Early Labeled and Small Loss Selection Semi-Supervised Learning Method for Remote Sensing Image Scene Classification

Ye Tian, Yuxin Dong * and Guisheng Yin

College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China; tianye168@hrbeu.edu.cn (Y.T.); yinguisheng@hrbeu.edu.cn (G.Y.)
* Correspondence: dongyuxin@hrbeu.edu.cn

Abstract: The classification of aerial scenes has been extensively studied as the basic work of remote sensing image processing and interpretation. However, the performance of remote sensing image scene classification based on deep neural networks is limited by the number of labeled samples. In order to alleviate the demand for massive labeled samples, various methods have been proposed to apply semi-supervised learning to train the classifier using labeled and unlabeled samples. However, considering the complex contextual relationship and huge spatial differences, the existing semi-supervised learning methods bring different degrees of incorrectly labeled samples when pseudo-labeling unlabeled data. In particular, when the number of labeled samples is small, it affects the generalization performance of the model. In this article, we propose a novel semi-supervised learning method with early labeled and small loss selection. First, the model learns the characteristics of simple samples in the early stage and uses multiple early models to screen out a small number of unlabeled samples for pseudo-labeling based on this characteristic. Then, the model is trained in a semi-supervised manner by combining labeled samples, pseudo-labeled samples, and unlabeled samples. In the training process of the model, small loss selection is used to further eliminate some of the noisy labeled samples to improve the recognition accuracy of the model. Finally, in order to verify the effectiveness of the proposed method, it is compared with several state-of-the-art semi-supervised classification methods. The results show that when there are only a few labeled samples in remote sensing image scene classification, our method is always better than previous methods.

Keywords: remote sensing images; scene classification; semi-supervised classification; small loss selection

1. Introduction

With advances in drone technology and high-resolution vision sensors, remote sensing plays a key role in obtaining all data without on-site inspections [1]. Hundreds of remote sensing satellites are now in orbit, acquiring a vast amount of information about the Earth’s surface every day. In this sense, remote sensing data processing may be considered a big data problem because of the large amount of data to be processed, diversity [2–4], and generation speed. The recent emergence of cloud computing has expanded the possibilities of remote sensing. In the field of high-resolution remote sensing (HRRS) image processing, scene classification methods that can be used to solve practical problems, such as maps and monitoring land types and urban planning, have become active research hotspots [5–7].

During the past few years, deep learning models, especially convolutional neural networks (CNNs), have received extensive attention in the field of scene classification [8–10]. However, CNNs usually require a large number of high-quality labeled samples in the training phase. Unfortunately, collecting the labeled data of training scene images is time- and energy-consuming [11]. In contrast, the acquisition of unlabeled images is much easier compared to acquiring a manually annotated dataset by experts and engineers.
In this case, the semi-supervised learning (SSL) methods have been introduced to jointly utilize labeled and unlabeled data in the context of HRRS images. For example, a semi-supervised generative framework is proposed in [12], which uses a residual network (Resnet) [13] and very deep CNNs (VGG) [14] as the feature extractors, uses the co-training-based self-labeled method to select and identify unlabeled data, and uses discriminatory evaluation to enhance the classification of the confusion classes with similar visualized features. In [15], the authors propose an SSL method for HRRS classification based on CNNs and ensemble learning, the effective ResNet is adopted to extract preliminary HRRS image features, and the strategy of ensemble learning is utilized to establish discriminative image representations by exploring the intrinsic information of all labeled and unlabeled data; finally, supervised learning is performed for scene classification. Although the above methods have made progress as well in semi-supervised scene classification, they need to use network ensembles to train multiple networks instead of one. It cannot be ignored that these methods still require a certain number of labeled samples.

It is well known in the machine learning community that SSL methods based on consistency regularization [16,17] and mixing regularization [18,19] have proven to be simple while effective, achieving a number of state-of-the-art results in the field of natural images over the last few years. Consistency regularization is driven by encouraging consistent predictions that two different augmentations of the same unsupervised image should lead to similar prediction probabilities. Mixing regularization is inspired by MixUp [20], which uses a blending factor from the beta distribution to blend pairs of images and corresponding ground truth labels. Interpolation Consistency Training (ICT) [18] uses MixUp on a pair of pseudo-label unlabeled images, whose class probabilities are predicted by the exponential moving average (EMA) of the training model and through consistent regularization to ensure that the prediction results of the training model on the mixed images are the same as the EMA model. MixMatch [19] works by estimating low-extropy labels for data-augmented unlabeled examples, mixing labeled and unlabeled data using MixUp, and then training an SSL classifier to output consistent predictions about the linear interpolation of the data.

Contrary to natural images, however, HRRS images contain complex contextual relationships and large differences in object scale and are often affected by the camera angle, the direction of objects, illumination, and atmospheric conditions, which can result in high intra-class variations and in low inter-class variations [21,22]. Therefore, SSL techniques based on consistent regularization methods are unable to achieve good generalization performance in remote sensing images with few labeled samples (for example, only one or five samples per category). Moreover, as the training process deepens, the neural network will memorize the unlabeled data together with the false pseudo-labels, which affects the recognition accuracy of the model [23,24].

To cope with this problem, in this paper, we propose an early labeled and small loss selection semi-supervised learning method for aerial scene classification, namely ELSLS-SSL. Early pre-training models are used to label partially unlabeled samples for HRRS images, inspired by the early-learning regularization [23]. First, we initialize multiple independent ResNet networks with different parameters, combine the MixMatch SSL method to independently train the model through only a few epochs, and obtain multiple ResNet network models with different parameters after training, which are used to labeled unlabeled samples. Then, the pseudo-labeled unlabeled samples are divided into low-noise labeled datasets and unlabeled datasets through high-probability sample selection. Finally, SSL is carried out with the labeled data, pseudo-labeled data, and unlabeled data based on small loss selection by combination with a pseudo-labeled loss function. The experiments on the AID and NWPU-RESISC45 datasets show that by adding the pseudo-labeled data, filtered and labeled by multiple ResNet models trained in the early stage to SSL, the trained neural network can achieve higher classification accuracy.

