Convolutional Neural Networks Based Remote Sensing Scene Classification under Clear and Cloudy Environments

Huiming Sun[1], Yuewei Lin[2], Qin Zou[3], Shaoyue Song[4], Jianwu Fang[5], Hongkai Yu[1]*
[1]Cleveland State University, [2]Brookhaven National Laboratory
[3]Wuhan University, [4]Beijing Jiaotong University, [5]Chang'an University

Abstract

Remote sensing (RS) scene classification has wide applications in the environmental monitoring and geological survey. In the real-world applications, the RS scene images taken by the satellite might have two scenarios: clear and cloudy environments. However, most of existing methods did not consider these two environments simultaneously. In this paper, we assume that the global and local features are discriminative in either clear or cloudy environments. Many existing Convolution Neural Networks (CNN) based models have made excellent achievements in the image classification, however they somewhat ignored the global and local features in their network structure. In this paper, we propose a new CNN based network (named GLNet) with the Global Encoder and Local Encoder to extract the discriminative global and local features for the RS scene classification, where the constraints for inter-class dispersion and intra-class compactness are embedded in the GLNet training. The experimental results on two publicized RS scene classification datasets show that the proposed GLNet could achieve better performance based on many existing CNN backbones under both clear and cloudy environments.

1. Introduction

Remote sensing (RS) scene classification is a fundamental task in the remote sensing research and it is widely used in agricultural exploration, urban planning, environmental monitoring, etc. The RS scene images are normally taken by the satellite or UAV (Unmanned Aerial Vehicle). In the real-world applications, the RS scene images might have two scenarios: clear and cloudy environments. However, most of existing methods, such as [2,4,6,8,23,24,26,28,30], did not consider these two environments simultaneously. In this paper, we would like to develop a new CNN based method to accurately classify the RS scene images under both clear and cloudy environments[1].

To solve this RS scene classification task, several approaches have been developed, which could be roughly divided into two types: traditional machine learning based methods and deep learning based methods. Early traditional machine learning based methods [3, 16, 18, 21, 27] mainly use low-level hand-crafted features (e.g., color, shape, texture) and regular classifiers (e.g., Support Vector Machine) to classify the images. Because the representation capacity of hand-crafted features might be not enough to fully describe the complex RS scene images, the traditional machine learning based methods did not perform very well even on the clear RS scene images. Recently, with the tremendous development of CNN, the deep learning based methods [1,2,14,23,29,30] show progresses in the RS scene classification by designing some end-to-end deep neural networks. As shown in [23], many CNN based image classification methods like AlexNet [11], VGG16 [22], and ResNet [7] could be used to classify the RS scene images with the relatively good performance. These well-known CNN architectures can be used as backbones for the deep learning based methodology development.

However, all these two kinds of methods assume that the input RS scene image is clear without the degradation by clouds. The previous work [3] introduced that the global and local features can be fused to improve the performance for the clear RS scene classification. In this paper, we assume that the global and local features are discriminative under either clear or cloudy environments, which could be used to classify the RS scene image under either clear or cloudy environments. However, many existing CNN based models somewhat ignored the global and local features in their network structure. Therefore, we propose a new CNN based network (named GLNet) under either clear or cloudy environments that utilizes Global Encoder and Local Encoder to extract the discriminative global and local features for the RS scene classification. In addition, the constraints for inter-class dispersion and intra-class compactness are

*Corresponding author: Hongkai Yu (h.yu19@csuohio.edu)

[1]In this paper, "cloud" represents the "thin cloud" in remote sensing.

Figure 1. Remote Sensing scene image classification under clear and cloudy environments. Top row: clear images from the public RSSCN7 dataset (7 classes in total) [30]; Bottom row: corresponding cloudy images from the synthetic dataset.

embedded in the network training of the proposed GLNet. Besides, to test the proposed GLNet, we synthesize the cloudy RS scene images based on the clear RS scene images of two publicized RS datasets [28, 30]. On the both clear and synthetic cloudy RS scene image recognition datasets, the extensive experimental results show that the proposed GLNet could obtain the best performance over the comparison methods.

