IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 6, 2019, accepted May 21, 2019, date of publication May 23, 2019, date of current version June 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918732

Global-Local Attention Network for Aerial Scene Classification

YIYOU GUO[1], JINSHENG JI[1], XIANKAI LU[1], HONG HUO[1], TAO FANG[1], AND DEREN LI[2], (Senior Member, IEEE)

[1]Department of Automation, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Corresponding author: Tao Fang (tfang@sjtu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0505000, in part by the National Science and Technology Major Project under Grant 21-Y20A06-9001-17/18, in part by the National Natural Science Foundation of China under Grant 41571402, and in part by the Science Fund for Creative Research Groups of the National Natural Science Foundation of China under Grant 61221003.

ABSTRACT The classification performance of aerial scenes relies heavily on the discriminative power of feature representation from high-spatial resolution remotely sensed imagery. The convolutional neural networks (CNNs) have recently been applied to adaptively learn image features at different levels of abstraction rather than requiring handcrafted features and achieved state-of-the-art performance. However, most of these networks focus on multi-stage global feature learning yet neglect the local information, which plays an important role in scene recognition. To address this issue, a novel end-to-end global-local attention network (GLANet) is proposed to capture both global and local information for aerial scene classification. FC layers in the VGGNet are replaced by the global attention (GA) branch and local attention (LA) branch, one of which learns the global information while the other learns the local semantic information via attention mechanisms. During each training, the labels of input images can be predicted by the local, global, and their concatenated features using softmax. According to different predicted labels, two auxiliary loss functions are further computed and imposed on the proposed network to enhance the supervision for network learning. The experimental results on three challenging large-scale scene datasets demonstrate the effectiveness of the proposed global-local attention network.

INDEX TERMS Scene classification, global-local attention network, deep learning, remote sensing.

I. INTRODUCTION

With the rapid development of satellite sensors, a huge amount of earth observation images have become readily available [1]–[4]. Due to the sharply increasing spatial resolution, the data volume of images grows dramatically, providing more detailed information including shape, texture, and so on. It is highly desirable to interpret remote sensing images with high spatial resolution in intelligent and automatic approaches [5]–[7]. Given this situation, aerial scene classification has become an increasingly important topic of image understanding for high spatial resolution remote sensing. It has attracted remarkable attention in academia and facilitated a wide range of applications, such as precise land-use/land-cover investigation, hazard detection, environmental monitoring, urban planning, traffic supervision, smart city, and weapon guidance [8]–[12].

In terms of image content, aerial scene classification attempts to automatically assign a semantic category label (e.g., residential area or commercial area) to each remote sensing scene containing multiple objects (e.g., grass, water, buildings) [13]–[15]. As these images may be obtained under different conditions (e.g., locations, times, sensors, weather and so on), the identical objects (e.g., buildings) even in the same scene category(e.g., residential area) frequently appear with large variations in orientations, sizes, colors, and angles, leading to high intraclass variance and low interclass variance. Moreover, a scene of the same category often consists of different objects while scenes of different categories may share some identical objects. For instance, building, road, and tree widely exist both in the residential area and the commercial area [16]. In short, the complexities of aerial

The associate editor coordinating the review of this manuscript and approving it for publication was Geng-Ming Jiang.

scene images lie in the high diversity of appearances, the great dissimilarity of geometries, and the complexities in scene semantics [16]. How to precisely represent the image content of these complex scenes is an important challenge for aerial scene classification [17], [18]. An effective feature representation plays a key role in aerial scene classification. There have been many efforts over the years to solve this problem and numerous approaches have been proposed. Existing aerial scene classification methods can generally be divided into two aspects according to feature generation: (a) methods using handcrafted features; and (b) methods using learned features, especially deep-learned features.

In the early days, most of the aerial scene classification methods are based on handcrafted features, including low-level and mid-level features. The former assumes that the same category of images should share certain statistically holistic attributes, and mainly concentrated on constructing various handcrafted features, such as color, texture, scale invariant feature transform (SIFT), GIST, and histograms of oriented gradients (HOG) or their combination [19]–[22]. In practical applications, these methods may not well capture semantic information contained in complex aerial scene images, and are of limited performance. To overcome this issue, the latter mainly attempt to develop a set of feature encoding methods to aggregate low-level features. As a popular mid-level approach, bag-of-words (BOW) model has attracted growing attention, many extension methods have been proposed for aerial scene classification [23]–[26]. In addition, the other category of mid-level representation is the topic model, such as latent Dirichlet allocation (LDA) and probabilistic latent semantic analysis (pLSA) [27]–[30]. However, it is worth noticing that mid-level features heavily depend on the extraction of low-level features. In other words, the aforementioned feature representations, no matter low-level or mid-level, require domain expert experiences to find the best design of a given dataset, which may lack the flexibility and adaptivity to different remote sensing datasets.

In recent years, much work has focused on automatically learning discriminative feature representations such as deep learning, so-called high-level features [31]. Convolutional neural networks (CNNs) have recently been applied to aerial scene classification and achieved state-of-the-art performance. Overcoming the limitation of feature extraction using deep convolutional neural networks at the same scale, Chen *et al.* [8] present a hybrid DNN to extract variable-scale features for aerial images. In order to fully explore the benefits of multilayer features for scene recognition, Li *et al.* [18] present a fusion strategy for integrating multilayer features of the pre-trained CNN model. Liu *et al.* [32] propose to fuse different pre-trained CNN models for representing scene images. Overall, these works usually leverage the output of fully connected (FC) layer as the global representation of images and neglect the local information. However, it has been proved that local part information plays an important role in scene classification [33]. Therefore, on the basis of CNN learning, how to further learn global context features and local semantic parts more accurately is an open issue for aerial scene classification.

Visual attention which comes from human perception not only tells where to focus but also improves the representation of interests [34]. Inspired by recent advances of attention mechanism in the deep neural network [35]–[37], an end-to-end global-local attention network (GLANet) is proposed to capture both global and local information for aerial scene classification. In order to alleviate the overfitting issue caused by excessive parameters in FC layers, FC layers in the backbone network are replaced by the global attention (GA) branch and local attention (LA) branch. First, the images are fed into convolution layers to generate feature maps. Second, based on the feature maps, global attention branch is applied to aggregate the global information on the feature channel level through a squeeze-excitation block [35]. This squeeze-excitation block learns the channel-wise relationships which can be regarded as a global attention procedure. On the other hand, a spatial residual attention branch is proposed to learn the semantic regions within the feature maps [38]. The learned spatial attention map focuses on the most discriminative part which corresponds to the semantic part in images (Fig.1). Then, the feature representations from two learned branches are further concatenated to obtain better image representation. In addition, according to different labels predicted by three different image representations each training, two auxiliary loss functions are computed and imposed on the proposed network to enhance the supervision for network learning. During network testing, the concatenated local and global features are directly applied to scene classification. Finally, the proposed global-local attention network can be trained with small training samples and achieves better performance on AID [39], NWPU-RESISC45 [40], and PatternNet [41].

The remainder of this paper is organized as follows. Section II describes the proposed end-to-end global-local attention network and its learning in detail. Section III is devoted to the experimental results and analysis. The main concluding remarks are mentioned in Section IV.

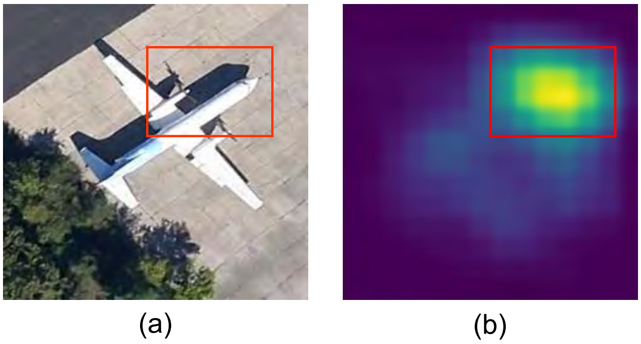


FIGURE 1. A local semantic part learned by GLANet. (a) Denotes the semantic part (the red box) in given image. (b) Denotes the attention region at the feature map level.

