remote sensing

MDPI

Article

UATNet: U-Shape Attention-Based Transformer Net for Meteorological Satellite Cloud Recognition

Zhanjie Wang [1,2], Jianghua Zhao [1], Ran Zhang [1,2], Zheng Li [1], Qinghui Lin [1] and Xuezhi Wang [1,*]

[1] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China; wangzhanjie@cnic.cn (Z.W.); zjh@cnic.cn (J.Z.); zhangran@cnic.cn (R.Z.); lizheng@cnic.cn (Z.L.); lqh@cnic.cn (Q.L.)

[2] University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: wxz@cnic.cn

Abstract: Cloud recognition is a basic task in ground meteorological observation. It is of great significance to accurately identify cloud types from long-time-series satellite cloud images for improving the reliability and accuracy of weather forecasting. However, different from ground-based cloud images with a small observation range and easy operation, satellite cloud images have a wider cloud coverage area and contain more surface features. Hence, it is difficult to effectively extract the structural shape, area size, contour shape, hue, shadow and texture of clouds through traditional deep learning methods. In order to analyze the regional cloud type characteristics effectively, we construct a China region meteorological satellite cloud image dataset named CRMSCD, which consists of nine cloud types and the clear sky (cloudless). In this paper, we propose a novel neural network model, UATNet, which can realize the pixel-level classification of meteorological satellite cloud images. Our model efficiently integrates the spatial and multi-channel information of clouds. Specifically, several transformer blocks with modified self-attention computation (swin transformer blocks) and patch merging operations are used to build a hierarchical transformer, and spatial displacement is introduced to construct long-distance cross-window connections. In addition, we introduce a Channel Cross fusion with Transformer (CCT) to guide the multi-scale channel fusion, and design an Attention-based Squeeze and Excitation (ASE) to effectively connect the fused multi-scale channel information to the decoder features. The experimental results demonstrate that the proposed model achieved 82.33% PA, 67.79% MPA, 54.51% MIoU and 70.96% FWIoU on CRMSCD. Compared with the existing models, our method produces more precise segmentation performance, which demonstrates its superiority on meteorological satellite cloud recognition tasks.

Keywords: cloud recognition; semantic segmentation; transformer; meteorological satellite cloud image; attention mechanism

Citation: Wang, Z.; Zhao, J.; Zhang, R.; Li, Z.; Lin, Q.; Wang, X. UATNet: U-Shape Attention-Based Transformer Net for Meteorological Satellite Cloud Recognition. *Remote Sens.* **2022**, *14*, 104. https://doi.org/10.3390/rs14010104

Academic Editor: Luis Gómez-Chova

Received: 15 November 2021
Accepted: 23 December 2021
Published: 26 December 2021

1. Introduction

According to the global cloud cover data provided by the International Satellite Cloud Climatology Project (ISCCP), more than 66% area above the earth is covered by a large number of clouds [1]. Cloud, an important member of the climate system, is the most common, extremely active and changeable weather phenomenon. It directly affects the radiation and water cycle of the earth-atmosphere system, and plays an important role in the global energy budget and water resources distribution [2,3]. Therefore, cloud observation is a significant content in meteorological work. It is fundamental for weather forecasting and climate research to correctly identifying such elements as cloud shape, cloud amount and cloud height, as well as the distribution and change of clouds, which also plays a key role in navigation and positioning, flight support and national economic development [4]. There are four main types of cloud observation: ground-based artificial observation, ground-based instrument observation, aircraft or balloon observation and

meteorological satellite observation. It is noted that the cloud classification standards of ground-based observation and meteorological satellite observation are different. For ground observation, the cloud type is determined according to the cloud base height and cloud shape. For satellite observation, cloud classification is usually based on the spectral characteristics, texture characteristics and spatio-temporal gradient of the cloud top. Ground-based artificial observation mainly relies on meteorologists, easily restricted by factors such as insufficient observation experience. Ground-based instrument observation also has the disadvantages of large nighttime errors and a limited observation area. Aircraft or balloon observation is too time-consuming and costly to apply to daily operations. Meteorological satellites are widely used for large-scale and continuous-time observations of clouds and the earth's surface. With the continuous development of satellite remote sensing technology and imaging technology, the quality, spatial resolution and timeliness of cloud images has greatly improved. The new generation of geostationary satellites, such as the Himawari-8 and Himawari-9 satellites [5] in Japan, the GEOS-R satellite [6] in the United States, and the FY-4A satellite [7] in China, can meet higher observation requirements. Since satellite cloud images cover a wide area and contain more surface features, they are more suitable for describing the cloud information and changes in a large range. Cloud recognition based on satellite cloud images has become an important application and research hotspot in the remote sensing field.

In our work, we focus on the overall cloud type distribution in China. Zhuo et al. built a ground-based cloud dataset of Beijing, China, which was collected from August 2010 to May 2011 and annotated by meteorologists from the China Meteorological Administration. It contains eight cloud types and the clear sky [8]. Zhang et al. used three ground-based cloud datasets, captured in Wuxi, Jiangsu Province and Yangjiang, Guangdong Province, China. The datasets were labeled by experts from the Meteorological Observation Center of China Meteorological Administration, Chinese Academy of Meteorological Sciences and the Institute of Atmospheric Physics, Chinese Academy of Sciences. They all contain six cloud types and the clear sky [9]. Fang et al. used the standard ground-based cloud dataset provided by Huayunshengda (Beijing) meteorological technology limited liability company, which was labeled as 10 cloud types and the clear sky [10]. Liu et al. used the multi-modal ground-based cloud dataset (MGCD), the first one composed of ground-based cloud images and multi-modal information. MGCD was annotated by meteorological experts and ground-based cloud-related researchers as six cloud types and the clear sky [11]. Liu et al. selected the FY-2 satellite cloud image dataset. FY-2 is the first operational geostationary meteorological satellite of China. The experimental dataset was collected from June to August 2007 and annotated by meteorological experts with richly educated and trained experience as six cloud types, ocean and land [12]. Bai et al. used Gao Fen-1 and Gao Fen-2 satellite cloud image datasets for cloud detection tasks. Gaofen is a series of Chinese high-resolution Earth imaging satellite for the state-sponsored program, China High-resolution Earth Observation System (CHEOS). The images and the manual cloud mask were acquired from the National Disaster Reduction Center of China. The majority of the images contain both cloud and non-cloud regions. Cloud regions include small-, medium-, and large-sized clouds. The backgrounds are common underlying surface environments including mountains, buildings, roads, agriculture, and rivers [13]. Cai et al. used the FY-2 satellite cloud image dataset, which was labeled as four cloud types and the clear sky, but it is only suitable for image-level classification, not pixel-level classification [14].

The coverage of the ground-based cloud image is relatively small, and it is impossible to show the overall cloud distribution of China in a single image. The existing China region satellite cloud image datasets have the limitations of fewer cloud types and a lack of pixel-level classification. Therefore, we construct a 4km resolution meteorological satellite cloud image dataset covering the entire China region and propose a novel deep learning method to accurately identify the cloud type distribution in China, so that researchers can analyze the temporal and spatial distribution characteristics of cloud amount, cloud water

path and cloud optical thickness in China from the perspective of different regions and different clouds.

Cloud recognition methods mainly include threshold-based methods, traditional machine learning methods, and deep learning methods. Threshold-based methods [15–17] determine appropriate thresholds for different sensors through specific channels of the image (reflectivity, brightness temperature, etc.) to identify cloud regions in a fast calculation speed. However, they ignore the structure and texture of the cloud, for which is difficult to determine an appropriate threshold for the situation with many cloud types. In contrast, machine learning methods [18–25] are more robust. They first classify regions with the same or similar pixels into one class, and then analyze the spectral, spatial and texture information of the image with pixel-based or object-oriented methods. The texture measurement, location information, brightness temperature, reflectance and NDVI index are fed into SVM, KNN and Adaboost algorithms as features to realize automatic cloud classification. However, the traditional machine learning methods also have great limitations. Most features are extracted manually, and the accuracy is comparatively low when processing high-resolution images, which makes it difficult to distinguish clouds from highly similar objects.

Many studies show that deep learning methods can adaptively learn the deep features of clouds and have higher detection accuracy than traditional machine learning methods [26–31]. Liu et al. introduced a neural network for satellite cloud detection tasks, and conducted experiments on the FY-2C satellite cloud image dataset. Its nadir spatial resolution is 1.25 km for visible channels, and 5 km for infrared channels. Their model improved the results greatly not only in pixel-level accuracy but also in cloud patch-level classification by more accurately identifying cloud types such as cumulonimbus, cirrus and clouds at high latitudes [12]. Cai et al. also constructed a convolution neural network for satellite cloud classification on the FY-2C satellite images, which could automatically learn features and obtain better classification results than those of traditional machine learning methods [14]. Liu et al. presented a novel joint fusion convolutional neural network (JFCNN) to integrate the multimodal information, which learns the heterogeneous features (visual features and multimodal features) from the cloud data for cloud classification with robustness to environmental factors [32]. Zhang et al. proposed transfer deep local binary patterns (TDLBP) and weighted metric learning (WML). The former can handle view shift well, and the latter solves the problem of an uneven number of different cloud types [9]. Zhang et al. developed a new convolutional neural network model, CloudNet, to perform cloud recognition tasks in a self-built dataset Cirrus Cumulus Stratus Nimbus, which can accurately identify 11 cloud types, including one cloud generated by human activities [33]. Lu et al. proposed two segnet-based architectures, P_Segnet and NP_Segnet, for the cloud recognition of remote sensing images, and adopted parallel structures in the architectures to improve the accuracy of cloud recognition [34]. Fang et al. trained five network models by fine-tuning network parameters and freezing weights of different network layers based on the cloud image dataset provided by standard weather stations after data enhancement, and used five network migration configurations on the enhanced dataset. Experiments showed that the fine-tuned densenet model achieved good results [10]. Liu et al. proposed a novel method named multi-evidence and multi-modal fusion network (MMFN) for cloud recognition, which can learn extended cloud information by fusing heterogeneous features in a unified framework [11]. Zhang et al. presented LCCNet, a lightweight convolutional neural network model, which has the lower parameter amount and operation complexity, stronger characterization ability and higher classification accuracy than the existing network models [35]. According to the analysis above, the existing research results still have the following two deficiencies: 1. Most cloud recognition methods are designed on ground-based cloud images, which can only be used to study local cloud distribution and changes without universality; 2. Current cloud recognition methods based on deep learning cannot fully capture the context information in images, and their feature extraction ability must be improved.

