NM-LinkNet: Cloud Detection from Remote Sensing Images with Non-local Operation and Multi-scale Feature Aggregation

Yongshi Jie[1], Anzhi Yue[2,3], Naijian Wang[4], Yan Wang[4], Xuejie Xu[5], Ding Ding[6], Wei Tan[1], Hongyan He[1], and Kun Xing[1(✉)]

[1] Beijing Institute of Space Mechanics and Electricity, China Academy of Space Technology, Beijing, China
xingkunfeixiang@163.com
[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China
[3] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Xian, China
[4] Bureau of Geophysical Prospecting INC., China National Petroleum Corporation, Zhuozhou, Hebei Province, China
[5] Department of Engineering Physics, Tsinghua University, Beijing, China
[6] Powerleader Computer Systems Co., Ltd, Shenzhen, Guangdong Province, China

Abstract. Cloud detection is an important preprocessing process in remote sensing applications. Cloud detection methods have developed from traditional methods to deep learning methods which are widely used at present. However, current cloud detection methods still have limitations in the detection of thin clouds and broken clouds, mainly because the thin clouds are sparsely distributed and the size of broken clouds is relatively small. To solve the above difficult problems, this paper proposes a cloud detection network NM-LinkNet, which combines non-local operation and multi-scale feature aggregation, to improve the detection ability of thin clouds and broken clouds. Based on LinkNet50, NM-LinkNet uses non-local operation to obtain the long-distance context information of sparse distributed thin clouds to enhance the features of thin clouds. The multi-scale feature aggregation module designed in this paper is used to extract the features of clouds of different scales to improve the detection ability of small broken clouds. The experimental results on SPARCS dataset show that the IoU and F1 of NM-LinkNet proposed in this paper reach 87.50% and 93.33% respectively, which exceeds the other five mainstream deep learning methods in quantitative data and visualization results.

Keywords: cloud detection · remote sensing · deep learning · non-local operation · multi-scale feature aggregation

1 Introduction

With the rapid development of remote sensing technology, optical remote sensing image plays an important role in semantic segmentation [1], object detection [2], change detection [3] and other remote sensing application tasks. However, more than 66% of the

H. P. Urbach and H. Jiang (Eds.): ISSOIA 2022, SPPHY 295, pp. 47–55, 2023.
https://doi.org/10.1007/978-981-99-4098-1_6

earth's surface area is covered by clouds [4], which has a negative effect on remote sensing surface information extraction. Optical satellite sensors cannot penetrate the clouds during imaging, which leads to the pollution of the acquired remote sensing images by clouds, results in the loss or weakening of the features of ground targets and limits the further remote sensing applications. Therefore, as an important preprocessing work, cloud detection has extremely important practical significance for remote sensing applications. However, the difficulty of cloud detection is to detect thin clouds and broken clouds. Thin clouds are sparsely distributed and relatively transparent compared with thick clouds, and the size of broken clouds is usually small, which makes cloud detection task challenging.

So far, researchers have developed a variety of cloud detection methods, which are mainly divided into threshold-based methods, machine learning-based methods and deep learning-based methods.

Threshold-based methods mainly construct spectral features from multi-band images, and then set a specific threshold to distinguish cloud pixels from non-cloud pixels. For example, ACCA proposed by Irish et al. [5] is an automatic cloud coverage evaluation algorithm for Landsat-7. The core idea of this algorithm is to combine the information of multiple spectral bands and use decision tree and threshold function to determine whether each pixel is cloud. The automatic cloud and shadow detection algorithm developed by Huang et al. [6] takes clear forest pixels as a reference, separates the cloud and reference pixels in the spectral-temperature space, estimates the height of the cloud on the basis of the cloud extraction results, and determines the position of the shadow in combination with the sun illumination. The FMask method proposed by Zhu et al. [7] sets multiple thresholds according to the spectral features of clouds in different bands of Landsat images, thus realizing automatic detection of clouds and cloud shadows. The threshold-based method has two disadvantages: first, it needs to obtain spectral feature information from multiple bands, which is limited by optical sensors and has poor universality in different satellite images. Second, the effect of this method is poor in complex background images, because there may be ground objects similar to clouds in the background, such as snow and ice, which makes it difficult to identify the threshold method [8].

