Classification of Remotely Sensed Images Using an Ensemble of Improved Convolutional Network

Li Wang, Yanjiang Wang, Yaqian Zhao, and Baodi Liu, *Member, IEEE*

***Abstract*— In the last few years, the deep learning methods, especially the residual neural network, have achieved impressive performance in remote sensing image recognition tasks. However, there are still specific problems that need to be addressed. It is well known that the first several layers of the network provide much discriminative information, and the ResNet reduces the size of the feature map so quickly that it failed to fully learn the information beneficial to classification in the early stage. Second, insufficient labeling data in remote sensing database may easily lead to overfitting and affect the final classification accuracy. Third, the optimal results cannot be achieved by relying solely on transfer learning. To overcome the problems mentioned earlier, we propose an enhanced residual neural network (ERNet) to improve the classification performance on remote sensing images. We moderately broadened the first several layers of the network, changed the size of the convolution filters, and made it learn more information of image features. Second, we add dropout layer to each residual unit of the proposed network to improve the accuracy and generalization power of ERNet. Finally, an ensemble of learning methods based on ERNet was introduced to improve the classification performance by fusing features of other baseline methods. Extensive experimental results on several benchmark data sets of remote sensing images demonstrate the superior performance of our proposed algorithm.**

***Index Terms*— Deep learning, remote sensing, residual network, image recognition.**

I. Introduction

RECENTLY, deep learning methods have shown superior performance in many tasks, such as image annotation [1], object detection [2], image classification [3], image reranking [4], and scene change detection [5]. Among them, the classification of remote sensing images [6] has also shown significant development prospects, attracting more and more attention from scholars.

Typically, there are two main classification methods for remote sensing images: one is based on the traditional machine learning method, and the other one is based on deep learning.

Traditional machine learning methods for remote sensing image classification mainly involve two stages: the feature extraction stage and the classification stage. An unsupervised feature learning framework for scene classification was presented by Zhang *et al.* [7], which is used to learn a set of robust and efficient feature extractors for scene classification. Lazebnik *et al.* [8] proposed an approximate global geometric correspondence method, which divides the images into many fine subregions and calculates the histograms of local features to represent the image features. However, relying only on experience or simple low-level features, traditional machine learning methods have limited ability to interpret the meaning of the scene.

The deep learning method has overcome the abovementioned problems [9], [10]. Many scholars use convolutional neural networks (CNN) as a feature extraction tool. Penatti *et al.* [11] proposed a classification framework based on transfer learning to verify the generalization power of CNN features. Liu *et al.* [12] extracted the features from pretrained CNN models and builds the collaborative-representation-based classifier for scene classification. However, extracting image features directly by using a pretrained model may not necessarily yield the best classification results, which depends on the similarity between the original data set and the target data set. Many researchers retrained the remote sensing data set by slightly optimizing the structure of neural network. Lin *et al.* [13] put forward a multiple-layer feature-matching generative adversarial network. Lu *et al.* [14] trained a shallow weighted deconvolution network to capture the abundant edge and texture information of high spatial resolution image. Castelluccio [15] employed pretrained CNNs and fine-tuned them on the scene data sets. In order to further improve the classification performance, researchers proposed methods that extracted the features from different CNNs to further improve the accuracy of remote sensing image classification. Yu *et al.* [16] proposed a convolutional feature fusion network to formulate an effective CNN representation of remote sensing images. Ji *et al.* [17] proposed a method of utilizing the attention network to localize multiscale discriminative regions of the scene images and combining features learned from the localized regions by a classification network. Nevertheless, how to design a more efficient classifier to fuse features is still a challenge.

Manuscript received April 21, 2019; revised August 15, 2019 and December 17, 2019; accepted April 8, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61671480, in part by the Natural Science Foundation of Shandong Province under Grant ZR2018MF017 and Grant ZR2018BF011, and in part by the Shandong Provincial Key Research and Development Program under Grant 2019JZZY011101. *(Corresponding author: Yanjiang Wang.)*

Li Wang is with the College of Control Science and Engineering, China University of Petroleum (East China), Qingdao 266580, China, and also with Inspur Electronic Information Industry Co., Ltd., Jinan 250001, China (e-mail: li.wang.upc@foxmail.com).