Context information is the key factor for improving image segmentation performance, and the receptive field roughly determines how much information can be utilized by the network. Existing deep learning cloud recognition methods are mostly implemented based on a convolutional neural network (CNN). However, relevant studies [36] show that the actual receptive field of CNN is much smaller than its theoretical receptive field. Therefore, the limited receptive field seriously restricts the representation ability of the model. To solve this problem, Transformer [37] is introduced into semantic segmentation tasks. Its characteristic is that it not only keeps the spatial resolution of input and output unchanged, but also effectively captures the global context information. Axial-DeepLab [38] is the first independent attention model with large or global receptive fields, which can make good use of location information without increasing the computational cost and serve as the backbone network for semantic segmentation tasks. However, Axial-DeepLab uses a specially designed axial attention, which has poor scalability to standard computing equipment. By comparison, SETR [39] is easier to use with standard self-attention. It adopts a structure similar to Vision Transformer (VIT) [40] for feature extraction and combines with a decoder to restore resolution, achieving good results on segmentation tasks. Given the balance between computation cost and performance, we build a hierarchical transformer in encoder and divide the input image into windows, so that self-attention can be calculated in sub-windows to ensure a linear relationship between computational complexity and the size of the input image. We conduct attention operations along the channel axis between the encoder and decoder, so that the decoder can better integrate the features of the encoder and reduce the semantic gap.

In summary, the research on deep learning satellite cloud recognition methods is of great significance and application value for improving the accuracy of weather forecast, the effectiveness of climate model prediction and the understanding of global climate change. We construct a China region meteorological satellite cloud image dataset (CRMSCD) based on the L1 data product of FY-4A satellite and the cloud classification results of Himawari-8 satellite. CRMSCD contains nine cloud types and the clear sky (cloudless) and conforms to the world Meteorological Organization standard. In this paper, we propose a cloud recognition network model based on the U-shaped architecture, in which the transformer is introduced to build the encoder and encoder–decoder connection, and the attention mechanism is designed to integrate the features of both the encoder and decoder. Consequently, we name it as U-shape Attention-based Transformer Net (UATNet). UATNet has more powerful extracting capabilities of spectrum and spatial information features and stronger adaptabilities to the changing characteristics of clouds. In addition, we propose two models of different sizes to fit varying requirements.

To summarize, we make the following major contributions in this work:

(1) We propose the UATNet model and introduce a transformer into meteorological satellite cloud recognition task, which solves the problem of CNN receptive field limitation and captures global context information effectively while ensuring the computing efficiency.
(2) We use two transformer structures in UATNet to perform attention operations along the patch axis and channel axis, respectively, which can effectively integrate the spatial information and multi-channel information of clouds, extract more targeted cloud features, and then obtain pixel-level cloud classification results.
(3) We construct a China region meteorological satellite cloud image dataset named CRMSCD and carry out experiments on it. Experimental results demonstrate that the proposed model achieved a significant performance improvement compared with the existing state-of-the-art methods.
(4) We discover that replacing batch normalization with switchable normalization in the convolution layers of a fully convolutional network and using encoder–decoder connection in the transformer model can significantly improve the effect of cloud recognition.

2. Materials and Methods

2.1. Data Introduction

In this paper, we use the images taken by the FY-4A meteorological satellite as the data source, and the cloud classification product of the Himawari-8 satellite as the label type, and construct the meteorological satellite cloud image dataset over China region through cropping and data enhancement. In the cloud classification product, there are 9 types of clouds: cirrus (Ci), cirrostratus (Cs), deep convection (Dc), altocumulus (Ac), altostratus (As), nimbostratus (Ns), cumulus (Cu), stratocumulus (Sc) and stratus (St). In addition, it also includes the clear sky, the mixed value and the invalid value.

FY-4 satellite is the second generation of China's geostationary meteorological satellite series used in quantitative remote sensing meteorology, with greatly enhanced capabilities for high-impact weather event monitoring, warning, and forecasting. Following on from the first generation, it offers several advances over the FY-2 [41]: detection efficiency, spectral band, spatial resolution, time resolution, radiation calibration and sensitivity [42]. It is equipped with multi-channel scanning imaging radiometer, interferometric atmospheric vertical sounder, lightning imager and space environment monitoring instruments. The scanning imaging radiometer mainly undertakes the task of obtaining cloud images. AGRI, a key optical sensor on FY-4A, has 14 spectral bands (increased from five bands on FY-2) and has significantly improved capabilities for cloud, convective system, land surface, environmental observations, and even data assimilation [43]. FY-4 not only observes clouds, water vapor, vegetation, and the land surface, just as the FY-2 can, but also has the ability to capture aerosols and snow. It can also clearly distinguish between different cloud forms and high/middle water vapor. Different from the limitation of the single visible light channel of FY-2, FY-4 generates color satellite cloud images for the first time, ideally generating regional observation images in one minute. In dataset construction, we select 4KML1 data of the China region taken by the imager of FY-4A satellite L1 data product. The band, spatial resolution, sensitivity, and usage of different channels are presented in Table 1, where ρ = reflectivity, S/N = signal to noise, NEΔT = noise equivalent differential.

Table 1. FY-4A AGRI specifications.

Channel	FY-4A AGRI			Main Application
	Band (μm)	Spatial Resolution (km)	Sensitivity	
Visible light and near infrared	0.45~0.49	1.0	S/N ≥ 90(ρ = 100%)	Aerosol, visibility
	0.55~0.75	0.5~1.0	S/N ≥ 200(ρ = 100%)	Vegetation, fog, cloud
	0.75~0.90	1.0	S/N ≥ 5 (ρ = 1%) @0.5K	Vegetation, aerosol
Shortwave infrared	1.36~1.39	2.0	S/N ≥ 200(ρ = 100%)	Cirrus
	1.58~1.64	2.0		Cloud, snow
	2.1~2.35	2.0~4.0		Cirrus, aerosol
Midwave infrared	3.5~4.0 (high)	2.0	NEΔT ≤ 0.7 K (300 K)	Cloud, fire
	3.5~4.0 (low)	4.0	NEΔT ≤ 0.2 K (300 K)	Land surface
Water vapor	5.8~6.7	4.0	NEΔT ≤ 0.3 K (260 K)	Upper-level water vapor
	6.9~7.3	4.0	NEΔT ≤ 0.3 K (260 K)	Mid-level water vapor
Longwave infrared	8.0~9.0	4.0	NEΔT ≤ 0.2 K (300 K)	Water vapor, cloud
	10.3~11.3	4.0	NEΔT ≤ 0.2 K (300 K)	Cloud, surface temperature
	11.5~12.5	4.0	NEΔT ≤ 0.2 K (300 K)	Cloud, water vapor, surface temperature
	13.2~13.8	4.0	NEΔT ≤ 0.2 K (300 K)	Cloud, water vapor

Himawari-8 is a new generation of Japanese geostationary meteorological satellites that carry state-of-the-art optical sensors with significantly higher radiometric, spectral, and spatial resolution than those previously available in geostationary orbit. It captures a full-disk image every 10 min, and the observation range is from 60°S to 60°N and from 80°E to 160°W. As listed in Table 2, the satellite has 16 observation spectral bands, including 3 visible light bands, 3 near-infrared bands and 10 infrared bands. The Himawari-8 cloud classification product we selected is based on meteorological properties, and is generated by threshold judgment of the fundamental cloud product and the advanced Himawari imager data. The fundamental cloud product contains cloud mask (including surface condition data), cloud type and cloud top height. The advanced Himawari imager data refers to the brightness temperature of bands 08, 10 and 13 [44].

Table 2. Himawari-8 specifications.

Channel	Himawari-8		Main Application
	Band(μm)	Spatial Resolution (km)	
Visible light	0.43~0.48	1.0	Vegetation, aerosol
	0.50~0.52	1.0	Vegetation, aerosol
	0.63~0.66	0.5	Low clouds, fog
Near infrared	0.85~0.87	1.0	Vegetation, aerosol, cirrus
	1.60~1.62	2.0	Cloud phase
	2.25~2.27	2.0	Particle size
Shortwave infrared	3.74~3.96	2.0	Low clouds, fog, fire, land
Water vapor	6.06~6.43	2.0	Upper-level water vapor
	6.89~7.01	2.0	Mid-level water vapor
	7.26~7.43	2.0	Low-level water vapor
Infrared	8.44~8.76	2.0	Cloud phase, SO_2
	9.54~9.72	2.0	O_3
	10.3~10.6	2.0	Cloud, cloud top information
	11.1~11.3	2.0	Cloud, sea-surface temperature
	12.2~12.5	2.0	Cloud, sea-surface temperature
	13.2~13.4	2.0	CO_2, cloud top height

In order to enrich the cloud types of the dataset, we observe the image data products of the FY-4A satellite, and then adjust the study area based on the territory of China. The study area ranges in longitude from 80°E to 139.95°E and in latitude from 5°N to 54°N. The location of the study area is illustrated in Figure 1.

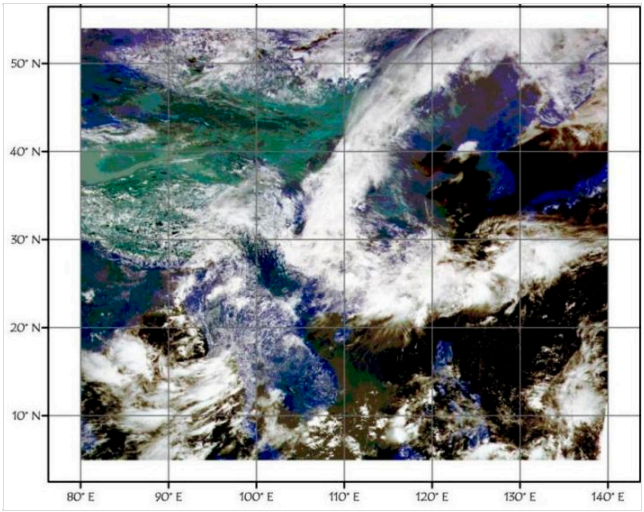


Figure 1. Location of the study area.

2.2. Data Processing

Based on FY-4A satellite data and Himawari-8 observation data, we carried out detailed preprocessing work. First, we screened FY-4A satellite data on a time scale based on solar illumination, and simultaneously performed projection transformation and geometric correction on a spatial scale. Then, the processed data were aligned with Himawari-8 observation data. Finally, the final dataset CRMSCD was obtained by cropping the China region and converting the file format. The detailed data processing flow is shown in Figure 2.

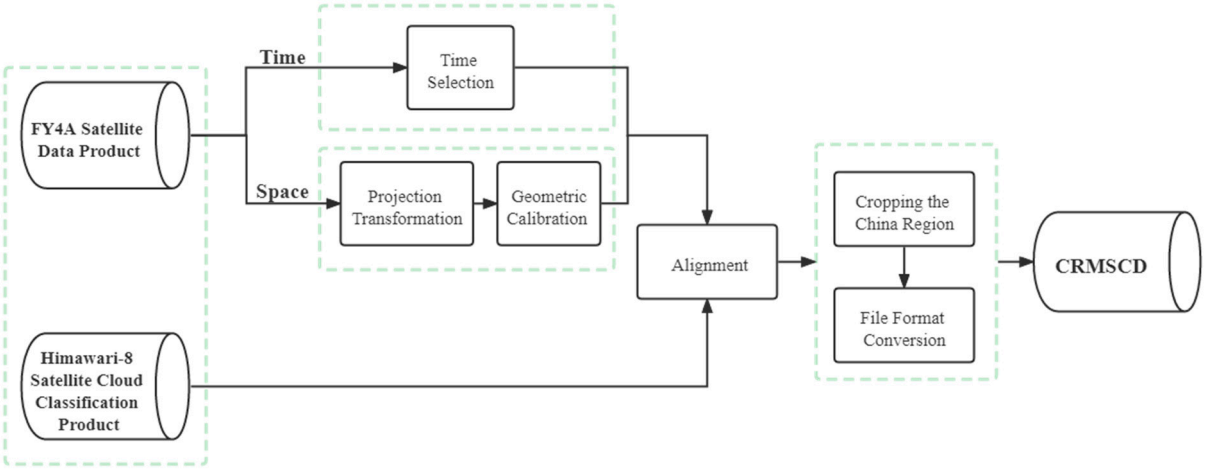


Figure 2. Data processing flow.

