GIScience & Remote Sensing

ISSN: 1548-1603 (Print) 1943-7226 (Online) Journal homepage: www.tandfonline.com/journals/tgrs20

Weakly supervised cloud detection combining spectral features and multi-scale deep network

Shaocong Zhu, Zhiwei Li, Xinghua Li & Huanfeng Shen

To cite this article: Shaocong Zhu, Zhiwei Li, Xinghua Li & Huanfeng Shen (2026) Weakly supervised cloud detection combining spectral features and multi-scale deep network, GIScience & Remote Sensing, 63:1, 2626022, DOI: 10.1080/15481603.2026.2626022

To link to this article: https://doi.org/10.1080/15481603.2026.2626022

Published online: 26 Feb 2026.

GISCIENCE & REMOTE SENSING
2026, VOL. 63, NO. 1, 2626022
https://doi.org/10.1080/15481603.2026.2626022

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Weakly supervised cloud detection combining spectral features and multi-scale deep network

Shaocong Zhu[a], Zhiwei Li[a,b], Xinghua Li[c] and Huanfeng Shen[a,d,e]

[a]School of Resource and Environmental Sciences, Wuhan University, Wuhan, People's Republic of China; [b]Department of Civil, Urban, Earth, and Environmental Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea; [c]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, People's Republic of China; [d]Key Laboratory of Geographic Information System, Ministry of Education, Wuhan, People's Republic of China; [e]Key Laboratory of Digital Cartography and Land Information Application, Ministry of Natural Resources, Wuhan, People's Republic of China

ABSTRACT

Clouds significantly affect the quality of optical satellite images, which seriously limits their precise application. Recently, deep learning has been widely applied to cloud detection and has achieved satisfactory results. However, the lack of distinctive features in thin clouds and the low quality of training samples limit the cloud detection accuracy of deep learning methods, leaving space for further improvements. In this paper, we propose a weakly supervised cloud detection method that combines spectral features and a multi-scale scene-level deep network (SpecMCD) to obtain highly accurate pixel-level cloud masks. The method first utilizes a progressive training framework with a multi-scale scene-level dataset to train the multi-scale scene-level cloud detection network. Pixel-level cloud probability maps are then obtained by combining the multi-scale probability maps and cloud thickness map based on the characteristics of clouds in dense-cloud and large cloud-area images. Finally, adaptive thresholds are generated based on the differentiated regions of the scene-level cloud masks at different scales and combined with distance-weighted optimization to obtain binary cloud masks. Two datasets (i.e. WDCD and GF1MS-WHU) comprising a total of 60 Gaofen-1 multispectral (GF1-MS) images, were used to verify the effectiveness of the proposed method. Compared to the other weakly supervised cloud detection methods such as WDCD and WSFNet, the F1-score of the proposed SpecMCD method shows an improvement of over 7.82%, highlighting the superiority and potential of the SpecMCD method for cloud detection under different cloud coverage conditions.

ARTICLE HISTORY
Received 4 October 2025
Accepted 29 January 2026

KEYWORDS
Cloud detection; weakly supervised learning; spectral feature; Gaofen-1

1. Introduction

High-resolution optical satellite imagery can be affected by varying degrees of clouds, resulting in different cases of surface information loss. Thick clouds result in a complete loss of information in some areas (Shen et al. 2015), while areas covered by thin clouds can suffer from spectral distortion (Wu et al. 2018). Therefore, many cloud detection techniques have been proposed to improve the usability of optical satellite imagery. The main objective of cloud detection is to identify and segment the cloud region in the image and to provide a mask for the subsequent interpretation and analysis of the image. An accurate cloud mask can minimise the impact of clouds on the subsequent applications of the imagery, such as image reconstruction (Zhu et al. 2023; Yun, Jung, and Han 2024) and land-cover mapping (Li et al. 2024).

Cloud detection methods based on multi-temporal images (Zhu et al. 2018; Zhang et al. 2021; Liang et al. 2024; Wang et al. 2024; Lee et al. 2025) achieve cloud detection by detecting the abrupt changes in time-series images, but the requirement for two or more images of different time periods limits the practical application of this approach (Zhai et al. 2018). Cloud detection methods based on single images (Zhu, Wang, and Woodcock 2015; Li et al. 2017;Ishida et al. 2018, Zhu, Li, and Shen 2024) can be categorised into two categories (Wang et al. 2021): 1) physical rule based methods; and 2) machine learning based methods.

CONTACT Huanfeng Shen shenhf@whu.edu.cn

Given the physical characteristics of clouds, such as the high reflectance and white colour (Zhu and Woodcock 2012), the physical rule based methods tend to achieve cloud detection by designing physical rules and segmentation thresholds. The physical rule based cloud detection methods have a robust performance and high efficiency (Sun et al. 2017). However, the selection of physical rules and thresholds needs to be performed manually, which makes it difficult to obtain the optimal parameters (Wang et al. 2024). This can result in detection leakage and misdetection problems, thereby reducing the usability of the cloud masks.

Machine learning based cloud detection methods train the image classification network with a large-scale dataset to obtain highly accurate cloud detection networks with automated segmentation capabilities. Many of the traditional machine learning methods have been applied to cloud detection tasks, including fuzzy clustering (Ping, Su, and Meng 2020), random forest (Fu et al. 2019; Wei et al. 2020), and support vector machine (Joshi, Wynne, and Thomas 2019; Ibrahim et al. 2021). Benefiting from the strong feature representation fitting ability of deep learning, deep learning based cloud detection methods trained with pixel-level labels have been widely applied because of the high accuracy that can be achieved. The pixel-level deep learning methods regard cloud detection as an image segmentation task and generate cloud masks via independent prediction on a pixel-by-pixel basis (Li et al. 2019; Chai et al. 2024; Li et al. 2024; Wright et al. 2024; Gbodjo et al. 2026). Many studies have been conducted to improve the accuracy of cloud detection by designing deep learning networks with different architectures (Chai et al. 2019; Yang et al. 2019; Zhao et al. 2023), introducing multiple image features (Wang et al. 2023; Li et al. 2022a) or incorporating image segmentation techniques (Xie et al. 2017; Zi, Xie, and Jiang 2018).

However, such methods often require a large amount of well-annotated pixel-level labels to achieve accurate cloud detection. To reduce the workload of manually annotated labels, weakly supervised methods have been proposed. The existing weakly supervised cloud detection methods can be categorised into two methods. The first method utilises the physical rules of clouds to generate pixel-level pseudo-labels for training pixel-level deep learning networks, thereby achieving pixel-level cloud detection (Liu et al. 2023; Yang et al. 2024; Zhu, Li, and Shen 2024; Li et al. 2022b). However, the difficulty in defining clear boundaries for thin clouds limits the accuracy of the pixel-level pseudo-labels generated by such methods, thereby compromising thin-cloud detection performance. The second method regards cloud detection as an image classification task. The images to be detected are segmented into separate small scenes, which are then classified as cloudy or cloudless by the scene-level deep learning network trained with scene-level samples. Compared to pixel-level cloud detection networks, scene-level cloud detection networks exhibit superior generalisability and performance (Shendryk et al. 2019), rendering them more suitable for large-area thin-cloud detection tasks. Some studies have employed a class activation map (Fu et al. 2018; Li et al. 2020) or a generative adversarial framework (Li et al. 2022b) to generate pixel-level cloud masks from scene-level cloud detection networks. Nevertheless, such methods tend to detect thick clouds with pronounced spectral signatures, which reduces the scene-level network's capability to identify thin clouds.

Although a large number of cloud detection algorithms have been proposed, the existing methods do have some weaknesses: 1) The accuracy of the deep learning based cloud detection methods relies on large-scale, high-quality training samples. However, as shown in Figure 1, the existing cloud detection datasets (Foga et al. 2017; Li et al. 2017; He et al. 2022; Zhu, Li, and Shen 2024) often do not cover thin clouds and fog, especially foggy thin clouds. This leads to the fact that most of the current cloud detection methods for high-resolution images can only achieve thick-cloud detection. 2) While the existing weakly supervised cloud detection methods incorporating spectral features can achieve accurate thick cloud detection, they often fail for thin clouds, due to their indistinct features. This limitation persists even when spectral features are integrated into the dataset construction or network optimisation. Moreover, although the spectral features (He, Sun, and Tang 2009; Liu et al. 2017) of thin clouds can be leveraged to directly generate relatively high-quality binary masks, the reliance on manual threshold selection and the difficulty in distinguishing clouds from bright surfaces greatly reduces the generality of the methods. 3) Scene-level deep networks are more suitable for large-area thin-cloud detection tasks, and single-size scene-level datasets are easy to construct. However, scene-level cloud detection methods based on single-size samples have difficulty in generating pixel-level cloud masks that cover both thick and thin clouds, along with their boundary details.

To address the challenges of high training data requirements for deep networks, the limited thin-cloud detection capability, and the single-scale nature of scene-level samples, we propose a weakly supervised

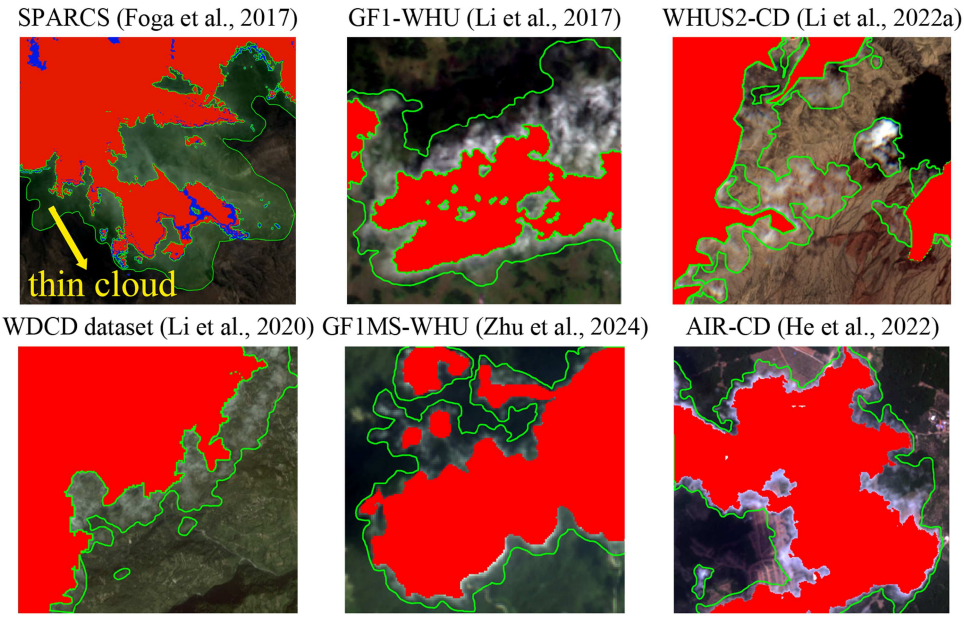

Figure 1. Examples of existing cloud detection datasets. The original thick-cloud labels are shown in red, the original thin-cloud labels are shown in blue, and the green outlines indicate omitted cloud regions.

cloud detection method that combines a cloud thickness map (CTM) and multi-scale scene-level deep network. Specifically, the proposed SpecMCD method constructs multi-scale scene-level labels through sample self-generation and manual supplementation, enabling the training of a multi-scale network to obtain multi-scale cloud probability maps. The cloud probability maps are then combined with the CTM to generate thick and thin-cloud probability maps according to the cloud distribution features across different coverages, followed by fusion based on the CTM gradients. A binary cloud mask is then automatically segmented using adaptive thresholding and distance weighting.

Since the proposed method relies exclusively on scene-level supervision to obtain pixel-level binary cloud masks, it is categorised as a weakly supervised cloud detection method. In contrast to fully supervised approaches that require dense pixel-wise labels, SpecMCD does not use pixel-level annotations for training. Compared with other weakly supervised methods, such as WDCD (Li et al. 2020) and TransMCD (Zhu, Li, and Shen 2024), SpecMCD achieves effective thin cloud and haze detection by embedding the CTM into the inference process, thereby overcoming both the difficulty of detecting thin clouds with ambiguous features and the need for manual threshold selection when incorporating spectral features. In summary, the contributions of this work are as follows:

1) A weakly supervised learning method for cloud detection is proposed (SpecMCD). By employing the multi-scale scene-level network to suppress bright surfaces and integrating the CTM to enhance thin-cloud features, SpecMCD generates cloud probability maps that effectively capture the cloud thickness distribution, thereby enabling high-precision pixel-level binary mask extraction.

2) Distinct probability maps are generated for dense and large-area clouds, fused via the CTM gradient, and refined using adaptive thresholding with distance-weighted optimisation, enabling automatic and accurate cloud detection.

3) To overcome the limitations of a single-scale network, we propose a progressive training framework that integrates multi-scale scene-level samples within a unified network. Furthermore, a local sliding window strategy is adopted to generate multi-scale scene-level cloud probability maps, thereby effectively reducing missed detections.

