symmetry

MDPI

Article

Cloud Detection for Satellite Imagery Using Attention-Based U-Net Convolutional Neural Network

Yanan Guo [1,2], Xiaoqun Cao [1,2,*], Bainian Liu [1,2] and Mei Gao [1,2]

[1] College of Meteorology and Oceanography, National University of Defense Technology, Changsha 410073, China; guoyanan@nudt.edu.cn (Y.G.); bnliu@nudt.edu.cn (B.L.); gaomei17@nudt.edu.cn (M.G.)

[2] College of Computer, National University of Defense Technology, Changsha 410073, China

* Correspondence: caoxiaoqun@nudt.edu.cn

Received: 7 June 2020; Accepted: 23 June 2020; Published: 25 June 2020

Abstract: Cloud detection is an important and difficult task in the pre-processing of satellite remote sensing data. The results of traditional cloud detection methods are often unsatisfactory in complex environments or the presence of various noise disturbances. With the rapid development of artificial intelligence technology, deep learning methods have achieved great success in many fields such as image processing, speech recognition, autonomous driving, etc. This study proposes a deep learning model suitable for cloud detection, Cloud-AttU, which is based on a U-Net network and incorporates an attention mechanism. The Cloud-AttU model adopts the symmetric Encoder-Decoder structure, which achieves the fusion of high-level features and low-level features through the skip-connection operation, making the output results contain richer multi-scale information. This symmetrical network structure is concise and stable, significantly enhancing the effect of image segmentation. Based on the characteristics of cloud detection, the model is improved by introducing an attention mechanism that allows model to learn more effective features and distinguish between cloud and non-cloud pixels more accurately. The experimental results show that the method proposed in this paper has a significant accuracy advantage over the traditional cloud detection method. The proposed method is also able to achieve great results in the presence of snow/ice disturbance and other bright non-cloud objects, with strong resistance to disturbance. The Cloud-AttU model proposed in this study has achieved excellent results in the cloud detection tasks, indicating that this symmetric network architecture has great potential for application in satellite image processing and deserves further research.

Keywords: cloud detection; remote sensing images; U-Net architecture; attention mechanism; deep learning; convolutional neural network

1. Introduction

With the development of remote sensing technology, a large number of high-resolution satellite data have been obtained, which can be applied to land cover monitoring, marine pollution monitoring, crop yield assessment, and other fields [1–4]. However, the presence of clouds, especially very thick ones, can contaminate captured satellite images and cause interference in the observation and identification of ground objects [5–7]. As a result, clouds create a lot of difficulties for tasks such as target identification, trajectory tracking, etc. On the other hand, clouds are the most uncertain factor in Earth's climate system. It is estimated that clouds cover about 67% of the Earth's surface [8,9]. Clouds can affect the energy and water cycles of global ecosystems at multiple scales by influencing solar irradiation transmission and precipitation, and thus have a significant impact on climate change

on Earth [10]. At the same time, cloud observations and forecasts are important for the management of the power sector, as cloud coverage affects the use of solar energy [11]. Therefore, cloud detection is of great research significance, both in various remote sensing applications and in fields such as earth science research.

Over the years, many cloud detection methods [12–14] have been proposed for satellite remote sensing data. Among the many methods, the threshold method is a relatively well-established cloud detection method. It accomplishes cloud detection task based on the radiative difference between cloud and non-cloud pixels. For example, the function of mask (Fmask) [15–17] is a typical threshold-based cloud detection method that uses a decision tree to mark each pixel as cloud or non-cloud. In each branch of the decision tree, decisions are given according to a threshold function. In practice, threshold-based methods are often not used in isolation and often rely on multi-spectral and multi-method combinations, such as adding other information (e.g., surface temperature, etc.) or combining other techniques (e.g., superpixel segmentation [18] and texture analysis [19], etc.) to improve the detection accuracy of threshold methods. Most of these methods are specific to specific bands of remote sensing data, and the selection of thresholds is often difficult and time-consuming. Although various threshold methods have achieved some success in their respective applications, most methods are not generalizable and require constant adjustment and optimization of the threshold selection.

In recent years, deep learning has made great breakthroughs in the field of computer vision [20–22], with remarkable results in areas such as face recognition, object detection and medical image analysis. Inspired by deep learning algorithms, many scholars have developed several approaches to cloud detection using deep learning. Convolutional neural network (CNN) is a widely used deep learning method that has unique advantages in the field of image processing. Shi et al. [5] utilized superpixel segmentation and deep Convolutional Neural Networks (CNNs) to mine the deep features of cloud. The experimental results show that their proposed model works well for both thin and thick clouds, and has good stability in complex scenarios. Chen et al. [23] implemented a multilevel cloud detection task for high-resolution remote sensing images based on Multiple Convolutional Neural Network s(MCNNs). Specifically, MCNNs architecture is used to extract multiscale information from each superpixel, next superpixels are classified as thin clouds, thick clouds, cloud shadows, and non-clouds. The results show that the proposed method has an excellent performance in the task of detecting multilevel cloud detection. Segal-Rozenhaimer et al. [24] proposed a novel domain adaptation CNN-based method, which utilizes the spectral and spatial information inherent in remote satellite imagery to extract the depth invariant features for cloud detection. The method can be better adapted to different satellite platforms in the prediction step without the need to train separately for each platform, improving the robustness of predictions from multiple remote sensing platforms. Ozkan [25] et al. proposed an efficient neural network model based on a deep pyramid network. In the task of cloud detection, the model can obtain very good classification results from a set of noisy labeled RGB color remote sensing images with accuracy up to pixel level. Francis [26] et al. proposed a CloudFCN model for cloud detection tasks based on the Fully Convolutional Network architecture. The experimental results show that this model works well in cloud detection, illustrating the great potential of the Fully Convolutional Network architecture for applications in satellite remote sensing data. These studies all show that deep learning methods have great advantages in satellite remote sensing data processing, and cloud detection methods that introduce convolutional neural networks are often superior to traditional cloud detection methods. U-Net [27–29] is a very effective image segmentation method that has a remarkable performance in many image segmentation tasks, especially medical image segmentation tasks. Many studies have found that models based on the U-Net architecture also show excellent performance in satellite remote sensing image segmentation [30–32]. For instance, Jeppesen [33] et al. propose a cloud detection algorithm (RS-Net) based on the U-net architecture and use it to detect clouds in satellite images. From the experimental results, it was found that the RS-Net performed better than the conventional method. Wieland et al. [34] presented an improved U-Net

