A Class-Incremental Detection Method of Remote Sensing Images Based on Selective Distillation

Hang Ruan, Jian Peng, Ye Chen, Silu He, Zhenshi Zhang and Haifeng Li

1 Beijing Institute of Tracking and Communication Technology, Beijing 100094, China
2 School of Geosciences and Info-Physics, Central South University, Changsha 410083, China
3 Undergraduate School, National University of Defense Technology, Changsha 410080, China
* Correspondence: pengj2017@csu.edu.cn

Abstract: With the rapid development of remote sensing technology and the growing demand for applications, the classical deep learning-based object detection model is bottlenecked in processing incremental data, especially in the increasing classes of detected objects. It requires models to sequentially learn new classes of objects based on the current model, while preserving old categories-related knowledge. Existing class-incremental detection methods achieve this goal mainly by constraining the optimization trajectory in the feature of output space. However, these works neglect the case where the previously learned background is a new category to learn, resulting in performance degradation in the new category because of the conflict between remaining the background-related knowledge or updating the background-related knowledge. This paper proposes a novel class-incremental detection method incorporated with the teacher-student architecture and the selective distillation (SDCID) strategy. Specifically, it is the asymmetry architecture, i.e., the teacher network temporarily stores historical knowledge of previously learned objects, and the student network integrates historical knowledge from the teacher network with the newly learned object-related knowledge, respectively. This asymmetry architecture reveals the significance of the distinct representation of history knowledge and new knowledge in incremental detection. Furthermore, SDCID selectively masks the shared subobject of new images to learn and previously learned background, while learning new categories of images and then transfers the classification results of the student model to the background class following the judgment model of the teacher model. This manner avoids interferences between the background category-related knowledge from a teacher model and the learning of other new classes of objects. In addition, we proposed a new incremental learning evaluation metric, C-SP, to comprehensively evaluate the incremental learning stability and plasticity performance. We verified the proposed method on two object detection datasets of remote sensing images, i.e., DIOR and DOTA. The experience results in accuracy and C-SP suggest that the proposed method surpasses existing class-incremental detection methods. We further analyzed the influence of the mask component in our method and the hyper-parameters sensitive to our method.

Keywords: object detection; class-incremental learning; selective distillation; remote-sensing image; deep learning

1. Introduction

The object detection task of the remote sensing image, one of the important tasks of remote sensing image interpretation, requires the detection of various targets describing surface objects on remote sensing images for the purpose of precise target location and category determination, and plays a critical role in the application domain such as object tracking and monitoring, activity recognition and scene analysis, urban planning, and military research [1,2]. In particular, with the development of deep learning techniques, the research on object detection driven by deep learning techniques has made great progress and has facilitated the development of different application fields.
Existing deep learning-based object detection is based on a typical learning paradigm where the model is trained on a pre-set fixed class and data. This setting is challenged in real-world applications because the classes and corresponding data volumes increase as the demand changes and increases acquired data. It is highly inefficient or even unrealistic to retrain the new data along with the historical data at each new inflow of data. As opposed to this joint training learning paradigm, it is significant to enable the model to train on new classes and data in an orderly incremental manner, i.e., incremental learning. It requires the model to satisfy three points: (1) Using no or as little historical data as possible, (2) achieving plasticity: the original detection model is able to continue learning new data and gain the ability to detect new classes, and (3) achieving stability: the original detection model should retain the learning performance on the historical data as much as possible and be able to detect the object of historical data, i.e., the well-known catastrophic forgetting issue [3–5]. In particular, the latter two are known as plasticity-stability balance problems.

To alleviate the plastic-stability problem, the typical manner uses a knowledge distillation strategy to retain the original detection capability for the detection of historical classes, while fine-tuning the original model [6,7]. This research achieves some success in contributing to the development of the computer vision research field, yet ignoring significant facts that remote sensing images capture a wide range of objects that exist on the Earth’s surface without human intervention. Specifically, since the background classes are richer and more diverse than those of natural images, the backgrounds of remote sensing images of the previous learning classes are likely to contain new objects that need to be learned in the future. For example, the original remote sensing image object detection model only needs to detect aircraft, and later needs to learn detecting vehicles in order to better analyze the airport vehicle scheduling situation. For the image data itself, the new class of vehicles, which will be concerned in the future, already exists in the aircraft class dataset and does not possess a label signal, so it will be learned as a background class during the learning process, as shown in Figure 1.

![Figure 1. Schematic of the incremental learning results for the class-incremental detection task of remote sensing.](image)

The existing approaches based on knowledge distillation fail to handle this situation. To be specific, the original learning experience of identifying unlabeled new classes as the background is also retained when implementing the knowledge distillation strategy, and this in turn hinders the effective learning of new class objects by the model. Figure 2 shows the changes of stability and plasticity of multiple class-incremental detection methods under the same object detection model on the DOTA [8] dataset as the training iterations of incremental learning increases. This metric is calculated in Section 4.3 and the specific experimental setup is shown in Section 4.4. As can be seen from the figure, the class-
incremental detection method ILOD [6] based on knowledge distillation has a smaller gap in stability performance and joint fine-tuning and fine-tuning compared to Faster ILOD [9], while the plasticity performance is poorer.

![Stability and plasticity performance of existing incremental detection methods on the DOTA dataset.](image)

**Figure 2.** Stability and plasticity performance of existing incremental detection methods on the DOTA dataset.

In this paper, we propose a remote sensing image class-incremental detection method based on selective distillation of knowledge (SDCID). It is implemented on the teacher-student asymmetry architecture, in which the teacher network stores historical objects related to knowledge, and the student network learns new objects and intergrades the teacher network’s knowledge. To avoid the conflict between plasticity and stability on common goals, we consider it essential to selectively retain the information related to the history classes in the original detection model for these shared objects while forgetting the learning experience about the new courses in the original detection model. Therefore, our approach configures a selective distillation strategy based on a dual network framework for knowledge distillation. In contrast to the general indiscriminate retention of historical classes and background information, the SDCID learns new knowledge from new images and distills out the knowledge irrelevant to the objects of new types while preserving previously learned knowledge corresponding to historical objects, avoiding different directions of guidance between new and historical knowledge for model parameter updates. Specifically, a selective mask is added to the objects of new classes to avoid the teacher model giving wrong guidance. A selective distillation loss function is designed to effectively align the teacher model with the student model in identifying the object of a new class to help the student model complete the learning of the objects of new classes independently. The selective distillation strategy and the fine-tuning learning strategy jointly complete the incremental learning of remote sensing image data of the new class so that the object detection model can detect remote sensing objects of the old and new classes simultaneously.

The main contributions of this paper are summarized as follows:

1. We propose a method of class-incremental detection of remote sensing images based on selective distillation, which can avoid knowledge conflicts on objects of new classes while retaining the ability to detect objects of historical classes, retaining important knowledge that is gainful for incremental learning and accomplishing effective integrated learning on old and new knowledge.

2. We propose a novel evaluation metric C-SP for incremental detection methods, which avoids the influence on the incremental learning result from the learning difficulty of the old and new classes themselves, directly evaluates the model stability and
3. The performance comparison experiments, visualization analysis, method validity analysis, and hyperparameter analysis experiments are conducted on the widely used remote sensing image object detection dataset, and the experimental results verify the superiority and validity of the method.

