Remote Sensing Scene Classification by Gated Bidirectional Network

Hao Sun, Siyuan Li, Xiangtao Zheng, and Xiaoqiang Lu, *Senior Member, IEEE*

***Abstract*— Remote sensing (RS) scene classification is a challenging task due to various land covers contained in RS scenes. Recent RS classification methods demonstrate that aggregating the multilayer convolutional features, which are extracted from different hierarchical layers of a convolutional neural network, can effectively improve classification accuracy. However, these methods treat the multilayer convolutional features as equally important and ignore the hierarchical structure of multilayer convolutional features. Multilayer convolutional features not only provide complementary information for classification but also bring some interference information (e.g., redundancy and mutual exclusion). In this paper, a gated bidirectional network is proposed to integrate the hierarchical feature aggregation and the interference information elimination into an end-to-end network. First, the performance of each convolutional feature is quantitatively analyzed and a superior combination of convolutional features is selected. Then, a bidirectional connection is proposed to hierarchically aggregate multilayer convolutional features. Both the top–down direction and the bottom–up direction are considered to aggregate multilayer convolutional features into the semantic-assist feature and appearance-assist feature, respectively, and a gated function is utilized to eliminate interference information in the bidirectional connection. Finally, the semantic-assist feature and appearance-assist feature are merged for classification. The proposed method can compete with the state-of-the-art methods on four RS scene classification data sets (AID, UC-Merced, WHU-RS19, and OPTIMAL-31).**

***Index Terms*— Feature aggregation, remote sensing (RS) image, scene classification.**

Manuscript received January 28, 2019; revised June 2, 2019 and July 7, 2019; accepted July 25, 2019. This work was supported in part by the National Key R&D Program of China under Grant 2017YFB0502900, in part by the National Natural Science Foundation of China under Grant 61806193 and Grant 61772510, in part by the Young Top-Notch Talent Program of Chinese Academy of Sciences under Grant QYZDB-SSW-JSC015, in part by the Open Research Fund of State Key Laboratory of Transient Optics and Photonics, Chinese Academy of Sciences, under Grant SKLST2017010, in part by the CAS "Light of West China" Program under Grant XAB2017B26, and in part by the Xi'an Postdoctoral Innovation Base Scientific Research Project. *(Corresponding author: Xiangtao Zheng.)*

H. Sun is with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China.

S. Li is with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China, also with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with Xi'an Jiaotong University, Xi'an 710049, China.

X. Zheng and X. Lu are with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xiangtaoz@gmail.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2019.2931801

I. Introduction

REMOTE sensing (RS) scene classification aims at inferring the correct category based on the content contained in the RS images. Since RS scenes contain complicated land covers, it is an intractable task to classify the RS scenes [1], [2]. RS scene classification has attracted much attention in the RS fields, such as target detection [3] and RS image retrieval [4], [5].

Recently, many methods have been proposed to classify the complex RS scenes with the convolutional neural networks (CNNs). Based on the way of using CNNs, the RS scene classification methods can be summarized into three categories: the direct CNN methods, the feature encoding methods, and the aggregation learning methods.

The direct CNN methods transfer the existing CNNs (e.g., AlexNet [6] and GoogLeNet [7]) to the RS scene classification task. As shown in Fig. 1(a), these works [8] directly employ a fine-tuned CNN to classify the RS scenes. However, the direct CNN methods only utilize the feature from the last layer of CNN to classify RS scenes. The features from different hierarchical layers of CNN, which are named multilayer convolutional features, are ignored in the direct CNN methods.

The feature encoding methods refer to first utilizing a pre-trained CNN as a feature extractor and then encoding features with traditional unsupervised feature encoding methods. Many works [9], [10] have demonstrated that the top convolutional layers can effectively capture semantic features and the bottom convolutional layers can extract appearance features. As shown in Fig. 1(b), multilayer convolutional features are extracted from the different hierarchical layers of CNN. Then, some unsupervised feature encoding methods are employed to encode convolutional features into an RS scene representation. However, the unsupervised feature encoding methods cannot exploit the label information to generate a discriminative representation for RS scene classification [11].

The aggregation learning methods refer to designing a new feature aggregation network. The feature aggregation is embedded in the convolutional network for the end-to-end learning. In the end-to-end network, the label information can be exploited to facilitate the feature learning and feature aggregation. As shown in Fig. 1(c), several works [12], [13] attempt to build an end-to-end network to aggregate multilayer convolutional features into a scene representation. It is demonstrated that aggregating multilayer convolutional features can achieve the superior performance in the RS scene classification [14]. Although multilayer convolutional features can provide complementary information [15], they

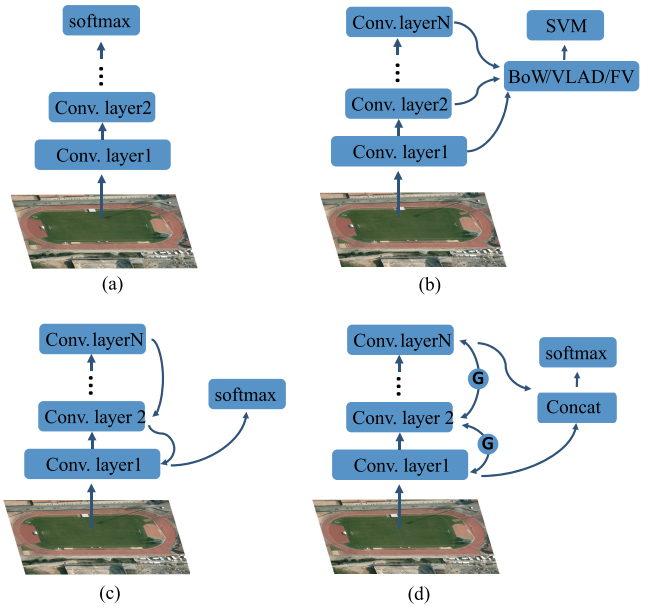

Fig. 1. Structures of recent RS scene classification methods. (a) Directly exploiting a fine-tuned CNN. (b) Employing the pre-trained CNN as a feature extractor and then exploiting traditional feature encoding methods to encode features. (c) Exploring new end-to-end feature aggregation convolutional networks for classification. (d) Proposed GBNet aggregates the different features at both the top–down direction and the bottom–up direction.

may also bring some interference information, such as feature redundancy and mutual exclusion among the convolutional features. The aggregation learning methods [16], [17] treat the multilayer convolutional features as equally important and ignore the hierarchical structure of multilayer convolutional features. A discriminative RS scene classification method should consider the hierarchical structure among different convolutional features [18].

In this paper, an end-to-end gated bidirectional network (GBNet) is proposed to take the hierarchical feature aggregation and the interference information elimination in a unified convolutional network. For the hierarchical feature aggregation, a bidirectional connection is proposed to hierarchically aggregate multilayer convolutional features layer-by-layer. For the interference information elimination, a gated function is exploited to assign a weight vector for each convolutional feature to determine whether the convolutional feature is passed forward or not. The glancing architecture of GBNet is shown in Fig. 1(d). The proposed method integrates the bidirectional connection and the gated function into a unified gated bidirectional connection. The gated bidirectional connection is composed of the top–down direction and the bottom–up direction. In the top–down direction, the semantic features from top layers are hierarchically fed into the bottom convolutional layers with the gated function to generate the semantic-assist features. In the bottom–up direction, the appearance features from bottom layers are hierarchically integrated into the top convolutional layers with the gated function to obtain the appearance-assist features. Finally, both the semantic-assist features and the appearance-assist features are integrated into a discriminative scene representation for the RS scene classification task.

The main contributions of this paper are summarized as follows.

1) To explore the hierarchical structure of multilayer convolutional features, a bidirectional connection is proposed for the hierarchical feature aggregation. The bidirectional connection takes both the top–down direction and bottom–up direction to aggregate multilayer convolutional features layer-by-layer.
2) To eliminate the interference information among multilayer convolutional features, a gated function is proposed to restrict the propagation of interference information. The gated function and bidirectional connection can be integrated into the end-to-end GBNet for joint learning.
3) To understand the performance of different convolutional features, each convolutional feature is quantitatively analyzed on the RS scene data sets. Based on the analysis, a superior combination of convolutional features is selected for aggregation.

The remaining sections are organized as follows. In Section II, the related RS scene classification works are reviewed. Next, the quantitative analysis of convolutional features is introduced in Section III. Then, the proposed GBNet is introduced in Section IV. The evaluations of GBNet are shown in Section V. Finally, the conclusion is presented in Section VI.

II. Related Work

Due to the extremely complex composition of RS scenes, the RS scene classification is still a challenging task. In the RS scene classification, the land covers contained in RS scenes provide very important clues to judge the category of scenes. However, the composition of RS scenes is usually very complex. An RS scene may contain multiple land-cover units that are quite different in size and appearance. It is difficult to represent the RS scenes with handcrafted descriptors [e.g., scale-invariant feature transform (SIFT)] [19], [20]. Early works exploited the sparse representation and low-rank representation to represent the RS scene images. For example, Wang *et al.* [21] improved the low-rank representation with the locality constraint and the structure constraint. Wang *et al.* [22] proposed a multiview clustering method to investigate the data correlation in crowd scenes. Recent works have attempted to classify RS scenes using the powerful CNNs [8], [23]. These works can be summarized into three categories: the direct CNN methods, the feature encoding methods, and the aggregation learning methods.

A. Direct CNN Methods

Early CNN-based works attempted to transfer the existing AlexNet [6], GoogLeNet [7], and VGG [24] to the RS scene classification. Liang *et al.* [8] adopted a transfer learning method to fine-tune the existing CNNs for RS scene classification. Cheng *et al.* [25] attempted to fine-tune the existing AlexNet, GoogLeNet and VGG for classification. Wang *et al.* [26] utilized the fully connected layer of the

pre-trained ResNet to extract the RS scene representations. However, the features from different hierarchical layers of CNN are ignored in the direct CNN methods.

