Multimodal water body extraction network based on remote sensing satellite images and digital surface model

Wenbo Ji[a], Weibin Li[*a,b], Xihui Feng[c], Tianyi Zhang[a], Chenhao Qin[a], Yi Ren[b]
[a]School of Artificial Intelligence, Xidian University, Xi'an 710071, Shaanxi, China; [b]Lab. of AI, Hangzhou Institute of Technology of Xidian University, Hangzhou 311231, Zhejiang, China; [c]Key Lab. of Coal Resources Exploration and Comprehensive Utilization, Ministry of Natural Resources, Xi'an 710021, Shaanxi, China

ABSTRACT

Multispectral remote sensing satellite images exhibit characteristics such as small objects, complex scenes, significant changes in object scale, and difficulty in distinguishing regions with similar spectral features. As the spectral reflection characteristics of water bodies vary with factors like season and geographical location, different background information affects the accuracy of the water body extraction. Therefore, extraction of broken and discontinuous water bodies is still challenging. Recent studies have shown that using multimodal information can represent the features of targets from different perspectives, thereby improving the robustness of semantic segmentation. To address these issues, this paper utilizes the spectral index's ability to recognize water to drive the extraction accuracy of neural networks for water recognition. A multimodal remote sensing semantic segmentation network (MRSSNET) is proposed, which integrates water index method to fuse images with Digital Surface Model (DSM) images. We use deep learning-based segmentation models to perform water segmentation, such as Fully Convolutional Networks (FCN), U-Net, Segformer, SegNext and Deeplabv3+, representatively. Experimental results demonstrate that MRSSNET outperforms the other four algorithms in identifying water bodies within complex and discontinuous geographical environments.

Keywords: Multimodal fusion, semantic segmentation, water body, remote sensing image

1. INTRODUCTION

Surface water is not only an important part of water resource, but also one of the important resources necessary for people's lives. Rapid and accurate acquisition of spatial distribution information of water bodies is crucial to the socio-economic and sustainable development of water resource. With remote sensing technology's rapid advancement, extracting information about water bodies has emerged as a significant research area[1]. With the continuously development of remote sensing technology[2,3] the use of satellite remote sensing images to extract water body area, geometric shape, water body ecological environment and other information has been applied in water resources survey, environmental protection and water resource macro-monitoring and other fields[4]. In the past few decades, surface water extraction methods based on remote sensing images have been mainly divided into three categories: threshold methods based on single-band images, identification methods based on spectral indexes, and image classification methods. As the spectral reflection characteristics of water bodies vary with factors like season, geographical location, and depth, non-water objects may exhibit similar spectral reflections to water body in certain bands. Therefore, the threshold method of single-band images to obtain the spatial distribution of surface water has certain limitations. The spectral index-driven water extraction method is based on the different spectral reflection characteristics of water in different bands and combines multiple bands to highlight the water body information in remote sensing images. For example, the Normalized Difference Water Index (NDWI) proposed by McFeeters[5] is based on the prominent reflection of water in the green band and the strong absorption in the infrared band, which highlights the water body information in the image and suppresses vegetation and soil. information. Xu[6] aimed at the problem that NDWI cannot suppresses the shadows of tall buildings very well. By combining the green band and the mid-infrared band, the water area of the entire image can be displayed in a balanced manner and the shadows of tall buildings can be suppressed to a certain extent. Feyisa et al.[7] extracted water bodies by constructing an automated water extraction index (AWEIsh) model, and selected water bodies under various backgrounds (such as black soil, shadows, etc.) around the world to conduct a large number of water

*weibinli@xidian.edu.cn; phone 18991899866

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 1339547 · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048285

extraction experiments. The results shows that the index can effectively amplify the difference between water bodies and non-water bodies, thereby achieving accurate extraction of water body information; the spectral index method is simple and effective, and is still an effective way to extract large-scale water bodies.

In recent years, the development of artificial intelligence technology has prompted some scholars to study the extraction of surface water bodies based on machine learning. For example, Abid et al.[8] used an unsupervised curriculum learning method based on convolutional neural networks to identify water bodies, which overcomes the challenges faced by remote sensing images. Chen et al.[9] used convolutional neural networks to extract water bodies. By comparing with traditional methods such as the normalized difference water index (NDWI), they proved the effectiveness of deep learning methods in water body extraction. Zhang et al.[10] proposed the MF-Segformer network based on multi-scale fusion technology, which has good performance in Weihe River Basin extraction.

