# Transferring Deep Models for Cloud Detection in Multisensor Images via Weakly Supervised Learning

Shaocong Zhu, Zhiwei Li, and Huanfeng Shen, *Senior Member, IEEE*

*Abstract*— Recently, deep learning has been widely used for cloud detection in satellite images; however, due to radiometric and spatial resolution differences in images from different sensors and time-consuming process of manually labeling cloud detection datasets, it is difficult to effectively generalize deep learning models for cloud detection in multisensor images. This article propose a weakly supervised learning method for transferring deep models for cloud detection in multisensor images (TransMCD), which leverages the generalization of deep models and the spectral features of clouds to construct pseudo-label dataset to improve the generalization of models. A deep model is first pretrained using a well-annotated cloud detection dataset, which is used to obtain a rough cloud mask of unlabeled target image. The rough mask can be used to determine the spectral threshold adaptively for cloud segmentation of target image. Block-level pseudo labels with high confidence in target image are selected using the rough mask and spectral mask. Unsupervised segmentation technique is used to construct a high-quality pixel-level pseudo-label dataset. Finally, the pseudo-label dataset is used as supervised information for transferring the pretrained model to target image. The TransMCD method was validated by transferring model trained on 16-m Gaofen-1 wide field of view(WFV)images to 8-m Gaofen-1, 4-m Gaofen-2, and 10-m Sentinel-2 images. The F1-score of the transferred models on target images achieves improvements of 1.23%–9.63% over the pretrained models, which is comparable to the fully-supervised models trained with well-annotated target images, suggesting the efficiency of the TransMCD method for cloud detection in multisensor images.

*Index Terms*— Cloud detection, deep learning, multisensor images, weakly supervised learning.

## I. INTRODUCTION

CLOUDS are inevitable contaminants in optical remote sensing images [1]. The presence of clouds leads to information loss [2], which affects the processing and precise application of optical remote sensing images [3], [4], [5], [6]. Cloud detection is a crucial preprocessing step in the processing of optical remote sensing images. Its objective is to identify and segment the cloudy regions within an image. Accurate cloud detection can minimize the negative impact of clouds on image applications and enhance the usability of the images. It is, therefore, necessary to perform cloud detection in remote sensing images.

In recent decades, several cloud detection methods for remote sensing images have been proposed, which can be grouped into four main categories [7]: 1) physical rule-based methods, 2) temporal change-based methods, 3) variational model-based methods, and 4) machine learning-based algorithms. The physical rule-based algorithms use the physical properties of clouds to determine the threshold for cloud detection. Among these methods, the Function of Mask (Fmask) [8], [9] identifies potential cloud regions using specific spectral rules, such as spectral variability and brightness temperature. Haze optimized transformation [10], [11] is also widely used to segment the cloud region in images. The physical rule-based algorithms are effective in some conditions [12], but they rely on manual experience for designing the physical rules, resulting in certain limitations for the physical rule-based algorithms. The temporal change-based algorithms achieve cloud detection by identifying the abrupt changes between time-series images of the same region [13]. Cirrus cloud mask (CMask) [14] uses a time series model to identify cloud regions when the top of the actual atmosphere reflectance of the Cirrus band is higher than the predicted value from the model. The temporal change-based algorithms require a specific quality and quantity of images to obtain an accurate cloud mask. The variational model-based algorithms construct a variational model based on a *priori* knowledge of the cloud and clear regions in the images to achieve cloud detection [15], [16]. The variational model-based algorithms perform well in combining cloud detection and cloud removal; however, these algorithms require high-quality images as input data, so their performance may not be optimal in some cases. The machine learning-based algorithms consider cloud detection as an image segmentation problem [17], enabling the model to learn cloud features from the sample set for accurate

Shaocong Zhu and Huanfeng Shen are with the School of Resource and Environmental Science, Wuhan University, Wuhan 430079, China (e-mail: zhushaocong@whu.edu.cn; shenhf@whu.edu.cn).

Zhiwei Li is with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: zhiwei. li@polyu.edu.hk).

cloud detection. Many machine learning algorithms have been widely applied to cloud detection, including random forest [18], support vector machine [19], [20], and neural networks [21], [22].

Deep learning, as a major branch of machine learning, has made significant progress in cloud detection. Among the different methods, the fully supervised learning methods using manually-labeled pixel-level datasets have been intensively studied, and their accuracy is continuously improving with the development of deep learning [23]. These methods typically use the spectral values of individual pixels or localized regions as input to the model in order to generate the corresponding results [24]. Numerous studies have designed models with different network architectures in order to enhance the accuracy of cloud detection [25], [26]. Adding diverse features to the model input, such as geographic information [27] and texture features [28], can improve the accuracy of the model in distinguishing clouds from other highlighted landscapes. Some researchers have combined image segmentation technology, such as simple linear iterative clustering (SLIC) [29] with a convolutional neural networks (CNNs) [30], [31] to improve the efficiency and accuracy of cloud detection.

The fully supervised learning methods, however, require a large amount of pixel-level labels to achieve a superior performance [32]. The satellite images from different sensors typically exhibit significant variations in spectral reflections and spatial resolution. Deep learning models trained on specific sensor images are difficult to apply to images from different sensors. This means that each sensor requires a large number of manually annotated labels in order to achieve accurate cloud detection, which is time-consuming and labor-intensive. As a result, weakly supervised learning methods have been widely applied in cloud detection to reduce the reliance on manual annotation of labels. The weakly supervised learning methods train a deep model using low-cost block-level label supervision [33], [34] and combined techniques such as cloud activation maps [35] or attention mechanisms [36] to achieve pixel-level cloud detection. Transfer learning, as a popular field in deep learning, has been used in cloud detection tasks. It enhances the performance of the model in the target domain by transferring knowledge from a relevant source domain [37]. Transfer learning-based algorithms can reduce the dependence of the model on target domain labels and can leverage existing cloud detection datasets for sensors that lack datasets [38]. Mateo-Garcia et al. [39] trained a model based on the cycle-consistent generative adversarial domain adaptation framework using Landsat 8 data and transferred the model to Proba-V data. Zhou et al. [40] proposed a transferable DeepLab V3+ model to achieve the cloud detection task by transferring the PASCAL VOC 2012 dataset to a ground-based cloud segmentation dataset.

Most of the existing cloud detection methods are, however, developed for images captured by a single sensor, such as Landsat [41], Sentinel-2 [42], [43], and Gaofen [44]. The acquisition of multisensor high-resolution images is becoming increasingly common, which makes it challenging to construct a well-annotated large-scale cloud detection dataset for each sensor. The high-resolution images have a higher spatial resolution and fewer spectral bands [red, blue, green, and near-infrared (NIR)], lacking efficient bands for cloud detection [45]. It is, therefore, necessary to develop a cloud detection method for multisensor high-resolution images that uses the existing cloud detection datasets to train the model and transfer it to sensors without labeled datasets. In this article, we present a novel approach called TransMCD, which is a weakly supervised learning method designed for transferring a deep model to effectively detect clouds in multisensor images. The proposed method pretrains a deep learning model using a well-annotated large-scale dataset. The spectral features of clouds with high reflectivity in the visible and NIR bands are then combined with the generalization of the pretrained model to generalize and automatically search for reliable training samples from an unlabeled image dataset. For samples with high confidence, the unsupervised image segmentation (UIS) technique [46] is employed to obtain pseudo labels for model fine-tuning.

In summary, the contributions of this article are as follows.
1) We propose a novel weakly supervised learning method that pretrains a deep learning model using an existing cloud detection dataset and transfers the model to unlabeled multisensor images. The method can automatically extract reliable information from unlabeled images, eliminating the need for manual annotation, and is applicable to most deep learning models for cloud detection.
2) We developed a framework that fuses model generalization with the physical features of clouds to construct a cloud detection dataset. When working with multisensor images, the manual adjustment of the spectral threshold is avoided by extracting the spectral threshold of the unlabeled image using the rough cloud masks obtained from the pretrained model.
3) Two novel datasets, GF1MS-WHU and GF2MS-WHU, are introduced for cloud detection. The GF1MS-WHU dataset consists of 141 unlabeled and 33 well-annotated 8-m Gaofen-1 multispectral (GF1-MS) images; furthermore, the GF2MS-WHU dataset includes 163 unlabeled and 29 well-annotated 4-m Gaofen-2 multispectral (GF2-MS) images. Based on the labeled images in the two datasets, a total of $10\,428$ and $21\,917$ fully labeled image patches have been made available at https://github.com/whu-ZSC/GF1-GF2MS-WHU.

## II. METHOD

We define a well-annotated large-scale dataset as the source domain $D_S$ and a newly acquired dataset without any labels as the target domain $D_T$. The purpose of the TransMCD method is to transfer the features learned by the CNN model from $D_S$ to $D_T$ so that the model can accurately detect clouds in the $D_T$ images. To transfer a CNN model for cloud detection in multisensor images with varying imaging conditions, the TransMCD method uses pseudo-label assignment [47] and model fine-tuning [4], [48] techniques. The proposed method consists of three main steps, as shown in Fig. 1. First, a well-annotated dataset $D_S$ is used to train the CNN model, with
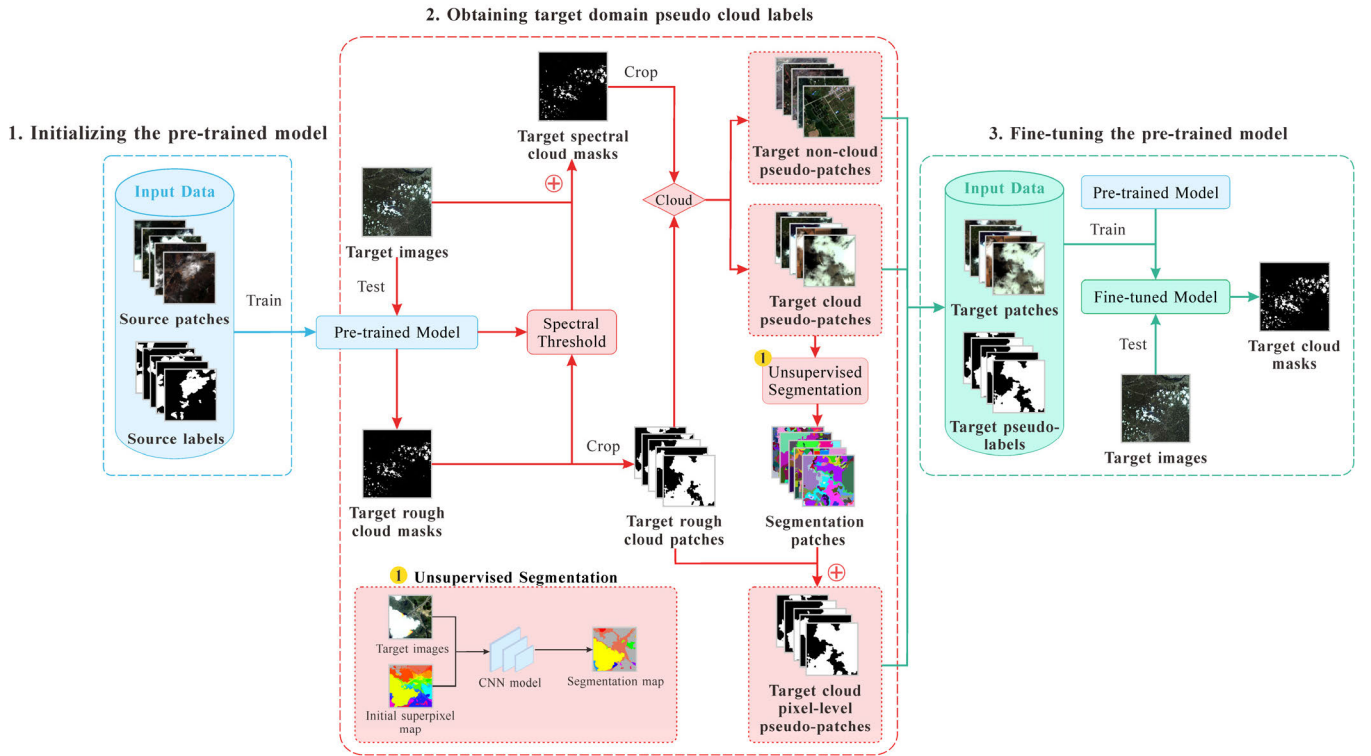
Fig. 1. Framework of the proposed weakly supervised learning method for transferring deep models for cloud detection.
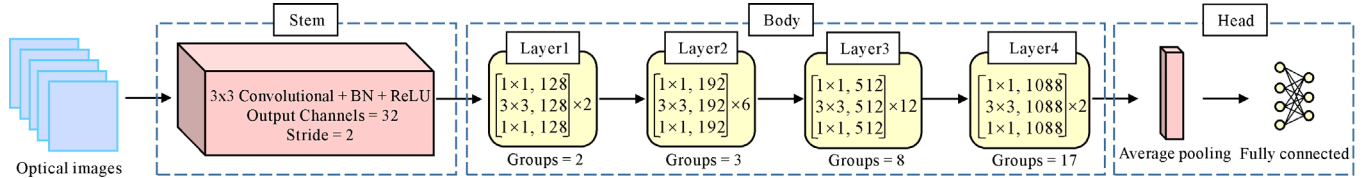


Fig. 2. Structure of the RegNetY-4.0GF model.

the objective of obtaining a well-trained pretrained model for the purpose of cloud detection. Second, the pretrained model is used to obtain a rough cloud mask $M_R$ for the $D_T$ images. Subsequently, the spectral cloud mask $M_S$ is obtained through the automatic extraction of the spectral thresholds by $M_R$. The $M_R$ and $M_S$ masks are combined to search for the $D_T$ block-level pseudo labels with high confidence, and the $D_T$ pixel-level pseudo labels are obtained by the optimization of $M_R$ using image superpixel segmentation and the UIS technique. Finally, the pretrained model is fine-tuned using the $D_T$ pixel-level pseudo labels, in order to transfer the model to the $D_T$ images.

