applied sciences

MDPI

Article

Exploiting Hierarchical Label Information in an Attention-Embedding, Multi-Task, Multi-Grained, Network for Scene Classification of Remote Sensing Imagery

Peng Zeng [1], Shixuan Lin [2,*], Hao Sun [3] and Dongbo Zhou [2]

[1] Hunan Institute of Land and Resources Planning, Changsha 410007, China
[2] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China
[3] School of Computer, Central China Normal University, Wuhan 430079, China
* Correspondence: lynnlsx@mails.ccnu.edu.cn

Abstract: Remote sensing scene classification aims to automatically assign proper labels to remote sensing images. Most of the existing deep learning based methods usually consider the interclass and intraclass relationships of the image content for classification. However, these methods rarely consider the hierarchical information of scene labels, as a scene label may belong to hierarchically multi-grained levels. For example, multi-grained level labels may indicate that a remote sensing scene image may belong to the coarse-grained label "transportation land" while also belonging to the fine-grained label "airport". In this paper, to exploit hierarchical label information, we propose an attention-embedding multi-task multi-grained network (AEMMN) for remote sensing scene classification. In the proposed AEMMN, we add a coarse-grained classifier as the first level and a fine-grained classifier as the second level to perform multi-task learning tasks. Additionally, a gradient control module is utilized to control the gradient propagation of two classifiers to suppress the negative transfer caused by the irrelevant features between tasks. In the feature extraction portion, the model uses an ECA module embedding Resnet50 to extract effective features with cross-channel interaction information. Furthermore, an external attention module is exploited to improve the discrimination of fine-grained and coarse-grained features. Experiments were conducted on the NWPU-RESISC45 and the Aerial Image Data Set (AID), and the overall accuracy of the proposed AEMMN is 92.07% on the NWPU-RESISC45 dataset and reached 94.96% on the AID. The results indicate that hierarchical label information can effectively improve the performance of scene classification tasks when categorizing remote sensing imagery.

Keywords: remote sensing imagery; scene classification; multi-task learning; attention mechanism

Citation: Zeng, P.; Lin, S.; Sun, H.; Zhou, D. Exploiting Hierarchical Label Information in an Attention-Embedding, Multi-Task, Multi-Grained, Network for Scene Classification of Remote Sensing Imagery. *Appl. Sci.* **2022**, *12*, 8705. https://doi.org/10.3390/app12178705

Academic Editor: Krzysztof Koszela

Received: 29 July 2022
Accepted: 28 August 2022
Published: 30 August 2022

1. Introduction

Remote sensing scene classification aims to automatically assign labels to remote sensing images and plays a huge role in land resource management, urban planning, etc. [1,2]. With the development of sensors and lens technology, remote sensing images have a high resolution, which means that the objects in the scene are clear and have abundant spatial detail information [3]. This means that remote sensing images have the characteristics of intraclass similarity and interclass diversity [4]. Because of this, objects in the same scene may vary in size and orientation, and different scenes may contain the same features due to different shooting angles and flight altitudes [5]. This creates challenges during scene classification.

There are three kinds of scene classification methods. The first kind is based on shallow-level features, such as color and texture. For these methods, oriented gradients [6] and local binary patterns [7] are commonly used feature descriptors. However, these methods perform poorly when the spatial distribution of the objects in the scene is not uniform or when the structure is complex [8]. The second kind is based on midlevel features, which

extract discriminative feature representations by encoding the shallow-level features. The bag of visual words [9], vectors of locally aggregated descriptors [10], and fisher vectors [11] are typical encoding methods. However, these methods still rely on hand-crafted features, which make it difficult for these methods to represent the semantic information of complex scenes [12].

Deep learning based scene classification methods are the third kind of scene classification method. Commonly used deep networks implemented for remote sensing scene classification include convolutional neural networks (CNNs) [13], graph convolutional networks (GCNs) [14], and generative adversarial networks (GANs) [15]. Among them, CNN is the most widely used. A CNN can effectively mine the abstract and discriminative semantic features of remote sensing images, and CNNs show a strong learning ability in remote sensing scene classification tasks. The basic CNN network structures are ResNet [16], ShuffleNet [17], and EfficientNet [18], among others. In recent years, many studies have performed deep feature fusion on the basis of these networks to achieve better performance during remote sensing scene classification [19,20]. Some methods improve the classification ability by fusing the global and local features of remote sensing images [21–23], and some methods extract the complementary features of different structures by combining multiple CNN networks [24]. Some studies use fine-tuned CNNs as feature extractors, and in the case of small datasets, the classification ability of fine-tuned networks is better than that of retrained networks [25]. Tian et al. [26] designed a scene classification network that can switch between small networks and deep networks according to the sample complexity and computational resource constraints. Wang et al. [27] introduced an attention mechanism and designed an ARCNet network that is able to focus on key information. These methods are fine-tuned on the basis of pretrained CNNs to achieve better performance. There are also some models that can be trained from scratch. Chen et al. [28] used knowledge distillation in a lightweight CNN to improve model performance. Additionally, Zhang et al. [29] added dilated convolution and channel attention to MobileNetV2. The above methods usually consider the interclass and intraclass relationships of the image content for classification. However, these methods do not take advantage of the hierarchical information provided by scene labels.

In this paper, the hierarchical information among scene labels refers to scene labels that may belong to hierarchically multi-grained levels. As shown in Figure 1, the multi-grained level labels indicate that a remote sensing scene image belongs to the coarse-grained label “transportation land” as well as to the fine-grained label “airport”. There are hierarchical connections and shared information between the two levels of labels. With reasonable use, the two labels can complement each other and promote the learning of the corresponding classifiers. As shown in Figure 1, the categories of the coarse-grained labels include “transportation land”, “public land”, “residential”, etc.; and the categories of the fine-grained labels include “airport”, “parking”, “school”, “square”, “dense residential”, etc. The coarse-grained labels include one or more fine-grained labels. For example, the coarse-grained label “residential” consists of three fine-grained labels: “dense residential”, “medium residential”, and “sparse residential”. The logical relationship between coarse- and fine-grained labels means that they contain common features. Some features that are difficult to learn in fine-grained classifiers can be easily learned in coarse-grained classifiers, and the model’s attention can be focused on features with greater influence [30].

To make full use of hierarchical information among multi-grained scene labels, this paper proposes an attention-embedding, multi-task, multi-grained, network (AEMMN) that uses a coarse-grained classifier as the first level and a fine-grained classifier as the second level to perform multi-task learning. Multi-task learning can improve the discrimination of the common features. A gradient control module is designed to control the gradient propagation of two classifiers so as to suppress the negative transfer caused by irrelevant features between tasks. In the feature extraction part, the proposed method uses an ECA module embedding Resnet50 to extract effective features with cross-channel interaction

information. Furthermore, the external attention modules are exploited to improve the discrimination of fine-grained and coarse-grained features.

