First and Second-order Information Fusion Networks for Remote Sensing Scene Classification

Erzhu Li, Alim Samat, *Member*, *IEEE*, Ce Zhang, Peijun Du, *Senior Member*, *IEEE* and Wei Liu

Abstract— **Deep convolutional networks have been the most competitive method in remote sensing scene classification. Due to the diversity and complexity of scene content, remote sensing scene classification still remains a challenging task. Recently, the second-order pooling method has attracted more interest because it can learn higher-order information and enhance the non-linear modeling ability of the networks. However, how to effectively learn second-order features and establish the discriminative feature representation of holistic images is still an open question. In this Letter, we propose a first and second-order information fusion networks (FSoI-Net) that can learn the first-order and second-order features at the same time, and construct the final feature representation by fusing the two types of features. Specifically, a self-attention-based second-order pooling (SaSoP) method based on covariance matrix is proposed to extract second-order features, and a fusion loss function is developed to jointly train the model and construct the final feature representation for the classification decision. The proposed networks have been thoroughly evaluated on three real remote sensing scene datasets and achieved better performance than the counterparts.**

Index Terms— **Deep learning, second-order pooling, self-attention mechanism, information fusion, scene classification.**

I. Introduction

SCENE classification is a classic research content of computer vision. Its purpose is to divide images into different categories according to their content [1]. Thus, scene classification provides an alternative solution to complete the land use and land cover (LULC) classification task using high-spatial-resolution (HSR) imagery. In this scheme, the basic processing unit becomes the scene image instead of pixels. Therefore, it is suitable to process HSR images with abundant spatial and structural patterns but few spectral features. In the past few years, scene classification has attracted wide attention in the field of remote sensing, and has been applied in different real-world applications, such as LULC classification [2], semantic annotation [3]. However, due to some common phenomena in remote sensing scene images, such as complex structure and diverse scales, it is still a challenge to classify remote sensing scenes with accuracy.

The core of scene classification is to build robust feature representations of images. The rise of deep learning has brought a milestone breakthrough in computer vision, and this has a transformative impact on remote sensing scene classification. Especially for convolutional neural networks (CNN), it has achieved impressive performance in various applications. Inspired by ImageNet Large Scale Visual Recognition Challenge (ILSVRC) , many state-of-the-art CNN models have been designed to challenge the object classification and detection on hundreds of categories and millions of images [4]. These models based on constantly improving network architectures yielded more accurate results for image recognition, and they are also widely used in the field of remote sensing image processing. Due to the limited training samples available for remote sensing scene classification, many pre-trained CNN models instead of fully trained new models from scratch are used as feature extractors or fine-tuned to complete the classification tasks [5]. Some research also attempted to use few-shot learning to address the issue of insufficient samples [6]. To handle the specific characteristics of remote sensing scenes, such as complex structure, various scales, and irregular rotation and scaling, some research has focused on extracting multi-scale features and employing order-less feature coding methods to overcome these issues [7]. However, such kind of methods often involve multiple processes, and cannot be trained end-to-end. This inevitably increases the computational complexity of the algorithm. Moreover, the pattern that separates the feature extraction and classification processes is challenging to apply to large-scale remote sensing image classification as a big data problem in remote sensing community. Therefore, it is crucial and of necessity to develop advanced network architecture to address remote sensing scene

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant No. 41801327 and 41071424, The Youth Innovation Promotion Association Foundation of the Chinese Academy of Sciences under Grant No. 2018476, and The Xuzhou Science and Technology Key R&D Program (Social Development) under Project KC20172 (*Corresponding author: Erzhu Li; Ce Zhang*).

E. Li and W. Liu is with School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou 221116, China (e-mail: lierzhu2008@126.com; liuw@jsnu.edu.cn).

A. Samat is with State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, CAS, Urumqi 830011, China (e-mail: alim_smt@ms.xjb.ac.cn).

C. Zhang is with Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, and UK Centre for Ecology & Hydrology, Library Avenue, Lancaster LA1 4AP, UK (e-mail: c.zhang9@lancaster.ac.uk).