The rest of this article is organized as follows. Section 2 introduces related work. The proposed method is described in Section 3. The experiments are described in Section 4,
which is followed by discussions of the method with further experiments in Section 5. Finally, Section 6 concludes this article.

2. Related Work

In this section, we briefly review existing works on semi-supervised learning methods and learning using noisy labels.

Semi-supervised learning is a kind of weakly supervised learning. Its main idea is to optimize the model by combining a large amount of unlabeled data with a small amount of labeled data in the process of training the model. Generally speaking, semi-supervised learning is a hybrid between supervised and unsupervised learning. It combines the advantages of the two. It can use labeled data for supervised learning, and can spontaneously generate labels for unlabeled data during the training process to optimize the model. In recent years, semi-supervised learning has made great progress. Interested readers can consult the following surveys and books [25–27].

During the past few years, Google Research has published a series of papers on semi-supervised learning methods, including MixMatch [19], ReMixMatch [28], and FixMatch [29]. MixMatch combines consistency regularization with data augmentation, entropy minimization, and MixUp. Based on MixMatch, ReMixMatch [28] adopts distribution alignment and augmentation anchoring strategies. It encourages the distribution of a model’s aggregated class predictions to match the marginal distribution of ground-truth class labels. For each given unlabeled input, it generates multiple strongly augmented versions and combines the pseudo-labels generated by the weakly augmented version to train the model; FixMatch [29] is the same as ReMixMatch in that it uses the model to predict weakly augmented unlabeled images to generate pseudo-labels. However, it will only be retained if the model produces high-confidence predictions. The model is then trained to predict pseudo-labels when inputting a strongly augmented version of the same image.

These semi-supervised learning methods have the following characteristics. A small number of labeled samples and a large number of unlabeled samples are used in the training process, and as the number of labeled samples increases, the recognition performance is significantly improved. In this case, this article screens unlabeled samples, and selects some samples for pseudo-labeling, in the hope of adding to the labeled samples, so as to improve the recognition accuracy of the model.

Considering how to screen samples is one of the main foci of this article. At the same time, in the process of semi-supervised learning, it is inevitable that pseudo-labeled samples are mislabeled, which also affects the recognition accuracy of the model. The study of false labels in training data belongs to the problem of noisy label learning. Most existing methods for training CNNs with noisy labels seek to correct the loss function. The most popular method can be understood as a relabeling method, such as modeling with directed graphical models [30], a knowledge graph [31], and improving the bootstrapping method by exploiting the dimensionality of the feature subspace [32]. The second type of method tends to clean and separate the training data and use the clean samples after separation for model training [33–35].

Since our main task is semi-supervised learning, a large number of unlabeled samples are optimized through dynamic labeling during the training process, so the above two main methods are not suitable for our task. In the training process, our method adopts a small loss selection method to dynamically select pseudo-labeled samples to filter out the incorrect samples as much as possible and improve the accuracy of the model.

3. Methodology

SSL methods aim to improve the model’s performance by leveraging unlabeled data. Current state-of-the-art SSL methods can be seen as noisy learning of pseudo-labeled data. When trained on noisy labels, deep neural networks have been observed to first fit the training data with clean labels during an early learning phase, before eventually
memorizing the examples with false labels [23]. Inspired by this idea, we propose an SSL method that uses early training models to label data. An overview of the method is shown in Figure 1, where the training procedure includes three phases: unlabeled sample labeling with early training multi-models, high-probability sample selection, and retraining. Here, \( f_1 \), \( f_2 \), and \( f_3 \) are three pre-trained networks using a labeled dataset and unlabeled dataset based on the MixMatch method in a few epochs. \( f_\theta \) is the final network retrained by using the labeled dataset, low-noise dataset, and sub-unlabeled dataset through small loss selection based on the MixMatch method.

### Figure 1. Framework of the proposed multi-screening unlabeled sample semi-supervised learning method.

#### 3.1. Early Training Multi-Models for Unlabeled Sample Labeling

For aerial scene classification, let \( \mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{N_L} \) denote the set of labeled training data, where \( x_i \) is the \( i \)-th sample, \( y_i \in \{0, 1\}^C \) is the one-hot label over \( C \) classes, and \( N_L \) is the total number of labeled samples. Similarly, the set of unlabeled data can be represented as \( \mathcal{D}_U = \{u_i\}_{i=1}^{N_U} \), where \( u_i \) is the \( i \)-th unlabeled sample, and \( N_U \) is the number of unlabeled samples. More formally, given a model with parameters \( \theta \), based on MixMatch [19], the combined loss \( \mathcal{L} \) for SSL is computed as:

\[
\mathcal{L} = \mathcal{L}_X + \lambda \mathcal{L}_U
\]

\[
\mathcal{L}_X = \frac{1}{|X'|} \sum_{x,y\in X'} H(y, p(x; \theta))
\]

\[
\mathcal{L}_U = \frac{1}{C|U'|} \sum_{u,q\in U'} \|q - p(u; \theta)\|^2_2
\]

where \( H(a,b) \) is the cross-entropy between distributions \( a \) and \( b \), \( p(x; \theta) \) is the model’s output softmax probability for class \( c \), \( X' \) and \( U' \) are transformed from a batch of labeled data and unlabeled data through MixUp [20], and \( \lambda \) are hyperparameters.