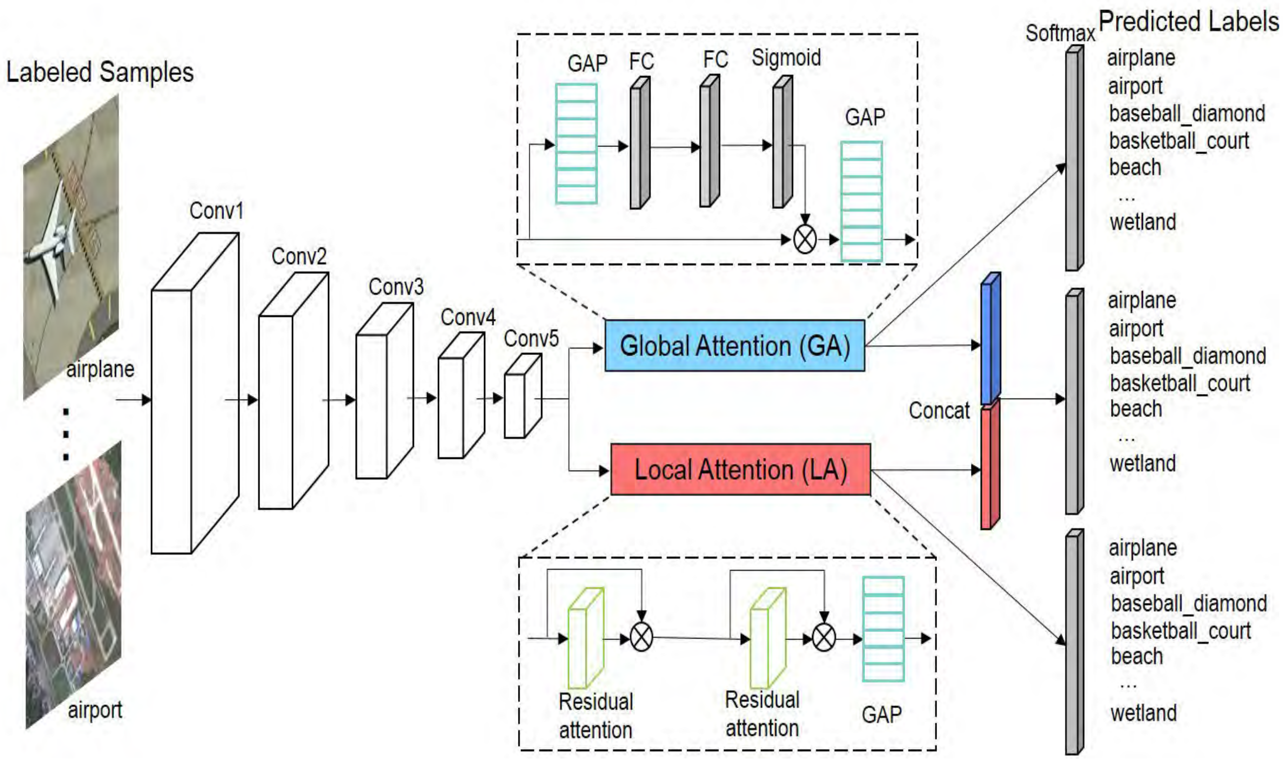


FIGURE 2. Framework of GLANet.

II. GLOBAL-LOCAL ATTENTION NETWORK

In this section, network architecture and its learning are introduced, as shown in Fig.2. The VGGNet up to *Conv*5 layer is leveraged as the backbone network for subsequent global and local feature extraction. Meanwhile, the two attention branches are used to obtain more discriminative information.

A. NETWORK ARCHITECTURE

1) BACKBONE NETWORK

The backbone network of our GALNet is a pre-trained fully convolutional neural network on ImageNet [42], which consists of the first five convolution blocks (i.e., *Conv*1-*Conv*5 in Fig.2) from VGGNet-16 [43]. Each block contains several convolutional layers and one max-pooling layer. In this way, the output feature maps at the end of the backbone network contain high-level semantic information as well as sufficient spatial information. Given an input image, we can extract the feature map X through the backbone network with size $H \times W \times C$. X is a 3D-tensor with the height H, width W and channels C. We substitute the three connected layers in origin VGGNet-16 with the proposed global attention branch and the local attention branch. We will introduce the detailed architecture of these two branches.

2) GLOBAL ATTENTION BRANCH

The top part in Fig.2 is a squeeze-excitation module [35] that performs global information abstraction. In the first GAP layer, the values of different channels can be regarded as the global context information of the image via global average pooling:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{X}_c(i,j) \tag{1}$$

where $\mathbf{X}_c \in \mathbb{R}^{H \times W}$ is a deep feature map with width W and height H for the *cth* channel, extracted by the backbone network of GLANet. $\mathbf{z} = \{z_c | c = 1, \ldots C\} \in \mathbb{R}^C$ is the global statistical representation of $\mathbf{X}$, generally, total number of channels $C = 512$. Then, after the first GAP, two small-node FC layers are constructed, and the corresponding connection weight matrices are $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively. In order to capture the channel-wise dependencies, after training the two FC layers, a weight factor $\mathbf{o}_c$ for the cth channel can be learned via a sigmoid layer using an attention mechanism [35]:

$$\mathbf{o}_c = \sigma(\mathbf{W}_2 c \eta(\mathbf{W}_1 c\, z_c)), \tag{2}$$

where $\mathbf{W}_1 c$ and $\mathbf{W}_2 c$ represent weight vectors of two FC layers corresponding to the cth channel, respectively, η is the relu activation function, and σ denotes the sigmoid function. The global branch is introduced to learn the channel-wise weights for feature maps and fuse information of them. As demonstrated in Fig.3, the global branch learns to assign different weights to different channels, the higher weight means the more important of that channel.

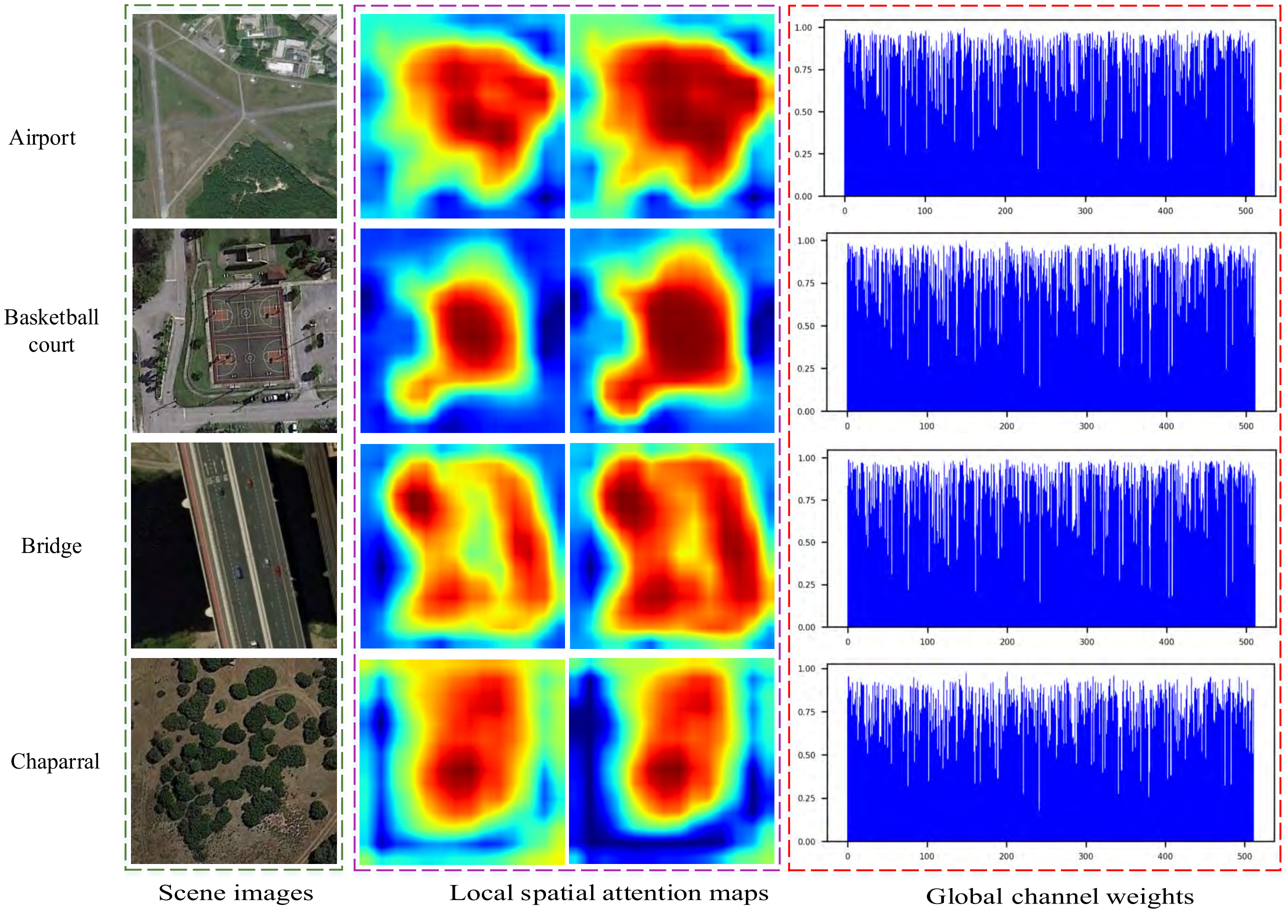


FIGURE 3. left column is the input image, the middle two columns are the local spatial attention maps from the two residual attention modules, and the right column means the global channel weights.

