remote sensing

MDPI

Article

Learning Lightweight and Superior Detectors with Feature Distillation for Onboard Remote Sensing Object Detection

Lingyun Gu [1], Qingyun Fang [2], Zhaokui Wang [2], Eugene Popov [1] and Ge Dong [2,*]

[1] Institute of Electronics and Telecommunications, Peter the Great Saint-Petersburg Polytechnic University, 195251 Saint Petersburg, Russia
[2] School of Aerospace Engineering, Tsinghua University, Beijing 100084, China
* Correspondence: dongge@tsinghua.edu.cn

Abstract: CubeSats provide a low-cost, convenient, and effective way of acquiring remote sensing data, and have great potential for remote sensing object detection. Although deep learning-based models have achieved excellent performance in object detection, they suffer from the problem of numerous parameters, making them difficult to deploy on CubeSats with limited memory and computational power. Existing approaches attempt to prune redundant parameters, but this inevitably causes a degradation in detection accuracy. In this paper, the novel Context-aware Dense Feature Distillation (CDFD) is proposed, guiding a small student network to integrate features extracted from multi-teacher networks to train a lightweight and superior detector for onboard remote sensing object detection. Specifically, a Contextual Feature Generation Module (CFGM) is designed to rebuild the non-local relationships between different pixels and transfer them from teacher to student, thus guiding students to extract rich contextual features to assist in remote sensing object detection. In addition, an Adaptive Dense Multi-teacher Distillation (ADMD) strategy is proposed, which performs adaptive weighted loss fusion of students with multiple well-trained teachers, guiding students to integrate the learning of helpful knowledge from multiple teachers. Extensive experiments were conducted on two large-scale remote sensing object detection datasets with various network structures; the results demonstrate that the trained lightweight network achieves auspicious performance. Our approach also shows good generality for existing state-of-the-art remote sensing object detectors. Furthermore, by experimenting on large general object datasets, we demonstrate that our approach is equally practical for general object detection distillation.

Keywords: CubeSat; remote sensing; object detection; lightweight network; contextual feature; multi-teacher distillation

Citation: Gu, L.; Fang, Q.; Wang, Z.; Popov, E.; Dong, G. Learning Lightweight and Superior Detectors with Feature Distillation for Onboard Remote Sensing Object Detection. *Remote Sens.* **2023**, *15*, 370. https://doi.org/10.3390/rs15020370

Academic Editor: Edoardo Pasolli

Received: 30 October 2022
Revised: 17 December 2022
Accepted: 5 January 2023
Published: 7 January 2023

1. Introduction

With the launch of numerous CubeSats, nanosatellites, and microsatellites, the cost of obtaining and processing remote sensing data has been decreased dramatically [1], which has greatly aided the advancement of Earth observation technology. Remote sensing object detection is an important task of Earth observation, which refers to identifying and locating specific objects from satellite images. It is essential in various missions such as ocean monitoring, disaster prevention, and environmental monitoring [2].

In recent years, deep learning-based object detection approaches have shown impressive performance in general scenes [3–5]. However, compared to ordinary scenes on the ground, remote sensing scenes include complicated backgrounds, dramatic changes in shooting angles and illumination, sharp scale changes, and small object sizes, making it difficult to obtain sufficient information about targets to perform accurate detection.

In order to obtain better performance, more complex network architectures are applied to perform remote sensing object detection tasks; however, these suffer from a massive number of parameters and excessive consumption of computational resources. Due to the

limited memory and computational power of CubeSat onboard platforms, it is difficult to deploy large-scale deep networks on them [6].

To deploy neural networks on small satellites, many studies have been conducted on the software and hardware dimensions. For the hardware dimension, Manning et al. [7] deployed Convolutional Neural Networks (CNNs) on FPGAs to classify images captured by the ISS SHREC platform. Arechiga et al. [8] implemented CNNs on the Nvidia Jetson TX1 for onboard image processing on small satellites. Bappyet al. [9] proposed onboard deep neural computation and machine learning models to analyse and process multiple spectral images for the 3U CubeSat. Giuffrida et al. [10] deployed a CloudScout CNN on the Myriad 2 vision processing unit for cloud detection on hyperspectral images. These studies explore the possibility of deploying CNNs on small satellites.

For the software dimension, previous work has focused on developing efficient and lightweight deep models. For example, MobileNets [11] designs depthwise separable convolution with few channels and small convolution kernels to reduce the number of network parameters. Parameter pruning [12] reduces the model size by removing unnecessary parameters from the deep neural network. Although the above approaches are able to compress the model and improve the speed of detection, they more or less lead to a reduction in model accuracy.

In recent years, knowledge distillation has received increasing attention. This refers to inheriting information from an extensive teacher network into a lightweight student network, thereby enhancing the performance of the lightweight network. Depending on the location of the distillation, it can be divided into two categories. The first is logits-based distillation [13,14], where teachers are distilled from the output level, while the second is feature-based distillation [15,16], where teachers are distilled from the middle feature layers. Compared to the logits-based approaches, feature-based distillation has demonstrated advantages for various tasks.

In general object detection scenes, most feature-based distillation forces students to mimic the teacher's output as closely as possible, as the teacher's features are representative. However, previous work [2,17] has demonstrated that contextual features can compensate for the little information available on remote sensing objects and effectively assist in detecting small objects, thus playing an important role in remote sensing detection tasks. Thus, this paper explores distilling the pixel distribution in the teacher's feature map and the contextual relationships in the teacher's feature map.

To explore the differences in student instruction with different teachers, we visualised the output of the middle feature layer of three teacher networks. As shown in Figure 1, it is clear that different teachers have different regions of interest. Therefore, learning the features from different teachers is vital for improving detection accuracy.

Based on the above observations, we propose a novel feature distillation approach, which considers contextual features and integrated knowledge from multiple teachers, to train lightweight and advanced detectors for onboard remote sensing object detection. We name this approach Context-aware Dense Feature Distillation (CDFD).

Specifically, multiple two-stage teachers are first trained offline and their weights are frozen at the end of training, followed by the teacher's neck and head parameters being inherited to the student. Then, the student's training starts. To address the problem of small object information in remote sensing images, a Contextual Feature Generation Module (CFGM) is proposed to learn context from the teacher's intermediate layer output as complementary information to enhance the student's capability to detect small objects. In addition, an Adaptive Dense Multi-teacher Distillation (ADMD) strategy is proposed, which performs adaptive weighted loss fusion of students with multiple well-trained teachers to combine knowledge from multiple teachers and improve students' feature representation. Extensive experiments were conducted on remote sensing datasets; the results show that our CDFD is able to reproduce or even exceed the performance of the large models with a small model without any additional computation. In summary, the contributions of this paper are:

(1) Context-aware Dense Feature Distillation (CDFD) is proposed, which takes multiple large-scale object detection networks as teacher models and distils their feature representation capabilities into lightweight student networks, thereby significantly improving the remote sensing object detection performance of the lightweight network without introducing any additional computational effort;
(2) To address the difficulty of detecting small objects in remote sensing images, a contextual feature generation module (CFGM) is designed, which enables the student network to not only mimic the generation of the teacher's feature map, but also to extract rich contextual features to assist in the remote sensing object detection;
(3) To integrate the strengths of multiple teachers, an Adaptive Dense Multi-teacher Distillation (ADMD) strategy is proposed, which calculates the adaptive weighted loss of students and multiple teachers to integrate the detection superiority of multiple teachers into the student detector;
(4) Extensive experiments were conducted on two large-scale remote sensing object detection datasets with various network structures. The results demonstrate that both CFGM and ADMD can improve students' detection performance effectively, while students with our CDFD not only obtained state-of-the-art performance, but even outperformed most teacher networks. In addition, the CDFD has good generalisation and is applicable to several two-stage remote sensing object detection approaches.