Deep networks tend to learn clean samples faster than noisy samples [36], and we assume that this phenomenon also exists in SSL. Although SSL technology can improve the generalization performance of the model to a certain extent, when there is little labeled data, the model will be affected by the noisy data in the unlabeled data. Our goal is to find a model or a combination of multiple models \( P = \{p(u; \theta_i)\}_{i=1}^{M} \) to label unlabeled samples. These models can not only learn clean data in SSL but also avoid overfitting noisy data during the early training process of the model. Then, they select relatively clean labeled samples from pseudo-labeled samples and convert them into one-hot labeled samples. Finally, the selected one-hot labeled samples are added to the SSL training process.
as low-noise labeled data, which are different from labeled data and unlabeled data, to train a new model with better generalization performance.

In order to select low-noise labeled data with high labeled quality, according to the early-learning phenomenon [23], we adopt the MixMatch SSL method, using labeled data and unlabeled data to train the model for a few epochs. However, using an independent model to select low-noise samples, and then combining low-noise samples to train a new model, may cause confirmation bias. Intuitively, two or multiple networks can filter different types of errors brought by noisy pseudo-labels since they have different learning abilities. Therefore, we use different initialization parameters and sample input order to independently and repeatedly train multiple models and use the predicted mean of these models for unlabeled samples as pseudo-labels for the samples. For an unlabeled sample \( u \in D_U \), we set

\[
\hat{y} = \frac{1}{M} \sum_{i=1}^{M} p(u; \theta_i)
\]

(4)

where \( \theta_i \) is the parameter of the early training model obtained by training only a few epochs using all data and different initialization parameters for the \( i \)-th time. Finally, we obtain the pseudo-labeled dataset \( D_{P_u} = \{(u_i, \hat{y}_i)\}_{i=1}^{N_U} \). The early training multi-models for the unlabeled sample labeling process is shown in Figure 2, where \( f_{\theta_1}, f_{\theta_2}, \) and \( f_{\theta_3} \) are three pre-trained networks using the labeled dataset and unlabeled dataset based on the MixMatch method in a few epochs. These three pre-trained models are used to pseudo-label unlabeled samples.

![Figure 2. Framework of unlabeled sample labeling with early training multi-models.](image)

**3.2. High-Probability Sample Selection**

In this section, we propose a method for screening pseudo-labeled samples. The specific process is shown in Figure 3. It uses the early training multi-model to pseudo-label the unlabeled samples, and according to the pseudo-labeling results, it sorts each category one by one according to the predicted probability. Finally, the top-ranked samples are selected as low-label noise samples to train the new model. In Section 3.1, through preliminary training, we obtained multiple scene classification models with different parameters and used these models to pseudo-label unlabeled samples. Intuitively, the pseudo-labeled dataset \( D_{P_u} = \{(u_i, \hat{y}_i)\}_{i=1}^{N_U} \) obtained in Section 3.1 contains a large amount of incorrectly labeled data and cannot be directly used to train the model. However, the performance of a CNN is better if the training data become less noisy. We aim to select some low-noise data in the pseudo-labeled dataset \( D_{P_u} \) to optimize the classification model. From the view of [35], CNNs tend to learn simple patterns first, then gradually memorize
all samples. For unlabeled data, if the pseudo-labeling results of most models are the same, they should be correctly labeled based on this observation; we select the low-noise pseudo-label samples from $D_{PU}$ as follows:

$$D_{Ps} = \arg \max_{D_{PU}} \left\{ \sum_{i=1}^{N_s} \sum_{j=1}^{C} \max(\hat{y}_i) \right\}$$

where $N_s$ is the number of samples selected in each category. In other words, in the pseudo-labeled dataset $D_{PU}$, we select the top $N_s$ samples with the highest predicted probability of $\hat{y}$ in each category to form the low-noise pseudo-labeled dataset $D_{Ps}$ and into a one-hot label.

![Figure 3. Framework of high-probability sample selection.](image)

### 3.3. Retraining and Small Loss Selection

After obtaining the low-noise pseudo-labeled dataset $D_{Ps}$, we use labeled $D_L$, low-noise pseudo-labeled $D_{Ps}$, and unlabeled datasets $D_{Pu}$ to train a new CNN model based on the MixMatch semi-supervised learning method. In order to make better use of the low-noise pseudo-labeled dataset, we rewrite the loss function as:

$$L = \mathcal{L}_X + \lambda_{Ps} \mathcal{L}_{Ps} + \lambda_U \mathcal{L}_U$$

$$\mathcal{L}_{Ps} = \frac{1}{|U_{Ps}|} \sum_{u, \hat{y} \in U_{Ps}} \text{H}(\hat{y}, p(u; \theta))$$

where $U_{Ps}$ is transformed from labeled data, low-noise pseudo-labeled data, and unlabeled data through MixUp [20], and $\lambda_{Ps}$ are hyperparameters. In order to ensure that the three datasets are mixed using MixUp, in the actual training of the model, we use the same number of labeled data and low-noise pseudo-labeled data in each mini-batch. The number of unlabeled data is the sum of the number of labeled data and low-noise pseudo-labeled data.

However, in the pseudo-labeled samples, there is still a certain amount of incorrectly labeled data. The addition of these mislabeled data to the training process can have a certain impact on the generalization performance of the model and reduce the accuracy of the model. Moreover, semi-supervised learning gradually labels unlabeled samples with the training process. This leads to the fact that the wrong pseudo-labeled samples are not always wrong during the training process, so the wrongly labeled samples cannot
be eliminated through a simple one-time screening. According to [35], small loss samples are likely to be ones that are correctly labelled. Thus, in the training process of our semi-supervised learning model, if we train our model using only small-loss pseudo-labeled samples in each batch of data, a certain number of incorrectly labeled samples would be eliminated.