2. Method

The method proposed in this paper consists of three main steps, as shown in Figure 2: 1) generation of a multi-scale scene-level network and cloud probability maps based on scene-level samples; 2) estimation of

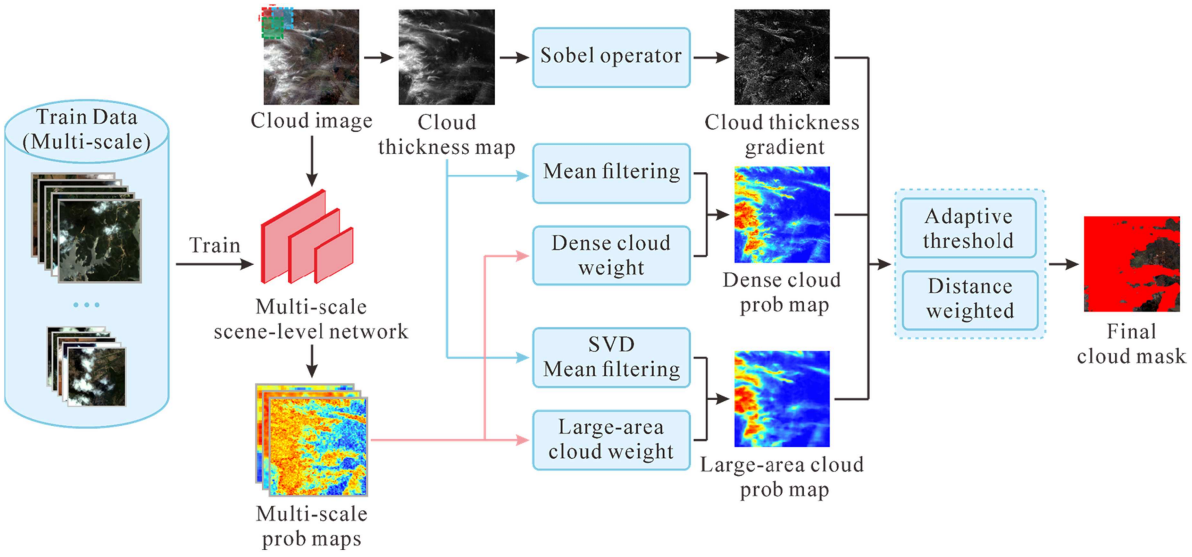

Figure 2. Framework of the proposed weakly supervised cloud detection method.

a CTM via singular value decomposition (SVD); 3) generation of a pixel-level cloud probability map by combining the multi-scale cloud probability maps and the CTM; and 4) extraction of a binary cloud mask using adaptive thresholding combined with distance-weighted optimisation.

2.1. Generation of a multi-scale scene-level network and cloud probability maps based on scene-level samples

Most of the existing scene-level datasets consist of image patches of a single fixed size, which limits the ability of cloud detection networks to accurately detect clouds with varying coverage. Large-scale cloud detection networks are prone to suffering from the loss of usable information, while small-scale cloud detection networks suffer from missed detections. Therefore, in the proposed method, multi-scale scene-level samples are utilised to train a multi-scale scene-level network. The resulting dataset includes images at three resolutions (256×256, 128×128, and 64×64) and comprises thick-cloud, thin-cloud, and cloud-free samples. Thin-cloud samples are obtained by manually outlining rough cloud masks for images containing large areas of thin clouds, whereas thick-cloud samples are generated using the scene-level pseudo-label generation strategy from the TransMCD method (Zhu, Li, and Shen 2024).

The multi-scale scene-level dataset is utilised to train the RegNetY network (Radosavovic et al. 2020), which comprises three main parts, as shown in Figure 3: 1) A stem incorporating a two-stride 3×3 convolutional layer with 32 output channels. 2) A body consisting of multiple downsampled stages. Each stage contains a series of blocks, which are composed of standard residual bottleneck blocks with group convolution (Xie et al. 2017) and a Squeeze-and-Excitation (Hu et al. 2020) attention mechanism. 3) A head component containing average pooling, fully connected, and dropout layers (Cao and Huang 2022). During training, each image block is assigned a single binary label indicating the presence or absence of clouds, and the network output is configured with two channels accordingly.

To fully leverage scene-level samples at different scales and generate multi-scale scene-level cloud probability maps using a single network, a progressive training framework is adopted, as illustrated in Figure 4. Firstly, the network scale is set to 256×256. Each 128×128 sample is replicated into four copies and each 64×64 sample is replicated into 16 copies to resize the multi-scale samples to match the target size. Since small-scale samples contain fewer thin cloud features, making them challenging for effective detection, the network is first trained with large-scale samples, and small-scale samples are gradually incorporated to enhance the network's capability in generating multi-scale scene-level cloud probability maps that capture thin clouds. The multi-scale probability maps are then generated using a local sliding window strategy (Li et al. 2020) to reduce missed detections. Window sizes are set to 256×256, 128×128,

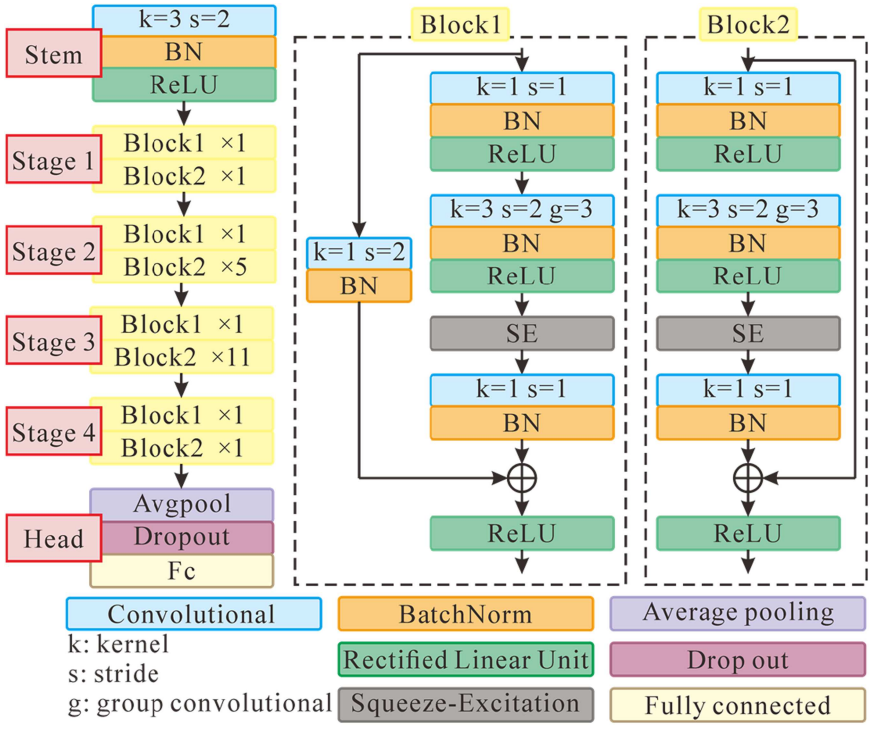

Figure 3. Structure of the RegNetY-040 network.

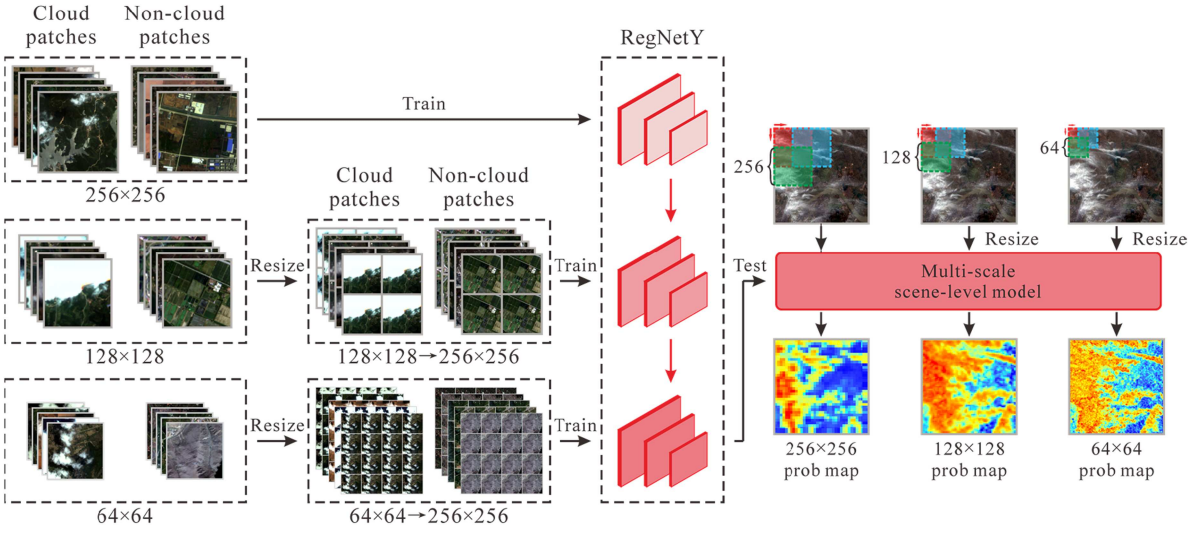

Figure 4. The progressive training framework for generating a multi-scale scene-level network and cloud probability maps. During training, image patches of different sizes are resized to 256 × 256 and progressively fed into the network, while during testing, images are cropped into patches using a sliding window and input into the network to generate multi-scale cloud probability maps.

and 64 × 64 according to the training sample sizes. Overlapping image blocks are generated by sliding the window from left to right and top to bottom with a step size of half the window size. All blocks are resized to 256 × 256 using the same progressive resizing strategy. The scene-level network classifies each block individually, and the predicted probability is assigned to the corresponding block. For overlapping regions, the maximum probability among overlapping blocks is taken as the final value to achieve block-level fusion.

Finally, a scene-level binary cloud mask is generated by applying an initial threshold of 0, with all regions exceeding this threshold classified as cloud-covered.

2.2. Generation of a CTM based on singular value decomposition

Although the blue band is most strongly affected by thin clouds, previous studies (Makarau et al. 2014; Liu et al. 2017) based on linear extrapolation have demonstrated that a synthetic band constructed from the blue and green bands exhibits a superior performance in image dehazing. Therefore, in the proposed method, this synthetic band is adopted to estimate an initialised CTM (Liu et al. 2017) capturing the spectral features of clouds. The initialised CTM is calculated as follows:

$$CTM = 2 * B - 0.95 * G \tag{1}$$

where B and G are the values of the image blue band and green band, respectively.

However, the initialised CTM obtained by the B and G bands is affected by highlighted surfaces, resulting in both cloud-covered areas and bright surfaces appearing as high-intensity regions. Therefore, regions with CTM values exceeding the median are identified as highlighted surfaces. Highlighted cloud-free regions are then obtained by comparing these surfaces with the intersections of the multi-scale scene-level cloud masks. These regions are subsequently refined by reducing the CTM values in the highlighted cloud-free regions to half of their original value, thereby mitigating the impact of bright surfaces on the initialised CTM.

Although the initialised CTM effectively highlights the spectral features of clouds, the distribution characteristics differ considerably between dense-cloud and large-area cloud images. Clouds in dense-cloud images are characterised by dispersed and irregular arrangements, while clouds in large-area cloud images are characterised by wide coverage, and the cloud boundaries are difficult to identify. Therefore, differentiated CTM optimisation strategies are applied to enhance the spectral features of both dense and large-area clouds.

For dense-cloud images, the CTM is smoothed using mean filtering to suppress noise. For large-area cloud images, where local details may reduce the global cloud features, the CTM is decomposed into two orthogonal matrices and one diagonal matrix by SVD. A low-rank approximation is then obtained by retaining the first 70 singular values, as described in Equation (2):

$$CTM_{SVD} = U_k \Sigma_k V_k^T \tag{2}$$

where CTM_{SVD} is the low-rank approximation of the CTM; U_k is the first k columns of the left singular value matrix; Σ_k is the first k singular values; and V_k^T is the transpose of the first k rows of the right singular matrix, for which k defaults to 70.

2.3. Generation pixel-level cloud probability maps by combining the CTM and multi-scale cloud probability maps

Scene-level networks are constrained by the scene size and often misclassify the cloud-free regions between dispersed cloud blocks as cloudy, resulting in a significant loss of usable information in dense cloud images. Furthermore, the indistinct spectral features of thin clouds hinder deep learning networks from achieving accurate detection, leading to a large number of omissions in large-area cloud images. To address these issues, the proposed method integrates multi-scale cloud probability maps with the CTM to generate pixel-level cloud probability maps based on the characteristics of both dense and large-area cloud images.

2.4. Generation of a large-area cloud probability map

To improve the ability of the proposed method to detect thin clouds in large-area cloud images, a large-area cloud-weighted probability map is first constructed by aggregating the multi-scale cloud probability maps using large-area cloud weight coefficients. This weighted probability map is then multiplied by the

low-rank CTM to obtain the large-area cloud probability map. The probability of each pixel in the large-area cloud probability map is calculated as follows:

$$\rho_{Large\,(i,j)} = (\mu_1 \cdot \rho_{256\,(i,j)} + \mu_2 \cdot \rho_{128\,(i,j)} + \mu_3 \cdot \rho_{64\,(i,j)}) \cdot CTM_{SVD\,(i,j)} \tag{3}$$

where $\rho_{Large\,(i,j)}$ is the large-area cloud probability of a pixel at row i, column j of the image; ρ_{256}, ρ_{128}, and ρ_{64} are the normalised cloud probabilities obtained by the three different-scale scene-level networks of 256×256, 128×128, and 64×64, respectively; CTM_{SVD} is the normalised low-rank CTM; and μ_1, μ_2, and μ_3 are constants, which default to 0.5, 0.4, and 0.1, respectively.

