." *Remote Sensing of Environment* 10 (3): 444. Art. no. https://doi.org/10.3390/rs10030444.

Liu, Y., Y. Zhong, and Q. Qin. 2018. "Scene Classification Based on Multiscale Convolutional Neural Network." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 56 (12): 7109–7121. https://doi.org/10.1109/TGRS.2018.2848473.

Liu, X., Y. Zhou, J. Zhao, R. Yao, B. Liu, and Y. Zheng. 2019. "Siamese Convolutional Neural Networks for Remote Sensing Scene Classification." *IEEE Geoscience and Remote Sensing Letters* 16 (8): 1200–1204. https://doi.org/10.1109/LGRS.2019.2894399.

Li, W., Z. Wang, Y. Wang, J. Wu, J. Wang, Y. Jia, and G. Gui. 2020. "Classification of High-Spatial-Resolution Remote Sensing Scenes Method Using Transfer Learning and Deep Convolutional Neural Network." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13:1986–1995. https://doi.org/10.1109/JSTARS.2020.2988477.

Lu, X., W. Ji, X. Li, and X. Zheng. 2019. "Bidirectional Adaptive Feature Fusion for Remote Sensing Scene Classification." *Neurocomputing* 328:135–146. https://doi.org/10.1016/j.neucom.2018.03.076.

Lu, X., H. Sun, and X. Zheng. 2019. "A Feature Aggregation Convolutional Neural Network for Remote Sensing Scene Classifification." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 57 (10): 7894–7906. https://doi.org/10.1109/TGRS.2019.2917161.

Maaten, L. V. D., and G. Hinton. 2008. "Visualizing Data using t-SNE." *Journal of Machine Learning Research* 9:2579–2605. https://doi.org/10.48550/arXiv.2108.01301.

Mahanand, S., M. D. Behera, P. S. Roy, P. Kumar, S. K. Barik, and P. K. Srivatava. 2021. "Satellite Based Fraction of Absorbed Photosynthetically Active Radiation is Congruent with Plant Diversity in India." *Remote Sensing* 13 (159): 1–18. https://doi.org/10.3390/rs13020159.

Najibi, M., P. Samangouei, R. Chellappa, and L. Davis. 2017. "SSH: Single Stage Headless Face Detector." *arXiv: 1708,03979*. https://doi.org/10.48550/arXiv.1708.03979.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra 2017. "Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization." Proceedings of the IEEE International Conference on Computer Vision(ICCV), Venice, Italy, 618–626.

Shi, C., T. Wang, and L. Wang. 2020. "Branch Feature Fusion Convolution Network for Remote Sensing Scene Classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13:5194–5210. https://doi.org/10.1109/JSTARS.2020.3018307.

Shi, C., X. Zhao, and L. Wang. 2021. "A Multi-Branch Feature Fusion Strategy Based on an Attention Mechanism for Remote Sensing Image Scene Classification." *Remote Sensing of Environment* 13 (10): 1950. https://doi.org/10.3390/rs13101950.

Simonyan, K., and A. Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *arXiv: 1409,1556*. [Online] https://arxiv.org/abs/1409.1556.

Sun, H., S. Li, X. Zheng, and X. Lu. 2020. "Remote Sensing Scene Classification by Gated Bidirectional Network." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 58 (1): 82–96. https://doi.org/10.1109/TGRS.2019.2931801.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. "Going Deeper with Convolutions." *arXiv: 1409,4842*. https://doi.org/10.48550/arXiv.1409.4842.

Tan, Q., Y. Liu, X. Chen, and G. Yu. 2017. "Multi-Label Classification Based on Low Rank Representation for Image Annotation." *Remote Sensing* 9 (2): 109. https://doi.org/10.3390/rs9020109.

Tasar, O., Y. Tarabalka, and P. Alliez. 2019. "Incremental Learning for Semantic Segmentation of Large-Scale Remote Sensing Data." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12 (9): 3524–3537. https://doi.org/10.1109/JSTARS.2019.2925416.

Wang, P., P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell 2018. "Understanding Convolution for Semantic Segmentation. In 2018 IEEE winter conference on applications of computer vision (WACV), Lake Tahoe, NV, USA, 1451–1460.