Yanjiang Wang and Baodi Liu are with the College of Control Science and Engineering, China University of Petroleum (East China), Qingdao 266580, China (e-mail: yjwang@upc.edu.cn).

Yaqian Zhao is with Inspur Electronic Information Industry Co., Ltd, Jinan 250001, China (e-mail: zhaoyaqian@inspur.com).

Digital Object Identifier 10.1109/LGRS.2020.2988934

To overcome the drawbacks mentioned earlier, we propose an enhanced residual neural network (ERNet), to improve the classification performance on sensing images. First, because the first several network layers can provide richer discriminant information, we moderately broadened the first several layers of the network, changed the size of the convolution filters, and made it learn more image feature information. Second, we added the dropout layer to each residual unit of the proposed network to improve the generalization power of ERNet. Over high dropout ratio can easily lead to network underfitting, so we have lowered the dropout ratio to improve the classification performance of the proposed algorithm. Finally, an ensemble of learning methods based on ERNet was used in the test stage to improve the classification performance by fusing features of other baseline methods. The experimental results on several benchmark data sets demonstrate the superior performance of our proposed algorithm.

The rest of this letter is organized as follows. The proposed framework is formulated in Section II. The experiment settings and experimental results are given in Section III. Finally, we conclude this letter in Section IV.

II. Methodology

A. Network Architectures

ResNet [18] has achieved remarkable achievements in the field of image processing, but it still has certain defects that need to be addressed. The first few layers of the original ResNet-18 [18] adopt the structure of Conv-batch normalization (BN)-Rectified Linear Unit (ReLU)-MaxPooling, in which the convolution layer uses 7×7 filters with a stride of 2, and the max-pooling layer downsampled the feature map. It is well known that first several layers of the network provide much discriminative information, and the ResNet-18 reduces the size of the feature map so quickly in the early stage that it cannot thoroughly learn the information beneficial to classification.

Therefore, in this letter, we propose an enhanced residual network model to improve the classification performance. Usually, the strategy is to increase the depth and width of the network, but an excessive increase of the depth and width of the network will greatly increase the network parameters, which tends to cause overfitting, especially when the training data set is small. Therefore, we moderately increase the width of the network first. The improved network structure is shown in Fig. 1, ERNet uses the two-channel convolution architecture in the initial network layers.

The first convolution unit uses 3×3 and 7×7 convolution filters to extract the image features, the Conv-BN-ReLU network structure is used for both channels, and the data of the last two channels are superimposed. The second convolution unit also has a two-channel structure. The first channel uses the Conv-BN-ReLU structure with 3×3 convolution filters. The second channel directly uses the max-pooling structure to obtain the image scale-invariant information. Finally, the feature maps of the two channels are added together. This design can prevent the explosive demand for computing resources with the increase of network layers. At the same time, it can also improve the identification performance of the network.

