




Cloud Detection from Remote Sensing Images by Cascaded U-shape Attention Networks

Ao Li¹, Jing Yang², and Xinghua Li¹(✉) 

¹ School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
lixinghua5540@whu.edu.cn

² CCCC Second Highway Consultants Co, Ltd, Wuhan, China

Abstract. Cloud is an important meteorological information in remote sensing applications as it plays a significant role in the Earth's climate and weather patterns, but it also brings difficulties to the information extraction from optical images, especially when the underlying surface features to be analyzed are obscured. Therefore, cloud detection is an indispensable step in optical remote sensing image processing. Different from low-spatial resolution images, medium and high-resolution images contain richer geographical features, and the distribution of clouds is more scattered, which makes it necessary to enhance the network's ability on detailed features extraction. Therefore, the two cascaded U-shape attention networks (CUA-Net) model is proposed to detect the cloud in Landsat 8 images. In the first U-shape network, the up-sampling layers in path expansion integrate the information from all previous layers to make full use of multi-scale features. Additionally, the attention modules in the skip connection are added to detect the position and edges of cloud accurately. After that, the second U-shape network is utilized to optimize the preliminary segmentations from the first network, thus obtaining results closer to the ground truth. In the experiments, CUA-Net was evaluated on 38-Cloud Dataset and compared with current mainstream networks, showing significant improvements both on visual effects and quantitative indicators.

Keywords: Cloud detection · Cascaded U-shape networks · Attention module

1 Introduction

Remote sensing images play a vital role in natural disaster detection, agricultural resources management, environmental monitoring, urbanization surveys and other research fields. However, a factor that cannot be ignored in optical satellite images is cloud cover. Cloud can interfere with the remote sensing data by reflecting and absorbing the electromagnetic radiation, which leads to difficulties in data interpretation. Consequently, it is a crucial part of remote sensing field to accurately identify the cloud coverage over images for subsequent applications [1].

Cloud detection methods can be roughly grouped into classical methods and pattern recognition methods [2]. The threshold-based methods are the earliest classical methods.

They mainly analyze individual pixels, such as the automatic cloud coverage evaluation [3] and Fmask [4], and they can segment cloud from images by multiple fixed thresholds. Especially, Sun et al. [5] proposed a general dynamic threshold cloud detection algorithm to solve the difficulty in fixed thresholds selection. Since those threshold-based methods are easily restricted by the spectrum, the Bayesian methods [6] and texture based methods [7] utilizing the spectral and geometric properties of cloud are proposed to leverage more features. Moreover, some methods based on statistical characteristics [8] are proposed for thin cloud detection. They mainly take advantage of the physical properties of clouds, so the results can be obtained quickly with the high-level characteristics of images ignored, which leads to detection difficulties when facing complex surface environments and ever-changing clouds.

With the development of computer hardware, pattern recognition technology has attracted the attentions. Many advanced machine learning methods to identify cloud are proposed. Among them, the early clustering [9], fuzzy clustering [10, 11] and SVM [12–14] have formed a mature system, however, the detection accuracy is relatively limited by their poor performance in large-scale training set. In recent years, artificial neural networks have emerged as a promising approach for cloud detection due to their ability to learn complex patterns and feature representations from multitudinous labeled training data. For example, the U-net [15, 16] uses a completely symmetrical network structure and skip connections to improve the accuracy of cloud detection with fewer training samples. MS-UNet [17] combines convolutions of different sizes to extract multi-scale features, thus identifying cloud of different sizes and shapes. Cloud-Net [18] proposed by Mohajerani et al. adds the residual structure to U-Net, and achieves superior results for Landsat 8 images. As time goes on, more advanced networks are proposed, Unet 3 + [19] uses full-scale skip connection to preserve spatial information and fuse features at different layers. Li et al. proposed global context-dense block U-Net (GCDB-UNet) [20] to enhance the detection capability of thin cloud. Lu et al. designed a mutual guidance module (MGM) [21] to solve the problem of rough segmentation boundaries. Although these methods have been able to detect most of cloud on remote sensing images, the thin cloud recognition and boundary identification capabilities still need to be further strengthened especially for medium and high-resolution images such as Landsat 8.

