Water-land segmentation based on the U-Net model in optical remote sensing images

Michał Gruba, Marcin Ciecholewski
Faculty of Electronics, Telecommunication and Informatics
Gdańsk University of Technology, Gdańsk, Poland
michal.gruba00@gmail.com, marcin.ciecholewski@pg.edu.pl

***Abstract*—The separation of an optical remote sensing image into water and land areas is a complex yet essential process for the extraction of coastlines and subsequent detection of objects. The accurate delineation of water-land boundaries based on optical remote sensing imagery represents a significant challenge to the conventional segmentation techniques. The advancement of deep learning has resulted in the dominance of convolutional neural networks (CNNs) in semantic segmentation, largely due to their robust local information extraction abilities. This paper presents the results of three models based on the U-Net architecture, for the segmentation of water and land areas based on optical satellite imagery using a publicly available dataset, namely the Sentinel-2 NOAA Water Edges Dataset (SNOWED). The paper proposes the optimisation of models using various loss functions, including cross-entropy, Sørensen-Dice, local regularisation, structured edge information and their the appropriate configurations. During the course of the research, the use of the Intersection over Union (IoU) metric yielded highly competitive results, with the best result obtained being 96.05%. The optimal model was developed using a loss function comprising two components: local regularisation and structured edge information.**

***Index Terms*—water–land segmentation, optical image, loss function, deep learning, U-Net.**

I. Introduction

In the context of remote sensing imagery, the objective of water–land segmentation is to accurately delineate the boundaries between the nearshore region and the land. The segmentation result is of paramount importance for the subsequent extraction of the coastline [1] and the detection of ships [2].

The utilisation of remote sensing approaches enhanced by deep learning is being developed for the monitoring of water bodies, due to the availability of publicly accessible data from programmes such as Copernicus [3] and Landsat [4]. The literature proposes methods that employ deep convolutional neural networks (DCNNs) for the semantic segmentation of satellite imagery, with the objective of identifying surface water regions and delineating water bodies [5]–[7]. The papers [5], [6] presented high-level results of the segmentations obtained. However, these studies were carried out only on private and unreleased datasets, which greatly limits the possibility of verifying the results obtained.

While DCNN has been shown to enhance water body extraction, there are still limitations to its capabilities. Convolutions have a restricted receptive field and are unable to model global information. Convolution gathers data from neighbouring pixels, which limits the accuracy of semantic segmentation by neglecting inter-pixel relationships. The integration of global information allows for more accurate pixel classification. Some researchers have proposed the incorporation of attention mechanisms with CNNs to address the limitations of these networks [7], [8]. Attention mechanisms assist in the weighting of salient features and have been employed to enhance water body extraction. A detailed paper on the development of image segmentation using DCNNs can be found in [8].

The transformer is a deep learning architecture [9]. It models the relationship between input tokens and deals with long-range dependencies. Unlike CNN, the Transformer processes one-dimensional sequence features from two-dimensional image features. The standard Transformer structure is composed of layer normalisation, multi-head self-attention, a multilayer perceptron, and skip connections. A survey of the extant literature reveals that certain studies have yielded satisfactory results with Transformer, although it should be noted that these are mostly based on extensive pre-training [10]. A comprehensive survey and evaluation of various segmentation methods employing Transformers can be found in [11].

In [12], the authors embedded MixFormer [13] into the U-Net model [14] and proposed a hybrid MixFormer architecture, designated as MU-Net. The MU-Net model [12] employs a combination of CNN and MixFormer to facilitate the capture of both local and global contextual information present in images. MixFormer is a Transformer structure that has been modified to enhance its capacity to capture global contextual information.

A recently annotated SNOWED dataset [15] of optical satellite imagery has been released. This dataset comprises 4334 images and it appears to be a valuable resource for training DCNN models to perform land-water segmentation. The paper [16] presents the results of a study to segment the river channel of the Po in Northern Italy using the SNOWED dataset, obtaining a metric of IoU=96.7%, which is the best and, at the time of writing, the only result based on the literature. The authors of the paper [16] used the basic U-Net model [14] in their research.