In order to reduce the influence of incorrectly labeled samples on the model, we apply a small loss criterion to select relatively correct pseudo-labeled samples, as shown in Figure 4. In our algorithm, we use training loss in Equation (6) to minimize the impact of incorrectly labeled samples on the model. Specifically, there is no wrong label in the labeled sample. During the entire training process, only the last two parts of Equation (6) have error flags, and the last part \( \mathcal{L}_U \) has more wrongly labeled samples. Specifically, we conduct small-loss selection in a batch as follows:

\[
\tilde{U}'_{ps} = \arg \min_{|U''_{ps}| \geq R(t_s)} |U''_{ps}| \mathcal{L}_{ps}
\]

\[
\tilde{U}'_{pu} = \arg \min_{|U''_{pu}| \geq R(t_u)} |U''_{pu}| \mathcal{L}_U
\]

where \( R(t_s) \) and \( R(t_u) \) are thresholds to control the number of incorrectly labeled samples to be screened. \( R(t) = 1 - \min \left\{ \frac{t}{T_k}, \tau \right\} \), \( T_k \) is the total training epoch, \( t \) is the currently training epoch, and \( \tau \) is a hyperparameter. At the begining of training, we keep more small-loss data (with a large \( R(t) \)) in each batch since deep networks would fit clean data first.

After obtaining the small-loss instances, we calculate the average loss on these examples for further backpropagation:

\[
\mathcal{L}_{ps} = \frac{1}{|\tilde{U}'_{ps}|} \sum_{u, \hat{y} \in \tilde{U}'_{ps}} H(\hat{y}, p(u; \theta))
\]

\[
\mathcal{L}_U = \frac{1}{|\tilde{U}'_{pu}|} \sum_{u, \hat{y} \in \tilde{U}'_{pu}} \| q - p(u; \theta) \|^2_2
\]

4. Experimental Results

In this section, we introduce the experimental setup, including the dataset, network architecture, training setup, and metrics. Then, we compare our proposed method with some state-of-the-art approaches by using the NWPU-RESISC45 and AID datasets.
4.1. Experimental Setup

4.1.1. Dataset

Two public aerial image datasets, NWPU-RESISC45 [37] and the Aerial Image Dataset (AID) [38], are used in the experimental section. NWPU-RESISC45 is a very large-scale benchmark for remote sensing scene classification that was created by Northwestern Polytechnical University (NWPU). AID contains samples of various resolutions from different sensors, which is extremely challenging and is one of the most commonly used datasets for evaluating scene classification algorithms.

4.1.2. Network Architecture

Resnet50 pre-trained on ImageNet was used as the backbone of our network architecture. The last 1000 dimensional fully connected (FC) layer of Resnet50 was replaced by a \( C \) dimensional FC layer, while \( C \) was the number of classes for the training dataset.

4.1.3. Training Setup

For our experiments, we used a batch of 16 images and 200 batches as an epoch. The early training multiple models had only trained 10 epochs and the final model had trained 120 epochs by using the labeled dataset, low-noise pseudo-labeled dataset, and unlabeled dataset, where an Adam optimizer was employed with a learning rate of \( 3 \times 10^{-5} \) for all models. The selecting number of samples \( N_s = 4 \), and the number of early training models \( M = 3 \). All the baseline SSL methods were trained by an Adam optimizer with a \( 3 \times 10^{-5} \) learning rate by 120 epochs, and the learning rate remained constant during the training phase. Finally, we conducted 3 independent experiments on each dataset and recorded the average accuracy of each independent experiment as the final recognition accuracy of the SLL method.

All experiments were carried out on a computer equipped with an Intel CPU i7 10700k, an NVIDIA GeForce RTX 2080Ti, and 16 GB DDR4 memory. The operating system was Ubuntu 18.04 and the running software was Python 3.7.

4.1.4. Metrics

Accuracy and the confusion matrix were used as evaluation metrics in our scene classification experiment. The accuracy was calculated as the number of correctly classified samples divided by the total number of samples. The advantage of the confusion matrix is that it can clearly show all the errors between different categories and the different degrees of confusion of the model to the samples.

4.2. Experiment on NWPU-RESISC45 Dataset

Our proposed method was evaluated on two large-scale datasets. The first dataset was the NWPU-RESISC45 dataset, which contains 31,500 remote sensing images of 45 categories, extracted from Google Earth by experts in the field of remote sensing, with a spatial resolution of approximately 30 m to 0.2 m per pixel. Each scene class in NWPU-RESISC45 contains 700 images, which are set to 256 × 256 pixels in the RGB color space. Figure 5 shows an example image of each class in the NWPU-RESISC45 dataset. For the NWPU-RESISC45 dataset, we first randomly selected 20% of samples from each category as the test set. Then, in order to verify the effectiveness of our semi-supervised learning method under differently labeled samples, we randomly selected 1, 2, 3, and 5 samples in each category as the labeled dataset, and the remaining samples as the unlabeled dataset.
The proposed method was compared with other state-of-the-art methods, including label propagation [15], EL + LR [15], Mean-teacher [16], ICT [18], and MixMatch [19]. Label propagation is a typical graph-based semi-supervised method, which can be flexibly migrated for different tasks and has good performance. EL + LR adopts ensemble learning (EL) to establish discriminative image representations by exploring the intrinsic information of all available data, and uses supervised learning to perform logistic regression (LR)-based scene classification. Mean-teacher [16] applies the moving average of model parameters to the teacher model, generates proxy labels for each unlabeled sample, and calculates consistency loss and supervision loss. Based on Mean-teacher, ICT [18], and MixMatch [19], we used MixUp mixed data in the training process of the semi-supervised learning of the model to improve the accuracy of the model. We experimented with the above method under the same training sample. Figures 6 and 7 and Table 1 show all the experimental results.
Figure 6. Confusion matrix obtained by our proposed method on NWPU-RESISC45 testing set.

