International Journal of
Geo-Information

MDPI

Article

Cloud and Snow Segmentation in Satellite Images Using an Encoder–Decoder Deep Convolutional Neural Networks

Kai Zheng [1], Jiansheng Li [1,*], Lei Ding [2], Jianfeng Yang [1], Xucheng Zhang [1] and Xun Zhang [1]

[1] Institute of Surveying and Mapping, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China; kaizheng90@foxmail.com (K.Z.); marsyjf@126.com (J.Y.); banditex1990@126.com (X.Z.); zx121777@163.com (X.Z.)

[2] Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy; lei.ding@unitn.it

* Correspondence: ljs2021@vip.henu.edu.cn

Abstract: The segmentation of cloud and snow in satellite images is a key step for subsequent image analysis, interpretation, and other applications. In this paper, a cloud and snow segmentation method based on a deep convolutional neural network (DCNN) with enhanced encoder–decoder architecture—ED-CNN—is proposed. In this method, the atrous spatial pyramid pooling (ASPP) module is used to enhance the encoder, while the decoder is enhanced with the fusion of features from different stages of the encoder, which improves the segmentation accuracy. Comparative experiments show that the proposed method is superior to DeepLabV3+ with Xception and ResNet50. Additionally, a rough-labeled dataset containing 23,520 images and fine-labeled data consisting of 310 images from the TH-1 satellite are created, where we studied the relationship between the quality and quantity of labels and the performance of cloud and snow segmentation. Through experiments on the same network with different datasets, we found that the cloud and snow segmentation performance is related more closely to the quantity of labels rather than their quality. Namely, under the same labeling consumption, using rough-labeled images only performs better than rough-labeled images plus 10% fine-labeled images.

Keywords: satellite image; semantic segmentation; encoder–decoder; CNN; TH-1; cloud and snow detection; label quality

check for updates

Citation: Zheng, K.; Li, J.; Ding, L.; Yang, J.; Zhang, X.; Zhang, X. Cloud and Snow Segmentation in Satellite Images Using an Encoder–Decoder Deep Convolutional Neural Networks. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 462. https://doi.org/10.3390/ijgi10070462

Academic Editors: Gloria Bordogna and Cristiano Fugazza

Received: 21 May 2021
Accepted: 1 July 2021
Published: 6 July 2021

1. Introduction

With satellites becoming indispensable infrastructures for the development of the national economy, the acquisition of remote sensing images (RSIs) has become easier. RSIs have been widely used in a variety of fields such as infrastructure, agriculture, forestry, geology, hydrology, transportation, disaster prediction, etc. However, 66.7% of the Earth's surface is covered by clouds [1], which is a major factor restricting the application of optical RSIs. Additionally, due to the similar characteristics (e.g., high reflectivity rate) of cloud and snow in optical bands, some traditional methods such like threshold-based methods cannot distinguish them and often lead to misjudgment. This greatly hinders the automatic processing of RSIs. Furthermore, there is a deeper need to detect cloud and snow such like the construction of the atmospheric reflectance database, which can be used to serve the retrieval of atmospheric aerosols [2]. Therefore, it is of great significance to segment cloud and snow quickly, accurately, and automatically.

A number of image segmentation methods have been proposed since the 1970s, among which the most classic one is the Otsu [3] proposed by Nobuyuki Otsu in 1979. It uses the exhaustive method to determine the threshold that results in the maximum variance between objects in images, thus segmenting images into foreground and background images. Due to the high reflectivity of clouds in optical bands, we can use the Otsu method to segment the cloud and background from the visible images. However, cloud and snow often share similar characteristics in optical bands, making it hard to distinguish them from each other, and it is almost impossible to segment cloud from other objects using Otsu all at once. Although many methods have been proposed to improve the Otsu (such as the multi-Otsu), they suffer from limitations such as large amount of calculations, low robustness, etc. Figure 1 shows the cloud segmentation result of the Otsu, where it produces totally different results on similar images.

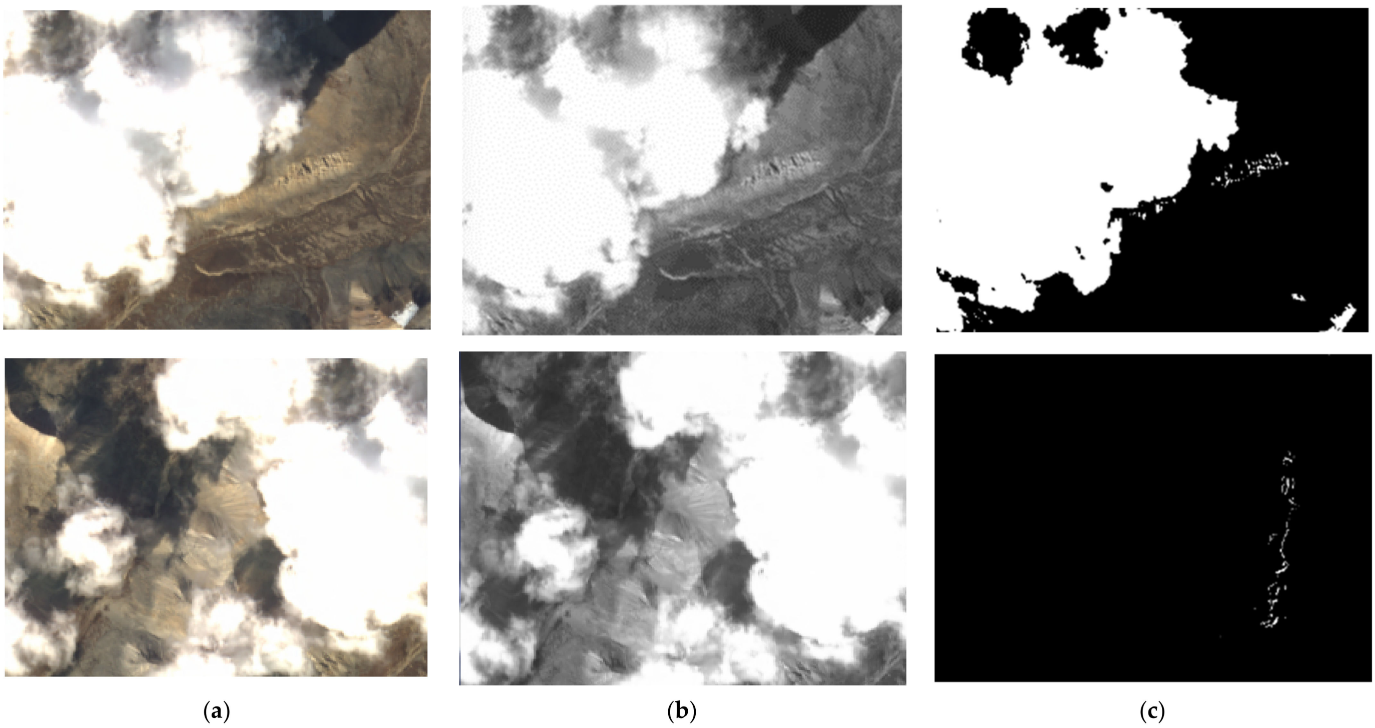

(**a**) (**b**) (**c**)

Figure 1. The segmentation examples of the Otsu algorithm. (**a**) Cloud; (**b**) binarization image; (**c**) segmentation result.