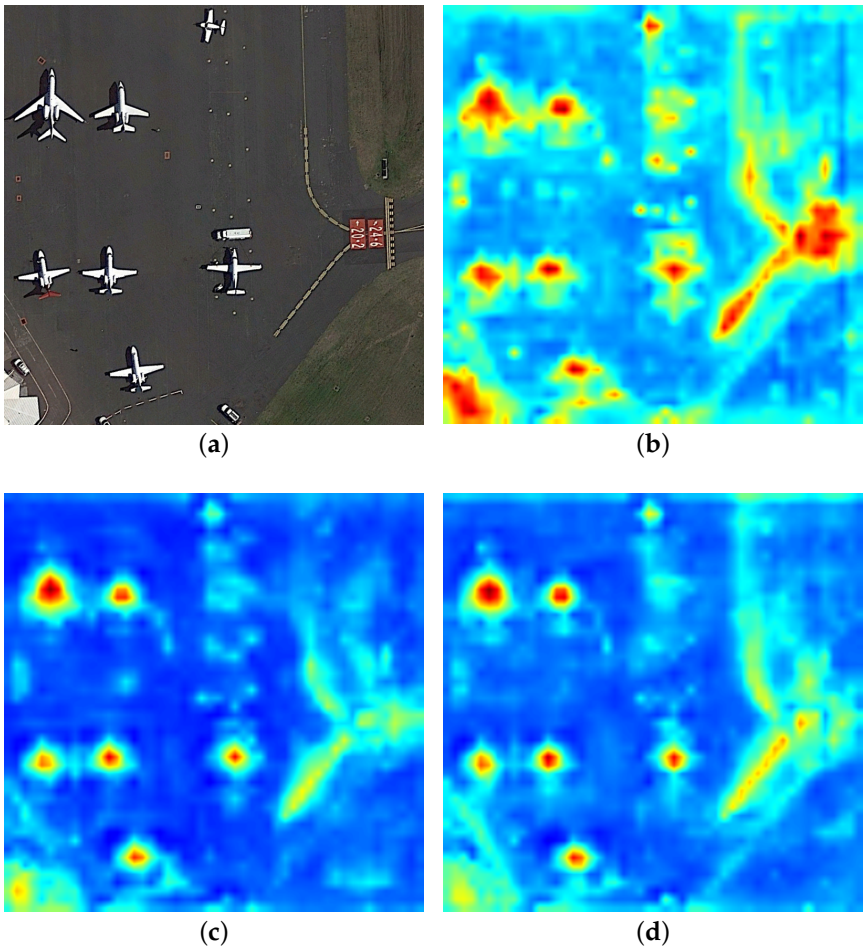


Figure 1. Visualisation of middle layer feature maps for different teacher networks. (**a**) Input image, (**b**) feature map of ResNet101, (**c**) feature map of ResNext32, (**d**) feature map of ResNext64.

2. Relation Works

2.1. CubeSats

CubeSats are small, low-mass satellites that can provide data and experimental platforms for scientific research at low cost. CubeSats were first proposed by Stanford University in 1999 and, over the past 20 years, have been developed considerably, with a wide range of applications in Earth remote sensing [18]. For example, Spire has deployed a constellation of LEO CubeSats called Lemur-2 for weather prediction and ship tracking missions [18]. Planet Labs has employed about 300 CubeSats to collect images at 3–5 m resolution for studies such as water tracking [19], vegetation monitoring [20], glacier investigation [21], permafrost monitoring [22], etc. The Dove Cluster [20,23] carries payloads including optical telescopes and high-resolution cameras to conduct Earth surface imaging, and the data obtained can be applied in the field of machine learning. Typical CubeSats have between 32KB and 8MB of on-board memory [6]; some CubeSats can carry up to 8GB of additional flash memory [24], but still cannot store excessive amounts of data. In addition, CubeSats have limited downlink capability, with most of them having a data transfer rate of 9600 BPS [6]. Deploying object-lightweight models on CubeSats for onboard real-time object recognition can solve these memory and communication problems.

2.2. Remote Sensing Object Detection

Most existing deep learning-based remote sensing object detection approaches are transferred from approaches designed for natural scene images. However, remote sensing images are very different from natural scene images, especially in terms of small objects, scale variations, and complex backgrounds. In order to obtain better detection performance, most studies use region proposal-based approaches [5,25,26] (also known as two-stage approaches) to detect remote sensing objects [27]. For example, the cross-scale fusion strategy [28,29] enhances the feature representation of objects by fusing features across layers and improves the detection of small objects. The attention module [17,30,31] models long-range dependence and is used to improve the model's representation of spatially non-local features. The frequency-domain convolution module [2] is used to extract global features as additional information to assist in the detection of remote sensing objects. Although the above approaches achieve excellent detection accuracy, they are all improved on the basis of two-stage networks, which have the disadvantages of high computational complexity and high computational resource consumption.

2.3. Knowledge Distillation

The core idea of knowledge distillation is to learn small student models from large teacher models and thus achieve competitive performance [32]. In general, a knowledge distillation system consists of three key components [33]: knowledge types, distillation strategies, and teacher–student architectures. Depending on the type of knowledge, knowledge distillation approaches can be divided into logits-based knowledge distillation [13,14] and feature-based knowledge distillation [15,16]. Logits-based knowledge distillation enables students to directly imitate the final predictions of the teacher model, and the student model trained by this approach generally relies on the output of the last layer. Feature-based knowledge distillation uses the intermediate network layer features of the teacher model as knowledge to supervise the training of the student model; this approach solves the supervision problem of the intermediate layer of the teacher model. Obviously, the feature-based distillation strategy enables students to learn a multi-layered feature representation approach. Therefore, this strategy is widely used in computer vision tasks, such as image classification [34,35], image segmentation [36,37], action recognition [38,39], and object detection [40,41]. Different teacher networks are able to provide their own useful knowledge to student networks, and the distillation strategy of multi-teacher networks is effective for training student models [39,42].

3. Methods

3.1. Context-Aware Dense Feature Distillation

In order to fully exploit multi-scale information in remote sensing images, detectors always utilise two-stage networks and feature pyramid networks [43], since the object detection in remote sensing tasks demands excellent detection accuracy [2,29]. Based on this premise, we propose a new novel approach named *Context-aware Dense Feature Distillation* (CDFD) for such detectors, as illustrated in Figure 2.