2. Related works

2.1. Object Detection for Remote Sensing Image

Remote sensing image object detection refers to the detection of the class and coordinate position of the object from the image obtained by remote sensing technology using certain algorithms. The methods of remote sensing image object detection are mainly divided into feature representation-based detection methods and deep learning-based detection methods according to the way to extract features. The feature representation-based detection methods have the problems of large computational effort and low localization accuracy, where the extracted regions largely overlap in the region selection process; in the feature extraction process, there are works using SIFT features for classification of ship objects [10], and works [11–13] using HOG features for ship, building, and vehicle detection, respectively. Other works [14] used BING features for aircraft detection and LBP features for ship detection [15], but the design of visual features relies on human a priori knowledge to manually design the underlying visual features for different classes of detection objects, which cannot contain multi-level visual semantic information and has poor robustness and a limited range of applicability. Traditional detection methods based on feature representation have low accuracy in classification and position localization, which cannot meet the application requirements well.

Due to the powerful performance of convolutional neural network (CNN) structures in deep learning for multi-layer visual feature extraction, a series of deep convolution-based object detection methods have emerged in the field of computer vision to extract multi-scale features of images, which outperform the detection methods of traditional feature representation. Deep convolution-based object detection methods use CNN-based backbone networks to extract features for images and generate feature maps for subsequent networks. The backbone networks are mostly derived from the networks used for image classification, from AlexNet [16], VGG [17], and GoogLeNet [18] to ResNet [19], with increasing network depth, width, and complexity to improve the performance of detection networks.

The depth convolution-based object detection methods can be further classified into one-stage detection methods and two-stage detection methods according to whether regions of interest (ROI) are extracted or not. The two-stage detection method first searches for possible foreground regions of interest and roughly locates the coordinates through a dedicated region proposal generation module, and then further classifies and adjusts the position of bounding box based on the previously generated region proposal. Common two-stage detection methods include R-CNN [20], Fast R-CNN [20], and Faster R-CNN [21]. One-stage detection methods directly use CNN networks for feature extraction of the full graph, and transform the problem of object boxes localization into a regression problem to deal with while completing the task of classifying and localizing the object boxes. Common one-stage detection methods include YOLO [22], SSD [23], etc.

The one-stage detection method, which uses only one network to uniformly achieve object classification and bounding box localization, is more computationally efficient and better meets real-time requirements than the two-stage method, but the one-stage detection method is still weak in terms of accuracy performance while the two-stage detection method can achieve higher detection accuracy in comparison. Among the two-stage detection methods, the Faster R-CNN method is a classical algorithm that is an improvement on the R-CNN method and Fast R-CNN method. Faster R-CNN is divided into four main components: (1) CNN backbone networks. (2) Region Proposal Networks (RPN), where
RPN replaces selective search to generate higher quality region proposal boxes. (3) ROI pooling layer. This layer collects the input feature maps and region proposal boxes, and extracts the uniform scale feature maps at the object level after combining this information. (4) Classifier and regressor. The classifier uses the objective-level feature map to calculate the class of the proposal box, and the regressor is used to determine the final precise location. The Faster R-CNN proposes an RPN network that subsequently connects the ROI pooling layer for further classification and localization and achieves a true end-to-end.

The problems to be solved for remote sensing image object detection tasks include multiple aspects. Improved models are incorporated to improve detection according to specific small objects, complex backgrounds, multiple scales, and shape differences. For example, the SCRDet++ method [24], for the first time, considers the instance-level feature map denoising to effectively separate small objects from the background and improve the detection of small objects. The feature map after instance-level denoising is feature decoupled from the original features, so that the object features in the spatial domain are enhanced and the background features in the spatial domain are weakened. Mask OBB method [25] designs a semantic attention network to provide semantic feature maps with background foreground discriminative power in a pixel-level classification way. The MFIAR-Net method [26] designed a multi-scale feature integration approach to generate salient features of appropriate size, achieving an effective balance between low-level localization features and high-level semantic features. The FADet method [27] introduced an adaptive multi-sensory field attention mechanism to generate high-quality region proposal boxes.

Remote sensing image object detection methods are developing rapidly, and many researchers have designed detection models with increasing accuracy according to the application scenarios and detection requirements. However, existing methods need to provide all of the training data to the detection models at once, and do not yet have the ability to continuously accept data stream and perform incremental learning.

2.2. Incremental Learning Based on Knowledge Distillation

Incremental learning is a learning paradigm where a learning system continues to learn new knowledge based on old knowledge. It requires progressive learning that emulates human learning behavior to effectively respond to a data stream. Incremental learning originated from cognitive neuroscience research in memory and forgetting [3] and needs to ensure that new knowledge is extracted by incremental learning of new samples in the presence of existing knowledge, fusing old and new knowledge to improve the accuracy of model prediction results.

Incremental learning has three requirements: (1) Using as little historical data as possible, which comes from limitations in data access, storage space, training resources, etc. (2) Maintaining model stability: the phenomenon of catastrophic forgetting of most of the previously learned old knowledge when neural networks learn only new knowledge. Catastrophic forgetting is a common problem present in different types of neural networks, from standard backpropagation neural networks to unsupervised neural networks. How to maintain model stability and overcome catastrophic forgetting as much as possible during incremental learning is an important issue of general concern. (3) Maintaining model stability: Incremental learning models continue to learn valid information about new data and have good understanding of new data.

Incremental learning methods based on knowledge distillation are about learning new tasks without using old data. Knowledge distillation [28] is a method designed for model compression tasks and has a bi-model structure of “teacher-student”. The teacher model, which has strong generalization ability in big data, is used as the store of features, and the decision result of the data can be used as the soft label to guide the student model to output consistent decision results, thus realizing the transfer of knowledge from teacher model to student model.
Existing incremental learning methods based on knowledge distillation have been firstly investigated on the basic vision analysis task of image classification. A representative approach in the image classification task is the Learning without Forgetting (LwF) approach [29]. The LwF approach introduces the idea of knowledge distillation [28], allowing the model to learn new knowledge in a fine-tuned manner while using the knowledge distillation loss as an additional regularization term, expecting the decision boundaries about the old knowledge to remain as constant as possible. The Learning without Memory (LwM) method [30], in addition to using the knowledge distillation loss, uses the attention distillation loss function to ensure as much as possible the feature stability of local regions with high relevance to the decision output during the convolution process. The PodNet approach [31] proposes a pooling distillation-based constraint method for convolutional networks starting from knowledge distillation, applying distillation on a spatial scale to the process of feature extraction, providing a representation for each class containing multiple agent vectors, and striking a balance between reducing forgetting old classes and learning new ones, which facilitates long-term incremental tasks.