The contributions of this paper are threefold: 1) We propose a new deep learning model to classify both clear and cloudy RS scene images; 2) We propose a new deep learning method named GLNet combining the learning of discriminative global and local features for the RS scene classification under clear and cloudy environments, where inter-class dispersion and intra-class compactness are embedded in the GLNet training; 3) Without publicized real cloudy RS scene image datasets, we propose a way to study this research problem from data synthesizing.

Our publicized code of the cloudy RS scene image synthesizing and the proposed GLNet can be found in https://github.com/wuchangsheng951/GLNET.

2. Related Work

RS scene classification: In the past few decades, many different approaches have been developed for remote sensing scene classification. They can be divided in two forms, i.e., traditional machine learning and deep learning based methods. The traditional machine learning methods [3, 16, 18, 21, 27] extract some handcrafted image features (like color histogram, texture) and then input them into some classifiers for the recognition, such as RF [19], SVM [20]. Although these traditional machine learning approaches could effectively deal with regular cases, they cannot accurately classify complex remote sensing scenes due to the lack of overall understanding of semantic information. To solve this problem, deep Convolution Neural Networks (CNNs), e.g., AlexNet [11], VGG16 [22], ResNet [7], have recently been applied for the remote sensing scene classification [1, 2, 9, 14, 15, 23, 24, 29, 30]. For example, [15] extracts the deep CNN features instead of handcrafted image features and applies the SVM classifier for the remote sensing scene classification; [9] stacks different layers of CNN feature maps and they show that the multilayer stacked covariance pooling is quite useful for the remote sensing scene classification; [24] uses a Gated Bidirectional Network to integrate the hierarchical feature aggregation for this task. These deep learning based methods might focus on the global image features and somewhat ignore the local image features, and they all assume that the input remote sensing images are clear without clouds.

RS scene classification under cloudy environment: To the best of our knowledge, RS scene classification under cloudy environment has not been systemically studied before. In the research area of RS scene classification, most of previous researches did not consider the cloudy environment, which actually happens in the real-world data collection. Previous researches try to remove clouds only for a better general visualization purpose [5, 12, 13, 17], but there are no related works to discuss the RS scene classification problem under cloudy environment so far as we know. In this paper, we propose a new deep learning based GLNet model for the RS scene classification under both clear and cloudy environments.

3. Method

3.1. Synthetic Cloudy RS Scene Image Data

We generate the synthetic cloud image I_c by using the summation of multi-scale random noise images. The detailed synthesizing is computed as follow:

$$I_c = \sum_s \Psi\left(Rand(2^s)\right)/2^s, \tag{1}$$

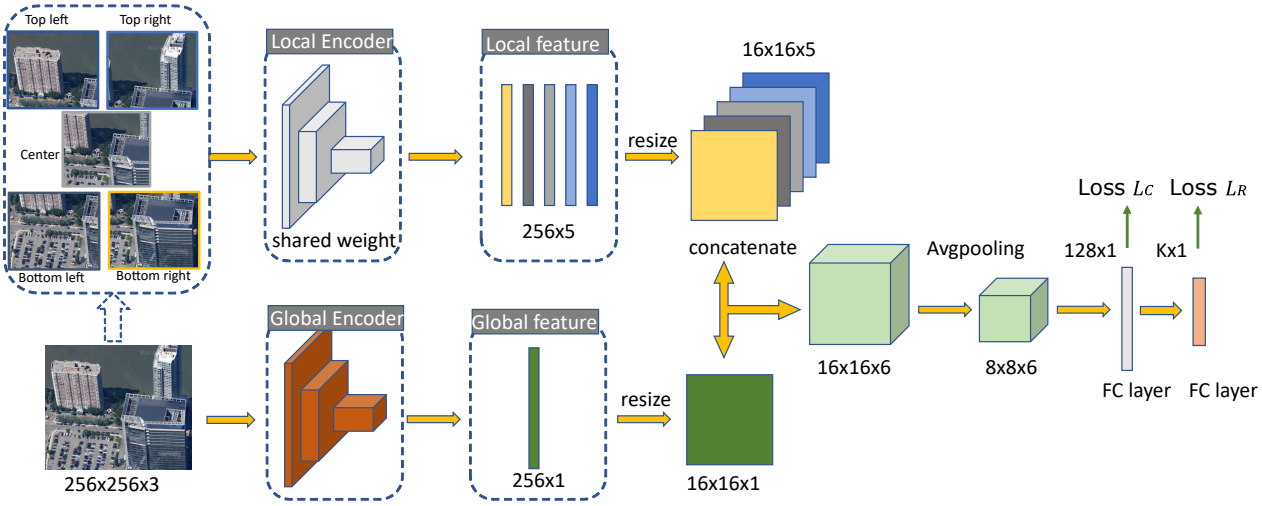

Figure 2. Overview architecture of the proposed GLNet for the RS scene classification under clear and cloudy environments. The proposed GLNet could learn the discriminative global and local features with the embedding of inter-class dispersion and intra-class compactness in the network training.

where $Rand(2^s)$ denotes a randomizing function which produces random noises with the image size of 2^s, and Ψ denotes the operator that resizes the random noise to the cloud image size, and s is the scale factor, which is the natural number with the range from 1 to $\log_2(N)$, where N is the cloud image size. The examples of synthetic cloud images are shown in Fig. 1.

3.2. Network Architecture

This section introduces the details of the proposed GLNet, whose network architecture is illustrated in Fig. 2. Different with many CNN models somewhat ignoring the local information, the proposed GLNet designed a two-branch CNN architecture to fully extract the global and local features simultaneously.

Given one input RGB RS scene image with the size of $H \times W \times 3$ as shown in Fig. 2, the proposed GLNet uses one branch as the Global Encoder to extract the global features $\mathbf{f}_g$ by feeding the whole image as the input, and simultaneously applies another branch as the Local Encoder to learn the local features $\{\mathbf{f}_{l_1}, \mathbf{f}_{l_2}, \mathbf{f}_{l_3}, \mathbf{f}_{l_4}, \mathbf{f}_{l_5}\}$ by feeding the local patches as the input. The local patches are the five local image regions divided from the whole image: top left, top right, bottom left, bottom right, and center. Each patch is a square, whose size is 36% of the whole image. It is worth mentioning that the local patches are spatially overlapped a little bit to maintain their hidden context relationship. Each local patch is then resized to $H \times W \times 3$ as the input to the Local Encoder. Since the global and local features are both discriminative under clear and cloudy environments, we fuse them by a simple concatenation after the feature extraction as the final discriminative features $\mathbf{f}$, followed by an average pooling layer and a Fully Connected (FC) layer to extract the deep features to learn the RS scene class centers and another FC layer to reduce the output dimension to the class number K.

The Global Encoder and Local Encoder can be replaced by some widely-used CNN as the backbones, such as AlexNet [11], VGG16 [22], and ResNet [7], etc.

3.3. Loss function

3.3.1 RS Scene Recognition Loss

The RS scene recognition loss is realized by the fully supervised cross-entropy loss for classification. Given one input image $\mathbf{x}$, its output for K classes by GLNet is defined as $o_1, o_2, \cdots, o_K$. The output is not normalized, so we use the "Softmax" function to normalize each output value as the probability into the range of $[0, 1]$, which is shown in the following equation:

$$p_i = \frac{\exp(o_i)}{\sum_{j=1}^{K} \exp(o_j)}, \tag{2}$$

where p_i is the probability to be the i-th class for the input image. Let us assume the predicted probability of $\mathbf{x}$ as the ground-truth class to be $p_{\mathbf{x}}$ and the ground-truth label is a K-dimensional one-hot vector $\mathbf{y}$, where $y_i = 1$ if the ground truth of $\mathbf{x}$ is class i. The RS scene recognition loss is defined the following equation:

$$L_R(\mathbf{x}, \mathbf{y}) = -y_i \cdot \log p_{\mathbf{x}}. \tag{3}$$

Minimizing the RS scene recognition loss during the network training will optimize the network to predict the consistent class as the ground truth label.