2.5. Generation of a dense-cloud probability map

To address the tendency of scene-level networks to misdetect in dense cloud images, the proposed method aggregates the multi-scale cloud probability maps using dense cloud weight coefficients to obtain a dense cloud weighted probability map. This weighted map is then multiplied by the smoothed CTM to generate the dense cloud probability map. The probability of each pixel in the dense cloud probability map is calculated as follows:

$$\rho_{Dense\,(i,j)} = (\mu_3 \cdot \rho_{256\,(i,j)} + \mu_2 \cdot \rho_{128\,(i,j)} + \mu_1 \cdot \rho_{64\,(i,j)}) \cdot CTM_{Mean\,(i,j)} \tag{4}$$

where $\rho_{Dense\,(i,j)}$ is the dense cloud probability of the pixel at row i, column j of the image; CTM_{Mean} is the normalised CTM smoothed by mean filtering; and the mean filtering window size is set to 29.

2.6. Fusion of the dense-cloud and large-area cloud probability maps

Since the proposed method generates two different types of cloud probability maps, manually determining the cloud type for each image remains time-consuming and labour-intensive. Thick clouds in the imagery typically have distinct boundaries, whereas thin clouds lack clear boundaries. Correspondingly, the CTM gradient is higher at thick cloud boundaries and lower at thin-cloud boundaries, as illustrated in Figure 5. Leveraging these characteristics, we propose a fusion strategy for dense and large-area cloud probability maps based on the CTM gradient, as illustrated in Figure 6. Specifically, the Sobel operator is applied to compute the CTM gradient and generate the binary gradient boundary mask as follows:

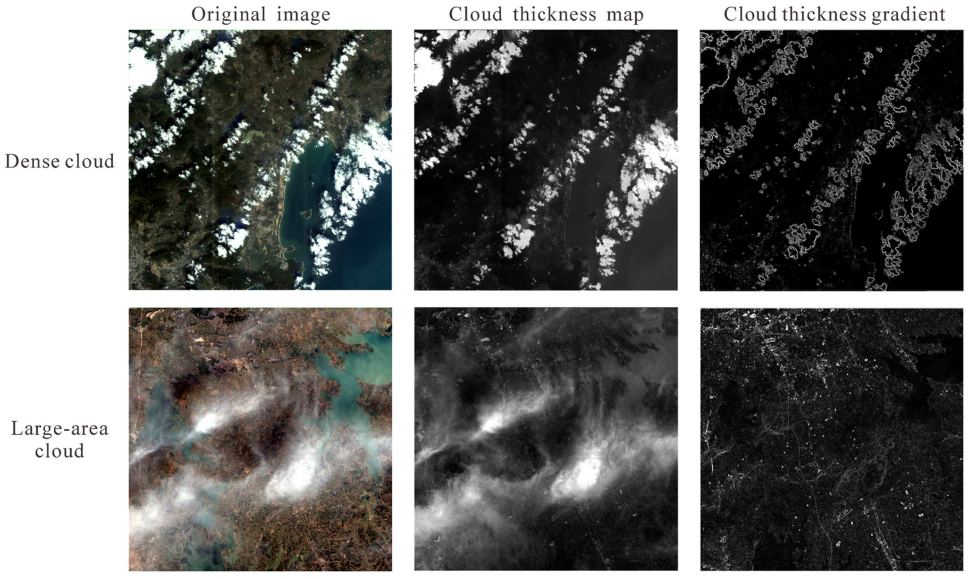

Figure 5. Examples of cloud thickness maps and cloud thickness gradients for different types of images.

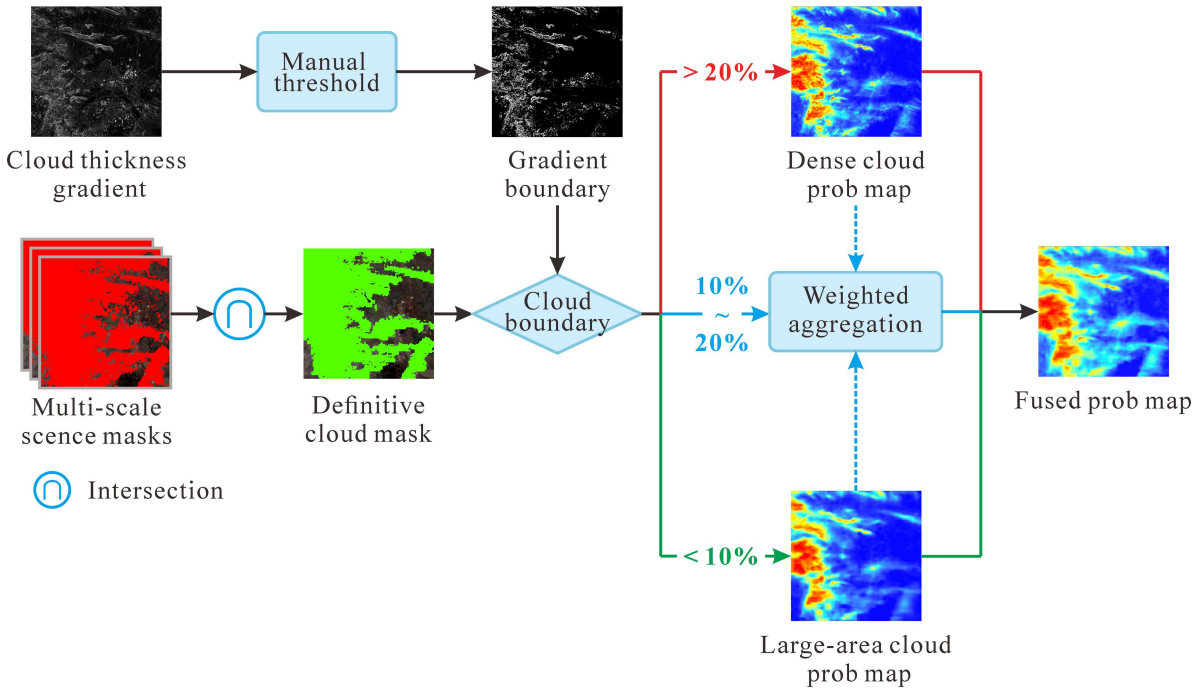

Figure 6. Flowchart for the fusing of dense-cloud and large-area cloud probability maps.

$$M_{Bound}(i,j) = \begin{cases} 1, & if\ Grad(i,j) > \mu_{Grad} \\ 0, & else \end{cases} \tag{5}$$

where M_{Bound} is the binary gradient boundary mask; $Grad(i,j)$ is the normalised CTM gradient of the pixel at row i, column j of the image; and μ_{Grad} is a constant, which defaults to 19.

By calculating the proportion of M_{Bound} within the intersections of the multi-scale cloud masks, the dense-cloud and large-area cloud probability maps are fused as follows:

$$\rho_{Fused(i,j)} = \begin{cases} \rho_{Dense(i,j)}, & if\ \ P \geq \mu_1 \\ k \cdot \rho_{Dense(i,j)} + (1-k) \cdot \rho_{Large(i,j)}, & if\ \mu_2 < P < \mu_1 \\ \rho_{Large(i,j)}, & if\ \ P \leq \mu_2 \end{cases} \tag{6}$$

where $\rho_{Fused(i,j)}$ is the fused cloud probability of the pixel at row i, column j of the image; P is the percentage of the M_{Bound} coverage area within M_{Cloud}; k is the fusion constant, calculated as $(P-\mu_2)/(\mu_1-\mu_2)$; and μ_1 and μ_2 are empirical constants, which default to 0.2 and 0.1, respectively.

2.7. Generation of binary cloud masks by combining adaptive thresholding with distance-weighted optimisation

To eliminate the need for manual threshold adjustment and enhance the robustness of the method, adaptive segmentation thresholds are automatically determined by analysing the differences in regions detected by the scene-level network across scales. As illustrated in Figure 7, the 256×256 scene-level cloud masks generally capture both thick and thin-cloud regions completely, whereas the 64×64 masks may partially miss thin-cloud areas. Consequently, the differences between these masks predominantly correspond to the undetected thin-cloud regions. Based on these differences, the thresholds for the different images are calculated as follows:

$$\mu_{Final} = \begin{cases} \mu_{Dense}, & if\ \ P \geq \mu_1 \\ k \cdot \mu_{Dense} + (1-k) \cdot \mu_{Large}, & if\ \ \mu_2 < P < \mu_1 \\ \mu_{Large}, & if\ \ P \leq \mu_2 \end{cases} \tag{7}$$

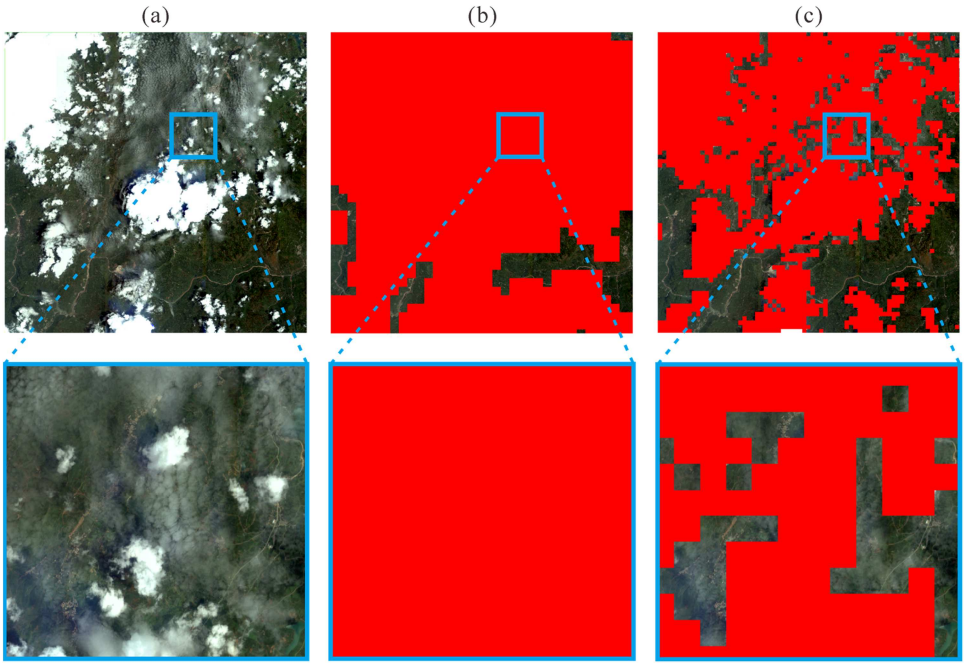

Figure 7. Examples of regional differences in cloud masks obtained by the scene-level network at different scales. (a) Original image; (b) 256×256 scene-level cloud mask; (c) 64×64 scene-level cloud mask.

where μ_{Final} denotes the adaptive threshold; μ_{Dense} is the dense-cloud threshold, obtained by averaging the probabilities within the region covered by the 256×256 scene-level cloud mask; and μ_{Large} represents the large-area cloud threshold, obtained by averaging the probabilities of the regions differing between the 256×256 and 64×64 scene-level cloud masks.

Based on μ_{Final}, the fused cloud probability map is segmented and the initialised binary cloud mask M_{Init} is generated as follows:

$$M_{Init\,(i,j)} = \begin{cases} 1, & if \quad \rho_{Fused\,(i,j)} > \mu_{Final} \\ 0, & else \end{cases} \tag{8}$$

M_{Init} is derived from μ_{Final} rather than manual thresholding, making it difficult to accurately detect thin clouds. Therefore, an adaptive expansion strategy is introduced by augmenting ρ_{Fused} around M_{Init} through distance-weighted optimisation to enhance the thin-cloud detection capability of the SpecMCD method, as shown in (9). Finally, the final binary cloud mask M_{Final} is obtained by segmenting the distance-weighted cloud probability map using μ_{Final}.

$$\rho_{Dist\,(i,j)} = \rho_{Fused\,(i,j)} + \frac{DistMax - Dist}{DistMax} \cdot \rho_{Mean} \tag{9}$$

where ρ_{Dist} is the cloud probability map after the distance-weighted optimisation; $DistMax$ is the adaptive distance constant, decreasing from 100 to 50, depending on P, calculated as $(150 - P * 500)$; $Dist$ is the distance from the pixel at row i, column j to the nearest cloud block; and ρ_{Mean} is the compensation probability, which is obtained by averaging the $\rho_{Fused\,(i,j)}$ of the region covered by the 128×128 scene-level cloud mask.

2.8. Experimental data and results

2.8.1. Experimental data and settings

Since a single-scale scene-level network cannot adequately capture cloud coverage across different distribution patterns, it is necessary to introduce multi-scale scene-level labels to train a multi-scale network

for cloud detection. In this study, 101 cloud images and 114 cloud-free images were used to generate multi-scale thick-cloud and cloud-free scene-level samples using the scene-level pseudo-label generation strategy from the TransMCD method (Zhu, Li, and Shen 2024). To address the difficulty of automatically generating thin-cloud samples, 10 images with large-area thin clouds were roughly annotated to provide approximate thin-cloud regions, which were used exclusively for constructing thin-cloud scene-level samples, as shown in Figure 8. It should be emphasised that the manual rough annotation only aims to ensure that the annotated regions are cloud-covered areas and is utilised solely to supplement the thin-cloud scene-level samples. No pixel-level annotations are used for model training. Based on these data, a multi-scale scene-level training dataset was constructed, comprising 21 699 image blocks of size 256×256, 84 845 image blocks of size 128×128, and 332 251 image blocks of size 64×64, for training the multiple scene-level network, as shown in Table 1. The bandwidth information for each band in the visible spectrum band is as follows: blue (0.45–0.52 μm), green (0.52–0.59 μm), and red (0.63–0.69 μm).

