remote sensing

MDPI

Article

Consecutive Pre-Training: A Knowledge Transfer Learning Strategy with Relevant Unlabeled Data for Remote Sensing Domain

Tong Zhang [1], Peng Gao [2], Hao Dong [3], Yin Zhuang [1,*], Guanqun Wang [1], Wei Zhang [4] and He Chen [1]

[1] Beijing Key Laboratory of Embedded Real-Time Information Processing Technology, Beijing Institute of Technology, Beijing 100081, China
[2] Shang Hai AI Laboratory, Shanghai 100024, China
[3] Center on Frontiers of Computing Studies (CFCS), School of Computer Science (CS), Peking University, Beijing 100871, China
[4] Advanced Research Institute of Multidisciplinary Sciences, Beijing Institute of Technology, Beijing 100081, China
* Correspondence: yzhuang@bit.edu.cn

check for updates

Citation: Zhang, T.; Gao, P.; Dong, H.; Zhuang, Y.; Wang, G.; Zhang, W.; Chen, H. Consecutive Pre-Training: A Knowledge Transfer Learning Strategy with Relevant Unlabeled Data for Remote Sensing Domain. *Remote Sens.* **2022**, *14*, 5675. https://doi.org/10.3390/rs14225675

Academic Editor: Stephan Havemann

Received: 17 September 2022
Accepted: 31 October 2022
Published: 10 November 2022

Abstract: Currently, under supervised learning, a model pre-trained by a large-scale nature scene dataset and then fine-tuned on a few specific task labeling data is the paradigm that has dominated knowledge transfer learning. Unfortunately, due to different categories of imaging data and stiff challenges of data annotation, there is not a large enough and uniform remote sensing dataset to support large-scale pre-training in the remote sensing domain (RSD). Moreover, pre-training models on large-scale nature scene datasets by supervised learning and then directly fine-tuning on diverse downstream tasks seems to be a crude method, which is easily affected by inevitable incorrect labeling, severe domain gaps and task-aware discrepancies. Thus, in this paper, considering the self-supervised pre-training and powerful vision transformer (ViT) architecture, a concise and effective knowledge transfer learning strategy called ConSecutive Pre-Training (CSPT) is proposed based on the idea of not stopping pre-training in natural language processing (NLP), which can gradually bridge the domain gap and transfer large-scale data knowledge to any specific domain (e.g., from nature scene domain to RSD) In addition, the proposed CSPT also can release the huge potential of unlabeled data for task-aware model training. Finally, extensive experiments were carried out on twelve remote sensing datasets involving three types of downstream tasks (e.g., scene classification, object detection and land cover classification) and two types of imaging data (e.g., optical and synthetic aperture radar (SAR)). The results show that by utilizing the proposed CSPT for task-aware model training, almost all downstream tasks in the RSD can outperform the previous knowledge transfer learning strategies based on model pre-training without any expensive manually labeling and even surpass the state-of-the-art (SOTA) performance without any careful network architecture designing.

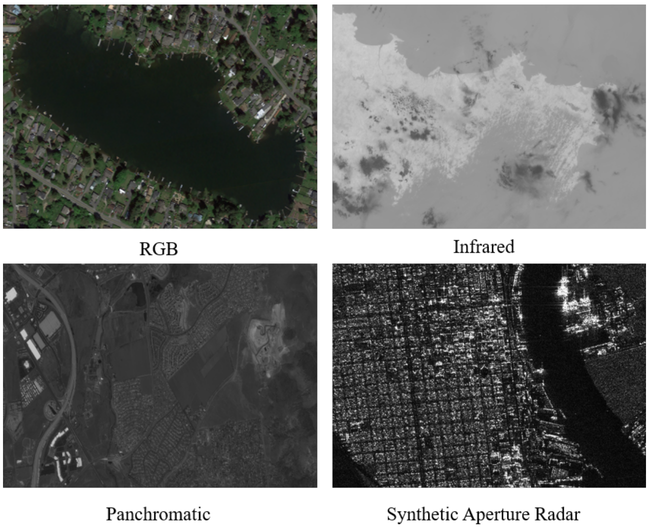
Keywords: knowledge transfer learning; remote sensing domain; self-supervised learning; vision transformer

1. Introduction

With the rapid development of remote sensing technology, there has been a gradual accumulation of available earth observation imaging data, which can be used for urban planning, resource investigation, military surveillance and rapid search and rescue in large-scale regions [1–5]. Consequently, how to convert the acquired large amount of remote sensing imaging data into valid information to support practical applications has become a very important question. At present, various data-hungry models, such as convolutional neural networks (CNNs) [6–13] and vision transformers (ViTs) [14–17], have emerged and been widely used for nature scene image interpretation tasks. Therefore, in order to apply these trained data-hungry models for other specific domains, knowledge transfer learning should be considered; specifically, when data-hungry models are applied for remote sensing domain (RSD), they have to utilize a large-scale dataset (e.g., ImageNet [18]) to sufficiently stimulate their potential and then better adapt for various downstream tasks (e.g., scene classification, object detection, and semantic segmentation) in the RSD. Until now, many efforts [19–34] have demonstrated that the consensus solution of supervised pre-training based knowledge transfer learning for model training has basically formed, which needs to pre-train the model on a large-scale dataset with manual annotation and then directly fine-tune pre-trained models on downstream task datasets in the RSD.

For example, Xia, G.S. et al. [19] pre-trained the CNN-based models (e.g., CaffeNet [6], VGG-VD-16 [7] and GoogLeNet [8]) on the large-scale labeled dataset ImageNet [18] by supervised learning and then fine-tuned them on scene classification dataset of AID [19], whose results show its great generalization ability compared with other traditional handcraft feature based methods. Related to land cover classification in the RSD, Liu, Y. et al. [32] utilized the PASCAL VOC [35] dataset to pre-train the encoder of their model, and then mitigated the pre-trained encoder on fewer specific land cover labeling data to provide acceptable pixel-wise prediction results. For the task of the object detection of SAR images, Li, J. et al. [29] considered to transfer the ZFNet [9] that is pre-trained on ImageNet [18] into SAR ship detection dataset SSDD. Specifically, to avoid overfitting on the small scale dataset, they froze the former three layers of pre-trained model and only fine-tuned the latter two layers on SSDD dataset to obtain better ship detection performance from SAR images. Like the above mentioned studies, not only supervised pre-training based knowledge transfer learning is widely employed on various remote sensing application scenes, but also different kinds of imaging data are involved and they have significant differences in imaging characteristics and spatial resolutions, as shown in Figure 1. Thus, it is difficult to set up a large enough and uniform dataset to support large-scale pre-training for various downstream tasks of different kinds of imaging data in the RSD, and it is also difficult and expensive to manually label a large-scale dataset with diverse imaging data for model pre-training. Instead, the relatively mature and large-scale nature scene datasets (e.g., ImageNet [18], Place365 [36], COCO [37] and PASCAL VOC [35]) as reported in Table 1 are usually employed for the data-hungry model training and generating transferable domain-level knowledge at pre-training step. Then, the transferable domain-level knowledge can be adapted into various downstream tasks to largely promote the fine-tuning performance comparing with training from scratch. Although it seems to be the right method for model training in the RSD, the supervised pre-training based knowledge transfer learning still suffers some problems regarding the inevitable incorrect labeling, severe domain gap and task-aware discrepancy, which constrain the further performance improvement of various downstream tasks in the RSD.

Currently, except for knowledge transfer learning via supervised pre-training methods [19,20,22–26,28–34,38] in the RSD, refs. [39–41] began to explore self-supervised pre-training based knowledge transfer learning, which can avoid massive manually labeling cost and release the potential of unlabeled data for model training in the RSD. In addition, they also pointed out that the self-supervised pre-training can generate more transferable feature representation than supervised pre-training method. According to studies of [39–42], the self-supervised pre-training based knowledge transfer learning exhibits two advantages: (1) no need to carefully label a large-scale and unified dataset; (2) more generalized and transferable feature representation, which are friendly to knowledge transfer learning in highly specialized fields such as RSD. For instance, Stojnic, V. et al. [43] analyzed the applicability of knowledge transfer learning with different numbers and domains of unlabeled images for self-supervised pre-training and then adapted them to scene classification task in the RSD. By analyzing the results, ref. [43] indicated that the self-supervised pre-training can provide better knowledge transfer ability than supervised pre-training, even when using significantly fewer unlabeled images for self-supervised pre-training. Then, considering the task-aware discrepancy between pre-training and fine-tuning processes, Li, H. et al. [44] specifically proposed a self-supervised pre-training pretext task of global style and local matching to adapt the requirement about pixel-level discrimination of land cover classification task, and the results indicated that the correlation between pretext and downstream tasks is important for mitigating task-aware discrepancy. Next, refs. [45–47] demonstrated that when collecting sufficient unlabeled data and narrowing the domain gap by collecting the cognate data for pre-traininig and fine-tuning steps, a self-supervised pre-training model would perform better on downstream tasks. However, according to the results of [46,47], they only can obtain limited performance improvement on multiple downstream tasks in the RSD than supervised pre-training based knowledge transfer learning from ImageNet [18]. In addition, Reed, C. J. et al. [48] provided a study of hierarchical self-supervised pre-training for knowledge transfer learning. They found that self-supervised pre-training on ImageNet [18] can improve the self-supervised pre-training on the unlabeled image data of RSD, and it can also reduce the convergence time and perform better on diverse downstream tasks than in-domain self-supervised pre-training carried out from scratch.


RGB

Infrared

Panchromatic

Synthetic Aperture Radar

Figure 1. The different types of imaging data in remote sensing domain.

Table 1. The description of natural scene and remote sensing datasets.

Domain	Payload	Task	Dataset	Resolution (m)	Classes	# Trainval	# Test
Natural Scene	RGB	Classification	ImageNet [18]	-	1000	1,331,167	100,000
			Place365 [36]	-	365	1,803,460	36,500
		Detection/Segmentation	COCO [37]	-	80	123,287	40,670
			PASCAL VOC [35]	-	20	11,530	-
Remote Sensing	Optical	Classification	AID [19]	0.5 to 8	30	2000	8000
			NWPU-RESISC45 [20]	0.2 to 30	45	9450	22,000
		Segmentation	POTSDAM [21]	0.05	6	3456	2016
			VAIHINGEN [21]	0.09	6	344	398
			GID [22]	0.8 to 3.24	15	4368	2912
		Detection	DIOR [23]	0.5 to 30	20	11,725	11,738
			NWPUVHR-10 [24]	0.5 to 2	10	1479	1279
			UCAS-AOD [25]	-	2	6489	2824
			HRSC2016 [26]	0.4 to 2	1	617	438
	SAR	Classification	MSTAR [27]	0.3	8	1890	7576
		Detection	SSDD [29]	1 to 15	1	812	348
			HRSID [28]	0.5 to 3	1	3642	1962

Therefore, the self-supervised pre-training based knowledge transfer learning is progressively becoming a mainstream to replace the supervised pre-training based knowledge transfer learning for data-hungry model training in the RSD. However, although above self-supervised pre-training based knowledge transfer learning can utilize large-scale unlabeled imaging data to obtain a more universal and transferable feature representation and then adapt it into the RSD, the severe domain gap from data difference between pre-training and fine-tuning steps should not be ignored, because it would affect the performance of knowledge transfer learning. Furthermore, due to current self-supervised pre-training is mainly relied on contrastive learning such as SimCLR [49], MoCo [50], BYOL [51] and SwAV [52], positive and negative pairs of unlabeled data still need to be carefully allocated for model pre-training, meanwhile the global decision information from the deep layer of CNNs is often employed for pretext tasks of instance level discrimination and invariant representation learning. Here, these pretext tasks work well for global decision task such as scene classification but not so well for dense prediction tasks (e.g., object detection and segmentation tasks) because there is the task-aware discrepancy between the pretext and downstream tasks. Accordingly, most CNN-based self-supervised pre-training methods for knowledge transfer learning can only achieve limited performance improvement over previous supervised pre-training method. Considering above issues, in order to achieve a unified and more effective self-supervised pre-training based knowledge transfer learning for model training to promote various downstream tasks in the RSD, there are still much room to be explored such as how to overcome the domain gap between pre-training data and fine-tuning data, select an effective network architecture for pre-training step and set an optimal pretext task. Recently, ViT based models have proven to be more powerful network architecture than CNNs, which interprets an image as a sequence of patches and then process them by a standard transformer structure as used in natural language processing (NLP). Obviously, the patch structure provides a condition for applying the idea of the pretext task of masked language modeling which achieved great success and unified the pre-training paradigm in NLP. Therefore, some works such as those on [53–57] have begun turning to masking partial image patches and then reconstructing the original image (i.e., masked image modeling (MIM) task) by an encoder-decoder structure for self-supervised pre-training in computer vision field.

In this article, inspired by the knowledge transfer learning idea of not stopping pre-training in NLP [58,59] and the pretext task of MIM, a concise and effective knowledge transfer learning strategy called ConSecutive Pre-Training (CSPT) is designed for model training based on self-supervised pre-training to promote almost all downstream tasks of RSD, as shown in Figure 2. Here, the ViT based encoder-decoder architecture is employed for model pre-training and then only using the pre-trained encoder for model fine-tuning on diverse downstream tasks. Different from the current existing knowledge transfer learning methods, we utilize the ViT model and the MIM of task-agnostic representation for the consecutive self-supervised pre-training process both on unlabeled large-scale data and task-related data, and knowledge transfer learning from nature scene to RSD is selected as the study scenario to prove that the designed CSPT can bridge the severe domain gap and establish a more effective and transferable feature representation to facilitate the fine-tuning step after pre-training on unlabeled natural and remote sensing data.