### A. Initializing the Pretrained Deep Models for Cloud Detection

CNN models are widely applied in cloud detection methods. A CNN model is made up of various layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers are responsible for the feature extraction, the pooling layers are responsible for the downsampling or spatial reduction, and the fully connected layers function as a classifier to generate the predicted probabilities. In this study, the RegNetY-4.0GF model [49] of the RegNet family was used as the cloud detection model. The structure of the

model is shown in Fig. 2. The RegNetY model consists of three main components: 1) the stem component of the model incorporates a stride two $3 \times 3$ convolutional layer with 32 output channels; 2) the body of the model is composed of a series of downsampled layers, with each layer containing a set of residual bottleneck blocks with group convolution; and 3) the head component consists of an average pooling layer and a fully connected layer. Because of cloud detection being a binary classification task, we configured the number of output channels of the prediction score from the model to 2 and normalized the prediction score to 0–1 using the sigmoid function to obtain the cloud probability. Subsequently, pixels with prediction scores higher than 0.5 were classified as cloud pixels.

Under the supervision of pixel-level labels, we integrate binary cross entropy (BCE) loss $L_{\text{BCE}}$, intersection over union (IoU) [50] loss $L_{\text{IoU}}$, and structural similarity index measure (SSIM) [51] loss $L_{\text{SSIM}}$ as the loss function $L_{\text{Combi}}$ to optimize the model parameters. $L_{\text{Combi}}$ is defined as follows:

$$L_{\text{Combi}} = L_{\text{BCE}} + L_{\text{IoU}} + L_{\text{SSIM}}. \tag{1}$$

BCE loss and IoU loss are commonly used in image classification and image segmentation tasks. The BCE loss measures the difference between the labels predicted by the model and the true labels, thereby encouraging the model to
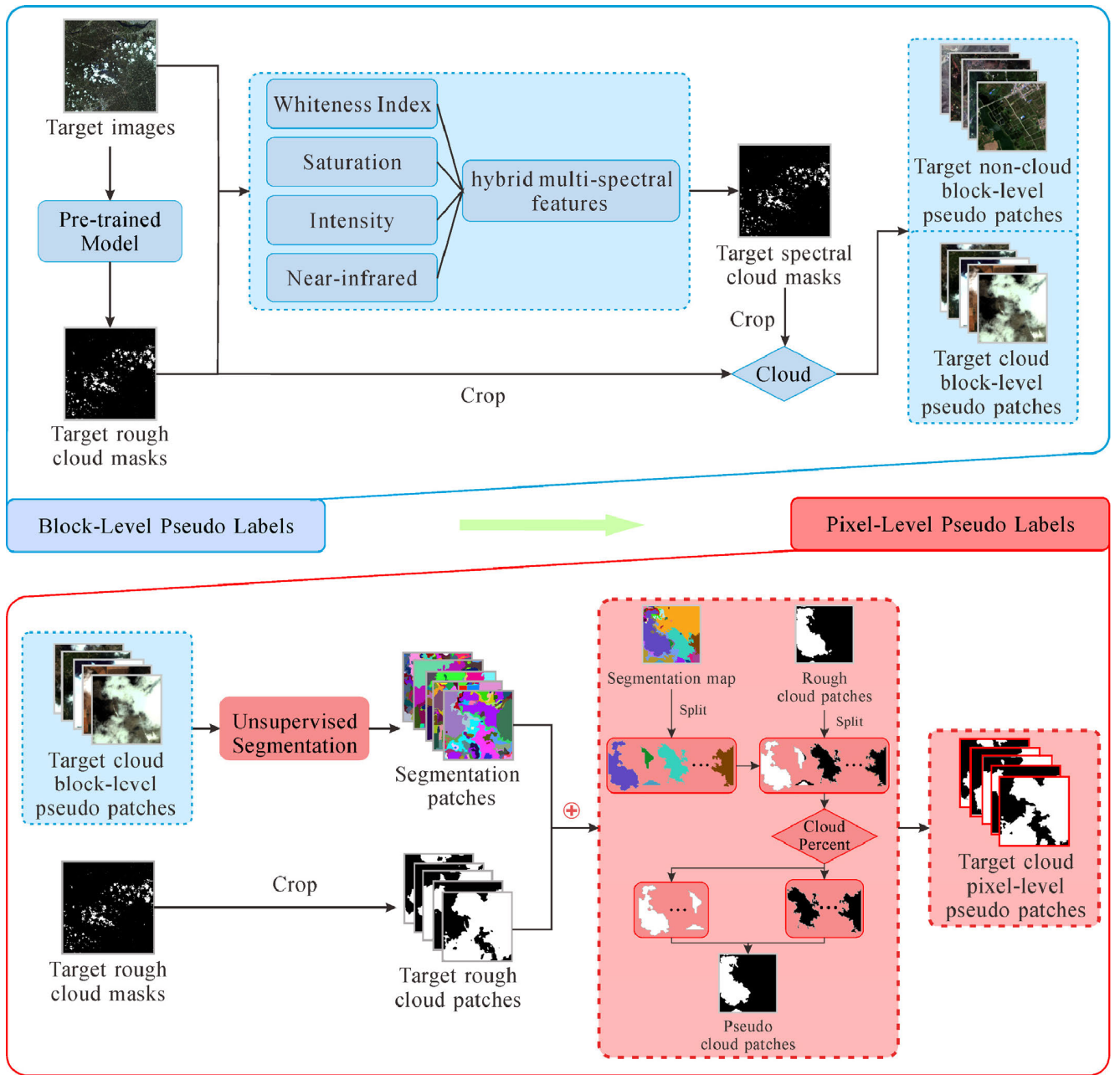
Fig. 3.    Workflow for obtaining the target-domain pseudo labels.

generate predicted labels that closely resemble the true labels. The IoU loss measures the extent of overlap between the predicted labels and the true labels, thereby prompting the model to learn the shape and position information of clouds in the true labels. The BCE loss and IoU loss can be calculated as follows:

$$L_{\text{BCE}} = -\left(y \log\left(\hat{y}\right) + (1-y)\log\left(1-\hat{y}\right)\right)(2) \qquad (2)$$

$$L_{\text{IoU}} = 1 - \frac{y \cdot \hat{y}}{y + \hat{y} - y \cdot \hat{y}} \qquad (3)$$

where $y$ is the true label (1 for cloud pixels and 0 for clear pixels), and $\hat{y}$ is the model-predicted probability of the pixel being a cloud.

SSIM loss is widely applied in image reconstruction, and it measures the degree of similarity between the predicted labels generated by the model and the true labels. In the cloud detection task, SSIM loss allows the model to focus on the variations occurring at the edges of the cloud, consequently enhancing the model's efficacy in detecting cloud edges. The calculation of the SSIM is conducted as follows:

$$L_{\text{SSIM}} = 1 - \frac{\left(2\mu_y\mu_{\hat{y}} + c_1\right)\left(2\sigma_{y\hat{y}} + c_2\right)}{\left(\mu_y^2 + \mu_{\hat{y}}^2 + c_1\right)\left(\sigma_y^2 + \sigma_{\hat{y}}^2 + c_2\right)} \qquad (4)$$

where $\mu$ is the mean value, $\sigma$ is the standard deviation, and $c_1$ and $c_2$ are constants, which default to $0.01^2$ and $0.03^2$.

### B. Obtaining the Pseudo Labels of the Target-Domain Images

Although CNN models demonstrate a certain generalization ability, the existence of diverse imaging conditions across different sensors often leads to radiometric differences and scale
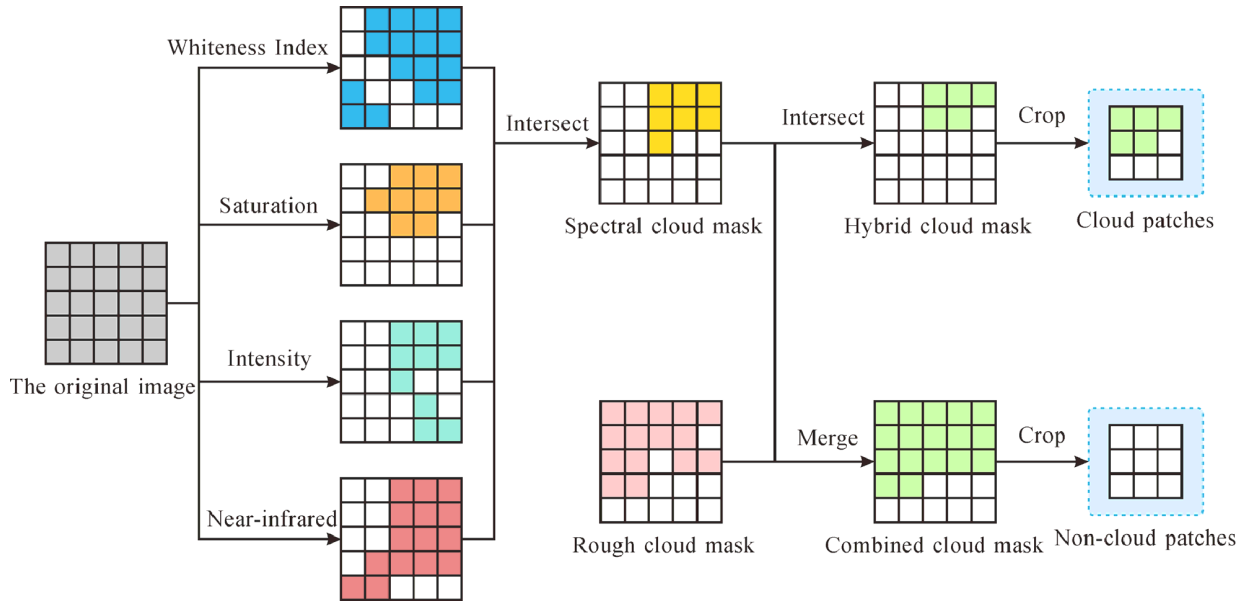
Fig. 4. Workflow for obtaining the target domain block-level pseudo labels.

variations among images. These factors ultimately diminish the superior performance of CNN models in the context of multisensor cloud detection. In order to attain accurate cloud detection results in the $D_T$ images using a deep learning model, we propose a framework for constructing the $D_T$ pseudo-label dataset. This framework combines the physical features of clouds with the generalization ability of the pretrained model, enabling the automatic extraction of highly reliable pseudo labels from the target domain for model transfer. The workflow for obtaining the $D_T$ pseudo labels is shown in Fig. 3, which is divided into two parts: 1) obtaining block-level pseudo labels based on adaptive multispectral (MS) thresholding and 2) obtaining pixel-level pseudo labels based on unsupervised superpixel segmentation.

*1) Obtaining Block-Level Pseudo Labels Based on Adaptive Multispectral Thresholding:* For the $D_T$ image, we use the generalization ability of the pretrained model to obtain the $D_T$ rough cloud mask $M_R$; however, $M_R$ typically exhibits a high rate of leakage recognition and an error recognition rate that hinders its ability to directly construct a pseudo-label dataset for model fine-tuning; thus, $M_R$ is cropped and the mask blocks with high confidence are chosen as the $D_T$ pseudo labels for the purpose of model fine-tuning. At first, we calculate the spectral threshold for the $D_T$ image based on $M_R$. The $D_T$ spectral threshold is obtained by counting the average spectral features of the $D_T$ image across the cloud coverage in $M_R$. In this study, three spectral feature parameters were chosen: 1) the whiteness index (WI) [4]; 2) hue, saturation, intensity (HSI) [52]; and 3) reflectance in the NIR band. Subsequently, we obtain the $D_T$ spectral cloud mask $M_S$ based on the hybrid MS thresholding. The WI uses the absolute difference between each visible band and the overall mean value to effectively capture the distinctive characteristics of clouds. The reflectance characteristics of clouds in the visible bands result in a near-zero value for the WI of cloud pixels; therefore, the WI can be used to identify

cloud pixels. The calculation of the WI is as follows:

$$\text{WI} = \sum_{i=1}^{3} \left| \frac{\text{Band}_i - \text{MeanVis}}{\text{MeanVis}} \right| \tag{5}$$

where $\text{Band}_i$ is the visible band (red band, blue band, and green band), MeanVis is the average of the visible bands. The closer the WI of the pixel is to zero, the higher the probability that the pixel is cloud.