The cloud and snow segmentation method based on elevation assistance [4] captures the difference of elevation among cloud, snow, and other objects. The three-dimensional geometric features of cloud are obtained through the dense matching of multiple images, after which the differences between clouds and background objects are exploited by comparing the existing elevation information. Although this kind of methods has high accuracy, it involves many conditional and time-consuming operations such as dense matching and digital elevation model registration.

In recent decades, with the development of pattern recognition and machine learning technologies, researchers have studied intelligent methods for the cloud segmentation and have achieved good results. Amato et al. [5] applied a principal component analysis (PCA) for image cloud detection based on statistical theory. Merchant et al. [6] proposed a cloud detection algorithm based on full probability Bayes theory. Zhao Xiao [7] used fuzzy C-Means clustering to complete sample iteration clustering by minimizing the objective function and used support vector machine (SVM) to perform the classification, which has the advantage of better segmenting results under the condition of empirical knowledge, while human intervention greatly hinders the segmentation efficiency. Additionally, there are sparse perceptual classifiers, automatic codec, and other methods. Generally speaking, the time cost of machine learning based algorithms increases linearly with the number of pixels in the image. Therefore, for large-scale RSIs, the operation time of such an algorithm is often very long and it is hard to satisfy the needs in real-world applications.

Recently, deep learning has made great progress in image classification, image segmentation, object detection, and other vision tasks. In 2015, Long et al. [8] proposed the fully convolution neural network (FCN) and applied it to image semantic segmentation. Unlike the classical CNN, which uses the fully connected layer after the convolution layer to obtain a fixed length feature vector for classification, FCN operates on images of flexible size and performs pixel-level segmentation. For example, Shao et al. [9] proposed a method based on the multiscale feature model MF-CNN, which can detect thin clouds and thick clouds in RSIs, which results in good detection accuracy on the Landsat 8 data. To summarize, the methods based on deep learning are emerging in cloud and snow segmentation in RSIs.

ResNet [10] is a feature extraction backbone proposed by He et al., which is built on the idea of residual learning to solve the problem of gradient disappearance/explosion and network degradation in traditional CNN (when the number of network increases). ResNet module bypasses the input information to the output through an additional connection channel to protect the integrity of the input. The whole network only needs to learn the part of the difference between input and output, which simplifies the difficulty of learning when the network deepens and helps to retain more original semantic information.

DeepLab [11–13] are deep learning semantic segmentation models proposed by Google, which contain encoder–decoder FCN structure and the ASPP module to fuse multiscale features and make better use of image features (compared to the plain FCN). The encoder–decoder structure is reintroduced into DeepLabV3. Combining the Xception backbone [14] and the ASPP module as the encoder, which uses a different expansion rate of perforated convolution to extract features. After feature fusion and 4× upsampling, it fuses with the low-level features extracted by Xception in the decoder and 4× upsampling are used to get the segmentation results. Due to the use of ASPP in DeepLabV3+, it improves the ability of multiscale access to semantic information and achieves state-of-the-art accuracy on multiple datasets.

TH-1 [15], which is the first generation of the transmission stereo mapping satellite in China, carries three types of five camera loads, including a three linear array CCD camera, two-meter resolution panchromatic camera, and ten-meter RGB camera. Our research data consist of 470 tiles of RSIs with RGB bands taken by the TH-1 satellite from 2018 to 2019. A total of 200 tiles were collected on 20 September, 2018, while the others were collected on 5 April 2019. In order to ensure the generalization ability of the network, the selected images cover various underlying surfaces, acquisition seasons and time phases, considering different geographical locations, climatic conditions, and cloud features. The geographic longitude and latitude range is 28°35′ E–120°05′ E, 5°25′ N–60°05′ N, respectively, covering different ground surfaces such as deserts, grasslands, cities, and mountains. Additionally, in order to segment clouds and snow simultaneously, 175 scene images with snow are selected.

The threshold-based methods are still used in the actual production. To solve the above problems, we propose a cloud and snow segmentation method based on DCNN with an encoder–decoder structure. Compared with traditional methods, it does not rely on prior knowledge in feature selection and extraction. We combine the advantages of ResNet50 [10] and encoder–decoder structure and improve the decoder to realize the simultaneous segmentation of cloud and snow in RSIs. By improving the network structure, using the exponential activation function (ELU) [16] and focal loss function [17], our objective is to get optimization of cloud edge segmentation and enhancement of the generalization ability of the network. On the other hand, with the DCNN becoming the mainstream of image semantic segmentation, the production of a high-quality pixel level label for semantic segmentation has attracted more and more attention. However, accurate image annotation requires a lot of manpower and the existing datasets are not all fine-labeled. At present, there are more and more DCNN networks, and the accuracy of image semantic segmentation is higher and higher, which, however, based on open-source datasets such as Microsoft COCO [18] and PASCAL-VOC-2012 [19], is mainly studied to innovate the network structure and improve the segmentation accuracy without analysis on the impact of label quality and quantity on the results. So, the other direction of this paper is to explore the influence of different data quality and quantity on the performance of cloud and snow segmentation on RSIs. Our major research contributions are summarized as follows:

First, we propose an end-to-end DCNN framework with encoder–decoder architecture, ED-CNN, which improves the decoder by fusion of features from different encoding stages. The outputs of ASPP, which after Conv 1 × 1 and 4 × upsampling, are concatenated with the enhanced low-level features from the enhanced decoder. Then, the concatenated feature maps are sent to Conv 3 × 3 and 4 × upsampling to recover their original size to segment the image pixels. Second, we present a TH-1 satellite dataset, which contains 23,520 coarse-labeled images with annotations. Additionally, a fine-labeled dataset of 300 images is added to support our experiments. Third, experiments have been conducted based on different datasets, including TH-1 images of a different temporal phase and Google Earth images, which demonstrates that the proposed network is superior to DeepLabV3+ with Xception and the ResNet50 and can be applied to multisource RSIs. Finally, we discuss the effects of labeling quality and quantity in the dataset through extensive experiments with the proposed network. It is demonstrated that the performance of cloud and snow segmentation is positively related mainly to the labeling quantity. Namely, the smaller rough-labeled dataset plus some fine-labeled images, which is 10% of the total images, is equal to the larger rough-labeled dataset with the same total image quantity. Furthermore, under the same labeling consumption, the larger rough-labeled dataset exceeds the smaller rough-labeled dataset plus a few fine-labeled images.