In order to better capture the complex semantic features and precisely segment the cloud in remote sensing images, the two cascaded U-shape attention networks (CUA-Net) model is proposed. Its innovations are as follows, (1) it enhances the connection between the network layers to preserve as much information as possible, (2) it makes use of the attention module to focus on relevant cloud features and to ignore irrelevant ones, which can improve the network's ability of identifying clouds in complex scenes with varying cloud and background noise, (3) a second U-shape network is designed to correct the inaccurate information gain from the previous steps. Via these structures, the features extracted from convolution blocks can be utilized effectively to recover sophisticated cloud masks and obtain higher accuracy.

2 Algorithm

The architecture is designed as two cascaded U-shape networks, as shown in Fig. 1. The first network is used to perform a preliminary segmentation by identifying the possible cloudy regions of the image. The output of the first network X_{En}^1 is then fed into the second network, which refines the edges and details by further segmenting the cloudy regions and removing false detections. After that, the preliminary results X_{En}^1 and the supplementary information X_{De}^1 are added and convolved once to obtain the final cloud detection results. The proposed CUA-Net will be introduced separately below.

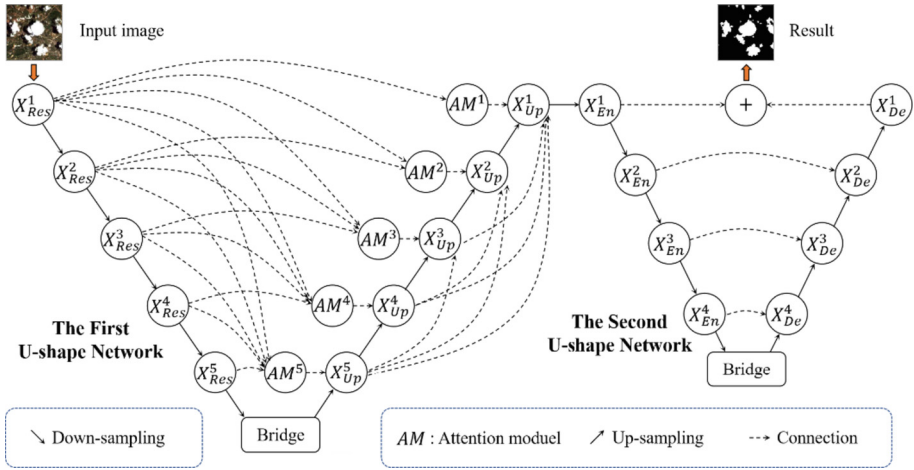


Fig. 1. The proposed Cascaded U-shape Attention Networks (CUA-Net).

2.1 The First U-shape Network

The first U-shape network consists of a contraction path for feature extraction and an expansion path for image recovery. The two parts are connected by the attention-based skip connection, which is used for transferring deep features from the contraction path to the expansion path to preserve spatial information.

Down-sampling Layer in Contraction Path. The down-sampling layer mainly uses residual structure shown in Fig. 2. Its branches on the above include two 3×3 convolutions to extract features from the input. The branches below use a small-scale skip connection, where the input firstly go through a 1×1 convolution, and then connected with itself. Finally, the results of the two branches are summed and put to a maximum pooling. This structure can avoid the gradient disappearance caused by the deep network, and make the encoder converge faster. Simultaneously, it allows the network to learn the residual mapping between the input and output feature maps, which helps to preserve the low-level features from upper layer.

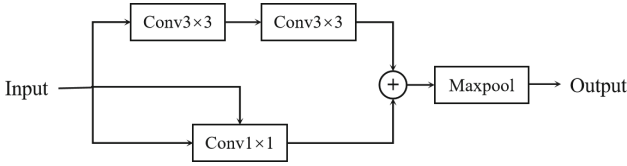


Fig. 2. Down-sampling layer in contraction path.

Attention-Based Skip Connection. In U-Net, the skip connections are used to preserve the features learned from the contraction path and improve the accuracy of segmentation. However, only layers with same depth are connected in the original U-Net architecture. To address this limitation, a modified skip connection shown in Fig. 3 is proposed, the features from all previous layers in the contracting path are concatenated and sent to the expansion path. In order to make the output from layer $X_{Res}^1, \dots, X_{Res}^{i-1}, X_{Res}^i$ able to be connected, multiple self-connections are used to make the dimension of $X_{Res}^1, \dots, X_{Res}^{i-1}$ as same as X_{Res}^i , and then feature graph size is unified by maximum pooling. After that, all the i layers are added and input to the subsequent attention module. This modified skip connection allows the network to capture more fine-grained details and improve cloud detection accuracy.