In this research, an attempt was made to obtain similar IoU metric results, which were presented in the paper [16], based

on the SNOWED dataset. However, this was not possible using the basic U-Net model [14], and cross-entropy as a loss function. Therefore, further research was carried out to obtain the best possible results, and for this purpose three distinct state-of-the-art approaches [5], [6], [12] were used. It should be emphasised that the most desirable in practice are universal classifiers that enable accurate segmentation of water-land areas for a wide class of cases and conditions at the water-land interface, e.g. taking into account sandy, stony, rocky coasts as well as shallows and visible waves on the water, river mouths and the presence of small, closed water reservoirs in the vicinity of the coast. Consequently, the present research concentrated on considering the most extensive possible range of cases, including both visible man-made infrastructure and non-urbanised areas. It is noteworthy that the SNOWED dataset contains images for the aforementioned cases and conditions that occur in water-land areas.

In summary, the main contributions of our research are as follows: (1) An optimisation of the solutions presented in the papers [5], [6], [12] on the SNOWED dataset was performed, taking into account various loss functions including cross-entropy, Sørensen-Dice, local regularisation, structured edge information. The loss functions employed facilitated the acquisition of both global and local information concerning the image pixels, a factor that is of great significance in achieving high segmentation accuracy. The optimal results were obtained by employing the sum of two loss functions in the following configurations: (a) cross-entropy and Sørensen-Dice, (b) local regularisation and structured edge information. (2) During the course of the research, the use of the IoU metric yielded highly competitive results, with the best result obtained being 96.05%. The most optimal results were achieved using the MU-Net model [12], which employed a sum of two loss functions, namely local regularisation and structured edge information. (3) Based on the results obtained from the evaluation metrics used, an investigation was conducted into the possibility of optimising the number of channels in the subsequent input layers of the models used. The study concluded that models with a simplified architecture can achieve results comparable to the best solutions.

II. Materials and Methods

A. Data

The SNOWED dataset [15], which was utilised in the present study, comprises 256 $\times$ 256 pixel resolution images of land and coastal areas. These images were obtained by the European Space Agency's (ESA) Sentinel-2 satellite. The dataset also includes ground-truth masks, in which water and land areas are assigned binary values. The SNOWED dataset contains 4334 images in total. During the course of the study, 2730 images were utilised as a training set, 1171 and 433 images were employed to validate and test the models, respectively.

Fig. 1 illustrates examples of satellite images from the SNOWED dataset, accompanied by their corresponding ground-truth masks. The colour yellow is employed to indicate water areas, while the colour purple is used for land areas.

B. Deep learning models used

Three state-of-the-art models were employed in the course of this research, namely: the Structured Edge Network for Sea-Land Segmentation, (SeNet) [5]; the Deep Fully Convolutional Network for Pixel-level Sea-Land Segmentation (DeepUNet) [6]; and the Embedding MixFormer into Unet to Extract Water Bodies from Remote Sensing Images (MU-Net) [12].

Fig. 1. Example satellite images from the SNOWED dataset [15]. The first and second columns represent the source image and the ground-truth mask, respectively.

1) SeNet model: The SeNet model [5] is based on the U-Net architecture [14], which is a widely used approach in the field of computer vision. The U-Net is a deep learning architecture that was originally developed for the purpose of semantic segmentation of medical images. The U-Net model comprises two paths, which collectively resemble the letter "U" in a diagrammatic representation. These are the contraction and expansion paths, which respectively correspond to the encoder and decoder components. The SeNet model introduces branching at the final layer, resulting in a network that is capable of performing two distinct tasks. One output of the network is responsible for image segmentation, while the other output is focused on edge detection. Consequently, the SeNet is capable of simultaneously performing sea-land segmentation and edge detection. Furthermore, the SeNet promotes the local regularised loss, which serves to reduce misclassification.

2) DeepUNet: The DeepUNet model [6] also draws its inspiration from the U-Net architecture [14]. The authors extended the U-Net model by incorporating DownBlocks (comprising two convolution layers concatenated through a ReLU layer) into the contracting path and UpBlocks (comprising two convolutional layers followed by an upsampling layer) into the

expansion path. The upsampled outputs are transformed into high-resolution feature maps, which are then combined by a subsequent convolutional layer. In the DownBlock layers, the inputs to a convolutional layer are combined with the outputs of the layer using an addition operation. The same strategy is employed in the UpBlock layers.