convolutional neural network for the task of multi-sensor cloud and cloud shadow segmentation. Their experimental results show that the model achieves great results on multiple satellite sensors with excellent generalization performance. Besides, adding shortwave-infrared bands can improve the accuracy of the semantic segmentation task of cloud and cloud shadow. In cloud detection tasks, the number and distribution of clouds tend to present very complex randomness. In order to achieve accurate cloud detection, attention should be focused on the areas with clouds during the cloud detection process. In the field of medical image classification, the object detection, etc., attention mechanism is a very effective method [35–38] that can allocate more processing resources to the target. Attention mechanism originates from human beings visual cognitive science. When reading text or looking at objects, humans tend to pay more attention to detailed information about the target and suppress other useless information. Similarly, the basic idea of the attention mechanism is that the model learns to focus on important information and ignore the unimportant information. Some studies have shown that attention mechanisms can improve classification effects [39–42]. Therefore, attentional mechanisms in computer vision should be introduced into the construction of cloud detection models. In this study, the Cloud-AttU model is proposed on the basis of U-Net, which introduces the attention mechanism. In the experiment, it was found that the results of cloud detection were significantly improved as the attention mechanism guided the model to learn more cloud-related features to detect clouds.

The structure of this paper is as follows: the U-Net architecture, the attention mechanism, and the Cloud-AttU model proposed in this paper are described in Section 2. Subsequently, the experiment design and results are given in Section 3. The results of the experiment are discussed in Section 4. Finally, Section 5 summarizes the work of this paper and lists the work to be continued in the future.

2. Methodology

To better implement cloud detection, this study applies the attention mechanism to the U-Net network. This improvement allows the model to better focus on areas with clouds and ignore areas without clouds during the cloud detection process, ultimately improving the accuracy of cloud detection. In the following, we describe the U-Net architecture, the attention mechanism, and the Cloud-AttU method proposed in our study, respectively.

2.1. U-Net Architecture

U-Net [27] was developed on the basis of the Fully Convolutional Network architecture [43] and was first applied to biomedical image segmentation in 2015. U-Net uses a symmetric encoder-decoder structure, which is one of the most popular methods for semantic medical image segmentation. The diagram of the basic U-Net architecture is shown in Figure 1. The left half of Figure 1 is the encoding path and the right half is the decoding path. Specifically, each block in the encoding path of the U-Net contains two convolutional layers and a max-pooling layer. At each step of the downsampling, the number of feature channels is doubled. In addition, the effect of max-pooling makes the size of the feature map smaller. Each block in the decoding path of U-Net contains the up-sampling, the fusion operation, as well as two convolutional layers. As shown in Figure 1, the skip-connection transfers information from the encoding path to the decoding path, thereby improving the ability of the U-Net to segment the image. Finally, for the multi-channel feature maps, a 1×1 convolution is used to obtain segmentation results.

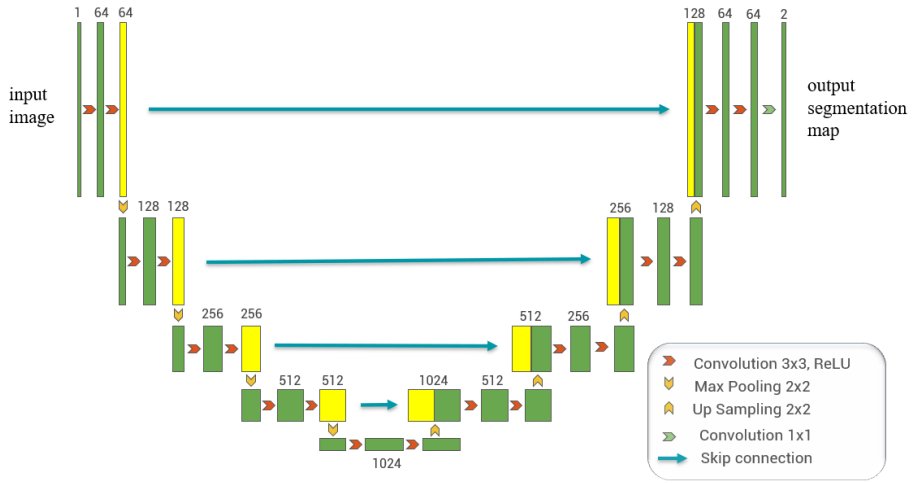

Figure 1. U-Net architecture diagram modified from the original study [27]. Green/yellow boxes indicate multi-channel feature maps; red arrows indicate 3 × 3 convolution for feature extraction; cyan arrows indicate skip-connection for feature fusion; downward orange arrows indicate max pooling for dimension reduction; upward orange arrows indicate up-sampling for dimension recovery.

2.2. Proposed Network Architecture

In this section, we describe in detail the proposed Cloud-AttU model for cloud detection tasks. The proposed architecture is obtained by modifying the original U-Net model, and the new model introduces the attention mechanism. The overall structure of Cloud-AttU model is shown in Figure 2, and like the original U-Net model, Cloud-AttU consists of two main paths: the contracting path that encodes the entire input image, and the extending path enables the feature maps to return to its original size by gradually step-wise up-sampling. Each step of the encoder consists of a structural dropout convolutional block and a 2 × 2 max pooling operation. As shown in Figure 2, each convolutional layer is followed by a DropBlock, a batch normalization (BN) layer, and a rectified linear unit (ReLU). The max pooling operation is then applied for down-sampling with a stride size of 2. At each down-sampling step, the number of feature channels is doubled. Each step of the decoder consists of a 2 × 2 transposition convolution operation which completes up-sampling, a concatenation operation with the matching feature map of the encoder, and then followed by a structural dropout convolutional block. Finally, the 1 × 1 convolution and Sigmoid activation functions are utilized to yield the final segmentation map.

