TSI-Siamnet: A Siamese network for cloud and shadow detection based on time-series cloudy images

Qunming Wang [a], Jiayi Li [a], Xiaohua Tong [a, *], Peter M. Atkinson [b, c]

[a] *College of Surveying and Geo-Informatics, Tongji University, 1239 Siping Road, Shanghai 200092, China*

[b] *Faculty of Science and Technology, Lancaster University, Lancaster LA1 4YR, UK*

[c] *Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK*

*Corresponding author. E-mail: xhtong@tongji.edu.cn.

Abstract: Accurate cloud and shadow detection is a crucial prerequisite for optical remote sensing image analysis and application. Multi-temporal-based cloud and shadow detection methods are a preferable choice to detect clouds in complex scenes (e.g., thin clouds, broken clouds and clouds with interference from artificial surfaces with high reflectivity). However, such methods commonly require cloud-free reference images, and this may be difficult to achieve in time-series data since clouds are often prevalent and of varying spatial distribution in optical remote sensing images. Furthermore, current multi-temporal-based methods have limited feature extraction capability and rely heavily on prior assumptions. To address these issues, this paper proposes a Siamese network (Siamnet) for cloud and shadow detection based on Time-Series cloudy Images, namely TSI-Siamnet, which consists of two steps: 1) low-rank and sparse component decomposition of time-series cloudy images is conducted to construct a composite reference image to cope with dynamic changes in the cloud distribution in time-series images; 2) an extended Siamnet with optimal difference calculation module (DM) and multi-scale difference features fusion module (MDFM) is constructed to extract reliable disparity features and alleviate semantic information feature dilution during the decoder part. TSI-Siamnet was tested extensively on seven land cover types in the well-known Landsat 8 Biome dataset. Compared to six state-of-the-art methods (including four deep learning-based methods and two classical non-deep learning-based methods), TSI-Siamnet produced the best performance with an overall accuracy of 95.05% and MIoU of 84.37%. In three more challenging experiments, TSI-Siamnet showed enhanced detection of thin and broken clouds and greater anti-interference to highly reflective surfaces. TSI-Siamnet provides a novel strategy to explore comprehensively the valid information in time-series cloudy images and integrate the extracted spectral-spatial-temporal features for reliable cloud and shadow detection.

Keywords: Cloud and shadow detection, deep learning, Siamese network (Siamnet), time-series.

1. Introduction

The common existence of clouds in optical remote sensing images greatly limits their application. To make more effective use of optical remote sensing images, for example, by removing and potentially replacing cloud pixels, cloud and shadow detection is first required.

Traditionally, cloud and shadow detection has been realized by manual annotation. This scheme, however, requires a large number of human and material resources, and is not suitable for large scale data processing. Meanwhile, there always exists a certain degree of subjectivity in manual annotation, and the assessments made by different experts can vary. To meet the requirements of large scale data processing, several automatic cloud and shadow detection algorithms have been developed. These algorithms can be divided into two categories: mono-temporal-based and multi-temporal-based (Zhu and Helmer, 2018).

Mono-temporal-based cloud detection algorithms are usually performed based on the physical characteristics of clouds, such as a high reflectance in the visible, near-infrared and mid-infrared bands and low brightness in the thermal infrared band. On the basis of these physical characteristics, clouds can be distinguished from the background by defining thresholds in spectral space. The USGS proposed the automated cloud-cover assessment algorithm based on 26 spectral thresholds for Landsat7 ETM+ images (Irish et al., 2006). The approach estimates the percentage of clouds in each image, but cannot obtain the accurate location of clouds. Luo et al. (2008) separated clouds from clear backgrounds by setting thresholds for spectral combinations of each MODIS band. Huang et al. (2010) considered adaptive thresholds to perform cloud detection based on the mean and standard deviation of each band of Landsat images. Choi (2004) proposed an adaptive normalized difference snow index (NDSI) threshold-based method for cloud detection in snow and ice covered areas. In addition to spectral features, cloud shape and texture features were also used in cloud and shadow detection. Li et al. (2017a) designed a multi-feature combined algorithm for cloud and shadow detection in GF-1 WFV data, which forms a preliminary cloud mask by segmentation based on spectral feature thresholds, and then optimizes the preliminary cloud mask with geometric and textural features of the target cloudy image. Zhu and Woodcock (2012) proposed an adaptive threshold-based method called Function of mask (Fmask) for cloud and shadow detection in Landsat images (Zhu et al., 2015; Qiu et al., 2019). Given its various benefits, the Fmask algorithm is now used widely by the USGS for quality assessment (QA) of Landsat 4-8 Level 1 and Level 2 products (Foga et al., 2017). In general, threshold-based methods can be effective for the identification of large thick clouds, but its performance can be greatly compromised for the detection of thin clouds and broken clouds. In addition, when brighter backgrounds (artificial surface, desert, snow, etc.) are involved, false detections are highly likely to occur (Jedlovec and Haines, 2007).

With advances in computer technology, machine learning-based methods have been applied widely for cloud and shadow detection. For example, Xu et al. (2013) employed decision trees to extract cloud boundaries from MODIS images. Random forests and support vector machines were also applied to cloud and shadow detection (Hu et al., 2015; Yuan and Hu, 2015). As a branch of machine learning, deep learning has received increasing attention in recent years due to its ability to extract features automatically. It was applied extensively to classification tasks for remote sensing images (Zhu et al., 2017; Mountrakis et al., 2018; Yuan et al., 2020; Zhang et al., 2018; Li et al., 2017b; Karakizi et al., 2018). As cloud and shadow detection is a typical classification task, deep learning is also applicable to cloud and shadow detection (Chai et al., 2019; Choubin et al., 2019; Ghassemi and Magli, 2019; Shendryk et al., 2019; Segal-Rozenhaimer et al., 2020; Wei et al., 2020; Wu et al., 2021). As an example, Mateo-García et al. (2017) designed a simple convolutional neural network (CNN) model to detect clouds in multi-spectral Proba-V images, which produced more accurate results than traditional machine learning-based algorithms (e.g., gradient boosting machines). Xie et al. (2017) further distinguished thin clouds, thick clouds and cloud shadows based on the CNN model. Zi et al. (2018) developed a PCANet-based algorithm for cloud and shadow detection of Landsat 8 images. Li et al. (2019) proposed a deep learning-based multi-scale convolutional feature fusion method, which is universal for multi-source sensors. The methods produced satisfactory cloud and shadow detection results in both GF-1 and Landsat 8 images. Jeppesen et al. (2019) developed the RS-Net with an encoder and decoder structure based on the existing U-net framework (Ronneberger et al., 2015). Wieland et al. (2019) also developed a multi-sensor cloud detection method based on the Unet (MUnet). Yu et al. (2020) proposed a new two-branch CNN structure, called multi-scale fusion gated network, to extract shallow and deep information by introducing pyramidal attention and spatial attention modules. Zhang et al. (2020) proposed a network based on the Gabor transformation and a dark channel subnet attention mechanism, which can learn texture feature information more effectively. Recently, Zhang et al. (2021) proposed a UD-Net, which introduces wavelet transform-based upsampling and downsampling blocks in a symmetric encoder and decoder structure to reduce information loss and enhance the texture features of clouds, which can effectively detect thin clouds. Recently, Chai et al. (2024) proposed a shallow CNN (SCNN) consisting of only three convolutional layers, without using pooling layers or normalization layers. The SCNN greatly reduces training costs, while achieving reliable cloud detection results. In addition, migration learning and weakly supervised learning strategies were also developed to address the limitations of deep learning algorithms that require large numbers of training data (Guo et al., 2022; Li et al., 2020; Zhao et al., 2022; Zou et al., 2019).

Unlike mono-temporal-based methods, multi-temporal-based methods treat cloud and shadow detection as a change detection problem. The values (e.g., reflectance or brightness, etc.) of cloud pixels usually change more dramatically than the background in a given time-series of images. Thus, cloud pixels can be identified by detecting the changed parts. Benefiting from the use of temporally neighboring images, this type of method can reduce missed detection of thin clouds and attenuate the interference of background with high brightness (Cayula and Cornillon, 1996; Ricciardelli et al., 2008). Wang et al. (1999) found that by using a cloud-free image of the same area as the reference, clouds in the Landsat image could be detected effectively by defining suitable thresholds. The multi-temporal cloud detection method proposed by Hagolle et al. (2010) detected clouds based on temporal changes of the blue band and spatial correlation between adjacent pixels. Chen et al. (2016) proposed an iterative optimal cloud transformation algorithm based on a cloud-free image to distinguish cloud pixels automatically from the background. Generally, these methods are highly dependent on the cloud-free reference images. For most optical satellite sensors (e.g., Landsat series), the valid (i.e., inherently cloud-free) temporally adjacent observations can be several months apart, limiting the applicability of these methods to some extent. Some studies synthesized relatively clean reference images for the target cloudy image by linear regression, which typically require at least three cloud-free images (Gómez-Chova et al., 2017; Goodwin et al., 2013; Mateo-García et al., 2018). Again, this type of strategy is influenced by the time interval between the cloud-free and target cloudy data. That is, the performance is compromised when the time interval is long. Different from the abovementioned methods, Zhu and Woodcock (2014) proposed the multiTemporal mask (Tmask) algorithm that does not require a cloud-free image as reference. It first generates an initial cloud mask for each image using the Fmask algorithm, and then simulates the change of pixel reflectance based on multi-temporal reflectance data. Finally, the cloud mask is optimized by comparing the model predictions with the actual observations. Compared with the Fmask algorithm based on a mono-temporal image, the accuracy of Tmask is increased obviously. For the Tmask method, however, each pixel needs at least 15 corresponding cloud-free pixels along the time-series, which can be demanding in some cases. An automatic method for screening clouds and cloud shadows (ATSA) proposed by Zhu and Helmer (2018) can deal with cloud in the time-series, which first highlights the cloud features by calculating the haze optimal transformation (HOT) index and then optimizes cloud detection results based on the HOT of the time-series. Additionally, some studies developed deep learning-based methods for detecting clouds in the time-series meteorological satellite data with very fine temporal resolution (up to minutes) (Tuia et al., 2018; Mateo-Garcia et al., 2019). These methods aim to detect the clouds in the time-series jointly, where the cloud labels for all images in the time-series are required in the training models.

In general, the trend is towards the utilization of multi-temporal images to increase cloud and shadow detection accuracy. Multi-temporal-based algorithms can effectively cope with cloud and shadow detection in complex scenes, such as the detection of thin and broken clouds and the interference of highly reflective artificial surfaces, thus, generating more accurate cloud masks than mono-temporal based algorithms. However, challenges remain for this type of algorithm. Specifically, it requires cloud-free reference images in the time-series. As mentioned above, however, cloud can exist persistently in a time-series, and the spatial location of clouds varies greatly along the time-series, making cloud-free temporal neighbors commonly difficult to obtain. A simple strategy is to look for cloud-free images with long intervals as reference, but this usually involves great land cover changes, introducing a new source of uncertainty into the cloud and shadow detection task. In addition, most of existing multi-temporal-based cloud and shadow detection algorithms identify clouds and shadows through human-extracted features with customized thresholds, limiting the accuracy of cloud and shadow detection.

To overcome the abovementioned issues, a new multi-temporal-based cloud and shadow detection algorithm, that is, a Siamese network (Siamnet) based on Time-Series cloudy Images (TSI-Siamnet), is proposed in this paper. The objectives are two-fold. The first is to deal with the prevalent cloud contamination (with dynamic spatial distribution) in the time-series and simulate a reference image for multi-temporal-based cloud and shadow detection. Accordingly, this paper synthesizes a composite reference image with suppressed cloud contamination based on the low-rank sparse decomposition method in the first step of TSI-Siamnet. The core idea is to fully utilize the valuable information in the partial cloud-free data in the time-series. Specifically, the non-cloud background and dynamically changed clouds in the time-series data are regarded as the low-rank and sparse components, respectively. The second objective is to develop a new deep learning-based method with more reliable feature extraction capability. Traditional model-based cloud and shadow detection methods always have limited feature extraction capability and depend heavily on prior assumptions. In contrast, deep learning is capable of extracting multi-scale and high-level features automatically without any specific assumptions. To this end, in the second step of TSI-Siamnet, we proposed an extended Siamnet-based cloud and shadow detection method, which extracts the features of the target cloudy image and the composite reference image separately by designing a dual branch with shared weights. An optimal difference calculation module (DM) was proposed to extract the optimal difference features, while a multi-scale difference features fusion module (MDFM) was designed to remedy the information loss during the fusion of multi-scale difference feature maps. Furthermore, the attention mechanism was also added to the convolutional layer to enhance the ability to extract reliable features.