After the weight factor acts on the feature map, then a global attention feature f_c^g for the cth channel is obtained via the second GAP layer:

$$f_c^g = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (o_c \cdot \mathbf{X}_c(i,j)) \tag{3}$$

where $\cdot$ denotes the channel-wise multiplication operation. The GAP layer acts in a manner native to the convolution structure by enforcing correspondences between feature maps and categories [44]. Finally, the global attention representation is described as follows:

$$\mathbf{F}_{GA} = (f_1^g, f_2^g, \ldots, f_{C-1}^g, f_C^g) \tag{4}$$

3) LOCAL ATTENTION BRANCH

In addition to the global information obtained through training global attention branch above, more discriminative semantic information are further learned from the same feature maps via a local attention branch (LA). As shown in the bottom part of Fig. 2, the LA branch utilizes two successive residual attention modules, one of which consists of a spatial attention layer and a residual connection. The spatial attention layer learns to apply weighted mask to the input feature maps. As shown in Fig. 4, the residual attention module is similar to identity mapping in the deep residual network. The feature map $\mathbf{X}$ performs convolution operation to obtain attention map $\phi(\mathbf{X})$, $\mathbf{W}$ represents connection weight the feature map $\mathbf{X}$ to the attention map $\phi(\mathbf{X})$. Then, a summarized map can be computed as follows:

$$s = g(\mathbf{W} * \mathbf{X} + b) \tag{5}$$

where $*$ denotes the convolution operation, b denotes bias on the deep features $\mathbf{X}$, g is a nonlinear function, and s is the attention map. Then attention map $\phi(\mathbf{X})$ can be obtained through further normalizing s to [0, 1]:

$$\phi(\mathbf{X}) = \frac{\exp(s(l))}{\sum_{l' \in L} \exp(s(l'))} \tag{6}$$

where $L = \{l' = (i,j) | i = 1, \ldots, W, j = 1, \ldots, H\}$. ϕ is the final spatial attention map applied to all channels.

For each convolutional layer, different convolution filters are sensitive to one certain character of the image [45], which can be formatted as local feature representations. In this paper, we build two residual modules to further enhance the feature of local attention regions as demonstrated in Fig.3. The two connected residual modules output the attention maps, which contains spatial information. Therefore, the attention region unusually achieves high responses

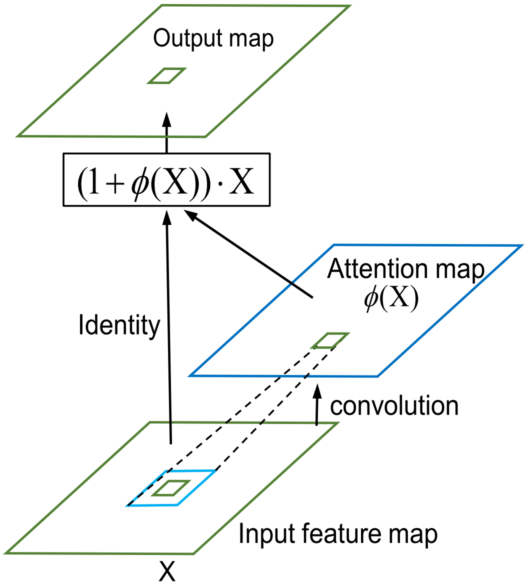


FIGURE 4. The proposed residual attention scheme.

(i.e., deeper color in Fig.3) whereas the irrelevant background achieves low responses on the local attention maps. From Fig.3, we can observe that these discriminative regions usually achieve high neural response on the two local attention maps.

Finally, the local attention representation $\mathbf{F}_{LA}$ of $\mathbf{X}$ can be obtained as follows:

$$\mathbf{F}_{LA} = (1 + \phi(\mathbf{X})) \cdot \mathbf{X} \tag{7}$$

where $\cdot$ denotes element-wise multiplication.

After stacking two residual attention modules, local features are extracted via GAP layer similar to Equation (1).

B. NETWORK LEARNING

During network training, the proposed global-local attention network is optimized through two types of supervision. The first supervision comes from standard classification loss. The three predicted labels are obtained based on global, local feature vectors and their concatenation in the softmax layer of this network. The overall classification loss between the predicted labels Y^* and the ground-truth label Y^S are computed as follows:

$$\mathcal{L}_{cls}(Y^S, Y^*) = -\frac{1}{N}\sum_{i=1}^{N}\langle y_i, \log y_i^S\rangle \tag{8}$$

where $Y^* \in \{Y^{(g)}, Y^{(l)}, Y^{(com)}\}$, $Y^{(g)}$, $Y^{(l)}$ and $Y^{(com)}$ denote the predicted labels from global features, local features, and their concatenation features, respectively. $\hat{y}_i \in \{Y^*\}$ is the predicted label for each image, and if $\hat{y}_i = y_i$ (y_i belongs to the ground-truth label), the corresponding label is 1 and the rest elements are 0. $\langle\rangle$ denotes the inner product and N is the number of labeled training samples.

Furthermore, in order to decouple supervision between the concatenated feature-related classification and the global or local classification, the second supervision called rank loss is further defined as follows:

$$\mathcal{L}_{rank}(p^{\triangle}, p^{com}) = \max\{0, p^{\triangle} - p^{com} + \gamma\} \tag{9}$$

where $p^{\triangle} \in \{p^{(g)}, p^{(l)}\}$ represents the predicted probability for the global and local attention branch, and γ is a margin parameter. This rank loss ensures that $p^{com} > p^{\triangle} + \gamma$. Thus, it means that the global attention branch and local attention branch focus on the respective attention regions. Therefore, the overall loss function is defined as follows:

$$\mathcal{L}(X) = \sum \mathcal{L}_{cls}(Y^S, Y^*) + \sum \mathcal{L}_{rank}(p^{\triangle}, p^{com}) \tag{10}$$

Through Eq.10, the gradient of the loss can be calculated. After that, the back propagation is utilized to update the parameters of each layer in the proposed global-local attention network.

III. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we adopt the proposed deep feature learning method GLANet for aerial scene classification. In order to facilitate this research, three challenging remote sensing scene datasets (e.g. AID [39], NWPU-RESISC45 [40], and PatternNet [41]) are utilized instead of the relatively small UC-Merced21 [23] and WHU-RS19 [46], whose classification performances are already saturated. These challenging datasets have diversity scenes, ranging from 30 to 45 classes, and each scene is with 300–800 examples. The classification performance of our GLANet is compared with that of the state-of-the-art methods with detailed experimental setup and reasonable analysis.

A. EXPERIMENTAL SETUP

In the global attention branch, two FC layers contain 64 and 512 nodes, respectively. Therefore, the size of the proposed global-local attention network is much smaller than the original VGGNet (56.73MB versus 528MB). The kernel size of the convolution layer in the residual attention scheme is set to 1 × 1. Stochastic Gradient Descent (SGD) with a weight decay of 0.0005 is applied to train the entire network (including both the backbone network and the newly added layers). Data augmentation is performed so that images of random size can be cropped to 224 × 224 pixels. The batch size is 32 for all three datasets, and the learning rate is 0.005. Different training-testing ratios are used to make a comprehensive comparison for each dataset, following the work of [39], [40], [47].

