

Article

MSACN: A Cloud Extraction Method from Satellite Image Using Multiscale Soft Attention Convolutional Neural Network

Lin Gao^{1,2,*}, Chenxi Gai¹, Sijun Lu³ and Jinyi Zhang^{1,4} 

¹ School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China; zhangjinyi@sylu.edu.cn (J.Z.)

² School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China

³ Shen Kan Engineering & Technology Corporation, MCC, Shenyang 110169, China

⁴ Faculty of Engineering, Gifu University, Gifu 501-1193, Japan

* Correspondence: gaolin324@sylu.edu.com; Tel.: +86-17801203241

Abstract: In satellite remote sensing images, the existence of clouds has an occlusion effect on ground information. Different degrees of clouds make it difficult for existing models to accurately detect clouds in images due to complex scenes. The detection and extraction of clouds is one of the most important problems to be solved in the further analysis and utilization of image information. In this article, we refined a multi-head soft attention convolutional neural network incorporating spatial information modeling (MSACN). During the encoder process, MSACN extracts cloud features through a concurrent dilated residual convolution module. In the part of the decoder, there is an aggregating feature module that uses a soft attention mechanism. It integrates the semantic information with spatial information to obtain the pixel-level semantic segmentation outputs. To assess the applicability of MSACN, we compare our network with Transform-based and other traditional CNN-based methods on the ZY-3 dataset. Experimental outputs including the other two datasets show that MSACN has a better overall performance for cloud extraction tasks, with an overall accuracy of 98.57%, a precision of 97.61%, a recall of 97.37%, and F1-score of 97.48% and an IOU of 95.10%.

Keywords: cloud extraction; concurrent dilated convolution; multiscale convolutional neural network; soft attention mechanism; ZY-3 satellite images



Citation: Gao, L.; Gai, C.; Lu, S.;

Zhang, J. MSACN: A Cloud

Extraction Method from Satellite

Image Using Multiscale Soft Attention

Convolutional Neural Network. *Appl.*

Sci. **2024**, *14*, 3285. [https://doi.org/](https://doi.org/10.3390/app14083285)

[10.3390/app14083285](https://doi.org/10.3390/app14083285)

Academic Editor: Andrea Prati

Received: 4 March 2024

Revised: 8 April 2024

Accepted: 9 April 2024

Published: 13 April 2024



Copyright: © 2024 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 ([https://](https://creativecommons.org/licenses/by/4.0/)

[creativecommons.org/licenses/by/](https://creativecommons.org/licenses/by/4.0/)

[4.0/](https://creativecommons.org/licenses/by/4.0/)).

1. Introduction

Cloud occlusion due to environmental factors has limited the performance of optical sensors, including domestic optical sensors. Therefore, it is of great significance to accurately extract cloud regions. Accurate extraction of cloud areas brings the following two benefits: First, low-quality images can be pre-checked under high traffic conditions, thereby reducing the amount of downlink data and improving image transmission efficiency; second, accurate extraction of cloud areas helps for land identification, and it provides the basis for subsequent cloud removal and reconstruction. Therefore, the cloud extraction task holds significant importance in image interpretation and quality inspection.

Nonetheless, there are still some difficulties in the aspect of cloud detection. They are summarized as follows: (1) **Cloud diversity and ambiguous boundaries:** the significant variations in cloud types and shapes, ranging from thin clouds to thick clouds, pose inherent challenges in cloud extraction tasks. These differences within the cloud class directly lead to substantial intra-class variability. Moreover, the imaging mechanisms associated with the portrayal of thinly cloud-covered regions result in unclear boundaries in the image, further complicating the interpretation process. The lack of distinct edges in thinly cloud-covered regions makes it difficult to discern and interpret these areas accurately, as shown in Figure 1a. (2) **Confounding of terrain and clouds in complex scenes:** depending on the influence of season and climate, local terrain is covered with snow, which reflects the

same information from the image as clouds. Therefore, cloud snow against a complex background is difficult to distinguish, as shown in Figure 1b.

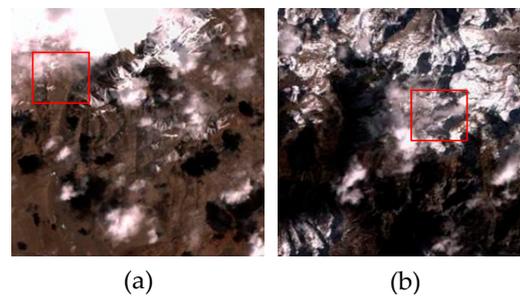


Figure 1. Confused cloud detection instance (red rectangles). (a) Boundaries of thin clouds are not clear; (b) Clouds and snow are confused.

At present, cloud detection and segmentation approaches are mainly bifurcated into two parts: threshold-based approaches and learning-based approaches [1–4]. The first type of algorithm is mainly based on cloud spectral characteristics (portions of the electromagnetic spectrum), brightness, texture characteristics and geometry by analyzing the spectral difference between the cloud and other surfaces, thresholds or rules are formulated to realize cloud extraction [5–8]. Quan Xiong et al. [9] employed a dynamic threshold hybrid multi-spectral feature (HMF) method for cloud extraction, which combines three kinds of spectra of normalized difference vegetation index (NDVI), whiteness and haze optimization transform (HOT) features to detect cloud pixels. One can utilize hybrid multi-spectral features. The pure threshold algorithm is straightforward, efficient and applicable for cloud detection; however, the method's impracticality stems from its sensitivity to both background variations and cloud coverage. To enhance the capability to recognize edge details, some people have proposed methods based on machine learning, such as using Support Vector Machine (SVM), and Random Forest (RF) to extract hyper-spectral images [10]. Sui Y et al. [11] used simple linear iterative clustering (SLIC) to divide optical satellite images into super-pixels, and then calculated the energy and spectral features using Gabor transformation by extracting the texture features. The characteristics of cloud super-pixels serve as the training samples for the SVM classifier. The SVM classifier is trained to establish the classification model. In addition, Shao M et al. [12] proposed a multi-dimensional and multi-granularity dense cascade forest (MDForest) for multi-spectral cloud detection. MDForest is a deep forest architecture that introduces a multi-dimensional and multi-granularity scanning mechanism, which enhances the cascade forest representation learning ability. At the same time, the spectral information of the multi-spectral satellite image was captured for cloud extraction. However, its recognition ability is not ideal, especially in complex backgrounds. In addition, the overall accuracy of the above cloud detection methods also is up to the number of image bands to effectively extract clouds. Recently, convolutional neural network (CNN) methods have also been used to detect clouds. For example, the M-type convolutional network model RM-Net uses atrous spatial pyramid pooling (ASPP). ASPP consists of atrous convolution and pyramid pooling [13]. When scaling features, the phenomenon that the information loss caused by multiple down-sampling is effectively reduced. It extracts multi-scale features of images without losing information and combines residual units to make the network less prone to degradation. The encoding and decoding modules extract the global context information of the image, judge the class probability of each pixel according to the fused features, and input it into the classifier for pixel-level cloud and non-cloud segmentation [14].