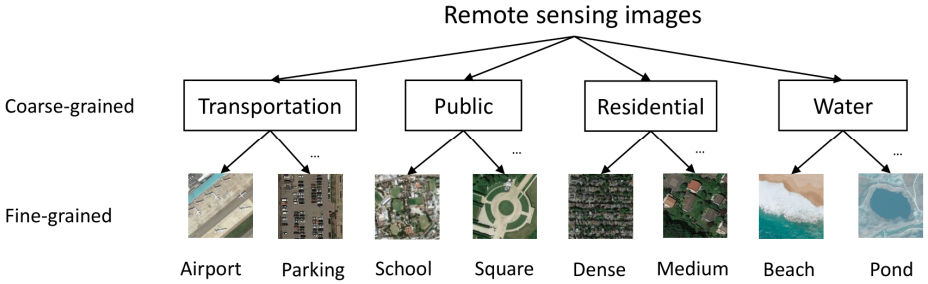

Figure 1. A multi-grained label diagram. Multi-grained labels include coarse-grained and fine-grained labels. A coarse-grained label is composed of several fine-grained labels.

Our contributions are as follows: (1) Different from traditional single-grained classification methods, this paper proposes a multi-grained classification scheme that starts from the coarse-grained level before moving to the fine-grained level. (2) This paper proposes an attention-embedding, multi-task, multi-grained, network to explore the hierarchical information in scene labels to improve the discrimination of common features. (3) A gradient control module is designed to facilitate the positive interaction of two-grained labels. (4) The experimental results show that by separating coarse-grained and fine-grained feature learning-gradient backpropagation the model can achieve excellent performance on remote sensing scene classification tasks.

2. Materials and Methods

In this paper, we propose an attention-embedding, multi-task, multi-grained, network (AEMMN) to exploit the inherent hierarchical information in multi-grained labels. The overall structures of the proposed AEMMN are shown in Figure 2. First, Resnet50 with the efficient channel attention (ECA) module is utilized to extract common features. Then the external attention modules are employed to extract the coarse-grained features and fine-grained features from the common features. Finally, the fine-grained features are input into the fine-grained classifier. The coarse-grained and fine-grained features are jointly input into the coarse-grained classifier. Additionally, to suppress the negative transfer caused by irrelevant features between tasks, a gradient control module is utilized to control the gradient propagation of the two classifiers. The classifier gradient flow only propagates along its own gradient dimension during backpropagation. The final scene classification result is determined by the fine-grained classifier, and the fine-grained and coarse-grained classifiers interact through the loss value and the parameters of the common part.

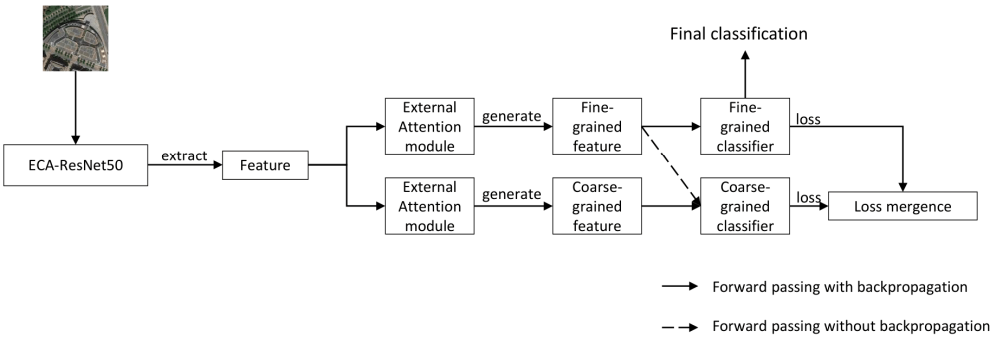

Figure 2. The overall structures of the proposed AEMMN. AEMMN consists of the feature extraction structure ECA-ResNet50, an external attention module, and coarse- and fine-grained classifiers.

2.1. Multi-Task Multi-Grained Network

The proposed AEMMN utilizes multi-task learning to exploit multi-grained labels to improve classification performance. Multi-task learning is a kind of joint learning in which multiple tasks are learned in parallel and the results affect each other. Relevant parts in tasks enable the network to focus on learning common useful features, and irrelevant parts can interact to escape local extrema [30–32]. There are two methods of multi-task learning parameter sharing: hard parameter sharing and soft parameter sharing. Hard parameter sharing is achieved by sharing the hidden layers among all of the tasks, while retaining several task-specific output layers. In soft parameter sharing, each task has its own parameters and models, and the distances between model parameters are regularized in order to encourage parameter similarity [33]. To reduce the number of parameters, we use hard parameter sharing in AEMMN.

To exploit the hierarchical information among multi-grained labels to, in turn, improve the discrimination of common features, coarse-grained classification tasks are designed to extract coarse-grained features, and fine-grained classification tasks are built to extract fine-grained features. The logical hierarchical relationship between coarse-grained labels and fine-grained labels makes coarse- and fine-grained classification tasks share common features. Multi-task learning is utilized to improve the discrimination of the common features. The addition of coarse-grained information allows fine-grained classifiers to better focus on important features, and some of the features that are difficult to learn in fine-grained classifiers can be easily learned from coarse-grained information. The equation to calculate the loss function after adding the coarse-grained classifier is as follows:

$$\text{Loss} = \alpha L_1 + \beta L_2 \tag{1}$$

where α and β determine the training tendency; the larger the value of α, the more inclined the training is to the coarse-grained classifier, and the larger the value of β, the more inclined the training is to the fine-grained classifier. L_1 is the cross entropy of the coarse-grained classifier, and L_2 is the cross entropy of the fine-grained classifier. The formulas for L_1 and L_2 calculation for a single sample are as follows:

$$L_1 = -\Sigma_{j=1}^{T} y_{1j} log P_{1j} \tag{2}$$

$$L_2 = -\Sigma_{j=1}^{T} y_{2j} log P_{2j} \tag{3}$$

where y_{1j} and y_{2j} are the jth value of the output vector y. P_{1j} and P_{2j} are the probability that this sample belongs to the jth category. T is the number of categories.

However, in multi-task learning, there may be irrelevant features between tasks, which can easily cause negative transfer and can affect the classification effect. Therefore, we propose a gradient control module to alleviate negative transfer and to make full use of different hierarchical label information.