Figure 7. Confusion matrix obtained by MixMatch on NWPU-RESISC45 testing set.
Table 1. Recognition accuracy (%) of label propagation, EL + LR, Mean-teacher, ICT, MixMatch, and our proposed method with differently labeled data for each category on NWPU-RESISC45 dataset.

| Method                  | NWPU-RESISC45 |
|-------------------------|---------------|
|                         | Num of Labeled Data for Each Category | 1 | 2 | 3 | 5 |
| Label Propagation [15]  | 20.56 | 38.66 | 53.42 | 69.37 |
| EL + LR [15]            | 25.36 | 40.62 | 68.57 | 70.82 |
| Mean-Teacher [16]       | 20.01 | 34.33 | 40.31 | 51.34 |
| ICT [18]                | 40.53 | 62.47 | 70.96 | 77.09 |
| MixMatch [19]           | 46.46 | 66.66 | 74.26 | 83.76 |
| R²S [24]                | 48.16 | 72.34 | 81.03 | 86.61 |
| Ours                    | 52.91 | 78.26 | 85.17 | 90.34 |

As can be seen from Table 1, in the case of a small number of labeled samples, the model recognition accuracy of all methods is greatly improved with the increase in labeled samples. The sufficient diversity of the NWPU-RESISC45 dataset and the characteristics of the variants have brought enough challenges to the accuracy of the model. Among all the comparison methods, only our method has an accuracy rate of more than 90% when there are five labeled samples in each category. Figures 6 and 7 show the confusion matrix obtained on NWPU-RESISC45 by our method and the MixMatch method, respectively, with three labeled samples for each category. From 45 different scene classes from a large confusion matrix, the problem of mislabeled samples occurs more frequently. It can be seen from Figure 7 that the accuracy of the MixMatch method for scene image recognition in several categories of airport, church, freeway, medium residential, palace, tennis court, and wetland is very low, and the accuracy of individual scenes is almost 0. However, after optimization by our proposed method, under the same training sample conditions, the model has greatly improved the recognition accuracy of several scene categories: airport, church, freeway, medium residential, and tennis courts. Among them, the improvement of medium residential and tennis courts is particularly apparent. However, the model still has obvious deficiencies in the identification of individual categories (palaces), and most of the palace samples are identified as commercial areas. We believe that this is due to the small number of labeled samples and the random selection of labeled samples for each category. In this case, the random sample selected for a single category is not representative, resulting in low classification accuracy. To verify this conjecture, we randomly selected samples in the same way and retrained the model. Experiments showed that, under this sample selection condition, there are always some categories that have low recognition accuracy. In future research, we will consider optimizing the classification effect by increasing the selection of a single sample for each category. In general, it is worthy of affirmation that the comparison between Figures 6 and 7 can fully show that our method has a higher accuracy rate in almost all categories compared to the MixMatch method.

4.3. Experiment on AID Dataset

The second dataset was the Aerial Image Dataset (AID). It has a number of 10,000 images and is divided into 30 classes, which are collected from Google Earth imagery, with the pixel resolution changing from approximately half a meter to 8 m, and the size was fixed as 600 × 600 pixels. The number of images in each category varies from 220 to 420. Figure 8 shows one example image in the AID dataset for each class. Table 2 shows the detailed information of the image numbers in each semantic class for the AID dataset. The samples in AID are collected from different remote sensing sensors, so the samples come from multiple sources. The pixel resolution of the samples in the dataset has changed from 8 m to approximately 0.5 m. Each image has a fixed size to cover scene categories of different resolutions, which increases the difficulty for the model to classify the sample scenes. For the AID dataset, we first randomly selected 70 samples from each category,
with a total of 2100 samples as the test set. Then, in order to verify the effectiveness of our semi-supervised learning method under differently labeled samples, as with the experiment on the NWPU-RESISC45 dataset, we randomly selected 1, 2, 3, and 5 samples in each category as the labeled dataset, and the remaining samples as the unlabeled dataset.

The proposed method was compared with other state-of-the-art methods, including label propagation [15], EL + LR [15], Mean-teacher [16], ICT [18], and MixMatch [19]. The results are given in Table 3. As can be seen from Table 3, for the AID dataset, our method performs much better than its comparisons with all labeled samples per category. From the perspective of recognition accuracy, our method has an accuracy of more than 90% in the case of three and five labeled samples in each category. Among the comparison methods, only the MixMatch method has an accuracy rate exceeding 90% when there are five labeled samples in each category. This is sufficient to show the superior performance of our method in the case of a small number of labeled samples.
Table 2. Different semantic categories and corresponding number of images in each type of AID dataset.

| Datasets   | Types         | Num | Types          | Num | Types        | Num |
|------------|---------------|-----|----------------|-----|--------------|-----|
| AID        | Airport       | 360 | Bare land      | 310 | Baseball field | 220 |
|            | Beach         | 400 | Bridge         | 360 | Center       | 260 |
|            | Church        | 240 | Commercial     | 350 | Dense        | 410 |
|            | Desert        | 300 | Farmland       | 370 | Forest       | 250 |
|            | Industrial    | 390 | Meadow         | 280 | Medium residential | 290 |
|            | Mountain      | 340 | Park           | 350 | Parking      | 390 |
|            | Play ground   | 370 | Pond           | 420 | Port         | 380 |
|            | Railway station | 260 | Resort        | 290 | River        | 410 |
|            | School        | 300 | Sparse residential | 300 | Square       | 330 |
|            | Stadium       | 290 | Storage tanks  | 360 | Viaduct      | 420 |

