*remote sensing*

MDPI

*Technical Note*

# Semi-Supervised Remote Sensing Image Semantic Segmentation via Consistency Regularization and Average Update of Pseudo-Label

**Jiaxin Wang [1], Chris H. Q. Ding [2], Sibao Chen [1,*], Chenggang He [1] and Bin Luo [1]**

[1] Key Lab of IC&SP of MOE, School of Computer Science and Technology, Anhui University, Hefei 230601, China; e18101006@stu.ahu.edu.cn (J.W.); e19101007@stu.ahu.edu.cn (C.H.); luobin@ahu.edu.cn (B.L.)

[2] Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA; chqding@uta.edu

* Correspondence: sbchen@ahu.edu.cn

Received: 14 October 2020; Accepted: 27 October 2020; Published: 3 November 2020




**Abstract:** Image segmentation has made great progress in recent years, but the annotation required for image segmentation is usually expensive, especially for remote sensing images. To solve this problem, we explore semi-supervised learning methods and appropriately utilize a large amount of unlabeled data to improve the performance of remote sensing image segmentation. This paper proposes a method for remote sensing image segmentation based on semi-supervised learning. We first design a Consistency Regularization (CR) training method for semi-supervised training, then employ the new learned model for Average Update of Pseudo-label (AUP), and finally combine pseudo labels and strong labels to train semantic segmentation network. We demonstrate the effectiveness of the proposed method on three remote sensing datasets, achieving better performance without more labeled data. Extensive experiments show that our semi-supervised method can learn the latent information from the unlabeled data to improve the segmentation performance.

**Keywords:** semi-supervised learning; remote sensing image segmentation; consistency training; pseudo label

## 1. Introduction

The semantic segmentation of remote sensing image plays an important role in urban planning, change detection, and the construction of geographic information systems. In the past few years, some researchers [1] have used SIFT information, texture information, and other features to classify the superpixel. They select appropriate superpixels on multiple scales to segment the remote sensing image. Based on the superpixel method, many methods can segment different areas of a remote sensing image. However, it is difficult to obtain a satisfactory segmentation result because the superpixel area has the same label.

In recent years, related methods of deep learning have made great progress in semantic image segmentation, such as FCN [2], UNet [3], DeepLab [4], and so on. Based on deep learning, many methods for remote sensing image segmentation have also been developed. They can obtain more precise segmentation results. However, the main obstacle is that adding high-quality pixel-level labels to train data manually is usually expensive, especially for remote sensing images of complex scenes. Zheng et al. [5] used pretraining to alleviate the problem of annotated data. For non-GRB remote sensing images, Wang et al. [6] also tried semi-supervised segmentation based on SVM. To overcome this obstacle, semi-supervised learning is a desirable method for remote sensing image semantic segmentation. This paper aims to contribute to this growing area of research by exploring semi-supervised segmentation in remote sensing images.

Semi-supervised learning has achieved good performance on image classification, such as mean teacher [7], VAT [8] and MixMatch [9]. In these methods, random translation and noise are usually adopted for regularization with unlabeled training data. However, too little work has been devoted to semi-supervised semantic image segmentation. Most existing methods need large amounts of data with pixel-level labels. Some other related studies have focused on pseudo labels. In our opinion, the most effective way is to choose an appropriate regularization method for semantic segmentation based on the semi-supervised classification algorithm.

In this paper, we propose a semi-supervised remote sensing image semantic segmentation method based on Consistent Regularization (CR) training and Average Update of Pseudo-label (AUP). Our main objective is to solve the problem of insufficient training data and high annotation cost in the task of remote sensing image segmentation. In the first part of our method, we first randomly perturb the unlabeled data, and then refer to the parameter updating mode in mean teacher for the model training. Through consistency regularization, which is composed of unlabeled output features and cross-entropy loss of labeled data, a better model can be obtained. For the use of pseudo labels, we first feed the unlabeled data into the previously trained model and obtain the initial features and their pseudo labels. Then, the combined pseudo labeled data and the labeled data are input into the network for training. Inspired by the temporal method [10], we constantly update features in the subsequent training epochs to get more accurate pseudo labels. With this iterative optimization approach, our method can eventually achieve significant improvement in remote sensing image semantic segmentation.

The contributions of this paper can be summarized as follows:

- We propose a semantic segmentation method (CRAUP) for remote sensing images using a small amount of labeled data and a large amount of unlabeled data.
- According to the characteristics of remote sensing images, we propose an improved Consistency Regularization (CR) training method.
- We propose Average Update of Pseudo-label (AUP) method to obtain more accurate pseudo labels and finally improve the segmentation performance of remote sensing images.

We introduce our approach in the following sections and evaluate our method through experiments with several remote sensing image datasets.

## 2. Related Works

Remote sensing image segmentation has always been an important research direction. It can be applied to many practical scenarios to analyze the characteristics of different areas. The early methods mainly use some traditional features to recognize remote sensing images. Zhong et al. [1] built a multiscale graph based on superpixels. They combined the spectral information, texture information, and structure information to enhance the expression ability of the graph. Based on the Markov random field (MRF) model, Zheng et al. [5] made full use of the hierarchical semantic information and further improved the segmentation performance for high spatial resolution remote sensing images.

Traditional image segmentation methods mainly include the following: superpixel segmentation, watershed segmentation, and level set segmentation. In terms of remote sensing image segmentation using superpixels, Wang et al. [6] combined superpixels with minimum spanning tree method. They classified the superpixels through the designed clustering algorithm. This method further optimized the segmentation results by using a more appropriate segmentation scale. Fengkai et al. [11] proposed a method by using the Generalized Mean Shift (GMS) algorithm for PolSAR image superpixel segmentation. They introduced a pre-sorting strategy and a post-processing step into the GMS algorithm. By combining a large number of superpixel algorithms, David et al. [12] presented a comprehensive evaluation and provided new insights relevant for applications. In the earlier watershed segmentation research, Jean et al. [13] introduced a notion of watershed in edge-weighted graphs. They showed that any watershed cut is a topological cut; this property plays an important role

for watershed-based segmentation methods. Watershed algorithm has been applied in some applied studies. Marcin et al. [14] used watershed segmentation and combined regions by maximizing the average contrast. Their approach is also a useful tool for monitoring changes. In addition, level set segmentation is a common image segmentation method. For multiregion segmentation of synthetic aperture radar (SAR) images, Alan et al. [15] proposed a hierarchical level set algorithm replaces the front propagation derivatives by morphological operations. Ruijin et al. [16] used the $\mathcal{L}$ distribution, which improves the segmentation performance for high-resolution polarimetric SAR images. Their proposed level set method can obtain more precise segmentation results.

