remote sensing

MDPI

Article

GF-1/6 Satellite Pixel-by-Pixel Quality Tagging Algorithm

Xin Fan [1,2], Hao Chang [1,2], Lianzhi Huo [1] and Changmiao Hu [1,*]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: hucm@aircas.ac.cn

Abstract: The Landsat and Sentinel series satellites contain their own quality tagging data products, marking the source image pixel by pixel with several specific semantic categories. These data products generally contain categories such as cloud, cloud shadow, land, water body, and snow. Due to the lack of mid-wave and thermal infrared bands, the accuracy of traditional cloud detection algorithm is unstable when facing Chinese Gaofen-1/6 (GF-1/6) data. Moreover, it is challenging to distinguish clouds from snow. In order to produce GF-1/6 satellite pixel-by-pixel quality tagging data products, this paper builds a training sample set of more than 100,000 image pairs, primarily using Sentinel-2 satellite data. Then, we adopt the Swin Transformer model with a self-attention mechanism for GF-1/6 satellite image quality tagging. Experiments show that the model's overall accuracy reaches the level of Fmask v4.6 with more than 10,000 training samples, and the model can distinguish between cloud and snow correctly. Our GF-1/6 quality tagging algorithm can meet the requirements of the "Analysis Ready Data (ARD) Technology Research for Domestic Satellite" project.

Keywords: cloud detection; Swin Transformer; GF-1; Sentinel-2; Fmask; Analysis Ready Data

1. Introduction

With the advancement of sensor hardware, the quality of remote sensing image standard data products has been improving. Meanwhile, with the improvement in quantification accuracy, data products with pixel-by-pixel marking quality and characteristics are gradually emerging, such as MODIS series satellites, Landsat series satellites, Sentinel-2, and other international satellites. These data products contain categories for clouds, shadows under clouds, land, water bodies, and snow. This helps users to filter interfering pixels when studying scenarios such as surface changes and classifications.

The core part of the remote sensing image quality tagging algorithm is cloud and under-cloud shadow detection, including detecting water bodies, snow, and other categories. Due to the greater number of wavebands of international satellite multispectral sensors, their band settings generally cover visible, mid-infrared, and thermal infrared wavebands. The spectral threshold method has become the mainstream algorithm in foreign countries, and is represented by the Automatic Cloud Cover Assessment (ACCA) algorithm [1] and the Fmask (Function of Mask) algorithm [2]. The ACCA algorithm establishes threshold rules for detection based on the physical properties of clouds and cloud shadows. It uses five bands and 26 band test rules, which are generally operable without significant systematic errors. The Fmask algorithm utilizes all Landsat bands, including thermal infrared bands, with higher accuracy than the ACCA algorithm [3]. Fmask is now one of the standard product algorithms for quality tagging Landsat series and Sentinel-2 satellite data [4]. The Fmask algorithm is also used to evaluate and improve the consistency of Landsat series satellite time series ARD data, with an overall accuracy of 96.4% for cloud detection [5]. Recent research on cloud detection from 2004 to 2018 was analyzed by Mahajan in their review paper [6]. Using a variety of machine learning and traditional algorithms, the

Citation: Fan, X.; Chang, H.; Huo, L.; Hu, C. GF-1/6 Satellite Pixel-by-Pixel Quality Tagging Algorithm. *Remote Sens.* **2023**, *15*, 1955. https://doi.org/10.3390/rs15071955

Academic Editor: Eufemia Tarantino

Received: 13 March 2023
Revised: 3 April 2023
Accepted: 5 April 2023
Published: 6 April 2023

researchers have investigated various cloud detection methods such as Cloud/No Cloud, Snow/Cloud, and Thin Cloud/Thick Cloud.

With the rapid development of deep learning, it has been reported that remote sensing applications of pixel-wise classification are gradually being incorporated into deep learning-based segmentation methods. Typical neural image classifiers emerge from the natural image classification tasks, such as VGG [7], ResNet [8,9], and DenseNet [10], and their convolutional modules have been applied to object detection and semantic segmentation. Neural semantic segmentation was initiated by the fully convolutional network (FCN) [11], which replaced fully connected layers with convolutional layers to accommodate arbitrary size segmentation. By using shortcut connections to combine multi-level feature maps with the same dimensions, UNet [12] popularized intermediate feature fusion, which made it possible to reuse features in an image segmentation task. Learning stronger feature representation was made possible by HRNet [13] through aggregating features from all parallel convolutions rather than just the high resolution convolutions. In order to improve both the efficiency and robustness, the DeepLab series [14–17] made use of novel backbones, atrous convolution, CRF post-processing, depth-wise separatable convolution, and atrous spatial pyramid pooling module. Recently, transformer-based segmentation models have emerged due to their long-range feature extraction [18–21].

Convolutional neural networks have become the mainstream for cloud detection and multiclass segmentation. As early as 2014, Hughes used neural networks to identify clouds and cloud shadows for automatic detection [22]. Chai used manually labeled cloud and cloud shadow mask data to train CNNs, segmenting Landsat images into four categories: clouds, thin clouds, cloud shadows, and no clouds [23]. Jeppesen proposed a novel deep learning model called Remote Sensing Network (RS-Net) for cloud detection in optical satellite imagery, based on the UNet architecture [24]. The Landsat 8 Biome and SPARCS datasets are used to train and evaluate the RS-Net model, and it gives a state of the art performance, particularly when used in biomes where the scenery is hard to tell apart, such as clouds over icy and snowy areas. Grabowski utilized the self-configuring nnU-Net to detect clouds in satellite images [25]. The nnU-Net is a self-reconfigurable framework able to perform meta-learning of a segmentation network over various datasets. Experiments on multispectral images from Landsat-8 and Sentinel-2 showed that nnU-Net perform the best cloud segmentation ever without any manual design. Jiao proposed a series of end-to-end Refined UNet cloud and cloud shadow detection models [26–29]. They introduced the edge-sensitive computation of conditional random fields in the neural network to achieve accurate edge segmentation of cloud and cloud shadow for Landsat 8 OLI data.

Focusing on achieving the UN 2030 Sustainable Development Goals (SDGs) and international sharing of Chinese GF-1/6 satellite data, the 2020 National Key R&D Program International Cooperation Project "Research on Analysis Ready Data (ARD) Technology for Domestic Satellites" has been launched. The project includes Chinese GF-1/6 satellite quality tagging technology as one of its four key technologies. Driven by the project, this research utilizes deep learning methods to study the Chinese GF-1/6 satellite quality tagging algorithm, aiming to produce standard quality tagging data products that contain five categories: clouds, cloud shadows, water bodies, land, and snow. The accuracy of standard data products should be close to that of international Landsat series and Sentinel-2 satellites.

In this paper, we utilize the Swin Transformer model, from the best paper from ICCV 2021, as the backbone network for our GF-1/6 quality tagging algorithm [30]. Because the currently available open-source cloud detection dataset is small and does not contain all categories that the ARD project requires, we designed a process of using Landsat-8/Sentinel-2 data and combining the Fmask v4.6 algorithm to produce the training sample dataset. Finally, a training sample set of more than 100,000 image pairs was created. Then, the well-trained Swin Transformer model was transferred to the quality tagging task of GF-1/6 satellites. Meanwhile, we also improved the model performance at distinguishing clouds from snow. Experiments show that the Swin Transformer's Large model (Swin-L)

achieves the quality tagging accuracy level of international Landsat series and Sentinel-2 satellites. The engineering value of this paper is to produce a complete flow and data product specification for a GF-1/6 satellite quality tagging algorithm. It employed grid offset processing to solve the problem of stitching seams in a large-size remote sensing image chunking process. Moreover, using DEM and GSWO auxiliary data to correct the quality tagging mask further improved the stability of the final standard data products.

2. Background

Chinese satellite multispectral images are clearly different from foreign satellites. Most of them, such as GF-1/6, only contain four bands from visible to near-infrared, lacking mid-wave and thermal infrared bands. This makes distinguishing clouds from snow, desert, and other high-brightness surfaces challenging by relying purely on the spectral threshold method. Chinese researchers have conducted detailed research on quality tagging algorithms for domestic satellite multispectral images.

The multi-feature combined (MFC) cloud and cloud shadow automatic detection algorithm uses guided filtering to improve the detection accuracy of cloud and cloud shadow edge regions from GF-1 WFV data [31]. Wang Mi from Wuhan University adopted the SLIC super-pixel segmentation algorithm for Chinese satellite in-orbit cloud detection to improve the accuracy of thick cloud edges [32]. The fractal dimension and mean gradient are applied to cloud detection in the ZY-3 satellite, and the effect of texture features on improving the detection accuracy is verified in the scene data with coexisting clouds and snow [33]. These studies provide good ideas for producing the quality tagging data products of Chinese satellite multispectral images. However, the challenge of distinguishing clouds from snow still exists. Although introducing image processing and auxiliary data can improve the accuracy of cloud detection based on the spectral threshold method, it fails to change the problem that clouds and snow are indistinguishable in the visible and near-infrared bands.

However, deep learning methods avoid the problem. Therefore, many researchers are now focusing on using deep learning algorithms to obtain models that can effectively distinguish between cloud, snow, and other target categories by training them on large-size sample datasets. For example, the Cloud-AttU [34] cloud detection method is based on the UNet network. Compared with international satellite quality tagging data products, the accuracy of Chinese satellite image cloud detection algorithms could be higher and more stable. In general, there are two ways to further improve the accuracy of deep learning models. On the one hand, we can select the segmentation model wisely and optimize it. On the other hand, enhancing the quantity and quality of training samples is also helpful.

The CNN-based segmentation method is still the mainstream backbone used by cloud detection algorithms use, which has advantages in multi-scale target semantic segmentation. However, the multi-scale characteristics of remote sensing images with a fixed resolution of the specific data source's cloud and cloud shadow are not apparent. The key to deciding the quality tag of a pixel at a particular location relies on the relationship between that pixel and the surrounding pixels, which is more important than the feature description of that pixel at different resolutions or scales. This is our major consideration: using a Transformer-based segmentation model with a self-attention mechanism for GF-1/6 quality tagging tasks.

Specifically, we select the Swin Transformer model, which contains sliding window operations and has a hierarchical design. One of the sliding-window operations includes a non-overlapping local window and an overlapping cross-window. It restricts the attention computation to a single window, which both introduces the local nature of the CNN by convolution operations and decreases the computation cost. The Swin Transformer performs well on all significant image tasks, and its mIoU on the natural image semantic segmentation dataset ADE20K reaches the state of the art performance. Compared with the CNN model, the Transformer-based model performs better in image tasks after pre-training on large dataset [35]. Large-size datasets for pre-training can break the limitation of the

Transformer's lack of inductive bias. For example, experiments on large datasets such as ImageNet-21K and Google JFT-200M have verified that the Vision Transformer model is better than the ResNet model [36].

In order to train the Swin Transformer for our GF-1/6 quality tagging task, we produced more than 100,000 sample images, coupled with a few manual accuracy checks and corrections. The training samples were produced according to the similarity of the RGB Chinese GF-1/6 images and Landsat-8/Sentinel-2 images. Each training sample is a Byte-type RGB image constructed in three visible bands, with a spatial resolution of 20 m, and pixel size 512 × 512. Additionally, the corresponding labels contained six categories of quality marker (cloud, cloud shadow, water, land, snow, and fill value). Therefore, the well-trained Swin Transformer model can be used directly for quality tagging the Chinese GF-1/6 images. Based on the model parameters obtained from the training, a quality tagging algorithm flow for the Chinese GF-1/6 satellite image is developed in this paper and applied to produce standard data products for the ARD project.

3. Methodology

3.1. Customized Dataset Preparation

The GF-1/6 image quality tagging algorithm model is best trained using the corresponding satellite's sample data. However, there are no quality tagging data products for GF-1/6 images. Additionally, the quality tagging accuracy using the existing algorithm is unstable. It is tough to distinguish between clouds and snow, making the manual correction work too difficult, and producing tens of thousands of data samples is not easy. The international Landsat-8 and Sentinel-2 satellite data contain quality tagging products, and many researchers have tested and proven their accuracy. The idea of constructing training samples in this paper is to use international data as the primary data source to produce a large sample dataset serving the quality tagging of Chinese GF-1/6 satellite images. In order to produce enough sample data for completing the quality tagging of GF-1/6 images, this paper utilized Landsat-8/Sentinel-2 satellite data to produce training samples based on the Fmask v4.6 algorithm. This section contains three parts: 1) sample image, which determines the selected band and resolution of the sample image by analyzing the difference between Landsat-8/Sentinel-2 data and GF-1/6 WFV data; 2) sample labeling, introducing the production of quality tagging data products for Landsat-8/Sentinel-2 data using the Fmask v4.6 algorithm, giving the class definition and producing a method of sample labeling; and 3) producing training samples, which introduces the specific technical process of producing training samples.