2. Methodology

2.1. Datasets Establishment

First, we produced a BMP image. The data file of TH-1 RSIs was saved from the TIF format with four channels to the “BMP” image with a resolution of 6000 pixels × 6000 pixels. Second, the dataset was divided and the images were cut. The 470 tiles of images were divided into the training set, the validation set, and the testing set according to the numeric ratio of 3:1:1. Since the original image was large in size and took up too much memory, it could not be trained directly in the network. Therefore, we cut the images into patches with 480 pixels × 360 pixels. A total of 23,520 images were generated, including 13,924 training images, 4798 validation images, and 4798 testing images. Third, images were rough-labeled. Labelme [20] was used to roughly mark the cloud area of each image, generate JSON files, and convert the JSON files into label marked images with the same label size in batches. The rough-labeled image and its mask are shown in Figure 1, where the red color represents the cloud, green represents the snow, and black represents the background. Fourth, some images were fine-labeled. In order to verify the influence of the fine annotation image and rough annotation image on the training results (a detailed description is in Section 2.3.2), 310 extra images were randomly selected and then labeled carefully to generate JSON files with the time of 6 times that rough-labeling costs on each image, which were then transformed into labeled images as Step 3. Fine-labeled images had more accurate edge marking (errors less than 5 pixels). The fine-labeled image and its mask are shown in Figure 2, where label 1 represents the cloud, label 2 represents the snow, and label 0 represents the background located in the lower right corner of the image. Fifth, image preprocessing was performed. In Reference [21] it is proved that the network performance can be effectively improved by data augmentation. In order to enhance the generalization ability of the network and prevent overfitting, the dataset was augmented. The augmentation operations include a vertical flip, horizontal flip, contrast change, etc. The original and augmented images are shown in Figures 3 and 4.

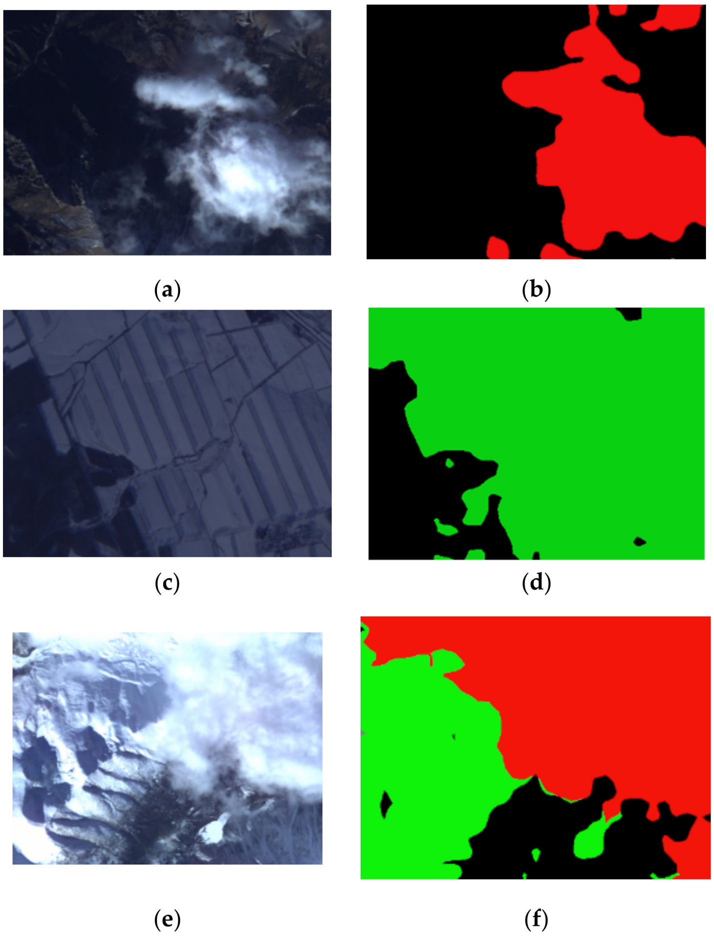


Figure 2. The coarse-labeled images and their masks. (**a**) Cloud; (**b**) mask; (**c**)snow; (**d**) mask; (**e**) cloud and snow; (**f**) mask.