With the great success of deep learning in nature image segmentation, more methods try to use neural network to segment remote sensing images. Nogueira et al. [17] studied the application of deep learning in remote sensing image segmentation. Due to the huge size of remote sensing image, they processed the image into patches of different sizes. Their method exploited a multicontext paradigm, and finally selected best patch size. In the design of network structure, a DeepResUnet structure [18] was proposed to complete the urban building segmentation of remote sensing images. They improved the network according to the characteristics of remote sensing images. After deepening the depth of the network, the model can effectively perform urban building segmentation at pixel scale from VHR imagery. The segmentation method based on deep learning achieves better segmentation results, but it also presents a problem that cannot be ignored. These methods often require enough annotated training data, especially for deeper network structures.

To solve the data problem of remote sensing image, Kaiser et al. [19] obtained satisfying performance by exploiting noisy large-scale training data. They improved the network model with more easily available weak labeled data. For multispectral image (MSI) and hyperspectral image (HSI) datasets, Kemker et al. [20] proposed an improved self-taught feature learning framework based on ladder work [21]. In the semantic segmentation of these two kinds of remote sensing image, they only used a few annotated data to achieve a good segmentation performance.

In the development of semi-supervised learning, semantic segmentation of natural imagery has also been further explored. Some researchers [22–24] tried to use box annotation or image-level annotation to replace pixel-level annotation. They studied different methods to reduce the reliance of neural network on pixel-level annotation data. Inspired by these methods, in the task of semantic segmentation of remote sensing images, Fu et al. [25] proposed a novel weakly supervised binary segmentation framework. They introduced a weakly supervised feature-fusion network (WSF-Net) to adapt to the unique characteristics of objects on remote sensing image. Their method also proved the applicability of the weak supervised method for remote sensing image. Based on the CAM method, Wang et al. [26] used a single labeled pixel per image or image-level labels train the segmentation network.

The above method mainly uses the weak label to do the further research on the remote sensing image segmentation. In recent years, some researchers also tried to use semi-supervised methods to process remote sensing images. For remote sensing image classification task, a semi-supervised method based on generative adversarial networks (GAN) was proposed by Yan et al. [27] to solve the data problem in scene classification. In the application of semi-supervised method, Protopapadakis et al. [28] designed a semi-supervised model to process land cover segmentation of remote sensing images. Their method improved the model by fine-tuning the model with unlabeled data, and their results showed that the semi-supervised method could effectively improved the segmentation performance. Some researchers [29] also explored the application of transfer learning in remote sensing image processing. For semi-supervised semantic segmentation, Wu et al. [30] proposed a classification algorithm for hyper spectral images. They made use of self-train method to train hyper spectral images, as well as introduced some auxiliary training of spatial regularization.

After reviewing the relevant methods of remote sensing image, we summarize the pros and cons of the relevant approaches. As shown in Table 1, general methods often perform well, but they also require a lot of labeled data for training. There are also some transfer learning methods, in which

they pretrain on other similar datasets, and then train remote sensing image datasets to achieve better performance. It should be noted that it is difficult to find a suitable match for these datasets. In terms of the direction of image labeling, some methods try to use weak labels to replace pixel-level labels, and they have also achieved good performance. However, they also need to do some labeling work while improving the model. Recently, there have also been some semi-supervised methods of remote sensing image; they propose semi-supervised fine-tuning for deep learning model. Their approach uses unlabeled images to fine-tune the model and does not take full advantage of the underlying information of unlabeled data. In summary, we find that too little work has been devoted to semi-supervised segmentation on remote sensing images.

**Table 1.** Comparison of image segmentation methods.

| Method | Pretrain | Additional Label | Advantages | Limitations |
|---|---|---|---|---|
| General method [11,18] | No | No | The more labeled data, the better performance | Depends on the amount of training data |
| Transfer learning method [19,29] | Yes | No | Learn useful information from similar datasets | Inadequate applicability |
| Weakly-supervised method [23,24,26] | No | Weakly annotation | Optimize the model with weak labels | There are still requirements for data annotation |
| Semi-supervised method [28] | No | No | Images without labels can improve the model | Complex training methods and the particularity of different datasets |

To solve the problem, we mainly combine the idea of consistency in semi-supervised classification learning, and then use the average updating method to generate pseudo labels, which further improves the segmentation performance on remote sensing images.

Comparing our method with other existing methods, our method can utilize a large amount of unlabeled data during the training process, which greatly reduces the cost of image annotation. Our method only needs a few pixel-level annotation data to obtain better segmentation results. By combining the characteristics of semi-supervised algorithm and remote sensing, the potential information of unlabeled data can be better learned. Besides, our method can be easily implemented on different datasets for different network architectures and has a good performance for one- or multi-class remote sensing image segmentation. However, our approach also has some shortcomings. While improving the model, our approach is also relatively more complex. Our training process is divided into two steps, and there are some extra parameters that need to be adjusted for different datasets. In general, although there are some disadvantages, our method still has obvious improvement compared with previous methods.

## 3. The Proposed Method

Due to the difference between image segmentation and image classification, we improve the consistency method appropriately. Firstly, according to the characteristics of remote sensing image, we choose color jitter as the preprocessing method, and the image is further changed without changing the label by adding random noise. Besides, for the design of consistency loss, image classification usually uses the consistency method for two probability distributions predicted by the model. In our method, consistency loss is directly calculated for their feature maps before the softmax layer.

Our method applies some new ideas from semi-supervised classification research. Consistency regularization (CR) of output features of unlabeled data has always been an important research direction in semi-supervised learning. The importance of consistency in semi-supervised methods has

been demonstrated by recent research. As for remote sensing image semantic segmentation, we first propose a regularization method which can make full use of unlabeled data as the first step. In addition, similar to iterative improvement labels in weakly supervised segmentation, our average update of pseudo-label (AUP) method also updates and improves the pseudo label on average. In the following, we describe each part of our approach.