3.1.1. Sample Image

Producing sample images is mainly based on the similarity of Landsat series, Sentinel-2, and GF-1/6 satellite data regarding band setting and spatial resolution. Landsat-8 and Sentinel-2 are more abundant than GF-1/6 WFV data in band settings. However, they are very similar in visible and near-infrared bands as shown in Table 1. The near-infrared band has an extensive spectral range, and imaging is more affected by the difference in the spectral response function curve than the visible band. Furthermore, the near-infrared band is less affected by atmospheric scattering, which is not conducive to identifying thin clouds. Therefore, the sample image is selected for three visible light bands to form an RGB format image. The spatial resolution is 10 m for the Sentinel-2 satellite, 30 m for the Landsat series satellite, and 16 m for the GF-1/6 satellite. All three data sources are unified to 20 m when producing the sampled images.

Table 1. Band setting and bandwidth comparison of Landsat-8, Sentinel-2, and GF-1/6 satellites.

Satellite	Landsat-8	Sentinel-2A	Sentinel-2B	GF-1	GF-6
Coastal Blue	0.433–0.453	0.432–0.453	0.432–0.453	—	0.40–0.45
Blue	**0.450–0.515**	**0.459–0.525**	**0.459–0.525**	**0.45–0.52**	**0.45–0.52**
Green	**0.525–0.600**	**0.542–0.578**	**0.541–0.577**	**0.52–0.59**	**0.52–0.59**
Yellow	—	—	—	—	0.59–0.63
Red	**0.630–0.680**	**0.649–0.680**	**0.650–0.681**	**0.63–0.69**	**0.63–0.69**
Red Edge 1	—	0.697–0.712	0.696–0.712	—	0.69–0.73
Red Edge 2		0.733–0.748	0.732–0.747		
Red Edge 3		0.773–0.793	0.770–0.790		0.73–0.77
NIR	0.845–0.885	0.780–0.886	0.780–0.886	0.77–0.89	0.77–0.89
Narrow NIR		0.854–0.875	0.853–0.875		
Water vapor	—	0.935–0.955	0.933–0.954	—	—
Cirrus	1.360–1.390	1.358–1.389	1.362–1.392	—	—
SWIR 1	1.560–1.660	1.568–1.659	1.563–1.657	—	—
SWIR 2	2.100–2.300	2.115–2.290	2.093–2.278	—	—
TIRS 1	10.60–11.19	—	—	—	—
TIRS 2	11.50–12.51	—	—	—	—

The bandwidth unit is μm.

Sample images were obtained using Landsat-8/Sentinel-2 Top of Atmosphere Reflectance (TOA) data with a fixed pixel size of 512 × 512. Since Landsat-8/Sentinel-2 TOA data are stored in the UInt16 format (their pixel values 0–65,535 are converted to TOA by dividing 10,000), the value needs to be compressed from 16-bit to 8-bit storage. In this paper, we use a fixed mapping method to compress the data for storage, and the specific mapping method is shown in Table 2. We intend to reduce the difference between various satellites by compressing the data. Meanwhile, the morphological difference, or the adjacency of pixels, is more important in the Transformer model than the numerical difference. The Byte-type sample is also more general, which is also conducive to the subsequent publication of the dataset and makes it convenient for more people to use. The boundary value is set according to experience. For example, 6000 represents the TOA of 0.6, which in reality reaches the lower boundary of bright surfaces such as clouds. The experiment proves that this step is effective and important.

Table 2. The fixed mapping transformation setting of sample images UInt16 to Byte format.

UInt16	Byte	Memo
0	0	Fill value
1–6000	1–250	Step length 24
6001–10,000	251–254	Step length 1000
>10,000	255	Saturation value

Compressed storage using the fixed mapping transformation described above is the core step of producing training samples. There are two advantages of such processing: First, it ensures the overall consistency of the sample image. The different image conversion storage processes are not related to the statistical histogram because they are converted according to a fixed mapping, and the converted data pixel values are consistent and reflect the relative magnitude of surface reflectance. Second, the complexity of data values is compressed, and the complex surface reflectance values are uniformly compressed from 0–10,000 to 0–255. This can help to reduce the complexity of a clear surface while emphasizing the contrast between clouds and snow. Dynamic stretching based on histogram statistics is commonly used in image transformation. It has become the default processing method in many algorithm toolkits, but dynamic stretching does not suit the sample image processing in this paper. A comparison of the fixed mapping and dynamic stretching transformation is shown in Figure 1. In the figure, two Sentinel-2 images of the same region are selected, with imaging times of 23 August 2020 and 12 September 2020, containing two extreme cases of large-scale cloud coverage and clear sky. It can be

seen that the fixed mapping transformation is not affected by the scale of cloud coverage. Fixed mapping has a consistent overall radiometric brightness, a low brightness for clear surfaces, a high brightness for cloud areas, and an approximate histogram distribution for the different surface coverage. While the dynamic stretching transformation is affected by the histogram difference, the radiation difference is more significant, the image surface with a low cloud amount has higher brightness, and the difference in the histogram distribution after stretching is significant.

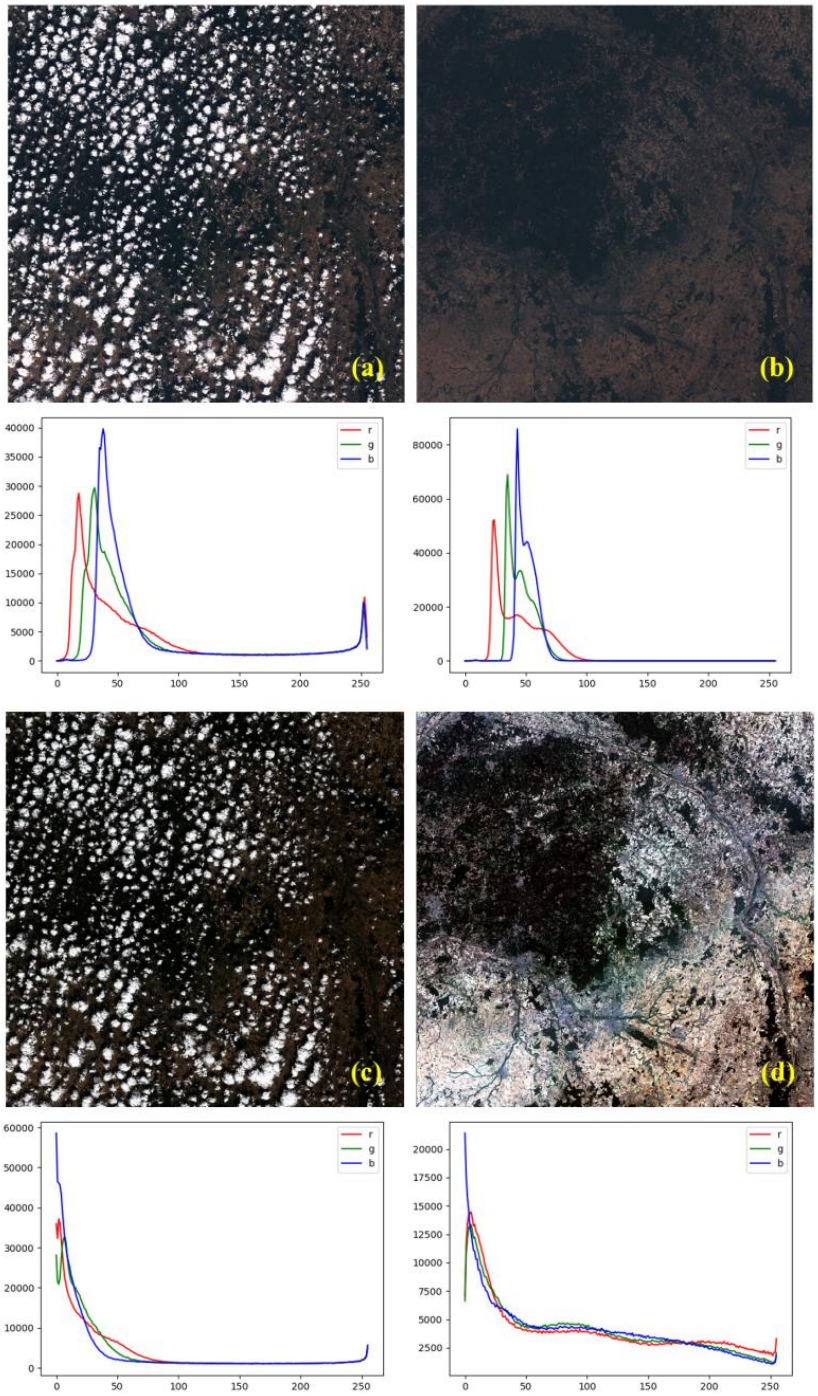


Figure 1. Comparison of fixed mapping transformation (**a**,**b**) and dynamic stretching transformation (**c**,**d**).

The sample images of Landsat 8, Sentinel-2, and GF-1/6 produced using our fixed mapping transformation have high similarity in spectral features and are difficult to distinguish visually. A typical example is shown in Figure 2. The Sentinel-2 and GF-1 images

were imaged in mid-June 2021. After conversion to Byte-type RGB images, the two images have a high similarity, and the difference is mainly caused by atmospheric aerosol optical thickness. Based on this similarity of sample images, a model trained on Landsat 8/Sentinel-2 data can be used to produce the GF-1/6 quality tagging standard data products directly.

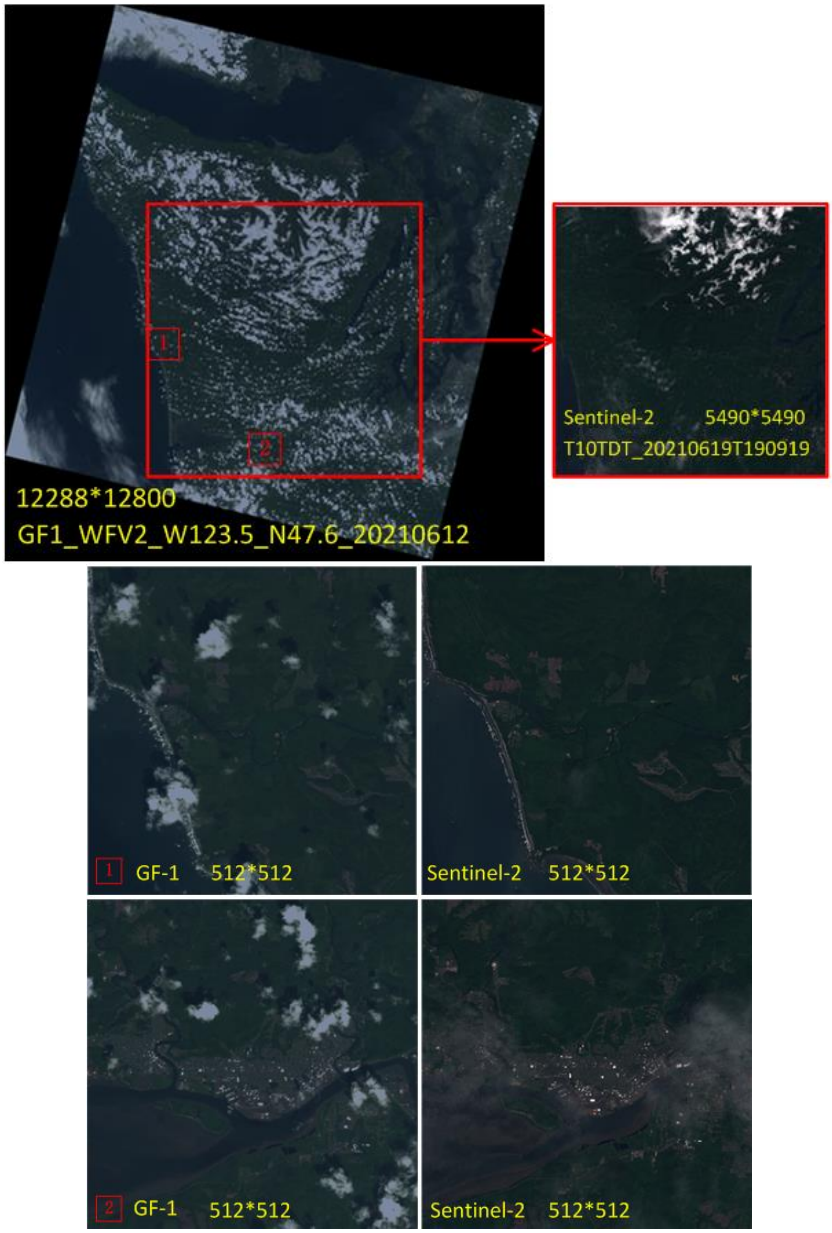


Figure 2. Comparison of Sentinel-2 and GF-1 sample image.

