Efficient recurrent attention network for remote sensing scene classification

Le Liang1,2 | Guoli Wang3

1 Data Communication Science and Technology Research Institute, Beijing, China
2 Xingtang Telecommunications Technology Co. Ltd., Beijing, China
3 Tsinghua University, Beijing, China

Correspondence
Guoli Wang, Tsinghua University, Beijing 100000, China.
Email: wangguoli1990@mail.tsinghua.edu.cn

Abstract
Scene classification for remote sensing is a popular topic, and many recent convolutional neural networks (CNNs)-based methods have shown the great model capacity and learning ability of highly discriminative features. Given a large number of training data, CNN can extract extensive features and learn to predict a remote sensing image. However, for supervised learning tasks, deep models often rely on a large number of labelled remote sensing images, which are difficult to pre-process. Thus, training a lightweight deep learning model is essential. Easy-classified and hard samples may also cause an imbalance of training set and lead the model to overwhelm the loss function. Accordingly, a novel Efficient Recurrent Attention Network (ERANet) for remote sensing scene classification is proposed. Different from traditional deep learning methods, Efficientnet-B0 is introduced as a lightweight backbone for the ARCNet framework, replacing the original one. By applying the modified efficient backbone, the low Floating Point Operations (FLOPs) and parameter numbers of the proposed ERANet are maintained. The significance of focal loss is determined and applied to address the sample imbalance problem and yield a desirable performance. Extensive experiments on several challenging remote sensing scene classification data sets prove the efficiency of the proposed ERANet.

1 | INTRODUCTION

Scene classification via high-resolution remote sensing images has become an increasingly heated topic in academic researches and industrial applications [1–3]. In recent years, given the rapid development and wide use of deep learning techniques, scene classification also turns to this field and has received desirable results [4]. However, when applying deep learning techniques in scene classification with high-resolution remote sensing images, the difficulty in extracting the most significant features from images through a convolutional neural network (CNN) becomes the main problem that limits the performance of algorithms. High-resolution satellite images are different from images collected from normal sources. These high-resolution images are usually taken from an overhead view covering a large area. Thus, these images include far more types of objects and features than normal images. The negative effects of redundant parts that are irrelevant to scene classification severely limit the performances of deep learning-based scene classification methods [5]. The solution lies in how one extracts useful and task-relevant features while masking out the useless and irrelevant features in feature extractions.

In the development of remote sensing scene classification by using different methods that extract features from images, previous works generally can be sorted into three categories [6]. The first category includes methods with handcrafted features. These methods commonly belong to extremely early works, including the following: the method that devices colour histograms and human-designed texture descriptor with global features that can be used by classifiers directly [7]; the method with local features and mid-term descriptors to produce the entire representations, such as scale-invariant feature transform (SIFT) [8]; and the method combining global and local features [9]. The second category is unsupervised learning features. Many unsupervised methods have achieved better performance than handcrafted feature-based methods [10, 11]. However, unsupervised learning remains not robust in terms of distinguishing the variations between scenes due to the lack of label information. The third category is deep learning features. In recent years, deep learning has been proven to be the most
effective way to conduct the feature extractions in almost all computer vision tasks, including remote sensing scene classification. Nearly all of the state-of-the-art methods in remote sensing scene classification use CNN to extract features. For example, Hu et al. [12] proposed two scenarios for producing features for high-resolution remote sensing images through extracting CNN features from different layers. Chaib et al. [13] proposed to use pre-trained VGG-Net as a deep feature extractor to acquire informative features from high-resolution remote sensing images.

Three significant problems that limit the performance of the large-scale industrial application of the previous methods can be concluded as follows. In the above works, the main problem limiting the performance of algorithms for remote sensing image scene classification remained unchanged and unsolved. All three categories, including deep learning feature extraction, which is proven as the most efficient type, still suffer from redundant information in the features extracted or fabricated. Regarding this problem, some investigations have been conducted, and some breakthroughs have been achieved. The first patch of investigation is relevant to allowing the models to select parts of inputs, rather than all of them used as saliency detection methods [14–16]. Although it has achieved some remarkable breakthroughs in making the model concentrate on useful parts of the remote sensing images, the performance is not desirable enough. Wang et al. [5] proposed ARCNet, which is a new effective method with desirable performance. ARCNet innovatively employs attention mechanisms in remote sensing scene classification. It is a relatively successful attempt to solve the first problem of existing scene classification methods. The second problem lies in the efficiency of ARCNet and other previous methods. The backbones of ARCNet and other deep learning methods are large, and the training could be slow and difficult. This problem hinders the large-scale industrial application of remote sensing scene classification because these bulky backbones cannot allow the models to achieve a desirable performance with limited computational resources. The third problem is the design of loss function. By making the ARCNet framework a little diverse from easy-classified samples and focusing on hard samples during the application, a satisfactory performance can be acquired. The study of Lin et al. [17] indicated that in many cases of training, the majority of the well-classified samples tend to overwhelm loss function and produce a negative influence on the optimization process in backpropagation.