3.3.2 RS Scene Center loss

In this paper, we apply the center loss [10, 25] to embed the inter-class dispersion and intra-class compactness in the network training. Given the m input images $\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_m}$ in one training mini-batch, we define the center of deep features for each of K classes. Let us assume that $\mathbf{x_i}$'s deep global-local feature center is $\mathbf{c_{yi}}$. The center loss is defined in the following equation:

$$L_C = \frac{1}{2}\sum_{i=1}^{m} ||\mathbf{x_i} - \mathbf{c_{yi}}||_2^2. \quad (4)$$

As defined in [25], minimizing the center loss means to learn each class's feature center and penalize the L_2 distances between the features and their corresponding class centers. The center $\mathbf{c_{yi}}$ is updated as the deep features changed during the each mini-batch training.

3.3.3 Overall loss

With the above two loss terms, the overall loss function of our approach can be written as:

$$L = L_R + \alpha L_C, \quad (5)$$

where α is a weight parameter to balance each loss term in the overall loss function.

4. Experiments

4.1. Datasets

Two popular remote sensing scene classification datasets (RSSCN7 and UC Merced) are used as the clear RS scene images to evaluate the proposed method. In addition, we use the method described in Section 3.1 to synthesize the cloudy RS scene images. The detailed dataset information is introduced in the following.

4.1.1 RSSCN7 and RSSCN7_cloud Datasets

The RSSCN7 [30] dataset is acquired from Google Earth. Each image has a size of 400×400 pixels in the RGB color space. These images contain seven classes: grassland, farmland, industrial and commercial regions, river and lake, forest field, residential region, and parking lot. Each category's image number is 400. Using each clear image in RSSCN7, we generate its corresponding cloudy image, leading to a new synthetic dataset named RSSCN7_cloud.

4.1.2 UC Merced and UC Merced_cloud Datasets

The UC Merced dataset [28] is composed of 21 classes and each class is consists of 100 images with a size of 256×256 pixels in the RGB color space. These 21 classes are agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court. In the same way, we generate its corresponding cloudy images, resulting in a new synthetic dataset named UC Merced_cloud.

For each of the RSSCN7 and RSSCN7_cloud datasets, we follow the default split for training and testing set in its original RSSCN7 work [30]: 50% for training and another 50% for testing. For each of the UC Merced and UC Merced_cloud datasets, we randomly select half images for training and another half images for testing same as that in [3].

4.2. Experimental Settings

We evaluate the classification performance of the proposed method on several classical CNN backbones. For example, Proposed$_A$, Proposed$_V$, and Proposed$_R$ indicate the proposed GLNet based on the CNN backbones of AlexNet [11], VGG16 [22], and ResNet50 [7], respectively. For the cloudy images, AlexNet$_c$, VGG16$_c$, and ResNet50$_c$ means directly testing the model pre-trained on the clear images. The input image of GLNet is resized to 256×256. We randomly apply the horizontal flip and the changes of the brightness, contrast and saturation for the data augmentation. We use the SGD optimization algorithm for the network training with the following hyper parameters: initiative learning rate as 0.0006, momentum as 0.5, batch size as 8 and training epoch as 50. All the experiments were run on a workstation with a NVIDIA Quadro P6000 GPU card (24G). We use PyTorch to implement the proposed GLNet.

