Looking Closer at the Scene: Multi-Scale Representation Learning for Remote Sensing Image Scene Classification

Qi Wang, *Senior Member, IEEE*, Wei Huang, *Student Member, IEEE*, Zhitong Xiong, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Remote sensing image scene classification has attracted great attention because of its wide applications. Although convolutional neural network (CNN) based methods for scene classification have achieved excellent results, the large scale variation of the features and objects in remote sensing images limits the further improvement of the classification performance. To address this issue, we present multi-scale representation for scene classification, which is realized by a global-local two-stream architecture. This architecture has two branches of global stream and local stream, which can individually extract the global features and local features from the whole image and the most important area. In order to locate the most important area in the whole image using only image-level labels, a weakly-supervised key area detection strategy of structured key area localization (SKAL) is specially designed to connect the above two streams. To verify the effectiveness of the proposed SKAL based two-stream architecture, we conduct comparative experiments based on three widely used CNN models, including AlexNet, GoogleNet and ResNet18, on four public remote sensing image scene classification data sets, and achieve the state-of-the-art results on all the four data sets. Our codes will be provided in https://github.com/hw2hwei/SKAL.**

***Index Terms*—remote sensing, scene classification, CNN, multi-scale representation, structured key area localization**

I. INTRODUCTION

BENEFITING from the remote sensing imaging equipments and technologies, in recent years, many semantic-level tasks of remote sensing images have developed rapidly, such as object detection [1], image retrieval [2], image captioning [3] [4], road extraction [5] and the others. As a basis for these tasks, remote sensing image scene classification [6]–[10] has become a research hotspot, which classifies remote sensing images into a set of scene classes according to the features and objects in the images.

There are plenty of similar and confusing features and objects in remote sensing images, and therefore it is cruical to extract discriminative features of remote sensing scences. According to feature extraction, there are two kinds of supervised features: handcrafted features and deep features. Compared with handcrafted features, deep features contain more high-level semantic information, which can be automatically learned by convolutional neural networks (CNNs). Due to the powerful ability of feature extraction, CNN-based methods [6], [8], [9], [11]–[14] have become the mainstream methods and achieved the state-of-the-art results in the field of remote sensing image scene classification.

Although the performance of CNN-based methods has improved significantly, scene classification of remote sensing images still suffers from the large scale variation of features and objects in the images. As shown in the images of Fig. 1, the most important areas occupy only a small part of the whole images, and they are usually surrounded by a large number of useless features and objects, which decreases the discrimination of the extracted features.

(a) airplane (b) lake (c) island (d) intersection

Fig. 1: Some samples of remote sensing scene images with the bounding boxes labeling the key area.

To tackle this issue, we present joint global and local feature representation for remote sensing image scene classification. As shown in Fig. 2, we design a global-local two-stream architecture. In this architecture, the blue branch network is the global feature extraction stream and the green branch network is the local feature extraction stream. The global area contains more global features such as contour and texture information, while the key local area enlarges the most important objects which can provide more fine-grained features and reduce background noise. Our global-local two-stream architecture

Q. Wang, W. Huang, Z. Xiong and X. Li are with the School of Computer Science, and with the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China (e-mail: crabwq@nwpu.edu.cn, hw2hwei@gmail.com, xiongzhitong@gmail.com, xuelong_li@nwpu.edu.cn).

X. Li is the corresponding author.

This work was supported by the National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, and 61871470.

can individually extract global features and local features from the input images of different scales, and finally fuse their classification scores.

In order to locate the most important local area in the whole image, we further propose a weakly-supervised key area detection strategy of structured key area localization (SKAL) as the yellow route in Fig. 2. The proposed SKAL defines an explicit local key area localization process based on the feature response degree of the patches in global feature maps. It can accurately locate the most important area, *i.e.* its boundary of $[x_1, x_2, y_1, y_2]$, using only image-level labels.

To verify the effectiveness of the proposed SKAL based global-local two-stream architecture in remote sensing image scene classification, the abundant comparative experiments based on three widely used CNN models (Alexnet [15], ResNet18 [16] and GoogleNet [17]) are conducted on four public large-scale remote sensing scene data sets including UC Merced data set [18], RSSCN7 data set [19], AID data set [20], and NWPU-RESISC45 data set [21]. We achieve the state-of-the-arts on all the four data sets.

The contributions of this paper can be summarized as the following four aspects:

(1) To deal with the problem of large scale variation in remote sensing images, we present joint global and local feature representation for remote sensing image scene classification from the perspective of multi-scale feature. Correspondingly, a global-local two-stream architecture is designed to individually extract global features and local features from the input images of different scales.

(2) To locate the most important area in the whole remote sensing scene image, a strategy of structured key area localization (SKAL) is specially proposed to connect the global and local streams. SKAL can accurately calculate the most important local area in the form of bounding box $[x_1, x_2, y_1, y_2]$.

(3) In order to prove the effectiveness of our SKAL based global-local two-stream architecture in remote sensing images, plenty of comparative experiments based on several widely used CNN models are conducted on four public scene classification data sets of remote sensing images. The state-of-the-art results demonstrate that our method can significantly improve the performance of remote sensing scene classification.

(4) As a multi-scale representation learning method, our global-local two-stream architecture can easily be applied in all kinds of CNN models, as with the advantages of simple implementation, fast operation and strong interpretability.

II. RELATED WORK

In this section, the related works of remote sensing image scene classification and weakly-supervised key object detection methods are reviewed in brief.

A. Remote Sensing Image Scene Classification

According to the feature extraction, the methods used in scene classification of remote sensing images can be roughly split into the following three types.

1) Handcrafted Features: The handcrafted feature based methods were first applied in remote sensing image scene classification. These methods rely on a series of manually designed feature descriptors including global feature descriptors (such as color histograms and texture descriptors [22], [23]), and local features descriptors (such as Histogram of Oriented Gradients (HOG) [24], [25] and Scale-Invariant Feature Transform (SIFT) [26], [27]). Global feature descriptors can generate the entire representation of a remote sensing image, which can be directly sent into the classifier. Local feature descriptors, which are usually the mid-level feature descriptors, need to be integrated into an global representation by feature combination technologies like bag-of-visual-words (BoVW) [28]. Further, Zhu *et al.* [29] propose a local-global feature fusion operation at the histogram level. These handcrafted features are well-designed, however, they cannot effectively deal with the challenges of intraclass diversity, interclass similarity and scale variation.

2) Unsupervised Features: Classification is intrinsically a supervised task but researchers also find ways to interpret unsupervised learning results as classes. Researchers have attempted unsupervised feature learning based methods [30]–[34], which aim at learning the feature encoding functions. For these unsupervised feature learning based methods, they take the handcrafted feature descriptors like SIFT as input, and generate the fused features. It is crucial to combine multiple features by using some encoding techniques. The typeical encoding methods used in remote sensing image scene classification include Principal Component Analysis (PCA), k-means clustering, sparse coding [32] and autoencoder [35]. What's more, BoVW [25], which can generate the visual dictionaries (codebooks) from the handcrafted features based on k-means clustering, is also one of the most popular unsupervised feature learning based methods. However, overall the unsupervised methods cannot generate the same discriminate features of different scene classes as the supervised features because of the lack of labels.

3) Deep Features: In recent years, deep learning methods, especially Convolutional Neural Networks (CNNs) [6], [8], [9], [11]–[14], [36]–[38], have dominated the most fields of natural images because of their powerful large-scale feature extraction. Similarly, deep feature based methods methods have become the mainstream of remote sensing image scene classification with the better classification performance than the handcrafted and unsupervised features. Compared with handcrafted feature-based methods that generally are determined by feature engineering skills and domain acknowledge, deep feature based methods can automatically learn the most discriminative semantic-level features from the raw images. Besides, the CNN models are the end-to-end trainable architectures instead of the complex multi-stage architectures, which are the main workflows of handcrafted feature based methods. Although deep feature based methods have obtained the excellent performance, the large scale variation is still one of the most difficult problems to solve.