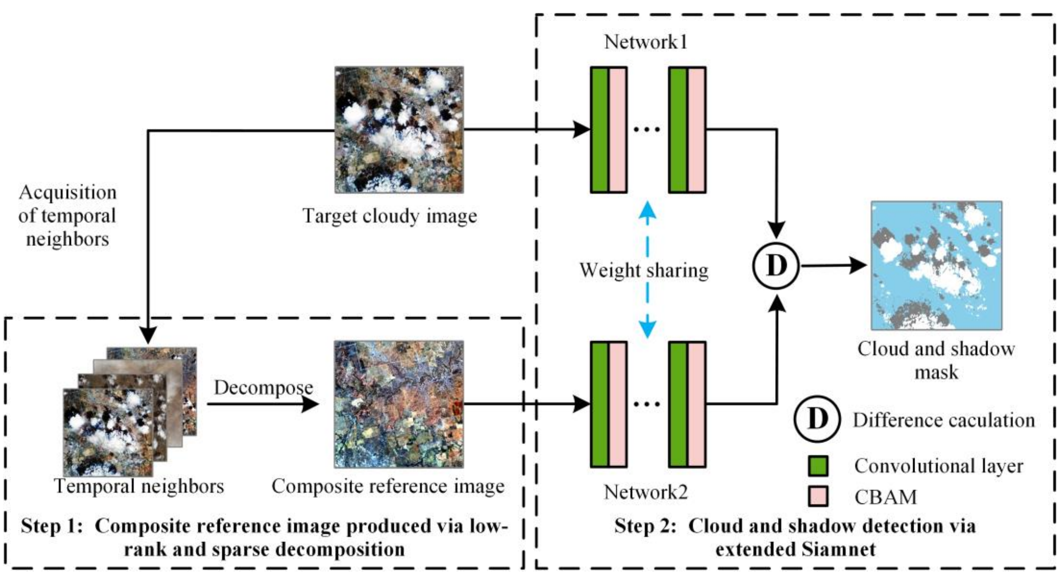


Fig. 1. The overall framework of the proposed TSI-Siamnet.

2. Methods

Fig. 1 illustrates the overall framework of the proposed TSI-Siamnet. TSI-Siamnet aims to create a usable reference image based on the time-series with prevalent cloud contamination and develop a deep learning-based method with powerful feature extraction capability, which are achieved by two steps. In the first step, for the target cloudy image, we constructed a time-series dataset by selecting images with low cloudiness within two years at the corresponding location. Then, a low-rank and sparse decomposition analysis was applied to the time-series data to produce the composite reference image with suppressed cloud contamination. In the second step, the extended Siamnet was developed, and the target cloudy image and the composite reference image were fed into the network as image pairs, where the cloud was identified by comparing the two input images using the network. Note that TSI-Siamnet detects the cloud and shadow in each target cloudy image separately, rather than simultaneously in the time-series cloudy images. The two steps will be described in detail in Sections 2.1 and 2.2.

2.1. Composite reference image construction via low-rank and sparse decomposition

Using a temporally neighboring image as reference, multi-temporal-based cloud and shadow detection can reduce the interference of background. However, cloud contamination is also a common issue in the temporally neighboring time-series. In applications, cloud-free images with long intervals are used as reference alternatively. That is, the temporally neighboring, cloudy images are always abandoned directly, although they may contain low percentage of cloud contamination. This scheme, however, usually involves new uncertainty, due to great land cover changes introduced by the temporally further, cloud-free images. The purpose of the first step of TSI-Siamnet is to simulate a more reliable reference image directly based on the temporally neighboring time-series with prevalent, dynamically changed clouds. This is realized by fully exploring the valuable information in the partial cloud-free data (also with dynamic spatial distribution) in the time-series.

As acknowledged widely, in the time-series data, the images are highly correlated with each other in both the spectral and temporal dimensions (Wang et al., 2016). Moreover, the cloud-free background usually remains constant or changes slightly in a short period, which means it has low-rank. In contrast, in the case of cloudy data with lower occupancy than the background, clouds and shadows induce significant variation in the time-series. That is, the assumption of a sparse prior is satisfied. Therefore, we can extract the cloud-free background from the time-series data by low-rank and sparse components decomposition.

In this paper, we adopted robust principal component analysis (RPCA) (Candes et al., 2009) to decompose the low-rank and sparse components. Specifically, for a set of time-series data with N images (with the same spatial size and number of bands), we constructed a new matrix $\mathbf{D}=\mathbf{R}^{HW\times CN}$, where H, W and C represent the height, width and number of bands of each image, respectively. In this matrix, each column represents a spectral band and each row corresponds to a target pixel with data from all bands and time-series at the geographic location. In the task of cloud and shadow detection, the matrix $\mathbf{D}$ is assumed to be composed of two parts: low-rank clear background and sparse clouds and shadows. Accordingly, the mathematical model is described as follows:

$$\mathbf{D}=\mathbf{L_b}+\mathbf{S_c} \tag{1}$$

where $\mathbf{L_b}$ and $\mathbf{S_c}$ represent the low-rank part due to clear background and sparse part due to clouds, respectively.

To decompose the low-rank and sparse parts, a corresponding prior restriction is imposed as follows:

$$\min_{\mathbf{L_b},\mathbf{S_c}} rank(\mathbf{L_b})+\lambda\|\mathbf{S_c}\|_0 \quad s.t.\ \mathbf{D}=\mathbf{L_b}+\mathbf{S_c} \tag{2}$$

where $rank(\mathbf{L_b})$ denotes the rank of the low-rank matrix $\mathbf{L_b}$ and $\|\mathbf{S_c}\|_0$ denotes the L0-norm of the sparse matrix $\mathbf{S_c}$. λ represents the weight of the sparse part, which is set to $\lambda=1/sqrt(max(HW,CN))$ as the default. When $rank(\mathbf{L_b})$ is small enough, $\mathbf{L_b}$ is considered to be ideally low-rank. L0-norm refers to the number of non-zero elements in the matrix $\mathbf{S_c}$, and fewer non-zero elements means that $\mathbf{S_c}$ is sparser.

In principle, low-rank sparse decomposition can be achieved by optimizing the two components in Eq. (2), $rank(\mathbf{L_b})$ and $\lambda\|\mathbf{S_c}\|_0$, under the sum constraint. However, Eq. (2) is a non-convex optimization problem that cannot be solved directly. Thus, we transferred Eq. (2), into a convex optimization problem. Specifically, $rank(\mathbf{L_b})$ is replaced by the nuclear norm of $\mathbf{L_b}$, which denotes the sum of singular values in the matrix. That is, when the nuclear norm is smaller, the rank can be approximated as lower. In addition, we substituted the L0-norm with the L1-norm. The L1-norm takes the maximum value of the sum of the absolute values of the matrix $\mathbf{S_c}$ along the column dimension. In the process of minimizing the L1-norm, when the element of matrix $\mathbf{S_c}$ is less than the threshold defined by λ, it will be assigned a value of zero to ensure the sparse property of matrix $\mathbf{S_c}$. The transformed convex optimization can be described as follows:

$$\min_{\mathbf{L_b},\mathbf{S_c}} \|\mathbf{L_b}\|_* + \lambda\|\mathbf{S_c}\|_1 \quad s.t.\ \mathbf{D}=\mathbf{L_b}+\mathbf{S_c} \tag{3}$$

where $\|\mathbf{L_b}\|_*$ denotes the nuclear norm of $\mathbf{L_b}$ and $\|\mathbf{S_c}\|_1$ denotes the L1-norm of $\mathbf{S_c}$. In this paper, we used the alternating direction method of multipliers (ADMM) to optimize this model (Boyd, 2010).

2.2. Cloud and shadow detection via extended Siamnet

Fig. 2 illustrates the structure of our extended Siamnet, where the features of input image pairs are extracted separately by a double branch with shared weights. To extract adaptively the refined features in the cloud and shadow detection task, we added an attention mechanism to each convolutional layer. Meanwhile, we proposed the optimal difference calculation module (DM) to derive the optimal disparity features. Furthermore, to alleviate the dilution of semantic information features during the decoder stage, we proposed the multi-scale difference features fusion module (MDFM) for enhancement.

As shown in Fig. 2, the target cloudy and composite reference images were fed into the extended Siamnet model as image pairs. The features at different scales were then extracted with five blocks (i.e., Blocks 1-5) equipped with the convolutional block attention module (CBAM) (Woo et al., 2018). The details of each block are shown in Table 1, where each block has a convolution layer (Conv2D) with a kernel size of 3×3 pixels and increasing kernel number layer-by-layer. The rectified linear unit (ReLU) was adopted as the activation function for each layer, batch normalization (BN) was adopted to prevent the gradient vanishing, and L2 regularization (L2) was utilized to avoid overfitting. The feature disparities of the two branches at multi-scales were obtained by the DM, which uses relative distance instead of absolute distance. Afterwards the multi-scale feature disparities were upsampled and fused by the MDFM. Finally, a softmax classifier was used to generate the final cloud mask. The three main modules, CBAM, DM and MDFM, are described in detail below.

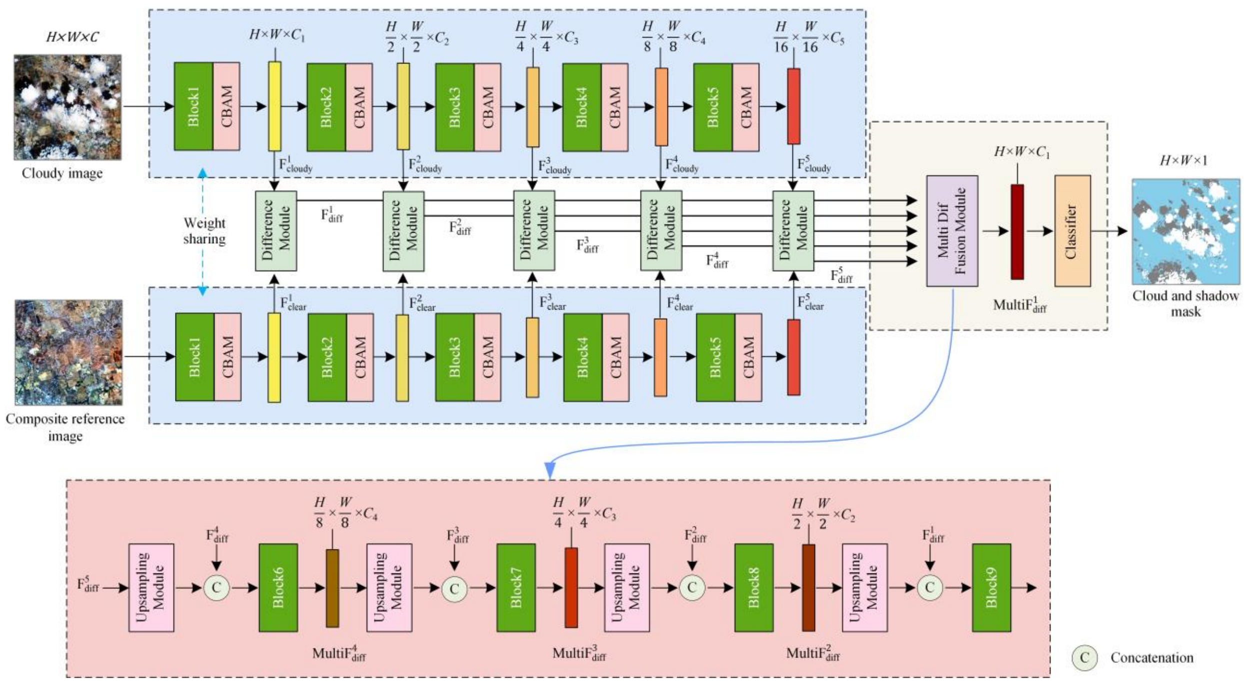


Fig. 2. The structure of the extended Siamnet (*H*, *W* and *C* represent the height, width and channels of the target cloudy image, $\mathbf{F}^{\mathbf{i}}_{\mathbf{cloud}}$ and $\mathbf{F}^{\mathbf{i}}_{\mathbf{clear}}$ represent the feature map of the target cloudy image and composite reference image in the *i*-th layer, $\mathbf{F}^{\mathbf{i}}_{\mathbf{diff}}$ represents difference between $\mathbf{F}^{\mathbf{i}}_{\mathbf{cloud}}$ and $\mathbf{F}^{\mathbf{i}}_{\mathbf{clear}}$ and $\mathbf{MultiF}^{\mathbf{i}}_{\mathbf{diff}}$ represents the fusion result of $\mathbf{F}^{\mathbf{i}}_{\mathbf{diff}}$ in the different layers).

Table 1 The structure of the basic convolutional Blocks (Conv2d represents a 2D convolution layer, ReLU represent rectified linear unit, BN represent bach normalization and L2 represents L2 regularization).