To comprehensively evaluate the performance of different cloud detection methods, the validation and test sets in the WDCD (Li et al. 2020) dataset and the GF1MS-WHU (Zhu, Li, and Shen 2024) dataset were combined to generate a new dataset containing 60 GF1-MS images. The images in these two datasets encompass a wide variety of land-cover types and have been widely utilised in cloud detection studies (Liu et al. 2023, 2025a, 2025b). Considering the issue of missing annotations in the original pixel-level labels, we refined the labels to enhance their accuracy for thin clouds, with examples shown in Figure 9. The refined labels provide a more accurate representation of cloud coverage, although minor misclassifications remain. Overall, the omission problem was effectively alleviated, making the refined labels more suitable for the subsequent applications. Finally, this dataset with pixel-level labels was randomly divided at a 7:3 ratio and used as training data for the fully supervised deep learning methods, as well as validation data for all the cloud detection methods.

In this study, all the RegNetY networks were implemented using the pre-trained RegNetY_040 (Radosavovic et al. 2020) network from the timm library (Wightman 2019) and optimised with the Adam optimiser. The hyper-parameters were configured as follows: β_1 was set to 0.9, β_2 was set to 0.999, and the loss function was cross-entropy loss with softmax. Training was conducted for 100 epochs with an initial learning rate of 1×10^{-4}, which was reduced by a factor of 0.1 after 25 epochs. In the progressive training framework, the 128×128 and 64×64 samples were incorporated at the 30th and 60th epochs, respectively. The hyper-parameters of other compared methods were set according to their original publications. Data normalisation was performed by dividing each pixel by the maximum value within its corresponding image block. All the networks were trained on a personal computer equipped with Windows, an Intel Core i7-10700 CPU @ 2.90 GHz, 32 GB RAM, and an NVIDIA GeForce RTX 3070Ti GPU with 8 GB of memory.

2.8.2. Comparison with weakly supervised cloud detection methods

To validate the effectiveness of the proposed SpecMCD method, we conducted a comparative analysis against weakly supervised cloud detection methods, including HCDNet (Liu et al. 2023), TransMCD (Zhu, Li, and Shen 2024), WSFNet (Fu et al. 2018), WDCD (Li et al. 2020), GAN-CDM (Li et al. 2022b), and three baseline scene-level RegNetY (Radosavovic et al. 2020) networks at different scales, namely, SL-256, SL-128, and SL-64. Among the different methods, WSFNet, WDCD, GAN-CDM, and the baseline RegNetY networks rely solely on scene-level labels, whereas HCDNet and TransMCD leverage physical rules to generate

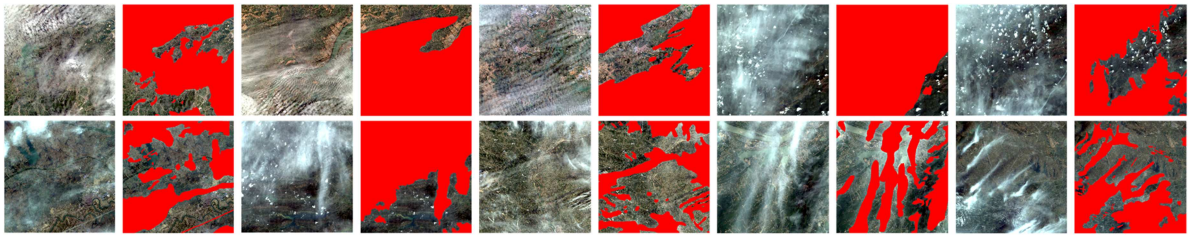

Figure 8. Examples of large-area thin-cloud images used to obtain multi-scale scene-level thin-cloud samples in the training dataset. Each image is followed by manually annotated labels, with the red regions indicating cloud areas.

Table 1. Summary of the experimental data utilised in this study.

Dataset source	Image source	Image size	Number of images	Label type	Usage
Multi-scale dataset	GF1-MS	256 × 256	61,336	Scene-level	Training of the multi-scale scene-level network
		128 × 128	224,885		
		64 × 64	826,705		
WDCD & GF1MS-WHU	GF1-MS	256 × 256	13,608	Pixel-level	Training of the pixel-level network
		>4500 × 4500	18	Pixel-level	Validation

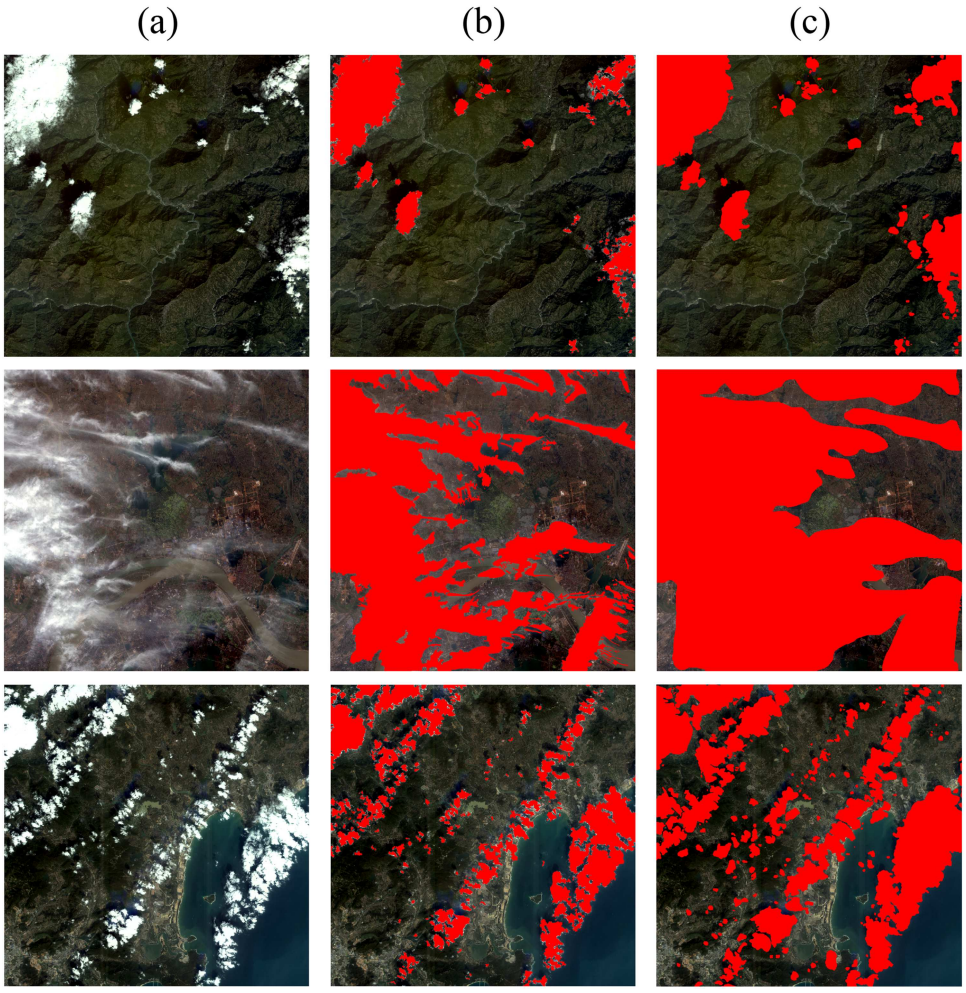

Figure 9. Visual examples of the original and manually optimised cloud labels. The red regions indicate cloud areas. (a) Gaofen-1 multispectral images; (b) original labels lacking thin-cloud samples; (c) manually optimised labels including thin-cloud samples.

pseudo-labels to improve cloud detection accuracy. The performance of all the methods was evaluated using the overall accuracy (OA), precision, recall, F1-score, and F2-score. Table 2 demonstrate that the physics rule-based HCDNet and TransMCD methods fail to generate reliable pseudo-labels for thin clouds, resulting in substantial under-detection of thin-cloud regions, with the recall lower than 0.46. The baseline scene-level networks—SL-256, SL-128, SL-64—can effectively avoid the problem of detection leakage, achieving a recall exceeding 0.95 in all the binary cloud masks. However, the baseline scene-level networks tend to exhibit significant misdetection, with the precision lower than 0.75. As the scale decreases, the misdetection occurrences of the scene-level networks are improved, but the detection leakage deteriorates, resulting in low OA, F1-score, and F2-score values for the baseline scene-level networks at different scales. The WSFNet and WDCD methods employ a class activation mapping-like mechanism to refine the scene-level network for generating pixel-level binary cloud masks. However, this refinement process leads to a reduction in recall, compared to the original scene-level networks. The GAN-CDM method leverages

Table 2. Quantitative evaluation of the binary cloud masks obtained by weakly supervised cloud detection methods.

Method	Supervision	OA	Precision	Recall	F1-score	F2-score
HCDNet	Rule	0.6847	0.9529	0.4474	0.5836	0.4916
TransMCD		0.6868	**0.9998**	0.4577	0.6049	0.5068
SL-256	Weakly supervised	0.7454	0.6393	**0.9715**	0.7330	0.8404
SL-128		0.8051	0.6960	0.9635	0.7807	0.8695
SL-64		0.8429	0.7464	0.9542	0.8215	0.8906
WSFNet		0.6787	0.7439	0.6231	0.6119	0.6086
WDCD		0.8028	0.7848	0.8599	0.7847	0.8185
GAN-CDM		0.8035	0.8164	0.8004	0.7853	0.7894
SpecMCD		**0.9126**	0.8815	0.9287	**0.8997**	**0.9156**

generative adversarial network (GAN) and cloud distortion model to achieve high-accuracy detection of thick clouds with pronounced physical characteristics, and its precision exhibits an improvement over WSFNet and WDCD. Nevertheless, in patches dominated by thin clouds and haze, GAN-CDM struggles to reliably distinguish subtle differences between thin-cloud regions and cloud-free surfaces, which leads to weaker thin-cloud detection performance compared with WDCD. Compared with the other weakly supervised methods, the SpecMCD method refines the scene-level cloud masks based on the spectral features of clouds, achieving significant improvements of over 6.97%, 7.82%, and 2.50% in OA, F1-score, and F2-score, respectively.

To further assess the performance of the different weakly supervised cloud detection methods, a visual comparison of the binary cloud masks was performed, as shown in Figure 10. The results show that the physics rule-based HCDNet and TransMCD methods are effective in thick cloud detection but exhibit a severely limited capability in detecting thin clouds under both dense and large-area cloud cover. The baseline scene-level networks at different scales demonstrate certain advantages in large-area cloud coverage regions but misclassify cloud-free regions between dense-cloud blocks, leading to significant errors. WSFNet, which emphasises thin-cloud features during training, tends to neglect thick-cloud structures, resulting in major omission errors in dense-cloud images. WDCD improves the precision in dense-cloud images through its class attention mechanism, but remains vulnerable to minor misdetections and severe detection leakage in large-area cloud images. Although GAN-CDM can achieve accurate detection of thick clouds through a GAN, the introduction of a large number of scene-level samples containing only thin clouds and haze interferes with its ability to discriminate between bright surface objects and clouds. As a result, numerous small-scale and fragmented misdetections occur in cloud-free regions. In contrast, the proposed SpecMCD method applies differentiated processing strategies for varying cloud coverage scenarios. The visual results confirm its superior performance in large-area cloud regions, enabling more comprehensive cloud detection. In dense-cloud scenes, SpecMCD also demonstrates superior results, although the performance near urban areas with high CTM values is similar to that of SL-64.

2.8.3. Comparison with fully supervised cloud detection methods

In this section, we compare the proposed weakly supervised SpecMCD method with the state-of-the-art fully supervised methods, including BoundaryNet (Zhao et al. 2023), HCDNet-Pixel (Liu et al. 2023) and RegNetY (Radosavovic et al. 2020). A quantitative evaluation is presented in Table 3. The results show that the fully supervised methods achieve a high accuracy, with F1-scores exceeding 0.88, and benefit from strong feature extraction capabilities that reduce misdetection. However, they still struggle with thin clouds and haze due to their indistinct spectral features. In contrast, SpecMCD achieves a superior performance, improving the OA and F2-score by more than 1.59% and 1.40%, respectively, compared with the fully supervised methods.