4.3. Experimental Results

This section will report the performance of the proposed GLNet on the benchmark datasets. Table 1 and Table 2 show the overall classification accuracy on the RSSCN7 and RSSCN7_cloud datatasets respectively. Table 3 and Table 4 show the overall classification accuracy on the UC Merced and UC Merced_cloud datatasets respectively. Compared to the baseline and comparison methods, it is obvious that the proposed method could achieve the highest classification accuracy for RS scene recognition under both clear and cloudy environments. With different CNN backbones, the proposed method could obtain better performance over the CNN baselines. Under the clear environment of RSSCN7, the Proposed$_V$ got 95.07% using VGG16 as the backbone, while the baseline VGG16 only got 93.57%. Under the cloudy environment, the pre-trained models on clear images like VGG16$_c$ got low performance of 78.50%, which indicates the difficulty of the cloudy environment. However, the Proposed$_V$ also got the highest accuracy of 94.79% on

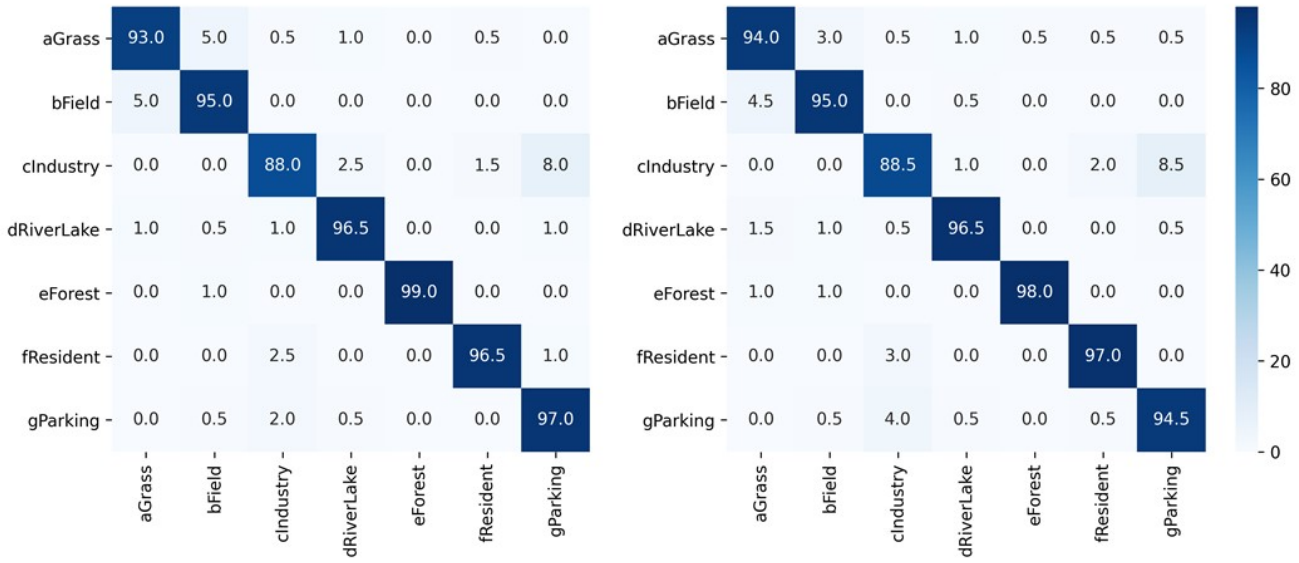

Figure 3. The confusion matrix of the classification result by the proposed GLNet on RSSCN7 dataset (Left) and RSSCN7_cloud dataset (Right) using VGG16 as backbone.