In the current study, a novel method is proposed as a solution to the above three problems. Three main measures are taken in the proposed method as the corresponding solution to each of the three problems, respectively. First, the framework of ARCNet [5] is employed to use the attention mechanism to distill useful information to extract effective, helpful, and relevant features from remote sensing images. Moreover, long short-term memory (LSTM) units are used to process recurrent attention features. Second, EfficientNet proposed by Tan and Le [18], specifically EfficientNet-B0, is employed as the backbone of the feature extraction network. EfficientNet is a newly proposed, light, efficient, and effective backbone that has been successfully adapted to many computer vision tasks concerning extracting features from images. With EfficientNet, the proposed method can acquire a state-of-the-art performance within a relatively tight limit of computational resources and the number of parameters. Third, in terms of the usage of loss functions, the focal loss proposed by Lin et al. [17] is employed to enhance further the performance of the proposed method. Focal loss reshapes the cross-entropy criterion by giving a factor to the standard cross-entropy loss. It downweights the loss assigned to well-classified samples and makes the model focus on sorting hard samples.

To sum up, this study contributes to the literature as follows:

1. By introducing the EfficientNet-B0 lightweight backbone into the ARCNet framework, the low FLOPs and parameter numbers are maintained, and the overall performance of the classification is improved.
2. Focal loss is selected to address the sample imbalance problem and yield a desirable performance.
3. The proposed model has achieved state-of-the-art performance on the three challenging remote sensing scene classification data sets: UC Merced (UCM) land-use data set, Aerial Image Data set (AID), and NWPU-RESISC45 (NWPU) data set.

The remainder of this study is structured as follows. Section 2 introduces some related works. Section 3 provides a detailed description of the proposed method. Section 4 reports the experimental results and analysis. Finally, Section 5 provides the conclusions of this study.

2 RELATED WORKS

In this section, related works on the application of attention mechanism in remote sensing and remote sensing scene classification methods are reviewed.

2.1 Attention mechanism

Although the attention mechanism had already been an overheated and widespread method to enhance performances in numerous types of tasks in computer vision, it was rarely used in the field of remote sensing image scene classification. Attention is divided into three kinds in accordance with the dimension on which they put selective weights. Channel-wise attention is for situations that place attentive weights along channels. Spatial-wise attention is for situations that place attentive weights inside each feature map. Recently, many works about remote sensing scene classification have employed attention mechanisms to enhance their performances and received satisfactory results. For instance, Tong et al. [19] devised a channel-wise attention mechanism in DenseNet as a backbone to construct a remote sensing scene classification network. Li et al. [20] proposed a set of augmentation operations over attention mechanism.
feature maps to compel the model to find class-discriminative features and reduce redundant information as much as possible. Gao et al. [21] proposed to use spatial-wise attention and channel-wise attention mechanisms simultaneously to explore the contextual dependencies from the channel and spatial directions separately. In the above studies, ARCNet was used with novel LSTM structures to process recurrent attention. The use of ARCNet has marked a milestone in the set of methods [5]. ARCNet not only uses attention mechanisms to extract useful features but also views the process of attention feature extraction in scene classification as a recurrent procedure. It also uses LSTM to address this issue and has received a significantly desirable performance. In some other related studies, attention mechanism was used in aerial image scene classification and has become quite mature that it had been integrated into unsupervised learning and deep learning [10, 22, 23]. The task and background of aerial scene classification are quite similar to that of remote sensing scene classification. Consequently, many techniques and inspirations from aerial scene classification can be adopted by remote sensing scene classification.

With the rapid development and widespread application of deep learning, methods with handcrafted features have been outrun by deep learning methods. Unsupervised learning also involves deep learning techniques. However, due to the lack of robust information from labels, unsupervised learning has been not as reliable as deep supervised learning. Currently, the mainstream method with the best performance is deep learning-based scene classification. Nearly all state-of-the-art remote sensing scene classification methods are based on deep learning [5].

3 | METHOD DESCRIPTION

The pipeline of the proposed method is shown in Figure 1. In this section, each procedure involved in the proposed method is elaborated. First, the backbone architecture for high-level feature extraction is explained. Second, a brief introduction of the used recurrent attention module is provided. Finally, the focal loss is used to further improve the classification performance of the proposed method.

2.2 | Remote sensing scene classification

Remote sensing high-resolution scene classification is a relevant topic in computer vision and remote sensing. As mentioned in the previous section, remote sensing scene classification has three kinds of approaches, namely, handcrafted features [7–9], unsupervised learning [10, 11], and deep learning [5, 12, 13].

3.1 | Efficient backbone

Most of the deep learning-based methods for scene classification [5] always utilize popular CNN architectures as a feature extractor, which is also called as ‘backbone’. These famous backbones have been compared on different data sets to select the one with the best performance. AlexNet [24], VGGNet [25],
and ResNet [26] are the three most frequently used models. Researchers can also apply their pre-trained models for fast training and high performance.