Table 3. Recognition accuracy (%) of label Propagation, EL + LR, Mean-teacher, ICT, MixMatch, $R^2S$, and our proposed method with different labeled data for each category on AID dataset.

| Method          | AID |
|-----------------|-----|
|                 | Num of Labeled Data for Each Category |
|                 | 1   | 2   | 3   | 5   |
| Label Propagation [15] | 31.24 | 40.21 | 65.71 | 73.42 |
| EL + LR [15]    | 29.27 | 45.32 | 73.63 | 79.41 |
| Mean-Teacher [16] | 19.38 | 31.31 | 40.02 | 51.66 |
| ICT [18]        | 44.70 | 69.98 | 80.33 | 85.24 |
| MixMatch [19]   | 48.66 | 74.72 | 85.80 | 91.63 |
| $R^2S$ [24]     | 54.13 | 78.92 | 89.13 | 91.82 |
| Ours            | 61.60 | 83.15 | 91.06 | 94.40 |

To further show the effectiveness of our proposed methods, Figure 9 shows the confusion matrix obtained on AID by our method, with three labeled samples for each category. The recognition accuracy comparison of each category of our method and MixMatch when there are three labeled data for each category is shown in Figure 10, which details the improvement of our method relative to the accuracy of MixMatch in each category. As shown in Figures 9 and 10, our method has a certain improvement in the recognition accuracy of almost every category. However, as with the NWPU-RESISC45 dataset, for individual categories, the recognition effect of our algorithm is not satisfactory. It can be seen from Figure 9 that the algorithm has a very poor recognition effect on the category of schools. From the comparison results in Figure 10, the recognition accuracy of our method for schools is slightly lower than that of MixMatch. In Figure 9, the model’s recognition accuracy of farmland and forest is 96% and 100%, respectively. The model has a very low recognition accuracy rate for schools, and most of the school samples are classified as medium residential by the model. From Figure 10, it can be found that the difference between farmland and forest is very large, while the difference between school and medium residential is very small. This shows that our model can fully learn and extract the characteristics of the two types of samples of farm and forest, and accurately classify the two types of samples when the labeled samples are limited. However, for schools and medium residential areas, with small differences between classes, there are still certain shortcomings, making the classification effect unsatisfactory. In follow-up research, we will explore ways to improve the model’s recognition of individual category samples.
Figure 9. Confusion matrix obtained by our proposed method on AID testing set.

Figure 10. Precision comparison for each class of our proposed method and the MixMatch method on AID dataset, where the Y-axis denotes per class classification accuracy improvement of our proposed method relative to MixMatch, and the X-axis denotes the class index of each category.

5. Discussion

In this section, we discuss the results of our method on two datasets. Finally, through sensitivity analysis, we analyze the effects of the number of training epochs for the early
training of multiple models \( N_E \), the number of early training models \( M \), and the number of samples of each category \( N_s \) selected from the pseudo-labeled dataset on the performance of the final model.

5.1. Discussion of Experimental Results

Experiments on two datasets show that our method can greatly improve the recognition performance of the semi-supervised learning model in scene classification in the case of a small number of labeled samples. It is undeniable that there are still poor recognition effects of individual categories in the two experiments. This is in contrast to the remote sensing image scene classification task, where some categories have a higher degree of similarity between classes and a greater degree of intra-class difference. Figure 11 shows samples of two categories, commercial areas and places in the NWPU-TESISC45 dataset, and samples of two categories in the AID dataset, medium residential and schools, and each category shows four samples. It can be seen that the inter-class differences between these samples are extremely low. This causes the model to produce errors in pseudo-labeling unlabeled data, and the existing semi-supervised learning methods (including our method) fail to detect such pseudo-labeling errors in time, which affects the model’s recognition accuracy of individual categories of samples.

![Figure 11. Sample images of some categories in NWPU-TESISC45 and AID datasets.](image)

Moreover, according to the experimental results, it is not difficult to find that, in the experiments on the two datasets, the accuracy of the Mean-teacher method is very low. This is also strongly related to the small number of labeled samples. However, the MixMatch method and the ICT method have a significant improvement over the Mean-teacher method. Through the overall comparison of these three methods, it is not difficult to see that MixUp has a great advantage in improving the accuracy of the model for the mixing of samples in the semi-supervised learning algorithm. Through sample mixing, the labels of pseudo-labeled samples can be smoothed so as to reduce the impact of incorrectly labeled samples on the performance of the model to a certain extent. This also shows that incorrectly labeled samples will have a certain impact on the model, especially when the number of labeled samples is very small. Thus, this also proves that our method has a positive effect on improving the classification performance of the model by adding some low-noise labeled samples and screening out some of the mislabeled samples with large losses.

5.2. Influence of Parameters on Performance of Proposed Method

In this section, the AID dataset is used as an example to analyze the influences of three important parameters, namely \( N_E \), \( M \), and \( N_s \), on the performance of the final training model under the condition of two labeled samples for each category.

Table 4 shows the changes in accuracy with the \( N_E \) (the number of training epochs for early training of multiple models) changing over a wide range of values when the other two are fixed, \( M = 3 \) and \( N_s = 4 \). It can be seen from Table 4 that the performance of the model is best when \( N_E = 10 \). Our method is to train multiple models in the first
stage, and then divide the unlabeled samples by these multiple models and select a small part of the samples to simply label them as low-noise labeled samples. Through labeled samples, low-noise labeled samples, and unlabeled samples, the model is retrained in a semi-supervised learning manner. The model obtained in the second stage is the final classification model. The $N_E$ number in Table 4 represents the number of early multi-model training and has nothing to do with the number of epochs of the second-stage model training. Our results show that when $N_E$ is smaller, the recognition result of the final model is better. After a few epochs of simple training, the models first learn some simple samples, and they learn for each category. With the deepening of the model training level, in the case of small labeled samples, the model tends to classify most unlabeled samples into several easy-to-learn categories, which affects the generalization performance of the model. Therefore, the number of iterations $N_E$ for early training of the multi-sample model should not be too great.