Convolutional block attention module (CBAM) [22] is a lightweight attention architecture composed of channel attention module (CAM) and spatial attention module (SAM). CAM focuses more on the category information. The input image will go through parallel MaxPool layer and AvgPool layer at first, and then pass by a single shared MLP to extract more comprehensive high-level features. SAM pays more attention on the spatial location of the target. It applies the average pooling and the maximum pooling along channel axis, which can effectively strengthen the spatial information.

The attention-based skip connection can preserve features extracted from all layers in contraction path and pay effective attention on the channel and spatial characteristics of the target. What's more, the number of parameters in this structural is small, which will not bring additional burden to the network.

Up-sampling Layer in Expansion Path. The up-sampling layer in the expansion path is used to increase the resolution of feature maps while reducing the number of channels, as shown in Fig. 4. The input X_{Up}^{i+1} is firstly up-sampled by a deconvolution, then combined with AM^i from corresponding skip connection and $X_{Up}^{i+2}, X_{Up}^{i+2}, \dots, X_{Up}^5$ from the lower up-sampling layers. By this way, not only the feature maps in contraction path are used, the maps in the layers in front of expansion path are also used. Their combination will go through two convolutions to recover the semantic details and be added to the deconvolved X_{Up}^{i+1} . More complex and detailed cloud properties from deep feature maps can be recovered due to the full use of multi-scale information.

2.2 The Second U-shape Network

The second U-Shape network is mainly utilized to refine the segmentation mask generated by the first network. Although most of the cloud information can be extracted after

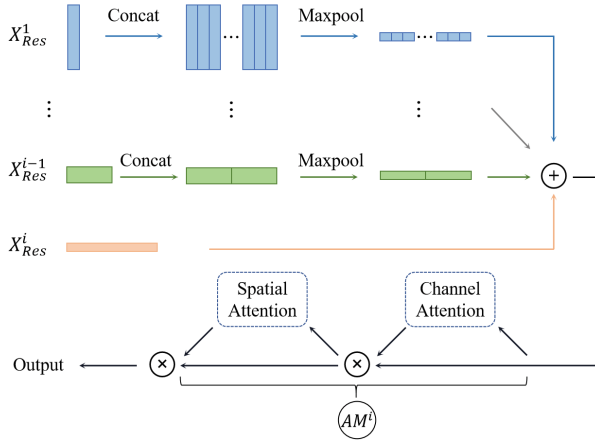


Fig. 3. Attention-based skip connection.

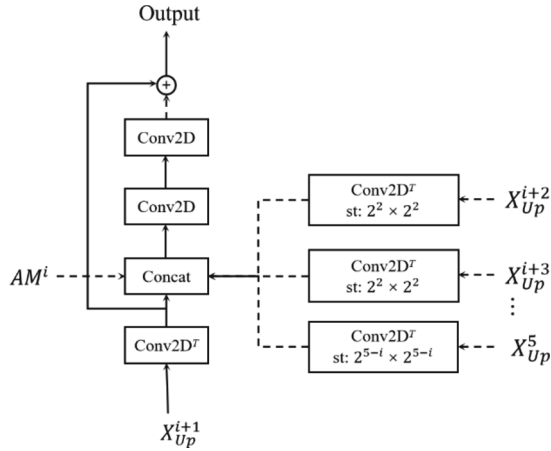


Fig. 4. Up-sampling layer in expansion path.

the anterior training, thin cloud and fragmentary cloud are easily failed to be detected, and some highlight surfaces can be mistaken as cloud. Therefore, the second U-shape network is designed to revise these incorrect detections. It consists of an encoder-decoder structure with skip connections between them, similar to a four-layer U-net. The difference is that the bridge layer in the middle takes advantage of dropout function to prevent the model from overfitting. No extra structures are added to the second network due to its complementary role and the expectation of lower network complexity.

2.3 Activation Function and Loss Function

ReLU is used as the activation function except from the last layers of the two U-shape networks and the attention module which has certain definition. It is a piecewise linear

function that produces an output of zero for negative inputs and a linear output for positive inputs. By introducing non-linearity, ReLU can avoid the network from gradient disappearance and overfitting with small cost. Sigmoid is used as the activation function after X_{Up}^1 and X_{De}^1 to map the output value between 0 and 1, thus determining the probability that each pixel is cloudy.