3.1.2. Sample Label

The sample label is a single-band image corresponding to the sample image, marked pixel by pixel with quality tagging categories. The Fmask (Function of mask) algorithm is used for automated clouds, cloud shadows, snow, and water masking for Landsat 4–9 and Sentinel-2 images. Fmask became available after version 1.6, and USGS integrated the C version of Fmask (CFmask) into quality-marked products for the Landsat series satellites to obtain higher accuracy than the ACCA algorithm. Fmask version 3.2 further improved the detection accuracy and started to support Sentinel-2 data [37]. Mountainous Fmask (MFmask) improves the detection accuracy of mountainous areas' clouds and cloud shad-

ows using a Digital Elevation Model (DEM) [38]. Fmask version 4.0 further optimized the algorithm and improved the detection accuracy using global auxiliary data, including DEMs and water layers. In 2022, the updated version of the Fmask algorithm in February was Fmask v4.6 (https://github.com/GERSL/Fmask, accessed on 1 November 2022).

Landsat-8 and Sentinel-2 standard data products already contain the Fmask detection results. However, there are problems with inconsistent versions of Fmask and significant differences in thin cloud detection boundaries. In order to ensure the consistency of the produced samples, this paper adopts the updated version of Fmask v4.6 from February 2022, coupled with the global elevation (SRTM, 90 m) and water body auxiliary data (GSWO), to regenerate the detection results. By comparing the original quality tagging results and adjusting the cloud confidence parameter correction, we obtain quality tagging results with higher accuracy and better relative consistency. The sample label is a single-band PNG format image with a pixel size of 512 × 512, the same as the sample image. The definition of label values for each category is shown in Table 3. Figure 3 shows the example of sample image and sample label.

Table 3. Quality label category setting table.

Class	Land	Water	Cloud Shadow	Snow	Cloud	Fill Value
Label	1	2	3	4	5	0
RGB	(105,111,43)	(0,0,255)	(122,122,130)	(208,225,246)	(255,255,255)	(0,0,0)

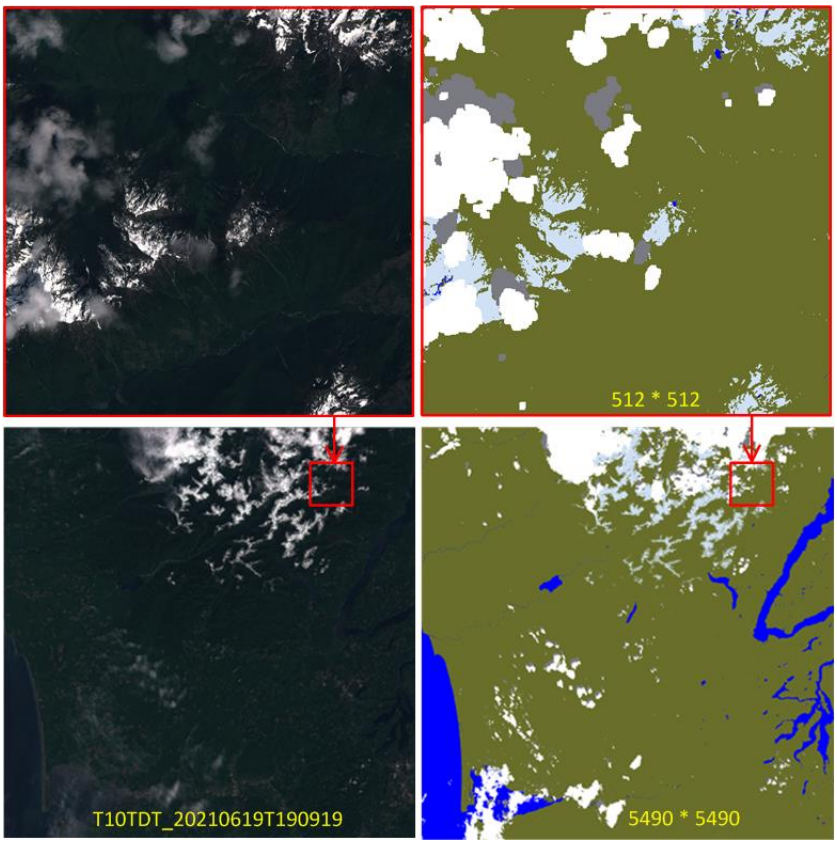


Figure 3. Example of sample image and sample label.

3.1.3. Producing Training Samples

Producing training samples aims to produce high-precision quality tagging labels by Landsat-8 and Sentinel-2 TOA data using the new version of Fmask, combined with a small amount of manual selection and quality corrections (Manual refinement). Then, the sample image and label pair are produced by cropping the full image into a 512 × 512 pixel size

(Sample pair cropping). For the samples cropped from the bottom or right edge, their size is less than 512 pixels. In this case, we fill them into 512 × 512 size with the fill value. The whole process is shown in Figure 4.

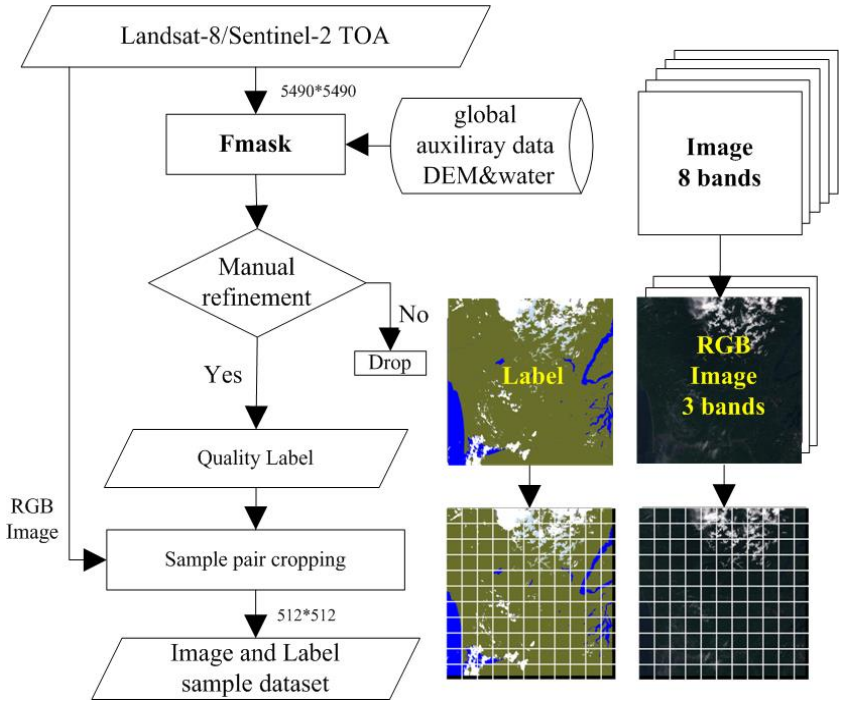


Figure 4. Flowchart of training sample production.

The manual refinement step is essential to guarantee sample quality. The Fmask algorithm mainly relies on the threshold method to detect clouds. However, the cases of "same spectrum for different objects" and "different objects for the same spectrum" in remote sensing images mean that false detection always exists. In the quality tagging task, some high-brightness surfaces and snowy mountains are often mislabeled as clouds. Two typical examples of false detection are shown in Appendix A Figure A1. During imaging time 19 June 2021, white beaches and waves on the west coast of the U.S. with coral reef components were mistakenly detected as clouds. During imaging time 29 December 2021, some high-reflectivity features in the urban areas of the Loess Plateau in China were mistakenly detected as clouds, and dark surfaces were mistakenly matched as cloud shadows. In order to improve the accuracy of the labeled samples, some manual participation is used to filter and refine the quality-labeled result. By simple visual comparison with the original image, we judged whether there were obvious misdetections, and if so, manually determined whether the wrong parts can be corrected quickly. If the correction can be performed within a few minutes using the brush-type tools, then we choose to finish the correction quickly. Otherwise, this sample was directly dropped.

The sample pair cropping step produced the sample image and label pairs. Figure 5 shows the global training sample distribution and quantity. We chose 1120 scenes of Sentinel-2 images distributed among the nine study areas of the "Analysis Ready Data (ARD) Technology Research for Domestic Satellites" project, covering all 12 months. Each scene of the Sentinel-2 image can cut out about 100 sample image pairs (samples with a full fill value are excluded). Finally, we produced more than 100,000 sample image and label pairs, and a certain percentage of them was randomly selected for validation.

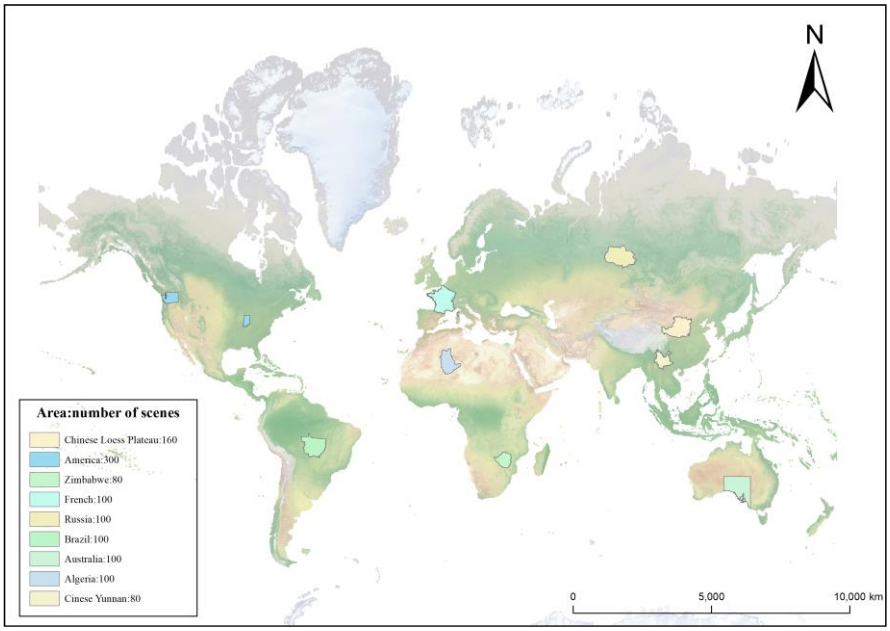


Figure 5. Map of training sample distribution and quantity.

3.2. Algorithm Flow Overview

This section introduces the quality tagging algorithm process for Chinese GF-1/6 satellite images, which consists of two parts: (1) backbone network selection and training, which introduces the segmentation model selection and the iterative update of the model parameters for improving the quality tagging accuracy of GF-1/6 data; and (2) GF-1/6 quality tagging algorithm flow, which introduces our proposed algorithm flow to meet the ARD project's engineering requirements.

3.2.1. Backbone Network Selection and Training

In this paper, we chose the Swin Transformer as our backbone network for the quality tagging algorithm. Meanwhile, two typical CNN-based segmentation models, HRNet and DeepLabv3 are also adopted to address the comparison. The architecture of the Swin Transformer model is shown in Figure 6.

