Multimodal model-based water body extraction algorithm on SAR imagery

Lin Y. [1], He L.[1*], Song Y.F.[1] and Zhong D.Q.[1]

[1]College of Surveying and Geo-informatics, Tongji University, Shanghai 200092, China

*helin19950830@gmail.com (*Corresponding author's email only)

Abstract**:** *The application of remote sensing technology for the extraction of inland water bodies has proven to be an effective method for large-scale water body investigations. Synthetic Aperture Radar (SAR) images offer the advantage of all-weather observation, unlike multispectral remote sensing images, which suffer from data quality degradation due to cloud cover. However, the efficient and accurate extraction of water bodies from SAR images, which is characterized by limited spectral information, remains a significant challenge. This article proposes an image segmentation algorithm based on a multimodal deep neural network model for extracting inland water bodies from SAR images. Initially, the method achieves feature alignment of SAR images and text based on the Transformer-based multimodal network. Subsequently, the image features of SAR images are encoded through a linear classifier applied to the scene classification task. Finally, an image decoder, a designed Convolutional Neural Network (CNN) structure, generates the results for the inland water body segmentation task. Experimental results using Sentinel-1 satellite data demonstrate that the multimodal feature encoder can effectively align SAR images with text. More than 90% precision and recall for the defined scene classification task can be achieved by the SAR image-text model. The proposed segmentation algorithm has also been validated for the extraction of inland water bodies.*

Keywords: SAR image, multimodal model, segmentation, inland water extraction

Introduction

Water body extraction is a crucial aspect of remote sensing applications, particularly in fields such as flood disaster monitoring, wetland conservation, and water resource management. While traditional optical remote sensing imagery can provide clear water body information under favorable weather conditions, its effectiveness significantly reduced in cloudy or rainy environments. In contrast, Synthetic Aperture Radar (SAR), as an active remote sensing technology (Franceschetti & Lanari, 2018), utilizes radar signals in the microwave spectrum to observe the Earth's surface. Unlike optical remote sensing, SAR provides reliable observational capabilities in all weather conditions and at any time of day, as active microwaves can penetrate clouds and fogs, rendering it impervious to variations in weather and lighting conditions. It makes SAR particularly valuable in extreme weather scenarios (Chen, Lv, Li, Qang, & Wang, 2020). SAR's pronounced reflective properties over water bodies enable it to effectively delineate the boundary between water and land, and it performs exceptionally well in complex environments. Consequently, SAR images are widely employed in tasks such as flood disaster monitoring (Iervolino, Guida, Iodice, et al., 2014), wetland management (Baghdadi, Bernier, Gauthier, et al., 2001), river extraction (Gasnier, Denis, Fjørtoft, Liège, & Tupin, 2021), and coastline extraction (Ding, Nunziata, Li, & Migliaccio, 2015). Nevertheless, the use of SAR imagery for water body extraction presents several challenges. Notably, speckle noise within the imagery can obscure water boundaries, reducing extraction accuracy. Furthermore, the variability in scattering characteristics among different surface types might lead to misclassification issues. To address these challenges, advanced algorithmic techniques, including machine learning and deep learning, have been increasingly applied to enhance water body extraction methods based on SAR imagery.

This paper presents a novel method that integrates Transformer-based multimodal model and Convolutional Neural Network (CNN) architectures to extract water body from SAR image segmentation tasks. Initially, the method aligns features from SAR images with those in descriptive text using a Transformer-based multimodal network designed for text-image integration. Subsequently, a linear classifier is employed to train the image features derived from SAR images for scene classification. Finally, the method utilizes a CNN-based image decoder to produce segmentation results for inland water bodies. This approach aims to leverage the strengths of both Transformer and CNN frameworks to explore the potential of multimodal models in water body extraction on SAR imagery.