Object detection models are more complex than image classification models and require semantic classification and bounding box localization for objects in images. Therefore, research on incremental object detection is more limited. To address the problem of catastrophic forgetting on historical classes, the ILOD approach [6] proposes an incremental object detector based on the idea of knowledge distillation without using historical data, but this approach uses a Fast R-CNN model, which is less advanced. In this method, the teacher model selects 128 region proposal boxes with the lowest probability of background class from all given region proposal boxes based on the input of new data, and then randomly selects 64 region proposal boxes, and outputs the class determinations and coordinate locations for these region proposal boxes as soft labels. These region proposal boxes are also provided to the student model to output the prediction results (classification probabilities and coordinate locations). The region proposal box determination result distillation loss function of the ILOD method calculates the gap between the soft labels and the prediction results. Later, a work proposed a Faster ILOD method [9], which makes improvements to ILOD on the structure of Faster R-CNN, and the distillation loss function contains three parts: feature distillation loss function, RPN distillation loss function, and region proposal box determination result distillation loss function. On the incremental detection task of remote sensing images, the FPN-IL [32] method implements incremental learning of classes under multi-feature fusion on the structure of Faster R-CNN and FPN.

According to the application objects of distillation strategies, distillation strategies can be classified into the following three types: (1) Full graph feature distillation strategy: the calculation object of distillation loss is the feature map of the backbone network. (2) RPN distillation strategy: the calculation object of distillation loss is the binary classification results and regression results of foreground and background output by the RPN network. (3) The region proposal box determination result distillation strategy: the calculation object of distillation loss is the determination result of the region proposal box. The ILOD method uses a one-fold distillation strategy, while the Faster ILOD method uses a three-fold distillation strategy. As can be seen from Figure 2, the three-fold distillation strategy is more stable than the one-fold distillation strategy, but it has more constraints on the model parameters, which is not conducive to the model to continue learning new knowledge for plasticity. The ILOD method does not have the two constraints of another two distillations and performs slightly better than the Faster ILOD method in terms of plasticity. Therefore, the incremental learning is performed by using the regional proposal box determination result distillation strategy to maintain model stability while facilitating model acquisition plasticity. For the region proposal box determination result distillation strategy, the two methods are more consistent in their calculations, but by distilling the entire knowledge without selection, the student model accepts the knowledge of categorizing the new classes as background, which eventually leads to the inability to take good care of stability and plasticity.
It should be noted that compared with the above methods, our proposed selective knowledge distillation strategy is based on the input image, filtering the local targets of the input image before retaining the relevant knowledge, which is not considered by previous methods. Furthermore, this strategy is based on the bottom and thus is universal for all three types of methods mentioned above.

3. Methods

3.1. Preliminary Knowledge

Dual network learning with knowledge distillation is based on the basic assumption that through knowledge distillation, the student network can use less network capacity to maintain knowledge of previous learning objects on the teacher network while adapting and learning new classes of objects. Based on this assumption, the idea of incremental learning for CNN models used for object detection is to generate a temporary copy of the model to preserve historic knowledge before learning new objects. Then, it freezes its parameters and constrains the learning process of the original model to learn new objects through distillation losses. It consists of two prominent roles: (1) selecting the proposal box corresponding to the old class, i.e., the proposal distillation candidate box and (2) calculating the two networks’ distillation losses. As shown in Figure 3, the student network is the original model used to learn to adapt to the new object class; in contrast, the teacher network is the model copy used to guide the student network to maintain historical knowledge.

Figure 3. Architecture of the SDCID method.

To prevent catastrophic forgetting and ensure that new object classes are learned while retaining the recognition of previous classes, the parameter updates are constrained here by introducing a knowledge distillation loss on the standard cross-entropy loss. This is achieved by generating a proposal box for new training samples from the model copy and defining the knowledge to be distilled based on the candidate frames, such as the activation or output distribution of each proposal box, and then minimizing the discrepancy between the two networks by distillation losses, such as the L2 distance loss function.

3.2. Selective Distillation Class Incremental Detection Framework

We propose a class incremental detection method using selective distillation based on a dual network learning model. It utilizes knowledge distillation to maintain the stability of the model based on the fine-tuning of the model to learn new knowledge. As shown in Figure 3, it consists of two sub-networks, i.e., the teacher network and the student network. The teacher and student models share the same initialization parameters, and they are both derived from the original detection model. During training, the parameters of the teacher model are frozen, while the student model is updated according to the overall loss function. The overall loss function consists of two components. One is from the fine-tuning strategy, which matches the labels of the new class data to the output of the student model...
to compute the original Faster R-CNN loss. The other is from the distillation strategy, which matches the output of the teacher model with the output of the student model to determine the results of the region suggestion frame. Compared with existing dual-network incremental detection frameworks, the main difference is the use of a selective distillation strategy, which is able to avoid the situation of sharing historical data background with new data learning objects and avoid plasticity-stability conflicts in incremental target detection by filtering shared objects in historical data background and selectively distilling local features of samples.

3.3. Selective Distillation Strategy

The selective distillation strategy consists of three components: selection of input image distillation, selection of region proposal box, and calculation of distillation loss.

3.3.1. Selection of Input Image Distillation

First, in order to avoid distillation of knowledge related to the objects of the new class, local masks are selectively added to the objects of the new class in the images to destroy their complete semantic information and avoid the teacher model to respond to the objects of the new class.

3.3.2. Selection of Region Proposal Box

The images with the added selective mask are input to the teacher model and student model, and the teacher model generates the region proposal boxes following the RPN structure. In order to prevent the object of the new classes from being selected, the top $K$ (with default value of 64) region proposal boxes with the highest foreground probability are selected to enter the region proposal box distillation pool, and these region proposal boxes are more likely to be determined as historical classes rather than new classes by the teacher model with historical knowledge.

3.3.3. Calculation of Distillation Loss

The selected region proposal boxes are fed to the ROI pooling layer of the teacher model and the student model to output the final determination results. The classification results and regression results are included in the determination results. The teacher model classifies each region proposal box with the results containing the probabilities on the background and historical classes, while the classification output of the student model contains the probabilities on the background classes, historic classes, and new classes. To align the classification results of the teacher and student models, the classification output of the student model was reconstructed and calculated as shown in Figure 4. The classification results of the student model on the new classes were integrated into the background class to fit the teacher model’s original decision pattern of only background and history classes, and to allow the student model to learn the correct class of the object on the decision pattern of background, history, and new classes. The reconstructed classification results of the student model for the regional candidate boxes are calculated as shown below.

$$\hat{p}_{st}(C_i) = \begin{cases} p_{st}(C_i), & \text{if } 0 < C_i \leq C_m \\ p_{st}(C_0) + \sum_{C_{m+1}}^{C_n} p_{st}(C_i), & \text{otherwise} \end{cases}$$  

(1)

where $C_i$ is the object of the $i$-th class; $m$ is the total number of historical classes; $n$ is the total number of new classes.
Figure 4. The schematic of calculating reconstructed classification results.

For non-added classes, in order to align the classification results of the teacher model and the reconstructed classification results of the student model, the cross-entropy loss of the two was calculated. The selective distillation loss, which was calculated together with the difference between the determination results of the teacher model and the student model, was calculated as follows:

\[
\text{Loss}_{\text{Sel Dist}} = \frac{1}{K \ast (m + 1)} \sum \left( -p_{teC_0 \sim C_m} \ast \log (p_{stC_0 \sim C_m}) + \left\| l_{teC_1 \sim C_m} - l_{stC_1 \sim C_m} \right\|_{2}^{2} \right)
\]

where \( K \) is the number of region candidate boxes and the default value is 64; \( m \) is the total number of historical classes; \( te \) is the teacher model; \( st \) is the student model; \( C_0 \) \( C_m \) are the background and historical classes; \( p_{teC_1 \sim C_m} \) is the \( m \) class classification result of the teacher model for the region candidate boxes; \( p_{stC_0 \sim C_m} \) is the \( m \) classes reconstruction classification result of the student model for the region candidate boxes; \( l_{teC_1 \sim C_m} \) and \( l_{stC_1 \sim C_m} \) are the regression results of the teacher model and the student model on the regional candidate boxes in \( m \) historical classes.