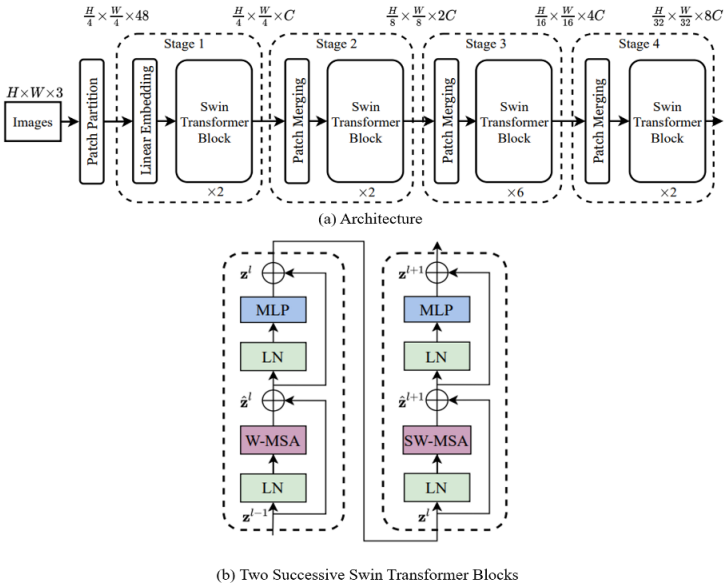


Figure 6. The architecture of the Swin Transformer model.

The entire model of the Swin Transformer adopts a hierarchical design, containing a total of four stages, each of which reduces the resolution of the input feature map and expands the field of perception layer by layer. At the beginning of the input, patch embedding is performed to slice the image into individual blocks and embed them. Each stage consists of patch merging and multiple blocks. The patch merging module mainly reduces the image resolution at the beginning of each stage. Each block consists mainly of LayerNorm (LN), Multilayer Perceptron (MLP), W-MSA, and SW-MSA. W-MSA and SW-MSA are multi-head self-attention modules with regular and shifted windowing configurations, respectively. The parameter C refers to the number of channels, which controls the different model sizes of the Swin Transformer. ViT encodes the embedding position at the input, and the Swin Transformer encodes the relative position at the time of computing attention. ViT adds a separate learnable parameter as the token for classification. However, the Swin Transformer performs the averaging operation directly and outputs the classification, similar to the global averaging pooling layer at the end of CNN.

Window attention is the key to the Swin Transformer model. Traditional Transformers are based on the global to calculate attention, and the computational complexity is very high. However, the Swin Transformer restricts the computation of attention to each window, reducing the amount of computation and lowering the order of self-attention computation. The main difference is that the relative position coding is added for attention calculation. Experiments have shown that adding relative position-coding improves the model performance. Window attention is computed under each window. To better interact with other windows, the Swin Transformer also introduces the shifted window operation, called shifted window attention.

The large Swin Transformer model (Swin-L) was trained using the sample data produced by Landsat-8/Sentinel-2 images, and the model was initialized with the pre-trained parameters from the ImageNet-21K dataset. Since the Landsat-8/Sentinel-2 sample data obtained according to the sample-producing process described above have very little difference from the GF-1/6 data, the model trained with Landsat-8/Sentinel-2 sample data are directly used for processing the GF-1/6 image to acquire the corresponding quality-labeled data products. Suppose the manual accuracy check results reach the quality requirements. In that case, the quality tagging results of GF-1/6 images are added to the training sample set after the sample-producing process. As the ratio of GF-1/6 images in the training sample set keeps increasing, the model parameters keep iterating, which is expected to further improve the accuracy of the GF-1/6 image quality tagging results. The model training and iteration process is shown in Figure 7.

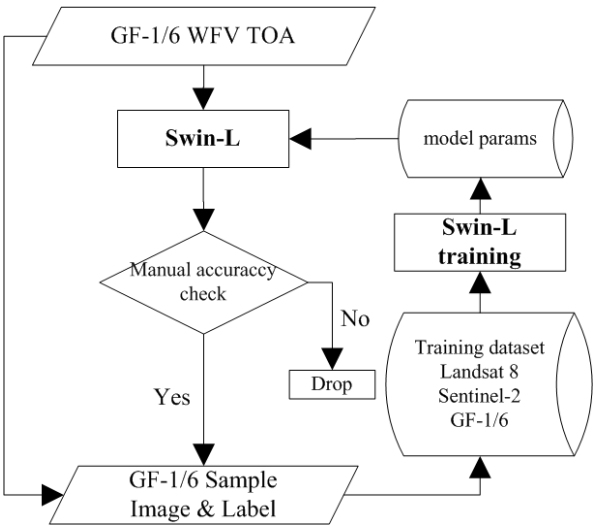


Figure 7. Model training process flowchart.

3.2.2. GF-1/6 Quality Tagging Algorithm Flow

This paper aims to produce quality tagging data products for Chinese GF-1/6 satellite WFV images, and the developed quality tagging algorithm flow needs to meet the ARD project's engineering requirements. Here, we mainly consider two crucial points: (1) The chunking process. Since the Swin Transformer model can only process images with 512 × 512 size, the inconsistency of results between adjacent image blocks may lead to the "seam" problem. We adopt an offset chunking process to cover the boundary area for solving the "seam" problem. (2) Automatic quality tagging correction. In order to further improve the quality tagging accuracy, the Shuttle Radar Topography Mission (SRTM) and Global Surface Water Occurrence (GSWO) data are introduced as auxiliary data. Then, we combine image processing methods to correct the quality tagging results of water bodies, shadows under clouds, and fill values. The algorithm flow of GF-1/6 image quality tagging is shown in Figure 8.

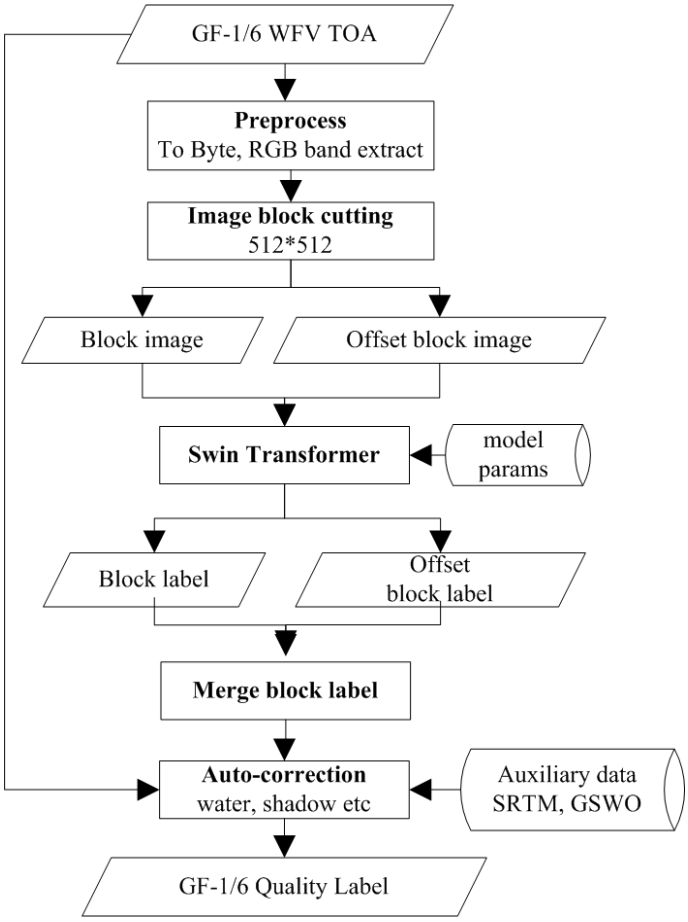


Figure 8. GF-1/6 quality tagging algorithm flowchart.

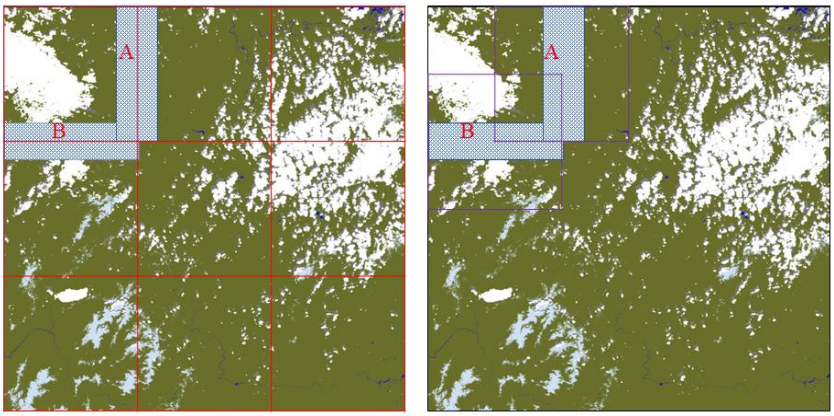
3.3. Details in Engineering Application of Quality Tagging

3.3.1. Seam Correction for Chunking Processed Full Image

The chunking process is a common method for loading large remote sensing images to deep learning algorithms since the Swin Transformer model can only process 512 × 512 fixed-size images simultaneously. After pre-processing the GF-1/6 WFV TOA data, three bands of visible light are selected and converted to Byte-type images according to the same fixed mapping transformation as the sample-producing process. Then, the images are cropped into 512-pixel steps starting from the upper left corner (0,0) pixel position of the source image, and the blocks less than 512 pixels in length are made up using a fill value of 0. The chunked images are pushed into the Swin Transformer model one by one

for processing. Finally, an entire image is created by stitching together the chunked results. Because of the chunking process, there may be inconsistencies between neighboring blocks. Especially in the Swin Transformer model, which includes the self-attention mechanism, the relationship between the boundary pixels and the surrounding pixels will impact the results, and the "seam" problem caused by the edges of neighboring blocks is perhaps more obvious.

In order to eliminate the "seam" problem, we adopt a simple and effective way for seam correction. In short, the idea is to reprocess and update the pixel values of the seam area between two adjacent block images. The specific implementation is to use an offset chunking process. For each vertical seam between two adjacent block images, we crop a 512×512 offset block image which takes this seam as the center line. Then, we produce the corresponding label of that cropped offset block image using the Swin Transformer. Finally, we merge the block label by updating the shadow area A in Figure 9 with the offset block image label. As for the horizontal seam, the operation is similar. After the offset chunking process, we merge the block label by updating the shadow area B in Figure 9 with the offset block image label. The width of shadow area could be set as a variable. In this paper, we choose 64 pixels as the width, based on experience. Finally, we can acquire the quality tagging result of a full image with the "seam" problem eliminated through this seam correction method.



(a) Block image label (b) Offset block image label

Figure 9. Offset chunking process to solve the seam problem.

3.3.2. Automatic Post-Processing Correction

The automatic quality tagging correction step is essential for helping the quality tagging algorithm process to further improve accuracy and stability. It is difficult to avoid the existence of false detections by relying solely on the Swin Transformer model's results. However, some false detections can be corrected by using auxiliary data. Fmask v4.6 utilizes the SRTM and GSWO data to correct the false detections. We adopt the same method here, but more simply and effectively, to correct for possible false detections of water bodies, shadows under clouds, and fill values.

The correction of water body misdetection uses the SRTM and GSWO data. First, for all water bodies' pixel-linked areas, it is determined whether there is a matching water body pixel marked by the GSWO data within a certain range around. If no matching pixel is queried, the SRTM data are used to calculate the average altitude and slope of the pixels in the water-body-linked area. If the average altitude is significantly greater than the average water level in this area, there is a significant slope. Thus, the water-body-linked area is considered as a false detection and should be corrected to the ground surface. The average water level in this region is also calculated using the SRTM and GSWO data. Experiments show that correction for water body misdetection is indispensable. The topographic shadows of mountainous areas are easily detected as water bodies, which can

be corrected easily. Figure 10 shows the example of mountainous topographic shadows falsely detected as water bodies and the results after correction.

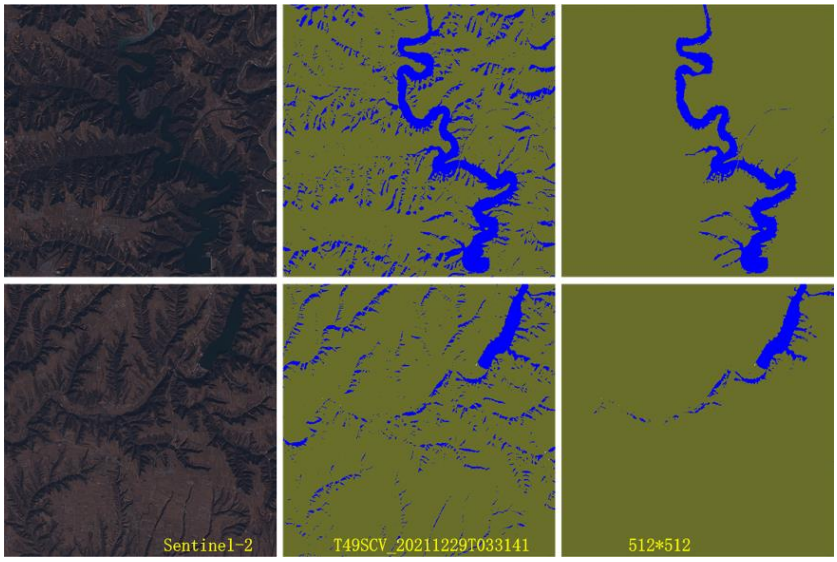


Figure 10. Example of water misdetection and the corresponding correction.