The proposed global-local attention network is compared with state-of-the-art classification methods for aerial scenes. Among these, CNN methods are most competitive, using transferred CNN features or Fine-tuned CNN features. For transferred CNN features, the CNN model pre-trained on ImageNet [42] is utilized as a feature extractor, and the second FC layer features are fed into a linear SVM for classification, such as 'CaffeNet', 'GoogleNet', 'VGGNet' in

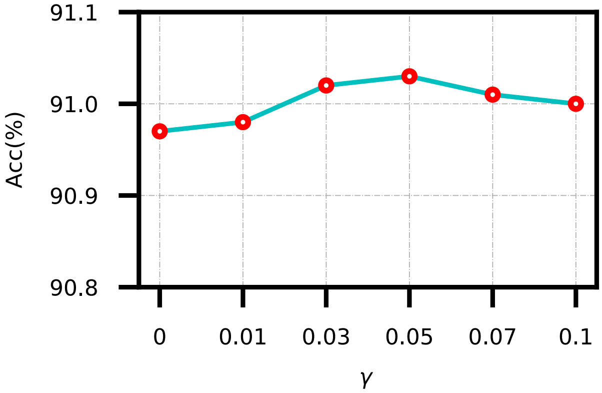


FIGURE 5. Classification accuracy (%) on NWPU-RESISC45 under 10% training ration with different γ.

this paper. For Fine-tuned CNN features, three different Fine-tuning strategies are selected for performance comparison. The first strategy involves Fine-tuned CNN model directly without changing the architecture (e.g. Fine-tuned VGGNet). In the second strategy, the FC layers are replaced by a single branch (GA or LA) to capture global context information or local semantic information, called GANet or LANet. Similarly, the third strategy combines GA branch with LA branch to simultaneously extract the global and local information, termed 'GLANet'. Finally, we also extract features using the proposed GLANet and trained an SVM classifier for classification, termed 'GLANet (SVM)'. It is worth mentioning that the linear SVM (Liblinear [48]) is utilized for supervised classification in this paper because it can quickly train a linear classifier on large-scale scene datasets from high resolution remote sensing images. The evaluation experiments are repeated ten times for a convincing performance comparison on an NVIDIA GTX TITAN GPU, whose mean and standard deviation are reported as the final results.

There exists a basic parameter of margin γ for the proposed GLANet. The margin γ influences the global attention branch and local attention branch focus on the respective attention regions. The effectiveness of parameter γ is further studied. We set the value of margin γ varies over the range of [0, 0.01, 0.03, 0.05, 0.07, 0.1] and run the experiment on NWPU-RESISC45. As shown in Fig.5, the accuracy is firstly improved and then decreased with the increasing of γ. But the variation is only very marginally. Specially, the classification accuracies of our GLANet are 90.97%, 90.97%, 91.02%, 91.03%, 91.01%, 91.00%, respectively. There exists a very small accuracy gap no more than 0.06%. From the above analysis, we can find that the accuracy for GLANet is not sensitive to the margin γ. Thus we set the value of γ to 0.05.

B. EXPERIMENT 1: AID

AID is collected from google earth imagery and contains 10, 000 remote sensing images of 600 × 600 pixels with 30 classes. Fig.6 shows representative images of each class: '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'. The image numbers of different aerial scenes vary from 200 to 400. In addition, the images in AID are from different sensors under different imaging conditions, which brings higher intra-class variations and smaller inter-class dissimilarity. For more information, see [39] and visit http://www.lmars.whu.edu.cn/xia/AIDproject.html.

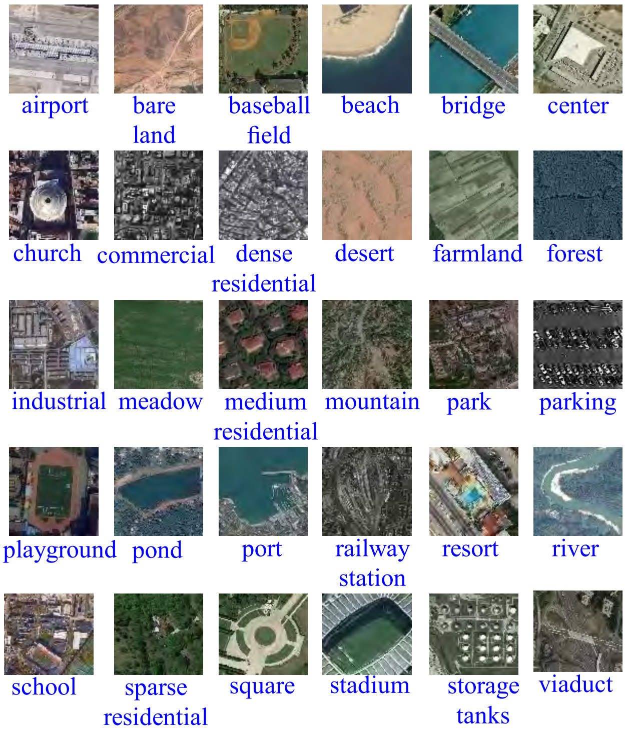


FIGURE 6. Example images from AID.

AID adopts 20%–80% and 50%–50%, where the dataset is randomly split into 20% (or 50%) for training and the rest 80% (or 50%) for testing. Classification results of our GLANet and previous representative methods are listed in Table 1. SIFT, BoWW (SIFT), CaffeNet, GoogleLeNet, VGGNet and Fine-tuned VGGNet are single-feature representations while salM^{3}LBP, DCNNs, TEX-Net-LF, Fusion by addition, Fusion by concatenation, DCA with concatenation, Two-Stream Fusion, Two-Stage Fusion, CNNs, Fine-tuned CNNs(LR), Fine-tuned CNNs (SVM) and the proposed GLANet are collaborative feature representations. Different from other collaborative feature representations based on CNN features, salM^{3}LBP-CLM combines global saliency-based multiscale multiresolution multistructure LBP (salM^{3}LBP) and local codebookless model (CLM). Among single-feature representations, the classification results of GoogleLeNet, VGGNet, and Fine-tuned VGGNet are greatly better than SIFT and BoWW (SIFT)

TABLE 1. Overall classification accuracy (%) comparison on AID. The highest accuracy appears in boldfaced.

Method	Accuracy	
	20%	50 %
SIFT [39]	13.50±0.67	16.76±0.65
BoWW(SIFT) [39]	61.40±0.41	67.65±0.49
CaffeNet [39]	86.86±0.47	89.53±0.31
GoogleLeNet [39]	83.44±0.40	86.39±0.55
VGGNet [39]	86.59±0.29	89.64±0.36
Fine-tuned VGGNet	88.27±0.29	92.49±0.30
salM^{3}LBP-CLM [14]	86.92±0.35	89.76±0.45
DCNNs [49]	90.82±0.16	**96.89±0.10**
TEX-Net-LF [50]	90.87±0.11	92.96±0.18
Fusion by addition [5]	–	91.87±0.36
Fusion by concatenation [5]	–	91.86±0.28
DCA with concatenation [5]	–	89.71±0.33
Two-Stream Fusion [51]	92.32 ±0.41	94.58±0.41
Two-Stage Fusion [32]	–	94.65±0.33
CNNs [11]	–	94.17±0.32
Fine-tuned CNNs(LR) [11]	–	94.93±0.28
Fine-tuned CNNs(SVM) [11]	–	95.36±0.22
GANet	92.84±0.22	96.02±0.22
LANet	93.72±0.14	96.52±0.18
GLANet + SVM	94.44±0.31	96.36± 0.23
GLANet	**95.02±0.28**	96.66±0.19

(at least 18% performance improvement), which confirms the powerful feature learning ability of the current-dominated learning methods. The experimental results of CaffeNet, GoogleLeNet and VGGNet are poorer than that of salM^{3}LBP, which indicates that handcrafted features are effectively fused may sometimes be more competitive than CNN feature based methods for a certain high resolution image dataset. As can be seen in Table 1, the accuracy can be further boosted by at least 2.2% by fusing CNN features, which indicates that different CNN features are also complementary for aerial scene classification. Under the training ratio of 50%, our GLANet (96.66%) and DCCNs (96.89%) obtain comparative results, which are better than that of other comparison methods. However, when the training ratio is 20%, the classification result of DCCNs decreases sharply and is only higher than that of handcrafted features based salM^{3}LBP while our GLANet (95.02%) is the best among these collaborative feature representations based methods. In addition, we have conducted experiments under other training-testing ratios, like 10%–90% and 30%–70%. Specifically, the GLANet achieves the accuracy of 92.49% and 95.64% respectively. The DCNNs gets the accuracy of 88.78% and 92.28%. It indicates GLANet can obtain gains of 3.71% and 3.36% under the training ratios of 10% and 30%.

Besides the above-mentioned comparison results, table 1 also represents the classification performance of the newly proposed GLANet and its variants, e.g. GANet, LANet and GLANet (SVM), to analyze the effectiveness of global context, local semantic and their collaborative representation. Compared with the baseline Fine-tuned VGGNet, GANet boosts the classification accuracy (6.25% and 3.6% improvement under the training ratios of 20% and 50%, respectively), which indicates that global average pooling strategy is more applicable than VGG16 to extract global features from AID. Similarly, the proposed LANet is also superior to the baseline method, obtaining 7.13% improvement under 20% training ratio and 4.03% improvement under 50% training ratio. It benefits from considering local semantic information in complex scenes of AID. As shown in table 1, our GLANet can further improve the classification accuracies (95.02% and 96.66% under the training ratios of 20% and 50%, respectively), which confirms that it is necessary to combine global context and local semantic information for collaborative representation. Additionally, it can be seen that GLANet using softmax is slightly superior to that of using SVM when the training ratio is small (0.2).