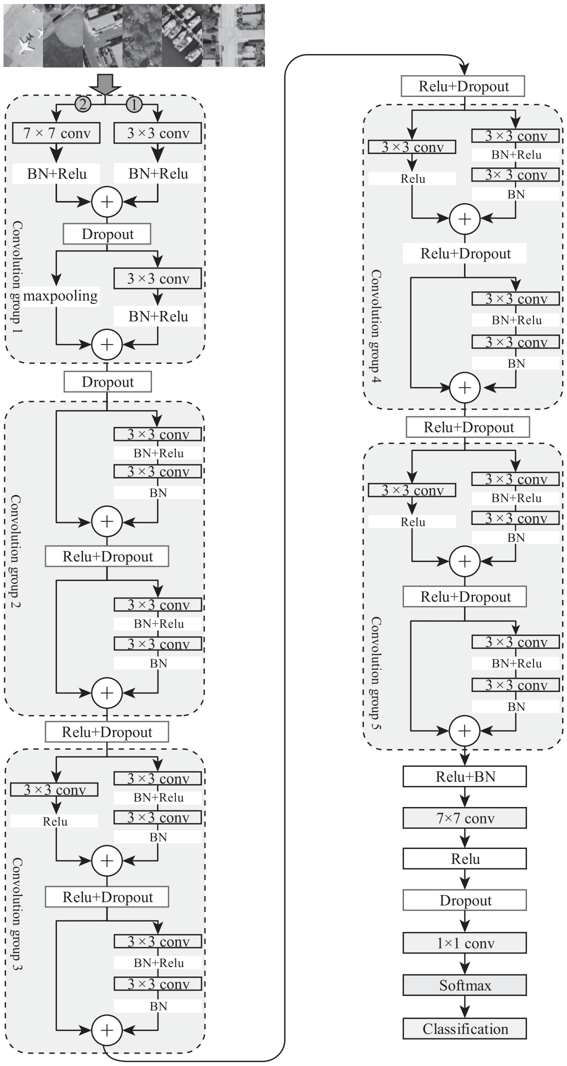


Fig. 1. ERNet structure that contains five convolution groups with more dropout layers.

Second, in order to prevent overfitting, the dropout layer is added to each residual unit in ERNet.

By applying dropout to all convolutional layers with a lower dropout ratio, the accuracy and generalization power can be improved in a certain degree.

The other convolution groups adopt the same network structure as the original ResNet. The softmax function is used on the output layer to complete the multiclassification task of remote sensing image, and the objective function built in this letter is as follows:

$$J = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} 1\{y^{(i)} = j\} \log\left(Y_j^{(i)}\right) \tag{1}$$

where i represents the ith learning sample, K represents the category of train samples, and $Y_j^{(i)}$ represents the output result of softmax layer.

B. Ensemble Learning

Multiclassifier integration can help reduce the risk of overfitting and improve the generalization ability of classifiers. In this letter, we use the ensemble learning method based on

Algorithm 1 Framework of Ensemble Learning for Our System

```
Input: Training data set $D_{tr} \in \mathbb{R}^{d \times N}$, and the test sample $x_{\mathrm{val}}$
Train $D_{tr}$ with the baseline method;
Train $D_{tr}$ with ERNet;
Extract corresponding features, and integrate into the ensemble learning training data set $X_{\mathrm{es}} \in \mathbb{R}^{m \times N}$
for $t = 1$; $t \leq T$; $t{+}{+}$ do
    Conduct random sampling of $X_{\mathrm{es}}$ with replacement, and obtain a data set $D_{\mathrm{es}}^{t} \in \mathbb{R}^{m \times k}$ containing $k$ samples.
    Use the SVM algorithm to train a base classifier $C_t$ on the $D_{\mathrm{es}}^{t}$.
    Place the base classifier $C_t$ in current ensemble learning system:$EL = EL \cup C_t$.
end for
Classifying $x_{\mathrm{val}}$ using base classifiers:
    $EL(x_{\mathrm{val}}) = \{C_1(x_{\mathrm{val}}), \cdots C_t(x_{\mathrm{val}}) \cdot \cdot, C_T(x_{\mathrm{val}})\}$
Combine the output results of each base classifier:
    $y_{\mathrm{va}l} = \arg\max_{\mathrm{id}} \sum_{t=1}^{T} I(C_t(x_{\mathrm{val}}))$
return $y_{\mathrm{va}l}$
```

bagging [19] to train the robust classifier and improve the final classification performance.

First, we use the trained networks to extract the features of the last convolution layers and integrate the corresponding features into a new matrix, X_{es} represents the data set of all integrated eigenvectors, and $X_{\mathrm{es}} \in \mathbb{R}^{m \times N}$, with m the dimension of the features and N the number of ensemble learning training samples.

Conduct random sampling of X_{es} with replacement. Extract k samples in total to form a new data set D_{es}, in which $D_{\mathrm{es}} \in \mathbb{R}^{m \times k}$. Through random sampling for T times, obtain T different data sets: $D_{\mathrm{es}}^{1}, D_{\mathrm{es}}^{2}, \ldots, D_{\mathrm{es}}^{t}, \ldots, D_{\mathrm{es}}^{T}$. Also, we use the support vector machine (SVM) method to train the classifier of each data set to obtain T diverse classifiers $EL = \{C_1, C_t, \ldots, C_T\}$.

In the process of testing, the simple majority voting rule is adopted to combine all classification results of EL system, and the final classification results are expressed by the following formula:

$$y_{\mathrm{va}l} = \arg\max_{\mathrm{id}} \sum_{t=1}^{T} I(C_t(x_{\mathrm{val}})) \tag{2}$$

where x_{val} represents the test sample, T represents the number of ensemble learning classifiers, $C_t(x_{\mathrm{val}})$ represents the classification result of x_{val} in the tth ensemble learning classifier, I represents a vector, which is 1 at the predicted label position and 0 at the rest. $\arg\max_{\mathrm{id}}$ represents the index of the maximum value of I.