Meanwhile, since self-supervised pre-training does not require manual annotation, a large amount of task-related unlabeled data can be employed on a consecutive self-supervised pre-training process, which can not only leverage the domain-level knowledge generated from large-scale nature scene data but also greatly release the potential of unlabeled data for model training in the RSD. Finally, extensive experiments are performed on a large-scale nature scene dataset (i.e., ImageNet [18]) and twelve remote sensing datasets (i.e., AID [19], NWPU-RESISC45 (NR45) [20], ISPRS POTSDAM and VAIHINGEN [21], GID [22], DIOR [23], NWPUVHR-10 [24], UCAS-AOD [25], HRSC2016 [26], MSTAR [27], HRSID [28] and SSDD [29]). These remote sensing datasets involve three downstream tasks (i.e., scene classification, object detection and land cover classification) and two categories of imaging data (i.e., optical RGB and synthetic aperture radar (SAR) images). From the experimental results, we find that when more task-related unlabeled data is joined into the further self-supervised pre-training step or waiting for more iterative epochs, the newly designed CSPT can achieve the promising performance even reaching the state-of-the-art (SOTA) result without any expensive labeling consumption and careful model design. In summary, the contributions of our study are summarized below:

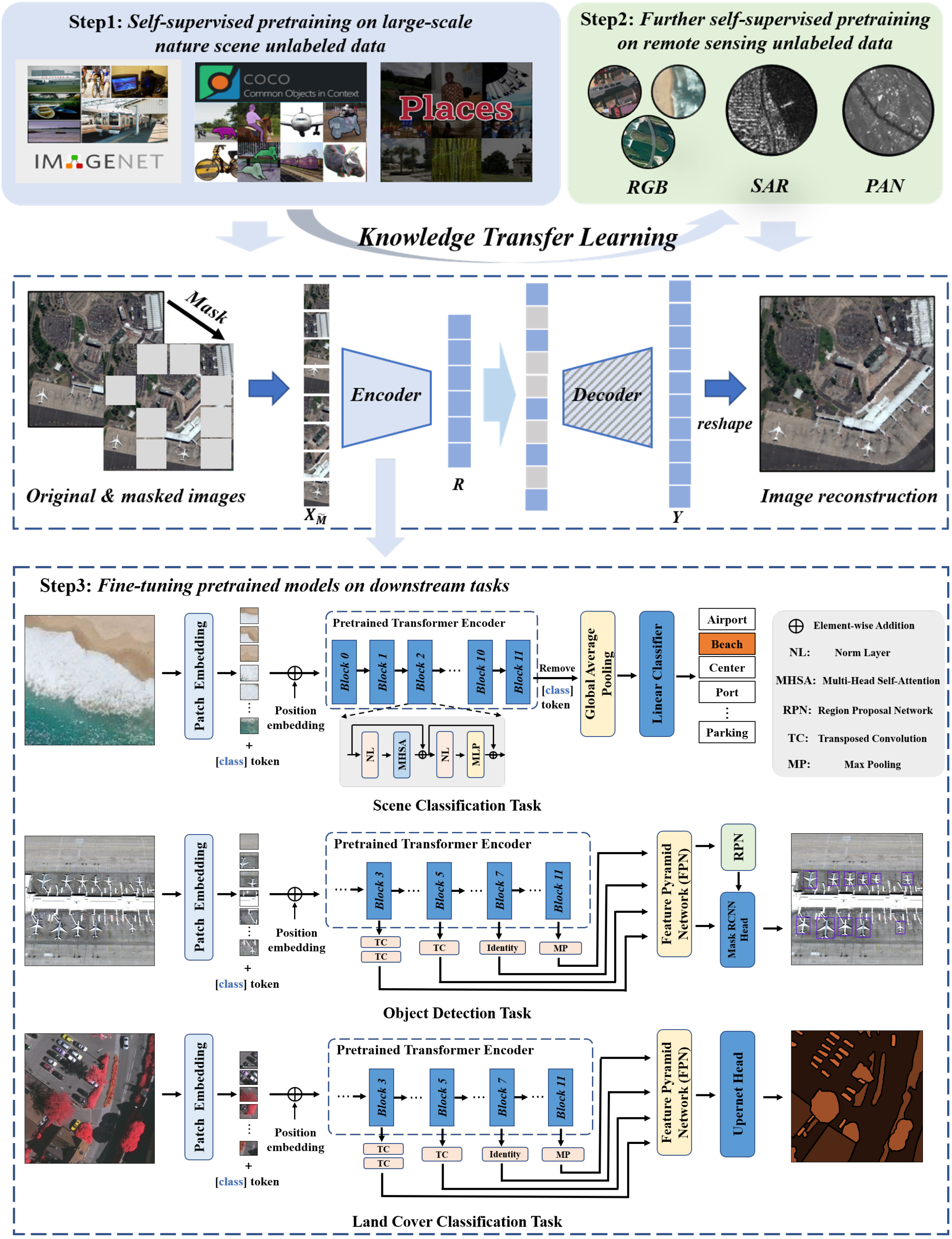

Step1: Self-supervised pretraining on large-scale nature scene unlabeled data
IMAGENET
COCO Common Objects in Context
Places
Step2: Further self-supervised pretraining on remote sensing unlabeled data
RGB
SAR
PAN
Knowledge Transfer Learning
Mask
Original & masked images
$X_{\tilde{M}}$
Encoder
R
Decoder
Y
reshape
Image reconstruction
Step3: Fine-tuning pretrained models on downstream tasks
Patch Embedding
Position embedding
+
[class] token
Pretrained Transformer Encoder
Block 0
Block 1
Block 2
Block 10
Block 11
NL
MHSA
NL
MLP
Remove [class] token
Global Average Pooling
Linear Classifier
Airport
Beach
Center
Port
Parking
⊕ Element-wise Addition
NL: Norm Layer
MHSA: Multi-Head Self-Attention
RPN: Region Proposal Network
TC: Transposed Convolution
MP: Max Pooling
Scene Classification Task
Block 3
Block 5
Block 7
Block 11
TC
TC
TC
Identity
MP
Feature Pyramid Network (FPN)
RPN
Mask RCNN Head
Object Detection Task
Upernet Head
Land Cover Classification Task

Figure 2. The knowledge transfer learning based model training process of our proposed CSPT.

1. For the current knowledge transfer learning strategy from the natural scene domain to the RSD, the severe domain gap is analyzed in detail. In addition, a concise and effective knowledge transfer learning strategy called CSPT is proposed to gradually bridge the domain gap and then efficiently transfer the domain-level knowledge of large-scale unlabeled data such as ImageNet [18] into the specific RSD. Meanwhile, the designed CSPT is a promising method to release the huge potential of unlabeled data for task-aware model training in the RSD.
2. Based on the MIM of task-agnostic representation, the impact of adding extra task-related unlabeled data and waiting for more iterative epochs on further self-supervised pre-training step of the proposed CSPT are studied. Then, we find that the designed CSPT can be a more unified and feasible way to promote the fine-tuning performance of various downstream tasks.
3. Extensive experiments were conducted, which include three downstream tasks (e.g., scene classification, object detection and land cover classification) and two kinds of imaging data (e.g., optical and SAR) in the RSD. The experimental results show that the designed CSPT can mitigate task-aware discrepancy and bridge the domain gap to advance the performance of diverse downstream tasks and reach competitive results in comparison with SOTA methods. Finally, we make the pre-trained model weights freely available at https://github.com/ZhAnGToNG1/transfer_learning_cspt (accessed on 11 August 2022) to the remote sensing community. In addition, the research can also follow the designed CSPT to train their own ViT model weights for specific downstream tasks in other imaging data or application scenarios.

The rest of this paper is organized as follows: Section 2 introduces the related works on knowledge transfer learning and self-supervised pre-training. Section 3 elaborates the problem analysis and newly designed knowledge transfer learning called CSPT in the RSD. Extensive experiments are reported in Section 4 with detailed discussions, and the conclusion is provided in Section 5.

2. Related Work

2.1. Knowledge Transfer Learning

Knowledge transfer learning is a fundamental study in computer vision field, and it is widely used for model training of downstream tasks in the RSD. Related to supervised pre-training based knowledge transfer learning methods, Long, Y. et al. [30] set up a new large scale aerial scene classification benchmark called Million-AID (M-AID), and made use of it to achieve the in-domain large-scale supervised pre-training in the RSD. According to their experimental analysis, fine-tuning CNN models pre-trained on M-AID would perform better than those pre-trained ImageNet [18] for scene classification tasks. In addition, Tong, X.Y. et al. [22] used knowledge transfer learning for improving land cover classification tasks. In detail, first, ResNet-50 [10] is pre-trained by 150,000 image patches of GID [22] to facilitate the generalization ability of pixel-wise land cover classification, and then, based on the pre-trained ResNet-50 [10], they also used 30,000 image patches for model fine-tuning to obtain the highest performance of pixel-wise land cover classification of 15 categories. Next, Li, K. et al. [23] set up an available benchmark for object detection in optical remote sensing images called DIOR, which contained 23,463 images and 192,472 object instances within 20 common object categories. Then, they pre-trained all detectors on large-scale nature scene datesets (e.g., ImageNet [18] and COCO [37]) and then fine-tuned them on DIOR [23] to achieve the knowledge transfer learning of object detection from natural scenes to RSD. In addition, by their experimental results on DIOR [23], they found that knowledge transfer learning also makes the superior detectors in natural scenes obtain competitive results in the RSD without any careful model design. Moreover, Wang, D. et al. [34] provided an empirical study of model pre-training in the RSD, which utilized the self-built optical remote sensing dataset M-AID [30] to pre-train models of CNNs [7,10] and ViTs [15,16] and then fine-tuned them on several downstream tasks to avoid the severe domain gap impact of transferring knowledge from the natural scenes to remote sensing scenes. From the experimental results of [34], they found that ViT-based models [15,16] are promising backbones to provide a stronger feature representation to facilitate downstream tasks in the RSD. Meanwhile, the study of [34] also indicated that utilizing a large-scale dataset, whether belonging to the nature or remote sensing domain, for model pre-training can also promote almost all downstream tasks in the RSD. In general, the previous knowledge transfer learning studies are mainly based on supervised pre-training paradigm, which needs to carefully set up a large-scale labeled dataset for model pre-training and then directly apply the pre-trained model into fine-tuning step. Therefore, the supervised pre-training based knowledge transfer learning method inevitably has the issues of expensive manual labelling, incorrect labeling existed in a large-scale dataset, domain gap and task-aware discrepancy between pre-training and fine-tuning steps. Therefore, in this article, we focus on the self-supervised pre-training based knowledge transfer learning method to break through the inherent paradigm of supervised pre-training based knowledge transfer learning and further improve the performance of various downstream tasks in the RSD.

2.2. Self-Supervised Pre-Training

Self-supervised pre-training is usually used to capture intrinsic patterns and semantic representations from original imaging data. Until now, it began to be used for knowledge transfer learning to facilitate model training and significantly boost performance of various downstream tasks. According the investigation and experimental analysis, the studies of [60,61] have demonstrated that adopting self-supervised pre-training to learn a transferable representation tightly relies on three elements: (1) the amount and domain of pre-training data collection; (2) the pretext task setting; and (3) the network architecture selection. Consequently, Tao, C. et al. [62] carefully designed a sample collection strategy to automatically capture unlabeled samples with class-balanced resampling both in natural and remote sensing scenes, and then employed these samples on the pretext task of contrastive learning to make the different augmented views (i.e., positive sample pairs) of the same images closer and separate views (i.e., negative sample pairs) of different images. Next, for different pretext task settings, Xu, Y. et al. [63] designed a novel unsupervised adversarial contrastive learning method to pre-train a CNN-based Siamese network, which minimized the feature similarity of augmented data and its corresponding unsupervised adversarial samples. Through the designed pretext task, [63] obtained competitive classification results on SAR target recognition datasets. In addition, to use prior information assisting pretext task setting in self-supervised pre-training, Ayush, K. et al. [64] introduced the geography-aware into the pretext task of invariant representation learning, specifically, it makes the positive pairs closer than typical unrelated negative pairs and meanwhile predicts the geo-location information of input images. By experimental analysis, the unlabeled remote sensing images with geo-location prior information would further promote the fine-tuning performance of the self-supervised pre-training method. Moreover, according to self-supervised pre-training of ViT based models, the studies of [53,55,56] demonstrated that the simple pretext task of MIM (i.e., predicting the raw pixels of RGB image values) for self-supervised pre-training can provide a strong transferable ability than previous pretext tasks. In addition, the study of [65] also has demonstrated that an extra branch of MIM-based pretext task in parallel with the existing contrastive learning can facilitate the self-supervised pre-training to mitigate task-aware discrepancy from diverse downstream tasks. Then, following these works that used ViT architecture and MIM-based pretext task in natural scenes, Wang, D. et al. [66] tried to pre-train ViT based models on large-scale remote sensing data (i.e., M-AID [30]) by MAE [53] to propose large vision models customized for remote sensing tasks. Similarly, Zhou, L. et al. [67] utilized MAE [53] for a self-supervised pre-training study on medical image analysis and showed that ViT based models pre-trained by MIM-based pretext task can significantly improve the fine-tuning performance of diverse medical downstream tasks. Thus, referring to the ViT [14] architecture and the MIM of task-agnostic representation, we would like to further explore a more effective and unified self-supervised pre-training strategy for knowledge transfer learning in the RSD.

3. Knowledge Transfer Learning Strategy

In this section, the analysis of domain gap and transfer ability from nature scene to RSD are first illustrated in Section 3.1. Next, in Section 3.2, we introduce the proposed knowledge transfer learning of CSPT in detail. Finally, the mechanism of the MIM-based pretext task is revisited in Section 3.3.