The SDCID method uses a multi-task learning approach with an overall loss function calculated as follows:

\[
\text{Loss}_{\text{total}} = \text{Loss}_{\text{Faster R-CNN}} + \lambda \ast \text{Loss}_{\text{Sel Dist}}
\]

where \( \text{Loss}_{\text{Faster R-CNN}} \) is the original Faster R-CNN loss, \( \lambda \) is the distillation hyperparameter that balances the original Faster R-CNN loss and the selective distillation loss, and the default is 1.

4. Experiment Settings
4.1. Baselines

1. Joint fine-tuning: learning all data, i.e., data of new class and historical class, to update the parameters of the original detection model.
2. Fine-tuning: Learning only the currently added data, i.e., data of new class, to update the parameters of the original detection model.
3. ILOD method: Class-incremental detection method based on knowledge distillation, which calculates distillation loss on the region candidate box determination results.
4. Faster ILOD method: Class-incremental detection method based on knowledge distillation, which calculates distillation loss on the full graph features, RPN output, and regional candidate box determination results.

4.2. Datasets

Experiments were conducted on the DOTA and DIOR datasets, and the datasets were divided by the class increment. The DOTA dataset was divided into 8 historical classes...
and 7 new classes, and the results are shown in Figure 5. All classes in the DIOR dataset were randomly divided into 10 historical classes and 10 new classes, and the results are shown in Figure 6. The historical category data, which does not have labels for the new class objects but actually contains the new class objects. When assessing the accuracy of the model before incremental learning, the accuracy of the historical classes was evaluated on the test data, and when assessing the accuracy of the model after incremental learning, the accuracy of all classes was evaluated on the test data.

Figure 5. Partition of DOTA dataset.

Figure 6. Partition of DIOR dataset.

4.3. Evaluation

4.3.1. The Detection Accuracy of Model after Incremental Learning

After the detection model has completed incremental learning, the performance of the class-incremental detection method can be evaluated using the AP50 accuracy metric of the trained model. Specifically, the AP50 accuracy on historical classes, new classes, and all classes can be calculated to measure the detection capability of the detection model on various classes.

The AP50 accuracy is derived from the evaluation metrics of object detection methods and is used to assess the detection accuracy of the model. Since the prediction result of object detection is a series of prediction boxes on the image, it needs to be compared with the real object to calculate the accuracy of the prediction result. When the prediction box and the true box do not exactly overlap, the Intersection over Union (IoU) between the prediction box and the true box needs to be calculated first to determine whether the prediction box and the true box can correspond. IoU, which calculates the ratio of intersection and concurrence between the prediction box and the true box, measures the degree of overlap between the two boxes. If the intersection ratio between the prediction box and the true box is greater than the predetermined IoU threshold, the prediction of
the position of the prediction box is determined to be accurate, while the opposite result is determined to be incorrect.

Precision calculates the percentage of correct results among all predicted results. Recall calculates the percentage of predicted objects among all objects present on the image. The formulas are as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (4)
\]
\[
\text{Recall} = \frac{TP}{TP + FN} \quad (5)
\]

where \(TP\) is the correct prediction result; \(FP\) is the incorrect prediction result; \(FN\) is the true object that was not predicted.

Average precision (AP), which combines a measure of precision and recall, is one of the most commonly used metrics for evaluating the accuracy of object detection methods. The prediction results are first ranked according to the confidence level, and multiple iterations of precision and recall are calculated to form the precision–recall curve. The result is better as it is closer to the upper right on the precision–recall curve. Therefore, the average precision is calculated as the area below the precision–recall curve and is the mean value of precision at different recall rates. \(AP^{50}\) is the area below the precision–recall curve at an IoU threshold of 0.5.

### 4.3.2. Stability and Plasticity of Model after Incremental Learning

Both the number and order of sessions of incremental learning affect the algorithm’s performance, while the accuracy metric is hard to objectively reflect the ability of the model to balance plasticity–stability before and after incremental learning. Considering the lack of relevant metrics in the field of incremental target detection, here, we propose a new metric to evaluate the model’s performance in plasticity–stability comprehensively. Specifically, we take the detection precision of the model that learns all data simultaneously as the upper limit. Furthermore, we compare it with the detection precision of the model after incremental learning, and use it to calculate the stability and plasticity of the model, corresponding to the stability and plasticity performance matrix, respectively.

The stability performance (\(S\)) metric calculates the ratio of two models—the detection model after incremental learning and the detection model that learns all data at once—in terms of historical class detection accuracy. The \(S\) metric is calculated as follows:

\[
S = \frac{AP^{50}_{\text{old class}}(\text{Detector}_{\text{IL}})}{AP^{50}_{\text{old class}}(\text{Detector}_{\text{LA}})}
\]

where \(\text{Detector}_{\text{IL}}\) is the detection model after incremental learning; \(\text{Detector}_{\text{LA}}\) is the detection model that learns all data at once; \(AP^{50}_{\text{old class}}\) is the detection accuracy in the historical class.

The plasticity performance (\(P\)) metric calculates the ratio of two models—the detection model after incremental learning and the detection model that learns all the data at once—in terms of the detection accuracy of the new classes. The formula for the \(P\) metric is as follows:

\[
P = \frac{AP^{50}_{\text{new class}}(\text{Detector}_{\text{IL}})}{AP^{50}_{\text{new class}}(\text{Detector}_{\text{LA}})}
\]

where \(AP^{50}_{\text{new class}}\) is the detection accuracy in the new class.

The combined stability–plasticity performance (C-SP) indicator calculates a harmonic mean of stability performance and plasticity performance. The harmonic mean is a statistical indicator for calculating the mean, which is more influenced by extreme values under the imbalance between the two, and is more influenced by the very small values than by the very large ones, focusing on the one that responds to the poor performance of the averaged
object, and is able to harshly combine the measures of stability and plasticity performance. The formula for the C-SP metric is as follows:

\[
C-SP = \frac{C_O + C_N}{S/C_O + P/C_N}
\]

where \(C_O\) is the number of historical classes; \(C_N\) is the number of new classes.

4.4. Training Settings

The base object detection model for class-incremental detection is Faster R-CNN, which is a model with good accuracy results that is widely used in remote sensing datasets, and is applied in four baseline methods and SDCID methods. The backbone network is Resnet-50 with initialization parameters from the pre-trained model on the ImageNet dataset, and the implementation is based on the maskrcnn_benchmark framework; the GPU configuration is NVIDIA GeForce GTX 2080. In terms of hyperparameter settings, the initial learning rate is set to 0.0001, the batch training size is set to 1, the number of iterations is set to 10 epochs. The distillation hyperparameters of ILOD method, Faster ILOD method and SDCID method are set to 1.