Literature Review

Water body extraction methods from SAR imagery have progressed from traditional techniques—including threshold-based segmentation, edge detection, and statistical and mathematical models—to more advanced machine learning algorithms. Threshold-based methods for water body extraction (An, Niu, Li, et al., 2010) rely on setting one or more thresholds to differentiate water from non-water regions. Commonly used thresholding techniques include the Otsu method, maximum entropy method, and backscatter thresholding. Additionally, edge detection algorithms such as the Canny and Sobel algorithms can exploit reflectance differences between water and land in SAR images to delineate water-land boundaries, thereby facilitating water body extraction (Marghany & Hobma, 2000). Statistical and mathematical models, such as the Gaussian Mixture Model (GMM), have also proven effective in this context (Hou, Tang, Jiao, et al., 2009). Moreover, traditional machine learning approaches, such as Support Vector Machines (SVM), have been applied to SAR-based water body extraction research (Lv, Yu, & Yu, 2010). However, both traditional and machine learning methods have limitations in robustness. Specifically, SAR images often contain significant coherent speckle noise, making it challenging to develop mapping models that adapt effectively to all pixels. This issue constrains the accuracy and reliability of water body extraction outcomes.

In the context of SAR image water body extraction, deep learning methods use multi-layer encoding to extract intricate features from images. These methods learn shared weights within the model to develop mapping functions that accommodate all pixels, thereby mitigating noise interference and enhancing algorithmic robustness. Convolutional Neural Networks (CNNs) have already been employed in SAR-based water body extraction tasks (Wang, Wang, Wang, et al., 2022), with advancements based on architectures like U-Net that address the specific characteristics of SAR images (Bai, Wu, Yang, et al., 2021). Such advancements include the integration of residual networks, spatial pyramid pooling, additional skip connections, and the development of more sophisticated loss functions. Furthermore, research has also investigated SAR image water body extraction algorithms utilizing Transformer encoder modules (Zhou, Yang, Ma, et al., 2022). Evidence suggests that the adaptive mapping relationships established by neural network architectures—independent of specific algorithms or assumptions—can surpass the performance of traditional methods in SAR image water body extraction (Guo, Wu, Huang, et al., 2022).

Deep learning methods have demonstrated significant promise for water body extraction in SAR imagery; however, several challenges persist. A major issue is that the current volume of SAR image data is inadequate to meet the extensive sample requirements necessary for training neural networks, particularly given the limited availability of publicly accessible datasets related to water body extraction (Bonafilia, Tellman, Anderson, et al., 2020). Consequently, developing more efficient feature extraction models and exploring techniques for multi-source data fusion to mitigate computational resource demands and data constraints represent crucial research avenues. Currently, most approaches predominantly focus on utilizing image information, such as the fusion of SAR and optical imagery (Zhang, Lin, Wang, et al., 2018), for water body extraction. Nevertheless, within the framework of deep neural networks, models incorporating multimodal data (Radford, Kim, Hallacy, et al., 2021) have already achieved promising results in classification tasks. Investigating whether the integration of SAR images with other multimodal data can enhance water body extraction performance remains a valuable area for further research.

Methodology

The subsequent sections will describe the proposed model in terms of two components: the Transformer-based multimodal encoder and the CNN-based decoder.

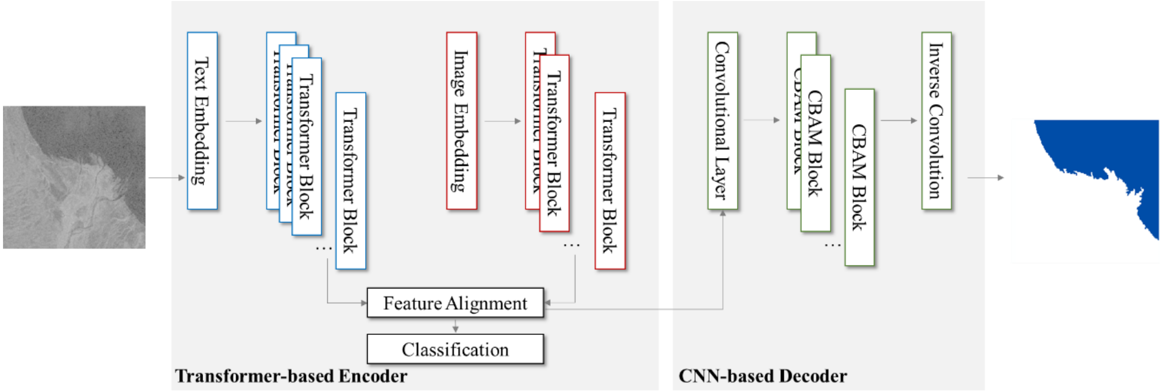

Figure 1: The proposed model structure for water body extraction from SAR images.