3.1. Problem Analysis

As mentioned in Sections 1 and 2, the domain gap is a troublesome issue in knowledge transfer learning which limits the performance of task-aware model training. Here, MAE [53] is utilized for self-supervised pre-training on the unlabeled nature scene dataset of ImageNet-1K (IN1K) [18] based on the ViT [14] architecture. For intuitive analysis, the pre-trained weights of ViT-B [14] from the unlabeled large-scale nature scene data are adopted to individually generate visualized attention scores from two unseen images, which is used to indicate the existing domain gap between nature and remote sensing scenes. The visualized attention scores were calculated from the last self-attention layer of ViT-B [14] via the query-key product, and the warmer color represents the higher attention scores in self-attention map. In Figure 3a, referring to the selected red rectangle area located in a vehicle of nature scene image, high attention scores reveal that the self-supervised pre-training weights of ViT-B [14] can accurately pay attention on relevant areas of the vehicle. However, when the same pre-trained weights of ViT-B [14] is directly applied to unseen remote sensing image, as shown Figure 3b, some high attention scores are distributed in areas which are irrelevant to the previously selected red rectangle area, and only a few high attention scores focus on relevant areas of vehicles in the remote sensing image. Accordingly, from these visualized results, it can be found that the domain gap does exist between nature and remote sensing scenes, and it would affect the understanding of remote sensing image. Subjectively, in Figure 3a,b, except for the difference in appearance of the vehicle in nature and remote sensing scenes, the context information of the vehicles is also different. For example, the vehicle in natural scene has more fixed context information because wheels are always on the ground and the top of the vehicle is toward the sky; however, these vehicles in remote sensing scenes would have more flexible context information because under the overlooking view, they can appear in any area with complex and various surroundings.

In addition, we further analyze the transfer ability of different pre-training methods, which directly affects the fine-tune process. Specifically, there are three curves of fine-tuning loss presented in Figure 3c,d, where, the x-axis represents the epoch, and the y-axis represents the corresponding fine-tuning loss value. The different color curves represent three different pre-training methods: (1) the blue curve represents the model that is self-supervised pre-trained on unlabeled IN1K [18] called SSP(IN1K); (2) the orange curve indicates the model that is supervised pre-trained on a ready-made IN1K [18] called SP(IN1K); (3) the green curve means the model that is self-supervised pre-trained on unlabeled IN1K [18] and then further self-supervised pre-trained on the training data of AID [19] and NR45 [20] called SSP(IN1K→Train). Notably, the above pre-training methods all adopt ViT-B [14] and fine-tune it on AID [19] and NR45 [20]. From Figure 3c,d, it can be seen that the blue curve of SSP(IN1K) can converge to a lower loss value than the orange curve of SP(IN1K). This indicates that the self-supervised pre-training method can indeed generate more transferable representation than the supervised pre-training method to leverage the fine-tuning step. Next, the further self-supervised pre-training process on the unlabeled task-related data is considered to bridge the domain gap so that it can generate better transfer ability for specific tasks. From the green curves of SSP(IN1K→Train), we can see that whether at the initial or the end of state, the fine-tuning loss values of SSP(IN1K→Train) are both lower than those of SSP(IN1K) and SP(IN1K) on AID [19] and NR45 [20]. Thus, the consecutive pre-training is a promising method to transfer domain-level knowledge into specific tasks in the RSD.

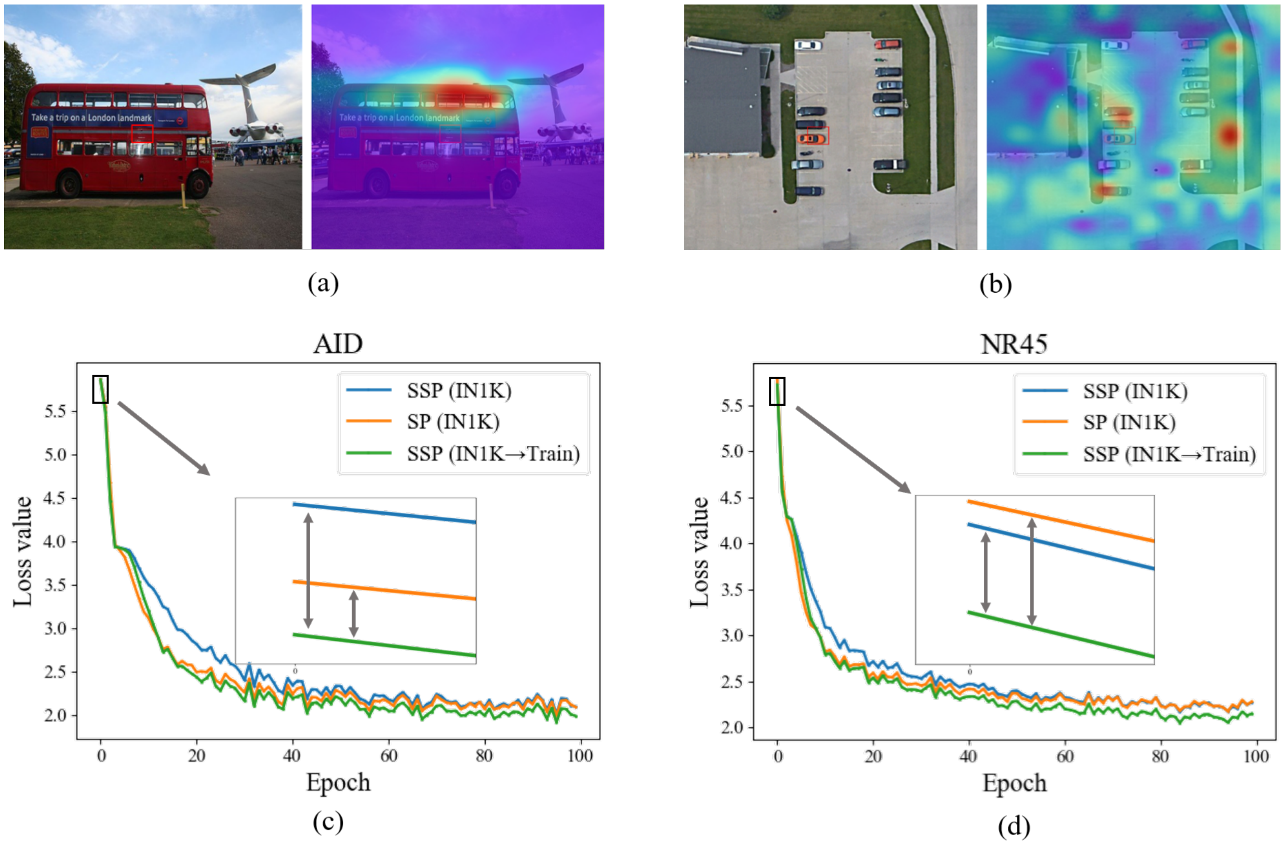


Figure 3. The analysis of domain gap and transfer ability from nature scene to RSD. (**a**) is the self-attention map for a vehicle of natural scene image; (**b**) is the self-attention map for the vehicles of remote sensing scene image; (**c**,**d**) represent the fine-tuning loss convergence curves of different pre-training methods on AID [19] and NR45 [20].

3.2. *Consecutive Pre-Training for Knowledge Transfer Learning*

Taking into account the problem analysis in Section 3.1 and inspired by [58,59] in NLP, a concise and effective self-supervised pre-training based knowledge transfer learning strategy called CSPT is proposed for bridging the domain gap and providing a unified model training paradigm for various downstream tasks in the RSD. As shown in Figure 2, the overall framework is composed of three steps: (1) self-supervised pre-training on a large-scale unlabeled dataset such as ImageNet [18]; (2) further self-supervised pre-training on task-related unlabeled remote sensing data; and (3) fine-tuning on diverse downstream tasks. Here, following the MAE [53], we adopt the MIM-based pretext task for self-supervised pre-training on unlabeled nature and remote sensing data both in steps of (1) and (2). Similarly, the benefits of not stopping pre-training model have been successfully verified in NLP [58]. The pretext task of masked language modeling (MLM) is employed for transformer-based model pre-training via masking and then completing tokens of words in each sentence. This enables the pre-trained language model to learn the vocabulary, sentence structure, semantic and even understand the context of large-scale unlabeled text data. Based on the pretext task of MLM, a generalist language model such as BERT [68] or GPT-3 [69] is firstly pre-trained by large-scale unlabeled text data to set up domain-level knowledge. Then, the generalist model in NLP can make the domain-level knowledge easily adapt to any sub-domain (e.g., biomedical, computer science publications, news, or reviewers) or downstream task by domain adaptive pre-training.

Along with the transformer structure of NLP being migrated into computer vision field, ViT [14] set up vision words (i.e., a group of images patches) from 2-D images, which is suitable for pre-training a model like the method of NLP. If ViT-based model can reconstruct randomly masked partial tokens of vision words, it indicates that the model has learned the pattern relation, structure, semantic and even understood the content of unlabeled images. Thus, as shown in step (1) of Figure 2, when the ViT-based model is pre-trained on a large enough unlabeled dataset such as ImageNet [18] by self-supervised learning, the various combinations of pixel-level pattern relation, structure, semantic and even content can be familiarized to set up domain-level knowledge and prepared for adapting to RSD. Although the domain-level knowledge has a powerful feature representation and transfer ability, directly transferring the domain-level knowledge of the nature scene into the RSD by a fine-tuning step is a non-optimal solution, as discussed in Section 3.1. Obviously, the basic characters (e.g., various combinations of pixel-level pattern relation and structure) of an image are domain-invariant schema. Then, driven by the abundant captured combinations of pattern relations and structures from large-scale nature scene data, further self-supervised pre-training the generalist model on task-related data can become quickly familiar with the semantic and content of image data for specific task, which can be regarded a reasonable transitional stage to bridge the domain gap between upstream and downstream data. Through the consecutive pre-training process, the pre-trained model is considered to fine-tune on downstream tasks, as shown in step (3) of Figure 2. In detail, only the encoder part is applied on the fine-tuning step, because the encoder part can learn the latent representation of images but the decoder part is mainly responsible for reconstructing the pixel value of images. Then, for scene classification task, the pre-trained encoder is employed as feature extractor, and then the extracted feature is fed into global average pooling and linear classifier to complete the classification task. For object detection and land cover classification tasks, firstly, replacing the original backbone of benchmark model by pre-trained encoder, and then the 3rd, 5th, 7th and 11th encoder blocks are adopted to make up the multi-scale feature description of feature pyramid network. Then, the subsequent components keep unchanged. Notably, the designed CSPT also can be considered for in-domain consecutive pre-training to achieve knowledge transfer learning when it is possible to establish domain-level knowledge by collecting a large-scale valid unlabeled data.

3.3. Revisiting Masked Image Modeling

As mentioned above, the MIM of task-agnostic representation is a very effective pretext task for the designed CSPT. It randomly masks partial tokens of vision words and then reconstructs their pixel values according to their corresponding ground truth. As shown in Figure 2, the ViT-based encoder and a lightweight transformer decoder are utilized for MIM pretext task. Next, through randomly masked region reconstruction by the decoder, the encoder can learn a powerful feature representation when the decoder can reconstruct clearer images as shown in Figure 4c–e from masked original images (i.e., in Figure 4b). Based on the encoder-decoder architecture, the whole reconstruction process is similar to the mechanism of gradually understanding remote sensing images by humans. Thus, the encoder-decoder architecture in Figure 2 can reasonably reveal whether the ViT-based model learned the content about input images well. In our study, the process of MIM-based pretext task follows [53], and it can be expressed by the following Equations (1)–(3):

$$R = encoder(X_{\tilde{M}}) \tag{1}$$

$$Y = decoder(R) \tag{2}$$

$$Loss = \frac{1}{\alpha(X_M)} \| Y_M - X_M \|_1 \tag{3}$$

In (1), X represents the image patches that are split from an input image. Then, aiming to capture the latent representation of R, $X_{\tilde{M}}$ can be encoded by $encoder(\cdot)$. Here, M is the index that indicates the masked patches of X. In contrast, $\tilde{M}$ is the index that represents the unmasked patches of X. Next, in (2), the encoded R can be decoded by $decoder(\cdot)$ and produce the reconstructed image Y, where X and $Y \in R^{H \times W \times 3}$. In (3), the L_1-loss is employed to evaluate the similarity of RGB pixel values between X_M and Y_M. Here, $\alpha(X_M)$ is the number of masked pixels. When the value of L_1-loss gradually decreases, the reconstructed results become clearer, which shows that the model has captured the basic pattern relation, structure, semantic and even understood the context of the input images. Subsequently, when the model can well understand the context of images, the ViT-based encoder can be applied for fine-tuning on diverse downstream tasks.