### *3.1. Consistency Regularization Training*

The regularization technique plays an important role in semi-supervised learning. Our new CR method applies consistency loss to the output feature of unlabeled data. In the semi-supervised semantic segmentation task, the input is divided into labeled image $x$ and unlabeled image $u_x$, and we only have the label $y$ for $x$. We use $X$ to denote the labeled dataset and $U_x$ to denote the unlabeled dataset. We train the student model using $X$ and $U_x$ and the update averaging model as the teacher model. Our method differs from semi-supervised classification in that we cannot translate and crop the input image because, in image segmentation, two random translation or cropping of an image will lead to changes of labels at the same pixel position, which will affect the consistency method. For supervised training, we perform color jittering and random flipping on the input images. Then, we use the general cross-entropy loss for labels and student model predictions as follows:

$$L_{ce} = \frac{1}{|X|} \sum_{x,p \in X} H(p, P_{model}(y|x; \theta_s)) \tag{1}$$

With respect to the unsupervised training, given an unlabeled image $u_x$, we randomly perturb $u_x$ twice. Unlike semi-supervised image classification, we cannot translate the image randomly as with mean teacher. In view of characteristics of remote sensing images, we choose the method of color jitter to make slight changes. As shown in Figure 1, the input generates $u_x1$ and $u_x2$ after twice performing color jitter, and we add random noise to them, respectively. For the unlabeled image, we randomly generate a normally distributed noise of the same size with mean 0 and standard deviation 1. Finally, we multiply this noise by an intensity factor of 0.001 and add it to the input image. Then, one gets output $P_u$ through the student model and the other gets output $P'_u$ through the teacher model. The definition of unsupervised loss is as follows:

$$L_{cons} = D_{x,\theta,\eta}[||f(u_x2, \theta', \eta') - f(u_x1, \theta, \eta)||^2], \tag{2}$$

$$\theta'_t = \alpha\theta'_{t-1} + (1 - \alpha)\theta_t, \tag{3}$$

where $\theta'$ is the parameter of the teacher model and $\theta$ is the parameter of the student model. $\eta$ and $\eta'$ are random noises added in the two perturbations, respectively. We define $D$ as mean-squared loss (MSE) in our experiment. In our approach, the parameters of the teacher's model do not take part in back propagation, but directly update the average according to the parameters of the student's model. After each training batch $t$, the parameters of the teacher model are updated with the parameters of the current student model.

We let supervised and unsupervised training work together in our approach. The full objective function is thus given by:

$$L(X, U_x, Y) = L_{ce}(X, Y, \theta) + \beta L_{cons}(U_x, \theta, \theta'). \tag{4}$$

To make the training of the model more stable, we set weight $\beta$ for the semi-supervised consistency loss as follows:

$$\beta = e^{-5K^2}, \text{ where } K = 1 - \frac{e_c}{e_m}. \tag{5}$$

Here, $e_c$ is the current epoch and $e_m$ is the max epoch. By learning the latent information in the unlabeled data through our consistent method, we obtain a more accurate model.

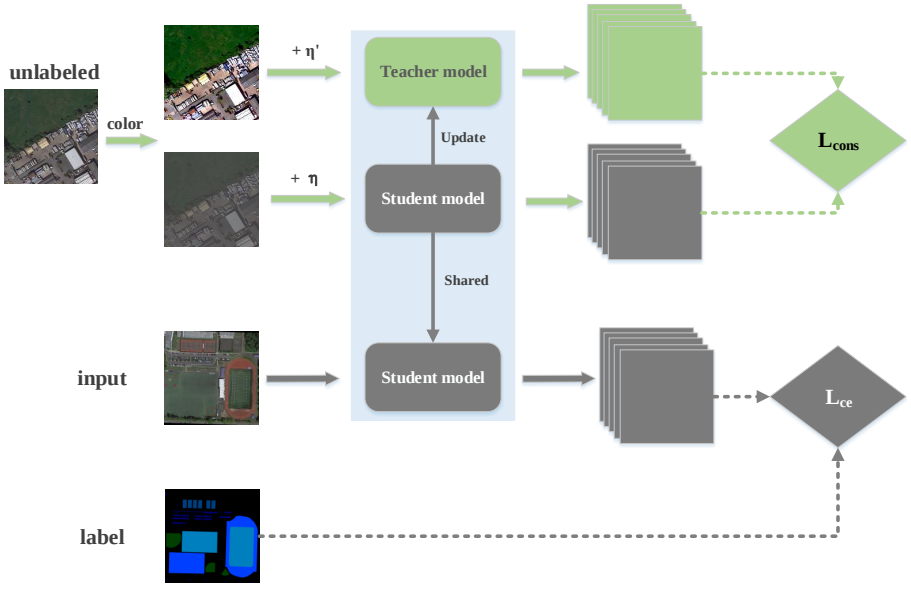



**Figure 1.** The main framework of our proposed consistency regularization (CR) training method. The parameters of two student models are shared between the supervised and unsupervised parts. The teacher model is updated according to the changes of student model. The final optimization loss consists of regularized loss and cross entropy loss.

### *3.2. Average Update of Pseudo-Label Feature*

Pseudo label is an important research direction in semi-supervised learning. Our second step method is also an improvement on the image segmentation based on this method.

The main framework of our AUP method is shown in Figure 2. At the beginning of this process, we have an improved CR model and two sets of data $X$ and $U_x$. As in the general method, we first input unlabeled images into the model and obtained their feature maps $F_u$. For these feature maps, we convert them to corresponding pseudo labels $y_p$ and save them. In comparison to traditional methods, we save the current output features of the unlabeled data. We take the labeled dataset $X(x, y)$ and the unlabeled dataset $U_x(u_x, y_p)$ as training set to the network and load the best model in the previous step. After adding pseudo labels to the training, we recalculate the output features corresponding to the unlabeled data after each epoch as $F_u^e$. Different from the usual pseudo label method, we do not directly use the output features to construct the pseudo labels, but use average update of pseudo labels to continuously improved it:

$$F_u^c = F_u^{e-1} + \gamma F_u^e. \tag{6}$$

where $F_u^e$ is the output of the current epoch and $F_u^{e-1}$ is the output of the previous epoch. By constantly updating during the training process, we get smoother output features $F_u^c$. Using this feature, we can obtain the segmentation results of unlabeled data after model prediction as pseudo labels $y_p$.

Combining the labeled data and unlabeled data with pseudo labels, our loss function is as follows:

$$L = \frac{1}{|X|} \sum_{x,p \in X} H(p, P_{model}(y|x)) + \frac{1}{|U_x|} \sum_{u_x, y_p \in U_x} H(p, P_{model}(y_p|u_x)). \tag{7}$$

After continuous iterative updates, we gradually acquire more accurate pseudo labels. We add them to the training set and train them to obtain a better model. The whole process of our semi-supervised remote sensing image semantic segmentation method based on CR and AUP is shown in Algorithm 1.

**Algorithm 1** Semi-supervised remote sensing image semantic segmentation by CR and AUP

**Input:** Training set $X = \{x, y\}$ and unlabeled set $U_x = \{u_x\}$. $x$ is a remote sensing image and $y$ is its segmentation label. $u_x$ is a remote sensing image without label.

**Output:** Label predicted for the test set $X_t$, $X_t$ contains remote sensing images to be predicted and their ground truth.