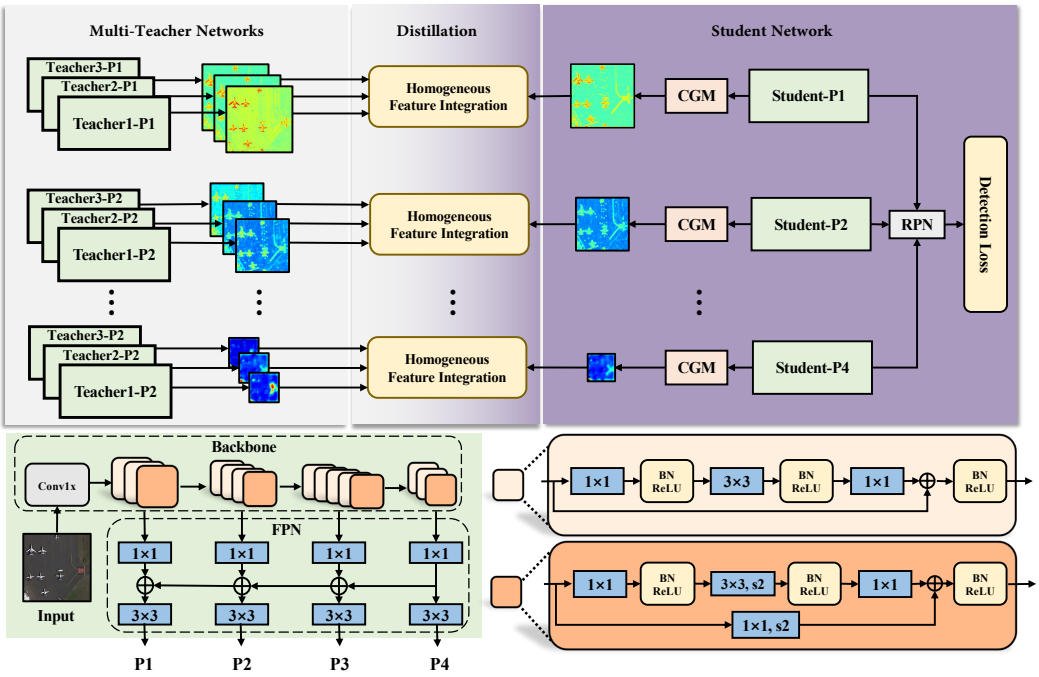


Figure 2. The overall framework of context-aware dense feature distillation.

First, the input images pass through the teacher and student backbone networks for feature extraction, followed by multi-scale features P1–P4 obtained through the neck. Then, the student network is divided into two branches. In one branch, the student features pass through the head network and then compute the detection loss of ground truth.

In the other branch, the alignment layer is applied to align the student feature maps with the teacher feature maps; then, the Contextual Feature Generation Module (CFGM) is employed to rebuild the non-local relationships between different pixels and transfer it from teacher to student. Finally, Adaptive Dense Multi-teacher Distillation (ADMD) uses an adaptive weight to balance the loss terms of each student–teacher pair.

The two main innovations of our approach are the Contextual Feature Generation Module (CFGM) and the Adaptive Dense Multi-teacher Distillation (ADMD) strategy, which are described in detail in Sections 3.2 and 3.3, respectively. In Section 3.4, the loss function of CDFD is described in detail.

3.2. Contextual Feature Generation Module

In remote sensing object detection tasks, small networks usually struggle to extract sufficient features. Non-local operations [30] have been shown to be significant for remote sensing object detection tasks by modelling the long-range dependence of different spatial locations [27]. As a result, existing detectors [2,29] introduce surrounding context as additional information about the remote sensing object to improve detection performance. Inspired by this, we designed a Contextual Feature Generation Module (CFGM) to help the student network extract global features, the framework of which is shown in Figure 3.

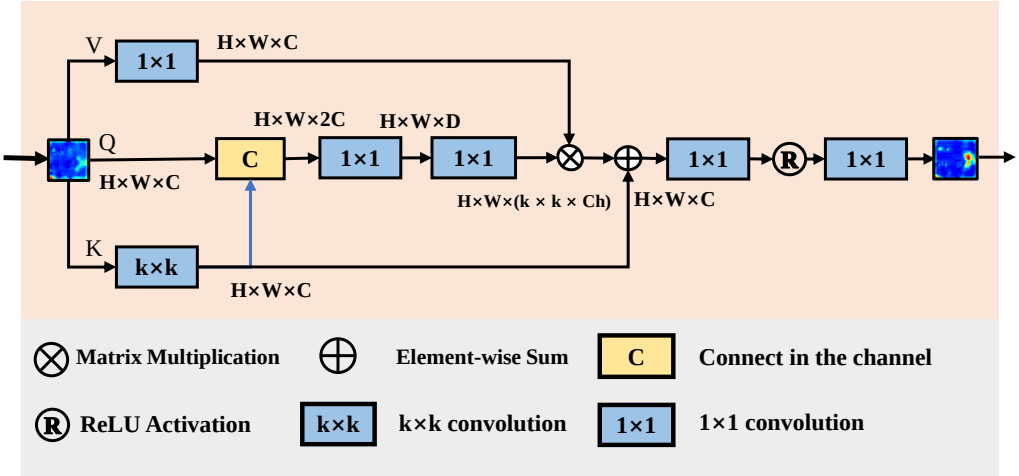


Figure 3. Description of Contextual Feature Generation Module.

Specifically, for the original student feature map $S_i \in \mathbb{R}^{H\times W\times C}$ in i-th stages, the query and key are defined as $Q = S_i$, $K = S_i$, and S_i is transformed into the value V by the embedding matrix W_v:

$$Q = S_i, \tag{1}$$

$$V = S_i W_v, \tag{2}$$

$$K = S_i, \tag{3}$$

$k \times k$ group convolution W_v is employed on all the neighbouring keys within the $k \times k$ grid to achieve a contextual representation of each key; thus, the contextual key K incorporates contextual information of each $k \times k$ grid and is called the local contextual representation:

$$K_1 = KW_{B\delta}, \tag{4}$$

where $W_{B\delta}$ represents the convolution–batch normalisation–activation layer. In this paper, the convolution kernel is 3 × 3 and the activation function is ReLU. Then, K_1 and Q are concatenated in the channel dimension and downscaled by a 1×1 convolution–normalisation–activation layer:

$$M = [K_1, Q]W_{B\delta}, \tag{5}$$

where $M \in= \mathbb{R}^{H\times W\times D}$, $D = 2C/factor$; factor equals 4 in this paper. An attention matrix is then obtained by an activation-free convolution operation:

$$A = MW_\theta, \tag{6}$$

where W_θ represents the convolution without the activation function, $A \in \mathbb{R}^{H\times W\times(k\times k\times C_h)}$ refers to the enhanced spatial-aware local relationship matrix, and C_h is the head number. Thus, each attention matrix A_j represents the j-th two-dimensional relative position embedding within each $k \times k$ grid and is shared among all C_h heads. For each head, the local attention matrix for each spatial location of A is learned based on query features and key features of the context, rather than isolated query-key pairs. Next, the attention matrix B is obtained by normalizing A and performing a softmax operation on each head along the channel dimension.

$$\widehat{A} = F_{softmax}(A), \tag{7}$$

Then, we obtain the global context G by aggregating all the values V based on the attention matrix A of the context:

$$G = V \circledast \widehat{A}, \tag{8}$$

where $\circledast$ denotes the local matrix multiplication operation that measures the pair-wise relations between each query and the corresponding keys within the local $k \times k$ grid in space; thus, global context information is obtained. Then, the local context K_1 and the global context G are fused to obtain the student feature map. Next, the student feature map attempts to generate the teacher's feature map by two convolution operations, which can be formulated as:

$$S'_i = K_1 + G, \tag{9}$$

$$f_{align}(W_{i2}(\sigma(W_{i1}(S'_i)))) \rightarrow T_i. \tag{10}$$

where S'_i is the learned student feature map in i-th stages and f_{align} denotes the align layer, which is a 1×1 convolutional layer with the same number of input channels as the student feature map and the same number of output channels as the teacher feature map. σ is the ReLU activation function and W_{i1} and W_{i2} are 3×3 convolution layers. T_i is the teacher feature map in i-th stages.