In our proposed Cloud-AttU model, the encoder shares information about its feature map with the decoder by skipping connections. By the skip connections from the lower to the higher layers, the model has more information to carry out its tasks. However, too much invalid or interfering information can affect the performance of the model. The attention module enables the network to learn the useful parts of the input image and what important features it should focus on to accomplish its task. We weight the feature maps using an attentional gate [44] to emphasize relevant features and ignore invalid features. In attention gates, the attention coefficient α_i highlights signal in the target region and weakens background signal. Thus attention gates do not need to cut the region of interest directly in the network but achieve suppression of information from irrelevant background regions. The architecture of attention gate which we used is shown in Figure 3. As shown in Figure 3, the attention gate has two inputs: the feature maps (g) in the decoder and the feature maps (f) in the encoder. The feature maps (g) are employed as gated signals to augment the learning of the feature maps (f). In summary, this gating feature (g) makes it possible to extract more useful features from the encoded features (f), while ignoring invalid features. The two inputs are merged pixel-by-pixel after convolutional operations (W_g, W_f) and Batch Normalization (b_g, b_f), respectively. The results obtained in the previous step are then activated using the Rectified Linear Unit ($ReLU, \sigma_1(x) = max(0, x)$).

For the new results, we perform the convolutional operations (W_θ) and Batch Normalization (b_θ) on them. Next we use the sigmoid activation function $\sigma_2(x) = \frac{1}{1+e^{(-x)}}$ to train the parameters in the gate and get the attention coefficient (α). Finally, the output of the attention gate is gained by multiplying the encoder feature by the attention coefficient (α). The feature selection process in the attention gate can be expressed by the following formula:

$$F = \sigma_1 \left[\left(W_f^T \times f + b_f \right) + \left(W_g^T \times g + b_g \right) \right] \tag{1}$$

$$\alpha = \sigma_2 \left(W_\theta^T \times F + b_\theta \right) \tag{2}$$

$$\text{output } = f \times \alpha \tag{3}$$

For image segmentation, the overall performance of the segmentation model is affected both by the architecture of the network and the loss function [45]. If the loss function is used differently during training, then the training model obtained may differ significantly. Especially in the case of high-class imbalance problems. Therefore, choosing a suitable loss function becomes more challenging. In many cases, only a small fraction of the pixels in satellite remote sensing images are clouds, where the ratio of cloud pixels to background pixels varies widely. If the loss function chosen does not adequately take into account the properties of cloud pixels, the model is prone to judge the background region as a cloud pixel during learning, resulting in incorrect pixel segmentation of the image. In the Cloud-AttU network, we choose the cross entropy loss function to train the model and optimize the loss function by Adam algorithm [46,47], and good results have been obtained. In future studies, we will choose other loss functions for further research and optimize the current loss function. The mathematical expression for the cross entropy loss function is defined as follows:

$$J(y, \hat{y}) = -\frac{1}{n} \sum_{i=1}^{n} y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i) \tag{4}$$

where n represents the number of pixels in each image, y_i represents the ground truth and $\hat{y}_i$ is the prediction result.

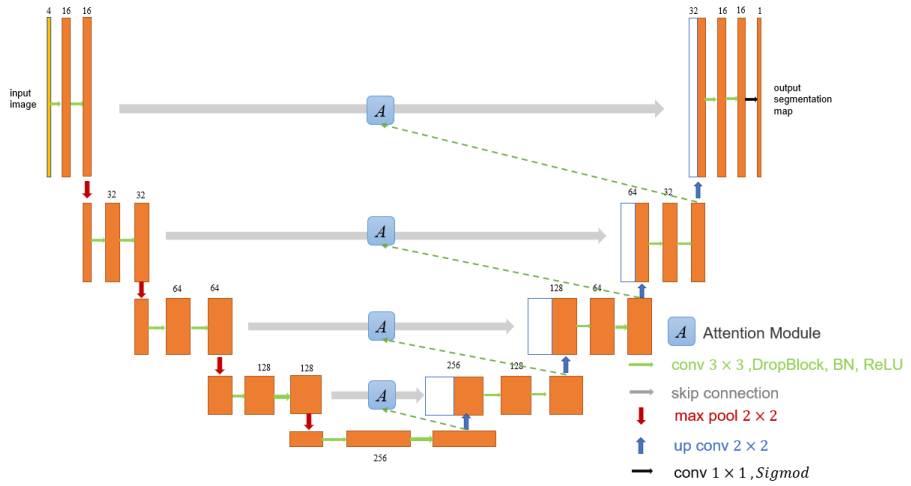

Figure 2. The structure of the Cloud-AttU model. All the orange/white boxes correspond to multi-channel feature maps. The Cloud-AttU is equipped with skip connections to adaptively rescale feature maps in the encoding path with weights learned from the correlation of feature maps in the decoding path.

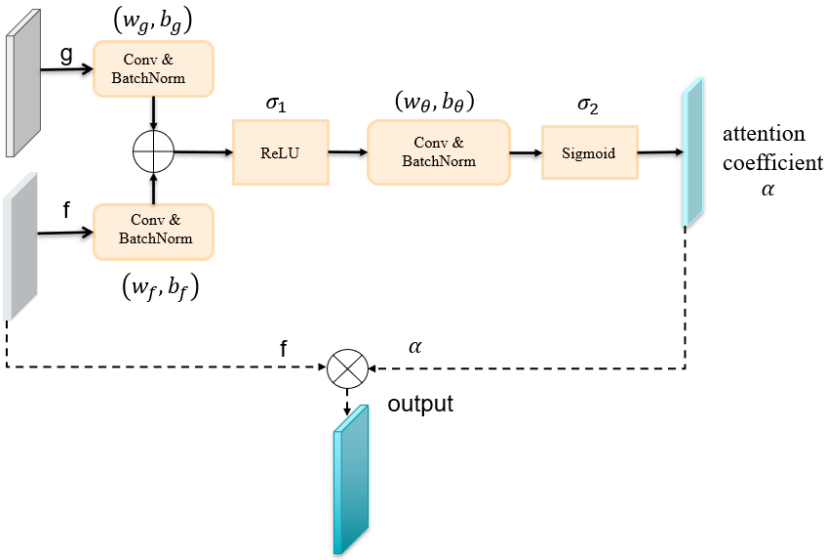

Figure 3. The diagram of attention gate in Cloud-AttU.