P. Du is with Department of Geographical Information Science, Nanjing University, Nanjing 210023, China (e-mail: dupjrs@126.com).

classification task in an intelligent fashion.

Recently, the construction of deep neural networks that can learn higher-order statistical features for image representations has attracted significant interest. Several second-order pooling methods are proposed to capture the second-order information of the final convolutional features. The pioneering work was bilinear CNN with excellent performance in fine-grained image recognition [8]. However, the bilinear pooling method produced very high-dimensional features, which led to a huge increase in model parameters. Another idea for modeling second-order features is to use covariance matrix of the final convolutional features. Some second-order pooling methods based on covariance matrix have been developed and achieved state-of-the-art results in various vision tasks, such as object recognition [9], action recognition[10] and fine-grained recognition [11]. For remote sensing scene classification, a few studies have also attempted to build a second-order pooling CNN model based on the covariance matrix, which directly uses the upper triangle elements of the covariance matrix as second-order features [12-14]. These methods obtained better results compared with the baseline CNN models. However, they still suffer from high pooling feature dimension and insufficient use of convolutional feature information. Although the second-order pooling CNN has successful in different tasks, many existing methods are proposed only for specific domains. How to effectively build higher-order feature representations for remote sensing scenes is still an open problem. Besides, existing methods pay too much attention to the construction of higher-order features, but ignore the comprehensive utilization of various features.

In this Letter, we pay much attention to the use of both first and second-order information to construct the feature representation of the remote sensing scene images. For this purpose, a self-attention-based second-order pooling (SaSoP) method based on covariance matrix is proposed to construct second-order features with low feature dimensions. Thereafter, a fusion cross-entropy loss is developed to train the first and second-order information fusion networks (FSoI-Net). This model can be used as an independent block plugged at the end of a network and trained end-to-end with entire network.

II. Proposed Method

A. Position-wise Attention Mechanism

For remote sensing scene images, many land use categories have complex structures such as industrial, school, commercial. Besides, they are extracted from different images covering a variety of areas with distinct spatial resolution, scale, rotation and scene size. Therefore, some key features that are invariant to scale, rotation and size play a pivotal role in identifying the category of scene images. To highlight these important features, and to integrate the importance of these features when constructing the covariance matrix, the position-wise attention features are designed based on attention mechanism, and then these features are used to construct the covariance matrix. As shown in Figure 1, the convolutional features extracted by the CNN backbone layers have firstly reduced the dimension using 1×1 convolution kernel. After the low-dimensional convolution features are obtained, a position-wise attention weight map is calculated based on the low-dimensional convolution features to emphasize the importance of depth features at each location for representing the scene image. Specifically, for convolutional features after dimension reduction $\mathbf{Z} \in \mathbb{R}^{h\times w\times c}$, we first aggregate channel information of $\mathbf{Z}$ across the channel using average-pooling and max-pooling operations, and acquire two different descriptors: $\mathbf{Z}_{avg} \in \mathbb{R}^{h\times w\times 1}$ and $\mathbf{Z}_{max} \in \mathbb{R}^{h\times w\times 1}$, representing the importance of each position of the convolutional features respectively. The descriptors are then concatenated and convolved by a convolution operation with a size of 3×3, generating a 2D map. Finally, the sigmoid activation function is followed to obtain the weight of different positions. The position-wise attention weight map is derived as:

$$\mathbf{S} = \sigma(\mathcal{F}(\mathbf{Z}_{avg}; \mathbf{Z}_{max})) \tag{1}$$

Where $\mathbf{S}$ is the weight map, $\mathcal{F}$ denotes a convolution operation, and σ represents the sigmoid function. The position-wise attention weight map describes the spatial distribution of important features. To use this information for final classification decision, the weight map is used to weight the convolutional features by position via a Hadamard product, which is expressed as follows:

$$\mathbf{Z}_{out} = F_{Hp}(\mathbf{Z}, \mathbf{S}) = \mathbf{X} \odot \mathbf{S} \tag{2}$$

Where Z_{out} is the weighted convolutional features.