2.2.1. Time Alignment

FY-4A satellite L1 data products are based on Coordinated Universal Time (UTC). According to the mapping between UTC and Beijing time, 03:30 and 04:30 in UTC correspond to 11:30 and 12:30 in Beijing time. During this period, the sun was fully illuminated, and the images could not be blocked by shadows, which was convenient for subsequent experimental research. Hence, we selected FY-4A images taken at 03:30 and 04:30. The sampling interval of FY-4A satellite was about 14 times per hour; the sampling interval of Himawari-8 satellite was 6 times per hour. Therefore, we selected the Himawari-8 satellite cloud classification results of the most recent earth observation time relative to the FY-4A image for time alignment. The data alignment method is illustrated in Table 3.

Table 3. Data alignment.

Geostationary Orbit Satellite	FY-4A	Himawari-8
Data Product Name	FY4A-_AGRI--_N_REGC_1047E_L1-_FDI-_MULT_NOM_20200710033000_20200710033417_4000M_V0001.HDF	NC_H08_20200710_0330_L2CLP010_FLDK.02401_02401.nc
Time alignment	20200710033000_20200710033417	20200710_0330

2.2.2. Projection Transformation

In spatial scale, we performed projection transformation on FY-4A data, and read FY-4A AGRI first-level data (HDF image), according to the specified latitude and longitude range and spatial resolution. Related indicators are presented in Table 4.

Table 4. FY-4A-related indicators of National Satellite Meteorological Center.

ea (km)	eb (km)	h (km)	λD	$LOFF$ / $COFF$		$LFAC$ / $CFAC$	
6378.1	6356.8	42164	$deg2rad(104.7)$	0500 M	10,991.5	0500 M	81,865,099
				1000 M	5495.5	1000 M	40,932,549
				2000 M	2747.5	2000 M	20,466,274
				4000 M	1373.5	4000 M	10,233,137

In the table, ea is the semi-major axis of the earth; eb is the short axis of the earth; h is the distance from the earth's center to the satellite's barycenter; λD is the longitude of the satellite's suborbital point; $LOFF / COFF$ is the line/column offset; $LFAC / CFAC$ is the line/column scaling factor; $deg2rad(104.7)$ indicates the radian corresponding to the angle of 104.7°; '0500 M', '1000 M', '2000 M', and '4000 M' indicate the file resolution.

According to the specified range of latitude and longitude, we convert geographic longitude and latitude into geocentric longitude and latitude. Moreover, the geocentric coordinate system is converted to a plane rectangular coordinate system, then x, y are converted into $line$ and $column$. The formula is as follows:

$$line = LOFF[resolution] + 2^{-16} y \cdot LFAC[resolution] \quad (1)$$

$$column = COFF[resolution] + 2^{-16} x \cdot CFAC[resolution] \quad (2)$$

where resolution is specified as 4000M.

2.2.3. Dataset Construction

Table 1 lists the 14 channels of FY-4A AGRI and their corresponding main applications. According to the main applications, the 9 channels, corresponding to visible light and near infrared, short-wave infrared, medium-wave infrared and long-wave infrared, play a crucial role in cloud recognition tasks. Therefore, we selected the data of 14 channels to make the standard dataset.

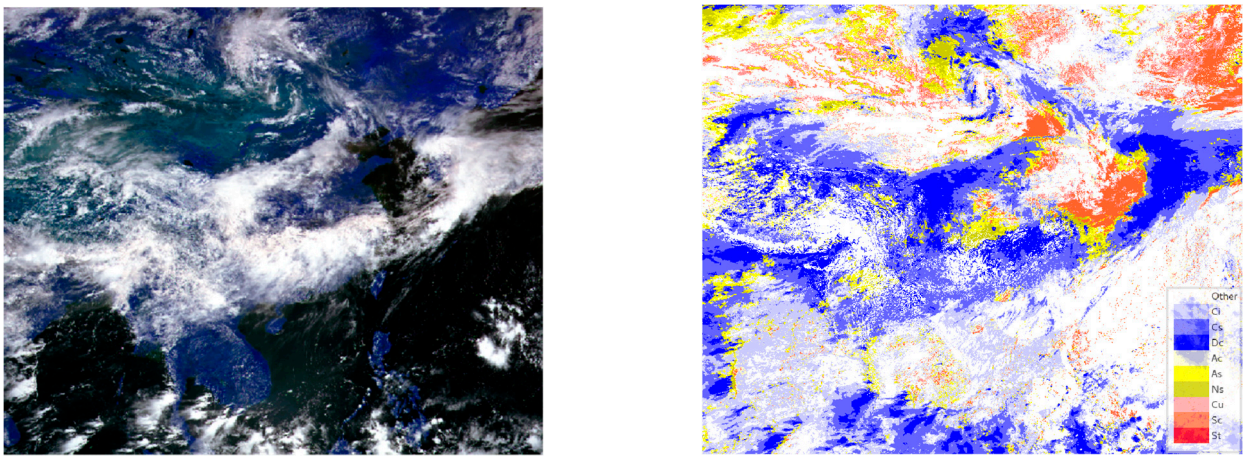
In a specific operation, according to the file channel number corresponding to 4000 M resolution and the line and column number of the full-disc data in the China region, we performed data extraction and geometric calibration. The geometric calibration values included digital quantization number (dn), reflectance, radiation brightness and brightness temperature. The channel number and the number of line and column corresponding to each resolution are listed in Table 5.

Table 5. The channel number and the number of line and column.

Resolution Ratio	Channel Number	The Line and Column Number of the Full-Disc Data
0500 M	Channel02	21,984
1000 M	Channel01, Channel02, Channel03	10,992
2000 M	Channel01–Channel07	5496
4000 M	Channel01–Channel14	2748

We referred to the official lookup table and coefficients to calculate the values of the different channels. According to the both of these, Channel01-Channel06 indicate the reflectivity. Channel07-Channel14 indicate the brightness temperature according to the lookup table, and radiation according to the coefficient. We made a geometric calibration for numerical indices, and converted the Hierarchical Data Format (HDF) file of FY-4A and the Network Common Data Format (NetCDF) file of Himawari-8 into Tag Image File Format (TIFF) files. The calibrated resolution of FY-4A's 3D scanner was 0.05°. According to the designated learning area and the calibrated resolution, we cropped the generated TIFF file to obtain the standard experimental dataset CRMSCD as the China region meteorological satellite cloud image dataset.

CRMSCD contains FY-4A daily images and corresponding Himawari-8 labels at 3:30 and 4:30 from 1 January 2020 to 31 July 2020. We select the data at 4:30 on 10 July 2020 to present the visualization results of our dataset. The meteorological satellite cloud image and the corresponding label are shown in Figure 3.



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

Figure 3. CRMSCD visualization results: (**a**) FY-4A images at 4:30 on July 10, 2020; (**b**) Himawari-8 labels at 4:30 on July 10, 2020. The color corresponding to the label value is located in the lower right corner of (**b**).

We take the data from 11 July to 31 July as the test set, and the remaining sample data are shuffled and randomly divided into the training set and verification set according to the ratio of 7 to 3. The final dataset contains 426 images of 1200 × 981 pixels.

2.3. UATNet

Deep learning cloud recognition methods are mostly improved based on the CNN architecture. The actual receptive field of CNN is much smaller than its theoretical receptive field, which seriously restricts the representation ability of the model. Therefore, we propose an end-to-end deep learning network UATNet to learn the mapping from image patch sequences to corresponding semantic labels. The specific structure of UATNet is

shown in Figure 4. Two different transformer structures, a hierarchical transformer computed with shifted windows (swin transformer) [45] and channel transformer, are introduced into the encoder and encoder–decoder connection of the U-shaped architecture, which conduct attention operations along the patch axis and the channel axis to effectively integrate spatial information and multi-channel information of clouds. Different from CNN, the transformer can obtain the global receptive field without stacking. In this section, we describe the two transformer structures in detail.

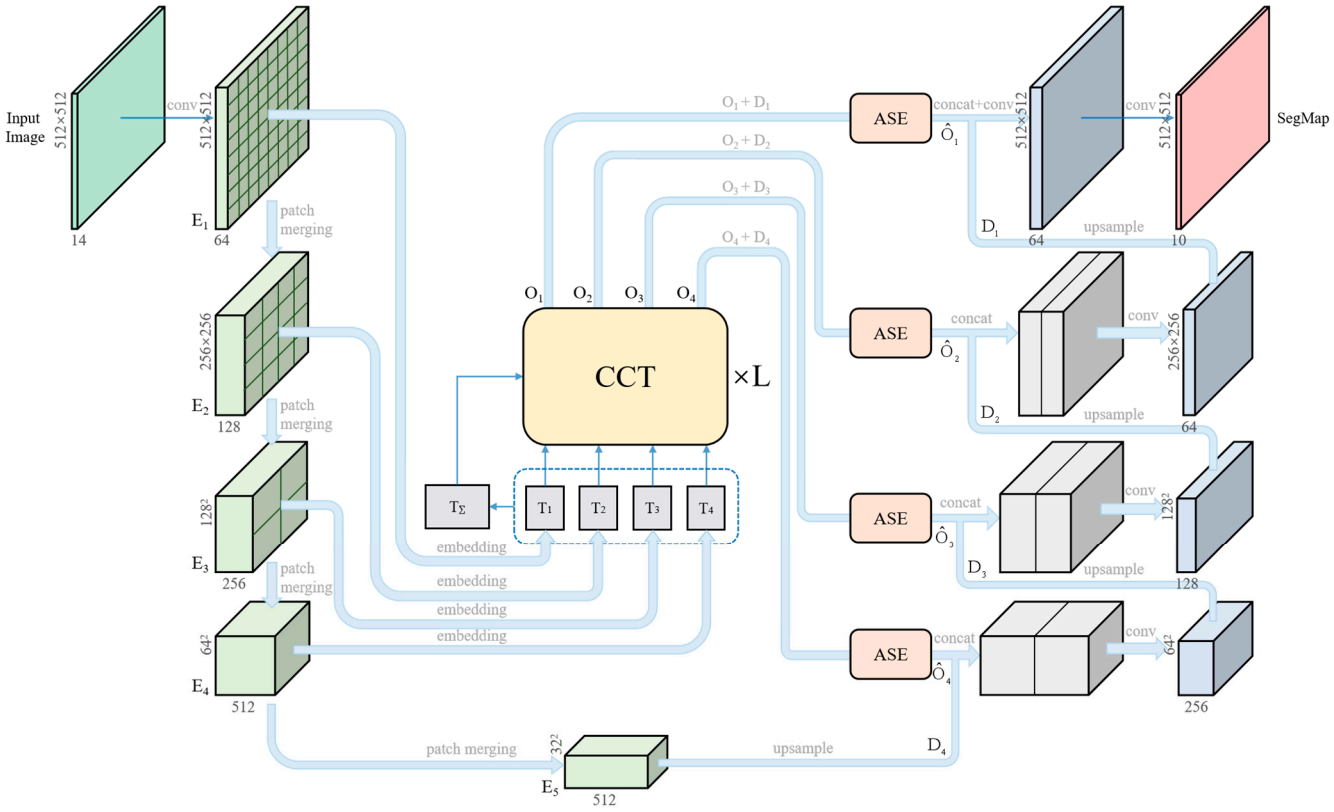


Figure 4. The architecture of UATNet. CCT and ASE are Channel-wise Cross fusion Transformer and Attention-based Squeeze and Excitation, respectively.