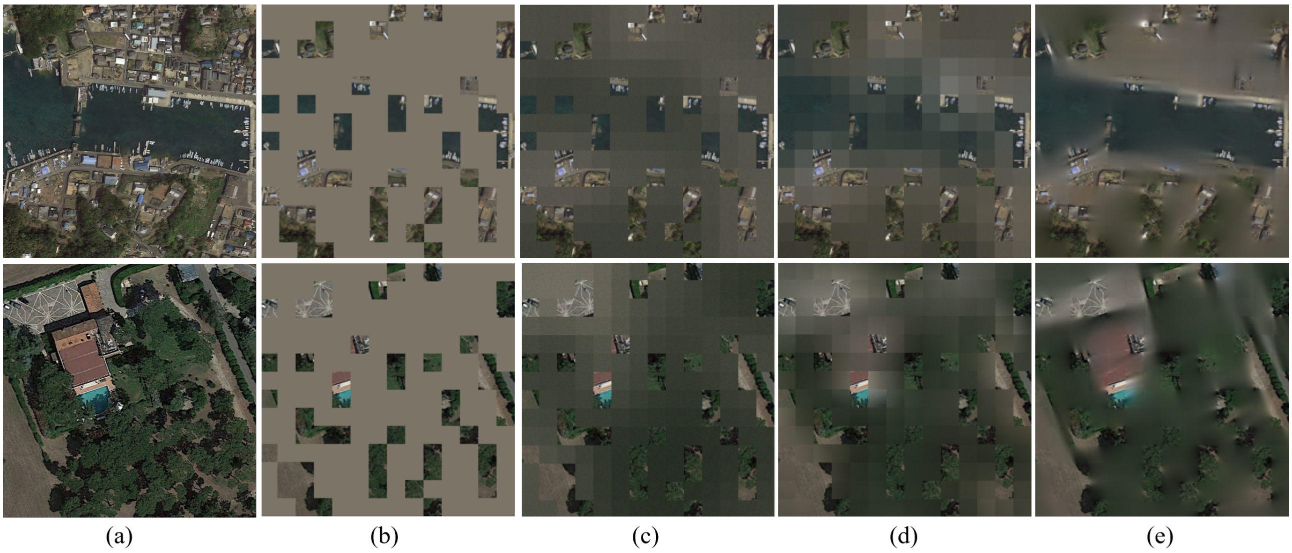


Figure 4. Reconstructed examples on AID [19] images. (**a**) denotes the original images; (**b**) denotes masked images; (**c**) denotes reconstructed results of pre-trained model at the 10th epoch; (**d**) denotes reconstructed results of pre-trained model at the 50th epoch; (**e**) denotes reconstructed results of pre-trained model at the 800th epoch. Obviously, the reconstruction quality from (**c**) to (**e**) becomes better and better.

4. Experiments and Analysis

In this section, extensive experiments were carried out to explore the impact of the designed CSPT on diverse downstream tasks in the RSD. Specifically, through these experiments, we prove three points: (1) Effectiveness: self-supervised pre-training is a more advanced method for knowledge transfer learning in the RSD. And, the further self-supervised pre-training step is a necessary step for consecutive pre-training to gradually bridge the domain gap between nature and remote sensing scenes and promote performance improvement of downstream tasks. (2) Robustness: the MIM of task-agnostic representation is applied for self-supervised pre-training steps can mitigate task-aware discrepancy and perform well on diverse downstream tasks in the RSD. (3) Scalability: when casting off the constraint of carefully manual annotation, how feasible it is to use different quantities of unlabeled data for further self-supervised pre-training, and how data quantity can impact the performance of the fine-tuning step are discussed in detail. In addition, SAR data that has more severe domain gap with nature scene data in the RSD is also introduced for discussing scalability and verifying the knowledge transfer ability of the designed CSPT. Finally, we also compared the results with other advanced model pre-training technologies while some SOTA methods are selected for comparison.

4.1. Datasets Description

To compare different knowledge transfer learning strategies, we adopt twelve public remote sensing datasets including nine optical remote sensing datasets and three SAR remote sensing datasets. These datasets involve three basic downstream tasks: scene classification, object detection and land cover classification. Further details about the dataset split, category number and data quantity are illustrated in Table 1. In particular, because the image sizes of ISPRS POTSDAM and VAIHINGEN [21], GID [22], NWPUVHR-10 [24] and UCAS-AOD [25] are irregular, we cropped them into 512 × 512 pixels. Specifically, the amount of datasets in Table 1 is listed after cropping. Then, it can be found that not only the amount of remote sensing datasets is extremely limited compared with large-scale natural scene datasets but also there exist different kinds of imaging data in the RSD. This allows us to reasonably verify the effect of knowledge transfer learning for downstream tasks with insufficient data.

4.2. Implementation Details

As mentioned in Section 3.2, our proposed CSPT involves three steps: two steps of self-supervised pre-training and one step of model fine-tuning. Next, we would elaborate the pre-training and fine-tuning settings in detail.

4.2.1. Pre-Training Setting

In our work, a ViT based encoder-decoder architecture is adopted to achieve MIM-based pretext task. Specifically, ViT-B/L [14] is chosen as the encoder, and a lightweight transformer model that only has eight self-attention blocks is adopted as the decoder. Then, all input images are regulated into 224 × 224 pixels and divided into non-overlapping 16 × 16 patches (i.e., vision words). Subsequently, to ensure effectiveness, we keep the same pre-training setting as MAE [53] by utilizing 75% tokens randomly masked and then reconstructed by the decoder. In addition, the original pixel values of masked tokens are regarded as supervised information to achieve the self-supervised pre-training. In practice, according to the first self-supervised pre-training step, the ViT-based model is pre-trained on a large-scale unlabeled dataset of IN1K [18] for 800 epochs, which can obtain a generalist model. Second, related to core idea of consecutive pre-training, the generalist model is further pre-trained on task-related unlabeled data to achieve knowledge transfer learning. For experimental details, the batchsize is set to 64 on one RTX 3090. AdamW [70] with momentum $\beta 1 = 0.9$ and $\beta 2 = 0.95$ is employed for optimization. The learning rate schedule adopts cosine decay with a base learning rate of 3.75×10^{-5}. Moreover, input images are augmented by random scale [0.2, 1.0], random crop and random horizontal flip. Then, to simplify the description about pre-training methods in the following experiment analysis, we defined the proposed CSPT as SSP(IN1K→Train) and SSP(IN1K→(Train + Test)), which individually mean that firstly self-supervised pre-training on IN1K [18] and then further pre-training on training set of target datasets and firstly self-supervised pre-training on IN1K [18] and then further pre-training on training and testing set of target datasets. In addition, some other different pre-training methods are selected as comparison, namely, supervised pre-training on IN1K [18], supervised pre-training on M-AID [30], self-supervised pre-training on M-AID [30], self-supervised pre-training on training set of target dataset, supervised pre-training on NR45 [20], self-supervised pre-training on Sentinel-2 [47] and self-supervised pre-training on IN1K [18], individually called SP(IN1K), SP(M-AID), SSP(M-AID), SSP(Train), SP(NR45), SSP(Sentinel-2) and SSP(IN1K).

4.2.2. Fine-Tuning Setting

In the fine-tuning step, we remove the decoder part of encoder-decoder architecture and integrate the encoder part into specific network framework of downstream tasks, as shown in step (3) of Figure 2. Notably, the basic network frameworks with plain design are chosen in our method. Next, the implementation details are elaborated according to different downstream tasks:

Scene classification: To migrate the pre-trained encoder into scene classification task, we firstly extract the latent feature from the encoder output. Then, the class token is removed from the latent feature. Next, we employ the global average pooling on remaining tokens to aggregate the representations, and feed the global representation to a linear classifier. Lastly, the CrossEntropy loss function is used for computing the loss value between the prediction of linear classifier and ground truth. For experimental parameter settings, we train all classification networks for 100 epochs with a batch size of 32. The AdamW [70] ($\beta 1 = 0.9$, $\beta 2 = 0.999$) is employed with an initial learning rate of 5×10^{-4}, and a learning rate schedule follows cosine decay. The input image size is set to 224 × 224 pixels. Augmentation technologies employ AutoAugment (rand-m9-mstd0.5-inc1), label smoothing (0.1), mixup (0.8) and cutmix (1.0). For result comparison, the mean average of Top-1 classification accuracy on the test set is reported.

Object detection: We select Mask-RCNN [71] in mmdetection framework [72] as the benchmark model. To fine-tune the pre-trained encoder network on object detection task, the original backbone network of benchmark model is replaced by the pre-trained encoder network. Then, about the neck network, feature pyramid network (FPN) is often used for fusing multi-scale features in object detection task. Accordingly, to feed multi-scale features into the FPN, the outputs of the 3rd, 5th, 7th and 11th blocks in encoder network are firstly transformed from sequence into 2-D spatial space by reshaping and permuting feature dimensions. Subsequently, the output of the 3rd block is upsampled by a factor of 4 via using two 2 × 2 transposed convolutions with stride = 2. The output of the 5th block is upsampled by a factor of 2 via using a single 2 × 2 transposed convolution with stride = 2. The output of the 7th block remains unchanged. The output of the 11th block is downsampled by a factor of 2 via 2 × 2 max pooling with stride = 2. Then, these four processed features are ready for FPN input. The other components (e.g., Region Proposal Network (RPN) and head network) of benchmark model still use default setting. For training details, the input image size is set as 512 × 512 pixels, and the total number of epochs is set to 12 with a batchsize of 8. Then, momentum = 0.9 and weight decay = 0.0001 are adopted for SGD optimizer, and the initial learning rate is set as 0.02 and then reduced by a factor of 10 times at the 8th and 11th epochs. In addition, random flipping and random resizing are employed for data augmentation. For result comparison, we evaluate the performance by using the mean average precision (mAP@0.5) of the PASCAL VOC object challenge [35].

Land cover classification: We make use of Upernet [73] within the mmsegmentation framework [74] as benchmark model, and the pre-trained encoder network is migrated in the same way as the object detection task. Meanwhile, the input image size is also set to 512 × 512 pixels. Random cropping and random flipping are used for data augmentation. AdamW [70] with momentum $\beta 1 = 0.9$, $\beta 2 = 0.999$ is employed for optimization. We perform fine-tuning for 96K iterations with a batch size of 2. The learning rate is set as 3×10^{-5} with poly scheduler. Finally, the mean Intersection of Union (mIoU) is employed for evaluation of all land cover classification performance.

4.3. Transfer Learning Ability Comparison

In this section, to evaluate the effectiveness of our designed CSPT, different knowledge transfer learning methods of SSP(Train), SSP(M-AID), SP(IN1K) and SSP(IN1K) are employed as comparison and then use the same fine-tuning setting on diverse downstream tasks. Specifically, we consider nine public optical remote sensing datasets (e.g., two scene classification datasets, namely, AID [19] and NR45 [20]; three land cover classification datasets, namely, POTSDAM [21], VAIHINGEN [21] and GID [22]; and four object detection datasets, namely, NWPUVHR10 [24], DIOR [23], UCAS-AOD [25] and HRSC2016 [26]).

Then, several experimental results are reported in Table 2. First, as shown in the 3rd column, compared with other knowledge transfer learning methods, the task-aware model training from scratch (i.e., without knowledge transfer learning) obtains the worst performance on all downstream tasks, which illustrates that insufficient data quantity of downstream tasks cannot support for data-hungry model (i.e., ViT-B [14]) training and knowledge transfer learning is indeed important for task-aware model training in the RSD. Second, as reported in the 4th column of Table 2, it can be found that if only using the training data of target dataset to pre-train the ViT-B [14] model (i.e., SSP(Train)) is hard to obtain the satisfactory performance compared with other model pre-training strategies with a large-scale data. Thus, pre-training on a large-scale data such as ImageNet [18] is very necessary to stimulate the learning potential of ViT-B [14]. Next, as reported in the 6th column of Table 2 (i.e., SP(IN1K)) and compared with SSP(IN1K) and SSP(IN1K→Train) in the 7th and 8th columns of Table 2, when SSP(IN1K) and SSP(IN1K→Train) are applied for model pre-training of ViT-B [14], the performance of almost all downstream tasks is improved than using the SP(IN1K) for model pre-training. In addition, as shown in Figure 5a,b, the convergence speed and accuracy of fine-tuning step according to SP(IN1K), SSP(IN1K) and SSP(IN1K→Train) are also analyzed on AID [19] and NR45 [20]. From Figure 5a,b, the blue line of SSP(IN1K) has a slightly faster convergence speed and higher accuracy for model fine-tuning step than the orange line of SP(IN1K), which also can be found in the 6th and 7th columns of Table 2. Then, related to the designed CSPT strategy, the green line of SSP(IN1K→Train) in Figure 5a,b can perform a prominently faster convergence speed and higher accuracy for model fine-tuning step than SP(IN1K) and SSP(IN1K). This indicated that when getting rid of the labeling constraint, the MIM-based self-supervised pre-training methods (e.g., SSP(IN1K) and SSP(IN1K→Train)) can set up more universal and transferable domain-level knowledge by task-agnostic representation of reconstructing image content so that better promoting downstream tasks. Moreover, our designed CSPT strategy of SSP(IN1K→Train) can be a better way of model pre-training, which can make the model rapidly converge and obtain the best Top-1 accuracy with fewer iterative epochs at fine-tuning step. Third, as discussed in Section 3.1, the existing domain gap between nature and remote sensing scenes would limit the performance of the fine-tuning step even for the model pre-training by SSP(IN1K). Thus, the in-domain self-supervised pre-training method of SSP(M-AID) is considered to be the comparison as reported in the 5th column of Table 2. Here, different from our proposed CSPT strategy (e.g., SSP(IN1K→Train) and SSP(IN1K→(Train + Test)), the SSP(M-AID) directly adopts the in-domain large-scale unlabeled data from M-AID [30] to pre-train the ViT-B [14], and it can avoid the domain gap because the fine-tuning performance of SSP(M-AID) indeed surpasses SSP(IN1K) on most downstream tasks, as shown in the 5th and 7th columns of Table 2. However, we found that though pre-training on an in-domain large-scale dataset can relieve the domain gap, there still exists certain difference between upstream and downstream data to constraint the model fine-tuning performance improvement, especially for different imaging data such as SAR images. Subsequently, as reported in the 8th, 9th, and 10th columns of Table 2, following the proposed CSPT strategy which allows to wait more epochs or add more task-related unlabeled data in the further pre-training step to gradually bridge the domain gap and transfer the domain-level knowledge into RSD, it can be found that the SSP(IN1K→Train) can get better model fine-tuning performance on most downstream tasks than SSP(M-AID). Next, to enlarge the data quantity for further self-supervised pre-training step, we can find that the SSP(IN1K→(Train + Test)) can further boost the model fine-tuning performance and respectively obtain 0.42∼3.73% and 0.62∼7.10% improvements on diverse downstream tasks comparing with SSP(M-AID) and SSP(IN1K), which are presented in the 11th, 12th, and 13th columns of Table 2.