Clouds have the feature of high intensity and low saturation; therefore, the conversion of the image from the red, green, and blue (RGB) color model to the HSI color model is conducted in order to accentuate the cloud regions. The process of converting RGB to HSI is defined as follows:

$$H = \begin{cases} \theta, & \text{Band}_G \geq \text{Band}_B \\ 360 - \theta, & \text{Band}_G < \text{Band}_B \end{cases} \tag{6}$$

$$S = 1 - \frac{3 \times \min(\text{Band}_R, \text{Band}_G, \text{Band}_B)}{\text{Band}_R + \text{Band}_G + \text{Band}_B} \tag{7}$$

$$I = \frac{\text{Band}_R + \text{Band}_G + \text{Band}_B}{3} \tag{8}$$

$$\theta = \cos^{-1} \left\{ \frac{\frac{[(\text{Band}_R - \text{Band}_G) + (\text{Band}_R - \text{Band}_B)]}{2}}{\sqrt{\begin{array}{c}(\text{Band}_R - \text{Band}_G)^2 \\ + (\text{Band}_R - \text{Band}_B)(\text{Band}_G - \text{Band}_B)\end{array}}} \right\} \tag{9}$$

where $\text{Band}_R$, $\text{Band}_G$, and $\text{Band}_B$ are the values of the RGB bands of the image. It can be seen that the cloud region exhibits a low saturation and a high intensity. The S and I components can, therefore, be used to differentiate the cloud region.

For each pixel ($\text{Pixel}_i$) in the image, we employ hybrid MS features for the cloud detection to minimize the cloud recognition error, as shown in Fig. 4. An intersection operation is then performed between $M_R$ and $M_S$ to derive the hybrid cloud mask, which is denoted as $M_H$. Subsequently, $M_R$, $M_S$, and $M_H$ are cropped into blocks. The blocks of $M_H$ with cloud regions exceeding 5% are classified as cloud image blocks so
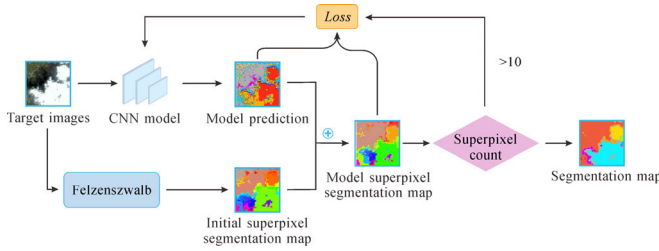
Fig. 5. Flowchart of the unsupervised image segmentation technique.

as to eliminate the misdetection of the pretrained model for parts of the highlighted surface in the urban area. Blocks that do not contain any cloud in both $M_R$ and $M_S$ are classified as noncloud image blocks. The afore-mentioned blocks are used for constructing the $D_T$ block-level pseudo-label dataset.

*2) Obtaining Pixel-Level Pseudo Labels Based on Unsupervised Segmentation:* For noncloud blocks, pixel-level pseudo labels are not required; therefore, in this section, we focus on cloud blocks. Pixel-level pseudo labels are obtained through the optimization of $M_R$ in $D_T$ using the UIS technique. The UIS technique consists of four steps, as shown in Fig. 5: 1) the cloud image blocks are segmented using the Felzenszwalb [53] image superpixel segmentation algorithm in order to generate an initial superpixel segmentation map; 2) the CNN model is applied to the cloud image blocks, in order to generate predicted labels; 3) the predicted label that occurs most frequently in each superpixel block is assigned as the label for that superpixel block; and 4) iteration steps 2 and 3 are used to constantly update the model parameters to merge small regions and obtain the final segmentation map.

### C. Transferring the Deep Model for Cloud Detection in Multisensor Images

In Section II-B, we described how we employ the generalization of the pretrained model and the hybrid spectral features of clouds to effectively identify and select the $D_T$ noncloud pseudo patches and cloud pseudo patches with high confidence. The pixel-level pseudo labels of the cloud patches are acquired using the UIS technique to construct the $D_T$ pseudo-label dataset. Afterward, fine-tuning is conducted on the pretrained model using the $D_T$ pseudo-label dataset to transfer the CNN model from the $D_S$ images to the $D_T$ images. In the process of model fine-tuning, we first mix the $D_S$ samples and the $D_T$ pseudo labels at a ratio of 1:1, and gradually remove the $D_S$ samples during the model iterations. In the last few iterations, the training of the model only relies on the $D_T$ pseudo labels to obtain the final fine-tuned model.

By incorporating the $D_T$ pseudo labels into the fine-tuning process for the pretrained model, it becomes possible to retain the generalized feature information of clouds from the $D_S$ dataset. Simultaneously, the fine-tuned model can also learn the distinctive features of clouds present in the $D_T$ dataset. As a result, the fine-tuned model demonstrates a high accuracy in detecting clouds in the $D_T$ images.

### III. EXPERIMENTAL DATA

To obtain a model with strong generalization ability, we pretrained the CNN model using 108 labeled global GF-1 wide
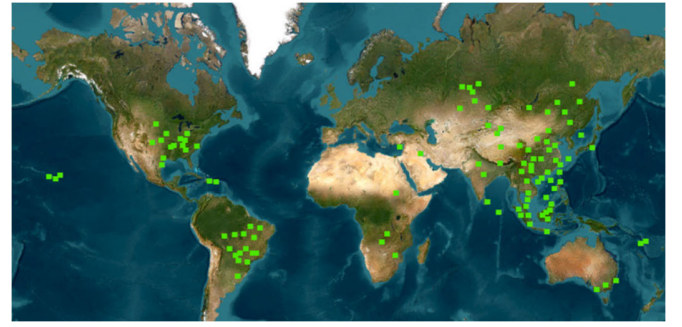


Fig. 6. Distribution of images in the GF1-WHU dataset.

TABLE I
COMPOSITION OF THE GF1-WHU, GF1MS-WHU, GF2MS-WHU, AND WHUS2-CD DATASETS

| Dataset | Image source | Spatial resolution | Labeled images | Unlabeled images |
|---|---|---|---|---|
| GF1-WHU | GF1-WFV | 16-m | 108 | 0 |
| GF1MS-WHU | GF1-MS | 8-m | 33 | 141 |
| GF2MS-WHU | GF2-MS | 4-m | 29 | 163 |
| WHUS2-CD | Sentinel-2 | 10-m | 32 | 0 |

field of view (WFV) (16 m) images covering different land cover types in the GF1-WHU dataset (see Fig. 6 and Table I) [44]. Two datasets were, furthermore, constructed in this article using 8-m GF1-MS and 4-m GF2-MS images, which are referred to as GF1MS-WHU and GF2MS-WHU (see Fig. 7 and Table I), respectively. These datasets were used to assess the transferability of the TransMCD method on multisensor images with different resolutions. In the process of annotating the cloud labels, we referred to the methodologies used in previous studies on cloud detection [1], [44], [54]. Concretely, the Lasso tool in Adobe Photoshop was used to manually mark the locations of clouds. The selected cloud pixel values were assigned the value of 255 to obtain accurate cloud labels as ground-truth samples for the accuracy evaluation. Afterward, we use 32 labeled 10-m Sentinel-2 images in the WHUS2-CD datasets (see Fig. 7 and Table I) [34], [42] to further validate the transferability of the TransMCD method on multisensor images with different resolutions and bandwidth. Note that this article only uses labeled images in the GF1-WHU dataset to pretrain the cloud detection model. The labeled images in the GF1MS-WHU, GF2MS-WHU, and WHUS2-CD datasets were used as the training data for the fully supervised deep learning methods, as well as the validation data for the different methods.

### A. GF1-WHU Dataset

The GF-1 satellite is configured with four WFV and two panchromatic and MS (PMS) sensors. The WFV sensors have four MS bands with a 16-m spatial resolution. The WFV MS bands span the visible to NIR spectral regions. The bandwidth information for each band is as follows: blue
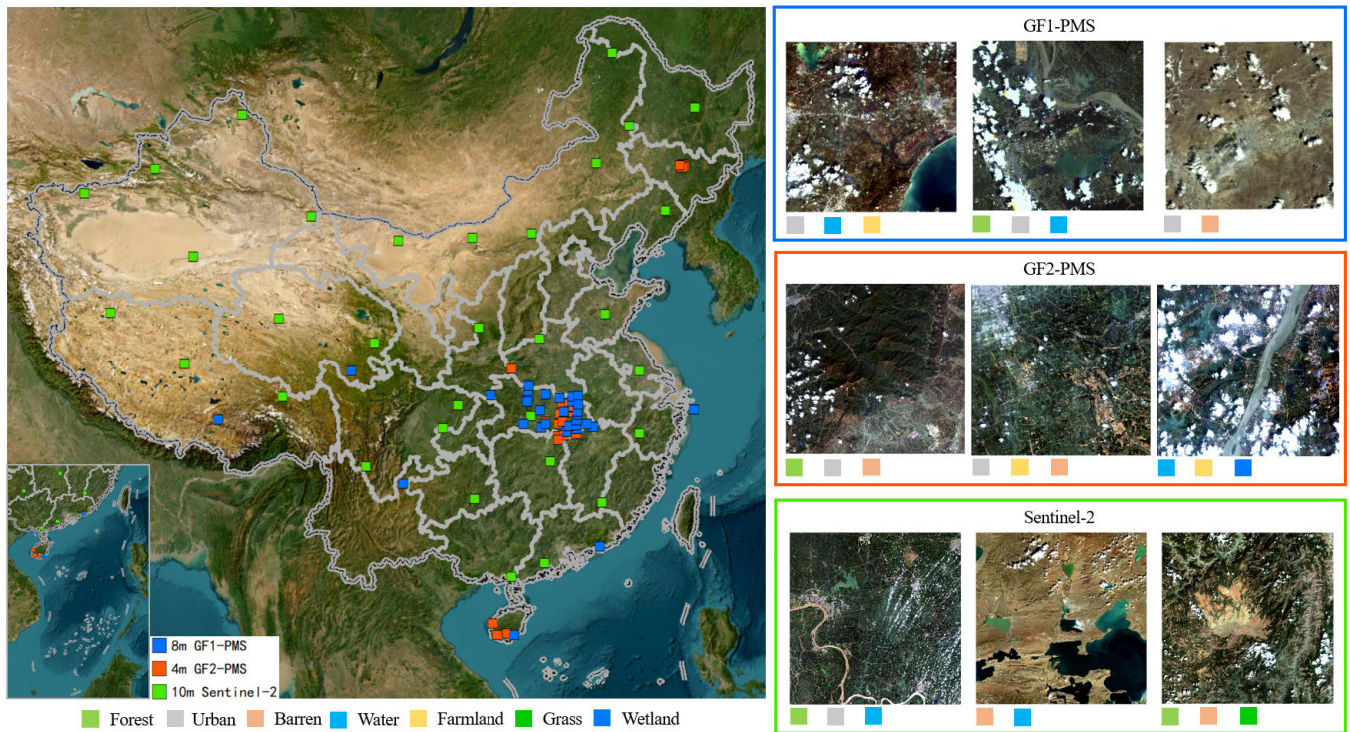
Fig. 7. Distribution of the labeled images in the GF1MS-WHU, GF2MS-WHU, and WHUS2-CD datasets and image examples of the different land-cover types. Note that the color labels at the bottom of each image show the main land-cover type.

(0.45–0.52 $\mu$m), green (0.52–0.59 $\mu$m), red (0.63–0.69 $\mu$m), and NIR (0.77–0.89 $\mu$m). The dataset consists of 108 labeled global GF-1 WFV (16 m) images, which were acquired between May 2013 and August 2016. These images encompass all four WFV sensors and were collected from different land cover types around the world under different cloud conditions. The diversity of images in the dataset has been demonstrated [55], [56] and has been provided online at http://sendimage.whu.edu.cn/en/mfc-validation-data/.

### B. GF1MS-WHU Dataset

The two GF-1 PMS sensors have four MS bands with an 8-m spatial resolution and a panchromatic (PAN) band with a higher spatial resolution of 2 m. The spectral range of the MS bands is identical to that of the WFV sensors. In this study, 141 unlabeled images collected from various regions in China were used as the training data for the proposed method. In addition, 33 labeled images were used as the training data for the fully supervised methods, as well as the validation data for the different methods. The acquisition of the images spanned from June 2014 to December 2020 and encompassed four MS bands in both PMS sensors. Note that Fig. 7 only presents the distribution regions of the labeled images.

### C. GF2MS-WHU Dataset

The GF-2 satellite is configured with two PMS sensors. Each sensor has four MS bands with a 4-m spatial resolution and a PAN band with a 1-m spatial resolution. The GF-2 PMS sensors have the same bandwidth as the GF-1 WFV sensors. In this study, 163 unlabeled images obtained from

Hubei, Jilin, and Hainan provinces were used as the training data for the proposed method, and 29 labeled images were used as the training data for the fully supervised methods, as well as the validation data for the different methods. The images were acquired from June 2014 to October 2020 and included four MS bands in both PMS sensors.

