GLOBAL CONTEXT AIDED SEMANTIC SEGMENTATION FOR CLOUD DETECTION OF REMOTE SENSING IMAGES

Fei Wen, Yongjun Zhang*, Bin Zhang

School of Remote Sensing and Information Engineering, Wuhan University, China - (wenfei, zhangyj, bin.zhang)@whu.edu.cn

Commission II, WG II/6

KEY WORDS: Cloud detection, CNN, Semantic segmentation, Context, Landsat-8

ABSTRACT:

Cloud detection is a vital preprocessing step for remote sensing image applications, which has been widely studied through Convolutional Neural Networks (CNNs) in recent years. However, the available CNN-based works only extract local/non-local features by stacked convolution and pooling layers, ignoring global contextual information of the input scenes. In this paper, a novel segmentation-based network is proposed for cloud detection of remote sensing images. We add a multi-class classification branch to a U-shaped semantic segmentation network. Through the encoder-decoder architecture, pixelwise classification of cloud, shadow and landcover can be obtained. Besides, the multi-class classification branch is built on top of the encoder module to extract global context by identifying what classes exist in the input scene. Linear representation encoded global contextual information is learned in the added branch, which is to be combined with featuremaps of the decoder and can help to selectively strengthen class-related features or weaken class-unrelated features at different scales. The whole network is trained and tested in an end-to-end fashion. Experiments on two Landsat-8 cloud detection datasets show better performance than other deep learning methods, which finally achieves 90.82% overall accuracy and 0.6992 mIoU on the SPARCS dataset, demonstrating the effectiveness of the proposed framework for cloud detection in remote sensing images.

1. INTRODUCTION

Due to their imaging mechanism, optical satellite sensors are inevitably influenced by cloud which can severely degrade image quality or even completely occlude land-covers. Remote sensing images acquired by such sensors may contaminated by cloud with high probability, hindering their downstream applications such as land-cover classification, change detection, environment monitoring and so on. Cloud detection, as a preprocessing step, plays an important role in remote sensing image utilization. Screening out cloud and the accompany shadow can facilitate the following image processing and analysis. Therefore, many researchers have been dedicated to cloud detection these years.

The available cloud detection methods can be roughly categorized into rule-based and machine learning approaches. Generally, the rule-based methods exploit reflectance variance in different bands and introduce sets of rules that threshold on single spectral band or combination of spectral bands to identify cloud in remote sensing images. Cloud shadow is then detected by considering solar geometry and its relative location to cloud. However, the rule-based methods are sensitive to image sensors and different scenes because of their empirical rule sets, which lower their generalization ability. In recent years, machine learning methods have been applied in many cloud detection tasks. Traditional machine learning methods train classifiers to identify cloud pixels or objects described by handcrafted features. Contrary to utilizing manually designed features, deep learning methods which employ CNNs to learn more representative features from training data has achieved superior performance. At the early stage, CNNs are trained to classify superpixels or sliding windows of the input image to generate a cloud probability map, cloud detection result is then obtained by setting proper threshold on that map. More straightforwardly, cloud and shadow can be screened out in semantic segmentation frameworks, by which pixel-wise classification can be directly generated. However, the available semantic segmentation networks used for cloud detection only utilize local or non-local features through stacked convolution and pooling layers while ignoring global contextual information of the whole input scene. In practice, for large remote sensing images, they have to be cropped into small blocks before being fed into CNNs because of limited computational resources, which may result in part of cropped blocks having no cloud pixels or no clear pixels at all. Examples are shown in Fig 1. In such cases, global contextual information about what classes presented in the input scene may help to prevent networks from

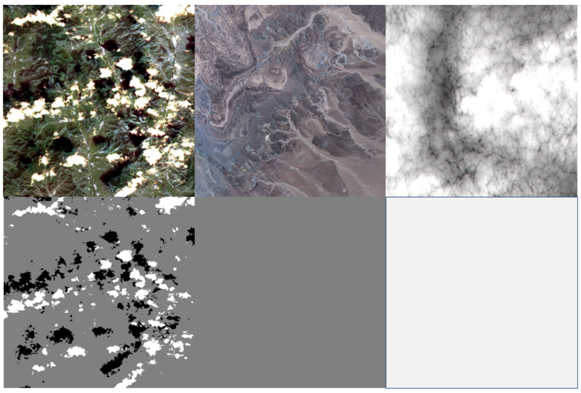

Figure 1. Examples of remote sensing scenes and their pixel-wise labels with three classes (white for cloud, black for shadow and gray for clear). For part of scenes having no cloud or clear pixels, we explore to capture global context of such scenes to prevent networks from falsely detecting cloud or clear pixels.