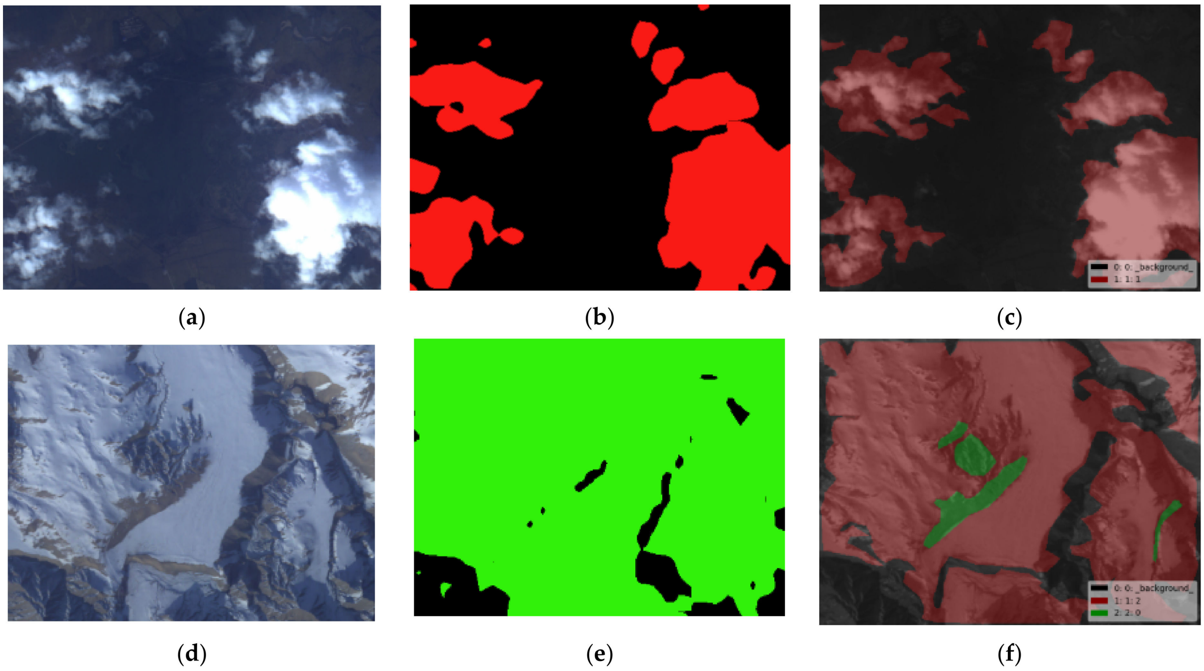


Figure 3. *Cont.*

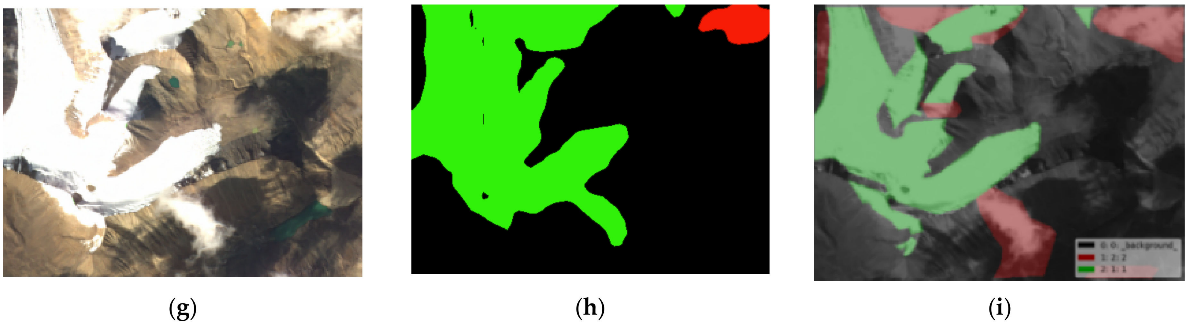


Figure 3. The fine-labeled images and their masks. (**a**) Cloud, (**b**) rough mask, (**c**) fine mask, (**d**) snow, (**e**) rough mask, (**f**) fine mask, (**g**) cloud and snow, (**h**) rough mask, and (**i**) fine mask.

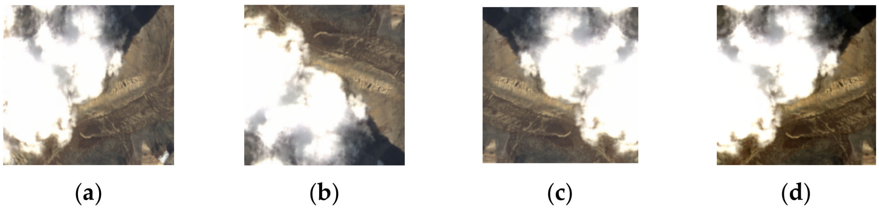


Figure 4. Original and augmented images. (**a**) Original image, (**b**) vertical flip, (**c**) horizontal flip, and (**d**) contrast change.

2.2. Methods

2.2.1. Network Backbone

How to extract features from a different kind of RSIs properly and effectively is a key problem. Cloud and snow in RSIs mostly present a planar structure. Their semantic information is simple, whereas their detailed information is rich, which puts forward a high demand for the detail extraction ability. For example, with the parameter amount of 22.8 M [14], Xception has a large number of parameters suitable for segmentation tasks with many kinds of objects, Meanwhile, it requires huge computational resources and is difficult to train, thus it is obviously not fully suitable for the task of cloud and snow segmentation. In this paper, Resnet50 backbone was selected as the encoder to extract features of cloud and snow (as shown in Figure 5). The parameter size of ResNet50 was only 0.85 M [10] and more direct connections were added in the network. Considering its advantages such as less parameters, easy training, and fast convergence, it is more suitable for cloud and snow segmentation compared to Xception.

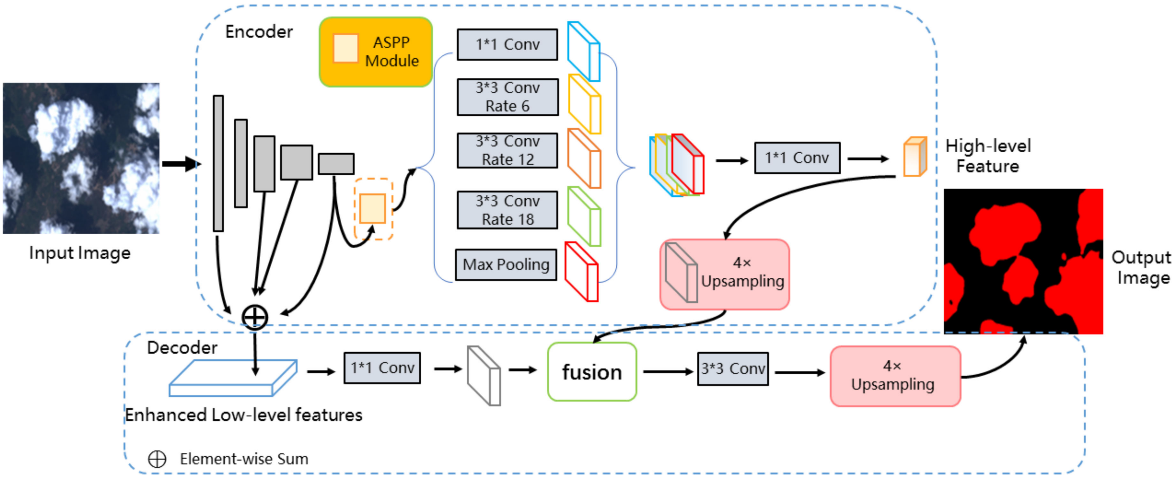


Figure 5. Network architecture of the proposed method.

2.2.2. Enhanced Decoder

In encoder–decoder architectures such like DeepLabV3+, the decoder subnet gradually recovers the spatial information, which is usually not as powerful as the encoder. In this regard, besides replacing the backbone to the ResNet50, we added skip connections in the decoder. To be specific, we selected features in the stages 1, stage 3, stage 4, and stage 5 of the ResNet50 to construct a top–down connection feature map pyramid, which enriched the semantic representation of low-level features to better utilize the spatial information as shown in Figure 5. The low-level feature maps with high resolution and high-level feature maps with rich semantic information were fused, which can quickly construct the decoder with better semantic information from 4 stages instead of a single stage without obvious cost increases.

