Water Body Extraction from High-resolution Remote Sensing images Based on Scaling EfficientNets

Li Lu*, Wen Qi2, Wang Bing3, Fan Shichao4, Lingling Li2, Liu Qingjie1

1Research center of Artificial Intelligence, Hangzhou Innovation Institute, Beihang University, Hangzhou China
2National Disaster Reduction Center of China, Ministry of Emergency Management, Beijing, China
3Beijing Research Institute of Space Mechanics and Electricity, Beijing, China

*Corresponding author: bh_lilu@126.com

Abstract. Effectively and accurately extracting water bodies from high-resolution remote sensing images is an important yet extremely challenging task due to ambiguities raised by spectral and texture similarities between water bodies and distractors such as shadows buildings, mountains and vegetations. Recent developed deep learning techniques have provided researchers powerful tools to defeat traditional hand-engineered features by learning powerful hierarchical representations, however, with a cost of very high model complexity and computational resource requirement. To conquer this issue, in this paper, we propose to build a deep learning model for water extraction based on the EfficientNet-B5. We evaluate possible variations of such design and compare them with several widely used approaches on GF-2 and Sentinel-2 satellites. Experimental results demonstrate that our model obtains much better performance than SVM and U-Net, which clearly shows the effectiveness and robustness. We hope this work will facilitate further research in this field and inspire researchers in the community to develop better models.

1. Introduction

Water bodies, such as lakes, rivers, and wetland, are defined as one of the major land cover elements, interactively connecting to atmosphere, biosphere, lithosphere, nature disaster and land hydrosphere [1-2]. Their distribution and changes will directly or indirectly impact the climate and biomass of regional and global areas, making this task of great significance to the remote sensing community.

In the past decades, many methodologies have been developed based on normalized difference water index (NDWI) or thresholding segmentation [3-7]. However, most of them rely on empirical spectral band selection and fine-tuned thresholds, which require rich experience and are hard to be decided in practical applications. Furthermore, the shadows of mountainous, buildings and vegetation regions post great challenges to these methods, because shadows and vegetation usually show similar spectral signatures to polluted rivers or artificial lakes [8-9]. Thanks to the powerful machine learning-based classifiers and sophisticated texture extractors [10-12], water body extraction has been improved significantly over the NDWI-based or threshold-based methods, yet challenges remain when facing shadow or vegetation regions, because low-level geometric and texture features are not discriminative enough to distinguish them from waters.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
Recently, inspired by achievements of using fully convolutional neural networks (FCNs) to solve building and road segmentation in remote sensing images [13-15], researchers attempt to apply deep learning methods [16-19] to water body extraction. For example, Wang X. et al. [16] used VGG16 as the basic network and then constructed FCN models with three different up-sampling structures for multiscale water body extraction. Experiments proved that FCN-8S could extract water information more accurately. To accelerate extraction in urban river areas, Wang Y. et al. [17] proposed a method combining FCN and the GEE platform for off-line learning and online prediction. The method was verified in 36 urban areas nationwide, among which the F1 score and Kappa coefficient of most urban extraction results reached 0.9. To extract fine details of narrow rivers, Liang Z. [18] introduced DenseNet to the U-shape FCN. The method was tested on Landsat8 images with Yangtze river basin, and the accuracy was up to 98.5%. Kim S. et al. [19] used U-Net to extract spectral features related to red tides from GOCI images of Korean Peninsula waters and proved that the learning effect of four-band training data set was 13% lower than that of six-band in terms of average target accuracy.

As known widely that waters in different topographic environments have significant differences in both physical appearances and bio-chemical components. This may bring difficulties for pixel-wise water segmentation. Additionally, the various multiscale designs used in the latest published worked inevitably result in extra model complexity. Slight accuracy increasing often require much more data and computation cost. Thus, to better balance the model complexity accuracy, this paper proposes to extract water bodies using the powerful EfficientNet. We tested our method on challenging urban and mountainous areas of Sentinel-2 and GF-2 optical images. Experiments demonstrated effectiveness of our method.

2. Methodology

2.1. Framework

Fig. 1 depicts the architecture of the proposed water body extraction framework. As shown, the framework consists of two main steps: firstly, a deep convolutional network was introduced as the backbone to extract hierarchical features of water bodies in remote sensing images. Note that we used a pre-trained model as the backbone because it encourages better performances [1]. Here, EfficientNet [20] with different compounding scaling strategies was employed as the backbone. Then, we constructed a U-shaped segmentation network and added those hierarchical features output by the backbone into the final segmentation network. Finally, at the end of the model, a binary mask representing water bodies was obtained. Our model was trained in an end-to-end way.