The correction of under-cloud shadow false detection utilizes the geometric relationship between cloud and cloud shadow. The Fmask algorithm uses a complex pixel-linked region geometry matching method to filter out shadow-linked regions that are not matched to clouds. Experimental results demonstrate that shadows under clouds identified by the Swin Transformer model have a good correlation with clouds, and the false detection rate is quite low. Therefore, we use a more straightforward method to correct the false detection of cloud shadows. Whether cloud pixels exist in a certain range around all pixel-linked regions of shadows under clouds is determined. If no cloud pixel exists, they are corrected to the surrounding quality tagging category, such as the ground surface or water bodies. Correction of fill-value misdetection is also necessary. The fill-value areas are four black corner areas formed by image rotation due to remote sensing image system geometric correction, marked as 0 values in the original image. The Swin Transformer model chunking process may lead to large dark areas of water bodies being detected as fill values by mistake. The correction method directly uses the fill value area in the original GF-1/6 WFV TOA image as the final result of the fill value in the quality tagging data products.

4. Experiment

The experiment was divided into two parts. In Section 4.1, we introduced validation experiments on the customized dataset using our proposed method. The effect of different Swin Transformer model size and training sample data volume was analyzed through comparison and quantitative accuracy evaluation. The large Swin Transformer model was then selected for training with a large number of samples on the Sentinel-2 training set. After this, the quality tagging experiment and quantitative accuracy evaluation were performed on the test set. We also compared the results between the Swin Transformer and Fmask algorithm. In Section 4.2, we applied the model trained on Sentinel-2 data transfer to the Chinese GF-1/6 image for quality tagging experiments and visual analysis.

4.1. Validation Experiment on Customized Dataset

4.1.1. Quantitative Evaluation

In this paper, the Swin Transformer was chosen as the backbone network of the quality tagging algorithm. The network architecture and major characteristics of the Swin Transformer have been described previously. The Swin Transformer contains tiny, small, base, and large models, each with different parameter sizes. We employed the small,

base, and large models of the Swin Transformer, but the tiny model was not used in our experiment. The size and computation complexity of the base Swin Transformer model are similar to the base ViT model, while the large and small Swin Transformer models are 2× and 0.5× versions of the base model size, respectively. Unlike CNN, the Swin Transformer does not control the model complexity by using network layers, and the number of Swin Transformer blocks for its four stages is fixed at 2,2,18,2. The control of different model parameters in the Swin Transformer is achieved by the hyperparameter C (number of channels) and the size of C in Swin-S (Small), Swin-B (Base), and Swin-L (Large) is 96, 128, and 192, respectively.

In order to analyze the effects of different Swin Transformer models and training samples of different data volume on the accuracy of quality tagging, several experiments for comparative analysis were designed. Sentinel-2 data were used to produce datasets containing 2k (small), 5k (base), and 10k (large) training samples to carry out the training of small, base, and large Swin Transformer models, respectively. We utilized UperNet [39] in MMSegmentation as our base framework due to its high efficiency. Models were trained on four GPUs with two images per GPU for 160k iterations. The small Swin Transformer model was pre-trained on ImageNet-1k, and the base and large Swin Transformer models were pre-trained on ImageNet-21k. Additionally, the patch and window size settings of the Swin Transformer in our experiment were 4 and 7.

During the training progress, the model accuracy was evaluated on the validation set every 16k iterations, and the accuracy evaluation metrics included the mean pixel accuracy (*mAcc*) and mean intersection over union (*mIoU*). P_{cm} is the confusion matrix, and p_{ij} denotes the number of observations that should actually belong to group i and are predicted to group j. The k refers to segmentation category number, which is 5 in our quality tagging task.

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

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{2}$$

$$P_{cm} = [p_{ij}]^{5\times 5}, i\&j \in [0,k] \tag{3}$$

We randomly selecteded two thousand Sentinel-2 sample images (512 × 512) as the test set. The experiment result showed that the model generalization effect was good. The *mIoU* of the small, base, and large Swin Transformer models reached 70.50%, 75.27%, and 76.53%, respectively. Meanwhile, the *mIoU* of HRNet and DeepLabv3 were 74.78% and 73.49%, respectively. The specific results of different models' *IoU* and *accuracy* are shown in the following Table 4.

Table 4. Quantitative evaluation of different backbone networks on our Sentinel-2 test set.

Model	mIoU (%)	mAcc (%)	Land (1)		Water (2)		Shadow (3)		Snow (4)		Cloud (5)	
			IoU (%)	Acc (%)	IoU (%)	Acc (%)	IoU (%)	Acc (%)	IoU (%)	Acc (%)	IoU (%)	Acc (%)
HRNet	74.78	81.34	90.01	97.19	86.54	90.15	54.77	67.58	53.49	59.79	89.09	91.98
DeepLabv3	73.49	81.13	89.68	96.62	86.88	90.74	51.74	62.51	51.58	63.53	87.60	91.93
Swin-S	70.50	77.63	86.71	96.85	72.13	74.11	51.19	64.79	54.33	60.68	88.12	91.72
Swin-B	75.27	82.46	90.08	97.28	87.37	90.70	52.90	64.62	57.25	67.66	88.76	92.02
Swin-L	76.53	83.72	90.86	97.39	89.22	93.61	54.30	64.82	58.82	70.23	89.47	92.54

In Table 4, the *mAcc* of the small Swin Transformer (Swin-S) model is 77.63%. The land, water, and cloud detection accuracy are higher among the five categories. In comparison,

the snow and shadow are relatively low. Compared with the small model, the *mAcc* of the base Swin Transformer (Swin-B) model is improved by 6.09%, and its *mAcc* reaches 82.46%. The large Swin Transformer (Swin-L) model achieved the highest *mAcc* in the experiment, 1.26% higher than the base model. Therefore, we chose the Swin-L model to perform the subsequent experiment. At the same time, for the selected Swin-L model, we compared the trend of model accuracy improvement during training on 5k, 10k, and 15k training sample datasets. The results can be shown in the Appendix A Figure A2. As the number of iterations increases, the model's accuracy improves and finally tends to a stable value. Comparing the results of 5k, 10k, and 15k training samples, it can be found that the model's accuracy improves with the increase in training sample data volume. At the same time, it has yet to be found that the model has reached saturation, and subsequent experiments could consider further increasing the training data to improve the quality tagging accuracy.

4.1.2. Visual Effect and Comparison with Fmask

Some results from the experiments on Sentinel-2 data using the Swin-L model are displayed below. Local 512 × 512 sample images containing different types of clouds and snowy mountains were mainly selected, and the quality tagging masks produced using the Swin-L model were superimposed on the source image for intuitive visual analysis.

Figure 11a,b,f shows that the Swin-L model is accurate at detecting clouds, and the cloud edge is also consistent with the manual visual interpretation. Images (c), (d), and (e) show that Swin-L not only detects clouds accurately but can also accurately identify shadows under clouds. There is also a good geometric matching relationship between cloud and cloud shadow. Images (g) and (h) show that the model can effectively distinguish between clouds and snow. Swin-L does not mistakenly detect snowy mountains as clouds.

Meanwhile, 20 Sentinel-2 images were selected from the global study area for testing. The quality tagging result of each whole scene image was obtained by the Swin-L model. Then, the quality tagging algorithm flow of the chunking and stitching process was applied, and the results were compared with those of the Fmask algorithm. Three typical scenes' quality tagging results are shown below.

Figures 12–14 show that the quality tagging results produced using Swin-L and Sentinel-2 data have a high agreement with the Fmask algorithm. The enlarged view of the local area in Figure 12 shows that the Fmask algorithm has isolated cloud and cloud shadow false detection noise, while the Swin-L model has almost no such false detection. For Figure 13, the Swin-L result is also better than Fmask. The enlarged view of the local area shows that Swin-L has more accurate edge recognition of thin clouds, while the edge of the cloud area in the Fmask algorithm is more expanded. At the same time, the Fmask algorithm also wrongly detects some cloud shadows and isolated water body noise. Figure 14 shows that Swin-L can effectively distinguish clouds from snow and achieve the same accuracy as the Fmask algorithm.

4.2. Producing GF-1/6 Image Quality Tagging Mask

GF-1/6 data and Sentinel-2 data have high similarity in the visible light band. This part of the experiment applied the model obtained from the Sentinel-2 data transfer to the Chinese GF-1/6 images for conducting quality tagging experiments. Since the Fmask algorithm currently only supports Sentinel-2 and Landsat series data, the accuracy verification of GF-1/6 data mainly relies on expert manual interpretation. We randomly selected 100 scenes of data from nine study areas, especially snowy mountain areas with seasonal snow line changes, such as Washington State in the United States. Some typical examples are shown in Figure 15 for visual analysis. Local 512 × 512 sample image quality tagging results showed that the Swin-L model can achieve high accuracy on GF-1/6 data. Our results have a good visual effect and can effectively distinguish clouds from snow.

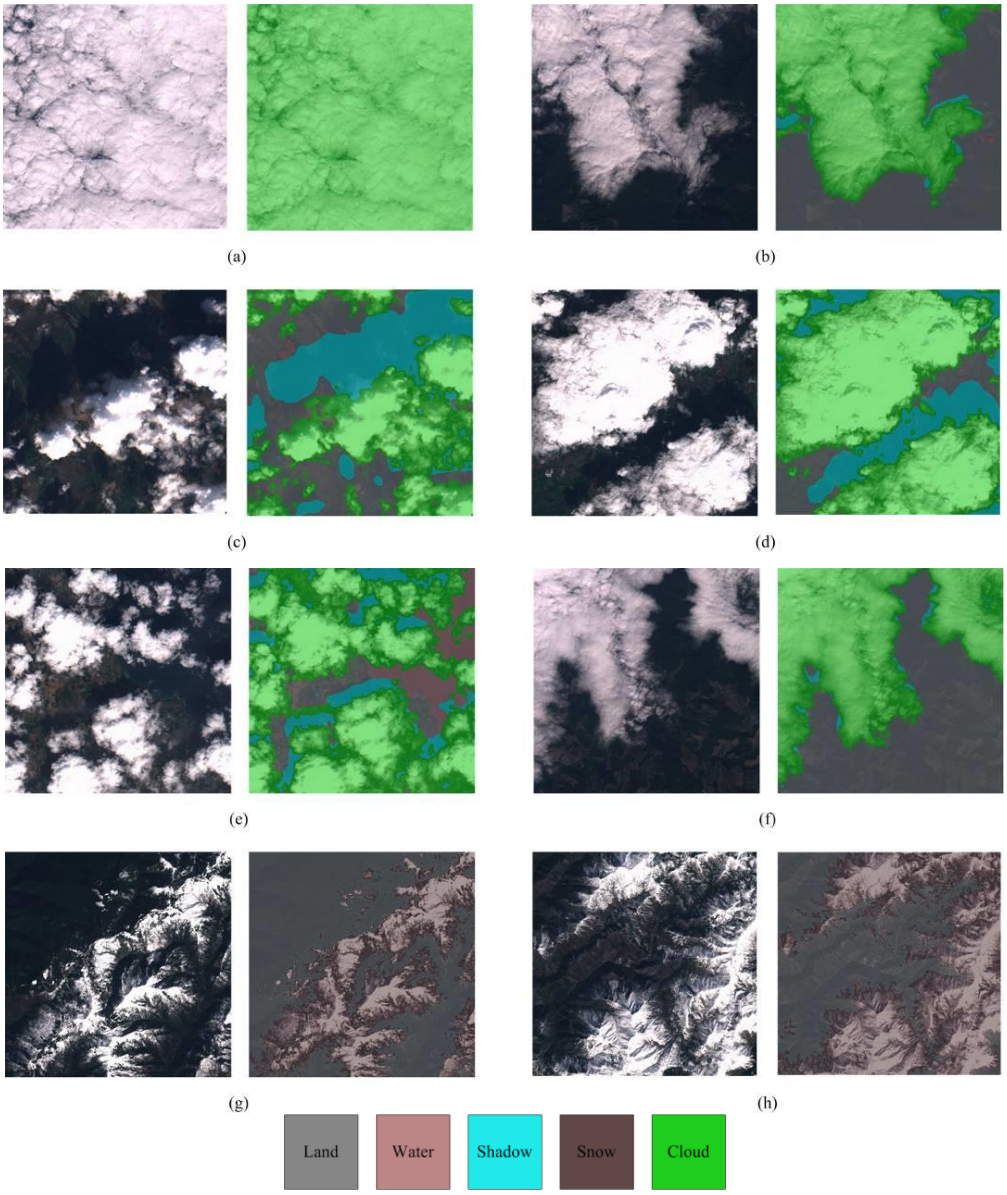


Figure 11. Typical visual examples of Sentinel-2 sample images and quality tagging masks (512 × 512) produced by Swin-L. (**a**–**f**) Cloud and cloud shadow detection. (**g**,**h**) Snow detection.