3.3. Adaptive Dense Multi-Teacher Distillation Strategy

As shown in Figure 1, different teacher networks have different regions of interest so, compared to mono-teacher models, multi-teacher models can contribute knowledge from multiple teachers to students and provide more useful knowledge. Figure 4a,b display two generic frameworks for sparse mono-teacher distillation and sparse multi-teacher distillation. Figure 4c shows dense mono-teacher distillation, while Figure 4d illustrates our proposed dense multi-teacher distillation scheme. It expands upon the original multi-teacher scheme by introducing the idea of dense distillation, i.e., knowledge transfer at different feature levels.

In the field of object detection, there is wide agreement that different-level features have distinct advantages [43,44]. Deep features are more important for a network to recognise large objects, since they contain the wider receptive fields and more semantic information than shallow features. On the other hand, shallow features are better at localizing small objects because they retain more spatial information. Guided by this view, our proposed approach can transfer knowledge from different teachers and different stages to the student model through a dense distillation scheme, which ultimately improves the detection accuracy of the student detector.

The straightforward approach to distilling knowledge from multiple teachers is to utilise the supervision signal as the average response from all teachers [32]. However, this approach is obviously oversimplified and cannot effectively utilise the unique knowledge of different teachers. To further efficiently perform knowledge transfer of multiple teachers, an adaptive dense multi-teacher distillation strategy based on adaptive weighted loss is proposed. Specifically, the feature representation learning of student networks is achieved through a well-designed distillation function.

For feature-based knowledge transfer, the distillation loss $\mathcal{L}_{Dis}$ can be formulated as

$$\mathcal{L}_{Dis} = \mathcal{L}_F(f_t(x), f_s(x)), \tag{11}$$

where $f_t(x)$ and $f_s(x)$ are the feature maps of the intermediate layers of the teacher and student models, respectively. The feature maps of the teacher and student models are matched using a similarity function, as indicated by $\mathcal{L}_F(.)$. The similarity function $\mathcal{L}_F(.)$ adopts the weighted Mean Squared Error (MSE) loss in this paper. Figure 5 depicts the procedures for our adaptive dense multi-teacher distillation strategy.

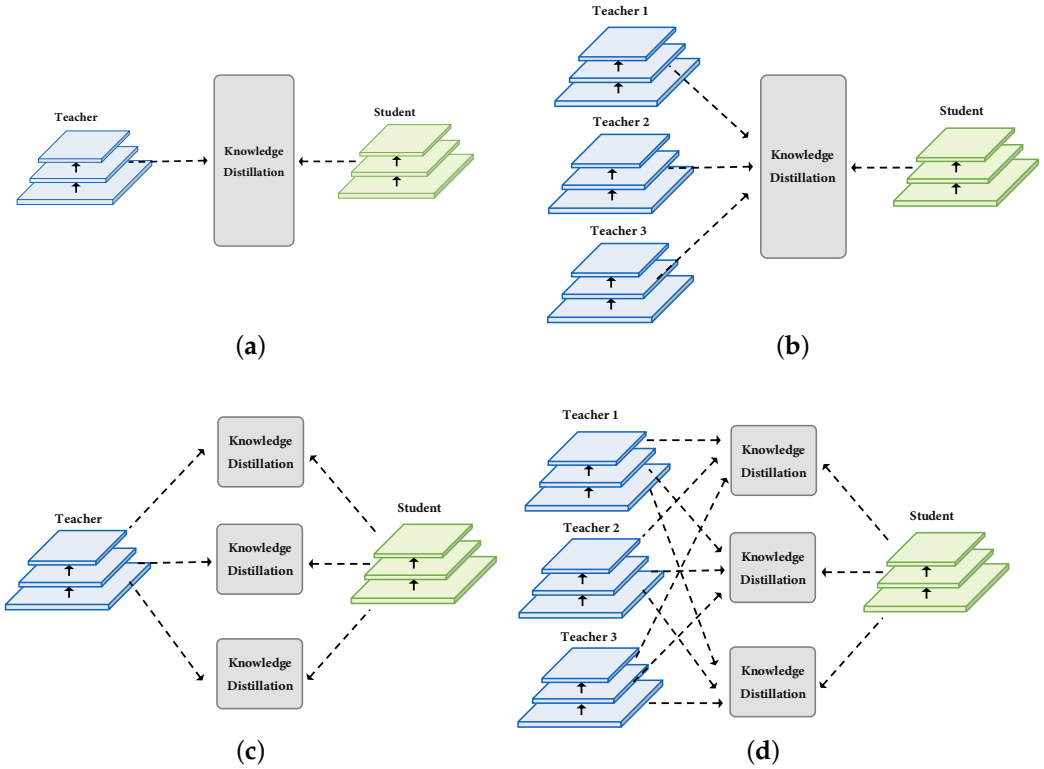


Figure 4. Four different distillation strategies. Due to the varied knowledge from different teachers and different stages, dense multi-teacher knowledge distillation can typically offer rich knowledge and fine-tune a better student model. (**a**) Sparse mono-teacher distillation, (**b**) sparse multi-teacher distillation, (**c**) dense mono-teacher distillation, (**d**) dense multi-teacher distillation.

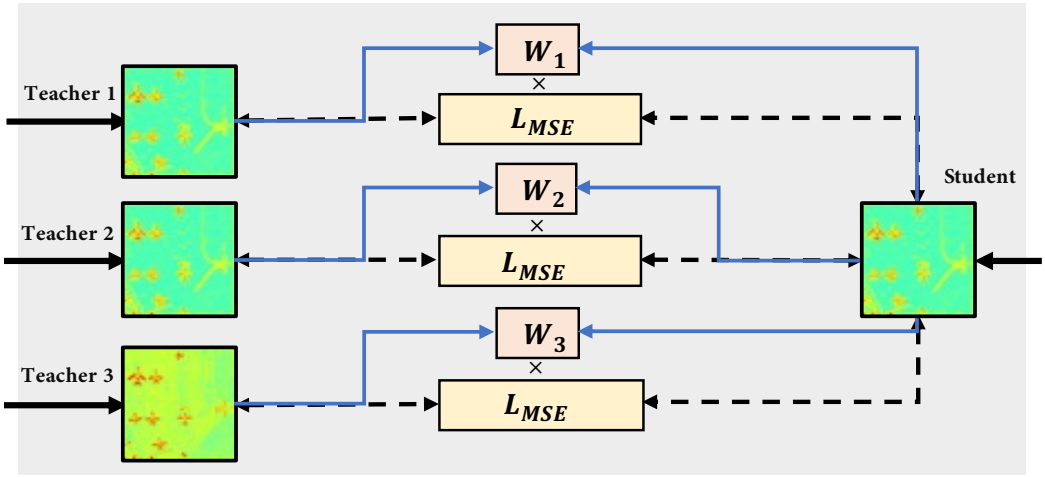


Figure 5. Adaptive dense multi-teacher distillation strategy.

It is important to note that the weight values can be dynamically adjusted by network learning rather than being fixed. However, since the weight is unbounded, it will increase the possibility of training instability. Therefore, to constrain the value range of each weight,

we resort to weight normalisation. Specifically, a softmax-based weighted loss function is designed for $\mathcal{L}_F(.)$, which can be formalised as:

$$\begin{aligned}\mathcal{L}_{Dis} &= \sum_{i=1}^{M}\sum_{j=1}^{N}\mathcal{L}_F\left(f_{t_{ij}}(x), f_{s_i}(x)\right)\\ &= \sum_{i=1}^{M}\sum_{j=1}^{N}\mathrm{F}_{\mathrm{softmax}}(w_{ij})\cdot\mathcal{L}_{MSE}\left(f_{t_{ij}}(x), f_{s_i}(x)\right)\\ &= \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{e^{w_{ij}}}{\sum_{j=1}^{N}e^{w_{ij}}}\cdot\mathcal{L}_{MSE}\left(f_{t_{ij}}(x), f_{s_i}(x)\right),\end{aligned} \tag{12}$$

where i is the i-th stage feature and M is the number of stages. j is the j-th teacher model and N is the number of teachers.