Step1: Consistence Regularization (CR)

Choose $x$ and $u_x$ from datasets $X$ and $U_x$.
Color jitter for $x \rightarrow x'$, $u_x \rightarrow u'_x$ and $u''_x$.
Calculate CE loss for the labeled data $x'$ and $y$.
Calculate MSE loss for $u'_x$ and $u''_x$.
Combine $L_{ce}$ and $L_{mse}$, update the parameters of the student model.
Obtain an improved teacher model as CR model.

Step2: Average Update of Pseudo-label (AUP)

$t \leftarrow 0$
**for** $t = 0$ to $e_m$ **do**
  **if** $t = 0$ **then**
    Obtain the feature maps of the unlabeled data using the current model as $F_{t0}$.
    Predict pseudo labels of the unlabeled sample by using $F_{t0}$.
  **else**
    Predict current feature $F_t$, and then the pseudo label feature $F_t$ is updated in combination with $F_{t-1}$.
    Update pseudo labels and continue with supervised model training.
  **end if**
**end for**
Use the final AUP model to predict the label of the test set $X_t$.

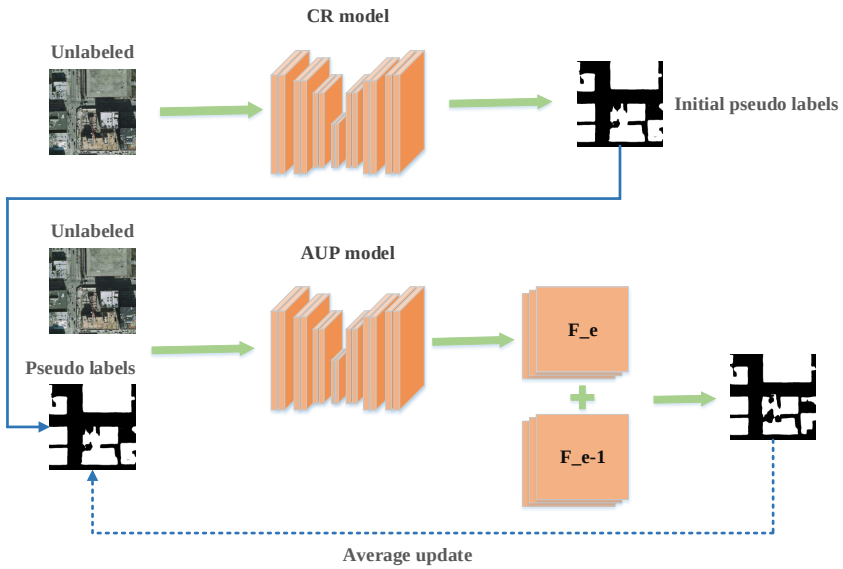



**Figure 2.** The main framework of our proposed average update of pseudo-label (AUP) method. We first obtain initial pseudo labels with the consistency regularization (CR) trained model, and then continue to train the model by updating feature maps and generate better pseudo labels after each epoch.

## 4. Experiments

We demonstrate the effectiveness of our approach on different remote sensing image datasets. In our experiment, we used different portions of labeled data in our comparative experiments, e.g., 1/8, 1/4, and 1/2. To verify the applicability of our method in various situations, we adopted several segmentation models as our base models. Simultaneously, we also made experimental comparisons on different types of datasets. We describe our experiments in detail in the following sections.

For remote sensing image segmentation, UNet [3] usually has good performance. Therefore, we first chose UNet as our base model. In addition, we conducted experiments with DeepLabV3 [31] and DeepLabV3+ [32] as our base models, where their classification backbone is ResNet50. The results indicate the universality of our method.

We tested the effectiveness of our method on three remote sensing datasets. The first dataset esd INRIA Aerial database [33], consisting of urban building aerial images collected by INRIA in 2017. The second one esd Road Extraction dataset [34], consisting of remote sensing images from road extraction competition in CVPR 2018. The last one was ISAID dataset [35], proposed by the remote sensing Laboratory of Wuhan University based on DOTA [36] database.

### *4.1. Implementation Details*

In our experiments, ADAM was used as training optimizer for all methods and the initial learning rate was set to 0.0001. For unsupervised image perturbations, we first randomly flipped the input image. Then, we added Gaussian noise to the original image, and we selected some images from ISAID dataset for comparison. As shown in Figure 3, when the coefficient $\xi$ is too large, the object in the image is too greatly affected, which is not conductive to model training. Thus, we chose a noise factor $\xi$ of 20 in our experiments. Finally, we conducted random color jitter for remote sensing images, and the coefficient range of dithering was usually 0.1–0.5. We conducted several groups of comparison experiments with different labels on INRIA Aerial Dataset when we used UNet. As shown in Figure 4, the model performs better when color perturbations are added. However, when the perturbation is too strong, the mIoU of the model decreases. Therefore, we finally conducted random color jitter within the 0.2 range of brightness, contrast, and hue, rather than using image translation.

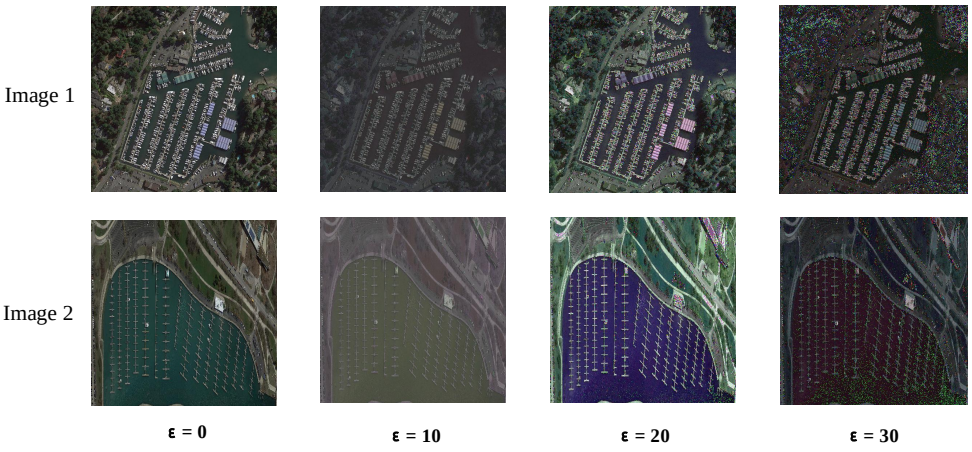



**Figure 3.** The influence of different noise coefficients on remote sensing images.

During model training, we used 0.999 as the update parameter $\theta$ for the teacher model as in the mean teacher method. In our semi-supervised experiment, 100 epochs were trained. The weight of unsupervised loss was increased from 0 to 1 after the first 30 epochs. Meanwhile, we defined the batch size of the labeled data to 2 and the batch size of the unlabeled data to 3, and then we mixed them together for training. To make the hidden information of pseudo labels more fully learned, our experiments updated feature maps at each epoch, but pseudo labels were updated every five

epochs. In addition, we used 0.99 as the update ratio $\gamma$ of the unlabeled data output feature. Finally, we report mean intersection-over-union (mIoU) for their test images, and we also make a visual analysis of the segmentation results to verify our hypothesis. Next, we detail our experiments on each datasets, showing comparative results and analyses.