[*] Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2020-583-2020 | © Authors 2020. CC BY 4.0 License.

falsely detecting nonexistent objects, thus improving segmentation accuracy.

To encode global context information of satellite image blocks, a novel global context aided semantic segmentation network (GCANet) is proposed in this paper. In addition to a U-shaped semantic segmentation network with encoder-decoder architecture, a multi-class classification branch is built on top of the encoder module to capture global context of the input scene. Linear representation learned in this branch is then combined with featuremaps of the decoder at different scales. In such a way, the global context information extracted by the added branch is injected into the semantic segmentation stream. Experiments on two Landsat-8 datasets show better performance compared to regular semantic segmentation networks, demonstrating the effectiveness of the proposed method.

2. RELATED WORKS

Depending on the type of image sources, cloud detection approaches can be divided into single-image-based and multi-temporal-based ones (Zhang et al., 2019). In this work, we discuss about the former category and discard the multi-temporal-based methods. As mentioned, methods for detecting cloud in a single image can be categorized into rule-based and machine learning ones. Further, to help understanding, we divide machine learning approaches into traditional and deep learning parts.

Rule-based methods. To mask cloud and the accompany shadow in remote sensing images, the most basic way is to introduce a set of thresholds on the original or derived products of image spectral bands. For example, two of the most well-known rule-based methods are Fmask (Zhu, Woodcock, 2012, Zhu et al., 2015) and Sen2Cor (Richter et al., 2012), which has been officially used for Landsat and Sentinel-2 images respectively. After that, some researchers have proposed improvements based on the two methods (Qiu et al., 2019, Zhai et al., 2018). The rule-based methods are easy to use once the set of thresholds have been carefully determined through plenty of validation. However, the predefined rules, namely those thresholds, are sensitive to image sensors, different scenes and illumination conditions, just to name a few. Besides, the rule-based methods often depend on certain spectral band that has good reaction to cloud, which is not always available to many sensors, lowering their generalization ability. Therefore, more intelligent cloud detection methods need to be investigated.

Traditional machine learning methods. Different from setting a set of thresholds on spectral bands of images to detect cloud, machine learning methods can learn from higher level features of images to distinguish cloud pixels from clear ones. Given well-labeled data pairs (images and the corresponding cloud masks), these methods train or learn a classifier from the training samples described by handcrafted features. (Scaramuzza et al., 2011) proposed two approaches to improve the ACCA (Hollingsworth et al., 1996) algorithm designed for Landsat TM images, one of which added a neural network to refine ambiguous results and another created a larger decision tree to improve accuracy. (Hughes, Hayes, 2014) developed the SPARCS (Spatial Procedures for Automated Removal of Cloud and Shadow) to identify and classify cloud and cloud shadow, which utilized a neural network approach to determine classification membership of each pixel in Landsat images. To explore the representation capability of off-the-shelf image features, several SVM (Support Vector Machine)-based methods were proposed to train linear classifiers to identify cloud superpixels or blocks. For instance, (Li et al., 2015) exploited gradient and gray level co-occurrence matrix to calculate descriptor for sub-blocks, (Tan et al., 2016) utilized latent semantic model to represent superpixels as training samples for the SVM. Despite machine learning methods can achieve good results in some cases, they are still not adaptable enough in practice due to the limited representation capability of hand-crafted features. With the development of data acquirement ability, it's easier to collect large dataset for our specific tasks, which can derive more flexible and data-driven methods.

Deep learning methods. In recent years, deep learning has achieved great success in computer vision tasks such as object detection, image classification, scene parsing, and so on. Through CNN, deep learning methods can learn more representative features directly from data and have shown significant superior performance compared to traditional machine learning methods. Therefore, such powerful tool has also been widely used in remote sensing society these years.

As for detecting cloud and cloud shadow in remote sensing images, the available methods involving deep learning can be grouped into classification-based and segmentation-based ones. From the point of classification, (Shendryk et al., 2019) proposed a multi-network ensemble strategy to perform multi-label classification for high-resolution satellite sub-scenes with cloud, shadow or land-cover. (Shi et al., 2016) trained a CNN to learn features of superpixels and generated cloud probability map by identifying each superpixel of the input image. More straightforwardly, cloud, cloud shadow and land-cover pixels can be directly labeled through semantic segmentation. (Jeppesen et al., 2019) introduced a U-net (Ronneberger et al., 2015) based network with a symmetric architecture linked by skip connection to screen out cloud and cloud shadow. Similarly, (Chai et al., 2019) applied the well-known SegNet (Badrinarayanan et al., 2017) architecture that consisted of a encoder and a decoder module for cloud detection, of which the encoder was based on VGG (Simonyan, Zisserman, 2014) while the decoder was made up of several up-sampling layers. To improve feature learning, (Shao et al., 2019) constructed a multiscale feature fusion network to detect thick and thin cloud and non-cloud pixels of remote sensing images. These networks are end-to-end and have been demonstrated to perform better than traditional methods. However, they only extract local or non-local features through stacked convolution and pooling operations or fuse multiscale features including global spatial features, the global context information of the input scene has not been fully utilized for remote sensing images.