In the current study, a novel neural network architecture is employed as the backbone. Different from the above traditional CNN networks [24–26], which try to find out a connection between layers or design a complicated structure from some well-designed handcrafted experiences, EfficientNet [18] takes an unusual way. Similar to [27], EfficientNet simultaneously optimizes accuracy and FLOPs by conducting a multi-objective neural architecture search. Suppose the accuracy and FLOPs of model $m$ are denoted as $\text{Accuracy (ACC)} (m)$ and $\text{FLOPS} (m)$, respectively, then the optimization goal is designed as:

$$\text{ACC} (m) \times \left[ \frac{\text{FLOPS} (m)}{T} \right]^{\omega},$$

where $T = 400M$ is the target FLOPs, and $\omega = -0.07$ is the trade-off hyper-parameter between accuracy and FLOPs. EfficientNet utilizes the same search space as MnasNet [27] and mainly builds on the mobile inverted bottleneck conv (MBConv) [27, 28] with the squeeze-and-excitation module [29]. It also uses Swish activation [30] instead of ReLU [24] for better performance. An example of MBConv6, k5×5, is shown in Figure 2. The searched network architecture is treated as the baseline and noted as EfficientNet-B0. Considering that EfficientNet-B0 can achieve a good trade-off between the accuracy and FLOPs, it is selected as the lightweight backbone for strong representative ability. This method is described in detail in Table 1. Based on EfficientNet-B0, the proposed ERANet only achieves 384.75 M FLOPs and 6.69 M parameters.

3.2 Recurrent attention module

High-resolution remote sensing images often include redundant information, which may go against the scene classification task. Therefore, Wang et al. [5] proposed ARCNet to generate attention features from high-level neural network features for better scene classification performance.

Suppose $F = \{f_1, f_2, \ldots, f_p\}$ is a feature map in the size of $P \times P$ with $D$ channels ($f_i \in R^D$). Each $f_i$ is the high-level feature representation of the corresponding receptive field. Assuming that $a_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,P}, f_i\}$, $i \in {1, \ldots, T}$ is a $P \times P$ sized mask matrix for the attention weights at $t$ time, the multiplication between $F$ and $a_i$ is computed as follows:

$$x_t = \sum_{i=1}^{p \times p} a_{i,j} f_i, \quad x_t \in R^D, f_i \in R^D,$$

where $x_t$ is a $P \times P \times D$ feature block that is subsequently processed by the recurrent attention module at $t$ time, as shown in Figure 1.

ARCNet employs a recurrent neural network (RNN) to obtain recurrent attention features and generate a series of sequential attention representations and persistent learning and update them during the training stage. The state update function can be calculated as follows:

$$RNN : h_{t-1}, x_t \rightarrow h_t,$$

where $h_{t-1}$ and $h_t$ are the hidden state at $t - 1$ and $t$ time, respectively. The LSTM unit is used as the processor in

### Table 1 EfficientNet-B0 backbone

| Input resolution | Operator               | $t$ | $c$ | $n$ | $s$ |
|------------------|------------------------|-----|-----|-----|-----|
| 224×224×3        | Conv3x3                | -   | 32  | 1   | 2   |
| 112×112×32       | MBConv; k3x3           | 1   | 16  | 1   | 1   |
| 112×112×16       | MBConv; k3x3           | 6   | 24  | 2   | 2   |
| 56×56×24         | MBConv; k5x5           | 6   | 40  | 2   | 2   |
| 28×28×40         | MBConv; k3x3           | 6   | 80  | 3   | 2   |
| 14×14×20         | MBConv; k3x5           | 6   | 112 | 3   | 1   |
| 14×14×112        | MBConv; k5x5           | 6   | 192 | 4   | 2   |
| 7×7×192          | MBConv; k3x3           | 6   | 320 | 1   | 1   |
| 7×7×220          | Conv1x1                | -   | 1280| 1   | 1   |

Note: Each line repeats the corresponding operator $s$ times with output channel $c$. $t$ denotes the sliding step and $c$ is the expansion factor.
Attention matrix size $7 \times 10$.

