remote sensing

Article

Collaborative Consistent Knowledge Distillation Framework for Remote Sensing Image Scene Classification Network

Shiyi Xing [1,†], Jinsheng Xing [2,*], Jianguo Ju [3,†], Qingshan Hou [4,†] and Xiurui Ding [5]

[1] James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK
[2] School of Mathematics and Computer Science, Shanxi Normal University, Taiyuan 030031, China
[3] Department of Information Science and Technology, Northwest University, Xi'an 710069, China
[4] School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China
[5] School of Natural Sciences, The University of Manchester, Manchester M15 4RB, UK
* Correspondence: xingjs@sxnu.edu.cn
† These authors contributed equally to this work.

Abstract: For remote sensing image scene classification tasks, the classification accuracy of the small-scale deep neural network tends to be low and fails to achieve accuracy in real-world application scenarios. However, although large deep neural networks can improve the classification accuracy of remote sensing image scenes to some extent, the corresponding deep neural networks also have more parameters and cannot be used on existing embedded devices. The main reason for this is that there are a large number of redundant parameters in large deep networks, which directly leads to the difficulty of application on embedded devices and also reduces the classification speed. Considering the contradiction between hardware equipment and classification accuracy requirements, we propose a collaborative consistent knowledge distillation method for improving the classification accuracy of remote sensing image scenes on embedded devices, called CKD. In essence, our method addresses two aspects: (1) We design a multi-branch fused redundant feature mapping module, which significantly improves the parameter redundancy problem. (2) To improve the classification accuracy of the deep model on embedded devices, we propose a knowledge distillation method based on mutually supervised learning. Experiments were conducted on two remote sensing image classification datasets, SIRI-WHU and NWPU-RESISC45, and the experimental results showed that our approach significantly reduced the number of redundant parameters in the deep network; the number of parameters decreased from 1.73 M to 0.90 M. In addition, compared to a series of student sub-networks obtained based on the existing different knowledge distillation methods, the performance of the student sub-networks obtained by CKD for remote sensing scene classification was significantly improved on two different datasets, with an average accuracy of 0.943 and 0.916, respectively.

Keywords: remote sensing image; parameter redundancy; knowledge distillation; mutual supervised learning; scene classification

check for updates

Citation: Xing, S.; Xing, J.; Ju, J.; Hou, Q.; Ding, X. Collaborative Consistent Knowledge Distillation Framework for Remote Sensing Image Scene Classification Network. *Remote Sens.* **2022**, *14*, 5186. https://doi.org/10.3390/rs14205186

Academic Editors: Junjun Jiang, Pia Addabbo, Silvia Liberata Ullo and Parameshachari Bidare Divakarachari

Received: 18 August 2022
Accepted: 14 October 2022
Published: 17 October 2022

1. Introduction

Over the past few years, deep neural networks have achieved state-of-the-art performance in computer vision [1–4], natural language processing [5–7], reinforcement learning [8–10], and various other fields [11–13]. However, with the increasing depth, as well as the width of the network, for example from the shallow LeNet to the wider Inception structure in GoogLeNet and deeper Resnet convolutional architecture, as well as the currently popular transformer architecture, the number of parameters of the deep model is constantly growing, which in turn, leads to a series of problems such as the redundancy of network parameters, more rigorous hardware requirements, and difficulty in training the model, and large deep models severely limit their applications in low-memory or high-real-time conditions. In recent years, the research [14,15] to develop faster and smaller models based on the idea of knowledge distillation to solve the above problems has been developing rapidly. In the traditional knowledge-distillation-based model compression methods [16–19], the smaller student network is typically guided by a larger teacher network. The primary purpose is to enable the student network to achieve competitive and even superior task performance by learning the prior knowledge of the teacher network. The key to achieving this goal is mainly related to two aspects: on the one hand, how to design the network structure of the teacher–student model; on the other hand, how to transfer important features from the large-sized teacher model to the small-sized student model in a more efficient way.

We observed the following phenomenon when performing model optimization with the standard knowledge distillation methodology:

I. When we trained a small-sized student network independently, it was usually more difficult to find the ideal model parameters to meet the relevant task requirements.

II. Compared with training a small-sized student network independently, when a large-sized teacher network was trained independently, although better task performance can be gained, the model parameters of the teacher network were not optimal due to the presence of a significant amount of parameter redundancy in the teacher network.

III. When jointly training teacher–student models, the parameter redundancy present in the teacher model was usually detrimental to the optimization of the student model, which may have a negative effect on the optimization of the student model.

In this paper, considering that current high-precision remote sensing image classification models require high-performance hardware devices, which are difficult to deploy on embedded devices with low performance, our goal was to solve the parameter redundancy problem in the teacher–student model and obtain a small deep neural network with powerful feature extraction capabilities that can be easily deployed on lower-performance hardware devices and meet the accuracy requirements for remote sensing image classification. To address these issues, we propose a collaborative consistency knowledge distillation framework.

Firstly, different from the previous convolutional neural networks, a plug-and-play redundant feature mapping module was designed for the redundant parameters in the teacher–student model. Specifically, this module contains both multi-branch feature extraction and fusion components, as well as redundant mapping convolution components. On the one hand, we can obtain an equivalent convolution kernel with stronger feature extraction capability with multi-branch feature extraction and fusion and utilize this equivalent convolution kernel to extract richer task-related high-level semantic information. On the other hand, the redundant mapping convolution component was used to generate the intrinsic feature maps of the inputs, and the redundant feature maps were further obtained by a series of low-cost linear operations, which greatly reduced the redundant parameters of the network.

Secondly, our CKD framework starts with a powerful and pre-trained teacher network and performs a one-way prior knowledge transfer to two untrained student sub-networks of different depths. In addition, for both student sub-networks, we propose that the student sub-networks not only absorb prior knowledge derived from the teacher, but also extract high-level semantic features that the other possesses via mutual supervised learning. The experimental results showed that the student sub-networks obtained by training in this way have better task-relevant model parameters.

In summary, our contributions are summarized as follows:

- To reduce the parameter redundancy of remote sensing image classification models and facilitate their deployment on embedded devices with low performance, we propose a plug-and-play multi-branch fused redundant feature mapping module. The equivalent convolutional kernel obtained by this module has a more powerful feature extraction capability and can more effectively optimize the parameter redundancy of the network.
- We propose a collaborative consistent knowledge distillation framework to obtain a more robust backbone network. In contrast to the traditional knowledge distillation

framework, we guided a pair of student sub-networks of different depths through a teacher model, where the student sub-networks not only learn prior knowledge deriving from the teacher network, but also acquire prior knowledge possessed by them by the way of mutual supervised learning.

- The experimental results on two benchmark datasets (SIRI-WHU, NWPU-RESISC45) showed that our approach provided a significant improvement over a series of existing depth models and the state-of-the-art knowledge distillation networks on the relevant remote sensing image scene classification task. In addition, the student sub-network obtained based on the CKD framework had a more powerful feature extraction capability, as well as a lower number of parameters, which can be widely used as a feature extraction network in various embedded devices.