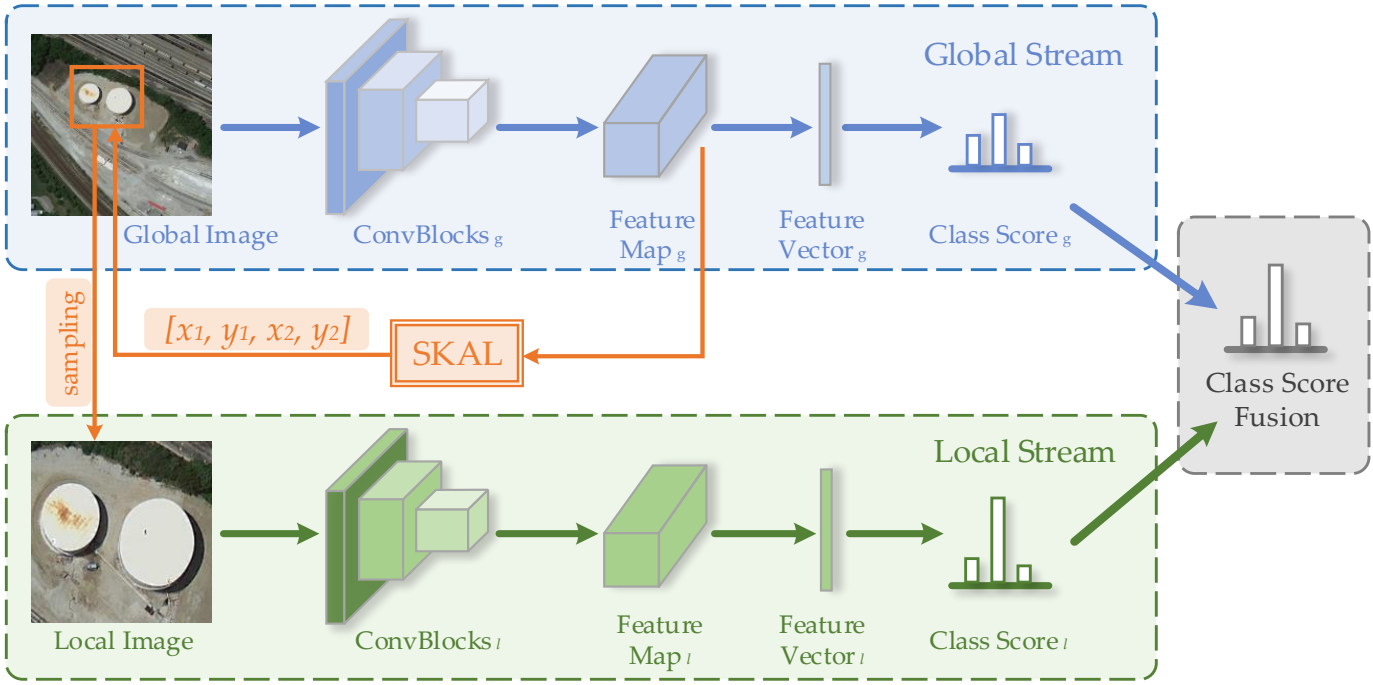


Fig. 2: SKAL based global-local two-stream architecture is designed for joint global and local feature representation in remote sensing scene image classification.

B. Weakly-Supervised Object Detection

Weakly-supervised object detection, which locates the most important local area using only image-level labels, is a meaningful research subject. It can not only increase the interpretability of our scene classification models, but also be used for further improving their performance. Pandey *et. al* [39] achieve weakly-supervised object detection based on the handcrafted feature of deformable part models (DPM). To make full use of the advantages of convolutional features, Bilen *et. al* [40] propose weakly-supervised deep detection networks to locate the key objects. And Cinbis *et. al* [41] introduce the multiple instance learning into weakly-supervised objection detection. Besides, Zhang *et. al* [42] bring the saliency detection into this field. Further, Fu *et. al* [43] realize a learnable key object localization network of Recurrent Attention Convolutional Neural Network (RA-CNN) for fine-grained image recognition and get significant gains. The clustering learning technology is also attempted in [44]. Recently, Yang *et. al* [45] proposed spatial prior for the object dependence for joint object detection and action classification.

For remote sensing images, however, there are lots of large-scale features and objects containing background noise. To deal with these complicated scene images, motivated by the idea of multi-scale feature representation in RA-CNN, we design a key area localization strategy of SKAL to generate the minimum area boundary to sample the most important area in the whole image. There are three main differences between the proposed SKAL and RA-CNN: (1) The proposed SKAL is a relatively interpretable key area localization srategy to some extent while RA-CNN utilizes an attention proposal sub-network (ATN), which also belongs to the black-box model, to predict the key area. (2) Size of the key area located by SKAL can be controlled artificially by adjusting a hyperparameter, while the RA-CNN is hard to realize it. Because the remote sensing image scene classification is the basic of other further remote sensing image processing tasks, the controllable size may be more suitable for the further image processing. (3) It is necessary for RA-CNN to set an extra inter-scale pairwise ranking loss which is used to constrain the location process and an subtle g alternative training strategy. For our SKAL, the training of two streams is independently and is easier to realize.

There is some connection between the proposed SKAL and a series of region-proposal objection detection methods [46]–[48] under the demand of area location. The difference is that these object detection methods need accurate bounding boxes of the interest objects while the proposed SKAL has no need for them.

III. METHOD

In this section, firstly, the compositions of convolutional neural networks are sequentially introduced. Then the proposed SKAL strategy based on the global multi-layer feature map is introduced in detail in Fig. 3. Finally, as shown in Fig. 2, we put forward the SKAL based global-local two-stream architecture to individually extract the global and local features, and fuse their classification scores.

A. CNN

Generally, an entire CNN model for image classification can be roughly divided into three parts: stacked convolutional layers, global average pooling layer and fully-connected layer.

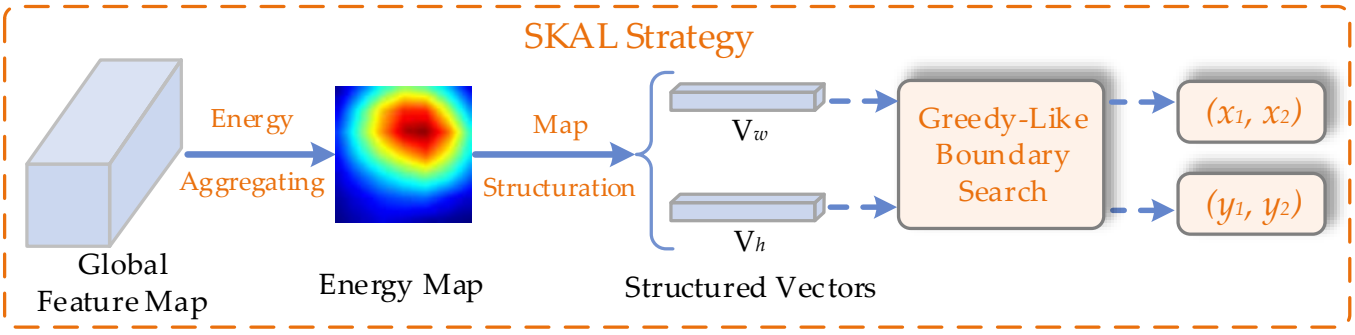


Fig. 3: Workflow of the proposed SKAL strategy.

Stacked Convolutional Layers. In CNNs, the multiple stacked convolutional layers are the most important parts used to extract features from low-level texture and color characteristics to high-level semantic information. Each convolutional layer normally consists of convolutional kernels and non-linear activation function (in some cases there is also batch normalization [49]). Because the convolutional layers of different models are stacked in different ways, it is hard to respectively describe their architectures. Thus, they are described as a unified $ConvBlocks$ to roughly represent the process of feature extraction in this paper. The input of these stacked convolutional layers is a RGB image $I \in \mathbb{R}^{3\times 224\times 224}$ and the output is the multi-layer feature map $M \in \mathbb{R}^{C\times H\times W}$ (C, H and W are the number of channels, spatial height and spatial width respectively). It is denoted as:

$$M = ConvBlocks(I), \tag{1}$$

Global Average Pooling Layer. Because the multi-layer feature map M distributes in spatial in units of patches, it is necessary to pool the feature map M into the corresponding feature vector $V \in \mathbb{R}^C$ for the next classifier (fully-connected layer). Thus, global pooling layer, mostly global average pooling (GAP) layer, is used to connect the convolutional layers and fully-connected layer, which is calculated by:

$$V(c) = \frac{1}{HW}\sum_{i=0}^{H}\sum_{j=0}^{W} M(c,i,j), \tag{2}$$

GAP can not only change the spatial dimension of features, but also decrease the overfitting of the trainable parameters.

Fully-connected Layer. Fully-connected (FC) layer plays the role of classifier in CNNs, which gives the classification score of each class based on the high-dimension feature vector V. It takes as input the V and takes as output a score vector of classification confidence denoted as $S \in \mathbb{R}^N$, where N is the number of classes in data set. It is calculated by:

$$S = W^T * V + b, \tag{3}$$

here $W \in \mathbb{R}^{C\times N}$ and $b \in \mathbb{R}^N$ are the weight and bias of the features V, respectively. The element of S is the classification score of each class. In order to scale the sum of S to 1, the operation of $softmax$ is added after the FC layer. It is formulated as:

$$\bar{S}_i = \frac{e^{S_i}}{\sum_{j=1}^{C} e^{S_j}}, \tag{4}$$

$\bar{S} \in \mathbb{R}^N$ is the scaled classification score.

The cascade of convolutional layers, GAP layer and FC layer is the entire CNN models for classification task. And it is also the structure of each stream in our global-local two-stream architecture.

B. SKAL

The most critical step is to localize the key area in the global image, which plays a role of a bridge between the global and local streams. As shown in Fig 3, we propose the SKAL strategy in detail in this subsection. Based on the multi-layer feature map of global image of M_g, which is calculated by Eqn (1)-(3) from the global image, the proposed SKAL generates a bounding box of $[x_1, x_2, y_1, y_2]$ to guide the sampling process. SKAL consists of the following three substeps: energy aggregation, energy map structuration and greedy-like boundary search.