Figure.1 Architecture of the proposed method for water body extraction.
2.2. Scaling EfficientNets

EfficientNet family (-B0 to -B7) had achieved state-of-the-art accuracy on the image classification task, yet been an order-of-magnitude smaller and faster than previous models. Fig.2 showed the base model of this network, named by EfficientNet B0. To investigate the relationships between classification accuracy and architecture of a network, Tan M. X. et. al. [20] summarized four kinds of scaling ways from the perspective of width, depth, resolution, and compound. By using the different scaling styles, they obtained EfficientNet-B1 to B7.

![Figure 2](image.png)

**Figure 2** The detailed architecture of Efficient-Net B0.

For the width-scaling network, it always enhanced the learning ability by increasing the channel dimension of the baseline. The network owning more channels at each layer would capture more fine-grained characteristics and is easier to be trained. For the depth-scaling network, it scaled up the network by increasing the depth of the baseline, that was, by increasing the number of convolutional layers. The deeper network could capture richer and more complex features. However, due to the following gradient vanishing problem, the network usually was difficult to be trained. Similarly, scaling up the network by increasing the resolution of the input image could capture more fine-grained local features. The compound scaling was a mixture of the above three strategies, which could collect advantages of the three derivations. However, with more scaling ways employed, the volume and parameters of networks would be increased, greatly. It required training more samples and iteration to train models.

In this study, to figure out which network was more suitable for water body extraction in remote sensing images, we used different EfficientNet models (from B1 to B7) as the backbone to extract multiple hierarchical features of input images. Then the features were concatenated into the corresponding decoder blocks of the U-Net (shown in Fig. 1) to train the model.

3. experiments

To validate the effectiveness and robustness of the proposed method, optical remote sensing images collected from GF-2 and Sentinel-2 were used as training and test data. To test generalization ability of our methods, we tested them on challenging urban and mountainous regions. Besides, the U-Net based with VGG16 and SVM were selected as comparisons to show the advantage of the proposed methods in water bodies extraction.

3.1. Datasets

GID (GaoFen Image Dataset) [21] dataset is mainly used for land use and land cover classification. It includes 150 GF-2 optical remote sensing images collecting from nearly 60 different cities in China, covering an overall area of more than 50,000 square kilometers. The dataset contains RGB images with a resolution of 10 m and multi-spectral images with a resolution of 4 m obtained by GF-2 satellite. Here we only use RGB images to generate training samples. The size of image is 6908×7300 pixels. In the following experiment, we cropped them into patches with a size of 256×256 pixels. The dataset can be download from http://captain.whu.edu.cn/GID/.

The original dataset has at 5 land cover types, but in this paper, we merely focused on the water body and treated other classes as negative samples during training. Table 1 showed the training and test samples used in experiments.

| Training | Testing | Resolution (meter) | Patch size | Sensor type |
|----------|---------|--------------------|------------|-------------|
| 5,692    | 1,146   | 10                 | 256×256    | GF-2        |

Table.1 Basic information of GID dataset
3.2. Settings
The experiments are conducted on a single Nvidia RTX 2080Ti GPU with 11GB memory onboard. During training, the initial learning rate is set to 0.0005 and the Adam optimizer is used to update the weights. The batch size was set to 16. The loss function is built by merging the focal loss [22] and dice loss [23], its mathematical formulation could be presented by the following equation:

\[ L = L_{\text{dice}} + \lambda \cdot L_{\text{focal}} \]

\[ = C \sum_{c=0}^{C} \frac{TP_p(c)}{TP_p(c) + \alpha FN_p(c) + \beta FP_p(c)} - \lambda \frac{1}{N} \sum_{c=0}^{C} \sum_{n=1}^{N} g_n(c)(1 - p_n(c))^2 \log(p_n(c)) \]

where \( N \) is the total number of samples, \( C \) is the number of classes, \( TP_p(c) \), \( FN_p(c) \) and \( FP_p(c) \) are true positive, fake negative and fake positive samples of the class \( c \), respectively. Values of \( \alpha \) and \( \beta \) are both set as 0.5, and \( \lambda \) is set by 1. The OA (Overall Accuracy), Kappa coefficient, PA (Producer Accuracy), UA (User Accuracy) are used in this paper to quantitatively evaluate the experimental results.