2. Related Work

2.1. Remote Sensing Image Scene Recognition

The remote sensing image recognition task realizes the recognition and classification of scene topics by analyzing the composition relationship of the targets in the image scene, which mainly contains methods based on mid-level features and deep learning methods. The main approaches based on mid-level features include visual word-packet models [20], combined with sparse representation [21], Flisher vector coding [22], and so on. However, these traditional methods can hardly meet the accuracy requirements for remote sensing image scene classification on embedded devices.

In recent years, deep learning models have performed well in remote sensing image recognition tasks. Yao et al. [23] utilized a pre-trained deep learning network for feature extraction of remote sensing scenes and adopted a random forest classifier for scene recognition of remote sensing images. Cheng et al. [24] combined deep learning with metric learning, and the problem of high similarity between remote sensing scenes and large intra-class differences was well solved by discriminative convolutional neural networks. Gong et al. [25] combined the attention mechanism with the deep learning model, which solved the overfitting problem of the deep learning model in remote sensing image processing to some extent. However, although these models can obtain better accuracy for remote sensing image scene classification, they are difficult to deploy on embedded devices with low performance due to having more model parameters.

2.2. Knowledge Distillation

In recent years, deep learning methods have achieved great success in the field of knowledge distillation [15,26,27]. In accordance with whether the teacher model is updated simultaneously with the student model, the learning schemes for knowledge distillation are mainly divided into two categories: offline knowledge distillation [28–33] and online knowledge distillation [34–37].

The training for the offline knowledge distillation method needs to be performed in stages; specifically, in the first stage, the large-scale teacher network is first trained based on the relevant training dataset until the network converges. In the second stage, based on the trained teacher network, the relevant features of the input data are extracted, and these features are then utilized to guide the training of the student network. The knowledge transfer from the pre-trained teacher network to the student network is enabled by the two stages. Bucilua et al. [16] advocated the use of knowledge transfer for compressing models as early as 2006, transferring knowledge from a large-scale complicated model to a lightweight model. The idea was adopted by Hinton et al. [19] in 2015, and the concept of knowledge distillation (KD) was formally defined, as well as a detailed training method for knowledge distillation networks given. FitNets [28] further extends the idea of knowledge distillation by adding an intermediate layer of knowledge distillation to the teacher network and boosts the training speed of the knowledge distillation network with the guidance of the intermediate layer feature map. Inspired by this, RKD [29] combines the output of multiple teacher models to produce structural units, which work together to guide student learning, driving better guidance for student models. CRD [30] introduces comparative learning for knowledge distillation and trains the student network to be able to learn more useful knowledge from the data representation of the teacher network. Li et al. proposed LKD [31], a local correlation exploration framework for knowledge distillation, which uses the intra-instance local relationship, the inter-instance relation on the same local location, and the inter-instance relation across different local locations for modeling. Xu et al. [32] proposed a feature-normalized knowledge distillation scheme by introducing a sample-specific correction factor instead of the uniform temperature T. Considering the ensemble knowledge distillation as a multi-objective optimization problem, Du et al. [33] investigated the diversity of teacher models in gradient spaces.

Unlike the process of offline distillation, the online knowledge distillation process updates the entire knowledge distillation framework simultaneously, that is the teacher model and the student model are updated in parallel. Over the last few years in particular, a series of online knowledge distillation methods have been proposed. For example, Lan et al. [34] proposed a learning framework for single-stage online distillation. Specifically, the framework establishes powerful online teacher models to enhance the learning of the target network while only training a single multi-branch network. Zhang et al. [35] proposed a deep mutual learning strategy that allows student models to learn collaboratively and teach each other throughout the training process. Yao et al. [36] designed an improved bidirectional knowledge distillation method, the dense cross-layer mutual distillation framework (DCM). Wu et al. [37] found that collaborative learning and mutual learning cannot build the online high-capacity teacher network, while the online integration ignores the collaboration between branches, which leads to the proposal of a novel peer collaborative learning approach for online knowledge distillation.

3. Methodology

For tasks associated with the field of computer vision, the number of parameters in the backbone network increases dramatically as the depth of the network is continuously deepened, which causes significant parameter redundancy from the backbone network, hence affecting the performance of several computer vision tasks. In order to reduce the redundant parameters of the backbone network and obtain a more powerful CNN feature extractor, we propose a collaborative consistency distillation framework, which can effectively deal with the parameter redundancy problem with the increasing depth of the network, while making the obtained backbone network have excellent feature extraction capability. As a result, it can better support various downstream computer vision tasks. The overall pipeline of the CKD framework is illustrated in Figure 1.

3.1. Redundant Feature Mapping Module

In contrast to the ordinary convolution in previous convolutional neural networks [38–43], our proposed redundant feature mapping module can be inserted into any network structure to improve the model structure, enhance the model feature extraction capability, and reduce the parameter redundancy and floating point operations of the model. The relevant structure of the redundant feature mapping module is shown in Figure 2, which mainly includes two aspects:

(1) Multi-branch feature extraction and fusion: For multi-branch feature extraction and fusion, different from the previous work, our objective was to obtain equivalent convolutional kernels with stronger feature extraction capability. In other words, the obtained single-branch $k \times k$ equivalent convolution kernel has multi-scale feature extraction capability. As shown in Figure 2, MRFM enhances the feature extraction capability of the CNN network with three parallel branches, and each branch employs the $k \times k$, $1 \times k$, and $k \times 1$ convolutional kernel sizes, respectively. When the network training is complete, the convolutional kernels of three sizes are fused into equivalent convolutional kernels of $k \times k$ with stronger extraction ability. The process of equivalent fusion mainly consists of two processes, BN fusion and branch fusion.