2.2. Gradient Control Module

Since there are related and irrelevant features in multi-task learning, some of the features that are useful in coarse-grained classifiers may be the noise in fine-grained classifiers. This noise causes negative transfer to affect the classification results, while related features will transfer positively and will improve the classification results. Therefore, it is necessary to inhibit negative transfer and to encourage positive transfer [34,35]. We achieve positive transfer by controlling gradient propagation, allowing fine-grained features to participate in coarse-grained predictions. Specifically, the external attention module is utilized to generate the coarse-grained feature $f1$ and the fine-grained feature $f2$. The $f2$ is the input for the fine-grained classifier, and $f1$ and $f2$ are the inputs for the coarse-grained classifier and are used for joint predictions. Both $f1$ and $f2$ are one-dimensional vectors, where $f1 \in \mathbb{R}^{n1}$ and $f2 \in \mathbb{R}^{n2}$. The selection of n1 and n2 has an impact on the experimental effect, and n2 needs n1 to have a larger value to keep fine-grained classification the main

task of the model. Here, when n1 is 100 and n2 is 500, experiments show that the overall model can better use coarse-grained information to improve the accuracy of fine-grained classifiers. To calculate the loss value of the corresponding grain classification, $f1$ and $f2$ are used for the classification of the coarse-grained and fine-grained classifiers, respectively.

In order to prevent fine-grained features from being biased towards coarse-grained recognition during learning, we introduced a gradient control module. The classifier gradient only propagates along its own features during the backpropagation process. The specific formula is as follows:

$$Input1 = \text{CONCAT}(f1, \Gamma(f2)) \tag{4}$$

$$Input2 = f2 \tag{5}$$

where $\Gamma(\cdot)$ means that the feature only participates in forward propagation and not in backpropagation. $Input1$ is the input feature of the coarse-grained classifier, and $Input2$ is the input feature of the fine-grained classifier.

2.3. Efficient Channel Attention Module

To provide the multi-grained classifiers with more effective common features, we embedded the ECA module into ResNet50 [34]. The ECA module can combine the information of each channel and its adjacent k channels without reducing the dimensionality, and its structure is shown in Figure 3. Alternatively, it realizes the information interaction of the adjacent k channels through a one-dimensional convolution with a kernel size of k [36]. In this paper, k is set to 5.

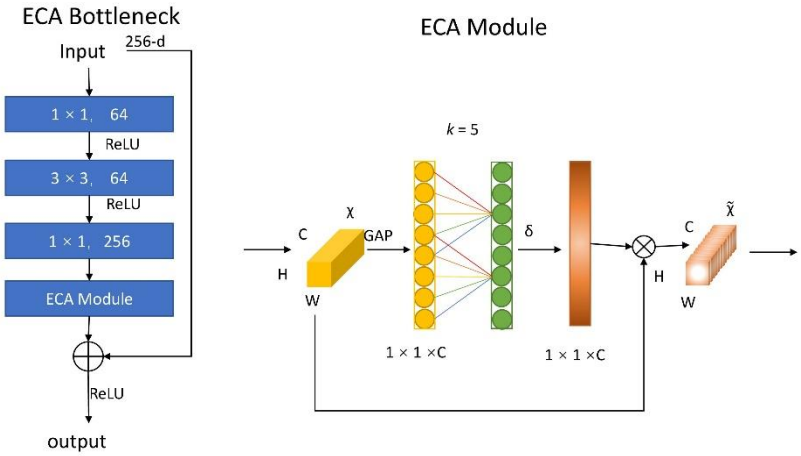

Figure 3. ECA-ResNet bottleneck structure diagram and ECA module structure diagram, where GAP is the global average pooling, and δ is the ReLU activation function. $X \in \mathbb{R}^{H \times W \times C}$ and $\widetilde{X} \in \mathbb{R}^{H \times W \times C}$.

The ECA module only considers the interaction of adjacent k channels, which not only avoids the complete independence of the channel parameters, but also reduces the model complexity compared to using a fully connected layer to allow the parameters of all of the channels to interact. For efficiency, all of the channels can share the same learning parameters. The learning weight calculation is as follows:

$$\omega_i = \sigma\left(\sum_{j=1}^{k} \omega_i^j y_i^j\right), \; y_i^j \in \Omega_i^k \tag{6}$$

where Ω_i^k is the aggregated feature set of the k adjacent channels, y is the aggregated feature, ω is the weights of the channels, and σ is a sigmoid function.

2.4. External Attention Module

In order to provide coarse- and fine-grained classifiers with specific features, we added external attention modules to the process of generating coarse-grained and fine-grained features from the common features. The external attention module joins the external input features to implicitly learn the features of the entire dataset so that the coarse- and fine-grained classifiers can be separately focused on the features that are useful to them. The structure of the external attention mechanism is shown in Figure 4.

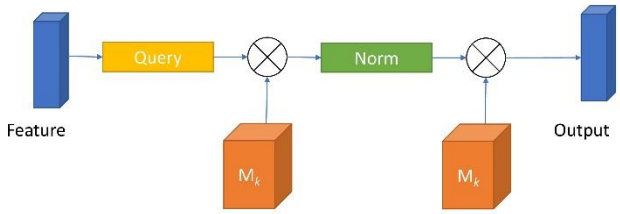

Figure 4. Structure of external attention module. The attention map is derived by calculating the pairwise associations in query and key. Then different weights are assigned to the value vector in the original image through the derived attention map to obtain a new feature map.

Compared to the self-attention mechanism, the external attention mechanism introduces an external spatial memory unit M to describe the most essential features of all samples; uses F to represent the input feature; A to represent the obtained attention matrix; $\alpha_{i,j}$ to represent the similarity between the i-th pixel and the j-th row; and M to represent the value matrix. The *Norm* is double normalization, which normalizes the columns and rows separately.

$$\begin{aligned} A = (\alpha_{i,j}) &= Norm\left(FM^T\right) \\ F_{out} &= AM \end{aligned} \tag{7}$$

In practical applications, M is divided into M_k and M_v, which represent the key and value, respectively, which can improve the fitting ability of the model.

$$\begin{aligned} A &= Norm\left(FM_K^T\right) \\ F_{out} &= AM_v \end{aligned} \tag{8}$$

3. Results

3.1. Dataset

We use the NWPU-RESISC45 dataset [37] and the Aerial Image Data Set (AID) [8] to conduct experiments. Figures 5 and 6 show sample images from the datasets.

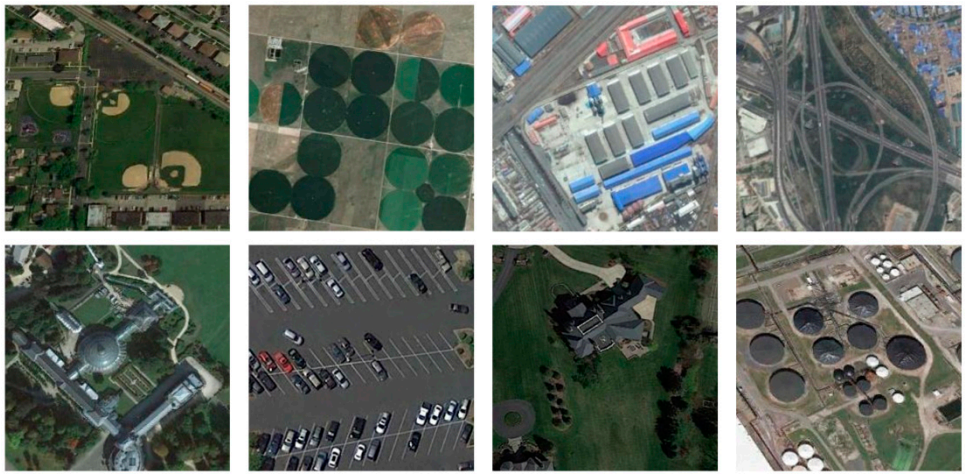

Figure 5. Example images from NWPU-RESISC45.