**Table 2.** Parameter settings in the recurrent attention module

| Parameter                  | Setting |
|----------------------------|---------|
| LSTM layer number          | 3       |
| Hidden layer size          | 256     |
| Recurrent number           | 10      |
| Input feature size         | $7 \times 7 \times 1280$ |
| Attention matrix size      | $7 \times 7 \times 10$ |

ARCNet as follows:

$$LSTM : h_{t-1}, c_{t-1}, x_t \rightarrow h_t, c_t,$$  \hspace{1cm} (4)

where $c_{t-1}$ and $c_t$ denote the memory cell state at $t - 1$ and $t$ time, respectively. Similar to ARCNet [5], three stacked LSTM layers whose hidden layers with $D$ memory cells and a softmax activation are used to generate the prediction vector as follows:

$$y_t = \langle y_{t,1}, y_{t,2}, \ldots, y_{t,L} \rangle, \quad t \in 1, \ldots, T,$$  \hspace{1cm} (5)

where $y_t$ denotes the prediction vector at $t$ time, and $T$ is the total number of recurrent time. $L$ denotes the category number, and thus $y_{t,l}$ means the probability belongs to the $l$th category at $t$ time. All prediction vectors in different time are directly summed up to obtain the final prediction vector $V$, as shown in Figure 1:

$$V = \sum_{t=1}^{T} y_t.$$  \hspace{1cm} (6)

The output of stacked LSTM $h_t$ is also input to another softmax, which outputs the probability revealing a significant degree of each pixel in the feature block at $t + 1$ time, i.e. the attention matrix $a_{t+1}$. The recurrent attention architecture used in the proposed method is the same as in ARCNet. The ablation experiments in ARCNet are directly used as a reference, and the parameters are finally set. The settings are elaborated in Table 2.

### 3.3 Loss function

For scene classification tasks, most of the state-of-the-arts (SOTAs) methods choose cross-entropy loss and its variants to compose its loss function [5]. Although traditional cross-entropy loss shows its performance on benchmarks, for actual scene classification problems, the imbalanced data and classes are likely to be tackled with high similarity. For instance, the AID data set includes 30 scene categories with $600 \times 600$ resolution, where the number of each class varies from 220 to 420. Moreover, the NWPU-RESISC45 data set contains classes of basketball courts and tennis courts, which intensifies the characteristic of small interclass variance.

The focal loss is applied in this study to release the imbalance of different classes of scene images and learn discriminative features from all classes of remote sensing images. Consequently, intraclass distance is reduced, and the interclass distance is enlarged. The focal loss [17] derives from cross-entropy loss, and its details are discussed in the subsequent sections.

The focal loss for the one-stage object detector is designed to obtain the outstanding performance of two-stage detectors, even if an extreme imbalance exists between foreground and background pixels. Consider the following binary classification problem for cross-entropy loss:

$$CE(p_t) = \begin{cases} -\log(p_t) & \text{if } y = 1 \\ -\log(1 - p_t) & \text{otherwise.} \end{cases}$$  \hspace{1cm} (7)

The two classes, namely, $y \in \{ \pm 1 \}$ and $\rho \in [0, 1]$, are considered as the probability for the ground-truth class ($y = 1$). $p_t$ can be defined as follows:

$$p_t = \begin{cases} \rho & \text{if } y = 1 \\ 1 - \rho & \text{otherwise.} \end{cases}$$  \hspace{1cm} (8)

Equation (7) can also be rewritten as $CE(p_t) = CE(p_t) = -\log(p_t)$ for convenience.

In this study, the cross-entropy loss is multiplied with a modulating factor $(1 - p_t)^\gamma$. $\gamma$ is a tunable focusing parameter restricted by $\gamma \geq 0$. Intuitively, $(1 - p_t)^\gamma$ reduces easy samples’ contribution to total loss and extends the range of receiving a low loss. By adding this modulating factor, $CE(p_t)$ is converted to the focal version as follows:

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t).$$  \hspace{1cm} (9)

As mentioned previously, a common sense is to bring a weighting factor $\alpha \in [0, 1]$ for the imbalance problem. In practice, $\alpha$ can be tuned in accordance with its inverse class frequency or adjusted as a hyper-parameter by cross-validation. By combining Equation (9) with the weighting factor, the final version of the focal loss function is defined by Equation (10) as follows:

$$FL(p_t) = -\alpha(1 - p_t)^\gamma \log(p_t).$$  \hspace{1cm} (10)

To sum up, the weighting factor $\alpha$ can balance the weights of positive and negative samples. However, it neglects the fact of easy and hard examples. By introducing the modulating factor $(1 - p_t)^\gamma$, $\gamma$ can tune the weight decay speed of easy examples. In particular cases, the focal loss degenerates to cross-entropy loss when $\gamma = 0$. For the scene classification tasks in this study, compared with cross-entropy loss, the imbalance problem can be solved, and the classes that have small interclass variance can be dealt with by tuning $\gamma$, where $\alpha$ is equal to 1.

### 4 EXPERIMENTS

In this section, experiments are carried out to demonstrate the efficiency of the proposed method. First, the data sets
TABLE 3 The statistics of data sets used for remote sensing scene categorization in this paper

| Data sets              | No. Images | No. Class | No. Class images |
|------------------------|------------|-----------|------------------|
| UCM land-use [31]      | 2100       | 21        | 100              |
| AID [32]               | 10,000     | 30        | 220–420          |
| NWPU-RESISC45 [1]      | 31,500     | 45        | 700              |

and evaluation metrics used to verify the proposed method are exhibited. Second, the detailed setting of the parameters is introduced. Finally, the results of comparison experiments with other state-of-the-art methods and an ablation study are displayed to illustrate the performance of the proposed method and the sources of enhancement.