3.4. Overall Loss

With the proposed distillation loss $\mathcal{L}_{Dis}$ for CDFD, all the models are trained with the total loss as follows:

$$\mathcal{L}_{All} = \mathcal{L}_{Dis} + \alpha\cdot\mathcal{L}_{Det}, \tag{13}$$

where $\mathcal{L}_{Det}$ denotes the detection loss for the original student models and α is a hyper-parameter to balance the two losses. In this paper, the hyper-parameter α is equal to 5×10^{-7} and the detection loss includes the classification loss and the regression loss [5]:

$$\mathcal{L}_{Det} = \frac{1}{N_{cls}}\sum_i\mathcal{L}_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i\mathcal{L}_{reg}(t_i, t_i^*), \tag{14}$$

where i refers to the index of an anchor in a mini-batch, p_i denotes the predicted probability of anchor i being an object, and p_i^* is the ground truth label. t_i denotes a vector representing the four parameterised coordinates of the predicted bounding box, while t_i^* is the ground truth box associated with a positive anchor. $\mathcal{L}_{cls}$ is the classification loss and $\mathcal{L}_{reg}$ is the regression loss. For classification losses, the cross-entropy loss function is used:

$$\mathcal{L}_{cls}(p_i, p_i^*) = -log[p_i^*p_i + (1-p_i^*)(1-p_i)], \tag{15}$$

For the regression loss, the smooth L1 function is used:

$$\mathcal{L}_{reg}(t_i, t_i^*) = R(t_i - t_i^*), \tag{16}$$

$$R = \mathcal{L}_{Smooth1}(x) = \begin{cases}0.5*x^2, |x|<1\\ |x|-0.5, otherwise\end{cases} \tag{17}$$

In this paper, the calculation of the detection loss is unchanged and the distillation loss is only calculated on the feature map, which can be obtained from the neck of the detector. Therefore, our CDFD is applicable to different detectors.

4. Experiment

4.1. Dataset

Since remote sensing datasets from Cubesats have not yet been released, our experiments were conducted on two public remote sensing datasets from regular satellites—DIOR [27] and DOTA [45]. In addition, to verify the effectiveness of CDFD on general object detection tasks, we conducted distillation experiments on the classical general object dataset COCO [46].

DIOR [27] is a large publicly available dataset for remote sensing image object detection, derived from Google Earth, with an image size of 800 × 800 pixels and a spatial resolution range of 0.5 to 30 m. Each instance is labelled using a horizontal bounding box. The DIOR dataset contains 23,463 images and 192,472 annotated instances, with 20 object

categories: airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, golf field, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill. The official guideline for splitting the DIOR dataset was followed for this paper, i.e., 11,725 remote sensing images as the training set and the remaining 11,738 images as the test set.

DOTA [45] is another large multi-scale optical remote sensing dataset. The data sources are mainly Google Earth and the JL-1 and GF-2 satellites of the China Resources Satellite Data and Applications Centre, with image sizes ranging from 800 × 800 to 4000 × 4000 pixels. For this paper, the DOTA-v1.5 dataset with horizontal bounding box annotation was used. Its training set consists of 1411 remote sensing images and the validation set consists of 458 remote sensing images. There are more than 400,000 object instances with 16 object classes: container crane, baseball field, basketball court, bridge, surface runway, harbour, helicopter, large vehicle, aircraft, roundabout, ship, small vehicle, football field, storage tank, swimming pool, and tennis court. Before training and testing, all images were cropped by sliding windows with size of 800 × 800 and overlap of 200. After the crop operation, our training set included 15,749 images, and the test set included 5297 images.

COCO [46] is one of the most widely used object detection datasets. It is achieved by collecting images of complex, everyday scenes containing common objects in the natural environment. Objects are labelled using a segmentation of each instance to aid accurate object localisation. COCO object detection dataset contains 80 object categories, with a total of 200,000 images. Following the official guideline, we used the default 120,000 images for training and 80,000 images for testing.

4.2. Evaluation Metrics

In order to quantitatively analyse the object detection results of the proposed approach, the final models were evaluated in terms of two dimensions—detection accuracy and computational complexity. Among them, detection accuracy uses the mean average precision of all categories (mAP) as a metric, which is calculated from the precision and recall,

$$precision = \frac{TP}{TP + FP}\,, \tag{18}$$

$$recall = \frac{TP}{TP + FN}\,, \tag{19}$$

where TP, FP, and FN refer to true positive, false positive, and false negative, respectively. The precision is used to describe the ability of the model to detect the correct object, while the recall describes the ability of the model to find all objects. The precision–recall curve (PRC) can be drawn from the recall and precision, and the average precision is the area under the PRC. The formulae for calculating average precision and mean average precision are as follows:

$$AP = \int_0^1 P(R)dR\,, \tag{20}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP\,, \tag{21}$$

where AP, mAP, P, and R denote average precision, mean average precision, precision, and recall, respectively, while N is the number of object categories.

Model size and giga-floating point operations (GFLOPs) were also used as metrics to evaluate the computational complexity of the model.

4.3. Implementation Details

All experiments were conducted on four RTX TITAN 24G with CUDA 10.2 and CuDNN 7.6.5 acceleration. The SGD optimiser with an initial learning rate of 0.02, momentum of 0.9, and weight decay of 0.0001 was used.

For the remote sensing datasets DIOR and DOTA, training lasted for 12 epochs; in the first 500 iterations, we used a warm-up approach to adjust the learning rate from 0.001 to 0.02. In the eighth and eleventh epochs, the learning rate was set to 0.1 times the previous one. For the general object dataset COCO, which was trained on for a total of 24 epochs, the learning rate was set to 0.1 times the previous one in the 16th and 22nd epochs, and the other settings were the same as for the remote sensing dataset. The training of three teachers used pre-trained weights from ImageNet and all new layers were initialised using Kaiming normal. The three teacher networks were trained first, and their weights were frozen at the end of training. When training the student network, an inheritance strategy [47,48] was used to initialise the students with the teacher's neck and head parameters to train the students with the same head structure. All experiments were based on the MMDetection toolbox [49] without any modifications.

4.4. Experimental Results

4.4.1. Comparison of Object Detection Results

Table 1 compares the performance of our student network with other detectors on the DIOR and DOTA datasets. The classical detectors, i.e., Faster RCNN [5], YOLOv3 [50], RetinaNet [51], and the three teacher models are compared in this experiment. Both the student and the three teacher networks used Faster RCNN networks with feature pyramid networks [43]. The backbones of the two students were ResNet18 and ResNet50 [52]; the backbones of the three teacher networks were ResNet101 [52], ResNext101 [53] with 32 groups (ResNext32), and ResNext101 with 64 groups (ResNext64). The backbone of YOLOv3 was Darknet53, and the backbone of RetinaNet was ResNet101.

Table 1. Comparison of performance on the DIOR dataset and DOTA dataset. The best results are marked by bold text. T indicates the teacher network.