a. Transformer-based multimodal encoder:

The proposed methodology based on image-text multimodal models establishes a framework (Li, Selvaraju, Gotmare, et al., 2021) for processing SAR data. This framework utilizes prompt text to indicate the presence of water bodies and land within SAR images. SAR images are categorized into distinct classes, each representing consistent information about water bodies and land, such that all textual descriptions associated with data in a given scene category remain uniform. For instance, SAR images containing both water bodies and land boundaries are described by the prompt text: "The image contains a boundary line between water bodies and land." In accordance with the Bidirectional Encoder Representations from Transformers text encoding method (BERT) (Devlin, 2018), textual data is digitized through tokenization, which converts written words into tokenized representations, which are then mapped into vector space. The text feature encoder comprises 12 layers of Transformer blocks. Similarly, SAR images are divided into patches using the Vision Transformer (Dosovitskiy, 2020) and mapped into vector space, with feature extraction performed by an image encoder that also consists of 12 Transformer layers.

The model leverages pre-trained parameters from BERT without further training. Instead, the focus is on the training of the image encoder. An adapter situated after the encoder is incorporated, including feature fine-tuning layers for both modalities and a classifier for scene classification tasks. During multimodal model training, image features are aligned with text encoding features through the fine-tuning layers. The loss function of the model is comprised of two components: the alignment loss between text and images and the classification loss associated with image classification. The classification task is defined based on the presence or absence of water bodies in the image. The image encoding features, optimized for this classification task, will improve the pre-trained parameters for SAR image water body extraction.

b. CNN-based decoder:

In the decoder section, the paper utilizes a convolution-based modular architecture. This structure incorporates modules consisting of two units based on the Convolutional Block Attention Module (CBAM) (Woo, Park, Lee, et al., 2018), with residual connections employed to concatenate these modules. CBAM, a lightweight attention mechanism, enhances the representational capacity of convolutional neural networks by integrating attention mechanisms across both spatial and channel dimensions. By separately processing and integrating input feature maps along these dimensions, CBAM facilitates a more comprehensive representation and decoding of SAR image features.

Residual connections (He, Zhang, Ren, et al., 2016) introduce skip connections that ensure effective gradient propagation from deeper to shallower layers within the deep network, thus mitigating the vanishing gradient problem. In the convolutional decoder, the incorporation of residual connections facilitates skip connections for the CBAM combination modules, preserving the feature representations of the shallower decoder layers and augmenting the network's capacity to represent features. This enhancement enables the model to capture complex patterns more effectively. Furthermore, the network's ability to approximate identity mappings is improved, which accelerates the convergence rate during training.

The process of reconstructing water body extraction results from encoded features to image dimensions predominantly relies on deconvolution operations for upsampling. During deconvolution, the network learns specific weights to refine the details of the output feature maps, progressively restoring high-level encoded features to their original image resolution. Consequently, the model adeptly converts abstract, low-resolution encoded features into high-resolution images, thereby achieving accurate reconstruction of SAR image water body extraction results.

Results and Discussion

The algorithm was evaluated using an experimental dataset derived from Sentinel-1 single-band SAR images including both water bodies and land. This dataset comprises 6,641 images, with water body reference samples that were manually annotated. The dataset was split into training and test sets with a ratio of 7:3. The experiments and subsequent analyses were conducted in accordance with the image encoding structure detailed below.

a. Multimodal coding features:

The optimization of visual encoder parameters is is achieved by solving the defined classification problem. Specifically, the training process utilizes labeled SAR image data, with model parameters being iteratively updated to minimize both classification loss and alignment loss. Employing the backpropagation algorithm, the visual encoder igradually adjusts its internal weights to improve its ability to extract features that differentiate between water bodies and land in SAR images.