Table 4. Recognition accuracy (%) of the aerial scene classification using different numbers of training epochs $N_E$ for early training of three models.

| $N_E$ | 10   | 20   | 30   | 40   |
|-------|------|------|------|------|
| Ours  | 83.15| 81.02| 79.75| 78.65|

For $N_S$ shown in Table 5, we can see that the number of low-noise samples selected by our method through the pre-training model is not as large as possible. When the number of samples selected by the model for each category is less than four, the performance of the model increases as the number of selected samples increases. When the model selects too many samples for each sample, there are too many falsely labeled samples in the selected pseudo-labeled samples, resulting in too much labeling noise of the pseudo-labeled samples, which affects the classification performance of the model. Therefore, for our method, the best classification performance can be achieved when four pseudo-labeled samples are selected for each category. At the same time, when the number of selected samples is 0, it can be regarded as an ablation experiment using only the small loss selection method. The results show that when only the small loss selection method is used, the classification accuracy is 75.87%, which is better than the MixMatch method. This proves that our small loss selection method is effective.

Table 5. Recognition accuracy (%) of the aerial scene classification using different selected numbers of samples $N_s$ for each category.

| $N_s$ | 0     | 1     | 2     | 3     | 4     | 5     | 6     |
|-------|-------|-------|-------|-------|-------|-------|-------|
| Ours  | 75.87 | 77.57 | 78.36 | 80.45 | 83.15 | 81.25 | 68.70 |

For $M$ shown in Table 6, the four early training models have the best results, and the three models yield the second-best. In order to reduce computing resources and ensure that the performance of the three models does not differ from that of the four models, we used the three models to label unlabeled data in the comparison experiment.

Table 6. Recognition accuracy (%) of the aerial scene classification using different numbers of early training models $M$.

| $M$ | 1     | 2     | 3     | 4     | 5     |
|-----|-------|-------|-------|-------|-------|
| Ours| 76.74 | 80.12 | 83.15 | 83.20 | 80.75 |

5.3. Discussion of the Computational Complexity

In this section, we discuss the computational complexity of our method and other semi-supervised learning methods. The main configuration of the computer used has
been explained in the experimental section. In the same environment, we intuitively compared the computational complexity by calculating the time spent in training and verification of different methods. The results are shown in Table 7. It can be seen from Table 7 that, compared with other comparison methods, our method is in the same order of magnitude as other methods in terms of time complexity and space complexity. Since our method involves the early training and screening of samples during the training process, the training time is increased, but compared with other methods, the overall training time is not significantly greater than other methods. The model is mainly used for image recognition, i.e., the model test time is more reasonable in actual use. All methods in this article are implemented through the Resnet50 model, so the test time is basically the same. Regarding the average time of each image, they all take around 4.60 ms.

| Method          | Training Time per Epoch | Testing Time per Image |
|-----------------|-------------------------|------------------------|
| Mean-teacher    | 98.04 s                 | 4.60 ms                |
| ICT             | 102.55 s                | 4.72 ms                |
| MixMatch        | 112.60 s                | 4.58 ms                |
| $R^2S$          | 123.46 s                | 4.76 ms                |
| Ours            | 131.87 s                | 4.83 ms                |

6. Conclusions

In this paper, we have presented an early labeled and small loss selection semi-supervised learning method to reduce the demand for labeled samples in remote sensing image scene classification. A simple method is used to select unlabeled data labeled with early pre-training models that only train in a few epochs, and the selected pseudo-labeled data are combined with labeled data and unlabeled data to train a new classification model under the small loss selection. This method can greatly improve the classification performance of the model. The experimental results on the AID and the NWPU-RESISC45 datasets show the superior performance of our method. In the experiment, we also found that, for a very small number of remote sensing image samples, because the difference between remote sensing image categories is not obvious, the existing semi-supervised learning methods cannot classify well in the case of a small number of labeled samples. In future research, we hope to explore the use of active learning to improve the generalization performance of the model for a very small number of samples in the case of selectively labeling a small number of samples in order to further improve the classification accuracy of the model and try to migrate the proposed method to polarimetric SAR images.

Author Contributions: Conceptualization, Data curation, Formal analysis, Methodology, Writing—original draft, Writing—review and editing: Y.T.; Writing—review and editing, Resources, Validation: Y.D.; Funding acquisition, Validation: G.Y. All authors have read and agreed to the published version of the manuscript.

Funding: Our research fund is funded by the Natural Science Foundation of Heilongjiang Province No. F2018006.

Institutional Review Board Statement: The research does not involve humans or animals.

Informed Consent Statement: The research does not involve humans.