The main content of the cloud detection algorithm based on machine learning is to label a large number of cloud pixels and non-cloud pixels as training samples, manually design some cloud features to build a feature set, and then use the feature set to train the classifier. Li et al. [9] proposed a cloud detection method based on SVM classifier, texture features, average gradient and other features. Bai et al. [10] integrated various features such as spectral features, texture features and NDVI, and used SVM classifier to carry out cloud detection. Shao et al. [11] designed a cloud detection method based on auto encoding network and fuzzy function. However, if machine learning methods are to be used to achieve better detection results, researchers are required to manually design better features.

Deep learning methods use deep convolutional neural networks to automatically learn target features from a large number of samples, which significantly improves the accuracy of image information extraction. In the field of Cloud detection, Mohajerani et al. [12] proposed Cloud-Net, which adopted a U-shaped encoder-decoder structure and

achieved good results on Landsat 8 datasets. Guo et al. [13] proposed a Cloud detection network called Cloud-AttU based on UNet [14] and attention mechanism. Liu et al. [15] improved the feature extraction network of D-LinkNet [16] and applied it in the cloud detection task. Although existing deep learning methods have significantly improved the accuracy of cloud detection, thin cloud and broken cloud detection is still a difficult problem.

The novel cloud detection network NM-LinkNet proposed in this paper adopts the encoder-decoder structure, introduces the non-local operation (NL) in the skip connection to capture the long-distance context information, and adds the multi-scale feature aggregation (MFA) module designed in this paper in the decoding process to aggregate the feature information of different scales in each stage of the decoder. The experimental results on SPARCS dataset show that NM-LinkNet improves the detection ability of thin clouds and broken clouds, and the IoU and F1 are 87.50% and 93.33%, respectively, which are better than many mainstream algorithms.

2 Method

2.1 Network Structure

The network structure of NM-LinkNet is shown in Fig. 1. NM-LinkNet uses LinkNet50 [17] as the basic network and uses the encoder-decoder structure. The encoder is the left branch of the network structure, which is composed of ResNet50 [18]. It uses a deep convolutional network to extract low-level features containing spatial details and high-level deep features rich in semantic information. NM-LinkNet adds non-local operations at each level of skip connection to aggregate long-distance context information at each stage. The function of the decoder is to gradually restore the features to their original size. Different from the original LinkNet decoder, this paper designs a multi-scale feature aggregation module in the decoding process to fuse the features of different scales in different stages of the decoder. The input of NM-LinkNet is an image with a dimension of $384 \times 384 \times 3$, and the output is a prediction result of $384 \times 384 \times 1$. Here, the prediction result is displayed with a color image for the convenience of display.

2.2 Non-local Operation

Compared with thick clouds, thin clouds are more transparent and sparsely distributed, which brings challenges to thin cloud detection. In order to deal with the problem that thin clouds are difficult to detect, this paper introduces non-local operation [19] into the network model to calculate the correlation between each pixel, obtain the context information of thin clouds, and enhance the features of thin cloud pixels with sparse distribution. The structure diagram of non-local operation is shown in Fig. 2. The dimension of the input feature is $H \times W \times 2C$. Firstly, the input features are passed through three 1×1 convolutional layers to reduce the number of channels by half, and three features with dimensions of $H \times W \times C$ are obtained, which are respectively represented by u, v and g. Then the dimensions of feature v and g are converted to $HW \times C$, and the dimensions of features u are transformed and transposed to $C \times HW$. Then, the two-dimensional feature of $HW \times HW$ is obtained by matrix multiplication of the feature