*1) **Energy Aggregation**:* It is prerequisite to quantitatively describe the importance degree of each patch in M_g. In this paper, the operation of energy aggregation, which takes M_g as input and takes the energy map of $M_E \in \mathbb{R}^{H\times W}$ as output, is applied as:

$$M_E = \sum_{i=0}^{C} M_g(i,H,W), \tag{5}$$

here it is notable that energy aggregation can be regarded as a kind of explicit attention mechanism without the requirement for training.

It is necessary to scale all the elements of M_E into the range of [0, 1] by min-max scaling, which can remove the interference from the negative element.

$$\hat{M}_E(i) = \frac{M_E(i) - min(M_E)}{max(M_E) - min(M_E)} \tag{6}$$

here $max(M_E)$ and $min(M_E)$ are the the values of the maximum and minimum elements in $M_E(i)$, respectively. $\hat{M}_E$ are the scaled energy map in the same dimension with M_E.

For more accurate localization, it is meaningful to upsample the energy map into a larger spatial size from $H\times W$ (normally

6×6 or 7×7), which is denoted as:

$$\bar{M}_E = bilinear(\hat{M}_E) \tag{7}$$

we used bilinear interpolation as the upsampling technique. And the size of $\bar{M}_E$ is set to 25×25 in all the CNN models in this paper.

*2) **Energy Map Structuration**:* After finishing the above preparations, we obtain an scaled energy map $\bar{M}_E$ that can quantitively describe the patch-wise importance degree of the global image. As well as we know, it is complicated to implement the search in 2-D space. Therefore, we further aggregate the scaled energy map into two 1-D structured energy vectors, $V_w \in \mathbb{R}^W$ and $V_h \in \mathbb{R}^H$, along the spatial height and width by:

$$\begin{cases} V_w = \sum_{i=0}^{H} \bar{M}_E(i, W), \\ V_h = \sum_{i=0}^{W} \bar{M}_E(H, i), \end{cases} \tag{8}$$

V_w and V_h are the structured interpretation of M_E. Energy map structuration has two advantages. On one hand, it greatly improves the search efficiency to translate the boundary search in the 2-D energy map into the search in two 1-D energy vectors. On the other hand, it can realize decoupling in 2-D space by the separate search along the spatial width and height.

*3) **Greedy-Like Boundary Search**:* Based on V_w and V_h, $[x_1, x_2]$ and $[y_1, y_2]$ can be calculated respectively. In order to quickly and accurately locate the most important 1-D area in the 1-D energy vector, we present a greedy-like boundary search method on the basis of energy.

Taking the V_w as an example, we present some concepts containing the energy of different elements in the width vector as:

$$\begin{cases} E_{[0:W]} = \sum_{i=0}^{W} V(i), \\ E_{[x_1:x_2]} = \sum_{i=x_1}^{x_2} V(i) \end{cases} \tag{9}$$

here $E_{[0:W]}$ is the energy sum of all the elements in the width vector, and $E_{[x_1:x_2]}$ contains the energy of the region along the spatial width from x_1 to x_2.

In this paper, the key area in the global image is defined as: the area occupies the smallest area but contains no less than a threshold of the total energy (ETr), *i.e.*, $E_{[x_1:x_2]}$ / $E_{[0:W]}$ $> ETr$. ETr is a hyper-parameter of energy proportion. On the basis of this definition, we can search the most key region using a greedy-like algorithm, which is summarized in Algorithm 1. The greedy-like boundary search algorithm can be subdivided into the following three steps.

Step A: Initializing the boundary. From Line 1 to 8, firstly, $[x_1, x_2]$ are initialized by the boundary of the half of V_w having the maximum energy.

Step B: Adjusting the boundary. Then the boundary of $[x_1, x_2]$ are adjusted iteratively to make its energy converge to a small neighbor of ETr. There are two states after *Step A*: from Line 9 to 16, energy of the initialized area of $[x_1, x_2]$ is higher than ETr; from Line 17 to 24, the energy is lower than ETr. When the energy is higher than ETr, the region of $[x_1, x_2]$ needs to shrink until the energy is no higher than ETr along the direction of the slowest energy drop. However, when the energy is lower than ETr, the region needs to enlarge until it is no lower than ETr along the direction of the fastest energy rise. Our greedy-like boundary search can find the smallest but most informative area quickly.

Step C: Scaling the boundary. Because the above boundaries are in the range of $[0, W]$, as shown from Line 26 to 27, it is necessary to scale them to $[0, 1]$ by the dimension of the energy vector.

The width boundary of the most key area in a global image, which is $[x_1, x_2]$, can be obtained by the above three steps. And the height boundary of $[y_1, y_2]$ can be obtained similarly by using the same algorithm. The entire boundary of $[x_1, x_2, y_1, y_2]$ is used to guide the key local area sampling process.

Algorithm 1 Greedy-Like Boundary Search.

Input: Structured width vector $V_w \in \mathbb{R}^W$;
Output: The scaled width boundary of the key area: [x_1, x_2];

1: $x_1 \leftarrow 0$
2: $x_2 \leftarrow W/2$
3: **for** $i = 0 \rightarrow W/2$ **do**
4: **if** $E_{[x_1:x_2]} < E_{[i:i+W/2]}$ **then**
5: $x_1 \leftarrow i$
6: $x_2 \leftarrow i + W/2$
7: **end if**
8: **end for**
9: **if** $E_{[x_1:x_2]}/E_{[0:W]} > ETr$ **then**
10: **while** $E_{[x_1:x_2]}/E_{[0:W]} > ETr$ **do**
11: **if** $V_w(x_1 + 1) < V_w(x_2 - 1)$ **then**
12: $x_1 \leftarrow x_1 + 1$
13: **else**
14: $x_2 \leftarrow x_2 - 1$
15: **end if**
16: **end while**
17: **else**
18: **while** $E_{[x_1:x_2]}/E_{[0:W]} < ETr$ **do**
19: **if** $V_w(x_1 - 1) > V_w(x_2 + 1)$ **then**
20: $x_1 \leftarrow x_1 - 1$
21: **else**
22: $x_2 \leftarrow x_2 + 1$
23: **end if**
24: **end while**
25: **end if**
26: $x_1 \leftarrow x_1/W \times 100\%$
27: $x_2 \leftarrow x_2/W \times 100\%$
28: **return** [x_1, x_2]

C. Global-Local Two-Stream Architecture

According to the scaled boundary of $[x_1, x_2, y_1, y_2]$, we can sample the key local area $I_l \in \mathbb{R}^{3\times224\times224}$ in the enlarged global image $I_g^{'} \in \mathbb{R}^{3\times448\times448}$ by bilinear interpolation technology, which is denoted as:

$$I_l = bilinear(I_g^{'}, (x_1, x_2, y_1, y_2)) \tag{10}$$

The global and local classification scores of $\bar{S}_g$ and $\bar{S}_l$ are calculated from the global image $I_g \in \mathbb{R}^{3\times224\times224}$ and the key local area I_l, which are formulated as:

$$\bar{S}_g = CNN_g(I_g) \tag{11}$$

$$\bar{S}_l = CNN_l(I_l) \tag{12}$$

Both CNN_g and CNN_l are the cascade of Eqn (1)-(4). These two streams have the same structure but do not share the parameters in order to extract the features of different scales. The fused calssification score is the average of $\bar{S}_g$ and $\bar{S}_l$ as:

$$\bar{S}_f = \frac{\bar{S}_g + \bar{S}_l}{2} \tag{13}$$

IV. EXPERIMENTS

In this section, the remote sensing image scene data sets and the evaluation metrics used in this paper are introduced firstly. Secondly, the experiment setup and training hyper-parameters are provided in detail. Following that, an example of the visualization of the proposed SKAL is shown for auxiliary interpretation. Finally, we report the experimental results on each data set with the comparison with some state-of-the-art methods, and analyse the performance of our SKAL based global-local two-stream architecture.

A. Data Sets and Evaluation Metrics

1) UC Merced Land-Use Data Set: The UC Merced (UCM) land-use data set [18] consists of 2,100 images that are split into 21 typical land-use scene classes of agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Each class contains 100 optical images measuring 256×256 pixels, and each pixel has a spatial solution of 30 cm in the RGB color space.

2) RSSCN7 Data Set: The RSSCN7 data set contains 2800 images that are made up of seven typical scene classes including the grassland, forest, farmland, industrial region, lake, parking lot, residential region, and river. Each class has 400 images collected from the global satellite map in Google Earth, which are individually sampled at four different scales. Images in RSSCN7 data set are 400×400 pixels. RSSCN7 is a challenging data set due to the changing seasons, varying weathers and scale diversity.

3) Aerial Image Data Set: The Aerial Image Data (AID) set [20] have 10,000 images split into 30 classes including airport, bare land, baseball field, beach, bridge, center, church,commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, and viaduct. Each class has hundreds of large-scale images measuring 600×600 pixels in the RGB space. Each pixel has the spatial resolution of the range from 800 cm/pixel to 50 cm/pixel.