Five types of selective masks in the SDCID method are set, as shown in Figure 7, where natural images are from the VOC2007 dataset [33] and remote sensing images are from the SIRI dataset [34], which are first transformed to the target size when used as masks, and the pixels are overlaid with equal weights.

![Figure 7. Schematic of selective mask.](image)

5. Results

5.1. Performance Comparison

5.1.1. Experimental Results on the DOTA Dataset

The experimental results of class-increment on the DOTA dataset are shown in Table 1. The SDCID method uses two selective masks: the Gaussian blur mask (M1) and the natural image mask (M2). Since only the joint fine-tuning method uses all historical class data among all methods, its experimental results are for reference only. The experimental results of the Faster ILOD method and ILOD method perform better in stability and worse in plasticity, contrary to the experimental results of the fine-tuning method. Among all methods, except for the joint fine-tuning method used as a reference, the SDCID (M1) method achieved the best results in both metrics, AP50 and C-SP, for all classes. The
model obtained by the SDCID (M1) method had optimal detection capability and balanced stability and plasticity well, which is closest to the joint fine-tuning method for all classes AP50 and C-SP results; the SDCID (M2) method was second.

Further analyzing the AP50 accuracy results on each class of the DOTA dataset, according to the results in Table 2, the ILOD method had the maximum value on one class, the Faster ILOD method had the maximum value on three classes, the SDCID (M2) method had the maximum value on six classes, and the SDCID (M1) method had the maximum value on five classes.

Table 1. Class-incremental experiment results on DOTA dataset. The bold value is the highest score of our method except to the upper bound.

| Model Setting       | AP50 | S (%) | P (%) | C-SP (%) |
|---------------------|------|-------|-------|----------|
|                     | The First 8 Class | The Last 7 Class | All Class |       |
| Learn all data at once | 55.54 | 63.63 |       |          |
| Incremental learning |       |       |       |          |
| Joint fine-tuning   | 57.21 | 54.94 | 56.15 | 103.01 | 86.34 | 94.50 |
| Fine-tuning         | 31.90 | 59.30 | 44.69 | 57.44 | 93.20 | 69.97 |
| ILOD + Faster R-CNN | 55.51 | 54.38 | 52.14 | 99.95 | 75.91 | 87.08 |
| Faster ILOD         | 56.07 | 47.27 | 51.96 | 100.95 | 74.29 | 86.47 |
| SDCID(M1)           | 55.44 | 53.53 | 54.55 | 54.55 | 99.82 | 84.13 | 91.83 |
| SDCID(M2)           | 53.63 | 54.42 | 54.00 | 96.56 | 85.53 | 91.08 |

Table 2. AP50 precision results on all classes of the DOTA dataset. The bold value is the highest score on each category.

| Class              | ILOD + Faster R-CNN | Faster ILOD | SDCID(M1) | SDCID(M2) |
|--------------------|----------------------|-------------|-----------|-----------|
| Large vehicle      | 67.14                | 63.53       | 67.79     | 67.53     |
| Track and field    | 63.17                | 49.48       | 60.47     | 61.35     |
| Helicopter         | 48.94                | 42.84       | 46.18     | 36.03     |
| Bridge             | 42.16                | 40.48       | 40.89     | 39.91     |
| Roundabout         | 59.79                | 64.13       | 42.43     | 39.03     |
| Soccer field       | 40.99                | 55.31       | 55.25     | 55.09     |
| Ship               | 66.51                | 55.31       | 55.25     | 55.09     |
| Oil tank           | 55.39                | 77.55       | 84.02     | 84.83     |
| Tennis court       | 78.84                | 74.21       | 79.46     | 79.96     |
| Aircraft           | 74.72                | 43.72       | 52.69     | 54.13     |
| Diamond            | 43.39                | 45.59       | 45.39     | 46.04     |
| Small vehicle      | 42.65                | 25.74       | 30.86     | 34.58     |
| Basketball Court   | 27.94                | 36.52       | 46.65     | 46.57     |
| Port               | 38.15                | 30.59       | 35.62     | 34.82     |

5.1.2. Experimental Results on the DIOR Dataset

The experimental results of class-increments on the DIOR dataset are shown in Table 3. Similar to the experimental results on the DOTA dataset, among the three baseline methods that do not use historical data, the fine-tuning method had the best results on stability performance and the Faster ILOD method had the best results on plasticity performance. The SDCID (M1) method had the best experimental results on all classes of AP50 and C-SP metrics. The AP50 accuracy results for each class of the DIOR dataset are shown in Table 4. The ILOD method had a maximum value on one class, the Faster ILOD method had a maximum value on four classes, the SDCID (M2) method had a maximum value on nine classes, and the SDCID (M1) method had a maximum value on six classes. The accuracy performance of the SDCID method on two data classes was better than that of the ILOD method and the Faster ILOD method.
Table 3. Class-incremental experiment results on DIOR dataset. The bold value is the highest score of our method except to the upper bound.

| Model Setting | AP50 | S (%) | P (%) | C-SP (%) |
|---------------|------|-------|-------|----------|
|               | The First 8 Class | The Last 7 Class | All Class |       |
| Learn all data at once | 59.74 | 51.21 |       |          |
| Joint fine-tuning | 61.97 | 39.40 | 50.69 | 103.74 | 76.93 | 88.35 |
| Fine-tuning | 48.81 | 44.56 | 46.68 | 81.70 | 87.02 | 84.27 |
| ILOD + Faster R-CNN | 59.68 | 31.13 | 45.41 | 99.91 | 60.79 | 75.58 |
| Faster ILOD | 60.10 | 30.14 | 45.12 | 100.60 | 58.86 | 74.27 |
| SDCID(M1) | 59.35 | 37.70 | 48.52 | 48.52 | 48.52 | 99.34 | 73.62 | 84.57 |
| SDCID(M2) | 59.27 | 37.54 | 48.41 | 99.22 | 73.31 | 84.32 |

Table 4. AP50 precision results on all classes of the DIOR dataset. The bold value is the highest score on each category.

| Class | ILOD + Faster R-CNN | Faster ILOD | SDCID(M1) | SDCID(M2) |
|-------|---------------------|------------|-----------|-----------|
| Aircraft | 48.43 | 49.01 | 48.16 | 49.56 |
| Airport | 73.19 | 73.45 | 71.94 |
| Baseball field | 65.59 | 65.61 | 65.88 |
| Basketball field | 83.82 | 83.34 | 83.70 |
| Bridge | 10.55 | 11.20 | 11.72 |
| Chimney | 73.95 | 74.01 | 74.26 |
| Dam | 50.36 | 50.41 | 50.36 |
| Expressway service area | 68.41 | 67.45 | 67.39 |
| Expressway toll-station | 51.35 | 47.67 | 48.43 |
| Golf course | 71.19 | 71.41 | 71.00 |
| Track and field | 41.08 | 51.69 | 51.69 |
| Port | 21.62 | 23.86 | 24.01 |
| Flyover | 33.75 | 35.96 | 36.81 |
| Ship | 27.19 | 26.51 | 27.96 |
| Court | 42.53 | 44.86 | 44.87 |
| Oil tank | 14.78 | 17.43 | 17.85 |
| Tennis court | 67.02 | 68.95 | 68.57 |
| Railway station | 10.66 | 18.37 | 18.37 |
| Vehicle | 16.98 | 21.37 | 20.65 |
| Windmill | 35.67 | 33.86 | 54.20 |

Furthermore, analyzing the algorithm’s performance on historical classes and new classes, we found that our algorithm learns some similar or shared classes better than other algorithms. For example, since ship and port objects have similar objects, ILOD and Faster ILOD perform poorly on new classes (port) after learning ship objects. Our method, either using the M1 or M2 mask approach, significantly improves. A similar phenomenon is supported by the results in Table 4, e.g., the existing method also failed to identify the class of port on the DIOR dataset, while SDCID based methods work well. The accuracy of our method on the old class is close to or exceeds that of similar methods while the accuracy on the new class is significantly better than other methods.

