Remote Sensing of Environment 225 (2019) 307–316

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks

Dengfeng Chai[a,b,*], Shawn Newsam[b], Hankui K. Zhang[c], Yifan Qiu[a], Jingfeng Huang[d,e]

[a] Institute of Spatial Information Technique, Zhejiang University, Hangzhou 310027, China
[b] Electrical Engineering and Computer Science, University of California, Merced, CA 95343, USA
[c] Geospatial Sciences Center of Excellence, South Dakota State University, Brookings, SD 57007, USA
[d] Institute of Applied Remote Sensing and Information Technology, Zhejiang University, Hangzhou 310058, China
[e] Key Laboratory of Agricultural Remote Sensing and Information Systems, Zhejiang University, Hangzhou 310058, China

ARTICLE INFO

Keywords:
Cloud detection
Semantic segmentation
Convolutional neural network (CNN)
Landsat

ABSTRACT

This paper formulates cloud and cloud shadow detection as a semantic segmentation problem and proposes a deep convolutional neural network (CNN) based method to detect them in Landsat imagery. Different from traditional machine learning methods, deep CNN-based methods convolve the entire input image to extract multi-level spatial and spectral features, and then deconvolve these features to produce the detailed segmentation. In this way, multi-level features from the whole image and all the bands are utilized to label each pixel as cloud, thin cloud, cloud shadow or clear. An adaption of SegNet with 13 convolutional layers and 13 deconvolution layers is proposed in this study. The method is applied to 38 Landsat 7 images and 32 Landsat 8 images which are globally distributed and have pixel-wise cloud and cloud shadow reference masks provided by the U.S. Geological Survey (USGS). In order to process such large images using the adapted SegNet model on a desktop computer, the Landsat Collection 1 scenes are split into non-overlapping 512 * 512 30 m pixel image blocks. 60% of these blocks are used to train the model using the backpropagation algorithm, 10% of the blocks are used to validate the model and tune its parameters, and the remaining 30% of the blocks are used for performance evaluation. Compared with the cloud and cloud shadow masks produced by CFMask, which are provided with the Landsat Collection 1 data, the overall accuracies are significantly improved from 89.88% and 84.58% to 95.26% and 95.47% for the Landsat 7 and Landsat 8 images respectively. The proposed method benefits from the multi-level spatial and spectral features, and results in more than a 40% increase in user's accuracy and in more than a 20% increase in producer's accuracy for cloud shadow detection in Landsat 8 imagery. The issues for operational implementation are discussed.

1. Introduction

The freely available Landsat imagery are supporting more and more applications due to their long record and global coverage (Roy et al., 2014; Wulder et al., 2016). However, the annual mean cloud cover of Landsat 5 and 7 images can reach as high as 40% as reported in Ju and Roy (2008). To support most terrestrial remote sensing applications, it is necessary to detect the clouds and cloud shadows in the imagery before further processing. Automatic cloud and cloud shadow detection is therefore an important issue for both data providers and users.

Clouds and cloud shadows in Landsat imagery can be detected using spectral tests, temporal differentiation and statistical methods. Spectral tests are based on the observation that the radiances (for thermal bands), the reflectances (for reflective bands) or other derived values (e.g. normalized difference vegetation index (NDVI)) for clouds or cloud shadows fall in limited ranges. One or two thresholds are thus employed for each original or derived band to test if the value is in the expected range. Results for different bands are typically fused using rules based on decision trees for example. In this way, the clouds and cloud shadows can be distinguished from clear pixels (Hollingsworth et al., 1996; Irish et al., 2006; Roy et al., 2010; Scaramuzza et al., 2012; Zhu and Woodcock, 2012; Vermote et al., 2016; Qiu et al., 2017; Sun et al., 2018). Temporal differentiation is based on the fact that clouds and cloud shadows are dynamic. Since they move, there are often

* CorrespondingPlease check the address for the corresponding author that has been captured, and confirm if correct.author at: Institute of Spatial Information Technique, Zhejiang University, Hangzhou 310027, China.
E-mail addresses: chaidf@zju.edu.cn (D. Chai), snewsam@ucmerced.edu (S. Newsam), hankui.zhang@sdstate.edu (H.K. Zhang), rs_qyf@zju.edu.cn (Y. Qiu), hjf@zju.edu.cn (J. Huang).

https://doi.org/10.1016/j.rse.2019.03.007
Received 8 September 2018; Received in revised form 9 January 2019; Accepted 4 March 2019

significant differences between images captured at different dates. By comparing multi-temporal images, the pixels with large differences are identified as either clouds or cloud shadows (Wang et al., 1999; Hagolle et al., 2010; Jin et al., 2013; Zhu and Woodcock, 2014; Frantz et al., 2015; Zhu and Helmer, 2018). Alternatively, statistics of spatial and spectral features of clouds are utilized by statistical methods to estimate cloud cover (Molnar and Coakley Jr, 1985) and to detect clouds (Ricciardelli et al., 2008; Amato et al., 2008). Usually, the pixel-wise detection is formulated as a classification problem and solved based using classification models such as support vector machines (SVM) (Lee et al., 2004), neural networks (NN) (Tian et al., 1999), etc.

The power of statistical methods can be enhanced by machine learning techniques which train a classifier as well as learn the features to be used in the classification. The training and classification are conducted in separate stages. In the first stage, both the features and classifiers are learned from training samples. In the second stage, clouds and cloud shadows are detected using the learned features and classifiers. Traditional neural networks are especially popular among existing classifiers (Lee et al., 1990; Scaramuzza et al., 2012; Hughes and Hayes, 2014). Although these methods can discover and exploit the distinguishing characteristics of clouds and cloud shadows hidden in the data, their performance is limited by the classification framework and the network structures and their capacities. First, traditional neural networks classifiers classify each pixel independently. Second, the networks have a simple structure and limited capacity as they consist of only one or two hidden layers between the input and output layers, and each layer consists of only a few nodes. Third, a limited number of features extracted from highly localized patches around a target pixel are input to the network and forwarded to the hidden and output layers. Due to these limitations, traditional neural networks are outperformed by the Fmask algorithm (Zhu and Woodcock, 2012) which is based on spectral tests.

Convolutional neural networks (CNNs) contain layers consisting of nodes arranged in three dimensions corresponding to the rows, columns and channels of an RGB image (LeCun et al., 1998a). A detailed description of CNNs can be found in Goodfellow et al. (2016). CNNs facilitate multi-channel image representation since each pixel can be represented by one node and each input image or feature map can be represented by one layer. An image is input to a CNN directly and is forwarded layer by layer based on a series of image convolutions. In this way, the context among neighboring pixels is captured by the convolutions, and an image is classified as a whole instead of pixel-by-pixel. Benefiting from the improved computing power of GPUs, deep CNNs with many layers have been developed by computer vision researchers in recent years. A shallow CNN can extract local features contained in a small field as the image convolved by only a few layers before the final representation is output. In contrast, a deep CNN can extract global features contained in a large field of the input image as this large field is convolved by many layers before the final representation is output by the final layer. Further, different features at multiple levels are also output as intermediate feature maps. This means that deep CNNs can extract features at multiple levels. Deep CNNs have been successfully applied to a variety of computer vision problems, including object detection and image segmentation (Girshick et al., 2014; Ren et al., 2015; Badrinarayanan et al., 2017) to name a few relevant to cloud detection. Key to the success of deep CNNs is their ability to learn and exploit such multi-level features in the data.

CNNs have become new tools for cloud detection in the remote sensing community. In particular, researchers have employed superpixel methods such as SLIC (Achanta et al., 2012) to group pixels into superpixels and then use CNNs to classify the superpixels into cloud, cloud shadow or clear classes (Xie et al., 2017; Zi et al., 2018). The cloud detection problem is thus formulated as a superpixel classification problem, which can be regarded as an extension of the pixel classification problem. Class scores are calculated for each superpixel and the final class assignment is based on these scores. Different superpixels are still classified independently. These methods also rely significantly on the superpixel grouping since they treat each superpixel as an atom. To correct for the errors resulting from the superpixel grouping, these methods often employ a post-processing step such as optimization based on Markov random fields (Zi et al., 2018). Such optimizations can be very time consuming during classification (Chen et al., 2018).