Semantic segmentation for single modality has achieved outstanding performance in the CV Community[11-14]. However, there is limited research on multimodal fusion tasks. Compared to single-mode remote sensing data, limited by resolution and spectrum, multimodal data can display the features of targets from different perspectives[15,16]. Therefore, utilizing complementary features of different modal data can better represent ground features and improve segmentation performance. Previous research on remote sensing image segmentation mostly focused on using visible light band combinations for land feature segmentation, while neglecting the rich band combinations of remote sensing images. Therefore, in terms of data, this article uses three spectral index methods to fuse images as RGB image inputs, and proposes an FEM module to fuse the features of spectral index images and DSM images. Finally, improvements are made in the loss function.

2. METHODS

The network structure of the proposed MRSSNET is presented in Figure 1. The network adopts a classic codec structure. The encoder is composed of a dual-branch ResNet50. At each layer, the optical branch and DSM branch features are fused in the FEM module. Finally, the high-level features are combined and sent to the DenseASPP and Concurrent Spatial and Channel Squeeze and Channel Excitation (scSE) modules[17].

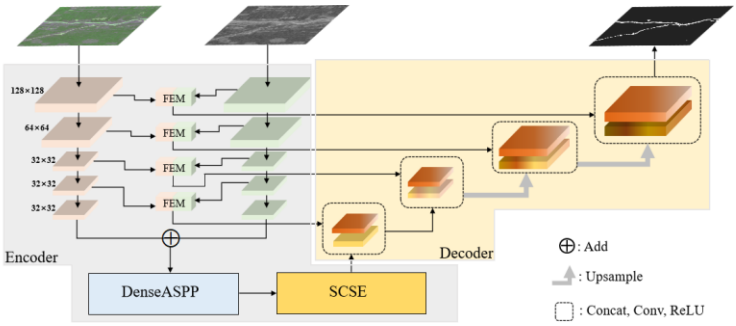

Figure 1. The MRSSNET network structure.

Since there are differences in spectral and texture information between water and its surroundings, it is necessary to make full use of these differences to accurately identify the water. Using RGB images and DSM images as dual inputs to the network, features are first extracted through ResNet50, and then the FEM module is used to fuse the features obtained at each step. The fused feature map is used as one of the inputs to the decoder. The cross-scale feature fusion module FEM can effectively prevent the problem of similar adjacent features and ensure the differentiation of fused features. It reduces information dispersion and interacts, and integrates global features to obtain more comprehensive and rich water body features. The DenseASPP enhances feature reuse and information flow through dense connections, and detailed features are fused layer by layer throughout the network, allowing for better fusion of multi-scale features. This enables the acquisition of richer contextual information, which helps the network improve segmentation performance. The SCSE module excels at capturing critical features of WBs in both the channel and spatial dimensions, effectively reducing information diffusion. By enhancing the ability to capture contextual information, it enables the acquisition of more comprehensive and detailed WB features. This module effectively addresses the challenge of capturing fine details that existing models often struggle with. In terms of the loss function, the auxiliary head is used to calculate the auxiliary loss value, which is conducive to the optimization learning of the overall model[18], thereby reducing the occurrence of missed extractions or discontinuous extractions.

Figure 2 shows the structure diagram of the FEM module. FEM can fuse these features, use semantic and detailed information to eliminate the interference of non-water features around the water body, and enhance the ability to extract edge information from WB. In addition, the FEM module can effectively prevent the problem of similar adjacent features and ensure the differentiation of fused features. The low-level feature maps $X_{\text{low}}^{\text{input}} \in (R^{\frac{H}{8}\times\frac{W}{8}\times C_2}, R^{\frac{H}{8}\times\frac{W}{8}\times C_3})$ are the feature output maps of Stage1 and Stage2 in the backbone respectively. The high-level feature maps $X_{\text{high}}^{\text{input}} \in (R^{\frac{H}{4}\times\frac{W}{4}\times C_4}, R^{\frac{H}{8}\times\frac{W}{8}\times C_5})$ are the feature output maps of Stage4 and Stage3 in the backbone respectively. The FEM module automatically resizes the high-level feature maps in the backbone to match the low-level feature map through upsampling, and performs feature fusion using cross-layer methods to ensure diversity in the fused features.