3. Experiments and Results

3.1. Datasets and Preparation

To train and test the proposed model, we have used a dataset derived from Landsat 8 satellite data. Landsat 8 is a satellite in the Landsat series launched by the National Aeronautics and Space Administration (NASA) on 11 February 2013, and has a finer band division than the previous Landsat satellites. Landsat 8 satellite has two optical sensors, operational land imager (OLI) and the thermal infrared sensor (TIRS). The data from Landsat 8 satellite provides data support for remote sensing applications in various fields. This study uses Landsat 8 Operational Land Imager (OLI) data, which contains nine spectral bands. Specifically, the spatial resolution for bands 1 to 7 and 9 is 30 m, while the resolution for band 8 is 15 m. The information for all bands of the operational land imager (OLI) on the Landsat 8 is shown in Table 1. In this study, four of these bands, Band 2 to Band 5, were utilized as they belong to the common bands provided by most remote sensing satellites, such as Sentinel-2, FY-4, GF-4, etc. The dataset used in this study is called the Landsat-Cloud dataset, which was originally created by Mohajerani et al. [48,49] from the Landsat 8 OLI scenes. The dataset contains 38 scenes, of which 18 are for training and 20 for testing. The ground truths of these scenes were extracted by manual methods. A normal spectral band of the Landsat 8 scenes is very large, about 9000 $\times$ 9000 pixels. It is very difficult to train a fully convolutional network directly with such a large input, as this leads to a large number of parameters and a large number of convolutional layers. This cannot be achieved in practice. To overcome this problem, large input images need to be cropped into multiple smaller patches. Therefore, each spectral band was cut into 384 $\times$ 384 non-overlapping segments. The total number of patches for the training set and test set was 8400 and 9201, respectively. In data pre-processing, we refer to various data augmentation methods including horizontal flipping, random rotation, and random scale scaling, etc. In this study, we have applied geometric translations to randomly augment the input patches.

Table 1. Landsat 8 OLI spectral bands.

Spectral Band	Wavelength (Micrometers)	Resolution (Meters)
Band 1—Coastal	0.433–0.453	30
Band 2—Blue	0.450–0.515	30
Band 3—Green	0.525–0.600	30
Band 4—Red	0.630–0.680	30
Band 5—Near Infrared (NIR)	0.845–0.885	30
Band 6—Short Wavelength Infrared (SWIR) 1	1.560–1.660	30
Band 7—Short Wavelength Infrared (SWIR) 2	2.100–2.300	30
Band 8—Panchromatic	0.500–0.680	15
Band 9—Cirrus	1.360–1.390	30

3.2. Training Methodology

In this study, all experiments were programmed and implemented on Ubuntu 16.04 using the PyTorch framework and trained using the NVIDIA GTX 1080 Ti GPU. Pycharm is used as the software environment of the experiment.

In the experiment, we used images of 384 $\times$ 384 pixels in size in the Landsat-Cloud dataset as input to the neural network. We set the training batch size as 4 and the maximum training epochs as 50. We use the Adam optimizer for optimization, with a learning rate of 0.01 for the first 40 epochs and 0.005 for the last 10 epochs. In the experiment, we set the initialization of the weights of the network to a uniform random distribution between $[-1, 1]$. Besides,we applied random rotation $(0^\circ, 90^\circ, 180^\circ, 270^\circ)$ as a data augmentation method before each training epoch.

3.3. Evaluation Metrics

In the experiment, when a cloud mask of a complete Landsat 8 scene was completed, it was compared to the corresponding Ground Truths (GT). The cloud mask obtained divides each pixel into either cloud or non-cloud categories. The performance of the model proposed in this paper was measured quantitatively by assessing the Jaccard Index, Precision, Recall, Specificity, and Overall Accuracy. The mathematical definition of these metrics is as follows:

$$\text{Jaccard Index} = \frac{TP}{TP+FN+FP} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{8}$$

$$\text{Overall Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

where FP, FN, TN, and TP, represent the total number of false positive, false negative, true negative, and true positive pixels, respectively. Since the Jaccard index is related to both recall and precision, it can measure the similarity between the two sets of images. Therefore, the Jaccard index is widely used to measure the performance of image segmentation algorithms [49,50].

3.4. Experimental Results

The designed neural network model was trained on the training dataset, and the unseen test scenes were predicted using the model obtained from the training. We evaluate the performance of the model through multiple sets of experiments. Figure 4 shows the cloud detection results for different kinds of test scenes. The first row of Figure 4 shows the RGB images, the second row shows the ground truths, and the third row shows the proposed model predictions. As can be seen from Figure 4, our proposed method can detect clouds regardless of whether the background is bright or

dark, and the segmentation results obtained are very close to the real cloud distribution. Furthermore, we find that the method proposed in this paper is also able to distinguish clouds from snow/ice and obtain segmentation results that are very close to the real scenes. These results prove that the combination of the U-Net architecture and the attention mechanism is effective and that this new neural network architecture performs well in cloud detection tasks.

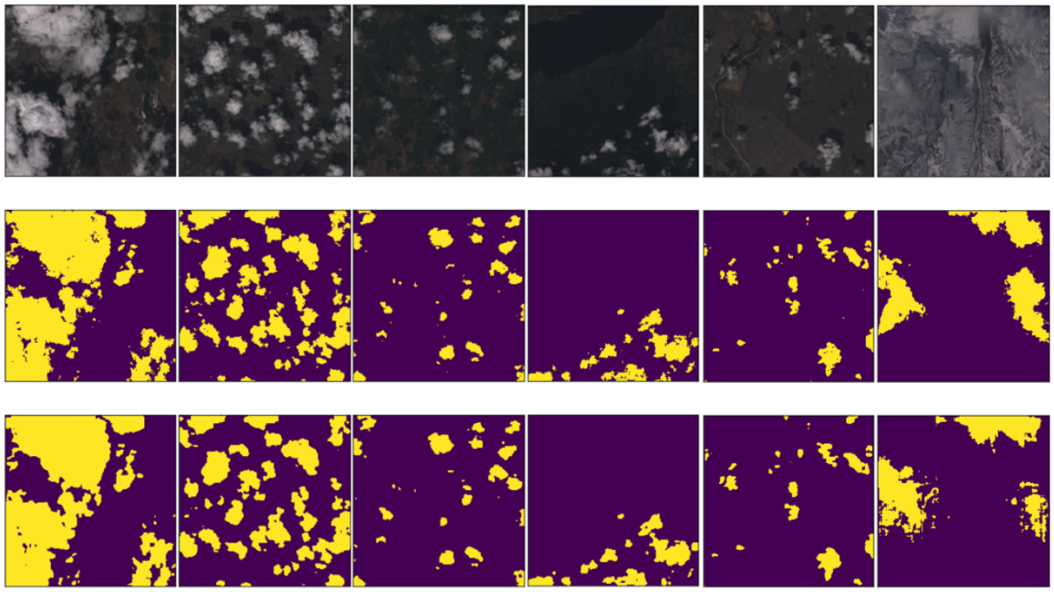

Figure 4. Cloud detection results of different scenes over Landsat-Cloud dataset [48]. The first row shows the RGB images (**top**), the second row shows the ground truths (**middle**) and the third row shows the predictions of Cloud-AttU model (**bottom**). The yellow in the figure indicates that cloud exists and the purple indicates that no cloud exists.