As Figure 6 shows, first the feature map from stage 5 was 2 × upsampled after 1 × 1 convolution, which was added with the output from stage 4, after 1 × 1 convolution too. The dimension numbers of these stages were all set to 256, ensuring that these feature maps could be added. Second, the added feature map was 2 × upsampled the second time to be the same size of the output from stage 3, then the process was repeated to get the fused feature map with output from stage 1. Third, the feature map of the enhanced decoder was concatenated with the feature map generated from the ASPP module after the 3 × 3 convolution, batch-normalization, and ELU operations. Finally, the segmentation map was obtained by 3 × 3 convolution and 4 × upsampling.

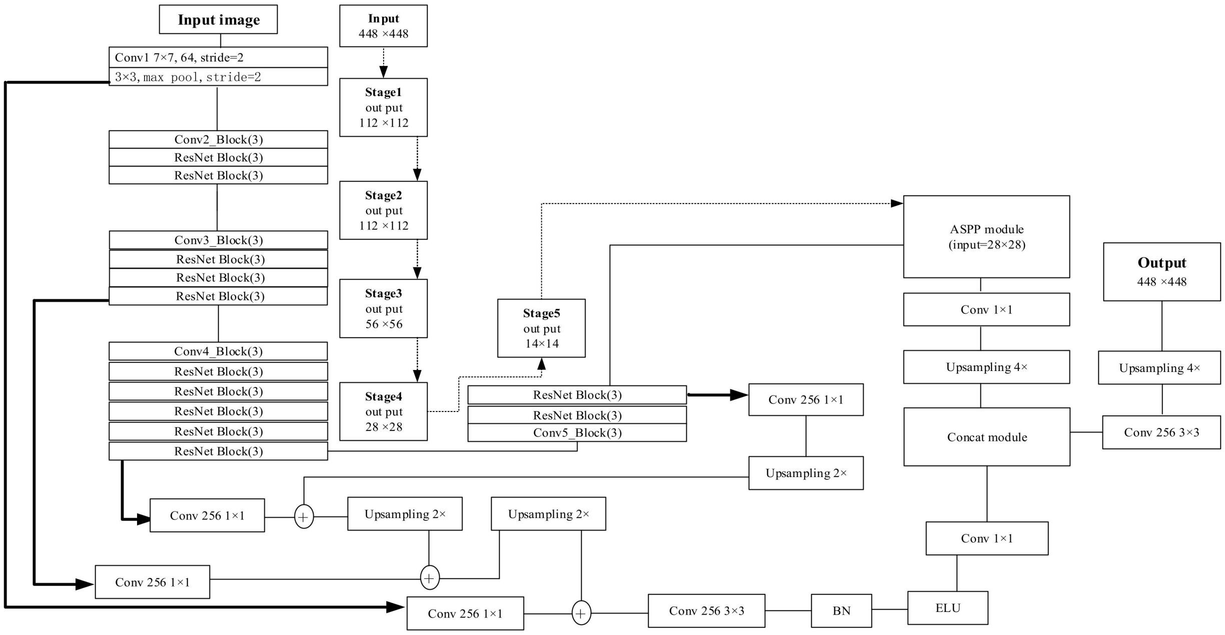


Figure 6. Network workflow of the proposed method (⊕ means element-wise sum).

2.2.3. Loss Function

In practice, there are many kinds of clouds with different shapes. Generally, the proportion of thin clouds and cirrus clouds is less than that of thick clouds. The training dataset in this paper also reflected the characteristics of less data of thin clouds, cirrus clouds, and snow. Cross entropy (CE) loss cannot balance the learning of fewer samples. Its formula is as follows,

$$CE(p_t) = -\log(p_t) \tag{1}$$

Through the combination of different parameters, focal loss [17] can solve the problem of sample imbalance in the semantic segmentation task, which is an improved version of the CE loss by adding a weight. Its formula is as follows:

$$FL(p_t) = -(\lambda - p_t)^{\gamma} \log(p_t) \tag{2}$$

where λ and γ are two hyperparameters and p_t is the prediction probability of the label. $(\lambda - p_t)$ can be regarded as the weight of Equation (1). The paper [17] sets γ = 2 and λ = 1, when the prediction of a certain category is accurate, that is, close to 1, the value is close to 0. The more inaccurate the prediction is, that is, close to 0, the closer to 1 it will be. For the samples that are easy to distinguish, the weight corresponding to the loss will be small, whereas for objects that are difficult to distinguish, their corresponding weight will be larger in order to retain the loss value of difficult samples and reduce the loss value of simple samples. We set γ = 2 and λ = 1.

2.2.4. Activation Function

The exponential linear unit (ELU) was proposed in the paper [16], which can make the mean value of output close to 0, thus accelerating the convergence of the network and effectively overcome problems such as gradient vanishing. If the output of a node is X, the output after passing through the ELU layer is illustrated in Equation (3). We adopted ELU activation after convolution layers.

$$f(x) = \begin{cases} x, x \geq 0 \\ \alpha(e^x - 1), x < 0. \end{cases} \tag{3}$$

2.2.5. Evaluation Metrics

Here we used pixel accuracy (PA) and the mean intersection over union (MIOU) as the evaluation metrics.

1. P.A: the simplest and direct indicator, which only calculates the ratio of the number of correctly classified pixels to the number of all pixels. The calculation is shown in Equation (4).

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{4}$$

2. MIOU: it calculates the coincidence ratio between the intersection and union of two sets, that is, the intersection union ratio between real segmentation and algorithm segmentation. This ratio can be redefined as the number of true positive cases (intersections) divided by the total number (including true positive cases, false negative cases, and false positive cases (Union)). MIOU is calculated by class, and then averaged. The calculation is shown in Equation (5).

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}.} \tag{5}$$

2.3. Experiments

2.3.1. Experimental Platform

The experimental hardware is the Lenovo workstation with 2.1 GHz CPU frequency of the Intel Xeon processor, and the GPU is NVIDIA Titan XP. We used the Tensorflow + Keras framework to build deep learning models. The network training was carried out on the basis of data augmentation. The initial learning rate was set to 3e^{-4} and the learning rate decreasing drop was set to 0.1. Using the ELU activation function and Adam [22] optimizer, the batch size was set to 8 according to the capability of the GPU.

2.3.2. Group Experiment Settings

We set the experiment into two stages numbered from 0 to 4 shown in Table 1. First, the performance of a different network was evaluated on dataset 1, which included 23,520 images. Second, another smaller dataset 2 including 6660 rough-labeled images and 310 fine-labeled images was chosen from dataset 1 to explore the influence of the quantity and quality of image labels on the accuracy of cloud and snow segmentation.

