**Simultaneous Cloud Detection and Removal From**

**Bitemporal Remote Sensing Images Using Cascade**

**Convolutional Neural Networks**

Shunping Ji , Member, IEEE, Peiyu Dai , Student Member, IEEE, Meng Lu , and Yongjun Zhang

Cloud detection

Cloud removal

Reconstruction: After the processes for cloud detection and removal, reconstructing the shaded regions with the neighboring pixels or multisource data.

(Partially different from the image restoration which is the process for recover the degraded image to the original appearance)

Clouds in the remotely sensed images highly influence the application of them, like the operation of localization, image interpretation, data fusion, as well as the obj detection.

Detect – Remove – Restore the shaded region

Difficulties: Fractal structures in geometry & the diversity in the spectrum of the clouds

**Methods for cloud and cloud shadow detection:**

1. Threshold methods: Sensitive spectral bands + Discriminative morphological, biological, or physical information

2. Change-detection-based methods: detect clouds from the multitemporal images

(change is random due to the multi-various imaging conditions, sensors, seasons, and so on; false changes (accidental) are inevitable.)

3. Conventional (traditional) supervised machine learning: SVM

4. Mainstream: deep-learning based cloud detection: CNN, FCN (with different optimization strategies)

**System requirements:**

Keras framework

Windows 10

NVIDIA 11G 1080-Ti GPU

## Proposed framework:

**Simultaneously detect and remove cloud and cloud shadow;**

**With higher detection accuracy;**

**With higher effects of image restoring performance.**

**Overall**

Cascade convolutional neural networks

Integrated cloud detection and removal framework

**1**

FCN (fully convolutional network) + Multiscale aggregation + Channel-attention mechanism

Cloud and cloud shadow detection

**2**

Another FCN + Detected cloud and shadow mask + cloudy image + temporal image

Cloud removal + missing-info reconstruction

(self-training strategy: learn the mapping between the clean-pixel pairs of the bitemporal image)

(Bypasses high demand of manual labels)

## Background Information:

### Cloud Removal

Missing-information reconstruction

1. Spatial-based methodology
2. Spectral-based methodology
3. Temporal-based methodology
4. Multisource-based methodology
5. **Spatial-based methodology:**

Clean pixels of an individual cloudy image

Recover the shaded regions

**Assumption:**

The pixel below the cloud & the shadow

possessing the similar textures with the cloud-free neighborhood.

The neighbor can restore them.

**Limitation:**

On the large size and complex RS images,

similarity is rarely met

(between the shaded area and its neighbor)

1. **Spectral-based methodology:**

Different spectral bands

Complementary info

**Prerequisite:**

parts of the multispectral bands can penetrate thin clouds

**Limitation:**

Incompetent in thick clouds situation

1. **Temporal-based methodology:**

Different time series at the same area

(not only use the cloudy image)

Additional observations

Reconstruct the corrupted region

**Advantages:**

More genetic and with better performance

The **use** of multitemporal images:

**Four categories:**

1. Temporal replacement
2. Temporal interpolation
3. Temporal learning

**Temporal replacement：**

Corresponding clean pixels in the temporal images

Replace the missing regions in the cloudy images

**Temporal interpolation:**

Similar to replacement

But with interpolation strategies.

**Temporal learning:**

Learning -based methods

Establish the relationship between the temporal images

E.g

Dictionary learning algorithm

Recover cloud and shadow regions from multitemporal images

RF, CNN

Different tasks of missing-information reconstruction

**Limitations**

1. Sensitive to geometrical registration errors
2. Abrupt spectral changes between the images

Worse image-to-image mapping

1. **Multisource-based methodology:**
2. Fusing the RS data from different types of sensors
3. Auxiliary by (SAR) data

(microwave signals are more capable of penetrating clouds)

## Model:

**Integrated spatial-temporal-spectral model**

**Input: different spectral bands and temporal image**

**Operate: CNNs (process spatial and spectral info)**

### Overall

Two cascaded FCNs

1. CDN: Detect cloud and cloud shadow

(only in the training stage)

Input 1: Cloud & cloud shadow mask

1. CRN

(Restoring the cloud-shaded region)

Input 2: Bitemporal images

### CDN Structure:

A Pixel to pixel FCN

A Dense Net-style building block in each scale of encoder

A multiscale feature-fusion strategy

A channel-attention mechanism

encoder: Dense-Net in each scale

Down-sampling larger: 2 max pooling

decoder: Up-sampling layer: 2 deconvolution

lateral connections

**Output of CDN:**

Cloud mask map

(same size with the input image)