4) NWPU-RESISC45 Data Set: The NWPU-RESISC45 data set contains 31,500 images split into 45 classes, including airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, dense residential, desert, forest, freeway, golf course, ground track field, harbor, industrial area, intersection, island, lake, meadow, medium residential, mobile home park, mountain, overpass, palace, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snowberg, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station, and wetland. Each class has 700 images of 256×256 pixels with the spatial resolution of the range from about 300 cm/pixel to 20 cm/pixel in the RGB color space. It is the largest remote sensing scene classification data set in terms of the number of images and classes so far, and is more challenging because of the higher between-class similarity.

B. Evaluation Metrics

The following two typical metrics are used to quantitatively evaluate the experimental results.

1) Overall Accuracy: The overall accuracy (OA) is defined as the number of the correctly classified images divided by the total number of images in the data set. The score of OA reflects the overall performance of classification models instead of per class.

2) Confusion Matrix: The confusion matrix (CM) is an two-dimension informative table which is used to analyze the between-class classification errors and confusion degree. Each row of the matrix represents all the image samples of a predicted class while each column represents the samples of a ground-truth class.

To obtain reliable experimental results, on UCM, RSSCN7 and AID data sets, we repeated the experiments for five times using the same training ratio to randomly split the data set, and report the mean value and standard deviation of these results. On NWPU-RESISC45 data set, the number of repetitions is three due to the large number of samples.

C. Experiment Setup

1) CNN Baselines: To evaluate the effectiveness and robustness of the SKAL based global-local two-stream architecture in scene classification of remote sensing images, the comparative experiments are conducted on the aforementioned four data sets based on three kinds of widely used CNN models, including AlexNet [15], GoogleNet [17] and ResNet18 [16] pretrained on ImageNet [50]. AlexNet is composed mainly of convolutional layers, GoogleNet concatenates filters with different sizes, and ResNets have residual connections. When they are used for the key area calculation, their classifiers (fully-connected layers) are removed and the final multi-layer feature maps are upsampled to $25 \times 25 \times C$ (25×25 is the spatial size and C is the dimension of feature map) for more accurate location. Here the technology of replacing the last layer of a model pretrained on ImageNet is widely used strategy [38], [51], [52].

2) Training Parameters: Adam algorithm [53] is selected as the optimizer and cross entropy loss is used as the loss function for all the models. All the models are trained for 50 epochs with the batch size set to 64. Learning rate of Adam is initialized to 1e-4, and it is divided by 10 every 20 epochs. As for the size of input images in global-local two-stream architecture, the input of global stream is the global image resized to 224×224, while the input of local stream is the local area, which is cropped from the enlarged global image of 448×448, and then resized to 224×224. For fair comparison, the state-of-the-art CNN models compared with our two-stream architecture are all based on the input size of 224×224.

3) Data Augmentation: Because of the scale difference of the global and local images, the data augmentation methods of these two streams are different. For global stream, we use random horizontal flipping to augment the global images. For local stream, besides random horizontal flipping, we randomly crop a square area from the global image for local image augmentation. In order to keep the scale of the local images consistent during training and testing, the side length proportion of the square area is set to 50%.

4) Training and Test Procedure: We adopt two-stage training strategy to individually train the global and local streams. The global stream is trained and tested firstly, and provide a preliminary scene classification score based on the global image. Then the local stream is trained by the augmented local samples, and give another score based on the key local area. Finally these two scores are fused by average operation.

In addition, all the experiments are implemented by Pytorch [54] 1.1.0 Version in the computing environment of 64-GB memory CPU and 1× 12-GB NVIDIA GeForce GTX 1080Ti GPU.

D. Visualization of SKAL

For intuitive understanding of our SKAL in Algorithm 1, a complete process of SKAL based on GoogleNet on a remote sensing image in UCM data set is shown in Fig. 4. For better explanation, all the coordinates of width and height are enlarged to the range of $[0, 25]$ from $[0, 1]$. ETr here is set to 60%.

In Fig. 4, the energy map is extracted from the original global image by Eqn (1), Eqn (5)-(7). And the initialized area of width and height, *i.e.* $[x_1, x_2, y_1, y_2]$=[3, 15, 7, 19], are obtained. Their initialized energy ratios are 69% and 52%, respectively. Because of the independence of the area searching process of spatial width and height, width $[x_1, x_2]$ and height $[y_1, y_2]$ are separately adjusted. Firstly, height area $[y_1, y_2]$ is iteratively enlarged to [5, 19] from [7, 19] with the energy ratio increasing from 52% to 62%. Secondly, the width area $[x_1, x_2]$ is iteratively shrunk from [3, 15] to [4, 14] with the energy ratio decreasing from 69% to 59%. Therefore, the final key area of $[x_{init}, x_{init}, y_{init}, y_{init}]$ is [4, 14, 5, 19]. After being scaled, this boundary is used for the guidance of the key area sampling.

Results in Fig. 4 indicate that the local image covers the most informative area in the global image, with the energy ratio of the structured vectors quickly converging to a small neighbor of 60%.

More localization samples are shown in Fig. 5 and Fig. 6. Fig. 5 shows some samples of discriminative classes which reflect the reasonable effect of key area location. Fig. 6 provides some samples of three ambiguous categories of "sparse_residential", "medium_residential" and "dense_residential". In Fig. 6, "sparse_residential" pays attention to the individual building, "medium_residential" focuses on the interface of buildings and trees, and "dense_residential" emphasizes a piece of houses next to each other which may be judged by the junction of houses.

E. In Comparison With Other Methods

1) UCM Data Set: UCM data set is the earliest and the most widely used remote sensing scene classification data set. Thus we first apply the SKAL based global-local two-stream architecture on UCM to explore the improvement of classification performance, and find a reasonable value of ETr. We make a hyperparameter tuning study based on AlexNet, ResNet18 and GoogleNet with *ETr* set to 60%, 70% and 80%, respectively. The ratios of training samples are 50% and 80%, which follow the splitting convention of UCM data set. Our results are shown in Table I. In the Table, the CNN models attached by the subscripts of $global$, $local$ and $global + local$ represent only the global stream (baseline), only the local stream and both of them, respectively.

According to the results, our SKAL based global-local two-stream architecture can provide a significant improvement for the classification of UCM data set under two kinds of splitting ratios. And it can be found that our method is not limited by the CNN models. Among the three CNN models, the performance of AlexNet obviously falls behind GoogleNet and ResNet18. Compared with ResNet18, there are some multi-scale convolutional blocks in GoogleNet, which is more suitable for our SKAL based two-stream method. Hence, when our two-stream architecture is applied, GoogleNet is slightly ahead of ResNet18.

We also make a comparison between the global stream and local stream as shown in Table I. Although the performance of local stream is far worse than the global stream, their fusion result is better than only the global stream. This phenomenon demonstrates the independence in feature extraction of the global stream and local stream, that is, the former extracts the global scene features while the latter mines the local key features. Besides, it is probably that the global stream plays a major role in the decision-making process and the local stream makes the compensation for it: if global stream has an clear classification judgment, local stream can enhance the result of global stream; on the contrary, if global stream hesitates between some ambiguous categories, local stream focusing on the enlarged key area can correct the possible mistaken classification results of global stream. As we can see in the table, the promotion effect decreases with the increase of local area, because the local features tend to be homogenous with the global features when the local area expands. Therefore it is necessary to limit the local area in a reasonable range,