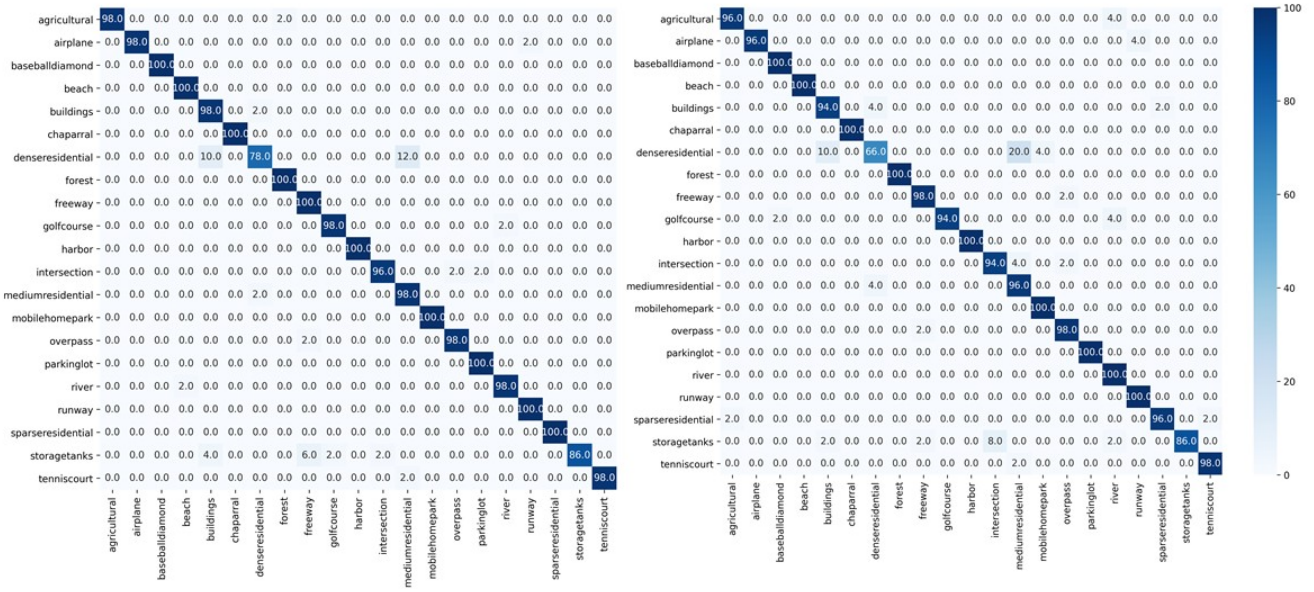

Figure 4. The confusion matrix of the classification result by the proposed GLNet on UC Merced dataset (Left) and UC Merced_cloud (Right) dataset using ResNet50 as backbone.

the RSSCN7_cloud datataset. The similar result is obtained on the UC Merced and UC Merced_cloud datatasets. The detailed confusion matrices for classification are shown in Fig. 3 and Fig 4, which also display the proposed GLNet could achieve better performance over the CNN baselines. On the UC Merced dataset, the traditional machine learning method salM^{3}LBP–CLM [3] combining global and local features got reasonable result of 94.21%, but the proposed method got better accuracy of 97.33% because it designed an advanced deep learning framework to learn global and local features embedding with the inter-class dispersion and intra-class compactness.

4.4. Discussion on the loss weight parameter α

This section will discuss the effects of different loss weight parameter α in the loss function of Eq. 5. Using the RSCCN7 dataset as an example, Fig. 5 shows the overall accuracy of the CNN baseline, the GLNet with $\alpha = 0$, and the GLNet with $\alpha = 0.5$. When $\alpha = 0$, the GLNet with only the RS scene recognition loss could get higher accu-

Table 1. Overall classification accuracy of RSSCN7 Dataset.

Method	Classification Accuracy (%)
AlexNet [11]	91.85
VGG16 [22]	93.57
ResNet50 [7]	93.64
DCNN [30]	77.00
DAC [23]	93.43
TEX-Net-LF [1]	94.00
TDFF [14]	92.37
Proposed$_A$	93.78
Proposed$_V$	**95.07**
Proposed$_R$	*94.71*

Table 2. Overall classification accuracy of RSSCN7_cloud Dataset.