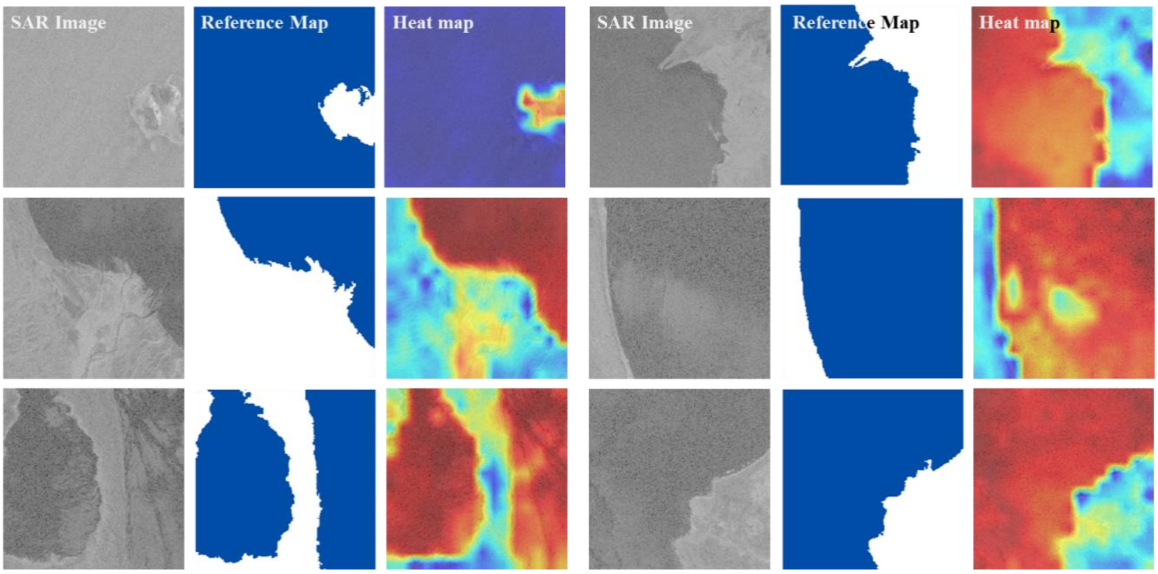

Figure 2: Visualization of the attention of multimodal model for SAR images.

To assess the effectiveness of the multimodal model in extracting features from SAR image data, heat maps were employed as a visualization tool. Heat maps, as a data visualization technique, offer an intuitive representation of the model's focus areas and feature intensity distribution across various spatial locations. This approach facilitates an analysis of the model's feature extraction performance by illustrating the regions to which the multimodal model pays the most attention during SAR image processing. By mapping high-dimensional feature data onto a two-dimensional space, heat maps visually depict the model's response intensity at different locations within the input image. These intensities reveal the spatial distribution of features learned by the model, highlighting which regions the Transformer-based model assigns higher weights during specific tasks.

The results indicate that, following training with the multimodal network model, the visual encoder for SAR images demonstrated satisfactory feature extraction performance. Compared to those in the original images, the areas of focus within the SAR images revealed significant distinctions between water and land, with clear delineation of water bodies and land boundaries. The model generally exhibited a greater focus on water bodies, although in certain instances (e.g., the first image), the model's attention was directed towards land. This varied attention distribution underscores the Transformer-based multimodal model’s capacity to differentiate between water and land in SAR images.

Table 1 presents metrics for training and validation accuracy, including recall and precision, for the three defined classification categories. The data show that the model performs robustly across all categories. It achieves the highest recall and precision in the category labeled "No land in the image." The performance for the categories "No water in the image" and "Water and land are in the image" is similar. Overall, with a classification accuracy nearing 90%, the multimodal model exhibits a low false positive rate.

Table 1: Validation accuracies of the multimodal encoder for classification.

	No water in the image	**No land in the image**	**Water and land are in the image**
Recall	0.8790	0.9187	0.8870
Precision	0.8955	0.9536	0.8955

b. Feature decoding results:

The Mean Intersection over Union (MIOU) metric was employed to evaluate the performance of the segmentation decoder for land and water bodies. After 1,000 training iterations, the highest training segmentation accuracy achieved was 0.9697, while the highest testing accuracy was 0.8485, which demonstrates the decoder's effective performance in distinguishing water and land boundaries. The Figure 3 illustrates the segmentation results as the model's loss converges during decoder training. Throughout the iterative training process, the decoder's performance showed progressive improvement, enhancing its ability to accurately capture the subtle features of water and land boundaries. Continuous optimization of network parameters enabled the decoder to refine its segmentation capabilities for various regions within the image.