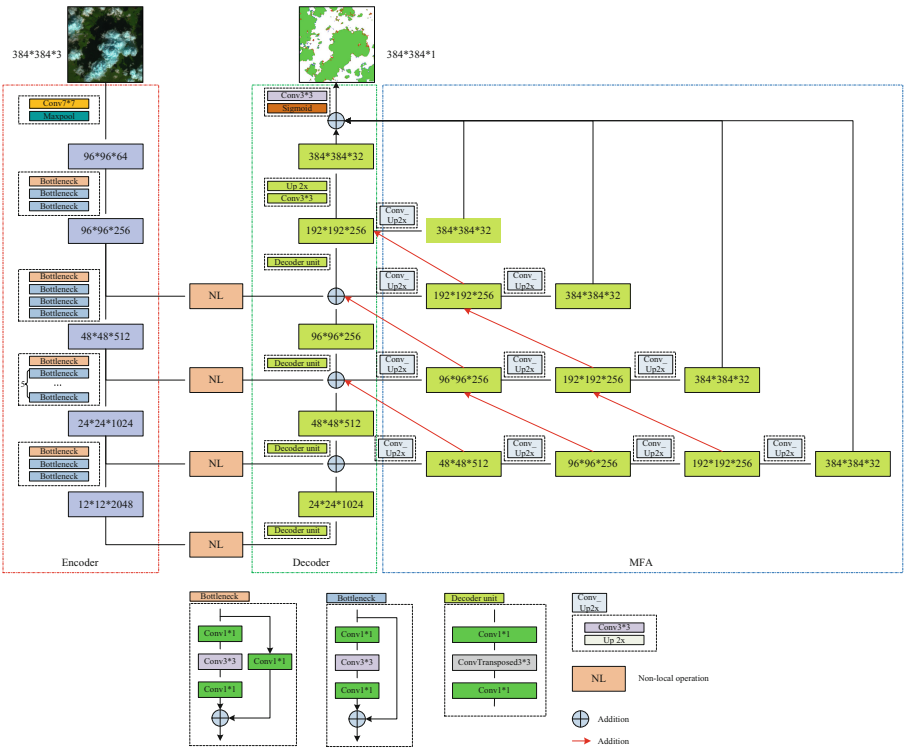

Fig. 1. Network structure of NM-LinkNet.

v and u after dimensional transformation, which represents the similarity between each pixel. The similarity matrix was normalized by softmax function, and then the matrix was multiplied with the feature g after dimension transformation, and then the feature with dimension H × W × C was generated after dimension transformation. Finally, it goes through 1 × 1 convolutional layer to double the number of channels, and then adds the input features, so as to finally get the output features of non-local operation.

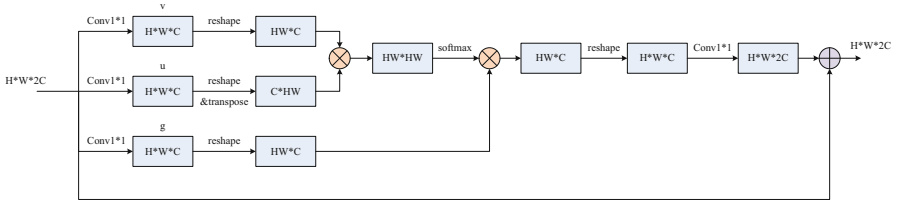

Fig. 2. Structure of non-local operation.

2.3 Multi-scale Feature Aggregation

The size of broken clouds in the image is generally small due to non-aggregation, and the detection ability of common methods for small broken clouds is often poor. To solve this problem, this paper proposes the MFA module to extract features of different scales to enhance the detection accuracy of the network for small broken clouds. The structure of the module is shown in Fig. 1. In the decoder of the original LinkNet, the output features of the decoding units of each stage, after being added with the features of the skip connection, will be used as the input of the next decoding unit. The output feature of the last decoding unit is then processed by up-sampling layer and convolution layer to get the prediction result. Different from LinkNet, the proposed NM-LinkNet incorporates the features of different scales at different stages of the decoder. In order to better preserve the spatial details of the features, MFA adopts a stepwise up-sampling method for the features of each stage of the decoder, and adds the features with the same dimension in adjacent stages to achieve the full fusion of features of different scales. Finally, MFA adds the $384 \times 384 \times 32$ features obtained after up-sampling at each stage of the decoder, and then obtains the cloud prediction result after up-sampling layer, convolution layer and sigmoid function.