B. Self-attention-based Second-order Pooling Block

The traditional CNN models usually extract the first-order statistical convolutional features as image representations right at the end of deep networks, such as ResNet [15]. To better characterize complex categories and build discriminative image representations, one solution is to learn high-order feature representations to enhance nonlinear modeling capabilities of the CNN model [16]. In this study, we propose a second-order pooling method based on self-attention mechanism. When a pre-trained CNN model is selected as the backbone, the feature map outputs at the final convolutional layer can be used as the input tensor of the SaSoP block. Such an input tensor can be expressed as $\mathbf{X} \in \mathbb{R}^{h\times w\times c}$, where h and w are spatial height and width and c is the number of channels of the convolutional features respectively. To compute the covariance matrix of $\mathbf{X}$, it is reshaped into a two-dimensional feature matrix $\mathbf{X}' \in \mathbb{R}^{d\times c}$, where $d = h \times w$. Thus, the covariance matrix can be calculated by equation (1).

$$\mathbf{Cov} = \mathbf{X}'^{\mathrm{T}}\, \hat{\mathbf{I}}\, \mathbf{X}' \tag{3}$$

$$\hat{\mathbf{I}} = \left(\frac{1}{d} - 1\right)\left(\mathbf{I} - \frac{1}{d}\mathbf{i}\,\mathbf{i}^{\mathrm{T}}\right) \tag{4}$$

Where $\mathbf{I}$ is a $d \times d$ identify matrix, and $\mathbf{i}$ is a d-dimensional column vector, all elements are set as 1.

The covariance matrix $\mathbf{Cov}$ is a symmetric matrix of size $c \times c$, which has a clear physical meaning. Each row of the covariance matrix represents the statistical correlation between one channel and all channels. Because it contains the mutual information between the convolutional features of different channels, higher-order information can be captured by the covariance matrix. Nevertheless, if the covariance matrix is used directly as the final feature representation of the scene image for classification, it will have a very high dimensionality even if half of the elements of the symmetric matrix are selected. In this research, a one-dimensional group convolution process is performed on the covariance matrix in the row direction to generate learning features of the covariance matrix, and these learning features are used to represent the higher-order information in the covariance matrix. Specifically, for *j-th* row of the covariance matrix, a convolution kernel is defined as<$\mathbf{W}_j$, b_j>, where $\mathbf{W}_j \in \mathbb{R}^{c\times 1}$is the learnable weight, b_j is the bias value. Thus, the convolution operation of *j-th* row of the covariance matrix ($\mathbf{Cov}_{j,:}$) is defined as,

$$\mathbf{Cov}_j' = \mathbf{Cov}_{j,:} * \mathbf{W}_j + \mathrm{b}_j \tag{5}$$

After processing the convolutional operation on each row of the covariance matrix, the learning features $\mathbf{Cov}' = [\mathbf{Cov}_1', \mathbf{Cov}_2', \cdots, \mathbf{Cov}_\mathrm{c}']^T$ are obtained. $\mathbf{Cov}'$ is the feature representation of the covariance matrix, and contains higher-order information of the input tensor. Our proposed self-attention uses the learning features $\mathbf{Cov}'$ as the attention weights, and the output of our SaSoP block is aggregated via a Hadamard product according to the following form:

$$\mathbf{Y} = F_{Hp}(\mathbf{X}, \mathbf{Cov}') = \mathbf{X} \odot \sigma(\,\mathbf{Cov}') \tag{6}$$

Where $\mathbf{Y}$ is the output tensor that has the same size as $\mathbf{X}$, $\odot$ denotes the Hadamard product, and σ is the sigmoid function. Followed by this process, the global average pooling (GAvP) method is used to generate the final second-order feature representation