B. Feature Encoding Methods

Recent CNN-based works exploited the pre-trained CNNs to extract deep features. Several works utilized the pre-trained CNN to extract features from multi-scale RS scene patches. Zhao and Du [27] employed a CNN to extract local spatial patterns from multi-scale patches and then encoded the local spatial patterns into an RS scene representation with the bag of words (BoW) [19]. Similarly, Hu *et al.* [28] also proposed to utilize a CNN pre-trained on ImageNet [29] as a deep feature extractor to represent multi-scale patches. Li *et al.* [30] utilized different convolutional layers of the pre-trained CNN to capture features from multi-scale patches. Then, the improved Fisher vector (FV) [31] was employed to aggregate convolutional features. Zheng *et al.* [32] proposed a multiscale pooling to extract the local information from the last convolutional layer and exploited the FV to generate the holistic scene representation. However, those methods need to pre-process RS scenes into many multi-scale scene patches. The number of multi-scale scene patches is usually very large. Extracting features from these similar image patches is time-consuming and redundant. On the other hand, several works attempted to exploit different kinds of CNNs to extract different features from RS scenes. Zhang *et al.* [12] exploited a gradient boosting random method to ensemble multiple different CNNs for RS scene classification. Lu *et al.* [33] exploited a recurrent neural network to fuse the features of CNN and the SIFT feature. Yu and Liu [14] exploited multiple convolutional networks with different receptive fields to capture different features. However, it is complicated to ensemble multiple CNNs for RS scene classification. Recent works on the CNNs indicated that the features from top convolutional layers contain high-level semantic information of large objects and the features of bottom convolutional layers can represent the small objects [13], [34], [35]. Lin *et al.* [13] proposed a top–down feature fusion strategy that exploited the high-level semantic information to complement features of bottom convolutional layers. Kong *et al.* [34] and Liu *et al.* [35] incorporated the features of top convolutional layers into features of bottom convolutional layers to represent small objects. Inspired by the complementarity of multilayer convolutional features, several works focused on aggregating multilayer convolutional features for RS scene classification. Wang *et al.* [9] employed a pre-trained CNN to extract multilayer features from different convolutional layers, and then, the vectors of locally aggregated descriptors (VLADs) [36] was utilized to generate the RS scene representation. He *et al.* [37] proposed a covariance pooling to integrate multilayer convolutional features. Chaib *et al.* [15] employed a discriminant correlation analysis [38] to merge multilayer convolutional features. However, these RS scene classification methods directly use traditional unsupervised feature encoding methods to integrate all information of multilayer convolutional features.

C. Aggregation Learning Methods

Recently, a mass of works focus on adjusting the structures of CNN to improve classification results. Yang and Ramanan [10] proposed to aggregate convolutional features of multiple layers with a directed acyclic graph for scene classification, because the convolutional features from different layers are interested in objects with different sizes. Lu *et al.* [39] proposed an end-to-end network to aggregate different convolutional features. In order to deal with the problems that the sizes and positions of objects are easy to change in the scene, Hayat *et al.* [40] designed a spatial unstructured layer embedded in the CNNs. DenseNet [41] connected each convolutional feature to other convolutional layers when data are propagated in the forward direction. SENet [42] recalibrated channel-wise features to emphasize the channels with effective information and suppress the useless channels. Kalantidis *et al.* [16] exploited a cross-dimensional weighting to aggregate multilayer convolutional features. Tang *et al.* [17] proposed a FisherNet to integrate the FV into the CNN for end-to-end training. Motivated by the success of recent end-to-end networks, an end-to-end GBNet is designed to aggregate multilayer convolutional features for RS scene classification. Wang *et al.* [43] proposed a recurrent attention structure to capture the features from key regions of RS scenes. Different from [43], the proposed GBNet focuses on aggregating multilayer convolutional features and enhancing the complementary information.

III. Motivation

Previous works [9], [30] usually select certain convolutional features empirically for feature aggregation. A few works have quantitatively analyzed the performance of each convolutional feature for RS scene classification. In this section, several experiments are conducted to quantitatively analyze the classification performance of different convolutional features. Based on the performance, several convolutional features are selected for feature aggregation. Furthermore, popular feature aggregation operations are introduced to integrate the complementary information among multilayer convolutional features. The details are introduced as follows.

A. Analyzing the Performance of Convolutional Features

The VGG-16 [24] is chosen as the basic network for experimental analysis due to its superior performance in the RS scene classification [14], [28], [44]. The features from different convolutional layers of VGG-16 are exploited to classify the RS scenes into proper categories with the linear softmax classifier. First, the weight parameters of VGG-16 are initialized with the weight parameters pre-trained on the ImageNet data set [45]. Then, the convolutional features are extracted from the different layers of initialized VGG-16 and flattened into vectors. To make the training of softmax classifier converges faster, the L2 normalization is utilized to normalize the feature vectors. Finally, the normalized feature vectors are fed to the linear softmax classifier for classification. In the training phase, the softmax classifier is trained with 50 epochs and the learning rate is 0.001 on each data set. The classification

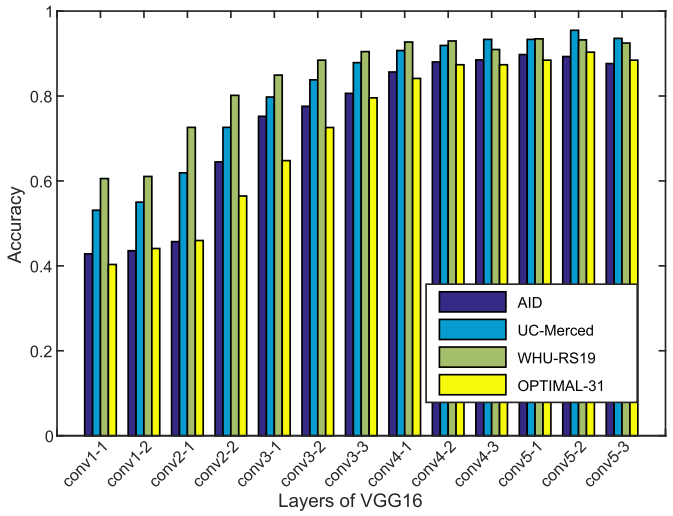

Fig. 2. Classification accuracy of each convolutional feature on the AID, UC-Merced, WHU-RS19, and OPTIMAL-31 data sets. The horizontal axis represents the feature of each convolutional layer and the vertical axis indicates the classification accuracy.

performance of each convolutional feature is shown in Fig. 2. The "conv$i$-j" in Fig. 2 means the convolutional feature from the $i-j$th convolutional layer. Reasonably, the features generated from different convolutional layers show different performance. It is not that the deeper the feature is, the better the classification performance is for the RS scenes. As shown in Fig. 2, the accuracies of conv5-2 are higher than those of conv5-3.

To indicate the relations between the convolutional features and the RS scene categories, the detailed RS scene classification results of each convolutional feature on the AID data set [46] are shown in Fig. 3. The row in Fig. 3 represents the classification results of a convolutional feature in different scene categories. The column shows the classification results of a scene under different convolutional features. In each column, the best result is marked with a black rectangular box. As shown in Fig. 3, the features from different convolutional layers are discriminative to different scene categories. The features from the bottom convolutional layers perform well in the dense residential, school, and square, where the composition is relatively complex. In these scenes, the size of land covers is very small and the spatial layout is compact. The features from the top convolutional layers show good performance in the bridge, center, playground, and pond, where the size of land covers is relatively large. The experimental results in Fig. 3 confirm the viewpoint from the latest works [9], [15] that aggregating multilayer convolutional features can effectively improve the accuracy for RS scene classification.

With in-depth studies of architectures of CNNs, several works [13], [34], [35] demonstrate that the top convolutional layers are well versed in harvesting the high-level semantic information for large-scale land covers and the bottom convolutional layers can effectively extract the appearance information for small-scale land covers. Therefore, it is natural to aggregate the features from different convolutional layers for the classification task. However, previous works empirically select certain convolutional features. In this section, the best combination of convolutional features is selected based on the quantitative experimental analysis. In fact, the CNN is a hierarchical structure with many convolutional features. Similar to [37], the number of convolutional features for feature aggregation is empirically set to three. There are 13 convolutional features totally in VGG-16, and three features are selected from them. There are 286 combinations. In order to quantify the performance of different combinations, an ideal assumption is defined as follows. Taking the combination of conv1-1, conv1-2, and conv2-1 as an example, if a scene is correctly classified with at least one of conv1-1, conv1-2, and conv2-1, it is considered that this combination can correctly classify this scene. Based on this ideal assumption, the classification results of all combinations are shown in Fig. 4. On each data set, the top three combinations are shown in Table I. As can be seen from Table I, shallow convolution features (conv1-1, conv2-1, and conv2-2) are helpful to RS scene classification on the UC-Merced, WHU-RS19, and OPTIMAL-31 data sets. However, the interference information among these features can decrease the accuracy in fact. The ideal assumption is only used for the purpose of convolutional feature selection. In Section III-B, various feature aggregation operations are introduced for the classification task.

B. Feature Aggregation

As shown in Fig. 3, the convolutional features of different layers are complementary for RS scene classification. Therefore, designing a feature aggregation operation is critical to improve classification accuracy. Although many of the existing works on the RS scene classification have been devoted to the study of convolutional feature aggregation [9], [30]. However, in recent works, the traditional feature encoding methods [VLAD [36] and improved Fisher kernel (IFK) [31]] are still utilized to merge convolutional features. The traditional feature aggregation operations usually aggregate convolutional features in an unsupervised manner and cannot effectively use labels to supervise the aggregation process [11], [47]. With the rapid development of deep networks, a variety of convolutional feature aggregation operations have been designed for the end-to-end training in several visual tasks [13], [48]. As shown in Fig. 5, the classic and effective feature aggregation operations can be divided into two categories: the concatenated aggregation and the arithmetic aggregation. The concatenated aggregation refers to stacking the convolutional feature maps across the feature channels. The arithmetic aggregation means computing the sum (or other operation) of the convolutional features at the same spatial positions and channels. As shown in Section III-A, the convolutional features conv3-3, conv5-1, and conv5-3 are chosen for the analytical experiment of convolutional feature aggregation on the AID data set. Let $x_i, i = 1, 2, 3$ represent the conv3-3, conv5-1, and conv5-3, respectively, and $\hat{x}$ represent the aggregated features. Two feature aggregation operations are introduced as follows.