3 Experiment

3.1 Dataset

In this paper, experiments are carried out on SPARCS dataset. The SPARCS dataset contains 80 Landsat8 images with 1000×1000. In this paper, 80 original images are seamlessly cropped into 720 384×384 image tiles, and the data are divided according to the ratio of 7:1:2. Finally, the training set contains 504 image tiles, the validation set contains 72 image tiles, and the testing set contains 144 image tiles.

3.2 Experimental Setup

In this paper, Pytorch 1.8.0 [20] is used to complete all the experiments. The hardware environment of the experiments is NVIDIA GeForce RTX 3060 (12G), and the operating system is Windows 10. The epochs and bach size of the experiments are 200 and 4, respectively. The optimizer is SGD, initial learning rate is 0.01 and the learning rate reduction strategy is "poly". The loss function used in the experiments is binary cross-entropy (BCE), and the predicted threshold is 0.5. In addition, this paper adopts random horizontal flip, vertical flip, Gaussian blur and rotation to expand the training samples.

3.3 Evaluation Metrics

In this paper, IoU and F1, which are commonly used in cloud detection, are used to evaluate the performance of each method. The formulas of IoU and F1 are as follows. The meaning of IoU is the ratio of intersection and union of prediction image and label image. F1 is an index to comprehensively evaluate Precision and Recall, which is often used in remote sensing semantic segmentation tasks. True Positive (TP) indicates

the number of pixels that are correctly predicted to be the cloud. False Positive (FP) indicates the number of pixels that are predicted to be cloud, but are predicted incorrectly. False Negative (FN) indicates the number of pixels that are incorrectly predicted as the background.

$$\text{IoU} = \frac{\text{TP}}{\text{TP+FP+FN}} \tag{1}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{4}$$

3.4 Result Analysis

In order to verify the effectiveness of NM-LinkNet, this paper conducted comparative experiments with FCN [21], UNet [14], LinkNet50 [17], DeepLabv3+ [22] and Cloud-Net [12] on SPARCS datasets. The quantitative comparison results are shown in Table 1. The accuracy of NM-LinkNet is the highest, with IoU and F1 reaching 87.50% and 93.33%, respectively, which exceeds other comparison methods. The IoU and F1 of NM-LinkNet are 5.53% and 3.24% higher than that of FCN, and 0.77% and 0.44% higher than that of LinkNet50, respectively. Compared with the basic network, the reason for the accuracy improvement of NM-LinkNet is the addition of non-local operation and multi-scale feature aggregation module, which improves the detection ability of thin clouds and broken clouds, thus improving the performance on SPARCS dataset.

Table 1. Comparison of results of different methods on SPARCS dataset (%)

Method	IoU	F1
FCN	81.97	90.09
Cloud-Net	85.03	91.91
DeepLabv3 +	85.11	91.95
UNet	85.39	92.12
LinkNet50	86.73	92.89
NM-LinkNet	**87.50**	**93.33**

In addition to quantitative comparison, this paper also compares the methods in terms of qualitative visualization results, as shown in Fig. 3. Thin clouds in the first column of images are sparsely distributed and relatively transparent. FCN detected only a small fraction of thin clouds. Compared with other methods, NM-LinkNet detected

the most thin cloud regions. Similarly, the images in the second and third columns are also thin clouds. NM-LinkNet obtains long-distance context information through non-local operation, so it has a better detection effect on thin clouds. The fourth to sixth columns contain images and results of broken clouds. FCN has many missed detection, and Cloud-Net, DeepLabv3+, UNet and LinkNet50 also have different degrees of missed detection. However, NM-LinkNet has a stronger ability to detect broken clouds due to