Dataset	Model	Backbone	mAP	Model Size (MB)	GFLOPs
DIOR	YOLOv3 [50]	Darknet53	57.0	61.6	121.4
	RetinaNet [51]	ResNet101 [52]	66.1	55.4	179.5
	Faster RCNN [5]	ResNet18	68.4	**28.2**	**101.5**
	Faster RCNN	ResNet50	70.6	41.2	134.5
	Faster RCNN(T1)	ResNet101	70.9	60.2	182.0
	Faster RCNN (T2)	ResNext32 [53]	73.0	59.9	184.4
	Faster RCNN(T3)	ResNext64	73.0	98.9	280.3
	+Ours	ResNet18	70.9	**28.2**	**101.5**
	+Ours	ResNet50	**73.1**	41.2	134.5
DOTA	YOLOv3 [50]	Darknet53	55.1	61.6	121.4
	RetinaNet [51]	ResNet101 [52]	42.4	55.4	179.5
	Faster RCNN [5]	ResNet18	48.4	**28.2**	**101.5**
	Faster RCNN	ResNet50	56.6	41.2	134.5
	Faster RCNN(T1)	ResNet101	57.4	60.2	182.0
	Faster RCNN (T2)	ResNext32 [53]	**58.1**	59.9	184.4
	Faster RCNN(T3)	ResNext64	57.8	98.9	280.3
	+Ours	ResNet18	56.9	**28.2**	**101.5**
	+Ours	ResNet50	58.0	41.2	134.5

The experimental results show that the performance of the student networks was significantly improved with the use of CDFD in various datasets and backbones. For example, Faster RCNN-ResNet50 obtained a state-of-the-art performance of 73.1% mAP on the DIOR dataset, which not only improved the mAP by 2.5% compared to the original network, but also outperformed all the teacher networks. Furthermore, in the Faster RCNN-ResNet50 setting, the student detector performance (58.0%) was comparable to the best teacher detector (58.1%) by training in CDFD, while greatly reducing model parameters and improving computing time. Faster RCNN-ResNet50 was the most lightweight model of all the detectors, and with CDFD training, the student detector obtained mAPs of 70.9%

and 56.9% on the DIOR and DOTA datasets, respectively, improving the mAP by 2.5% and 8.5% over the original network.

Notably, with the CDFD training, the student detector even outperformed most of the teacher detectors, demonstrating that the student detectors obtained better features by learning the teacher's context and comprehensive knowledge.

4.4.2. Ablation Studies

To further demonstrate the effectiveness of the proposed CFGM and ADMD, ablation experiments were conducted on the DIOR and DOTA datasets. Both the teacher and student networks were Faster RCNNs with feature pyramid networks, while the three teacher backbones were ResNet101, ResNext32, and ResNext64, respectively. The student backbones were ResNet50 and ResNet18. The results are shown in Table 2.

Table 2. Comparisons of the impact of the CFGM and the ADMD. The best results are marked by bold text.

Dataset	Student	Baseline	CFGM	ADMD	mAP
DIOR	Faster RCNN-Res50	✓			70.6
		✓	✓		72.3
		✓		✓	72.5
		✓	✓	✓	**73.1**
	Faster RCNN-Res18	✓			68.4
		✓	✓		70.6
		✓		✓	69.5
		✓	✓	✓	**70.9**
DOTA	Faster RCNN-Res50	✓			56.6
		✓	✓		57.0
		✓		✓	57.2
		✓	✓	✓	**58.0**
	Faster RCNN-Res18	✓			48.4
		✓	✓		56.7
		✓		✓	56.3
		✓	✓	✓	**56.9**

When the CFGM was introduced on the DIOR dataset, the mAP of the student networks Faster RCNN-Res50 and Faster RCNN-Res18 improved from the original scores of 70.6% and 68.4% to 72.3% and 70.6%, respectively. As for DOTA dataset, the mAP of Faster RCNN-Res50 and Faster RCNN-Res18 were improved from the original 56.6% and 48.4% to 57.0% and 56.7%, respectively. The experiment results show that the student detectors obtained better features by learning the context of the teacher detectors, thus improving detection performance for the remote sensing objects.

Compared to the baseline network, the introduction of the ADMD strategy resulted in 1.9% and 2.3% higher mAP performance compared to the baseline on the DIOR dataset for Faster RCNN-Res50 and Faster RCNN-Res18, respectively. On the DOTA dataset for Faster RCNN-Res50 and Faster RCNN-Res18, the mAP was 0.6% and 7.9% higher than baseline. This demonstrates that our ADMD strategy guides the student to integrate the detection superiority of multiple teachers in order to learn better features, thereby achieving excellent performance in object detection.

Each component (CFGM and ADMD) provided additional significant gains for the different student detectors on various datasets and backbones. The joint consideration of CDFD for CFGM and ADMD provided the best performance for both Faster RCNN-Res50 and Faster RCNN-Res18 on the DIOR (73.1%, 58.0%) and DOTA (70.9%, 56.9%) datasets. From Table 2, it can be seen that the two modules proposed in this paper—CFGM and ADMD—are both effective in enhancing the performance of the student detector, and the combination of both modules further improves detection performance.

By comparing the performance of CFGM and CDFD on students, we noticed that CDFD further improved student performance with Faster RCNN-ResNet50, while the performance difference was negligible with Faster RCNN-ResNet18. This is due to the fact that compared to ResNet18, the student backbone ResNet50 is more similar to the teacher backbones ResNext32 and ResNext64, and it is easier to learn useful knowledge from similar teachers.

4.4.3. Comparison with State-of-the-Art Distillation Approaches on RS Datasets

Four sets of comparison experiments were conducted with ResNet50 and ResNet18 on the DIOR and DOTA datasets to compare our CDFD with several state-of-the-art distillation approaches. The framework for both students and teachers was Faster RCNN, and the backbone of the students is ResNet50. For distillation approaches FGD [48], MGD [54], and our approach without the multi-teacher strategy, the backbone of teachers was ResNet101, while for our approach with multi-teacher strategy, three teachers were employed with backbones ResNet101, ResNext32, and ResNext64.

It can be seen that our approach has state-of-the-art performance on all datasets and backbone networks, with all students achieving significant accuracy improvements with our CDFD, as shown in Table 3. With Faster RCNN-ResNet50 without the dense multi-teacher training strategy, our approach performed comparably to the high-performance distillation approach MGD; when the dense multi-teacher training strategy was introduced, it outperformed MGD on both the DIOR and DOTA datasets. With Faster RCNN-ResNet50 without the dense multi-teacher training strategy, our approach was slightly weaker than MGD, while with the dense multi-teacher training strategy, the performance was comparable to MGD on both datasets.

Table 3. Comparison with state-of-the-art distillation approaches. The best results are marked by bold text. † means without multi-teacher strategy.

Dataset	Student	Approach	mAP
DIOR	Faster RCNN-Res50	No-KD	70.6
		+FGD [48]	71.9
		+MGD [54]	72.2
		+Ours †	72.3
		+Ours	**73.1**
		Our gain	**+2.5**
	Faster RCNN-Res18	No-KD	68.4
		+FGD [48]	68.8
		+MGD [54]	70.9
		+Ours †	70.6
		+Ours	**70.9**
		Our gain	**+2.5**
DOTA	Faster RCNN-Res50	No-KD	56.6
		+FGD [48]	53.6
		+MGD [54]	57.0
		+Ours †	57.0
		+Ours	**58.0**
		Our gain	**+1.4**
	Faster RCNN-Res18	No-KD	48.4
		+FGD [48]	51.6
		+MGD [54]	56.9
		+Ours †	56.7
		+Ours	**56.9**
		Our gain	**+8.5**

4.4.4. Distillation of State-of-the-Art Detectors

In order to verify the generality of CDFD, we conducted experiments on more detectors with stronger students and teachers. We tested two state-of-the-art remote sensing detectors—AFPN [29] and FFPF [2]—which are two representative object detection networks. AFPN uses an aware feature pyramid network, while FFPF uses a frequency domain-aware backbone network and a bilateral frequency domain-aware feature pyramid network.