3.3. Results on GID
Here, the following Fig.3 and Fig.4 respectively show the water body extraction results and accuracy histograms in test samples of GID dataset, by using the proposed method with the backbone from EfficientNet-B1, B3, B5 and B7.

Seen from the mapping results of the seven scenes (Fig.3(a1) to Fig.3(g1)), the network with EfficientNet-B5 (referred by '-B5' in the following part) obtains better water body OA than other three backbones, as well as reaching to higher PA, UA and Kappa. That means that the network based on '-B5' could accurately extract pixels belonging to water, while taking fewer false alarms. This indicates that the discriminative feature learning ability of the backbone had firstly increased on the baseline -B0, then performance started to decrease as the continue additional complexity of the model.

|                  | EfficientNet-B1 | EfficientNet-B3 | EfficientNet-B5 | EfficientNet-B7 | U-Net |
|------------------|------------------|------------------|------------------|------------------|-------|
| **Time**         | 4.5272           | 4.9393           | 6.2261           | 8.0137           | 2.3   |
| **Size**         | 7.8M             | 12M              | 30M              | 66M              | 20M   |

Figure 3 Extraction accuracy histograms of the proposed methods in GID. X-axis corresponding to 7 scenes, namely (a1) to (g1). (Zoom in to see clearly.)
3.4. Comparisons

Fig. 5 and Tab. 3 respectively show extraction results and average statistical precision of the proposed, U-Net and SVM in Sentinel-2 images. From the Ave.OA (average overall accuracy), the proposed method obviously performs best in the five scenes by 0.9613. Specially, in Fig.5 (b) and (d), our method could separate the narrow road from water area, while the U-Net based on VGG16 had discontinuity parts in these scenes. Therefore, it could be convinced that the EfficientNet was more robust and effective than VGG16 for feature extraction when dealing with water body extraction in optical remote images. Besides, the Tab.2 calculates the costs of time consuming and memory usage, when a test sample was predicted by each loaded network. It is obviously seen that the backbone ‘-B3’ and ‘-B5’ had better efficient extraction performance and lower cost than others. So combining the water extraction precision and costs results, we prefer to use EfficientNet-B5 as the backbone.
Figure 5 Comparison results of the three methods on Sentinel-2 images

Table 3 accuracy on Sentinel-2

| Methods      | Ave.OA | Ave.PA | Ave.UA | Ave.Kappa |
|--------------|--------|--------|--------|-----------|
| EfficientNet-B5 | 0.9613 | 0.9251 | 0.9045 | 0.9159    |
| U-Net        | 0.8472 | 0.8112 | 0.8209 | 0.7992    |
| SVM          | 0.6021 | 0.6723 | 0.6654 | 0.5539    |

4. Conclusion
This paper has proposed a robust and accurate water body extraction method by introducing the scaling EfficientNet as the backbone to build a segmentation network. We conduct extensive experiments on GF-2 and Sentinel-2 images and compare it with the other two methods to demonstrate the effectiveness and superiority. The experimental results show that our methods have shown significant advantages over other learning methods, in terms of extraction accuracy, as well as costs of running time and memory capacity. Finally, according to the comparison analysis of different backbone settings of using EfficientNet, it can draw a conclusion that the backbone ‘-B5’ model could extract finer water bodies in both urban and mountainous areas.

Acknowledgment
This paper is supported by the National Key Research and Development Program of China (No. 2016YFC0803109).

This work was supported by NSFC Nos. 41871283 and 61601011.

References
[1] N. Qayyum, S. Ghuffar, H. M. Ahmad, A. Yousaf, and I. Shahid, “Glacial Lakes Mapping Using Multi Satellite PlanetScope Imagery and Deep Learning,” ISPRS Int. J. Geo-Inf., vol. 9, 2020, pp. 560-583.
[2] Y. D. Fan, Q. Wen, W. Wang, P. Wang, L. L. Li, P. Zhang, “Quantifying disaster physical damage using remote sensing data—a technical work flow and case study of the 2014 Ludian earthquake in China,” Int. J. Disaster Risk ence, vol. 8, 2017, pp: 1-18.
[3] H. Q. Xu, “A study on information extraction of water body with the modified normalized difference water index,” Int. J. Remote Sens., vol. 9, 2005, pp. 589-595.
[4] J. C. Luo, Y. W. Sheng, Z. F. Shen, J. L. Li, L. J. Gao, “Automatic and high-precise extraction
for water information from multispectral images with the step-by-step iterative transformation mechanism,” Int. J. Remote Sens., vol.13, 2009, pp. 610-622.