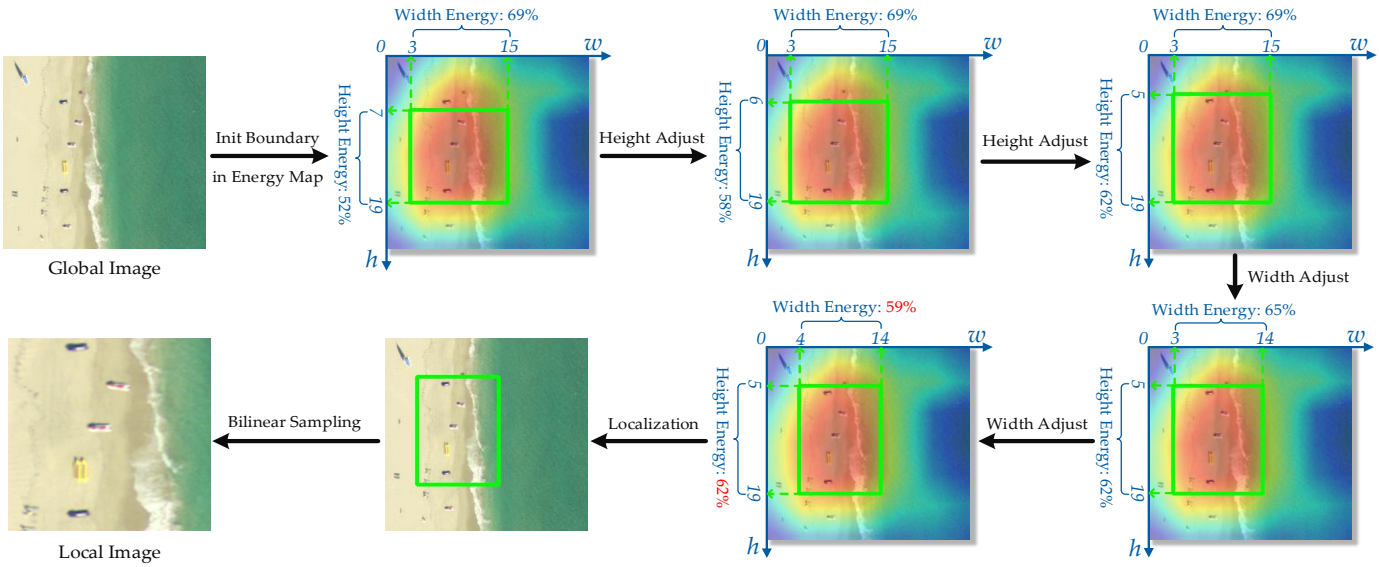
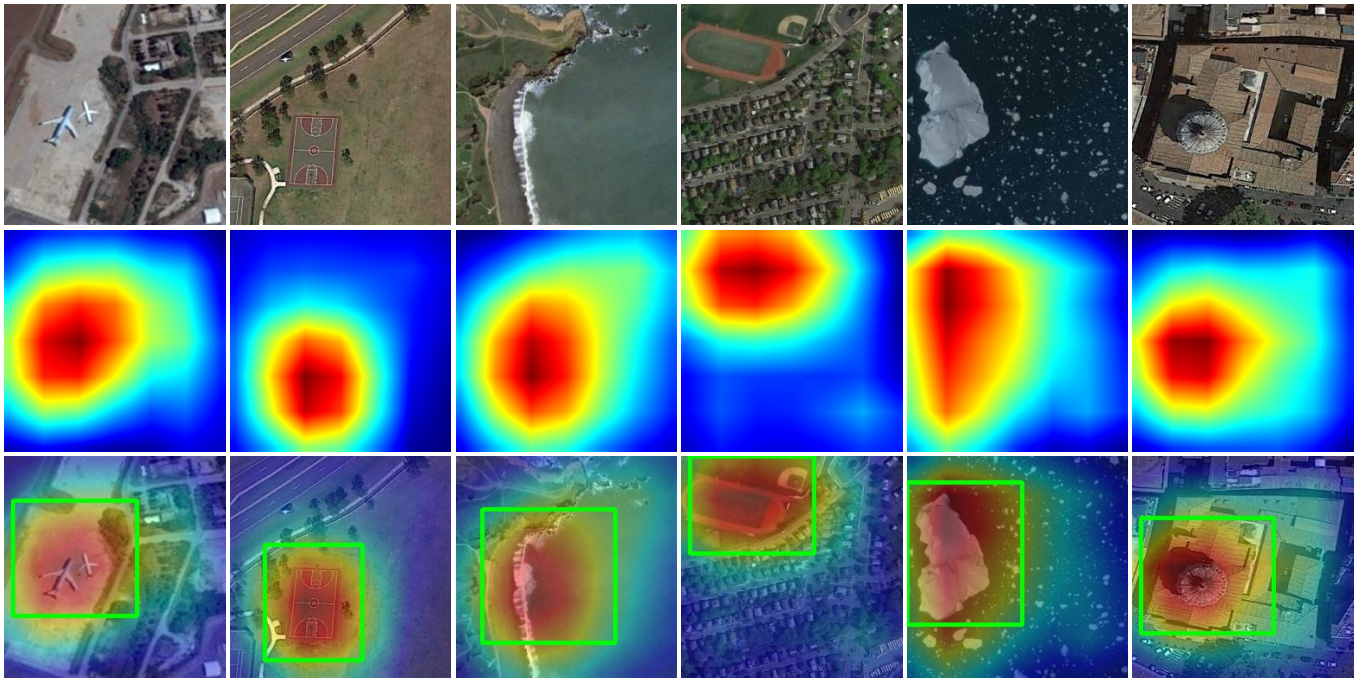


Fig. 4: Visualization of SKAL with ETr as 60%.

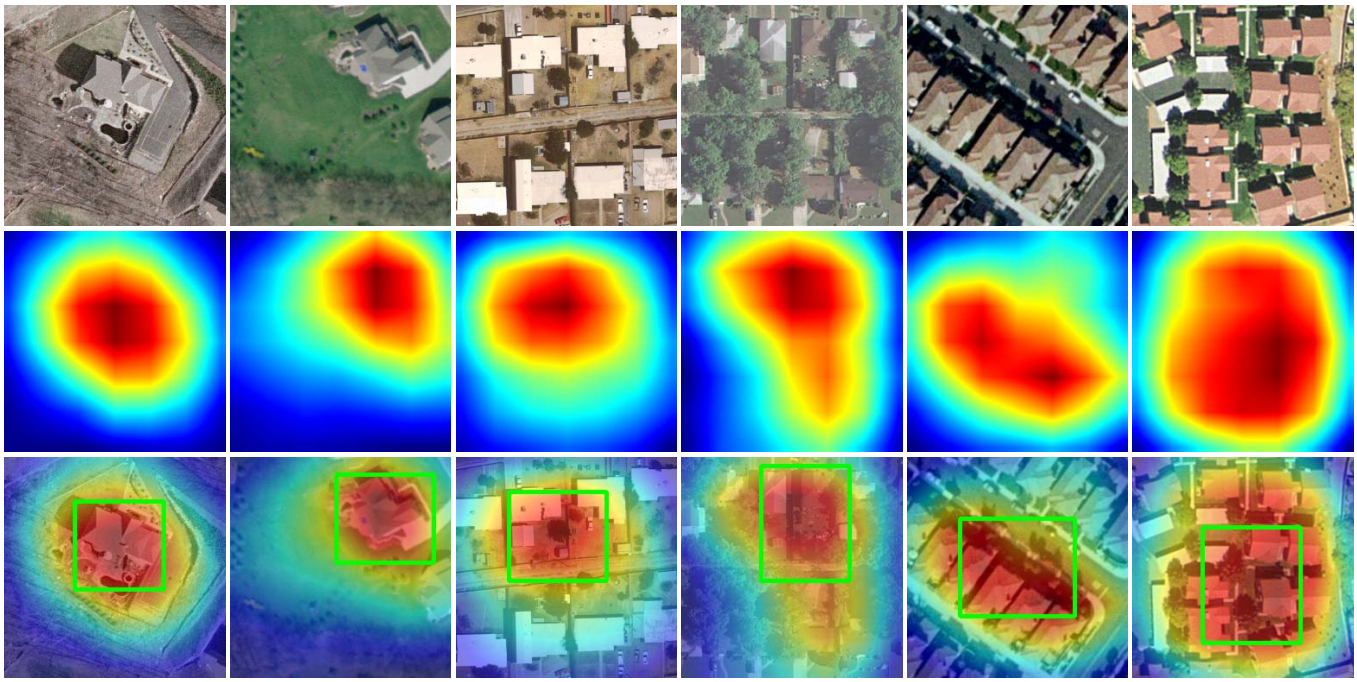


(a) airplane (b) basketbal_court (c) beach (d) ground_track_field (e) sea_ice (f) church

Fig. 5: Some samples of key area localization based on SKAL in NWPU-RESISC45 data set. The images from top to bottom are the following: original image, energy map and the fusion image labeled by the bounding box.

which is determined by *ETr*. Too small threshold leads the loss of useful local features while too big threshold decreases the discrimination of the local features. It could be found that all of 60%, 70% and 80% work well, but 70% is the best in most cases. In the following experiments, *ETr* is set to 70% by default.

Based on the comparison experiments of $global$ and $global + local$, the training and test time of the proposed SKAL based two-stream architecture on UCM data set is explored, and the results are provide in Table II. Because $local$ stream needs the bounding box of key area which depends on $global$ stream, there is no the corresponding time cost of only $local$. During the training stage when images are trained in a mini batch of 64 for 50 epochs, the training time of $global$ mainly contains feature extraction and gradient backward propagation in only $global$ while the training time of $global+local$ includes the feature extraction in $global$, SKAL, and the feature extraction and gradient backward propagation in $local$ stream. According to this table, it could be found that the training time of $global + local$ is just a few more



(a) sparse_residential_1 (b) sparse_residential_2 (c) medium_residential_1 (d) medium_residential_2 (e) dense_residential_1 (f) dense_residential_2

Fig. 6: Some samples of three ambiguous categories, including "sparse_residential", "medium_residential" and "dense_residential", in UCM data set. The images from top to bottom are the following: original image, energy map and the fusion image labeled by the bounding box.

TABLE I: OA(%) based on different CNN baselines with different training ratio and *ETr* on the UCM data set.