### D. WHUS2-CD Dataset

The Sentinel-2 mission provides images at three different spatial resolutions of 10, 20, and 60-m, which have been widely used for cloud detection. The Sentinel-2 images were acquired by two satellites covering 13 spectral bands. In this study, only the visible and NIR bands of 10-m Sentinel-2 satellite are used for cloud detection. The bandwidth information for visible and NIR bands are as follows: blue (0.458–0.523 $\mu$m), green (0.543–0.578 $\mu$m), red (0.650–0.680 $\mu$m), and NIR (0.785–0.900 $\mu$m). The dataset consists of 32 labeled Sentinel-2 (10-m) images distributed over mainland China and contains a variety of land cover types. The acquisition of the images spanned from April 2018 to May 2020. The dataset has been provided online at https://github.com/Neooolee/WHUS2-CD.

## IV. EXPERIMENTS AND RESULTS

### A. Experimental Setting

We conducted an experiment to evaluate the effectiveness of the TransMCD method with four different deep learning models, namely, ResNet-34 [57], RegNetY-040, DeepLab V3+ [58], and BoundaryNet [59]. We conducted a comparison between the models generated by the TransMCD method and

TABLE II
QUANTITATIVE EVALUATION OF RESNET, REGNETY, DEEPLAB
V3+, AND BOUNDARYNET ON THE GF1-WHU DATASET

| Model | OA | Precision | Recall | F1-score |
|---|---|---|---|---|
| ResNet | 0.9799 | 0.9388 | 0.9124 | 0.9226 |
| RegNetY | 0.9795 | 0.9311 | 0.9161 | 0.9209 |
| DeepLab V3+ | 0.9777 | **0.9480** | 0.8986 | 0.9208 |
| Boundary Net | **0.9800** | 0.9148 | **0.9357** | **0.9235** |

the pretrained models, as well as the models trained using manual labels in the fully supervised mode (ML-FS). The ResNet-34, RegNetY-040, and DeepLab V3+ models were trained with the pretrained weights from ImageNet [60] and the Adam optimizer [61]. The hyper-parameters were configured as follows: $\beta_1$ was set to 0.9, $\beta_2$ was set to 0.999, and $\varepsilon$ was set to $10^{-8}$. The number of epochs was set to 50, and the initial learning rate was set to $10^{-5}$ for the pretraining and $10^{-4}$ for the fine-tuning. The learning rate was decreased by 0.1 at the 15th and 35th epochs. The parameters of BoundaryNet were set according to the original article [59]. We cropped the training data to 250 × 250 and then expanded this to 256 × 256 by padding the image edges with 0 values to minimize the impact of black edges in practical applications. The image enhancement strategy [62] was then used to mitigate the issue of overfitting the model. To standardize the images obtained from the multisensor satellites, a normalization technique was applied where each pixel was divided by the maximum value present in the image. During the pretraining and fine-tuning of all models, the input are the original spectral bands (blue, green, red, and NIR) of the images. During the training of the fully supervised models, we partitioned the GF1MS-WHU and GF2MS-WHU datasets into 70% training data and 30% test data. The division of training data and test data in the GF1-WHU [44] and WHUS2-CD [42] datasets were consistent with the original paper. In the training and testing, all the deep models were trained on a desktop computer (Windows operating system, Intel Core i7-10700 CPU @ 2.90 GHZ, 32 GB RAM, and an NVIDIA GeForce RTX 3070Ti GPU with 8 GB of memory).

### B. Accuracy Evaluation Metrics

To evaluate the performance of the cloud detection models, the experimental results are assessed here using the following metrics: overall accuracy (OA), recall, precision, and F1-score. These metrics can be calculated as follows:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$F1score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{13}$$

where TP represents the number of pixels correctly predicted as cloud, FP represents the number of pixels incorrectly predicted as cloud, TN represents the number of pixels correctly predicted as noncloud, and FN represents the number of pixels incorrectly predicted as noncloud.

### C. Accuracy Assessment of the Pretrained Model for Cloud Detection in Gaofen-1 WFV Images

To obtain a pretrained model with strong generalization ability, we used the GF1-WHU dataset for model pretraining. The quantitative evaluation of the four models on the GF1-WHU dataset is shown in Table II and Fig. 8 shows some examples of the cloud detection results.

From the examples shown in Fig. 8, it can be seen that the results of the four models are visually satisfactory, and the quantitative evaluation results are high, achieving an F1-score of over 0.92. It is evident that all four models demonstrate high accuracy in cloud detection when trained under the fully supervised mode. Upon closer examination of the details, it becomes evident that both ResNet and DeepLab V3+ exhibit notable issues in recognizing cloud boundaries, which leads to significant leakage. BoundaryNet demonstrates a higher accuracy in detecting cloud boundaries, although it does experience some instances of misdetection in local regions. RegNetY achieves balanced results in cloud detection.

### D. Accuracy Assessment of the Transferred Models

First, we performed fine-tuning on the pretrained models using pseudo labels obtained from 141 unlabeled 8-m GF1-MS images in the GF1MS-WHU dataset through the TransMCD method. We used the pretrained model and the model acquired by the ML-FS mode as the comparison models. The quantitative evaluation of the pretrained models, the ML-FS models, and the TransMCD models is shown in Table III. From the results, it is evident that despite the F1-score of the four TransMCD models being lower than that of the ML-FS models, all of the TransMCD models exhibit an improvement of more than 1.2% when compared to the pretrained models. The precision values of the TransMCD-ResNet and TransMCD-BoundaryNet models are 0.8893 and 0.8474, respectively, which are higher values than the pretrained models but significantly lower than the other methods, indicating that the two models may have some misdetection occurrings. The recall of the TransMCD-DeepLab V3+ model is 0.8885, indicating that the model exhibits significant leakage recognition. The TransMCD-RegNetY model demonstrates a balanced performance in cloud detection. For the ResNet, RegNetY, and DeepLab V3+ models, the F1-score of the pretrained and TransMCD models is higher than that of the BoundaryNet model. This indicates that state-of-the-art models may not be more appropriate for the transfer learning task. Although the F1-score of the pretrained BoundaryNet model is relatively low, the optimization of the proposed method results in a 2.81% improvement. This finding suggests that the TransMCD approach is suitable for enhancing the performance of various cloud detection models. The accuracy of the TransMCD models, however, is influenced by the generalization of the pretrained models.
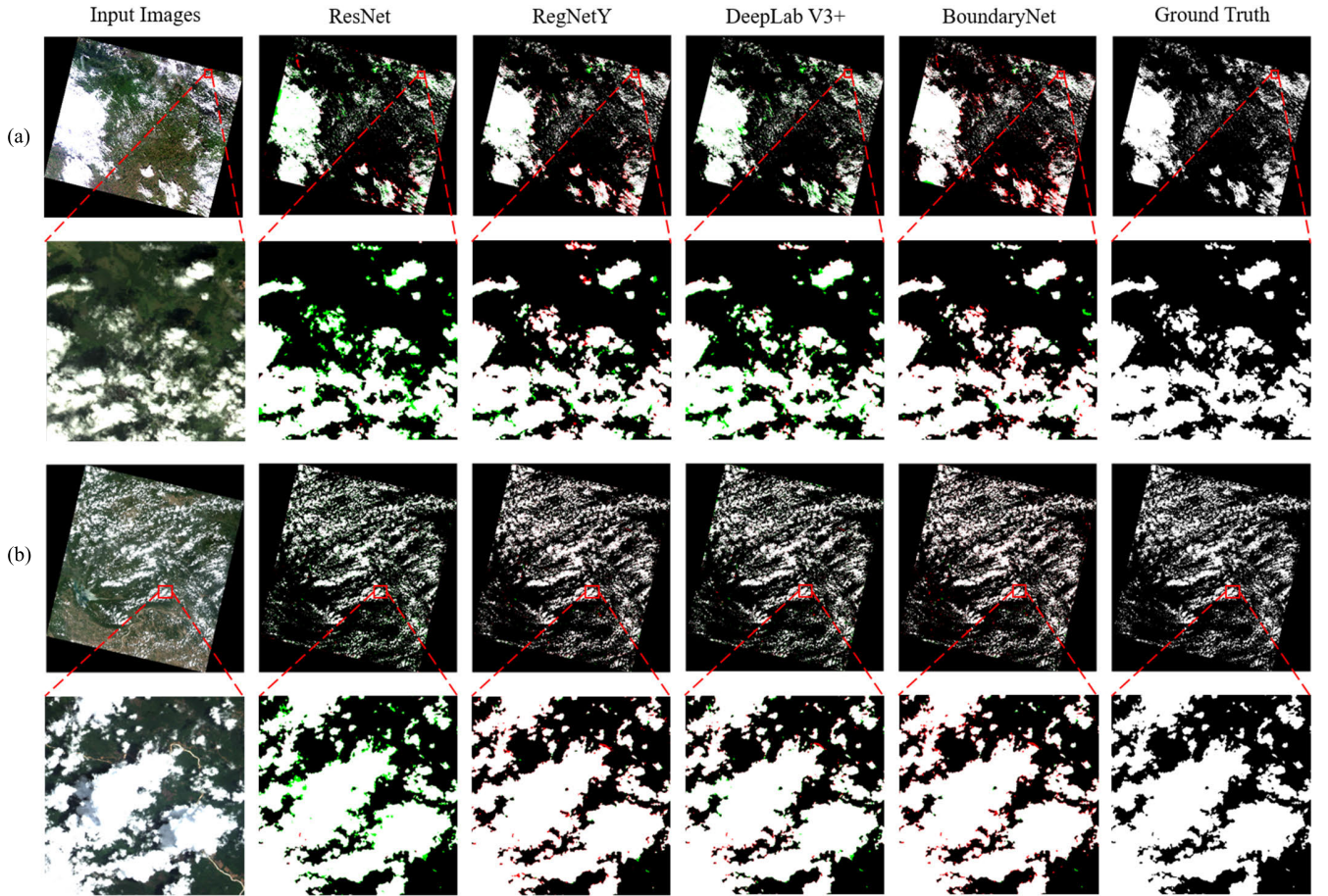
Fig. 8. Examples of cloud detection results for GF1-WFV images obtained by ResNet, RegNetY, DeepLab V3+, and BoundaryNet. (a) Grass. (b) Forest with barren land.

TABLE III
QUANTITATIVE EVALUATION OF RESNET, REGNETY, DEEPLAB V3+, AND BOUNDARYNET ON THE GF1MS-WHU DATASET

| Model | Mode | OA | Precision | Recall | F1-score | Improvement(%) |
|---|---|---|---|---|---|---|
| ResNet (10 test scenes) | Pre-trained | 0.9853 | 0.8691 | 0.9092 | 0.8874 | - |
| | ML-FS | 0.9917 | 0.9383 | 0.9158 | 0.9258 | 3.84 |
| | TransMCD | 0.9894 | 0.8893 | 0.9253 | 0.9057 | 1.83 |
| RegNetY (10 test scenes) | Pre-trained | 0.9869 | 0.8657 | 0.9334 | 0.8973 | - |
| | ML-FS | 0.9919 | 0.9258 | 0.9250 | 0.9242 | 2.69 |
| | TransMCD | **0.9910** | 0.9184 | 0.9195 | **0.9179** | 2.06 |
| DeepLab V3+ (10 test scenes) | Pre-trained | 0.9872 | 0.908 | 0.8842 | 0.8944 | - |
| | ML-FS | 0.9912 | 0.9226 | 0.9234 | 0.9224 | 2.80 |
| | TransMCD | 0.9895 | **0.9275** | 0.8885 | 0.9067 | 1.23 |
| BoundaryNet (10 test scenes) | Pre-trained | 0.9822 | 0.8209 | 0.9028 | 0.8549 | - |
| | ML-FS | 0.9906 | 0.8965 | 0.9452 | 0.9190 | 6.41 |
| | TransMCD | 0.9861 | 0.8474 | **0.9295** | 0.8830 | **2.81** |

Fig. 9 gives cloud detection examples for the pretrained mode, ML-FS mode, and TransMCD mode with the Reg-NetY model. We can see from the results that the pretrained model exhibits notable misdetection. Both the ML-FS model and TransMCD model show satisfactory results overall. The TransMCD model, however, has leakage recognition in the small, thin cloud region. From the details, it can be observed that the pretrained model still results in many misdetections occurring. The ML-FS model and the TransMCD model demonstrate superior performance in thick cloud regions, offering more precise delineation of cloud boundaries; however, the ML-FS model achieves a better cloud detection performance in the thin cloud regions compared to the TransMCD model.