C. Fusion Loss Function

For common deep CNN models, the first-order statistics (i.e., mean vector or fully-connected features) extracted by the global average pooling (GAvP) or fully-connected layer are often used as the final image representations for classification. We propose to make full use of first and second-order information to describe the holistic image context, and the first and second-order features are learned together in our proposed module and determine the category of the input image jointly. Specifically, the first-order feature are summarized by GAvP, and the second-order feature are extracted by our proposed SaSoP block. In this case, a fusion loss function is established to fuse the first and second-order features and train the proposed model jointly. As shown in equation (7), the loss fusion contains three input variables x_1, x_2, y, representing the first-order features, second-order features and corresponding category label. It is a cross-entropy loss function with two inputs, where j is the sample index in each mini-batch.

$$\mathrm{loss}(x_1, x_2, y) = -\log\left(\frac{\exp(x_1[y])}{\sum_j \exp(x_1[j])}\right) - \log\left(\frac{\exp(x_2[y])}{\sum_j \exp(x_2[j])}\right) \tag{7}$$

As shown in Eq. (7), the first and second-order features share the same role in the model training process. The outputs of the two streams are fused through addition as the decision rule to decide the final category.

D. Networks Implementation

Summarized as Fig.1, the proposed convolutional network model consists of four parts: (a) CNN backbone architecture, as the backbone of the network to complete convolution feature extraction; (b) the first-order feature extractor based on GAvP layer; (c) the proposed position-wise attention SaSoP block to construct second-order feature representation for classification; (d) the fusion loss function integrates first and second-order features for joint training. Our work focuses on the final three steps to enhance the nonlinear representation ability of the CNN model. Thus, the pre-trained CNN models can be used as the backbone architecture in this study. Here, ResNet-50 and ResNet-101 [15] are employed as the backbone architecture of the proposed network. For this purpose, all layers after the final convolutional layer are removed to construct a convolution process module, which is used to extract the convolutional features of the input scene images. We reduce the number of channel (C) into 128 for both CNN backbone architectures by the 1 × 1 convolution operation. If the input size is changed, the output size of dimension reduction block, and position-wise attention will be changed accordingly.

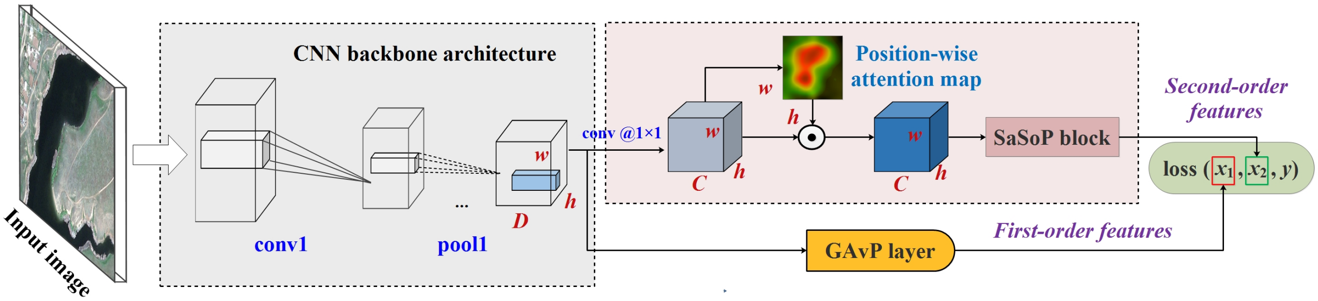

Fig.1. Example architecture of the first and second-order information fusion networks.

III. Experiments and Discussion

A. Experimental Datasets and Setup

In this study, three real remote sensing scene image data sets were used to test the proposed method. The first data is the UC Merced (UCM) data set [17], which is extracted from aerial remote sensing images. This data set is an early open data set containing 21 categories, each category contains 100 scene images with a fixed size of 256×256×3.

Another data set, the Aerial Image Dataset (AID) [5], was generated using satellite images from Google Earth. It is a large data set containing 100000 images with a size of 600×600×3, which were divided into 30 categories. In addition, the sample size of each category ranges from 220 to 420, and the spatial resolution of the scene images is varied, since they were captured by different satellite sensors.

The last data set is the NWPU-RESISC45 (NWPU) [18]. It is a large data set widely used as the large-scale benchmark data for testing deep learning methods. Similar to AID data set, this data set collected 31500 scene images (size of 256×256×3) from Google Earth, including 45 categories. Each category contains the same number of scene images in this data set.