4.1 | Experiment data set

The experiments are conducted on three well-known data sets that are widely used for evaluating remote sensing scene classification tasks. These data sets are the UCM land-use data set, AID, and NWPU-RESISC45 data set. All the details are illustrated in Table 3.

1. UCM land-use data set: The UCM land-use data set [31] is the earliest data with ground truth developed from a public high-resolution overhead image. It is processed from aerial orthography by humans, and it can be accessed from the United States Geological Survey National Map. The land-use scene has 21 categories, including aeroplane, baseball court, beach, agricultural land, buildings, chaparral, dense residence, forest, expressway, golf course, harbour, intersection, medium-density residence, overpass, mobile home park, river, runway, parking lot, sparse residence, storage tanks, and tennis court. Each category has 100 images with a size of 256 × 256 pixels, and each pixel is of a resolution of 30 cm in RGB space. The interclass similarity among classes is extremely high, thereby making UCM one of the most challenging remote sensing classification data set.

2. AID: The AID [32] is a data set for aerial scene classification with 10,000 images whose sizes are 600 × 600 pixels. It consists of 30 aerial scene types: viaduct, storage tanks, stadium, square, sparse residence, school, river, resort, railway station, port, pond, playground, parking lot, park, mountain, medium residence, meadow, industrial land, forest, farmland, centre, baseball field, bridge, commercial areas, church, beach, and airport. The ground-truth labels are made by experts in remote sensing image interpretation.

3. NWPU-RESISC45: The NWPU-RESISC45 data set [1] is a free and publicly available data set for remote sensing scene classification. It was created by NWPU. It has 31,500 images with 45 scene classes, thereby implying that each class has 700 images. Compared with other data sets for remote sensing image scene classification, NWPU-RESISC45 is of great large scale in terms of categories and the total number of images. It is also a highly challenging data set for remote sensing scene classification for its in-class diversity and interclass similarity.

4.2 | Evaluation metrics

In scene classification, the most widely used evaluation metrics are OA (overall accuracy) and CM (confusion matrix). We apply OA and CM to the above three data sets for evaluation.

1. OA: By not considering which class of scene the images belong to, the number of all correctly classified images is taken and divided by the number of all the images in the whole data set.

In training, the testing results on the training set are usually extremely good. However, when placing the trained model on some samples that have never occurred in the training process, the performance may fall drastically. Therefore, not all images are taken in the data set for training. Only 10–80% of the whole data set is used as the training set, and the corresponding remaining data are used as a validation set. Cross-validation is also conducted. For each of the three data sets, cross-validation is repeated five times. For each experiment on each data set, every model is trained and tested five times with a different train-validation split each time.

2. CM: It is a well-known method to illustrate the performance of supervised learning algorithms visually, especially for classification tasks. For a N classes data set, its N × N CM can be computed using truth labels and corresponding predicted labels and normalized to a clear version. In a CM, every row is considered its ground-truth categories, and every column is predicted categories. Therefore, it can be viewed as a category-level classification result.

4.3 | Training details

In this section, the detailed training parameter settings are exhibited. The experiments on the focal factor in the focal loss are explained.

1. Training Parameter: In the experimental part, we first resize all the input images to 256 × 256, and then we randomly crop them in the size of 224 × 224 for training and centrally crop them in the size of 224 × 224 for testing. In the three data sets, the batch size is set as 64, with a learning rate of 0.001. The optimization method in updating parameters is Adam with a weight decay of 0.0001, a momentum of 0.9, and patience of 5. In training, a total of 120 epochs are covered. The implementation of the proposed method is based on Pytorch. The training process is conducted on one NVIDIA Titan X.

2. Focal Loss Parameter: The parameter γ in the exponential position of the mathematical equation of the focal loss controls how much the model will focus on hard samples, given that samples are easy to be misclassified. During experiments on the AID data set, a range of different γ is provided to
TABLE 4 Overall accuracy (%) under the training ratios of 20% and alpha = 1 with different gamma on the aid data set

| gamma | 0.05 | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | 1.0 |
|-------|------|-----|-----|-----|-----|-----|-----|
| OA (%) | 0.9576 | 0.9580 | 0.9593 | 0.9589 | 0.9582 | 0.9582 | 0.9587 |

Note: The best results are highlighted in bold.

TABLE 5 Overall accuracy (%) of our proposed method and the comparison methods under the training ratios of 80% and 50% on the UCM data set

| OA (%) | 80% images for training | 50% images for training |
|--------|-------------------------|-------------------------|
| Methods |                        |                         |
| ARCNet [5] | 99.12 ± 0.40 | 96.81 ± 0.14 |
| GoogLeNet [32] | 94.31 ± 0.89 | 92.70 ± 0.60 |
| CaffeNet [32] | 95.02 ± 0.81 | 93.98 ± 0.67 |
| SalM3LBPCLM [33] | 95.75 ± 0.80 | 94.21 ± 0.75 |
| MS-CLBP+FV [34] | 93.00 ± 1.20 | 88.76 ± 0.76 |
| MSDFF [35] | 99.76 | 98.85 |
| ERANet (Ours) | 100.00 ± 0.00 | 99.30 ± 0.10 |

Note: The best results are highlighted in bold.