With the advancement of artificial intelligence, deep learning algorithms exhibit remarkable performance in image interpretation, particularly in the domain of optical satellite remote sensing imagery. Different from natural images, satellite images from optical sensors have a larger scale, more coverage and richer ground truth details. From the perspective

of deep learning, we categorize cloud extraction methods into the following two classes: CNN-based and CNN-Trans-based methods. (a) CNN-based: U-Net [15] is a classic image segmentation method with excellent performance in many binary classification tasks. Numerous studies have demonstrated that methods of the U-shape structure [16,17] showcase outstanding performance in the segmentation of optical satellite images [18]. As an illustration, CloudU-Net, which is a structure derived from U-Net, uses Atrous Convolution instead of the traditional convolution layer to enlarge the field of view. It increases the training speed through batch normalization, which can also prevent over-fitting of the model [19]. Although the algorithm performs well in the case of dense small objects, it does not take into account the diversity of clouds, because the characteristics of clouds are uncertain. (b) CNN-Trans-based: Self-attention mechanism is a core part of the transformer algorithm. The attention mechanism draws inspiration from human visual cognition science, where individuals naturally concentrate on detailed information related to a target while suppressing irrelevant details when reading text or observing objects. It is a process that goes from coarse to fine. The integration of the attention mechanism into CNN networks was introduced by researchers in 2017 [20]. Since then, attention-based mechanisms in CNNs have found widespread application across various domains. The CNN-Trans attention module comprises two key components: the channel and spatial attention modules, respectively. The former accentuates the correlation among the dimensions of each layer, which forces the attention on the interested feature information and suppresses the useless channel. The latter can retain high-frequency feature information through spatial operation. On one hand, Hu et al. [21] introduced a novel CNN unit called the squeeze-and-excitation block, which dynamically adjusts the feature response value by modeling inter-channel relationships. In comparison, CBAM [22] not only incorporates spatial attention but also employs a concurrent structure involving multiscale pyramid pooling within the channel attention. Experimental validation attests to its effectiveness. Spatial attention, on the other hand, directs focus to regions of interest in the spatial aspect. While the attention mechanism has demonstrated considerable utility since its integration into CNNs, the issue of redundancy remains a common challenge. Therefore, numerous researchers have employed transformer models for cloud extraction tasks in optical satellite imagery [23–26]. Zhang J. et al. [27] proposed a CNN cloud detection algorithm for GF-1 satellite images. Through cascading the channel and spatial attention, it introduced a probabilistic upsampling module to merge the downsampling channels through the entire network structure. Then, dark channel transformation based on dark channel prior technology and NSCT was added to the above network [28]. Even though the attention mechanism in transformer models has demonstrated excellent performance in various domains, the direct transplantation of transformer to cloud extraction tasks does not yield satisfactory results. Some CNN-Trans-based methods connecting traditional CNN with a Transformer can increase the complexity of the model, making it challenging to find the optimal solution through optimization methods. Therefore, we design a network that integrates the strengths of deep CNN and attention, making it more suitable for imagery obtained from satellite sensors.

In this article, we refined a Multiscale Soft Attention Convolutional Neural Network (MSACN) which is a multiscale deep convolutional network structure with a soft attention mechanism incorporating spatial information. Taking inspiration from the U-Net [15], ResNet [29] and attention mechanism [20], MSACN consists of two parts: a deep feature encoder module and a multi-head soft attention decoder module for cloud prediction. Compared with other networks, MSACN exploits the shallow-level information and high-level features of cloudy/non-cloudy pixels, which improves the extraction decision without any manual specific spectral information processing, since the pre-trained network can be visual objects in images to extract rich and unique high-level representations. Summarily, the contributions can be succinctly outlined as follows:

- (1) We expand the scale of the ZY-3 satellite remote sensing cloud extraction dataset. To test in more complex scenarios, we augmented the dataset by incorporating a subset of clouds with snow images. The raw data are pre-extracted using the model, followed

- by manual refinement using Photoshop. Each image underwent meticulous selection to ensure the training data accuracy.
- (2) We refine a multiscale deep convolutional neural network with soft attention and spatial information for cloud segmentation from satellite remote sensing images. Following the encoder–decoder architecture, our primary enhancement lies in the incorporation of a concurrent dilated residual convolution module and a multi-head soft attention fusion between the encoding and decoding processes, respectively.
 - (3) To evidence the validation, we employ comparative analyses with similar methods, including traditional CNN-based approaches, as well as dissimilar methods such as transformer-based models, all within the same datasets and training environment. In terms of the overall accuracy, precision, recall, F1-score and IoU, the performance of MSACN outperformed with other networks, showcasing its superior effectiveness. Meanwhile, we transplant the model to other datasets to assess its adaptability.

2. Materials and Methods

In this paper, MSACN consists of two parts: concurrent dilated residual convolution module, and multi-head soft attention module. To adaptively accommodate the diversity of cloud-shaped dimensions, the concurrent dilated convolutional module establishes pyramid-like features via multi-scale dilated factors in the front of the architecture. The residual convolution units are used for the remaining of the encoding process. The multi-head soft attention module is integrated into the process of decoding for extracted results restoration. By fusing spatial features at different feature resolution levels, a multi-head spatial fusion attention module is established based on depth semantic features, and then the dots of every head-attention channel are connected to achieve the prominent cloud linear distinguishable features. By concatenating the encoding features and upsampling channels, the soft attention module alleviates the cloud boundaries problem caused by the roughness of the architecture.

2.1. The Overall Structure

As shown in Figure 2, MSACN is a U-shape structure including an encoder and decoder components, which are multi-scale Concurrent Dilated Residual Convolution Module (CDRCM) and Multi-head Soft Attention Module (MSAM), respectively. The MSACN is taking inspiration from U-Net as a whole. In addition to the core modules parts (CDRCM and MSAM), other backbone parts refer to ResNet50. Its expansion path and contraction path also have a corresponding relationship. Thus, the feature extraction part uses the residual module to deepen the model without loss of resolution and learn more complex features to reinforce the representation ability of the model. Before inputting the feature extractor part to the cloud semantic prediction module through skip connection, the shape is changed through convolution to adapt to the prediction module, and CDRCM is used in the first convolution to expand the field of view and preserve the spatial resolution. The input layer has the richest and most primitive features, and the reflected features are not lost due to processing. Therefore, the dilated pyramid-like unit is built for the original input image to control the scale difference by the dilation–convolution factor in order to improve the shape-scale in-variance performance. A series of multi-head attention modules are added to the cloud semantic prediction module. The multi-head attention of the decoding process belongs to the soft attention mechanism. Through the combination of multiple heads soft attention blocks, the depth semantic features of the cloud are strengthened and make them more approximate to be linear separable features. The multi-head soft attention module consists of these white blocks, as shown on the right half of the structure in Figure 2. Each multi-head attention module is based on concatenated features. It receives two inputs from the encoder and decoder parts, and performs splicing and convolution operations on its output and upsampling results to ensure the depth of the network. The attention mechanism believes that features at different levels in the network have different importance, and by assigning greater weight to important features, it can

inform subsequent layers to focus more on the interested information and suppress useless information. This improved method trains the network to more accurately capture the spatial location information and boundary details of clouds, thereby enhancing the model's accuracy. Through multiple operations of upsampling and attention modules, the network is able to gradually restore spatial details and perform fine boundary segmentation to improve the accuracy and clarity of cloud boundaries.