In contrast, we propose a semantic segmentation network combined with global contextual information, in which an auxiliary multi-class classification branch is added to encode contextual information of the whole input scene. The advantage is intuitive: if the input scene is completely clear without any cloud pixels, the features learned from the classification branch should help to selectively strengthen land-cover related features while weakening cloud related features, reducing the probability of falsely detecting some pixels as cloud in such scene, thus improve the semantic segmentation accuracy.

3. METHODOLOGY

With the development of deep learning, some researchers have altered to apply CNNs on cloud detection tasks. The detection

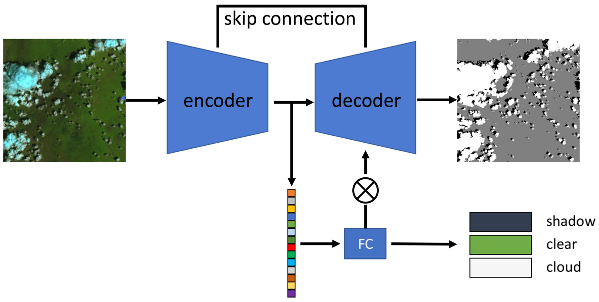

Figure 2. Flowchart of the proposed GCANet.

of cloud and its shadow in remote sensing images is treated as pixel-wise labeling problem, which is solved using a variety of semantic segmentation networks. As shown in Fig 2, the proposed framework in this paper also utilizes a semantic segmentation architecture as skeleton, while additionally a new branch is added to assist segmentation. The skeleton is consisted of encoder and decoder module linked by skip connection. The added branch performs multi-class classification on top of featuremaps of the encoder. Besides, the linear representation learned from this branch is combined with featuremaps of the decoder at each scale. The whole framework is end-to-end for training and testing. Details of the skeleton and the added branch are described in the following sections.

3.1 U-shaped semantic segmentation architecture

Depending on whether the featuremaps are up-sampled back to original resolution directly or step by step, the architectures of semantic segmentation networks can be roughly categorized as FCN-like or U-shaped. The FCN-like architectures are often with larger model size, while the U-shape architectures are more lightweight and have been proven to be effective for tasks with limited training samples. Inspired by U-Net (Ronneberger et al., 2015) which has been successfully used in medical image processing, we adopt a U-shaped architecture as the skeleton of our network for cloud detection of remote sensing images. The skeleton extracts features through a encoder module and up-samples the featuremaps through a decoder module. For each feature layer at the same scale in the encoder and decoder module, a skip connection is applied to fuse features as similar with the U-Net. Different from the original U-Net, the feature extractors of other widely used classification networks can be adopted as encoder in the proposed architecture. To upsample featuremaps back to the same size as the input image, the decoder module either applies deconvolution or bilinear upsampling. To trade off between model size and segmentation accuracy, lightweight networks, such as VGG (Simonyan, Zisserman, 2014), can be used as the backbone of encoder module. Bilinear upsampling is often preferred when building decoder module because of parameter free.

3.2 Multi-class classification

As mentioned, we add a multi-class classification branch to the semantic segmentation skeleton. Taking the last feature layer of the encoder module as input, the added branch firstly squeezes the multi-channel featuremap to channel-wise factors by applying global average pooling. Details about the squeeze will be described later. These factors are channel-wise statistics and have two meaningful usage. On one hand, these factors can be rescaled to combine with feature layers in the decoder module, which can recalibrate featuremaps for better segmentation. On the other hand, a fully connected layer with Sigmoid function is built on top of these factors to predict what kind of classes are presented in the input scene. The motivation is straightforward. For a subimage cropped from one large original remote sensing image, if we can identify what categories presented in the input scene, it should help to prevent the network from falsely detecting nonexistent objects. When training with the added branch, multi-class labels of the training samples can be directly generated from segmentation labels by checking what categories exist in those samples. Finally, binary cross entropy loss can be utilized to regularize this auxiliary branch.

3.3 Global context aided semantic segmentation

3.3.1 Global context pooling.

In fact, global information has been widely used in semantic segmentation networks these years. Most of these methods extract global features at the deep layer of networks by global average pooling. Then the global features are fused with other featuremaps before being fed into the following layers (Chen et al., 2017, Zhao et al., 2017). However, these global features have no explicit contextual meanings due to the lack of additional supervised information on the input scene. To cover this shortage, (Zhang et al., 2018) designed a context encoding network (EncNet) to capture global contextual information by identifying what classes in the input scene, thus improved semantic segmentation performance.