5.2. Prediction Results Analysis

To visually compare the detection performance of the fine-tuning, ILOD method, Faster ILOD method, and SDCID method, the images in the DOTA validation set are predicted with the detection models trained by the four methods. The prediction results are
shown in Figure 8, there are two classes of objects in the figure, the bridge is the historical class and the vehicle is the new class.

![Figure 8. Prediction result on DOTA validation set.](image)

Among all methods, the SDCID method proposed in this paper can detect objects of both new and historical classes in remote sensing images with the best detection results, which shows that the method can adapt to both historical and new classes appearing in different training periods. In contrast, the fine-tuning algorithm has poor detection for the historical class and cannot detect objects with the class of bridges in remote sensing images, while it has better detection for the new class and can detect vehicles in remote sensing images better, and does not have a good ability to detect objects of the historic and new classes at the same time. Unlike the performance of the fine-tuning algorithm, the ILOD and Faster ILOD methods are able to detect bridge objects on remote sensing images containing the historic class, but cannot detect vehicles well on remote sensing images containing the new class; only a few vehicle objects can be detected, and the ability of detecting historic class objects is stronger than that of detecting the new class.

5.3. Efficiency Analysis

In order to verify the effectiveness of the SDCID method in the three stages, distillation input, selection of region proposal box, and calculation of distillation loss, experiments were conducted on the DOTA dataset to validate each of the three stages.

5.3.1. Distillation Input Step

First, for the adopted masks, five selective masks were set, and their experimental results are shown in Table 5. The other settings of the SDCID method were kept constant in the experiments. The best results were obtained under the Gaussian blur mask (M1) setting.
on AP50, S, and C-SP metrics, and the best results were obtained under the natural image mask (M2) setting on the P metric. Overall, the AP50 results and C-SP results of the SDCID method with the five selective mask settings were better than those of the fine-tuning method, ILOD method, and Faster ILOD method.

Table 5. Experiment result—validation of the SDCID method for distillation input step. The bold value is the highest score on each metric.

| Selective Mask                          | AP50 | S (%) | P (%) | C-SP (%) |
|-----------------------------------------|------|-------|-------|----------|
| Gaussian blur mask (M1)                 | 54.55| 99.82 | 84.13 | 91.83    |
| Natural image mask (M2)                 | 54.00| 96.56 | 85.53 | 91.08    |
| Object overlay with natural images (M3) | 53.50| 96.45 | 83.97 | 90.20    |
| Remote sensing image mask (M4)          | 53.34| 94.98 | 84.90 | 89.99    |
| Object overlay with remote sensing images (M5) | 53.74| 96.65 | 84.55 | 90.60    |

5.3.2. Region Proposal Box Selection Step

The SDCID method selected all the candidate boxes from the teacher model when selecting the region proposal box for distillation, and validation experiments were conducted for this part of the settings. Seven types of region suggestion box selection methods were set, the other settings of the SDCID method were kept unchanged, the selective mask was selected as M1, and its experimental results are shown in Table 6. It can be seen that the best results on three indicators, AP50, P, and C-SP, were obtained with the setting that all the region proposal boxes come from the teacher model, and the best results on the S indicator were obtained with the settings that the number of region proposal boxes from the teacher model and the student model were 8 and 56, respectively, but the results on other indicators were poor.

Table 6. Experiment result—validation of the SDCID (M1) method for region proposal. The bold value is the highest score on each metric.

| Number of Region Proposal Boxes From Teacher Model | AP50 | S (%) | P (%) | C-SP (%) |
|---------------------------------------------------|------|-------|-------|----------|
| 0                                                 | 64   | 54.13 | 100.31| 82.24    | 90.98    |
| 8                                                 | 56   | 53.61 | 101.85| 101.85   | 78.95    | 89.71    |
| 16                                                | 48   | 53.43 | 100.96| 79.21    | 89.49    |
| 32                                                | 32   | 53.58 | 101.58| 79.10    | 89.69    |
| 48                                                | 16   | 53.51 | 100.66| 78.90    | 89.72    |
| 56                                                | 8    | 53.27 | 100.34| 79.29    | 89.28    |
| 64                                                | 0    | 54.55 | 99.82 | 84.13    | 91.83    |

The experimental results for the selective mask selection M2 are displayed in Table 7, and it can be seen that the best results on the three metrics AP50, P, and C-SP were obtained with the setting where all the region proposal boxes come from the teacher model. The best results on the S metric were obtained with the settings where the number of region proposal boxes from the teacher model and the student model were 56 and 8, respectively, but the results on the other metrics were poor.
Table 7. Experiment result—validation of the SDCID (M2) method for region proposal box selection step. The bold value is the highest score on each metric.

| Number of Region Proposal Boxes | From Teacher Model | From Student Model | AP50 | S (%) | P (%) | C-SP (%) |
|-------------------------------|--------------------|--------------------|------|-------|-------|---------|
| 0                             | 64                 | 53.88              | 96.96| 84.72 | 90.84 |
| 8                             | 56                 | 53.01              | 96.91| 81.86 | 92.25 |
| 16                            | 48                 | 53.41              | 98.01| 82.10 | 89.88 |
| 32                            | 32                 | 53.46              | 98.60| 81.68 | 89.90 |
| 48                            | 16                 | 53.41              | 98.01| 82.10 | 89.88 |
| 56                            | 8                  | 53.27              | 100.34| 79.29 | 89.28 |
| 64                            | 0                  | 54.00              | 96.56| 85.53 | 91.08 |

Two validation experiments under selective mask selection settings show that the SDCID approach to selecting the region proposal box (all from the teacher model) has superior model accuracy as well as combined stability–malleability performance compared to other region proposal box selection approaches.

5.3.3. Distillation Loss Calculation Step

The SDCID method reconstructs the classification results of the student model when calculating distillation loss. If the reconstruction approach is not taken, the original approach is calculated for the classification results of the non-added classes of the student model. In order to verify the effectiveness of the SDCID method using the reconstruction approach when calculating distillation loss, validation experiments were conducted on the DOTA dataset. The other settings of the SDCID method were kept unchanged, the selectivity masks were M1 and M2, and the experimental results after the original approach and the reconstruction approach to calculate selective distillation loss are shown in Table 8 and Table 9, respectively. From the experimental results, it can be seen that the reconstruction approach outperforms the original approach in every index, which shows the effectiveness for the distribution of its classification results of the SDCID method to calculate distillation loss by transferring the classification information of the student model to the background class.