Denote the true value as t , the predicted value as p , and the total number of pixels as N , the loss function used can be denoted as Eq. (1).

$$Loss(t, p) = 1 - \frac{(1 + \beta^2) \times \sum_{i=1}^N t(i)p(i) + \epsilon}{\sum_{i=1}^N t(i) + \beta^2 \times \sum_{i=1}^N t(i)p(i) + \epsilon} \quad (1)$$

where i means the i th pixel in the image, β is a constant which controls the weight of recall relative to precision. In the experiments, β is taken as 2 to give more weight to recall, making it more suitable for cloud detection datasets where the positive class is smaller than the negative class. ϵ is assigned as 10^{-7} to avoid any division by zero.

3 Data and Experiments

3.1 Data and Environment

The experimental data set is 38-Cloud Dataset [18] made by Sorour Mohajerani, including 18 scenes for training and 20 scenes for testing, and each scene is cut to 384×384 patches. The source of the dataset is Landsat 8 images with the resolution of 30 m, and their red, green, blue and near-infrared bands are chosen for cloud detection.

The experiments were performed on a Linux system with Python 3.6, configured with GPU versions of Tensorflow 1.12.0, Keras 2.2.4 and skimage 0.15.0. A Quadro RTX 5000 graphics card was used as the driver for training and prediction. The Adam optimizer with an initial learning rate of 1×10^{-4} was used during training, and when the learning rate was reduced to 1×10^{-8} , the training was finished.

3.2 Experiments Results

In order to verify the ability of the proposed CUA-Net, the comparison experiments and ablation experiments were conducted. The comparison experiments involve the performance of CUA-Net with state-of-the-art networks. On the other hand, the ablation experiments were conducted to evaluate the effectiveness of the second U-shape network and CBAM in skip connections.

Comparison Experiments. U-net [16], MS-UNet [17], Cloud-Net [18] and Unet 3 + [19] are selected for comparison, and the experimental results are shown in Fig. 5, where the black and white refers to the correctly identified clear and cloudy area, respectively, while the red means it is cloudy but falsely detected as clear, and the blue means it is clear but falsely detected as a cloudy area.

The visual effects of cloud detection from whole scene image by different methods are shown in Fig. 5(a). It can be seen that these methods can detect majority of