Table 2. The comparison of knowledge transfer learning strategies on optical remote sensing downstream tasks.

Task	Datasets	Transfer Learning Strategies (ViT-B [14])										
		-	Train	M-AID	IN1K		IN1K→Train			IN1K→(Train + Test)		
		From Scratch	Self-Sup.	Self-Sup.	Sup.	Self-Sup.	Self-Sup.			Self-Sup.		
		-	ep800	ep1600	-	ep800	ep800	ep1600	ep2400	ep800	ep1600	ep2400
Scene Classification	NR45 [20]	67.35	88.40	94.69	94.10	93.94	94.23	94.21	94.16	**95.11**	94.90	94.84
	AID [19]	63.15	86.53	96.10	94.04	95.00	95.78	96.05	96.00	96.69	96.69	**96.75**
Land Cover Classification	POTSDAM [21]	60.97	63.32	76.85	76.43	78.08	78.36	78.14	78.09	**78.70**	77.98	78.19
	VAIHINGEN [21]	60.43	59.62	73.82	69.21	71.05	72.34	72.04	72.90	**74.69**	74.19	73.07
	GID [22]	44.70	46.08	60.96	62.64	62.93	**64.97**	62.82	63.58	63.31	64.69	64.55

Table 2. *Cont.*

Task	Datasets	Transfer Learning Strategies (ViT-B [14])										
		-	Train	M-AID	IN1K		IN1K→Train			IN1K→(Train + Test)		
		From Scratch	Self-Sup.	Self-Sup.	Sup.	Self-Sup.	Self-Sup.			Self-Sup.		
		-	ep800	ep1600	-	ep800	ep800	ep1600	ep2400	ep800	ep1600	ep2400
Object Detection	NWPUVHR-10 [24]	54.80	66.50	88.10	68.20	86.00	87.10	87.20	87.50	88.40	88.30	**88.90**
	DIOR [23]	36.90	56.00	68.20	52.70	66.80	68.30	68.20	67.60	**69.80**	69.20	68.50
	UCAS-AOD [25]	49.00	59.40	89.60	83.30	88.70	89.40	90.00	89.30	90.00	90.10	**90.30**
	HRSC2016 [26]	30.00	49.30	86.50	82.60	83.00	89.00	89.40	89.20	89.60	89.90	**90.10**

Note: (1) Evaluation metric: mean average of Top-1 classification accuracy (%) for scene classification; mean Intersection of Union mIoU (%) for land cover classification; mean average precision mAP@0.5 (%) for object detection. (2) The best results are marked in bold.

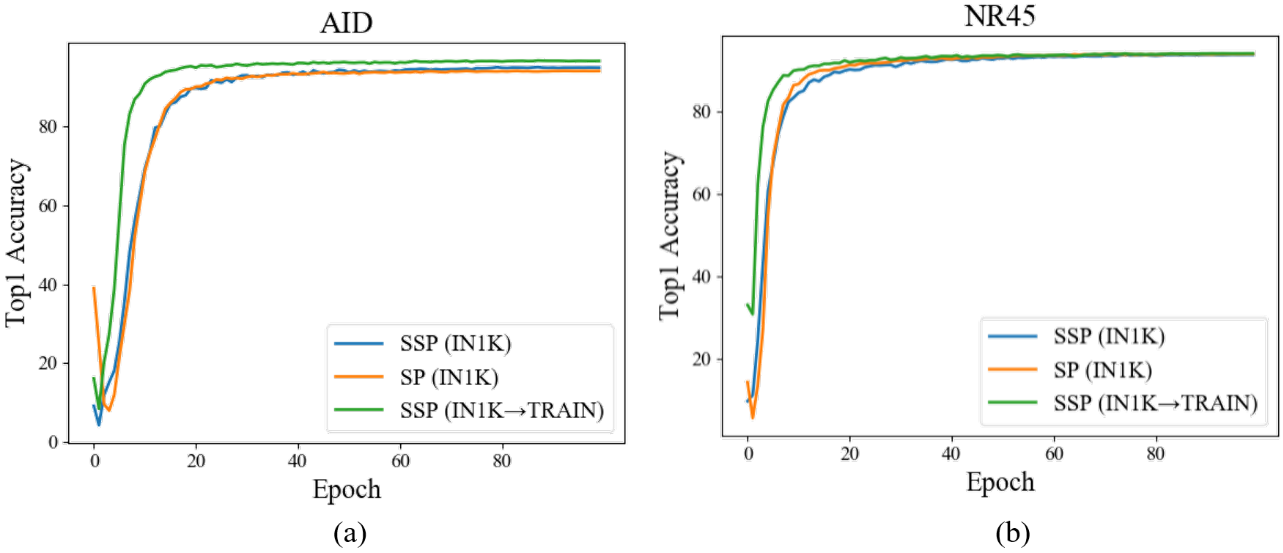


Figure 5. The accuracy curves of three different model pre-training strategies. (**a**) represents the accuracy curves on AID [19]; (**b**) represents the accuracy curves on NR45 [20].

In addition, related to the analysis of domain gap, several visualized attention score maps are shown in Figure 6. Comparing with SSP(M-AID) and SSP(IN1K) as shown in Figure 6c,d,g,h, it can be seen that the designed CSPT in Figure 6b,f can easily attend to any category of imaging data (e.g., optical RGB or SAR images) and correctly identify the selected area (i.e., the red areas with higher attention scores). To sum up, the proposed CSPT strategy is a better method than in-domain self-supervised pre-training (i.e., SSP(M-AID)), and it also can be a uniform model training method based on knowledge transfer learning for RSD.

4.4. Scalability of Data Quantity

According to Section 4.3, the further self-supervised pre-training step of the designed CSPT has been proved to be very effective for knowledge transfer learning. Meanwhile, we also find that in Table 2, the different unlabeled data quantity applied for the further self-supervised pre-training step can affect the model fine-tuning performance. Thus, the scalability of data quantity is discussed in this section. Specifically, two unlabeled data expansion settings are formulated, namely, (1) domain relevant data (DRD) and (2) domain irrelevant data (DID) to objectively analyze the impact of different unlabeled data joining in the further self-supervised pre-training step. In addition, from the view of fine-tuning step, RSD generally possesses insufficient data quantity for fine-tuning on low-resource downstream tasks, as shown in Table 1. Thus, to verify that our proposed CSPT also can adapt to low-resource downstream tasks, we further analyze the impact of fewer labeled samples applied in the fine-tuning step under different pre-training methods and models. Next, we discuss these two views as follows.

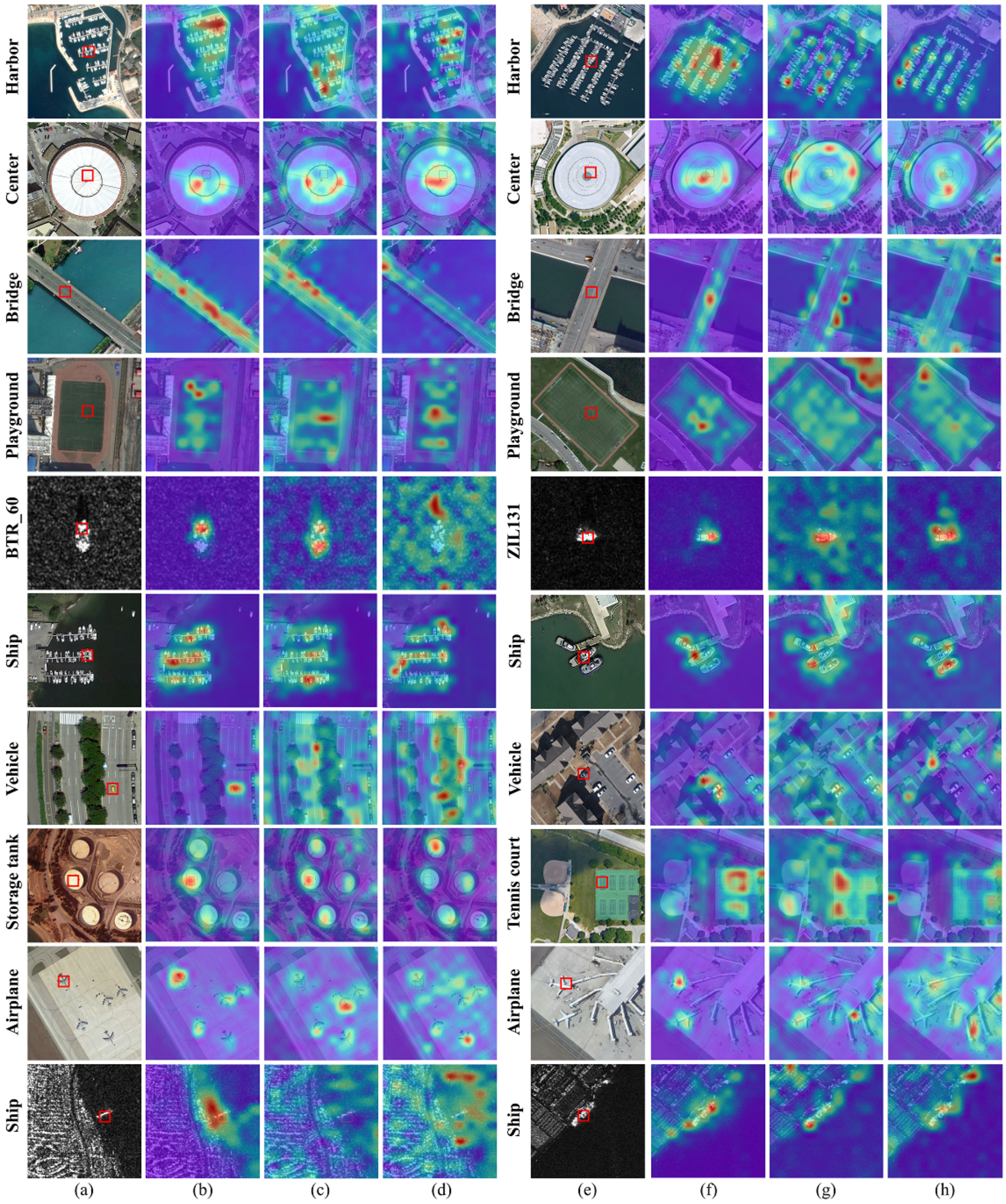


Figure 6. The comparison of self-attention maps. (**a**,**e**) columns represent original images which are from optical and SAR images; (**b**,**f**) columns are score maps produced by our proposed CSPT; (**c**,**g**) columns are score maps produced by self-supervised pre-training on M-AID [30] and then fine-tuning; (**d**,**h**) columns denote score maps produced by self-supervised pre-training on IN1K [18] and then fine-tuning.

4.4.1. Discussion on Further Self-Supervised Pre-Training Step

As shown in Figure 7, to verify the impact of adding more task-related unlabeled data for further self-supervised pre-training step, we set up a large-scale unlabeled dataset of DRD by gathering three dataset images in the RSD, including DOTA [75], DIOR [23] and NR45 [20], which contains 66,593 images and covers various kinds of remote sensing scenes. Next, as the blue and gray bars shown in Figure 7, when further self-supervised pre-training on the combination of an unlabeled training set of target datasets and DRD (i.e., SSP(IN1K→(Train + DRD))) for 800 epochs, these fine-tuning results obtain more performance improvement than only using unlabeled training data of target datasets (i.e., SSP(IN1K→Train)). The reason is that the DRD involves relevant unlabeled data of most downstream tasks, thus a comprehensive data distribution can facilitate the generalist model from IN1K [18] to further learn the general knowledge representation for RSD, which can then easily adapt to diverse downstream tasks.

As for the data expansion of setting (2), the impact of adding DID into further self-supervised pre-training step is discussed. Referring to the training sets of target dataset (e.g., AID [19], NWPUVHR-10 [24] and HRSC2016 [26]), the same amount of unlabeled data from natural scene dataset of Place365 [36] is considered to form DID. Then, the combination of DID and given training sets is applied on the further self-supervised pre-training step (i.e., SSP(IN1K→(Train + DID))) for 800 epochs. As the yellow and gray bars shown in Figure 7, we find that adding more unlabeled data of DID into further self-supervised pre-training step would cause accuracy reductions of 0.53%, 0.1% and 0.3% compared with only using unlabeled training data of target datasets (i.e., SSP(IN1K→Train)) on AID [19], NWPUVHR-10 [24] and HRSC2016 [26], respectively. This further demonstrates the existence of the domain gap between nature and remote sensing scenes would severely affect the performance improvement of knowledge transfer learning. In general, adding more task-related unlabeled data is very important for the further self-supervised pre-training step of the proposed CSPT, which can help the proposed CSPT to release a huge potential of unlabeled data for promoting the model fine-tuning performance in the RSD.