Convolution Blocks	Type	Kernel number	Kernel size	Padding
Block1	Conv2D + ReLU + BN + L2 (0.0005)	32	3×3	valid
Block2	MaxPooling2D	-	2×2	valid
	Conv2D + ReLU + BN + L2 (0.0005)	64	3×3	valid

Block3	MaxPooling2D	-	2×2	valid
	Conv2D + ReLU + BN + L2 (0.0005)	128	3×3	valid
Block4	MaxPooling2D	-	2×2	valid
	Conv2D + ReLU + BN + L2 (0.0005)	256	3×3	valid
Block5	MaxPooling2D	-	2×2	valid
	Conv2D + ReLU + BN + L2 (0.0005)	256	3×3	valid

2.2.1. The Convolutional Block Attention Module (CBAM)

CBAM is a simple and effective attention module for use in a feed-forward CNN. It infers attention results sequentially along the channel and spatial dimensions. Specifically, the input feature map is passed through the channel attention module, and then the output is fed into the spatial attention module.

The channel attention module compresses the input feature map in the spatial dimension, and then global max pooling and global average pooling are applied based on the size of the feature map. Average pooling generates feedback for all image elements, while max pooling produces feedback only for the position with the largest response. The outputs of average pooling and max pooling are then concatenated and multiplied with the input feature map. Generally, the channel attention module can be expressed as follows:

$$\mathbf{F_c} = \mathrm{ReLU}(\mathrm{Add}(\mathrm{MLP}(\mathrm{GAP}(\mathbf{F})), \mathrm{MLP}(\mathrm{GMP}(\mathbf{F})))) \otimes \mathbf{F} \quad (4)$$

where $\mathbf{F}$ represents the input feature map, $\mathbf{F_c}$ represents the result of the channel attention module, MLP represents the multi-layer perception and is used to control the number of channels of the output, and GAP and GMP stand for global average pooling and global max pooling, respectively.

The output of the channel attention module is then fed into the spatial attention module, in which the channel dimension is compressed based on the channels of the feature map. In the spatial attention module, global average pooling and global max pooling are performed to extract the average and maximum values in the channels, and the two pooling results are then fused and multiplied with the channel attention result. The spatial attention module can be expressed as follows:

$$\mathbf{F_s}=\mathrm{ReLU}(\mathrm{Conv2D}_{3\times3}(\mathrm{Concatenation}(\mathrm{GAP}(\mathbf{F_c}),\ \mathrm{GMP}(\mathbf{F_c}))))\otimes\mathbf{F_c} \tag{5}$$

where $\mathbf{F_c}$ represents the output of the channel attention module and $\mathbf{F_s}$ represents the result of the spatial attention module.

2.2.2. The Difference Module (DM)

In traditional approaches to change detection, the absolute distance (e.g., Euclidean distance) is commonly used to calculate the disparity between images at different times. However, the absolute distance result is a two-dimensional feature map with limited information for subsequent feature extraction. Therefore, in this paper, we developed a relative distance module DM to extract optimal feature maps representing differences at various scales. Specifically, in each layer, DM first concatenates the feature maps of the target cloudy and composite reference images, after which a convolutional layer with a kernel number of 64 and a kernel size of 3×3 pixels is used. Through the DM, the optimal features of the difference are learnt and extracted by the network instead of straightforward distance calculation. The DM can be expressed as follows:

$$\mathbf{F^i_{diff}}=\mathrm{BN}(\mathrm{ReLU}(\mathrm{Conv2D}_{3\times3}(\mathrm{Concatenation}(\mathbf{F^i_{cloud}},\ \mathbf{F^i_{clear}})))) \tag{6}$$

where $\mathbf{F^i_{diff}}$, $\mathbf{F^i_{cloud}}$ and $\mathbf{F^i_{clear}}$ denote the feature maps of the difference feature map, target cloudy and composite reference image in the *i*-th stratified layer, respectively.

2.2.3. The Multi-scale Feature Difference Maps Fusion Module (MDFM)

Unlike typical segmentation algorithms (e.g., PSPnet (Zhao et al., 2017)) that directly upsample the final feature map multiple times until it matches the size of the input image, we proposed MDFM, in which the difference feature maps at different scales are first upsampled and then concatenated several times. Specifically, for five difference feature maps with various scales, the difference feature map in the deepest layer (i.e., the coarsest difference feature map) is upsampled and then concatenated with the map of the previous layer. The concatenation result is then upsampled and concatenated with the previous one iteratively until it matches the size of the target cloudy image. Block 6-9 are convolution layers with a kernel size of 3×3 pixels and a kernel number of 64. We adopted a 2×2 bilinear sample, followed by a convolutional layer of size 3×3 pixels to implement the upsampling. The details are as follows:

$$\mathbf{Up}(\mathbf{F^i_{diff}})=\mathrm{BN}(\mathrm{ReLU}(\mathrm{Conv2D}_{3\times3}(\mathrm{Upsample}(\mathbf{F^i_{diff}}, (2H, 2W), \text{"bilinear"})))) \tag{7}$$

where $\mathbf{Up}(\mathbf{F^i_{diff}})$ represents the upsampling result of the difference feature map $\mathbf{F^i_{diff}}$ and (H, W) represents the original size of the difference feature map $\mathbf{F^i_{diff}}$.

To fully fuse the information in difference features representing the different scales, difference feature maps are concatenated with the corresponding upsampled maps according to Eqs. (8) and (9):

$$\mathbf{MultiF^4_{diff}}=\mathrm{BN}(\mathrm{ReLU}(\mathrm{Conv2D}_{3\times3}(\mathrm{Concatenation}(\mathbf{Up}(\mathbf{F^5_{diff}}), \mathbf{F^4_{diff}})))) \tag{8}$$

$$\mathbf{MultiF}_{\mathbf{diff}}^{\mathbf{i}}=\mathrm{BN}(\mathrm{ReLU}(\mathrm{Conv2D}_{3\times 3}(\mathrm{Concatenation}(\mathbf{Up}(\mathbf{MultiF}_{\mathbf{diff}}^{\mathbf{i+1}}), \mathbf{F}_{\mathbf{diff}}^{\boldsymbol{i}})))) \quad s.t.\ i=3, 2, 1 \quad (9)$$

where $\mathbf{MultiF}_{\mathbf{diff}}^{\mathbf{4}}$ represents the fusion of the upsampled disparity feature map in the 5-th layer and the original disparity feature map in the 4-th layer, $\mathbf{MultiF}_{\mathbf{diff}}^{\mathbf{i}}$ represents the fusion of the upsampled $\mathbf{MultiF}_{\mathbf{diff}}^{\mathbf{i+1}}$ and the original difference feature map in the *i*-th layer.

3. Experiments

3.1. Datasets

In the experiments, the popular Landsat 8 Biome data were used for demonstration. The data were provided by Foga et al. (2017) and have been used widely for training and testing deep learning models for cloud and shadow detection. The mask for the Landsat 8 Biome data contains thick clouds, thin clouds, cloud shadows and background. In this study, thick and thin cloud are uniformly classified as cloud. The original Landsat 8 Biome data consist of 96 images that are evenly distributed across the globe and cover eight land cover types, including barren, forest, grass/crops, shrubland, urban, water, wetlands and snow/ice. As acknowledged widely, it is a very challenging task to detect the cloud in the snow and ice covered areas, the data used in the experiments do not cover snow/ice. The used Landsat 8 Biome data are TOA-corrected, and include seven bands (i.e., bands 1-7). Training and testing data are evenly distributed among seven land cover types without any overlap, and the training data consist of 756 images with a spatial size of 256×256 pixels, while the testing data contain 336 images with a size of 256×256 pixels. For each image, we selected 10-to-15 (within two years) temporally closest images with relatively low cloudiness (less the 50%) for RPCA to construct the composite reference images.

3.2. Experimental setup

3.2.1. Benchmark Methods

In this paper, three mono-temporal deep learning methods (i.e., MUnet, DeepLabV3+ (Chen et al., 2018) and PSPnet) and one multi-temporal deep learning method (i.e., CDUnet++ (Peng et al., 2019)) were compared with the proposed TSI-Siamnet method. Moreover, two non-deep learning methods were also used as benchmark methods, including one multi-temporal-based method (i.e., ATSA), and one mono-temporal-based method (i.e., the classical thresholding method Fmask). MUnet is a typical deep learning network with an encoding-decoding structure that achieved satisfactory results in cloud and shadow detection task. DeepLabV3+ and PSPnet are representative networks for image segmentation, which are also fully applicable for cloud and shadow detection. CDUnet++ was originally designed for change detection. In this paper, to facilitate its application in cloud and shadow detection, the input data are consistent with that for TSI-Siamnet. The ATSA algorithm produced reliable results for Landsat-8 OLI, Landsat-4 MSS and Sentinel-2 data. The Fmask method was applied widely by the USGS to produce cloud masks for Landsat data. All benchmark methods were implemented using publicly available codes, and were adjusted accordingly to accommodate multi-band remote sensing images. The traditional physics-based algorithms employed default thresholds, while deep learning-based algorithms used uniform parameter settings as described in Section 3.2.3. It should be noted that no pretrained models were used in the experiments.

3.2.2. Accuracy metrics

In this study, the user accuracy (UA), producer accuracy (PA), and intersection of union (IoU) were used to evaluate the accuracy of detection of each identified class (i.e., clear background, cloud and shadow). Among these, UA and PA correspond to omission and commission errors, respectively. In addition, the mean IoU (MIoU) and overall accuracy (OA) were also used for comprehensive accuracy evaluation of all classes. All the metrics were calculated by referring to the reference labels of the Landsat 8 Biome data.

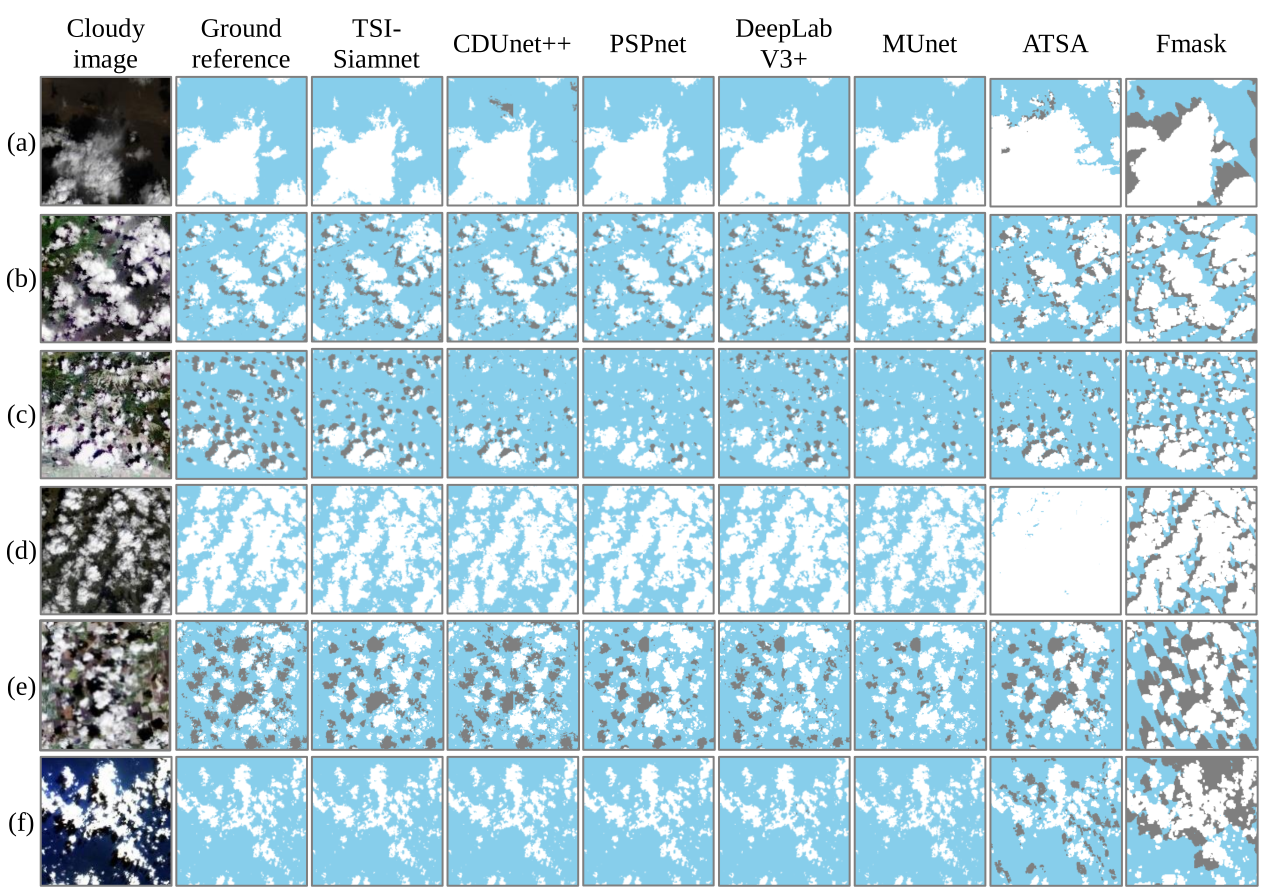
3.2.3. Hyperparameters

All the deep learning-based methods applied here adopted the same hyperparameter settings. More precisely, the batch size and epoch were set to 8 and 200, respectively. The Adam optimizer was used to optimize the parameters of all the networks. The learning rate was set to 0.001 in the first 100 epochs and 0.0001 in the second 100 epochs. All the deep learning-based methods were implemented using TensorFlow version 2.6.0 on a single NVIDIA GTX 3060Ti GPU with 32-GB memory.