2.3.1. Encoder Architecture

UATNet encoder extracts high-level semantic features and transforms low-dimensional features into abstract high-dimensional feature vectors. E1~E5 in Figure 4 illustrate the specific framework of UATNet encoder. In our encoder architecture, there are four stages, which are denoted as stage1~stage4 according to the increase in network depth. Each stage contains two parts: patch merging and an even number of consecutive swin transformer blocks, in which swin transformer blocks are several transformer blocks with modified self-attention computation applied on the patch tokens.

The size of the input image is 14 × H × W. We change the channel dimension of the input image to C dimension (C is set to 64 in implementation) by 3 × 3 convolution, and then input it into stage1. In stage1, the images are divided with the size of C × H × W into a set of non-overlapping image patches through patch partition, where the size of each image patch is 2 × 2, the characteristic dimension of each image patch is 14 × 2 × 2 = 56, and the number of image patches is $\frac{H}{2}\times\frac{W}{2}$. Then, we stretch the grid containing $\frac{H}{2}\times\frac{W}{2}$ patches into a sequence and call a trainable linear projection function to further map each vectorized patch to a potential 2C-dimensional embedding space to obtain the

one-dimensional image patch tokens of the image and input them into an even number of consecutive transformer blocks. The operation of stage2~stage4 is roughly the same as stage1. In stage2~stage4, patch merging is used instead of patch partition to merge the input according to 2 × 2 adjacent patches, and the dimensions of the linear projection function mapping are 4C, 8C and 8C, respectively. Since the patch partition of stage1 is equivalent to the merging operation of adjacent patches on a single pixel (1 × 1 image patch), we also mark this operation as patch merging in Figure 4. As the network depth increases, patch merging can be used to construct a hierarchical transformer.

Tokens are fed into a Multi-head Self Attention (MSA), followed by a Multi-Layer Perceptron (MLP). We conduct the Layer Normalization (LN) before each MSA and MLP, and residual joins after each MSA and MLP. The sliding window mechanism is adopted in each MSA module to divide the input feature maps into non-overlapping windows, and then perform self-attention calculation in different windows, each of which contains 16 × 16 patches. To encode the spatial information of each position, we learn a special position embedding based on the relative position bias of the patch in the window, and add it to the image patch tokens. The relative position along each axis lies in the range [−15, 15]. Although the self-attention mechanism in the transformer is disordered, spatial location information related to the original location information is embedded in the input.

As illustrated in Figure 5, we used a shifted window partition similar to swin transformer to solve the information exchange problem of different windows, and alternately used Window-based Multi-head Self-Attention (W-MSA) and Window-based Multi-head Self Attention with spatial displacement (SD W-MSA) in two consecutive transformer structures. W-MSA uses regular window partitioning from the upper left corner. SD W-MSA adopts a windowing configuration that is shifted from that of the preceding layer, by displacing the windows by (8, 8) pixels from the regularly partitioned windows. Since the number of patches in the window is much smaller than that in picture, the computational complexity of W-MSA has a linear relationship with the image size. The partition method of shifting windows introduces connections between adjacent non-coincidence windows of the upper layer, greatly increasing the receptive field.

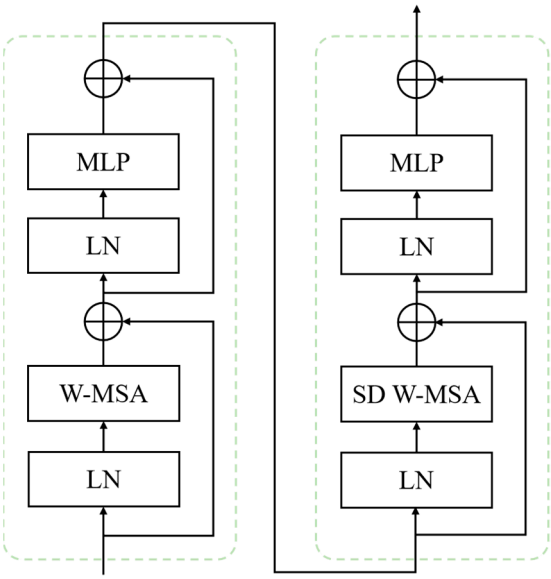


Figure 5. Two successive swin transformer blocks in encoder. W-MSA and SD W-MSA are Window-based Multi-head Self-Attention and Window-based Multi-head Self Attention with spatial displacement, respectively.

Spatial displacement [46] reshapes the spatial dimension based on the window size and the number of tokens, and stretches it into a sequence. We introduce spatial displace-

ment in SD W-MSA to package inputs from different windows by spatial dimension transformation, and then establish a long-distance cross-window connection. After the self-attention operation, we use spatial alignment [46] to adjust tokens to the original position so that the features and the image are spatially aligned. To reduce the number of calculations, we combine window division with spatial displacement, and combine window-to-image conversion with spatial alignment.

2.3.2. Feature Fusion of Encoder and Decoder

Figure 4 shows the visualization process of encoder–decoder feature fusion. E1~E5 are feature maps of encoder, and D1~D4 are feature maps of decoder. In this section, we introduce the process in detail.

We tokenized the output feature maps of the transformer layers at 4 different scales corresponding to E1~E4 in encoder, and reshaped the features into flattened 2D patches with patch sizes $\left\{P, \frac{P}{2}, \frac{P}{4}, \frac{P}{8}\right\}$, respectively, so that the patches can be mapped to the same areas of the encoder features at four scales. We used the 4 layers of tokens $T_k(k=1,2,3,4)$, $T_k \in R^{\frac{HW}{k^2} \times C_k}$ as queries, and then concatenate the tokens of four layers T_k to obtain $T_\Sigma = Concat\left(T_1, T_2, T_3, T_4\right)$ as the key and value. We input the $T_k(k=1,2,3,4)$ and T_Σ into the Channel-wise Cross Fusion Transformer (CCT) [47], and used the transformer's long-distance dependent modeling advantages to fuse four different scale encoder features. In order to better assign attention weights and make the gradient smoothly propagated, we conducted the attention operation along the channel axis rather than the patch axis.

In the implementation, we set the number of attention heads to 4 and built a 4-layer CCT. The four outputs of the 4th layer O_1, O_2, O_3 and O_4, are reconstructed though an upsampling operation followed by a convolution layer, and then concatenated with the decoder features D_1, D_2, D_3 and D_4, respectively.

We used the *k*-th channel transformer output $O_k \in R^{C \times H \times W}, k=(1,2,3,4)$ and the *k*-th decoder feature map $D_k \in R^{C \times H \times W}, k=(1,2,3,4)$ as the input of Attention-based Squeeze and Excitation (ASE). Figure 6 shows the architecture of ASE, where r is the dimension reduction coefficient.

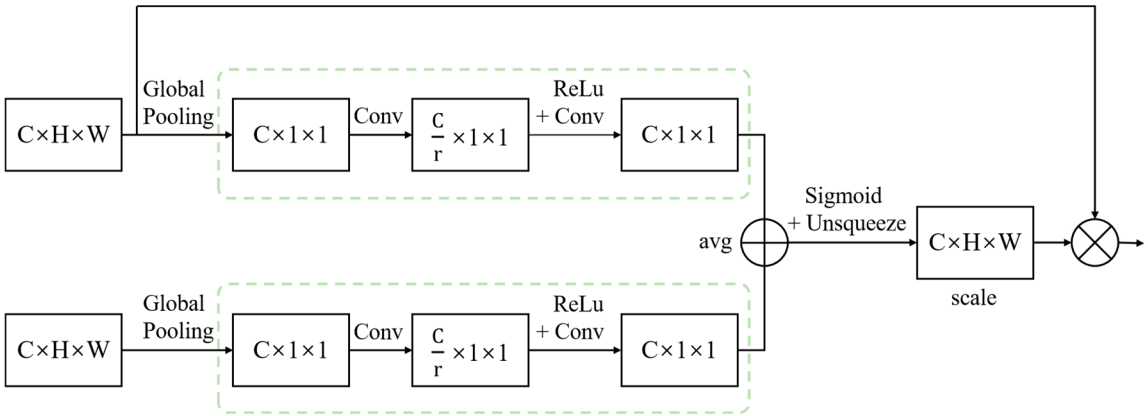


Figure 6. The architecture of Attention-based Squeeze and Excitation (ASE).

In ASE, we conducted feature compression of O_k and D_k along the spatial dimension through a squeeze operation [48] to obtain the global receptive field. Spatial squeeze is performed by a global average pooling (GAP) layer:

$$G(O_k^c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} O_k^c(i,j), \quad G(O_k^c) \in R^C \tag{3}$$

$$G(D_k^c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} D_k^c(i,j), \quad G(D_k^c) \in R^C \tag{4}$$

where c is the *c*-th channel, k is the *k*-th layer. Then, in order to improve the generalization ability of the model and better learn the nonlinear relationship between different channels, we performed an excitation operation [48] to reduce the dimension of feature channels, and restore the original dimensions to obtain O_k' and D_k' after using Rectified Linear Unit (ReLU) activation function. We use 1 × 1 convolution in specific operations:

$$O_k' = Conv_{1\times1}(\mathrm{Re}\,LU(Conv_{1\times1}(G(O_k)))) \tag{5}$$

$$D_k' = Conv_{1\times1}(\mathrm{Re}\,LU(Conv_{1\times1}(G(D_k)))) \tag{6}$$

We calculated the average of O_k' and D_k' to obtain Z_k. Here, we use a gating mechanism in the form of a sigmoid to generate weights for each characteristic channel of Z, which explicitly models the importance of different channels. After unsqueezing the result vector Z_k, we multiplied it by O_k to obtain $\hat{O}_k$ through ReLU activation function:

$$scale_k = unsqueeze(\sigma(Z_k)) \tag{7}$$

$$\hat{O}_k = \mathrm{Re}\,LU(O_k \cdot scale_k) \tag{8}$$

where $\sigma(\cdot)$ is denoted as the sigmoid function; $\hat{O}_k$ and D_k are concatenated in the channel dimension. After 3 × 3 convolution operation in series, D_{k-1} can be obtained.

ASE automatically obtains the importance of each feature channel of Z_k through learning, and then enhances useful features and suppresses features that are not useful for the current task according to their importance, thus achieving adaptive calibration of feature channels. Therefore, ASE can integrate encoder fusion features with decoder features to solve the problem of semantic inconsistency between encoder and decoder.