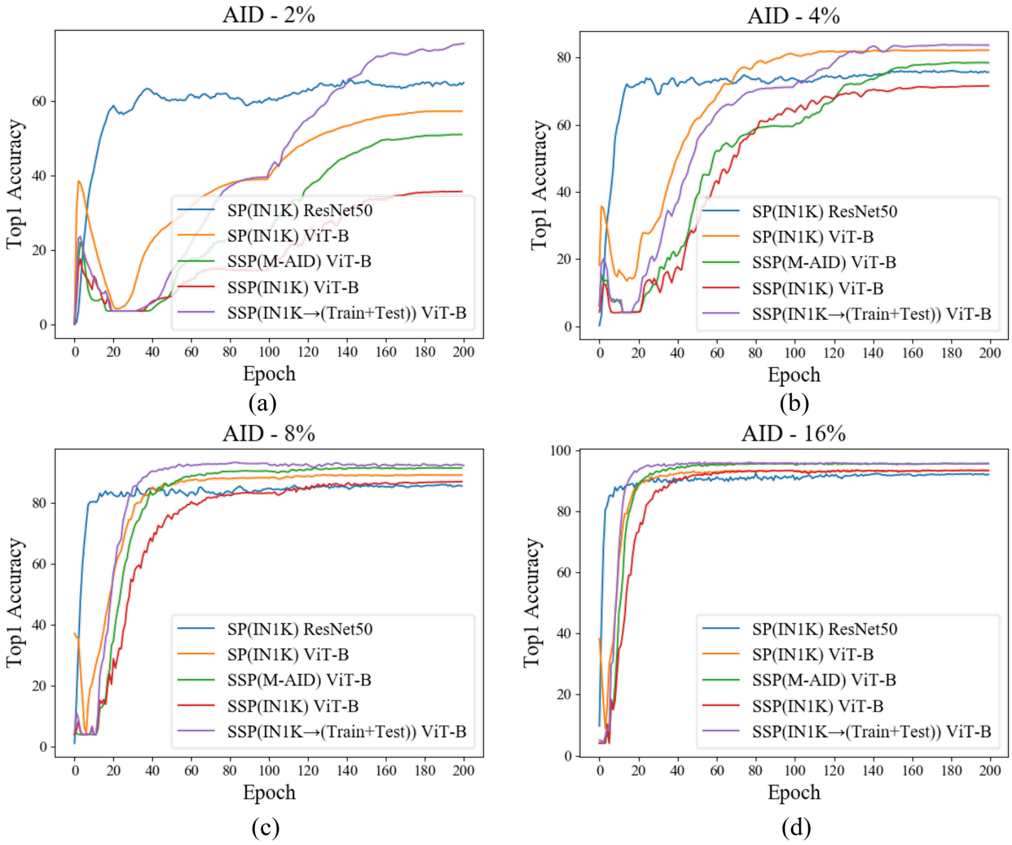


Figure 7. The Top 1 accuracy curves of fine-tuning on training data of AID [19] with different traning set ratios. (**a**) represents the Top 1 accuracy curves under the training set ratio of 2%; (**b**) represents the Top 1 accuracy curves under the training set ratio of 4%; (**c**) represents the Top 1 accuracy curves under the training set ratio of 8%; (**d**) represents the Top 1 accuracy curves under the training set ratio of 16%.

4.4.2. Discussion on Fine-Tuning Step

As reported in Table 3, to explore the scalability of our proposed CSPT on low-resource downstream tasks, different pre-training methods (e.g., train from scratch, SP(IN1K), SSP(M-AID) and SSP(IN1K)) with different network architectures (e.g., ViT-B [14] and ResNet-50 [10]) are selected for comparison. Notably, in our designed CSPT, all images of target datasets (i.e., AID [19] and NR45 [20]) can be regarded as task-related unlabeled data and employed on the further self-supervised pre-training step (i.e., SSP(IN1K→(Train + Test))) for 800 epochs. Then, for fair comparison, we fine-tuned all pre-trained models on the training dataset of AID [19] and NR45 [20] with training set ratios of 2%, 4%, 8%, 16% for 100 epochs. According to Table 3, one observation is that our proposed CSPT produces the best results in all training set ratios of NR45 [20] and 8% and 16% training set ratios of AID [19], but it only obtains suboptimal results on 2% and 4% training set ratios of AID [19]. It is worthy noting that from Table 1, AID [19] has less data quantity than NR45 [20] so that when using 2% or 4% training set ratios of AID [19], there are 5∼9 samples per class used for training in fine-tuning step. Thus, our proposed CSPT for ViT-B [14] does not seem to handle well on fine-tuning with fewer labeled samples. However, as shown in Figure 8, by analyzing the accuracy trends, it can be observed that the blue curve of SP(IN1K) for ResNet-50 [10] reaches the performance bottleneck quickly, but the purple curve of our proposed CSPT for ViT-B [14] has been growing in the default epoch setting (i.e., 100 epochs). Accordingly, the fine-tuning epochs are expanded to study the performance bottleneck of our proposed CSPT for ViT-B [14]. As shown in Figure 8a,b, the blue curves remain at a constant performance level, while the purple curves exceed the blue curves and reach the best performance with the increase in epochs. We conclude that the difference of convergence speed is rooted in network architecture. Because ViT [14] has much less image-specific inductive biases than CNNs, it requires a longer time to learn the relevant patterns from fewer labeled samples. In addition, according to Figure 8a, with the increase in the number of training epochs, the orange, green and red curves of SP(IN1K), SSP(M-AID), SSP(IN1K) for ViT-B [14] are still lower than the blue curve of SP(IN1K) for ResNet-50 [10], differently, the purple curve of our proposed CSPT for ViT-B much exceeds other methods. To sum up, the results demonstrate that our proposed CSPT can advance model fine-tuning performance improvement on low-resource downstream tasks.

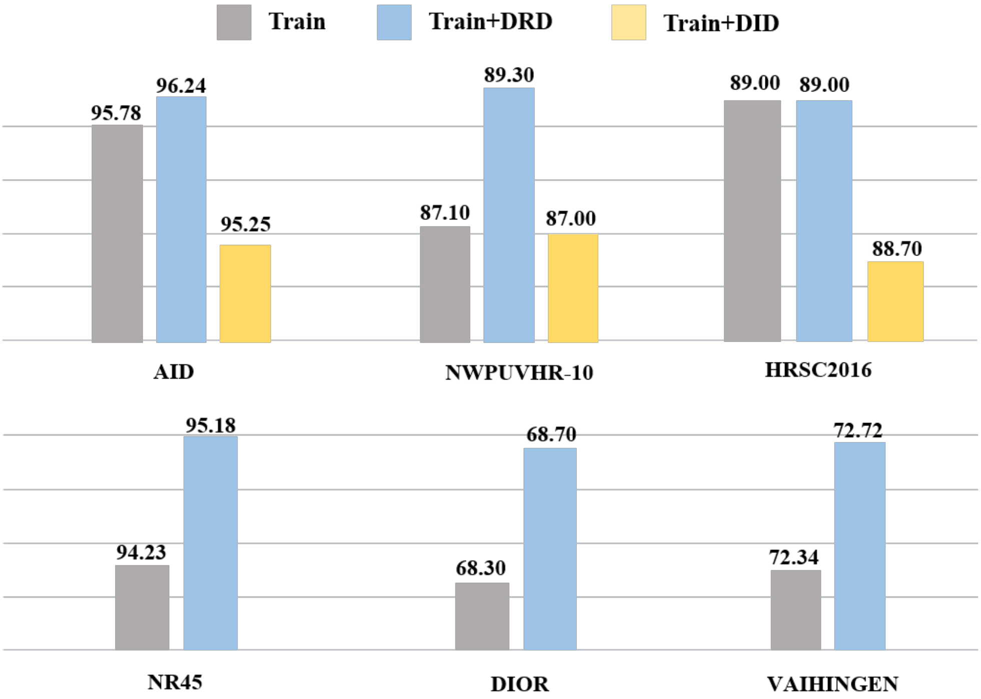


Figure 8. Discussion on different data expansion applied for the further self-supervised pre-training step. Notably, AID [19] and NR45 [20] are reported in Top-1 accuracy; NWPUVHR-10 [24], HRSC2016 [26] and DIOR [23] are reported in mAP@0.5; VAIHINGEN [21] is reported in mIoU.

Table 3. The comparison of fine-tuning on few labeled samples.

Pre-Training Method	**Architecture**	**NR45**				**AID**			
		2%	**4%**	**8%**	**16%**	**2%**	**4%**	**8%**	**16%**
Train from scratch	ResNet-50	27.48	44.30	54.62	63.73	21.80	26.11	46.15	59.78
	ViT-B	27.50	38.73	51.05	66.19	23.17	32.05	43.97	59.65
SP(IN1K)	ResNet-50	73.45	80.92	86.77	91.34	**65.60**	**75.70**	86.10	92.62
	ViT-B	75.80	81.51	90.24	93.08	39.03	66.72	88.19	93.47
SSP(IN1K)	ViT-B	59.36	85.92	88.38	92.58	17.59	42.38	83.29	93.44
SSP(M-AID)	ViT-B	70.77	87.17	91.26	93.51	25.26	59.63	90.45	95.70
SSP(IN1K→(Train + Test))	ViT-B	**80.43**	**89.66**	**92.56**	**94.33**	42.79	73.69	**93.23**	**96.10**

Note: (1) The average of Top-1 classification accuracy (%) is used for reporting above results. (2) The best results are marked in bold. (3) SSP(IN1K→(Train + Test)) represents our designed CSPT.

4.5. Scalability on SAR Imaging Data

It can be seen that remote sensing images possess the characteristics of multi-payload and multi-platform. Apart from optical remote sensing images, SAR satellites are also essential payloads. However, not only the number of available SAR images is limited, but also the SAR images have significant imaging difference with optical images. Consequently, to verify the scalability of our proposed CSPT, we also discuss the performance of the designed CSPT strategy for model training on SAR images. Here, we conduct experiments on one target classification dataset (e.g., MSTAR [27]), and two ship detection datasets (e.g., HRSID [28] and SSDD [29]). Their data descriptions are listed in Table 1. In detail, the unlabeled training data and the combination of unlabeled training and testing data of target datasets are respectively adopted on the further self-supervised pre-training step for 800, 1600 and 2400 epochs. The experimental results are reported in Table 4. Compared with SSP(M-AID) and SSP(IN1K) in the 3rd and 10th columns of Table 4, the designed CSPT of SSP(IN1K→(Train + Test)) achieves the best results, and individually boosts the accuracy by 1.94% and 1.96% Top-1 accuracy on MSTAR [27], 0.9% and 1.2% mAP on SSDD [29] and 1.5% and 1.9% mAP on HRSID [28]. The results show that even on remote sensing data with great imaging difference with nature scene image data, our proposed CSPT can still effectively transfer domain-level knowledge of large-scale nature scene data into diverse downstream tasks. Moreover, the model fine-tuning performance of SSP(M-AID) obtains bad results, even lower than the model fine-tuning performance of SSP(IN1K). This reflects that when there exists significant imaging difference between in-domain large-scale data and specific downstream dataset, it is insufficient to eliminate the domain gap by pre-training on an in-domain large-scale dataset because it cannot be compatible with all imaging data in the RSD. Thus, compared with directly pre-training on an in-domain large-scale dataset, our CSPT strategy has the advantage of flexibility and scalability for downstream tasks with special imaging data.

Table 4. The knowledge transfer ability comparison on SAR images.

Task	**Dataset**	**Self-Supervised Pre-Training**													
		M-AID	**M-AID→Train**			**M-AID→(Train + Test)**			**IN1K**	**IN1K→Train**			**IN1K→(Train + Test)**		
		ep1600	**ep800**	**ep1600**	**ep2400**	**ep800**	**ep1600**	**ep2400**	**ep800**	**ep800**	**ep1600**	**ep2400**	**ep800**	**ep1600**	**ep2400**
Target Classification	MSTAR [27]	98.01	99.67	99.79	99.80	99.93	99.95	99.96	98.03	99.82	99.80	99.80	99.96	99.96	**99.97**
Ship Detection	SSDD [29]	90.60	91.10	91.10	90.90	91.60	91.40	91.20	90.90	91.80	91.50	91.30	91.00	91.70	**91.80**
	HRSID [28]	68.30	68.60	68.90	69.30	68.60	69.60	69.60	68.70	68.90	69.10	69.90	69.00	69.70	**70.20**

Note: (1) Network architecture uses ViT-B [14]. (2) Evaluation metric: the average of Top-1 classification accuracy (%) for target classification; the average precision AP@0.5 (%) for ship detection. (3) The best results are marked in bold.

Based on the conclusion that SSP(M-AID) cannot perform well on downstream tasks of SAR images, we replace the dataset used for the first step of our CSPT strategy with M-AID [30] to further verify whether our CSPT strategy can adapt different large-scale datasets to achieve knowledge transfer learning for promoting performance of downstream tasks. The results are reported in 4th∼9th of Table 4. It can be found that when utilizing large-scale dataset of M-AID [30] to achieve CSPT strategy (i.e., SSP(M-AID→Train) and SSP(M-AID→(Train + Test))), it can obtain 1∼1.95% performance gain compared with SSP(M-AID) in the 3rd column of Table 4 and achieve similar results with SSP(IN1K→Train) and SSP(IN1K→(Train + Test)) in the 11th∼16th of Table 4. The results indicates that our CSPT strategy would not severely depend on the choice of large-scale dataset used for the first step of CSPT strategy. Meanwhile, it also shows further self-supervised pre-training on task-related data is very flexible, which not only can make full use of the domain-level knowledge from different large-scale datasets but also effectively learn the task-related knowledge so that stably bringing performance gain in various downstream tasks.

4.6. Comparison Experiment Analysis

In this section, to demonstrate the potential and superiority of the designed CSPT, the proposed CSPT is applied to train plain networks in the RSD to compare with recently proposed SOTA methods and outstanding model pre-training technologies. In addition, four public remote sensing datasets (i.e., AID [19], NR45 [20], DIOR [23] and ISPRS POTSDAM [21]) are employed as benchmark datasets which involve three downstream tasks (e.g., scene classification, object detection and land cover classification).