Different networks can learn different characteristics of images, which present both similarities and differences. The stronger the complementarity of different features, the better the performance of feature classification after fusion.

III. Experiments

In this section, we show our experimental results on two remote sensing image data sets. To illustrate the efficiency of our approach, we compare the performance of our method with that of several state-of-the-art methods.

TABLE I

Classification Rate of ERNet and Baseline Methods in the UC-Merced Data Set (%)

Models\Datasets	Accuracy	FLOPS	Params
Inception-v2	95.58	1.94B [22]	11.2M [22]
Squeezenet	95.34	1.7B [23]	1.25M [23]
ResNet-18	95.91	1.8B	11.74M
ERNet-w/o	96.02	1.92B	11.76M
ERNet-part	96.14	1.92B	11.76M
ERNet	96.23	1.92B	11.76M
ResNet-18 + Inception-v2	96.64	–	–
ResNet-18 + Squeezenet	96.42	–	–
Squeezenet + Inception-v2	95.9	–	–
ERNet + Squeezenet	96.47	–	–
ERNet + Inception-v2	97.02	–	–

A. Experiment on UC Merced Land Use Data Set

In this letter, we use the UC-Merced data set [20], which consists of the high-resolution remote sensing images of 21 distinctive land-use classes. Each category contains 100 images of 256 × 256 pixels in size.

In order to achieve better classification results, first, we used the ImageNet database [21] to pretrain the ERNet, which can obtain better classification results in transfer learning. We used the data augmentation technique to increase the size of the UC-Merced data set, which can improve the recognition performance to some extent. The samples were flipped horizontally and then rotated by 90°, 180°, and 270°. The training set was extended by eightfold with these data augmentation methods. We randomly chose 60 images per class as training samples and 40 images per class as testing samples. To eliminate randomness, we randomly split the data set into a train set and test set for ten times.

We trained the ERNet with a mini-batch size of 32, a momentum of 0.9, and a weight decay of 0.005. We used a learning rate of 0.001 and decayed every two epochs by using an exponential rate of 0.9. All the experiments were carried out on Nvidia GTX1080TI with 11-GB memory. Several baseline methods, including Inception-V2, Squeezenet, and ResNet-18, were used as the benchmarks. The experimental results are shown in Table I.

From Table I, we can see that ERNet achieves a recognition accuracy of 96.23%, which is 0.32 higher than the accuracy obtained with ResNet-18. ERNet improves the classification performance by extracting richer features of the early layers. We can also see that the ensemble learning method can improve the performance of the original classification method, and ERNet combined with Inception-V2 classification method achieved 97.02% recognition rate. We also verified the role of dropout layers in ERNet. ERNet-w/o represents that only the dropout layer before the last convolutional layer is kept. ERNet-part represents that the dropout layers between the residual units are added based on the previous step. ERNet-w/o

TABLE II

EXPERIMENT ON THE UC-MERCED DATA SET (%)

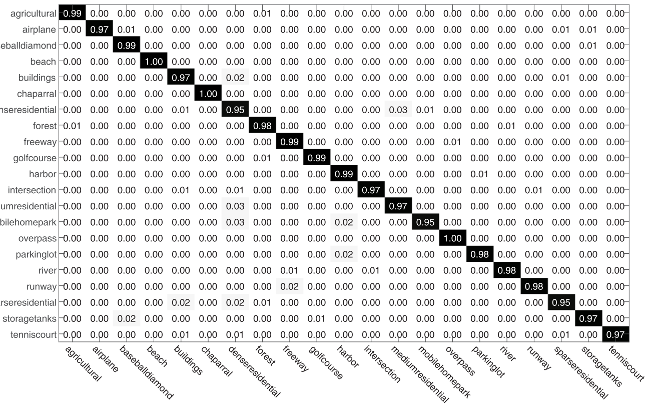
Methods	Year	Accuracy
CaffeNet [24] + VLAD	2015	95.39
MAGANs [13]	2017	87.69
WDM [14]	2017	95.71
UCFFN [16]	2018	87.83
CNN-W + VLAD with SVM [25]	2018	95.61
CNN-R + VLAD with SVM [25]	2018	95.85
VGG19 + liblinear [12]	2019	95.05
VGG19 + CRC [12]	2019	94.67
VGG19 + CS-CRC [12]	2019	95.26
ResNet+ CRC [12]	2019	96.9
ResNet-18		96.78
Inception-V2		96.51
Squeezenet		96.24
ERNet		96.95
Squeezenet + Inception-V2		96.67
Squeezenet + ResNet-18		96.89
Inception-V2 + ResNet-18		97.05
ERNet + Inception-V2		97.27
ERNet + Squeezenet		97.11