3.3. Results

3.3.1. Qualitative evaluation

Fig. 3 shows visually the results for all seven methods. We selected one image from each of the seven land cover types for display. It can be seen from Fig. 3 that Fmask detects more cloud pixels than the other methods, which is mainly related to the constructed 3×3 buffer in the method. ATSA shows greater consistency with the ground reference data than Fmask, but its performance depends upon the quality of the used time-series data, and it exhibits more false positives in barren and shrubland scenes. Overall, the performances of Fmask and ATSA are not as satisfactory as for the five deep learning-based methods (e.g., the results of the barren and urban scenes). Furthermore, in the five deep learning-based methods, TSI-Siamnet is more accurate than the other four methods, especially in the barren, grass/crops and wetlands scenes, with obviously fewer omission and commission errors.



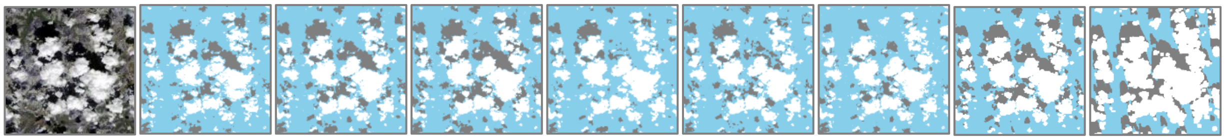
(g) 

Fig. 3. Cloud and shadow detection results of the seven different methods for a part of the testing images (one area was selected for each of the seven land cover types). (a)–(g) refer to the results of barren, forest, grass/crops, shrubland, urban, water and wetlands, respectively. True color composites (R: 4, G: 3 and B: 2) of the testing images are shown in the first column. White, gray and blue represent cloud, cloud shadow and background, respectively.

3.3.2. Quantitative evaluation

Table 2 lists the quantitative evaluation of the results of TSI-Siamnet against the six benchmark methods for all 336 testing images (the mean value of all 336 images was taken for each metric). The most accurate value under each metric is marked in bold. As can be seen from the table, Fmask produces the smallest IoU for both clouds and shadows, and has larger UA than PA for cloud and shadow, which is consistent with the visual results. Benefitting from the use of multi-temporal data, most of the metrics for ATSA are superior to those of Fmask. With respect to the five deep learning-based methods, they present greater accuracy than the traditional methods. For example, the OA and MIoU of all the five deep learning-based methods are at least 13.51% and 16.65% larger, respectively. Furthermore, compared with the three mono-temporal deep learning methods (MUnet, PSPnet and DeepLabV3+), CDUnet++ produces greater accuracay, indicating that the use of the composite reference image is beneficial. Finally, TSI-Siamnet is more accuate than CDUnet++. For example, the OA and MIoU of TSI-Siamnet are 0.96% and 3.46% larger, respectively, suggesting the advanatage of the developed extend Siamnet.

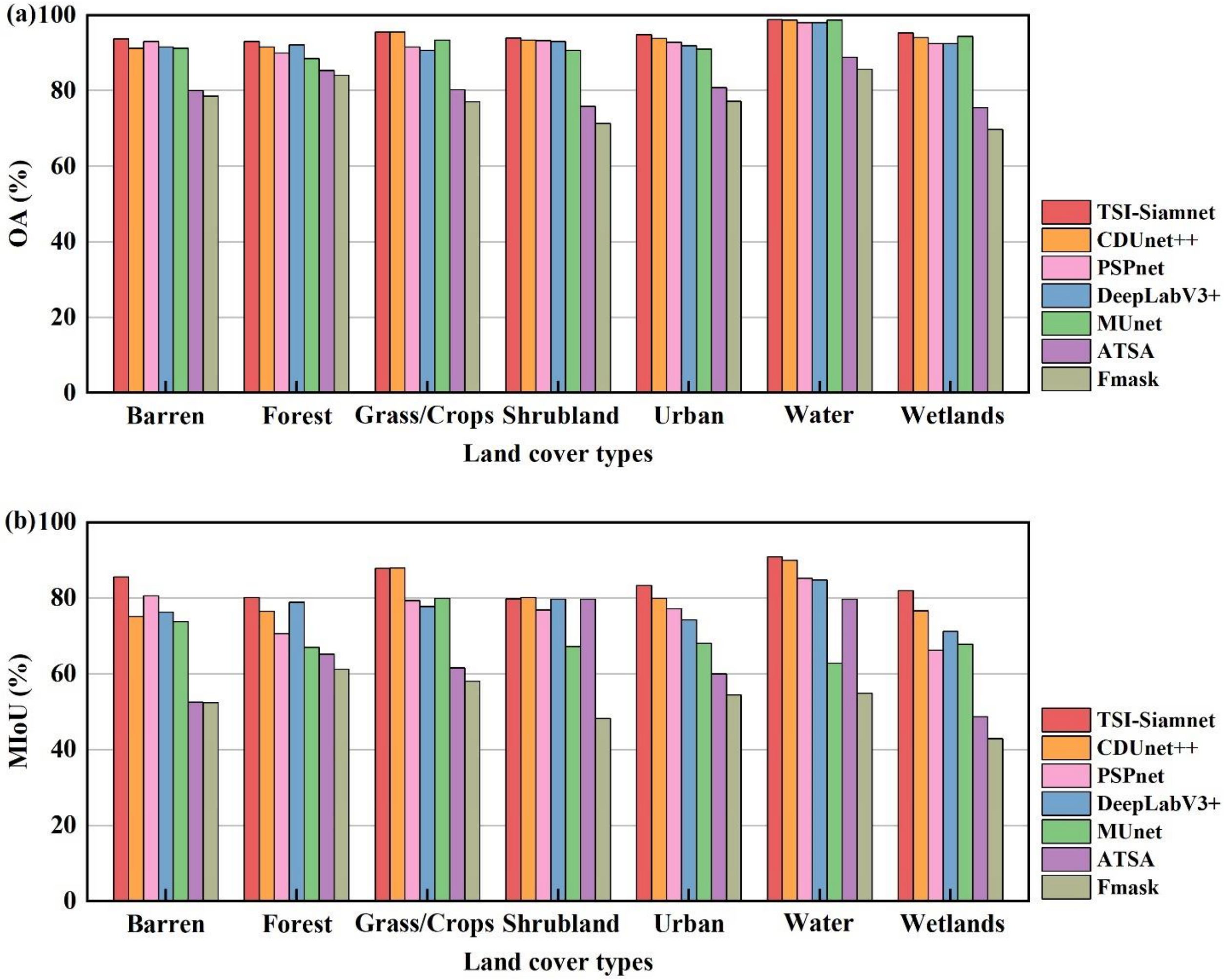


Fig. 4. The accuracy of the seven different methods for each land cover type. (a) OA. (b) MIoU.

Table 2 Accuracy metrics of the seven different methods for all 336 testing images (values of all 336 images were averaged for each metric; the **bold** value means the most accurate result under each metric).

		PA (%)	UA (%)	IoU (%)	OA (%)	MIoU (%)
	Clear	71.02	**96.16**	69.06		
Fmask	Cloud	96.66	68.84	67.24	77.68	54.98
	Shadow	57.23	36.43	28.63		
	Clear	75.03	90.44	69.52		
ATSA	Cloud	91.45	79.39	73.91	79.14	56.14
	Shadow	59.07	38.17	31.99		
MUnet	Clear	**98.22**	91.45	89.96	92.65	72.79

	Cloud	94.27	**95.31**	90.09		
	Shadow	39.39	**93.34**	38.32		
	Clear	95.75	93.90	90.14		
DeepLabV3+	Cloud	93.39	92.80	87.09	92.86	78.40
	Shadow	66.39	82.03	57.96		
	Clear	95.60	94.22	90.30		
PSPnet	Cloud	96.44	91.38	88.40	93.04	77.75
	Shadow	57.82	88.47	53.77		
	Clear	96.58	94.63	91.56		
CDUnet++	Cloud	95.30	94.95	90.70	94.09	80.91
	Shadow	68.28	84.11	60.48		
	Clear	96.42	96.14	**92.83**		
TSI-Siamnet	Cloud	**96.88**	94.48	**91.69**	**95.05**	**84.37**
	Shadow	**76.22**	87.24	**68.57**		

To quantitatively analyze the performance across different land cover types, we calculated the OA and MIoU of the methods for each land cover type separately, as shown in Fig. 4. Both the OA and MIoU of the deep learning-based methods are larger than those of Fmask and ATSA. Moreover, TSI-Siamnet produces the largest OA and MIoU in almost all land cover types, especially in barren and urban scenes. We further analyzed the stability of each method in Fig. 5. It can be seen that the deep learning-based methods tend to be more stable, but it must be noted that neither Fmask nor ATSA require data for training. Noticeably, TSI-Siamnet presents the most satisfactory performance in almost all land cover types.

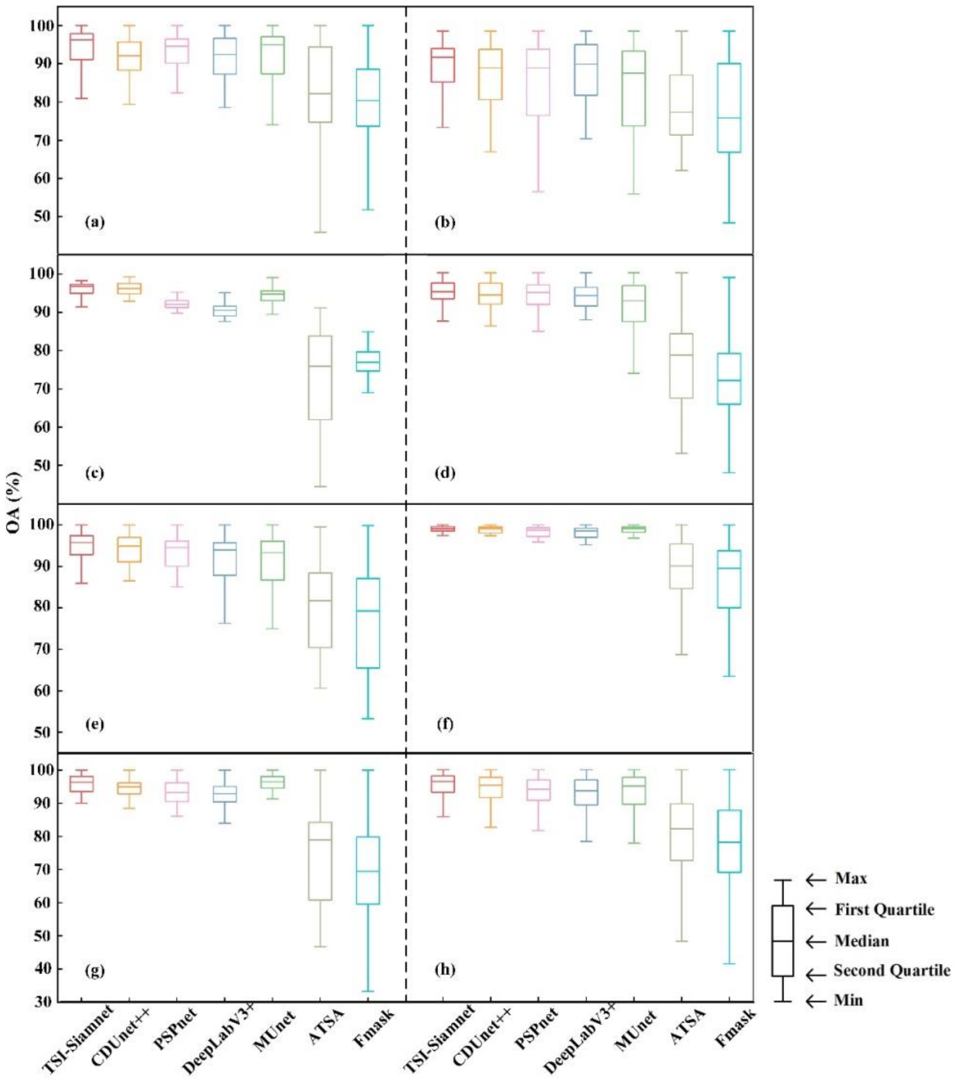


Fig. 5. The stability of the seven methods for different land cover types. (a) Barren. (b) Forest. (c) Grass/Crops. (d) Shrubland. (e) Urban. (f) Water. (g) Wetlands. (h) All classes.