To further validate the transferability of the TransMCD method on images with significant resolution differences,
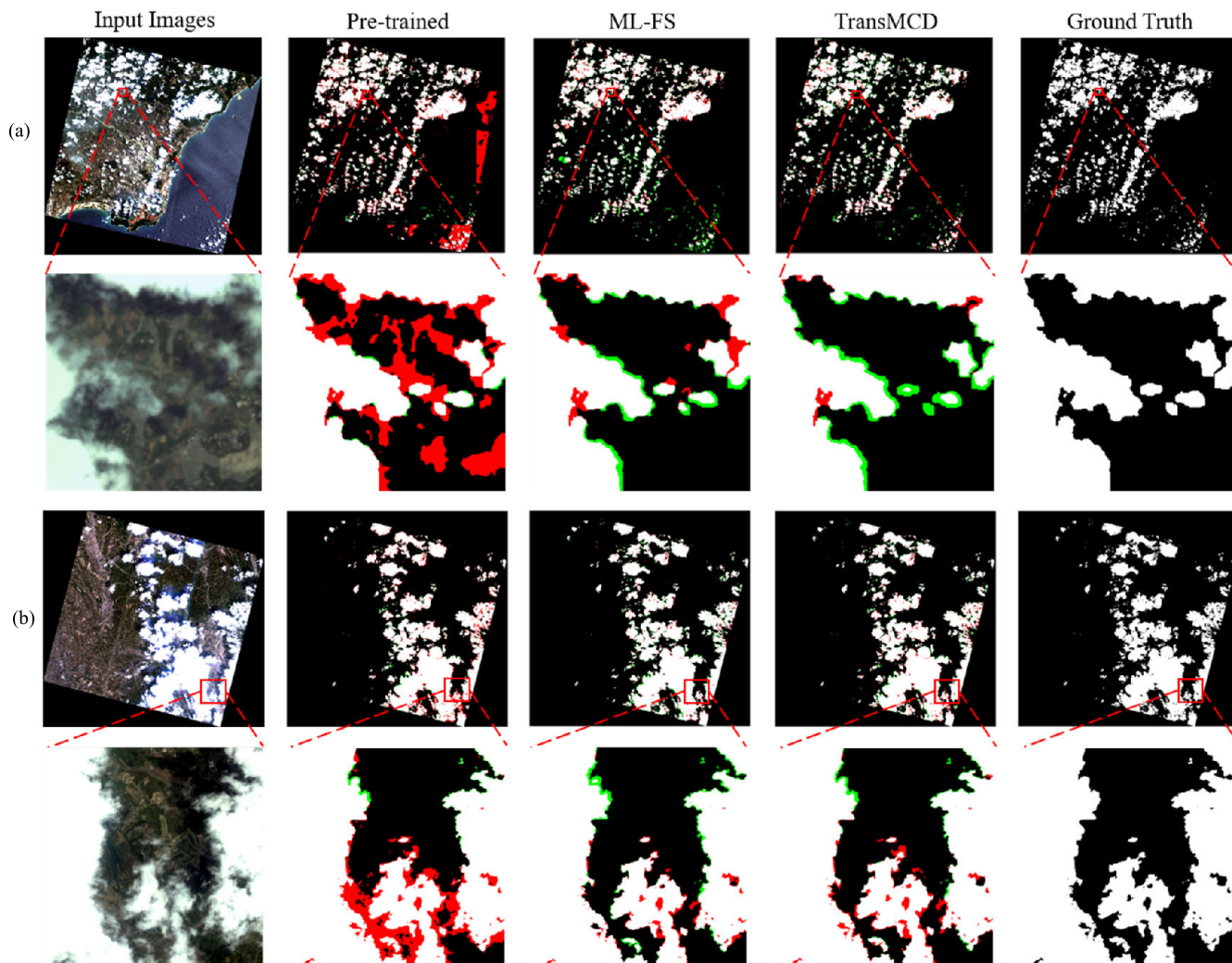
Fig. 9. Examples of cloud detection results for the GF1-MS images obtained by the RegNetY model. (a) Water with farmland. (b) Forest with farmland.

TABLE IV
QUANTITATIVE EVALUATION OF RESNET, REGNETY, DEEPLAB V3+, AND BOUNDARYNET ON THE GF2MS-WHU DATASET

| Model | Mode | OA | Precision | Recall | F1-score | Improvement(%) |
|---|---|---|---|---|---|---|
| ResNet (9 test scenes) | Pre-trained | 0.9437 | 0.9513 | 0.7849 | 0.8569 | - |
| | ML-FS | 0.9619 | 0.9207 | 0.9174 | 0.9179 | 6.10 |
| | TransMCD | 0.9550 | **0.9484** | 0.8454 | 0.8900 | 3.31 |
| RegNetY (9 test scenes) | Pre-trained | 0.9429 | 0.9144 | 0.8380 | 0.8707 | - |
| | ML-FS | 0.9611 | 0.9057 | 0.9408 | 0.9216 | 5.09 |
| | TransMCD | **0.9559** | 0.9122 | **0.9025** | **0.9052** | 3.45 |
| DeepLab V3+ (9 test scenes) | Pre-trained | 0.9429 | 0.9716 | 0.7629 | 0.8469 | - |
| | ML-FS | 0.9603 | 0.9168 | 0.9089 | 0.9134 | 6.65 |
| | TransMCD | 0.9550 | 0.9018 | 0.8861 | 0.8985 | 5.16 |
| BoundaryNet (9 test scenes) | Pre-trained | 0.9331 | 0.9044 | 0.7932 | 0.8421 | - |
| | ML-FS | 0.9542 | 0.8970 | 0.9282 | 0.9103 | 6.82 |
| | TransMCD | 0.9551 | 0.9030 | 0.8942 | 0.8961 | **5.40** |

we used 134 unlabeled 4-m GF2-MS images to fine-tune the pretrained model using the proposed method. The quantitative evaluation results of the pretrained model, the ML-FS model, and the TransMCD model are listed in Table IV. It can be seen that, as the image resolution increases, the task of cloud detection becomes more challenging. In addition, the cloud detection accuracy of the four models in the GF2-MS images can be observed to decrease in comparison to the GF1-MS images. The recall values of the four pretrained models are below 0.84, which indicates that there is a lot
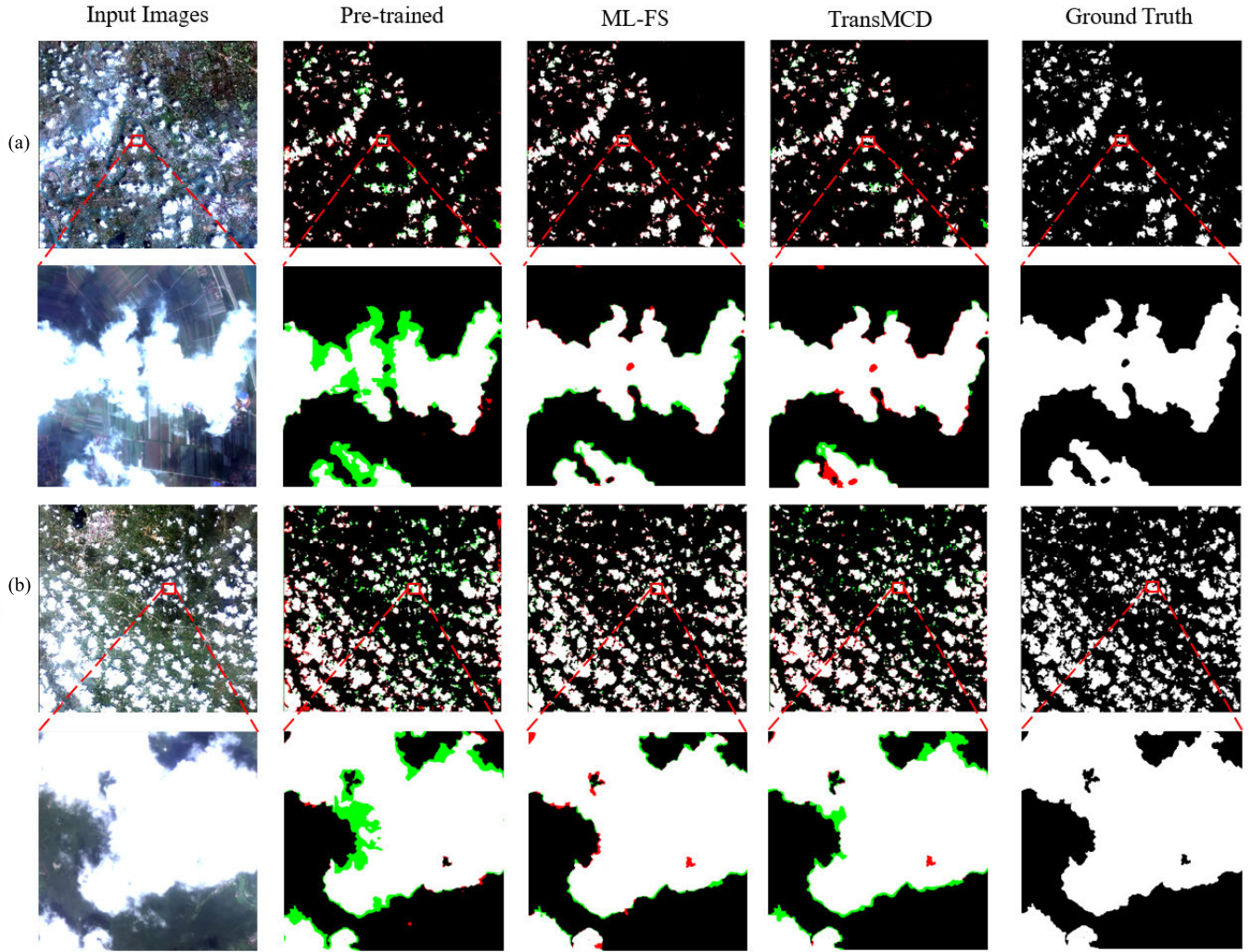
Fig. 10. Examples of cloud detection results for the GF2-MS images obtained by the RegNetY model. (a) Wetland, urban, and farmland. (b) Urban with forest.

TABLE V
QUANTITATIVE EVALUATION OF RESNET, REGNETY, DEEPLAB V3+, AND BOUNDARYNET ON THE WHUS2-CD DATASET

| Model | Mode | OA | Precision | Recall | F1-score | Improvement(%) |
|---|---|---|---|---|---|---|
| ResNet (8 test scenes) | Pre-trained | 0.9758 | 0.9210 | 0.7427 | 0.8183 | - |
| | ML-FS | 0.9874 | 0.9359 | 0.8934 | 0.9135 | 9.52 |
| | TransMCD | 0.9821 | **0.9112** | 0.8473 | 0.8777 | 5.94 |
| RegNetY (8 test scenes) | Pre-trained | 0.9714 | 0.9275 | 0.7589 | 0.8325 | - |
| | ML-FS | 0.9881 | 0.9450 | 0.8870 | 0.9141 | 8.16 |
| | TransMCD | **0.9829** | 0.8942 | **0.8776** | **0.8851** | 5.26 |
| DeepLab V3+ (8 test scenes) | Pre-trained | 0.9700 | 0.9304 | 0.6803 | 0.7711 | - |
| | ML-FS | 0.9855 | 0.9198 | 0.8823 | 0.9002 | 12.91 |
| | TransMCD | 0.9812 | 0.8990 | 0.8390 | 0.8674 | **9.63** |
| BoundaryNet (8 test scenes) | Pre-trained | 0.9739 | 0.8300 | 0.7715 | 0.7870 | - |
| | ML-FS | 0.9872 | 0.9262 | 0.9088 | 0.9170 | 13.00 |
| | TransMCD | 0.9804 | 0.8761 | 0.8748 | 0.8735 | 8.65 |