The experiments were conducted on the DIOR dataset. For AFPN, the student network was AFPN-ResNet50, and the teacher networks were Faster RCNN-ResNext32, Faster RCNN-ResNext64, and AFPN-ResNet101. For FFPF, the student network was FFPF-ResNet50, and the teacher networks were Faster RCNN-ResNext32, Faster RCNN-ResNext64, and FFPF-ResNet101.

Table 4 shows the experimental results. Without any additional computational effort, the mAP of FFPF increased from 72.8% to 73.6% and the mAP of AFPN increased from 71.8% to 72.6% with CDFD. The experiment results show that student detectors obtain better feature from stronger teacher detectors, demonstrating that CDFD has good generalisation and is applicable to various SOTA remote sensing detectors.

Table 4. Results of more detectors with stronger students and teacher detectors on the DIOR dataset. The best results are marked by bold text.

Student	Model	mAP
FFPF [2]	No-KD	72.8
	+ours	**73.6**
AFPN [29]	No-KD	71.8
	+ours	**72.6**

4.4.5. Comparison with State-of-the-Art Distillation Approaches on General Object Datasets

To verify that CDFD is also effective for general object detection tasks, we compared CDFD to several of the state-of-the-art detection distillation approaches on the COCO dataset. Faster RCNN frameworks were used for both students and teachers, with ResNet50 for the student backbone and ResNet101 as the teacher backbone of FGFI [55], GID [56], and FGD [48], while ResNet101, ResNext32, ResNext64 were used for the teacher backbone of CDFD.

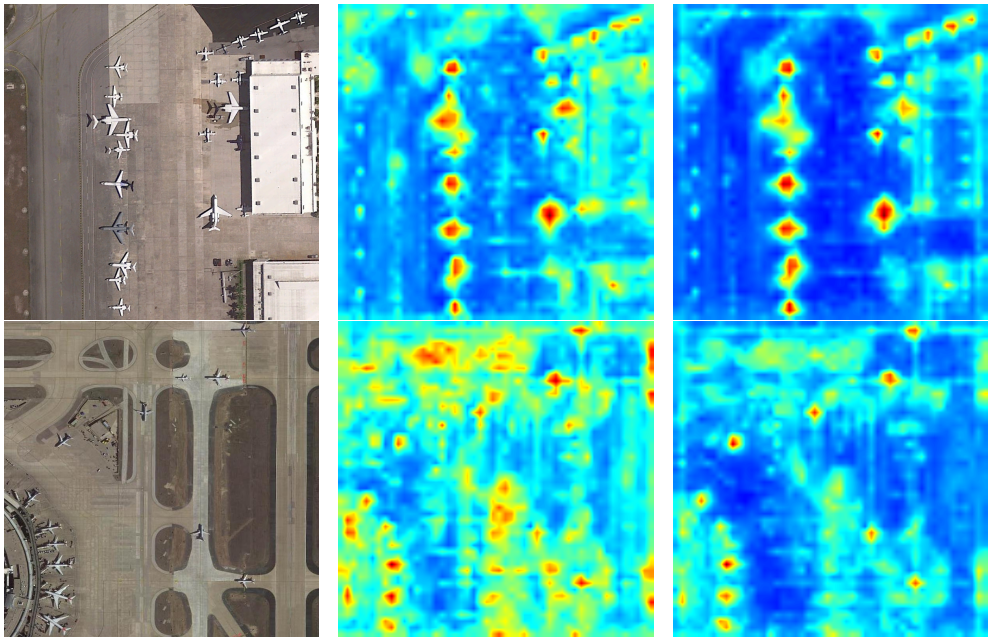
Table 5 shows that our approach significantly outperformed previous SOTA approaches, and students obtained significant performance gains from the three teachers using the proposed CDFD. The student model improved from 38.4% to 40.9% on mAP, completely eliminating the performance decrease due to the lightweight backbone. Thus, the experimental results confirm that the proposed CDFD approach is equally applicable to common object detection tasks.

Table 5. Comparison with state-of-the-art distillation approaches on COCO dataset.

Model	mAP	AP_S	AP_M	AP_L
Faster RCNN-ResNext64 (T1)	42.1	24.8	46.2	55.3
Faster RCNN-ResNext32 (T2)	41.2	24.0	45.5	53.5
Faster RCNN-Res101 (T3)	39.8	22.5	43.6	52.8
Faster RCNN-Res50 (S)	38.4	21.5	42.1	50.3
+FGFI [55]	39.3	22.5	42.3	52.2
+GID [56]	40.2	22.7	44.0	53.2
+FGD [48]	40.5	22.6	44.7	53.2
+Ours	40.9	23.4	44.8	53.7
Our gain	**+2.5**	**+1.9**	**+2.7**	**+3.4**

4.5. Visualisation Results

To more intuitively understand the role of our proposed CDFD, we visualised detection heat maps of the original model and of the distilled student model on the DIOR dataset; the regions with warm colours are the regions of greater interest to the model. The framework for both the original and student models was the Faster RCNN-ResNet50 detector. As shown in Figure 6, the student models trained with CDFD were able to focus more precisely on the regions where the objects are located, avoiding interference from the background and therefore achieving better detection performance.



(**a**) (**b**) (**c**)

Figure 6. Examples of heatmap visualisations on the DIOR dataset. (**a**) Input image, (**b**) heat map of original model, (**c**) heat map of student model.

5. Conclusions

To create lightweight and superior detectors for onboard remote sensing object detection, we propose the novel Context-aware Dense Feature Distillation (CDFD), guiding a lightweight student detector to fully learn the superiority of multiple large teacher detectors. Our approach includes novel components: a Contextual Feature Generation Module (CFGM) for learning teacher context, and an Adaptive Dense Multi-teacher Distillation Strategy (ADMD) for learning multi-teacher feature maps, allowing lightweight detectors to obtain better features without additional computation, resulting in considerable performance gains. Extensive experiments on different datasets and network structures demonstrate that the proposed CDFD effectively improves the performance of lightweight detectors with good generalisation. Furthermore, experiments on an extensive general object dataset demonstrate that our CDFD is equally effective for general object detection distillation. However, there are limitations to the application of CDFD, i.e., it is not suitable for detectors without a feature pyramid network. Therefore, in further research, it is necessary to investigate distillation methods that are more general and suitable for all detectors.