Figure 10 also presents the binary cloud masks obtained by BoundaryNet, HCDNet-Pixel, RegNetY, and SpecMCD. From the visualisation results, it can be seen that the BoundaryNet method over-emphasises the boundary details of cloud, which enables it to generate relatively precise cloud masks in dense cloud regions. However, it is not suitable for identifying large-area cloud regions. HCDNet-Pixel exhibits stronger thin-cloud detection than the other fully supervised methods, although it does introduce minor misdetections. RegNetY achieves a more balanced overall performance. Despite training with optimised pixel-level labels, the fully supervised methods still struggle with accurate detection in large-area cloud regions,

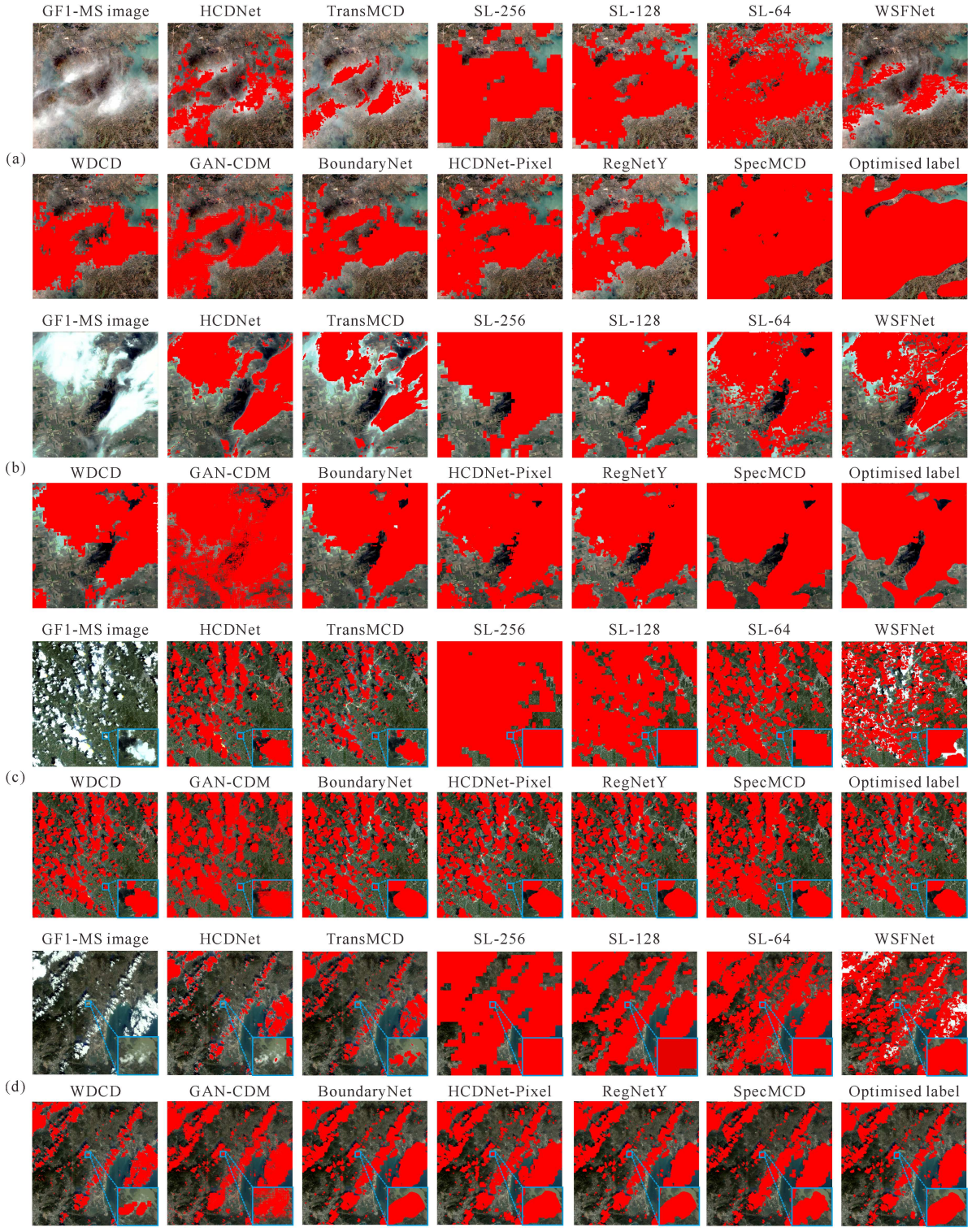

Figure 10. Examples of binary cloud masks obtained by weakly and fully supervised cloud detection methods for Gaofen-1 multispectral images. The red regions indicate cloud areas. (a) and (b) Large-area cloud images. (c) and (d) Dense-cloud images.

particularly for thin clouds. In contrast, SpecMCD demonstrates a superior performance in large-area cloud detection, enabling more comprehensive cloud coverage identification. However, in dense cloud regions, SpecMCD shows some misdetection, and its ability to capture fine cloud details remains weaker than that of the fully supervised methods.

Table 3. Quantitative evaluation of the binary cloud masks generated by fully supervised cloud detection methods.

Method	Supervision	OA	Precision	Recall	F1-score	F2-score
BoundaryNet	Fully supervised	0.8734	0.9221	0.8689	0.8813	0.8717
HCDNet-Pixel		0.8878	0.8842	0.9126	0.8899	0.9016
RegNetY		0.8967	**0.9461**	0.8781	**0.9029**	0.8866
SpecMCD	Weakly supervised	**0.9126**	0.8815	**0.9287**	0.8997	**0.9156**

Table 4. Quantitative evaluation of the binary cloud masks obtained by single-scale (SL-256, SL-128, SL-64) and progressively trained multi-scale (SL-Stack-256, SL-Stack-128, SL-Stack-64) scene-level networks and their corresponding SpecMCD method.

Method	OA	Precision	Recall	F1-score	F2-score
SL-256	0.7454	0.6393	0.9715	0.7330	0.8404
SL-128	0.8051	0.6960	0.9635	0.7807	0.8695
SL-64	0.8429	0.7464	0.9542	0.8215	0.8906
SpecMCD-Base	0.9040	0.8702	0.9309	0.8945	0.9147
SL-Stack-256	0.7505	0.6287	**0.9913**	0.7375	0.8540
SL-Stack-128	0.8282	0.7038	0.9815	0.7963	0.8878
SL-Stack-64	0.8822	0.7848	0.9674	0.8536	0.9139
SpecMCD	**0.9126**	**0.8815**	0.9287	**0.8997**	**0.9156**

2.9. Discussion

2.9.1. Effectiveness of the major components in SpecMCD

Firstly, we evaluated the effectiveness of the proposed progressive training framework by comparing the binary masks obtained by the multi-scale scene-level networks trained using the progressive framework (denoted as SL-Stack-256, SL-Stack-128, and SL-Stack-64) with those generated by baseline scene-level networks trained solely on single-scale samples. In addition, we evaluated the accuracy of the pixel-level binary masks obtained from both the progressively trained multi-scale network and the single-scale baseline network (SpecMCD-Base) within the proposed method, as shown in Table 4. The results show that the progressive framework consistently improves the OA, recall, F1-score, and F2-score across cloud masks at different scales, compared to single-scale training. These improvements confirm that the proposed progressive training framework effectively enhances the ability of the scene-level network to obtain accurate multi-scale cloud masks. Notably, while the scene-level network tends to exhibit higher recall due to it over-classifying entire patches as clouds—particularly in dense-cloud images—this also introduces a substantial number of misdetections. The proposed SpecMCD method mitigates these misdetections, achieving a more balanced detection performance. Furthermore, the comparative experiments between SpecMCD-Base and SpecMCD demonstrate that incorporating a high-precision multi-scale scene-level network further improves the accuracy of the pixel-level cloud masks within the proposed method.

Subsequently, we conducted an ablation study on the SpecMCD method to evaluate the contributions of its major components, including the use of the CTM in generating pixel-level probability maps, the dense-cloud branch only, the large cloud branch only, and the distance-weighted optimisation module. Table 5 presents the quantitative evaluation of the ablation results for the SpecMCD method. The full SpecMCD method achieves the best overall performance, demonstrating that the collaborative integration of all components is essential for accurate binary cloud mask generation. Removing the CTM results in a pronounced reduction in precision (0.8368), indicating that incorporating the CTM into the pixel-level cloud probability map generation substantially enhances the model's capability to capture fine-cloud structures and effectively suppresses misdetections along cloud boundaries. When only the dense-cloud branch or the large-cloud branch is retained, the OA drops by more than 7%, which confirms that the differentiated generation of pixel-level cloud probability maps for distinct cloud characteristics significantly improves the adaptability and robustness of SpecMCD across varying cloud coverage images. Moreover, since the proposed method adopts an adaptive thresholding strategy instead of a manual threshold, removing the distance-weighted optimisation module yields a relatively conservative cloud detection mask. Consequently, the final binary masks exhibit an extremely high precision of 0.9830. However, this conservative behaviour significantly limits the model's ability to accurately capture thin cloud regions surrounding thick cloud, resulting in a pronounced increase in omission errors and a low recall of only 0.7230. Overall, these results confirm that each component contributes synergistically to the final performance of the

Table 5. Quantitative evaluation of the binary cloud masks in the ablation study of the SpecMCD method.

Component removed	OA	Precision	Recall	F1-score	F2-score
Without cloud thickness map	0.8890	0.8368	0.9240	0.8699	0.8996
Dense cloud branch only	0.8394	0.9075	0.8053	0.8366	0.8141
Large-area cloud branch only	0.7817	0.6684	**0.9798**	0.7601	0.8604
Without distance-weighting	0.8723	**0.9830**	0.7230	0.8264	0.7600
SpecMCD	**0.9126**	0.8815	0.9287	**0.8997**	**0.9156**

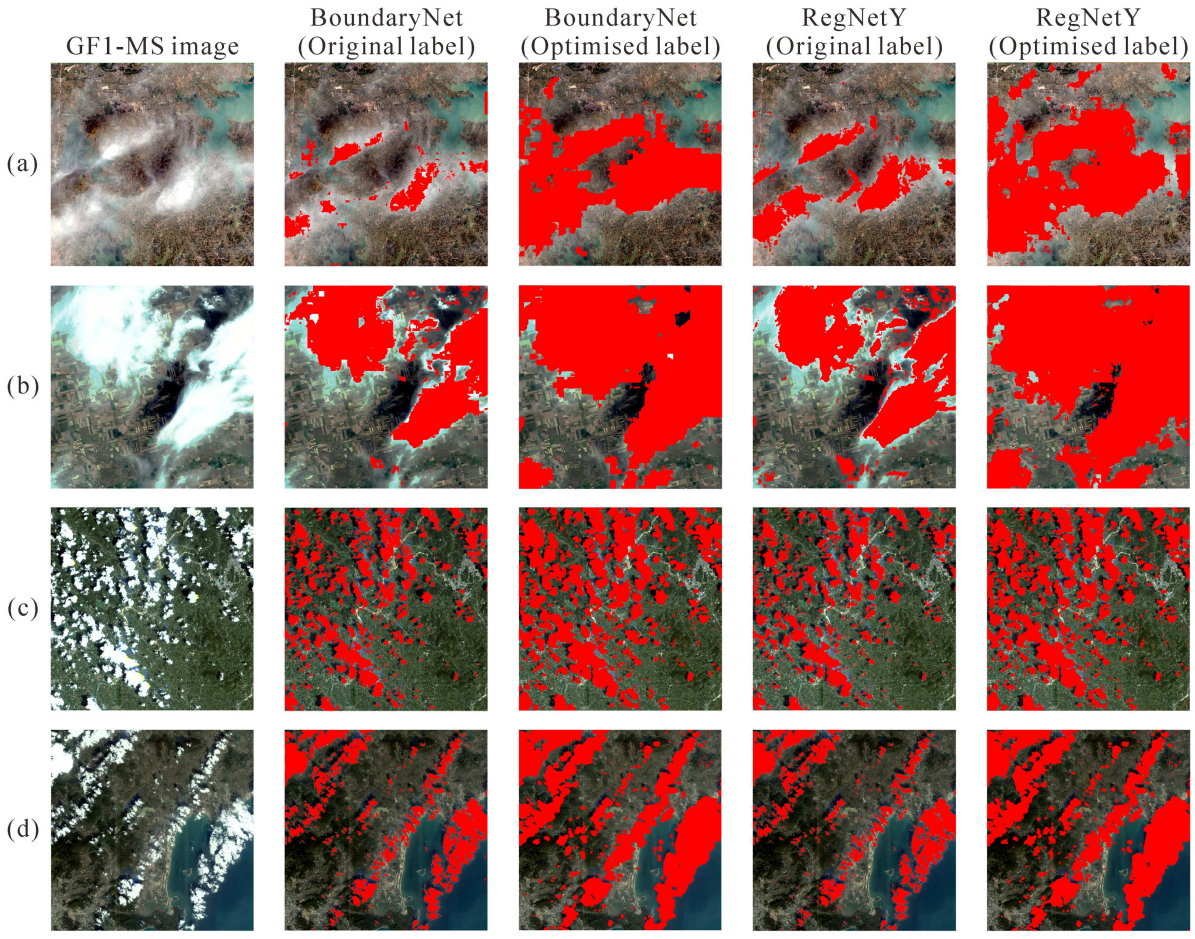

Figure 11. Examples of fully supervised cloud detection results using different manually annotated pixel-level labels. The red regions indicate cloud areas. (a) and (b) Large-area cloud images. (c) and (d) Dense-cloud images.

SpecMCD method, and that their joint optimisation is critical for achieving a balance between precision and recall under diverse cloud distribution images.