Initially, the decoder's segmentation results exhibited considerable boundary blurring and smoothness, attributable to the network's insufficient learning of the specific features of water bodies and land. However, as training advanced, particularly after 100 epochs, the encoder adjusted its internal parameters through extensive data input and backpropagation. This adjustment led to a gradual reduction in discrepancies between predicted results and reference samples. The convergence observed during training is reflected in the increasingly precise and refined segmentation boundaries. As training progressed, the model's segmentation results became more stable, with a marked reduction in boundary prediction errors, indicating the decoder's improved ability to capture the nuanced differences between water bodies and land.

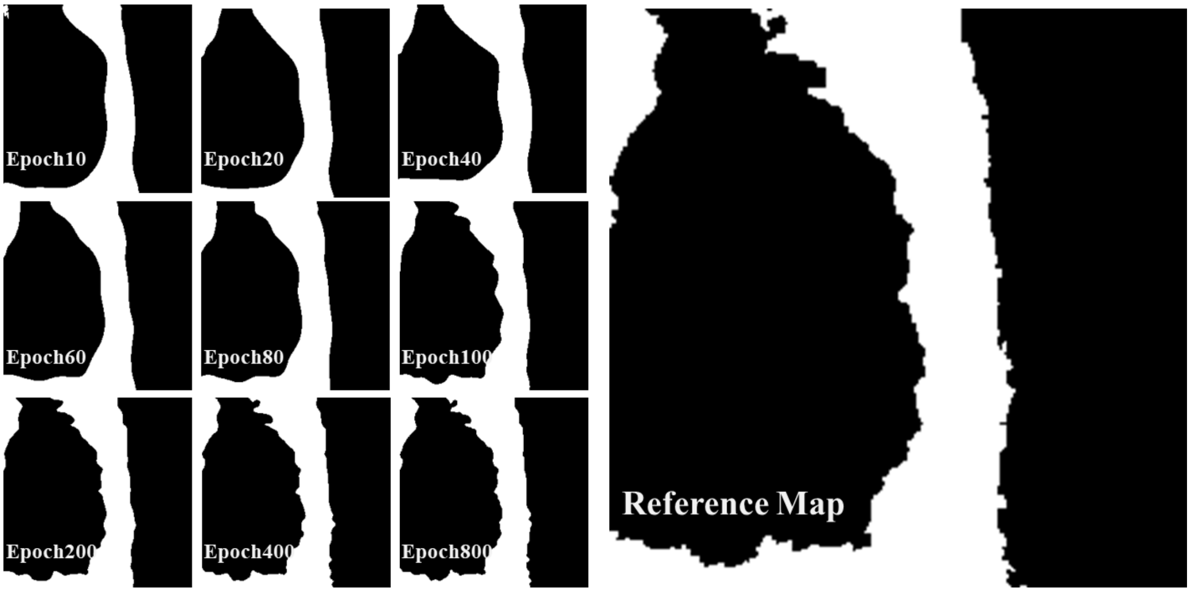

Figure 3: Visualization of the convergence of segmentation predictions during iterations.

Conclusion and Recommendation

In summary, we designed and trained a multimodal encoder for pre-training and a decoder for generating segmentation results to address the challenge of water extraction in SAR images. During the pre-training phase, we employed a Transformer-based multimodal encoder that integrates multimodal data, allowing the model to leverage complementary information from different data sources. This integration enhances the model's capacity to distinguish between water bodies and land from SAR images. The multimodal encoder effectively extracts features from SAR images by minimizing classification and alignment losses during training.

For the segmentation task, we developed a convolutional neural network-based decoder proficient in handling the spatial detail features of SAR images. Through convolutional operations, the decoder progressively reconstructs high-dimensional features extracted by the multimodal encoder, producing high-precision segmentation maps of water bodies and land. The joint training of the multimodal encoder and decoder ensures a cohesive synergy in both feature extraction and segmentation result generation.

Future research may focus on incorporating more sophisticated model architectures or utilizing richer training datasets to further enhance the model's segmentation accuracy and robustness. Specifically, optimizing the model's generalization capability for complex scenes and diverse data sets presents a significant and promising area for further investigation.