Next, the quality tagging algorithm flow was applied to the chunking and stitching process for the selected GF-1/6 scene images, aimed at obtaining the quality tagging label of each whole scene image. As the training sample dataset continues to expand, the results of quality tagging gradually improve. Figure 16 shows the process of distinguishing clouds and snow, first with lots of error, then with less error, and finally becoming completely separable. At the 2k sample data volume, the quality tagging result showed a large number of cases where snow was mistakenly detected as clouds and water bodies. When the amount of data reaches 5k, the cloud and snow are basically separated, but there are still a small number of cases where snow is mistakenly detected as water bodies. When the data volume reaches 10k, cloud and snow are completely separable, and there is no apparent false detection. The quality tagging results meets the requirements of the ARD project's engineering applications.

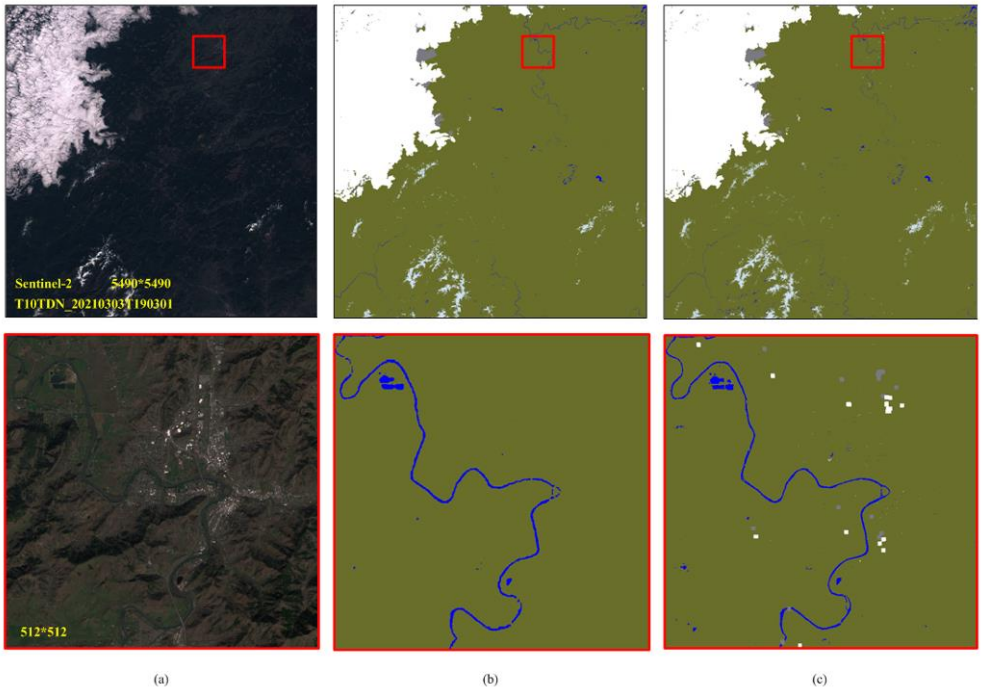


Figure 12. Comparison of Swin-L and Fmask quality tagging result (1). (**a**) RGB source image. (**b**) Swin-L label. (**c**) Fmask label.

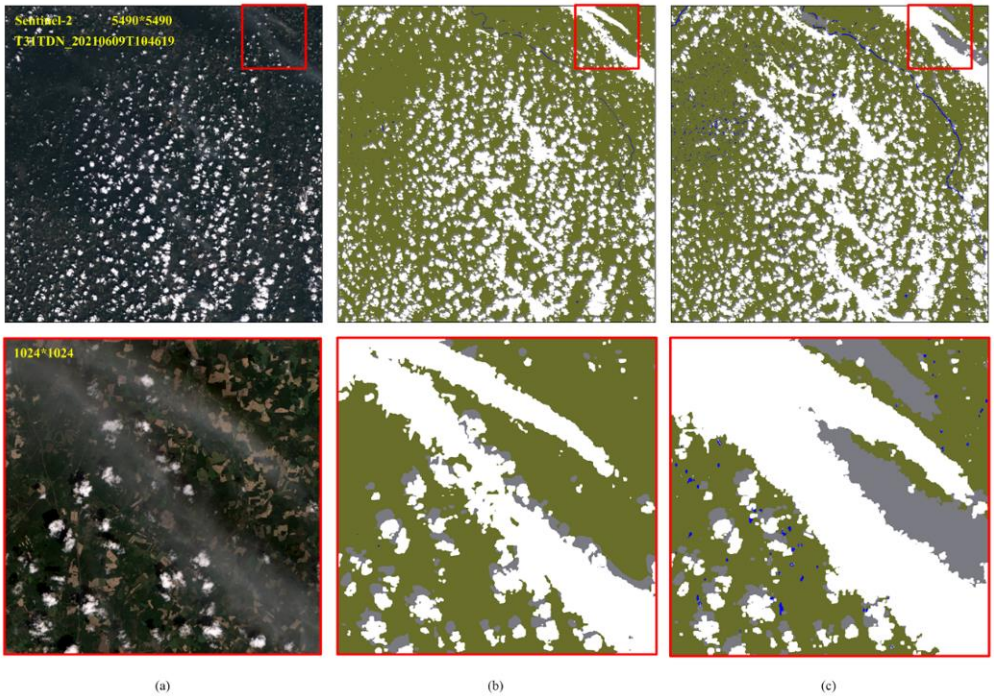


Figure 13. Comparison of Swin-L and Fmask quality tagging result (2). (**a**) RGB source image. (**b**) Swin-L label. (**c**) Fmask label.

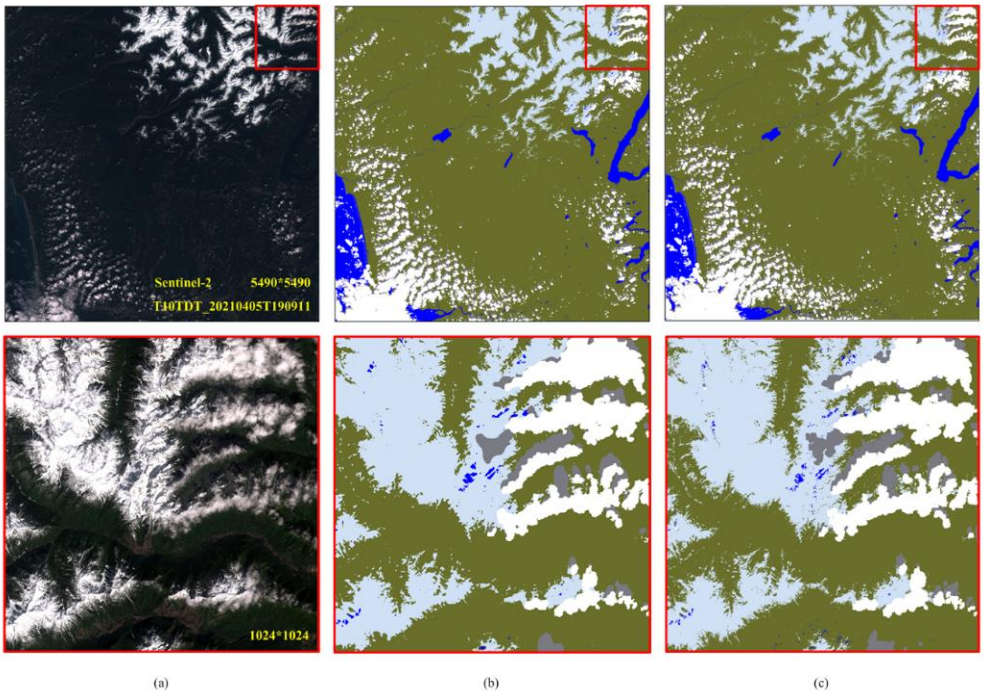


Figure 14. Comparison of Swin-L and Fmask quality tagging result (3). (**a**) RGB source image. (**b**) Swin-L label. (**c**) Fmask label.

Figure 17 shows several representative results of the GF-1/6 image quality tagging masks produced using the final well-trained Swin-L model. The quality tagging label is generally visually good. The Swin Transformer can detect thick and broken clouds more accurately. The local area enlargement shows that its recognition results of shadows under clouds also have an excellent geometric matching relationship with the cloud area. Swin Transformer can also accurately identify snowy mountains on GF-1/6 images and does not mistakenly detect snow as clouds. There are more typical quality tagging results of GF-1/6 images in Appendix A Figure A3.

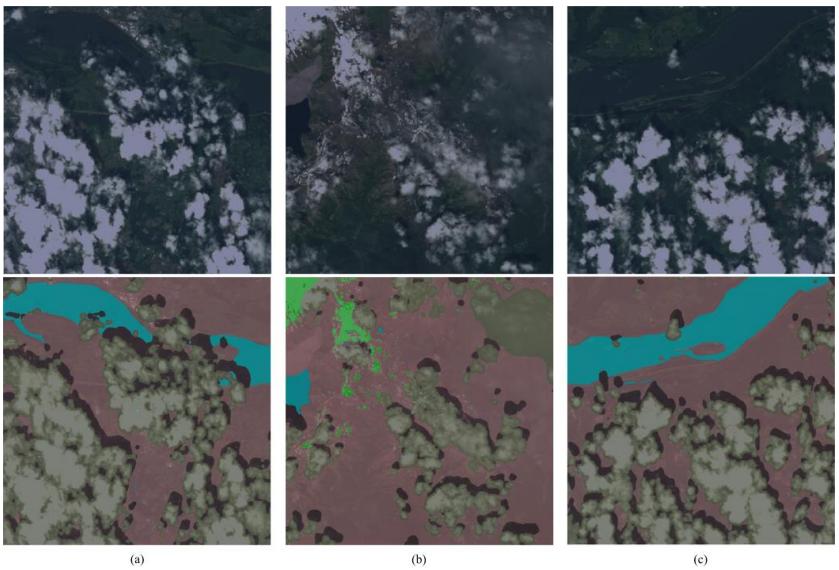


Figure 15. *Cont.*

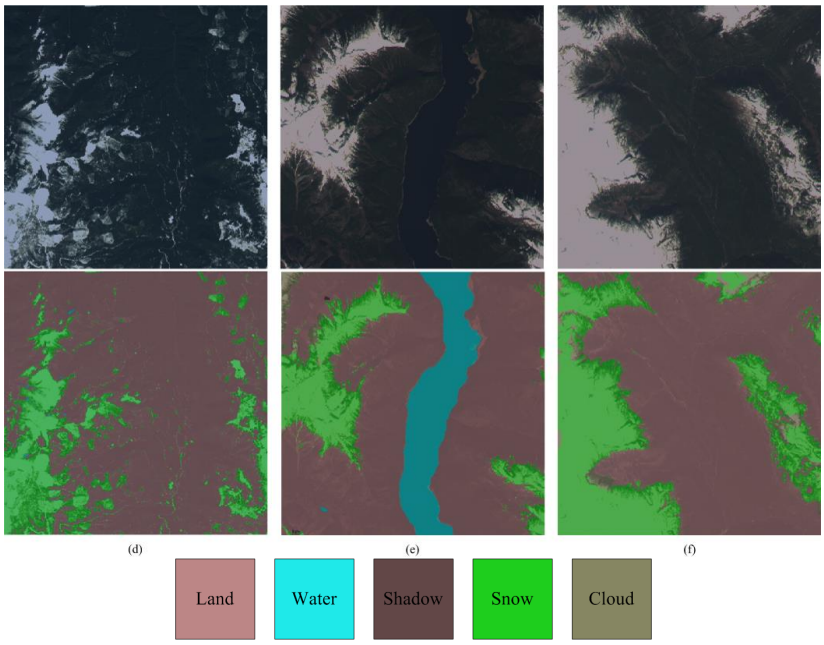


Figure 15. Typical visual examples of GF-1/6 sample images and quality tagging masks (512 × 512) produced by Swin-L. (**a**–**c**) Cloud and cloud shadow detection. (**d**–**f**) Snow detection.