1) Concatenated Aggregation: As shown in Fig. 5(b), due to the inherent hierarchy of CNNs, the different convolutional features x_i can be concatenated to generate the robust features $\hat{x}$ [see Fig. 5(b)] for the RS scene classification task.

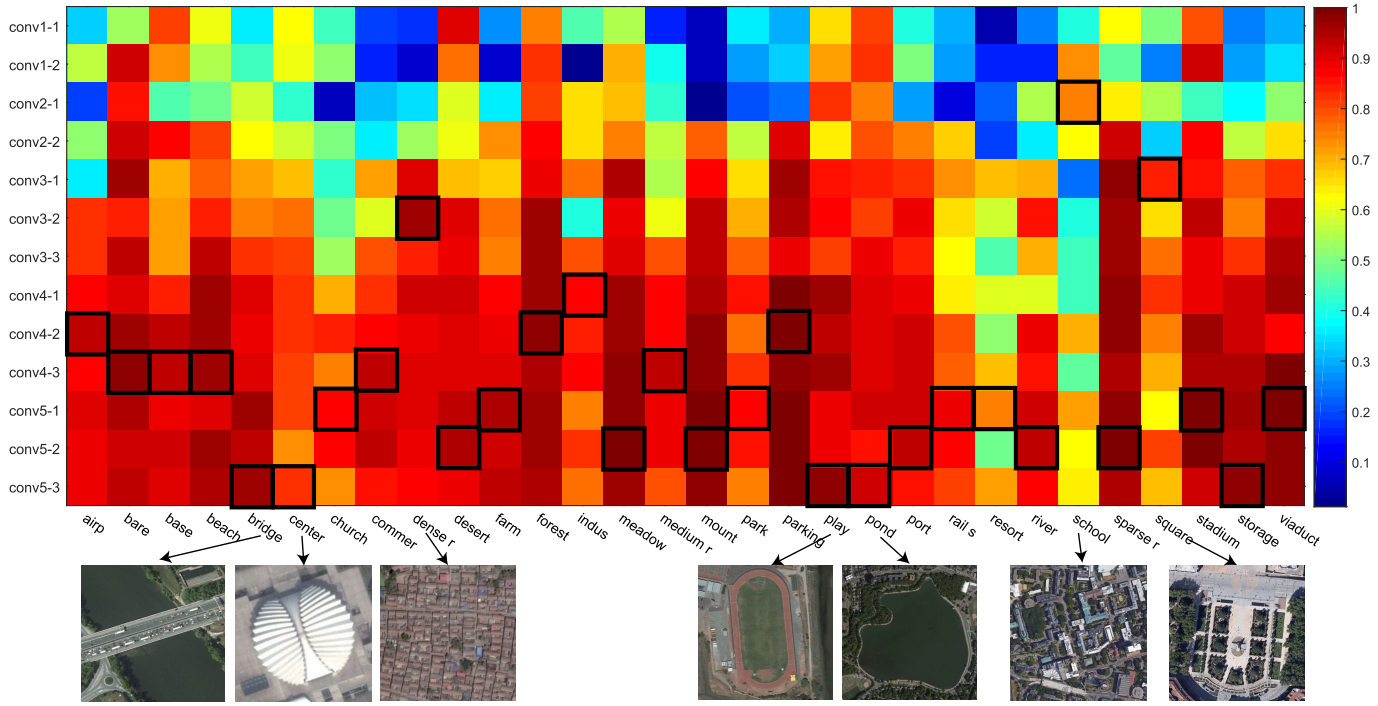

Fig. 3. Detailed classification results of various convolutional features in each RS scene category. The redder color indicates the higher classification accuracy and the more blue color represents the lower classification accuracy. As shown in the black rectangular box, the different convolutional features are discriminative for different RS scene categories.

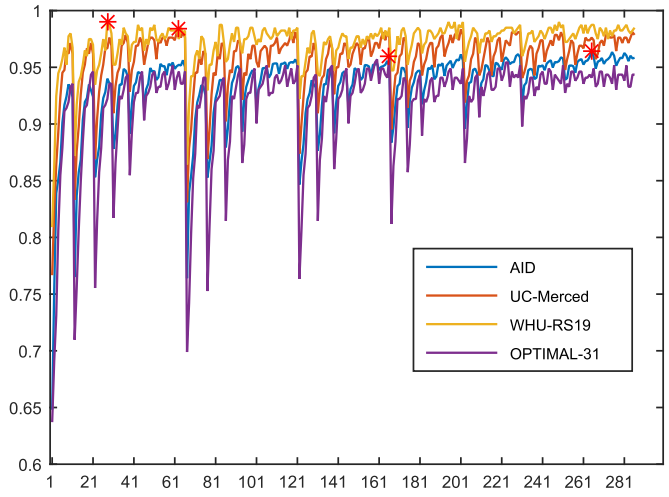

Fig. 4. Classification results of different combinations of convolutional features. "*" represents the best result. Y-axis is the classification accuracy. X-axis represents different combinations. On the x-axis, scale 1 denotes the combination of conv1-1, conv1-2, and conv2-1, scale 2 denotes the combination of conv1-1, conv1-2, and conv2-2, . . ., scale 11 denotes the combination of conv1-1, conv1-2, and conv5-3, scale 12 denotes the combination of conv1-1, conv2-1, and conv2-2, . . ., and scale 286 denotes the combination of conv5-1, conv5-2, and conv5-3.

TABLE I

OAs OF THE TOP THREE COMBINATIONS ON EACH DATA SET

	Different combinations	Accuracy
AID Dataset (50% for training)		
1	conv3-3, conv5-1 and conv5-3	96.38%
2	conv4-1, conv5-1 and conv5-3	96.34%
3	conv4-2, conv5-1 and conv5-3	96.26%
UC-Merced Dataset (80% for training)		
1	conv1-1, conv4-3 and conv5-3	98.33%
2	conv2-2, conv5-1 and conv5-3	98.33%
3	conv1-1, conv4-3 and conv5-2	98.10%
WHU-RS19 Dataset (60% for training)		
1	conv1-1, conv2-2 and conv5-1	98.99%
2	conv2-2, conv4-2 and conv5-2	98.99%
3	conv2-2, conv4-3 and conv5-2	98.99%
OPTIMAL-31 Dataset (80% for training)		
1	conv2-1, conv5-2 and conv5-3	95.97%
2	conv2-1, conv4-2 and conv5-3	95.70%
3	conv2-1, conv4-2 and conv5-2	95.43%

Before the concatenated operation, all convolutional features of different layers should be resized to the same size $[M, N]$ and the number of feature channels C_i can be arbitrary. The concatenated aggregation can be formulated as

$$\hat{x} = \text{cat}[x_1, x_2, x_3] \tag{1}$$

where *cat* represents the concatenated operation and $x_i \in \mathbb{R}^{M\times N\times C_i}$, $\hat{x} \in \mathbb{R}^{M\times N\times(C_1+C_2+C_3)}$.

2) Arithmetic Aggregation: The aggregated features $\hat{x}$ [see Fig. 5(c)] are computed by the element-wise arithmetic operations from each convolutional feature x_i. Before the arithmetic aggregation, the sizes and number of channels of each convolutional feature x_i should be unified to $[M, N, C]$ by the pooling and 1×1 convolution operation, respectively. M and N represent the width and height of features, respectively, and C represents the number of feature channels. The arithmetic aggregation can be formulated as

$$\hat{x} = \text{arith}(x_1, x_2, x_3) \tag{2}$$

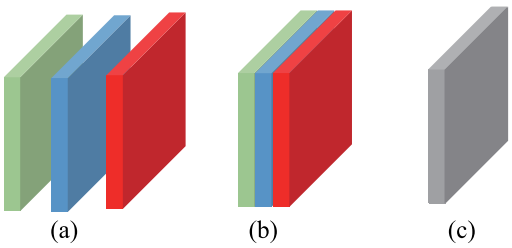

Fig. 5. Different feature aggregation methods. (a) Convolutional features from different layers of VGG-16, where the green, blue, and red represent the features of conv3-3, conv5-1, and conv5-3, respectively. (b) Aggregated feature generated with the concatenated aggregation method. (c) Aggregated feature generated with the arithmetic aggregation method.

where the arith represents the element-wise sum (or product, maximum) operation and $x_i, \hat{x} \in \mathbb{R}^{M\times N\times C}$. When $arith$ is the sum operation, $\hat{x} = x_1 + x_2 + x_3$, the arithmetic aggregation method is similar to the FPN [13].

The aforementioned feature aggregation operations, which simply merge multilayer convolutional features together, do not consider whether the context of the convolutional features is effective for the task. As shown in Fig. 3, conv5-1 performs better than conv5-3 and conv3-3 in seven RS scene categories on the AID data set. In aggregating conv3-3, conv5-1, and conv5-3, eliminating the interference information of conv5-3 and conv3-3 in these seven RS scene categories is helpful for RS scene classification. In this paper, a GBNet is designed for effective integration of the complementary information and elimination of the interference information among multilayer convolutional features. The details of the proposed method are introduced in Section IV. In Section V-C, the performance of the aforementioned feature aggregation operations is compared with the proposed method on the AID data set.

IV. Proposed Method

As demonstrated in Section III-A, the convolutional features from different convolutional layers are discriminative to different RS scene categories. The top convolutional features are versed in harvesting the semantic features. The bottom convolutional features are adept in extracting the appearance features. For RS scene classification, because the composition of RS scenes is very complex, both the semantic features and the appearance features are essential. In this paper, a GBNet is proposed to integrate the complementarity of multilayer convolutional features for RS scene classification. Taking the combination of conv3-3, conv5-1, and conv5-3 as an example, the details of the proposed method are introduced as follows. In Section V, we also demonstrate that the GBNet with the combination of conv3-3, conv5-1, and conv5-3 performs better than other combinations.

A. Overview

As shown in Fig. 6, the proposed GBNet consists of three modules: CNN module, gated bidirectionally connected module, and classification module. The CNN module is exploited to harvest the convolutional features. In this paper, VGG-16 [24] is employed to extract multiple convolutional features from different convolutional layers. Then, the convolutional features from multiple layers are fed into the gated bidirectionally connected module for hierarchical feature aggregation. The gated bidirectional connection module can generate the semantic-assist features and appearance-assist features. Finally, the semantic-assist features and appearance-assist features are integrated by the global pooling and concatenation operations, and the linear softmax classifier is utilized for RS scene classification.