[5] Z. F. Shen, L. G. Xia, J. L. Li, J. C. Luo, X. D. Hu, “Automatic and high precision extraction of rivers from remotely sensed images with gaussian normalized water index,” Int. J. Image and Graph, vol.18, 2013, pp. 421-428.

[6] Q. D. Guo, R. L. Pu, J. L. Li, J. Cheng, “A weighted normalized difference water index for water extraction using Landsat imagery,” Int J. Remote Sens., vol.38, 2017, pp. 5430-5445.

[7] Y. Zhou, H. R. Zhao, H. Hao, C. J. Wang, “A new multi-spectral threshold normalized difference water index (MST-NDWI) water extraction method-a case study in YanHe watershed,” In Proceedings of the ISPRS Technical Commission III Midterm Symposium on "Developments, Technologies and Applications in Remote Sensing", May, 2018, China, Beijing.

[8] H. Wang, F. Qin, “Summary of the research on water body extraction and application from remote sensing image,” Sci. Surv. Map., vol.43, 2018, pp. 23-32.

[9] D. Li, B. S. Wu, B. W. Chen, Y. Xue, Y. Zhang, “Review of water body information extraction based on satellite remote sensing,” J. Tsinghua University (Science and Technology), vol.60, 2020, pp.147-161.

[10] B. Li, X. Z. Zhou, “Research on extraction method of river water-body in mountain area of western Sichuan based on MODIS data,” J. Subtrop. Resour. Environ., vol.25, 2009, pp. 59-62.

[11] Y. Y. Mao, Y. F. Chen, “Water Body Extraction Method Based on Object-oriented in Mountainous Area with Landsat 8 Image,” J. Subtrop. Resour. Environ., vol.10, 2015, pp. 86-92.

[12] C. Z. Liu, J. C. Shi, S. Gao, L. Chen, “The study on extracting of water body from MODIS image using an improved linear mixture model,” Remote Sens. Inf., vol.1, 2010, pp. 84-88.

[13] J. Du, “Deep learning based UAV remote sensing image water body identification,” Jiangxi Science, vol.35, 2017, pp: 158-161.

[14] T. C. Zuo, J. T. Feng, X. J. Chen, “HF-FCN: hierarchically fused fully convolutional network for robust building extraction,” In Asian Conference on Computer Vision, 2016, pp. 291-302.

[15] Z. X. Zhang, Q. J. Liu, Y. H. Wang, “Road extraction by deep residual U-Net,” IEEE Geosci. Remote Sens. Lett., vol.15, 2018, pp. 749-753.

[16] X. Wang, L. C. Sui, M. Q. Zhong, D. M. Li, L. L. Dang, “Fully convolution neural networks for water extraction of remote sensing images,” Bull. Surv. Map., 2018, vol.6, pp. 41-45.

[17] Y. D. Wang, Z. W. Li, C. Zeng, G. S. Xia, H. Shen, “An urban water extraction method combining deep learning and google earth engine,” IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens., vol.13, 2020, pp. 768-781.

[18] Z. Y. Liang, “Research on water information extraction method of multi-source remote sensing based on deep learning and its application”, 2019, China, Anhui University.

[19] S. M. Kim, J. Shin, S. Baek, J. H. Ryu, “U-Net convolutional neural network model for deep red tide learning using GOCI,” J. Coas. Resea., 2019, vol.90, pp: 302-.

[20] M. X. Tan, V. L. Quoc, “EfficientNet: rethinking model scaling for convolutional neural networks,” 2019, arXiv:1905.11946.

[21] Y. M. Tong, G. S. Xia, Q. Lu, “Learning transferable deep models for land-use classification with high-resolution remote sensing images, 2018, arXiv preprint arXiv:1807.05713.

[22] Y. Y. Lin, P. Goyal, R. Girshick, K. M. He, P. Dollar, “Focal loss for dense object detection,” Int. Conf. Comput. Vision, 2017.

[23] F. Milletari, N. Navab, S. A. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,” 2016 Fourth International Conference on 3D Vision (3DV), pp. 565-571