To further validate the proposed neural network architecture, we compare the proposed cloud detection architecture with FCN [51], Fmask [16], the original U-Net architecture and Cloud-Net [48]. Table 2 shows the experimental results of the different methods on the Landsat-Cloud dataset. As shown in Table 2, the Jaccard index of Cloud-AttU is 4.5% higher than the Jaccard index of FCN and 3.8% higher than the Jaccard index of Fmask. Fmask is the most widely used cloud detection method at present, with an overall accuracy rate of 94.26%. As can be seen from the table, the performance of our proposed model also exceeds that of Fmask. We can find that the Cloud-AttU method proposed in this study is also superior to the advanced U-Net and Cloud-Net method, suggesting that our incorporation of attention mechanisms into the U-net architecture for cloud detection is very effective. Since we train the model on the same training set and test it on the same test set, the experimental results of this study can prove the obvious superiority of our proposed architecture.

Table 2. Comparison of performance of different methods on Landsat-Cloud dataset.

Model	Jaccard index (%)	Precision (%)	Recall (%)	Specificity (%)	Overall Accuracy (%)
FCN [51]	84.90	95.17	87.65	97.10	94.91
Fmask [16]	85.45	89.26	96.57	94.07	94.26
Unet	86.06	95.14	89.73	97.45	95.80
Cloud-Net [48]	87.25	96.60	90.04	98.03	96.13
Cloud-AttU	88.72	97.16	91.30	98.24	97.05

From the above experimental results, it can be concluded that the Cloud-AttU network and Cloud-Net network have significant advantages compared to other methods. To visually analyze the performance of our proposed method, Figure 5 compares the cloud detection results of the

Cloud-AttU network and the Cloud-Net network. As shown in Figure 5, both methods yield good detection results in a general land context, and the results of Cloud-AttU are superior to those of Cloud-Net. In addition, in cases where there are ice and snow interference on land, the Cloud-Net model is susceptible to ice and snow interference, and Cloud-AttU is more resistant to interference. As shown in the third row of Figure 5, the Cloud-AttU network can identify ice and snow without misclassifying it as clouds, while the Cloud-Net network misclassifies ice and snow on the ground as clouds. From the experimental results, it can be concluded that the Cloud-AttU model proposed in this study has excellent cloud detection capability and strong anti-interference capability in complex ground background.

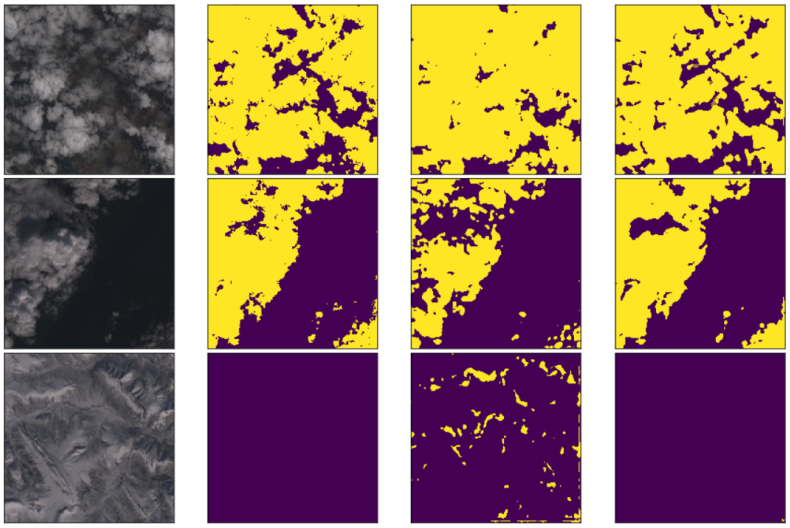

Figure 5. Cloud detection results of different scenes over Landsat-Cloud dataset [48]. The first column is the RGB image (**left**), the second column is the ground truth (**center left**), the third column is the predictions of Cloud-Net model (**center right**) and the fourth column is the predictions of Cloud-AttU model (**right**). The yellow in the figure indicates that cloud exists and the purple indicates that no cloud exists.

From the above experimental results, it can be concluded that cloud detection is susceptible to interference from factors such as ice and snow. In Figure 5, it is shown that the Cloud-AttU model proposed in this paper has obvious advantages over Cloud-Net. To further analyze the performance of the Cloud-AttU model under snow/ice disturbance, we have conducted several experiments on cloud detection in ice and snow environments. At the same time, the same experiments have been conducted using the Cloud-Net model. The results of the two sets of experiments have been compared and analyzed. Figure 6 shows a portion of the experimental results, where the first column is the RGB images. As shown in Figure 6, the snow/ice scenes that cause the interferences have been marked using the red boxes, which are useful to analyze the performance of two cloud detection methods in detail. As can be seen in Figure 6, the Cloud-AttU model can distinguish clouds from snow/ice in complex backgrounds. However, Cloud-Net has poor ability to distinguish between cloud and snow/ice and usually treats snow/ice as clouds in many cases. From the experimental results, it can be concluded that the Cloud-AttU model proposed in this paper has excellent cloud detection capability to resist interference from factors such as ice and snow in complex environments. Therefore, this model has great potential and deserves further optimization for early application to real business.