B. Convolutional Neural Network

Since the CNN shows much better performance than handcrafted feature descriptors [8], [23], it has been widely applied to the RS scene classification [49]. In particular, VGG-16 [24] has shown excellent performance in the recent works [14], [43], [50]. In the proposed GBNet, the VGG-16 is also selected as the basic CNN for RS scene classification. VGG-16 is a typical hierarchical network architecture, which consists of multiple convolutional layers. Conv3-3, conv5-1, and conv5-3 of VGG-16 are exploited to extract the convolutional features. Different convolutional features can distinguish different scene categories. Aggregating multiple convolutional features is beneficial to RS scene classification [9], [15], [30]. Although multilayer convolutional features can provide complementary information [15], they may also bring some interference information, such as feature redundancy and mutual exclusion among the convolutional features. Interference information among multiple convolutional features can reduce the classification accuracy [15]. In this paper, a gated bidirectional connection is designed to aggregate features. The complementary information among different convolutional features can be adaptively merged, and the interference information can be eliminated by the GBNet.

C. Gated Bidirectional Connection

To make full use of the complementary information and eliminate the interference information among different convolutional features, a gated bidirectional connection module is proposed. The bidirectional connection can be divided into the top–down direction and bottom–up direction. On the one hand, for the top–down direction, the semantic features of top convolutional features (conv5-3) are hierarchically fed into the bottom convolutional features (conv3-3) with the gated function to generate the semantic-assist features. The gated function is utilized to control the passing of the complementary information contained in conv5-3 that is beneficial to classification. In other words, the interference information can be eliminated by the gated function. As shown in Fig. 3, the convolutional features of conv5-1 outperform those of conv5-3 in specific RS scene categories. Therefore, it is essential to exploit the gated function to restrain the passing of information contained in conv5-3 that interferes with those RS scene classification. On the other hand, for the bottom–up direction, the appearance features of bottom convolutional layers are hierarchically aggregated into the top convolutional features (conv5-3) with the gated function to generate

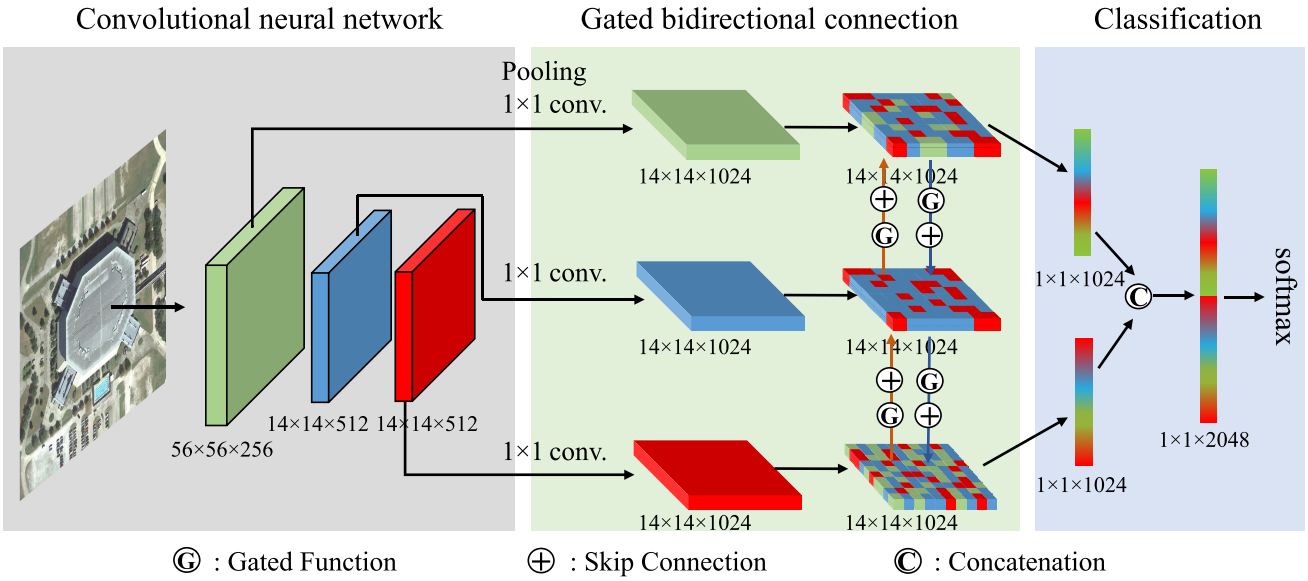

Fig. 6. Proposed GBNet consists of three modules: CNN module, gated bidirectional connection module, and classification module. First, the RS scenes are fed to a CNN module to extract the convolutional features from different convolutional layers. Then, the gated bidirectional connection module is exploited to aggregate the context information of features, useful to the RS scene classification task, into the semantic-assist features and appearance-assist features. Finally, the semantic-assist features and appearance-assist features pooled by a global average pooling are concatenated for the classification.

the appearance-assist features. Before the gated bidirectional connection, all convolutional features are normalized by the L2 normalization operation across each spatial position. Then, a 4×4 average pooling is exploited to pool the size of conv3-3 to $M \times N$, which is the same as the size of conv5-1 and conv5-3. The details of the gated bidirectional connection are introduced as follows.

1) Bottom–Up Direction: In the bottom–up direction, the bottom convolutional features are hierarchically integrated to the top convolutional features. Due to a lot of convolution and pooling operations, the top convolutional features are weak in extracting the appearance features of the small-scale objects, which are beneficial to the RS scene classification task, as shown in Fig. 3. Therefore, it is natural to incorporate appearance features of bottom convolutional layers into top convolutional features. The detailed structure of the bottom–up connection is shown in Fig. 7. Similar to Section III-B, the normalized conv3-3, conv5-1, and conv5-3 are defined as $x_1 \in \mathbb{R}^{M \times N \times C_1}$, $x_2 \in \mathbb{R}^{M \times N \times C_2}$, and $x_3 \in \mathbb{R}^{M \times N \times C_3}$, respectively.

First, x_1 is fed to a 1×1 convolutional layer to resize the number of channels to C. After the 1×1 convolutional layer, $x_1 \in \mathbb{R}^{M \times N \times C_1}$ is converted to $x_1^b \in \mathbb{R}^{M \times N \times C}$. In detail, this 1×1 convolutional layer contains C convolutional kernels $w_1^{b,t} \in \mathbb{R}^{1 \times 1 \times C_1}, t = 1, 2, \ldots, C$. x_1 is convoluted with each convolutional kernel $w_1^{b,t}$ to generate $x_1^{b,t} \in \mathbb{R}^{M \times N}$. Then, $x_1^{b,t}$ is stacked by the channel to generate x_1^b

$$x_1^{b,t} = \sigma\left(w_1^{b,t} * x_1\right) \tag{3}$$

$$x_1^b = \left[x_1^{b,1}, x_1^{b,2}, \ldots, x_1^{b,C}\right] \tag{4}$$

where $[\cdot]$ represents stacking by the channel and $\sigma(\cdot)$ represents the rectified linear unit (ReLU) activation function. For convenience, the formula of convolutional layer is simplified as follows:

$$x_1^b = \sigma\left(w_1^b * x_1\right) \tag{5}$$

where w_1^b is the weight parameter of convolutional layer, $*$ represents the convolution, and the bias parameter of convolution is omitted in this paper.

Then, the gated function is utilized to control the passing of complementary information of x_1^b to x_2^b. The structure of gated function is motivated by [42] to generate a C-dimensional gated vector $g_1^b \in \mathbb{R}^{1 \times C}$ (C is the channel number of x_1^b) with values of 0–1 for each element. The ith channel of x_1^b is multiplied by the ith element of g_1^b to eliminate the interference information. As shown in Fig. 7, the components in the black dotted rectangles represent the gated function. The gated function is composed of a global average pooling layer and two fully connected layers with the activation functions. x_1^b is fed to a global average pooling layer to generate a $1 \times 1 \times C$ vector, and this $1 \times 1 \times C$ vector is fed to a fully connected layer combined with a ReLU activation function and a fully connected layer combined with a sigmoid activation function to generate the g_1^b. g_1^b can be formulated as

$$g_1^b = \mathrm{sig}\left(fc\left(\sigma\left(fc\left(\mathrm{pool}\left(x_1^b\right)\right)\right)\right)\right) \tag{6}$$

where $\mathrm{sig}(\cdot)$ represents the sigmoid activation function, $\mathrm{sig}(x) = (1 + \exp(-x))^{-1}$, $\sigma(\cdot)$ represents the ReLU activation function, $fc(\cdot)$ represents the fully connected layer, and $\mathrm{pool}(\cdot)$ represents the global average pooling layer. With the gated function, the complementary information of x_1^b is fed to x_2^b. x_2^b is composed of x_2 and the complementary information of x_1^b

$$x_2^b = \sigma\left(w_2^b * x_2\right) + x_1^b \cdot g_1^b \tag{7}$$

where $\cdot$ represents that the ith channel of x_1^b is multiplied by the ith element of g_1^b and w_2^b is the weight parameter of the

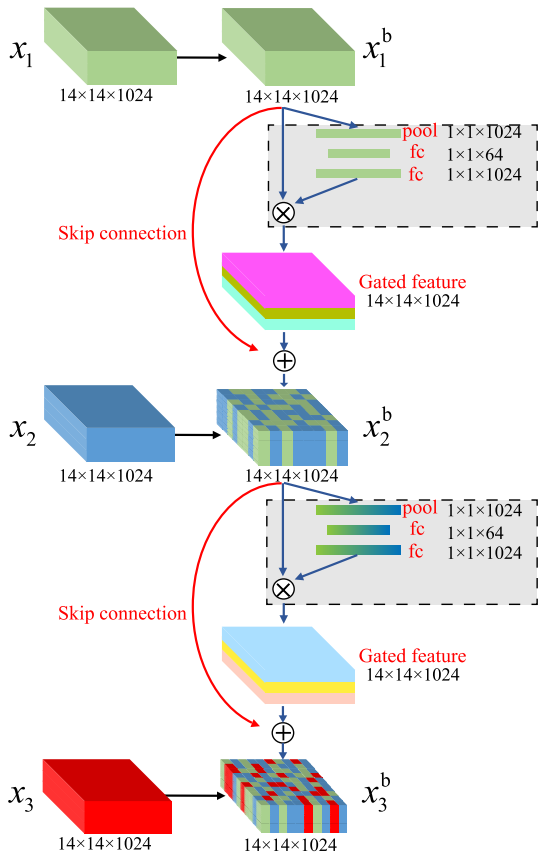

Fig. 7. Gated bottom–up connection. x_1^b, x_2^b, and x_3^b refer to the unified convolutional features of conv3-3, conv5-1, and conv5-3 from VGG-16, respectively. The components in the black dotted rectangle represent the gated function.