Inspired by the performance of deep neural networks for the semantic segmentation in other computer vision tasks, this paper proposes a novel method to detect clouds and cloud shadows in Landsat imagery. First, cloud and cloud shadow detection is formulated as a semantic segmentation problem instead of a pixel-by-pixel classification problem. As shown in Fig. 1, spatial context from the whole image and spectral information from all the channels is exploited to segment an image as a whole, while only spatial information within a limited neighborhood (e.g., 3 * 3 pixels) is utilized by methods that perform pixel classification. Second, the semantic segmentation is achieved via a deep convolutional neural network (CNN) that convolves an input image many times to output a confidence map for each class. The confidence of each pixel in each map indicates the probability of the pixel belonging to the corresponding class. As shown in Fig. 2, a multi-channel input image is convolved by the encoder (left part) to extract features at 6 different levels, and then these features are deconvolved by the decoder (right part) to produce 4 detailed confidence maps with the same resolution as the input image. Finally, each pixel is classified into the class with the highest confidence, and all pixels are classified collaboratively and simultaneously. The proposed method is evaluated through extensive experiments on two Landsat data sets. It is shown to significantly improve upon the state-of-the-art methods for cloud and cloud shadow detection.

The rest of this paper is organized as follows. The Landsat images and the cloud and cloud shadow reference data are described in Section 2, cloud and cloud shadow detection is formulated as a semantic segmentation problem in Section 3, the deep convolution neural network used to perform the semantic segmentation is described in Section 4, experiments with evaluations and comparisons are presented in Section 5, and the conclusion along with a discussion is presented in Section 6.

2. Data description

2.1. Landsat 7 and Landsat 8 cloud and cloud shadow reference data

Both the Landsat 7 cloud reference dataset selected by Irish et al. (2006) (**L7 Irish**) and Landsat 8 cloud reference dataset derived by Foga et al. (2017) (**L8 Biome**) are employed in this study (please refer to Fig. 1 in Foga et al., 2017). The L7 Irish and L8 Biome datasets consist of 206 and 96 scenes respectively, and they are evenly distributed over nine latitude zones and eight biomes respectively. Ground truth cloud masks have been manually labeled for each scene by analysts at the U.S. Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center for validating cloud detection algorithms. These Landsat cloud cover assessment validation datasets are publicly available from the USGS website[1]. The differences among interpretations by different analysts were reported to be 7% (Scaramuzza et al., 2012). In the cloud masks for 41 of the L7 Irish scenes, each pixel is labeled as cloud, thin cloud, cloud shadow or clear. In the cloud masks for the other L7 Irish scenes, cloud shadows are not distinguished from clear pixels mainly due to the labor intensive nature of this task. Only those scenes with both cloud masks and cloud shadow masks are used in our experiments below. Three of these scenes located in Antarctica are excluded from our experiments since the cloud mask reference images are defined in the Universal Transverse Mercator (UTM)

[1] https://landsat.usgs.gov/landsat-cca-validation-datasets.

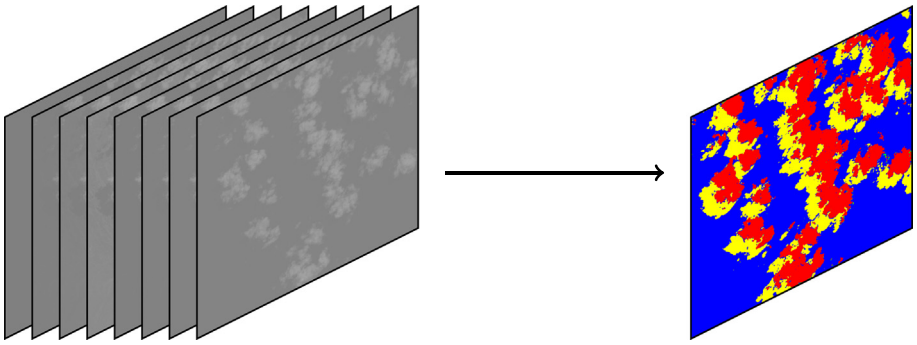

Fig. 1. Cloud and cloud shadow detection is formulated as a semantic segmentation problem. On the left is the input, a multi-channel image in which clouds and cloud shadows are to be detected, and on the right is the output, a cloud mask image in which each pixel is assigned a label denoting its class. Cloud, thin cloud, clear and cloud shadow are illustrated as red, green, blue and yellow respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

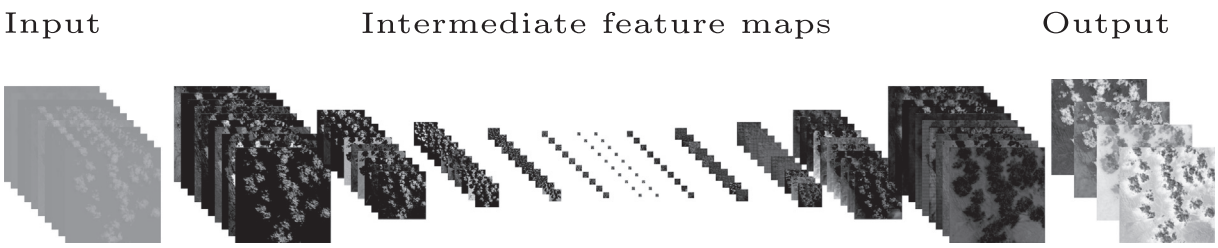

Fig. 2. Encoding and decoding procedure. The intermediate feature maps in between the input and output demonstrate features extracted at different levels from fine to coarse and then back to fine. This figure depicts only 6 convolved and 6 deconvolved feature maps at different resolutions among the 13 convolved and 13 deconvolved feature maps of our model. It also only depicts 10 instead of the all channels for each feature map. All the features are normalized to [0,⋯ ,255].

projection while the Landsat Collection 1 images (Section 2.2) are defined in the polar stereographic projection. Similarly, cloud shadows are distinguished from clear pixels in only 32 of the L8 Biome scenes. In total, the 38 L7 Irish and 32 L8 Biome scenes that have both cloud masks and cloud shadow masks are used in our experiments. These scenes are also evenly distributed over latitude zones or biomes. Their manually labeled cloud and cloud shadow masks serve as the ground truth in our experiments.

2.2. Landsat 7 and Landsat 8 Collection 1 data and pre-processing

The 38 Landsat 7 and 32 Landsat 8 Collection 1 data corresponding to the scenes in L7 Irish and L8 Biome are downloaded from the USGS Earth Resources Observation and Science (EROS) Center archive. All the bands except for one from Landsat 8 are resampled to 30 m spatial resolution in the Collection 1 data. Since multi-channel images are required by the proposed neural network to have the same spatial resolution in each channel, eight Landsat 7 bands and ten Landsat 8 bands of 30 m resolution are employed to detect clouds and cloud shadows in out experiments.

The digital number images are converted to top of atmosphere reflectance or brightness temperature to be input to the neural networks. A solar zenith angle is derived for each pixel using the Landsat Angles Creation Tool provided by the USGS. Then, using the derived solar zenith angle and the scaling factors stored in the metadata, the digital numbers are converted to top of atmosphere reflectance for the solar reflective bands and to brightness temperature for the thermal bands.

A cloud mask is available in the Collection 1 data for each scene. A pixel is set as cloud when its cloud bit is 1, and it is set as cloud shadow when its cloud shadow bit is 1. A pixel is set as clear when both its cloud bit and cloud shadow bit are 0. These masks are generated by the CFMask algorithm (Zhu and Woodcock, 2012; Foga et al., 2017). In these masks, thin clouds are also recognized as clouds. We compare the results of our method to these CFMask generated masks in the experiments using the manually created ground truth.