Based on framework of the EncNet, the global contextual information can be obtained in a simpler way. To help understanding, we briefly introduce the key context encoding (CE) module in EncNet. The CE gathers dictionary learning and residual encoding together to generate robust representations for the input image. Given an input featuremap with the shape of $C \times H \times W$, The CE considers it as a set of $C-$dimensional input features $X = \{x_1, ..., x_N\}$, where N is equal to $H \times W$, which learns a dictionary $D = \{d_1, ..., d_K\}$ containing K codewords of $C-$dimensional and a set of smoothing factors $S = \{s_1, ..., s_K\}$. Then a fixed length representation for the input featuremap, denoted as $E = \{e_1, ..., e_K\}$ and e_i is $C-$dimensional, can be generated. Finally, the encoded representation E was aggregated by $e = \Sigma_{k=1}^{K} \phi(e_k)$, where ϕ denotes Batch Normalization with ReLU activation. The CE was modified based on the module originally proposed in (Zhang et al., 2017) for texture recognition which used all entries of the encoded representations for classification. However, the CE reduces to a pooling layer after the encoded representation being aggregated to a linear representation so that it may be reasonable to simplify such a complex module. In this work, we directly apply global average pooling on deep features of the encoder to replace CE, which we name it context pooling (CP). For one thing, a pooling layer is parameter free compared to the CE module. For another thing, output of the global average pooling has direct statistic relationship with the input features.

Similar to CE module, a linear representation of the featuremap output by the encoder can be generated through CP, which will be used in two ways. On one hand, followed by a fully connected layer, the linear representation can be further transferred to channel-wise featuremap scaling factors through $\gamma = \delta(We)$, where W denotes weights of the fully connected layer and δ is the sigmoid function. The scaling factors will be combined with featuremaps of the decoder by channel-wise multiplication, aiming to selectively strengthen or weaken certain features. On the other hand, additional fully connected layer with sigmoid function is built on top of this linear representation

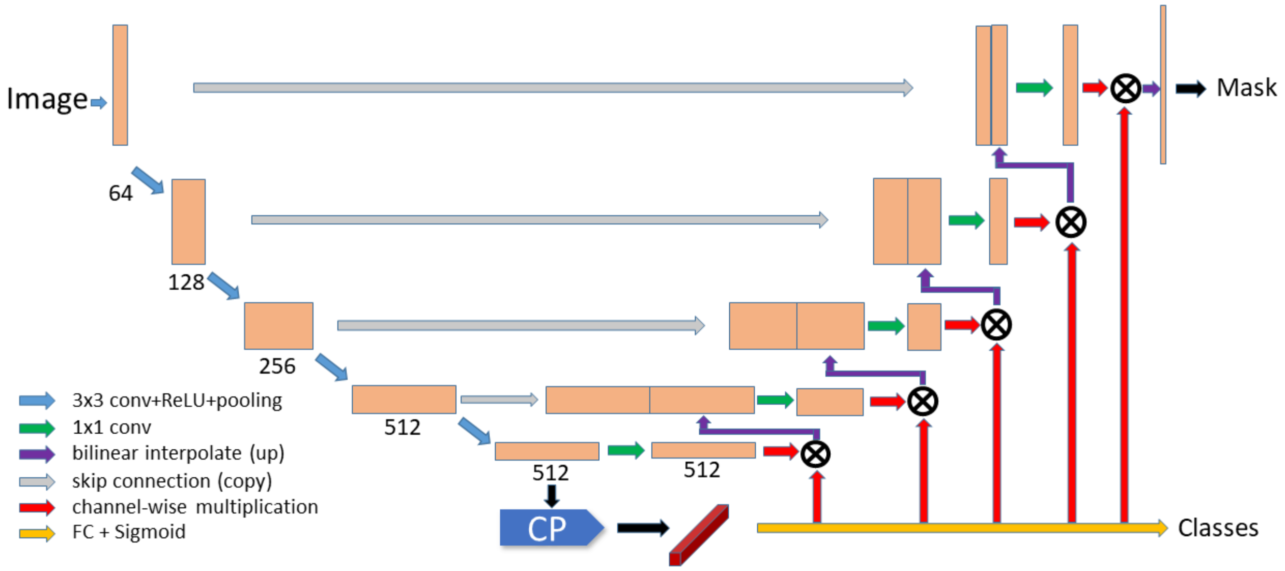

Figure 3. Architecture of the proposed GCANet. The skeleton is a U-shaped architecture similar to the UNet. A multi-class classification branch is added on top of the encoder. The new branch learns linear representation of the input featuremap. Through fully connected layers, the linear representation is, on one way, used to identify classes presented in the input scenes, and is combined with featuremaps of the encoder by channel-wise multiplication on the other way.

to predict categories presented in the input scene, which acts as auxiliary supervised contextual information to aid semantic segmentation.