Methods	50% images for training			80% images for training		
	ETr=60%	*ETr*=70%	*ETr*=80%	*ETr*=60%	*ETr*=70%	*ETr*=80%
AlexNet$_{global}$	91.38±0.52	91.38±0.52	91.38±0.52	96.31±0.12	96.31±0.12	96.31±0.12
AlexNet$_{local}$	82.72±0.34	86.81±0.14	86.86±0.09	85.60±0.36	90.36±1.07	88.69±0.12
AlexNet$_{global+local}$	93.10±1.00	**93.77±1.28**	93.48±0.71	97.38±0.24	**97.38±0.48**	97.02±0.12
ResNet18$_{global}$	97.43±0.19	97.43±0.19	97.43±0.19	99.05±0.24	99.05±0.24	99.05±0.24
ResNet18$_{local}$	92.57±0.19	95.48±0.05	94.66±0.66	93.45±1.31	96.43±0.24	97.03±0.59
ResNet18$_{global+local}$	97.81±0.10	97.95±0.05	**98.15±0.05**	99.52±0.24	**99.52±0.24**	99.28±0.24
GoogleNet$_{global}$	97.76±0.19	97.76±0.19	97.76±0.19	98.81±0.71	98.81±0.71	98.81±0.71
GoogleNet$_{local}$	91.81±0.48	95.53±0.09	95.24±0.19	94.29±0.71	96.08±0.59	96.43±0.48
GoogleNet$_{global+local}$	97.90±0.10	**98.19±0.05**	98.14±0.24	99.41±0.36	**99.70±0.30**	99.41±0.36

than $local$. If the time of feature extraction in $local$ is taken into consideration, the time cost of SKAL in $global + local$ could be relatively ignored. During the testing stage when the images are test one by one, the test time of $global$ mainly consists of image loading feature extraction of $global$ while there are extra time cost of the proposed SKAL and feature extraction in $local$. Although the model size of $global + local$ is twice as big as $global$, test time of the former is less than twice time cost of the latter. It probably is caused by the image loading, and it also suggests that the proposed SKAL does not take the noticeable running time. Such results indicate that the proposed SKAL has a low computational complexity.

To objectively evaluate the performance of the proposed SKAL based global-local two-stream method, as shown in Table III, we make a comparison of OA(%) with some state-

TABLE II: Training time of 50% images of UCM data set and test time of the rest 50% images.

Methods	Training Time (Second)	Test Time (Second)
AlexNet$_{lobal}$	123	2.62
AlexNet$_{global+local}$	144	4.30
ResNet18$_{lobal}$	196	3.75
ResNet18$_{global+local}$	236	6.40
GoogleNet$_{lobal}$	209	3.81
GoogleNet$_{global+local}$	255	6.44

of-the-art methods on UCM data set, including handcrafted feature based methods, unsupervised feature based methods and deep feature based methods. As we can see in Table III, our method significantly outperforms all the other state-of-the-art scene classification methods. When 50% images are used for training, our two-stream method wins the first place with an obvious increase of 1.38% over the second method of ARCNet-VGG16 [8]. When the training ratio increases to 80%, compared with the third method of ARCNet-VGG16 [8], our GoogleNet based two-stream architecture has an significant gain of 0.58%, in consideration of the near 100% OA. Although Resnet101-FSL [6] is much deeper and bigger than GoogleNet, our global-local two-stream architecture based on GoogleNet still outperforms Resnet101-FSL. It can also be found that our method has huge advantages in terms of classification accuracy than the handcrafted feature methods [55], [56] and unsupervised methods [30].

TABLE III: Comparison of OA(%) with some state-of-the-art results on UCM data set.

Methods	50% training	80% training
AlexNet	91.38±0.52	96.31±0.12
AlexNet$_{global+local}$	93.77±1.28	97.38±0.48
ResNet18	97.43±0.19	99.05±0.24
ResNet18$_{global+local}$	97.95±0.05	99.52±0.24
GoogleNet	97.76±0.19	98.81±0.71
GoogleNet$_{global+local}$	**98.19±0.05**	**99.70±0.30**
Resnet101-FSL [6]	—	**99.52**
ARCNet-VGG16 [8]	**96.81±0.14**	99.12±0.40
DDRL-AM [9]	—	99.05±0.08
SF-CNN with VGGNet [13]	—	99.05±0.27
MSCP [57]	—	98.36±0.58
ELM based Two-Stream [11]	96.97±0.75	98.02±1.03
TEX-Net-LF [12]	96.91±0.36	97.72±0.54
Combing Scenarios I and II [38]	—	98.49
Fusion by Addition [58]	—	97.42±1.79
CNN-NN [59]	—	97.19
SalM3LBPCLM [55]	94.21±0.75	95.75±0.80
VGG-VD-16 [20]	94.14±0.69	95.21±1.20
MS-CLBP+FV [56]	88.76±0.76	93.00±1.20
Unsupervised Feature Learning [30]	—	81.67±1.23

Moreover, as shown in Fig. 7, we make two CMs of GoogleNet and the corresponding two-stream architecture to further analyze the improvement of each class of UCM data set. As it can be observed in Fig. 7, there are some misclassified samples among the scenes of medium_residential, dense_residential, buildings, storage_tanks and intersection. When the proposed two-stream architecture is applied, the number of the misclassified samples and scenes decreases rapidly, which proves the effectiveness of our method.

2) RSSCN7 Data Set: RSSCN7 data set is a difficult remote sensing scene data set, affected by the changing seasons, various weathers and scale diversity. These problems challenge the stability and robustness of the proposed global-local two-stream architecture in key local area localization. To study the performance of our method in dealing with these problems, the comparative experiments are conducted based on AlexNet, ResNet18 and GoogleNet (with *ETr* set to the default value of 70%), and the experimental results of OA(%) are reported in Table IV. In the table, the subscript of *global* is removed and the results of the local streams are not provided for more clear comparison.

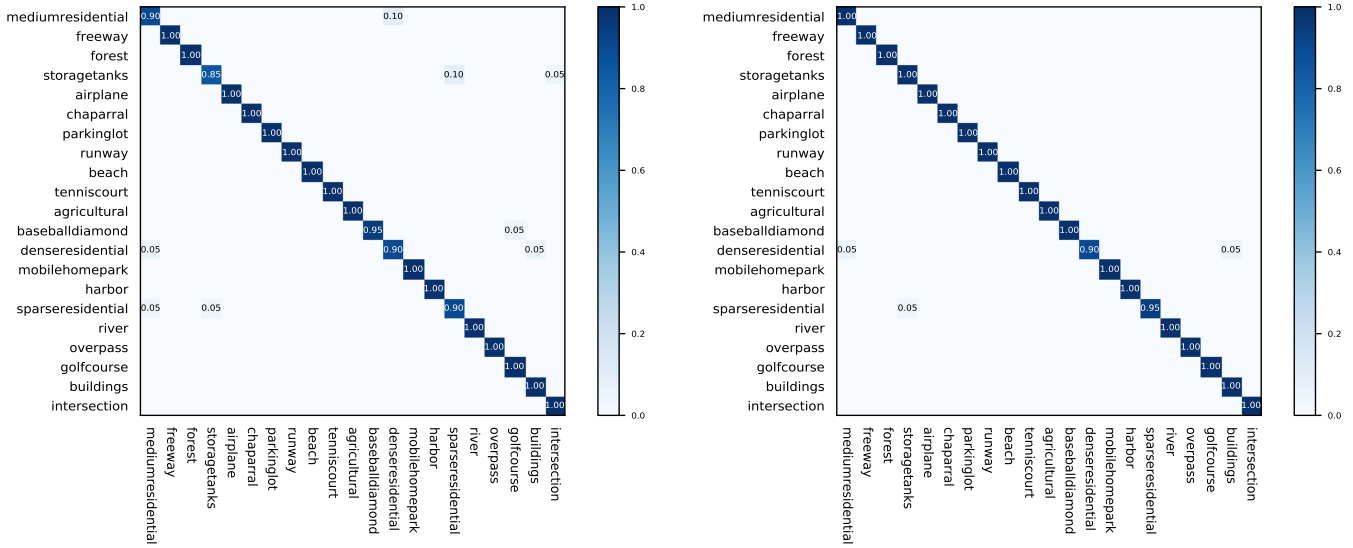
Table IV results indicate that the increase of about 2% is obtained when the proposed two-stream architecture is applied over these three CNN baselines. The wide and meaningful improvement powerfully supports the stability and effectiveness of our method. Especially, when the CNN baselines are limited by the lack of training data, the local images can also be regarded as the extra training samples from the perspective of data augmentation. Compared with all the state-of-the-art methods, the proposed method has a huge advantage in the results of OA. Our global-local two-stream architecture based on ResNet18 wins the first place and has the accuracy gain of 1.44% under the training ratio of 20%, and 2.04% under the training ratio of 50%, over the second method of Resnet50 based TEX-Net-LF [12].

To study the performance of each class in RSSCN7 data set based on the proposed two-stream architecture, the CM is made as shown in Fig 8. According to this CM, it can be found that the missclassified samples are mainly distributed in the scenes of industry, parking, grass and field. There are large intraclass diversity and high interclass similarity in these scenes, which need to be solved in the future work.

TABLE IV: Comparison of OA(%) with some state-of-the-art results on RSSCN7 data set.