4.6.1. Scene Classification

Following [30], the unified data split (i.e., training set: testing set = 2:8) is adopted to evaluate AID [19] and NR45 [20]. In Table 5, different pre-training methods and networks of these comparison methods are listed in the 3rd and 4th columns. Firstly, many well-designed modules have been adopted for improving performance to achieve SOTA results, such as advanced attention mechanisms (e.g., CAD [76], EAM [77], MBLANet [78], ESD-MBENet [79] and MSA-Net [80]) and powerful feature fusion modules (e.g., MG-CAP [81], F2BRBM [82] and KFB-Net [83]). From the experimental results of Table 5, these well-designed models cannot obtain competitive results. Second, except for special module designs, some works also focus on studying powerful pre-training technologies. For example, MoCov3 [84], MAE [53] and SimMIM [55] are designed for self-supervised pre-training on natural scene data. Here, we also transfer these pre-trained models from large-scale nature scene data into the AID [19] and NR45 [20] datasets. From the 11th, 17th and 18th rows of Table 5, it can be seen that MoCov3 [84] based on contrastive learning (CL) obtains slightly worse results than MAE [53] and SimMIM [55]. This is since CL is decision-making based on deep semantic features that would lose much information from original images. In addition, CL needs to carefully construct positive and negative sample pairs; Otherwise, unsuitable positive and negative sample pairs would affect the performance of CL [44,85,86]. In contrast, MIM-based pretext task applied for MAE [53] and SimMIM [55] directly reconstructs the pixels of the original images, which can not only perceive general and detailed image information but also avoid positive and negative sample allocation. Thus, the reconstruction of randomly masked tokens is beneficial for evoking the cognition of underlying knowledge from images. In addition, to avoid the domain gap, some researchers have used self-built large-scale remote sensing data to pre-train their models (e.g., SeCo [47], ASP [30] and RSP [34]). From the 9th, 10th and 16th rows of Table 5, it can be seen that these methods pre-trained on the RSD dataset mostly suppress the above methods pre-trained on nature scene dataset, which illustrates that domain gap would restrict the performance supremum. Although these methods achieve good performance by eliminating domain gap, the expensive collection cost of the self-built unlabeled or labeled large-scale dataset is inevitable. Finally, according to the results of our proposed CSPT as shown in the 21rd and 22th rows, it obtain the SOTA results of 96.75% on

AID [19] and 95.62% on NR45 [20], which indicates that the effectiveness of our proposed CSPT on scene classification task.

Table 5. The comparison results on AID [19] and NR45 [20]. The best results are marked in bold.

Method	Publication	Setting	Network	Top-1 Accuracy (%)	
				AID (2:8)	NR45 (2:8)
MG-CAP [81]	TIP2020	SP(IN1K)	VGG-16	93.34	92.95
CAD [76]	JSTAR2020	SP(IN1K)	DenseNet-121	95.73	94.58
KFBNet [83]	TGRS2020	SP(IN1K)	DenseNet-121	95.50	95.11
F2BRBM [82]	JSTAR2021	SP(IN1K)	ResNet-50	96.05	94.87
MBLANet [78]	TIP2021	SP(IN1K)	ResNet-50	95.60	94.66
EAM [77]	GRSL2021	SP(IN1K)	ResNet-101	94.26	94.29
MSA-Net [80]	JSTAR2021	SP(IN1K)	ResNet-101	93.53	93.52
ESD-MBENet [79]	TGRS2021	SP(IN1K)	DenseNet-121	96.39	95.36
ASP [30]	arXiv2022	SP(M-AID)	ResNet-101	95.40	94.20
SeCo [47]	ICCV2021	SSP(Sentinel-2)	ResNet-50	93.47	92.91
MoCov3 [84]	ICCV2021	SSP(IN1K)	ResNet-50	92.51	91.79
Swin Transformer [15]	ICCV2021	SP(IN1K)	Swin-T	96.55	94.70
Vision Transformer [14]	ICLR2021	SP(IN1K)	ViT-B	94.04	94.10
CTNet [87]	GRSL2021	SP(IN1K)	MobileNet-v2+ViT-B	96.25	95.40
ViTAEv2 [16]	arXiv2022	SP(IN1K)	ViTAEv2-S	96.61	95.29
RSP [34]	arXiv2022	SP(M-AID)	ViTAEv2-S-E40	96.72	95.35
SimMIM [55]	CVPR2022	SSP(IN1K)	ViT-B	93.08	92.57
MAE [53]	CVPR2022	SSP(IN1K)	ViT-B	95.00	93.94
MAE [53]	CVPR2022	SSP(IN1K)	ViT-L	94.92	94.34
CSPT	-	SSP(IN1K→(Train + DRD))	ViT-B	96.24	95.18
CSPT	-	SSP(IN1K→(Train + Test))	ViT-B	**96.75**	95.11
CSPT	-	SSP(IN1K→(Train + Test))	ViT-L	96.30	**95.62**

4.6.2. Object Detection

Object detection belongs to the dense prediction task, which has more complex network structure, including backbone, neck and head networks. Here, we only attend to the replacement of backbone network that is pre-trained from our CSPT or other pre-training strategies. As shown in Section 4.2.2, the Mask-RCNN [71] is selected as the benchmark model. In addition, the public remote sensing dataset DIOR [23] is adopted to evaluate the performance. It contains 23,463 images with 192,472 instances involving 20 object categories such as airplane (AL), airport (AT), baseball field (BF), basketball court (BC), bridge (B), chimney (C), dam (D), expressway service area (ESA), expressway toll station (ETS), golf course (GC), ground track field (GTF), harbour (HB), overpass (O), ship (S), stadium (SD), storage tank (ST), tennis court (TC), train station (TS), vehicle (V), and windmill (W). Following to the data split rule of DIOR [23], we compared our strategy with other SOTA detectors and advanced pre-training technologies. Among these methods, some classical detectors, such as Faster-RCNN [88], YOLOv5 [89], Mask-RCNN [71], PANet [90] and CenterNet [91] are selected. Meanwhile, we also compare with some SOTA detectors from RSD including MSFC-Net [92], CANet [93] and FSoD [94]. As reported in the 9th and 14th rows of Table 6, compared with the original version of Mask-RCNN [71], our method achieves 3% mAP improvement without bells and whistles based on ViT-B [14]. In addition, when using ViT-L [14] as backbone, we obtain 71.7% mAP in plain Mask-RCNN detector, which is competitive with the SOTA detectors and even exceeds some of them. Moreover, some self-supervised pre-training methods (e.g., MAE [53], MoCov3 [84] and SimMIM [55]) are compared, as shown in the 9th to 12th rows of Table 6. It can be observed that the proposed CSPT still achieves the best performance. Through above analysis, it can be demonstrated that the simple further self-supervised pre-training on task-related unlabeled data can effectively avoid both large-scale data annotation and bridge the domain gap, and then bring significant performance gain in remote sensing object detection task.

Table 6. The comparison results on DIOR [23]. The best results are marked in bold.

Method	Setting	Backbone	mAP (%)	AL	AT	BF	BC	B	C	D	ESA	ETS	GC	GTF	HB	O	S	SD	ST	TC	TS	V	WM
Faster-RCNN [88]	SP(IN1K)	ResNet-101	53.6	51.3	61.6	62.2	80.6	26.9	74.2	37.3	53.4	45.1	69.6	61.8	43.7	48.9	56.1	41.8	39.5	73.8	44.7	33.9	65.3
Mask-RCNN [71]	SP(IN1K)	ResNet-101	65.2	53.9	76.6	63.2	80.9	40.2	72.5	60.4	76.3	62.5	76.0	75.9	46.5	57.4	71.8	68.3	53.7	81.0	62.3	53.0	81.0
YOLOv5 [89]	SP(IN1K)	CSPdarknet-53	68.5	87.3	61.7	73.7	90.0	42.6	77.5	55.2	63.8	63.2	66.9	78.0	58.1	58.1	87.8	54.3	**79.3**	89.7	50.2	53.9	79.6
CenterNet [91]	SP(IN1K)	DLA-34	63.2	78.6	56.5	76.1	88.1	33.2	77.1	41.0	47.4	55.5	71.4	72.5	23.0	52.7	**89.8**	54.0	78.6	86.2	46.1	**57.8**	77.4
CANet [93]	SP(IN1K)	ResNet-101	**74.3**	70.3	82.4	72.0	87.8	**55.7**	79.9	67.7	83.5	**77.2**	77.3	**83.6**	56.0	**63.6**	81.0	79.8	70.8	88.2	67.6	51.2	89.6
PANet [90]	SP(IN1K)	ResNet-101	66.1	60.2	72.0	70.6	80.5	43.6	72.3	61.4	72.1	66.7	72.0	73.4	45.3	56.9	71.7	70.4	62.0	80.9	57.0	47.2	84.5
MSFC-Net [92]	SP(IN1K)	ResNeSt-101	70.1	85.8	76.2	74.3	90.1	44.1	78.1	55.5	60.9	59.5	76.9	73.6	49.5	57.2	89.6	69.2	76.5	86.7	51.8	55.2	84.3
FSoD [94]	SP(NR45)	MSE-Net	71.8	**88.9**	66.9	**86.8**	**90.2**	45.5	79.6	48.2	86.9	75.5	67.0	77.3	53.6	59.7	78.3	69.9	75.0	**91.4**	52.3	52.0	**90.6**
Mask-RCNN(MAE) [53]	SSP(IN1K)	ViT-B	66.8	58.9	85.6	69.4	80.7	37.8	78.5	70.2	85.0	55.4	80.7	77.4	58.7	57.1	44.3	79.2	44.3	83.1	70.9	27.5	74.8
Mask-RCNN(MAE) [53]	SSP(IN1K)	ViT-L	68.3	66.1	86.5	73.3	83.6	41.4	81.6	72.2	86.2	58.3	79.2	78.7	60.3	61.1	60.1	73.4	42.1	83.3	71.3	28.9	78.7
Mask-RCNN(MoCov3) [84]	SSP(IN1K)	ResNet-50	62.5	57.9	75.1	65.1	85.3	36.2	71.9	59.2	66.4	51.6	74.0	75.8	58.8	54.8	67.8	67.8	44.2	83.0	58.4	27.6	76.6
Mask-RCNN(SimMIM) [55]	SSP(IN1K)	ResNet-50	63.5	59.6	80.4	69.7	77.0	34.5	77.5	64.9	77.6	52.4	76.8	74.4	52.0	55.5	59.6	70.8	40.5	80.2	64.4	27.1	75.0
Mask-RCNN(CSPT)	SSP(IN1K→(Train + DRD))	ViT-B	68.7	69.9	87.7	70.8	81.2	41.6	80.5	74.8	86.0	58.8	78.9	75.6	60.6	58.9	60.8	78.3	44.6	84.1	76.2	29.0	76.4
Mask-RCNN(CSPT)	SSP(IN1K→(Train + Test))	ViT-B	69.8	69.8	89.1	74.7	82.6	42.2	80.5	**76.9**	86.4	58.8	80.7	77.7	**61.9**	60.2	60.9	79.2	46.1	84.3	77.2	29.0	77.3
Mask-RCNN(CSPT)	SSP(IN1K→(Train + Test))	ViT-L	71.7	74.1	**89.9**	81.2	86.2	44.5	**81.9**	74.8	**90.1**	61.3	**81.9**	79.6	61.6	61.0	61.0	**83.7**	44.5	88.1	**78.9**	29.2	79.9

4.6.3. Land Cover Classification

For the land cover classification task, we adopt Upernet [73] as the benchmark model and then replace its backbone with the model pre-trained by our proposed CSPT. Then, the ISPRS POTSDAM [21] is selected as our benchmark dataset to evaluate our proposed CSPT. As shown in Table 1, there are six categories of land cover (e.g., impervious surface, building, low vegetation, tree, car and clutter) used for evaluating the performance (i.e., mIoU) of the models. Next, some advanced methods are selected for comparison such as Deeplabv3+ [95], GCNet [96] and BES-Net [97] from nature and remote sensing scenes. As reported in the 5th and 7th rows of Table 7, based on ViT-B [14], our proposed CSPT brings 2.27% mIoU gain compared with employing SP(IN1K) on Upernet [73]. Moreover, our result is very competitive with other SOTA methods.

Table 7. The comparison results on ISPRS POTSDAM [21]. The best results are marked in bold.

Method	Setting	Backbone	mIoU(%)
BES-Net [97]	SP(IN1K)	ResNet-18	78.21
Deeplabv3+ [95]	SP(IN1K)	ResNet-50	75.21
Upernet [73]	SP(IN1K)	ResNet-50	75.86
GCNet [96]	SP(IN1K)	ResNet-101	75.38
Upernet [73]	SP(IN1K)	ViT-B	76.43
Upernet(CSPT)	SSP(IN1K→Train)	ViT-B	78.36
Upernet(CSPT)	SSP(IN1K→(Train + Test))	ViT-B	**78.70**

5. Conclusions

In this paper, first, we provided an empirical analysis of knowledge transfer learning and illustrated some limitations and problems of traditional knowledge transfer learning. Second, a concise and effective knowledge transfer learning strategy called CSPT based on the idea of not stopping pre-training in NLP is proposed to gradually narrow the domain gap and better transfer domain-level knowledge from the natural scene domain to the RSD. The further self-supervised pre-training step adopted in CSPT can release the potential of unlabeled data for model pre-training and then facilitate fine-tuning step. In addition, the MIM-based pretext task of task-agnostic representation that is utilized for self-supervised pre-training can mitigate the task-aware discrepancy from diverse downstream tasks. Finally, through extensive experiments, our proposed CSPT has been shown to bring significant performance improvements on various downstream tasks in the RSD. Meanwhile, the comparison also shows that the designed CSPT can achieve the competitive results compared with SOTA methods on diverse downstream tasks without bells and whistles. In the future, we plan to explore more reasonable knowledge transfer learning strategies for specific downstream tasks.