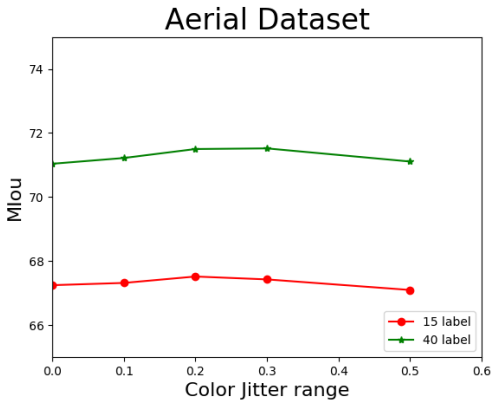



**Figure 4.** Semi-supervised experimental results of different color jitter ranges on INRIA Aerial dataset.

### *4.2. Image Semantic Segmentation on Inria Aerial Dataset*

The INRIA Aerial dataset contains aerial color imagery with a spatial resolution of 0.3 m. The image covers remote sensing images from densely populated cities to alpine towns. The objective of segmentation is divided into two semantic classes: building and no building. The size of each image is 5000 $\times$ 5000 pixels. Since the direct training would lead to the excessive occupation of GPU memory, we cropped each image into 100 patches of 500 $\times$ 500 pixels to train the model. In this way, we did not lose the details of the image. We separated the data into 155 images in the training set and 25 images in the test set. For semi-supervised setting, we used a small portion, from 15 to 80, as labeled data.

Table 2 shows comparative experimental results of segmentation accuracy of CR and AUP in our method on INRIA Aerial dataset when different network structures were adopted. As can be seen in the table, when training with only a small amount of labeled data, after adding unlabeled data and using our CR and AUP method, the mIoU value of segmentation can be significantly improved, especially when there are few labeled data available, such as only 15 labeled images. CR increases mIoU value from 67.52 to 71.51. In addition, when 155 images were labeled and we also used CR for training, the segmentation accuracy could be improved. However, because AUP uses the pseudo labels of unlabeled data, we did not do this experiment when all training data have labels.

**Table 2.** Segmentation accuracy (mIoU, %) comparison of CR and AUP in our method on INRIA Aerial dataset when different network structures are adopted.

| INRIA Aerial Dataset | 15 Labels | 40 Labels | 80 Labels | 155 Labels |
|---|---|---|---|---|
| UNet [3] | 67.52 | 71.50 | 73.81 | 75.82 |
| UNet + CR | 71.51 | 73.18 | 74.42 | 76.23 |
| UNet + CR + AUP | 72.54 | 73.63 | 74.95 | - |
| DeepLab V3 [31] | 71.97 | 74.39 | 75.52 | 77.61 |
| DeepLab V3 + CR | 73.32 | 75.86 | 76.43 | 78.19 |
| DeepLab V3 + CR + AUP | 74.27 | 76.21 | 76.92 | - |
| DeepLab V3+ [32] | 72.30 | 75.31 | 76.19 | 78.35 |
| DeepLab V3+ + CR | 73.83 | 76.08 | 76.82 | **78.61** |
| DeepLab V3+ + CR + AUP | **74.40** | **76.64** | **77.12** | - |

To further test our AUP method, we saved pseudo labels in different epochs of experiments for comparison. With the semi-supervised experiments, the output features of our unlabeled data

are constantly updated to produce more accurate pseudo labels. As shown in Figure 5, the pseudo label quality corresponding to the unlabeled data is gradually improved. After adding these more accurate pseudo labels into the training procedure, better remote sensing semantic segmentation results are obtained.

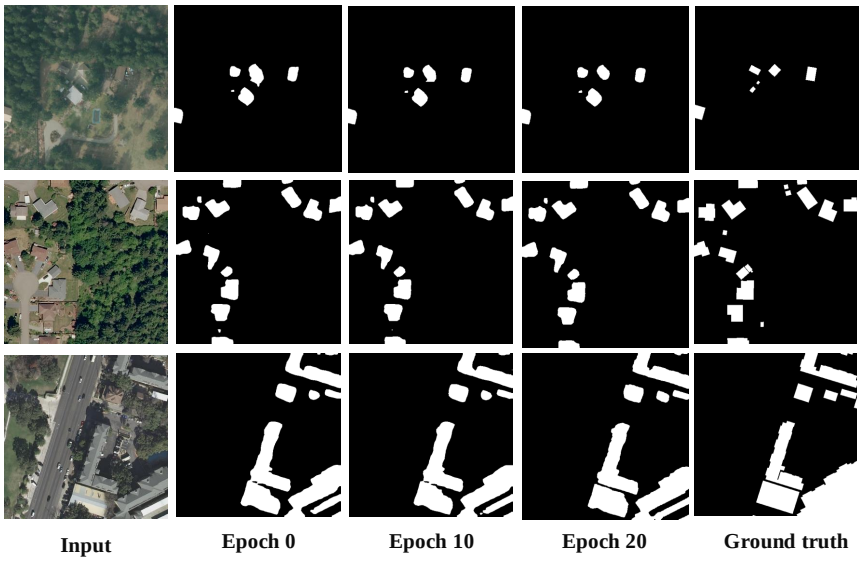



**Figure 5.** Some pseudo-label examples of remote sensing images semantic segmentation obtained by our AUP method. From left to right are the pseudo labels in different epochs when using AUP.

### *4.3. Image Semantic Segmentation on Road Extraction Dataset*

After the previous experiments, we wanted to verify whether our method can also perform well when both the input image and the segmentation target are small. Therefore, we also conducted experiments on Road Extraction dataset, which contains a total of 6226 remote sensing images with a resolution of 1024 $\times$ 1024 pixels. The main task is to extract the road area in the image, and the classes to be segmented are road and non-road. During the experiments, we resized the training set to 512 $\times$ 512 pixels. The training set consisted of 5400 images and the test set of 826 images. The number of labeled images in the semi-supervised experiments was set as 400, 1350, and 2700.

The experimental results of segmentation accuracy of CR and AUP in our method on Road Extraction dataset are shown in Table 3 when different network structures are adopted. From the results, we can see that, when CR and AUP were used, mIoU values increase by about 2% on average. In addition, for the task of road extraction, we can also find that the network structure of UNet is better than DeepLab v3. This is one reason many remote sensing image semantic segmentation algorithms use UNet as the baseline. The network of UNet is more concise and efficient, thus it is easier to improve and optimize.