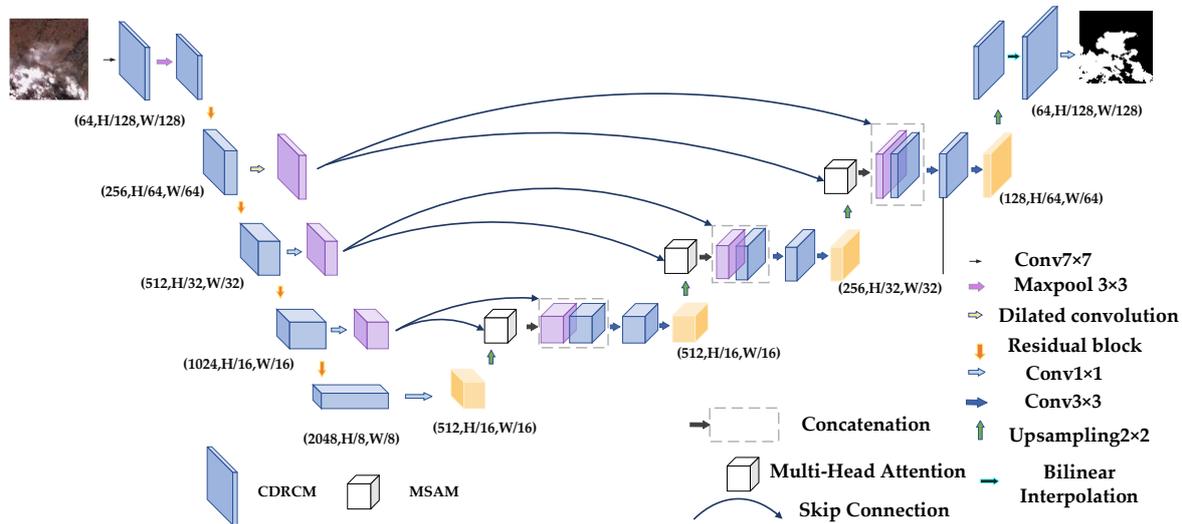


Figure 2. Illustration of MSACN with encoding and decoding structures.

2.2. Concurrent Dilated Residual Convolution Module

The concurrent dilated residual convolution module is a core component of the encoding process in MSACN. CDRCM plays a pivotal role in the enhancement of our proposed method. During the encoding phase, it systematically incorporates concurrent dilated convolutions with varying dilation factors, forming a pyramid-like structure. This design is strategically implemented to capture features at multiple scales and enrich the model's receptive field. Typically, the concurrent dilated residual convolution module is used to be integrated into the residual basic block. The concatenated output from these dilated convolutional channels is further processed through residual modules to facilitate effective feature encoding. The design of dilation factors considers the relatively high spatial resolution of the raw images, making it unsuitable to employ smaller dilation sizes for establishing sparse convolutional kernels. The expansion of the basic block is set to 1, so the shape of the feature channel is the same as the input. The bottleneck has an extra convolution layer on the right side of the basic block, and expansion is set to 4, which means that the size of input and output maps is different at this time. As illustrated in Figure 2, after the skip connection is matched with the shape of the cloud semantic prediction module, a full convolution layer is introduced to the middle part of ResNet50 to change the channel-matching shape. To better receive the validated information of the cloud, we enlarge the field of view and maintain the resolution features of the input channel, and replace the first convolution with the dilation convolution of dilation = 4, as illustrated in Figure 3a, kernel1. It is the dilated convolution of dilation = 2, while kernel2 is an ordinary convolution. Through a series of residual blocks and convolutional layers, the dimensions of channels gradually increase, the shape gradually decreases, and finally a preliminary effective feature layer with a shape of [2048, 8, 8] is obtained.

In the part of the remaining encoding process, apart from a series of convolutions, BN, ReLU, and MaxPooling to obtain output, the basic residual block includes multiple bottlenecks and the basic block module is shown in (b) and (c) of Figure 3. The inclusion of these two modules addresses the vanishing gradient problem inherent in DCNNs by introducing cross-layer shortcut connections, making the network easier to train and optimize.

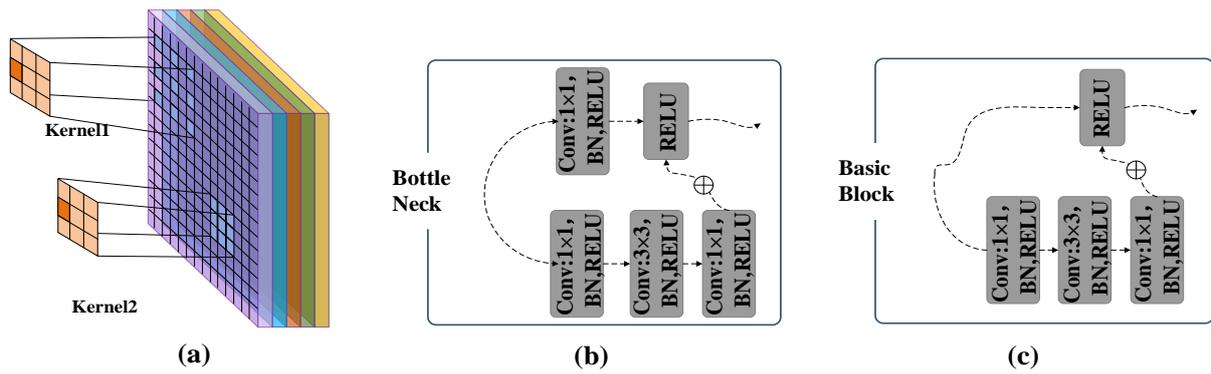


Figure 3. (a) Illustration of CDRCM; (b) Illustration of the bottleneck module; (c) Illustration of the basic block module.

2.3. Multi-Head Soft Attention Module

During the encoding process, we employ the multi-head soft attention module, as illustrated in Figure 4. Each attention head within this module functions as a soft attention. Inspired by [30,31], it encompasses various convolutional operations culminating in a final attention score. This design empowers the model to dynamically allocate attention across features, adapting to the diverse inputs. Unlike conventional attention mechanisms, multi-head soft attention modules can process input representations with spatial dimensions, such as images, feature maps, or other numerical data types. The x_1 input from the skip-connection and the x_2 input generated by the previous layer are fed into the 1×1 convolution, turning them into the same number of channels, because x_2 comes from the next layer of x_1 , and the size is $x_1 = 1/2$, so x_1 is downsampled. After that, they are accumulated and passed through ReLU via another 1×1 convolution and sigmoid activation function. The process involves assigning an importance score, ranging from 0 to 1, to every segment of the feature map. Subsequently, this attention channel map is multiplied by the input of the skip connection, leading to the generation of the final output for the attention block. Following this, the outputs of identical 4 or 8 soft attention modules are concatenated, followed by subsequent average pooling across different attention heads. This ensemble ultimately produces the final output.

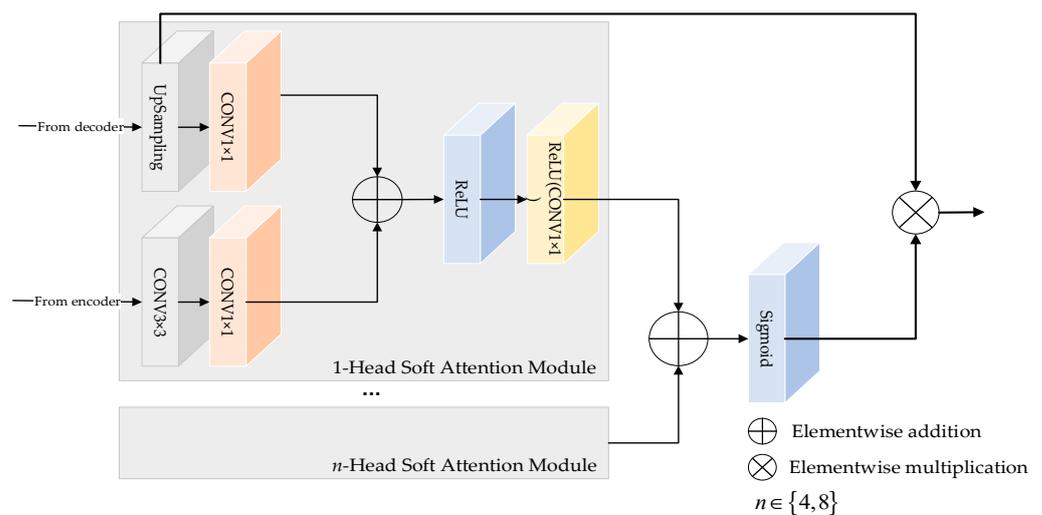


Figure 4. The structure of MSAM.