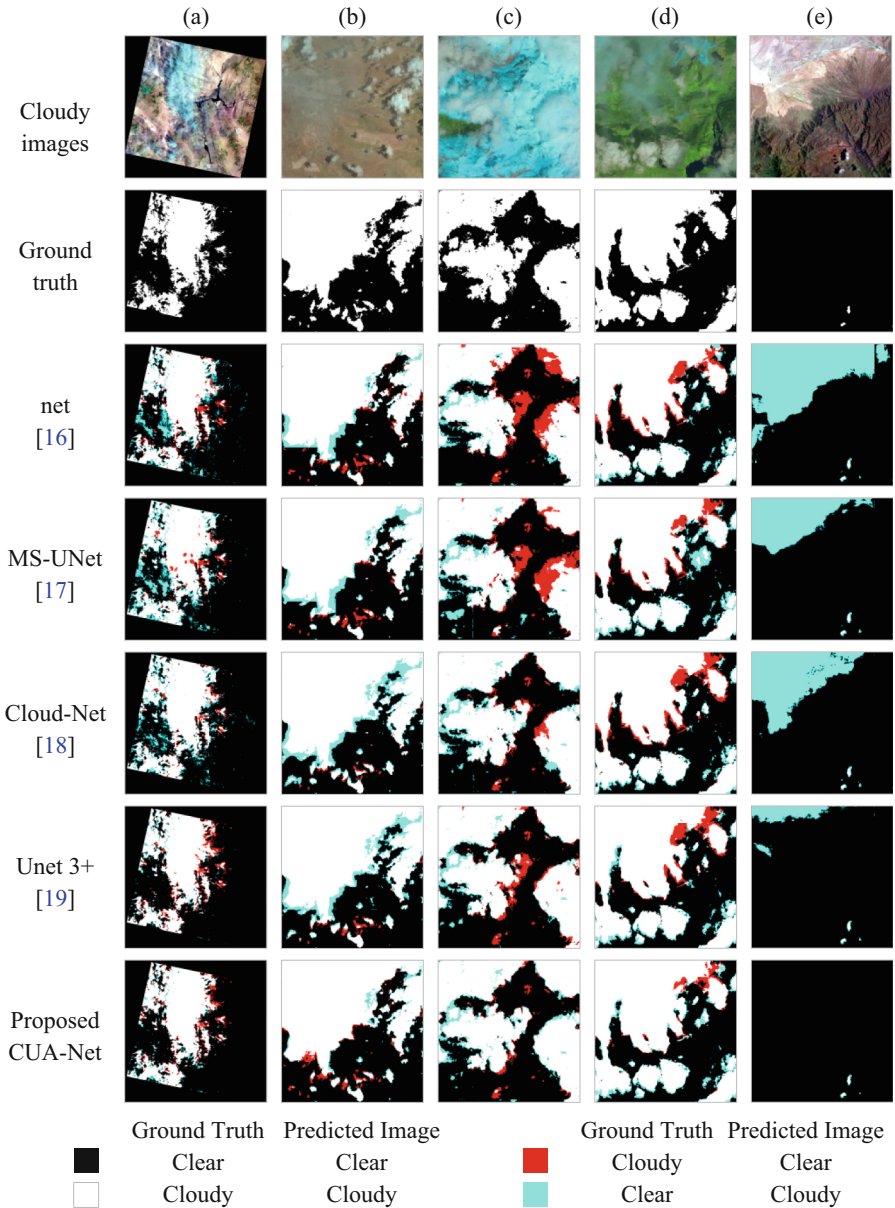


Fig. 5. Visual results of cloud detection in comparison experiments.

cloudy area, but U-net, MS-UNet and Cloud-Net have more mistakes, especially for the highlighted regions in lower right corner. Although Unet 3 + can achieve better results, the performance on boundaries is still worse and the missing cloud information is more compared with CUA-Net. Figure 5(b)–Fig. 5(e) is the visual effect of local details, representing four different types of landcovers: bare land, ice land, vegetation and mountains.

Results indicate that CUA-Net can achieve better visual effect with less confusion and more clear boundaries under different surface conditions. For example, in Fig. 5(b) and Fig. 5(d) covering both thin and thick cloud, all methods can accurately detect the main cloud, but for edges and details, the results gained from CUA-Net is most consistent with the ground truth. As for Fig. 5(c) covered with ice and snow, U-net and MS-UNet have many omissions on the boundary, Cloud-Net and Unet 3 + perform better but the capability of detail extraction still need to be strengthened, while the CUA-Net can accurately distinguish between ice land and cloud due to its advantageous structures. For highlighted ground shown in the above of Fig. 5(e), all the other four methods detect it as cloud more or less except CUA-Net. Through the visual interpretation, it can be confirmed that CUA-Net can achieve more detailed edges and superior cloud detection results than other methods.

To evaluate the cloud detection accuracy more objectively, Precision, Recall, Specificity, Intersection over Union (IoU), Overall Accuracy (OA) and F1 score are selected for quantitative evaluation. High precision indicates that the detected cloud is generally true, while high recall means that the model can detect most cloud. Specificity is used to measure the negative predictions, IoU to measure the overlap between the predicted result and ground truth, and OA for the correctly classified instances. F1 score is the harmonic mean of precision and recall to measure their balance. They are defined as Eqs. (2)–(7).

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (5)$$

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \quad (6)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

where TP (true positive) indicates the total amounts of correctly detected cloud pixels, TN (true negative) represents the number of correctly detected clear pixels, FP (false positive) means the amounts of clear pixels incorrectly detected as cloud pixels and the FN (false negative) on the contrary. The quantitative evaluation results are shown in Table 1.

Table 1. Accuracy evaluation results in comparison experiments (%).

Method	Precision	Recall	Specificity	IoU	OA	F1
U-net [16]	78.27	89.73	95.52	72.87	94.32	83.61
Ms-UNet [17]	78.53	89.76	95.14	71.09	94.52	83.77
Cloud-Net [18]	80.80	89.83	96.16	72.70	95.32	85.07
Unet 3 + [19]	87.33	90.68	97.52	79.77	96.13	88.97
Proposed CUA-Net	88.58	91.10	97.80	80.94	96.72	89.82

Table 1 shows that the proposed method achieves higher accuracy than the other four networks in Precision, Recall, Specificity, IoU, OA and F1, which is consistent with the judgment of visual interpretation, indicating that the proposed method performs better in most of remote sensing scenes.