Figure 6. Example images from AID.

NWPU-RESISC45: This dataset consists of 31,500 images, comprising 256 × 256 RGB images extracted from Google Earth that cover more than 100 countries around the world. The dataset is divided into 45 classes and each class contains 700 images. Most of the classes, with the exception of islands, lakes, etc., have a resolution of about 30-0.2 m/pixel [37].

AID: This dataset consists of 10,000 images, comprising RGB images at a size of 600 × 600 that were mainly extracted from Google Earth images of seven countries including China, the United States, and the United Kingdom. The dataset includes images taken under different time and imaging conditions and is divided into 30 classes; each class contains 220–420 images. Pixel resolutions vary from 8 m to 0.5 m.

3.2. Experimental Setup

In order to make full use of the dataset, it was necessary to preprocess the images. The images were randomly cropped to a size of 224 × 224 and were preprocessed using horizontal flipping and other operations. In order to reasonably evaluate the experimental results, we followed the dataset split ratio used in recent papers. For AID, 20% and 50% of the images were randomly selected as the training set, and the rest were used as the test set. A total of 5% of the training set images were randomly selected as the validation set. For the NWPU-RESISC45 dataset, 10% and 20% of the images were randomly selected as the training set, and the rest were used as the test set. A total of 5% of the training set images were randomly selected as the validation set.

The equation accuracy (OA) and confusion matrix were used to evaluate the performance of the network. The formula for calculating the overall accuracy is as follows:

$$\mathrm{OA} = \frac{1}{N}\sum_{i=1}^{r} x_{ii} \tag{9}$$

In the experiments, the datasets were randomly divided into the training set, validation set, and test set to evaluate the performance of the proposed method. To avoid randomness, we repeated the experiment five times. The average classification results in the test set are reported in the Experimental Results section. The results are used to calculate the mean and variance. We used the PyTorch framework to build the network, used ImageNet pretrained weights for weight initialization and used stochastic gradient descent (SGD) combined with cosine annealing [38] to train the network. The learning rate calculation equation for cosine annealing is as follows:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + cos(\frac{T_{cur}}{T_{max}}\pi)\right) \tag{10}$$

where η_t is the learning rate. η_{max} is the maximum learning rate, and here, η_{max} is set to 0.1. η_{min} is the minimum learning rate, and here, η_{min} is set to 0. T_{cur} is the current number of iterations, and T_{max} is the maximum number of iterations.

We implemented our method using the ResNet50 and ECA modules pretrained on ImageNet images. The initial learning rate of the external attention module and ResNet50 with ECA modules in AEMMN was set to 0.005; the rest of the initial learning rate was set to 0.05; and the batch size was 48. The momentum of the SGD optimizer is set to 0.9, and the weight decay was set to 5×10^{-4}. We used a NVIDIA GeForce GTX 3080 GPU for training. In the experiment, the α parameter of the gradient loss function in Equation (1) was 1, and the β parameter was 0.7. There were 100 hidden coarse-grained feature points that were generated by the ResNet50 with ECA module, and there were 500 hidden fine-grained feature points.

The coarse-grained and fine-grained labels of the NWPU-RESISC45 and AID datasets used in this paper are listed in Tables 1 and 2.

Table 1. Coarse-grained and fine-grained labels in the NWPU-RESISC45 dataset.

Coarse-Grained Label	Fine-Grained Label
Cultivated	Circular farmland
	Rectangular farmland
	Terrace
Woodland	Chaparral
	Forest
	Wetland
Grassland	Meadow
Commercial Service	Commercial area
Industrial and Mining	Industrial area
	Thermal power station
Residential	Dense residential
	Medium residential
	Sparse residential
Public	Baseball diamond
	Basketball court
	Golf course
	Ground track field
	Mobile home park
	Runway
	Tennis court
	Stadium
Special	Church
	Palace
	Storage tanks

Table 1. *Cont.*

Coarse-Grained Label	Fine-Grained Label
Transportation land	Airport
	Airplane
	Bridge
	Freeway
	Harbor
	Intersection
	Overpass
	Parking lot
	railway
	Railway station
	Roundabout
Water	Beach
	Island
	Lake
	River
	Sea ice
	Ship
	Iceberg
Other	Cloud
	Desert
	Mountain

Table 2. Coarse-grained and fine-grained labels in AID.

Coarse-Grained Label	Fine-Grained Label
Cultivated	Farmland
Woodland	Forest
Grassland	Meadow
Commercial Service	Commercial
Industrial and Mining	Industrial
Residential	Dense residential
	Medium residential
	Sparse residential
Public	Baseball Field
	Center
	Park
	Playground
	School
	Square
	Stadium

Table 2. *Cont.*

Coarse-Grained Label	Fine-Grained Label
Special	Church
	Resort
	Storage tanks
Transportation land	Airport
	Bridge
	Parking
	Port
	Railway station
	Viaduct
Water	Beach
	Pond
	River
Other	Bare land
	Desert
	Mountain

3.3. Experimental Results

Table 3 shows the experimental results of some of the methods used on the NWPU-RESISC45 dataset, where the training set proportions were 10% and 20%. When the training set ratio was 10%, the overall accuracy of the proposed method was 92.07%, which is 3.59% higher than the original backbone network ResNet50 and 0.42% higher than BMDF-LCNN, achieving the highest accuracy among the methods in the table. When the training set ratio was 20%, it was 2.67% higher than the original backbone network ResNet50 and 0.96% higher than BMDF-LCNN, demonstrating the highest accuracy among the methods in the table.

Table 3. Overall accuracy (%) of different methods with training ratios of 50% and 20% in the NWPU-RESISC45 dataset.

Methods	Overall Accuracy	
	10% Training Ratio	20% Training Ratio
(Fine-tuning) GoogleNet [39]	82.57 ± 0.14	86.02 ± 0.18
(Fine-tuning) VGG-16 [40]	87.15 ± 0.45	90.36 ± 0.18
(Fine-tuning) ResNet50 [16]	88.48 ± 0.21	91.86 ± 0.19
D-CNN [41]	89.22 ± 0.50	91.89 ± 0.22
MG-CAP [42]	90.83 ± 0.12	92.85 ± 0.11
BMDF-LCNN [43]	91.65 ± 0.15	93.5 ± 70.22
SCCov [44]	89.30 ± 0.35	92.10 ± 0.25
MLFF [45]	90.01 ± 0.33	92.45 ± 0.20
T-CNN [46]	90.25 ± 0.14	93.05 ± 0.12
Ours	92.07 ± 0.14	94.53 ± 0.11

Figures 7 and 8 show the confusion matrix of the proposed method on the NWPURE-SISC45 dataset with training set ratios of 10% and 20%.