2.9.2. Effectiveness of manually optimised pixel-level labels

Due to the prevalent under-labelling of thin clouds in the WDCD and GF1MS-WHU datasets, all the pixel-level cloud labels in this study were manually optimised to ensure that the training and validation data more comprehensively and accurately reflected the distribution of both thick and thin clouds. To evaluate the effectiveness of the optimised pixel-level labels, we trained two fully supervised cloud detection networks—Boundary and RegNetY—using both the original dataset labels and the manually optimised labels, and compared their detection results. As illustrated in Figure 11, the Boundary and RegNetY networks trained with the optimised labels can effectively learn thin-cloud features from the training samples and exhibit significantly improved thin-cloud detection capabilities in large-area cloud images. In dense-cloud images, the optimised labels also enhance the networks' ability to detect thin clouds around thick cloud, yielding cloud masks that more accurately reflect the cloud coverage and spatial distribution. In

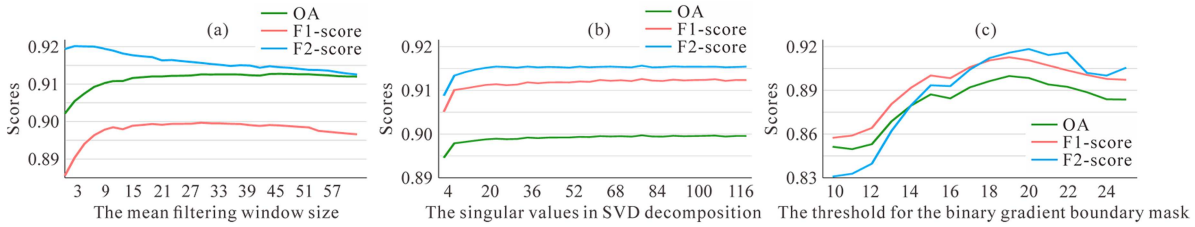

Figure 12. Sensitivity analysis of the hyperparameters for generating pixel-level cloud probability maps. (a) The mean filtering window size for smoothing the cloud thickness map. (b) The number of singular values k in singular value decomposition. (c) The threshold μ_{Grad} for obtaining the binary gradient boundary mask.

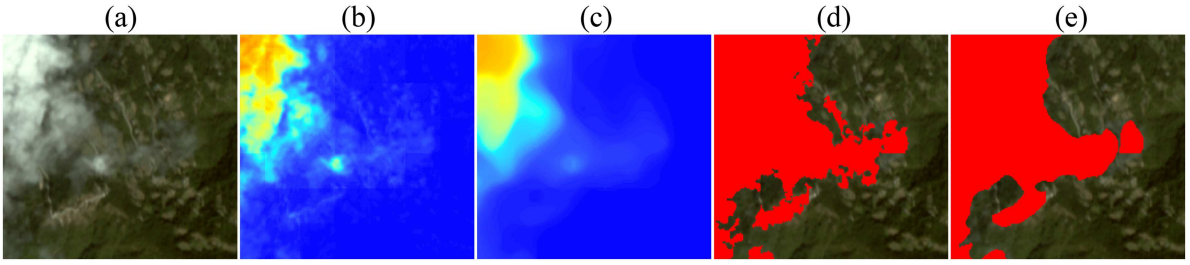

Figure 13. Visual examples of mean filtering on dense cloud probability maps and the corresponding binary cloud masks. (a) Gaofen-1 multispectral images. (b) Original cloud probability map without mean filtering. (c) Cloud probability map after mean filtering with a 29×29 window. (d) Original binary cloud mask without mean filtering. (e) Binary cloud mask after mean filtering with a 29×29 window.

contrast, networks trained with the original labels can detect thick cloud but fail to capture thin cloud features, due to the lack of thin-cloud samples in the original datasets. Consequently, the original networks exhibit notable thin-cloud omissions in both large-area cloud regions and thick-cloud boundaries, which is insufficient for the thin and thick-cloud detection addressed in this study. Overall, the optimised pixel-level labels substantially improve the detection of thin-cloud regions by the fully supervised network, thereby demonstrating the necessity and effectiveness of the cloud label optimisation conducted in this study.

2.9.3. Sensitivity analysis of the hyperparameters

The SpecMCD method introduces several hyperparameters for generating pixel-level cloud probability maps after incorporating the spectral features. To validate the robustness of the proposed method, we conducted a sensitivity analysis of the mean filtering window size, the selected singular values in SVD decomposition and the threshold for obtaining the binary gradient boundary mask.

Figure 12(a) illustrates the impact of mean filtering window size on smoothing the CTM in dense-cloud images and its influence on the accuracy of the final binary cloud mask. Figure 13 further compares the dense-cloud probability maps and binary cloud masks before and after mean filtering. The results demonstrate that mean filtering tends to distort dense-cloud shapes, causing the omission of fine-scale details in the binary masks and leading to a decline in the F2-score. Nevertheless, an appropriately sized mean filter can effectively suppress noise and reduce false positives from small bright non-cloud surfaces, thereby improving the OA and F1-scores and yielding a more balanced detection performance.

As shown in Figure 12(b), when extracting cloud features from large-area cloud images using SVD decomposition, selecting singular values greater than 8 has little effect on the final cloud mask. Furthermore, Figure 12(c) presents a sensitivity analysis of the threshold μ_{Grad} for obtaining the binary gradient boundary mask. Since the CTM gradient tends to be higher at thick-cloud boundaries, an excessively small μ_{Grad} may cause boundary extraction errors, hereby reducing the accuracy and stability of the final mask. When μ_{Grad} is set within the range of 17–21, its influence on the accuracy of the binary cloud mask is minimal, indicating that SpecMCD is robust to this hyperparameter, within a reasonable

range. However, setting μ_{Grad} beyond this range leads to degraded thick-cloud boundary detection and consequently a lower mask accuracy.

2.9.4. Analysis of the pixel-level cloud probability maps

Spectral feature based methods can achieve a satisfactory performance in certain images (Foga et al. 2017). However, they require manual threshold adjustments across different sensors, or even across scenes from the same sensor, which limits their robustness and automation, compared with machine learning based methods. While the proposed SpecMCD method incorporates spectral features, this integration may reduce the automatic segmentation capability of the deep learning network. Therefore, we employ adaptive thresholding to generate the binary cloud mask, rather than an artificially optimal threshold. Nevertheless, this mask still struggles to accurately delineate thin clouds surrounding thick-cloud regions. Even with distance weighting optimisation, the absence of a fixed scope for thin clouds leads to misdetections in dense-cloud areas and detection leakage in large-area cloud regions.

To further investigate this limitation, we compared the cloud probability maps generated by different methods without binarization. Figure 14 shows that the cloud probability maps from HCDNet, TransMCD, WSFNet, WDCD, and GAN-CDM effectively capture the thick-cloud distribution but show limited representation of thin clouds, hindering accurate binary mask generation. The baseline scene-level networks display strong capability in representing cloud probabilities over large areas; however, their ability diminishes as the network scale decreases. The fully-supervised methods, including BoundaryNet, HCDNet-Pixel, and RegNetY, achieve reliable cloud probability estimation in dense-cloud regions. However, despite being trained with thin cloud samples, they fail to capture the gradual transition from thick to thin clouds, producing probability maps that lean toward binarization. The SpecMCD method generates cloud probability maps in dense cloud regions that surpass those of the other weakly supervised methods and approach the quality of RegNetY. Nevertheless, SpecMCD still loses fine-scale details, particularly at cloud boundaries, resulting in a slightly inferior performance, compared with the fully supervised methods. In large-area cloud images, SpecMCD more effectively represents the cloud thickness and can better represent the cloud distribution in the imagery.

2.9.5. Effectiveness of SpecMCD for downstream applications

Since the proposed SpecMCD method can achieve a more comprehensive detection of both thick and thin clouds, it has the potential to provide more reliable surface information for downstream applications. To assess this capability, we selected the FRARC (Zhu et al. 2023) thick cloud removal method as a representative downstream task and evaluated the reconstruction performance using different cloud masks, as shown in Figure 15. The experimental results indicate that reconstructions based on RegNetY masks can effectively remove thick cloud. However, they remain highly sensitive to undetected thin clouds, leading to radiance overestimation in the reconstructed regions. In contrast, the reconstructions guided by SpecMCD masks benefit from the improved detectability of thin clouds, enabling the concurrent removal of both thick and thin cloud and producing visually more coherent and realistic results.

Despite these advantages, limitations remain. In regions affected by extremely thin cloud cover, SpecMCD still encounters challenges in achieving fully accurate detection, as exemplified by the area highlighted by the blue box in Figure 15(c). This observation suggests that further improvements in the discrimination of extremely thin clouds are necessary for certain complex atmospheric conditions, and it provides a meaningful direction for future methodological enhancements.

Considering the requirements on both time efficiency and memory usage for deployment in downstream tasks, we further analyse the time consumption and maximum memory usage of different cloud detection methods, as summarised in Table 6. In terms of time consumption, the proposed SpecMCD method is higher than other methods, as it relies on a sliding-window strategy to generate multi-scale scene-level cloud probability maps and further integrates SVD decomposition and distance-weighted optimisation to obtain pixel-level binary cloud masks. Nevertheless, for images with a size of approximately 4600×4600 pixels, the overall time consumption of SpecMCD is within 2 minutes, which is generally acceptable for practical applications. Regarding maximum memory usage, SpecMCD achieves the lowest memory usage among the compared methods, with the maximum memory usage constrained to within 6 GB. Overall, although the proposed SpecMCD method involves several computation-intensive

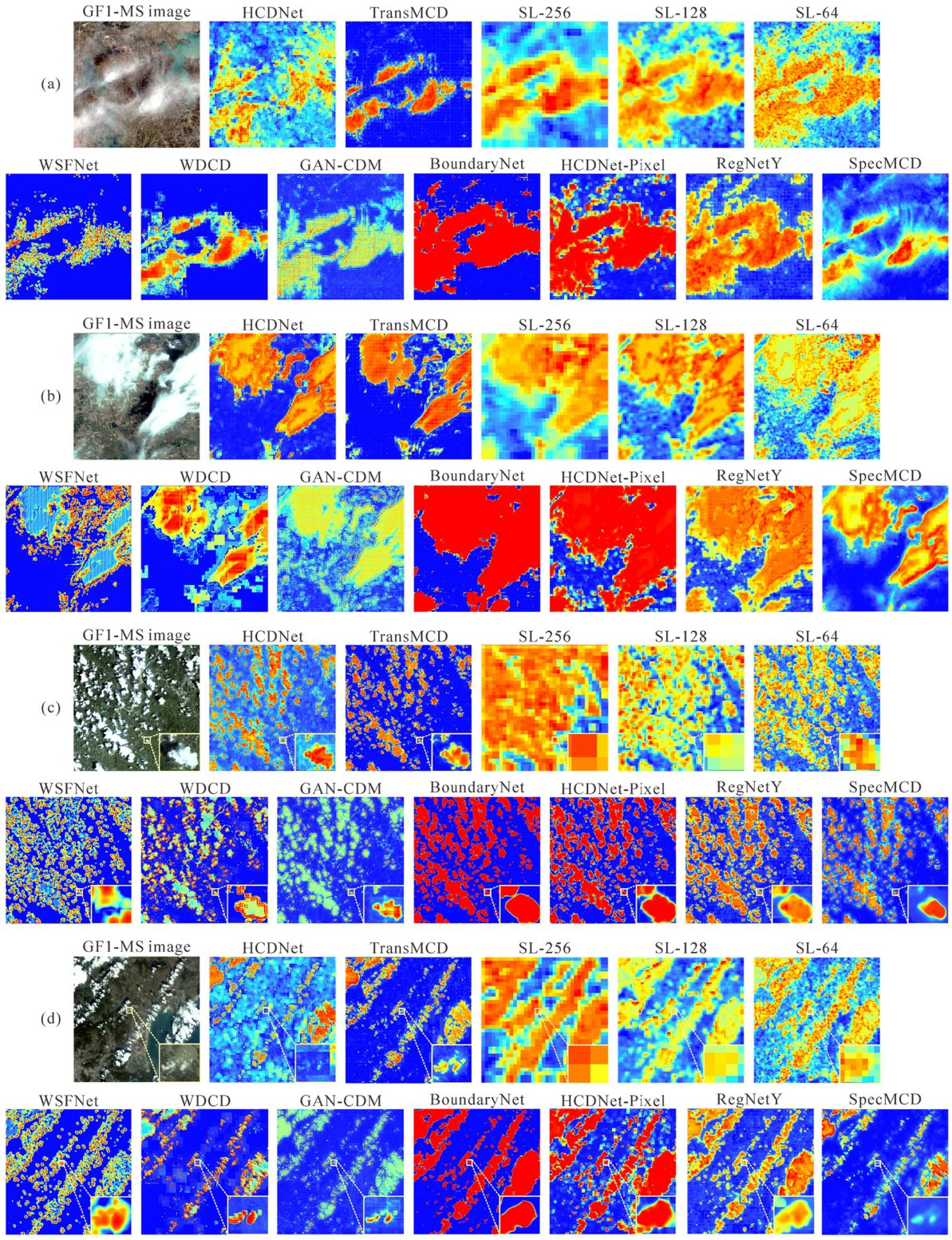

Figure 14. Examples of cloud probability maps obtained by weakly and fully supervised cloud detection methods for Gaofen-1 multispectral images. The cloud probability maps use a colour scale from blue, indicating low cloud probability, to red, indicating high cloud probability. (a) and (b) Large-area cloud images. (c) and (d) Dense-cloud images.

components that incur relatively high computational cost, its overall time consumption and memory usage remain generally acceptable for practical applications. Nevertheless, in future studies, we will investigate more efficient strategies for accelerating the generation of multi-scale scene-level cloud probability maps, with the aim of further improving the practicality of the proposed method.