of leakage recognition. The four ML-FS models achieve a high accuracy, with an F1-score exceeding 0.91. Although the F1-score of the four TransMCD models is lower than that of the ML-FS models, the TransMCD models demonstrate a significant improvement in the F1-score of more than 3.3% when compared to the pretrained models. From the quantitative evaluation results of the GF1MS-WHU and GF2MS-WHU datasets, it can be seen that the pretrained model can obtain relatively satisfactory results on the GF1-PMS image due to the small difference between the 8-m
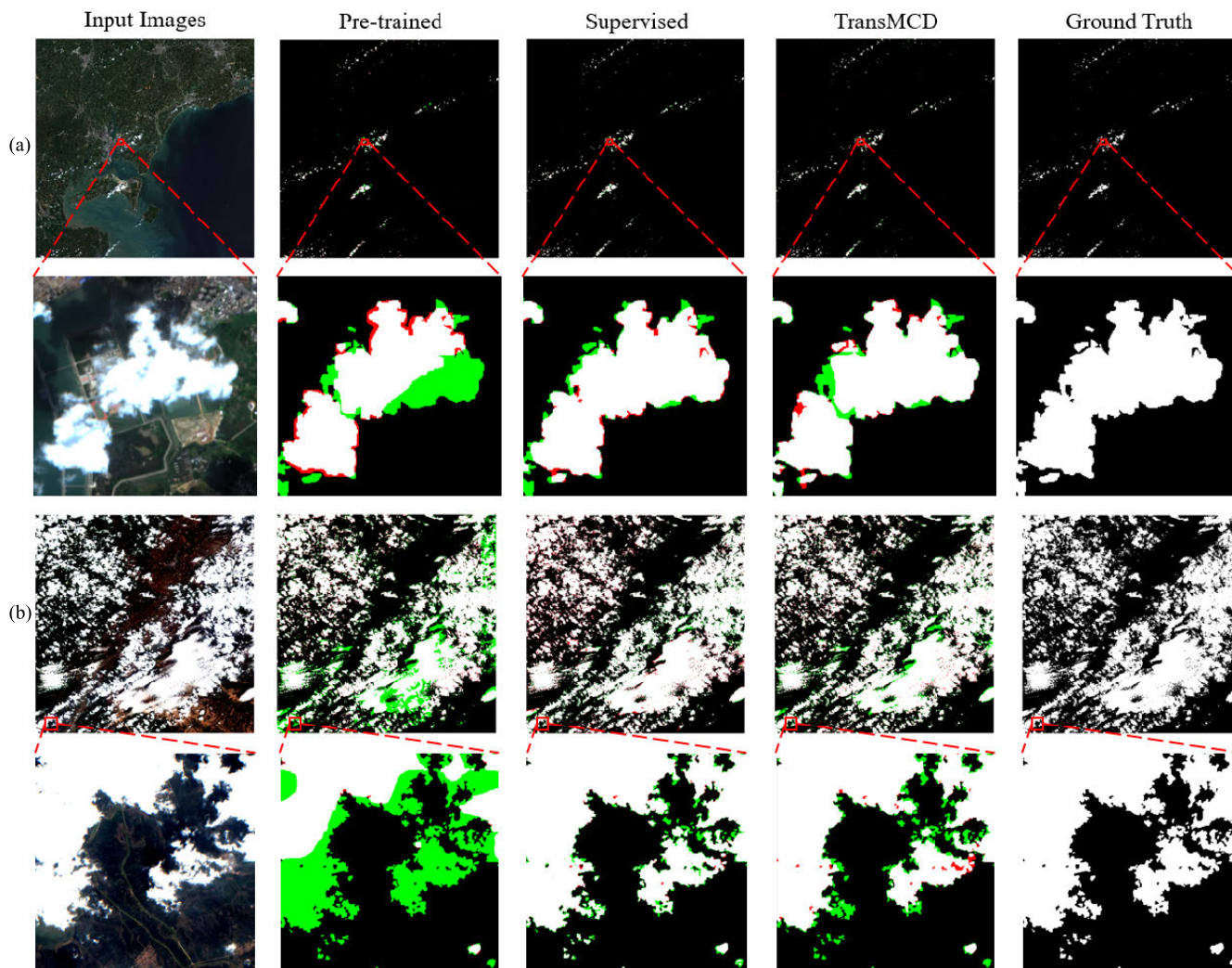
Fig. 11. Examples of cloud detection results for the Sentinel-2 images obtained by the RegNetY model. (a) Water, urban, and farmland. (b) Barren with forest.

TABLE VI

QUANTITATIVE EVALUATION OF THE SPECTRAL CLOUD MASK OBTAINED BY DIFFERENT SPECTRAL FEATURES

| Spectral feature | OA | Precision | Recall | F1-score |
|---|---|---|---|---|
| WI | 0.9492 | 0.9344 | 0.5979 | 0.7216 |
| HSI | 0.9363 | 0.9384 | 0.4681 | 0.6175 |
| NIR | 0.8985 | 0.6780 | 0.5087 | 0.5226 |
| WI-HSI | 0.9352 | 0.9387 | 0.4589 | 0.6077 |
| WI-NIR | 0.9389 | 0.9759 | 0.4835 | 0.6389 |
| HSI-NIR | 0.9361 | 0.9753 | 0.4387 | 0.5996 |
| WI-HSI-NIR | 0.9308 | **0.9776** | 0.4254 | 0.5856 |

TABLE VII

QUANTITATIVE EVALUATION OF THE ROUGH CLOUD MASK, SPECTRAL CLOUD MASK, AND HYBRID CLOUD MASK

| Cloud mask | OA | Precision | Recall | F1-score |
|---|---|---|---|---|
| Rough cloud mask | 0.9728 | 0.8706 | 0.9029 | 0.8816 |
| Spectral cloud mask | 0.9308 | 0.9776 | 0.4254 | 0.5856 |
| Hybrid cloud mask | 0.9307 | **0.9956** | 0.4225 | 0.5862 |

TABLE VIII

SAMPLE NUMBERS OF THE HYBRID CLOUD MASK AND GROUND TRUTH

| Cloud mask | Number of cloudy image blocks | Number of cloud-free image blocks |
|---|---|---|
| Hybrid cloud mask | 6626 | 10746 |
| Ground truth | 11124 | 16770 |

GF1-PMS image and the 16-m GF1-WFV image, which results in the accuracy improvement of the TransMCD method on the GF1MS-WHU dataset not as significant as that on the GF2MS-WHU dataset. It also indicates that the TransMCD method has a greater advantage in transferring images with large resolution disparities.

Fig. 10 presents cloud detection examples for the GF2-MS images obtained using three different methods—pretrained,

ML-FS, and TransMCD—with the RegNetY model. From the visualization results, it can be seen that the pretrained model exhibits a limitation in accurately detecting thin clouds at the

TABLE IX
QUANTITATIVE EVALUATION OF ROUGH CLOUD MASK AND PIXEL-LEVEL PSEUDO LABEL OBTAINED BY REGNETY

| Dataset | Cloud mask | OA | Precision | Recall | F1-score | Improvement (%) |
|---|---|---|---|---|---|---|
| GF1MS-WHU | Rough cloud mask | 0.9326 | 0.8981 | 0.9198 | 0.8990 | - |
| | Pseudo-label | **0.9455** | **0.9070** | **0.9476** | **0.9209** | **2.19** |
| GF2MS-WHU | Rough cloud mask | 0.9108 | 0.8769 | 0.9411 | 0.8978 | - |
| | Pseudo-label | 0.9253 | 0.8952 | 0.9473 | 0.9131 | 1.53 |
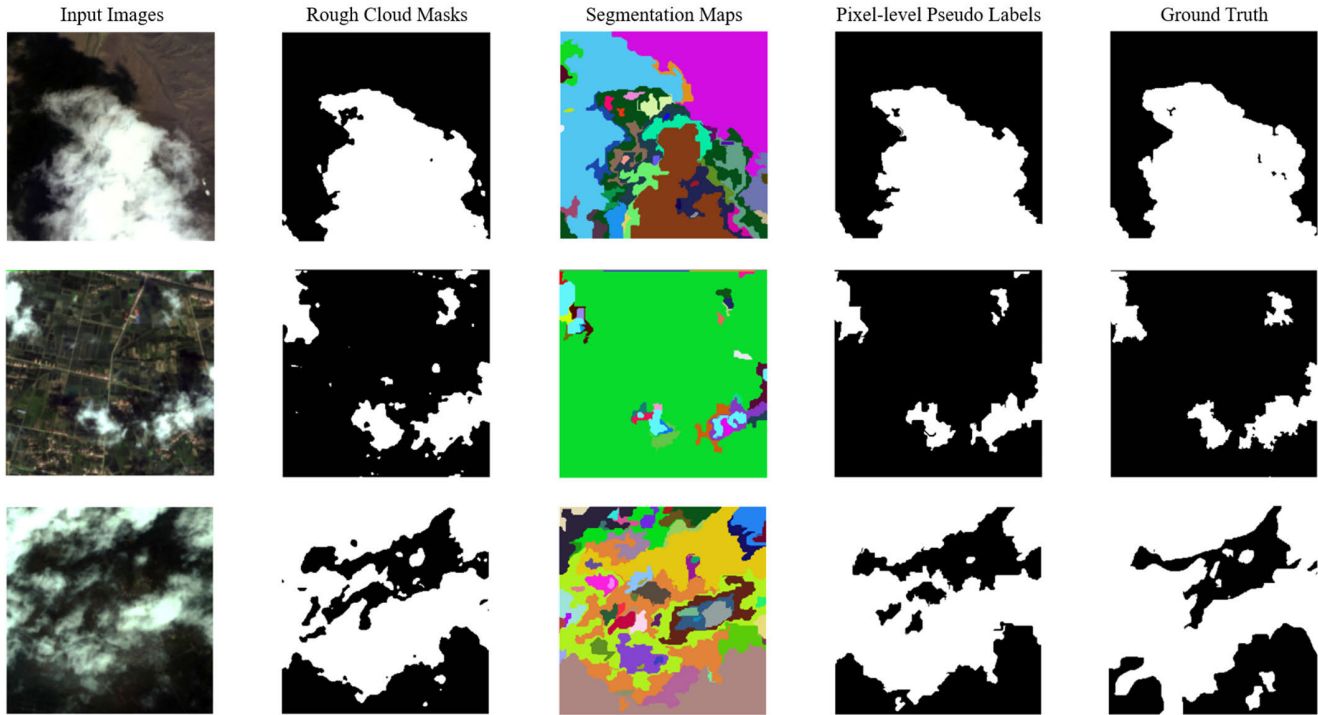


Fig. 12. Examples of the rough cloud masks, segmentation maps, and pixel-level pseudo labels for the GF1-MS images.

cloud boundaries and misdetection occurs in local regions. Compared to the pretrained model, the TransMCD model demonstrates enhanced accuracy in detecting cloud boundaries and effectively mitigates the misdetection; however, the TransMCD model may lose certain detailed information and may fail to detect some small clouds in local regions, leading to less accurate cloud detection results compared to the ML-FS model.

Then, we use 32 Sentinel-2 images in the WHUS2-CD dataset with significant bandwidth differences from the pretraining data to validate the transferability of the TransMCD method. The quantitative evaluation results of the pretrained model, the ML-FS model, and the TransMCD model are listed in Table V. From the results, it can be seen that although the resolution of the Sentinel-2 images is relatively similar to the GF1-WFV images, there are bandwidth differences. It leads to low accuracy of four pretrained models, with F1-Score below 0.84, indicating that the bandwidth difference of the multisensor images has a significant impact on the accuracy of deep models in the transfer task of cloud detection. Although the accuracy of the pretrained models is low, each TransMCD model achieved a high improvement in F1-Score of more than 5.2%. It shows that the TransMCD method still has advantages in transferring images with significant

differences. Fig. 11 presents cloud detection examples for the pretrained mode, ML-FS mode, and TransMCD mode with the RegNetY model. From the results, we can see that exhibits of notable omission occur in most regions. The models fine-tuned by the TransMCD method can mitigate the majority of the omission that occurs, but still have leakage recognition in local thin cloud regions compared to the ML-FS model. From the transfer experiments on three different datasets, it can be seen that the TransMCD method can perform well in the transfer task of multisensor cloud detection. For the either misdetection or omission of the pretrained models in the $D_T$ images, the TransMCD method is effective in inhibiting the problems; however, the accuracy of the cloud detection model fine-tuned by the TransMCD method will be affected by both the transferability of the deep model and the difference between the $D_S$ images and the $D_T$ images.

## V. DISCUSSION

### A. Accuracy Validation of the Produced Pseudo-Label Dataset

In the process of obtaining the $D_T$ block-level pseudo labels, we used a hybrid MS threshold to extract the spectral cloud mask. We first validate the hybrid multispectral features

TABLE X
QUANTITATIVE EVALUATION OF REGNETY IN THE REGION TRANSFERABILITY EXPERIMENT

| Dataset | Mode | OA | Precision | Recall | F1-score | Improvement(%) |
|---------|------|-----|-----------|--------|----------|----------------|
| GF1MS-WHU | Pre-trained | 0.9786 | 0.8632 | 0.9201 | 0.8751 | - |
| | ML-FS | 0.9804 | 0.8898 | 0.8976 | 0.8922 | 1.71 |
| | TransMCD | **0.9827** | **0.9076** | 0.8938 | **0.8999** | 2.48 |
| GF2MS-WHU | Pre-trained | 0.9598 | 0.8917 | 0.8289 | 0.8565 | - |
| | ML-FS | 0.9670 | 0.8195 | **0.9547** | 0.8802 | 2.37 |
| | TransMCD | 0.9661 | 0.8533 | 0.9237 | 0.8873 | **3.08** |

TABLE XI
TIME EFFICIENCY FOR CONSTRUCTING THE CLOUD DETECTION DATASET

| Data type | Image size | Sample generation strategy | Time consumption (min) | Sample number | |
|-----------|------------|----------------------------|------------------------|---------------|---|
| | | | | Cloud | Cloud-free |
| GF1-MS | 4548×4544 | Manual labeling | 500–600 | 175 | 149 |
| | | TransMCD | 6–7 | 100 | 107 |
| GF2-MS | 7300×6908 | Manual labeling | 800–900 | 312 | 444 |
| | | TransMCD | 9–10 | 152 | 300 |

used for constructing the block-level pseudo labels, and the quantitative evaluation results are shown in Table VI. Based on the evaluation results, it is evident that $M_S$ extracted using hybrid multispectral features combining WI, HSI, and NIR reflectance has the highest precision of 0.9776, which can effectively reduce the misdetection occurrings in $M_S$, and fewer misdetections are essential for constructing block-level pseudo-label datasets with high confidence. Subsequently, by conducting an evaluation of $M_R$, $M_S$, and $M_H$, as shown in Table VII, we can see that the precision of $M_H$ is 0.9956, which indicates the absence of misdetection. The results show that for the remaining portion of misdetection in $M_S$, it can be effectively eliminated by intersecting $M_R$ and $M_S$.

We also analyzed the quantity of samples acquired using $M_H$ and compared it to the number of samples selected by the ground truth. The results are listed in Table VIII, which shows that the number of samples selected by $M_H$ is 62.3% of the total samples. The framework for automatically constructing datasets for cloud detection allows for the selection of samples from unlabeled images, thereby enabling the acquisition of the necessary number of samples for model training.