achieves a recognition accuracy of 96.02% and ERNet-part achieves a recognition accuracy of 96.14%. We found that accuracy and generalization power can be improved to some extent by applying dropout to all convolutional layers. This can be attributed to that the dropout operation affects the behavior of neurons in different layers, dropout operation causes the network to adapt to new feature combinations and thus improves its robustness, and it also improves the generalization power by allowing better sparsity.

We analyze the parameters and time complexity of the proposed network and compare it with the other classic methods. In time complexity calculation, only the FLOPs of convolution operations for computation complexity comparison are considered, and this is because other operations, such as BN and pooling, are insignificant compared to convolution operations. In Table I, we can see that our methods have 1.92 billion flops (multiply-adds) and 11.76M parameters. ERNet's parameters are slightly lower than Inception-V2, and the computation complexity is higher than Resnet-18. By focusing on the model compression, squeezenet has the lowest amount of parameters and computation.

Second, we increase the proportion of training samples, randomly select 80% of the images from each category as the training set, and use the rest as the test set. Several baseline methods, as well as some state-of-the-art classification methods [12]–[14], [16], [24], [25] for remote sensing images, are utilized as benchmarks.

From Table II, we can see that ERNet achieves the recognition rate of 96.95%, whereas ResNet-18 obtained the recognition rate of 96.78%. With the increase of the training samples, the recognition rate of the algorithm was improved. Based on the ensemble learning algorithm, ERNet combined with Inception-V2 classification method achieved the highest recognition rate of 97.27%.



	agricultural	airplane	baseballdiamond	beach	buildings	chaparral	denseresidential	forest	freeway	golfcourse	harbor	intersection	mediumresidential	mobilehomepark	overpass	parkinglot	river	runway	sparseresidential	storagetanks	tenniscourt
agricultural	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
airplane	0.00	0.97	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00
baseballdiamond	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
beach	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
buildings	0.00	0.00	0.00	0.00	0.97	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
chaparral	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
denseresidential	0.00	0.00	0.00	0.00	0.01	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
forest	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
freeway	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
golfcourse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
harbor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
intersection	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
mediumresidential	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mobilehomepark	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.02	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00
overpass	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
parkinglot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00
river	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00
runway	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00
sparseresidential	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00
storagetanks	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00
tenniscourt	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.97

Fig. 2. Confusion matrix of the ERNet on the UC Merced data set.

To further illustrate the performance of our proposed ERNet method, we evaluated the classification rate per class of our method on the UC-Merced data set using a confusion matrix. The confusion matrix of the classification results in Fig. 2 shows that our method can lead to over 95% classification accuracy for most categories. We can also see that the most confused classes are dense residential and medium residential. This confusion is caused by that the scenes from the dense residential areas and the medium residential areas that are too similar to distinguish.

B. Experiment on WHU-RS19 Data Set

The WHU-RS data set [26] consists of 19 types of samples, each type contains about 50 images, and each image has a size of 600 × 600. The resolution, scale, and azimuth of the image in this data set vary significantly, which causes challenges in image classification.

In order to achieve better classification performance, we use the ERNet pretrained with the ImageNet to fine-tune the WHU-RS database; 50% of each subject are randomly selected for training, and the remainder for testing. We use the same data enhancement methods as those used in the UC-Merced Land experiment to extend the data set. In experiments with the squeezenet, Inception-V2, and ResNet-18 methods, the pretrained models were also employed to fine-tune the WHU-RS database. We set a batch size of 16, a momentum of 0.9, a decay of 0.0005, and a learning rate of 0.001. We compared our results with various methods that have reported classification accuracy on the WHU-RS data set. We repeated the experiments ten times and calculated the average testing accuracy rate as the final performance. The recognition accuracy is shown in Table III.