2.3. Training, validation and test sets

It is not possible to process entire Landsat images at once on a standard desktop computer with limited GPU memory since the size of the deep neural network increases dramatically with input image size. Therefore, as shown in Fig. 3, each Landsat image is partitioned into a set of small non-overlapping images, each of which is of 512 * 512. Since the Landsat images are the (rotated) georeferenced versions, not the original ones, there are fill pixels as indicated by the white areas. We only use the images without fill pixels. This results in around 120 512 * 512 images for each Landsat scene. These 512 * 512 images are the input to the deep neural network.

As is standard in the computer vision community, the collection of

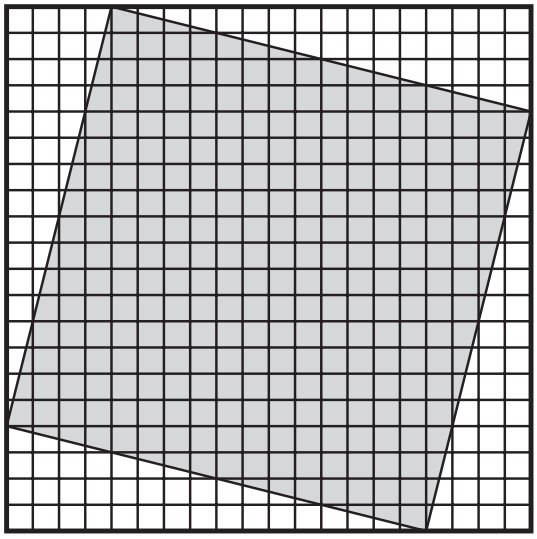

Fig. 3. Image partition for a typical Landsat image. The original image is indicated by the gray area, the georeferenced image is indicated by the outer rectangle. As shown, some areas are filled by white (null) pixels in the georeferenced image. According to the imposed grid, the georeferenced image is partitioned into a set of small 512 * 512 images.

Table 1
Training, validation and test sets for L7 Irish and L8 Biome. The number of 512 * 512 30 m images in each set is as follows.

	Images (512 * 512)		
	Train 60%	Val. 10%	Test 30%
L7 Irish	2732	420	1328
L8 Biome	2410	378	1178

512 * 512 images is partitioned into three sets: a training set, a validation set and a test set. 60%, 10% and 30% of the images are randomly grouped into the training, validation and test sets. Their sizes in terms of the number of images are listed in Table 1.

The images in the training set are used to train the deep neural networks. Those in validation set are used for experiments such as parameter tuning. And, those in test set are used only for the final evaluation—they are not involved in training or tuning. Note that while we only apply our trained model to the images in the test set, it could be applied to any other Landsat images without retraining.

3. Cloud detection as semantic segmentation

As illustrated in Fig. 1, given a multi-band input image, cloud and cloud shadow detection consists of labeling each pixel as cloud, cloud shadow, clear or other category. In some cases, thick cloud and thin cloud need to be distinguished from each other. In other cases, water, snow and ice also need to be recognized. Such a pixel-wise labeling leads to cloud and cloud shadow detection at the pixel-level.

Previous NN-based approaches to cloud detection perform pixel-by-pixel classification. Given a target pixel, a feature vector is extracted from a small (e.g. 3 * 3) patch around the target pixel and input to a neural network, which outputs a vector of class confidences indicating a pixel classification. Each pixel in an image is classified independently given its feature vector. The contextual relationships between neighboring pixels are not explicitly modeled but encapsulated implicitly in the features extracted from the local patches around the target pixels (Lee et al., 1990; Scaramuzza et al., 2012; Hughes and Hayes, 2014). Limited contextual information is sometimes also used in a post-processing step to improve the classification results (Hughes and Hayes, 2014). Since global features of an image cannot be extracted from a small local patch, these approaches have limited performance even with the post-processing step. In contrast, deep CNNs can extract both global and local features. Further, since an image is convolved layer by layer in a CNN as a whole, the contextual information among neighboring pixels is captured by a series of convolutions instead of a single pre-processing step for feature extraction. Although CNNs have been applied to cloud detection, they were previously used to convolve a patch covering a superpixel and output a class for the superpixel but not its individual pixels (Xie et al., 2017; Zi et al., 2018). As in pixel classification, different superpixels are classified independently.

This paper instead formulates cloud and cloud shadow detection as a semantic segmentation problem. By utilizing both the spatial context from the whole image and spectral features from all the bands, an image is partitioned into a set of disjoint regions and each region is labeled as cloud shadow, clear, thin cloud or cloud. Our formulation does not necessarily seek to label the regions as belonging to different objects as in standard scene parsing but instead performs cloud detection in which each pixel is labeled with a single class. However, the segmentation, and therefore the classification, is performed for the image as a whole rather than in a pixel-by-pixel manner.

4. Cloud detection based on deep CNN

The semantic segmentation is achieved via a fully convolutional neural network (FCNN) (Long et al., 2015), more specifically, an encoder-decoder network (Badrinarayanan et al., 2017) as depicted in Fig. 2. The left/right part is called an encoder/decoder, which consists of a sequence of layers (only a subset of layers are depicted). To represent pixels, the nodes in each layer are organized in 3 dimensions corresponding to the rows, columns and channels of an image. The left and right layers denote the input and output images respectively, and the layers in between denote a sequence of intermediate feature maps. An input image, from left to right, undergoes a sequence of convolutions (left part) and deconvolutions (right part) before the final representation is output.

Without loss of generality, let us focus on convolving a local volume of a previous layer to output a single pixel whose coordinates in the current layer are (h,w,d). The convolution is carried out as follows:

$$f'_{h,w,d} = \left(\sum_{i=-1}^{1} \sum_{j=-1}^{1} \sum_{k=1}^{c} W^{d}_{i,j,k} f_{h+i,w+j,k} \right) + b^{d}, \quad (1)$$

where $W^{d}_{i,j,k}$ are the weights of the convolution filter and b^{d} is the bias of the filter. In this case, there are 3 * 3 * c weights and one bias for each filter. 3 * 3 is the filter size in spatial dimension, and k is the filter size in spectral dimension. As c is the number of channels of the previous layer, this convolution covers all the channels. When d takes $1,\cdots,c'$, this local volume is convolved to output c' features corresponding to c' channels of current layer. A rectified linear function $\max(0, f'_{h,w,d})$ is applied to each output of the convolution. This nonlinear function enables the deep CNNs learn more complex features than cannot be derived using linear transformations.

As depicted in Fig. 2, the encoder/decoder has the effect of down/up sampling, which is achieved through the pooling/unpooling layers. Pooling is developed to down-sample the feature maps along the rows and columns but not channels. Specifically, 2 * 2 max pooling is employed to select the maximum value of a 2 * 2 window so that 2 rows by 2 columns are merged to generate one output. In this way, the feature map is down-sampled to half its resolution in each spatial dimension. Unpooling is developed as a counterpart to a pooling for up-sampling, which is based on the indices calculated in its corresponding pooling operation.

Let us focus on convolution in the spatial dimension as every convolution covers all bands. One pixel in the final encoding layer is convolved from a 3 * 3 window of its previous layer, and this 3 * 3 window is convolved from a 5 * 5 window of its previous layer. Taking account of all the convolution layers and pooling layers as listed in Table 2, a pixel in the final encoding layer is convolved from a 212 * 212 window of the input image. In total, 512 features (i.e., 512 channels) are output by the final encoding layer. These features are global features. As depicted in Fig. 2, feature maps at 6 levels are extracted from the input image by the encoding layers. The left feature map stores local features, the right one stores global features, and the middle feature maps store features at middle levels. These features are magnified by the decoding layers to generate a detailed output. Since the decoding and encoding layers are arranged in a pairwise fashion, the input and output have the same resolution, which assures a pixel-wise classification.