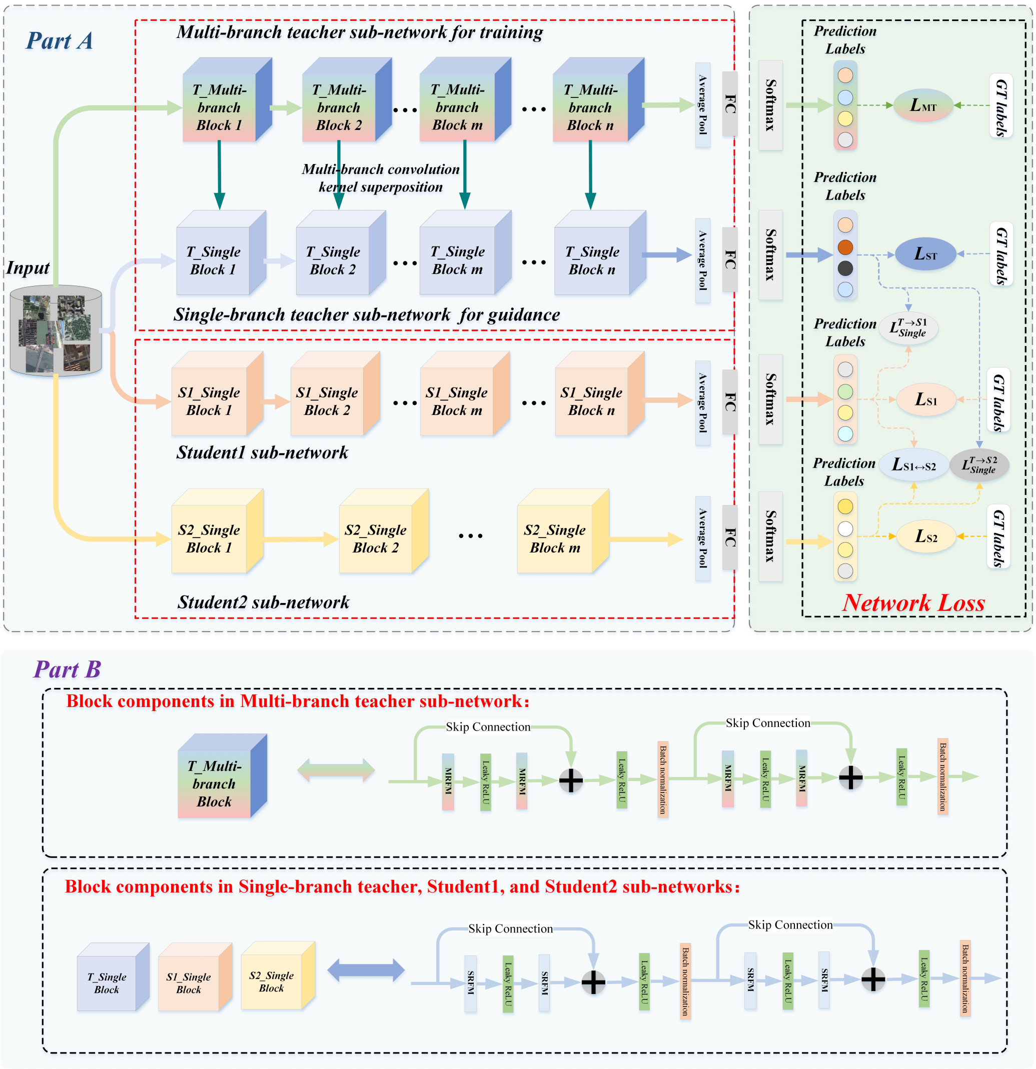

Figure 1. Illustration of the collaborative consistency distillation framework. For the network input, deep semantic features of the remote sensing image scene classification can be obtained in two aspects: On the one hand, we used the teacher sub-networks to extract deep semantic features for guiding the student sub-networks to extract more refined classification feature information. On the other hand, the classification feature information was reinforced by mutual supervision among student sub-networks to enable the student sub-networks to obtain higher classification results. *Part A* demonstrates in detail the components of our proposed the CKD framework. Specifically, it contains a multi-branch teacher sub-network, a single-branch teacher sub-network, and a pair of student sub-networks of different depths. *Part B* presents the basic block components of the student sub-networks and the teacher sub-networks in the CKD framework.

BN fusion: In order to prevent the overfitting and accelerate the training speed of the network and for the MRFM module, it is necessary to perform the BN operation as shown in Equation (1) after each branch performs the convolution operation.

$$O_{:,:,j} = \left(\sum_{k=1}^{C} M_{:,:,k} * F^{(j)}_{:,:,k} - \mu_j\right)\frac{\gamma_j}{\sigma_j} + \beta_j \tag{1}$$

where $M \in \mathbb{R}^{U\times V\times C}$ denotes the input feature maps of size $U \times V$ and the number of channels C and k denotes the input feature map of the k-th channel. $F \in \mathbb{R}^{H\times W\times C}$ indicates a convolution kernel of size $H \times W$ and the number of channels C. The output feature map $O \in \mathbb{R}^{R\times T\times D}$ of size $R \times T$ and number of channels D is obtained after the convolution operation $*$. μ_j and σ_j are the mean and standard deviation of the BN operation, and γ_j and β_j are the scaling factor and offset, respectively.

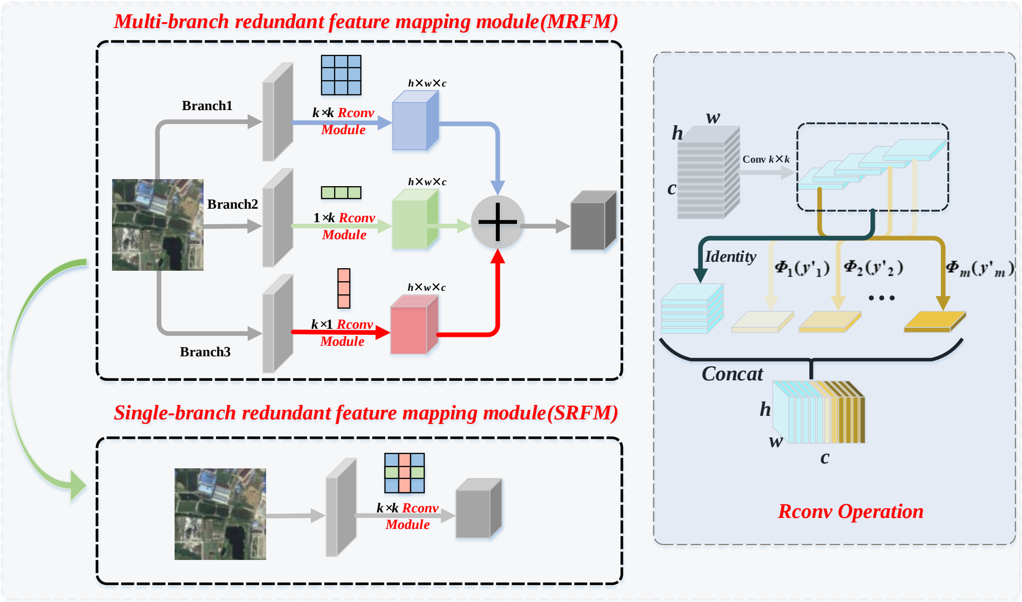

Figure 2. Illustration of the redundant feature mapping module. Specifically, the redundancy mapping module consists of three aspects: the multi-branch redundancy mapping module (MRFM), the single-branch redundancy mapping module (SRFM), and the redundancy mapping convolution (Rconv) operation. It is worth noting that the single-branch redundancy mapping module is generated from the multi-branch redundancy mapping module after BN fusion and branch fusion operations.

After the above BN operation, the convolutional kernels of different sizes are fused based on the principle of additivity between 2D convolutional kernels to produce an equivalent convolutional kernel with the same feature output, and the associated process can be represented by Equation (2).

$$I * K^{(1)} + I * K^{(2)} = I * \left(K^{(1)} \oplus K^{(2)}\right) \tag{2}$$

where I indicates a matrix that can be cropped or filled. $K^{(1)}$ and $K^{(2)}$ are two 2D convolution kernels with compatible dimensions, and $\oplus$ refers to the summation operation at the corresponding positions.