3.3.3. Detection results in three challenging cases

To further examine TSI-Siamnet, we present the visual results for three challenging cases in Figs. 6-8. Fig. 6 shows the detection results for thin clouds. Thin cloud detection has always been a challenging issue due to the complexity of cloud information mixed with the background. As can be seen from Fig. 6, TSI-Siamnet shows greater ability to detect thin clouds than the benchmark methods, which can detect more thin clouds correctly than the other methods. For example, in Fig. 6(a) the detected thin cloud is much more consistent with that of the ground reference, especially in the part marked in yellow. CDUnet++, DeepLabV3+, MUnet and PSPnet result in more omission errors in thin cloud detection, while ATSA and Fmask present more commission errors.

Fig. 7 exhibits the detection results for broken clouds. Since the difference in brightness between broken clouds and background is generally small, the benchmark methods present noticeable omission errors. In general, TSI-Siamnet still produces more accurate broken cloud and shadow detection results. Checking the results marked in yellow in Fig. 7(b), TSI-Siamnet produces obviously smaller omission errors.

Highly reflective surfaces are a common interfering factor in cloud detection. That is, due to similar spectral characteristics, highly reflective surfaces are susceptible to being incorrectly detected as cloud. As shown in Fig. 8, the benchmark methods incorrectly detect several background pixels as cloud pixels, especially in the area marked in yellow. In contrast, the TSI-Siamnet method presents far fewer commission errors. Overall, TSI-Siamnet produces the most reliable performances for cloud and shadow detection in these three challenging cases.

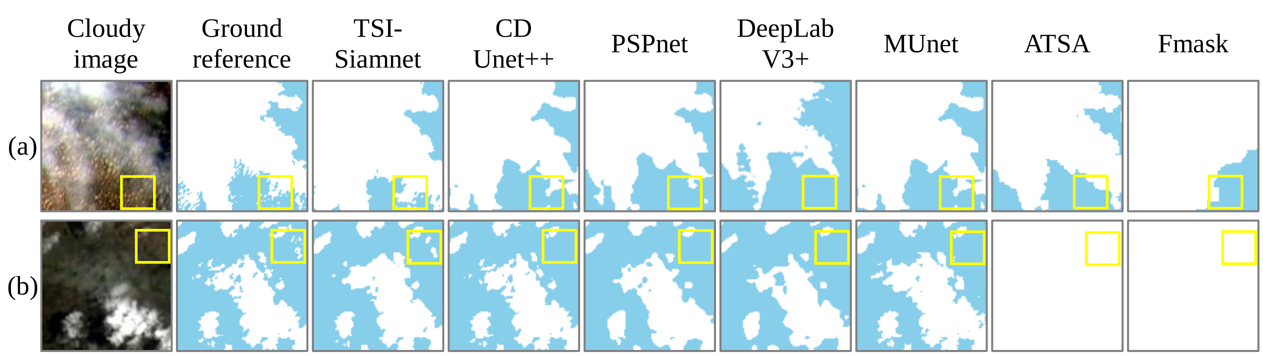


Fig. 6. Detection results for thin clouds. (a) Thin clouds over barren. (b) Thin clouds over wetlands. True color composites (R: 4, G: 3 and B: 2) of testing images are shown in the first column. White, gray and blue represent cloud, cloud shadow and background, respectively.

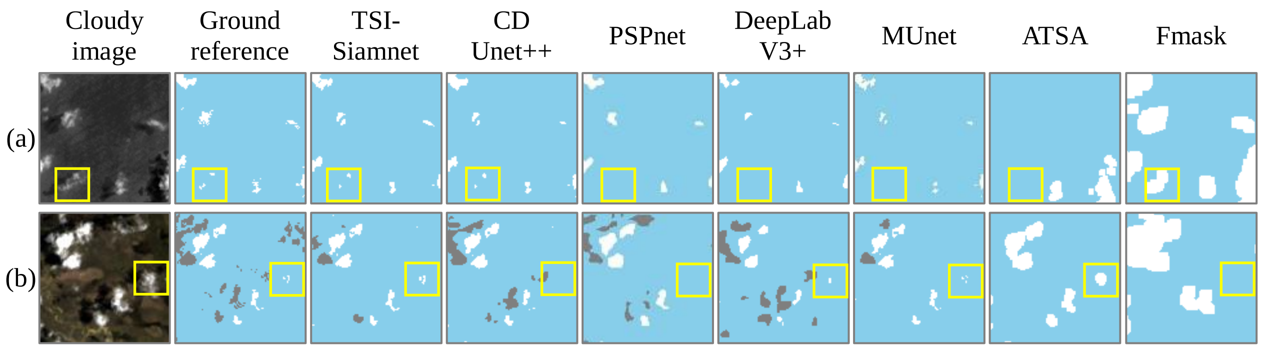


Fig. 7. Detection results for broken clouds. (a) Broken clouds above water. (b) Broken clouds over forest. True color composites (R: 4, G: 3 and B: 2) of testing images are shown in the first column. White, gray and blue represent cloud, cloud shadow and background, respectively.

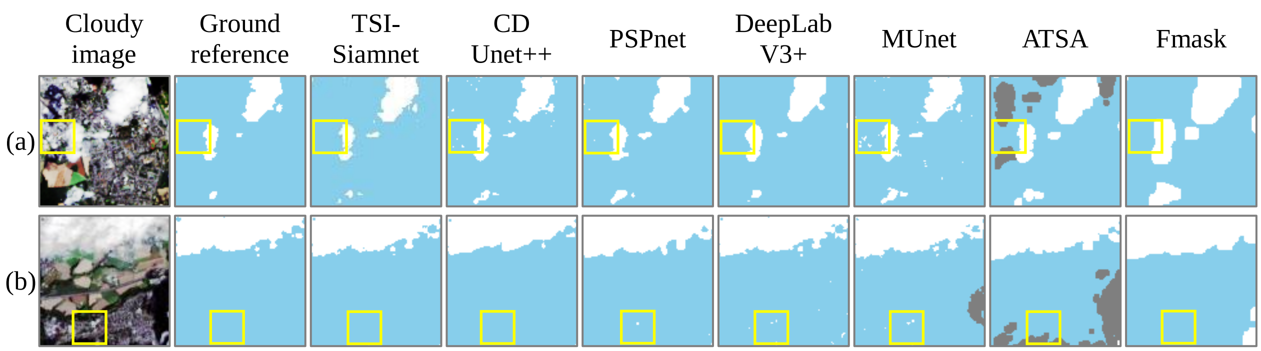
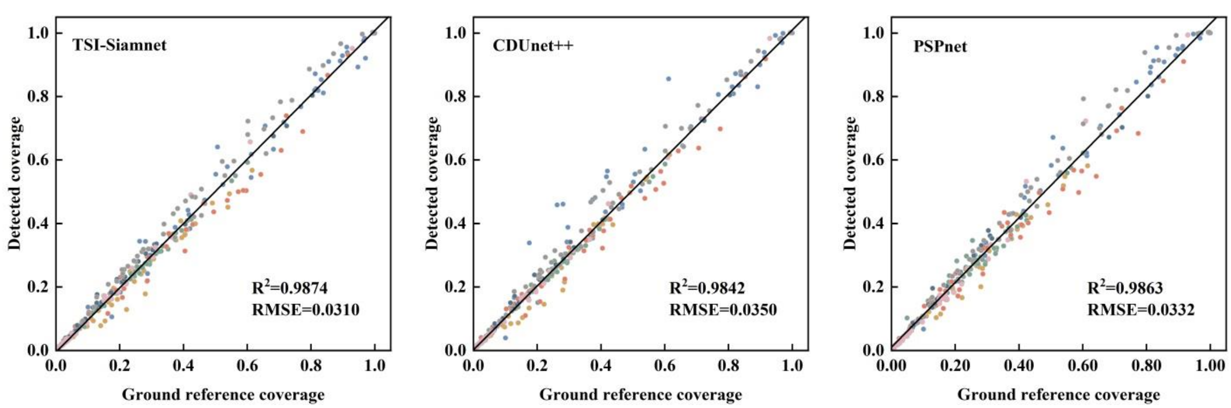


Fig. 8. Detection results for clouds above artificial surface with large reflectance. (a) and (b) are both clouds above surface with large reflectance. True color composites (R: 4, G: 3 and B: 2) of testing images are shown in the first column. White, gray and blue represent cloud, cloud shadow and background, respectively.



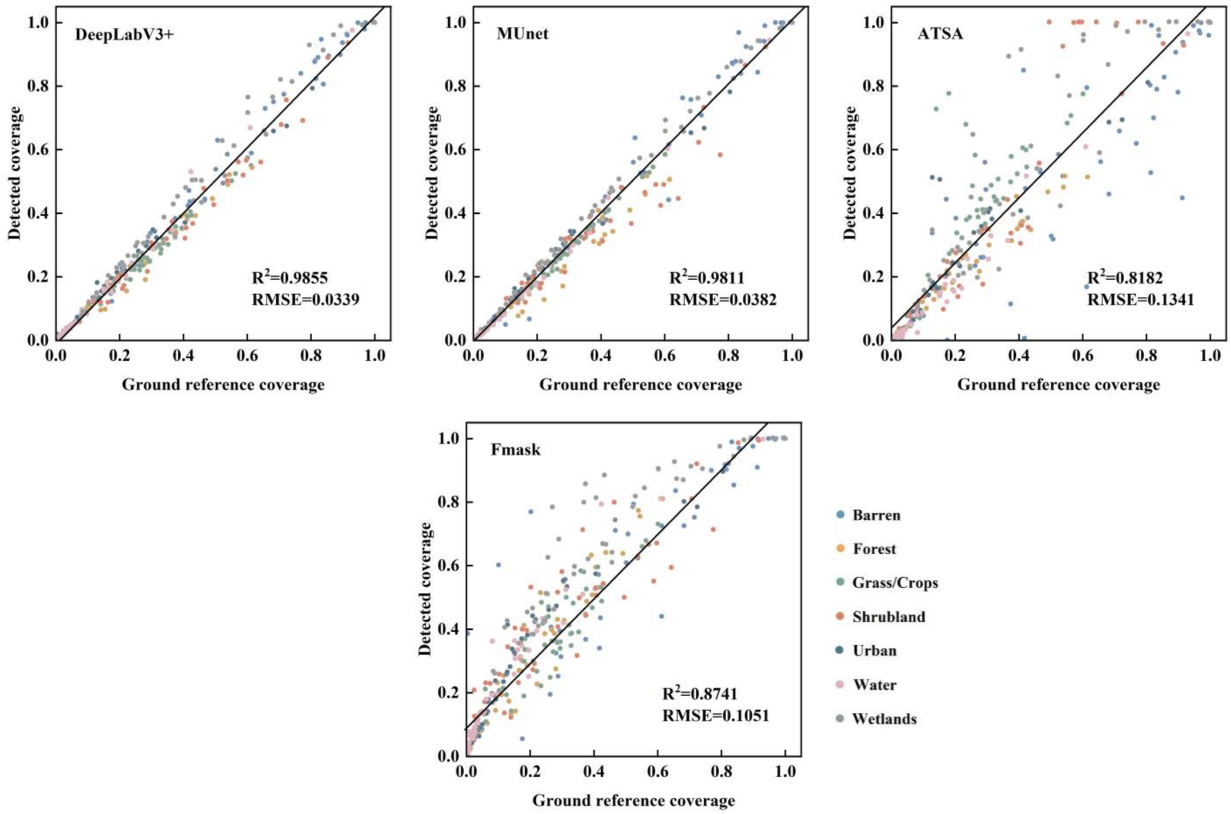


Fig. 9. The ground reference cloud coverage plotted against the detected cloud coverage for the seven different methods.