For image blocks with clouds, the UIS technique is employed to enhance the precision of the pseudo labels. Table IX shows the quantitative evaluation of the $M_R$ and the pixel-level pseudo labels for cloudy image blocks obtained by RegNetY. The results show that the pseudo labels optimized by the UIS technique have 2.19% and 1.53% improvement in F1-score in the two datasets compared to the $M_R$, which indicates that the UIS technique is effective in optimizing the rough cloud mask. The examples of $M_R$ optimized based on the UIS technique are shown in Fig. 12. In comparison to $M_R$, it can be seen that the pixel-level pseudo labels exhibit greater precision at the boundaries of clouds, reducing the misdetection and the recognition of leakage within the cloud interior; therefore, the pixel-level pseudo labels acquired through the UIS technique are appropriate for fine-tuning the
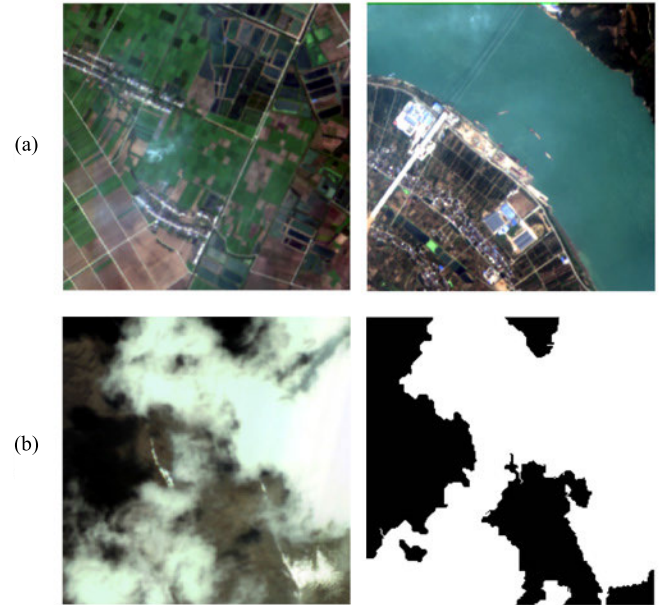


Fig. 13. (a) Examples of the image blocks not selected as a pseudo label. (b) Example of the image block with cloud and snow and its pixel-level pseudo label.

pretrained model, in order to attain satisfactory cloud detection results in $D_T$.

### B. Generalization Ability Analysis for the Transferred Models

To access the effectiveness of the TransMCD method in enhancing the generalization of the models across various regions, we deliberately partitioned the training data and validation data. For the GF1-MS images, 23 images from Hubei province were used to train the fully supervised model, while ten images from other provinces were used as the validation data. For the GF2-MS images, 16 images from Hubei province were used to train the fully supervised model, while 13 images from Jilin and Hainan provinces were used as the validation

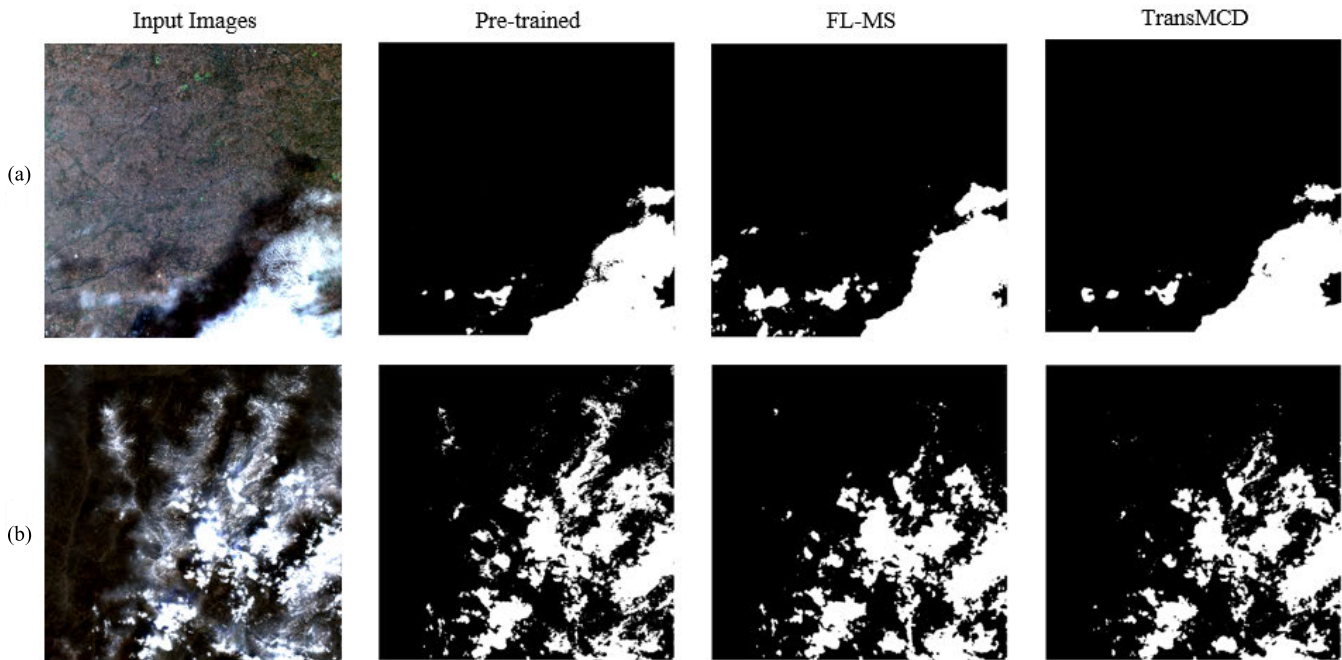| Input Images | Pre-trained | FL-MS | TransMCD |
|---|---|---|---|



Fig. 14. Examples of the cloud detection results for the GF1-MS images obtained using the three different modes of the RegNetY model. (a) Thin cloud area. (b) Snow/ice area.

data. The distribution of the data is shown in Fig. 7, and the quantitative evaluation of the RegNetY model is provided in Table X.

The quantitative evaluation results show that the ML-FS model is incapable of acquiring information about the region where the validation data are situated when the training data and validation data are from separate regions. The TransMCD model has the capability to extract specific information from the validation region, enabling it to achieve a higher F1-score in the region transferability experiment. Overall, our argument is that the TransMCD method can extract valuable information from unlabeled images, thereby enhancing the transferability of the model.

### C. Time Efficiency of the Constructed Dataset

In constructing datasets for cloud detection, the manual annotation of samples often consumes a large amount of time. The proposed framework for automatically constructing datasets for cloud detection is, however, able to generate image blocks for model training within only a few minutes. As shown in Table XI, the process of manually labeling a $4548 \times 4544$ image and dividing it into blocks with a size of $256 \times 256$, which includes 175 cloud blocks and 149 cloud-free blocks, takes approximately 10 h to obtain a well-annotated image. In contrast, the proposed framework can efficiently acquire pixel-level pseudo labels from an unlabeled image in a time span of only 9 min. The proposed framework, in spite of obtaining a smaller number of samples from a single image compared to manual labeling, has the capability to annotate any image and automatically extract samples with high confidence. The proposed framework, therefore, has the capability to acquire an ample amount of training samples in a short period of time when a large number of

images are available. It is apparent that the framework for automatically constructing datasets in the TransMCD method has the potential to significantly reduce the time required to construct a pseudo-label dataset from unlabeled images.

### D. Limitations

The TransMCD method proposed in this article can automatically construct the $D_T$ pseudo-label dataset for fine-tuning the model. This enables the fine-tuned model to achieve a similar effect to the fully supervised model in $D_T$. The deep learning model, however, learns the features in the dataset. When the dataset lacks samples of snow and thin clouds, it cannot provide important prior information for the model to accurately detect thin clouds and distinguish between clouds and snow in the imagery. The TransMCD method uses the model generalization and cloud spectral features to construct a pseudo-label dataset; however, the TransMCD method constructs high-confidence cloud pseudo labels based on the $M_H$. From the quantitative evaluation results in Table VII, we can see that the $M_H$ has a lot of leakage recognition, and its recall is only 0.4225. Many image blocks containing thin cloud features will not be selected as pseudo labels, as the sample shown in Fig. 13(a). The pseudo-labeled dataset constructed in this way often lack image blocks containing thin cloud features, which results in the model obtained by the TransMCD method cannot accurately identify thin clouds in the image, as shown in Fig. 14(a), and the TransMCD method constructed pixel-level pseudo labels based on $M_R$ obtained by the pretrained model. As can be seen in Fig. 13(b), the highlighted surfaces, such as snow, around the cloud are labeled as cloud in the pseudo labels due to the misdetection occurs in the pretrained model; therefore, the TransMCD method is unable to accurately identify snow in the images,

as shown in Fig. 14(b). In summary, we believe that the TransMCD method is suitable for cloud detection in images without snow or large areas of thin clouds.

It is, however, not possible to accurately identify thin clouds and snow in the images using only the cloud spectral features. The capability of the TransMCD model to differentiate between clouds and snow is, therefore, significantly influenced by the generalization of the pretrained model. Although the TransMCD model shows some improvement in identifying thin clouds and reducing the misdetection of snow, compared to the pretrained model, its effectiveness is not as high as that of the ML-FS model. To overcome this shortcoming, we will consider using low-cost manually annotated block-level labels in the target domain that contain information about snow and thin clouds, to fine-tune the pretrained model, and achieve accurate detection of snow and thin clouds.

Because of the significant disparities in the generalization ability of various deep learning models, this article does not test the effectiveness of the TransMCD method in many state-of-the-art models. In future work, we plan to evaluate the effectiveness of the TransMCD method across a broader range of models and present a deep learning model tailored for multisensor cloud detection tasks. The model in this article uses initial weights from the ImageNet dataset, which are suitable for RGB images. In this article, we do not investigate the impact of diverse initial weights on the final model, as the two-stage training approach employed in the TransMCD method reduces the impact of the initial weights on the final model. In the future, we will use initial weights that are more suitable for remote sensing images, such as BigEarthNet [63], to assess the impact of initial weights on the TransMCD method.

## VI. CONCLUSION

In this article, we have proposed the TransMCD method, which is a weakly supervised learning method for transferring deep models to detect clouds in multisensor images. The TransMCD method pretrains a deep learning model using an existing cloud detection dataset and transfers the models to multisensor images, particularly high-resolution images that do not have cloud labels. In addition, a framework is proposed for automatically constructing a cloud detection dataset by combining the physical features of clouds and the model generalization. The framework for constructing pseudo labels uses the model generalization to adaptively determine the spectral threshold for cloud segmentation to obtain high-quality block-level pseudo labels. This eliminates the need for manual adjustment of the spectral threshold in multisensor images and leverages the UIS technology to generate pixel-level pseudo labels, which significantly reduces the burden of manually annotating cloud samples. The results of the transfer experiment conducted on the GF1MS-WHU, GF2MS-WHU, and WHUS2-CD datasets, using four segmentation models, indicated that the models transferred by TransMCD demonstrated a 1.23%–9.63% improvement in F1-score for cloud detection on target domain images, compared to the pretrained models. We argue that the TransMCD

method has good potential in cloud detection for multisensor images and can be used to integrate the state-of-the-art segmentation models. Meanwhile, the TransMCD method still exhibits certain limitations when compared to the fully supervised method using well-annotated labels. For instance, the proposed method lacks the ability to distinguish between clouds and snow, and it also struggles to achieve accurate segmentation of thin clouds. In our future work, we will consider fusing geometric and textural features of clouds in order to distinguish noncloud objects that have similar spectral features to clouds. In addition, we will incorporate low-cost manually annotated block-level labels with thin cloud information into the proposed method to achieve accurate thin cloud detection.

## REFERENCES

[1] R. R. Irish, J. L. Barker, S. N. Goward, and T. Arvidson, "Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 10, pp. 1179–1188, Oct. 2006, doi: 10.14358/pers.72.10.1179.

[2] H. Shen et al., "Missing information reconstruction of remote sensing data: A technical review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 61–85, Sep. 2015, doi: 10.1109/MGRS.2015.2441912.

[3] Z. Zhu and C. E. Woodcock, "Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change," *Remote Sens. Environ.*, vol. 152, pp. 217–234, Jul. 2014, doi: 10.1016/j.rse.2014.06.012.

[4] L. Gomez-Chova, G. Camps-Valls, J. Calpe-Maravilla, L. Guanter, and J. Moreno, "Cloud-screening algorithm for ENVISAT/MERIS multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4105–4118, Dec. 2007, doi: 10.1109/TGRS.2007.905312.

[5] H. Ma, S. Liang, H. Shi, and Y. Zhang, "An optimization approach for estimating multiple land surface and atmospheric variables from the geostationary advanced Himawari imager top-of-atmosphere observations," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 2888–2908, Apr. 2021, doi: 10.1109/TGRS.2020.3007118.

[6] P. Wu et al., "A two-step deep learning framework for mapping gapless all-weather Land Surface Temperature using thermal infrared and passive microwave data," *Remote Sens. Environ.*, vol. 277, Aug. 2022, Art. no. 113070, doi: 10.1016/j.rse.2022.113070.