Branch fusion: As shown in Figure 2, the three feature extraction branches are reduced to one feature extraction branch, and the feature extraction is completed based on the equivalent convolutional kernel obtained after BN fusion. After such an operation, the features we extracted are equivalent to the extraction results of multiple feature extraction branches. In other words, this operation enhances the feature extraction ability of the network and reduces the network parameters, which improve the performance of the network. For the j-th convolution kernel, $F'(j)$ represents the fused convolution kernel, b_j represents the bias, $F^{(j)}$, $\bar{F}^{(j)}$, and $\hat{F}^{(j)}$ represent the outputs of the $k \times k$, $1 \times k$, and $k \times 1$ convolution kernels, respectively, and the result after branching fusion can be expressed as:

$$F'(j) = \frac{\gamma_j}{\sigma_j} F^{(j)} \oplus \frac{\bar{\gamma}_j}{\bar{\sigma}_j} \bar{F}^{(j)} \oplus \frac{\hat{\gamma}_j}{\hat{\sigma}_j} \hat{F}^{(j)} \tag{3}$$

$$b_j = -\frac{\mu_j \gamma_j}{\sigma_j} - \frac{\bar{\mu}_j \bar{\gamma}_j}{\bar{\sigma}_j} - \frac{\hat{\mu}_j \hat{\gamma}_j}{\hat{\sigma}_j} + \beta_j + \bar{\beta}_j + \hat{\beta}_j \tag{4}$$

$$O_{:,:,j} + \bar{O}_{:,:,j} + \hat{O}_{:,:,j} = \sum_{k=1}^{C} M_{:,:,k} * F'(j)_{:,:,k} + b_j \quad (5)$$

where μ and σ are the mean and standard deviation of the BN operation, γ and β are the scaling factor and offset, and $O_{:,:,j}$, $\bar{O}_{:,:,j}$, and $\hat{O}_{:,:,j}$ are the output feature maps of the $k \times k$, $1 \times k$, and $k \times 1$ convolution kernels, respectively.

(2) Redundant mapping convolution operation (Rconv): Due to the significant redundancy in the feature maps extracted by the existing backbone network, to address this problem, the ordinary convolution layer is divided into two parts, as shown in the Rconv module in Figure 2, which fully combines the ordinary convolution operation, as well as the linear transformation operation. Specifically, we first obtained the intrinsic feature maps by ordinary convolutional operations; second, we performed the identical transformation and a series of simple linear transformations on the intrinsic feature maps. The two operate in parallel: On the one hand, the intrinsic feature maps are preserved, and the computational burden of the network is reduced. On the other hand, the redundant information in the feature maps is preserved with the inexpensive linear mapping, which obtains the redundant feature maps.

The equation for the ordinary convolution operation to generate n feature maps is expressed as:

$$Y = X * f + b \quad (6)$$

For the Rconv module, m feature maps are first generated by ordinary convolution. It can be expressed as follows.

$$Y' = X * f' \quad (7)$$

where $X \in \mathbb{R}^{c \times h \times w}$ is the input feature maps, c, h and w are the number of channels and the height, and width of the input feature map, respectively, and $*$ denotes the convolution operation. $f \in \mathbb{R}^{c \times k \times k \times n}$ and $f' \in \mathbb{R}^{c \times k \times k \times m}$ denote the convolution kernel. $k \times k$ is the kernel size. b is the bias term. For simplicity, the bias term is neglected in Equation (7). $Y \in \mathbb{R}^{n \times h' \times w'}$ and $Y' \in \mathbb{R}^{m \times h' \times w'}$ denote the output feature maps with n and m channels, respectively, and $m \ll n$. h', and w' represent the height and width of the output feature maps. In addition, to obtain the required n feature maps, redundant feature information is generated by adding linear operations to the inherent feature maps in Y'.

$$y_i = \Phi_i(y'_i) \qquad \forall i = 1, \ldots, m \quad (8)$$

where y'_i denotes the i-th intrinsic element map in Y', y_i denotes the redundant feature map of y'_i, and $\Phi_i(\cdot)$ is an inexpensive linear operation on y'_i.

3.2. Cooperative Consistency Distillation Algorithm

The main purpose of the collaborative consistency knowledge distillation algorithm is to obtain remote sensing image scene classification models that are convenient for deployment on embedded devices. Therefore, to achieve the above goal, in contrast to the previous work, our approach consists of two main aspects. On the one hand, a single-teacher multi-student knowledge distillation model is constructed based on the proposed redundant feature mapping module, and the two student sub-networks with fewer parameters and higher accuracy are obtained based on this architecture. On the other hand, the accuracy of each student sub-network is further improved by a collaborative consistency strategy between the student sub-networks.

As shown in Figure 1 *Part A*, the teacher network consists of two parts: multi-branch and single-branch teacher networks. We first trained the multi-branch teacher network, which is composed of multiple multi-branch blocks. When the multi-branch teacher network is trained, multiple feature extraction branches are transformed into a single feature extraction branch by the multi-branch feature fusion operation, which not only drastically reduces the number of parameters of the network, but also obtains a single-branch teacher network with equivalent feature extraction capability. The single-branch teacher network is composed of multiple single blocks, and the structure of the multi-branch block and single block is shown in Figure 1 *Part B*. Second, we used the pre-trained single-branch teacher sub-network to guide the feature learning of the Student1 and Student2 sub-networks, so that the student sub-network learns as much prior knowledge as possible from the single-branch teacher sub-network. As a result, the student network can not only achieve the purpose of model compression, but also achieve an accuracy similar to the teacher sub-network. Note that the student sub-networks have different specifications, and the student sub-network S_1 holds a deeper network structure. For the *i*-th data sample, the loss of the student sub-network S_1, as well as S_2 can be expressed as:

$$L_{Single}^{T\rightarrow S_1} = L_{CE}\left(y_i, p_i^{S_1}\right) + \lambda_1 L_{CE}\left(p_i^{S_1}, q_i\right) \tag{9}$$

$$L_{Single}^{T\rightarrow S_2} = L_{CE}\left(y_i, p_i^{S_2}\right) + \lambda_2 L_{CE}\left(p_i^{S_2}, q_i\right) \tag{10}$$

where $L_{CE}(p_i, q_i)$ is the cross-entropy loss between the predicted value q_i of the single-branch teacher network T and the predicted value p_i of the student network S and $L_{CE}(y_i, p_i)$ is the cross-entropy loss between the predicted value p_i of the student network S and the true label y_i. λ indicates the regularization weight, which balances the losses of different components. Through fitting the predicted labels of the single-branch teacher network T, the student network S is able to learn as much prior knowledge from the teacher network T as possible.