3.3.4. Evaluation based on cloud coverage

To evaluate the accuracy of TSI-Siamnet in cloud coverage estimation, a scatterplot of the detected cloud coverage in 336 images was compared against the ground reference (Fig. 9). It can be seen that TSI-Siamnet produces greater accuracy in cloud estimation, especially in the barren, urban and shrubland scenes. Specifically, the R^2 coefficient of TSI-Siamnet is the largest (i.e., 0.9874) and the root mean square error (RMSE) is the smallest (i.e., 0.0310). Table 3 shows the stability of cloud coverage estimation by calculating the mean absolute deviation and standard deviation of the estimation errors. It can be seen that the mean absolute deviation and standard deviation of TSI-Siamnet are obviously smaller than those of the benchmark methods, indicating that TSI-Siamnet is more stable for cloud coverage estimation.

Table 3 Statistical results of cloud coverage detection error in terms of mean absolute deviation and standard deviation.

	Mean absolute deviation	Standard deviation
TSI-Siamnet	0.0191	0.0237
CDUnet++	0.0201	0.0305
PSPnet	0.0251	0.0291
DeepLabV3+	0.0230	0.0265
MUnet	0.0229	0.0309
ATSA	0.0827	0.1156
Fmask	0.1253	0.1003

3.3.5. Cloud detection results in thick cloud areas

In regions with persistent cloud cover, images are often extensively obscured by clouds, and time-series images may not provide useful information for reference. That is, the reference image synthesized by RPCA may not provide effective information. To examine the performance of the proposed TSI-Siamnet method in this case, cloud detection for two thick cloud areas was performed and the results are shown in Fig. 10. It is seen that TSI-Siamnet still produces satisfactory detection results. More precisely, TSI-Siamnet produces an average OA of 98.05% for the two areas, which is 1.13%, 2.01%, 2.20%, 0.20%, 5.58% and 4.35% larger than CDUnet++, PSPnet, DeepLabV3+, MUnet, ATSA and Fmask, respectively. The reason is that TSI-Siamnet contains a dual-branch structure. Although the branch responsible for the synthesized reference image struggles to provide effective information, the branch for the target cloudy image can still achieve reliable detection by the feature extraction module composed of multiple convolutional layers with appended CBAM module and feature fusion module with skip connections (a process analogous to mono-temporal-based cloud detection in this case).

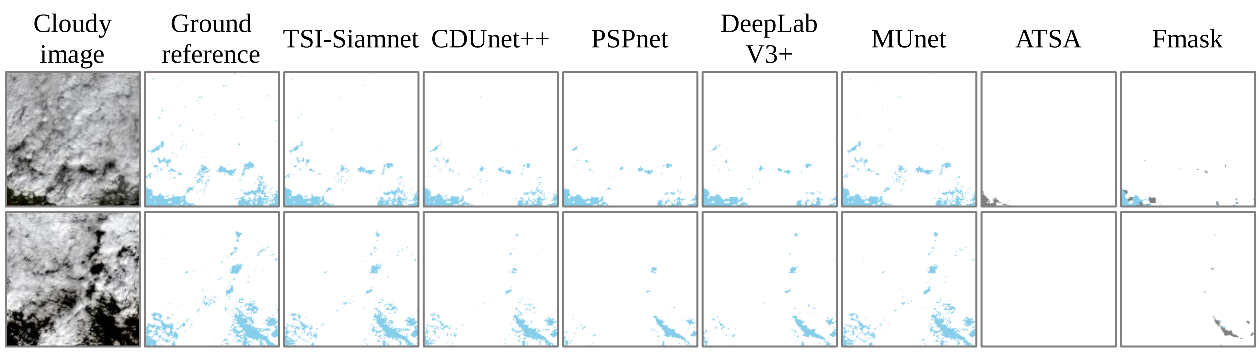


Fig. 10. Detection results for two thick cloud areas. True color composites (R: 4, G: 3 and B: 2) of testing images are shown in the first column. White, gray and blue represent cloud, cloud shadow and background, respectively.

3.3.6. RPCA-composite image versus clean image as reference image

Although cloud-free images are commonly difficult to obtain, the likelihood of being such images being available increases for longer intervals. However, longer intervals can lead to the inclusion of more dramatic changes in the background. Thus, we conducted an experiment to test the effect of using, as the reference image, an RPCA-composite image and a temporally distant cloud-free image. As shown in Fig. 11, the target cloudy image was acquired on August 6, 2013. We chose two cloud-free images from September 15, 2016 and May 29, 2017 for comparison of the cloud detection performance. This scene mainly covers urban, bare land and vegetation, and the land cover changes can be seen clearly by comparing Fig. 11(b) and Fig. 11(c).

The RPCA-composite image and the two cloud-free images were, respectively, fed into TSI-Siamnet as the reference image, and the three results are shown in Fig. 12. It can be seen that the cloud detection result obtained using the cloud-free image in 2016 is closer to the result of the proposed method (i.e., the RPCA-composite image as a reference). Alternatively, when using the cloud-free reference image in 2017, more detection errors are produced, especially in the marked yellow part, which corresponds to the areas experiencing intensive land cover changes in Fig. 11. The corresponding quantitative assessment results are shown in Table 4. Generally, the cloud-free reference in 2017 with greater land cover changes leads to the lowest accuracy, which is consistent with the visual result. The result obtained using the cloud-free reference image in 2016 shows larger metrics than that for 2017, but the accuracy is still lower than for the proposed RPCA-based strategy.

(a) (b) (c)

Fig. 11. The selected testing image for validating the benefit of using the RPCA-composite reference image (true color composite (R: 4, G: 3 and B: 2) images are shown). (a) Target cloudy image acquired on August 6, 2013. (b) Clean image acquired on September 15, 2016. (c) Clean image acquired on May 29, 2017.

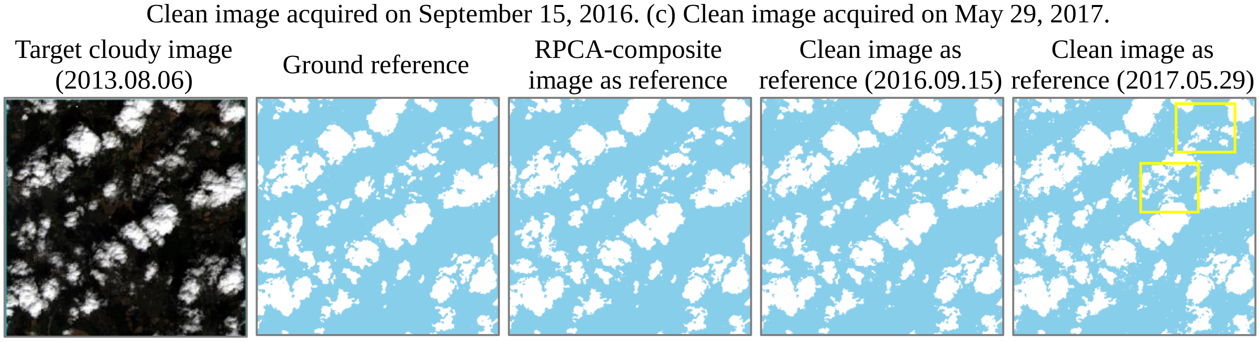


Fig. 12. The detection results of TSI-Siamnet with difference reference images in Fig. 11. True color composite (R: 4, G: 3 and B: 2) of the testing image is shown in the first column. White and blue represent cloud and background, respectively.

Table 4 Accuracy evaluation results of TSI-Siamnet with different reference images (the **bold** value means the most accurate result under each metric).

	OA %	IoU %
RPCA-composite image	**98.08**	**93.73**
Clear image on 2016.09.15	98.02	92.74
Clear image on 2017.05.29	95.09	85.59

3.3.7. Validation of the RPCA method

In this paper, we composited reference image from time-series cloudy images by RPCA. To demonstrate the advantage of the RPCA method, we also constructed reference image by averaging the time-series cloudy images, referred to as the Ave-reference image. Then, the extended Siamnet (or the change detection method CDUnet++) was performed using the RPCA-composite reference image and Ave-reference image as auxiliary data separately. That is, four different versions were implemented, and the corresponding accuraices are shown in Table 5. The results indicate that the RPCA-composite reference image leads to more accurate results for both TSI-Siamnet and CDUnet++. Additionally, the accuracy of TSI-Siamnet is greater than that of CDUnet++.

Table 5 Accuracies of using different reference images (RPCA-composite or Ave-reference) based on different networks (TSI-Siamnet or CDUnet++) (the **bold** value means the most accurate result under each metric).

Network	Reference image	OA (%)	MIoU (%)
TSI-Siamnet	RPCA-composite reference image	**95.05**	**84.37**
	Ave-reference image	94.66	82.73
CDUnet++	RPCA-composite reference image	94.09	80.91
	Ave-reference image	93.75	80.68

3.3.8. Ablation studies

We performed three ablation studies to evaluate the effectiveness of the TSI-Siamnet modules, including the DM, MDFM and CBAM. First, we used the Euclidean distance instead of the DM to calculate the difference in multi-scale features to validate the effectiveness of the DM. Second, the advantages of the MDFM were validated by fusing the upsampled difference features directly according to Eqs. (10) and (11):

$$\mathbf{MultiF_{diff}}=\mathrm{Concatenation}(\mathbf{Up}(\mathbf{F^{1}_{diff}}), \mathbf{Up}(\mathbf{F^{2}_{diff}}), \mathbf{Up}(\mathbf{F^{3}_{diff}}), \mathbf{Up}(\mathbf{F^{4}_{diff}}), \mathbf{Up}(\mathbf{F^{5}_{diff}})) \tag{10}$$

$$\mathbf{MultiF_{diff}}=\mathrm{BN}(\mathrm{ReLU}(\mathrm{Conv2D}_{3\times3}(\mathbf{MultiF_{diff}}))) \tag{11}$$

where $\mathbf{MultiF_{diff}}$ represents the fusion result of multi-scale disparity feature maps.

Third, we compared TSI-Siamnet with TSI-Siamnet without CBAM to demonstrate the effectiveness of the CBAM module. The accuracies of the various cases are shown in Table 6. Moreover, Fig. 13 shows the visual results for seven land cover types. It can be seen that TSI-Siamnet without using any of the DM, MDFM and CBAM results in more commission or omission errors, especially in the absence of the DM and MDFM. With the aid of the CBAM, TSI-Siamnet can further increase the accuracy.

Table 6 Ablation study of the three blocks in TSI-Siamnet (the **bold** value means the most accurate result under each metric).