Wang, X., S. Wang, C. Ning, and H. Zhou. 2021. "Enhanced Feature Pyramid Network with Deep Semantic Embedding for Remote Sensing Scene Classification." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 59 (9): 7918–7932. https://doi.org/10.1109/TGRS.2020.3044655.

Wang, X., X. Xiong, and C. Ning. 2019. "Multi-Label Remote Sensing Scene Classification Using Multi-Bag Integration." *Institute of Electrical and Electronics Engineers Access* 7:120399–120410. https://doi.org/10.1109/ACCESS.2019.2937188.

Wu, X., D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao. 2019. "ORSIm Detector: A Novel Object Detection Framework in Optical Remote Sensing Imagery Using Spatial-Frequency Channel Features." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 57 (7): 5146–5158. https://doi.org/10.1109/TGRS.2019.2897139.

Xia, G. S., J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, and L. Zhang. 2017. "AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classifification." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 55 (7): 3965–3981. https://doi.org/10.1109/TGRS.2017.2685945.

Xia, G. S., W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maitre 2018. "Structural High-Resolution Satellite Image Indexing." Proceeding ISPRS Commission VII Mid-Term Symposium '100 Years ISPRS, Vienna, Austria 38:298–303.

Yang, Y., and S. Newsam 2010. "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classifification." In Proceedings of the 18th SIGSPA-TIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 3–5 November, 270.

Zhang, W., P. Tang, and L. Zhao. 2019. "Remote Sensing Image Scene Classification Using CNN-Capsnet." *Remote Sensing of Environment* 11 (5): 494. Art. no. https://doi.org/10.3390/rs11050494.

Zhang, K., Z. Zhang, Z. Li, and Y. Qiao. 2016. "Joint Face Detection and Alignment Using Multi-Task Cascaded Convolutional Networks." *arXiv: 1604,02878v1* 23:1499–1503. https://doi.org/10.1109/LSP.2016.2603342.

Zhang, B., Y. Zhang, and S. Wang. 2019. "A Lightweight and Discriminative Model for Remote Sensing Scene Classification with Multidilation Pooling Module." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12 (8): 2636–2653. https://doi.org/10.1109/JSTARS.2019.2919317.

Zhao, B., Y. Zhong, G. S. Xia, and L. Zhang. 2016. "Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 54 (4): 2108–2123. https://doi.org/10.1109/TGRS.2015.2496185.

Zhong, Y., F. Fei, and L. Zhang. 2016. "Large Patch Convolutional Neural Networks for the Scene Classification of High Spatial Resolution Imagery." *Journal of Applied Remote Sensing* 10 (2): 25006. Art. no. https://doi.org/10.1117/1.JRS.10.025006.

Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. 2014. "Object Detectors Emerge in Deep Scene Cnns." *arXiv* 1412 (3): 6856.

Zhou, Z. H., M. L. Zhang, S. J. Huang, and Y. F. Li. 2012. "Multi-Instance Multi-Label Learning." *Artificial Intelligence* 176 (1): 2291–2320. https://doi.org/10.1016/j.artint.2011.10.002.

报告编号：**L24Y2023-0633**

检　索　报　告

一、检索要求

1. 委 托 人：石翠萍 Shi, CP (Shi, Cuiping)
2. 委托单位：齐齐哈尔大学
3. 检索目的：论文被 SCI-E 收录情况

二、检索范围

Science Citation Index Expanded (SCI-EXPANDED)	1990-present	网络版
JCR-(Journal Citation Reports)	2022	网络版
中国科学院文献情报中心期刊分区表(升级版)	2022	网络版

三、检索结果

委托人提供的**1**篇论文被SCI-E收录，论文被收录、所在期刊的JCR影响因子、中科院期刊分区(升级版)情况见附件一。

特此证明！

检索报告人：杨茗惠

东北师范大学科技查新咨询中心

教育部科技查新工作站(L24)