3) MU-Net: The MU-Net model [12] is a hybrid MixFormer architecture [13], in which the MixFormer module is embedded within the U-Net to enable the capture of both local and global contextual information present in images. To obtain feature maps at different scales, the original image was downsampled using a max-pooling operation. The extraction of local information from deep features was first achieved using convolutional layers, with the aim of accurately identifying water features in complex backgrounds. Global contextual information was then modelled using a MixFormer block to extract deeper semantic features of water bodies. The features generated by the encoder were then refined by the Attention Mechanism Module (AMM). The AMM suppresses non-water features and noise by weighting features related to water bodies. Finally, bilinear interpolation and skip connection were utilised to recover the resolution and detail information of the image, thereby producing the final water body extraction results.

C. Loss functions

Cross-entropy (CE) loss [17], a widely employed loss function, involves the comparison of the predicted class with the target class. This is achieved by examining each pixel in the image. However, preliminary experiments demonstrated that the CE loss function imposes limitations on learning potential when features are complex, small or linear. Consequently, subsequent experiments employed several additional loss functions. One such loss function is the Sørensen-Dice (SD) loss [17], which is based on a metric for model evaluation known as the Sørensen-Dice coefficient.

$$L_{CE} = -\sum_{i=1}^{t} y_i log(p_i) \tag{1}$$

$$L_{SD} = 1 - \frac{\sum_{i=1}^{t} y_i p_i + \epsilon}{\sum_{i=1}^{t} y_i + p_i + \epsilon} \tag{2}$$

where t is the total number of pixels, y_i is the ground-truth value of the pixel, and p_i is the predicted probability value of the pixel returned by the model, and ϵ is a small value equal to 10^{-5} to ensure that division by zero does not occur. It is possible to define an alternative loss function on the basis of (1) and (2), expressed by their sum.

$$L_{CE+SD} = \alpha \cdot L_{CE} + \beta \cdot L_{SD} \tag{3}$$

Preliminary studies have confirmed that the use of the sum of two different loss functions, expressed by (3), can produce better results than the use of single loss functions based on (1) or (2). We set $\alpha = 1.0$, $\beta = 0.7$ in our experiments.

It is evident that the topography of land regions is frequently characterised by intricate texture and intensity distribution. This is attributed to the interplay of sunlight, altitude, and the presence of objects on the ground. Furthermore, the presence of waves in maritime regions can impede the attainment of optimal segmentation results, thereby exacerbating the challenges associated with accurate delineation. The softmax loss employed in semantic segmentation is confined to pixel-wise loss, disregarding the interrelationship between adjacent pixels that exhibit analogous colour values. In order to enhance the utilisation of the local relationship between adjacent pixels, the local regularised loss was adopted [5], [18]. The loss function can be expressed as follows:

$$L_{Seg} = \frac{1}{N}\sum_{i=1}^{N}\Big\{ -logp_{l_i,i} + \frac{\lambda}{2}\sum_{j\in Nb(i)}[(p_{0,i}-p_{0,j})^2 + (p_{1,i}-p_{1,j})^2]e^{\frac{-(x_i-x_j)^2}{\sigma}}\Big\} \tag{4}$$

where l_i denotes the ground-truth label of pixel i, with the value of l_i set to 1 for land pixels and 0 for sea pixels; $p_{l_i,i}$ represents the probability of the ground-truth label assigned to i; N is the total number of points in the batch, while x_i is the colour value of i; $Nb(i)$ refers to the eight neighbours of i. The variance term, denoted by σ, is calculated as the average squared distance between all neighbouring pixels in each image, i.e. $\sigma = \langle (x_i - x_j)^2 \rangle$. The second term in (4) forces neighbouring pixels with similar colour values to have similar label probabilities. The value of λ is important for the segmentation results. A large value of λ leads to under-segmentation and reduces the edge accuracy. Satisfactory results on the validation set are obtained when λ is in the range $[10, 50]$.

In the context of certain sophisticated structures, such as wharfs and ships, the segmentation network frequently encounters difficulties in achieving precise results in the vicinity of the shoreline and terrestrial boundaries. Moreover, in the absence of a discrete point between adjacent sea and land edge pixels, it is necessary to consider two sets of edges: $E(L)$ is the set of edges in the land area, while $E(S)$ denotes the set of edges in the water area. The edge-based loss function is defined as follows:

$$\begin{aligned} L_{Edge} = -\frac{1}{N}\Big\{ &\sum_{i\in E(L)} \frac{\sum_{j=1}^{8} SN_i(j)[logp_{1,i} + log(1-p_{1,i(j)})]}{\sum_{j=1}^{8} SN_i(j)} \\ &+ \sum_{i\in E(S)} \frac{\sum_{j=1}^{8} SN_i(j)[log(1-p_{1,i}) + logp_{1,i(j)}]}{\sum_{j=1}^{8} SN_i(j)} \\ &+ \sum_{i\notin E(L)\cup E(S)} (log(1 - \sum_{j=1}^{8} \frac{\mid p_{1,i} - p_{1,i(j)} \mid}{8})\Big\} \end{aligned} \tag{5}$$