3.3.2 Featuremap recalibration. Through CE module in the EncNet, the featuremap was adaptively recalculated before upsampling by conducting channel-wise multiplication with scaling factors. Similarly, (Hu et al., 2018) proposed a "Squeeze-and-Excitation" (SE) block as a plug and play module to adaptively recalibrate channel-wise feature response. The SE block firstly squeezed the featuremap to a linear representation through global average pooling. Then, two fully connected layers were built on top of it to learn a nonlinear transformation that models interaction between channels. Finally, the SE block output recalibrated featuremap after performing channel-wise multiplication with the transformed linear representation. Intuitively, the CE module and SE block act as a featuremap recalibration module which selectively emphasize or de-emphasize certain feature channels to improve feature representation ability.

Through CP module, a linear representation encoded global contextual information can be learned, then it should be transferred to featuremap scaling factors and applied to rescale featuremaps similar to CE and SE. In this paper, we adopt such featuremap recalibration strategy in the proposed U-shaped architecture. Since the decoder up-samples featuremaps at different scales step by step, we introduce multiple groups of channel-wise scaling factors for featuremaps of the decoder at each scale. Each group scaling factors is generated from the same linear representation through one fully connected layer to match the channel dimension of each featuremap of the decoder. Different from recalibrating featuremap only once in CE, we obtain several groups of scaling factors and recalibrate all the featuremaps of the decoder at different scales.

3.3.3 Network architecture. To combine global context pooling and featuremap recalibration together, we propose a Global Context Aided semantic segmentation Network (GCANet). The architecture of GCANet is shown in Fig 3. The skeleton is a U-shaped structure which is consisted of an encoder and a decoder. The encoder extracts multiscale features of the input image while the decoder up-samples the featuremaps back to the input scale. Similar to other U-shape networks, a skip connection is built between the encoder and the decoder to fuse features at the same scale. To trade-off between model size and accuracy, we utilize VGG16 as the backbone of the encoder. For the decoder, bilinear upsampling is used to replace deconvolution operators in the original U-Net.

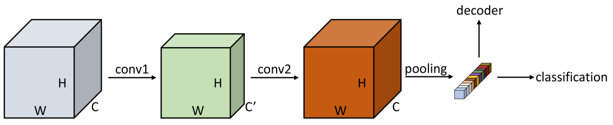

Figure 4. Details of the CP module in GCANet. Two convolution layers are built to increase nonlinearity before global average pooling. The same strategy is applied in EncNet

Apart from the skeleton, a multi-class classification branch is added as in Fig 3. The branch Firstly squeezes the featuremap output from the encoder into a linear representation through the CP module. Architecture of the CP module is shown in Fig 4. Similar to the EncNet, we build two convolution layers before the global average pooling layer. After that, on one way, the linear representation is transformed by five different fully connected layers with sigmoid activation function, which generates five groups of channel-wise scaling factors with different length. Each group of scaling factors is corresponding to a feature layer of the decoder at certain scale. Then channel-wise multiplication between featuremaps at five different scales of the encoder and their related scaling factors are conducted to selectively strengthen and weaken certain features. On the other way, the linear representation is fed into another fully connected layer followed by sigmoid activation function to predict the categories presented in the input scene. There are two outputs of the GCANet. The semantic segmentation skeleton is regularized by cross entropy loss (CE loss) and the multi-class classification branch is regularized by binary cross entropy loss (BCE loss). During training, the final loss (f_loss) of the GCANet is calculated by weighted summing the segmentation loss (S_loss) and the classification loss (C_loss):