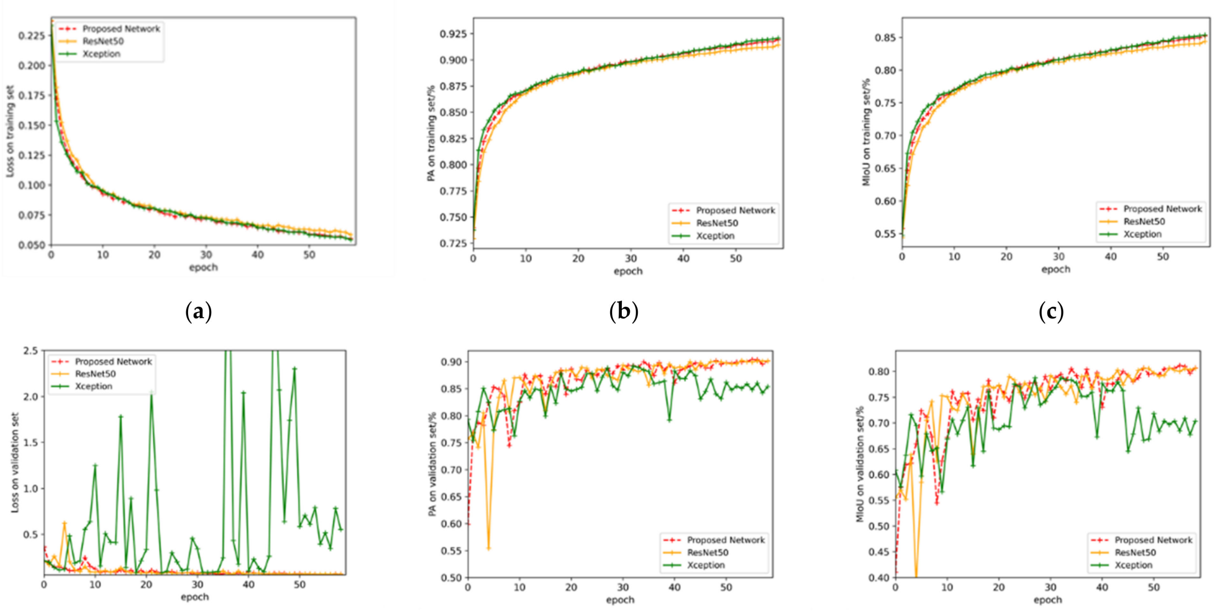
Table 1. Experiment setting.

Group	Group Number	Data	Total	Train	Validation	Test
Stage 1 Dataset 1	0	Rough-labeled	23,520	13,924	4798	4798
Stage 2 Dataset 2	1	Rough-labeled	5000	3000	1000	1000
		Fine-labeled		0	0	0
	2	Rough-labeled	4700	2700	1000	1000
		Fine-labeled		0	0	0
	3	Rough-labeled	5010	2700	1000	1000
		Fine-labeled		310	0	0
	4	Rough-labeled	6660	4660	1000	1000
		Fine-labeled		0	0	0

3. Results and Analysis

3.1. Training Process and Results

First, we used dataset 1 to train the proposed network, ResNet50, and Xception for 60 epochs, respectively. Figure 7 shows the loss and accuracy comparison between three networks on the training set and the validation set. The red line represents the proposed method, while the yellow and green lines represent ResNet50 and Xception, respectively.

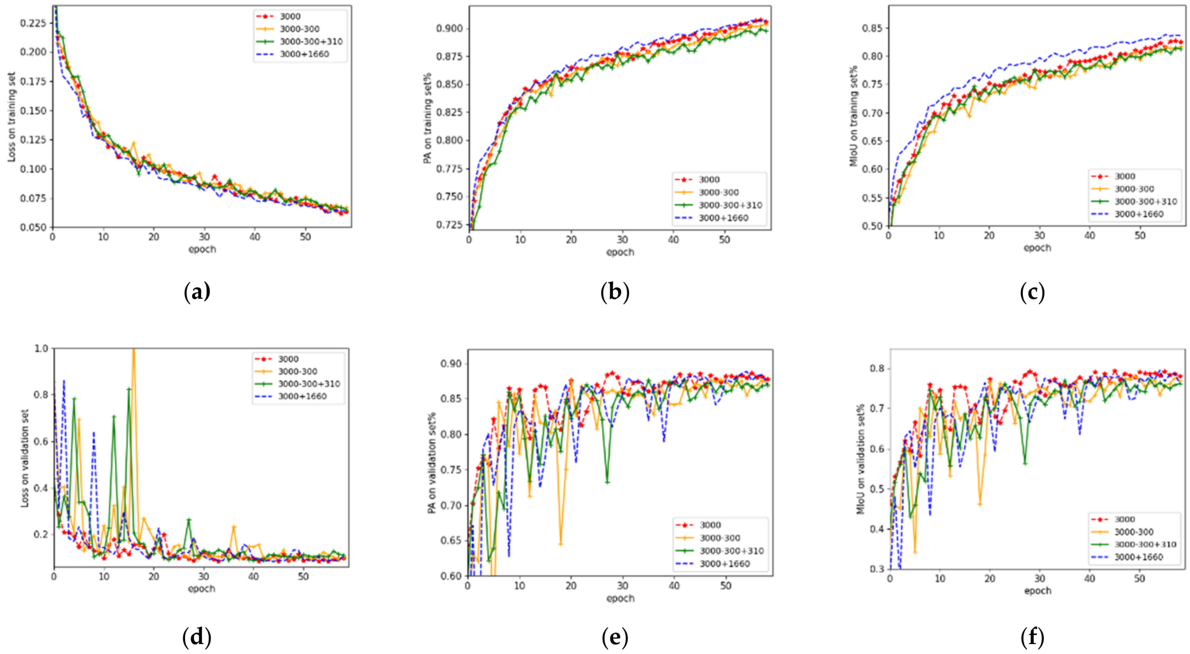


(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 7. Results of three networks on dataset 1. (**a**) Training loss; (**b**) training PA; (**c**) training MIoU; (**d**) validation loss; (**e**) validation PA; (**f**) validation MIoU.

Second, dataset 2 of 6660 images was randomly chosen from set 1, dividing into 4660 for the training set, 1000 for the validation set, and 1000 for the testing set. Another 310 fine-labeled images were added to the training set. Four group of experiments were performed to explore the influence of data use strategy on the proposed network. Figure 8 shows the results.