4.4 Experimental results

In this part, the comparison experiments between the proposed method and some other state-of-the-art methods are explained. In all these experiments, the evaluation metrics are OA and CM. All comparison results are covered and analysed in accordance with different data sets.

1. UCM land-use data set: Comparative experiments with other state-of-the-art methods in remote sensing scene classification are conducted on the UCM data set. Two experiment settings are used, as shown in Table 5. In the first setting, 80% of all data are used as a training set and 20% as a testing set. In the second setting, 50% is used as a training set, and the other 50% as a testing set. The proposed method is compared with ARCNet [5], GoogLeNet [32], CaffeNet [32], SalM3LBPCLM [33], MS-CLBP+FV [34], and MSDFF [35]. Among all SOTAs, the proposed ERANet has achieved the best scores with 80% and 50% training set. ERANet even has 100% OA when it is trained with 80% images. The CM of 50% of training images is shown in Figure 3 to analyse the results in detail. All the categories are predicted correctly, and only the classes of ‘aeroplane’, ‘building’, ‘harbour’, and ‘sparseresidential’ have some difficult samples that are wrong. This result proves the relatively good performance of the proposed ERANet.

2. AID: Comparative experiments with other state-of-the-art methods in remote sensing scene classification are conducted on the AID data set. Two experiment settings are used, as shown in Table 6. In the first setting, 50% of data is used as a training set and the other 50% as a testing set. In the second setting, 20% is used as a training set and 80% as a testing set. The proposed method is compared with ARCNet [5], MG-CAP [36], DCNN [37], MSCP [38], Combined CNN&GCN [39], MSITL [40], ADFF [41], and MSDFF [35]. Among all SOTAs, the proposed ERANet has achieved the best scores with 50% and 20% training set. ERANet has boosted the OA to 98% and 95% with 50% and 20% training images, respectively.

FIGURE 3 Confusion matrix for UCM with 50% training data (OA = 99.24%)

TABLE 6 Overall accuracy (%) of our proposed method and the comparison methods under the training ratios of 50% and 20% on the AID data set

| OA (%) | 50% images for training | 20% images for training |
|--------|-------------------------|-------------------------|
| Methods |                        |                         |
| ARCNet [5] | 93.10 ± 0.55 | 88.75 ± 0.40 |
| MG-CAP [36] | 96.12 ± 0.12 | 93.34 ± 0.18 |
| DCNN [37] | 96.89 ± 0.10 | 90.82 ± 0.16 |
| MSCP [38] | 96.56 ± 0.18 | 92.21 ± 0.17 |
| Combined CNN&GCN [39] | 96.89 ± 0.10 | 94.93 ± 0.31 |
| MSITL [40] | 95.46 ± 0.24 | 92.65 ± 0.29 |
| ADFF [41] | 94.75 ± 0.24 | 93.68 ± 0.29 |
| MSDFF [35] | 96.74 | 93.47 |
| ERANet (Ours) | 98.39 ± 0.16 | 95.93 ± 0.13 |

Note: The best results are highlighted in bold.

explore the most effective $\gamma$ value that represents the most suitable intensity of focus. $\gamma$ is set to 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, and 1.0 in each experiment, as shown in Table 4. In this serial experiment about $\gamma$ on AID, 20% of the images from AID are used in training and 80% in testing. We finally set $\gamma$ as 0.3.
FIGURE 4 Confusion matrix for AID with 20% training data (OA = 95.96%)

TABLE 7 Overall accuracy (%) of our proposed method and the comparison methods under the training ratios of 20% and 10% on the NWPU data set

| Methods         | 20% images for training | 10% images for training |
|-----------------|-------------------------|-------------------------|
| MG-CAP [36]     | 92.95 ± 0.13            | 90.83 ± 0.12            |
| RTN [42]        | 92.71                   | 89.90                   |
| DCNN [37]       | 91.89 ± 0.22            | 89.22 ± 0.50            |
| MSCP [38]       | 88.93 ± 0.14            | 85.33 ± 0.17            |
| Combined        | 92.87 ± 0.13            | 90.75 ± 0.21            |
| CNN&GCN [39]    |                         |                         |
| MSITL [40]      | 91.62 ± 0.35            | 88.30 ± 0.24            |
| ADFF [41]       | 91.91 ± 0.23            | 90.58 ± 0.19            |
| MSDFF [35]      | 93.35                   | 91.56                   |
| ERANet (Ours)   | 95.12 ± 0.17            | 91.95 ± 0.19            |

Note: The best results are highlighted in bold.