$$f\_loss = S\_loss + \lambda C\_loss \tag{1}$$

4. EXPERIMENTS

4.1 Datasets

To test the performance of the proposed framework, two Landsat-8 OLI (Operational Land Imager) datasets are used in our experiments. One is the Landsat Boime dateset (Foga et al., 2017) which contains 96 Landsat-8 Level-1 products, and another is the Spatial Procedures for Automated Removal of Cloud and Shadow (SPARCS) dataset consisted of 80 well labeled Landsat-8 scenes with 1000 x 1000 pixels. Since the purpose is to detect both cloud and cloud shadow in Landsat images, while only 32 products of the Biome have cloud shadow labeled, we use this subset of Biome dataset and the whole SPARCS dataset in our experiments. The original labels are rearranged to have only three classes that represent cloud shadow, clear and cloud respectively. To normalize pixel value into 0 $\sim$ 1, all images are converted to TOA reflectance products. Only four bands (blue, green, red and NIR) are used in order to make our framework more adaptive to regular remote sensing images with limited bands. We construct two groups of experiments to demonstrate the effectiveness of the proposed method. For the first group, only Biome subset is used, of which 24 Level-1 products are for training and the rest are for testing. For the second group, all images of the Biome subset are used for training while the SPARCS scenes are used for test. During training, all Landsat images are cropped into blocks with the size of 256 x 256 pixels. When testing, inference is conducted on partly overlapped sliding windows with the size of 256 x 256 pixels of the input images.

4.2 Implementation details

The proposed framework is implemented in PyTorch. For the skeleton architecture, we use pre-trained VGG16 with Batch Normalization (BN) as the backbone of our encoder module. The decoder is made up of five bilinear up-sampling layers. Similar to UNet, featuremaps of encoder and decoder at the same scale are concatenated through skip connection and remapped by depth convolution before up-sampling as shown in Fig 3. The network is learned from the weighted sum of segmentation loss and classification loss. When training, Adam optimizer with base learning rate 0.0001 is used. The momentum is set to 0.9 and the weight decay is set to 0.0001. The networks are trained for 100 epochs on the two groups of experiments. The weight of classification loss is set to 0.4. Data augmentation, including random blur, gaussian noise, flip and rotation, is applied during training.

To demonstrate the effectiveness of the proposed network, one rule-based method (Fmask) and several deep learning methods (UNet, FCN (Long et al., 2015), Deeplabv3 (Chen et al., 2017)) are compared. We use the latest Fmask4.0 to detect cloud and cloud shadow with recommended parameters. The UNet, acts as the baseline of our method, has exactly the same structure as the skeleton of our network, and the other two networks are implemented on the basis of PytorchEncoding (Zhang et al., 2018). For quantitative comparison, we adopt typical metrics used in semantic segmentation tasks for evaluation, which are pixel accuracy (pixAcc) and mean Intersection of Union (mIoU).

4.3 Experimental results

Two groups of experiments were conducted to demonstrate the effectiveness of the proposed framework. For the first group of experiments, we were aiming at verifying the superior performance of U-shaped architecture compared to FCN-like architecture for cloud detection task. In computer vision, state-of-the-art semantic segmentation methods all utilize FCN-like architecture. These networks adopt ResNet as backbone, which often have bigger model size and achieve better accuracy than U-shaped networks. The reason may lie in that enough data is available for training and testing. However, this is not the case in remote sensing.

The results are shown in table 1, Fmask achieved much worse results than deep learning methods on the Biome subset. The Fmask indeed can detect cloud well in some cases, it is not robust enough due to complicated image scenes and illuminance condition. Among deep learning methods, networks with U-shaped architecture performed better than FCN-like ones. On one hand, the image number and variety of our dataset is much smaller than that of computer vision datasets such as ImageNet (Deng et al., 2009). On the other hand, the segmented classes of cloud detection task are much less than that in computer vision tasks. FCN-like networks are too complex for cloud detection in remote sensing because of their high probability of overfitting. As a result, they are prone to converge to a inferior results. On the contrary, U-shaped networks are more suitable for cloud detection task due to their moderate model size. At last, as can be seen in table 1, the proposed framework shows better performance than UNet because of the added classification branch.

Table 1. Results of the first group of experiments. The 2-4 columns denote the IoU of each class respectively.

Method	shadow	clear	cloud	mIoU	picAcc
Fmask	0.1894	0.8011	0.5788	0.5231	0.8234
UNet	0.5779	0.8750	0.6906	0.7145	0.8991
FCN	0.5162	0.8665	0.6852	0.6893	0.8857
Deeplab	0.4980	0.8562	0.6258	0.6600	0.8733
GCANet	**0.6003**	**0.8929**	**0.7205**	**0.7379**	**0.9116**

To test generality of the proposed framework, we conducted the second group of experiments, which trained the networks on the whole Biome subset and tested on the SPARCS. The 80 scenes of the SPARCS are cropped from 80 different original Landsat images distributed at all over the world. Images in the SPARCS are of better variety and labeling quality than those in the Biome. The results are shown in table 2. The Fmask achieved the best IoU of cloud but it's overall accuracy and mIoU were still worse than deep learning methods. Different from deep learning methods that used only 4 bands of Landsat-8 images, the Fmask used all bands to identify pixel classes. As we found in our experiments, even though the Fmask could obtain acceptable detection results of cloud, it was often unable to detect cloud shadow well. Among deep learning methods, the U-shaped networks still performed better than the FCN-like ones, and the proposed framework outperformed the UNet.