The semantic concatenation layer serves the purpose of connecting the outputs generated by the multi-head soft attention modules. This structure aligns with the framework of the encoder. Utilizing these five initial effective feature channels, the output from the

preceding layer is concatenated, followed by feature fusion. The method of feature fusion involves upsampling and stacking the feature layers.

$$x = \text{merge}(x_1, x_2) \quad (1)$$

In Formula (1), x_1 is the corresponding positions in the encoding process, and x_2 is the feature from the upper layer convolution output. We upsample the output of the fifth convolutional pooling block to obtain a feature layer of [16, 16, 512], and then change the channel through 1×1 convolution and connect the output x_1 to the fifth convolution pool through jump x_2 after block upsampling is used as the two inputs of the multi-head soft attention module, and then the output is concated with x_1 to obtain a feature layer of [32, 32, 256], and so on, and finally through the bilinear interpolation method restores the feature layer shape back to the input image size, uses a 1×1 convolution to adjust the channels, and adjusts the dimension of channels of the final feature layer to num_classes (cloud and non-cloud pixels).

2.4. MSACN Deformation

In this part, the depth of the network is categorized based on both relative depth and absolute depth. The absolute network depth generally refers to all layers of the network. Relative depth primarily denotes the dimensions of pooling layers in the network. As the characteristic resolution in satellite imagery is often associated with pooling layers, especially for medium-resolution remote sensing images, an excessively deep network with numerous pooling layers not only results in information resolution loss but also causes irreversible loss of resolution, leading to sub-optimal cloud extraction performance. Additionally, training such deep networks becomes challenging. Thus, considering the strong correlation between the depth of the network structure and the resolution characteristics of input images, we undertake a local deformation of the refined method: MSACN-small. We divide the entire network into four segments, using pooling layers and upsampling layers as boundaries for both encoding and decoding parts. In the context of medium-resolution remote sensing images, where pruning the network to reduce its scale is necessary, we conduct experiments and analyses with a downsized model, denoted as MSACN-small, feathering three pooling layers. This compact variant participates in the comparative analysis of our proposed method.

3. Experiments and Results

To assess the effectiveness of MSACN, extensive experiments are evident on the ZY-3 optical satellite image dataset. In the experiment, we strictly set a series of hyper-parameters and conditions to ensure the accuracy and comparability of the experiment. We compare the MSACN with other methods using convolutional neural network (CNN) architectures and Transformer-based methods to evaluate the performance advantages of MSACN, and try to explore the different impacts of architecture on remote sensing image processing. In terms of experimental results, we record various performance indicators in detail, to comprehensively assess the performance of various methods. A detailed presentation of our experimental results will follow in the subsequent sections.

3.1. Experiments Setup

3.1.1. Data Augmentation and Preprocess

We utilize the ZY-3 cloud dataset provided by [31]. In terms of data conditions, the quality of cloud datasets plays a crucial role in the training and processing of cloud detection. The variety and comprehensiveness of cloud datasets are of great importance. To enrich the dataset, we augment 1160 images with dimensions of $1024 \times 1020 \times 3$. The newly added images predominantly feature cloud and snow instances within the same scene, as depicted in Figure 5. Notably, these cloud and snow patterns exhibit strikingly similar characteristics in the images, increasing the complexity of the training scenarios. This augmentation aligns more closely with real-world situations encountered in satellite

imagery quality assessments. The dataset distribution after augmentation is outlined in Table 1. To ensure the accuracy of data, a meticulous labeling process is employed for the newly added images. In this process, an existing model is used for preliminary feature extraction, followed by manual refinement for each image using Photoshop. This iterative approach ensures the precision of the labels, as human expertise is applied to correct and validate the automatically generated labels. Regardless of cloud thickness, we label pixels as “1” and non-clouds as “0”.

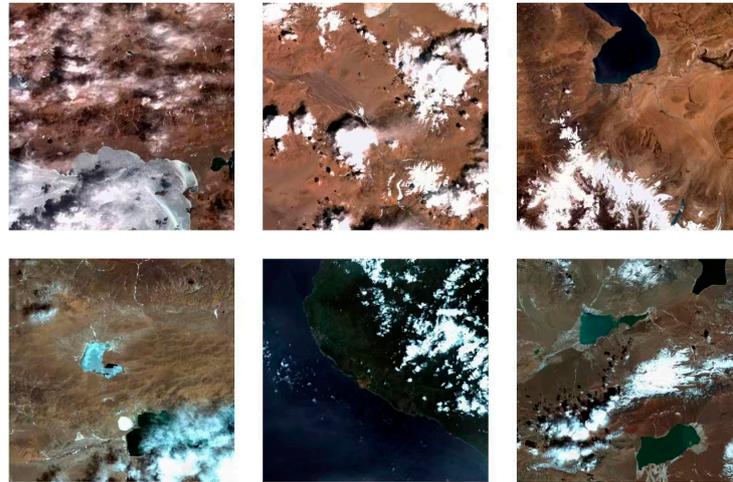


Figure 5. Augmented example images in complex scenarios (with snow and clouds).

Table 1. Distribution of datasets.

	Train Set (70%)	Validation Set (20%)	Test Set (10%)
No. Of Original ZY-3 Images	3351	957	479
No. Of Added ZY-3 Images	4163	1189	595
38-cloud of Landsat OLI 8 dataset	27	8	3
GF-1 WFV dataset	76	22	10

Our augmented training data come from the ZY-3 satellite, which comes from the Land satellite Remote Sensing Application Center, Ministry of Natural Resources. The website is <http://www.lasac.cn/> (accessed on 5 April 2024). We use a square color pan-sharpened version of the image with a resolution of 5.8 m.

In the discussion section, we use two datasets for further exploration, the 38-cloud dataset [32,33] and the GF1_WHU dataset [34]. All data in the 38-cloud dataset were compiled by the Robot Vision Laboratory (LRV). It contains 38 Landsat OLI 8 scene images and their manually labeled ground truth for remote sensing cloud extraction. For the consistency of experimental comparison, the unified size of all pictures in the dataset is adjusted to 256×256 . The SENDIMAGE laboratory released this validation dataset, which includes 108 GF-1 Wide Field of View Level 2A scenes and their reference clouds and cloud shadow masks. Ground truth images are obtained by manually labeling cloud boundaries after a visual inspection conducted by an experienced user. The GF1_WHU dataset contains two labels: cloud and shadow. To make the dataset labels consistent, the cloud shadow labels summarized in the dataset are preprocessed and marked as the background. The scene image is converted from tif format to png format by ArcMap 10.3 software.

3.1.2. Training and Implementation

In terms of hardware environment, this experiment is carried out under the pytorch deep learning framework of the Windows10 system, whose Intel (R) Core (TM) i5-11400F CPU@2.60 GHz GPU NVIDIA 3060 12G memory.

In the training phase, we use AdamW optimizer [35] to train on 256×256 images with some hyper-parameters: learning rate (lr) = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$. On top of our network, we set the Sigmoid function to tidy up the results. The loss function adopts the Dice loss function. Training optimization indicators include overall accuracy, recall, precision, F1, and IoU. It took our network about 200 epochs to converge.

As mentioned above, we use Pytorch framework to build and train our model. All training is performed with a batch size of 16 and a validation set is used to evaluate performance during training. The main software packages used include python 3.7, CUDA 11.6, cuDNN 8.5, pytorch1.12, etc. The experimental loss function is shown in Figure 6.