Author Contributions: All of the authors made significant contributions to the work. L.G. provided the original idea for the study and conducted experiments. L.G. and Q.F. proposed the framework and wrote the manuscript. Z.W., E.P. and G.D. provided suggestions and reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The study was realised with the support of the China Scholarship Council (Grant No. 202106210056) and was supported in part by the program: "Best International Grant for PhD" of Peter the Great, St. Petersburg Polytechnic University.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Maskey, A.; Cho, M. CubeSatNet: Ultralight Convolutional Neural Network designed for on-orbit binary image classification on a 1U CubeSat. *Eng. Appl. Artif. Intell.* **2020**, *96*, 103952. [CrossRef]
2. Lingyun, G.; Popov, E.; Ge, D. Fast Fourier Convolution Based Remote Sensor Image Object Detection for Earth Observation. *arXiv* **2022**, arXiv:2209.00551.
3. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
4. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef] [PubMed]
6. Selva, D.; Krejci, D. A survey and assessment of the capabilities of Cubesats for Earth observation. *Acta Astronaut.* **2012**, *74*, 50–68. [CrossRef]
7. Manning, J.; Langerman, D.; Ramesh, B.; Gretok, E.; Wilson, C.; George, A.; MacKinnon, J.; Crum, G. *Machine-Learning Space Applications on Smallsat Platforms with Tensorflow*; NASA Goddard Space Flight Center: Greenbelt, MD, USA, 2018.
8. Arechiga, A.P.; Michaels, A.J.; Black, J.T. Onboard image processing for small satellites. In Proceedings of the NAECON 2018-IEEE National Aerospace and Electronics Conference, Dayton, OH, USA, 11 July 2018; IEEE: Piscatway, NJ, USA, 2018; pp. 234–240.
9. Huq, R.; Islam, M.; Siddique, S. AI-OBC: Conceptual Design of a Deep Neural Network based Next Generation Onboard Computing Architecture for Satellite Systems. In Proceedings of the 1st China Microsatellite Symposium, Xi'an, China, 18–20 November 2018.
10. Giuffrida, G.; Diana, L.; de Gioia, F.; Benelli, G.; Meoni, G.; Donati, M.; Fanucci, L. Cloudscout: A deep neural network for on-board cloud detection on hyperspectral images. *Remote Sens.* **2020**, *12*, 2205. [CrossRef]
11. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
12. Wu, J.; Leng, C.; Wang, Y.; Hu, Q.; Cheng, J. Quantized convolutional neural networks for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4820–4828.
13. Kim, J.; Park, S.; Kwak, N. Paraphrasing complex network: Network compression via factor. transfer. In Proceedings of th Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, QC, Canada, 3–8 December 2018.
14. Mirzadeh, S.I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; Ghasemzadeh, H. Improved knowledge distillation via teacher assistant. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 5191–5198.
15. Lee, S.; Song, B.C. Graph-based knowledge distillation by multi-head attention network. *arXiv* **2019**, arXiv:1907.02226.
16. Tung, F.; Mori, G. Similarity-preserving knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1365–1374.
17. Cheng, G.; Si, Y.; Hong, H.; Yao, X.; Guo, L. Cross-scale feature fusion for object detection in optical remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* **2020**, *18*, 431–435. [CrossRef]
18. Saeed, N.; Elzanaty, A.; Almorad, H.; Dahrouj, H.; Al-Naffouri, T.Y.; Alouini, M.S. Cubesat communications: Recent advances and future challenges. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1839–1862. [CrossRef]
19. Cooley, S.W.; Smith, L.C.; Ryan, J.C.; Pitcher, L.H.; Pavelsky, T.M. Arctic-Boreal lake dynamics revealed using CubeSat imagery. *Geophys. Res. Lett.* **2019**, *46*, 2111–2120. [CrossRef]
20. Houborg, R.; McCabe, M.F. Daily Retrieval of NDVI and LAI at 3 m Resolution via the Fusion of CubeSat, Landsat, and MODIS Data. *Remote Sens.* **2018**, *10*, 890. [CrossRef]
21. Altena, B.; Kääb, A. Glacier ice loss monitored through the Planet cubesat constellation. In Proceedings of the 2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp), Bruges, Belgium, 27–29 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
22. Huang, L.; Luo, J.; Lin, Z.; Niu, F.; Liu, L. Using deep learning to map retrogressive thaw slumps in the Beiluhe region (Tibetan Plateau) from CubeSat images. *Remote Sens. Environ.* **2020**, *237*, 111534. [CrossRef]
23. Ghuffar, S. DEM generation from multi satellite PlanetScope imagery. *Remote Sens.* **2018**, *10*, 1462. [CrossRef]

24. Toorian, A.; Diaz, K.; Lee, S. The cubesat approach to space access. In Proceedings of the 2008 IEEE Aerospace Conference, Tampa, FL, USA, 8–11 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–14.
25. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
26. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
27. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
28. Zhang, K.; Shen, H. Multi-Stage Feature Enhancement Pyramid Network for Detecting Objects in Optical Remote Sensing Images. *Remote Sens.* **2022**, *14*, 579. [CrossRef]
29. Cheng, G.; He, M.; Hong, H.; Yao, X.; Qian, X.; Guo, L. Guiding clean features for object detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 801920. [CrossRef]
30. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**. [CrossRef] [PubMed]
31. Qingyun, F.; Zhaokui, W. Cross-Modality Attentive Feature Fusion for Object Detection in Multispectral Remote Sensing Imagery. *Pattern Recognit.* **2022**, *130*, 108786. [CrossRef]
32. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
33. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge distillation: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
34. Peng, Z.; Li, Z.; Zhang, J.; Li, Y.; Qi, G.J.; Tang, J. Few-shot image recognition with knowledge transfer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 441–449.
35. Wang, J.; Gou, L.; Zhang, W.; Yang, H.; Shen, H.W. Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 2168–2180. [CrossRef]
36. Dou, Q.; Liu, Q.; Heng, P.A.; Glocker, B. Unpaired multi-modal segmentation via knowledge distillation. *IEEE Trans. Med. Imaging* **2020**, *39*, 2415–2425. [CrossRef] [PubMed]
37. Hou, Y.; Ma, Z.; Liu, C.; Hui, T.W.; Loy, C.C. Inter-region affinity distillation for road marking segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12486–12495.
38. Wang, X.; Hu, J.F.; Lai, J.H.; Zhang, J.; Zheng, W.S. Progressive teacher-student learning for early action prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3556–3565.
39. Wu, M.C.; Chiu, C.T.; Wu, K.H. Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscatway, NJ, USA, 2019; pp. 2202–2206.
40. Cun, X.; Pun, C.M. Defocus blur detection via depth distillation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 747–763.
41. Chawla, A.; Yin, H.; Molchanov, P.; Alvarez, J. Data-free knowledge distillation for object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 3289–3298.
42. Zhao, H.; Sun, X.; Dong, J.; Chen, C.; Dong, Z. Highlight every step: Knowledge distillation via collaborative teaching. *IEEE Trans. Cybern.* **2020**. [CrossRef] [PubMed]
43. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
44. Qingyun, F.; Lin, Z.; Zhaokui, W. An efficient feature pyramid network for object detection in remote sensing imagery. *IEEE Access* **2020**, *8*, 93058–93068. [CrossRef]
45. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
46. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014, pp. 740–755.
47. Kang, Z.; Zhang, P.; Zhang, X.; Sun, J.; Zheng, N. Instance-conditional knowledge distillation for object detection. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16468–16480.
48. Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; Yuan, C. Focal and global knowledge distillation for detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4643–4652.
49. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.
50. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
51. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

53. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
54. Yang, Z.; Li, Z.; Shao, M.; Shi, D.; Yuan, Z.; Yuan, C. Masked Generative Distillation. *arXiv* **2022**, arXiv:2205.01529.
55. Wang, T.; Yuan, L.; Zhang, X.; Feng, J. Distilling object detectors with fine-grained feature imitation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4933–4942.
56. Dai, X.; Jiang, Z.; Wu, Z.; Bao, Y.; Wang, Z.; Liu, S.; Zhou, E. General instance distillation for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20 June 2021; pp. 7842–7851.