4.1. Network architecture

Our encoder-decoder network is an adaption of SegNet (Badrinarayanan et al., 2017), whose encoder is based on VGG (Simonyan and Zisserman, 2014). The different types of layers in the network are described below.

Input layer stores the input image. It is a 512 * 512 * c volume, where, c is the number of channels (bands): $c = 8$ for Landsat 7 images and $c = 10$ for Landsat 8 images.

Conv layer convolves the previous layer to derive the current layer. The number of weights involved in the convolutions are listed for all layers in Table 2. The four numbers indicate the

Table 2
Encoder-decoder configuration. The left and right columns correspond to the encoder and decoder respectively. The encoding, convolutional layers progress downwards and the decoding, deconvolutional layers progress upwards. Each conv (or deconv) layer is followed by a RELU layer (not shown). The arrows indicate skip connections. Four numbers are listed for each convolutional/deconvolutional layer. The last one indicates the number of filters for this layer, and the other three indicates the filter size in the spatial and spectral dimensions.

Input		Output
conv3 * 3 * 10 * 96		deconv3 * 3 * 96 * 96
conv3 * 3 * 96 * 96		deconv3 * 3 * 128 * 96
Pooling	→	Unpooling
conv3 * 3 * 96 * 128		deconv3 * 3 * 128 * 128
conv3 * 3 * 128 * 128		deconv3 * 3 * 256 * 128
Pooling	→	Unpooling
conv3 * 3 * 128 * 256		deconv3 * 3 * 256 * 256
conv3 * 3 * 256 * 256		deconv3 * 3 * 256 * 256
conv3 * 3 * 256 * 256		deconv3 * 3 * 512 * 256
Pooling	→	Unpooling
conv3 * 3 * 256 * 512		deconv3 * 3 * 512 * 512
conv3 * 3 * 512 * 512		deconv3 * 3 * 512 * 512
conv3 * 3 * 512 * 512		deconv3 * 3 * 512 * 512
Pooling	→	Unpooling
conv3 * 3 * 512 * 512		deconv3 * 3 * 512 * 512
conv3 * 3 * 512 * 512		deconv3 * 3 * 512 * 512
conv3 * 3 * 512 * 512		deconv3 * 3 * 512 * 512
Pooling	→	Unpooling

filter size in the spatial and spectral dimensions, and the number of filters respectively. They correspond to i,j,k,d of $W^{d}_{i,j,k}$ in Eq. (1). This table is for Landsat 8 images. 10 should be replaced by 8 when the input images are Landsat 7 images. Our model contains 13 convolutional layers as listed in Table 2.

Deconv layer is a counterpart to a conv layer. It has the same structure as a conv layer but has different weights. Our model contains 13 deconvolutional layers as listed in Table 2.

RELU layer applies a rectified linear function to each output of the conv/deconv layer. It generates the element-wise activations to form a feature map.

Pool layer down-samples the feature maps along the rows and columns to produce a feature map with half the resolution in each spatial dimension.

Unpool layer is a counterpart to a pool layer for up-sampling.

Output layer is a class confidence map of 512 * 512 * 4, where 4 is the number of classes (cloud, thin cloud, clear and cloud shadow). Each pixel thus has a confidence value for each class.

Our encoder and decoder are described in the left and right columns of Table 2 respectively. Each conv or deconv layer is followed by a RELU layer, which is merged into its corresponding conv or deconv layer and not explicitly listed in the table. Starting from the input image, the features flow downwards through the left layers and upwards through the right layers, and finally reach the output layer. As indicated by the arrows in Table 2, skip connections allow features to flow directly from the encoding layer to the decoding layer of the same level without a reduction in spatial resolution. By skipping encoding layers, feature details are retained to assure a precise segmentation.

4.2. Training

No parameters are involved in the RELU, pooling and unpooling layers. However, 3 * 3 * c weights and a bias are involved in each filter for each conv/deconv layer. These weights and biases need to be learned from the training samples. This is achieved via optimizing these free parameters so that the predicted class scores for the training samples match the ground truth labels as well as possible.

Without loss of generality, let an input image and its corresponding (ground truth) label image be $x = \{x_{ij}\}$ and $y = \{y_{ij}\}$, where i,j are row index and column index respectively, and let $f = \{f_k\}$ be the vector of class scores, where k is the class index. The objective function to be minimized in training is the cross-entropy loss

$$L(x; \Theta) = \sum_{ij} l'(x_{ij}; \Theta) \tag{2}$$

$$= \sum_{ij} -\log\left(\frac{\exp(f_{y_{ij}})}{\sum_k \exp(f_k)}\right) \tag{3}$$

where Θ is the set of parameters to be optimized, $l'(x_{ij};\Theta)$ is the loss for pixel x_{ij}, and the total loss is the sum of losses over all pixels. If the predicted label is the same as ground truth, i.e. $f_k = 1$ when $k = y_{ij}$ and $f_k = 0$ when $k \neq y_{ij}$, then $\exp(f_{y_{ij}}) = \sum_k \exp(f_k)$ and $l'(x_{ij};\Theta) = 0$. Otherwise, $\exp(f_{y_{ij}}) < \sum_k \exp(f_k)$ and $l'(x_{ij};\Theta) > 0$. By minimizing the cross-entropy loss, the predicted labels become consistent with the ground truth labels.

The backpropagation algorithm is employed to minimize the loss (LeCun et al., 1998b). It consists of a forward pass and a backward pass. The forward pass convolves an input image to calculate the loss. In this pass, the local gradients of the outputs with respect to the parameters are also computed for each layer. From the final layer to input layer, the backward pass propagates the gradients layer by layer by multiplying the local gradients of the current layer with those propagated to this layer. When this reaches the input layer, the gradients of the loss with respect to all the parameters are calculated. Based on these gradients, the loss function can be minimized by updating the parameters based on their gradients.

The RMSProp schema has been found to have a beneficial equalizing effect and so we use it to update the parameters (Tieleman and Hinton, 2012):

$$\gamma = \beta*\gamma + (1 - \beta)*d\theta^2 \tag{4}$$

$$\theta = \theta - \alpha*d\theta/(\sqrt{\gamma}) \tag{5}$$

where $\theta \in \Theta$ is a parameter to be updated, the learning rate α is set to 0.001, the decay rate β is set to 0.995, and γ is initialized to 0 and updated iteratively.

A technique called dropout is employed to prevent overfitting during training. It activates a neuron with a certain probability, which is set to be 0.5 in this paper. Dropout effectively complements other regularization methods such as L1 and L2 (Srivastava et al., 2014).

A typical example of training for 100 epochs (iterations) is illustrated in Fig. 4. After each epoch, the loss function and overall accuracy are computed using the validation set. As depicted in the figure, the loss decreases and the accuracy increases as the training progresses. Although the accuracy continues increasing with the number of iterations, a good accuracy is achieved after 20 epochs. The maximal, minimal and mean accuracy in the range of 20 to 100 epochs is 93.18%, 90.78% and 92.15% respectively and the standard deviation is 0.62%. It is possible to reduce the training time by reducing the number of epochs. However, the training time is not crucial since training is only performed once. Since inference does not require iteration, the cloud detection is very fast. 50 epochs are employed for training in the experiments.

4.3. Inference

Once the network has been trained, that is, all the parameter values have been learned, inference (classification) is performed by a simple forward pass through the network: a given image is fed to the input layer, it is transformed layer-by-layer by the encoder-decoder, and the per pixel class scores are output. This is an efficient computation consisting of filtering over local volumes.