To intuitively show the improvement of our method compared to other methods, we visualize some cloud detection results of

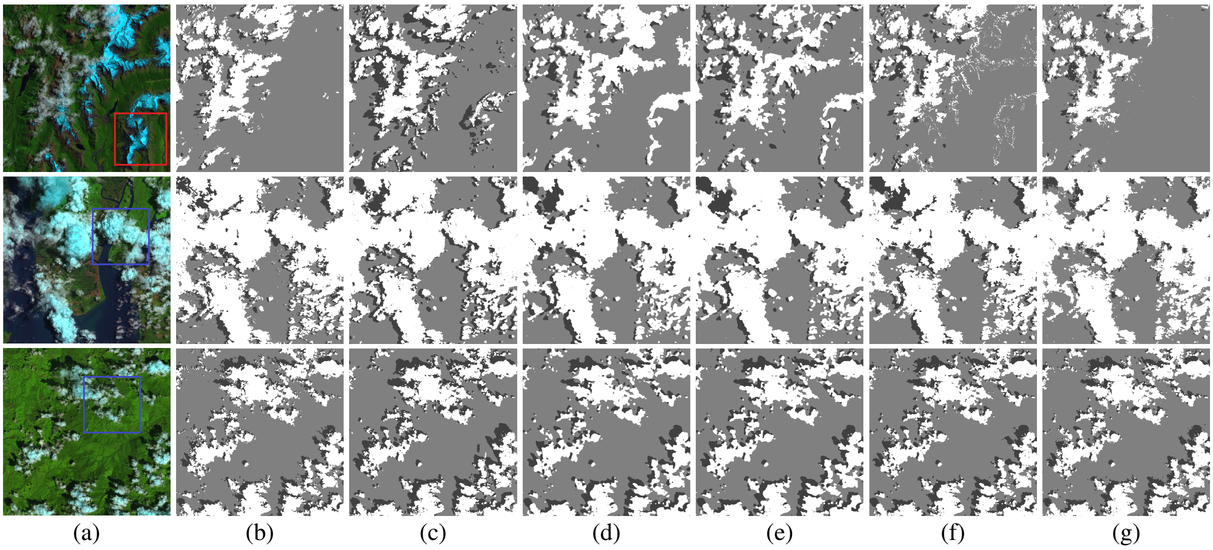

Figure 5. Three cloud detection examples of the SPARCS. Column (a) is the original Landsat-8 images; column (b) is the ground-truth labels; column (c)(d)(e)(f)(g) are the results of Fmask, FCN, Deeplab, UNet and the proposed GCANet respectively.

Table 2. Results of the second group of experiments. The 2-4 columns denote the IoU of each class respectively.

Method	shadow	clear	cloud	mIoU	picAcc
Fmask	0.3626	0.8678	**0.7535**	0.6613	0.8822
UNet	0.4457	0.8812	0.7144	0.6804	0.8977
FCN	0.4014	0.8739	0.6822	0.6525	0.8894
Deeplab	0.4207	0.8777	0.6718	0.6567	0.8891
GCANet	**0.4625**	**0.8978**	0.7372	**0.6992**	**0.9082**

the SPARCS as shown in Fig 5. Thanks to global contextual information and featuremap recalibration, GCANet has lower probability of falsely detecting non-cloud objects as cloud. As can be seen in the first row of Fig 5, all other methods except GCANet have falsely detected red box region which is covered with snow as cloud region. Snow has always been a challenge to cloud detection because of its similarity to cloud. Even though Fmask uses more bands than GCANet, it is still prone to misclassify snow pixels. Apart from higher detection accuracy, results of U-shaped networks are finer than those of FCN-like networks. Further, compared to the baseline network, the GCANet outperforms the UNet in both accuracy and fineness. We visualize two zoom views marked as blue boxes in Fig 5. As shown in Fig 6, GCANet obtains slightly finer results than UNet. Above all, both quantitative and qualitative result outperforms the compared methods, demonstrating the effectiveness of proposed method.