For the student sub-networks, the predicted outputs of the two student sub-networks S_1 and S_2 on the *i*-th sample data are denoted as $p_i^{S_1}$ and $p_i^{S_2}$, respectively. The two student sub-networks are updated simultaneously. During the training of the student sub-network S_1, the student sub-network S_2 helps S_1 converge by using learned classification characteristics to guide S_1. To measure the variance between the predictions $p_i^{S_1}$ and $p_i^{S_2}$ of the two student sub-networks, the Kullback–Leibler (KL) loss is used for calculation. Then, the KL loss of S_2 to S_1 can be expressed as:

$$L_{S2\rightarrow S1}\left(p_i^{S_2} \| p_i^{S_1}\right) = p_i^{S_2} \log \frac{p_i^{S_2}}{p_i^{S_1}} \tag{11}$$

Similarly, during the training of the student sub-network S_2, the student sub-network S_1 guides S_2 with the learned classification information, and the KL loss of S_1 to S_2 can be expressed as:

$$L_{S2\rightarrow S1}\left(p_i^{S_1} \| p_i^{S_2}\right) = p_i^{S_1} \log \frac{p_i^{S_1}}{p_i^{S_2}} \tag{12}$$

In summary, the student sub-network is required to fit not only the truth label y, but also the prediction label q of the single-branch teacher sub-network T and the prediction label of another student sub-network. Therefore, the overall loss of the student sub-network includes the traditional supervised loss L_{S_1}, L_{S_2}, the mutual supervised loss $L_{S2\rightarrow S1}$, $L_{S1\rightarrow S2}$ among the student sub-networks, and the distillation loss $L_{Single}^{T\rightarrow S1}$, $L_{Single}^{T\rightarrow S2}$ with the single-branch teacher sub-network. The final loss of the student sub-network can be expressed by the following equation.

$$L_{final_{S1}} = L_{Single}^{T\rightarrow S1} + \alpha_1 L_{S2\rightarrow S1}\left(p_i^{S_2} \| p_i^{S_1}\right) + L_{S_1} \tag{13}$$

$$L_{final_{S2}} = L_{Single}^{T\rightarrow S2} + \alpha_2 L_{S1\rightarrow S2}\left(p_i^{S_1} \| p_i^{S_2}\right) + L_{S_2} \tag{14}$$

The collaborative consistency distillation algorithm guides multiple student sub-networks through the teacher sub-network, while maintaining the collaborative consistency among student sub-networks through mutual supervised learning. The final goal of optimizing the parameter redundancy and improving the classification performance of student sub-networks was accomplished, and the related process is described in Algorithm 1.

Algorithm 1 Collaborative consistency distillation algorithm

```
Input: training set $D_{train}$, label set $Y$, learning rate $lr$
Initialization parameters: $\theta_{stu1}$ for Student Sub-network 1, $\theta_{stu2}$ for Student Sub-network 2
Repeat:
  1: Randomly selected data $X$ from the training set $D_{train}$.
  2: Pre-trained multi-branch teacher sub-model $T_m$.
  3: Generate single-branch teacher sub-networks $T_s$ based on multi-branch teacher sub-networks $T_m$.
  4: Update the parameter $\theta_{stu1}$ of Student Sub-network 1:
       $\theta_{stu1} \leftarrow \theta_{stu1} + lr \frac{\partial L_{final_{S1}}}{\partial \theta_{stu1}}$   (15)
  5: Update the parameter $\theta_{stu2}$ of Student Sub-network 2:
       $\theta_{stu2} \leftarrow \theta_{stu2} + lr \frac{\partial L_{final_{S2}}}{\partial \theta_{stu2}}$   (16)
End: Student sub-networks $S_1$ and $S_2$ converge.
```

4. Experimentation and Results Discussion

In this section, we perform several sets of comparative experiments and rigorously analyze the experimental results of the CKD framework on NWPU-RESISC45 [44] and SIRI-WHU [45].

4.1. Datasets

NWPU-RESISC45: The NWPU-RESISC45 dataset contains a total of 45 remote sensing scenes, and each scene consists of 700 images with a size of 256 × 256 pixels. The NWPU-RESISC45 dataset exhibits rich variation in appearance, spatial resolution, illumination, background, and occlusion.

SIRI-WHU: The SIRI-WHU dataset is composed of 12 categories of remote sensing scene images, with a total of 2400 images, and each category consists of 200 images with a size of 200 × 200 pixels. The data were obtained from Google Earth and mainly cover urban areas in China.

4.2. Implementation Details

We developed our proposed collaborative consistent distillation framework based on Pytorch and conducted the related experiments with 6 NVIDIA GeForce GTX 3080Ti GPUs. In our experiments, we used the Adam optimizer [46] to optimize the parameters of the network, setting the initial learning rate to 0.01, the momentum factor to 0.9, the weight decay rate to 10^{-4}, and the batch size to 256. The model was trained for a total of 300 epochs, and the learning rate decreased to 1/10 of the previous learning rate for each 60 epoch iteration.

Since the fully connected layer in the convolutional neural network restricts the input image size, therefore it is necessary to pre-process the images in the dataset when the relevant model is trained on the NWPU-RESISC45 and SIRI-WHU datasets. For the training set, firstly, the input images based on NWPU-RESISC45 dataset were randomly cropped to 200 × 200 after mirror filling to standardize the size of the input images. It is worth noting that for the remote sensing image size in the SIRI-WHU dataset, we still kept 200 × 200. Secondly, in order to enrich the training set and improve the generalization ability of the model, a simple random left–right flip operation was performed on the training set. Finally, the images were processed by normalization. For the testing set, the input images were center cropped, and the size was set to 200 × 200, which unified the size of the input image between the training set and the testing set. Similarly, the normalization operation was performed on the testing set images.

4.3. Comparison of Remote Sensing Image Scene Classification Methods on SIRI-WHU and NWPU-RESISC45 Datasets

To evaluate the remote sensing image scene classification performance of the CKD student sub-networks, the network model was compared with other deep learning models based on the SIRI-WHU and NWPU-RESISC45 datasets. The average classification accuracy in the experiments and a series of other evaluation metrics are shown in Table 1. The CKD student sub-networks proposed in this paper had the highest classification accuracy of 0.916 for the NWPU-RESISC45 dataset and 0.943 for the SIRI-WHU dataset.

Table 1. The evaluation metrics results while using different deep learning models on NWPU-RESISC45 and SIRI-WHU.