For the proposed network, the size of the input image is flexible. In order to reduce the computational burden in model training, the images in UCM and NWPU data sets are adjusted to $224 \times 224 \times 3$, and the images in AID data set are adjusted to $299 \times 299 \times 3$. In the training phase, the networks are optimized using stochastic gradient decent (SGD) with a momentum of 0.9, a weight decay rate of 10^{-4} and a mini-batch of 32. The initial learning rate is set to 0.005, and the cosine annealing schedule is used to set the learning rate of each parameter group, and the maximum number of iterations is set to 10. Finally, the networks are trained with a total of 45 epochs. The networks were developed on PyTorch, and all the experiments were performed on a 64 bits Intel Xeon silver 4114 machine with 64GB RAM memory and a single GeForce GTX 1080ti GPU.

To evaluate the proposed networks, different training ratios (Tr) are set to test the models, and two metrics of overall accuracy (OA) and Kappa coefficient (κ) are used to measure the classification accuracy.

B. Experimental Results and Analysis

Two types of statistic features describing first and second-order information are considered in the constructed networks. When evaluating the fusion networks, other networks based on first-order or second-order features are also evaluated accordingly.

We first test the 50-layer and 101-layer plain nets (ResNet-50/101), as well as the second-order pooling networks (ResNet-50/101+SaSoP) and the fusion nets (ResNet-50/101+FSoI-Net1/2) on the UCM dataset. ResNet-50/101+FSoI-Net1 does not use the position-wise attention mechanism, and ResNet-50/101+FSoI-Net2 is the fused networks with the position-wise attention. The results listed in Table I show that ResNet-101 has higher classification accuracy than the shallower ResNet-50. When the second-order pooling method is used to model the holistic image, the performances have been further increased for ResNet-50+SaSoP with Tr=50%, while other results have declined. For the ResNet-50/101+FSoI-Net1, the first and second-order features are fused for holistic image modeling, the highest classification accuracy is obtained. This indicates that second-order pooling based on covariance matrix has the advantage of improving the modeling performance. In addition, we have two major observations from Table I. First, second-order features extracted by second-order pooling block can provide different statistical information, but they do not necessarily have obvious advantages compared with first-order features. Second, the fusion of first and second-order features has obvious superimposing advantages, which demonstrated that both first and second-order features contribute to enhancing the feature representation ability of images. When the position-wise attention block is inserted, the classification accuracy of the fusion networks is increased further. This shows the importance of position, which is helpful to enhance the representation capability of second-order features.

The results of experiments performed on the other two datasets are shown in Tables II and III, respectively. From the reports in Table II, in most cases, the plain net ResNet-50 inserted by the second-order pooling block has slightly higher accuracy than the original, but the result of ResNet-101 is the opposite. However, compared with first-order features or second-order features alone, the fusion networks FSoI-Net brings clear improvements in the experiments on the AID dataset. For the NWPU dataset, similar experimental results were obtained, as shown in Table III. These results suggest that the two kinds of features (first and second-order features) are complementary, and fusion of the two can improve the generalization capability of the networks.

TABLE I
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE UCM DATASET

Method	Tr=50%		Tr=80%	
	OA (%)	κ	OA (%)	κ
ResNet-50	97.59±0.65	0.975	98.97±0.77	0.989
ResNet-101	98.03±0.95	0.979	99.37±0.14	0.993
ResNet-50+SaSoP	97.84±0.57	0.977	98.81±0.71	0.988
ResNet-101+ SaSoP	97.68±0.55	0.976	98.73±0.27	0.987
ResNet-50+FSoI-Net1	98.67±0.17	0.986	99.68±0.27	0.997
ResNet-101+FSoI-Net1	98.92±0.31	0.989	99.37±0.27	0.993
ResNet-50+FSoI-Net2	98.51±0.40	0.984	**99.68±0.14**	**0.997**
ResNet-101+FSoI-Net2	**98.70±0.43**	**0.986**	99.60±0.12	0.996