Method	Classification Accuracy (%)
AlexNet$_c$	65.50
VGG16$_c$	78.50
ResNet50$_c$	75.78
AlexNet [11]	88.85
VGG16 [22]	93.14
ResNet50 [7]	91.57
Proposed$_A$	92.07
Proposed$_V$	**94.79**
Proposed$_R$	*93.71*

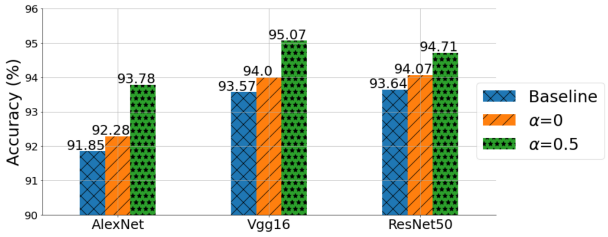

Figure 5. Experimental result of the proposed method with different α on the RSSCN7 dataset.

racy over the CNN baseline. When $\alpha = 0.5$, the GLNet with both the RS scene recognition loss and RS scene center loss could get the highest accuracy. This experimental result verifies the effects of each loss term in the loss function of Eq. 5.

4.5. Discussion on the failure cases

Using ResNet50 backbone as example, we show the failure cases of the proposed method on the RSSCN7 and RSSCN7_cloud datasets in Fig. 6. Some images might be quite similar, leading to recognition confusions. For example, the Industry image (as shown in first column of Fig. 6) contains some parked vehicles, so the proposed method may confuse it as the class of Parking. In addition, the clouds might cause more difficulties in scene classification. For

Table 3. Overall classification accuracy of UC Merced Dataset.

Method	Classification Accuracy (%)
AlexNet [11]	91.62
VGG16 [22]	96.38
ResNet50 [7]	96.76
salM^{3}LBP–CLM [3]	94.21
Two-Stream Fusion [29]	*96.97*
TEX-Net-LF [1]	96.98
Proposed$_A$	95.24
Proposed$_V$	96.76
Proposed$_R$	**97.33**

Table 4. Overall classification accuracy of UC Merced_cloud Dataset.

Method	Classification Accuracy (%)
AlexNet$_c$	82.48
VGG16$_c$	88.95
ResNet50$_c$	93.24
AlexNet [11]	90.10
VGG16 [22]	93.52
ResNet50 [7]	*96.10*
Proposed$_A$	94.57
Proposed$_V$	95.81
Proposed$_R$	**97.33**

example, the Resident image (as shown in sixth column of Fig. 6) contains many small rectangle-shaped buildings under cloudy environment, then the proposed method may confuse the buildings as vehicles, so the proposed method classifies it as the class of Parking. Under the cloudy environment, some discriminative image features might be partially occluded or blurred by the clouds, which causes more recognition difficulties.

5. Conclusions

In this paper, we proposed a new deep learning network for remote sensing scene image classification under clear and cloudy environments. By combing the global and local features and embedding the inter-class dispersion and intra-class compactness, our proposed method is more robust and accurate than the normal CNN networks. The experimental results on two public remote sensing clear image datasets and two synthetic cloudy datasets verified the effectiveness and accuracy of the proposed method.

6. Acknowledgement

Huiming Sun and Hongkai Yu are supported by NCHRP-225. Qin Zou is supported by NSFC 61872277. Jianwu Fang is supported by NSFC U1713217.

Figure 6. Failure cases of the proposed method (using ResNet50 backbone as example) on the RSSCN7 and RSSCN7_cloud datasets. The prediction and ground truth are shown under each remote sensing image in red and green colors respectively.