2023 年 **11** 月 **30** 日

附件一：SCI-E收录情况

标题: A lightweight convolution neural network based on joint features for Remote Sensing scene image classification
作者: **Shi, CP (Shi, Cuiping)**; Zhang, XL (Zhang, Xinlei); Wang, LG (Wang, Liguo); Jin, Z (Jin, Zhan)
来源出版物: INTERNATIONAL JOURNAL OF REMOTE SENSING 卷: 44 期: 21 页: 6615-6641 DOI: 10.1080/01431161.2023.2273246 出版年: NOV 2 2023
Web of Science 核心合集中的 "被引频次": 0
被引频次合计: 0
使用次数 (最近 180 天): 0
使用次数 (2013 年至今): 0
引用的参考文献数: 53
摘要: Unlike natural images, remote sensing scene images usually contain one scene label and many object labels, and many object labels are arranged dispersedly, which brings great difficulties to feature extraction of scene label. To accurately identify scene labels from remote sensing scene images with multiple object labels, it is important to fully understand the global context of the image. In order to solve the challenges of multi-label scene images and improve the classification performance, a global context feature extraction module is proposed in this paper. The module combines the semantics information of different regions through a global pooling and three different scale sub-regions pooling, which makes the module have stronger ability of global feature representation. In addition, in order to fully understand the semantic content of remote sensing images, a three branch joint feature extraction module is constructed, which consists of the global context feature module, 3 x 3 convolution branch and identity branch are fused. Finally, a lightweight convolution neural network based on joint features (LCNN-JF) is constructed using traditional convolution, depthwise separable convolution, joint feature extraction module and classifier for remote sensing scene image classification. A series of experimental results on four datasets, UCM, AID, RSSCN and NWPU, demonstrate that the proposed method has better feature representation ability and can achieve better classification of remote sensing scene images.
入藏号: WOS:001099139200001
语言: English
文献类型: Article
作者关键词: global context feature; joint feature; lightweight convolutional neural network (LCNN); remote sensing scene image
KeyWords Plus: FEATURE FUSION
地址: [Shi, Cuiping] Huzhou Univ, Coll Informat Engn, Huzhou, Peoples R China.
[Zhang, Xinlei; Jin, Zhan] Qiqihar Univ, Coll Elect & Commun Engn, Qiqihar, Peoples R China.
[Wang, Liguo] Dalian Nationalities Univ, Coll Informat & Commun Engn, Dalian, Peoples R China.
通讯作者地址: Shi, CP (通讯作者)，Huzhou Univ, Coll Informat Engn, Huzhou, Peoples R China.
电子邮件地址: shicuiping@qqhru.edu.cn
Affiliations: Huzhou University; Qiqihar University; Dalian Minzu University
出版商: TAYLOR & FRANCIS LTD
出版商地址: 2-4 PARK SQUARE, MILTON PARK, ABINGDON OR14 4RN, OXON, ENGLAND
Web of Science Index: Science Citation Index Expanded (SCI-EXPANDED)
Web of Science 类别: Remote Sensing; Imaging Science & Photographic Technology
研究方向: Remote Sensing; Imaging Science & Photographic Technology
IDS 号: X5VX9
ISSN: 0143-1161
eISSN: 1366-5901
29 字符的来源出版物名称缩写: INT J REMOTE SENS
ISO 来源出版物缩写: Int. J. Remote Sens.
来源出版物页码计数: 27
基金资助致谢:
基金资助机构 授权号
The authors would like to thank the editors and the reviewers for their help and suggestion.
The authors would like to thank the editors and the reviewers for their help and suggestion.

输出日期: 2023-11-30

期刊影响因子 ™ 2022：3.4

中国科学院文献情报中心期刊分区(升级版, 2022)截图如下:

INTERNATIONAL JOURNAL OF REMOTE SENSING

刊名	INTERNATIONAL JOURNAL OF REMOTE SENSING
年份	2022
ISSN	0143-1161
Review	否
Open Access	否
Web of Science	SCIE

	学科	分区	Top期刊
大类	工程技术	3	否
小类	IMAGING SCIENCE & PHOTOGRAPHIC TECHNOLOGY 成像科学与照相技术	3	-
	REMOTE SENSING 遥感	3	

The End