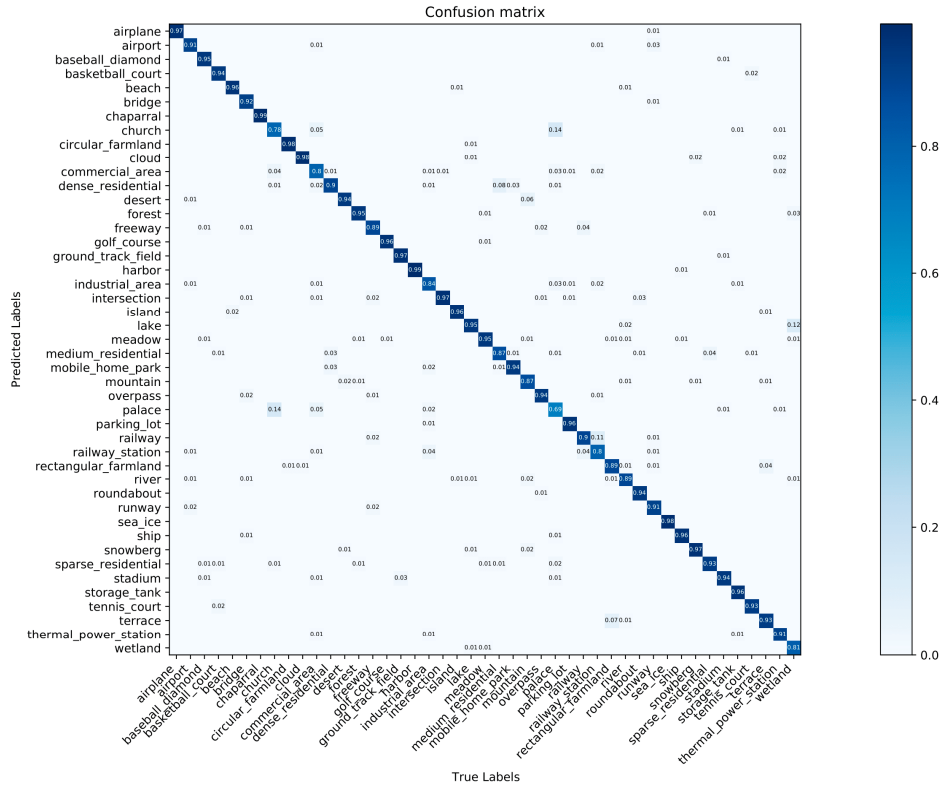

Figure 7. Confusion matrix of the proposed method on the NWPU-RESISC45 dataset with a training ratio of 10%. The content of the matrix is the accuracy of the predicted category within each category.

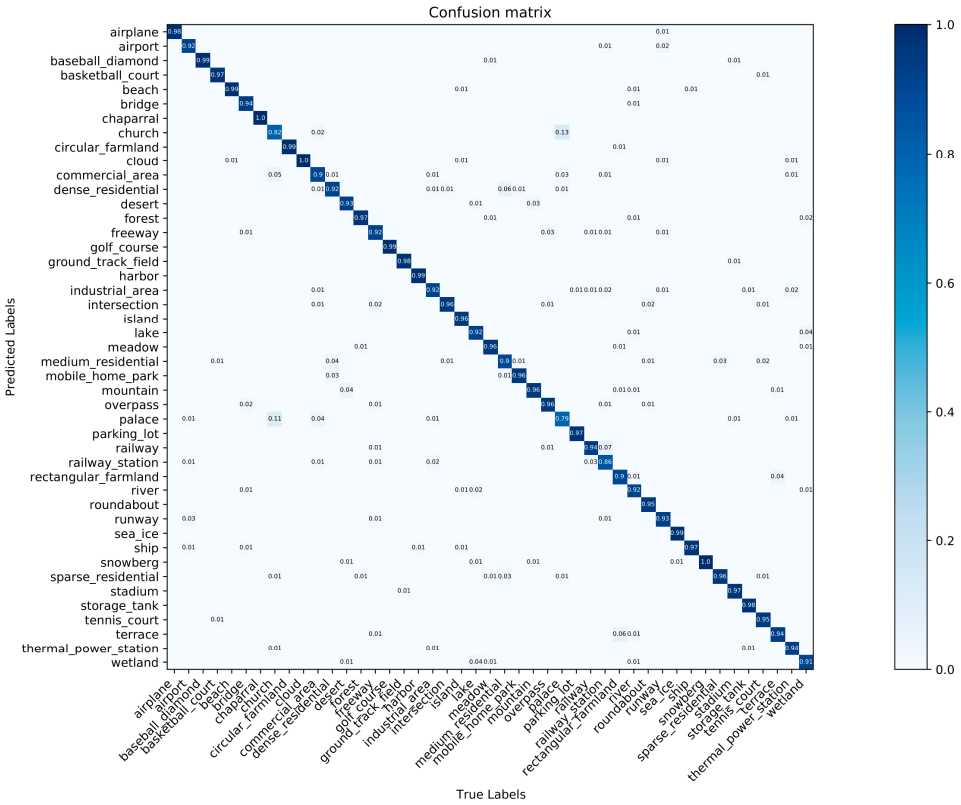

Figure 8. Confusion matrix of the proposed method on the NWPU-RESISC45 dataset with a training ratio of 20%. The content of the matrix is the accuracy of the predicted category within each category.

With a training ratio of 10%, although no scene class had 100% accuracy, most classes demonstrated an accuracy higher than 90%. Some scenes, such as those of churches, had low accuracy, as seen from the confusion matrix, and this could be because the proposed method confuses it with the palace and the commercial area labels.

For the training ratio of 20%, the accuracy of the cloud, chaparral, and iceberg categories was 100%; the accuracy of most of the categories was above 95%, with the exception of some categories such as those for churches and railway stations. Some scenes, such as churches and palaces are confused with one another due to the similar building shapes, resulting in low accuracy.

Table 4 shows the comparison of experimental results of some of the methods on the AID when the training set ratios were set to 20% and 50%. When the training ratio was 20%, the overall accuracy of the proposed method reached 94.96%, which is 0.41% higher than the T-CNN method, demonstrating the highest accuracy among the latest methods in the table, and 2.39% higher than the original backbone network ResNet50. When the training set ratio was 50%, the overall accuracy of the proposed method was 96.72%, which is 0.76% higher than the original backbone network ResNet50.

Table 4. Overall accuracy (%) of different methods with the training ratios of 50% and 20% on AID.