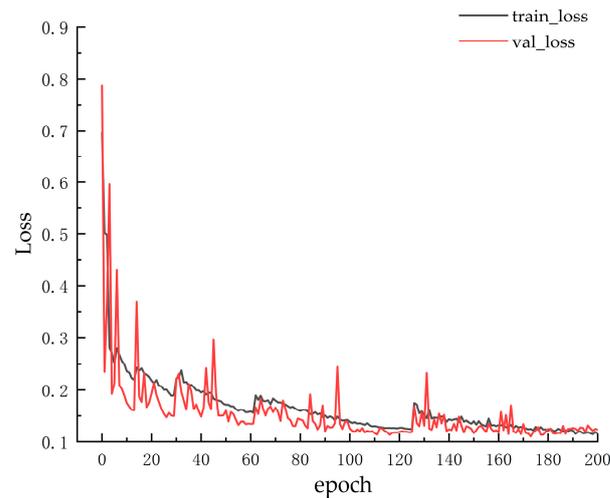


Figure 6. Loss function Plot in the ZY-3 satellite cloud dataset.

3.1.3. Estimate Metrics

To assess the computational feasibility of MSACN in the domain of cloud extraction, we adopted a variety of metrics, including P (Precision), R (Recall), F1 score based on confusion matrix, accuracy and cross-union ratio (IoU), expressed as follows.

$$P = \frac{TP}{TP + FP} \quad (2)$$

$$R = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (4)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (6)$$

In the case of cloud extraction, P (precision) stands for the ratio of true cloud pixels in the output results, while recall refers to how many pixels are detected as clouds out of all cloud pixels in the image. Both are important metrics for evaluating the performance of classification models. TP means True Cases, which is the part of pixels correctly labeled as clouds; FN stands for False Negatives, which is the number of cloud pixels incorrectly marked as belonging to non-clouds; FP represents False Positives, which is the number of cloud pixels identified instead of labeled, TN represents true negative examples, the pixels of non-clouds marked as non-clouds. For F1, it better reflects multifaceted cloud extraction errors. The F1 score is figured by precision and recall, and is a comprehensive indicator considering P and R. The accuracy rate is expressed as the ratio of cloud pixels correctly detected as clouds and non-clouds to the total number of pixels, which can reflect

the ability of MSACN in correctly classifying pixels. IoU is used to measure the ratio of the overlapping area between the model-predicted cloud area and the truth cloud area.

3.2. Experiments Analysis

As shown in Table 2, a total of Transform-based and CNN-based models are compared in the same experimental environment. TransUNet [36] and SwinUNet [37] are improved models based on the Transformer structure. TransUNet is a network combined with the Transformer module. The Transformer module is used to establish global dependencies, and the detailed information is retained through the U-Net structure. Among them, a Transformer is a neural network structure, which captures long-distance dependencies through the global self-attention mechanism, so as to better understand the context information of the input sequence. The multi-head soft attention mechanism performs self-attention and weight calculation on different positions of the input, allowing the model to strengthen or weaken the degree of attention according to different positions. TransUNet may perform better for small-size images or low-resolution images, but requires enormous data for training, and has a poor ability to extract local detail information. SwinUNet is a network structure that combines Swin Transformer and U-Net. Swin Transformer is a network structure based on Transformer that introduces Swin Block and window attention mechanism. Swin Block as a basic building block includes a displacement layer and a window attention layer. The traditional self-attention mechanism requires the entire input sequence when calculating the attention weight. Perform global operations, which is computationally expensive for large-scale images. Swin Transformer divides large-scale image input into multiple fixed-size image blocks (called windows), and calculates self-attention weights at each window level to reduce computational complexity. SwinUNet may encounter resource constraints when processing high-resolution images or large-scale datasets.

Table 2. Various verification indicators of MSACN and various models.

Method	Accuracy	IoU	Precision	Recall	F1
TransUNet	0.9653	0.8988	0.9528	0.9409	0.9464
SwinUNet	0.9189	0.8147	0.8908	0.9011	0.8941
U-Net	0.9704	0.9285	0.9595	0.9665	0.9628
DeeplabV3+	0.9714	0.9218	0.9589	0.9559	0.9574
MSACN-small	0.9746	0.9403	0.9703	0.9689	0.9699
MSACN	0.9857	0.9510	0.9761	0.9737	0.9748

In order to compare with the traditional cloud extraction model based on CNN, the added CNN-based includes the combination network of U-Net, and DeeplabV3+ for comparative experiments. U-Net is a classic structure for image semantic segmentation. Owing to the limitations of the local receptive field and upsampling layer of the network structure, U-Net often performs poorly in the segmentation of fine boundaries, including important global context information. For more demanding tasks, the performance of U-Net is limited. And DeeplabV3+ [31] is a convolutional neural network image segmentation model based on pyramid pooling. This model improves the traditional DeepLabV3+ by introducing atrous convolution and decoder modules, but for small-sized target objects, its detection and segmentation accuracy may be reduced. MSACN-small is a deformation of the proposed method. MSACN is the main network proposed in this paper. The number of multi-head modules of MSACN in Table 2 is 8. In the ablation experiment in Section 3.3, the relationship between the heads of soft attention modules and the model accuracy will be further explored.

In Table 3, IoU and accuracy take the optimal values in all rounds, while Precision, Recall and F1 take the values in the round where the IoU optimal value is located (hereinafter MSACN refers to MSACN-8 head). Figure 6 is the trend chart of the five evaluation indicators in 200 epochs. As illustrated in Table 3, the MSACN-8 head performs outstandingly in the ZY-3 satellite image dataset. The accuracy is 2.03% higher than the Transformer-

based:TransUNet and 1.32% higher than the CNN-based:U-Net; the IoU ratio of TransUNet is 5.32% higher and 2.35% higher than U-Net. The qualitative and quantitative analysis reveals that MSACN demonstrates excellent performance and yields satisfactory extraction results in complex scenes characterized by significant variations in cloud distribution and content. The pre-trained weights of ResNet50 are used for transfer learning, and the model undergoes improvement through fine-tuning so that the model can swiftly adapt and converge in the domain of cloud extraction from satellite imagery.

Table 3. Comparison of various verification indicators between MSACN and each model on the same cloud dataset.

Image	Method	Accuracy	IoU	Precision	Recall	F1
Image1	TransUNet	0.9171	0.9033	0.6792	0.9024	0.7751
	SwinUNet	0.9306	0.9217	0.8159	0.7256	0.7681
	U-Net	0.9508	0.8159	0.8867	0.7905	0.8358
	DeeplabV3+	0.9504	0.9400	0.7993	0.9172	0.8542
	MSACN-small	0.9387	0.9297	0.8026	0.8129	0.8077
	MSACN	0.9641	0.9580	0.9210	0.8617	0.8904
Image2	TransUNet	0.9246	0.8968	0.8637	0.8914	0.8773
	SwinUNet	0.9319	0.9088	0.9321	0.8357	0.8812
	U-Net	0.9401	0.9197	0.9575	0.8392	0.8944
	DeeplabV3+	0.9505	0.9322	0.9392	0.8944	0.9162
	MSACN-small	0.9487	0.9297	0.9328	0.8949	0.9135
	MSACN	0.9594	0.9441	0.9575	0.9061	0.9311
Image3	TransUNet	0.8675	0.7517	0.9068	0.8147	0.8582
	SwinUNet	0.8880	0.7763	0.9223	0.8307	0.8741
	U-Net	0.8764	0.7574	0.9267	0.8057	0.8619
	DeeplabV3+	0.8995	0.7925	0.8851	0.9390	0.9113
	MSACN-small	0.7070	0.7851	0.9128	0.8488	0.8796
	MSACN	0.9238	0.8446	0.9334	0.9274	0.9304

In addition to the numerical results in the table above, for the case of cloud coverage under complex backgrounds, three remote sensing cloud coverage images with different characteristics are selected from the ZY-3 satellite images for analysis. These images have different degrees of cloud cover and background, and the experimental results of MSACN are demonstrated in Figures 7–9. The figures below illustrate the raw satellite image, and the outputs of TransUNet, SwinUNet, U-Net, DeeplabV3+, MSACN-small, MSACN and Ground Truth, respectively. Upon visual inspection of the results, the result of MSACN is closer to the labeled image, where white represents cloud pixels.