References

[1] Rao Muhammad Anwer, Fahad Shahbaz Khan, Joost van de Weijer, Matthieu Molinier, and Jorma Laaksonen. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 138:74–85, 2018. 1, 2, 6

[2] Qi Bi, Kun Qin, Han Zhang, and Gui-Song Xia. Local semantic enhanced convnet for aerial scene recognition. *IEEE Transactions on Image Processing*, 2021. 1, 2

[3] Xiaoyong Bian, Chen Chen, Long Tian, and Qian Du. Fusing local and global features for high-resolution scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(6):2889–2901, 2017. 1, 2, 4, 5, 6

[4] Souleyman Chaib, Huan Liu, Yanfeng Gu, and Hongxun Yao. Deep feature fusion for vhr remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(8):4775–4784, 2017. 1

[5] Yang Chen, Luliang Tang, Xue Yang, Rongshuang Fan, Muhammad Bilal, and Qingquan Li. Thick clouds removal from multitemporal zy-3 satellite images using deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:143–153, 2019. 2

[6] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2849–2858, 2019. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1, 2, 3, 4, 6

[8] Nanjun He, Leyuan Fang, Shutao Li, Antonio Plaza, and Javier Plaza. Remote sensing scene classification using multilayer stacked covariance pooling. *IEEE Transactions on Geoscience and Remote Sensing*, 56(12):6899–6910, 2018. 1

[9] Nanjun He, Leyuan Fang, Shutao Li, Antonio Plaza, and Javier Plaza. Remote sensing scene classification using multilayer stacked covariance pooling. *IEEE Transactions on Geoscience and Remote Sensing*, 56(12):6899–6910, 2018. 2

[10] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1945–1954, 2018. 4

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1, 2, 3, 4, 6

[12] Xinghua Li, Liyuan Wang, Qing Cheng, Penghai Wu, Wenxia Gan, and Lina Fang. Cloud removal in remote sensing images using nonnegative matrix factorization and error correction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 148:103–113, 2019. 2

[13] Qi Liu, Xinbo Gao, Lihuo He, and Wen Lu. Haze removal for a single visible remote sensing image. *Signal Processing*, 137:33–43, 2017. 2

[14] Yishu Liu, Yingbin Liu, and Liwang Ding. Scene classification based on two-stage deep feature fusion. *IEEE Geoscience and Remote Sensing Letters*, 15(2):183–186, 2017. 1, 2, 6

[15] Xiaoqiang Lu, Xiangtao Zheng, and Yuan Yuan. Remote sensing scene classification by unsupervised representation learning. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5148–5157, 2017. 2

[16] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:166–177, 2019. 1, 2

[17] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346, 2020. 2

[18] Mercy W Mwaniki, Moeller S Matthias, and Gerhard Schellmann. Application of remote sensing technologies to map the structural geology of central region of kenya. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(4):1855–1867, 2015. 1, 2

[19] Mahesh Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005. 2

[20] Mahesh Pal and PM Mather. Support vector machines for classification in remote sensing. *International journal of remote sensing*, 26(5):1007–1011, 2005. 2

[21] Muhammad Mazhar Ullah Rathore, Anand Paul, Awais Ahmad, Bo-Wei Chen, Bormin Huang, and Wen Ji. Real-time big data analytical architecture for remote sensing application. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(10):4610–4621, 2015. 1, 2

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2, 3, 4, 6

[23] Shaoyue Song, Hongkai Yu, Zhenjiang Miao, Qiang Zhang, Yuewei Lin, and Song Wang. Domain adaptation for convolutional neural networks-based remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 16(8):1324–1328, 2019. 1, 2, 6

[24] Hao Sun, Siyuan Li, Xiangtao Zheng, and Xiaoqiang Lu. Remote sensing scene classification by gated bidirectional network. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):82–96, 2020. 1, 2

[25] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016. 4

[26] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 1

[27] Guanhua Xu, Qinhuo Liu, Liangfu Chen, and Liangyun Liu. Remote sensing for china's sustainable development: Opportunities and challenges. *Journal of Remote Sensing*, 20(5):679–688, 2016. 1, 2

[28] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279, 2010. 1, 2, 4

[29] Yunlong Yu and Fuxian Liu. A two-stream deep fusion framework for high-resolution aerial scene classification. *Computational Intelligence and Neuroscience*, 2018. 1, 2, 6

[30] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 12(11):2321–2325, 2015. 1, 2, 4, 6