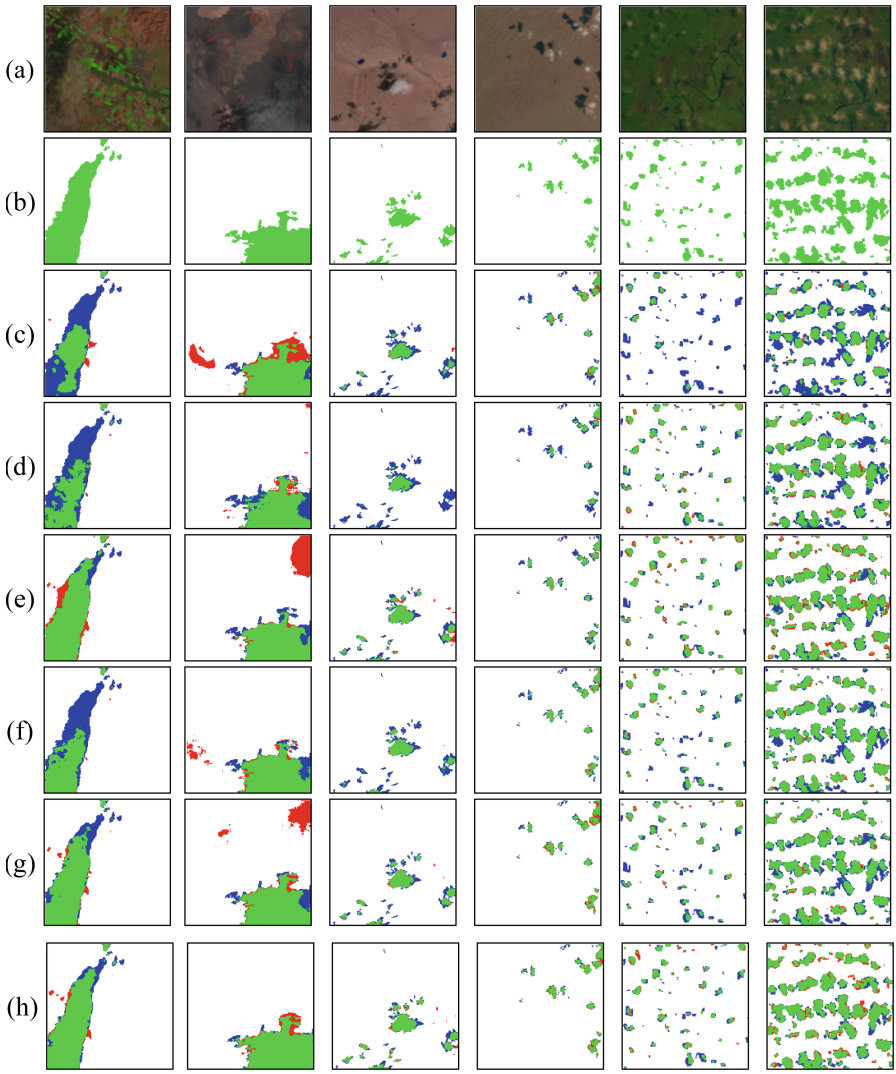

Fig. 3. Visualization results of each method on SPARCS dataset. (a) Image; (b) Label; (c) FCN results; (d) Cloud-Net results; (e) DeepLabv3 + results; (f) UNet results; (g) LinkNet50 results; (h) NM-LinkNet results; Green: TP; White: TN; Blue:FN; Red:FP.

the full aggregation of multi-scale features using the MFA module, with relatively less missed detection.

4 Discussion

In order to verify the effectiveness of each improved module, the ablation experiment of NM-LinkNet is carried out in this paper. In this paper, non-local operation and MFA are added to LinkNet50 successively to evaluate the performance of the model. The results of the ablation experiment are shown in Table 2. The IoU and F1 of LinkNet50 are 86.73% and 92.89%, respectively. After adding non-local operation on the basis of LinkNet50, IoU and F1 are improved by 0.07% and 0.04%, respectively. After adding the MFA module, IoU and F1 increased by 0.7% and 0.4%, respectively. After adding non-local operation and MFA module respectively, the accuracy of the network is improved, indicating that these two modules can improve the accuracy of cloud detection.