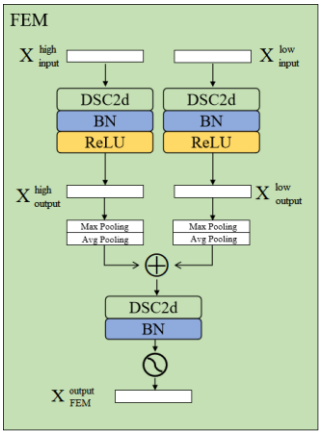

Figure 2. The FEM module structure diagram.

The loss function of MRSSNET is Loss_{seg}, which consists of focal loss and dice loss with an equal weight ratio of 1:1. The formula is as follows:

$$\text{Loss}_{\text{seg}} = \text{Loss}_{\text{Focal}} + \text{Loss}_{\text{Dice}} \tag{1}$$

where $\text{Loss}_{\text{Focal}}$ is the focal loss value and $\text{Loss}_{\text{Dice}}$ is the dice coefficient loss value. During the task of extracting water bodies, only a small area in a large number of sample images contains water bodies. When segmenting small targets, Dice Loss can alleviate the negative effects caused by the imbalance between the foreground and background in the samples. Focal Loss is an improvement based on cross-entropy loss. When there is an imbalance between positive and negative samples in the dataset, Focal Loss helps improve model performance. Therefore, in this task, Focal Loss and Dice Loss are used together to make full use of the advantages of both. Focal Loss can handle category imbalance, while Dice Loss can ensure accurate boundary segmentation.

SCSE is composed of Spatial Squeeze and Channel Excitation (cSE) and Channel Squeeze and Spatial Excitation (sSE). sSE compresses the feature map in the channel dimension through convolution, aggregates the channel information onto the spatial representation, and then generates the weight coefficient of each position through the Sigmoid activation function. Finally, the weight coefficient is compared with the original feature map. Position-by-position multiplication implements Spatially Recalibrate, cSE aggregates spatial information into the global representation of each channel through global average pooling, then obtains the importance of the channel through two fully connected layers and a Sigmoid activation function, and finally uses the weight coefficient to compare it with the original feature map channel by channel, multiplying to achieve channel-wise recalibration.

3. EXPERIMENTS

3.1 Dataset and experimental details

Since there are differences in spectral and texture information between water and its surroundings, it is necessary to make full use of this difference to accurately identify the water. In terms of data sources, True color images are composed of red, green and blue bands, which are 4, 3, and 2 bands respectively (RGB 432). True color images are no

longer used as training data, but NDWI, MNDWI, and AWEI are used as R, G and B band input to enhance the ability to extract water body edge information.

In this study, we selected the 2016 Landsat 8 OLI 30 m resolution image of the Weihe River Basin for dataset production. The corresponding ground truth is obtained using visual interpretation. The DSM data uses the ALOS World 3D-30 m dataset, which is a global DSM dataset with a resolution of 30 m. Next, we used the characteristics of different spectra in multispectral data and use a spectral calculator to calculate three spectral index images. For example, NDWI uses the green band and near-mid-infrared band, MNDWI uses the green band and mid-infrared band, and AWEI uses blue, green, near-mid-infrared, mid-infrared, and thermal infrared bands. We generated the corresponding NDWI, MNDWI, and AWEI images and merge them. We crop the merged image and corresponding label into a 256×256 small image, obtaining a total of 16610 pairs of sample data. We divided all data into training dataset, verification dataset and test dataset according to the ratio of 6:2:2, that is randomly selected 9966 pairs of samples as the training dataset, 3322 pairs of samples as the verification dataset, and the remaining 3322 pairs of samples as a test dataset. In order to reduce the overfitting of the model and improve the accuracy of the model, we used a series of methods to increase the training sample set, including flipping, rotating, cropping, deforming, and scaling. Finally, we expanded the training dataset to 24,000 pairs of samples. The NDWI, MNDWI, AWEI formula is as follows:

$$\mathrm{NDWI} = (Green - NIR)/(Green + NIR) \tag{2}$$

$$\mathrm{MNDWI} = (Green - MIR)/(Green + MIR) \tag{3}$$

$$\mathrm{AWEI} = BLUE + 2.5 * Green - 1.5 * (NIR + MIR) - 0.25 * SWIR \tag{4}$$

3.2 Results and discussion

In order to verify the performance of our proposed algorithm, we selected Fully Convolutional Networks (FCN), U-Net, Segformer, SegNext and DeeplabV3+ for comparison. Table 1 shows the F1_score, Precision, Recall and mIoU of different algorithms on the test dataset.

Table 1. Comparison of water body extraction performance of different models on the visible light RGB 432 dataset.