3. Results

3.1. Implementation Details

We employ all experiments on CentOS release 7.9 system and use NVIDIA A100-PCIE GPU card with 40 GB memory for graphics acceleration. Experiments are implemented with PyTorch, an open-source Python machine learning library. We do not use any pre-trained weights to train UATNet. For CRMSCD, we set the batch size to 16. The input resolution and patch size are set as 512 × 512 and 4.

To obtain a fast convergence, we also employ the AdamW optimizer, Adam with decoupled weight decay, to train our models, where the initial learning rate is set to 0.002, combined with polynomial (poly) learning rate decay strategy. The specific formula is as follows:

$$lr = base\_lr \times \left(1 - \frac{iter}{\max iter}\right)^{power} \tag{9}$$

where $base\_lr$ is the initial learning rate, $power$ is the decay index, $iter$ is the current training step size, and $\max iter$ is the maximum training step size. When we establish $power$ to be equal to 1, the learning rate decay curve is a straight line. When

$power > 1$, the decay curve of learning rate becomes concave inward. When $power < 1$, the decay curve of learning rate becomes convex outward. Figure 7 shows the decay curves under different power values. We compare four power values in the figure, the effect is better when the power is 1.5. In practice, we set $power = 1.5$ for parameter optimization.

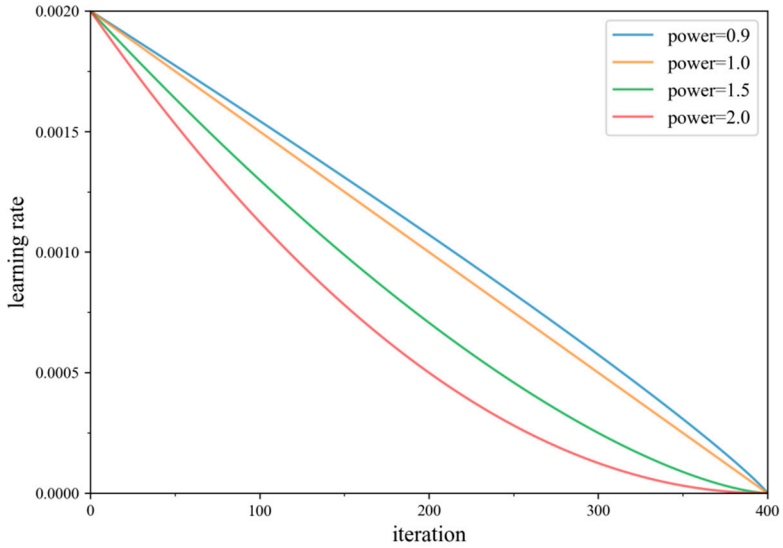


Figure 7. The decay curves under different power values.

In practice, we set the epoch number to 400 and employ the focal loss [49] as our loss function to train our network. Focal loss can address the class imbalance by reshaping the standard cross entropy loss such that it down-weights the loss assigned to well-classified examples, the formula is as follows:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \quad p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \tag{10}$$

where $p \in [0,1]$ is the model's estimated probability for the class with label $y = 1$, α_t is a weighting factor to address class imbalance, γ is a tunable focusing parameter. We set $\alpha_t = 0.4$ and $\gamma = 2$ in our experiments. Note that we use the same settings and loss functions to train all the baselines. In convolution layers, we uniformly use batch normalization and ReLU activation function.

3.2. Evaluation Indicator

We use Pixel Accuracy (PA), Mean Pixel Accuracy (MPA), Mean Intersection over Union (MIoU), and Frequency Weighted Intersection over Union (FWIoU) as evaluation indicators in order to evaluate the performance of UATNet on CRMSCD. The higher the evaluation indicators, the better the effect of our models.

Assuming that there are k + 1 classes of samples (including k target classes and 1 background class), p_{ij} represents the total number of pixels that are labeled as class i but classified as class j. That is, p_{ii} represents the total number of pixels both classified and labeled as class i, p_{ij} and p_{ji} are false positives and false negatives, respectively.

(1) Pixel Accuracy (PA) and Mean Pixel Accuracy (MPA)

Pixel accuracy (PA) can express the classification accuracy of pixel points and calculate the ratio between the amount of properly classified pixels and their total number. The formula is as follows:

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{11}$$

Mean pixel accuracy (MPA), is an improved version of PA which computes the ratio of correct pixels on a per-class basis. MPA is also referred to as class average accuracy. The formula is as follows:

$$MPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{12}$$

(2) Mean Intersection over Union (MIoU) and Frequency Weighted Intersection over Union (FWIoU)

Intersection over Union (IoU) is a standard metric used for comparing the similarity and diversity of sample sets. In semantic segmentation, it is the ratio of the intersection of the predicted segmentation with the ground truth, to their union. Mean Intersection over Union (MIoU) is the class-averaged IoU. The definition is given in (13):

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{13}$$

Frequency Weighted Intersection over Union (FWIoU) is an enhancement of MIoU. It sets weights for each class according to its frequency. The weights are multiplied by the IoU of each class and summed up. The formula of FWIoU is given in (14):

$$FWIoU = \frac{1}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{14}$$

3.3. Data Augmentation

The CLTYPE variable of the Himawari-8 NetCDF file stores 12 different cloud storage types. In addition to nine cloud types and the clear sky, it also contains the mixed value and the invalid value. The mixed value indicates that the cloud type of the pixel cannot be determined, and the invalid value indicates that the cloud image data of the pixel is missing. In order to better complete the cloud recognition task, we calculate the number of mixed pixels and invalid pixels in the CRMSCD, and the two types do not exist in the training set, validation set, and test set after experimental verification. The corresponding value and color information for all types are shown in Table 6.

Table 6. Annotation and color information of cloud types.

Cloud Type	Annotation	Color Information
Cirrus (Ci)	1	(204 204 255)
Cirrostratus (Cs)	2	(102 102 255)
Deep convection (Dc)	3	(0 0 255)
Altocumulus (Ac)	4	(255 255 204)
Altostratus (As)	5	(255 255 0)
Nimbostratus (Ns)	6	(204 204 0)

Cumulus (Cu)	7	(255 153 153)
Stratocumulus (Sc)	8	(255 102 51)
Stratus (St)	9	(255 0 0)
Clear	255	(255 255 255)

To reduce the number of parameters and calculations, we perform image cropping on CRMSCD, with the cropping size of 512 × 512 and randomly selected positions. The size of 512 × 512 matches the clipping size in test, which can reduce the boundary error caused by the stitching results, and randomly selected positions are helpful for improving the generalization ability of the model. In addition, random flipping (including horizontal flipping and vertical flipping) is performed on the training set to avoid overfitting. The pixel value range of the input image is adjusted to 0~255. According to Table 6, we convert the annotations and then normalize each channel of the image.

3.4. Experimental Analysis

3.4.1. UATNet Variants

To study the relationship between model size and model performance, we propose two models of different sizes, UATNet-S and UATNet-B. Swin-T, Swin-S, Swin-B, and Swin-L are four different-sized models of swin transformer. The channel numbers of UATNet-S and UATNet-B are similar to those of Swin-B, and the layer numbers of UATNet-S and UATNet-B are similar to those of Swin-T and Swin-S, respectively.

In the two versions, head dimension is set to 32, channel number is set to 128, and window size is set to 16. The difference between them is the number of encoder layers and the number of attention heads. In UATNet-S, the number of network layers of four stages is {2,2,6,2}, and the number of attention heads is {3,6,12,24}. In UATNet-B, the number of network layers of four stages is {2,2,18,2}, and the number of attention heads is {4,8,16,32}. Table 7 lists the specific settings of the two versions of UATNet.

Table 7. UATNet variants.

Model	Channel Number	Window Size	Layers	Head Dimension	Heads
UATNet-S	128	16	{2, 2, 6, 2}	32	{3, 6, 12, 24}
UATNet-B	128	16	{2, 2, 18, 2}	32	{4, 8, 16, 32}

3.4.2. Comparison on CRMSCD

In the test, we crop each input image to 512 × 512 with a total of 12 image blocks and the overlap size of 256. We input the generated image blocks to UATNet for prediction, and then stitch the output results together to obtain a 1200 × 981 image, and finally save it in HDF format. In our evaluation, we focus on the recognition of cirrostratus, deep convection, altostratus, nimbostratus and stratocumulus, which is more helpful to meteorological research.

We compare UATNet with two types of methods for comprehensive evaluation, covering eight mainstream fully convolutional networks: U-Net [50], UNet3+ [51], PSPNet [52], SegNet [53], DeepLabv3+ [54], GSCNN [55], HRNetV2 [56] and HRNetV2 + OCR [57] and three state-of-the-art transformer-based segmentation methods, including SETR, swin transformer, and UCTransNet. Note that the decoder of swin transformer is implemented with Progressive Upsampling (PUP). The experimental results are reported in Table 8, where the best results are in bold.

Table 8. Results of semantic segmentation on CRMSCD test set. The best results are in bold.

Method	CRMSCD			
	PA (%)	MPA (%)	MIoU (%)	FWIoU (%)

U-Net	80.50	65.45	50.56	68.85
UNet3+	79.59	60.01	47.25	67.45
PSPNet	74.55	52.37	39.44	61.17
SegNet	78.41	60.35	45.98	66.23
DeepLabv3+	79.32	62.21	48.20	67.17
GSCNN	80.96	64.92	51.37	69.22
HRNetV2	79.62	62.84	48.73	67.55
HRNetV2 + OCR	80.49	64.23	50.48	68.60
SETR	69.24	40.34	29.50	54.74
Swin-T	72.01	47.73	35.53	58.21
Swin-S	71.65	46.53	34.57	57.71
Swin-B	71.67	46.46	34.52	57.76
Swin-L	71.75	47.87	35.30	58.01
UCTransNet	81.49	66.26	52.62	69.87
UATNet-S	81.82	67.14	53.36	70.35
UATNet-B	**82.33**	**67.79**	**54.51**	**70.96**