ERNet achieved a recognition rate of 97.84%, which is 0.25% and 0.36% higher than the recognition rates of ResNet-18 network and Inception-V2 network, respectively. we can also see that the ensemble learning method can effectively improve the recognition rate. The highest recognition rate of 98.27% was achieved when ERNet was combined with the Inception-V2 classification method, 0.23% points higher

TABLE III
EXPERIMENT ON THE WHU-RS19 DATA SET (%)

Models\Datasets	WHU-RS19
VGG19 + CRC [12]	94.58
ResNet-152 + CRC [12]	97.11
VGG19 + SPM-CRC [12]	96.68
VGG19 + WSPM-CRC [12]	96.76
ResNet-152 + SPM-CRC [12]	97.76
ResNet-152 + WSPM-CRC [12]	97.74
ResNet-18	97.59
Inception-V2	97.48
Squeezenet	97.02
ERNet	97.84
Squeezenet + Inception-V2	97.53
Squeezenet + ResNet-18	97.71
Inception-V2 + ResNet-18	97.97
ERNet + Squeezenet	98.04
ERNet + Inception-V2	98.27

than the second-highest recognition rate. The ensemble learning method combines the features of multiple networks can improve image classification performance.

IV. CONCLUSION

In this letter, we present an ERNet to improve the classification performance of sensing images. At first, we moderately broadened the first several layers of the network, changed the size of the convolution filters, and made it learn more information of the image. Second, we added a dropout layer to each residual unit to improve the accuracy and generalization power of ERNet. Third, an ensemble learning method based on ERNet was used to improve the classification ability by fusing features of other baseline methods. The experimental results on several public data sets demonstrate that ERNet can provide superior performance to the classical approaches.

REFERENCES

[1] W. Liu, D. Tao, J. Cheng, and Y. Tang, "Multiview Hessian discriminative sparse coding for image annotation," *Comput. Vis. Image Understand.*, vol. 118, pp. 50–60, Jan. 2014.

[2] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[3] W. Liu, X. Ma, Y. Zhou, D. Tao, and J. Cheng, "p-Laplacian regularization for scene recognition," *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 2927–2940, Aug. 2019.

[4] J. Yu, Y. Rui, and D. Tao, "Click prediction for Web image reranking using multimodal sparse coding," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2019–2032, May 2014.

[5] B. Du, Y. Wang, C. Wu, and L. Zhang, "Unsupervised scene change detection via latent Dirichlet allocation and multivariate alteration detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4676–4689, Dec. 2018.

[6] 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.

[7] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[8] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2169–2178.

[9] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 2, pp. 661–674, Feb. 2020, doi: 10.1109/TNNLS.2019.2908982.

[10] J. Su, Q. Wang, S. Chen, and X. Li, "An introspective learning strategy for remote sensing scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 533–536.

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

[12] B.-D. Liu, J. Meng, W.-Y. Xie, S. Shao, Y. Li, and Y. Wang, "Weighted spatial pyramid matching collaborative representation for remote-sensing-image scene classification," *Remote Sens.*, vol. 11, no. 5, p. 518, 2019.

[13] 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.

[14] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[15] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," 2015, *arXiv:1508.00092*. [Online]. Available: http://arxiv.org/abs/1508.00092

[16] Y. Yu, Z. Gong, C. Wang, and P. Zhong, "An unsupervised convolutional feature fusion network for deep representation of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 1, pp. 23–27, Jan. 2017.

[17] J. Ji, T. Zhang, L. Jiang, W. Zhong, and H. Xiong, "Combining multilevel features for remote sensing image scene classification with attention model," *IEEE Geosci. Remote Sens. Lett.*, early access, Nov. 1, 2019, doi: 10.1109/LGRS.2019.2949253.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[19] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, Boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

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

[21] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[22] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.

[23] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[24] 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, 2015.

[25] P. Li, P. Ren, X. Zhang, Q. Wang, X. Zhu, and L. Wang, "Region-wise deep feature representation for remote sensing images," *Remote Sens.*, vol. 10, no. 6, p. 871, 2018.

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