Dataset	Methods	Image Size	Acc	Precision	Recall	F1
NWPU-RESISC45	AlexNet	200 × 200	0.872	0.876	0.869	0.869
	GoogLeNet	200 × 200	0.886	0.897	0.893	0.892
	ResNet 50	200 × 200	0.874	0.879	0.875	0.875
	Inception V1	200 × 200	0.813	0.824	0.817	0.815
	Inception V2	200 × 200	0.887	0.894	0.891	0.891
	MobileNet	200 × 200	0.882	0.887	0.884	0.885
	VGG16	200 × 200	0.879	0.884	0.882	0.881
	Xception	200 × 200	0.872	0.879	0.874	0.875
	Ours	200 × 200	0.916	0.923	0.917	0.917
SIRI-WHU	AlexNet	200 × 200	0.887	0.892	0.889	0.882
	GoogLeNet	200 × 200	0.916	0.921	0.917	0.915
	ResNet 50	200 × 200	0.912	0.918	0.913	0.914
	Inception V1	200 × 200	0.873	0.882	0.875	0.876
	Inception V2	200 × 200	0.928	0.932	0.924	0.926
	MobileNet	200 × 200	0.908	0.917	0.912	0.914
	VGG16	200 × 200	0.903	0.915	0.908	0.912
	Xception	200 × 200	0.914	0.923	0.916	0.917
	Ours	200 × 200	0.943	0.948	0.945	0.942

4.4. Comparison with the State-of-the-Art Knowledge Distillation Methods on the NWPU-RESISC45 and SIRI-WHU Datasets

To more comprehensively evaluate our CKD framework, we also compared CKD with recent state-of-the art knowledge distillation methods reported on the SIRI-WHU and NWPU-RESISC45 datasets in Tables 2 and 3. We used two baselines to evaluate the classification performance of our CKD-RPO framework. Specifically, the first type of baseline employs a series of offline distillation methods, including KD [19], FN [32], AE-KD [33], LKD [31], RKD [29], and CRD [30]. The second kind of the baseline is the online knowledge distillation methods, which were DML [35], ONE [34], DCM [36], and PCL [37].

The experimental results are reported in Table 2, where the best results are marked in bold. Experimental results on the SIRI-WHU and NWPU-RESISC45 datasets showed that the proposed CKD achieved the best performance not only on offline distillation methods, but also on online knowledge distillation methods compared to other state-of-the-art methods. It also demonstrates that our CKD was capable of enhancing classification tasks. Despite the fact that the experimental setup in these references varied slightly, it appears that our strategy outperformed previous state-of-the-art methods.

Table 2. The accuracy of ResNet20 while using different knowledge distillation approaches on SIRI-WHU and NWPU-RESISC45.

Methods	Types	SIRI-WHU	NWPU-RESISC45
DML [35]	online	91.3%	86.9%
KD [19]	offline	91.7%	87.3%
RKD [29]	offline	91.2%	86.4%
CRD [30]	offline	91.4%	87.6%
FN [32]	offline	90.8%	-
LKD [31]	offline	-	88.4%
AE-KD [33]	offline	-	87.1%
Ours	offline	92.0%	90.5%

Table 3. The accuracy of ResNet110 while using different knowledge distillation approaches on SIRI-WHU and NWPU-RESISC45.

Methods	Types	SIRI-WHU	NWPU-RESISC45
DML [35]	online	92.6%	85.7%
KD [19]	offline	91.4%	84.2%
RKD [29]	offline	92.1%	85.3%
CRD [30]	offline	91.8%	86.3%
DCM [36]	online	-	87.9%
ONE [34]	online	92.3%	88.5%
PCL [37]	online	92.5%	90.3%
Ours	offline	94.3%	91.6%

When using the larger student sub-network the Resnet110, as shown in Table 3, we performed relevant comparison experiments on the SIRI-WHU and NWPU-RESISC45 datasets in order to further evaluate the performance of the CKD framework by the top-1 accuracy. We can obviously observe that our model exhibited significant advantages against the state-of-the-art methods. With the CKD framework, Resnet110 continued to be able to achieve an appreciable performance improvement for classification tasks.

4.5. Comparison of the Number of Parameters among CKD Student Sub-Networks and Resnet Networks

The parameters' variability between different student sub-networks in the CKD framework for the backbone networks Resnet20, Rconv_Res20, Resnet32, Rconv_Res32, Resent 56, Rconv_Res56, Resnet110, and Rconv_Res110 is compared in this section. The number of parameters of different student sub-networks is shown in Table 4.

Table 4. The ablation experiment results of different student sub-networks in the CKD framework.

Network Types		Params		Gain (↑)
Resnet 20	Rconv_Res20	0.27 M	0.15 M	0.12 M
Resnet 32	Rconv_Res32	0.46 M	0.24 M	0.22 M
Resnet 56	Rconv_Res56	0.85 M	0.47 M	0.38 M
Resnet 110	Rconv_Res110	1.73 M	0.90 M	0.83 M

From Table 4, we can find that the Rconv_Res-series student sub-networks maintained comparatively fewer parameters. As the depth of the network deepened, the number of parameters in the Resnet-series student sub-networks increased more significantly compared to the Rconv_Res-series student sub-networks. This also demonstrates that the redundancy parameters of the Resnet-series student sub-networks increased dramatically with the increasing depth of the student sub-networks. In contrast, the Rconv_Res-series student sub-networks in the CKD framework were able to effectively eliminate redundant parameters.

4.6. Ablation Experimental

4.6.1. The Performance of the Student Sub-Networks with Different Depths in CKD Based on the SIRI-WHU Dataset

The backbone networks that we employed in our experiments included typical student-level backbone networks: Resnet20 [47], Resnet32, Resnet56, and Resnet110 and large-scale backbone networks at the teacher level: Resnet110 and Densenet121 [48].

Table 5 compares the top-1 accuracy [49] on the SIRI-WHU dataset obtained by various architectures under the two-student sub-network condition. We can observe the following conclusions from Table 5:

(1) For Student Sub-networks 1 and 2, the collaborative consistency distillation algorithm (CKD) significantly improved the classification accuracy of each student sub-network, and the gain values indicate the gains of each student sub-network.
(2) Although Rconv_Res110 is a much larger backbone network than Rconv_Res32, it still benefited from being trained with a smaller student sub-network.
(3) The smaller student sub-networks can usually gain more from the collaborative consistency distillation algorithm.

Table 5. Accuracy (%) on the SIRI-WHU dataset. CKD measures the difference in accuracy between the network learned with CKD and the same network learned independently. "Gain" indicates the percentage improvement of the accuracy rate.