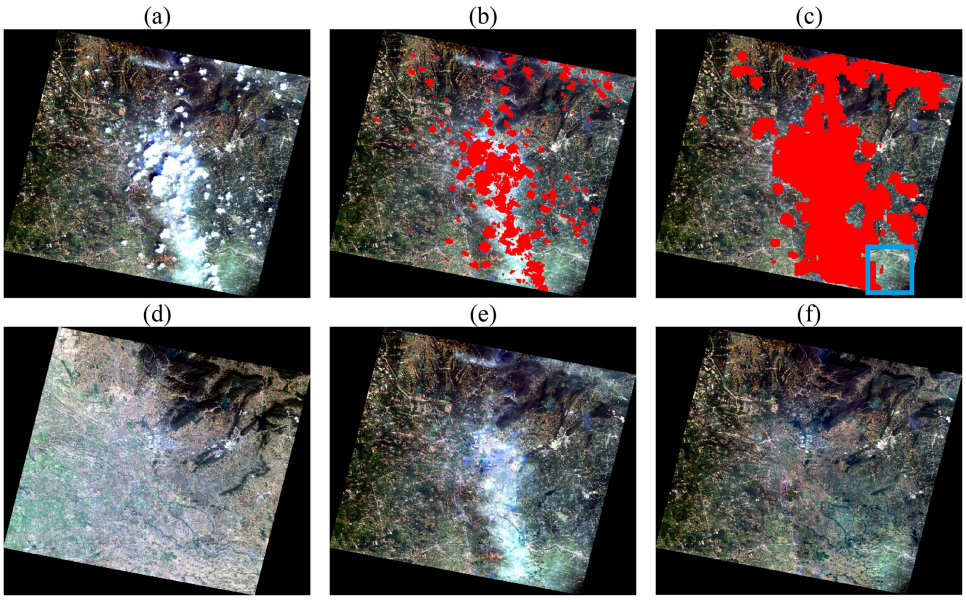

Figure 15. Comparison of image reconstruction results using different cloud masks. (a) Target image. (b) Cloud mask generated by the RegNetY method. (c) Cloud mask generated by the SpecMCD method. (d) Reference image. (e) Reconstruction based on the RegNetY mask. (f) Reconstruction based on the SpecMCD mask.

Table 6. Time consumption and maximum memory usage of different methods on the validation dataset.

Average imge size	Method		Time consumption (seconds)		Maximum memory usage (MB)
>4694 × 4694	WDCD		102.2		7342.2
	GAN-CDM		18.6		6654.2
	BoundaryNet		40.0		6232.8
	RegNetY		**18.1**		6357.7
	SpecMCD	256 × 256 Prob. map	9.1	119.9	**6031.4**
		128 × 128 Prob. map	22.7		
		64 × 64 Prob. map	65.7		
		Others	22.4		

2.9.6. Limitations and future perspectives

The SpecMCD method can generate pixel-level cloud masks using scene-level labels. While the SpecMCD method can achieve superior cloud detection in large-area images compared to the fully supervised methods, its performance is inferior to that of the fully supervised methods in dense-cloud images. Since distinguishing clouds from snow/ice using only visible and near-infra-red bands is inherently challenging (Li et al. 2022a), the reliance of SpecMCD on visible bands for spectral feature extraction makes it heavily dependent on the multi-scale scene-level network. As shown by the blue boxed region in Figure 16, misclassification of snow as cloud occurs when the multi-scale network fails to correctly differentiate the two.

To evaluate the cross-sensor robustness of the proposed SpecMCD method, the multi-scale scene-level network and the SpecMCD method trained on the GF1-MS dataset were directly applied—without any fine-tuning—to cloud detection on Gaofen-2 multispectral (GF2-MS) images, as shown in Figure 17. The experimental results indicate that the scene-level network exhibits a cross-sensor generalisation capability, enabling the proposed SpecMCD method to generate visually satisfactory binary cloud masks on GF2-MS images. Nevertheless, although 215 original images were utilised to generate multi-scale scene-level samples for training the scene-level networks, the size of the training dataset remains limited compared with the large-scale pre-training datasets employed by remote sensing foundation models. Therefore, future research will aim to incorporate multi-source optical imagery to further expand the dataset and adopt advanced data augmentation strategies to enhance the cross-sensor robustness of the SpecMCD method. In addition, we will focus on constructing and optimising cloud detection datasets across different sensors, enabling a comprehensive assessment of the cross-sensor robustness of the proposed method.

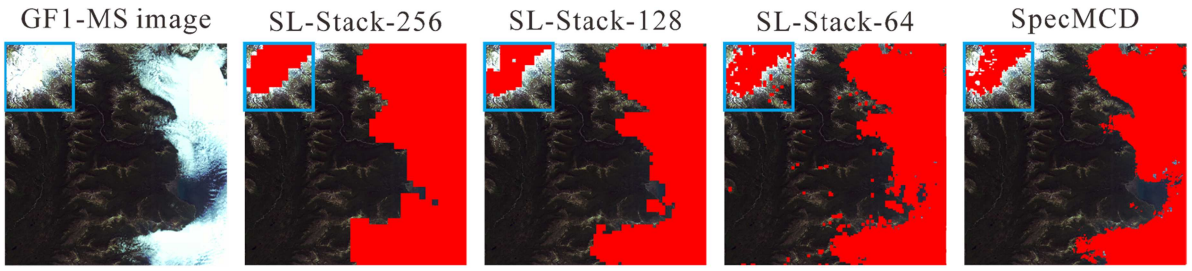

Figure 16. Examples of binary cloud masks obtained by multi-scale scene-level networks (SL-Stack-256, SL-Stack-128, SL-Stack-64) and the SpecMCD method for Gaofen-1 multispectral images. The red regions indicate detected cloud areas.

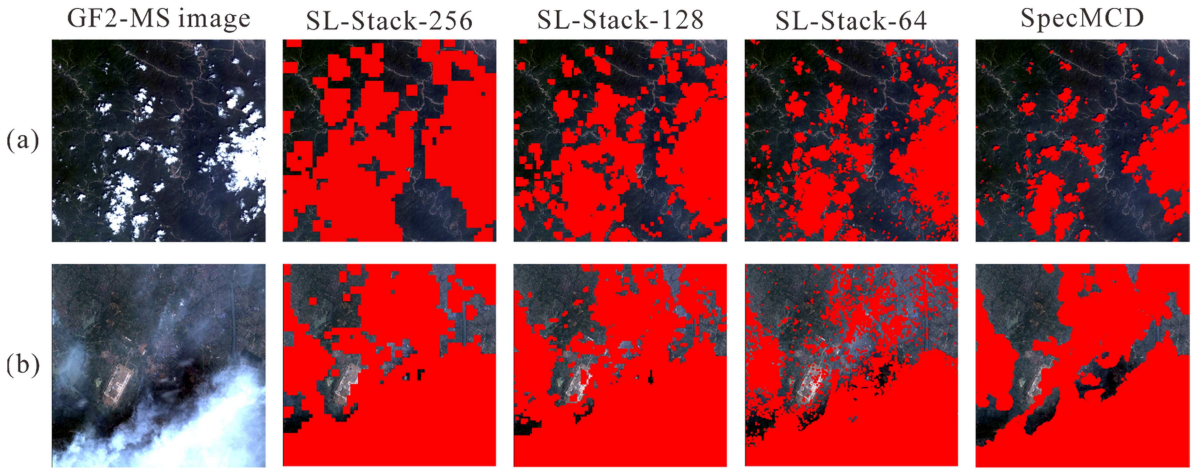

Figure 17. Examples of binary cloud masks obtained by multi-scale scene-level networks (SL-Stack-256, SL-Stack-128, SL-Stack-64) and the SpecMCD method for Gaofen-2 multispectral images. The red regions indicate cloud areas. (a) Dense cloud image. (b) Large-area cloud image.

In future work, we aim to combine pixel-level and scene-level networks to improve the cloud detection accuracy in dense-cloud regions. To address the misclassification between cloud and snow, we will incorporate manually labelled cloud-snow samples in a data-driven framework to strengthen the discriminative capability of the multi-scale network. Moreover, since the proposed method cannot be directly applied to cloud shadow detection, we will explore combining a fully supervised cloud shadow network with morphological constraints from cloud-shadow relationships to achieve shadow detection. In addition, we will investigate the use of advanced and robust edge detection methods to further enhance thick-cloud boundary extraction, which is expected to improve the accuracy of the fused pixel-level cloud probability maps.

2.10. Conclusion

In this paper, we have proposed a weakly supervised cloud detection method (SpecMCD) that combines spectral features and a multi-scale scene-level deep network. SpecMCD achieves accurate detection of both thick and thin clouds, which significantly reduces the omissions occurring in the binary cloud masks. Overall, the proposed method has the following advantages: 1) A progressive training framework is proposed to integrate multi-scale scene-level samples into a single network, which can generate highly accurate multi-scale scene-level cloud probability maps. 2) A differentiated processing strategy is employed to combine the multi-scale scene-level network with the CTM according to the distribution characteristics of dense and large-area clouds. Furthermore, the probability maps are fused based on the CTM gradient to improve the detection performance for dense and large-area clouds. 3) By leveraging the differences

among multi-scale scene-level masks, adaptive thresholding is employed to reduce the need for manual threshold adjustment, while distance-weighted optimisation further refines the binary masks.

The results of the experiments combining two datasets (i.e. WDCD and GF1MS-WHU) demonstrated that SpecMCD improves the F1-score by more than 7.82%, compared with the other weakly supervised methods, and is effective in reducing the omission errors in cloud detection, especially for thin clouds. Nevertheless, the accurate differentiation between clouds and snow remains a challenging task. Future work will explore combining scene-level and pixel-level networks to further enhance detection in dense-cloud regions and improve cloud-snow discrimination through a data-driven framework.

Author contributions

CRediT: **Shaocong Zhu:** Conceptualization, Data curation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Zhiwei Li:** Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing; **Xinghua Li:** Data curation, Investigation, Resources, Supervision; **Huanfeng Shen:** Conceptualization, Formal analysis, Funding acquisition, Project administration, Resources, Software, Supervision, Writing – review & editing.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This study was supported by the National Natural Science Foundation of China (No. 42130108 and 42101357).

ORCID

Shaocong Zhu 0009-0003-4334-8373
Zhiwei Li 0000-0001-5635-8499
Huanfeng Shen 0000-0002-4140-1869

Data availability statement

The data that support the findings of this study are available from the corresponding author, S.Z, upon reasonable request and are available from https://doi.org/10.5281/zenodo.17265317 (Zhu 2025).

Preprint

https://doi.org/10.48550/arXiv.2510.00654.

References

Cao, Y., and X. Huang. 2022. "A Coarse-To-Fine Weakly Supervised Learning Method for Green Plastic Cover Segmentation Using High-Resolution Remote Sensing Images." *ISPRS Journal of Photogrammetry and Remote Sensing* 188: 157–176. https://doi.org/10.1016/j.isprsjprs.2022.04.012.

Chai, D., J. Huang, M. Wu, X. Yang, and R. Wang. 2024. "Remote Sensing Image Cloud Detection Using a Shallow Convolutional Neural Network." *ISPRS Journal of Photogrammetry and Remote Sensing* 209: 66–84. https://doi.org/10.1016/j.isprsjprs.2024.01.026.

Chai, D., S. Newsam, H. Zhang, Y. Qiu, and J. Huang. 2019. "Cloud and Cloud Shadow Detection in Landsat Imagery Based on Deep Convolutional Neural Networks." *Remote Sensing of Environment* 225: 307–316. https://doi.org/10.1016/j.rse.2019.03.007.

Foga, S., P. Scaramuzza, S. Guo, Z. Zhu, R. Dilley, T. Beckmann, G. Schmidt, J. Dwyer, M. Hughes, and B. Laue. 2017. "Cloud Detection Algorithm Comparison and Validation for Operational Landsat Data Products." *Remote Sensing of Environment* 194: 379–390. https://doi.org/10.1016/j.rse.2017.03.026.

Fu, H., Y. Shen, J. Liu, G. He, J. Chen, P. Liu, J. Qian, and J. Li. 2019. "Cloud Detection for FY Meteorology Satellite Based on Ensemble Thresholds and Random Forests Approach." *Remote Sensing* 11: 44. https://doi.org/10.3390/rs11010044.

Fu, K., W. Lu, W. Diao, M. Yan, H. Sun, Y. Zhang, and X. Sun. 2018. "WSF-NET: Weakly Supervised Feature-Fusion Network for Binary Segmentation in Remote Sensing Image." *Remote Sensing* 10: 1970. https://doi.org/10.3390/rs10121970.

Gbodjo, Y., L. Hughes, M. Molinier, D. Tuia, and J. Li. 2026. "Self-Supervised Representation Learning for Cloud Detection Using sentinel-2 Images." *Remote Sensing of Environment* 334: 115205. https://doi.org/10.1016/j.rse.2025.115205.

He, K., J. Sun, and X. Tang. 2009. Single Image Haze Removal Using Dark Channel Prior. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 1956–1963, Miami, FL, USA: IEEE. https://doi.org/10.1109/CVPR.2009.5206515.

He, Q., X. Sun, Z. Yan, and K. Fu. 2022. "DABNet: Deformable Contextual and Boundary-Weighted Network for Cloud Detection in Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–16. https://doi.org/10.1109/TGRS.2020.3045474.

Hu, J., L. Shen, S. Albanie, G. Sun, and E. Wu. 2020. "Squeeze-And-Excitation Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42: 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372.