Ablation Experiments. In order to verify the effect of second U-shape network (denoted as S-UNet) and CBAM in skip connections, we designed four ablation experiments: (1) only the first U-shape network used (denoted as F-UNet only), (2) the second U-shape network used without CBAM (denoted as +S-UNet), (3) the CBAM used without the second U-shape network (denoted as +CBAM), (4) both the second U-shape network and the CBAM used (CUA-Net). Their visual effect and accuracy evaluation results are shown in Fig. 6 and Table 2, respectively.

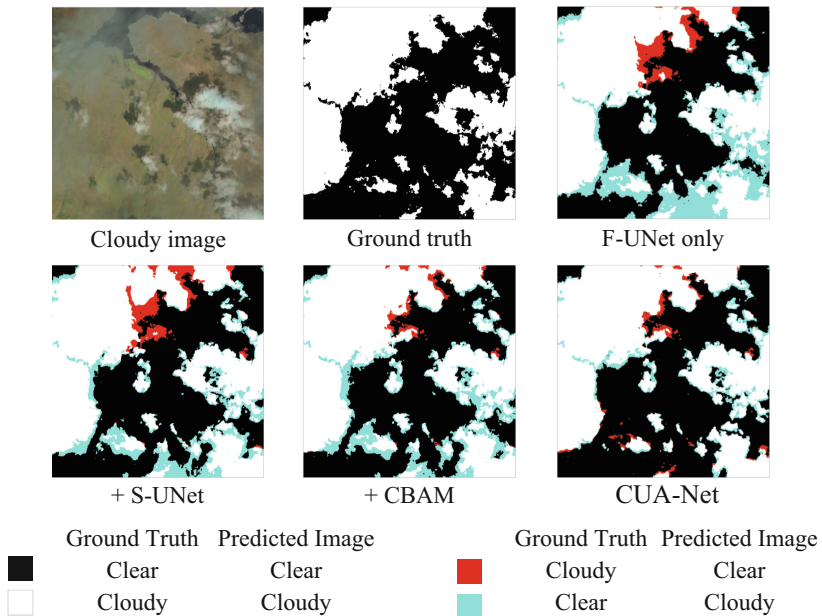
**Fig. 6.** Cloud detection visual results of ablation experiments.

Table 2. Accuracy evaluation results of ablation experiments (%).

Method	Precision	Recall	Specificity	IoU	OA	F1
F-UNet only	82.79	90.09	96.97	74.61	96.18	86.29
+S-UNet	85.87	88.74	97.57	76.47	95.85	87.28
+CBAM	84.66	92.47	96.95	78.49	96.26	88.39
CUA-Net	88.58	91.10	97.80	80.94	96.72	89.82

Comparing the results in groups F-UNet only and +S-UNet combined with +CBAM and CUA-Net, it is found that S-UNet leads to a slight decrease in Recall, but the Specificity, IoU and F1 scores are higher than the experiments without S-UNet, and the Precision is remarkably improved. The visual interpretation also shows that the addition of S-UNet can achieve results closer to the ground truth, as it can be a good complement to the edges and details for cloud. Comparing the results in groups F-UNet only and +CBAM combined with +S-UNet and CUA-Net, it can be confirmed that CBAM can focus well on the attributes and locations of cloud, which can improve the detection accuracy comprehensively, and reduce the probability of confusing cloudy and clear area. The overall results show that better cloud detection results can be achieved with both S-UNet and CBAM.

4 Conclusion

In conclusion, the proposed CUA-Net for cloud detection has shown promising results. The second U-shape network helps to supplement the details and cloud boundaries, thus obtaining more refined and truth-related results. The dense connections and the attention model help the network preserve and focus on important features and suppress irrelevant features, contributing to higher accuracy. The CUA-Net has been evaluated on 38-Cloud dataset compared with four representative networks. The results show that it performs better than other methods in terms of quantitative evaluation and visual effect. Overall, the proposed method has potential to be applied in remote sensing fields where cloud detection is essential, and further research can be conducted to optimize the model for better performance.

Acknowledgement. The authors are grateful to the reviewers for their attention and comments on our paper. This research is supported by the National Natural Science Foundation of China (NSFC) under Grant no. 42171302 and the Key R&D Program of Hubei Province, China (2021BAA185).