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

(**d**) (**e**) (**f**)

Figure 8. Results of the proposed network on dataset 2. (**a**) Training loss; (**b**) training PA; (**c**) training MIoU; (**d**) validation loss; (**e**) validation PA; (**f**) validation MIoU.

3.2. Comparison and Analysis

First, the testing results of the three methods on dataset 1 are shown in Table 2. In the testing set of 4798 images, the PA of cloud and snow segmentation obtained by the proposed method was 90.3%, while the PA of Xception, ResNet50, and DANet [23] were 87.4%, 89.2%, and 88.9% respectively. The MIoU of the proposed method was 81.1%, exceeding Xception and ResNet50 by 4.2%, 0.6%, and 0.8% respectively. When the proposed enhanced decoder was used to enhance low-level features, the PA and MIoU of the network increased obviously.

Table 2. Results of the three networks on dataset 1.

Method	PA	MIoU
Xception	87.4	76.9
ResNet50	89.2	80.5
DANet	88.9	80.3
ED-CNN	90.3	81.1

Second, four groups of experiments were conducted on dataset 2 to explore the influence of different data quality and quantity on the proposed network. The results are shown in Table 3, in which group 1 and group 2 proved that a 10% reduction in the number of training data will reduce the overall performance. On the contrary, group 2 and group 3 show that a 10% increase of fine-labeled data replacing the reduction had a slight side effect on performance. In our opinion, adding different features of data when the training set is not large enough will lead to the above problem. Additionally, on average, it took six times as long to label a fine-labeled image than making a rough-labeled image. However, group 4 shows that 4660 rough-labeled images resulted in better results compared to the other three groups.

Table 3. Results of a different data use strategy on dataset 2.

Number	Data Use			PA	MIoU
	Rough	Fine	Testing		
1	3000	0	1000	88.5	79.3
2	2700	0		87.4	78.0
3	2700	310		87.2	77.4
4	4660	0		88.8	79.4

Finally, the qualitative analysis of typical images was conducted to compare the results of the proposed method with the Otsu and another two networks (Xception and ResNet50). Figure 9 shows the comparison results of the traditional Otsu method versus the three networks, and the red part is the cloud while the green part is snow. It can be found that when the image illumination conditions were not ideal and the underlying surface contrast was not high, the Otsu created a false alarm in Figure 9b and it could not segment cloud and snow at the same time (as in Figure 9d). The reason is that the Otsu algorithm does not consider the neighborhood information in segmentation and is sensitive to noise. In Figure 9a,c, Xception missed some pixels and got the wrong result in Figure 9b, while getting it totally wrong in Figure 9d. The performance of the proposed network was slightly better than that of the ResNet50 in preserving the spatial details.

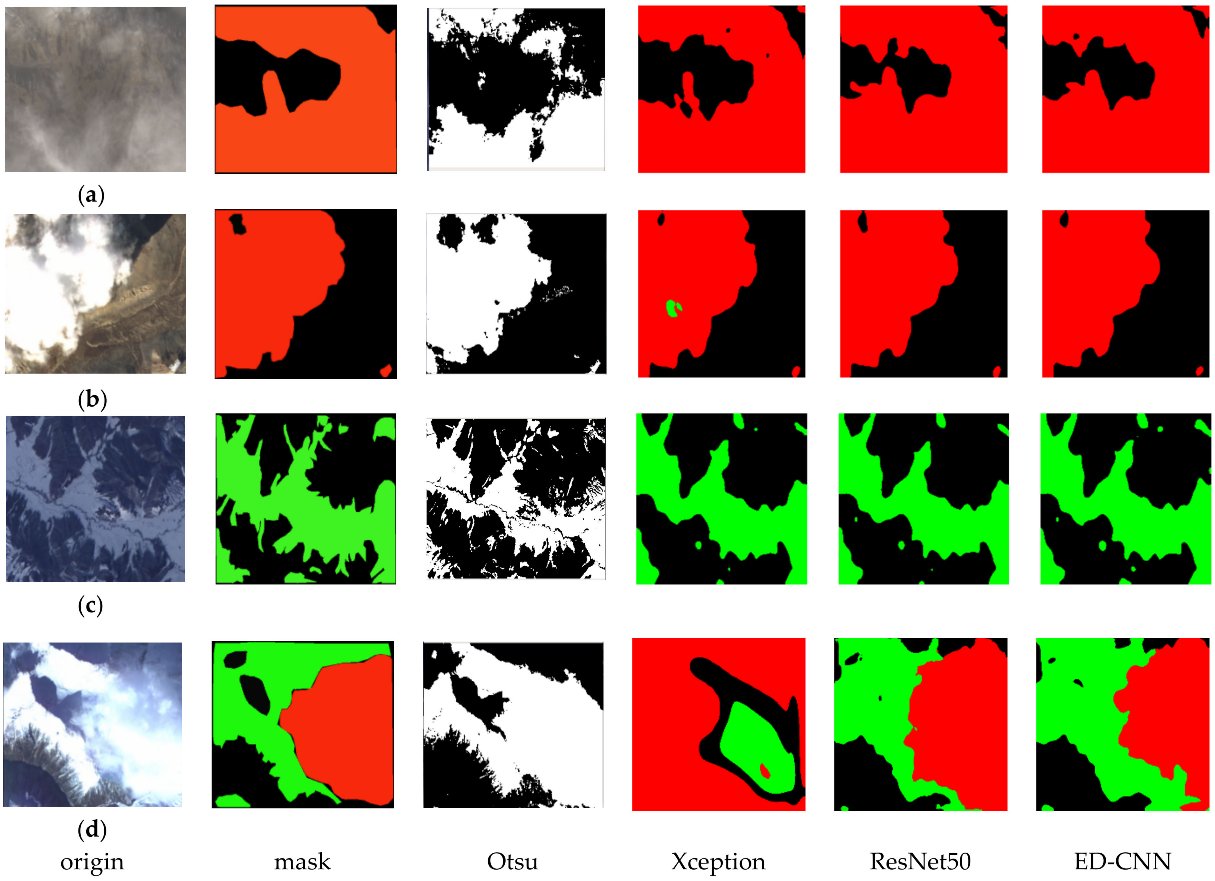


Figure 9. Testing results of the proposed method on the TH-1 image acquired from different temporal phases.

Furthermore, TH-1 images from different time-phases are chosen to verify the generalization performance of the proposed method. The results are shown in Figure 10. It is demonstrated that the proposed method could accurately segment cloud and snow, showing a better generalization ability.