**Multiscale strategy:**

**Concatenation of the features from different scales**

Up-sample last features of each scale -- original resolution

Compression: apply a 1 × 1 convolution

Activation by Re-LU

Concatenate them channel-wisely

multiscale feature map

(The concatenated four gray layers)

(Inconsistence between channels since from different scales)

(CAM) channel-attention module

Global consistency

**Structure of CAM**

Multiheaded self-attention strategy + between channel interdependence calculation

Input: Feature map (from CNN)

3 independent 3 × 3convolutions

K, Q, V

(3 parallel feature maps)

……

Channel-attention boosted feature map

**Multicategory cross-entropy loss function**

Probability map

(how likely a pixel belongs to the cloud and cloud shadow.)

low threshold to translate (=high recall)

Cloud Mask map

(most of the cloud and cloud shadow regions are covered with a high recall)

### CRP: Cloud Removal Process

Training stage (CRN)

Prediction stage (CRN)

Self-trained CRN (based on an FCN)

Simultaneously handled

**Dual task of CRP:**

1. Radiometric transformation

Temporal image -- current image

(ensure radiometric consistency)

Learned transformation parameters

1. Restoration of the shaded area

**What is so called ‘self-trained CRN’?**

Another FCN

Supervised deep learning-based method

Self-training strategy

Without any manually labeled samples

No need large amount of real samples

Map a temporal image --- a current cloudy image

Use relations of clean-pixel pairs between the bitemporal images

(the invalid pixels: cloud and shadow have been removed by CDN)

**Training stage:**

**Overall:**

CDN

+

Bitemporal images

Bitemporal images with Automatically removed shaded pixels

(clean image: the cloudy areas are none)

Corresponding pixels of the two-time images

CRN

Learn the relative radiometric transformation between them

**Kernel:**

Randomly simulate the cloud regions

(no need for true samples)

Training samples:

Train…

Highly accurate recovering model

Restore an image

Approach the original cloudy image with real clouds excluded

**Details:**

Current cloudy image

CDN mask

Current clean images

Randomly simulate new cloud in large amount

Simulated damaged image

(Current simulated image)

+

Temporal image

Bitemporal images

(input)

CDN mask

Temporal clean images

Reconstruct image

Restore images

**Prediction Stage**

Original bitemporal image

CRN

Repaired image

**Advantages:**

1. Pixels in the simulated cloud region = a part of ground truth

assess quantitively the performance of an algorithm

1. The simulation of a large amount of random clouds – more robust model

recover arbitrary cloud-shaded area in the real RS images

1. No need to know the cloud location and region since the model has learned how to fix arbitrary clouds in an image through the training process with simulated clouds

## Data set & Experiment setting

### Data set introduction

WHU Cloud data set

6 image pairs:

cloudy + cloud free

(From Landsat-8 data set)

+

6 Bitemporal images

(Corresponding)

(temporal reference data, acquired at a similar time)

(Avoid real land cover changes)

Ground truth: manually delineated cloud and cloud shadow

Auxiliary evaluation: Additional GF cloud-detection data set

(10 GF cloudy images, covering multifarious landcovers)

To comprehensively evaluate the method

Red, Green, Blue

(B4, B3, B2)

(Landsat8 + GF)

To make the algorithm compatible

Crop image

(512 × 512 patches)

Training & Prediction

(GPU memory capacity)

### Experimental setting:

Cloud-detection task:

Train: 680 patches

WHU Validate: 50 patches

Test: 129 patches

Train: 1604 patches

GF Validate: 1200 patches

Test: 1046 patches

Adaptive moment estimation (Adam) optimization algorithm

Iterated 1000 epochs

## Quantitative Assessment

### CDN

**Compared Methods:**

DeeplabV3+, U-Net, MF-CNN, MSCN

**Indicators:**

IoU, Recall, Precision (foreground detection)

OA (pixel-level accuracy; foreground & background detection)

**Results:**

**WHU cloud data set:**

CDN > U-net > DeeplabV3+ > MFCNN > MSCN

CDN: close to ground truth

U-net: over-smooth (lack of multiscale aggregation strategy)

DeeplabV3+: weak in shadow detection (data less; up-sampling process coarse)

MFCNN: oversharp boundaries

MSCN: batch size small (2) Batch normalization is poor poorer performance

**GF-2 Dataset:**

CDN > DeeplabV3+ MSCN, MF-CNN > U-net

Better than other methods at various terrain and atmospheric conditions

### CRN

**Compared with other methods**

1. Algorithm (model/network) ability:

Different model with different algorithm, possessing different training results (predicting results)

Compared under the same training samples (simulated cloud mask (black pixels))