The CM of 20% of training images is shown in Figure 4 to analyse the results in detail. Most categories are predicted with 90% precision. Only the classes of ‘Resort’, ‘School’, and ‘Square’ are lower than 90%. These categories have some confusing features and even cannot be classified by humans easily. To sum up, the proposed ERANet outperforms the others on the AID data set.

3. NWPU-RESISC45: Comparative experiments with other state-of-the-art methods in remote sensing scene classification are conducted on the NWPU data set. Two experiment settings are used, as shown in Table 7. In the first setting, 20% of all data is used as a training set and 80% as a testing set. In the second setting, 10% is used as a training set, and the other 90% as a testing set. The proposed method is compared with MG-CAP [36], RTN [42], DCNN [37], MSCP [38], Combined CNN and GCN [39], MSITL [40], ADFF [41], and MSDFF [35]. Among all SOTAs, the proposed ERANet has achieved the best score with 20% and 10% training set. ERANet is 1.57% higher than the others in OA with 20% training images. It is still better than MSDFF when a small training set is used.

The CM of 10% of training images is shown in Figure 5 to analyse the results in detail. As for the hardest training setting in the experiments, the proposed model is only trained with 10% images, and the data set has up to 45 confusable classes. Although the best scores are still achieved by the proposed method, it has still performed relatively poorly on the classes of ‘church’, ‘palace’, and ‘railway_station’. On this difficult NWPU data set, the superiority of the proposed ERANet over the other SOTAs is still evident.

4.5 Ablation study

In this part, we address and discuss an ablation study on the AID data set with 20% training data to elaborate the influence of each composition. Based on the experimental results, we analyse the effect of every components.

In Table 8, each line represents a combination of several components, and ✓ means the corresponding part is applied in that model. Model 1 gets the poorest performance in the
ablation study. By adding the recurrent attention module, ACC increases from 94.86% to 95.78%. The above results prove the essential of the recurrent attention module. The difference between Models 2 and 3 is the choice of loss function. Model 3 predicts better than Model 2 with an increase of 0.15%. Although traditional cross-entropy loss shows its performance on benchmarks, the imbalanced data and classes with high similarities are tough to be deal with for actual scene classification problems. Focal loss can be a good choice for this kind of problem.

5 | CONCLUSIONS

In this study, a novel ERANet is proposed for remote sensing scene classification. Different from traditional deep learning methods, EfficientNet-B0 is introduced as a lightweight backbone for the ARCNet framework, replacing the original one. By applying the modified efficient backbone, the low FLOPs and parameter numbers of ERANet are maintained, and the overall performance of the classification is improved. Moreover, the significance of the focal loss is discovered and applied to address the sample imbalance problem and yield a desirable performance. Finally, extensive experiments are conducted on several challenging remote sensing scene classification data sets, and the results prove that the proposed ERANet has achieved a relatively good SOTA performance.