convolutional layer. The gated function is the key component in the proposed method. However, the gated function introduces a plenty of weight parameters, which makes it difficult to train the proposed GBNet, as shown in Section V-C. To make training easy to converge, a skip connection is exploited to combine with the gated function. The formulation of x_2^b becomes

$$x_2^b = \sigma\left(w_2^b * x_2\right) + x_1^b \cdot g_1^b + x_1^b. \tag{8}$$

The skip connection connects x_1^b to x_2^b, including the interference information that has been eliminated by the gated function. The skip connection is a compromise for training difficulties. With the introduction of the skip connection, the purpose of the gated bidirectional connection has changed from eliminating the interference information to enhancing the complementary information. If the gated vector g_1^b in 7 cannot eliminate the interference information and retain the complementary information, the gated vector g_1^b in 8 cannot enhance the complementary information. In other words, if the classification accuracy is improved after utilizing the skip connection, it demonstrates that complementary information can be enhanced by 8. g_1^b in 8 can retain the complementary information and eliminate the interference information.

Finally, the appearance-assist feature x_3^b is generated with the gated function and skip connection similar to x_2^b

$$x_3^b = \sigma\left(w_3^b * x_3\right) + x_2^b \cdot g_2^b + x_2^b \tag{9}$$

$$g_2^b = \mathrm{sig}\left(fc\left(\sigma\left(fc\left(\mathrm{pool}\left(x_2^b\right)\right)\right)\right)\right) \tag{10}$$

where w_3^b is the weight parameter of convolution.

2) Top–Down Direction: In the top–down direction, the top convolutional features (x_3) are hierarchically integrated into the bottom convolutional features (x_1) layer-by-layer. Due to the few convolutional operations, the convolutional features from bottom layers have small receptive fields and capture weak semantic information from RS scenes. Additionally, the gated function and skip connection are utilized to control the passing of complementary information of x_3. The details of top–down direction are similar to the bottom–up direction. Finally, the semantic-assist feature x_1^t is generated in the top–down direction. The top–down direction can be formulated as

$$x_i^t = \sigma\left(w_i^t * x_i\right) + x_{i+1}^t \cdot g_{i+1}^t + x_{i+1}^t, \quad i = 1, 2 \tag{11}$$

$$g_{i+1}^t = \mathrm{sig}\left(fc\left(\sigma\left(fc\left(\mathrm{pool}\left(x_{i+1}^t\right)\right)\right)\right)\right), \quad i = 1, 2 \tag{12}$$

where $x_3^t = \sigma(w_3^t * x_3)$, $x_i^t \in \mathbb{R}^{M\times N\times C}$ denotes the features aggregated in the top–down direction, and w_i^t is the weight parameter of convolution.

The proposed gated bidirectional connection is different from the GRU [51]. First, the GRU is an effective way to process sequence data. The proposed gated bidirectional connection can be embedded into VGG-16 to aggregate multilayer convolutional features. Second, the hidden layers of GRU share the same weight parameters. In the proposed gated bidirectional connection, each 1×1 convolution layer does not share weight parameters.

D. Classification

In this section, the detailed steps of integrating the appearance-assist feature (x_3^b) and semantic-assist feature (x_1^t) for RS scene classification are introduced. First, the global pooling operation is utilized to pool x_3^b and x_1^t into feature vectors. Because the RS scene data set often contains only a small amount of training data, it is difficult to train the CNN with a large number of parameters. For training the end-to-end GBNet easily, the global pooling operation is employed to reduce the dimension of x_3^b and x_1^t from $\mathbb{R}^{M\times N\times C}$ to $\mathbb{R}^{1\times 1\times C}$. The number of parameters in the classifier is greatly reduced. Then, the concatenation operation is exploited to merge the pooled x_3^b and x_1^t into x_{con}. Finally, a linear softmax classifier is employed for classification. The objective function of GBNet is the cross entropy as follows:

$$\mathrm{Loss}_{\mathrm{conv}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{K}\mathbf{1}\{y^n = j\}\log\frac{e^{\theta_j^T x_{\mathrm{con}}^n}}{\sum_{i=1}^{K} e^{\theta_i^T x_{\mathrm{con}}^n}} \tag{13}$$

where x_{con} is the concatenation of the appearance-assist feature and semantic-assist feature, y is the scene label, θ is the parameter of linear softmax, K is the number of RS scene category, N is the size of the training batch, and $\mathbf{1}\{\cdot\}$ is an indicator function (if y^n is equal to j, $\mathbf{1}\{y^n = j\} = 1$; otherwise, $\mathbf{1}\{y^n = j\} = 0$).

V. Experiments

The GBNet can effectively harvest the discriminative scene representation for the RS scene classification task. In this section, the proposed GBNet is evaluated on the public RS scene classification data sets. First, the data sets are introduced in detail in Section V-A. Second, Section V-B introduces the experimental settings. Third, ablation experiments are conducted to demonstrate the performance of the different feature aggregation operations in Section V-C. Fourth, the GBNet combined with the global feature is explored for RS scene classification in Section V-D. Fifth, the performance of GBNet combined with the global feature is compared with several state-of-the-art methods in Section V-E. Next, the performance of different convolutional feature combinations is explored in Section V-F. Finally, the training and testing time is reported in Section V-G.

A. Data Set

The data sets are defined in the following.

1) AID data set[1] [46] is made up of 30 RS scene categories, each of which has at least 200 images. All images are extracted from Google Earth imagery. There are 10 000 images in total, the size of which is 600×600. Following the data partition in [46], 50% and 20% of the images in each RS scene category are randomly selected for training.
2) UC-Merced data set[2] [19] is composed of 21 RS urban scene categories around the country. The images are extracted from the USGS National Map. There are 100 images in each RS scene category, the size of which is 256×256. Following the data partition in [19], 80% and 50% of the images in each scene category are randomly selected to train the proposed method.
3) WHU-RS19 data set[3] [52] consists of 19 RS scene categories, each of which contains 50 images at least. The images are extracted from Google Earth imagery. The image size is 600×600. Following the data partition in [52], 60% and 40% of the images in each category are randomly selected as the training set.
4) OPTIMAL-31 data set[4] [43] is composed of 31 RS scene categories, each of which contains 60 RS images. The images are also extracted from Google Earth imagery. There are 1860 images in the data set, each of which has a size of 256×256. Following the data partition in [43], 80% of the images in each RS scene category are randomly selected as the training set.

B. Experimental Settings

1) *Evaluation Measures:* In the RS scene classification, the overall accuracy (OA) and average accuracy (AA) are the most commonly used evaluation measures to quantify the performance of RS scene classification methods. OA refers to the ratio of the number of correctly classified images to the total number of testing images. AA refers to the mean of the accuracy of each RS scene category. In addition, confusion matrix (CM) is also a common measure to evaluate classification methods. CM is a particular matrix to visualize the classification results of each RS scene category. In this paper, the OA, AA, and CM are employed to quantify the performance of the proposed GBNet.
2) *Training Settings:* Due to the superior performance of VGG-16 [24] in recent RS scene classification methods [14], [28], [44], VGG-16 is chosen as the feature extractor to extract multiple convolutional features. The weight parameters of VGG-16 are initialized with the weight pre-trained on ImageNet [29].[5] Other weight parameters of GBNet are initialized randomly by the Gaussian distribution with the mean of 0 and the variance of 0.001. The stochastic gradient descent (SGD) is exploited to optimize the weight parameters of GBNet. The value of weight decay is 0.0005 and the value of momentum is 0.9. The batch size is 50. The learning rate in the training phasing is 0.001. All experiments are conducted with the Ubuntu 14.04 system, CAFFE[6] toolbox, Inter Core i7-5930K, GeForce GTX Titan X, and 64-GB RAM.

C. Ablation Experiments

In this paper, a gated bidirectional connection is proposed for the convolutional feature aggregation. Different from traditional concatenated and arithmetic aggregation as shown in Section III-B, the gated bidirectional connection focuses on simultaneously aggregating complementary information and eliminating interference information. The gated bidirectional connection is composed of the top–down direction and bottom–up direction. In each pathway, the gated function is utilized to eliminate interference information and the skip connection is used to promote network convergence. In total, several components are exploited for feature aggregation in the proposed gated bidirectional connection. In order to understand the performance of each component clearly, the ablation experiments are conducted on the AID data set. In the ablation experiments, 50% of the images in each RS scene category are randomly selected for training.

1) *GBNet:* The proposed gated bidirectional neural network.
2) *Bottom–Up:* The bottom–up direction with a gated function and skip connection of GBNet. A global pooling is employed to pool the appearance-assist features into a feature vector and the linear softmax is exploited.
3) *Bottom–Up Without Skip:* Remove the skip connection from the bottom–up direction for classification.
4) *Bottom–Up Without Gating and Skip:* A 1×1 convolution is utilized to replace the skip connection and gated function of the bottom–up direction.