Table 8. Experiment result—validation of the SDCID (M1) method for distillation loss calculation step. The bold value is the highest score on each metric.

| Calculation of Classification Results for Student Models in Selective Distillation Loss | AP50 | S (%) | P (%) | C-SP (%) |
|-------------------------------------|------|-------|-------|---------|
| Original                            | 53.89| 98.04 | 83.69 | 90.77   |
| Reconstructed                       | 54.55| 99.82 | 84.13 | 91.83   |

Table 9. Experiment result—validation of the SDCID (M2) method for distillation loss calculation step. The bold value is the highest score on each metric.

| Calculation of Classification Results for Student Models in Selective Distillation Loss | AP50 | S (%) | P (%) | C-SP (%) |
|-------------------------------------|------|-------|-------|---------|
| Original                            | 53.41| 95.25 | 84.85 | 90.09   |
| Reconstructed                       | 54.00| 96.56 | 85.53 | 91.08   |

5.4. Hyperparameter Analysis

In the SDCID method, there is a distillation hyperparameter, $\lambda$, whose value is set to 1 by default in the performance comparison and efficiency analysis experiments. We analyzed the distillation hyperparameters by investigating the effect of different distillation hyperparameters on the performance of the SDCID method. The values of distillation hyperparameters were set to seven values in addition to the default value. A distillation
hyperparameter value greater than 1 indicates that the overall loss function guides the original detection model to focus more on knowledge distillation to maintain the knowledge on the object of historical classes and reduce the learning effort on the data of new classes. When the distillation hyperparameter value is less than 1, the overall loss function has less weight to maintain the knowledge on objects of historical classes according to the knowledge selective distillation strategy, and the parameters related to the distribution of new class data are continuously updated by fine-tuning.

The experiments were conducted on the DOTA dataset with the setting of the selective mask of M1, and the AP50 accuracy results for the historical class, the added class, and all classes are shown in Figure 9. The experiments were conducted on the DOTA dataset with the setting of selective mask of M1, and the AP50 accuracy results for the historical class, the added class, and all classes are shown in Figure 9. In addition, the accuracy upper limit of incremental learning, which comes from the detection model after training all data at once, is also shown in the figure. It can be seen from the figure that the AP50 accuracy on the new class gradually decreases as the distillation hyperparameter value increases, and the maximum value has a large gap with the accuracy upper limit on the new class, while the AP50 accuracy on the historical class gradually increases and already exceeds the accuracy upper limit on the historical class from setting the distillation hyperparameter value to 1. The SDCID (M1) method maintains the detection accuracy on the historical class well by the distillation strategy; in contrast, the learning performance of the added class still needs to be improved.

Figure 9. Prediction result on DOTA validation set.

The AP50 accuracy curves for the historic and new classes with different distillation hyperparameters showed opposite trends, with the two distillation hyperparameters before and after the intersection point of the two curves being 0.5 and 1. The highest accuracy result for all classes with $\lambda = 1$ was obtained with an AP50 value of 54.55, which well balanced the detection accuracy on the historic and new classes. The second-highest was when $\lambda = 0.5$ with an AP50 value of 54.51. The accuracy of the former (55.44) was higher than that of the latter (53.71), while the accuracy of the former (53.53) was lower than that of the latter (55.43), and the overall learning performance of the method for the new classes was not satisfactory. Thus, it needs to be further analyzed based on the stability and plasticity performance to further analyze the influence of distillation hyperparameters on the SDCID method.

The stability and plasticity performance of the SDCID (M1) method on the DOTA dataset for different distillation hyperparameter settings are shown in Figure 10. The three curves in the figure are stability performance (S), plasticity performance (P), and combined stability-plasticity performance (C-SP), respectively. The S and P curves have opposite
trends with the change of distillation hyperparameter values. Among all distillation hyperparameters, the highest C-SP value of 91.98% was obtained when $\lambda = 0.5$, which very well balanced stability and plasticity. The C-SP value at $\lambda = 1$ was the second-highest at 91.82%, and the difference between S and P values was greater than that between S and P values at $\lambda = 0.5$. Therefore, the SDCID (M1) method can balance the learning weights of the target knowledge of the historical class and the new class by setting the distillation hyperparameter value to 0.5 when choosing the distillation hyperparameter according to the stability and plasticity performance in this experiment.

![Figure 10. Prediction result on DOTA validation set.](image)

6. Discussion
6.1. Advantages
1. Remote sensing images capture a wide range of surface objects, and existing class-incremental detection methods only consider objects of new class in new data, lacking insight into the phenomenon that objects of a new class may exist in historic data; this paper fills the research gap in this part. The ultimate goal of the class-incremental detection method is that the target model should maintain the ability to correctly detect historic classes in previous data when data increments occur, but also correct the previous learning experience of learning new classes as background classes based on the label information of new classes in new data, and detect objects of new class as the corresponding foreground classes.

2. The distillation strategies of existing knowledge distillation-based class-incremental detection methods and the stability and plasticity performance of the methods are summarized as the basis of the proposed method in this paper. In the existing class-incremental detection methods based on knowledge distillation, the student model learns the knowledge of new class while receiving the knowledge distilled from the teacher model. When the remote sensing objects of the new class appear in the historical class data in an unlabeled manner, the teacher model determines the new class objects as background rather than as foreground with semantic information. The undifferentiated distillation strategy on the teacher model, on the other hand, leads to the problem of inefficient learning of knowledge of new class objects.

3. The proposed method utilizes a selective distillation strategy to limit the scope of knowledge distillation, which well solves the negative effect of the original knowledge distillation strategy on the model plasticity. From the experimental results of the performance comparison, we can see that the method can simultaneously take care of the model plasticity and stability, and can achieve the highest detection accuracy on all classes, including historical classes and new classes.
6.2. Deficiency and Prospect

In this paper, for the scenario of new class data appearing in remote sensing images, the class-incremental detection method of remote sensing images that can continue to learn new knowledge on top of the existing knowledge is investigated to meet the demand for the detection of dynamically changing remote sensing image objects. However, there are still some shortcomings to be further investigated and improved, including the following aspects:

1. For algorithm evaluation, we only evaluated the algorithm on two publicly available remote sensing image object detection datasets, which suffer from two problems: on the one hand, the data size is small, while we believe that the public class incremental conflict situation will be more severe in larger scale image object detection tasks; on the other hand, there are fewer conflicting classes in the dataset, which also limits the performance of the algorithm.

2. Regarding algorithm theory, we found that the mask’s design has an essential impact on the algorithm. However, the intrinsic factors of the mask affecting the algorithm performance are unclear and need further exploration.

3. The class-incremental detection method of remote sensing based on knowledge selective distillation proposed in this paper is oriented to the incremental data scenario of new class data. The incremental data in the real world is complex and diverse, and there may be both similar instance data and new class data. It is a very important research direction for the future to study the incremental detection method of adaptive multi-type remote sensing object data increment to meet the detection demand more comprehensively.

4. The research on remote sensing incremental detection method in this paper focuses on one-time incremental learning under the current moment. However, the increment of remote sensing data exists for a long time, and if the existing knowledge fusion mechanism for one-time incremental learning is used to complete multiple incremental learning, each incremental learning requires knowledge fusion, which may lead to the problem of low stability of the learned knowledge when the data differences of multiple increments are too large. Therefore, it is necessary to study remote sensing increment detection methods that can be oriented to long-range increments of remote sensing images.