Through experiments on this dataset, the results show that our method is not only applicable to specific datasets, but also can achieve good performance in different remote sensing image segmentation tasks. It is worth noting that, in this experiment, we did not make too many adjustments and modifications to the experimental hyper parameters, so our method can be easily migrated to different data scenarios and obtain meaningful results.

**Table 3.** Segmentation accuracy (mIoU, %) comparison of CR and AUP in our method on Road Extraction dataset when different network structures are adopted.

| Road Extraction Dataset | 400 Labels | 1350 Labels | 2700 Labels | 5400 Labels |
|---|---|---|---|---|
| UNet [3] | 48.70 | 53.55 | 57.34 | 59.81 |
| UNet + CR | 51.53 | 55.57 | 57.82 | 60.02 |
| UNet + CR + AUP | 52.76 | 56.12 | 58.23 | - |
| DeepLab V3 [31] | 48.42 | 52.40 | 54.66 | 57.23 |
| DeepLab V3 + CR | 50.36 | 53.24 | 55.13 | 57.67 |
| DeepLab V3 + CR + AUP | 51.44 | 53.68 | 55.54 | - |
| DeepLab V3+ [32] | 51.50 | 56.21 | 59.11 | 62.41 |
| DeepLab V3+ + CR | 52.54 | 57.09 | 59.62 | **62.63** |
| DeepLab V3+ + CR + AUP | **53.63** | **58.01** | **60.22** | - |

### *4.4. Multi-Class Image Semantic Segmentation on Isaid Dataset*

The datasets of the previous two experiments are image semantic segmentation of two classes, and the region to be segmented is a target class and background. Therefore, we further conducted experiments on ISAID dataset to explore the performance of our method in multi-class image segmentation. ISAID dataset consists of 15 target classes, and our task is to identify the different classes from the input images and to segment the correct areas. The dataset provides 1411 remote sensing images of various sizes, including various targets: basketball court, airplane, port, etc. During the training, we uniformly resized input images to 512 $\times$ 512 pixels, and divided the training set into 1200 images and the test set into 211 images. The data for the semi-supervised experiment were set to 100, 300 and 600 labeled data randomly selected from all training data.

We still adopted mIoU as the evaluation criterion for semantic segmentation of multiple classes. The experimental results of different segmentation methods on multi-class ISAID dataset are shown in Table 4. As can be seen in the table, due to the increased difficulty of the task, the scene of remote sensing image is more complex, and the segmentation results of mIoU of several models are relatively low. Comparing the results of the three network models, we can also find that DeepLab-related model has obvious advantages in multi-class image segmentation. In addition, combined with semi-supervised experiments of three models, our approach performs well in the case of few labeled data, and the final results are significantly improved. However, due to the complex label information of multi-class segmentation, our AUP method has only been slightly improved. In general, our semi-supervised method can obtain better results in the semantic segmentation of multi-class remote sensing images.

**Table 4.** Segmentation accuracy (mIoU, %) comparison of CR and AUP in our method on multi-class ISAID dataset.

| ISAID Dataset | 100 Labels | 300 Labels | 600 Labels | 1200 Labels |
|---|---|---|---|---|
| UNet [3] | 25.13 | 27.59 | 33.50 | 36.92 |
| UNet + CR | 28.46 | 30.61 | 34.78 | 37.24 |
| UNet + CR + AUP | 29.51 | 31.12 | 35.17 | - |
| DeepLab V3 [31] | 36.13 | 39.12 | 41.85 | 44.86 |
| DeepLab V3 + CR | 38.35 | 40.49 | 42.55 | 45.13 |
| DeepLab V3 + CR + AUP | 38.92 | 40.82 | 42.95 | - |
| DeepLab V3+ [32] | 36.76 | 41.46 | 43.92 | 47.45 |
| DeepLab V3+ + CR | 39.31 | 42.56 | 44.86 | **47.73** |
| DeepLab V3+ + CR + AUP | **39.85** | **42.88** | **45.02** | - |

We also visualized the output of our CR and AUP model to compare the contribution of different parts of our method. Some examples of segmentation results are shown in Figure 6. As can be seen from the figure, the segmentation results of the model are seriously affected when there are few labeled data

available. However, when we added unlabeled data to the training dataset and used our semi-supervised method, we could significantly improved the segmentation results of the model. The comparison in different experimental stages also shows that both of the two parts in our method can improve the segmentation accuracy step by step, and finally our model can obtain satisfactory results.

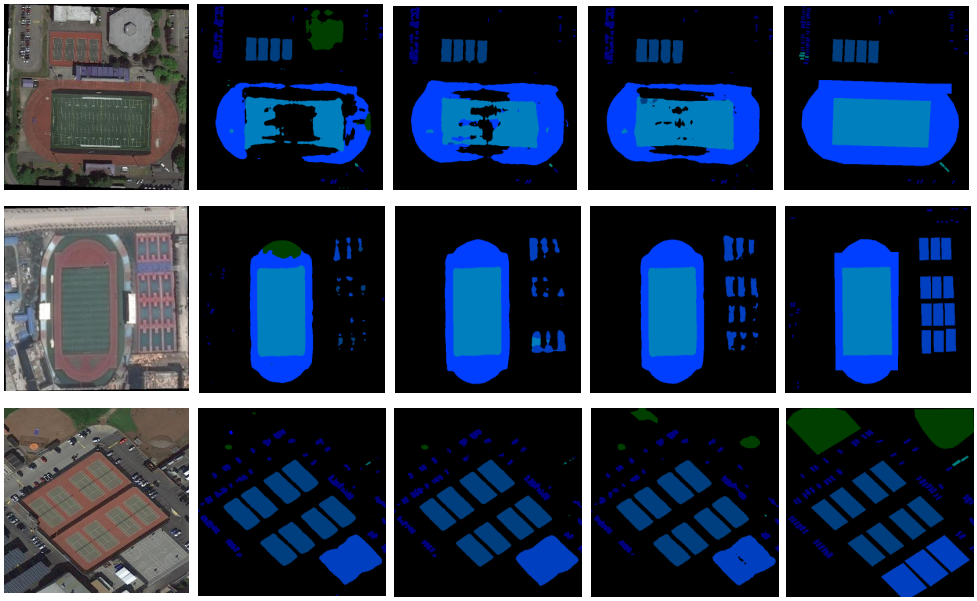



**Figure 6.** Some examples of remote sensing image semantic segmentation when the proposed CR and AUP are adopted on validation set of ISAID dataset.