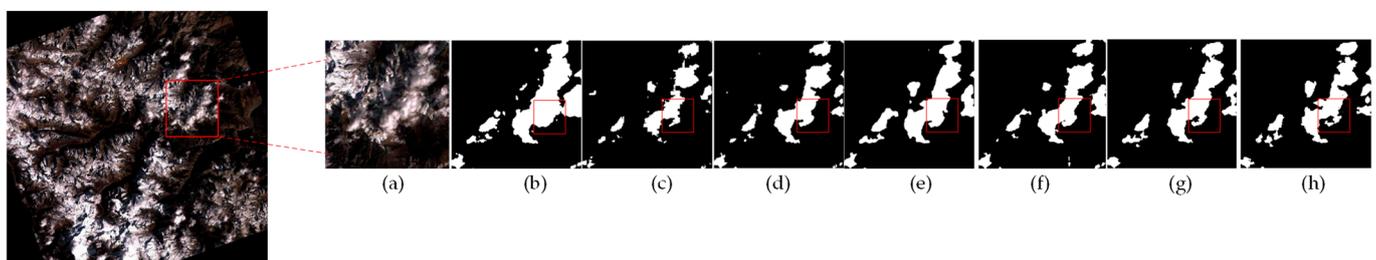


Figure 7. Experimental outputs of Cloud Extraction Results using Different models in Image1 (local enlarged in the red rectangle). (a) Raw image; (b) TransUNet; (c) SwinUNet; (d) U-Net; (e) DeeplabV3+; (f) MSACN-small; (g) MSACN; (h) Ground Truth.

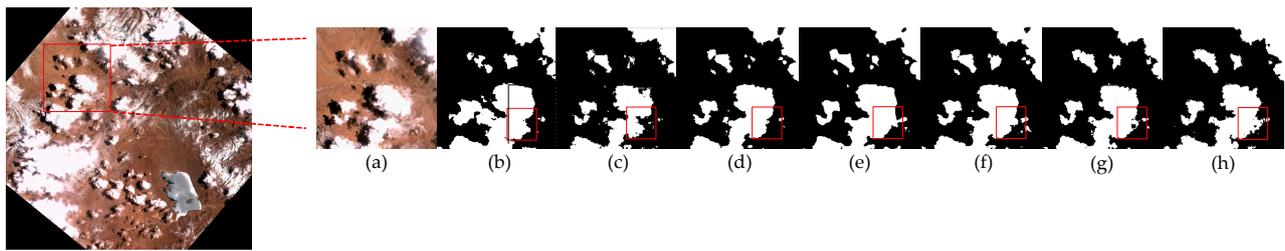


Figure 8. Experimental outputs of Cloud Extraction Results using Different models in Image2 (local enlarged in the red rectangle). (a) Raw image; (b) TransUNet; (c) SwinUNet; (d) U-Net; (e) DeeplabV3+; (f) MSACN-small; (g) MSACN; (h) Ground Truth.

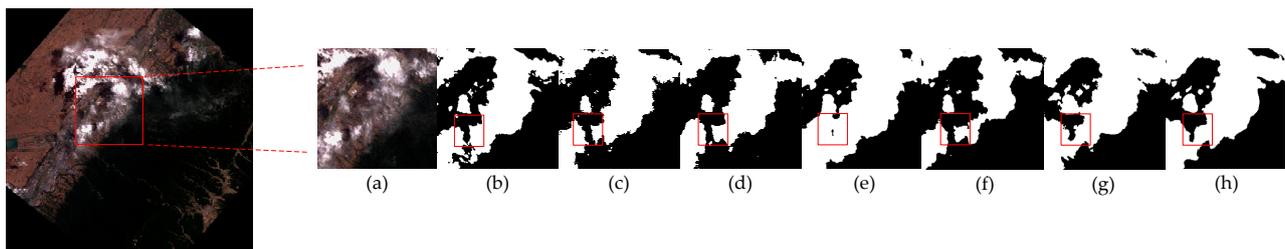


Figure 9. Experimental outputs of Cloud Extraction Results using Different models in Image3 (local enlarged in the red rectangle). (a) Raw image; (b) TransUNet; (c) SwinUNet; (d) U-Net; (e) DeeplabV3+; (f) MSACN-small; (g) MSACN; (h) Ground Truth.

As can be seen in Figure 7, clouds cover a snowy mountain. The reflection between clouds and snow is very similar, and the extraction of clouds in satellite images is more susceptible to interference from such high-reflectivity ground objects such as snow or ice. Figure 7h shows the results of MSACN. Compared with the Transformer-based methods in Figure 7b,c and the CNN-based approaches in Figure 7d,e, MSACN is shown in Figure 7g. In the red rectangle area, MSACN can clearly distinguish clouds from the complex background containing snow. From this point of view, MSACN has excellent cloud detection capabilities and can resist the interference of snow factors in complex environments. From the data indicators of image1 in Table 3, we can see that the accuracy of MSACN is 4.7% higher than TransUNet and 1.33% higher than U-Net; the F1 of MSACN is 12.23% higher than SwinUNet with Swin Transformer as the backbone, and higher than U-Net is 5.46% higher. Although MSACN pays more attention to the boundaries and details of cloud layers, its focus on details results in poor performance of the Recall indicator.

Figure 8 shows an image of thick clouds. The underlying surface of the entire image includes a body of water, large areas of land, and hillsides. These make the cloud detection task difficult. As exhibited in Figure 8, the process of extracting cloud details by MSACN is more effective than the other two methods. The results of MSACN are better than U-Net, but the thin cloud segments of those neighborhood pixels are very blurry. There are two reasons for this. First, cloud images are of high resolution and contain a variety of objects. Second, cloud segmentation is different from natural image segmentation. The task is a pixel-to-pixel binary classification problem, which focuses on retaining edge details. However, the network we proposed obtains inter-class distance information by combining the advantages of multi-head soft attention modules and CNN, and strengthens the control of important features, thereby making the model have better spatial adaptability. This integrated method provides an effective solution for the fine division of cloud layer boundaries and brings significant improvements and enhancements to cloud detection tasks. In the area chosen by the red rectangle in Figure 8, the advantages of MSACN can be clearly seen. As shown in the indicators in Table 3, MSACN's accuracy is 2.75% higher than SwinUNet and 1.93% higher than U-Net; MSACN's IoU is 4.74% higher than Transformer-based SwinUNet and 1.19% higher than CNN-based DeeplabV3+.

In Figure 9, there is haze and clouds of varying thickness covering half of the image. In contrast, U-Net and DeeplabV3+ perform poorly in distinguishing thin clouds and cloud-free areas. We use the image3 part in Figure 9 to evaluate the performance of MSACN when facing various boundary thin and thick clouds. As shown in Figure 9e, it can be seen that DeeplabV3+ cannot process detailed information, resulting in difficulties in distinguishing fine clouds. According to the indicators of image3 in Table 3, MSACN's accuracy is 3.58% higher than that based on Transformer and 2.43% higher than DeeplabV3+; MSACN's F1 is 5.63% higher than SwinUNet and 3.01% higher than U-Net. Compared results show that MSACN can obtain excellent results regardless of complex backgrounds or unevenly thick clouds. The results of various indicators evident the feasibility and effectiveness of MSACN architecture.