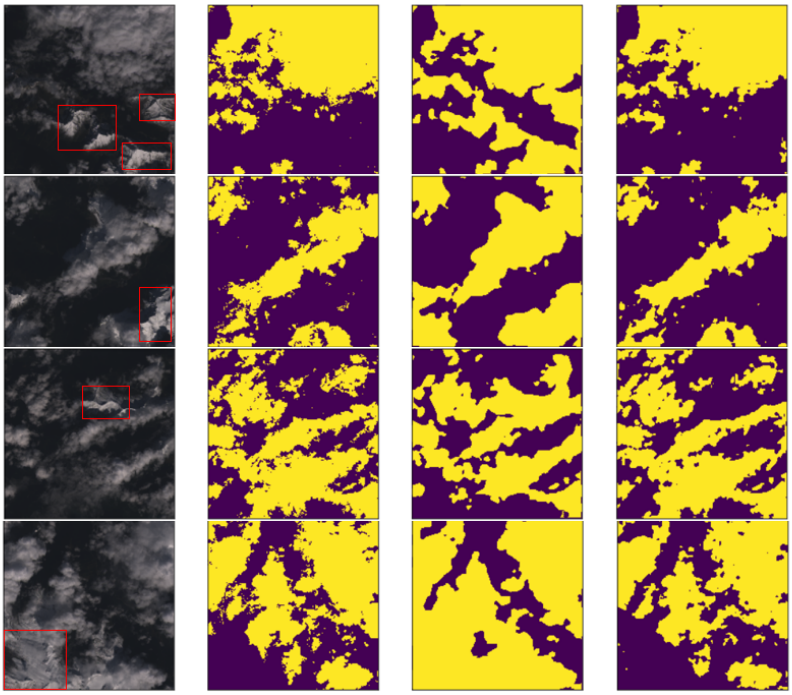

Figure 6. Cloud detection results under the influence of snow and ice ground over Landsat-Cloud dataset [48], the first column is RGB image (**left**), the second column is ground truth (**middle left**), the third column is the prediction of the Cloud-Net model (**center right**), and the fourth column is the prediction of the Cloud-AttU model (**right**). The yellow in the figure indicates the presence of clouds and the purple indicates the absence of clouds.

Not only can snow/ice interfere with cloud detection, but other factors such as surface rivers, bright surfaces, and lakes can also interfere with cloud detection. Therefore, to analyze the model's ability to discriminate against other confounding factors, we added validation cases for further analysis. Figure 7 shows the cloud detection results under the interference of factors such as surface rivers and bright surfaces. It can be seen from the figure that under the interference of these factors, the Cloud-AttU model proposed in this study still has a good effect. From the experimental results, it can be concluded that the model proposed in this study is highly resistant to multiple types of interferers, with great application prospects and worthy of further study. At the same time, the consumption of computer resources during the training process is also a key issue to consider, where the time required to complete the training of the neural network model is a very important parameter. In this study, we need about 7.2 h to complete the training. Besides, the size of our saved training model is about 430 M. In this study, due to our primary focus on the combination of attention mechanism and U-Net architecture, insufficient attention was paid to training techniques, resulting in training less efficient. Therefore, our future work will focus more on training acceleration issues to improve training efficiency. For example, we plan to reduce training time and improve training efficiency by reducing training data and introducing appropriate regularization methods. We aim to apply the model to real business by continuously optimizing the model structure and training methods.

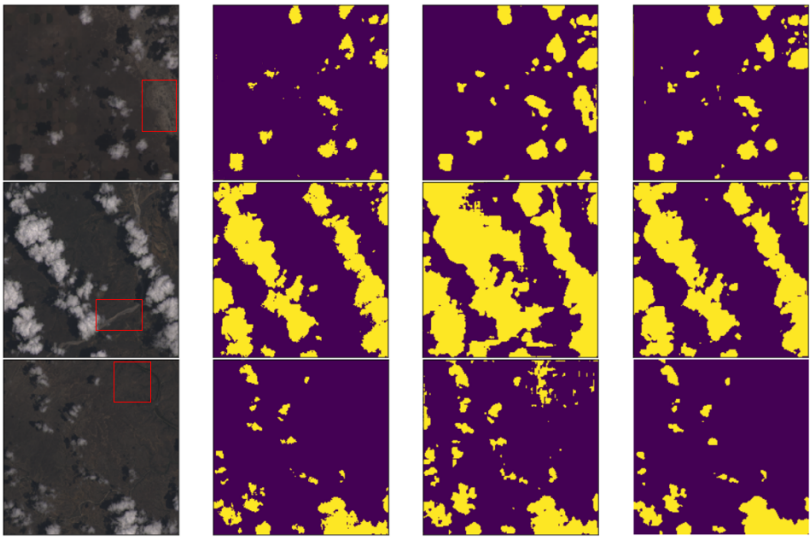

Figure 7. Cloud detection results under the influence of other factors over Landsat-Cloud dataset [48], the first column is RGB image (**left**), the second column is ground truth (**middle left**), the third column is the prediction of the Cloud-Net model (**center right**), and the fourth column is the prediction of the Cloud-AttU model (**right**). The yellow in the figure indicates the presence of clouds and the purple indicates the absence of clouds.

4. Discussions

In this study, we add attention mechanism to the U-Net network, adjust the network architecture according to the characteristics of the satellite data, choose the appropriate Loss function for optimization, and finally obtain a neural network model suitable for cloud detection of satellite data. We compared the model with other well-established models and the following discussions are obtained from the above experimental results.

(1) From the experimental results, we can find that the U-Net network, the Cloud-Net network and Cloud-AttU network based on the U-Net architecture are significantly better than Fmask. The U-Net network adopts the symmetric Encoder-Decoder structure, which achieves the fusion of high-level features and low-level features through the skip-connection operation, making the output results contain richer multi-scale information. This symmetrical network structure is concise and stable, significantly enhancing the effect of image segmentation. The results of this study demonstrate the good performance of the U-Net architecture in cloud detection tasks, indicating that this symmetrical network architecture, which fuses multi-scale information, has great potential for applications in satellite image processing and deserves further research.

(2) From the experimental results, it was found that U-Net with the attention mechanism can achieve better cloud detection results than the original U-Net. This performance boost should benefit from the attention gate. In the attention module, the output is obtained by multiplying the feature map by the attention coefficient in the attention gate. The attention coefficients tend to get larger values in the clouded region and smaller values in the cloudless region. This mechanism makes the value of the cloudless region of the feature map smaller and the value of the target region of the feature map larger, thus improving the performance of cloud detection.

(3) From the experimental results, it can be concluded that Cloud-AttU with the attention mechanism has a stronger cloud detection capability compared to Cloud-Net and single U-Net. Cloud-AttU can better resist the interference of snow and ice, and has a stronger identification ability. It is well known that satellite remote sensing data are susceptible to interference from various noises, so data processing methods that are resistant to interference are highly desirable for satellite data. Attentional mechanisms have a clear advantage in recognizing and resisting

noise interference, and thus hold great potential and research promise in numerous areas of satellite data processing.