Fig.7 presents the confusion matrix for GLANet on AID under 50% training ratio. From this confusion matrix, we can see that most scene categories (90%) can achieve the classification accuracy more than 0.9 using our method and thus easy to be categorized. Among these 30 scene categories, resort (0.84) obtain the poorest performance around 0.85, which no longer belongs to the difficult scene to be categorized. Compared with the previous confusion matrix in [39], school (0.67 versus 0.93), square (0.67 versus 0.88), resort (0.7 versus 0.84), and center (0.7 versus 0.86) can be improved to a great extent, exceeding 14%. As shown in Fig.7, the most notable confusion occurs between resort and park. Specifically, 6% of images from 'resort' are mistakenly classified as 'park' while 3% of images from 'park' are mistakenly classified as 'resort'. It can attribute to the similar appearances and the common objects shared by resort and park scenes such as green belts. DCNNs introduces a discriminative loss term into pre-trained CNN models and then Fine-tune the whole CNN model to match AID, achieving slightly better classification results than our GLANet under 50% training ratio. We further compared GLANet with DCNNs in terms of the confusion matrix. It can be observed that our GLANet can discriminate most classes accurately. For GLANet and DCNNs, the accuracies for center, forest and meadow differ to a great extent. In details, DCNNs achieves better results for center with the improvement of 14% while our GLANet achieved better results for forest and meadow with the improvement of 8% and 6%, respectively. The overall results prove the superiority of our GLANet. It employs GA and LA simultaneously and is more suitable for aerial scene classification on AID.

C. EXPERIMENT 2: NWPU-RESISC45

NWPU-RESISC45 is created by Northwestern Polytechnical University from Google Earth (Google Inc.), which is available online from http://www.escience.cn/people/JunweiHan/NWPU-RESISC45. This dataset consists of 31, 500 remote sensing images, covering more than one hundred countries and regions all over the world, including developing, transitional, and highly developed economies [40]. These images can be categorized into 45 classes and Fig.8 shows representative images of each class: 'airplane', 'airport', 'baseball diamond', 'basketball court', 'beach', 'bridge',

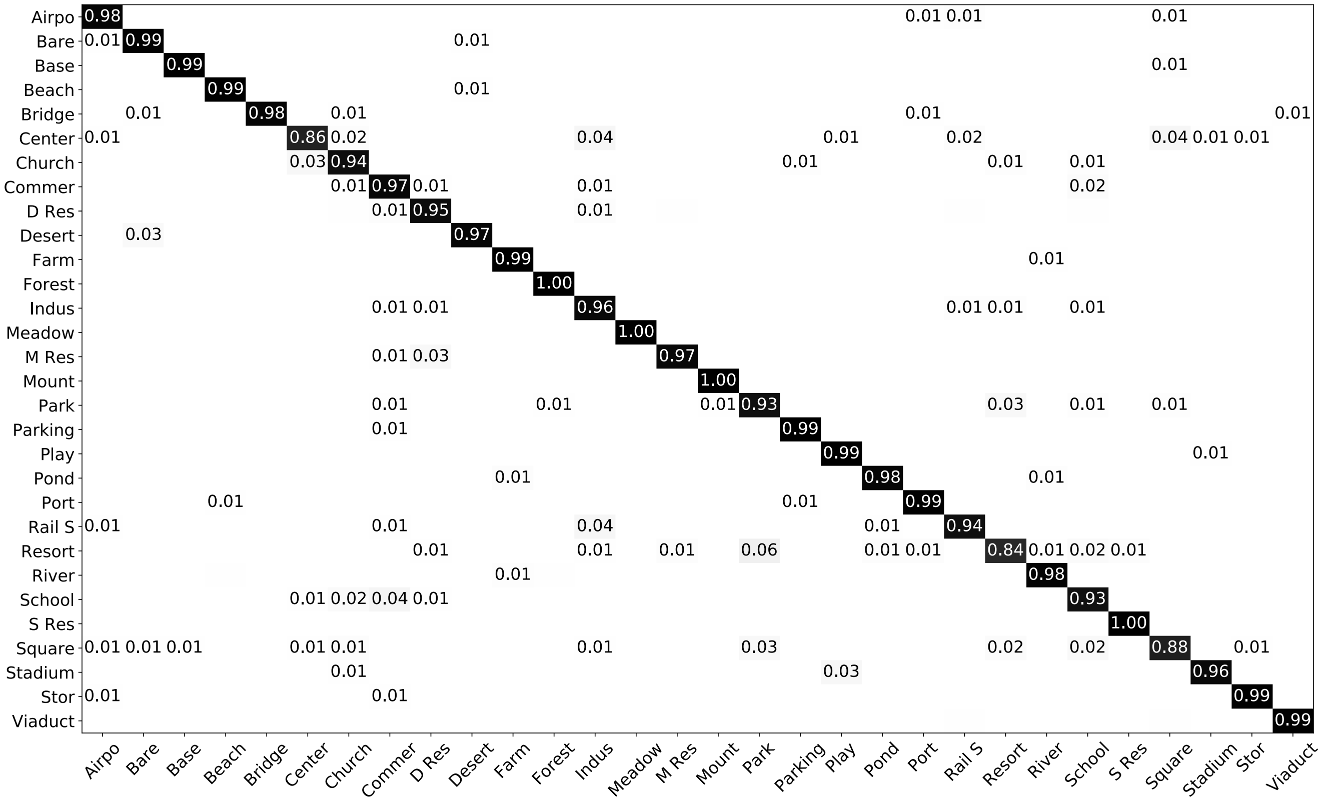


FIGURE 7. Confusion matrix for the results on AID under 50% training ratio using GLANet.

'chaparral', 'church', 'circular farmland', 'cloud', 'commercial', 'dense residential', 'desert', 'forest', 'freeway', 'golf course', 'ground track field', 'harbor', 'industrial', '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 contains 700 images of 256×256 pixels. Except for the island, lake, mountain, and snowberg, most of the scene classes have spatial resolutions vary from about 30 to 0.2 m. To the best of our knowledge, NWPU-RESISC45 is the most challenging large-scale scene dataset from high resolution remote sensing images.

NWPU-RESISC45 is used according to 10%–90% and 20%–80%. The classification performance of GLANet and several representative experimental results of state-of-the-art approaches for NWPU-RESISC45 are listed in Table 2. They use various features including low-level feature (LBP), mid-level feature (BoVW-dense SIFT) and CNN features. Among these CNN features, Alex, GoogleLeNet, VGGNet and Fine-tuned VGGNet are the off-the-shelf deep learning features; BoCF (Alex), BoCF (GoogleLeNet) and BoCF (VGGNet) are mid-level CNN features; LASC-CNN (single-scale), LASC-CNN (multiscale), TEX-TS-Net, SAL-TS-Net and DCNNs are multiple CNN features, which have fused multiscale CNN features, different layers and even CNN features of different models to represent the images. As can be seen in Table 2, CNN features based methods outperform low-level and mid-level features based methods

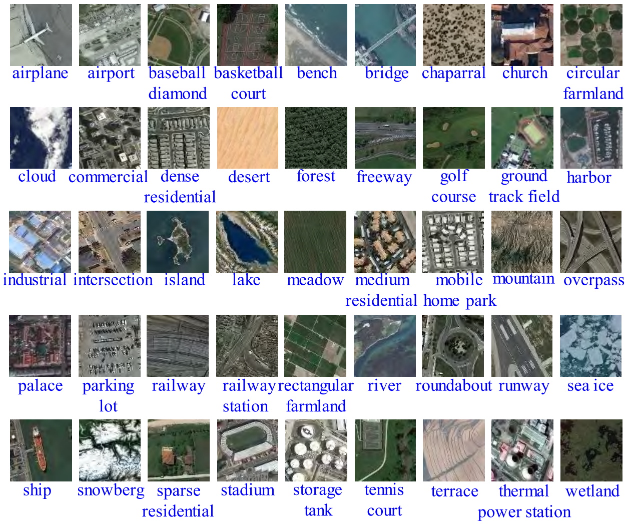


FIGURE 8. Example images from NWPU-RESISC45.

TABLE 2. Overall classification accuracy (%) comparison on NWPU-RESISC45. The highest accuracy appears in boldfaced.