TABLE II
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE AID DATASET

Method	Tr=20%		Tr=50%	
	OA (%)	κ	OA (%)	κ
ResNet-50	94.26±0.16	0.941	96.50±0.31	0.964
ResNet-101	94.73±0.44	0.945	96.68±0.10	0.966
ResNet-50+SaSoP	94.52±0.21	0.943	96.55±0.11	0.964
ResNet-101+ SaSoP	94.64±0.30	0.944	96.77±0.11	0.967
ResNet-50+FSoI-Net1	95.43±0.25	0.953	97.03±0.17	0.969
ResNet-101+FSoI-Net1	95.88±0.26	0.957	97.31±0.28	0.972
ResNet-50+FSoI-Net2	95.49±0.31	0.953	97.16±0.07	0.971
ResNet-101+FSoI-Net2	**95.91±0.10**	**0.958**	**97.42±0.10**	**0.973**

TABLE III
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE NWPU DATASET

Method	Tr=10%		Tr=20%	
	OA (%)	κ	OA (%)	κ
ResNet-50	91.00±0.15	0.908	93.68±0.25	0.935
ResNet-101	92.28±0.24	0.921	94.18±0.16	0.941
ResNet-50+SaSoP	91.32±0.34	0.911	93.77±0.15	0.963
ResNet-101+ SaSoP	91.67±0.38	0.915	94.01±0.06	0.939
ResNet-50+FSoI-Net1	92.38±0.22	0.942	94.38±0.10	0.943
ResNet-101+FSoI-Net1	92.83±0.13	0.927	94.74±0.10	0.946
ResNet-50+FSoI-Net2	92.57±0.10	0.924	94.40±0.21	0.943
ResNet-101+FSoI-Net2	**92.91±0.17**	**0.926**	**94.76±0.18**	**0.946**

TABLE IV
PARAMETER SIZE AND COMPUTATION COMPLEXITY COMPARISON AMONG DIFFERENT METHODS ON THE UCM DATASET

Method	Parameter size (MB)	Flops (G)
ResNet-50/101	23.55/42.54	4.11/7.83
ResNet-50/101+SaSoP	23.79/42.78	4.12/7.85
ResNet-50/101+FSoI-Net1	23.84/42.83	4.12/7.85
ResNet-50/101+FSoI-Net2	23.84/42.83	4.12/7.85

Table IV shows a comparison of the parameters and computational complexity of different models. The number of parameters of ResNet-50/101+SaSoP is comparable to that of the vanilla ResNet-50/101. The increased parameters of ResNet-50/101+SaSoP are mainly attributed to the group convolutional layer. The fusion models also slightly increase the number of parameters. For theoretical amount of floating point (Flops), the computation of ResNet-50/101+SaSoP increased slightly due to the increase of parameters, but there is no obvious difference compared with the fused models.

Here, we compare the FSoI-Net with several related and state-of-the-art methods, as shown in Table V. Siamese ResNet-50 is a metric learning method. The skip-connected covariance (SCCov) network[12] directly uses covariance matrix of different convolution features as the image representation. Other works include PANet50 [19], ResNet-101+EAM [20] and ACNet [21], all focusing on the self-attention-based fusion strategies to enhance feature representations, and have achieved competitive performance. Amongst these methods, SCCov, as a typical second-order pooling method, cannot achieve better performances than other methods. In contrast, our networks that fuse first and second-order information achieves better performance than other networks. This comprehensive comparison demonstrates that both first and second-order information can help improve the network performance, and the fusion of the two kinds of features can enhance the representational learning capability of deep networks.