Data Availability Statement: The experiment in this paper uses public datasets, so no data are reported in this work.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.
References

1. Abdelwahab, S.; Hamdaoui, B.; Guizani, M.; Rayes, A. Enabling Smart Cloud Services Through Remote Sensing: An Internet of Everything Enabler. IEEE Internet Things J. 2014, 1, 276–288. [CrossRef]

2. Dong, Y.; Zhang, Q. A Combined Deep Learning Model for the Scene Classification of High-Resolution Remote Sensing Image. IEEE Geosci. Remote Sens. Lett. 2019, 16, 1540–1544. [CrossRef]

3. Qin, R.; Fu, X.; Lang, P. PolSAR Image Classification Based on Low-Frequency and Contour Subbands-Driven Polarimetric SENet. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2020, 13, 4760–4773. [CrossRef]

4. Pallotta, L.; Clemente, C.; De Maio, A.; Soraghan, J.J. Detecting Covariance Symmetries in Polarimetric SAR Images. IEEE Trans. Geosci. Remote Sens. 2017, 55, 80–95. [CrossRef]

5. Zhang, W.; Tang, P.; Zhao, L. Remote sensing image scene classification using CNN-CapsNet. Remote Sens. 2019, 11, 494. [CrossRef]

6. Cheng, G.; Li, Z.; Yao, X.; Guo, L.; Wei, Z. Remote Sensing Image Scene Classification Using Bag of Convolutional Features. IEEE Geosci. Remote Sens. Lett. 2017, 14, 1735–1739. [CrossRef]

7. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.S. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2020, 13, 3735–3756. [CrossRef]

8. Dede, M.A.; Aptoula, E.; Genc, Y. Deep network ensembles for aerial scene classification. IEEE Geosci. Remote Sens. Lett. 2018, 16, 732–735. [CrossRef]

9. Liu, Y.; Liu, Y.; Ding, L. Scene classification based on two-stage deep feature fusion. IEEE Geosci. Remote Sens. Lett. 2017, 14, 183–186. [CrossRef]

10. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. IEEE Trans. Geosci. Remote Sens. 2018, 56, 2811–2821. [CrossRef]

11. Nogueira, K.; Penatti, O.A.; Dos Santos, J.A. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recognit. 2017, 61, 539–556. [CrossRef]

12. Han, W.; Feng, R.; Wang, L.; Cheng, Y. A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification. ISPRS-J. Photogramm. Remote Sens. 2018, 145, 23–43. [CrossRef]

13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

14. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.

15. Dai, X.; Wu, X.; Wang, B.; Zhang, L. Semisupervised scene classification for remote sensing images: A method based on convolutional neural networks and ensemble learning. IEEE Geosci. Remote Sens. Lett. 2019, 16, 869–873. [CrossRef]

16. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 1195–1204.

17. Oliver, A.; Odena, A.; Raffel, C.A.; Cubuk, E.D.; Goodfellow, I. Realistic evaluation of deep semi-supervised learning algorithms. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, QC, Canada, 3–8 December 2018; pp. 3235–3246.

18. Verma, V.; Lamb, A.; Kannala, J.; Bengio, Y.; Lopez-Paz, D. Interpolation consistency training for semi-supervised learning. In Proceedings of the 2019 International Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 3635–3641.

19. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C.A. MixMatch: A holistic approach to semi-supervised learning. In Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; pp. 5049–5059.

20. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, 30 April–3 May 2018.

21. Li, E.; Xia, J.; Du, P.; Lin, C.; Samat, A. Integrating Multilayer Features of Convolutional Neural Networks for Remote Sensing Scene Classification. IEEE Trans. Geosci. Remote Sens. 2017, 55, 5653–5665. [CrossRef]

22. Pires de Lima, R.; Marfurt, K. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. Remote Sens. 2020, 12, 86. [CrossRef]

23. Liu, S.; Niles-Weed, J.; Razavian, N.; Fernandez-Granda, C. Early-Learning Regularization Prevents Memorization of Noisy Labels. arXiv 2020, arXiv:2007.00151.

24. Tian, Y.; Li, J.; Zhang, L.; Sun, J.; Yin, G. Deep residual learning for image recognition. In Proceedings of the the CAAI International Conference on Artificial Intelligence (CICAI), Hangzhou, China, 29–30 May 2021.

25. Zhu, X.J. Semi-Supervised Learning Literature Survey; Department of Computer Sciences, University of Wisconsin-Madison: Madison, WI, USA, 2005.

26. Van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. Mach. Learn. 2020, 109, 373–440. [CrossRef]

27. Chapelle, O.; Scholkopf, B.; Zien, A. Semi-supervised learning (chapelle, O. et al., eds.; 2006) [book reviews]. IEEE Trans. Neural Netw. 2009, 20, 542. [CrossRef]

28. Berthelot, D.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Sohn, K.; Zhang, H.; Raffel, C. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv 2019, arXiv:1911.09785.
29. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.L. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *arXiv* 2020, arXiv:2001.07685.

30. Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; Wang, X. Learning from massive noisy labeled data for image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 7–12 June 2015; pp. 2691–2699.

31. Li, Y.; Yang, J.; Song, Y.; Cao, L.; Luo, J.; Li, L.J. Learning from noisy labels with distillation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1910–1918.

32. Ma, X.; Wang, Y.; Houle, M.E.; Zhou, S.; Erfani, S.; Xia, S.; Wijewickrema, S.; Bailey, J. Dimensionality-driven learning with noisy labels. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 3355–3364.

33. Arazo, E.; Ortego, D.; Albert, P.; O’Connor, N.; McGuinness, K. Unsupervised label noise modeling and loss correction. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 312–321.

34. Thulasidasan, S.; Bhattacharya, T.; Bilmes, J.; Chennupati, G.; Mohd-Yusof, J. Combating label noise in deep learning using abstention. *arXiv* 2019, arXiv:1905.10964.

35. Wei, H.; Feng, L.; Chen, X.; An, B. Combating noisy labels by agreement: A joint training method with co-regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13726–13735.

36. Li, J.; Socher, R.; Hoi, S.C. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In Proceedings of the International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May 2019.

37. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* 2017, 105, 1865–1883. [CrossRef]

38. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* 2017, 55, 3965–3981. [CrossRef]