5. Conclusions

Cloud detection is an important step in the pre-processing of satellite remote sensing data, and accurate cloud detection results are of great significance to improve the utilization of satellite data. With the development of artificial intelligence technology, deep learning methods have made great breakthroughs in the fields of image processing and computer vision. In this study, we propose an effective cloud detection method, Cloud-AttU neural network model, drawing on deep learning methods. It is a new neural network model based on the U-net architecture with the attention gate. In the process of designing the model, we modified and optimized the network structure considering the advantages of the U-net architecture in computer vision and the characteristics of the cloud in remote sensing data. At the same time, the attention gate improves the learning of target regions associated with the segmentation task while suppressing regions that were not associated with the task. Therefore, the attention gate is integrated into the proposed model to enhance the efficiency of semantic information dissemination by skipping connections. We have conducted a series of experiments using the Landsat 8 data and confirmed that the Cloud-AttU model that introduces the attention mechanism works well in the cloud detection task, and its performance is significantly better than other previous methods. The Cloud-AttU model proposed in this paper can still achieve excellent cloud detection results when there is interference information such as ice and snow. Given the success in cloud detection, the architecture proposed in this paper can also be applied to other areas of remote sensing imagery with appropriate modifications. In future studies, we will further optimize the proposed model and apply it to other satellite remote sensing data such as GF-4, FY-4A satellites.

Author Contributions: Conceptualization, X.C. and Y.G.; methodology, Y.G. and X.C.; validation, M.G.; investigation, Y.G. and M.G; writing—original draft preparation, Y.G.; writing—review and editing, B.L., M.G. and X.C.; visualization, Y.G. and M.G.; supervision, X.C. and B.L.; project administration, X.C. and B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (Grant No.41475094) and the National Key R&D Program of China (Grant No.2018YFC1506704).

Acknowledgments: The authors would like to thank the reviewers for their very useful and detailed comments, which have greatly improved the content of this paper.

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

Abbreviations

The following abbreviations are used in this manuscript:

CNN	convolutional neural network
FCN	full convolution network
AG	attention gate
FMask	Function of mask
BN	Batch Normalization
ReLU	Rectified Linear Unit
NASA	National Aeronautics and Space Administration
OLI	operational land imager
TIRS	Thermal Infrared Sensor
GF4	GaoFen-4
FY-4	FengYun-4
MCNNs	Multiple Convolutional Neural Networks
Adam	Adaptive Moment Optimization

References

1. Leprince, S.; Barbot, S.; Ayoub, F.; Avouac, J. Automatic and Precise Orthorectification, Coregistration, and Subpixel Correlation of Satellite Images, Application to Ground Deformation Measurements. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 1529–1558. [CrossRef]
2. Leitloff, J.; Hinz, S.; Stilla, U. Vehicle Detection in Very High Resolution Satellite Images of City Areas. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 2795–2806. [CrossRef]
3. Wei, J.; Li, Z.; Peng, Y.; Sun, L. MODIS Collection 6.1 aerosol optical depth products over land and ocean: validation and comparison. *Atmos. Environ.* **2019**, *201*, 428–440. [CrossRef]
4. Xie, F.; Shi, M.; Shi, Z.; Yin, J.; Zhao, D. Multilevel Cloud Detection in Remote Sensing Images Based on Deep Learning. *IEEE J. Select. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3631–3640. [CrossRef]
5. Shi, M.; Xie, F.; Zi, Y.; Yin, J. Cloud detection of remote sensing images by deep learning. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 701–704.
6. Shen, X.; Li, Q.; Tian, Y.; Shen, L. An uneven illumination correction algorithm for optical remote sensing images covered with thin clouds. *Remote Sens.* **2015**, *7*, 11848–11862. [CrossRef]
7. Li, X.; Wang, L.; Cheng, Q.; Wu, P.; Gan, W.; Fang, L. Cloud removal in remote sensing images using nonnegative matrix factorization and error correction. *ISPRS J. Photogramm. Remote Sens.* **2019**, *148*, 103–113. [CrossRef]
8. King, M.D.; Platnick, S.; Menzel, W.P.; Ackerman, S.A.; Hubanks, P.A. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 3826–3852. [CrossRef]
9. Sun, L.; Yang, X.; Jia, S.; Jia, C.; Wang, Q.; Liu, X.; Wei, J.; Zhou, X. Satellite data cloud detection using deep learning supported by hyperspectral data. *Int. J. Remote Sens.* **2020**, *41*, 1349–1371. [CrossRef]
10. Ceppi, P.; Brient, F.; Zelinka, M.D.; Hartmann, D.L. Cloud feedback mechanisms and their representation in global climate models. *WIREs Clim. Chang.* **2017**, *8*, e465. [CrossRef]
11. Barbieri, F.; Rajakaruna, S.; Ghosh, A. Very short-term photovoltaic power forecasting with cloud modeling: A review. *Renew. Sustain. Energy Rev.* **2017**, *75*, 242–263. [CrossRef]
12. Chen, P.Y.; Srinivasan, R.; Fedosejevs, G.; Narasimhan, B. An automated cloud detection method for daily NOAA-14 AVHRR data for Texas, USA. *Int. J. Remote Sens.* **2002**, *23*, 2939–2950. [CrossRef]
13. Kostornaya, A.A.; Saprykin, E.I.; Zakhvatov, M.G.; Tokareva, Y.V. A method of cloud detection from satellite data. *Russ. Meteorol. Hydrol.* **2017**, *42*, 753–758. [CrossRef]
14. Tang, H.; Yu, K.; Hagolle, O.; Jiang, K.; Geng, X.; Zhao, Y. A cloud detection method based on a time series of MODIS surface reflectance images. *Int. J. Digit. Earth* **2013**, *6*, 157–171. [CrossRef]
15. Zhu, Z.; Woodcock, C.E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **2012**, *118*, 83–94. [CrossRef]
16. Zhu, Z.; Wang, S.; Woodcock, C.E. Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote Sens. Environ.* **2015**, *159*, 269–277. [CrossRef]
17. Qiu, S.; Zhu, Z.; He, B. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. *Remote Sens. Environ.* **2019**, *231*, 111205. [CrossRef]
18. Liu, S.; Zhang, L.; Zhang, Z.; Wang, C.; Xiao, B. Automatic cloud detection for all-sky images using superpixel segmentation. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 354–358.
19. Long-fei, L.; Yun-hao, C.; Jing, L. Texture analysis methods used in remote sensing images. *Remote Sens. Technology and Application* **2011**, *18*, 441–447.
20. Ioannidou, A.; Chatzilari, E.; Nikolopoulos, S.; Kompatsiaris, I. Deep Learning Advances in Computer Vision with 3D Data: A Survey. *ACM Comput. Surv.* **2017**, *50*. [CrossRef]
21. Villalba-Diez, J.; Schmidt, D.; Gevers, R.; Ordieres-Meré, J.; Buchwitz, M.; Wellbrock, W. Deep Learning for Industrial Computer Vision Quality Control in the Printing Industry 4.0. *Sensors* **2019**, *19*, 3987. [CrossRef]
22. Li, W.; Fu, H.; Yu, L.; Cracknell, A. Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. *Remote Sens.* **2017**, *9*, 22. [CrossRef]
23. Chen, Y.; Fan, R.; Bilal, M.; Yang, X.; Wang, J.; Li, W. Multilevel cloud detection for high-resolution remote sensing imagery using multiple convolutional neural networks. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 181. [CrossRef]