[1] http://captain.whu.edu.cn/project/AID/
[2] http://weegee.vision.ucmerced.edu/datasets/landuse.html
[3] http://www.escience.cn/people/yangwen/whu-rs19.html
[4] http://crabwq.github.io/
[5] http://www.robots.ox.ac.uk/~vgg/software/very_deep/caffe/VGG_ILSVRC_16_layers.caffemodel
[6] http://caffe.berkeleyvision.org/

5) *Top–Down:* The top–down direction with a gated function and skip connection of GBNet. A global pooling is employed to pool the semantic-assist features into a feature vector and the linear softmax is exploited.
6) *Top–Down Without Skip:* Remove the skip connection from the top–down direction for classification.
7) *Top–Down Without Gating and Skip:* A 1×1 convolution is utilized to replace the skip connection and gated function of the top–down direction.
8) *Direct Aggregation With Product:* Aggregate multilayer convolutional features without direction as the introduction in Section III-B. First, the pooling and 1×1 convolution are applied to each convolutional feature. Then, an element-wise product is utilized to merge the features.
9) *Direct Aggregation With Maximum:* Similar to the previous point, an element-wise maximum is utilized to merge the features.
10) *Direct Aggregation With Concatenation:* Similar to the previous point, a concatenation is utilized to integrate the features.
11) *Direct Aggregation With Summation:* Similar to the previous point, an element-wise summation is utilized to merge the features.
12) *Direct Aggregation With Gating and Summation:* The gated function is applied to each convolutional feature separately, and then, multilayer convolutional features are directly merged for classification.
13) *Direct Aggregation With Gating, Skip, and Summation:* The gated function and skip connection are applied to each convolutional feature separately, and then, multilayer convolutional features are directly merged for classification.
14) *GBNet With Product:* An element-wise product is utilized to merge the appearance-assist feature and semantic-assist feature in GBNet.
15) *GBNet With Maximum:* An element-wise maximum is utilized to merge the appearance-assist feature and semantic-assist feature in GBNet.
16) *GBNet With Summation:* An element-wise summation is utilized to merge the appearance-assist feature and semantic-assist feature in GBNet.

In this section, the effects of different components contained in the GBNet are studied. The experimental results are shown in Table II. First, the proposed GBNet obtains 94.12% OA on the AID data set. Then, the GBNet is decomposed into the bottom–up direction and top–down direction. The effects of gated function and skip connection on the performance of GBNet are studied. The skip connection improves the classification accuracy by 0.54% and 1.96% at the bottom–up direction and top–down direction, respectively. The gated function combined with skip connection improves the classification accuracy by 0.82% and 2.56% at the bottom–up direction and top–down direction, respectively. The classification results show that the gated function can remove interference information for feature aggregation. The skip connection promotes the convergence of GBNet. In addition, the other classic feature aggregation operations are explored for RS scene classification, as the introduction in Section III-B. The directly feature aggregation refers to aggregating multilayer convolutional features without direction. The difference between the gated bidirectional connection and directly feature aggregation is that one is a direct feature aggregation and the other is a feature fusion from the top–down direction and bottom–up direction. As shown in Table II, the arithmetic aggregation with summation achieves 93.24% classification accuracy. Furthermore, the gated function and skip connection are applied to arithmetic aggregation with summation. For the gated function only, the accuracy of arithmetic aggregation with summation is reduced by 1.20%. This is because the gated function introduces too many weight parameters, which makes the GBNet difficult to train. The gated function and skip connection improve the classification accuracy by 0.24% for the arithmetic aggregation with summation. Furthermore, the arithmetic aggregation is explored for merging the appearance-assist feature and semantic-assist feature. As shown in Table II, the concatenated aggregation shows better performance than the arithmetic aggregation.

TABLE II
ABLATION EXPERIMENTS ON THE AID DATA SET

Method	50% for training
conv3-3	80.62%
conv5-1	89.76%
conv5-3	87.64%
Bottom-up	93.28%
Bottom-up without skip	92.74%
Bottom-up without gating and skip	92.46%
Top-down	93.58%
Top-down without skip	91.62%
Top-down without gating and skip	91.02%
Direct aggregation with product	89.74%
Direct aggregation with maximum	92.30%
Direct aggregation with concatenation	93.08%
Direct aggregation with summation	93.24%
Direct aggregation with gating and summation	92.04%
Direct aggregation with gating, skip and summation	93.48%
GBNet with product	93.76%
GBNet with maximum	93.68%
GBNet with summation	93.94%
GBNet	**94.12%**

D. GBNet Combined With Global Features

The proposed GBNet focuses on aggregating multilayer convolutional features for RS scene classification. However, VGG-16 is a hierarchical network, and its fully connected layer can be used to extract the global feature from the RS scenes. In this section, the global feature is further integrated to GBNet. The details are shown in Fig. 8. Similar to GoogLeNet [7], extra two auxiliary linear softmax classifiers are added into the proposed network to promote the convergence of network. Following GoogLeNet, to control the tradeoff among three loss functions, a hyperparameter $\lambda \in [0, 1]$ is applied to the auxiliary loss functions. The loss function of GBNet combined with the global feature is formulated as follows:

$$\text{Loss} = \text{Loss}_{\text{conv}} + \lambda(\text{Loss}_{\text{aux1}} + \text{Loss}_{\text{aux2}}) \tag{14}$$

where $\text{Loss}_{\text{aux1}}$ and $\text{Loss}_{\text{aux2}}$ are also the cross entropy.

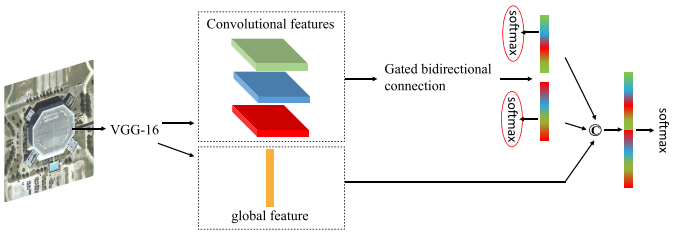

Fig. 8. GBNet combined with the global feature. As shown in the red circle, two auxiliary linear softmax functions are added to this network to promote network convergence.

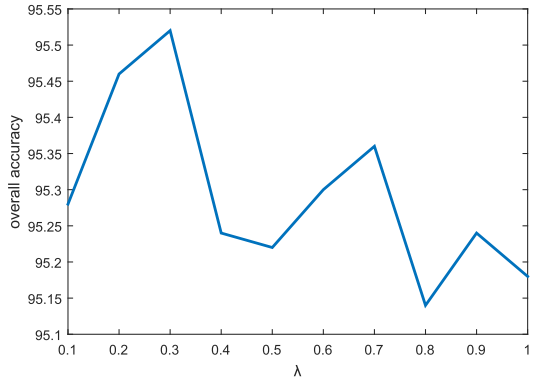

Fig. 9. OAs of GBNet combined with the global feature at different λ ratios on the AID data set.

TABLE III

OAs OF GBNET COMBINED THE GLOBAL FEATURE AND AUXILIARY CLASSIFIERS ON THE AID DATA SET (50% FOR TRAINING)

Method	Accuracy
GBNet + global feature	**95.52%**
GBNet + global feature (without auxiliary classifiers)	94.98%
GBNet	94.12%

In order to determine the value of hyperparameter λ, several experiments are conducted with λ ranging from 0.1 to 1. The experimental results are shown in Fig. 9. The hyperparameter λ is set to 0.3 for training the GBNet combined with the global feature. Furthermore, in order to show the performance of auxiliary linear softmax, the auxiliary linear softmax classifiers are removed to carry out experiments. As shown in Table III, the auxiliary softmax classifiers improve the OA by 0.54% effectively on the AID data set.

E. Comparison With the State of the Art

In this section, the performance of the proposed method with the combination of conv3-3, conv5-1, and conv5-3 is compared with some state-of-the-art methods on four public RS scene data sets. In Section V-F, the experiments demonstrate that the GBNet with the combination of conv3-3, conv5-1, and conv5-3 performs better than other combinations. The experiments are repeated five times and the average results are reported.

1) AID Data Set: The AID data set is a relatively large RS scene data set with at least 100 images for training models in each scene category. As shown in Section V-C, the proposed

TABLE IV

OAs ON THE AID DATA SET

Method	50% for training	20% for training
GoogLeNet [46]	86.39 ± 0.55%	83.44 ± 0.40%
CaffeNet [46]	89.53 ± 0.31%	86.86 ± 0.47%
VGG-16 [46]	89.64 ± 0.36%	86.59 ± 0.29%
MCNN [53]	91.80 ± 0.22%	-
Fusion by Addition [15]	91.87 ± 0.36%	-
ARCNet-VGGNet16 [43]	93.10 ± 0.55%	88.75 ± 0.40%
VGG-16 + MSCP [37]	94.42 ± 0.17%	91.52 ± 0.21%
Multilevel Fusion [14]	95.36 ± 0.22%	-
VGG-16 (fine-tuning)	93.60 ± 0.64%	89.49 ± 0.34%
GBNet	93.72 ± 0.34%	90.16 ± 0.24%
GBNet + global feature	**95.48 ± 0.12%**	**92.20 ± 0.23%**

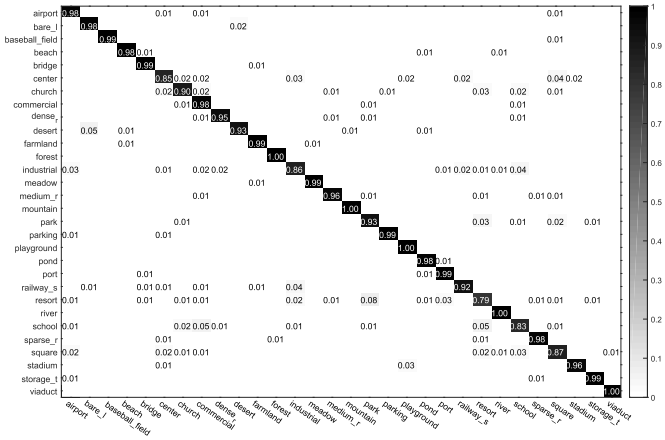

Fig. 10. CM of GBNet under the 50% training ratio on the AID data set. The AA is 95.21%.

gated bidirectional connection and skip connection are effective to aggregate multiple convolutional features. All compared methods are based on the deep convolutional networks, and most of the compared methods are based on the feature aggregation. As shown in Table IV, the proposed method achieves a 95.48% ± 0.12% and 92.20% ± 0.23% accuracy under the 50% and 20% training ratio, respectively. Especially, a multilevel fusion method [14] that integrated three different convolutional networks made a superior performance on the AID data set. The proposed GBNet is slightly higher than the multilevel fusion method by 0.12% percentage points under the 50% training ratio. ARCNet-VGGNet16 [43] is one of the latest RS scene classification methods. The proposed method achieves 2.38% and 3.45% points higher than ARCNet-VGGNet16 in terms of OA under the 50% and 20% training ratio, respectively. The CM of GBNet under the 50% training ratio is shown in Fig. 10. Each row of the CM represents the classification results of each RS scene category.