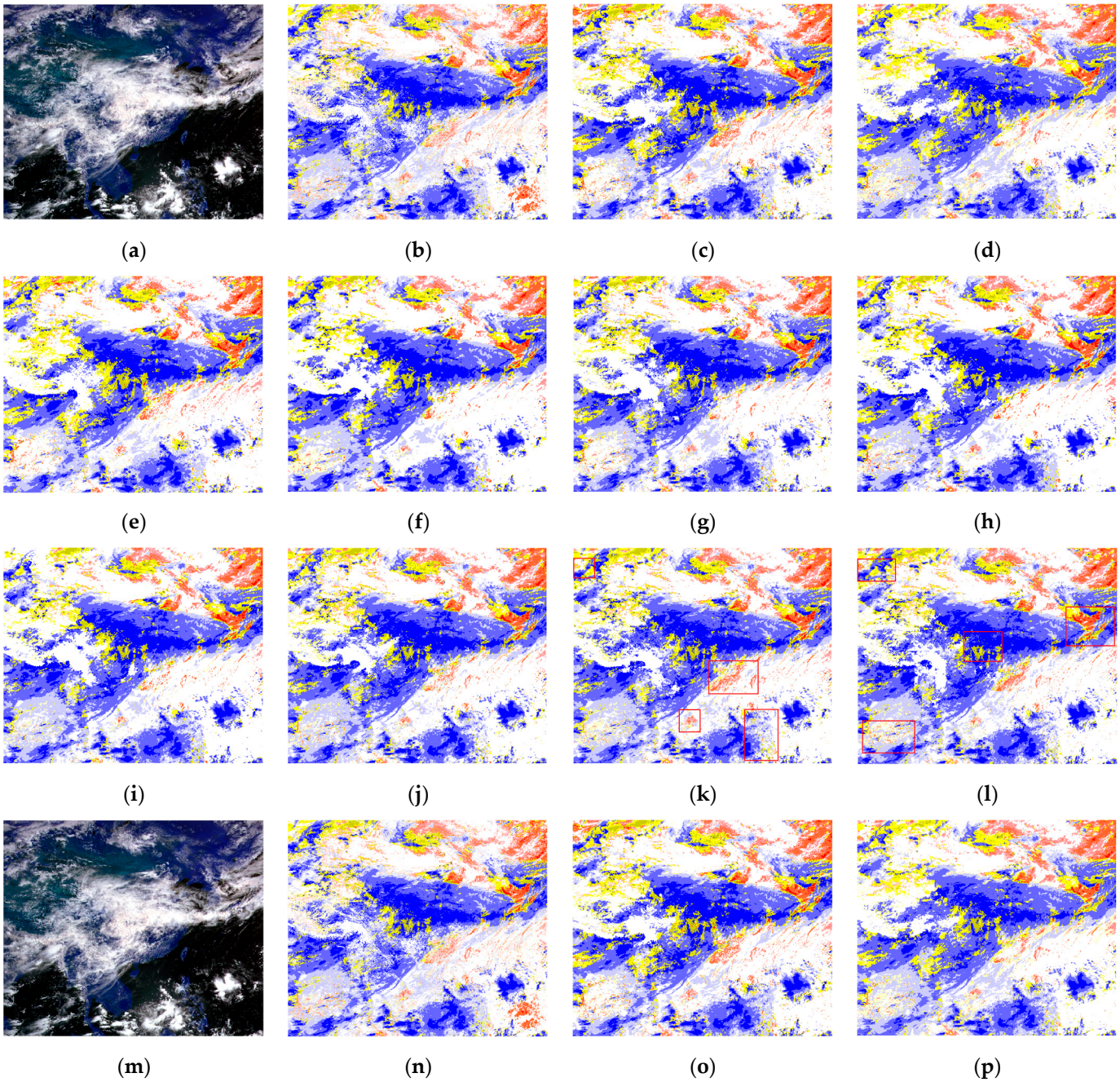
Table 8 illustrates the overall performance and cloud recognition effect of UATNet on CRMSCD. Compared with the fully convolutional networks, PA, MPA, MIoU and FWIoU gain of UATNet-B range from 1.37% to 7.78%, from 2.34% to 15.42%, from 3.14% to 15.07%, from 1.74% to 9.79%, respectively. Compared with the transformer-based models, PA, MPA, MIoU and FWIoU gain of UATNet-B range from 0.84% to 13.09%, from 1.53% to 27.45%, from 1.89% to 25.01%, from 1.09% to 16.22%, respectively. The above results show that UATNet is effective for pixel-level classification of meteorological satellite cloud images, and all of the evaluation indicators are better than the current state-of-the-art neural network methods. Furthermore, we summarize the causes for the difference between the experimental results and the annotation of Himawari-8: 1. The sampling interval of FY-4A satellite is about 14 times per hour, while the sampling interval of Himawari-8 satellite is 6 times per hour. In the time alignment phase, we select the adjacent time for data alignment based on sampling interval of the two satellites, as shown in Table 3, which causes a certain time error. 2. The Himawari-8 satellite cloud classification product (Cloud Type) is classified based on a threshold, and it has noise and errors [44]. 3. Due to the large time overhead of the segmentation tasks, we make a trade-off between the dataset size and the model performance. In our experiments, the epoch is set to 400, but our models are not fully converged when the epoch reaches 400. In view of the above points, we consider adding more kinds of cloud images in the follow-up research so that the network will have a stronger generalization ability and persuasiveness. Hierarchical classification is attempted and is firstly divided into four types: the clear sky, high clouds, medium clouds, low clouds, and then subdivided each cloud type, which could improve the cloud recognition effect.

To understand the recognition effect of each cloud type, we analyze the class pixel accuracy (CPA) of UATNet on the test set, that is, the proportion of correctly classified pixels for each class. According to the results, the CPA of the clear sky, cirrostratus and deep convection is high—85.18%, 69.91%, and 69.73%, respectively—while the CPA of cirrus, cumulus, altocumulus and stratocumulus is low—45.59%, 40.65%, 28.66%, and 54.61%, respectively. Cirrostratus and deep convection are high and thick clouds with a well-defined spectral signature, and thus have a high CPA. The main reasons for the low CPA of cirrus, cumulus, altocumulus and stratocumulus clouds are as follows: 1. Samples of each cloud type are not representative enough, which makes it difficult to capture key features for classification. In fact, the causes of formation of cirrus are diverse. Cirrus clouds may be formed by a high-altitude convection, and thus often have the shape of cumulus. Cirrus clouds can also be transformed by the uplift of altocumulus or from the remaining snow virga of altocumulus in the air. In addition, cumulus and stratocumulus

clouds have similar shapes and structures. It is easy to misclassify these associated cloud types without capturing the core features; 2. The types of samples are not distributed homogeneously. The clear sky has the largest sample size, and the average sample size in the test set is 437,488, so the clear sky has the best recognition effect. The sample sizes of cumulus, altocumulus and stratocumulus are relatively small, and the average sample sizes in the test set are 60,060, 76,161, 43,212, respectively, so the corresponding recognition effect is poor. For this problem, we adopt a focal loss to improve the influence of uneven data sample distribution to a certain extent.

We also train the lighter-capacity UATNet-S, which has a smaller size, fewer parameters and a faster speed with little loss above four evaluation indicators. UATNet-S is more suitable for a wide range of deployments and applications.

We visualize the segmentation results of the comparable models in Figure 8, and analyze the experimental results in detail according to Table 8 and Figure 8.



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

(**e**) (**f**) (**g**) (**h**)

(**i**) (**j**) (**k**) (**l**)

(**m**) (**n**) (**o**) (**p**)

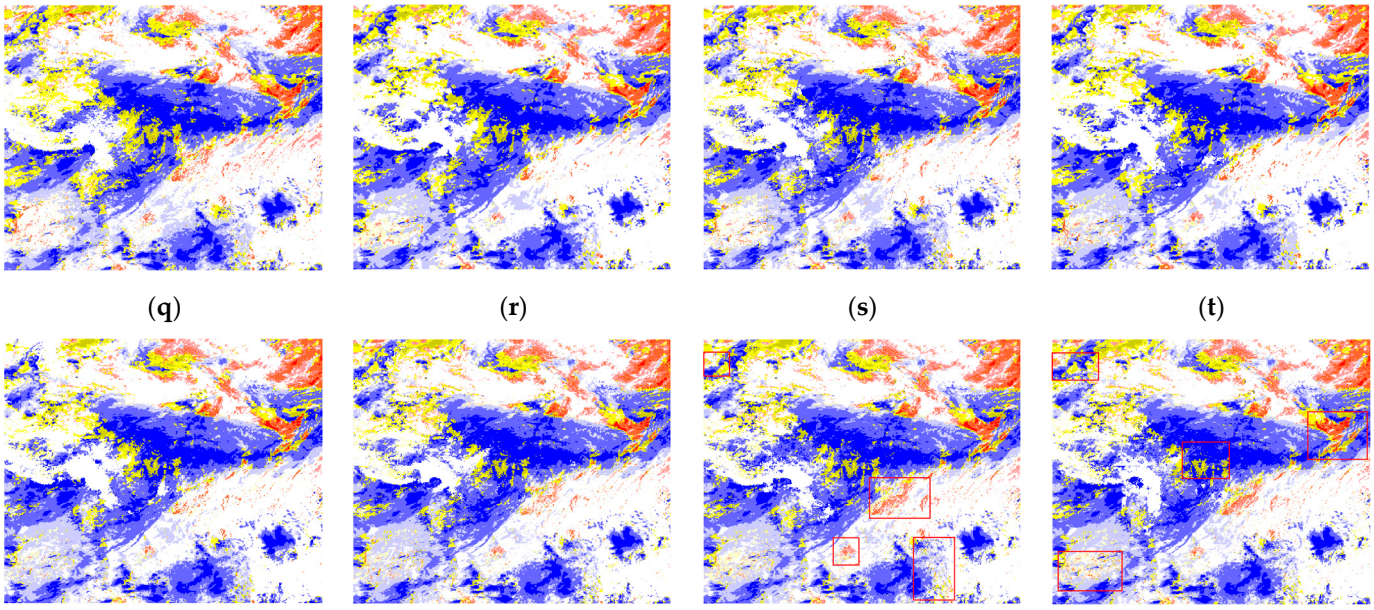


Figure 8. The visual comparison on CRMSCD at 3:30 on 11 July 2020 of (**a**) Original Image, (**b**) Ground Truth, (**c**) U-Net, (**d**) UNet3+, (**e**) SegNet, (**f**) DeepLabv3+, (**g**) GSCNN, (**h**) HRNetV2, (**i**) HRNetV2+OCR, (**j**) UCTransNet, (**k**) UATNet-S and (**l**) UATNet-B, and the visual comparison on CRMSCD at 3:30 on 18 July 2020 of (**m**) Original Image, (**n**) Ground Truth, (**o**) U-Net, (**p**) UNet3+, (**q**) SegNet, (**r**) DeepLabv3+, (**s**) GSCNN, (**t**) HRNetV2, (**u**) HRNetV2+OCR, (**v**) UCTransNet, (**w**) UATNet-S and (**x**) UATNet-B. The color corresponding to the label value is located in the lower right corner of Figure 3b.

We first discuss the fully convolutional neural network models. PSPNet uses the pyramid pooling module to aggregate global contextual information in the top layer of the encoder, but it does not combine the spatial information of the underlying features in the decoder, causing a relatively rough segmentation boundary. SegNet is a lightweight network with fewer parameters based on a symmetrical encoder–decoder structure. It replaces upsampling with unpooling in the decoder, but the segmentation results are not accurate enough without considering the relationship between pixels. DeepLabv3+ uses xception as the backbone network. Maxpooling is replaced by deep separable convolution, which is applied to the atrous spatial pyramid pooling and the decoder. Although DeepLabv3+ combines multi-scale context information, it does not generate a sufficiently refined segmentation boundary. U-Net model is simple, but it makes full use of the underlying features to make up for the loss of the upsampling information. The recognition efficiency of U-Net is better than the more complex deeplabv3+. Although UNet3+ introduces the full-scale skip connection, it does not pay different attention to the information of different scales, resulting in the slightly worse recognition effect than U-Net. HRNetV2 augments the high-resolution representation by aggregating the (upsampling) representations from all the parallel convolutions. On the basis of HRNetV2, HRNetV2+OCR selects the pixels around the target as the context, uses the representation of surrounding pixels to obtain the target representation, and achieves better segmentation results. GSCNN designs two parallel CNN structures to perform conventional extraction and extraction of image boundary-related information. Due to its sufficient capture of boundary information, it obtains the best performance in fully convolutional neural networks.