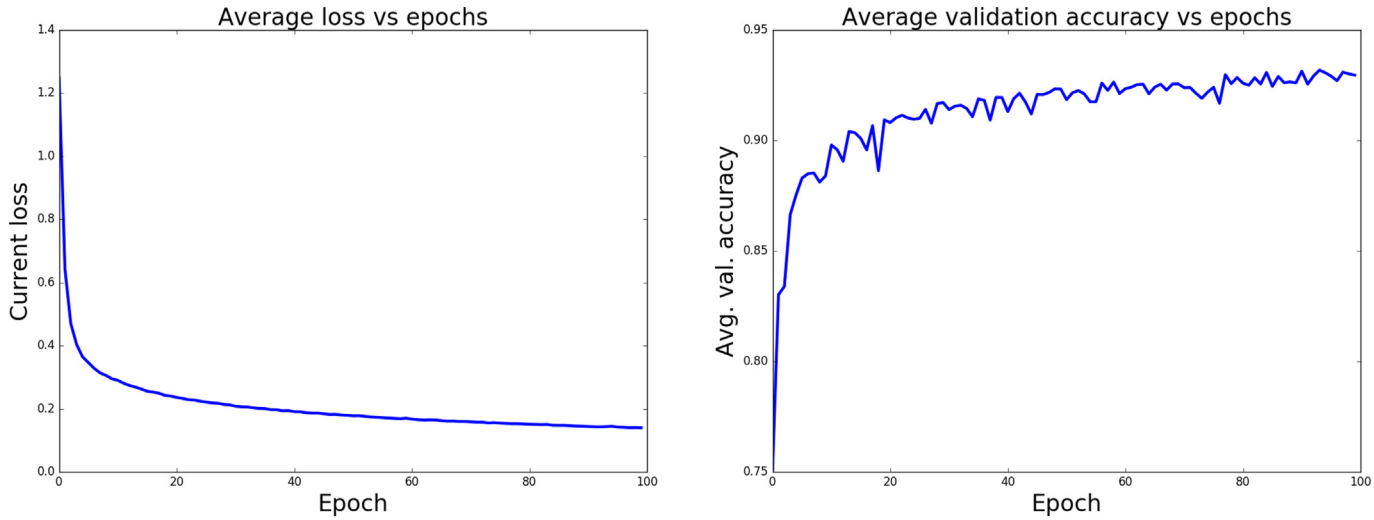

Fig. 4. Training procedure. Left and right are the loss function and segmentation accuracy versus the number of epochs.

Fig. 2 illustrates the forward pass by showing features at different levels. As shown, the encoder and decoder down-sample and up-sample the features respectively. Since a large volume of the input image is convolved to one single feature at the highest level, high-level global features covering a large volume are extracted by the encoder.

As depicted in Fig. 5, the output layer consists of 4 maps, each of which specifies the confidence for a specific class for all pixels. A class label map can be induced from this output layer. Without loss of generality, let the class score map and class label map be $f = \{f_{ijk}\}$ and $C = \{C_{ij}\}$, where i,j,k are the row, column and class indices respectively. A label is assigned to pixel (i,j) based on the maximum class score:

$$C_{ij} = \arg\max_k f_{ijk}. \tag{6}$$

In this way, a class label map of size 512 * 512 is produced. This label map is the final result of the semantic segmentation and serves as a cloud and cloud shadow mask.

5. Experimental results

This section demonstrates the advantages of the proposed cloud detection method by a series of comparisons. First, the base model SegNet is compared with two widely used deep CNNs (DeepLab and PSPNet) for semantic segmentation based on RGB bands. Then, different bands and types of data are input to the adapted SegNet for further comparisons. Finally, the proposed method is evaluated and compared with CFMask, which has been used by the U.S. Geological Survey (USGS) to release the official Landsat products.

In each experiment, a model is trained using a training set and evaluated using a test set. The generated cloud and cloud shadow masks are compared with the ground truth and evaluated based on the confusion matrix and accuracies for the cloud shadow (CShadow), clear (Clear), thin cloud (TCloud) and cloud (Cloud) classes. In order to compare with the CFMask algorithm, the cloud and thin cloud classes are merged into one class (MCloud) to create another set of confusion matrix and accuracies.

The overall accuracy A^O, the producer's accuracy A^P and the user's accuracy A^U are calculated as:

$$A^O = \frac{\sum_i N_{ii}}{\sum_{ij} N_{ij}}. \tag{7}$$

$$A_j^P = \frac{N_{jj}}{\sum_i N_{ij}} \tag{8}$$

$$A_i^U = \frac{N_{ii}}{\sum_j N_{ij}} \tag{9}$$

where, A_j^P is the producer's accuracy for class j, A_i^U is the user's accuracy for class i, and N_{ij} is the number of pixels that come from class j and predicted as class i. The producer's and user's accuracies are the complements of omission and commission errors, respectively, which are alternatives for evaluation in the literature (Foga et al., 2017).

5.1. SegNet vs. the other networks

We compare SegNet with DeepLab (Chen et al., 2018) and PSPNet (Zhao et al., 2017). These CNNs are widely used for semantic segmentation by the computer vision community. Since they are developed to segment RGB images, three bands corresponding to Red, Green and Blue are input to the neural networks. More specifically, Digital Numbers (DNs) of Bands 4, 3 and 2 of Landsat 8 Operational Land Imager (OLI) are input to the CNNs. As reported in Table 3, SegNet performs as good as PSPNet, and they both outperform DeepLab.

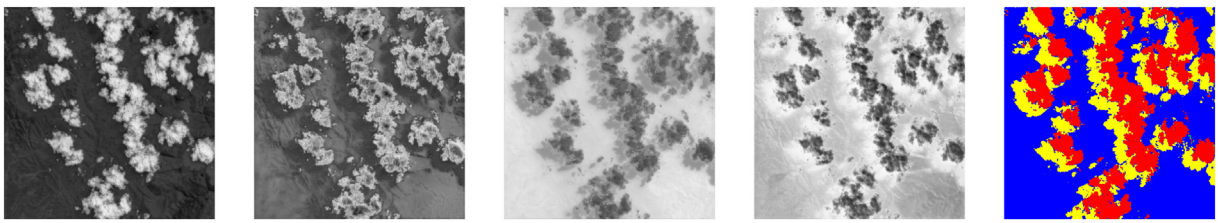

Fig. 5. Class confidence maps and class label map. The left 4 gray images represent the 4 confidence maps for cloud, thin cloud, clear and cloud shadow respectively. The right color image indicates a label map, in which cloud, thin cloud, clear and cloud shadow are illustrated as red, green, blue and yellow respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Overall accuracy A^O(%), producer's accuracy A^P(%) and user's accuracy A^U(%) are reported for different experiments based on two types of input data (DN and TOA), two band compositions (RGB and 10 bands) and three kinds of CNN networks (DeepLab, PSP and SegNet). Two versions of A^O are reported: the left and right values correspond to the classification of four classes (CShadow, Clear, TCloud, and Cloud) and three classes (CShadow, Clear, and MCloud) respectively.

Input	Bands	Network	Accuracy	CShadow	Clear	TCloud	Cloud	MCloud
DN	RGB	DeepLab	A^P	33.67	95.45	51.07	84.30	84.54
			A^U	64.29	92.47	57.54	79.70	85.16
			A^O			87.29		90.17
		PSPNet	A^P	56.40	96.54	64.89	90.61	90.95
			A^U	73.38	95.17	65.87	89.69	90.89
			A^O			91.18		93.44
		SegNet	A^P	56.31	95.73	72.31	90.07	93.89
			A^U	71.99	96.05	62.16	90.95	88.89
			A^O			91.14		93.45
	10 bands		A^P	71.48	96.78	73.86	93.69	93.71
			A^U	72.20	96.88	73.93	92.81	93.20
			A^O			93.23		94.93
TOA			A^P	72.45	96.67	67.97	96.08	94.69
			A^U	73.94	97.17	74.55	88.25	92.73
			A^O			93.03		95.11

5.2. RGB vs. ten bands

The distinguishing characteristics hidden in data can be exploited by CNNs to improve classification. Better performance is expected when more data are incorporated since more cues can be utilized to improve classification. To this end, the DNs of 10 bands of the Landsat 8 bands of 30 m resolution are input to the adapted SegNet described in Table 2. As reported in Table 3, when the RGB bands are replaced by 10 bands, the overall accuracy of four class (CShadow, Clear, TCloud, and Cloud) classification is improved from 91.14% to 93.23%, and the overall accuracy of three class (CShadow, Clear, and MCloud) classification is improved from 93.45% to 94.93%. These improvements support the above expectation. Therefore, we employ all bands for cloud detection.