Network Types		**Independent Acc %**		**CKD-RPO Acc %**		**Gain (↑)**	
Sub_Stu1	**Sub_Stu2**	**Sub_Stu1**	**Sub_Stu2**	**Sub_Stu1**	**Sub_Stu2**	**Sub_Stu1**	**Sub_Stu2**
Resnet 32	Resnet 20	91.3	90.8	91.6	91.4	0.3	0.6
Resnet 32	Resnet 32	91.3	91.3	92.0	92.0	0.7	0.7
Resnet 56	Resnet 32	91.8	91.3	92.7	91.8	0.9	0.5
Resnet 110	Resnet 32	93.2	91.3	93.6	91.9	0.4	0.6
Rconv_Res32	Rconv_Res20	92.1	91.4	92.3	92.0	0.2	0.6
Rconv_Res32	Rconv_Res32	92.1	92.1	92.9	92.9	0.8	0.8
Rconv_Res56	Rconv_Res32	92.7	92.1	93.4	92.6	0.7	0.5
Rconv_Res110	Rconv_Res32	93.8	92.1	94.3	92.8	0.5	0.7

4.6.2. The Effectiveness of Each Component in the Redundant Feature Mapping Operation

To more comprehensively evaluate our CKD framework, we conducted ablation studies to analyze the correlation between different components in the redundant feature mapping operation. The redundant feature mapping operation of the CKD framework mainly involves three components: Rconv module, MRFM, and SRFM. To investigate the impact of each component on the redundancy mapping module on the CKD framework, based on Resnet, we set a series of student sub-networks of different depths and their corresponding variants and compared the performance between the student sub-networks of different depths and the corresponding variants.

Resnet: A series of image classification models was constructed based on Resnet20, Resnet32, Resnet56, and Resnet110.

Resnet with Rconv module: Only the Rconv module was applied to a series of image classification models.

Resnet with MRFM: A series of classification models was reconstructed based on the MRFM module to predict the categories of remote sensing image scenes in the SIRI-WHU and NWPU-RESISC45 datasets.

Resnet with SRFM: First, we obtained the SRFM module through the MRFM module. Then, based on the SRFM module, a series of classification models was redesigned to predict the categories of remote sensing image scenes in the SIRI-WHU and NWPU-RESISC45 datasets.

The results of the ablation experiments of the redundant feature mapping operation in the CKD framework are shown in Figure 3.

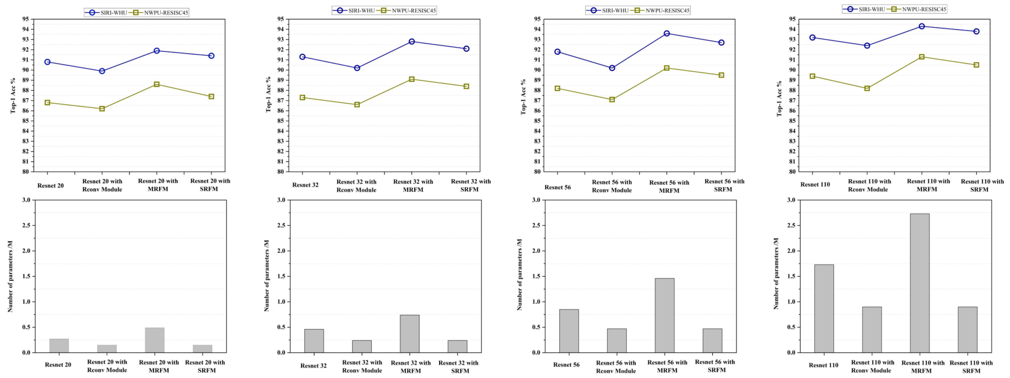

Figure 3. In order to investigate the impact of each component in the redundancy mapping module on the CKD framework, we compared the performance of a series of student sub-networks (Resnet20, Resnet32, Resnet56, and Resnet110) of different depths and the corresponding variants.

From Figure 3, we can draw the following conclusions:

(1) Resnet with the Rconv module showed the worst classification performance among the methods for all datasets. This shows that reconstructing the Resent model with only the simple Rconv module, although it can reduce the parameter redundancy of the networks, can also lead to a degradation of the model classification performance.

(2) Resnet with MRFM achieved the best classification performance. However, the number of parameters of the models was relatively more compared to Resnet with SRFM. At the same time, the improvement in classification accuracy of the models was insignificant, and we believe that it is not worthwhile to gain a slight improvement in the classification performance through such a scale of the number of parameters.

(3) With the number of parameters keeping consistent, Resnet with SRFM possessed better classification performance compared to Resnet with the Rconv module. This indicates that the equivalent convolutional kernel obtained by the multi-branch fusion operation exhibited a more powerful feature extraction ability, which effectively improved the classification performance of the model.

5. Conclusions

In this work, we proposed a collaborative consistent knowledge distillation framework in order to reduce the parameters of the remote sensing image scene classification model and further facilitate the deployment on embedded devices with poor hardware conditions. Our framework consisted of two main aspects: the redundant feature mapping module and the collaborative consistency distillation algorithm. The experimental results on two benchmark datasets, SIRI-WHU and NWPU-RESISC45, showed that our framework significantly improved the remote sensing image scene classification performance of the student sub-network and substantially reduced the redundant parameters of the backbone network. In addition, the pre-trained student sub-networks obtained by the CKD framework had a powerful feature extraction ability with fewer parameters, which can be widely used in various embedded devices.

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

Funding: This research was supported by the Natural Science Foundation of Shanxi Province Grants (20210302124257).

Data Availability Statement: The data presented in this study are available in NWPU-RESISC45 and SIRI-WHU. The details of the datasets are described in Section 4.1 and the relevant references correspond to [44,45].