ORCID

Gwoi Wang http://orcid.org/0000-0001-8685-3968

REFERENCES

1. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proc. IEEE 105(10), 1865–1883 (2017)
2. Cheng, G., et al: Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. Int. J. Remote Sens. 34(1), 45–59 (2013)
3. Yao, X., et al.: Semantic annotation of high-resolution satellite images via weakly supervised learning. IEEE Trans. Geosci. Remote Sens. 54(6), 3660–3671 (2016)
4. Cheng, G., et al.: Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE J. Sel. Top Appl. Earth Obs. Remote Sens. 13, 3735–3756 (2020)
5. Wang, Qi, et al.: Scene classification with recurrent attention of VHR remote sensing images. IEEE Trans. Geosci. Remote Sens. 57(2), 1155–1167 (2019)
6. Zhang, L., Zhang, L., Du, B.: Deep learning for remote sensing data: A technical tutorial on the state of the art. IEEE Geosci. Remote Sens. Mag. 4(2), 22–40 (2016)
7. Bhagavathy, S., Manjunath, B.S.: Modeling and detection of geospatial objects using texture motifs. IEEE Trans. Geosci. Remote Sens. 44(12), 3706–3715 (2006)
8. Yang, Y., Newsam, S.D.: Geographic image retrieval using local invariant features. IEEE Trans. Geosci. Remote Sens. 51(2), 818–832 (2013)
9. Risiovec, V., Babic, Z.: Fusion of global and local descriptors for remote sensing image classification. IEEE Geosci. Remote Sens. Lett. 10(4), 836–840 (2013)
10. Yu, Y., Li, X., Liu, F.: Attention GANs: Unsupervised deep feature learning for aerial scene classification. IEEE Trans. Geosci. Remote Sens. 58(1), 519–531 (2020)
11. Sheng, G., et al.: High-resolution satellite scene classification using a sparse coding based multiple feature combination. Int. J. Remote Sens. 33(8), 2395–2412 (2012)
12. Hu, F., et al.: Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 7(11), 14,680–14,707 (2015)
13. Chabli, S., et al.: Deep feature fusion for VHR remote sensing scene classification. IEEE Trans. Geosci. Remote Sens. 55(8), 4775–4784 (2017)
14. Zhang, F., Du, B., Zhang, L.: Saliently-guided unsupervised feature learning for scene classification. IEEE Trans. Geosci. Remote Sens. 54(3), 2175–2184 (2015)
15. Zhang, H., Zhang, J., Xu, F.: Land use and land cover classification base on scene saliency map cooperated coding. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 2616–2620 (2015)
16. Zhang, L., Zhang, Y.: Airport detection and aircraft recognition based on two-layer saliency model in high spatial resolution remote-sensing images. IEEE J. Sel. Top Appl. Earth Obs. Remote Sens. 10(4), 1511–1524 (2017)
17. Lin, T., et al.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, pp. 2999–3007. IEEE Computer Society (2017)
18. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June, Long Beach, California, USA, pp. 6105–6114. Vol. 97 of Proceedings of Machine Learning Research, PMLR (2019)
19. Tong, W., et al.: Channel-attention-based DenseNet network for remote sensing scene classification. IEEE J. Sel. Top Appl. Earth Obs. Remote Sens. 13, 4121–4132 (2020)
20. Li, F., et al.: An augmentation attention mechanism for high-spatial-resolution remote sensing image scene classification. IEEE J. Sel. Top Appl. Earth Obs. Remote Sens. 13, 3862–3878 (2020)
21. Gao, Y., et al.: Remote sensing scene classification with dual attention-aware network. In: IEEE 5th International Conference on Image, Vision and Computing (ICIVC), pp. 171–175 (2020)
22. Guo, Y., et al.: Global-local attention network for aerial scene classification. IEEE Access 7, 67,200–67,212 (2019)
23. Lan, M., et al.: Global context based automatic road segmentation via dilated convolutional neural network. Int. J. Sci. 535, 156–171 (2020)
24. Križevski, A., Sunskive, I., Hintoun, G.E.: ImageNet classification with deep convolutional neural networks. In: Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems. Proceedings of a meeting held December 3–6, Lake Tahoe, NV, pp. 1106–1114 (2012)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, Conference Track Proceedings (2015)
26. He, K., et al.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, pp. 770–778. IEEE Computer Society (2016)
27. Tan, M., et al.: MnasNet: Platform-aware neural architecture search for mobile. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, pp. 2820–2828 (2019)
28. Sandler, M., et al.: Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. CoRR abs/1801.04381 (2018)
29. Hu, J., Shen, L., Sun, G: Squeeze-and-excitation networks. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, Workshop Track Proceedings (2018)
30. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, Workshop Track Proceedings (2018)
31. Yang, Y., Newsam, S.D.: Bag-of-visual-words and spatial extensions for land-use classification. In: Agrawal, D., Zhang, P., Abbadi, A.E., Mokbel, M.F. (eds.) Proceedings of 18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, November 3–5, San Jose, CA, USA, pp. 270–279. ACM (2010)
32. Xia, G., et al.: AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens. 55(7), 3965–3981 (2017)
33. Bian, X., et al.: Fusing local and global features for high-resolution scene classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10(6), 2889–2901 (2017)
34. Huang, L., et al.: Remote sensing image scene classification using multi-scale completed local binary patterns and fisher vectors. Remote Sens. 8(6), 483 (2016)
35. Xue, W., Dai, X., Liu, L.: Remote sensing scene classification based on multi-structure deep features fusion. IEEE Access 8, 28,746–28,755 (2020)
36. Wang, S., Guan, Y., Shao, L.: Multi-granularity canonical appearance pooling for remote sensing scene classification. IEEE Trans. Image Process. 29, 5396–5407 (2020)
37. Cheng, G., et al.: When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. IEEE Trans. Geosci. Remote Sens. 56(5), 2811–2821 (2018)
38. He, N., et al.: Remote sensing scene classification using multilayer stacked covariance pooling. IEEE Trans. Geosci. Remote Sens. 56(12), 6899–6910 (2018)
39. Liang, J., Deng, Y., Zeng, D.: A deep neural network combined CNN and GCN for remote sensing scene classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 4325–4338 (2020)
40. Zhang, J., et al.: Training convolutional neural networks with multi-size images and triplet loss for remote sensing scene classification. Sensors 20(4), 1188 (2020)
41. Zhu, R., et al.: Attention-based deep feature fusion for the scene classification of high-resolution remote sensing images. Remote Sens. 11, 11 (2019). https://www.mdpi.com/2072-4292/11/17/1996
42. Chen, Z., et al.: Recurrent transformer network for remote sensing scene categorisation. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3–6, p. 266. BMVA Press (2018)

How to cite this article: Liang L., Wang G. Efficient recurrent attention network for remote sensing scene classification. IET Image Process. 2021;15:1712–1721. https://doi.org/10.1049/ipr2.12139