5.3. DN vs. TOA

The Digital Number (DN) images of Landsat 8 reflective bands and thermal bands are converted to Top of Atmosphere (TOA) reflectance and brightness temperature respectively, and they are input to the adapted SegNet described in Table 2. As reported in Table 3, the segmentation quality based on DN and TOA reflectance or brightness temperature are similar. While the overall accuracy based on DN is better than that based on TOA reflectance or brightness temperature for the four class classification, the opposite is true for the three class classification. This interesting difference can be explained as follows. On one hand, certain cues distinguishing thin clouds from thick clouds hidden in DN images may be lost in the conversion from DN to TOA reflectance or brightness temperature. On the other hand, TOA reflectance or brightness temperature is more stable with respect to different ground scenes, and this characteristic is helpful for detecting thin and thick clouds as a whole. Due to these stable characteristics, we input TOA reflectance or brightness temperature to the adapted SegNet for cloud detection just as they are input to CFMask.

5.4. Cloud detection using adapted SegNet and ten bands TOA reflectance or brightness temperature

Based on the above experiments, we input ten bands TOA reflectance or brightness temperature to the adapted SegNet to detect clouds and cloud shadows. We compare this with CFMask, which is used to generate the official Landsat products released by the U.S. Geological Survey (USGS). CFMask is the production state-of-the-art and represents an informative benchmark as it is familiar to the Landsat cloud detection community as well as the Landsat user community more broadly. These users have a good sense of its performance.

To carry out a fair comparison in terms of the biomes of ground scenes, 60%, 10% and 30% 512 * 512 images of each scene are randomly grouped into the training, validation and test sets such that the number of images from different biomes are balanced in all three sets. In contrast, 60%, 10% and 30% 512 * 512 images from all scenes are randomly grouped into training, validation and test sets to carry out the experiments listed in Table 3.

5.4.1. Qualitative evaluation and comparison

Two L7 Irish examples from polar and subtropical regions together with three L8 Biome examples of snow, grass and water scenes are presented in Fig. 6 for visual demonstration. These examples represent various land surfaces, cloud sizes and cloud shapes. They also represent different performances of the proposed algorithm as their overall accuracies are 96.80%, 76.65%, 89.20%, 88.64% and 91.81% from top to bottom. The ground truth masks as well as the masks generated by CFMask are also presented for visual comparison.

Since multi-level spatial and spectral features covering different regions of the input image are captured by the deep CNN, the proposed method detects clouds and cloud shadows successfully as demonstrated by the five examples. The detected clouds and cloud shadows match the ground truth well. The Landsat 8 results match better than the Landsat 7 results as the small clouds and thin clouds in the last row are better preserved than those in the second row. This is possibly the result of two factors. First is that 10 bands of Landsat 8 as opposed to 8 bands of Landsat 7 are input to the neural network. Second is that the Landsat 8 OLI has 12-bit radiometric resolution as opposed to the 8-bit of the Landsat 7 ETM+. This allows the Landsat 8 images to capture more detailed characteristics of the scenes, which can then be fully exploited by the deep CNN. Some thin cloud pixels are confused with cloud pixels; however, this is not critical for Landsat data applications as both cloud and thin cloud are considered to be invalid for land surface monitoring.

In comparison, thin clouds are not distinguished by CFMask, and the clouds and cloud shadows produced by CFMask are not similar to those in ground truth. Overall, the results of CFMask are not as good as those of the proposed method. This is because multi-level spatial and spectral features covering large regions and all channels of the input image are not fully utilized by CFMask. Instead, it relies heavily on the TOA values. For example, the pixels with extremely low TOA values over water are incorrectly classified as cloud shadow in the first row and fourth column, and the pixels with extremely high TOA values over snow/ice are incorrectly classified as cloud in the third row and fourth column. Such errors are avoided by the proposed method as it exploits the multi-level spectral and spatial features.

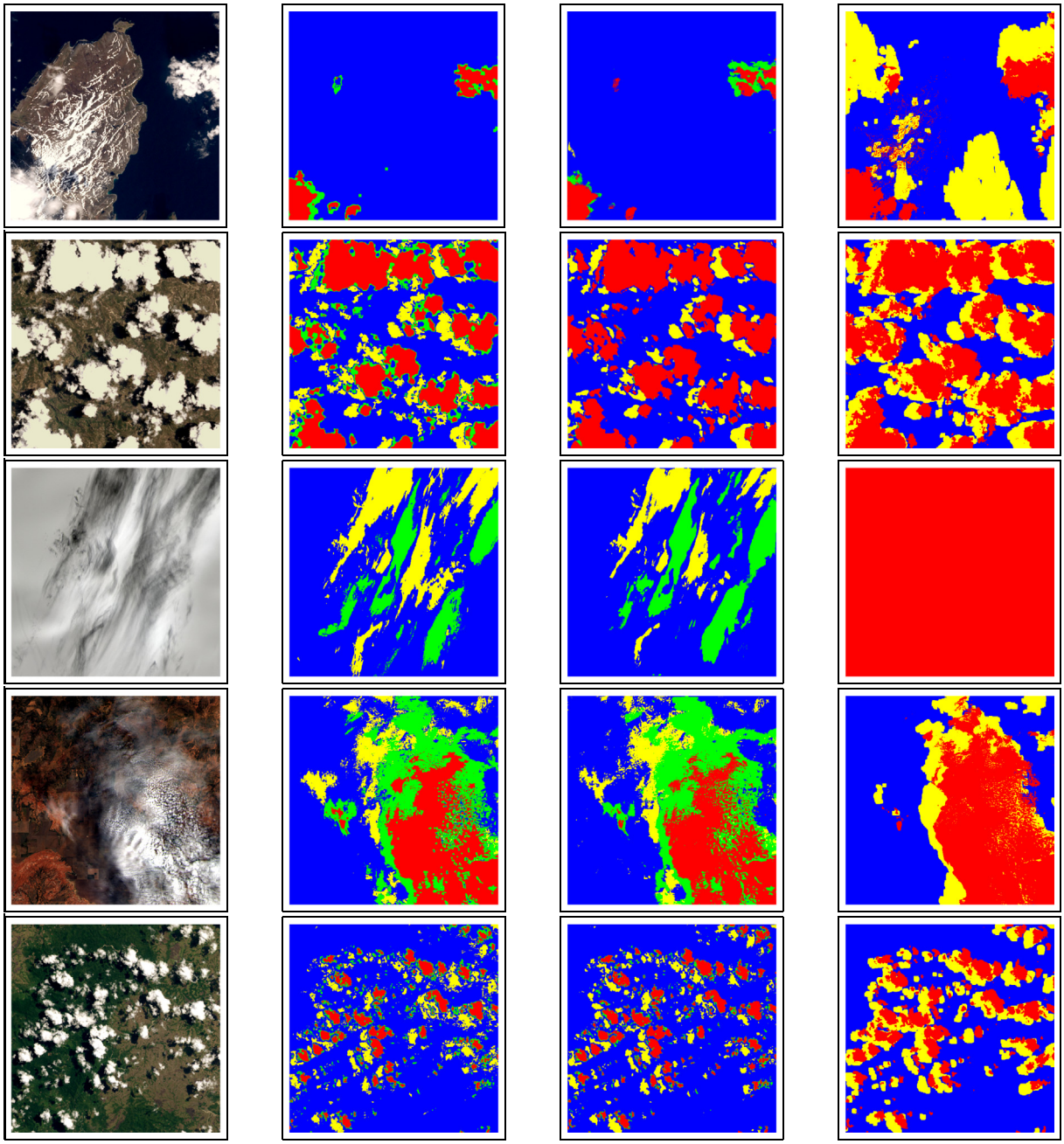

Fig. 6. Examples of cloud and cloud shadow detection in Landsat 7 and 8 images. Each image is 512 * 512 30 m pixels. From top to bottom, the examples are from scenes of (L7, Path 195, Row 10), (L7, Path 31, Row43), (L8, Path 1, Row 11), (L8, Path 29, Row 37) and (L8, Path 113, Row 63) respectively. From left to right are the input images, the ground truth cloud masks, the results of our method and the results of CFMask. The input images are true color composites of RGB bands. Cloud, thin cloud, clear and cloud shadow are illustrated as red, green, blue and yellow respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.4.2. Quantitative evaluation