Author Contributions: Conceptualization, Y.Z., P.G. and H.D.; methodology, Y.Z., P.G. and T.Z.; software and validation, T.Z., G.W. and W.Z.; formal analysis, Y.Z., P.G., and T.Z.; investigation, Y.Z., P.G., and T.Z.; resources, H.C., H.D., Y.Z. and P.G.; writing—original draft preparation, Y.Z. and T.Z.; writing—review and editing, Y.Z., P.G., H.D. and T.Z.; visualization, T.Z.; supervision, H.C., Y.Z., P.G. and H.D.; funding acquisition, H.C., Y.Z. and H.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chang Jiang Scholars Program under grant T2012122 and in part by the Civil Aviation Program under grant B0201, the Space-based on orbit real-time processing technology program under grant 2018-JCJQ-ZQ-046, the Low energy consumption hardware accelerated computing technology for multi-source satellite data under grant 2021YFA0715204, and in part by the National Science Foundation for Young Scientists of China under grant 62101046 as well as by the National Natural Science Foundation of China (No. 62136001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Liu, Y.; Li, Z.; Wei, B.; Li, X.; Fu, B. Seismic vulnerability assessment at urban scale using data mining and GIScience technology: Application to Urumqi (China). *Geomat. Nat. Hazards Risk* **2019**, *10*, 958–985. [CrossRef]
2. Rathore, M.M.; Ahmad, A.; Paul, A.; Rho, S. Urban planning and building smart cities based on the Internet of Things using Big Data analytics. *Comput. Netw. Int. J. Comput. Telecommun. Netw.* **2016**, *101*, 63–80. [CrossRef]
3. Ozdarici-Ok, A.; Ok, A.O.; Schindler, K. Mapping of Agricultural Crops from Single High-Resolution Multispectral Images—Data-Driven Smoothing vs. Parcel-Based Smoothing. *Remote Sens.* **2015**, *7*, 5611. [CrossRef]
4. Sadgrove, E.J.; Falzon, G.; Miron, D.; Lamb, D.W. Real-time object detection in agricultural/remote environments using the multiple-expert colour feature extreme learning machine (MEC-ELM). *Comput. Ind.* **2018**, *98*, 183–191. [CrossRef]
5. Reilly, V.; Idrees, H.; Shah, M. Detection and tracking of large number of targets in wide area surveillance. In *Computer Vision—ECCV 2010. ECCV 2010*; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6313, pp. 186–199.
6. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
7. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
8. 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 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
9. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
11. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
12. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. Resnest: Split-attention networks. *arXiv* **2020**, arXiv:2004.08955.
13. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
15. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
16. Zhang, Q.; Xu, Y.; Zhang, J.; Tao, D. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *arXiv* **2022**, arXiv:2202.10108.

17. Gao, P.; Lu, J.; Li, H.; Mottaghi, R.; Kembhavi, A. Container: Context aggregation network. *arXiv* **2021**, arXiv:2106.01401.
18. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis. (IJCV)* **2015**, *115*, 211–252. [CrossRef]
19. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
20. Gong, C.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883.
21. Rottensteiner, F.; Sohn, G.; Gerke, M.; Wegner, J.D. ISPRS semantic labeling contest. *ISPRS Leopoldshöhe Ger.* **2014**, *1*, 4.
22. Tong, X.Y.; Xia, G.S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens. Environ.* **2020**, *237*, 111322. [CrossRef]
23. 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]
24. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
25. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In Proceedings of the IEEE International Conference on Image Processing, Quebec City, QC, Canada, 27–30 September 2015; pp. 3735–3739.
26. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International Conference on Pattern Recognition Applications and Methods*; SciTePress: Porto, Portugal, 2017; Volume 2, pp. 324–331.
27. Ross, T.D.; Worrell, S.W.; Velten, V.J.; Mossing, J.C.; Bryant, M.L. Standard SAR ATR evaluation experiments using the MSTAR public release data set. In *Algorithms for Synthetic Aperture Radar Imagery*; International Society for Optics and Photonics: Bellingham, WA, USA, 1998; Volume 3370, pp. 566–573.
28. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation. *IEEE Access* **2020**, *8*, 120234–120254. [CrossRef]
29. Li, J.; Qu, C.; Shao, J. Ship detection in SAR images based on an improved faster R-CNN. In Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), Beijing, China, 13–14 November 2017; pp. 1–6.
30. Long, Y.; Xia, G.S.; Zhang, L.; Cheng, G.; Li, D. Aerial Scene Parsing: From Tile-level Scene Classification to Pixel-wise Semantic Labeling. *arXiv* **2022**, arXiv:2201.01953.
31. Ranjan, P.; Patil, S.; Ansari, R.A. Building Footprint Extraction from Aerial Images using Multiresolution Analysis Based Transfer Learning. In Proceedings of the 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 10–13 December 2020; pp. 1–6.
32. Liu, Y.; Fan, B.; Wang, L.; Bai, J.; Xiang, S.; Pan, C. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 78–95. [CrossRef]
33. Chen, Z.; Zhang, T.; Ouyang, C. End-to-end airplane detection using transfer learning in remote sensing images. *Remote Sens.* **2018**, *10*, 139. [CrossRef]
34. Wang, D.; Zhang, J.; Du, B.; Xia, G.S.; Tao, D. An Empirical Study of Remote Sensing Pre-Training. *arXiv* **2022**, arXiv:2204.02825.
35. Everingham, M.; Eslami, S.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
36. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1452–1464. [CrossRef] [PubMed]
37. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision–ECCV 2014. ECCV 2014*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2014; pp. 740–755.
38. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–21 June 2021; pp. 8748–8763.
39. Chakraborty, S.; Uzkent, B.; Ayush, K.; Tanmay, K.; Sheehan, E.; Ermon, S. Efficient conditional pre-training for transfer learning. *arXiv* **2020**, arXiv:2011.10231.
40. Ericsson, L.; Gouk, H.; Hospedales, T.M. How well do self-supervised models transfer? In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5414–5423.
41. Kotar, K.; Ilharco, G.; Schmidt, L.; Ehsani, K.; Mottaghi, R. Contrasting contrastive self-supervised representation learning pipelines. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9949–9959.
42. Asano, Y.M.; Rupprecht, C.; Zisserman, A.; Vedaldi, A. PASS: An ImageNet replacement for self-supervised pre-training without humans. *arXiv* **2021**, arXiv:2109.13228.
43. Stojnic, V.; Risojevic, V. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1182–1191.
44. Li, H.; Li, Y.; Zhang, G.; Liu, R.; Huang, H.; Zhu, Q.; Tao, C. Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]

45. Tao, C.; Qi, J.; Lu, W.; Wang, H.; Li, H. Remote sensing image scene classification with self-supervised paradigm under limited labeled samples. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 1–5. [CrossRef]
46. Li, W.; Chen, K.; Chen, H.; Shi, Z. Geographical Knowledge-Driven Representation Learning for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
47. Manas, O.; Lacoste, A.; Giro-i Nieto, X.; Vazquez, D.; Rodriguez, P. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9414–9423.
48. Reed, C.J.; Yue, X.; Nrusimha, A.; Ebrahimi, S.; Vijaykumar, V.; Mao, R.; Li, B.; Zhang, S.; Guillory, D.; Metzger, S.; et al. Self-supervised pre-training improves self-supervised pre-training. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 2584–2594.
49. Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; Hinton, G.E. Big self-supervised models are strong semi-supervised learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22243–22255.
50. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
51. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
52. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9912–9924.
53. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. *arXiv* **2021**, arXiv:2111.06377.
54. Bao, H.; Dong, L.; Wei, F. Beit: Bert pre-training of image transformers. *arXiv* **2021**, arXiv:2106.08254.
55. Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. Simmim: A simple framework for masked image modeling. *arXiv* **2021**, arXiv:2111.09886.
56. Gao, P.; Ma, T.; Li, H.; Dai, J.; Qiao, Y. ConvMAE: Masked Convolution Meets Masked Autoencoders. *arXiv* **2022**, arXiv:2205.03892.
57. Zhang, R.; Guo, Z.; Gao, P.; Fang, R.; Zhao, B.; Wang, D.; Qiao, Y.; Li, H. Point-M2AE: Multi-scale Masked Autoencoders for Hierarchical Point Cloud Pre-training. *arXiv* **2022**, arXiv:2205.14401.
58. Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; Smith, N.A. Don't stop pre-training: Adapt language models to domains and tasks. *arXiv* **2020**, arXiv:2004.10964.
59. Dery, L.M.; Michel, P.; Talwalkar, A.; Neubig, G. Should we be pre-training? an argument for end-task aware training as an alternative. *arXiv* **2021**, arXiv:2109.07437.
60. Anand, M.; Garg, A. Recent advancements in self-supervised paradigms for visual feature representation. *arXiv* **2021**, arXiv:2111.02042.
61. Ericsson, L.; Gouk, H.; Loy, C.C.; Hospedales, T.M. Self-Supervised Representation Learning: Introduction, Advances and Challenges. *arXiv* **2021**, arXiv:2110.09327.
62. Tao, C.; Qia, J.; Zhang, G.; Zhu, Q.; Lu, W.; Li, H. TOV: The Original Vision Model for Optical Remote Sensing Image Understanding via Self-supervised Learning. *arXiv* **2022**, arXiv:2204.04716.
63. Xu, Y.; Sun, H.; Chen, J.; Lei, L.; Ji, K.; Kuang, G. Adversarial Self-Supervised Learning for Robust SAR Target Recognition. *Remote Sens.* **2021**, *13*, 4158. [CrossRef]
64. Ayush, K.; Uzkent, B.; Meng, C.; Tanmay, K.; Burke, M.; Lobell, D.; Ermon, S. Geography-aware self-supervised learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10181–10190.
65. Wang, L.; Liang, F.; Li, Y.; Ouyang, W.; Zhang, H.; Shao, J. RePre: Improving Self-Supervised Vision Transformer with Reconstructive Pre-training. *arXiv* **2022**, arXiv:2201.06857.
66. Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Advancing Plain Vision Transformer Towards Remote Sensing Foundation Model. *arXiv* **2022**, arXiv:2208.03987.
67. Zhou, L.; Liu, H.; Bae, J.; He, J.; Samaras, D.; Prasanna, P. Self Pre-training with Masked Autoencoders for Medical Image Analysis. *arXiv* **2022**, arXiv:2203.05573.
68. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
69. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.141655.
70. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
71. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
72. 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.
73. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.

74. Contributors, M. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. 2020. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 22 January 2022).
75. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
76. Tong, W.; Chen, W.; Han, W.; Li, X.; Wang, L. Channel-Attention-Based DenseNet Network for Remote Sensing Image Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4121–4132. [CrossRef]
77. Zhao, Z.; Li, J.; Luo, Z.; Li, J.; Chen, C. Remote sensing image scene classification based on an enhanced attention module. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1926–1930. [CrossRef]
78. Chen, S.B.; Wei, Q.S.; Wang, W.Z.; Tang, J.; Luo, B.; Wang, Z.Y. Remote Sensing Scene Classification via Multi-Branch Local Attention Network. *IEEE Trans. Image Process.* **2021**, *31*, 99–109. [CrossRef]
79. Zhao, Q.; Ma, Y.; Lyu, S.; Chen, L. Embedded Self-Distillation in Compact Multi-Branch Ensemble Network for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15.
80. Zhang, G.; Xu, W.; Zhao, W.; Huang, C.; Yk, E.N.; Chen, Y.; Su, J. A Multiscale Attention Network for Remote Sensing Scene Images Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 9530–9545. [CrossRef]
81. Wang, S.; Guan, Y.; Shao, L. Multi-Granularity Canonical Appearance Pooling for Remote Sensing Scene Classification. *IEEE Trans. Image Process.* **2020**, *29*, 5396–5407. [CrossRef] [PubMed]
82. Zhang, X.; An, W.; Sun, J.; Wu, H.; Zhang, W.; Du, Y. Best representation branch model for remote sensing image scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 9768–9780. [CrossRef]
83. Li, F.; Feng, R.; Han, W.; Wang, L. High-Resolution Remote Sensing Image Scene Classification via Key Filter Bank Based on Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8077–8092. [CrossRef]
84. Chen, X.; Xie, S.; He, K. An empirical study of training self-supervised visual transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, QC, Canada, 11–17 October 2021; pp. 9640–9649.
85. Guo, Y.; Xu, M.; Li, J.; Ni, B.; Zhu, X.; Sun, Z.; Xu, Y. HCSC: Hierarchical Contrastive Selective Coding. *arXiv* **2022**, arXiv:2202.00455.
86. Peng, X.; Wang, K.; Zhu, Z.; You, Y. Crafting Better Contrastive Views for Siamese Representation Learning. *arXiv* **2022**, arXiv:2202.03278.
87. Deng, P.; Xu, K.; Huang, H. When CNNs meet vision transformer: A joint framework for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
88. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
89. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; NanoCode012.; Kwon, Y.; TaoXie.; Michael, K.; Fang, J.; imyhxy.; et al. ultralytics/yolov5: V6.2—YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai Integrations. 2022. Available online: https://doi.org/10.5281/zenodo.7002879 (accessed on 17 August 2022).
90. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
91. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
92. Zhang, T.; Zhuang, Y.; Wang, G.; Dong, S.; Chen, H.; Li, L. Multiscale Semantic Fusion-Guided Fractal Convolutional Object Detection Network for Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–20. [CrossRef]
93. Li, Y.; Huang, Q.; Pei, X.; Chen, Y.; Jiao, L.; Shang, R. Cross-layer attention network for small object detection in remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 2148–2161. [CrossRef]
94. Wang, G.; Zhuang, Y.; Chen, H.; Liu, X.; Zhang, T.; Li, L.; Dong, S.; Sang, Q. FSoD-Net: Full-scale object detection from optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–18. [CrossRef]
95. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
96. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October 2019–2 November 2019.
97. Chen, F.; Liu, H.; Zeng, Z.; Zhou, X.; Tan, X. BES-Net: Boundary Enhancing Semantic Context Network for High-Resolution Image Semantic Segmentation. *Remote Sens.* **2022**, *14*, 1638. [CrossRef]