## 5. Discussion

After a large number of experiments, we can conclude that our method CR and AUP is effective for semi-supervised segmentation of remote sensing images. As can be seen from the above experiments, when 1/8, 1/4, and 1/2 of the labeled data are used, respectively, our results are significantly improved compared with the baseline models. Therefore, we further explored the differences in the degree of improvement of our methods for different proportions of labeled data in the training data. We selected the best baseline model DeepLabV3+ and compared its results across different datasets. The segmentation performance of the model corresponding to variation of number of labeled images is shown in Figure 7.

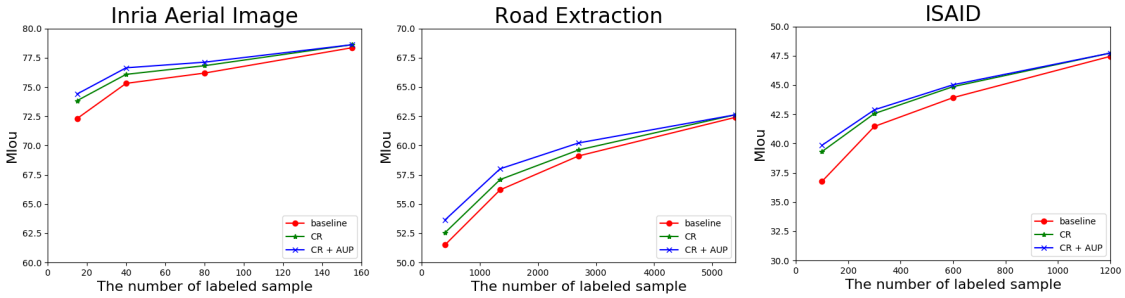



**Figure 7.** Variation of segmentation performance of the proposed CR and AUP method with different numbers of labeled training samples on three datasets.

As can be seen in the subfigures, the accuracy of the model is improved by 2–3% on all datasets when our CR and AUP methods were applied, especially when there were few labeled data. When only 1/8 of labeled data is available, our approach can achieve the accuracy of baseline model with 1/4 of

labeled images. When all data are labeled, the performance of the model can be further improved by using our CR method.

To better compare the difference between the different models, we summarize the performance of our method when using 100 labeled data on ISAID dataset. All experiments were performed on a 1080ti GPU. As shown in Table 5, UNet has the minimum trainable parameter of 17.2 M and the fastest training time, but the accuracy of the model is much lower than the other models. In addition, compared to DeepLab V3 and DeepLab V3+, DeepLab V3+ has a slightly increased parameter, but the training time is shorter and the model has better performance. Finally, our CR and AUP methods only increase the training time without changing the time consumption of the test phase, which we think is acceptable for our experimental purposes.

**Table 5.** The proposed method's performance in terms of trainable parameters (time and space). We compared the performance of different methods using 100 labels on the ISAID dataset.

| Architecture | Params | MIoU | Base Time | CR + AUP Time |
|---|---|---|---|---|
| UNet [3] | 17.2 M | 29.51 | 32 min | 2 h 11 min + 1 h 20 min |
| DeepLab V3 [31] | 39.6 M | 38.92 | 1 h | 4 h 19 min + 2 h 20 min |
| DeepLab V3+ [32] | 39.7 M | 39.85 | 55 min | 4 h + 2 h 8 min |

Comparing to the previous semi-supervised classification methods such as mean teacher [7], which used 1000, 2000, and 4000 as the number of labeled images on CIFAR10, as can be seen from their experimental results, their method achieved the test error rate of 21.55% on 1000 labels, 15.73% on 2000 labels, and 12.31% on 4000 labels. From their results, we can also learn that the increase in the amount of labeled data has a great impact on the deep learning method. Our method produced similar results with different numbers of labeled data. However, we can also infer from this result that the semi-supervised method still has great potential. For image segmentation task, the semi-supervised method is not improved as much as the image classification. We will study the differences between image segmentation and image classification to explore more effective semi-supervised segmentation methods.

Our method AUP does not work as well on the multi-class segmentation dataset as on two-class segmentation datasets. Through analyzing the update status of pseudo label and true label, we believe that the main reason is that the scene of multi-classification is relatively complex and the accuracy of pseudo label is not good, so it is difficult to improve the pseudo label by average update. In future work, we will further explore a more appropriate use of pseudo labels for multi-class semantic segmentation tasks and try to conduct semi-supervised research on object detection.

In addition, it is worth noting that, in our pseudo label update method, we can observe slight changes of pseudo labels by visualizing them. Compared with the ground truth, we learn that our pseudo labels constantly reclassify the misidentified pixels as backgrounds during the updating process. After reviewing different unlabeled images, we found that the main problem is that the boundary of the target was not easy to be divided into the correct class. Therefore, strengthening the training weight of the object boundary in semi-supervised learning may further improve the segmentation performance of the model.

Through experimental analysis, we can find that our method can make better use of unlabeled images by applying the semi-supervised consistency loss and pseudo label updating method, which obviously improved the model performance on remote sensing image segmentation. In addition, there are some weaknesses in our approach. When we use pseudo label update method, the pseudo label does not always update for the better, and sometimes it gets worse.

We also have some ideas about future research directions. First, since the unlabeled data of remote sensing images are relatively easy to obtain, we can try to continuously increase the unlabeled data to explore the limit of the semi-supervised algorithm. Besides, due to the instability of pseudo label

update method, we will try to explore the Fixmatch [37] method in the future and combine pseudo label with the consistency method to participate in training at the same time. Thus, we can simplify the model training process and obtain better performance.

## 6. Conclusions

In this paper, we explore the semantic segmentation of remote sensing images in the absence of labeled data and propose a semi-supervised approach to solve the problem. Combining the characteristics of semi-supervised method and remote sensing image, we propose two semi-supervised method (CR and AUP) that can be used together. Through extensive experiments on three remote sensing image datasets, it is shown that our methods can significantly improve the segmentation performance of the model by adding unlabeled data. For remote sensing images, the acquisition of a large amount of unlabeled data is relatively simple. Therefore, our research also has very important practical significance.

Our research shows that recent semi-supervised classification methods can also be applied to image segmentation or object detection after appropriate improvement. For remote sensing images, it is also worth further studying the semi-supervised approach because the unlabeled data are easy to obtain. In the future, we will explore the theoretical details of the latest semi-supervised methods, such as MixMatch and UDA, and then try to further improve the semi-supervised segmentation performance of remote sensing images.