Next, we analyze transformer-based segmentation methods. SETR and swin transformer require large-scale datasets for pre-training to achieve superior results, while the experimental effects of direct training are poor. UCTransNet introduces the channel transformer into U-Net structure, which makes up for the semantic and resolution gap between low-level and high-level features through more effective feature fusion and multi-scale

channel cross-attention, in order to capture more complex channel dependencies, showing good performance on cloud recognition tasks.

Compared with the above models, our models achieve better results in meteorological satellite cloud recognition tasks. The red boxes in Figures 8 k,l,w,x highlight regions where the two versions of UATNet perform better than other methods, which shows that UATNet generates better segmentation results that are more similar to the ground truth than other methods. Our proposed method not only focuses on significant regions of different cloud types, but also produces clear boundaries. In other words, UATNet is capable of finer segmentation while retaining detailed shape information.

3.4.3. Ablation Studies

To further study the relative contribution of each component of the model, we conduct a series of experiments on CRMSCD by removing and modifying different blocks of UATNet.

As shown in Table 9, we conduct an ablation experiment based on the structure of the encoder. Compared with "Baseline (encoder like Swin-B)", "Baseline + spatial displacement" has performance improvements in PA, MPA, MIoU and FWIoU of 0.70%, 0.99%, 1.42% and 0.90%, respectively. Research results reveal that introducing spatial displacement operation to our encoder can improve segmentation performance to a certain extent.

Table 9. Ablation experiments of spatial displacement on encoder.

Method	CRMSCD			
	PA (%)	MPA (%)	MIoU (%)	FWIoU (%)
Baseline (encoder like Swin-B)	81.63	66.80	53.09	70.06
Baseline + spatial displacement	82.33	67.79	54.51	70.96

As shown in Table 10, we also make an ablation experiment based on the feature fusion of encoder and decoder. "Baseline + ASE" has a significant improvement in all evaluation indicators compared to "Baseline (CCT + skip connection)". ASE is designed to fuse encoder and decoder features, which improves 0.63%, 2.56%, 1.86% and 1.02% in PA, MPA, MIoU and FWIoU, respectively, indicating the effectiveness of ASE block.

Table 10. Ablation experiments of ASE on encoder–decoder connection.

Method	CRMSCD			
	PA (%)	MPA (%)	MIoU (%)	FWIoU (%)
Baseline (CCT + skip connection)	81.70	65.23	52.65	69.94
Baseline + ASE	82.33	67.79	54.51	70.96

The CCT block can effectively fuse multi-scale and multi-channel features, but simply connecting the features of encoder and decoder through the CCT and skip connection does not consider the importance of different feature channels, which may damage the final performance of the model. According to the importance, the ASE block assigns different weights to different channels by enhancing useful features and suppressing useless features. Simultaneously, ASE also introduces nonlinear activation function to improve the generalization ability of the model. The results demonstrate that the ASE block can effectively mine spatial information and semantic information in images and capture non-local semantic dependencies.

4. Discussion

4.1. Findings and Implications

Meteorological satellite cloud images reflect the characteristics and changing processes of all kinds of cloud systems comprehensively, timely and dynamically. They also become an indispensable reference for meteorological and water conservancy departments in the decision-making process. Therefore, we construct a China region meteorological satellite cloud image dataset, CRMSCD, based on FY-4A satellite. It contains nine cloud types and the clear sky (cloudless). CRMSCD expands the satellite cloud image datasets in China, making it easier for researchers to make full use of the meteorological satellite cloud image information with its wide coverage, high timeliness and high resolution for carrying out cloud recognition research tasks.

According to the evaluation of experimental results and a visual analysis, our method has a higher recognition accuracy with smoother and clearer boundaries than existing image segmentation methods. At the same time, we introduce transformer structure in the encoder and encoder–decoder connection, demonstrating the excellent performance of transformer. The transformer model shows an excellent performance in semantic feature extraction, long-distance feature capture, comprehensive feature extraction and other aspects of natural language processing [58–62], which overturns the architecture of traditional neural network models and makes up for the shortcomings of CNN and recursive neural network (RNN). Recently, the use of a transformer to complete visual tasks has become a new research direction, which can significantly improve the scalability and training efficiency of the model.

The FY-4A meteorological satellite cloud image contains 14 channels including visible light and near external light, shortwave infrared, midwave infrared, water vapor and longwave infrared, which carry a lot of rich semantic information. We innovatively introduce a transformer into meteorological satellite cloud recognition tasks, and use its powerful global feature extraction capabilities to enhance the overall perception and macro understanding of images. The transformer can capture the key features of different channels and explore the structure, range, and boundary of different cloud types. Features such as shape, hue, shadow, and texture are distinguished to achieve more accurate and efficient cloud recognition results. Existing visual transformer models, such as SETR and swin transformer, etc., require pre-training on large-scale data to obtain comparable or even better results than CNN. Compared with the above models, our models obtain excellent experimental results without pre-training on CRMSCD. Efficient and accurate cloud recognition with its high timeliness and strong objectivity provides a basis for weather analysis and weather forecasting. In areas lacking surface meteorological observation stations, such as oceans, deserts, and plateaus, meteorological satellite cloud recognition makes up for the lack of conventional detection data and plays an important role in improving the accuracy of weather forecasting, navigation and positioning.

4.2. Other Findings

During the experiments, we also find two skills, which can efficiently improve the effect of cloud recognition in meteorological satellite cloud images. Therefore, we conduct two groups of comparative experiments, and the experimental results are shown in Tables 11 and 12. For demonstration purposes, we only select a fully convolutional neural network (U-Net) and a transformer-based model (Swin-B).

Table 11. Experiment of switchable normalization in convolution layers.

Method	CRMSCD			
	PA (%)	**MPA (%)**	**MIoU (%)**	**FWIoU (%)**
U-Net (batch normalization)	80.50	65.45	50.56	68.85
U-Net (switchable normalization)	81.36	65.35	51.98	69.65

We use switchable normalization to replace the batch normalization in each convolution layer of U-Net. Experimental results show that the use of switchable normalization can significantly improve the effect of U-Net on cloud recognition tasks. The improvement is still applicable in other fully convolutional neural networks.

Table 12. Experiment of skip connection on transformer-based model.

Method	CRMSCD			
	PA (%)	MPA (%)	MIoU (%)	FWIoU (%)
Swin-B	71.67	46.46	34.52	57.76
Swin-B + skip connection	79.29	61.23	47.72	67.08

According to the experimental results in Table 12, the introduction of encoder–decoder connection in Swin-B greatly improves the effect of cloud recognition.

4.3. Limitations

Although UATNet has shown superior performance in experiments, it still has two limitations.

First, we only trained and tested UATNet on CRMSCD without verifying it for other satellite cloud image datasets, which merely demonstrates the excellent cloud segmentation effect of UATNet on CRMSCD. The features captured on a single dataset may be relatively limited, and the versatility and universality of UATNet were not well verified in cloud recognition.

Second, we calculated the computational complexity of all the above models, including the total number of floating-point operations (FLOP) and the parameters of the model. As shown in Table 13, the number of calculations and parameters of UATNet were higher than most mainstream image segmentation methods compared in the experiment. In UATNet, encoder and the feature fusion of encoder and decoder are completed by the transformer, and the encoder has a large number of transformer layers, which makes the total number of parameters large. Larger calculation scale and total number of parameters are a great challenge to computing resources and training duration.

Table 13. Computational complexity on CRMSCD.

Method	Input Size	Parameters (M)	Flop (GFLOPS)
U-Net	14 × 512 × 512	29.0	195.7
UNet3+		21.1	800.4
PSPNet		65.7	78.9
SegNet		29.5	163.4
DeepLabv3+		54.5	82.9
GSCNN		28.3	569.7
HRNetV2		65.9	223.8
HRNetV2+OCR		70.4	1319.4
SETR		31.5	27.7
Swin-T		40.8	46.0
Swin-S		62.1	67.8
Swin-B		102.3	103.0
Swin-L		215.2	203.1
UCTransNet		66.2	173.5
UATNet-S		97.0	222.1
UATNet-B		137.6	388.5

4.4. Future Work

We analyzed the advantages and disadvantages of UATNet in detail. In this section, we discuss four possible research directions in the future.

In future experiments, we will expand the size of the dataset and train our model on other satellite cloud image datasets. It can not only enhance the ability of UATNet to learn more abundant and more critical features, but also verify the generalization ability and universal applicability of UATNet in cloud recognition.

UATNet has a relatively large number of calculations and parameters, resulting in a large consumption of computing resources and a long time for model convergence, which is not conducive to model deployment. How to reduce network parameters without losing model accuracy and realize the trade-off between calculation speed and calculation accuracy is a meaningful research direction. We will try to integrate the ideas of SqueezeNet [63], MobileNet [64] and other methods into our model for more refined model compression. Transfer learning [65,66] and knowledge distillation [67,68] can also be introduced into our cloud recognition work.

The convolution operation in CNN is ideal for extracting local features, but it has some limitations in capturing global feature representation. The UATNet encoder uses transformer blocks, whose self-attention mechanism and MLP block can reflect complex spatial transformation and long-distance feature dependence. The transformer focuses on global features such as contour representation and shape description, while ignoring local features. In cloud recognition, the fusion of transformer and CNN to improve the local sensitivity and global awareness of the model plays an important role in capturing rich and complex cloud feature information.

Finally, we will compare with physics-based algorithms and explore the robustness to noise in radiance data and to shifts in co-registration between channels.

5. Conclusions

In this paper, we constructed the China region meteorological satellite cloud image dataset, CRMSCD, and proposed a meteorological satellite cloud recognition method, UATNet, to obtain pixel-level cloud classification results. We innovatively introduced a transformer into cloud recognition tasks and solved the problem of limited receptive field of CNN-based cloud recognition models, which can effectively capture the global information of multi-scale and multi-channel satellite cloud images, while ensuring computational efficiency. Specifically, UATNet constructs a hierarchical transformer based on swin transformer block and patch merging and introduces spatial displacement to build long-distance cross-window connections. Simultaneously, CCT and ASE blocks are used jointly to adaptively fuse the features of the encoder and decoder to bridge the semantic gap of feature mapping. UATNet integrates the two transformer structures together to perform attention operations along the patch axis and channel axis, respectively, which not only effectively integrates the spatial information and multi-channel information of clouds, but also smooths the gradient and extract more targeted cloud features. Extensive experiments including cloud recognition and ablation studies were conducted, which demonstrate the effectiveness of the proposed model.