DM	×	√	√	√
MDFM	√	×	√	√
CBAM	√	√	×	√
OA (%)	94.00	93.81	94.13	**95.05**
MIoU (%)	81.76	79.47	82.21	**84.37**

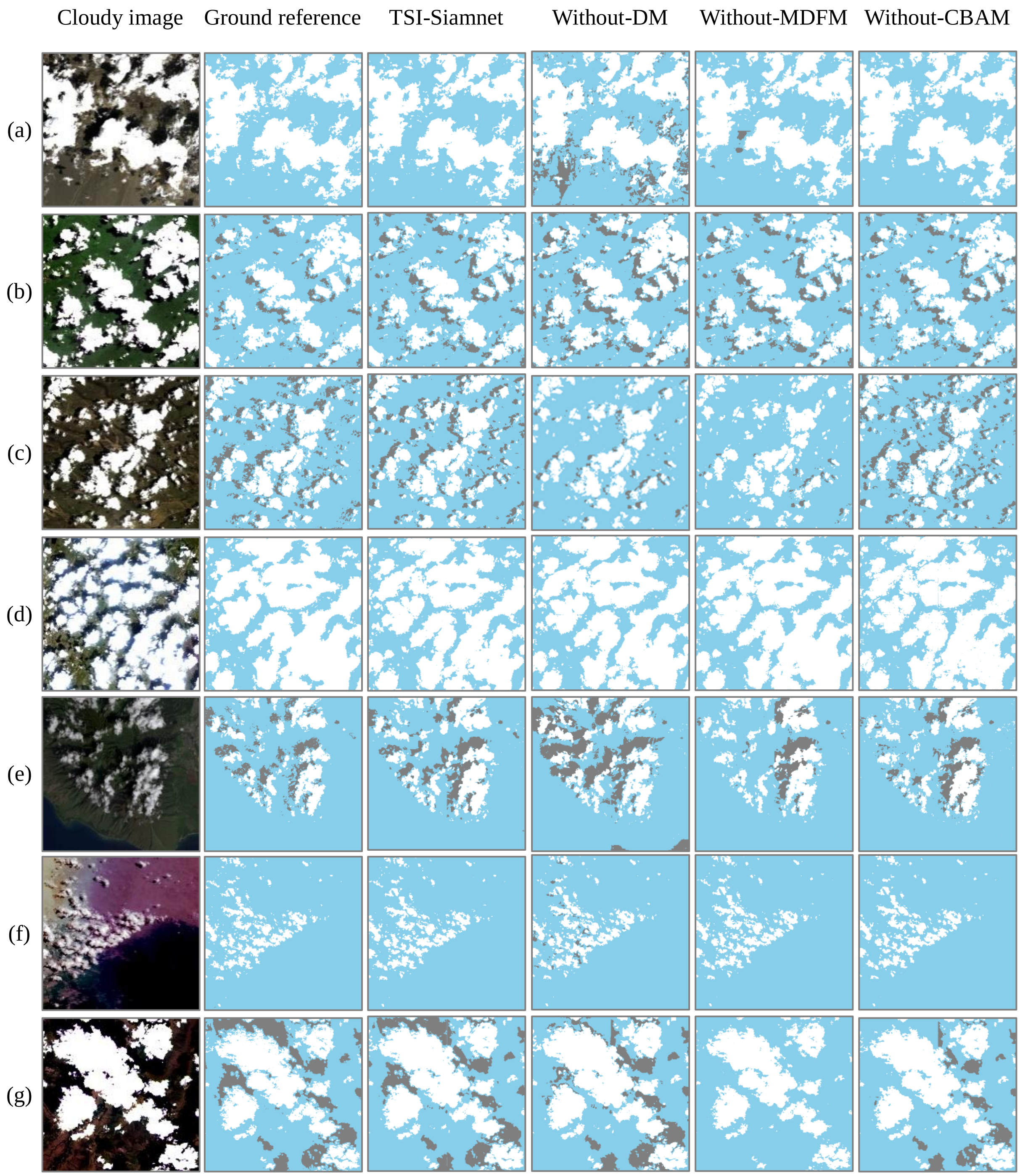
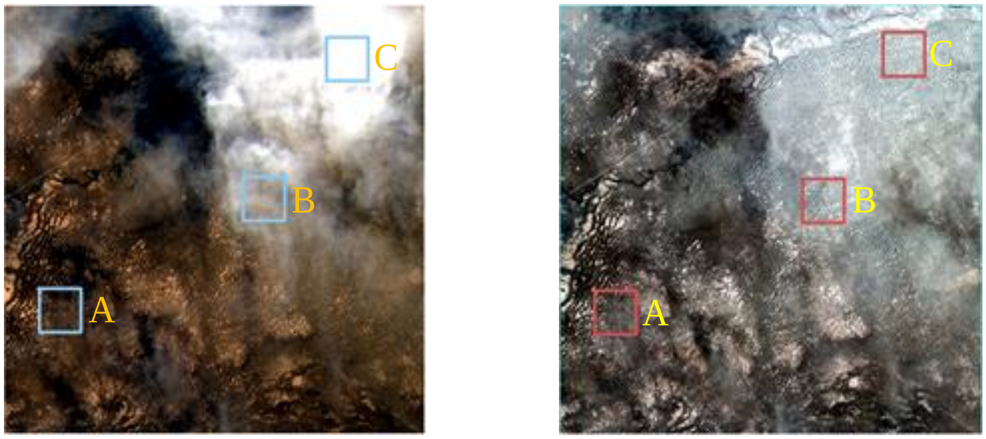


Fig. 13. Cloud and shadow detection results of TSI-Siamnet with different modules removed or altered. (a)–(g) refers to the main land cover types of barren, forest, grass/crops, shrubland, urban, water and wetlands, respectively. True color composites (R:4, G:3 and B:2) of the testing images are shown in the first column. White, gray and blue represent cloud, cloud shadow and background, respectively.

4. Discussion

4.1. The rationale behind RPCA

The proposed TSI-Siamnet method identifies cloud pixels by comparing the difference between the target cloudy image and the RPCA-composite reference image. To analyze the rationale of using RPCA to construct a composite reference image, we selected a scene containing a clear background, and thin and thick clouds simultaneously. As shown in Fig. 14, the interference of both thin and thick clouds is suppressed after the RPCA process, and the background information is revealed to some extent. To quantitatively analyze the difference between the target cloudy image and the corresponding RPCA-composite image, we, respectively, selected three blocks of size 50×50 pixels from these two images to provide scatterplots in the seven bands. As shown in Fig. 15, the thick cloud pixels present a large difference and separation compared with the corresponding pixels of the RPCA results in all bands. For the thin cloud pixels, the RPCA-composite image is partially different from the target cloudy image, especially for the blue, green and red bands. It should be noted that some of the thin cloud pixels do not present obvious differences after the RPCA process, and these pixels are also difficult for non-deep learning-based methods to detect. For the clear background pixels, there is no significant difference before and after the RPCA process, indicating that the background interference can be effectively removed during cloud detection.



(a) (b)

Fig. 14. An example of a cloudy image with RPCA-based processing (A: clear background; B: thin cloud; and C: thick cloud). (a) Target cloudy image. (b) RPCA-composite reference image.

4.2. Computational complexity

In Table 7, we evaluated the complexity and efficiency of the deep learning-based methods by the number of parameters and the inference time. The inference time is counted for an image of size 1k × 1k pixels. As can be seen from the table, our method has only 3.2 million parameters, which is far fewer than the other models, indicating that TSI-Siamnet is a relatively lightweight model. Since TSI-Siamnet has to extract features from both the target cloudy image and the corresponding RPCA-composite image, the inference time is slightly longer than the other methods, but this sacrifice is acceptable for much greater accuracy.

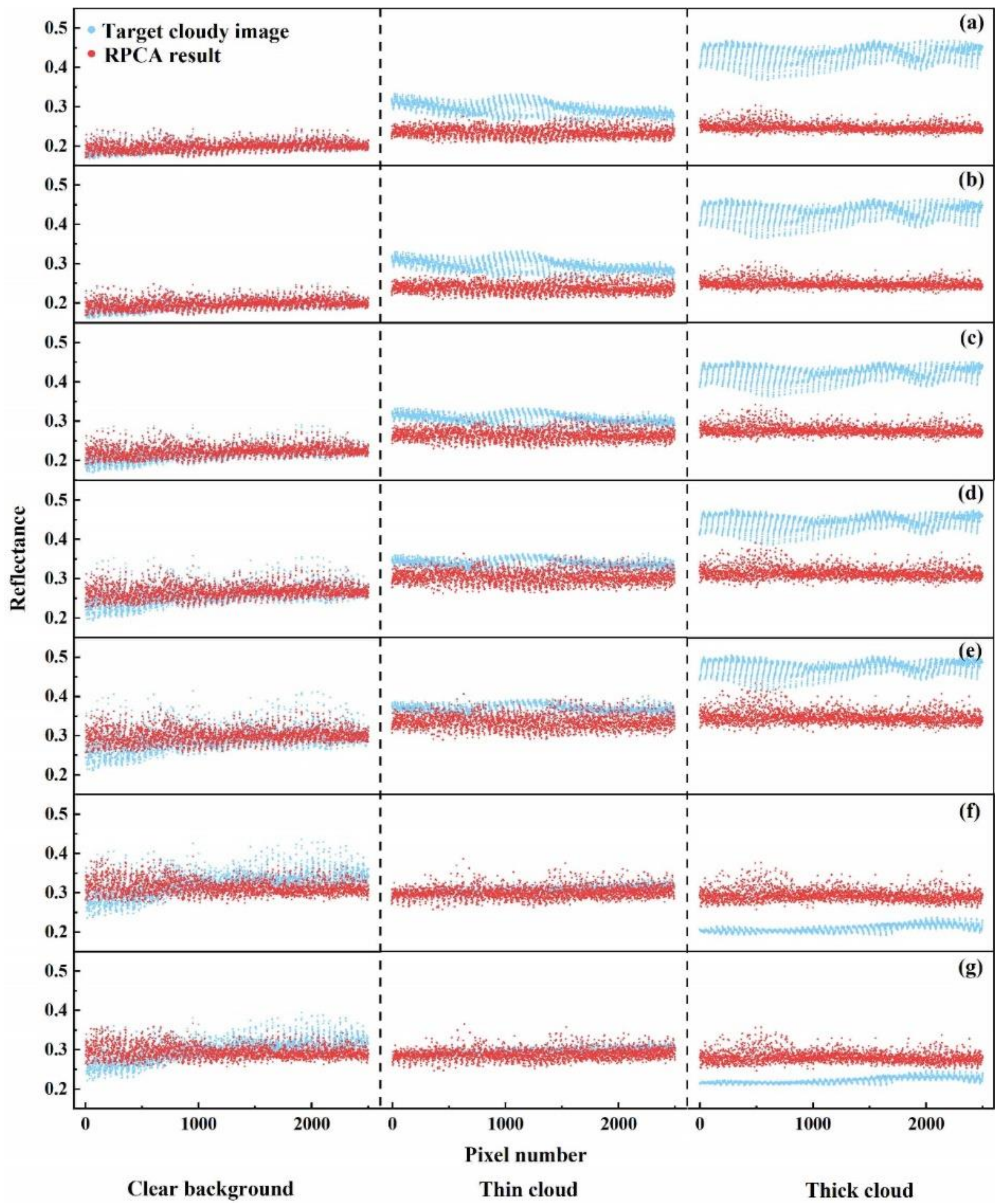


Fig. 15. Comparison between the target cloudy image and the RPCA-composite reference image. (a)-(g) are results for bands 1 to 7 for Landsat 8.

Table 7 Computational complexity analysis of different deep learning-based methods.

	Parameters	Inference time (s) (1k × 1k)
TSI-Siamnet	3.2×10^6	0.70
CDUnet++	9.1×10^6	0.68
PSPnet	46.7×10^6	0.53
DeepLabV3+	41.1×10^6	0.53
MUnet	8.6×10^6	0.50

4.3. Future research

Although TSI-Siamnet achieves promising cloud and shadow detection results, there is still room for further enhancement. First, our algorithm does not consider cloud and shadow detection in snow/ice covered areas. It would be worthwhile research to undertake research to identify the differences in physical characteristics between snow/ice and cloud, and develop corresponding modules in TSI-Siamnet to effectively reduce the interference caused by snow/ice in cloud and shadow detection. Second, in this paper, the RPCA algorithm was used to synthesize a single auxiliary reference image by integrating the valid information in the available time-series data, which inevitably leads to a certain degree of information loss. In future research, it would be interesting to develop models that can more exploit the remaining information in time-series data to synthesize more reliable reference images, such as to provide more reliable input to Siamnet.

5. Conclusion

In this paper, we proposed a new multi-temporal-based method called TSI-Siamnet for cloud and shadow detection in optical remote sensing images. The algorithm implements cloud and shadow detection from the perspective of change detection, reducing the interference of complex backgrounds and increasing cloud and shadow detection accuracy. TSI-Siamnet consists of two main parts: (i) cloud-free reference image construction based on RPCA and (ii) cloud and shadow detection via the extended Siamnet. The developed RPCA method mines effectively the valid information in time-series cloudy images to synthesize reliable reference images, suppressing the interference of cloud contamination in the time-series data. The developed extended Siamnet, including the construction of DM and MDFM modules, utilizes fully the spectral-spatial-temporal features of the available images and extracts reliable feature differences.

TSI-Siamnet was tested with the Landsat 8 Biome dataset (including 336 images covering seven land cover types) and compared with four deep learning-based methods (i.e., CDUnet++, PSPnet, DeepLabV3+ and MUnet) and two classical non-deep learning-based methods (i.e., ATSA and Fmask). The key findings are summarized as follows.

1) TSI-Siamnet produced the greatest cloud and shadow detection accuracy amongst the seven methods, with an OA of 95.05% and MIoU of 84.37%.
2) The advantage of CDUnet++ over three mono-temporal deep learning methods validated the effectiveness of the composite reference image, while the advantage of TSI-Siamnet over CDUnet++ validated the advanatage of the developed extend Siamnet.
3) TSI-Siamnet also outperformed the benchmark methods in terms of stability and cloud coverage estimation.
4) For the three more challenging cases, TSI-Siamnet also demonstrated noticeable

advantages. Specifically, TSI-Siamnet produced the most accurate boundaries of thin clouds, and produced obviously fewer omission errors when detecting broken clouds. Moreover, for the interference of highly reflective artificial surfaces, the benchmark methods are susceptible to incorrectly detecting background pixels as cloud pixels, while TSI-Siamnet produced far fewer commission errors.

Acknowledgements

This research was supported by the National Natural Science Foundation of China under Grants 42222108, 42221002 and 42171345.

References

Boyd, S., 2010. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. 3, 1–122. https://doi.org/10.1561/2200000016.

Cayula, J.-F., Cornillon, P., 1996. Cloud detection from a sequence of SST images. Remote Sens. Environ. 55, 80–88. https://doi.org/10.1016/0034-4257(95)00199-9.

Candes, E.J., Li, X., Ma, Y., Wright, J., 2009. Robust Principal Component Analysis? J. ACM. 58, 1-37. https://doi.org/10.1145/1970392.1970395

Chai, D., Newsam, S., Zhang, H.K., Qiu, Y., Huang, J., 2019. Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks. Remote Sens. Environ.