Method	Accuracy	
	10%	20 %
LBP [40]	19.20±0.41	21.74± 0.18
BoVW(dense SIFT) [40]	41.72±0.21	44.97± 0.28
Alex [40]	76.69±0.21	79.85±0.13
GoogleLeNet [40]	76.19±0.38	78.48±0.26
VGGNet [40]	76.47±0.18	79.79±0.15
Fine-tuned VGGNet [40]	86.75±0.27	89.43±0.31
BoCF(Alex) [52]	55.22±0.39	59.22±0.18
BoCF(GoogleLeNet) [52]	78.92±0.17	80.97±0.17
BoCF(VGGNet) [52]	82.65±0.31	84.32±0.17
LASC-CNN(single-scale) [53]	80.69	83.64
LASC-CNN(multiscale) [53]	81.37	84.30
TEX-TS-Net [54]	84.77±0.24	86.36±0.19
SAL-TS-Net [54]	85.02±0.25	87.01±0.19
DCNNs [49]	89.22±0.50	91.89±0.22
GANet	87.96±0.23	91.36±0.18
LANet	89.41±0.26	92.35±0.24
GLANet (SVM)	89.52±0.23	92.35±0.19
GLANet	**91.03±0.18**	**93.45±0.17**

in very big margins under the training ratios of 10% and 20%, which demonstrates that the huge superiority of CNNs in capturing the discriminative representations for remote sensing scenes from NWPU-RESISC45. Among these CNN features based comparison methods, the pre-trained CNNs have the lowest accuracy except for BoCF (Alex), which indicates that Fine-tuned, mid-level representation and feature combination are commonly used strategies to further boost the discriminative ability of CNN feature representations. However, our GLANet, which is based on an elegant architecture, performs much better (91.03% and 93.45% under the training ratios of 10% and 20%, respectively) than all the comparison methods. It indicates that the strong discriminative power of our GLANet compared with the previous state-of-the-art methods, providing a more semantic and robust representation for aerial scenes from NWPU-RESISC45.

In addition, the experimental results of the GANet, LANet and GLANet(SVM) are also displayed in table 2 to evaluate the effectiveness of different modules for GLANet. Compared with Fine-tuned VGGNet, GANet obtains 1.21% and 1.93% improvement while LANet obtains 2.66% and 2.93% improvement under the training ratios of 10% and 20%, respectively. It benefits from the attention mechanism. From table 2, it can be seen that the GLANet using softmax provides better classification performances than using SVM, 91.03% versus 89.52% under 10% training ration and 93.45% versus 92.35% under 20% training ration. Last but not least, the newly proposed GLANet can achieve the best performance (91.03% and 93.45% under the training ratios of 10% and 20%, respectively) compared with only using one branch (GANet or LANet) or Fine-tuned VGGNet, which is a result of combining both GA and LA branches. In short, each module in the GLANet does indeed improve the performance of scene classification on NWPU-RESISC45.

The result in terms of confusion matrix obtained by GLANet for NWPU-RESISC45 under 20% training ratio is shown in Fig.9. From this confusion matrix, it can be seen that 96% of the 45 categories can achieve the classification accuracy of over 86%, which again indicates that GLANet is reasonable. As illustrated in Fig.8, the scenes of dense residential, medium residential and residential have similar spatial distribution and identical objects(e.g. building, tree and road). Fig.9 shows that the proposed GLANet could also accurately classify these small inter-class dissimilarity scenes (0.89, 0.91 and 0.92 for dense residential, medium residential and sparse residential, respectively), reducing misclassification for NWPU-RESISC45. Nevertheless, both church (0.79) and palace (0.74) are still difficult to recognize. Even so, our GLANet can achieve a substantial improvement with the accuracies (0.58 and 0.52) of the same scenes from the confusion matrix of [40], which directly used the VGGNet. This result may be explained by the fact that the integration of GA and LA helps in learning discriminative features. On the whole, GLANet can achieve accurate scene reasoning for NWPU-RESISC45.

FIGURE 9. Confusion matrix for the results on NWPU-RESISC45 under 20% training ratio using GLANet.

D. EXPERIMENT 3: PATTERNNET

The images in PatternNet [41] are also collected by Wuhan University from Google Earth imagery or via the Google Map API for US cities, which is available online from https://sites.google.com/view/zhouwx/dataset. It contains 38 classes with 800 images for each class and each image has a fixed size of 256 × 256 pixels. Fig.10 shows representative images of each class: 'airplane', 'baseball field', 'basketball court', 'beach', 'bridge', 'cemetery', 'chaparral', 'christmas tree farm', 'closed road', 'coastal mansion', 'crosswalk', 'dense residential', 'ferry terminal', 'football field', 'forest', 'freeway', 'golf course', 'harbor',

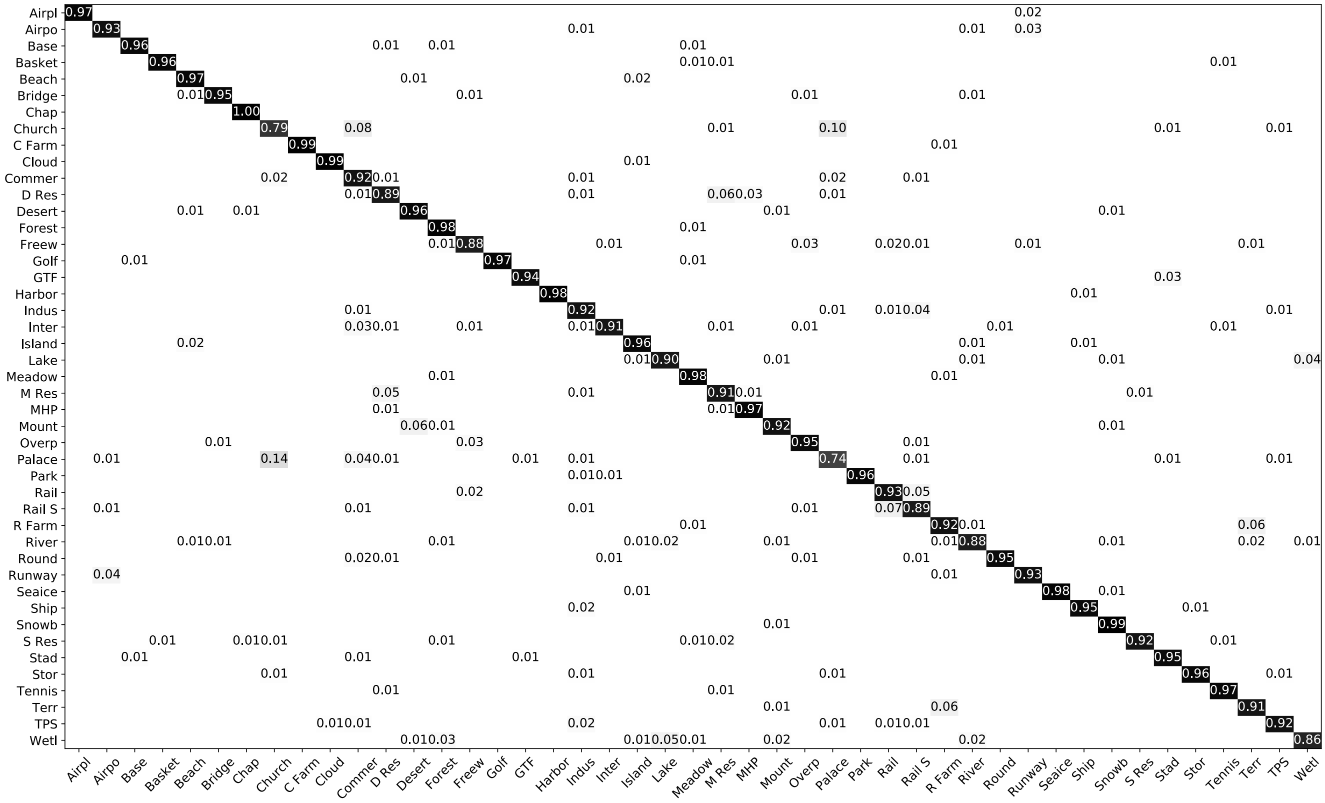


FIGURE 10. Example images from PatternNet.

'intersection', 'mobile home park', 'nursing home', 'oil gas field', 'oil well', 'overpass', 'parking lot', 'parking space', 'railway, river', 'runway', 'runway marking', 'shipping yard', 'solar panel', 'sparse residential', 'storage tank', 'swimming pool', 'tennis court', 'transformer station' and 'wastewater treatment plant'. This dataset contains images with the varying resolution, ranging from 0. 062m to 4.693m. Generally, it is hard to compare the same methods on all three datasets as lots of current methods only report their performance on a specific dataset and do not release their codes. We clarify that PatternNet is a recently released dataset, and we compare our GLANet with all available results in the existing literature [47], [55].