3.3. Ablation Analysis

In this part, to explore the availability of MSACN, we explore whether the difference in the heads of soft attention modules in the backbone affects the overall model accuracy from two aspects: the difference in the backbone and the number of attention modules. As can be seen from Table 4, the backbone uses ResNet50 and the number of head soft attention modules is set to 8, which works best. So, the results indicate a nuanced trade-off between the complexity of the model structure, as defined by the multiple heads, and its overall performance. While a higher number of heads can enhance accuracy, it might incur an additional computational burden. At the same time, the ResNet50 backbone, especially with eight heads, emerges as particularly effective, achieving the highest quantitative evaluation indexes like accuracy, IoU, Precision, Recall, and F1 among the configurations tested.

Table 4. Accuracy evaluation of different network structure complexities and different backbones.

Backbone	Heads	Accuracy	IoU	Precision	Recall	F1
VGG16	4	0.9724	0.9317	0.9679	0.9612	0.9645
	8	0.9684	0.9267	0.9643	0.9592	0.9617
ResNet50	4	0.9667	0.9283	0.9693	0.9569	0.9626
	8	0.9857	0.9510	0.9761	0.9737	0.9748

4. Discussion

In this section, in order to assess the adaptability of MSACN, we further discuss the results of training and validation from other satellites of two different sensors, Landsat OLI 8 and GF-1. From the results, we try to discuss the relationship between the resolution of satellite images and the depth of CNN.

For the test image collected from Landsat OLI 8, we present the comparative experimental results using different methods, shown in Table 5. Overall, the results suggest that MSACN-small stands out in achieving a balance between accuracy, IoU, and precision-recall trade-offs. In contemplating the observed results, it is plausible that the 30m resolution of Landsat OLI 8 imagery might be a contributing factor. Lower resolutions often necessitate less complex neural networks to achieve satisfactory results. Additionally, as depicted in Table 5, for lower-resolution images, various cloud extraction methods exhibit minimal discrepancies in their outcomes. MSACN-small achieved the highest accuracy (97.18%), closely followed by TransUNet (97.16%). MSACN-small exhibits the highest IoU (94.50%), indicating superior spatial overlap between predicted and ground truth cloud images. TransUNet and DeeplabV3+ also show competitive IoU values, reflecting their strong segmentation performance. Meanwhile, MSACN-small consistently performs well, balancing precision and recall effectively.

Table 5. Comparative experimental results of 38-cloud dataset with different lr (learning rate).

lr	Method	Accuracy	IoU	Precision	Recall	F1
1×10^{-3}	TransUNet	0.9716	0.9429	0.9712	0.9700	0.9706
	SwinUNet	0.9518	0.9045	0.9508	0.9489	0.9494
	U-Net	0.9688	0.9384	0.9595	0.9665	0.9628
	DeeplabV3+	0.9697	0.9393	0.9719	0.9661	0.9743
	MSACN-small	0.9718	0.9450	0.9735	0.9707	0.9717
	MSACN	0.9635	0.9268	0.9678	0.9575	0.9619
7×10^{-5}	TransUNet	0.9699	0.9381	0.9685	0.9680	0.9681
	SwinUNet	0.9453	0.8937	0.9450	0.9427	0.9434
	U-Net	0.9670	0.9353	0.9674	0.9666	0.9666
	DeeplabV3+	0.9696	0.9409	0.9702	0.9695	0.9703
	MSACN-small	0.9641	0.9278	0.9648	0.9604	0.9625
	MSACN	0.9691	0.9391	0.9687	0.9684	0.9686

In Table 6, we present the comparative experimental results for the GF1-HWU dataset, evaluating various cloud-extracting methods. Across all evaluated metrics, the cloud extraction methods exhibit consistently high performance. MSACN-small stands out with the highest scores, indicating robust performance in terms of accuracy, IoU, and precision-recall balance. In the GF-1 dataset, both TransUNet and SwinUNet exhibit oscillating results with comparatively lower IoU scores of 94.84% and 93.84%, respectively. Contrarily, conventional CNN methods like U-Net and DeeplabV3+ exhibit more stable and robust outcomes in the given context. The MSACN series methods consistently showcase outstanding performance across different datasets. This dual accomplishment not only affirms the effectiveness of the method but also indicates a certain level of robustness in its methodology.

Table 6. Comparative experimental results of GF-HWU dataset with different lr (learning rate).

lr	Method	Accuracy	IoU	Precision	Recall	F1
1×10^{-3}	TransUNet	0.9734	0.9484	0.9870	0.9734	0.9865
	SwinUNet	0.9730	0.9384	0.9662	0.9697	0.9678
	U-Net	0.9878	0.9757	0.9881	0.9874	0.9877
	DeeplabV3+	0.9885	0.9759	0.9878	0.9878	0.9879
	MSACN-small	0.9907	0.9794	0.9896	0.9896	0.9896
	MSACN	0.9889	0.9749	0.9884	0.9861	0.9873
7×10^{-5}	TransUNet	0.9824	0.9656	0.9809	0.9824	0.9811
	SwinUNet	0.9778	0.9491	0.9724	0.9751	0.9737
	U-Net	0.9662	0.9335	0.9649	0.9664	0.9656
	DeeplabV3+	0.9886	0.9752	0.9879	0.9870	0.9837
	MSACN-small	0.9907	0.9810	0.9897	0.9913	0.9904
	MSACN	0.9872	0.9742	0.9864	0.9874	0.9869

To assess the performance of MSACN at different learning rates and the linear relationship between the learning rate and the image resolution, we try to use different learning rates for experimental comparison. In many experiments on the 38-cloud dataset, by comparing the model indicators in the table, it can be seen that the accuracy of the large learning rate is higher. In particular, the lr has a great effect on the Transform-based model. In the CNN-based model, MSACN-small, which removes the low-level convolution-pooling block, can explore the relationship between network structure and image resolution. In the GF-WHU dataset, regardless of the lr, the accuracy of MSACN-small is 0.29% higher than that of U-Net. It can be seen that the GF-WHU dataset works better in the shallow network. The 38-cloud dataset is affected by the hyper-parameter and learning rate, and the shallow network MSACN-small only performs well in large learning rates. This is all due to the fact that the 38-cloud dataset and the GF-WHU dataset have the characteristics of spatial-

resolution images. The lower-resolution satellite images have fewer details, and the shallow network makes it easier to capture the texture features in the image. The 38-cloud where the low-resolution images are located is more susceptible to the impact of the learning rate, so it shows better results at a larger learning rate. Both Tables 5 and 6 illustrate that the optimal models in all cases are based on CNN, so the Transformer may not be able to make full use of its self-attention when processing these images due to the low-resolution remote sensing image datasets with less-detailed mechanism and sequence modeling capabilities. In contrast, CNNs excel at extracting image features and capturing local and global context. The following conclusions can be depicted from the experimental data:

- For high-resolution complex scenes, there is a certain relationship between model accuracy and attention module;
- The size of lr: (a) When the data are low-resolution images, a large learning rate works better. (b) The learning rate has a greater impact on the fusion attention module, and a small learning rate is better.

5. Conclusions

In conclusion, a Multiscale Soft Attention Convolutional Neural Network structure is proposed to alleviate the mixture-pixel problem of cloud extraction in optical satellite images. MSACN is an end-to-end structure that consists of a concurrent dilated convolution module and a multi-head soft attention module. Experiments prove that for high-resolution images, MSACN achieves remarkable results. There is a certain relationship between the performance of the CNN structure and the attention mechanism for the complex scenes. For ZY-3 satellite images, the combined dilated convolution module and multi-head soft attention model significantly enhance the extraction accuracy of complex scenarios in satellite imagery with cloud and snow. To further assess the effectiveness of MSACN, we compared TransUnet, SwinUnet, UNet, and DeepLabV3+. The experimental outputs demonstrate the outstanding performance of MSACN. On the ZY-3 dataset, the accuracy of MSACN achieves 98.79%. Although MSACN has good performance on cloud extracting tasks, there are still several directions for further improvement, including optimization of hyper-parameters and network structure. Furthermore, we will make more improvements in proposing the novel cloud extraction method integrating prior knowledge as a post-processing step.