2) UC-Merced Data Set: The UC-Merced data set is a classic RS scene data set. As shown in Table V, the classification accuracy of ARCNet-VGGNet16 reaches 99.12% under the 80% training ratio. It is because the hyperparameters of ARCNet-VGGNet16 are optimally tuned on this UC-Merced data set. The proposed method achieves a 98.57%±0.48% and 97.05% ± 0.19% accuracy under the 80% and 50% training ratio, respectively, which can compete with the state-of-the-art

TABLE V
OAs ON THE UC-MERCED DATA SET

Method	80% for training	50% for training
GoogLeNet [46]	94.31 ± 0.89%	92.70 ± 0.60%
AlexNet [8]	95.00 ± 1.74%	-
CaffeNet [46]	95.02 ± 0.81%	93.98 ± 0.67%
VGG-16 [46]	95.21 ± 1.20%	94.14 ± 0.69%
VGG-F-ImageNet [8]	95.76 ± 1.70%	-
MCNN [53]	96.66 ± 0.90%	-
D-DSML-CaffeNet [54]	96.76 ± 0.36%	-
MDDC [44]	96.92 ± 0.57%	-
ResNet [8]	97.19 ± 0.57%	-
Fusion by Addition [15]	97.42 ± 1.79%	-
VGG-16 + EMR [9]	98.14%	-
VGG-16 + MSCP [37]	98.36 ± 0.58%	-
VGG-S + VGG-16 + IFK [28]	98.49%	-
VGG-16 + IFK [30]	98.57 ± 0.34%	-
ARCNet-VGGNet16 [43]	**99.12 ± 0.40%**	96.81 ± 0.14%
VGG-16 (fine-tuning)	97.14 ± 0.48%	96.57 ± 0.38%
GBNet	96.90 ± 0.23%	95.71 ± 0.19%
GBNet + global feature	98.57 ± 0.48%	**97.05 ± 0.19%**

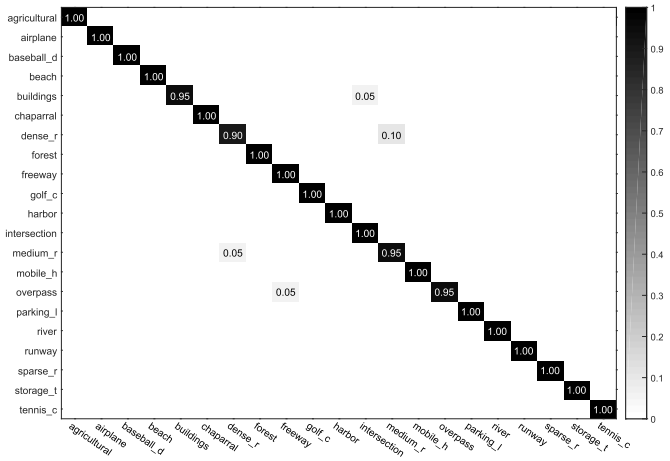

Fig. 11. CM of GBNet under the 80% training ratio on the UC-Merced data set. The AA is 98.81%.

methods. The CM of GBNet under the 80% training ratio is shown in Fig. 11.

3) WHU-RS19 Data Set: The WHU-RS19 is a relatively small RS scene data set. As shown in Table VI, the WHU-RS19 data set is less difficult than the other three data sets. The ARCNet-VGGNet16 [43] achieved a surprising 99.75% ± 0.25% accuracy under the 60% training ratio. ARCNet-VGGNet16 is a complex RS scene classification method, which is composed of a VGG-16 and a multilayer long short-term memory network [55]. It is quite complicated to classify scenes using two kinds of deep network structures. The proposed method, which is simply composed of a VGG-16 and a gated bidirectional connection, achieves a 99.25 ± 0.50% classification accuracy under the 60% training ratio. Our method has a slightly lower accuracy of 0.5% than the ARCNet-VGGNet16 method. In fact, the proposed GBNet misclassifies two more images than the ARCNet-VGGNet16 under the 60% training ratio. The CM of GBNet under the 60% training ratio is shown in Fig. 12.

4) OPTIMAL-31 Data Set: The OPTIMAL-31 data set is a new RS scene classification data set. The scenes contained

TABLE VI
OAs ON THE WHU-RS19 DATA SET

Method	60% for training	40% for training
GoogLeNet [46]	94.71 ± 1.33%	93.12 ± 0.82%
CaffeNet [46]	96.24 ± 0.56%	95.11 ± 1.20%
VGG-16 [46]	96.05 ± 0.91%	95.44 ± 0.60%
D-DSML-CaffeNet [54]	96.64 ± 0.68%	-
MDDC [44]	98.27 ± 0.53%	-
Fusion by Addition [15]	98.65 ± 0.43%	-
VGG-S + VGG-16 + IFK [28]	98.89%	-
ARCNet-VGGNet16 [43]	**99.75 ± 0.25%**	**97.50 ± 0.49%**
VGG-16 (fine-tuning)	96.88 ± 0.61%	96.74 ± 0.57%
GBNet	98.34 ± 0.40%	95.38 ± 0.26%
GBNet + global feature	99.25 ± 0.50%	97.32 ± 0.32%

TABLE VII
OAs ON THE OPTIMAL-31 DATA SET

Method	80% for training
Fine-tuning AlexNet [43]	81.22 ± 0.19%
Fine-tuning GoogLeNet [43]	82.57 ± 0.12%
ARCNet-AlexNet [43]	85.75 ± 0.35%
Fine-tuning VGGNet16 [43]	87.45 ± 0.45%
VGG-16 [46]	89.12 ± 0.35%
ARCNet-ResNet34 [43]	91.28 ± 0.45%
ARCNet-VGGNet16 [43]	92.70 ± 0.35%
VGG-16 (fine-tuning)	89.52 ± 0.26%
GBNet	91.40 ± 0.27%
GBNet + global feature	**93.28 ± 0.27%**

TABLE VIII
OAs OF DIFFERENT COMBINATIONS ON THE AID DATA SET (50% FOR TRAINING)

Method	Accuracy
GBNet (conv3-3, conv5-1, conv5-3) + global feature	**95.52%**
GBNet (conv4-1, conv5-1, conv5-3) + global feature	95.34%
GBNet (conv4-2, conv5-1, conv5-3) + global feature	95.22%

TABLE IX
PERFORMANCE OF DIFFERENT COMBINATIONS ON DIFFERENT DATA SETS

UC-Merced Dataset (80% for training)	
GBNet (conv3-3, conv5-1, conv5-3) + global feature	**99.05%**
GBNet (conv1-1, conv4-3, conv5-3) + global feature	97.86%
GBNet (conv2-2, conv5-1, conv5-3) + global feature	98.81%
GBNet (conv1-1, conv4-3, conv5-2) + global feature	97.62%
WHU-RS19 Dataset (60% for training)	
GBNet (conv3-3, conv5-1, conv5-3) + global feature	**99.75%**
GBNet (conv1-1, conv4-3, conv5-3) + global feature	98.24%
GBNet (conv2-2, conv5-1, conv5-3) + global feature	98.74%
GBNet (conv1-1, conv4-3, conv5-2) + global feature	98.49%
OPTIMAL-31 Dataset (80% for training)	
GBNet (conv3-3, conv5-1, conv5-3) + global feature	**93.55%**
GBNet (conv2-1, conv5-2, conv5-3) + global feature	93.01%
GBNet (conv2-1, conv4-2, conv5-3) + global feature	92.74%
GBNet (conv2-1, conv4-2, conv5-2) + global feature	92.74%

in this data set are very complicated. As shown in Table VII, the fine-tuned VGG-16 achieved only 89.52% ± 0.26% classification accuracy, which is much lower than on other RS scene data sets. The proposed GBNet combined with the global feature obtains a 93.28% ± 0.27% classification accuracy on the OPTIMAL-31 data set. The accuracy of the proposed method is 0.58% higher than that of the ARCNet-VGGNet16. The CM of GBNet is shown in Fig. 13. The proposed GBNet

TABLE X
TRAINING AND TESTING TIME OF DIFFERENT NETWORKS ON FOUR RS SCENE DATA SETS

	VGG-16		GBNet		GBNet + global feature	
	Train.(m)	Test.(s)	Train.(m)	Test.(s)	Train.(m)	Test.(s)
AID (50% for training)	786.56	62.18	801.78	62.81	860.56	74.37
AID (20% for training)	314.78	99.57	320.22	100.79	343.89	119.23
UC-Merced (80% for training)	263.92	5.08	269.40	5.31	289.22	6.46
UC-Merced (50% for training)	165.26	12.63	167.93	13.25	180.84	16.17
WHU-RS19 (60% for training)	95.33	5.20	97.42	5.28	104.52	6.38
WHU-RS19 (40% for training)	64.05	7.83	65.46	7.92	70.27	9.55
OPTIMAL-31 (80% for training)	234.04	4.56	238.78	4.67	256.15	6.39

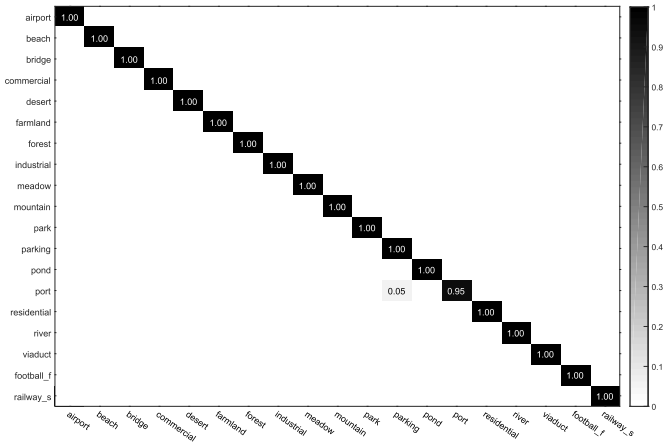

Fig. 12. CM of GBNet under the 60% training ratio on the WHU-RS19 data set. The AA is 99.75%.