Experiments on PatternNet are conducted by 20%–80%, 50%–50%, and 80%-20%. The classification performances of VGGNet, Fine-tuned VGGNet, Product Rule Combination (PRC) [55], Enhance Fusion Network (EFNet) [47], and the newly proposed GLANet with its variants, e. g. GANet, LANet, and GLANet (SVM), are reported in Table 3. In general, PatternNet is quite easy to classify for these deep learning architectures with high performances, exceeding 94% average accuracy. It is because PatternNet is more homogeneous within class and distinct between classes [47], compared with AID and NWPU-RESISC45. Table 3 shows that Fine-tuned VGGNet has the lowest accuracy, even lower than VGGNet. This result demonstrates that an excessive number of parameters in the FC layers may be harmful to network learning. As shown in Table 3, Product Rule Combination (99.52% under 0.5 training ratio) is superior to all single feature-based methods, which indicates that feature combination is a good strategy to further boost classification performance, but still inferior to our GLANet(SVM) and GLANet. In addition, the proposed single-feature methods replace the FC layers with the global or local attention branch and boost the classification accuracy drastically: (1) 3.39% under 0.2 training ratio and 1.23% under 0.5 training ratio for GANet; (2) 4.53% under 0.2 training ratio and 1.27% under 0.5 training ratio for LANet. Combining GA branch and LA branch simultaneously, the final proposed GLANet achieves the best classification performances (99.46% and 99.65% under 0.2 and 0.5 training ratios, respectively), which illustrates the importance of feature complementation between global context and local semantic information. (3) Using more training samples, our GLANet and its variants can obtain competitive results, with a small accuracy gap no more than 0.09%. It is noted that Enhance Fusion Network (EFNet) is the recently published method [47], using four DCNNs (e.g, CaffeNet, GoogLeNet, ResNet-50, and ResNet-101) and multiple network fusion techniques. It can obtain the same classification result (99.70%) as our GLANet under 80% training ratio, which is only very marginally better than our GLANet (99.65%) with fewer training samples used (50%). It is obvious the classification accuracy on PatternNet tends to saturation.

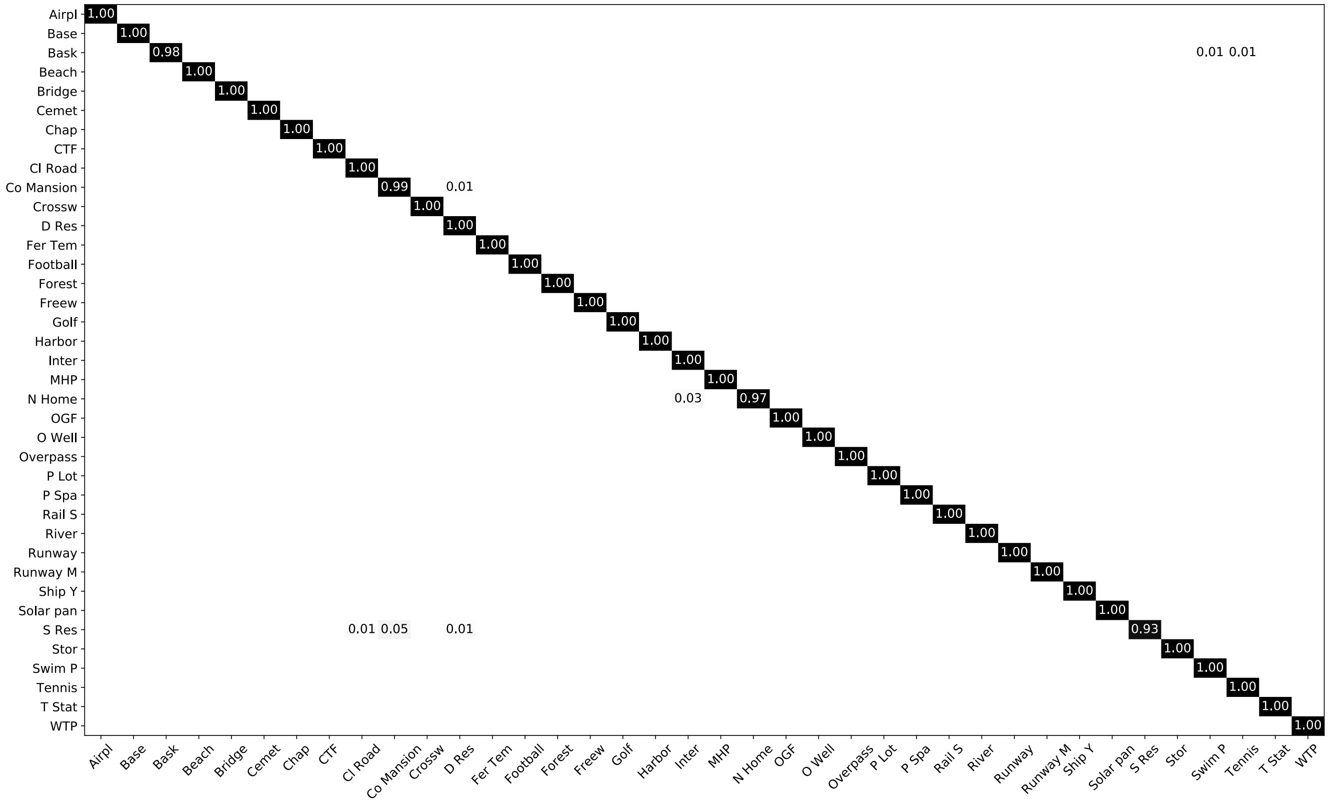


FIGURE 11. Confusion matrix for the results on PatternNet under 50% training ratio using GLANet.

TABLE 3. Overall classification accuracy (%) comparison on PatternNet. The highest accuracy appears in boldfaced.

Method	Accuracy		
	20%	50 %	80 %
VGGNet	95.07±0.20	98.46±0.13	–
Fine-VGGNet	94.11±0.20	97.55±0.21	–
PRC [55]	–	99.52±0.17	–
EFNet [47]	–	–	**99.70**
GANet	97.50±0.26	98.78±0.28	**99.70±0.26**
LANet	98.64±0.23	98.82±0.24	99.61 ±0.23
GLANet (SVM)	98.91±0.19	99.40±0.21	**99.70±0.20**
GLANet	**99.46±0.13**	**99.65±0.11**	**99.70±0.15**

For PatternNet, an overview of the performance of GLANet is further presented in the confusion matrix under 50% training ratio in Fig.11. Most of the scene categories (34 of 38) can be fully recognized by GLANet, as shown in Fig.11. Among these scenes, the 'sparse residential' is most confused using GLANet with the lowest accuracy of 93%, which indicates that it still can be classified satisfactorily. This demonstrates that the GLANet can obtain discriminative feature representation and depicted remote sensing scenes from PatternNet very well.

IV. CONCLUSION

In this paper, we propose an end-to-end global-local attention network for aerial scene classification. The GLANet incorporates both global and local information to generate an effective collaborative representation. The squeeze-excitation attention mechanism and spatial residual attention are utilized to learn global context information and local semantic information, respectively. In addition to the standard cross-entropy loss, a rank loss function is introduced to further enhance the discriminative ability of the GLANet. On the whole, our GLANet can well explore the complementary attributes globally and locally, and provides a comprehensive description for aerial scenes, leading to substantially improve the discriminative ability of collaborative feature representation. The experimental results on three challenging large-scale scene datasets (e.g. AID, NWPU-RESISC45, and PatternNet) demonstrate that the newly proposed global-local attention network can achieve competitive performance compared with the current state-of-the-art approaches.

Learning an effective collaborative representation with global and local deep features is strongly appealing, and this paper provides preliminary results for further research. In our future work, further experiments will be conducted on multi-temporal high resolution remote sensing images, hyperspectral images, even other non-optical images of remote sensing, such as polarimetric synthetic aperture radar(PolSAR). On the other hand, we will extend our GLANet in other network architectures apart from VGGNet, and develop an improved global-local attention network, which fuses multiple GLANets of different backbones.

REFERENCES

[1] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using via VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[2] Q. Hu, W. Wu, T. Xia, Q. Yu, P. Yang, Z. Li, and Q. Song, "Exploring the use of google earth imagery and objectbased methods in land use/cover mapping," *Remote Sens.*, vol. 5, no. 11, pp. 6026–6042, Nov. 2013.

[3] H. Pu, Z. Chen, B. Wang, and G. Jiang, "A novel spatial-spectral similarity measure for dimensionality reduction and classification of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7008–7022, Nov. 2014.

[4] Z. Chen, H. Pu, B. Wang, and G. Jiang, "Fusion of hyperspectral and multispectral images: A novel framework based on generalization of pan-sharpening methods," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 8, pp. 1418–1422, Aug. 2014.

[5] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.

[6] O. A. B. Penatti, K. Nogueira, and J. A. D. Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 44–51.