Ibrahim, E., J. Jiang, L. Lema, P. Barnabé, G. Giuliani, P. Lacroix, and E. Pirard. 2021. "Cloud and Cloud-Shadow Detection for Applications in Mapping Small-Scale Mining in Colombia Using sentinel-2 Imagery." *Remote Sensing* 13: 736. https://doi.org/10.3390/rs13040736.

Ishida, H., Y. Oishi, K. Morita, K. Moriwaki, and T. Nakajima. 2018. "Development of a Support Vector Machine Based Cloud Detection Method for MODIS with the Adjustability to Various Conditions." *Remote Sensing of Environment* 205: 390–407. https://doi.org/10.1016/j.rse.2017.11.003.

Joshi, P. P., R. H. Wynne, and V. A. Thomas. 2019. "Cloud Detection Algorithm Using SVM with SWIR2 and Tasseled Cap Applied to Landsat 8." *International Journal of Applied Earth Observation and Geoinformation* 82: 101898. https://doi.org/10.1016/j.jag.2019.101898.

Lee, Y., S. Min, J. Yoon, J. Ha, S. Jeong, S. Ryu, and M. Ahn. 2025. "Application of Deep Learning in Cloud Cover Prediction Using Geostationary Satellite Images." *GIScience Remote Sensing* 62: 2440506. https://doi.org/10.1080/15481603.2024.2440506.

Li, J., C. Hu, Q. Sheng, J. Xu, C. Zhu, and W. Zhang. 2024. "A Multi-Scale Features-Based Cloud Detection Method for suomi-NPP VIIRS Day and Night Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* X-1-2024: 115–122. https://doi.org/10.5194/isprs-annals-X-1-2024-115-2024.

Li, J., Z. Wu, Q. Sheng, B. Wang, Z. Hu, S. Zheng, G. Campus-Valls, and M. Molinier. 2022b. "A Hybrid Generative Adversarial Network for Weakly-Supervised Cloud Detection in Multispectral Images." *Remote Sensing of Environment* 280: 113197. https://doi.org/10.1016/j.rse.2022.113197.

Li, J., Z. Wu, Z. Hu, C. Jian, S. Luo, L. Mou, X. Zhu, and M. Molinier. 2022a. "A Lightweight Deep Learning-Based Cloud Detection Method for Sentinel-2A Imagery Fusing Multiscale Spectral and Spatial Features." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–19. https://doi.org/10.1109/TGRS.2021.3069641.

Li, Y., W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan. 2020. "Accurate Cloud Detection in High-Resolution Remote Sensing Imagery By Weakly Supervised Deep Learning." *Remote Sensing of Environment* 250: 112045. https://doi.org/10.1016/j.rse.2020.112045.

Li, Z., H. Shen, H. Li, G. Xia, P. Gamba, and L. Zhang. 2017. "Multi-Feature Combined Cloud and Cloud Shadow Detection in GaoFen-1 Wide Field of View Imagery." *Remote Sensing of Environment* 191: 342–358. https://doi.org/10.1016/j.rse.2017.01.026.

Li, Z., H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He. 2019. "Deep Learning Based Cloud Detection for Medium and High Resolution Remote Sensing Images of Different Sensors." *ISPRS Journal of Photogrammetry and Remote Sensing* 150: 197–212. https://doi.org/10.1016/j.isprsjprs.2019.02.017.

Li, Z., Q. Weng, Y. Zhou, P. Dou, and X. Ding. 2024. "Learning Spectral-Indices-Fused Deep Models for Time-Series Land Use and Land Cover Mapping in Cloud-Prone Areas: The Case of Pearl River Delta." *Remote Sensing of Environment* 308: 114190. https://doi.org/10.1016/j.rse.2024.114190.

Liang, K., G. Yang, Y. Zuo, J. Chen, W. Sun, X. Meng, and B. Chen. 2024. "A Novel Method for Cloud and Cloud Shadow Detection Based on the Maximum and Minimum Values of sentinel-2 Time Series Images." *Remote Sensing* 16: 1392. https://doi.org/10.3390/rs16081392.

Liu, Q., X. Gao, L. He, and W. Lu. 2017. "Haze Removal for a Single Visible Remote Sensing Image." *Signal Process* 137: 33–43. https://doi.org/10.1016/j.sigpro.2017.01.036.

Liu, W., B. Luo, J. Liu, H. Nie, and X. Su. 2025a. "SCTNet: A Shallow CNN-transformer Network with Statistics-Driven Modules for Cloud Detection." *IEEE Geoscience and Remote Sensing Letters* 22: 1–5. https://doi.org/10.1109/LGRS.2025.3561004.

Liu, W., B. Luo, J. Liu, H. Nie, and X. Su. 2025b. "FEMNet: A Feature-Enriched Mamba Network for Cloud Detection in Remote Sensing Imagery." *Remote Sensing* 17: 2639. https://doi.org/10.3390/rs17152639.

Liu, Y., Q. Li, X. Li, S. He, F. Liang, Z. Yao, J. Jiang, and W. Wang. 2023. "Leveraging Physical Rules for Weakly Supervised Cloud Detection in Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 61: 1–18. https://doi.org/10.1109/TGRS.2023.3294817.

Makarau, A., R. Richter, R. Müller, and P. Reinartz. 2014. "Haze Detection and Removal in Remotely Sensed Multispectral Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 52: 5895–5905. https://doi.org/10.1109/TGRS.2013.2293662.

Ping, B., F. Su, and Y. Meng. 2020. "A Cloud and Cloud Shadow Detection Method Based on Fuzzy c-means Algorithm." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13: 1714–1727. https://doi.org/10.1109/JSTARS.2020.2987844.

Radosavovic, I., R. Kosaraju, R. Girshick, K. He, and P. Dollár. 2020. Designing Network Design Spaces. In *2020 IEEE/CVF Conference On Computer Vision And Pattern Recognition*10425–10433, Seattle, WA, USA: IEEE. https://doi.org/10.1109/CVPR42600.2020.01044.

Shen, H., X. Li, Q. Cheng, C. Zeng, G. Yang, H. Li, and L. Zhang. 2015. "Missing Information Reconstruction of Remote Sensing Data: a Technical Review." *IEEE Geoscience and Remote Sensing Magazine* 3: 61–85. https://doi.org/10.1109/MGRS.2015.2441912.

Shendryk, Y., Y. Rist, C. Ticehurst, and P. Thorburn. 2019. "Deep Learning for Multi-Modal Classification of Cloud, Shadow and Land Cover Scenes in PlanetScope and sentinel-2 Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 157: 124–136. https://doi.org/10.1016/j.isprsjprs.2019.08.018.

Sun, L., X. Mi, J. Wei, J. Wang, X. Tian, H. Yu, and P. Gan. 2017. "A Cloud Detection Algorithm-Generating Method for Remote Sensing Data at Visible to Short-Wave Infrared Wavelengths." *ISPRS Journal of Photogrammetry and Remote Sensing* 124: 70–88. https://doi.org/10.1016/j.isprsjprs.2016.12.005.

Wang, J., D. Yang, S. Chen, X. Zhu, S. Wu, M. Bogonovich, Z. Guo, Z. Zhu, and J. Wu. 2021. "Automatic Cloud and Cloud Shadow Detection in Tropical Areas for PlanetScope Satellite Images." *Remote Sensing of Environment* 264: 112604. https://doi.org/10.1016/j.rse.2021.112604.

Wang, Q., J. Li, X. Tong, and P. Atkinson. 2024. "TSI-siamnet: A Siamese Network for Cloud and Shadow Detection Based on Time-Series Cloudy Images." *ISPRS Journal of Photogrammetry and Remote Sensing* 213: 107–123. https://doi.org/10.1016/j.isprsjprs.2024.05.022.

Wang, Y., L. Gu, X. Li, F. Gao, and T. Jiang. 2023. "Coexisting Cloud and Snow Detection Based on a Hybrid Features Network Applied to Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 61: 1–15. https://doi.org/10.1109/TGRS.2023.3299617.

Wang, Z., L. Zhao, J. Meng, Y. Han, X. Li, R. Jiang, J. Chen, and H. Li. 2024. "Deep Learning-Based Cloud Detection for Optical Remote Sensing Images: A Survey." *Remote Sensing* 16: 4583. https://doi.org/10.3390/rs16234583.

Wei, J., W. Huang, Z. Li, L. Sun, X. Zhu, Q. Yuan, L. Liu, and M. Cribb. 2020. "Cloud Detection for Landsat Imagery By Combining the Random Forest and Superpixels Extracted Via Energy-Driven Sampling Segmentation Approaches." *Remote Sensing of Environment* 248: 112005. https://doi.org/10.1016/j.rse.2020.112005.

Wightman, R. 2019. PyTorch Image Models. GitHub Repos. https://doi.org/10.5281/zenodo.4414861.

Wright, N., J. Duncan, J. Callow, S. Thompson, and R. George. 2024. "CloudS2Mask: A Novel Deep Learning Approach for Improved Cloud and Cloud Shadow Masking in sentinel-2 Imagery." *Remote Sensing of Environment* 306: 114122. https://doi.org/10.1016/j.rse.2024.114122.

Wu, W., J. Luo, X. Hu, H. Yang, and Y. Yang. 2018. "A Thin-Cloud Mask Method for Remote Sensing Images Based on Sparse Dark Pixel Region Detection." *Remote Sensing* 10: 617. https://doi.org/10.3390/rs10040617.

Xie, F., M. Shi, Z. Shi, J. Yin, and D. Zhao. 2017. "Multilevel Cloud Detection in Remote Sensing Images Based on Deep Learning." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10: 3631–3640. https://doi.org/10.1109/JSTARS.2017.2686488.

Xie, S., R. Girshick, P. Dollár, Z. Tu, and K. He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference On Computer Vision And Pattern Recognition*5987–5995, Honolulu, HI, USA: IEEE. https://doi.org/10.1109/CVPR.2017.634.

Yang, J., J. Guo, H. Yue, Z. Liu, H. Hu, and K. Li. 2019. "CDnet: CNN-based Cloud Detection for Remote Sensing Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 57: 6195–6211. https://doi.org/10.1109/TGRS.2019.2904868.

Yang, J., W. Li, K. Chen, Z. Liu, Z. Shi, and Z. Zou. 2024. "Weakly Supervised Adversarial Training for Remote Sensing Image Cloud and Snow Detection." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17: 15206–15221. https://doi.org/10.1109/JSTARS.2024.3448356.

Yun, Y., J. Jung, and Y. Han. 2024. "Cloud Restoration of Optical Satellite Imagery Using Time-Series Spectral Similarity Group." *GIScience Remote Sens* 61: 2324553. https://doi.org/10.1080/15481603.2024.2324553.

Zhai, H., H. Zhang, L. Zhang, and P. Li. 2018. "Cloud/Shadow Detection Based on Spectral Indices for Multi/Hyperspectral Optical Remote Sensing Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 144: 235–253. https://doi.org/10.1016/j.isprsjprs.2018.07.006.

Zhang, H., Q. Huang, H. Zhai, and L. Zhang. 2021. "Multi-Temporal Cloud Detection Based on Robust PCA for Optical Remote Sensing Imagery." *Computers and Electronics in Agriculture* 188: 106342. https://doi.org/10.1016/j.compag.2021.106342.

Zhao, C., X. Zhang, N. Kuang, H. Luo, S. Zhong, and J. Fan. 2023. "Boundary-Aware Bilateral Fusion Network for Cloud Detection." *IEEE Transactions on Geoscience and Remote Sensing* 61: 1–14. https://doi.org/10.1109/TGRS.2023.3276750.

Zhu, S. 2025. Dataset-of-SpecMCD. https://doi.org/10.5281/zenodo.17265317.

Zhu, S., Z. Li, and H. Shen. 2024. "Transferring Deep Models for Cloud Detection in Multisensor Images Via Weakly Supervised Learning." *IEEE Transactions on Geoscience and Remote Sensing* 62: 1–18. https://doi.org/10.1109/TGRS.2024.3358824.

Zhu, S., Z. Li, H. Shen, and D. Lin. 2023. "A Fast Two-Step Algorithm for Large-Area Thick Cloud Removal in High-Resolution Images." *Remote Sensing Letters* 14: 1–9. https://doi.org/10.1080/2150704X.2022.2152753.

Zhu, Z., and C. Woodcock. 2012. "Object-Based Cloud and Cloud Shadow Detection in Landsat Imagery." *Remote Sensing of Environment* 118: 83–94. https://doi.org/10.1016/j.rse.2011.10.028.

Zhu, Z., S. Qiu, B. He, and C. Deng. 2018. Cloud and Cloud Shadow Detection for Landsat Images: The Fundamental Basis for Analyzing Landsat Time Series. In *Remote Sensing Time Series Image Processing* 1st 3–19. https://doi.org/10.1201/9781315166636-1.

Zhu, Z., S. Wang, and C. Woodcock. 2015. "Improvement and Expansion of the Fmask Algorithm: Cloud, Cloud Shadow, and Snow Detection for Landsats 4-7, 8, and Sentinel 2 Images." *Remote Sensing of Environment* 159: 269–277. https://doi.org/10.1016/j.rse.2014.12.014.

Zi, Y., F. Xie, and Z. Jiang. 2018. "A Cloud Detection Method for Landsat 8 Images Based on Pcanet." *Remote Sensing* 10: 877. https://doi.org/10.3390/rs10060877.