Author Contributions: Conceptualization, L.G. and C.G.; methodology, C.G.; investigation, S.L.; resources, S.L.; writing—original draft preparation, C.G.; writing—review and editing, L.G.; supervision, J.Z.; project administration, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by Liaoning Provincial Department of Education Youth Project (No. 1030040000560), National Natural Science Foundation of China (No. 42071428), China Scholarship Council (No. 202208210120), Liaoning Province Applied Basic Research Program (Youth Special Project) (2023JH2/101600038), Shenyang Youth Science and Technology Innovation Talent Support Program (RC220458).

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest: Author Sijun Lu was employed by Shen Kan Engineering & Technology Corporation, MCC. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Miroszewski, A.; Mielczarek, J.; Czelusta, G.; Szczepanek, F.; Grabowski, B.; Le Saux, B.; Nalepa, J. Detecting Clouds in Multispectral Satellite Images Using Quantum-Kernel Support Vector Machines. *arXiv* **2023**, arXiv:2302.08270. [[CrossRef](#)]
2. Li, W.; Zhang, F.; Lin, H.; Chen, X.; Li, J.; Han, W. Cloud Detection and Classification Algorithms for Himawari-8 Imager Measurements Based on Deep Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4107117. [[CrossRef](#)]

3. 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; pp. 6939–6942.
4. Liu, C.; Yang, S.; Di, D.; Yang, Y.; Zhou, C.; Hu, X.; Sohn, B.-J. A Machine Learning-Based Cloud Detection Algorithm for the Himawari-8 Spectral Image. *Adv. Atmos. Sci.* **2022**, *39*, 1994–2007. [[CrossRef](#)]
5. Massetti, L.; Materassi, A.; Sabatini, F. NSKY-CD: A System for Cloud Detection Based on Night Sky Brightness and Sky Temperature. *Remote Sens.* **2023**, *15*, 3063. [[CrossRef](#)]
6. Zekoll, V.; de los Reyes, R.; Richter, R. A Newly Developed Algorithm for Cloud Shadow Detection—TIP Method. *Remote Sens.* **2022**, *14*, 2922. [[CrossRef](#)]
7. Wang, J.; Yang, D.; Chen, S.; Zhu, X.; Wu, S.; Bogonovich, M.; Guo, Z.; Zhu, Z.; Wu, J. Automatic Cloud and Cloud Shadow Detection in Tropical Areas for PlanetScope Satellite Images. *Remote Sens. Environ.* **2021**, *264*, 112604. [[CrossRef](#)]
8. Kang, X.; Gao, G.; Hao, Q.; Li, S. A Coarse-to-Fine Method for Cloud Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 110–114. [[CrossRef](#)]
9. Xiong, Q.; Wang, Y.; Liu, D.; Ye, S.; Du, Z.; Liu, W.; Huang, J.; Su, W.; Zhu, D.; Yao, X.; et al. A Cloud Detection Approach Based on Hybrid Multispectral Features with Dynamic Thresholds for GF-1 Remote Sensing Images. *Remote Sens.* **2020**, *12*, 450. [[CrossRef](#)]
10. Singh, R.; Biswas, M.; Pal, M. Cloud Detection Using Sentinel 2 Imageries: A Comparison of XGBoost, RF, SVM, and CNN Algorithms. *Geocarto Int.* **2022**, *38*, 1–32. [[CrossRef](#)]
11. Sui, Y.; He, B.; Fu, T. Energy-Based Cloud Detection in Multispectral Images Based on the SVM Technique. *Int. J. Remote Sens.* **2019**, *40*, 5530–5543. [[CrossRef](#)]
12. Shao, M.; Zou, Y. Multi-Spectral Cloud Detection Based on a Multi-Dimensional and Multi-Grained Dense Cascade Forest. *J. Appl. Remote Sens.* **2021**, *15*, 028507. [[CrossRef](#)]
13. Duta, I.C.; Liu, L.; Zhu, F.; Shao, L. Pyramidal Convolution: Rethinking Convolutional Neural Networks for Visual Recognition. *arXiv* **2020**, arXiv:2006.11538.
14. Hu, J.; Zhang, X.; Yang, C. Cloud detection in RGB color remote sensing images based on improved M-type convolutional network. *Adv. Laser Optoelectron.* **2019**, *56*, 229–238.
15. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Springer Int. Publ.* **2015**, *9351*, 234–241.
16. 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.* **2019**, *16*, 618–622. [[CrossRef](#)]
17. 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](#)]
18. Cao, K.; Zhang, X. An Improved Res-UNet Model for Tree Species Classification Using Airborne High-Resolution Images. *Remote Sens.* **2020**, *12*, 1128. [[CrossRef](#)]
19. Shi, C.; Zhou, Y.; Qiu, B.; Guo, D.; Li, M. CloudU-Net: A Deep Convolutional Neural Network Architecture for Daytime and Nighttime Cloud Images' Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1688–1692. [[CrossRef](#)]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Proc. Conf. Neural Inf. Proc. Syst.* **2017**, *30*, 5998–6008.
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
22. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
23. Ghaffarian, S.; Valente, J.; Van Der Voort, M.; Tekinerdogan, B. Effect of Attention Mechanism in Deep Learning-Based Remote Sensing Image Processing: A Systematic Literature Review. *Remote Sens.* **2021**, *13*, 2965. [[CrossRef](#)]
24. Zhang, L.; Sun, J.; Yang, X.; Jiang, R.; Ye, Q. Improving Deep Learning-Based Cloud Detection for Satellite Images With Attention Mechanism. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 6005505. [[CrossRef](#)]
25. Yao, X.; Guo, Q.; Li, A. Light-Weight Cloud Detection Network for Optical Remote Sensing Images with Attention-Based DeeplabV3+ Architecture. *Remote Sens.* **2021**, *13*, 3617. [[CrossRef](#)]
26. Yao, Z.; Jia, J.; Qian, Y. MCNet: Multi-Scale Feature Extraction and Content-Aware Reassembly Cloud Detection Model for Remote Sensing Images. *Symmetry* **2021**, *13*, 28. [[CrossRef](#)]
27. Zhang, J.; Shi, X.; Wu, J.; Song, L.; Li, Y. Cloud Detection Method Based on Spatial-Spectral Features and Encoder-Decoder Feature Fusion. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5407915. [[CrossRef](#)]
28. Zhang, J.; Wu, J.; Wang, H.; Wang, Y.; Li, Y. Cloud Detection Method Using CNN Based on Cascaded Feature Attention and Channel Attention. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4104717. [[CrossRef](#)]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. 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.

31. Gao, L.; Song, W.; Tan, H.; Liu, Y. Cloud Detection Based on Multi-Scale Dilation Convolutional Neural Network for ZY-3 Satellite Remote Sensing Imagery. *Acta Opt. Sin.* **2019**, *39*, 0104002.
32. Mohajerani, S.; Krammer, T.A.; Saeedi, P. Cloud Detection Algorithm for Remote Sensing Images Using Fully Convolutional Neural Networks. *arXiv* **2018**, arXiv:1810.05782.
33. Mohajerani, S.; Saeedi, P. Cloud-Net: An End-to-End Cloud Detection Algorithm for Landsat 8 Imagery. *arXiv* **2019**, arXiv:1901.10077.
34. Li, Z.; Shen, H.; Li, H.; Xia, G.; Gamba, P.; Zhang, L. Multi-Feature Combined Cloud and Cloud Shadow Detection in GaoFen-1 Wide Field of View Imagery. *Remote Sens. Environ.* **2017**. [[CrossRef](#)]
35. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv* **2018**, arXiv:1802.02611.
36. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
37. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv* **2021**, arXiv:2105.05537.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.