Methods	Overall Accuracy	
	10% Training Ratio	**20% Training Ratio**
(Fine-tuning) GoogleNet [39]	83.44 ± 0.40	86.39 ± 0.55
(Fine-tuning) VGG-16 [40]	86.59 ± 0.29	89.64 ± 0.36
(Fine-tuning) ResNet50 [16]	92.57 ± 0.21	95.96 ± 0.17
D-CNN [41]	90.82 ± 0.16	96.89 ± 0.10
MG-CAP [42]	93.34 ± 0.18	96.12 ± 0.12
ACGLNet [47]	94.44 ± 0.09	96.10 ± 0.10
BMDF-LCNN [43]	94.46 ± 0.15	96.76 ± 0.18
SCCov [44]	93.12 ± 0.25	96.89 ± 0.10
MLFF [45]	92.73 ± 0.12	95.06 ± 0.33
T-CNN [46]	94.55 ± 0.27	96.72 ± 0.23
Ours	94.96 ± 0.13	96.72 ± 0.12

Figures 9 and 10 show the confusion matrix of the proposed method on the AID with training set ratios of 20% and 50%.

For the training ratio of 20%, it can be seen from the confusion evidence in the above figure that most of the categories achieved an accuracy of more than 95%. Among them, the accuracy of the resort category is low. According to the confusion matrix, the reason for this is that the proposed method misjudges some resort scenes as park scenes.

For a training ratio of 50%, it can be seen from the confusion matrix in the figure above that the accuracy rate of most categories is higher than 95%, and the accuracy rates of the beach, mountain, pond, and viaduct categories all reach 100%. When 20% of the images are divided into the training set, the proposed method has low accuracy in the resort category because some samples are misjudged as belonging to the park category.

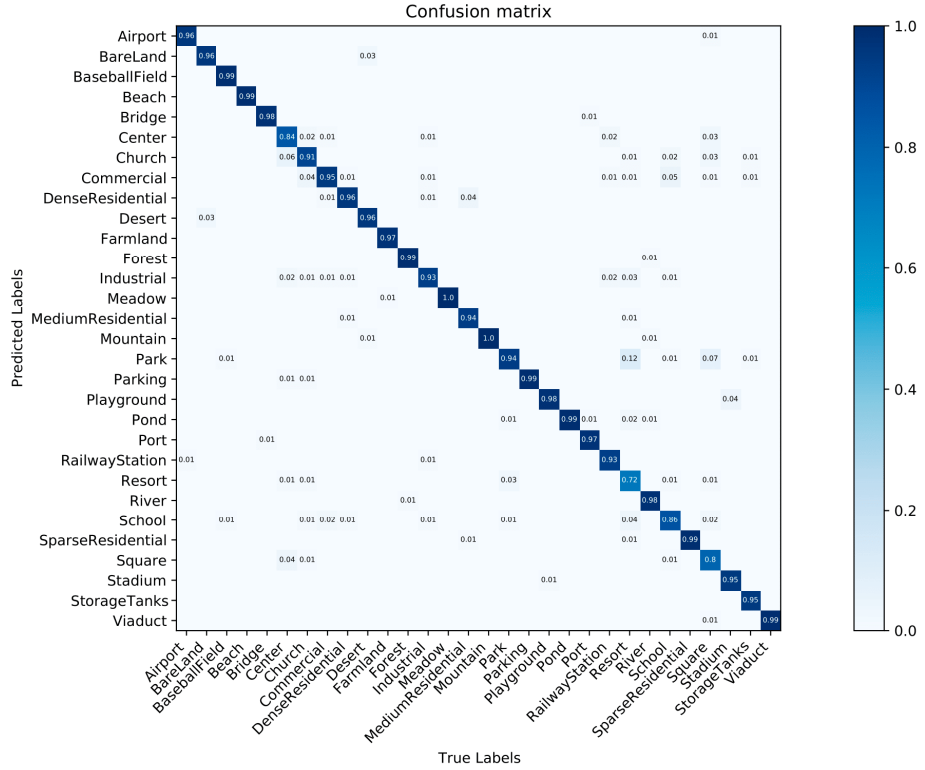

Figure 9. Confusion matrix of the proposed method on the AID with a training ratio of 20%. The content of the matrix is the accuracy of the predicted category within each category.

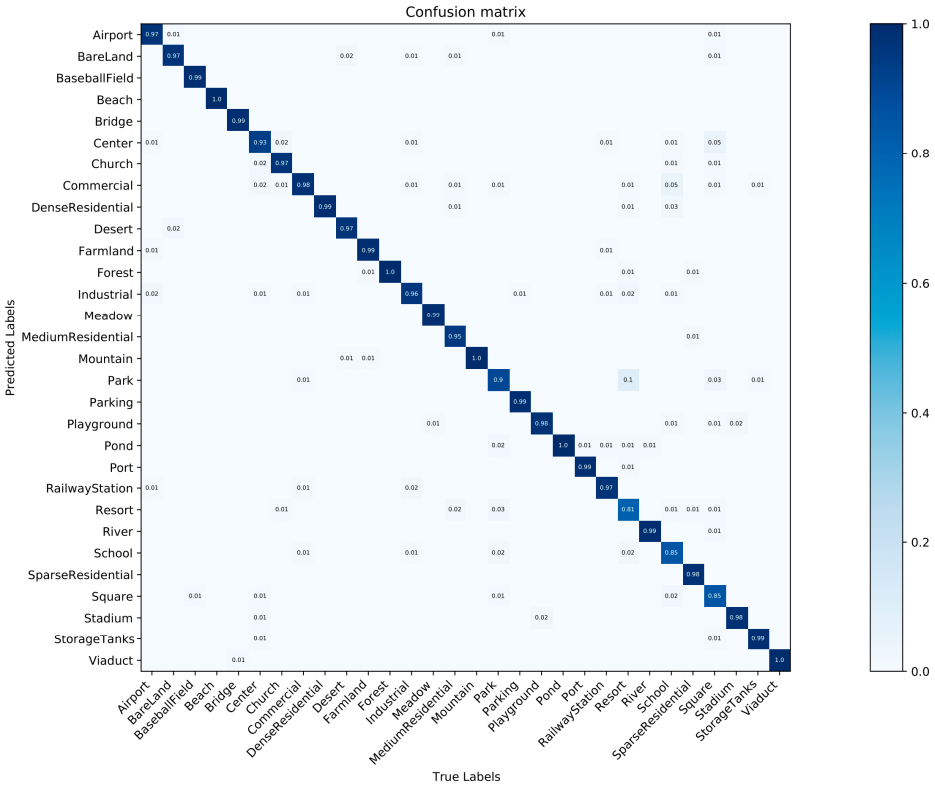

Figure 10. Confusion matrix of the proposed method on the AID with a training ratio of 50%. The content of the matrix is the accuracy of the predicted category within each category.