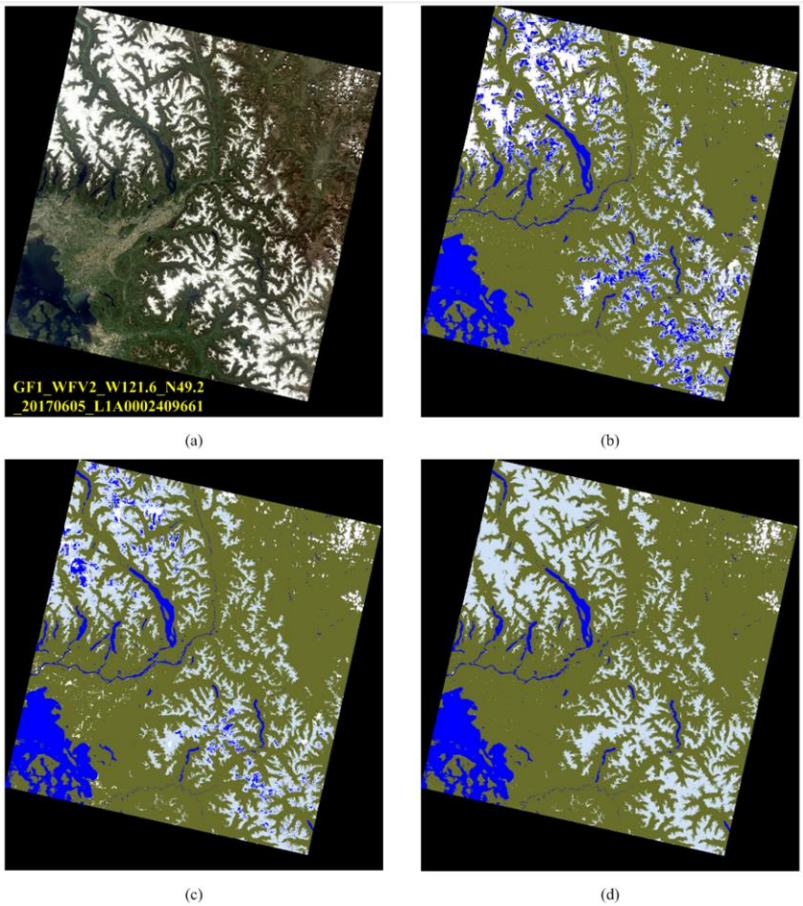


Figure 16. Example GF-1 WFV image and quality tagging masks produced by different training sample numbers of Swin-L models. (**a**) RGB source image; (**b**) 2k samples based on Swin-L mask; (**c**) 5k samples based on Swin-L mask; (**d**)10k samples based on Swin-L mask.

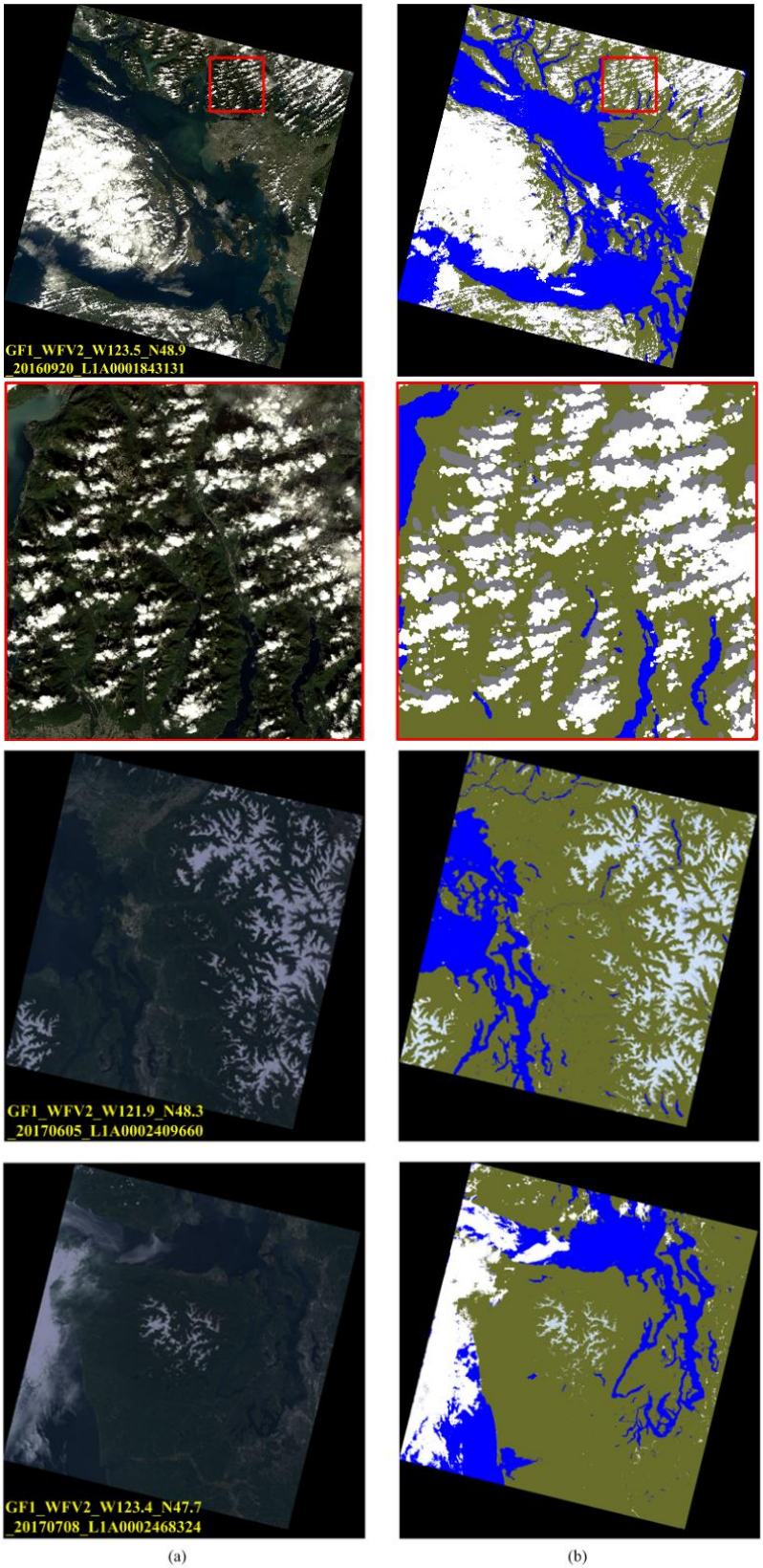


Figure 17. Examples of GF-1 WFV full images and quality tagging masks produced by Swin-L. (**a**) RGB source image. (**b**) Swin-L label.

5. Discussion

Our method's GF-1/6 image quality tagging accuracy can be close to the results of Sentinel-2 images using the Fmask algorithm. It indicates that the Swin Transformer's Large model can learn the features of six quality tagging categories (land, water, cloud shadow, snow, cloud, and fill value) in the visible band through a large-size training sample set. The experiments show that when the number of training samples reaches 100,000, our method's accuracy improves significantly, the mean accuracy is 86.25%, and the overall accuracy consistency with Fmask v4.6 is 85.52%. Meanwhile, our method for the regional edge of cloud and cloud shadow is usually better than Fmask.

Moreover, our GF-1/6 image quality tagging algorithm can effectively distinguish clouds from snow. With over 10,000 training samples, the case of snowy mountains being falsely detected as clouds was significantly reduced, and the overall accuracy was 83.72%. The Sentinel-2 data can distinguish clouds from snow, usually by using spectral features. However, the Swin Transformer model can also distinguish by shape and texture features in visible RGB images. In remote sensing images with a 20 m spatial resolution, the manual visual interpretation can intuitively distinguish clouds and snow by experience. On the premise of sufficient training samples, our experiment indicated that the Swin Transformer model can obtain the same capability using only the RGB band.

6. Conclusions

In this paper, we proposed a novel pixel-by-pixel quality tagging algorithm flow for Chinese GF-1/6 satellite WFV images. It aims to achieve the requirements of the "Analysis Ready Data (ARD) Technology Research for Domestic Satellites" project and resolve the lack of Chinese satellite quality tagging data products. Considering the similarity of Landsat-8/Sentinel-2 and GF-1/6 multispectral images in band and resolution, a generalizable training sample set was constructed. Utilizing the Fmask algorithm with 1120 scenes of Sentinel-2 images, over 100,000 training samples were produced with a small number of manual corrections. Then, the training results of the Swin Transformer's Large model are directly used for pixel-by-pixel quality tagging of GF-1/6 images. Aiming to analyze the applicability of the Transformer model's different characteristics from the CNN model in the remote sensing image segmentation field, this paper carried out experiments using the original Swin Transformer model and corresponding pretrained parameters. Compared with the multi-scale feature of CNN, the self-attention mechanism of the Transformer model is more suitable for the semantic segmentation problem of remote sensing images with a fixed spatial resolution.

The preliminary results of this study reflect some advantages of the Transformer model, such as being more suitable for large-size training sample sets and less prone to saturation problems. Meanwhile, we also developed important methods including fixed mapping transformation, seam correction for full chunking processed images, and automatic post-processing correction. We provide a complete GF-1/6 satellite quality tagging algorithm flow and data product specification using the proposed method in this paper. However, there is still some work left to do. In the following work, we will modify and optimize the original Swin Transformer model to make it more suitable for remote sensing image semantic segmentation problems.

Author Contributions: Conceptualization, X.F. and C.H.; methodology, X.F. and C.H.; software, X.F. and C.H.; data curation, H.C.; writing—original draft preparation, X.F. and H.C.; writing—review and editing, C.H. and L.H.; visualization, X.F.; supervision, C.H and L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partly supported by the National Key Research and Development Program of China (grant number 2019YFE0197800) and National Natural Science Foundation of China (grant number 41971396).

Acknowledgments: We would like to thank OpenMMLab for their semantic segmentation toolbox and benchmark [40]. An open-source implementation of our GF-1/6 quality tagging algorithm-based Swin Transformer with customized dataset is publicly available at https://github.com/fanxin-cug/mmsegmentation-cloud (accessed on 1 April 2023).

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

Appendix A

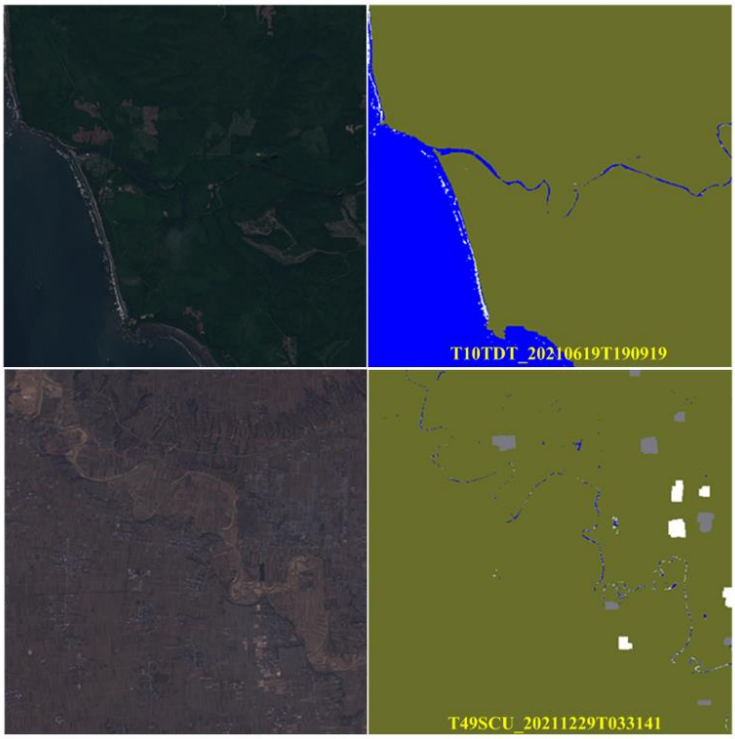


Figure A1. Example of training samples requiring manual refinement.