225, 307–316. https://doi.org/10.1016/j.rse.2019.03.007.

Chai, D., Huang, J., Wu, M., Yang, X., Wang, R., 2024. Remote sensing image cloud detection using a shallow convolutional neural network. ISPRS J. Photogramm. Remote Sens. 209, 66–84. https://doi.org/10.1016/j.isprsjprs.2024.01.026.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: 2018 Proceedings of the European conference on computer vision (ECCV), pp. 801–818. https://doi.org/10.48550/arXiv.1802.02611.

Chen, S., Chen, X., Chen, J., Jia, P., Cao, X., Liu, C., 2016. An iterative haze optimized transformation for automatic cloud/haze detection of Landsat imagery. IEEE Trans. Geosci. Remote Sens. 54, 2682–2694. https://doi.org/10.1109/TGRS.2015.2504369.

Choi, H., 2004. Cloud detection in Landsat imagery of ice sheets using shadow matching technique and automatic normalized difference snow index threshold value decision. Remote Sens. Environ. 91, 237–242. https://doi.org/10.1016/j.rse.2004.03.007.

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., Mosavi, A., 2019. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Sci. Total Environ. 651, 2087–2096. https://doi.org/10.1016/j.scitotenv.2018.10.064.

Foga, S., Scaramuzza, P.L., Guo, S., Zhu, Z., Dilley, R.D., Beckmann, T., Schmidt, G.L., Dwyer, J.L., Joseph Hughes, M., Laue, B., 2017. Cloud detection algorithm comparison and validation for operational Landsat data products. Remote Sens. Environ. 194, 379–390.

https://doi.org/10.1016/j.rse.2017.03.026.

Ghassemi, S., Magli, E., 2019. Convolutional neural networks for on-board cloud screening. Remote Sens. 11, 1417. https://doi.org/10.3390/rs11121417.

Gómez-Chova, L., Amorós-López, J., Mateo-García, G., Muñoz-Marí, J., Camps-Valls, G., 2017. Cloud masking and removal in remote sensing image time series. J. Appl. Remote Sens. 11, 015005. https://doi.org/10.1117/1.JRS.11.015005.

Goodwin, N.R., Collett, L.J., Denham, R.J., Flood, N., Tindall, D., 2013. Cloud and cloud shadow screening across Queensland, Australia: An automated method for Landsat TM/ETM+ time series. Remote Sens. Environ. 134, 50–65. https://doi.org/10.1016/j.rse.2013.02.019.

Guo, J., Xu, Q., Zeng, Y., Liu, Z., Zhu, X., 2022. Semi-Supervised Cloud detection in Satellite Images by Considering the Domain Shift Problem. Remote Sens. 14(11), 2641. https://doi.org/10.3390/rs14112641.

Hagolle, O., Huc, M., Pascual, D.V., Dedieu, G., 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. Remote Sens. Environ. 114, 1747–1755. https://doi.org/10.1016/j.rse.2010.03.002.

Hu, X., Wang, Y., Shan, J., 2015. Automatic recognition of cloud images by using visual saliency features. IEEE Geosci. Remote Sens. Lett. 12, 1760–1764. https://doi.org/10.1109/LGRS.2015.2424531.

Huang, C., Thomas, N., Goward, S.N., Masek, J.G., Zhu, Z., Townshend, J.R.G., Vogelmann, J.E., 2010. Automated masking of cloud and cloud shadow for forest change analysis using

Landsat images. Int. J. Remote Sens. 31, 5449–5464. https://doi.org/10.1080/01431160903369642.

Jedlovec, G., Haines, S., 2007. Spatial and temporal varying thresholds for cloud detection in satellite imagery. In: 2007 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 3329–3332. https://doi.org/10.1109/IGARSS.2007.4423557.

Jeppesen, J.H., Jacobsen, R.H., Inceoglu, F., Toftegaard, T.S., 2019. A cloud detection algorithm for satellite imagery based on deep learning. Remote Sens. Environ. 229, 247–259. https://doi.org/10.1016/j.rse.2019.03.039.

Karakizi, C., Karantzalos, K., Vakalopoulou, M., Antoniou, G., 2018. Detailed land cover mapping from multitemporal Landsat-8 data of different cloud cover. Remote Sens. 10, 1214. https://doi.org/10.3390/rs10081214.

Li, R., Liu, W., Yang, L., Sun, S., Hu, W., Zhang, F., Li, W., 2017b. DeepUNet: A Deep Fully Convolutional Network for Pixel-level Sea-Land Segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 3954-3962. https://doi.org/10.48550/arXiv.1709.00201.

Li, Y., Chen, W., Zhang, Y., Tao, C., Xiao, R., Tan, Y., 2020. Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning. Remote Sens. Environ. 250, 112045. https://doi.org/10.1016/j.rse.2020.112045.

Li, Z., Shen, H., Cheng, Q., Liu, Y., You, S., He, Z., 2019. Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. ISPRS J. Photogramm. Remote Sens. 150, 197–212. https://doi.org/10.1016/j.isprsjprs.2019.02.017.

Li, Z., Shen, H., Li, H., Xia, G., Gamba, P., Zhang, L., 2017a. Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery. Remote Sens. Environ. 191, 342–358. https://doi.org/10.1016/j.rse.2017.01.026.

Luo, Y., Trishchenko, A., Khlopenkov, K., 2008. Developing clear-sky, cloud and cloud shadow mask for producing clear-sky composites at 250-meter spatial resolution for the seven MODIS land bands over Canada and North America. Remote Sens. Environ. 112, 4167–4185. https://doi.org/10.1016/j.rse.2008.06.010.

Mateo-García, G., Gómez-Chova, L., Camps-Valls, G., 2017. Convolutional neural networks for multispectral image cloud masking. In: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, pp. 2255–2258. https://doi.org/10.1109/IGARSS.2017.8127438.

Mateo-García, G., Gómez-Chova, L., Amorós-López, J., Muñoz-Marí, J., Camps-Valls, G., 2018. Multitemporal Cloud Masking in the Google Earth Engine. Remote Sens. 10, 1079. https://doi.org/10.3390/rs10071079.

Mateo-Garcia, G., Adsuara, J.E., Perez-Suay, A., Gomez-Chova, L., 2019. Convolutional Long Short-Term Memory Network for Multitemporal Cloud Detection Over Landmarks. In: 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, Yokohama, Japan, pp. 210–213. https://doi.org/10.1109/IGARSS.2019.8897832.

Mountrakis, G., Li, J., Lu, X., Hellwich, O., 2018. Deep learning for remotely sensed data. ISPRS J. Photogramm. Remote Sens. 145, 1–2. https://doi.org/10.1109/MGRS.2016.2540798.

Peng, D., Zhang, Y., Guan, H., 2019. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. Remote Sens. 11, 1382. https://doi.org/10.3390/rs11111382.

Qiu, S., Zhu, Z., He, B., 2019. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. Remote Sens. Environ. 231, 111205. https://doi.org/10.1016/j.rse.2019.05.024.

Ricciardelli, E., Romano, F., Cuomo, V., 2008. Physical and statistical approaches for cloud identification using Meteosat Second Generation-Spinning Enhanced Visible and Infrared Imager Data. Remote Sens. Environ. 112, 2741–2760. https://doi.org/10.1016/j.rse.2008.01.015.

Irish, R. R., Barker, J. L., Goward, S. N., Arvidson T., 2006. Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. Photogramm. Eng. Remote Sens. 72, 1179–1188. https://doi.org/10.14358/PERS.72.10.1179.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: 2015 Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234-241. https://doi.org/10.48550/arXiv.1505.04597.

Segal-Rozenhaimer, M., Li, A., Das, K., Chirayath, V., 2020. Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN). Remote Sens. Environ. 237, 111446. https://doi.org/10.1016/j.rse.2019.111446.

Shendryk, Y., Rist, Y., Ticehurst, C., Thorburn, P., 2019. Deep learning for multi-modal classification of cloud, shadow and land cover scenes in PlanetScope and Sentinel-2 imagery. ISPRS J. Photogramm. Remote Sens. 157, 124–136. https://doi.org/10.1016/j.isprsjprs.2019.08.018.

Tuia, D., Kellenberger, B., Perez-Suey, A., Camps-Valls, G., 2018. A Deep Network Approach to Multitemporal Cloud Detection. In: 2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, Valencia, pp. 4351–4354. https://doi.org/10.1109/IGARSS.2018.8517312.

Wang, B., Ono, A., Muramatsu, K., Fujiwara, N., 1999. Automated detection and removal of clouds and their shadows from landsat TM images. IEEE Trans. Inform. Syst. 82, 453–460.

Wang, J., Olsen, P.A., Conn, A.R., Lozano, A.C., 2016. Removing clouds and recovering ground observations in satellite image sequences via temporally contiguous robust matrix completion. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 2754–2763. https://doi.org/10.1109/CVPR.2016.301.

Wei, J., Huang, W., Li, Z., Sun, L., Zhu, X., Yuan, Q., Liu, L., Cribb, M., 2020. Cloud detection for Landsat imagery by combining the random forest and superpixels extracted via energy-driven sampling segmentation approaches. Remote Sens. Environ. 248, 112005 https://doi.org/10.1016/j.rse.2020.112005.

Wieland, M., Li, Y., Martinis, S., 2019. Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network. Remote Sens. Environ. 230, 111203. https://doi.org/10.1016/j.rse.2019.05.022.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: convolutional block attention module. In:

Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19. https://doi.org/10.48550/arXiv.1807.06521.

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

Xie, F., Shi, M., Shi, Z., Yin, J., Zhao, D., 2017. Multilevel cloud detection in remote sensing images based on deep learning. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10, 3631–3640. https://doi.org/10.1109/JSTARS.2017.2686488.

Xu, L., Niu, R., Fang, S., Dong, Y., 2013. Cloud detection based on decision tree over Tibetan Plateau with MODIS data. In: Tian, J., Ma, J. (Eds.), MIPPR 2013: Remote Sensing Image Processing, Geographic Information Systems, and Other Applications. International Society for Optics and Photonics SPIE volume 8921. pp. 107-112 https://doi.org/10.1117/12.2030399. URL.

Yu, J., Li, Y., Zheng, X., Zhong, Y., He, P., 2020. An effective cloud detection method for Gaofen-5 images via deep learning. Remote Sens. 12, 2106. https://doi.org/10.3390/rs12132106.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 111716. https://doi.org/10.1016/j.rse.2020.111716.

Yuan, Y., Hu, X., 2015. Bag-of-words and object-based classification for cloud extraction from satellite imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 4197–4205. https://doi.org/10.1109/JSTARS.2015.2431676.

Zhang, J., Wang, H., Wang, Y., Zhou, Q., Li, Y., 2021. Deep network based on up and down blocks using wavelet transform and successive multi-scale spatial attention for cloud detection. Remote Sens. Environ. 261, 112483. https://doi.org/10.1016/j.rse.2021.112483.

Zhang, J., Zhou, Q., Wu, J., Wang, Y., Wang, H., Li, Y., Chai, Y., Liu, Y., 2020. A Cloud Detection Method Using Convolutional Neural Network Based on Gabor Transform and Attention Mechanism with Dark Channel Subnet for Remote Sensing Image. Remote Sens. 12, 3261. https://doi.org/10.3390/rs12193261.

Zhang, Z., Liu, Q., Wang, Y., 2018. Road Extraction by Deep Residual U-Net. IEEE Geosci. Remote Sens. Lett. 15, 749–753. https://doi.org/10.1109/LGRS.2018.2802944.

Zhao, C., Zhang, X., Luo H., Zhong, S., Tang, Lei., Peng, J., Fan, J., 2022. Semi-Supervised Cloud detection for Remote Sensing Imagery via Self-Training. In: 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE, pp. 311-316. https://doi.org/10.1109/ICAICA54878.2022.9844616.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2881-2890. https://doi.org/10.48550/arXiv.1612.01105.

Zhu, X., Helmer, E.H., 2018. An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions. Remote Sens. Environ. 214, 135–153. https://doi.org/10.1016/j.rse.2018.05.024.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 5, 8–36. https://doi.org/10.1109/MGRS.2017.2762307.

Zhu, Z., Wang, S., Woodcock, C.E., 2015. 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. 159, 269–277. https://doi.org/10.1016/j.rse.2014.12.014.

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

Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94. https://doi.org/10.1016/j.rse.2011.10.028.

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

Zou, Z., Li, W., Shi, T., Shi, Z., Ye, J., 2019. Generative Adversarial Training for Weakly Supervised Cloud Matting. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, pp. 201–210. https://doi.org/10.1109/ICCV.2019.00029.