24. Segal-Rozenhaimer, M.; Li, A.; Das, K.; Chirayath, V. Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN). *Remote Sens. Environ.* **2020**, *237*, 111446. [CrossRef]
25. Ozkan, S.; Efendioglu, M.; Demirpolat, C. Cloud detection from RGB color remote sensing images with deep pyramid networks. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 6939–6942.
26. Francis, A.; Sidiropoulos, P.; Muller, J.P. CloudFCN: Accurate and Robust Cloud Detection for Satellite Imagery with Deep Learning. *Remote Sens.* **2019**, *11*, 2312. [CrossRef]
27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
28. Feng, W.; Sui, H.; Huang, W.; Xu, C.; An, K. Water body extraction from very high-resolution remote sensing imagery using deep U-Net and a superpixel-based conditional random field model. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 618–622. [CrossRef]
29. Wei, S.; Zhang, H.; Wang, C.; Wang, Y.; Xu, L. Multi-temporal SAR data large-scale crop mapping based on U-Net model. *Remote Sens.* **2019**, *11*, 68. [CrossRef]
30. Cao, K.; Zhang, X. An Improved Res-UNet Model for Tree Species Classification Using Airborne High-Resolution Images. *Remote Sens.* **2020**, *12*, 1128. [CrossRef]
31. Cui, B.; Zhang, Y.; Li, X.; Wu, J.; Lu, Y. WetlandNet: Semantic Segmentation for Remote Sensing Images of Coastal Wetlands via Improved UNet with Deconvolution. In *Genetic and Evolutionary Computing*; Springer: Singapore, 2020; pp. 281–292.
32. Soni, A.; Koner, R.; Villuri, V.G.K. M-UNet: Modified U-Net Segmentation Framework with Satellite Imagery. In Proceedings of the Global AI Congress,Kolkata, India, 12–14 September 2019; Springer: Singapore, 3 April 2020; pp. 47–59.
33. Jeppesen, J.H.; Jacobsen, R.H.; Inceoglu, F.; Toftegaard, T.S. A cloud detection algorithm for satellite imagery based on deep learning. *Remote Sens. Environ.* **2019**, *229*, 247–259. [CrossRef]
34. Wieland, M.; Li, Y.; Martinis, S. Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network. *Remote Sens. Environ.* **2019**, *230*, 111203. [CrossRef]
35. Lian, S.; Luo, Z.; Zhong, Z.; Lin, X.; Su, S.; Li, S. Attention guided U-Net for accurate iris segmentation. *J. Vis. Commun. Image Represent.* **2018**, *56*, 296–304. [CrossRef]
36. He, N.; Fang, L.; Plaza, A. Hybrid first and second order attention Unet for building segmentation in remote sensing images. *Inf. Sci.* **2020**, *63*, 140305. [CrossRef]
37. Zhang, X.; Wang, X.; Tang, X.; Zhou, H.; Li, C. Description generation for remote sensing images using attribute attention mechanism. *Remote Sens.* **2019**, *11*, 612. [CrossRef]
38. Guo, M.; Zhang, D.; Sun, J.; Wu, Y. Symmetry Encoder-Decoder Network with Attention Mechanism for Fast Video Object Segmentation. *Symmetry* **2019**, *11*, 1006. [CrossRef]
39. Xu, R.; Tao, Y.; Lu, Z.; Zhong, Y. Attention-Mechanism-Containing Neural Networks for High-Resolution Remote Sensing Image Classification. *Remote Sens.* **2018**, *10*, 1602. [CrossRef]
40. Ma, W.; Yang, Q.; Wu, Y.; Zhao, W.; Zhang, X. Double-Branch Multi-Attention Mechanism Network for Hyperspectral Image Classification. *Remote Sens.* **2019**, *11*, 1307. [CrossRef]
41. Huang, Y.C.; Chang, J.R.; Chen, L.F.; Chen, Y.S. Deep Neural Network with Attention Mechanism for Classification of Motor Imagery EEG. In Proceedings of the 9th International IEEE/EMBS Conference on Neural Engineering (NER), San Francisco, CA, USA, 20–23 March 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1130–1133.
42. Xiang, X.; Yu, Z.; Lv, N.; Kong, X.; El Saddik, A. Attention-Based Generative Adversarial Network for Semi-supervised Image Classification. *Neural Proc. Lett.* **2019**, 1–14. [CrossRef]
43. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
44. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
45. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Cardoso, M.J. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 240–248.

46. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
47. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
48. Mohajerani, S.; Saeedi, P. Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1029–1032.
49. Mohajerani, S.; Saeedi, P. CPNet: A Context Preserver Convolutional Neural Network for Detecting Shadows in Single RGB Images. In Proceedings of the 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), Vancouver, BC, Canada, 29–31 August 2018; pp. 1–5.
50. Mohajerani, S.; Saeedi, P. Shadow Detection in Single RGB Images Using a Context Preserver Convolutional Neural Network Trained by Multiple Adversarial Examples. *IEEE Trans. Image Proc.* **2019**, *28*, 4117–4129. [CrossRef]
51. Mohajerani, S.; Krammer, T.A.; Saeedi, P. Cloud detection algorithm for remote sensing images using fully convolutional neural networks. *arXiv* **2018**, arXiv:1810.05782.