3.4. Discussion

We used Grad-CAM to visualize the feature extraction capabilities of the model. Grad-CAM calculates the weighted sum of each feature map by calculating the weight of each feature map in the last convolutional layer to the picture category and maps it to the original picture. Some images were randomly selected from the NWPU-RESISC45 dataset and were compared with the proposed method and methods without the addition of a multi-grained classifier. The results are shown in the following Figure 11, where the color gradient from blue to red indicates the contribution to the classification results from small to significant. It can be seen that the proposed method can better cover the main target compared to the heatmap generated when the multi-granularity classifier has not been added. This proves that adding a multi-granularity classifier can better extract features that are important to the classification results.

3.5. Computational Complexity

FLOPs represent the number of floating point operations and are used to measure the computational complexity of a model, shows in Table 5. They are often used as an indirect measure of the speed of the neural network model. Compared to the baseline model, the proposed model increases FLOPs by 0.85 G due to the addition of certain modules. However, with a slight increase in the amount of computation, the OA of the model is significantly improved on different splits of the two datasets. Compared to the recently developed T-CNN method, the FLOPs of the proposed method are only 0.32 G higher, but there was a large improvement in accuracy. In conclusion, the proposed method can improve the classification accuracy without significantly increasing the computational complexity.

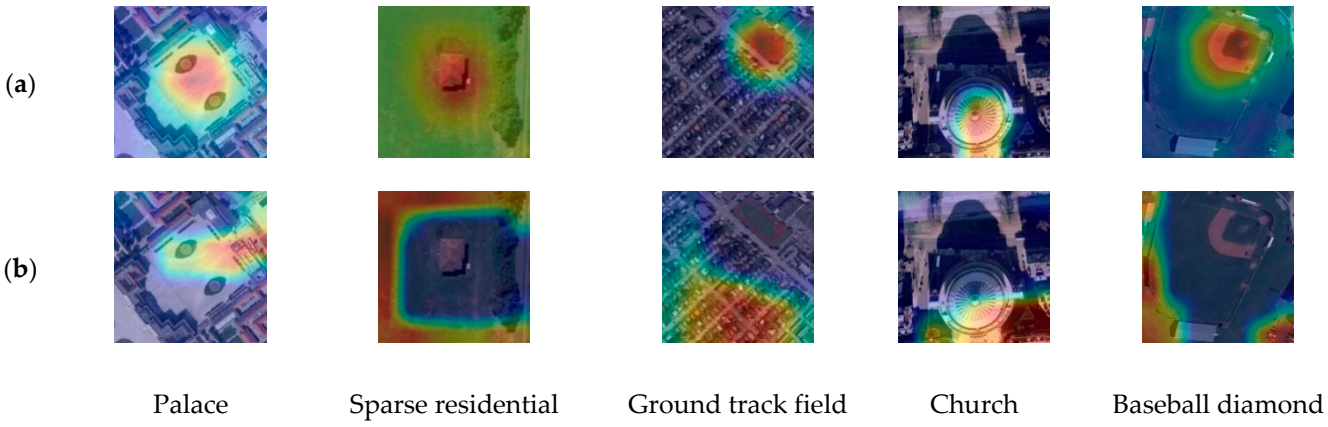

Figure 11. Grad-CAM visualization results of (**a**) with the addition of the multi-grained label and (**b**) without the addition of multi-grained label relationships on the NWPU-RESISC45 dataset.

Table 5. FLOPs (G) of different methods.

Method	FLOPs(G)
VGG-16 [40]	15.62
ResNet50 [16]	4.13
T-CNN [46]	4.67
Ours	4.99

4. Conclusions

To exploit the hierarchical label information of labels, this paper proposed an attention embedding, multi-task, multi-grained, network for scene classification, which used fine-grained labels and coarse-grained labels in different tasks and controlled the gradient

propagation process of the coarse- and fine-grained classifiers. The proposed AEMMN can take advantage of hierarchical relationships to focus more on important features and better learn what is otherwise difficult for single granularity labels, thereby benefiting granularity classifiers and improving performance.

The proposed AEMMN can be further extended to remote sensing image retrieval tasks. Image retrieval requires granularities with different labels under different conditions. In this case, the output label of the classifier corresponding to the granularity can be selected for use. In addition, the required granularity of the label can be related to the user's area of interest. The finer the label, the smaller the area of interest.

Experiments were performed using the NWPU-RESISC45 dataset and the AID. We set the generated coarse-grained and fine-grained feature dimensions. In this experiment, the generated coarse-grained feature dimension was set to 100, and the fine-grained feature dimension was set to 500. The settings of fine-grained and coarse-grained feature dimensions not only needed to ensure that the fine-grained features would not lose information due to the low dimensions being too low, but they also needed to allow the coarse-grained features to occupy a certain proportion of the input features for the coarse-grained classifiers. In addition, the weights of the coarse- and fine-grained loss in the total loss were adjusted to determine how the model parameters were prioritized for calculation during coarse-grained tasks and fine-grained tasks. Through experiments, we have found that the proposed method can achieve a high level of performance when the coarse-grained loss weight is 1 and the fine-grained loss weight is 0.7.

In future scene classification studies, we can improve the feature processing aspect of the coarse- and fine-grained module to make it more suitable for the corresponding classifier. Moreover, we can use simulated data technology to enrich our dataset. Simulated data can synthesize data pictures under different weather and lighting conditions. Simulated data can reduce the manual annotation of human resources and can enrich the dataset, which is beneficial to the performance of the model [48]. In addition, simulated data can also be used to validate experiments [49].

Author Contributions: Conceptualization, P.Z., S.L., H.S. and D.Z.; methodology, P.Z., S.L., H.S. and D.Z.; software, P.Z. and S.L.; validation, P.Z., H.S. and S.L.; formal analysis, H.S.; investigation, P.Z. and D.Z.; resources, S.L. and D.Z.; data curation, P.Z. and D.Z.; writing—original draft preparation, P.Z. and S.L.; writing—review and editing, P.Z., S.L., H.S. and D.Z.; visualization, P.Z., S.L. and H.S.; supervision, D.Z. and H.S.; project administration, P.Z., S.L. and D.Z.; funding acquisition, D.Z. and H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China (62177017, 41671377) and the Fundamental Research Funds for the Central Universities (KJ02502022-0169, CCNU22QN014).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are public image dataset [8,37].