Same cloud simulator

1. Reconstruction results

**Compared with self**

Algorithm robustness

**Compared with other methods:**

Cloud-removal results:

Training results (prediction results)

Reconstruction effect

Comparison between 3 cloud-removal methods:

CRN, STSCNN, U-Net

Representative indicators:

SSIM, PSNR, SAM, CC

Randomly simulate the missing pixels in a current image without clouds

Train a network with the input of bitemporal images

Produce a repaired image approaching to the current image

Ground truth is whole current image

Evaluate comprehensively

**Process:**

Current cloudy image + Temporal image

CRN model

(Trained with the simulated masks)

Compare the prediction results of 3 methods

**U-net**

U-net & CRN share similar FCN structure

1. U-net no multiscale strategy

CRN with multiscale strategy

(concatenate features from multiscale)

Multiscale Strategy Improve the performance of FCN

1. Only 1 up-sampling path (prediction)

**STSCNN**

Dilated convolutions

Extract features with different-size receptive fields

Inferior

1. Fixed size of all features map;

Shallow Feature channels

Time- & Memory- consuming CNN

1. No scale robustness

Lack aggregation of multiscale info

**Results:**

CRN > U-net >STSCNN

CRN exceeded

1. STCNN more like temporal image

Contains obvious raw textures from the temporal image

Radiometric transformation between bitemporal image is not good

1. CRN & U-net more like ground truth

Better in textural and spectral info preservation

CRN reconstruct more details, preserve finer and clearer textures

CRN closest to the ground truth

U-net blur the image

**Compared with self:**

Algorithm ability & robustness

Vary simulated mask’s sizes & types

Stay same

Model (Algorithm) robust

## Discussion

**Done**

Integrate the CD, CSD & CR tasks

Improve model from recent methods

Provide a general framework for accommodating different algorithm or tasks

E.g. spectral-based cloud detection; missing info reconstruction problems

Advance CDN: FCN + multiscale strategy + a densely connected encoder + CAM

**Building blocks**

(for a CNN-based feature extractor)

ResNet-style

VGG

DenseNet block

For Cloud Detection:

1. BN layer has side effect when the batch size is small

CDN+BN < CDN no BN

1. Multiscale strategy is important

U-net (no multiscale strategy) perform worst (oversmooth )

1. CRN (DenseNet) with least params marginally advantageous than other models
2. Use all features at different levels in the encoder perform the best
3. CAM is effective

Boosted with CAM reorganize the between channel consistency of the multiscale features

More details on individual cloud and shadow with precise boundaries

Best: CRN(DenseNet256) + CAM = CDN

1. CRN(DenseNet) =CRN(VGG)=CRN(ResNet)

**Growth Rate**

Dimension of the feature maps in each DenseNet block

(feature dimension & total number of params)

1. A small growth rate is sufficient to obtain perfect detection performance
2. CRN (DenseNet) with 256 growth rate achieves the most comprehensive and complete detection

**FCN structure**

**For Cloud Removal**

Performance of different building blocks & CAM

1. Opposite to results in Cloud Detection
2. CRN (VGG) is the best
3. Simplest VGG performs well
4. CAM is not necessary

(uncommon, CR is a pixel-pixel mapping from the temporal image to the current image;

Concentrate on parts of pixels seems unnecessary)

1. Multiscale strategy is necessary

**For Cloud Detection & Cloud Removal**

1. the performance of a CNN on a specific task is not fully dependent on the complex structures and extensive parameters, even if samples are sufficient
2. empirical network structure designing is important for a specific remote sensing data processing task, just as handcraft feature designing for a conventional machine learning method
3. Automated machine learning (AutoML) technology: automatically find optimal models for a specific task

**Land Type**

1. CD worse in snow, ice regions

(similar patterns shared between snow, ice, and clouds)

1. Bare land also obtained a lower score

(inadequate training samples)

**Limitations:**

1. OK for multi or hyperspectral, medium- or low-resolution RS images

(RS data are structured, which can be easily regularized)

1. Limited on high-resolution images

(registration accuracy is hard to reach subpixels)

(DSM/crop the image smaller can solve the problem)

1. There are various types of clouds in RS dataset, however the deep learning-based methods are generic.
2. The percentage of cloudage would not influence the result, therefore is not a significant params to CRN, since one the model is robust, in addition the image containing more than 30% or 50% would be removed rather than repaired
3. Results of CRN will be affected by performance of CDN

The performance of CDN would be influenced by the diversity of RS data, different imaging situation and labeled training samples.

Improve the recall rate and consider the precision-recall curve of the background,

Which means preferentially using the clean pixels with a high degree of confidence.