5. There have been many studies on designing improvement modules for detecting small objects, multi-scale, complex backgrounds, dense distribution, shape differences, and other detection difficulties on remote sensing images, and their model accuracy is higher than the classical Faster R-CNN model. In this paper, the object detection models are all classical Faster R-CNN models, and the improvement modules that can improve the detection accuracy are not added to the incremental detection methods. It is also one of the future research directions to study incremental detection methods for remote sensing image detection difficulties.

7. Conclusions

Class-incremental detection is significant in extending the detection range to the new class. Existing class-incremental detection research fails to distinguish well between the new and old knowledge that identifies the objects of the new class as the background. To solve this knowledge fusion, we propose a new remote sensing class-incremental detection method based on selective knowledge distillation based on the asymmetry of teacher-student architecture. It selectively distills the knowledge related to historical classes while learning new knowledge. This method firstly adds the images input to the teacher model and the student model in the distillation strategy with selective masks for the objects of a new class, then obtains the determination results based on the region candidate boxes extracted by the teacher model, and finally calculates the selective distillation loss by reconstructing the classification results of the student model, effectively aligning the learning results of the teacher model and the student model, and transferring the knowledge
unrelated to the objects of a new class from the teacher model to the student model. We evaluate the proposed method on two object detection datasets for remote sensing images, i.e., DIOR and DOTA. The results of the performance comparison experiments fully demonstrate the excellent performance of the method proposed in terms of object detection accuracy and the combined performance of model stability and plasticity, which has surpassed the existing methods. We also analyzed our method’s advantages and the existing methods’ disadvantages according to the prediction results on remote sensing images. In addition, we further explored the influence of distillation hyperparameters and the mask’s design.

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

Funding: This work was supported in part by the National Natural Science Foundation of China, Grant/Award Numbers: 42271481 and 41861048; the High Performance Computing Platform of Central South University and HPC Central of Department of GIS, in providing HPC resources.

Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: The authors declare that there is no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

- AlexNet: ImageNet classification with deep convolutional neural networks
- AP: Average precision
- AP50: The AP below the threshold 0.5 of IoU
- CNN: Convolution neural network
- C-SP: Combined stability-plasticity performance
- Fast R-CNN: Upgraded version of R-CNN
- Faster R-CNN: Upgraded version of Fast R-CNN
- GoogleNet: Going deeper with convolutions
- GPU: Graphic processing unit
- LwF: Learning without forgetting
- R-CNN: Region-based CNN
- ResNet: Residual network
- RPN: Region Proposal Networks
- SDCID: Class-incremental Detection method based on Selective Distillation
- VGG: Visual geometry group

References

1. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* 2016, 117, 11–28.
2. Qin, W.; Song, T.; Liu, J.; Wang, H.; liang, Z. Remote Sensing Military Target Detection Algorithm Based on Lightweight YOLOv3. *Comput. Eng. Appl.* 2021, 57, 7.
3. McCloskey, M.; Cohen, N.J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*; Elsevier: Amsterdam, The Netherlands, 1989; Volume 24, pp. 109–165.
4. Ratcliff, R. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychol. Rev.* 1990, 97, 285.
5. French, R.M. Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* 1999, 3, 128–135.
6. Shmelkov, K.; Schmid, C.; Alahari, K. Incremental learning of object detectors without catastrophic forgetting. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3400–3409.
7. Li, D.; Tasci, S.; Ghosh, S.; Zhu, J.; Zhang, J.; Heck, L. RILOD: Near real-time incremental learning for object detection at the edge. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing, Arlington, VI, USA, 7–9 November 2019*; pp. 113–126.
8. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Daturu, 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.

9. Peng, C.; Zhao, K.; Lovell, B.C. Faster ilod: Incremental learning for object detectors based on faster rcnn. Pattern Recognit. Lett. 2020, 140, 109–115.

10. Shuai, T.; Sun, K.; Shi, B.; Chen, J. A ship target automatic recognition method for sub-meter remote sensing images. In Proceedings of the 2016 4th International Workshop on Earth Observation and Remote Sensing Applications (EORSASA), Guangzhou, China, 4–6 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 153–156.

11. Shi, Z.; Yu, X.; Jiang, Z.; Li, B. Ship detection in high-resolution optical imagery based on anomaly detector and local shape feature. IEEE Trans. Geosci. Remote Sens. 2013, 52, 4511–4523.

12. Konstantinidis, D.; Stathaki, T.; Argyriou, V.; Grammalidis, N. Building detection using enhanced HOG–LBP features and region refinement processes. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2016, 10, 888–905.

13. Tuermier, S.; Kurz, F.; Reinartz, P.; Stilla, U. Airborne vehicle detection in dense urban areas using HoG features and disparity maps. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2013, 6, 2327–2337.

14. Zheng, J.; Xi, Y.; Feng, M.; Li, X.; Li, N. Object detection based on BING in optical remote sensing images. In Proceedings of the 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Dantong, China, 15–17 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 504–509.

15. Song, Z.; Sui, H.; Wang, Y. Automatic ship detection for optical satellite images based on visual attention model and LBP. In Proceedings of the 2014 IEEE Workshop on Electronics, Computer and Applications, Ottawa, ON, Canada, 8–9 May 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 722–725.

16. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Commun. ACM 2017, 60, 84–90.

17. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.

18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.

19. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

20. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.

21. 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.

22. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.

23. 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: Cham, Switzerland, 2016, pp. 21–37.

24. Yang, X.; Yan, J.; Liao, W.; Yang, X.; Tang, J.; He, T. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. IEEE Trans. Pattern Anal. Mach. Intell. 2022, https://doi.org/10.1109/TPAMI.2022.3166956.

25. Wang, J.; Ding, J.; Guo, H.; Cheng, W.; Pan, T.; Yang, W. Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. Remote Sens. 2019, 11, 2930.

26. Yang, F.; Li, W.; Hu, H.; Li, W.; Wang, P. Multi-scale feature integrated attention-based rotation network for object detection in VHR aerial images. Sensors 2020, 20, 1686.

27. Li, C.; Xu, C.; Cui, Z.; Wang, D.; Zhang, T.; Yang, J. Feature-attentioned object detection in remote sensing imagery. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3886–3890.

28. Hinton, G.; Vinyals, O.; Dean, J.; et al. Distilling the knowledge in a neural network. arXiv 2015, 2, arXiv:1503.02531.

29. Li, Z.; Hoiem, D. Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 40, 2935–2947.

30. Dhar, P.; Singh, R.V.; Peng, K.C.; Wu, Z.; Chellappa, R. Learning without memorizing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5138–5146.

31. Douillard, A.; Cord, M.; Ollion, C.; Robert, T.; Valle, E. Podnet: Pooled outputs distillation for small-tasks incremental learning. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 86–102.

32. Chen, Q.; Wang, S.; Chen, L. Incremental detection of remote sensing objects with feature pyramid and knowledge distillation. IEEE Geosci. Remote Sens. 2020, 60, 1–13.
33. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* 2010, 88, 303–338.

34. Zhu, Q.; Zhong, Y.; Zhao, B.; Xia, G.S.; Zhang, L. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. *IEEE Geosci. Remote Sens. Lett.* 2016, 13, 747–751.