Quantitative evaluations of the cloud and cloud shadow detection on L7 Irish and L8 Biome are presented in Table 4. The overall accuracies for both data sets are over 94%. For Landsat 7 images, the producer's accuracy and user's accuracy of cloud detection, thin cloud detection and cloud shadow detection are over 89%, 59% and 71% respectively. For Landsat 8 images, the producer's accuracy and user's accuracy of cloud detection, thin cloud detection and cloud shadow detection are over 92%, 74% and 71% respectively. Even for the challenging task of cloud shadow detection, both the producer's accuracy and user's accuracy are over 71%. Due to the two factors stated in Section 5.4.1, the performance on Landsat 8 is better than on Landsat 7. The performance of thin cloud detection is improved significantly. This is also demonstrated by the examples in Fig. 6, where the green pixels in the bottom three rows are more accurate than those in the upper two rows.

As described in Section 2.1, scenes in L8 dataset are evenly distributed over eight biomes respectively. A detailed evaluation in terms of these biomes is presented in Table 5. Only the accuracy for snow scenes is below 90%. This is due to the fact that snow is similar to cloud in the images. Benefiting from CNN's ability to explore both spectral and spatial cues to distinguish clouds from snow, the overall accuracy is over 86%. In contrast, CFMask completely fails to distinguish clouds from snow as illustrated by the third row of Fig. 6.

Table 4
Overall accuracy A^O(%), producer's accuracy A^P(%) and user's accuracy A^U(%) are reported for the adapted SegNet and CFMask on L7 Irish and L8 Biome respectively. Two versions of A^O are reported: the left and right values correspond to classification of four classes and three classes respectively.

Dataset	Network	Accuracy	CShadow	Clear	TCloud	Cloud	MCloud
L7	SegNet	A^P	71.43	97.81	59.92	89.78	86.51
		A^U	74.73	96.71	72.27	89.55	91.31
		A^O			94.33		95.26
	CFmask	A^P	78.75	91.47			83.60
		A^U	31.34	97.40			83.18
		A^O					89.88
L8	SegNet	A^P	71.54	97.76	74.58	94.94	93.07
		A^U	81.10	96.58	80.43	92.54	94.47
		A^O			94.00		95.47
	CFmask	A^P	49.33	87.47			82.66
		A^U	36.29	92.83			74.31
		A^O					84.58

Table 5
Overall accuracy A^O(%), producer's accuracy A^P(%) and user's accuracy A^U(%) are reported for the adapted SegNet on L8 Biome. They are evaluated in terms of different land covers.

Biome	Accuracy	CShadow	Clear	TCloud	Cloud	MCloud
Barren	A^P	70.88	98.26	72.77	95.24	95.24
	A^U	87.35	96.58	68.79	96.35	94.73
	A^O			93.55		95.45
Forest	A^P	82.86	98.25	71.79	96.36	93.94
	A^U	86.48	96.41	85.46	97.53	98.56
	A^O			95.41		96.16
Grass	A^P	77.60	98.31	71.76	95.22	91.04
	A^U	80.01	97.30	82.27	90.83	94.54
	A^O			94.86		96.22
Shrubland	A^P	64.98	97.99	78.09	96.52	97.38
	A^U	79.59	98.71	80.39	89.99	93.63
	A^O			95.95		97.37
Snow	A^P	57.65	93.98	78.03	77.51	87.73
	A^U	74.49	90.68	81.11	68.12	88.45
	A^O			86.19		88.92
Urban	A^P	66.75	99.11	68.72	90.49	90.33
	A^U	80.44	98.12	73.36	94.63	95.23
	A^O			96.70		97.52
Water	A^P	73.99	98.99	68.30	95.97	93.88
	A^U	78.46	98.34	83.07	92.46	96.83
	A^O			96.79		97.76
Wetlands	A^P	81.17	87.85	73.92	97.90	98.90
	A^U	79.66	93.26	74.37	95.73	97.15
	A^O			92.05		95.60

5.4.3. Quantitative comparison

The comparisons with the CFMask algorithm for L7 Irish and L8 Biome are presented in Table 4. Since thin clouds are not distinguished from clouds in the Landsat Collection 1, thin clouds and clouds in the results of the proposed method are merged into one class for a fair comparison.

The proposed method significantly outperforms the CFMask algorithm in terms of both the producer's and user's accuracies. The overall accuracy is improved by more than 5% and 10% for L7 Irish and L8 Biome respectively.

Again the improvement is more obvious for Landsat 8 due to the two factors stated in Section 5.4.1. The most significant improvement is for the Landsat 8 cloud shadow detection where the proposed method results in more than a 40% increase in the user's accuracy and a 20% increase in the producer's accuracy over CFMask. Basically, a pixel is identified by CFMask as a potential cloud pixel based on a set of spectral tests such as $B7 > 0.03$ and $BT < 27$ and $NDSI < 0.8$ and $NDVI < 0.8$, where, $NDSI = (B2 - B5)/(B2 + B5)$, $NDVI = (B4 - B3)/(B4 + B3)$, Bi is the TOA reflectance for Band i and BT is TOA Brightness Temperature for the thermal band. Although the view angle of the satellite sensor and the illuminating angle are used to match clouds and cloud shadows, CFMask relies significantly on spectral tests rather than complex spatial and spectral features at multiple levels. Without such multi-level features, it is difficult to distinguish cloud shadows from certain land surface classes, e.g., water. All the user's and producer's accuracies of all the classes for both Landsat 7 and Landsat 8 are higher for the proposed method than for CFMask except for the producer's accuracy for Landsat 7 cloud shadow. This exception may be because CFMask intentionally buffers the cloud shadow masks to boost the producer's accuracy; that is all pixels neighboring a cloud shadow pixel are identified as cloud shadow pixels. The developers of the method argue that commission error is preferred over omission error (Zhu and Woodcock, 2012). Note that this commission error preference can be easily achieved by applying the same buffering techniques to the generated cloud masks. Moreover, it can be introduced into the loss function so that a CNN model can learn the preference.

5.5. Efficiency

The main computation in both the convolution and deconvolution stages is filtering, which is implemented as a dot product of two vectors. This computation is independent for each pixel and so can be easily parallelized, particularly on GPUs.

It takes around 15 h to train a model. However, only 0.45 s is needed to segment a 512 * 512 image. All the tiled images for a typical Landsat scene can be segmented in less than 2 min, and they can be assembled together based on the partition grid to produce a label map for the entire scene.

All the experiments are carried out on a desktop computer with an Intel Core i7-7700 K CPU (4-Cores, 8 MB Cache, Turbo Boost 2.0, Overclocked up to 4.4 GHz on all four cores), an NVDIA GeForce GTX 1080Ti GPU (with 11GB GDDR5X) and 32 G (2 * 16 G) 2400 MHz DDR4 Memory. This indicates that fast processing of Landsat images can be achieved on a standard desktop computer with a single GPU.