References

1. Long, C., Li, X., Jing, Y., Shen, H.: Bishift networks for thick cloud removal with multitemporal remote sensing images. *Int. J. Intell. Syst. Intell. Syst.* **2023**, 9953198 (2023)
2. Gupta, R., Nanda, S.J.: Cloud detection in satellite images with classical and deep neural network approach: a review. *Multimed. Tools Appl.* **81**(22), 31847–31880 (2022)
3. Irish, R.R., Barker, J.L., Goward, S.N., Arvidson, T.: Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. *Photogramm. Eng. Remote Sens.. Eng. Remote Sens.* **72**(10), 1179–1188 (2006)
4. Zhu, Z., Woodcock, C.E.: Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **118**, 83–94 (2012)
5. Sun, L., et al.: A universal dynamic threshold cloud detection algorithm (UDTCDA) supported by a prior surface reflectance database. *J. Geophys. Res. Atmospheres* **121**(12), 7172–7196 (2016)
6. Xu, L., Wong, A., Clausi, D.A.: A novel bayesian spatial-temporal random field model applied to cloud detection from remotely sensed imagery. *IEEE Trans. Geosci. Remote Sens. Geosci. Remote Sens.* **55**(9), 4913–4924 (2017)
7. Başeski, E., Cenaras, Ç.: Texture and color based cloud detection. In: 7th International Conference on Recent Advances in Space Technologies, pp. 311–315. Istanbul, Turkey (2015)
8. He, X.Y., Hu, J.B., Chen, W., Li, X.Y.: Haze removal based on advanced haze-optimized transformation (AHOT) for multispectral imagery. *Int. J. Remote Sens.* **31**(20), 5331–5348 (2010)
9. Gómez-Chova, L., et al.: Cloud detection for CHRIS/Proba hyperspectral images. In: 10th Remote Sensing of Clouds and the Atmosphere, pp. 508–519. International Society for Optics and Photonics, Bruges, Belgium (2005)
10. Surya, S., Simon, P.: Automatic cloud detection using spectral rationing and fuzzy clustering. In: 2nd International Conference on Advanced Computing, Networking and Security, pp. 90–95. Mangalore, India (2013)
11. Bo, P., Fenzhen, S., Yunshan, M.: A cloud and cloud shadow detection method based on fuzzy c-means algorithm. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **13**, 1714–1727 (2020)
12. Li, P., Dong, L., Xiao, H., Xu, M.: A cloud image detection method based on SVM vector machine. *Neurocomputing* **169**, 34–42 (2015)
13. Sui, Y., He, B., Fu, T.: Energy-based cloud detection in multispectral images based on the SVM technique. *Int. J. Remote Sens.* **40**(14), 5530–5543 (2019)
14. Latry, C., Panem, C., Dejean, P.: Cloud detection with SVM technique. In: International Geoscience and Remote Sensing Symposium, pp. 448–451. Barcelona Spain (2007)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: 18th International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241. Springer, Munich, Germany (2015)
16. Mohajerani, S., Krammer, T.A., Saeedi, P.: A cloud detection algorithm for remote sensing images using fully convolutional neural networks. In: 20th International Workshop on Multimedia Signal Processing, pp. 1–5. Vancouver, Canada (2018)
17. Kushnure, D.T., Talbar, S.N.: MS-UNet: a multi-scale UNet with feature recalibration approach for automatic liver and tumor segmentation in CT images. *Comput. Med. Imaging Graph.. Med. Imaging Graph.* **89**, 101885 (2021)
18. Mohajerani, S., Saeedi, P.: Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery. In: 39th International Geoscience and Remote Sensing Symposium, pp. 1029–1032. IEEE, Yokohama, Japan (2019)

19. Huang, H., et al.: Unet 3+: a full-scale connected unet for medical image segmentation. In: 45th International Conference on Acoustics, Speech and Signal Processing, pp. 1055–1059. Barcelona, Spain (2020)
20. Li, X., Yang, X., Li, X., Lu, S., Ye, Y., Ban, Y.: GCDB-UNet: a novel robust cloud detection approach for remote sensing images. *Knowl.-Based Syst.-Based Syst.* **238**, 107890 (2022)
21. Lu, C., Xia, M., Qian, M., Chen, B.: Dual-branch network for cloud and cloud shadow segmentation. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–12 (2022)
22. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: convolutional block attention module. In: 15th European Conference on Computer Vision, pp. 3–19. Munich, Germany (2018)