[7] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, "Exploiting convolutional neural networks with deeply local description for remote sensing image classification," *IEEE Access*, vol. 6, pp. 11215–11228, Mar. 2018.

[8] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.

[9] Z. Lv, X. Li, B. Zhang, W. Wang, Y. Zhu, and J. Hu, and S. Feng, "Managing big city information based on WebVRGIS," *IEEE Access*, vol. 4, pp. 407–415, Jan. 2018.

[10] Z. Lv, T. Yin, X. Zhang, H. Song, and G. Chen, "Virtual reality smart city based on WebVRGIS," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1015–1024, Dec. 2016.

[11] Y. Yu and F. Liu, "Aerial scene classification via multilevel fusion based on deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 287–291, Feb. 2018.

[12] W. Yang, X. Yin, and G.-S. Xia, "Learning high-level features for satellite image classification with limited labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4472–4482, Aug. 2015.

[13] S. Ozkan, T. Ates, E. Tola, M. Soysal, and E. Esen, "Performance analysis of state-of-the-art representation methods for geographical image retrieval and categorization," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 11, pp. 1996–2000, Nov. 2017.

[14] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2889–2901, Jun. 2017.

[15] A. Qayyum, A. S. Malik, N. M. Saad, M. Iqbal, M. F. Abdullah, W. Rasheed, T. AB R. Abdullah, and M. Y. B. Jafaar, "Scene classification for aerial images based on CNN using sparse coding technique," *Int. J. Remote Sens.*, vol. 38, nos. 8–10, pp. 2662–2685, 2017.

[16] F. Hu, G.-S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, May 2015.

[17] Y. Zhang, X. S. Wei, J. Wu, J. Cai, J. Lu, V. A. Nguyen, and M. N. Do, "Weakly supervised fine-grained categorization with part-based image representation," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1713–1725, Apr. 2016.

[18] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[19] Y. Yang and S. Newsam, "Comparing sift descriptors and Gabor texture features for classification of remote sensed imagery," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 1852–1855.

[20] B. Luo, S. Jiang, and L. Zhang, "Indexing of remote sensing images with different resolutions by multiple features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 4, pp. 1899–1912, Aug. 2013.

[21] P. Tokarczyk, J. D. Wegner, S. Walk, and K. Schindler, "Features, color spaces, and boosting: New insights on semantic classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 280–295, Jun. 2014.

[22] H. Sun, S. Liu, and S. Zhou, "Combining low level features and visual attributes for VHR remote sensing image classification," in *Proc. Int. Symp. Multispectral Image Process. Pattern Recognit.*, 2015, pp. 1–8.

[23] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, 2010, pp. 270–279.

[24] L. Zhao, P. Tang, and L. Huo, "A 2-D wavelet decomposition-based bag-of-visual-words model for land-use scene classification," *Int. J. Remote Sens.*, vol. 35, no. 6, pp. 2296–2310, Mar. 2015.

[25] H. Sridharan and A. Cheriyadat, "Bag of lines BoL for improved aerial scene representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 676–680, Mar. 2015.

[26] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.

[27] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 28–32, Jan. 2010.

[28] C. Vaduva, I. Gavat, and M. Datcu, "Latent Dirichlet allocation for spatial analysis of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2770–2786, May 2013.

[29] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, Nov. 2015.

[30] J. Shi, X. Tian, Z. Jiang, D. Zhao, and M. Liu, "Sparsity-constrained probabilistic latent semantic analysis for land cover classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2016, pp. 5453–5456.

[31] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Oct. 2015.

[32] Y. Liu, Y. Liu, and L. Ding, "Scene classification based on two-stage deep feature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 183–186, Feb. 2018.

[33] Q. Zhu, L. Zhang, Y. Liu, Y. Zhong, and D. Li, "A deep-local-global feature fusion framework for high spatial resolution imagery," *Remote Sens.*, vol. 10, no. 4, pp. 1–22, Apr. 2018.

[34] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Rev. Neurosci.*, vol. 3, no. 3, pp. 201–215, 2002.

[35] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. doi: 10.1109/TPAMI.2019.2913372.

[36] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6450–6458.

[37] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6298–6306.

[38] G. G. Abel, D. Modolo, and V. Ferrari, "Do semantic parts emerge in convolutional neural networks?" *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 476–494, May 2018.

[39] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

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

[41] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatterNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogram. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.

[42] J. Deng, W. Wang, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 6450–6458.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[44] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400

[45] B. Zhou, A. Khosla, A. Laperdriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2921–2929.

[46] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, 2012.

[47] G. J. Scott, K. C. Hagan, R. A. Marcum, J. A. Hurt, D. T. Anderson, and C. H. Davis, "Enhanced fusion of deep neural networks for classification of benchmark high-resolution image data sets," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1451–1455, Sep. 2018.

[48] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[49] 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 Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 1865–1883, May 2018.

[50] 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 J. Photogramm. Remote Sens.*, vol. 138, pp. 74–85, Apr. 2018.

[51] Y. Yu and F. Liu, "A two-stream deep fusion framework for high-resolution aerial scene classification," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, Jan. 2018.

[52] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 183–186, Aug. 2017.

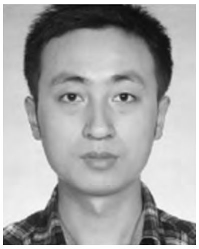
[53] B. Yuan, S. Li, and N. Li, "Multiscale deep features learning for land-use scene recognition," *J. Appl. Remote Sens.*, vol. 12, no. 1, pp. 1–25, Feb. 2018.

[54] Y. Yun and F. Liu, "Dense connectivity based two-stream deep feature fusion framework for aerial scene classification," *Remote Sens.*, vol. 10, no. 7, pp. 1–12, Jun. 2018.

[55] M. Altaei, S. Ahmed, and H. Ayad, "Effect of texture feature combination on satellite image classification," *Int. J. Adv. Res. Comput. Sci.*, vol. 9, no. 2, pp. 675–683, Mar. 2018.

YIYOU GUO received the B.S. degree in geomatics engineering and the M.S. degree in GIS from Wuhan University, China, in 2005 and 2010, respectively. He is currently pursuing the Ph.D. degree with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. His research interests include feature selection/learning, deep learning, image processing, and object tracking.

JINSHENG JI received the B.S. degree in automation from Nanjing Agricultural University and the M.S. degree in control science and engineering from Shanghai Jiao Tong University, China, where he is currently pursuing the Ph.D. degree with the Department of Automation, China. His research interests include computer version and machine learning.



XIANKAI LU received the B.S. degree in automation from Shandong University, Jinan, China, in 2012. He is currently pursuing the Ph.D. degree with Shanghai Jiao Tong University, Shanghai, China. His research interests include image processing, object tracking, and deep learning.

HONG HUO received the B.S. degree in computer application and the M.S. degree in computer applications from the Jilin University of Technology (now merged into Jilin University), Changchun, China, in 1995 and 1998, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from Shanghai Jiao Tong University, Shanghai, China, in 2014, where she has been an Instructor with the Institute of Image Processing and Pattern Recognition, since 2000. Her research interests include image analysis and interpretation, machine learning, and visual perception with applications to remote sensing imagery.

TAO FANG received the B.S. and M.S. degrees in geology and survey from the Xian University of Science and Technology, Xi'an, China, in 1988 and 1991, respectively, and the Ph.D. degree in remote sensing and geographical information system from the China University of Mining and Technology, Beijing, China, in 1996. From 1996 to 1998, he was a Postdoctoral Research Fellow of remote sensing and geographical information systems with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China. From 1999 to 2000, he was an Assistant Professor with the Department of Electronic Engineering and a Postdoctoral Research Fellow of electronics with the HDTV Laboratory, University of Electronic Science and Technology of China, Chengdu, China, respectively. He is currently a Professor with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. His research interests include neural representation and brain-like computing, deep learning with applications to remote sensing classification, and object recognition.

DEREN LI (M'02–SM'03) received the M.Sc. degree in photogrammetry and remote sensing from the Wuhan Technical University of Surveying and Mapping, Wuhan University, Wuhan, China, in 1981, and the Dr.Eng. degree in photogrammetry and remote sensing from Stuttgart University, Stuttgart, Germany, in 1985. He was elected as an Academician of the Chinese Academy of Sciences, Beijing, China, in 1991, and the Chinese Academy of Engineering, Beijing, and the Euro–Asia Academy of Sciences, Beijing, in 1995. He is currently the Academic Committee Chairman of the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include spatial information science and technology, such as remote sensing, GPS and geographic information system (GIS), and their integration. He was the President of the International Society for Photogrammetry and Remote Sensing Commissions III and VI and the first President of the Asia GIS Association, from 2002 to 2006.