Methods	20% training	50% training
AlexNet	88.59±0.36	91.97±0.17
AlexNet$_{global+local}$	90.81±0.15	93.35±0.35
ResNet18	91.81±0.40	94.61±0.75
ResNet18$_{global+local}$	**93.89±0.52**	**96.04±0.68**
GoogleNet	90.73±0.27	93.97±0.30
GoogleNet$_{global+local}$	93.41±0.26	95.75±0.21
Resnet50 [12]	90.23±0.43	93.12Q±0.55
Resnet50 based TEX-Net-LF [12]	**92.45±0.45**	**94.00±0.57**
VGG-M [12]	86.00±0.63	88.80±0.55
VGG-M based TEX-Net-LF [12]	88.61±0.46	91.25±0.58
Deep filter banks [60]	—	90.40±0.60
CaffeNet [20]	85.57±0.95	88.25±0.62
VGG-VD-16 [20]	83.98±0.87	87.18±0.94
GoogleNet [20]	82.55±1.11	85.84±0.92
HCV [60]	—	84.70±0.70
DBN based feature selection [19]	—	77.0

3) AID Data Set: AID is a high-resolution large-scale remote sensing scene data set covering a lot of background noise. The comparative experiments are conducted on AID data set based on the aforementioned three CNN baselines under the default training settings and ETr. The results achieved by our two-stream architecture of OA(%) are provided in Table V, with the comparison with some state-of-the-art results. It is notable that the state-of-the-art results of CNN-based methods on AID data set reported in this paper are based on the image



(a) GoogleNet (b) GoogleNet based two-stream architecture

Fig. 7: Confusion matrix of UCM Data Set under the training ratio of 80% using the following two methods: GoogleNet (global stream) and the proposed SKAL based global-local two-stream architecture based on GoogleNet.

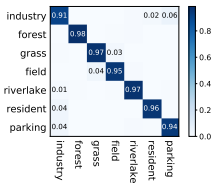


Fig. 8: Confusion matrix of RSSCN7 Data Set under the training ratio of 50% using the proposed SKAL based global-local two-stream architecture based on ResNet18.

size of 224×224 within a small variance, because image size has an important influence on the classification accuracy.

Results of Table V indicate that the proposed global-local two-stream architecture provides two kinds of advantages for CNN baselines. On one hand, our two-stream architecture significantly and widely improves the performance of all the three CNN baselines. And our global-local two-stream architecture based on ResNet18 outperforms all the current state-of-the-art methods, with the advantage of at least 0.78% in OA under the training ratio of 50% and 0.57% under the training ratio of 20%. On the other hand, our two-stream architecture reduces the standard deviation of experimental results and provides more stable classification accuracies although the training images are randomly selected from the data set for several times.

In order ot evaluate the performance of the proposed method on AID data set, we make a CM in Fig. 9, which is achieved by the global-local two-stream architecture based on ResNet18. As shown in this CM, the most easily misclassified scenes are park, resort, railway station, square, school and center. All of them are strongly related to the plenty of buildings and vegetation cover. These highly similar features and objects limits the further improvement on AID data set, and this problem could be improved by deeper and more complex feature representation.

TABLE V: Comparison of OA(%) with some state-of-the-art results on Aerial Image Data Set.

Methods	20% training	50% training
AlexNet	85.30±0.48	90.75±0.41
AlexNet$_{global+local}$	88.26±0.27	92.54±0.12
ResNet18	93.36±0.12	95.51±0.31
ResNet18$_{global+local}$	**94.38±0.10**	**96.76±0.20**
GoogleNet	93.12±0.31	95.64±0.18
GoogleNet$_{global+local}$	94.26±0.15	96.65±0.11
Resnet101-FSL [6]	—	**95.88**
Resnet50 based TEX-Net-LF [12]	**93.81±0.12**	95.73±0.16
ELM based Two-Stream [11]	92.32±0.41	94.58±0.25
VGG-VD16 + MSCP [57]	91.52±0.21	94.42±0.17
ARCNet-VGG16 [8]	88.75±0.40	93.10±0.55
VGG-M based TEX-Net-LF [12]	90.87±0.11	92.96±0.18
Fusion by Addition [58]	—	91.87±0.36
SalM3LBPCLM [55]	86.92±0.35	89.76±0.45
VGG-VD-16 [20]	86.59±0.29	89.64±0.36
CaffeNet [20]	86.86±0.47	89.53±0.31
MS-CLBP+FV [55]	86.48±0.27	—
GoogLeNet [20]	83.44±0.40	86.39±0.55

4) NWPU-RESISC45 Data Set: NWPU-RESISC45 is the biggest remote sensing data set of scene classification with 45 challenging scenes. Benefiting from the large amount of training data, the classification results of NWPU-RESISC45 are more stable and convincing. We conduct the ablation experiments on NWPU-RESISC45 under the same experimental conditions of CNN baselines, training settings and *ETr* as the previous three data sets.

We make the comparative experiments with/without the

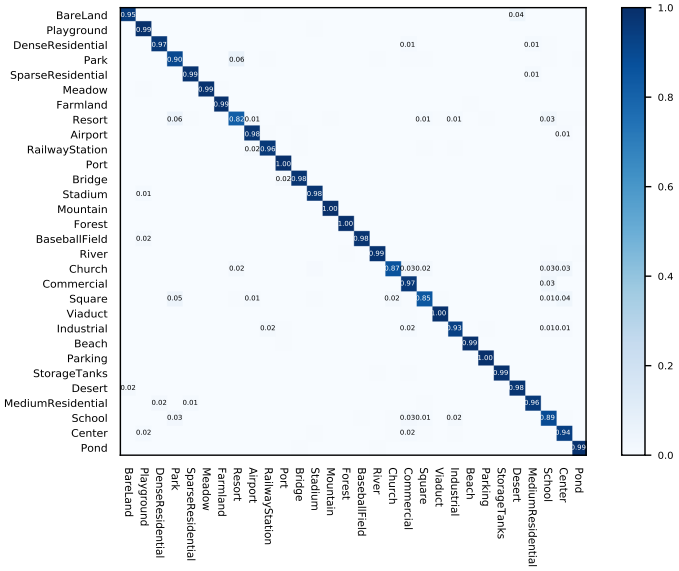


Fig. 9: Confusion matrix of Aerial Image Data Set under the training ratio of 50% using the proposed SKAL based global-local two-stream architecture based on ResNet18.

TABLE VI: Comparison of OA(%) with some state-of-the-art results on NWPU-RESISC45 Data Set.

Methods	10% training	20% training
AlexNet	77.44±0.28	83.69±0.25
AlexNet$_{global+local}$	80.28±0.16	85.34±0.14
ResNet18	88.91±0.23	91.77±0.18
ResNet18$_{global+local}$	90.04±0.15	92.79±0.11
GoogleNet	89.40±0.25	91.93±0.16
GoogleNet$_{global+local}$	**90.41±0.12**	**92.95±0.09**
SF-CNN with VGGNet [13]	89.89±0.16	**92.55±0.14**
ResNet-18 + AM + CL [9]	**92.17±0.08**	92.46±0.09
VGGNet-16 + RIFD [61]	90.12	92.27
D-CNN with VGGNet-16 [14]	89.22±0.50	91.89±0.22
VGG-VD16 + MSCP + MRA [57]	88.07±0.18	90.81±0.13
SAL-TS-Net [62]	85.02±0.25	87.01±0.19
TEX-TS-Net [62]	84.77±0.24	86.36±0.19
ELM based Two-Stream [11]	80.22±0.22	83.16±0.18
AlexNet [21]	76.69±0.21	79.85±0.13
BoVW + SPM [21]	27.83±0.61	32.96±0.47
LBP [21]	19.20±0.41	21.74±0.18

proposed two-stream architecture, and make a comparison of OA(%) with some state-of-the-art methods, which are shown in Table VI. According to the results, our GoogleNet based global-local two-stream architecture wins the second place under the training ratio of 10% and the first place under 20%. The current best method of ResNet18 + AM + CL [9] has an OA gain of 0.29% when training ratio increases from 10% to 20%. However, our GoogleNet based two-stream architecture has an OA gain of 2.54%, which indicates that our method has better potential of OA with the increase of training samples.

The CM of our ResNet18 based two-stream architecture is reported in Fig 10. The samples most likely to be misclassified mostly belong to the classes of freeway, church, railway station, industrial area, palace, commercial area, wetland, river and medium residential. There are lots of confusing objects and features among these scene classes that limits the OA.

V. Conclusion

In this paper, we propose a structured key area localization (SKAL) strategy to localize the most important area in remote sensing scene images. Based on SKAL, a global-local two-stream architecture, which can individually extract the global and local features, is further presented for scene classification of remote sensing images. To verify the effectiveness and robustness of the proposed SKAL based global-local two-stream architecture, we conduct a lot of comparative experiments based on three kinds of widely used CNN models, including AlexNet, ResNet18 and GoogleNet, on four popular remote sensing scene data sets, and achieve all the state-of-the-art results of these data sets. The experimental results demonstrate the powerful capability of the joint global and local feature representation of the proposed method, which can solve the problem of large scale variance in remote sensing scene images to some extent.

References

[1] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[2] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 950–965, 2017.

[3] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2017.

[4] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-scale cropping mechanism for remote sensing image captioning," in *2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2019.

[5] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[6] W. Huang, Q. Wang, and X. Li, "Feature sparsity in convolutional neural networks for scene classification of remote sensing image," in *2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2019.