TABLE V
ACCURACY COMPARISON OF OUR METHODS WITH OTHER MODELS

Method	OA (%)		
	UCM (Tr=80%)	AID (Tr=50%)	NWPU (Tr=20%)
Siamese ResNet-50 [22]	94.29	-	92.28
SCCov [12]	99.05±0.25	96.10±0.16	92.10±0.25
PANet50 [19]	99.21±0.18	97.05±0.30	92.61±0.25
ResNet-101+EAM [20]	99.21±0.26	97.06±0.19	94.29±0.09
ACNet [21]	99.76±0.10	95.38±0.29	92.42±0.16
ResNet-50+FSoI-Net2	**99.68±0.14**	97.16±0.07	94.40±0.21
ResNet-101+FSoI-Net2	99.60±0.12	**97.42±0.10**	**94.76±0.18**

IV. CONCLUSION

In this Letter, we explored the effectiveness of various feature representation methods to improve network performance. By exploiting the first and second-order information at the same time, the proposed model can learn more discriminative feature representations and achieve the highest classification accuracy compared with other methods. Besides, our numerical experiments revealed two useful findings. First, the first-order and second-order information are complementary, and both provide effective expression for image recognition. Second, the second-order information has nonlinear characteristics and can achieve high accuracy with few samples. Therefore, these findings suggest that fusing first and second-order information is a better solution to enhance the network performance.

REFERENCES

[1] P. Du, X. Bai, K. Tan *et al.*, "Advances of four machine learning methods for spatial data handling: A review," *Journal of Geovisualization Spatial Analysis,* vol. 4, pp. 1-25, 2020.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012, pp. 1097-1105.

[3] X. Yao, J. Han, G. Cheng *et al.*, "Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 54, no. 6, pp. 3660-3671, 2016.

[4] A. Sengupta, Y. Ye, R. Wang *et al.*, "Going deeper in spiking neural networks: VGG and residual architectures," vol. 13, pp. 95, 2019.

[5] G.-S. Xia, J. Hu, F. Hu *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 55, no. 7, pp. 3965-3981, 2017.

[6] L. Li, J. Han, X. Yao *et al.*, "DLA-MatchNet for Few-Shot Remote Sensing Image Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1-10, 2020.

[7] E. Li, J. Xia, P. Du *et al.*, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 55, no. 10, pp. 5653-5665, 2017.

[8] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1449-1457.

[9] P. Li, J. Xie, Q. Wang *et al.*, "Towards faster training of global covariance pooling networks by iterative matrix square root normalization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 947-955.

[10] A. Cherian, and S. Gould, "Second-order temporal pooling for action recognition," *International Journal of Computer Vision,* vol. 127, no. 4, pp. 340-362, 2019.

[11] Z. Gao, J. Xie, Q. Wang *et al.*, "Global second-order pooling convolutional networks." pp. 3024-3033.

[12] N. He, L. Fang, S. Li *et al.*, "Skip-connected covariance network for remote sensing scene classification," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 31, no. 5, pp. 1461-1474, 2019.

[13] N. He, L. Fang, S. Li *et al.*, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 56, no. 12, pp. 6899-6910, 2018.

[14] S. Akodad, S. Vilfroy, L. Bombrun *et al.*, "An ensemble learning approach for the classification of remote sensing scenes based on covariance pooling of CNN features," in 2019 27th European Signal Processing Conference (EUSIPCO), 2019, pp. 1-5.

[15] K. He, X. Zhang, S. Ren *et al.*, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[16] Z. Gao, J. Xie, Q. Wang *et al.*, "Global Second-Order Pooling Convolutional Networks," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3019-3028.

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

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

[19] D. Zhang, N. Li, and Q. Ye, "Positional Context Aggregation Network for Remote Sensing Scene Classification," *IEEE Geoscience and Remote Sensing Letters,* vol. 17, no. 6, pp. 943-947, 2020.

[20] Z. Zhao, J. Li, Z. Luo *et al.*, "Remote Sensing Image Scene Classification Based on an Enhanced Attention Module," *IEEE Geoscience and Remote Sensing Letters*, pp. 1-5, 2020.

[21] X. Tang, Q. Ma, X. Zhang *et al.*, "Attention Consistent Network for Remote Sensing Scene Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 14, pp. 2030-2045, 2021.

[22] X. Liu, Y. Zhou, J. Zhao *et al.*, "Siamese Convolutional Neural Networks for Remote Sensing Scene Classification," *IEEE Geoscience and Remote Sensing Letters,* vol. 16, no. 8, pp. 1200-1204, 2019.