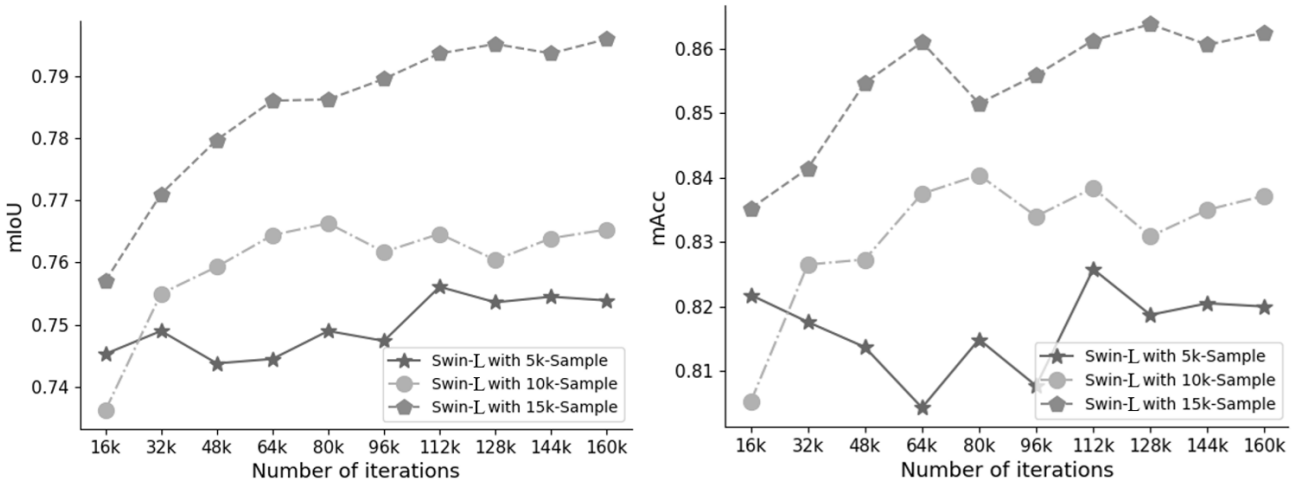


Figure A2. The Swin-L's mIoU and mAcc performance with different training sample data volume during training progress.

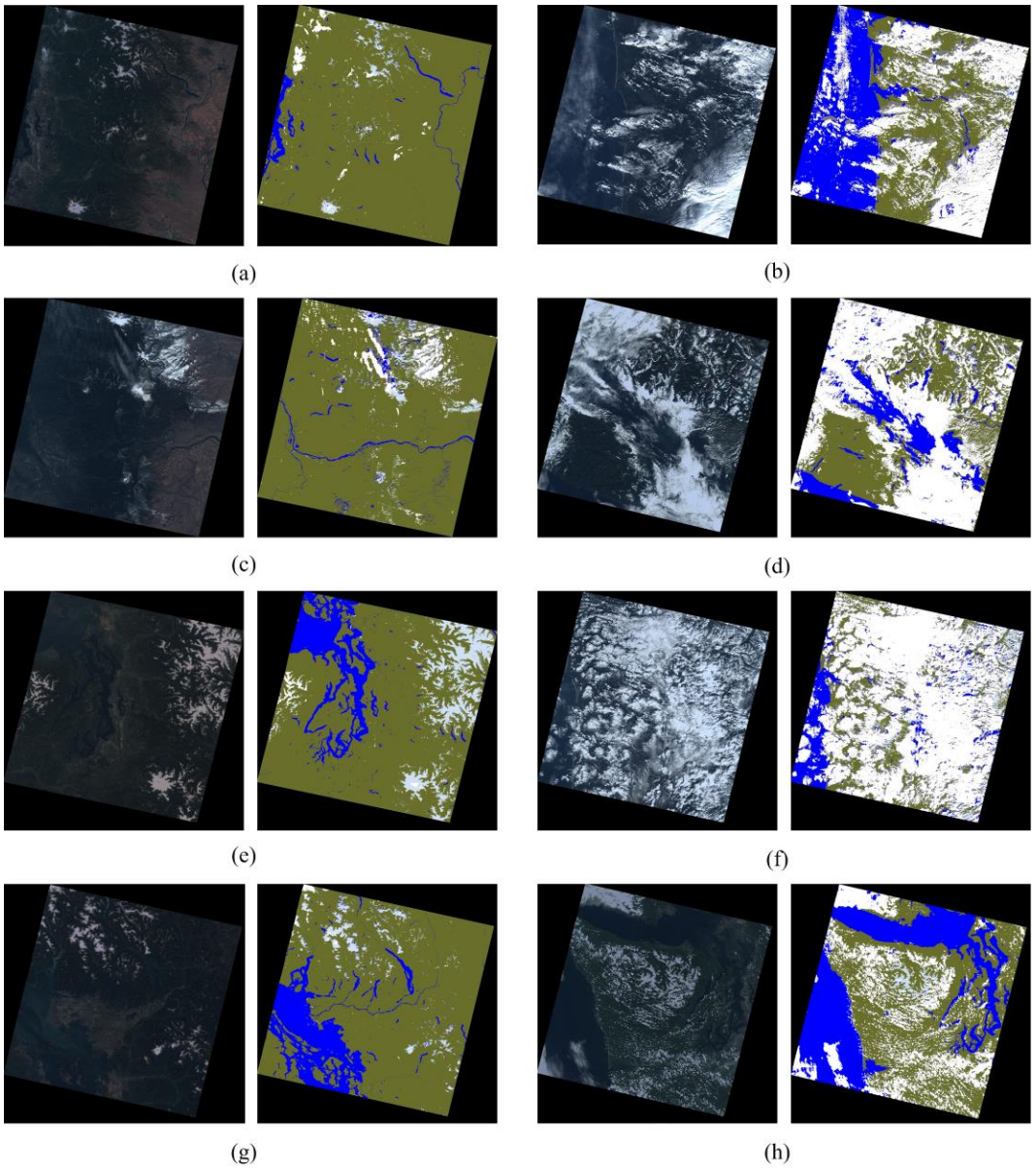


Figure A3. More examples of GF-1/6 WFV full images and quality tagging masks produced by Swin-L. (**a**) W121.1_N47.6_20170724. (**b**) W123.5_N45.9_20201123. (**c**) W121.7_N45.9_20201025. (**d**) W123.5_N49.2_20201005. (**e**) W122.2_N47.6_20150418. (**f**) W122.9_N45.9_20210217. (**g**) W122.3_N49.3_20160818. (**h**) W123.5_N47.6_20210612.

References

1. 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. [CrossRef]
2. Zhu, Z.; Woodcock, C.E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **2012**, *118*, 83–94. [CrossRef]
3. Foga, S.; Scaramuzza, P.L.; Guo, S.; Zhu, Z.; Dilley, R.D., Jr.; Beckmann, T.; Schmidt, G.L.; Dwyer, J.L.; Hughes, M.J.; Laue, B. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sens. Environ.* **2017**, *194*, 379–390. [CrossRef]
4. Qiu, S.; Zhu, Z.; He, B. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4-8 and Sentinel-2 imagery. *Remote Sens. Environ.* **2019**, *231*, 111205. [CrossRef]
5. Qiu, S.; Lin, Y.; Shang, R.; Zhang, J.; Ma, L.; Zhu, Z. Making Landsat Time Series Consistent: Evaluating and Improving Landsat Analysis Ready Data. *Remote Sens.* **2019**, *11*, 51. [CrossRef]
6. Mahajan, S.; Fataniya, B. Cloud detection methodologies: Variants and development—A review. *Complex Intell. Syst.* **2020**, *6*, 251–261. [CrossRef]

7. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
8. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 630–645.
10. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
11. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
12. Ronneberger, O.; Fischer, P.; Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*; Springer: Cham, Switzerland, 2015; pp. 234–241.
13. 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.
14. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
15. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
16. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
17. 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.
18. Strudel, R.; Pinel, R.G.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. *arXiv* **2021**, arXiv:2105.05633.
19. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. *arXiv* **2020**, arXiv:2012.15840.
20. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203.
21. Petit, O.; Thome, N.; Rambour, C.; Soler, L. U-Net Transformer: Self and Cross Attention for Medical Image Segmentation. *arXiv* **2021**, arXiv:2103.06104.
22. Hughes, M.; Hayes, D.; Oak Ridge National Lab; Ornl ORTU. Automated Detection of Cloud and Cloud Shadow in Single-Date Landsat Imagery Using Neural Networks and Spatial Post-Processing. *Remote Sens.* **2014**, *6*, 4907–4926. [CrossRef]
23. Chai, D.; Newsam, S.; Zhang, H.K.; Qiu, Y.; Huang, J. Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks. *Remote Sens. Environ.* **2019**, *225*, 307–316. [CrossRef]
24. Jeppesen, J.H.; Jacobsen, R.H.; Inceoglu, F.; Toftegaard, T.S. A cloud detection algorithm for satellite imagery based on deep learning. *Remote Sens. Environ.* **2019**, *229*, 247–259. [CrossRef]
25. Grabowski, B.; Ziaja, M.; Kawulok, M.; Longépé, N.; Saux, B.L.; Nalepa, J. Self-Configuring nnU-Nets Detect Clouds in Satellite Images. *arXiv* **2022**, arXiv:2210.13659.
26. Jiao, L.; Huo, L.; Hu, C.; Tang, P. Refined UNet: UNet-Based Refinement Network for Cloud and Shadow Precise Segmentation. *Remote Sens.* **2020**, *12*, 2001. [CrossRef]
27. Jiao, L.; Huo, L.; Hu, C.; Tang, P. Refined UNet V2: End-to-End Patch-Wise Network for Noise-Free Cloud and Shadow Segmentation. *Remote Sens.* **2020**, *12*, 3530.
28. Jiao, L.; Huo, L.; Hu, C.; Tang, P. Refined unet v3: Efficient end-to-end patch-wise network for cloud and shadow segmentation with multi-channel spectral features. *Neural Networks* **2021**, *143*, 767–782.
29. Jiao, L.; Huo, L.; Hu, C.; Tang, P.; Zhang, Z. Refined UNet V4, End-to-End Patch-Wise Network for Cloud and Shadow Segmentation with Bilateral Grid. *Remote Sens.* **2022**, *14*, 358. [CrossRef]
30. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
31. 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**, *191*, 342–358. [CrossRef]
32. Wang, M.; Zhang, Z.; Dong, Z.; Jin, S.; Hongbo, S. Stream-computing Based High Accuracy On-board Real-time Cloud Detection for High Resolution Optical Satellite Imagery. *Acta Geodaet. Cartogr. Sin.* **2018**, *47*, 76–769. [CrossRef]
33. Li, T.T.; Tang, X.M.; Gao, X.M. Research on separation of snow and cloud in ZY-3 image cloud recognition. *Bull. Survey. Mapp.* **2016**, *0*, 46–49+68.
34. Guo, Y.; Xiaoqun, C.; Bainian, L.; Mei, G. Cloud Detection for Satellite Imagery Using Attention-Based U-Net Convolutional Neural Network. *Symmetry* **2020**, *12*, 1056. [CrossRef]
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.

36. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale; International Conference on Learning Representations. *arXiv* **2021**, arXiv:2010.11929.
37. Zhu, Z.; Woodcock, C.E. Improvement and Expansion of the Fmask Algorithm: Cloud, Cloud Shadow, and Snow Detection for Landsats 4-7, 8, and Sentinel 2 images. *Remote Sens. Environ.* **2014**, *159*, 269–277. [CrossRef]
38. Qiu, S.; He, B.; Zhu, Z.; Liao, Z.; Quan, X. Improving Fmask cloud and cloud shadow detection in mountainous area for Landsats 4–8 images. *Remote Sens. Environ.* **2017**, *199*, 107–119. [CrossRef]
39. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding; In Proceedings of the European Conference on Computer Vision (ECCV). *arXiv* **2018**, arXiv:1807.1022; pages 418–434, 418–434.
40. MMSegmentation Contributors. MMSegmentation: Openmmlab Semantic Segmentation Toolbox and Benchmark. 2020. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 10 November 2022).