Model	OURS	DeepLabV3+	SegNext	Segformer	U-Net	FCN
mIoU (%)	**85.27**	83.41	82.67	81.91	81.52	80.34
Precision (%)	**83.37**	82.13	81.32	77.35	81.83	79.25
Recall (%)	82.29	78.85	73.61	**82.37**	79.04	76.36
F1_score (%)	**82.36**	80.22	75.84	80.04	81.22	77.89

In order to better verify the extraction effect of the model on water bodies, we conducted a qualitative analysis of the results of all algorithms. As shown in Figure 3, each column represents a different scenario, and we selected a total of 5 different scenarios. The first row is a pseudo-color composite image of a real scene, the second row is a labeled image of each scene, and the third to sixth rows are the prediction results of the comparison algorithm.

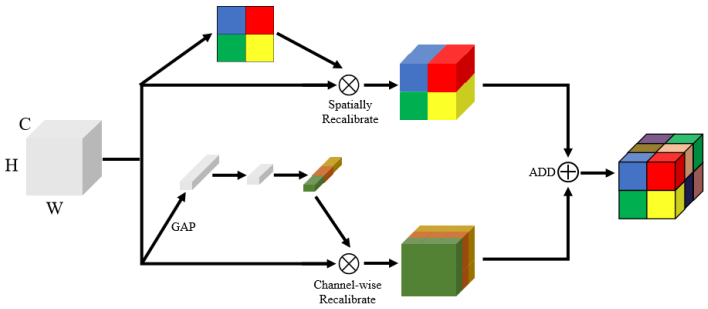

Figure 3. The SCSE module structure diagram.

As shown in Table 1, show the performance of all models on the visible light dataset. we found that MRSSNET has the highest F1_Score, Precision and mIoU values (82.36%, 83.37%, 85.27%). Segformer has the highest Recall.

As shown in Table 2, show the performance of all models on the Spectral Index fusion Images dataset. we found the spectral index fusion image dataset has significantly improved performance, mainly due to its ability to recognize water, which improves the accuracy of neural networks in water recognition. MRSSNET achieved the highest F1_score, Precision, recall and mIoU values (85.86%, 86.51%, 85.22%, 88.08%).

Table 2. Comparison of water body extraction performance of different models on spectral index fusion images dataset.

Model	OURS	DeepLabV3+	SegNext	Segformer	U-Net	FCN
mIoU (%)	**88.08**	86.02	83.31	84.93	84.78	83.57
Precision (%)	**86.51**	85.83	82.69	79.06	83.27	81.12
Recall (%)	**85.22**	80.41	75.75	84.03	80.09	77.87
F1_score (%)	**85.86**	83.02	77.06	81.46	82.83	79.46

As shown in Figure 4, the green solid line is a key focus area. The solid yellow line represents water bodies that have not been extracted. The solid red line indicates that the extracted water body is not continuous and there is a phenomenon of fracture. It can be seen from the results that each model exhibits varying degrees of incompleteness. There are fewer incomplete phenomena in MRSSNET. For small water bodies, the shadow of the mountain will become the main factor affecting the extraction accuracy. SETR and Segformer will smooth out sharp water body boundaries, thereby losing part of the water body information. Each algorithm performs well in the reservoir extraction.

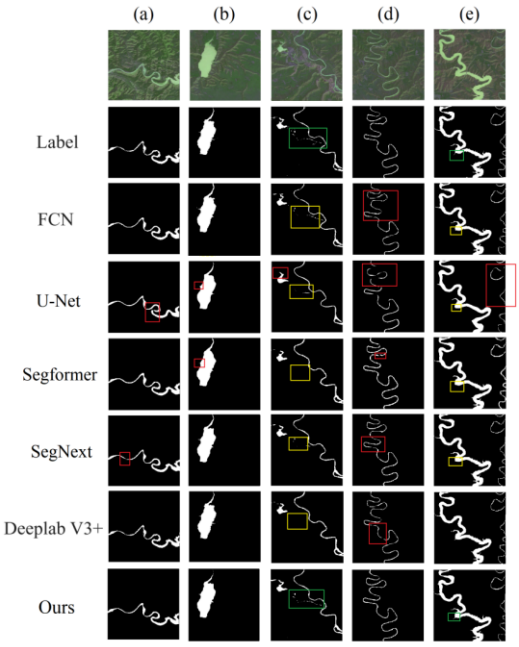

Figure 4. The extraction results comparison of five kinds of water bodies. (a, e): The scenario containing small water bodies; (b): The scenario containing the reservoir; (c, d): The scenario containing small and continuous water bodies.