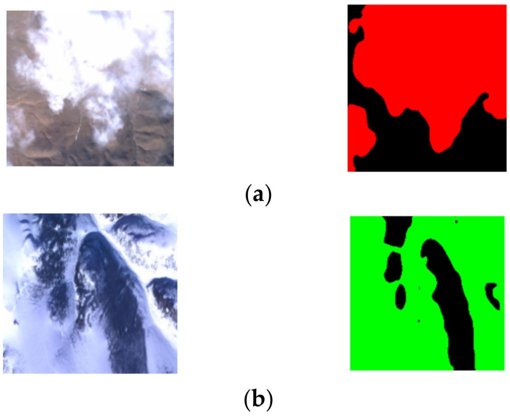


Figure 10. Testing results of the proposed method on the TH-1 image acquired from different temporal phases. (**a**) Cloud; (**b**) snow ground.

4. Conclusions

4.1. The Proposed Method for Cloud and Snow Segmentation

In this paper, we proposed an end-to-end cloud and snow segmentation network for TH-1 RSIs, which combined the advantages of the encoder–decoder architecture and the enhanced decoder. On the one hand, it avoided the shortcomings of traditional cloud detection algorithms (such as they are parameter-dependent, time-consuming, and the scope of application is limited). It achieved the mIoU of 81.1% on 4798 testing images and reduced the segmenting time (on a single 480 × 360 image) to 49.2 ms, which could basically meet the requirements of image preprocessing. On the other hand, the enhancement of the decoder proved to be a useful way to improve the segmentation performance through exploiting features at different encoder stages, which bridges the gap between different levels of features. Additional experiments show that this method can be used on images acquired from different sensors, as shown in Figure 11.

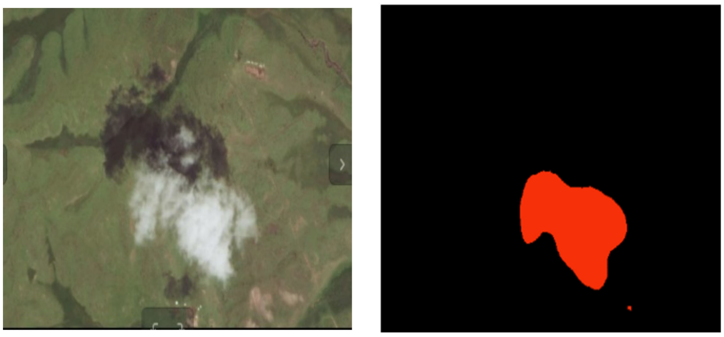


Figure 11. Testing results of the proposed method on a Google Earth image.

There is still room for further improvement in the cloud and snow segmentation method proposed in this paper, such as improving the training accuracy and using additional multisource RSIs for transfer learning. Another research direction is to use multispectral information to solve the problem of distinguishing cloud, fog, and snow in mixed regions.

4.2. Influence of Different Datasets on Segmentation Performance

Given a certain network, we found that its segmentation performance was positively related mainly to the number of training images and labels. Specifically, when the training time was sufficient, more training images led to higher accuracy, whereas a 10% increase of fine-labeled data replacing the original rough-labeled reduction had a slight side effect on performance. Considering fewer categories and a lower complexity of cloud and snow segmentation, our conclusion was that when the same labeling time was considered, we achieved better results by only roughly labeling the data. Instead of spending more manual resources to make fine-labeled masks, roughly labeling more data can lead to the same segmentation accuracy.

There is a margin for further research on the effects of label quality and quantity, such as clarifying the pixel error of coarse-marking labels and exploring the effect of error types and sizes on cloud and snow segmentation results.

Author Contributions: Conceptualization, Kai Zheng and Jiansheng Li; methodology, Kai Zheng; software, Kai Zheng and Lei Ding; validation, Kai Zheng and Jianfeng Yang; formal analysis, Xun Zhang; investigation, Xucheng Zhang; resources, Xucheng Zhang; data curation, Jianfeng Yang; writing—original draft preparation, Kai Zheng; writing—review and editing, Jiansheng Li; visualization, Lei Ding; supervision, Jiansheng Li. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Exclude.

Informed Consent Statement: Exclude.

Data Availability Statement: Not available.

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

References

1. Zhu, Z.; Woodcock, C.E. Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change. *Remote Sens. Environ.* **2014**, *152*, 217–234. [CrossRef]
2. Jian, H. Cloud and Snow Detection Algorithm and Surface Reflectance Database Construction Algorithm of Remote Sensing. Ph.D. Thesis, Henan University, Kaifeng, China, 2017.
3. Xianjun, G.; Youchuan, W.; Shunyi, Z.; Yuan-wei, Y. Real time automatic detection of clouds in aerial photography. *Spectrosc. Spectr. Anal.* **2014**, *34*, 1909–1913. [CrossRef]
4. Shi, T.; Yu, B.; Clothiaux, E.E.; Braverman, A.J. Daytime Arctic Cloud Detection Based on Multi-Angle Satellite Data with Case Studies. *J. Am. Stat. Assoc.* **2008**, *103*, 584–593. [CrossRef]
5. Amato, U.; Antoniadis, A.; Cuomo, V.; Cutillo, L.; Franzese, M.; Murino, L.; Serio, C. Statistical cloud detection from SEVIRI multispectral images. *Remote Sens. Environ.* **2008**, *112*, 750–766. [CrossRef]
6. Merchant, C.J.; Harris, A.R.; Maturi, E.; Maccallum, S. Probabilistic physically based cloud screening of satellite infrared imagery for operational sea surface temperature retrieval. *Q. J. R. Meteorol. Soc.* **2005**, *131*, 2735–2755. [CrossRef]
7. Xiao, Z. Research on Cloud Detection Method of High Resolution Satellite Remote Sensing. Ph.D. Thesis, Harbin Institute of Technology, Harbin, China, 2013.
8. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651.
9. Shao, Z.; Pan, Y.; Diao, C.; Cai, J. Cloud Detection in Remote Sensing Images Based on Multiscale Features-Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4062–4076. [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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
11. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
12. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
13. 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), 2018—15th European Conference, Munich, Germany, 8–14 September 2018; pp. 801–818.
14. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
15. Songming, L.; Yan, L.; Jindong, L. TH-1 Transmission Photogrammetry and Remote Sensing Satellite. *Acta Remote Sens. Sin.* **2012**, *16*, 10–16. [CrossRef]
16. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv* **2015**, arXiv:1511.07289.

17. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007. [CrossRef]
18. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
19. Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
20. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [CrossRef]
21. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
22. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
23. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.