**Author Contributions:** Methodology, J.W. and C.H.Q.D.; Resources, B.L.; Validation, C.H.; Writing—original draft, J.W.; and Writing—review and editing, S.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by NSFC Key Project of International (Regional) Cooperation and Exchanges (No. 61860206004), National Natural Science Foundation of China (NO. 61976004, 61572030), Key Project of Chinese National Programs for Fundamental Research and Development (973 Program) under Grant 2015CB351705, and Collegiate Natural Science Fund of Anhui Province (NO. KJ2017A014).

**Acknowledgments:** The authors thank the editor and anonymous reviewers for their valuable comments and suggestions, which were very helpful in improving this paper.

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

## References

1. Zhong, Y.; Gao, R.; Zhang, L. Multiscale and Multifeature Normalized Cut Segmentation for High Spatial Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6061–6075. [CrossRef]
2. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
3. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, 234–241, arXiv:1505.04597.
4. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
5. Zheng, C.; Zhang, Y.; Wang, L. Semantic Segmentation of Remote Sensing Imagery Using an Object-Based Markov Random Field Model With Auxiliary Label Fields. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3015–3028. [CrossRef]
6. Wang, M.; Dong, Z.; Cheng, Y.; Li, D. Optimal Segmentation of High-Resolution Remote Sensing Image by Combining Superpixels With the Minimum Spanning Tree. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 228–238. [CrossRef]
7. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017; pp. 1195–1204.

8. Miyato, T.; Maeda, S.i.; Koyama, M.; Ishii, S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1979–1993. [CrossRef]
9. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C. MixMatch: A Holistic Approach to Semi-Supervised Learning. *arXiv* **2019**, arXiv:1905.02249.
10. Laine, S.; Aila, T. Temporal Ensembling for Semi-Supervised Learning. *arXiv* **2017**, arXiv:1610.02242.
11. Lang, F.; Yang, J.; Yan, S.; Qin, F. Superpixel Segmentation of Polarimetric Synthetic Aperture Radar (SAR) Images Based on Generalized Mean Shift. *Remote Sens.* **2018**, *10*, 1592. [CrossRef]
12. Stutz, D.; Hermans, A.; Leibe, B. Superpixels: An evaluation of the state-of-the-art. *Comput. Vis. Image Underst.* **2018**, *166*, 1–27. [CrossRef]
13. Cousty, J.; Bertrand, G.; Najman, L.; Couprie, M. Watershed Cuts: Thinnings, Shortest Path Forests, and Topological Watersheds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 925–939. [CrossRef] [PubMed]
14. Ciecholewski, M. River channel segmentation in polarimetric SAR images: Watershed transform combined with average contrast maximisation. *Expert Syst. Appl.* **2017**, *82*, 196–215. [CrossRef]
15. Braga, A.M.; Marques, R.C.P.; Rodrigues, F.A.A.; de Medeiros, F.N.S. A Median Regularized Level Set for Hierarchical Segmentation of SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1171–1175. [CrossRef]
16. Jin, R.; Yin, J.; Zhou, W.; Yang, J. Level Set Segmentation Algorithm for High-Resolution Polarimetric SAR Images Based on a Heterogeneous Clutter Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 4565–4579. [CrossRef]
17. Nogueira, K.; Dalla Mura, M.; Chanussot, J.; Schwartz, W.R.; dos Santos, J.A. Dynamic Multicontext Segmentation of Remote Sensing Images Based on Convolutional Networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7503–7520. [CrossRef]
18. Yi, Y.; Zhang, Z.; Zhang, W.; Zhang, C.; Li, W.; Zhao, T. Semantic Segmentation of Urban Buildings from VHR Remote Sensing Imagery Using a Deep Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 1774. [CrossRef]
19. Kaiser, P.; Wegner, J.D.; Lucchi, A.; Jaggi, M.; Hofmann, T.; Schindler, K. Learning Aerial Image Segmentation From Online Maps. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6054–6068. [CrossRef]
20. Kemker, R.; Luu, R.; Kanan, C. Low-Shot Learning for the Semantic Segmentation of Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6214–6223. [CrossRef]
21. Rasmus, A.; Valpola, H.; Honkala, M.; Berglund, M.; Raiko, T. Semi-Supervised Learning with Ladder Networks. *arXiv* **2015**, arXiv:1507.02672.
22. Souly, N.; Spampinato, C.; Shah, M. Semi Supervised Semantic Segmentation Using Generative Adversarial Network. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5689–5697.
23. Chang, Y.T.; Wang, Q.; Hung, W.C.; Piramuthu, R.; Tsai, Y.H.; Yang, M.H. Weakly-Supervised Semantic Segmentation via Sub-Category Exploration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8991–9000.
24. Ouali, Y.; Hudelot, C.; Tami, M. Semi-Supervised Semantic Segmentation with Cross-Consistency Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12674–12684.
25. Fu, K.; Lu, W.; Diao, W.; Yan, M.; Sun, H.; Zhang, Y.; Sun, X. WSF-NET: Weakly Supervised Feature-Fusion Network for Binary Segmentation in Remote Sensing Image. *Remote Sens.* **2018**, *10*, 1970. [CrossRef]
26. Wang, S.; Chen, W.; Xie, S.M.; Azzari, G.; Lobell, D.B. Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery. *Remote Sens.* **2020**, *12*, 207. [CrossRef]
27. Yan, P.; He, F.; Yang, Y.; Hu, F. Semi-Supervised Representation Learning for Remote Sensing Image Classification Based on Generative Adversarial Networks. *IEEE Access* **2020**, *8*, 54135–54144. [CrossRef]
28. Protopapadakis, E.; Doulamis, A.D.; Doulamis, N.; Maltezos, E. Semi-Supervised Fine-Tuning for Deep Learning Models in Remote Sensing Applications. *arXiv* **2020**, arXiv:2006.00345.
29. Hong, D.; Yokoya, N.; Xia, G.; Chanussot, J.; Zhu, X.X. X-ModalNet: A Semi-Supervised Deep Cross-Modal Network for Classification of Remote Sensing Data. *arXiv* **2020**, arXiv:2006.13806.
30. Wu, Y.; Mu, G.; Qin, C.; Miao, Q.; Ma, W.; Zhang, X. Semi-Supervised Hyperspectral Image Classification via Spatial-Regulated Self-Training. *Remote Sens.* **2020**, *12*, 159. [CrossRef]

31. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
32. Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv* **2018**, arXiv:1802.02611.
33. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
34. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raska, R. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. *arXiv* **2018**, arXiv:1805.06561.
35. Zamir, S.W.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Khan, F.S.; Zhu, F.; Shao, L.; Xia, G.; Bai, X. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images. *arXiv* **2019**, arXiv:1905.12886.
36. Xia, G.; 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.
37. Sohn, K.; Berthelot, D.; Li, C.; Zhang, Z.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Zhang, H.; Raffel, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *arXiv* **2020**, arXiv:2001.07685.


**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).