4. CONCLUSIONS

This study introduces a multimodal fusion semantic segmentation network MRSSNET, that uses spectral index fusion images and DSM images as inputs to the network on the dataset. Unlike traditional methods that input true color images, the network utilizes the spectral index's ability to recognize water and drives the accuracy of the neural network in extracting water bodies. DSM images have excellent surface feature information, which is used to assist networks in learning structured information between water bodies and non-water bodies. It can effectively extract various types of water bodies in complex scenes. It makes full use of the backbone network to extract water body semantic information, uses the FEM module to enhance the water body edge segmentation effect, uses the DenseASPP module to enhance the

small water body extraction effect, and finally uses the auxiliary head to optimize the model learning process. This study utilized Landsat 8 OLI images for experimentation and compared the results with other mainstream segmentation algorithms. The results show that MRSSNET has better performance.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 62176196), the Shaanxi Provincial Water Conservancy Development Fund Project (Project No. 2024SLKJ-06), and the Research Project of Shaanxi Coal Geology Group Co., Ltd., (No. SMDZ-2023CX-14).

REFERENCES

[1] Scanlon, B. R., Fakhreddine, S., Rateb, A., de Graaf, I., Famiglietti, J., Gleeson, T., et al., "Global water resources and the role of groundwater in a resilient water future," Nature Reviews Earth & Environment 4(2), 87-101 (2023).
[2] Zhang, T., Su, H., Yang, X. and Yan, X. H., "Remote sensing prediction of global subsurface thermohaline based on lightgbm," Journal of Remote Sensing 24(10), 1255-1269 (2020).
[3] Su, H., et al., AI-Based Subsurface Thermohaline Structure Retrieval from Remote Sensing Observations, [Artificial Intelligence Oceanography], Springer Nature Singapore, Singapore: Singapore, 105-123 (2023).
[4] Wieland, M., Martinis, S., Kiefl, R. and Gstaiger, V., "Semantic segmentation of water bodies in very high-resolution satellite and aerial images," Remote Sensing of Environment 287, 113452 (2023).
[5] McFeeters, S. K., "The use of the normalized difference water index (NDWI) in the delineation of open water features," International Journal of Remote Sensing, Taylor & Francis 17, 1425-1432 (1996).
[6] Xu, H., "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," International Journal of Remote Sensing, Taylor & Francis 27, 3025-3033 (2006).
[7] Feyisa, G. L., Meilby, H., Fensholt, R. and Proud, S. R., "Automated water extraction index: A new technique for surface water mapping using landsat imagery," Remote Sensing of Environment 140, 23-35 (2014).
[8] Abid, N., Shahzad, M., Malik, M. I., et al., "UCL: Unsupervised Curriculum Learning for water body classification from remote sensing imagery," International Journal of Applied Earth Observation and Geoinformation 105, 102568 (2021).
[9] Chen, Y., Fan, R., Yang, X., et al., "Extraction of urban water bodies from high-resolution remote-sensing imagery using deep learning," Water 10(5), 585 (2018).
[10] Zhang, T., Qin, C., Li, W., et al., "Water body extraction of the Weihe River Basin based on MF-SegFormer applied to Landsat8 OLI data," Remote Sensing 15(19), 4697 (2023).
[11] Long, J., Shelhamer, E. and Darrell, T., "Fully convolutional networks for semantic segmentation," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 3431-3440 (2015).
[12] Ronneberger, O., Fischer, P. and Brox, T., "U-Net: Convolutional networks for biomedical image segmentation," Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent., 234-241 (2015).
[13] Chen, J., et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv arXiv:2102.04306, (2021).
[14] Cao, H., et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," Proc. Eur. Conf. Comput. Vis., 205-218 (2022).
[15] Ma, X., Zhang, X., Pun, M. O., et al., "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," IEEE Transactions on Geoscience and Remote Sensing, (2024).
[16] Hong, D., et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," IEEE Trans. Geosci. Remote Sens. 59(5), 4340-4354 (2020).
[17] Roy, A. G., Navab, N. and Wachinger, C., "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," Medical Image Computing and Computer Assisted Intervention-MICCAI 2018: 21st International Conference, Granada, Spain, 421-429 (2018).
[18] Zhao, H., Shi, J., Qi, X., et al., "Pyramid scene parsing network," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2881-2890 (2017).