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

References

1. Alam, E.; Sufian, A.; Das, A.K.; Bhattacharya, A.; Ali, M.F.; Rahman, M.M.H. Leveraging Deep Learning for Computer Vision: A Review. In Proceedings of the 22nd International Arab Conference on Information Technology, ACIT 2021, Muscat, Oman, 21–23 December 2021; IEEE: New York, NY, USA, 2021; pp. 1–8. [CrossRef]
2. Hassan, H.; Ren, Z.; Zhao, H.; Huang, S.; Li, D.; Xiang, S.; Kang, Y.; Chen, S.; Huang, B. Review and classification of AI-enabled COVID-19 CT imaging models based on computer vision tasks. *Comput. Biol. Med.* **2022**, *141*, 105123. [CrossRef] [PubMed]
3. Wong, P.K.; Luo, H.; Wang, M.; Leung, P.H.; Cheng, J.C.P. Recognition of pedestrian trajectories and attributes with computer vision and deep learning techniques. *Adv. Eng. Inform.* **2021**, *49*, 101356. [CrossRef]
4. Shen, W.; Chen, L.; Liu, S.; Zhang, Y. An image enhancement algorithm of video surveillance scene based on deep learning. *IET Image Process.* **2022**, *16*, 681–690. [CrossRef]
5. Cohen, K.; Fort, K.; Mieskes, M.; Névéol, A.; Rogers, A. Reviewing Natural Language Processing Research. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts, EACL 2021, Online, 19–20 April 2021; Augenstein, I., Habernal, I., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 14–16. [CrossRef]
6. Jiang, H. Reducing Human Labor Cost in Deep Learning for Natural Language Processing. Ph.D. Thesis, Georgia Institute of Technology, Atlanta, GA, USA, 2021.
7. Lauriola, I.; Lavelli, A.; Aiolli, F. An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools. *Neurocomputing* **2022**, *470*, 443–456. [CrossRef]
8. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef] [PubMed]
9. Agarwal, R.; Schwarzer, M.; Castro, P.S.; Courville, A.C.; Bellemare, M.G. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In Proceedings of the Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, Virtual, 6–14 December 2021; pp. 29304–29320.
10. Curi, S.; Bogunovic, I.; Krause, A. Combining Pessimism with Optimism for Robust and Efficient Model-Based Deep Reinforcement Learning. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Virtual Event, 18–24 July 2021; Volume 139, pp. 2254–2264.
11. El Kaid, A.; Brazey, D.; Barra, V.; Baïna, K. Top-Down System for Multi-Person 3D Absolute Pose Estimation from Monocular Videos. *Sensors* **2022**, *22*, 4109. [CrossRef] [PubMed]
12. Qiu, J.; Yin, J.; Qian, W.; Liu, J.; Huang, Z.; Yu, H.; Ji, L.; Zeng, X. A Novel Multiresolution-Statistical Texture Analysis Architecture: Radiomics-Aided Diagnosis of PDAC Based on Plain CT Images. *IEEE Trans. Med. Imaging* **2021**, *40*, 12–25. [CrossRef] [PubMed]
13. Huang, F.; Cao, Z.; Jiang, S.H.; Zhou, C.; Huang, J.; Guo, Z. Landslide susceptibility prediction based on a semi-supervised multiple-layer perceptron model. *Landslides* **2020**, *17*, 2919–2930. [CrossRef]
14. Wang, L.; Yoon, K. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3048–3068. [CrossRef] [PubMed]
15. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge Distillation: A Survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
16. Bucilua, C.; Caruana, R.; Niculescu-Mizil, A. Model compression. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006.
17. Ba, J.; Caruana, R. Do Deep Nets Really Need to be Deep? In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 2654–2662.
18. Urban, G.; Geras, K.J.; Kahou, S.E.; Aslan, Ö.; Wang, S.; Mohamed, A.; Philipose, M.; Richardson, M.; Caruana, R. Do Deep Convolutional Nets Really Need to be Deep and Convolutional? In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
19. Hinton, G.E.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.
20. Yang, Y.; Newsam, S.D. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, San Jose, CA, USA, 3–5 November 2010; Agrawal, D.; Zhang, P.; Abbadi, A.E.; Mokbel, M.F., Eds.; ACM: New York, NY, USA, 2010; pp. 270–279. [CrossRef]
21. Zhu, Q.; Zhong, Y.; Wu, S.; Zhang, L.; Li, D. Scene Classification Based on the Sparse Homogeneous-Heterogeneous Topic Feature Model. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2689–2703. [CrossRef]
22. Liu, L.; Wang, P.; Shen, C.; Wang, L.; van den Hengel, A.; Wang, C.; Shen, H.T. Compositional Model Based Fisher Vector Coding for Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2335–2348. [CrossRef] [PubMed]
23. Yao, Y.; Liang, H.; Li, X.; Zhang, J.; He, J. Sensing Urban Land-Use Patterns by Integrating Google Tensorflow and Scene-Classification Models. *arXiv* **2017**, arXiv:1708.01580.
24. 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]

25. Gong, X.; Xie, Z.; Liu, Y.; Shi, X.; Zheng, Z. Deep Salient Feature Based Anti-Noise Transfer Network for Scene Classification of Remote Sensing Imagery. *Remote Sens.* **2018**, *10*, 410. [CrossRef]
26. Alkhulaifi, A.; Alsahli, F.; Ahmad, I. Knowledge distillation in deep learning and its applications. *PeerJ Comput. Sci.* **2021**, *7*, e474. [CrossRef] [PubMed]
27. Blakeney, C.; Li, X.; Yan, Y.; Zong, Z. Parallel blockwise knowledge distillation for deep neural network compression. *IEEE Trans. Parallel Distrib. Syst.* **2020**, *32*, 1765–1776. [CrossRef]
28. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. FitNets: Hints for Thin Deep Nets. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
29. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational Knowledge Distillation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; Computer Vision Foundation/IEEE: New York, NY, USA, 2019; pp. 3967–3976. [CrossRef]
30. Tian, Y.; Krishnan, D.; Isola, P. Contrastive Representation Distillation. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
31. Li, X.; Wu, J.; Fang, H.; Liao, Y.; Wang, F.; Qian, C. Local Correlation Consistency for Knowledge Distillation. In Proceedings of the Computer Vision—ECCV 2020—16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XII; Vedaldi, A., Bischof, H., Brox, T., Frahm, J., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12357, pp. 18–33. [CrossRef]
32. Xu, K.; Rui, L.; Li, Y.; Gu, L. Feature Normalized Knowledge Distillation for Image Classification. In Proceedings of the Computer Vision—ECCV 2020—16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXV; Vedaldi, A., Bischof, H., Brox, T., Frahm, J., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12370, pp. 664–680. [CrossRef]
33. Du, S.; You, S.; Li, X.; Wu, J.; Wang, F.; Qian, C.; Zhang, C. Agree to Disagree: Adaptive Ensemble Knowledge Distillation in Gradient Space. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual, 6–12 December 2020.
34. Lan, X.; Zhu, X.; Gong, S. Knowledge Distillation by On-the-Fly Native Ensemble. In Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montreal, QC, Canada, 3–8 December 2018; pp. 7528–7538.
35. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep Mutual Learning. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; Computer Vision Foundation/IEEE Computer Society: Washington, DC, USA, 2018; pp. 4320–4328. [CrossRef]
36. Yao, A.; Sun, D. Knowledge Transfer via Dense Cross-Layer Mutual-Distillation. In Proceedings of the Computer Vision—ECCV 2020—16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XV; Vedaldi, A., Bischof, H., Brox, T., Frahm, J., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12360, pp. 294–311. [CrossRef]
37. Wu, G.; Gong, S. Peer Collaborative Learning for Online Knowledge Distillation. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021, 33rd Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, 2–9 February 2021; AAAI Press: Menlo Park, CA, USA, 2021; pp. 10302–10310.
38. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 3431–3440. [CrossRef]
39. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016.
40. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 764–773. [CrossRef]
41. Ding, X.; Guo, Y.; Ding, G.; Han, J. ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, 27 October–2 November 2019; IEEE: New York, NY, USA, 2019; pp. 1911–1920. [CrossRef]
42. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; Computer Vision Foundation/IEEE Computer Society: Washington, DC, USA, 2018; pp. 7132–7141. [CrossRef]
43. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features From Cheap Operations. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; Computer Vision Foundation/IEEE: New York, NY, USA, 2020; pp. 1577–1586. [CrossRef]
44. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
45. Zhao, B.; Zhong, Y.; Xia, G.; Zhang, L. Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 2108–2123. [CrossRef]

46. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington, DC, USA, 2016; pp. 770–778. [CrossRef]
48. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 2261–2269. [CrossRef]
49. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable Person Re-identification: A Benchmark. In Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 1116–1124. [CrossRef]