6. Conclusion and discussion

This paper proposes a method based on deep CNNs for cloud and cloud shadow detection in Landsat imagery. The problem is formulated as one of semantic segmentation. The CNN based semantic segmentation allows spatial and spectral features computed over large spatial regions to be used to classify pixels as cloud, thin cloud, cloud shadow or clear. The semantic segmentation is achieved via a transformation from an input image to a label map based on an adapted SegNet, which has proven effective in the field of computer vision (Badrinarayanan et al., 2017). The revised number of channels allows the network to deal with multi-band Landsat imagery, and the skip connections introduced between the convolution and deconvolution layers allows detailed features to be utilized in the deconvolution. Extensive experiments demonstrate that state-of-the-art performance is achieved by the proposed method with overall accuracies of 94% for both the Landsat 7 and Landsat 8 imagery.

Although significant advantages were found using the proposed deep CNN based method for Landsat cloud and cloud shadow detection, *operational* Landsat cloud detection needs to consider the following issues. First is dealing with the fill pixels in the Landsat imagery. In our current framework, a Landsat scene is partitioned into a set of 512 * 512 30 m non-overlapping images to be fed into the deep CNN model. Only image blocks without fill pixels are currently considered. However, image blocks at the boundaries of Landsat scenes usually contain fill pixels. Moreover, the Landsat 7 ETM + Scan Line Corrector (SLC) images contain fill strips. This issue can be addressed by treating fill pixels as an additional class. The proposed method can then learn and distinguish fill pixels from clouds, thin clouds, cloud shadows and clear pixels. The second issue concerns cirrus cloud detection. Since the Landsat 8 band design has the capability to detect cirrus clouds

(Kovalskyy and Roy, 2015), it is expected that the proposed method can detect cirrus clouds in Landsat 8 imagery. This issue can also be addressed by treating cirrus clouds as an additional class. By collecting reference data with labeled cirrus clouds, the proposed method can learn to distinguish cirrus clouds from the other classes. We expect that the proposed CNN based method will perform better than the commonly used spectral test based methods for cirrus cloud detection. The third issue concerns the generalization of the model. Since the training samples used in this study are limited, they may not represent all kinds of land surface, clouds and cloud shadows. It is not clear that the trained model is universal enough for operational Landsat cloud and cloud shadow detection. This issue can be addressed by collecting more and varied cloud and cloud shadow reference data. Generally, good prediction for new images can be achieved by a model trained using a wide variety of training samples. More reference data can also provide enough testing images to confirm that a universal model is achieved.

The proposed method takes multiple bands as input for cloud and cloud shadow detection and can be easily extended to similar sensors provided that the training data are collected. For example, the Sentinel-2 satellite (Drusch et al., 2012) is similar to Landsat, however, thermal bands are replaced by red edge bands. It is straightforward for the proposed neural network to deal with the Sentinel-2 multi-band images. Since the information hidden in the data can be exploited by deep CNN based methods, it is expected that the proposed framework can detect clouds and cloud shadows in Sentinel-2 multi-band images.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 41571335). It was also supported by the China Scholarship Council for visiting research at the University of California, Merced.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., et al., 2012. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34 (11), 2274–2282.

Amato, U., Antoniadis, A., Cuomo, V., Cutillo, L., Franzese, M., Murino, L., Serio, C., 2008. Statistical cloud detection from seviri multispectral images. Remote Sens. Environ. 112 (3), 750–766.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens. Environ. 120, 25–36.

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

Frantz, D., Röder, A., Udelhoven, T., Schmidt, M., 2015. Enhancing the detectability of clouds and their shadows in multitemporal dryland Landsat imagery: extending Fmask. IEEE Geosci. Remote Sens. Lett. 12 (6), 1242–1246.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

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 (8), 1747–1755.

Hollingsworth, B.V., Chen, L., Reichenbach, S.E., Irish, R.R., 1996. Automated cloud cover assessment for Landsat TM images. In: Imaging Spectrometry II. vol. 2819. pp. 170–180.

Hughes, M.J., Hayes, D.J., 2014. Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing. Remote Sens. 6 (6), 4907–4926.

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 (10), 1179–1188.

Jin, S., Homer, C., Yang, L., Xian, G., Fry, J., Danielson, P., Townsend, P.A., 2013. Automated cloud and shadow detection and filling using two-date Landsat imagery in the USA. Int. J. Remote Sens. 34 (5), 1540–1560.

Ju, J., Roy, D.P., 2008. The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally. Remote Sens. Environ. 112 (3), 1196–1211.

Kovalskyy, V., Roy, D.P., 2015. A one year Landsat 8 conterminous United States study of cirrus and non-cirrus clouds. Remote Sens. 7 (1), 564–578.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998a. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

LeCun, Y., Bottou, L., Orr, G.B., Müller, K.-R., 1998b. Efficient backprop. In: Neural Networks: Tricks of the Trade. Springer, pp. 9–50.

Lee, J., Weger, R.C., Sengupta, S.K., Welch, R.M., 1990. A neural network approach to cloud classification. IEEE Trans. Geosci. Remote Sens. 28 (5), 846–855.

Lee, Y., Wahba, G., Ackerman, S.A., 2004. Cloud classification of satellite radiance data by multicategory support vector machines. J. Atmos. Ocean. Technol. 21 (2), 159–169.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Molnar, G., Coakley Jr, J., 1985. Retrieval of cloud cover from satellite imagery data: a statistical approach. J. Geophys. Res. Atmos. 90 (D7), 12960–12970.

Qiu, S., He, B., Zhu, Z., Liao, Z., Quan, X., 2017. Improving Fmask cloud and cloud shadow detection in mountainous area for Landsats 4–8 images. Remote Sens. Environ. 199, 107–119.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99.

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 (6), 2741–2760.

Roy, D.P., Ju, J., Kline, K., Scaramuzza, P.L., Kovalskyy, V., Hansen, M., Loveland, T.R., Vermote, E., Zhang, C., 2010. Web-enabled Landsat data (WELD): Landsat ETM+ composited mosaics of the conterminous United States. Remote Sens. Environ. 114 (1), 35–49.

Roy, D.P., Wulder, M., Loveland, T.R., Woodcock, C., Allen, R., Anderson, M., Helder, D., Irons, J., Johnson, D., Kennedy, R., et al., 2014. Landsat-8: science and product vision for terrestrial global change research. Remote Sens. Environ. 145, 154–172.

Scaramuzza, P.L., Bouchard, M.A., Dwyer, J.L., 2012. Development of the Landsat data continuity mission cloud-cover assessment algorithms. IEEE Trans. Geosci. Remote Sens. 50 (4), 1140–1154.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. arXiv preprint. https://arXiv:1409.1556.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Sun, L., Liu, X., Yang, Y., Chen, T., Wang, Q., Zhou, X., 2018. A cloud shadow detection method combined with cloud height iteration and spectral analysis for Landsat 8 OLI data. ISPRS J. Photogramm. Remote Sens. 138, 193–207.

Tian, B., Shaikh, M.A., Azimi-Sadjadi, M.R., Haar, T.H.V., Reinke, D.L., 1999. A study of cloud classification with neural networks using spectral and textural features. IEEE Trans. Neural Netw. 10 (1), 138–151.

Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA Neural Netw. Mach. Learn. 4 (2), 26–31.

Vermote, E., Justice, C., Claverie, M., Franch, B., 2016. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. Remote Sens. Environ. 185, 46–56.

Wang, B., Ono, A., Muramatsu, K., Fujiwara, N., 1999. Automated detection and removal of clouds and their shadows from Landsat TM images. IEICE Trans. Inf. Syst. 82 (2), 453–460.

Wulder, M.A., White, J.C., Loveland, T.R., Woodcock, C.E., Belward, A.S., Cohen, W.B., Fosnight, E.A., Shaw, J., Masek, J.G., Roy, D.P., 2016. The global Landsat archive: status, consolidation, and direction. Remote Sens. Environ. 185, 271–283.

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 (8), 3631–3640.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890.

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.

Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94.

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.

Zi, Y., Xie, F., Jiang, Z., 2018. A cloud detection method for Landsat 8 images based on PCANet. Remote Sens. 10 (6), 877.