[7] Z. Li, H. Shen, Q. Weng, Y. Zhang, P. Dou, and L. Zhang, "Cloud and cloud shadow detection for optical satellite imagery: Features, algorithms, validation, and prospects," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 89–108, Jun. 2022, doi: 10.1016/j.isprsjprs.2022.03.020.

[8] Z. Zhu and C. E. Woodcock, "Object-based cloud and cloud shadow detection in Landsat imagery," *Remote Sens. Environ.*, vol. 118, pp. 83–94, Mar. 2012, doi: 10.1016/j.rse.2011.10.028.

[9] Z. Zhu, S. Wang, and C. E. Woodcock, "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.*, vol. 159, pp. 269–277, Mar. 2015, doi: 10.1016/j.rse.2014.12.014.

[10] S. Chen, X. Chen, J. Chen, P. Jia, X. Cao, and C. Liu, "An iterative haze optimized transformation for automatic cloud/haze detection of Landsat imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2682–2694, May 2016, doi: 10.1109/TGRS.2015.2504369.

[11] Y. Zhang, B. Guindon, and X. Li, "A robust approach for object-based detection and radiometric characterization of cloud shadow using haze optimized transformation," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5540–5547, Sep. 2014, doi: 10.1109/TGRS.2013.2290237.

[12] S. Foga et al., "Cloud detection algorithm comparison and validation for operational Landsat data products," *Remote Sens. Environ.*, vol. 194, pp. 379–390, Jun. 2017, doi: 10.1016/j.rse.2017.03.026.

[13] R. Liu and Y. Liu, "Generation of new cloud masks from MODIS land surface reflectance products," *Remote Sens. Environ.*, vol. 133, pp. 21–37, Jun. 2013, doi: 10.1016/j.rse.2013.01.019.

[14] S. Qiu, Z. Zhu, and C. E. Woodcock, "Cirrus clouds that adversely affect Landsat 8 images: What are they and how to detect them?" *Remote Sens. Environ.*, vol. 246, Sep. 2020, Art. no. 111884, doi: 10.1016/j.rse.2020.111884.

[15] Y. Chen, W. He, N. Yokoya, and T.-Z. Huang, "Blind cloud and cloud shadow removal of multitemporal images based on total variation regularized low-rank sparsity decomposition," *ISPRS J. Photogramm. Remote Sens.*, vol. 157, pp. 93–107, Nov. 2019, doi: 10.1016/j.isprsjprs.2019.09.003.

[16] C. Duan, J. Pan, and R. Li, "Thick cloud removal of remote sensing images using temporal smoothness and sparsity regularized tensor optimization," *Remote Sens.*, vol. 12, no. 20, p. 3446, Oct. 2020, doi: 10.3390/rs12203446.

[17] X. Hu, Y. Wang, and J. Shan, "Automatic recognition of cloud images by using visual saliency features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 8, pp. 1760–1764, Aug. 2015, doi: 10.1109/LGRS.2015.2424531.

[18] N. Ghasemian and M. Akhoondzadeh, "Introducing two random forest based methods for cloud detection in remote sensing images," *Adv. Space Res.*, vol. 62, no. 2, pp. 288–303, Jul. 2018, doi: 10.1016/j.asr.2018.04.030.

[19] H. Ishida, Y. Oishi, K. Morita, K. Moriwaki, and T. Y. Nakajima, "Development of a support vector machine based cloud detection method for MODIS with the adjustability to various conditions," *Remote Sens. Environ.*, vol. 205, pp. 390–407, Feb. 2018, doi: 10.1016/j.rse.2017.11.003.

[20] Y. Yuan and X. Hu, "Bag-of-words and object-based classification for cloud extraction from satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 4197–4205, Aug. 2015, doi: 10.1109/JSTARS.2015.2431676.

[21] J. Jang, A. A. Viau, F. Anctil, and E. Bartholomé, "Neural network application for cloud detection in SPOT VEGETATION images," *Int. J. Remote Sens.*, vol. 27, no. 4, pp. 719–736, Feb. 2006, doi: 10.1080/01431160500106892.

[22] C. Poulsen, U. Egede, D. Robbins, B. Sandeford, K. Tazi, and T. Zhu, "Evaluation and comparison of a machine learning cloud identification algorithm for the SLSTR in polar regions," *Remote Sens. Environ.*, vol. 248, Oct. 2020, Art. no. 111999, doi: 10.1016/j.rse.2020.111999.

[23] Z. Li, H. Shen, Y. Wei, Q. Cheng, and Q. Yuan, "Cloud detection by fusing multi-scale convolutional features," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 149–152, Apr. 2018, doi: 10.5194/isprs-annals-iv-3-149-2018.

[24] J. Yang, J. Guo, H. Yue, Z. Liu, H. Hu, and K. Li, "CDNet: CNN-based cloud detection for remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6195–6211, Aug. 2019, doi: 10.1109/TGRS.2019.2904868.

[25] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4062–4076, Jun. 2019, doi: 10.1109/TGRS.2018.2889677.

[26] M. Domnich et al., "KappaMask: AI-based cloudmask processor for Sentinel-2," *Remote Sens.*, vol. 13, no. 20, p. 4100, Oct. 2021, doi: 10.3390/rs13204100.

[27] X. Wu, Z. Shi, and Z. Zou, "A geographic information-driven method and a new large scale dataset for remote sensing cloud/snow detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 174, pp. 87–104, Apr. 2021, doi: 10.1016/j.isprsjprs.2021.01.023.

[28] Y. Wang, L. Gu, X. Li, F. Gao, and T. Jiang, "Coexisting cloud and snow detection based on a hybrid features network applied to remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5405515, doi: 10.1109/TGRS.2023.3299617.

[29] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012, doi: 10.1109/TPAMI.2012.120.

[30] Y. Zi, F. Xie, and Z. Jiang, "A cloud detection method for Landsat 8 images based on PCANet," *Remote Sens.*, vol. 10, no. 6, p. 877, Jun. 2018, doi: 10.3390/rs10060877.

[31] F. Xie, M. Shi, Z. Shi, J. Yin, and D. Zhao, "Multilevel cloud detection in remote sensing images based on deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3631–3640, Aug. 2017.

[32] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5689–5697 doi: 10.1109/ICCV.2017.606.

[33] J. Li et al., "A hybrid generative adversarial network for weakly-supervised cloud detection in multispectral images," *Remote Sens. Environ.*, vol. 280, Oct. 2022, Art. no. 113197, doi: 10.1016/j.rse.2022.113197.

[34] Z. Wu, J. Li, Y. Wang, Z. Hu, and M. Molinier, "Self-attentive generative adversarial network for cloud detection in high resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1792–1796, Oct. 2020, doi: 10.1109/LGRS.2019.2955071.

[35] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112045, doi: 10.1016/j.rse.2020.112045.

[36] J. Zhang, Y. Wang, H. Wang, J. Wu, and Y. Li, "CNN cloud detection algorithm based on channel and spatial attention and probabilistic upsampling for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5404613, doi: 10.1109/TGRS.2021.3105424.

[37] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

[38] G. Mateo-García, V. Laparra, D. López-Puigdollers, and L. Gómez-Chova, "Transferring deep learning models for cloud detection between Landsat-8 and Proba-V," *ISPRS J. Photogramm. Remote Sens.*, vol. 160, pp. 1–17, Feb. 2020, doi: 10.1016/j.isprsjprs.2019.11.024.

[39] G. Mateo-García, V. Laparra, D. López-Puigdollers, and L. Gómez-Chova, "Cross-sensor adversarial domain adaptation of Landsat-8 and Proba-V images for cloud detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 747–761, 2021, doi: 10.1109/JSTARS.2020.3031741.

[40] Z. Zhou et al., "A novel ground-based cloud image segmentation method by using deep transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3072618.

[41] D. Chai, S. Newsam, H. K. Zhang, Y. Qiu, and J. Huang, "Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks," *Remote Sens. Environ.*, vol. 225, pp. 307–316, May 2019, doi: 10.1016/j.rse.2019.03.007.

[42] J. Li et al., "A lightweight deep learning-based cloud detection method for Sentinel-2A imagery fusing multiscale spectral and spatial features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5401219, doi: 10.1109/TGRS.2021.3069641.

[43] C.-C. Liu et al., "Clouds classification from Sentinel-2 imagery with deep residual learning and semantic image segmentation," *Remote Sens.*, vol. 11, no. 2, p. 119, Jan. 2019, doi: 10.3390/rs11020119.

[44] Z. Li, H. Shen, H. Li, G. Xia, P. Gamba, and L. Zhang, "Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery," *Remote Sens. Environ.*, vol. 191, pp. 342–358, Mar. 2017, doi: 10.1016/J.RSE.2017.01.026.

[45] C. Huang et al., "Automated masking of cloud and cloud shadow for forest change analysis using Landsat images," *Int. J. Remote Sens.*, vol. 31, no. 20, pp. 5449–5464, Oct. 2010, doi: 10.1080/01431160903369642.

[46] A. Kanezaki, "Unsupervised image segmentation by backpropagation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1543–1547, doi: 10.1109/ICASSP.2018.8462533.

[47] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, vol. 3, no. 2, p. 896.

[48] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/tkde.2009.191.

[49] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10425–10433.

[50] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 658–666, doi: 10.1109/CVPR.2019.00075.

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[52] E. Welch, R. Moorhead, and J. K. Owens, "Image processing using the HSI color space," in *Proc. IEEE SOUTHEASTCON*, Mar. 1991, pp. 722–725.

[53] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004, doi: 10.1023/b:Visi.0000022288.19776.77.

[54] P. L. Scaramuzza, M. A. Bouchard, and J. L. Dwyer, "Development of the Landsat data continuity mission cloud-cover assessment algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1140–1154, Apr. 2012, doi: 10.1109/TGRS.2011.2164087.

[55] Z. Zhang, S. Tan, and Y. Zhou, "CloudformerV3: Multi-scale adapter and multi-level large window attention for cloud detection," *Appl. Sci.*, vol. 13, no. 23, p. 12857, Nov. 2023, doi: 10.3390/app132312857.

[56] X. Zhang, J. Li, C. Yang, and C. Zhang, "CIFNet: Context information fusion network for cloud and cloud shadow detection in optical remote sensing imagery," *J. Appl. Remote Sens.*, vol. 17, no. 1, Feb. 2023, Art. no. 016506, doi: 10.1117/1.jrs.17.016506.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[58] L. C. E. Chen, Y. K. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, vol. 11211, Sep. 2018, pp. 833–851., doi: 10.1007/978-3-030-01234-2_49.

[59] K. Wu, Z. Xu, X. Lyu, and P. Ren, "Cloud detection with boundary nets," *ISPRS J. Photogramm. Remote Sens.*, vol. 186, pp. 218–231, Apr. 2022, doi: 10.1016/j.isprsjprs.2022.02.010.

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[61] D. Kinga and J. B. Adam, "A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, vol. 5, San Diego, CA, USA, 2015, p. 6.

[62] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[63] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigearthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5901–5904 doi: 10.1109/IGARSS.2019.8900532.

**Zhiwei Li** received the B.Eng. degree in geoinformation science and technology from the China University of Geosciences, Wuhan, China, in 2015, and the Ph.D. degree in cartography and geographic information engineering from Wuhan University, Wuhan, in 2020.

Since July 2022, he has been a Research Assistant Professor with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong. Prior to his current position, he worked as a Postdoctoral Fellow with Wuhan University from July 2020 to June 2022. His research interests include urban remote sensing, cloud detection and removal, multisource data fusion, land cover/use mapping, and flood monitoring. He was or is the Principal Investigator of one project funded by the National Natural Science Foundation of China and two projects funded by the China Postdoctoral Science Foundation.
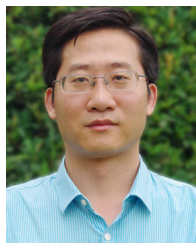
**Huanfeng Shen** (Senior Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

In 2007, he joined the School of Resource and Environmental Sciences (SRES), Wuhan University, where he is currently a Luojia Distinguished Professor and the Associate Dean of SRES. He has authored over 100 research articles in peer-reviewed international journals. His research interests include remote sensing image processing, multisource data fusion, and intelligent environmental sensing. He was or is the Principal Investigator of two projects supported by the National Key Research and Development Program of China and six projects supported by the National Natural Science Foundation of China.

Dr. Shen is a Council Member of China Association of Remote Sensing Application, an Education Committee Member of the Chinese Society for Geodesy Photogrammetry and Cartography, and a Theory Committee Member of the Chinese Society for Geospatial Information Society. He is currently a member of the Editorial Board of the *Journal of Applied Remote Sensing*, the *Journal of Geography*, and the *Journal of Geo-Information Science*.

**Shaocong Zhu** received the bachelor's degree in geographic information science from Wuhan University, Wuhan, China, in 2021, where he is currently pursuing the Ph.D. degree majoring in cartography and geographic information engineering.

His research interests mainly include cloud detection, cloud removal, and mosaic of multisource high-resolution remote-sensing images.