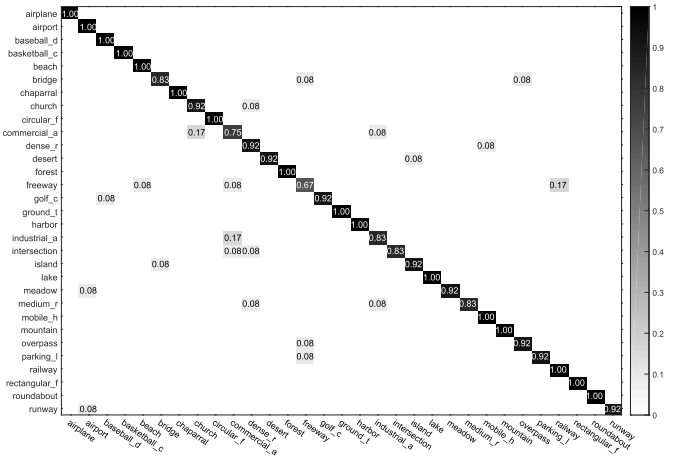

Fig. 13. CM of GBNet under the 80% training ratio on the OPTIMAL-31 data set. The AA is 93.55%.

performs poorly in the freeway and commercial area scenes, where the size and angle of the land-cover units are very variable.

F. Different Combinations of Convolutional Features

In this section, the performance of different feature combinations is discussed. First, as shown in Fig. 4 and Table I, classification accuracies under the other combinations achieve very close to the accuracy of conv3-3, conv5-1, and conv5-3 on the AID data set. The performance of other two combinations is explored in the proposed method. As shown in Table VIII, the GBNet + global feature with the combination of conv3-3, conv5-1, and conv5-3 achieves the best classification accuracy on the AID data set.

Second, the same as Section III, the feature combinations are selected on each data set. The performance of GBNet+global feature with different combinations on the UC-Merced, WHU-RS19, and OPTIMAL-31 data sets is reported in Table IX. As shown in Table IX, the top three combinations of the UC-Merced, WHU-RS19, and OPTIMAL-31 data sets perform worse than the combination of conv3-3, conv5-1, and conv5-3. This is because the top three combinations on the UC-Merced, WHU-RS19, and OPTIMAL-31 data sets contain the shallow convolution features (conv1-1, conv1-2, conv2-1, or conv2-2). The sizes of conv1-1 and conv1-2 are 224×224. The sizes of conv2-1 and conv2-2 are 112×112. In the proposed GBNet, the sizes of different convolution features need to be unified to 14×14 by the average pooling operation. The size of the shallow convolutional features is too large. In the process of unifying feature sizes, too much information of shallow convolution features is lost. The proposed GBNet cannot fully exploit the shallow convolutional features for RS scene classification.

G. Training and Testing Time

The training and testing time can directly reflect the computational efficiency of the proposed method. VGG-16 is the backbone of GBNet and GBNet + global features. Therefore, VGG-16 is employed as a baseline to analyze the time-computing of GBNet and GBNet + global feature. The training and testing time is reported in Table X. All networks are trained with 200 epochs on each data set. As shown in Table X, the training time of GBNet is very close to that of VGG-16. VGG-16 for RS scene classification contains approximately 134 million weight parameters, most of which are introduced by the fully connected layers. In the GBNet, the fully connected layers of VGG-16 are removed. The number of weight parameters included in GBNet is approximately 18 million. The weight parameters of GBNet are less than those of VGG-16. However, the training time of GBNet is more than that of VGG-16. This is because, in the CAFFE toolbox, the computational efficiency of the fully connected layer is very high. However, the computational efficiency of the convolutional layer is relatively low. The GBNet introduces several 1×1 convolutional layers in the gated bidirectional connection, which reduces the computational efficiency. In GBNet+global feature, the fully connected

layers of VGG-16 are utilized to extract the global feature. GBNet + global feature contains approximately 138 million weight parameters. Therefore, training GBNet + global feature takes more time than training VGG-16.

VI. Conclusion

In this paper, a GBNet is proposed for the RS scene classification. The proposed method focuses on hierarchically aggregating the complementary information and eliminating the interference information among different convolutional features. The experiments demonstrate that the proposed method can compete with the state-of-the-art methods. However, in the process of unifying feature sizes, too much information of shallow convolutional features is lost. The proposed GBNet cannot fully exploit the performance of shallow convolutional features. In further works, we will focus on capturing the complementary information from convolutional features of arbitrary size.

References

[1] J. Li, J. A. Benediktsson, B. Zhang, T. Yang, and A. Plaza, "Spatial technology and social media in remote sensing: A survey," *Proc. IEEE*, vol. 105, no. 10, pp. 1855–1864, Oct. 2017.

[2] F. Luo, H. Huang, Y. Duan, J. Liu, and Y. Liao, "Local geometric structure feature for dimensionality reduction of hyperspectral imagery," *Remote Sens.*, vol. 9, no. 8, p. 790, Aug. 2017.

[3] M. Chen, D. Liu, K. Qian, J. Li, M. Lei, and Y. Zhou, "Lunar crater detection based on terrain analysis and mathematical morphology methods using digital elevation models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 3681–3692, Jul. 2018.

[4] O. E. Dai, B. Demir, B. Sankur, and L. Bruzzone, "A novel system for content-based retrieval of single and multi-label high-dimensional remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2473–2490, Jul. 2018.

[5] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo, and W. Min, "Remote sensing image retrieval using convolutional neural network features and weighted distance," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 10, pp. 1535–1539, Oct. 2018.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2012, pp. 1097–1105.

[7] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[8] Y. Liang, S. T. Monteiro, and E. S. Saber, "Transfer learning for high resolution aerial image classification," in *Proc. IEEE Appl. Imagery Pattern Recognit. Workshop*, Oct. 2016, pp. 1–8.

[9] G. Wang, B. Fan, S. Xiang, and C. Pan, "Aggregating rich hierarchical features for scene classification in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4104–4115, Sep. 2017.

[10] S. Yang and D. Ramanan, "Multi-scale recognition with DAG-CNNs," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1215–1223.

[11] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "ActionVLAD: Learning spatio-temporal aggregation for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3165–3174.

[12] 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, Mar. 2016.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 936–944.

[14] 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.

[15] 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.

[16] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *Proc. Eur. Conf. Comput. Vis. Workshops*, Sep. 2016, pp. 685–701.

[17] P. Tang, X. Wang, B. Shi, X. Bai, W. Liu, and Z. Tu, "Deep FisherNet for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 2244–2250, Jul. 2019.

[18] F. Shi, E. Petriu, and R. Laganière, "Sampling strategies for real-time action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2595–2602.

[19] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2010, pp. 270–279.

[20] 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.

[21] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 911–923, Feb. 2019.

[22] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[23] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2015, pp. 1–13.

[25] 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.

[26] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019.

[27] W. Zhao and S. Du, "Scene classification using multi-scale deeply described visual words," *Int. J. Remote Sens.*, vol. 37, no. 17, pp. 4119–4131, Sep. 2016.

[28] 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 Sens.*, vol. 7, no. 11, pp. 14680–14707, Jan. 2015.

[29] J. Deng, W. Dong, 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. 248–255.

[30] 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.

[31] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2010, pp. 143–156.

[32] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.

[33] X. Lu, W. Ji, X. Li, and X. Zheng, "Bidirectional adaptive feature fusion for remote sensing scene classification," *Neurocomputing*, vol. 328, pp. 135–146, Feb. 2019.

[34] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "Ron: Reverse connection with objectness prior networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5244–5252.

[35] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37.

[36] R. Negrel, D. Picard, and P.-H. Gosselin, "Evaluation of second-order visual features for land-use classification," in *Proc. 12th Int. Workshop Content-Based Multimedia Indexing*, Jun. 2014, pp. 1–5.

[37] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899–6910, Dec. 2018.

[38] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 9, pp. 1984–1996, Sep. 2016.

[39] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, to be published.

[40] M. Hayat, S. H. Khan, M. Bennamoun, and S. An, "A spatial layout and scale invariant feature representation for indoor scene classification," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4829–4841, Oct. 2016.
[41] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.
[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[43] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.
[44] K. Qi, C. Yang, Q. Guan, H. Wu, and J. Gong, "A multiscale deeply described correlatons-based model for land-use scene classification," *Remote Sens.*, vol. 9, no. 9, pp. 917–933, Sep. 2017.
[45] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[46] G.-S. Xia *et al.*, "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.
[47] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1437–1451, Jun. 2018.
[48] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1933–1941.
[49] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "MARTA GANs: Unsupervised representation learning for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2092–2096, Nov. 2017.
[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] A. Yuan, X. Li, and X. Lu, "3G structure for image caption generation," *Neurocomputing*, vol. 330, pp. 17–28, Feb. 2019.
[52] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 173–176, Jan. 2011.
[53] Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 12, pp. 7109–7121, Dec. 2018.
[54] Z. Gong, P. Zhong, Y. Yu, and W. Hu, "Diversity-promoting deep structural metric learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 371–390, Jan. 2018.
[55] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

Hao Sun is currently pursuing the Ph.D. degree with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, and the University of Chinese Academy of Sciences, Beijing, China.

His research interests include pattern recognition, computer vision, and machine learning.

Siyuan Li is currently with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, the University of Chinese Academy of Sciences, Beijing, China, and Xi'an Jiaotong University, Xi'an, China. His research interests include pattern recognition, machine learning, hyperspectral image analysis, and medical imaging.

Xiangtao Zheng received the M.Sc. and Ph.D. degrees in signal and information processing from the Chinese Academy of Sciences, Xi'an, China, in 2014 and 2017, respectively.

He is currently an Assistant Professor with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His research interests include computer vision and pattern recognition.

Xiaoqiang Lu (M'14–SM'15) is currently a Full Professor with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His research interests include pattern recognition, machine learning, hyperspectral image analysis, cellular automata, and medical imaging.