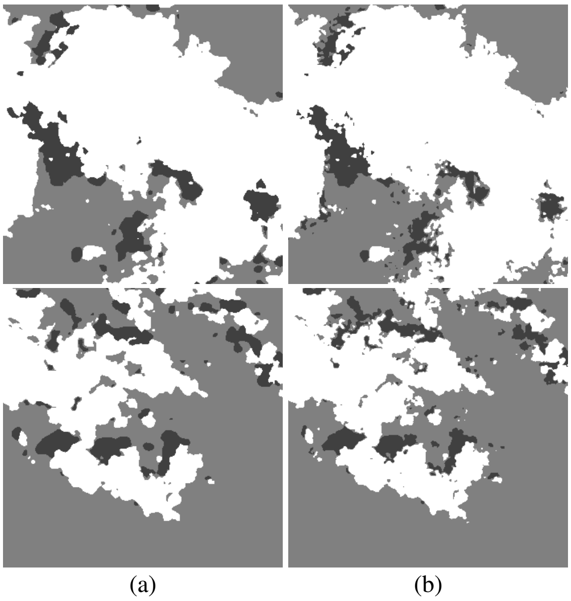

Figure 6. Zoom views of two sites in cloud detection results of the SPARCS. Column (a) is the results of UNet; Column (b) is the result of GCANet.

5. CONCLUSION

In order to capture global contextual information of remote sensing sub-scenes, a global context aided semantic segmentation network named as GCANet is proposed for cloud and cloud shadow detection in this paper. we add a multi-class classification branch to a U-shaped network with encoder and decoder structure. The added branch is built on top of the featuremap output from the encoder, which is aimed at capture contextual information by identifying what categories exist in the input scene. The linear representation learned from this branch is combined with featuremaps of the decoder at all scales to selectively strengthen class related features or weaken class unrelated features. By explicitly employing global context of the input scene in such a supervised way, the proposed network can achieve better results than its counterpart without the classification branch and other deep learning methods. Besides, the multi-class labels of training samples can be directly obtained from semantic segmentation labels, which is easy to fulfill. The experiments have demonstrated the effectiveness of the proposed method.

Though the introduction of global contextual information is intuitive, the improvement on quantitative results is not significant enough compared to the existed U-shaped network as shown in our experiments. Future works may lie in two folds. For one thing, the diversity and labeling quality of image data need to be improved. For another, more datasets including different remote sensing images, such as sentinel-2 or other high resolution images, need to be investigated and tested in order to verity the effectiveness and robustness of the proposed framework.

ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB050-5003 and National Natural Science Foundation under Grant 413-22010.

REFERENCES

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12), 2481–2495.

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 sensing of environment*, 225, 307–316.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 248–255.

Foga, S., Scaramuzza, P. L., Guo, S., Zhu, Z., Dilley Jr, 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 sensing of environment*, 194, 379–390.

Hollingsworth, B. V., Chen, L., Reichenbach, S. E., Irish, R. R., 1996. Automated cloud cover assessment for landsat tm images. *Imaging Spectrometry II*, 2819, International Society for Optics and Photonics, 170–179.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

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

Jeppesen, J. H., Jacobsen, R. H., Inceoglu, F., Toftegaard, T. S., 2019. A cloud detection algorithm for satellite imagery based on deep learning. *Remote Sensing of Environment*, 229, 247–259.

Li, P., Dong, L., Xiao, H., Xu, M., 2015. A cloud image detection method based on SVM vector machine. *Neurocomputing*, 169, 34–42.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

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 Sensing of Environment*, 231, 111205.

Richter, R., Louis, J., Müller-Wilm, U., 2012. Sentinel-2 msi–level 2a products algorithm theoretical basis document. *European Space Agency,(Special Publication) ESA SP*, 49(0), 1–72.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Scaramuzza, P. L., Bouchard, M. A., Dwyer, J. L., 2011. Development of the Landsat data continuity mission cloud-cover assessment algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 50(4), 1140–1154.

Shao, Z., Pan, Y., Diao, C., Cai, J., 2019. Cloud Detection in Remote Sensing Images Based on Multiscale Features-Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6), 4062–4076.

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 Journal of Photogrammetry and Remote Sensing*, 157, 124–136.

Shi, M., Xie, F., Zi, Y., Yin, J., 2016. Cloud detection of remote sensing images by deep learning. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 701–704.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tan, K., Zhang, Y., Tong, X., 2016. Cloud extraction from chinese high resolution satellite imagery by probabilistic latent semantic analysis and object-based machine learning. *Remote Sensing*, 8(11), 963.

Zhai, H., Zhang, H., Zhang, L., Li, P., 2018. Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery. *ISPRS journal of photogrammetry and remote sensing*, 144, 235–253.

Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A., 2018. Context encoding for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7151–7160.

Zhang, H., Xue, J., Dana, K., 2017. Deep ten: Texture encoding network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 708–717.

Zhang, Y., Wen, F., Gao, Z., Ling, X., 2019. A Coarse-to-Fine Framework for Cloud Removal in Remote Sensing Image Sequence. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8), 5963–5974.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

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 Sensing of Environment*, 159, 269–277.

Zhu, Z., Woodcock, C. E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote sensing of environment*, 118, 83–94.