[7] L. Yan, R. Zhu, N. Mo, and Y. Liu, "Improved class-specific codebook with two-step classification for scene-level classification of high resolution remote sensing images," *Remote Sensing*, vol. 9, p. 223, 03 2017.

[8] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, 2018.

[9] J. Li, D. Lin, Y. Wang, G. Xu, and C. Ding, "Deep discriminative representation learning with attention map for scene classification," *arXiv preprint arXiv:1902.07967*, 2019.

[10] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE transactions on neural networks and learning systems*, 2019.

[11] Y. Yu and F. Liu, "A two-stream deep fusion framework for high-resolution aerial scene classification," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[12] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS journal of photogrammetry and remote sensing*, vol. 138, pp. 74–85, 2018.

[13] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2019.

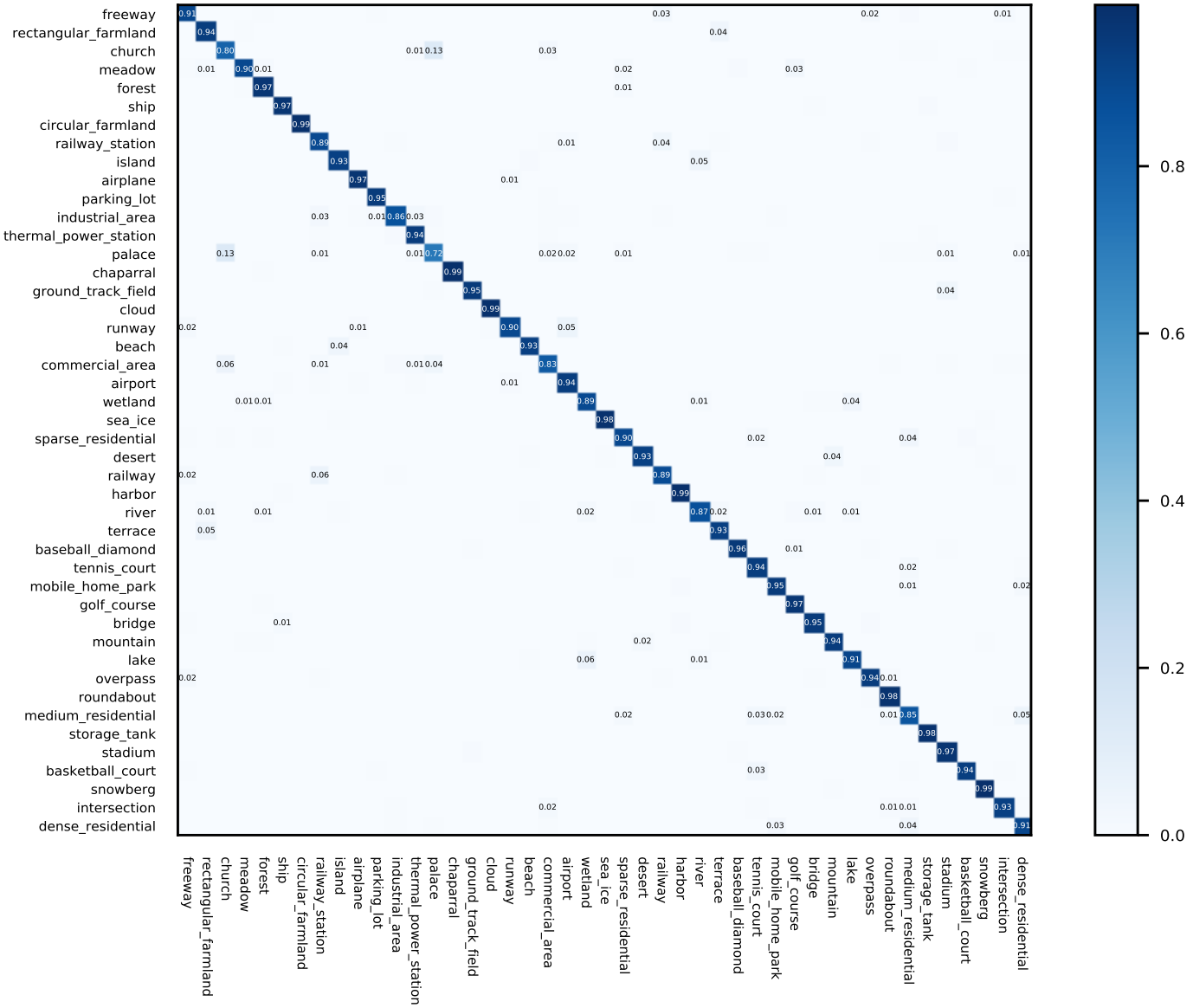


Fig. 10: Confusion matrix of NWPU-RESISC45 data set under the training ratio of 50% using the proposed SKAL based global-local two-stream architecture based on ResNet18.

[14] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," *IEEE transactions on geoscience and remote sensing*, vol. 56, no. 5, pp. 2811–2821, 2018.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[18] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. ACM, 2010, pp. 270–279.

[19] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2321–2325, 2015.

[20] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[21] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: mark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[22] J. A. dos Santos, O. A. B. Penatti, and R. da Silva Torres, "Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification." in *VISAPP (2)*, 2010, pp. 203–208.

[23] S. Bhagavathy and B. S. Manjunath, "Modeling and detection of geospatial objects using texture motifs," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 12, pp. 3706–3715, 2006.

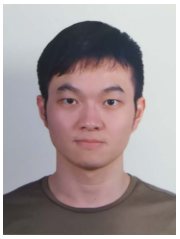
[24] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 98, pp. 119–132, 2014.

[25] G. Cheng, P. Zhou, J. Han, L. Guo, and J. Han, "Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images," *IET Computer Vision*, vol. 9, no. 5, pp. 639–647, 2015.

[26] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Transactions on Geoscience and Remote Sensing*,

vol. 51, no. 2, pp. 818–832, 2012.
[27] V. Risojević and Z. Babić, "Fusion of global and local descriptors for remote sensing image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 4, pp. 836–840, 2012.
[28] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 524–531.
[29] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 6, pp. 747–751, 2016.
[30] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 439–451, 2013.
[31] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *International journal of remote sensing*, vol. 33, no. 8, pp. 2395–2412, 2012.
[32] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, pp. 173–176, 2010.
[33] Y. Li, C. Tao, Y. Tan, K. Shang, and J. Tian, "Unsupervised multilayer feature learning for satellite image scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 2, pp. 157–161, 2016.
[34] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5148–5157, 2017.
[35] W. Zhou, Z. Shao, C. Diao, and Q. Cheng, "High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder," *Remote sensing letters*, vol. 6, no. 10, pp. 775–783, 2015.
[36] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.
[37] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5653–5665, 2017.
[38] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14 680–14 707, 2015.
[39] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1307–1314.
[40] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2846–2854.
[41] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 189–203, 2016.
[42] D. Zhang, D. Meng, L. Zhao, and J. Han, "Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning," *arXiv preprint arXiv:1703.01290*, 2017.
[43] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4438–4446.
[44] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. L. Yuille, "Pcl: Proposal cluster learning for weakly supervised object detection," *IEEE transactions on pattern analysis and machine intelligence*, 2018.
[45] Z. Yang, D. Mahajan, D. Ghadiyaram, R. Nevatia, and V. Ramanathan, "Activity driven weakly supervised object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2917–2926.
[46] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
[47] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
[48] T. Hoeser and C. Kuenzer, "Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends," *Remote Sensing*, vol. 12, no. 10, p. 1667, 2020.
[49] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
[51] R. Pires de Lima and K. Marfurt, "Convolutional neural network for remote-sensing scene classification: Transfer learning analysis," *Remote Sensing*, vol. 12, no. 1, p. 86, 2020.
[52] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.
[55] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 6, pp. 2889–2901, 2017.
[56] L. Huang, C. Chen, W. Li, and Q. Du, "Remote sensing image scene classification using multi-scale completed local binary patterns and fisher vectors," *Remote Sensing*, vol. 8, no. 6, p. 483, 2016.
[57] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 6899–6910, 2018.
[58] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for vhr remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4775–4784, 2017.
[59] E. Othman, Y. Bazi, N. Alajlan, H. Alhichri, and F. Melgani, "Using convolutional features and a sparse autoencoder for land-use scene classification," *International Journal of Remote Sensing*, vol. 37, no. 10, pp. 2149–2167, 2016.
[60] H. Wu, B. Liu, W. Su, W. Zhang, and J. Sun, "Deep filter banks for land-use scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1895–1899, 2016.
[61] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 265–278, 2018.
[62] Y. Yu and F. Liu, "Dense connectivity based two-stream deep feature fusion framework for aerial scene classification," *Remote Sensing*, vol. 10, no. 7, p. 1158, 2018.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science, with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.



Wei Huang received the B.E. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently working toward the M.S. degree in computer science in the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include deep learning and computer vision.

Zhitong Xiong received the M.E. degree from the Northwestern Polytechnical University, Xian, China, where he is currently pursuing the Ph.D. degree with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL). His research interests include computer vision and machine learning.

Xuelong Li (M'02-SM'07-F'12) is currently a Professor with the School of Computer Science, with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China.