References

An, C. J., Niu, Z. D., Li, Z. J., et al. (2010). Otsu threshold comparison and SAR water segmentation result analysis. Journal of Electronics & Information Technology, 32(9), 2215-2219. https://doi.org/10.1007/s11767-010-0272-5

Baghdadi, N., Bernier, M., Gauthier, R., et al. (2001). Evaluation of C-band SAR data for wetlands mapping. International Journal of Remote Sensing, 22(1), 71-88. https://doi.org/10.1080/01431160120087

Bai, Y., Wu, W., Yang, Z., et al. (2021). Enhancement of detecting permanent water and temporary water in flood disasters by fusing Sentinel-1 and Sentinel-2 imagery using deep learning algorithms: Demonstration of Sen1Floods11 benchmark datasets. Remote Sensing, 13(11), 2220. https://doi.org/10.3390/rs13112220

Bonafilia, D., Tellman, B., Anderson, T., et al. (2020). Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 210-211). https://doi.org/10.1109/CVPRW50498.2020.00028

Chen, J., Lv, J., Li, N., Qang, Q., & Wang, J. (2020). External groundwater alleviates the degradation of closed lakes in semi-arid regions of China. Remote Sensing, 12(1), 45. https://doi.org/10.3390/rs12010045

Devlin, J. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

Ding, X., Nunziata, F., Li, X., & Migliaccio, M. (2015). Performance analysis and validation of waterline extraction approaches using single- and dual-polarimetric SAR data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8, 1019-1027. https://doi.org/10.1109/JSTARS.2015.2390949

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://arxiv.org/abs/2010.11929

Franceschetti, G., & Lanari, R. (2018). Synthetic Aperture Radar Processing. CRC Press.

Gasnier, N., Denis, L., Fjørtoft, R., Liège, F., & Tupin, F. (2021). Narrow river extraction from SAR images using exogenous information. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 5720-5734. https://doi.org/10.1109/JSTARS.2021.3070612

Guo, Z., Wu, L., Huang, Y., et al. (2022). Water-body segmentation for SAR images: Past, current, and future. Remote Sensing, 14(7), 1752. https://doi.org/10.3390/rs14071752

He, K., Zhang, X., Ren, S., et al. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

Hou, B., Tang, X., Jiao, L., et al. (2009). SAR image retrieval based on Gaussian mixture model classification. In 2nd Asian-Pacific Conference on Synthetic Aperture Radar (pp. 796-799). IEEE. https://doi.org/10.1109/APSAR.2009.5279194

Iervolino, P., Guida, R., Iodice, A., et al. (2014). Flooding water depth estimation with high-resolution SAR. IEEE Transactions on Geoscience and Remote Sensing, 53(5), 2295-2307. https://doi.org/10.1109/TGRS.2013.2283715

Li, J., Selvaraju, R., Gotmare, A., et al. (2021). Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems, 34, 9694-9705. https://doi.org/10.5555/3495724.3495725

Lv, W., Yu, Q., & Yu, W. (2010). Water extraction in SAR images using GLCM and support vector machine. In IEEE 10th International Conference on Signal Processing Proceedings (pp. 796-799). https://doi.org/10.1109/ICSP.2010.5485178

Marghany, M., & Hobma, T. W. (2000). Operationalization of SAR polarized data for assessment of coastal erosion. International Archives of Photogrammetry and Remote Sensing, 33(B1; PART 1), 201-208. https://doi.org/10.5194/isprs-archives-XXXIII-B1-201-2000

Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (pp. 8748-8763). https://proceedings.mlr.press/v139/radford21a.html

Wang, J., Wang, S., Wang, F., et al. (2022). FWENet: A deep convolutional neural network for flood water body extraction based on SAR images. International Journal of Digital Earth, 15(1), 345-361. https://doi.org/10.1080/17538947.2021.1995002

Woo, S., Park, J., Lee, J. Y., et al. (2018). CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (pp. 3-19). https://doi.org/10.1007/978-3-030-01234-2_2

Zhang, M., Lin, H., Wang, G., et al. (2018). Mapping paddy rice using a convolutional neural network (CNN) with Landsat 8 datasets in the Dongting Lake area, China. Remote Sensing, 10(11), 1840. https://doi.org/10.3390/rs10111840

Zhou, Y., Yang, K., Ma, F., et al. (2022). Water-land segmentation via structure-aware CNN-transformer network on large-scale SAR data. IEEE Sensors Journal, 23(2), 1408-1422. https://doi.org/10.1109/JSEN.2022.3140020