Table 2. Results of ablation experiment (%)

Method	IoU	F1
LinkNet50	86.73	92.89
LinkNet50 + NL	86.80	92.93
LinkNet50 + NL + MFA	87.50	93.33

5 Conclusion

In order to solve the problem that thin clouds and broken clouds are difficult to detect, a new cloud detection network NM-LinkNet is proposed in this paper. NM-LinkNet adopts encoder-decoder structure and introduces non-local operation to obtain long-distance context information, so as to improve the detection ability of sparse thin clouds. In addition, a multi-scale feature aggregation module is designed to fully aggregate the features of different scales of broken clouds, so as to improve the detection ability of small broken clouds. The experimental results on the public SPARCS dataset show that NM-LinkNet outperforms many mainstream methods, and improves the detection ability of thin and broken clouds, indicating that NM-LinkNet is effective for improving the accuracy of cloud detection.

Funding. This research was supported by the National Natural Science Foundation of China (42050202).

References

1. Yuan, X., Shi, J., Gu, L.: A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Syst. Appl. **169**, 114417 (2021)

2. Cheng, G., Han, J.: A survey on object detection in optical remote sensing images. ISPRS J. Photogramm. Remote. Sens. **117**, 11–28 (2016)
3. Asokan, A., Anitha, J.: Change detection techniques for remote sensing applications: a survey. Earth Sci. Inf. **12**(2), 143–160 (2019)
4. Zhang, Y., Rossow, W.B., Lacis, A.A., et al.: Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: Refinements of the radiative transfer model and the input data. J. Geophys. Res. Atmos. **109**(D19) (2004)
5. Irish, R.R., Barker, J.L., Goward, S.N., et al.: Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. Photogramm. Eng. Remote. Sens. **72**(10), 1179–1188 (2006)
6. Huang, C., Thomas, N., Goward, S.N., et al.: Automated masking of cloud and cloud shadow for forest change analysis using Landsat images. Int. J. Remote Sens. **31**(20), 5449–5464 (2010)
7. Zhu, Z., Woodcock, C.E.: Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. **118**, 83–94 (2012)
8. Qiu, S., He, B., Zhu, Z., et al.: Improving Fmask cloud and cloud shadow detection in mountainous area for Landsats 4–8 images. Remote Sens. Environ. **199**, 107–119 (2017)
9. Li, P., Dong, L., Xiao, H., et al.: A cloud image detection method based on SVM vector machine. Neurocomputing **169**, 34–42 (2015)
10. Bai, T., Li, D., Sun, K., et al.: Cloud detection for high-resolution satellite imagery using machine learning and multi-feature fusion. Remote Sens. **8**(9), 715 (2016)
11. Shao, Z., Deng, J., Wang, L., et al.: Fuzzy autoencode based cloud detection for remote sensing imagery. Remote Sens. **9**(4), 311 (2017)
12. Mohajerani, S., Saeedi, P.: Cloud-Net: an end-to-end cloud detection algorithm for Landsat 8 imagery. In: IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 1029–1032. IEEE (2019)
13. Guo, Y., Cao, X., Liu, B., et al.: Cloud detection for satellite imagery using attention-based U-Net convolutional neural network. Symmetry **12**(6), 1056 (2020)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J.., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Liu, G., Wang, G., Bi, W., et al.: Application of improved D-LinkNet model in cloud detection of domestic satellite image. Bull. Surv. Mapp., (11): 54–58, 64 (2021)
16. Zhou, L., Zhang, C., Wu, M.: D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 182–186 (2018)
17. Chaurasia, A., Culurciello, E.: LinkNet: exploiting encoder representations for efficient semantic segmentation. IEEE Vis. Commun. Image Process. (VCIP) **2017**, 1–4 (2017)
18. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
19. Wang, X., Girshick, R., Gupta, A., et al.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
20. Paszke, A., Gross, S., Massa, F., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
22. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer vision, pp. 801–818 (2018)