Author Contributions: Conceptualization, Z.W. and J.Z.; Methodology, Z.W.; Software, Z.W. and X.W.; Validation, Z.W. and Z.L.; Formal Analysis, Z.W.; Investigation, Z.W. and Z.L.; Resources, X.W.; Data Curation, Z.W., J.Z. and Q.L.; Writing—Original Draft Preparation, Z.W. and R.Z.; Writing—Review and Editing, Z.W., J.Z., R.Z. and X.W.; Visualization, Z.W. and R.Z.; Funding Acquisition, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA19020203 and XDA19020103), and also supported in part by Key Research Program of Frontier Sciences, CAS, and grant number ZDBS-LY-DQC016 and in part by the National Natural Science Foundation of China (NSFC) under grant 61836013.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Science Data Bank at https://datapid.cn/31253.11.sciencedb.01337 (accessed on 16 December 2021), https://www.doi.org/10.11922/sciencedb.01337 (accessed on 16 December 2021), reference number 10.11922/sciencedb.01337.

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

References

1. Zhang, Y.; Rossow, W.B.; Lacis, A.A.; Oinas, V.; Mishchenko, M.I. Calculation of radiative fluxes from the surface to top of atmosphere based on isccp and other global data sets: Refinements of the radiative transfer model and the input data. *J. Geophys. Res. Atmos.* **2004**, *109*, 109.
2. Carslaw, K.; Harrison, R.; Kirkby, J. Cosmic rays, clouds, and climate. *Science* **2002**, *298*, 1732–1737.
3. Stephens, G.L. Cloud feedbacks in the climate system: A critical review. *J. Clim.* **2005**, *18*, 237–273.
4. Cui, F.; Ju, R.R.; Ding, Y.Y.; Ding, H.; Cheng, X. Prediction of Regional Global Horizontal Irradiance Combining Ground-Based Cloud Observation and Numerical Weather Prediction. *Adv. Mater. Res.* **2014**, *1073-1076*, 388–394.
5. Bessho, K.; Date, K.; Hayashi, M.; Ikeda, A.; Imai, T.; Inoue, H.; Kumagai, Y.; Miyakawa, T.; Murata, H.; Ohno, T. An introduction to himawari-8/9—Japan's new-generation geostationary meteorological satellites. *J. Meteorol. Soc. Japan. Ser. II* **2016**, *94*, 151–183.
6. Schmit, T.J.; Griffith, P.; Gunshor, M.M.; Daniels, J.M.; Goodman, S.J.; Lebair, W.J. A closer look at the abi on the goes-r series. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 681–698.
7. Yang, J.; Zhang, Z.; Wei, C.; Lu, F.; Guo, Q. Introducing the new generation of chinese geostationary weather satellites, fengyun-4. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 1637–1658.
8. Zhuo, W.; Cao, Z.; Xiao, Y.; Technology, O. Cloud classification of ground-based images using texture–structure features. *J. Atmos. Ocean. Technol.* **2014**, *31*, 79–92.
9. Zhang, Z.; Li, D.; Liu, S.; Xiao, B.; Cao, X. Multi-view ground-based cloud recognition by transferring deep visual information. *Appl. Sci.* **2018**, *8*, 748.
10. Fang, C.; Jia, K.; Liu, P.; Zhang, L. Research on cloud recognition technology based on transfer learning. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 791–796.
11. Liu, S.; Li, M.; Zhang, Z.; Xiao, B.; Durrani, T.S. Multi-evidence and multi-modal fusion network for ground-based cloud recognition. *Remote Sens.* **2020**, *12*, 464.
12. Liu, Y.; Xia, J.; Shi, C.-X.; Hong, Y. An improved cloud classification algorithm for China's fy-2c multi-channel images using artificial neural network. *Sensors* **2009**, *9*, 5558–5579.
13. Bai, T.; Li, D.; Sun, K.; Chen, Y.; Li, W. Cloud detection for high-resolution satellite imagery using machine learning and multi-feature fusion. *Remote Sens.* **2016**, *8*, 715.
14. Cai, K.; Wang, H. Cloud classification of satellite image based on convolutional neural networks. In Proceedings of the 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 24–26 November 2017; pp. 874–877.
15. Rossow, W.B.; Garder, L.C. Cloud detection using satellite measurements of infrared and visible radiances for isccp. *J. Clim.* **1993**, *6*, 2341–2369.
16. Irish, R.R.; Barker, J.L.; Goward, S.N.; Arvidson, T. Characterization of the landsat-7 etm+ automated cloud-cover assessment (acca) algorithm. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 1179–1188.
17. Zhu, Z.; Woodcock, C.E. Object-based cloud and cloud shadow detection in landsat imagery. *Remote Sens. Environ.* **2012**, *118*, 83–94.
18. Azimi-Sadjadi, M.; Zekavat, S. Cloud classification using support vector machines. In Proceedings of the IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No. 00CH37120), Honolulu, HI, USA, 24–28 July 2000; Pp 669-671.
19. Christodoulou, C.I.; Michaelides, S.C.; Pattichis, C.S. Multifeature texture analysis for the classification of clouds in satellite imagery. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 2662–2668.
20. Li, J.; Menzel, W.P.; Yang, Z.; Frey, R.A.; Ackerman, S.A. High-spatial-resolution surface and cloud-type classification from modis multispectral band measurements. *J. Appl. Meteorol.* **2003**, *42*, 204–226.
21. Merchant, C.; Harris, A.; Maturi, E.; MacCallum, S. Probabilistic physically based cloud screening of satellite infrared imagery for operational sea surface temperature retrieval. *Q.J.R. Meteorol. Soc. A J. Atmos. Sci. Appl. Meteorol. Phys. Oceanogr.* **2005**, *131*, 2735–2755.
22. Amato, U.; Antoniadis, A.; Cuomo, V.; Cutillo, L.; Franzese, M.; Murino, L.; Serio, C. Statistical cloud detection from seviri multispectral images. *Remote Sens. Environ.* **2008**, *112*, 750–766.
23. Le Hégarat-Mascle, S.; André, C. Use of markov random fields for automatic cloud/shadow detection on high resolution optical images. *ISPRS J. Photogramm. Remote Sens.* **2009**, *64*, 351–366.

24. Li, P.; Dong, L.; Xiao, H.; Xu, M. A cloud image detection method based on svm vector machine. *Neurocomputing* **2015**, *169*, 34–42.
25. Yuan, Y.; Hu, X. Bag-of-words and object-based classification for cloud extraction from satellite imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 4197–4205.
26. Taravat, A.; Del Frate, F.; Cornaro, C.; Vergari, S. Neural networks and support vector machine algorithms for automatic cloud classification of whole-sky ground-based images. *Geosci. Remote Sens. Lett.* **2014**, *12*, 666–670.
27. 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.
28. Phung, V.H.; Rhee, E.J. A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Appl. Sci.* **2019**, *9*, 4500.
29. Zhang, J.; Liu, P.; Zhang, F.; Iwabuchi, H.; e Ayres, A.A.d.H.; De Albuquerque, V.H.C. Ensemble meteorological cloud classification meets internet of dependable and controllable things. *IEEE Internet Things J.* **2020**, *8*, 3323–3330.
30. Manzo, M.; Pellino, S. Voting in transfer learning system for ground-based cloud classification. *arXiv* **2021**, arXiv:2103.04667.
31. Li, G. A novel computer-aided cloud type classification method based on convolutional neural network with squeeze-and-excitation. *J. Phys. Conf. Ser.* **2021**, *1802*.
32. Liu, S.; Li, M.; Zhang, Z.; Xiao, B.; Cao, X. Multimodal ground-based cloud classification using joint fusion convolutional neural network. *Remote Sens.* **2018**, *10*, 822.
33. Zhang, J.; Liu, P.; Zhang, F.; Song, Q. Cloudnet: Ground-based cloud classification with deep convolutional neural network. *Geophys. Res. Lett.* **2018**, *45*, 8665–8672.
34. Lu, J.; Wang, Y.; Zhu, Y.; Ji, X.; Xing, T.; Li, W.; Zomaya, A.Y. P_segnet and np_segnet: New neural network architectures for cloud recognition of remote sensing images. *IEEE Access* **2019**, *7*, 87323–87333.
35. Zhang, L.; Jia, K.; Liu, P.; Fang, C. Cloud recognition based on lightweight neural network. In Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, 7–10 December 2020; pp. 1033–1042.
36. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4905–4913.
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems,* Long Beach, CA, USA, 4-8 December 2017; pp. 5998–6008.
38. Wang, H.; Zhu, Y.; Green, B.; Adam, H.; Yuille, A.; Chen, L.-C. *Axial-Deeplab: Stand-Alone Axial-Attention for Panoptic Segmentation. European Conference on Computer Vision*; Springer: New York, NY, USA, 2020; pp. 108–126.
39. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 21–24 June 2021; pp. 6881–6890.
40. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
41. Min, M.; Wu, C.; Li, C.; Liu, H.; Xu, N.; Wu, X.; Chen, L.; Wang, F.; Sun, F.; Qin, D. Developing the science product algorithm testbed for chinese next-generation geostationary meteorological satellites: Fengyun-4 series. *J. Meteorol. Res.* **2017**, *31*, 708–719.
42. Yaohai, D.; Xiaojie, C.; Qiang, C.; Wang, L.; Junfeng, S.; Lamei, C.; Feng, J. Fy-4 meteorological satellite. *China Aerosp.* **2019**, *18*, 31–39.
43. Wang, X.; Min, M.; Wang, F.; Guo, J.; Li, B.; Tang, S.; Sensing, R. Intercomparisons of cloud mask products among fengyun-4a, himawari-8, and modis. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8827–8839.
44. Suzue, H.; Imai, T.; Mouri, K. High-resolution cloud analysis information derived from himawari-8 data. *Meteorol. Satell. Cent. Tech. Note* **2016**, *61*, 43–51.
45. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv* **2021**, arXiv:2103.14030.
46. Huang, Z.; Ben, Y.; Luo, G.; Cheng, P.; Yu, G.; Fu, B. Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv* **2021**, arXiv:2106.03650 .
47. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.R. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. *arXiv* **2021**, arXiv:2109.04335.
48. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp 7132–7141.
49. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
50. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: New York, NY, USA, 2015; pp. 234–241.
51. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.

52. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881-2890.
53. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495.
54. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
55. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 5229–5238.
56. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
57. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. Computer Vision–ECCV 2020: In Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020; Part VI 16; Springer: New York, NY, USA, 2020; pp. 173–190.
58. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
59. Liu, Y.; Lapata, M. Text summarization with pretrained encoders. *arXiv* **2019**, arXiv:1908.08345.
60. Shao, T.; Guo, Y.; Chen, H.; Hao, Z. Transformer-based neural network for answer selection in question answering. *IEEE Access* **2019**, *7*, 26146–26156.
61. Wang, Q.; Li, B.; Xiao, T.; Zhu, J.; Li, C.; Wong, D.F.; Chao, L.S. Learning deep transformer models for machine translation. *arXiv* **2019**, arXiv:1906.01787.
62. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The long-document transformer. *arXiv* **2020**, arXiv:2004.05150.
63. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. Squeezenet: Alexnet-level accuracy with 50× fewer parameters and <0.5 mb model size. *arXiv* **2016**, arXiv:1602.07360.
64. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
65. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359.
66. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 1–40.
67. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
68. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge distillation: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819.