Acknowledgments: The authors would like to thank all of the anonymous reviewers for their helpful comments and suggestions to improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sun, H.; Li, S.; Zheng, X.; Lu, X. Remote Sensing Scene Classification by Gated Bidirectional Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 82–96. [CrossRef]
2. Zhang, L.; Zhang, L. Artificial Intelligence for Remote Sensing Data Analysis: A Review of Challenges and Opportunities. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 270–294. [CrossRef]
3. Shen, H.; Jiang, M.; Li, J.; Zhou, C.; Yuan, Q.; Zhang, L. Coupling Model- and Data-Driven Methods for Remote Sensing Image Restoration and Fusion: Improving Physical Interpretability. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 231–249. [CrossRef]

4. Chen, W.; Zheng, X.; Lu, X. Semisupervised Spectral Degradation Constrained Network for Spectral Super-Resolution. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
5. Zheng, X.; Sun, H.; Lu, X.; Xie, W. Rotation-Invariant Attention Network for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2022**, *31*, 4251–4265. [CrossRef] [PubMed]
6. Phung, M.T.; Tu, T.H. Scene Classification for Weak Devices Using Spatial Oriented Gradient Indexing. In *Eighth International Conference on Graphic and Image Processing (ICGIP 2016)*; SPIE: Bellingham, WA, USA, 2017; Volume 10225, p. 1022520.
7. Anwer, R.M.; Khan, F.S.; van de Weijer, J.; Molinier, M.; Laaksonen, J. Binary Patterns Encoded Convolutional Neural Networks for Texture Recognition and Remote Sensing Scene Classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 74–85. [CrossRef]
8. 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]
9. Li, F. Automatic Acquisition of Appropriate Codewords Number in BoVW Model and the Corresponding Scene Classification Performance. In Proceedings of the 37th Chinese Control Conference, Wuhan, China, 25–27 July 2018.
10. Jegou, H.; Douze, M.; Schmid, C.; Perez, P. Aggregating Local Descriptors into a Compact Image Representation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: San Francisco, CA, USA, 2010; pp. 3304–3311.
11. Perronin, F.; Larlus, D. Fisher Vectors Meet Neural Networks: A Hybrid Classification Architecture. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
12. Lu, X.; Sun, H.; Zheng, X. A Feature Aggregation Convolutional Neural Network for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7894–7906. [CrossRef]
13. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2323. [CrossRef]
14. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations, Copenhagen, Denmark, 9–11 September 2017.
15. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661. [CrossRef]
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
18. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 28 May 2019.
19. Wang, B.; Dong, G.; Zhao, Y.; Li, R.; Cao, Q.; Chao, Y. Non-Uniform Attention Network for Multi-Modal Sentiment Analysis. In *International Conference on Multimedia Modeling*; Springer: Cham, Switzerland, 2022; pp. 612–623.
20. Hao, S.; Zhang, H. Performance Analysis of PHY Layer for RIS-Assisted Wireless Communication Systems with Retransmission Protocols. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 5388–5404. [CrossRef]
21. Li, F.; Feng, R.; Han, W.; Wang, L. High-Resolution Remote Sensing Image Scene Classification via Key Filter Bank Based on Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8077–8092. [CrossRef]
22. Wang, Q.; Huang, W.; Xiong, Z.; Li, X. Looking Closer at the Scene: Multiscale Representation Learning for Remote Sensing Image Scene Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 1414–1428. [CrossRef] [PubMed]
23. Sun, H.; Zheng, X.; Lu, X. A Supervised Segmentation Network for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2021**, *30*, 2810–2825. [CrossRef] [PubMed]
24. Xue, W.; Dai, X.; Liu, L. Remote Sensing Scene Classification Based on Multi-Structure Deep Features Fusion. *IEEE Access* **2020**, *8*, 28746–28755. [CrossRef]
25. Castelluccio, M.; Poggi, G.; Sansone, C.; Verdoliva, L. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. *arXiv* **2015**, arXiv:1508.00092.
26. Tian, T.; Li, L.; Chen, W.; Zhou, H. SEMSDNet: A Multiscale Dense Network with Attention for Remote Sensing Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5501–5514. [CrossRef]
27. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene Classification with Recurrent Attention of VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 1155–1167. [CrossRef]
28. Chen, G.; Zhang, X.; Tan, X.; Cheng, Y.; Dai, F.; Zhu, K.; Gong, Y.; Wang, Q. Training Small Networks for Scene Classification of Remote Sensing Images via Knowledge Distillation. *Remote Sens.* **2018**, *10*, 719. [CrossRef]
29. Zhang, B.; Zhang, Y.; Wang, S. A Lightweight and Discriminative Model for Remote Sensing Scene Classification with Multidilation Pooling Module. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2636–2653. [CrossRef]
30. Ruder, S. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv* **2017**, arXiv:1706.05098.
31. Chang, D.; Pang, K.; Zheng, Y.; Ma, Z.; Song, Y.Z.; Guo, J. Your "Flamingo" Is My "Bird": Fine-Grained, or Not. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11471–11480.
32. Liu, P.; Qiu, X.; Huang, X. Recurrent Neural Network for Text Classification with Multi-Task Learning. *arXiv* **2016**, arXiv:1605.05101. [CrossRef]

33. Crawshaw, M. Multi-Task Learning with Deep Neural Networks: A Survey. *arXiv* **2020**, arXiv:2009.09796.
34. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11531–11539.
35. Kokkinos, I. UberNet: Training a 'Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision Using Diverse Datasets and Limited Memory. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
36. Zhang, W.; Tang, P.; Zhao, L. Remote Sensing Image Scene Classification Using CNN-CapsNet. *Remote Sens.* **2019**, *11*, 494. [CrossRef]
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. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the 5th International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016. [CrossRef]
39. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
40. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
41. 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]
42. Wang, S.; Guan, Y.; Shao, L. Multi-Granularity Canonical Appearance Pooling for Remote Sensing Scene Classification. *IEEE Trans. Image Process.* **2020**, *29*, 5396–5407. [CrossRef]
43. Shi, C.; Zhang, X.; Sun, J.; Wang, L. Remote Sensing Scene Image Classification Based on Dense Fusion of Multi-Level Features. *Remote Sens.* **2021**, *13*, 4379. [CrossRef]
44. He, N.; Fang, L.; Li, S.; Plaza, J.; Plaza, A. Skip-Connected Covariance Network for Remote Sensing Scene Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 1461–1474. [CrossRef]
45. Wang, X.; Duan, L.; Shi, A.; Zhou, H. Multilevel Feature Fusion Networks with Adaptive Channel Dimensionality Reduction for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
46. Wang, W.; Chen, Y.; Ghamisi, P. Transferring CNN with Adaptive Learning for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18. [CrossRef]
47. Shen, J.; Yu, T.; Yang, H.; Wang, R.; Wang, Q. An Attention Cascade Global–Local Network for Remote Sensing Scene Classification. *Remote Sens.* **2022**, *14*, 2042. [CrossRef]
48. Ciampi, L.; Messina, N.; Falchi, F.; Gennaro, C.; Amato, G. Virtual to Real Adaptation of Pedestrian Detectors. *Sensors* **2020**, *20*, 5250. [CrossRef] [PubMed]
49. Staniszewski, M.; Foszner, P.; Kostorz, K.; Michalczuk, A.; Wereszczyński, K.; Cogiel, M.; Golba, D.; Wojciechowski, K.; Polański, A. Application of Crowd Simulations in the Evaluation of Tracking Algorithms. *Sensors* **2020**, *20*, 4960. [CrossRef] [PubMed]