The classification of a pixel as an edge is determined by the presence of at least one of its eight neighbours exhibiting a distinct class. The notation $SN_i(j)$ employed in (5) denotes a function that returns $SN_i(j) = 1$ when i is an edge and its j-th

neighbour belongs to a different class, otherwise $SN_i(j) = 0$. The symbol $p_{1,i}$ denotes the probability that pixel i belongs to class 1, while $p_{1,i(j)}$ denotes the probability that the pixel j-th neighbour belongs to this class. Given that the probabilities of a given pixel belonging to each class add up to unity (i.e. $p_0 + p_1 = 1$), it is possible to use only one of them. In this case, p_1 is used.

Finally, the loss function is expressed as the sum of (4) and (5)

$$L_{Seg+Edge} = L_{Seg} + L_{Edge} \tag{6}$$

The integration of segmentation and edge detection facilitates the network in recognising distinguishable edge features.

If a different loss function is used in a particular model, the only modification required is to add an additional output layer, or two additional output layers if the sum of two loss functions is used, as in (3) and (6). This can be implemented for different models, i.e.: SeNet [5], DeepUNet [6] and MU-Net [12].

III. Experimental results and discussion

The SNOWED dataset was utilised in the course of the experiments conducted. Details pertaining to the division of this dataset into learning, validation and test subsets are presented in Section II-A. The experiments investigated the effect of utilising the loss function given by (3) and (6). The SeNet [5], DeepUNet [6] and MU-Net [12] models were tested in both basic configurations and in configurations with a half-reduced number of filters in subsequent convolution layers, and thus also with a half-reduced number of input channels in subsequent layers. These configurations are presented in Table I together with the results obtained for the SNOWED test set.

The quantitative evaluation of image segmentation is conducted using standard metrics, including F1-score and the IoU. The IoU measure can be calculated for each of the classes, so in the case of the water and land segmentation, the appropriate labels used were IoU_{water}, for water areas, and IoU_{land} for land areas, in addition, the average of these two values, denoted IoU_{mean}, was included.

Considering the results in Table I, it can be observed that training the SeNet model on the SNOWED dataset using the loss function $L_{Seg+Edge}$ (6) produces extremely poor results, although using this model in combination with the loss function L_{CE+SD} (3) or using the loss function $L_{Seg+Edge}$ (6) with the DeepUNet model produces very good results. In general, the values of the IoU_{water} metric are higher than the IoU_{land} metric, indicating superior segmentation of water areas by the models utilised. The best results for each metric are obtained by the configuration numbered 10, which is the MU-Net model trained with the loss function $L_{Seg+Edge}$ (6). This configuration achieves the highest values of the IoU_{mean} metric, which is 96.05%. Example results of the MU-Net model with configuration No. 10, together with source images and ground-truth images, are shown in Fig. 2.

It is evident from the results obtained after halving the number of filters utilised in the subsequent convolution layers that there is only a slight decrease of approximately 1% in both the F1-score and the IoU_{mean} metrics. However, for the SeNet model and the applied loss function $L_{Seg+Edge}$ (6), a substantially larger difference was observed, reaching up to 20% for the IoU metric and 8% for the F1-score metric. Moreover, it has been observed that for the DeepUNet and MU-Net models with configurations in which the number of filters was reduced and the loss function $L_{Seg+Edge}$ (6) was employed, almost identical metric results were obtained as for the basic configuration with the number of filters unchanged and the loss function L_{CE+SD} used (3). In light of the results obtained in Table I, it can be concluded that optimising the models to reduce their size is a worthwhile endeavour, since the results obtained are close to the maximum.

Experiments were performed using the PyTorch deep learning framework and the CUDA library on a PC with AMD Ryzen 7 5800@3.8 GHz, 16 GB RAM and NVIDIA RTX 3070ti graphics card, under Windows 11. Table II shows the values of the four hyperparameters, namely number of epochs and batch size, learning rate and optimiser used for the SeNet, DeepUnet and MU-Net models.

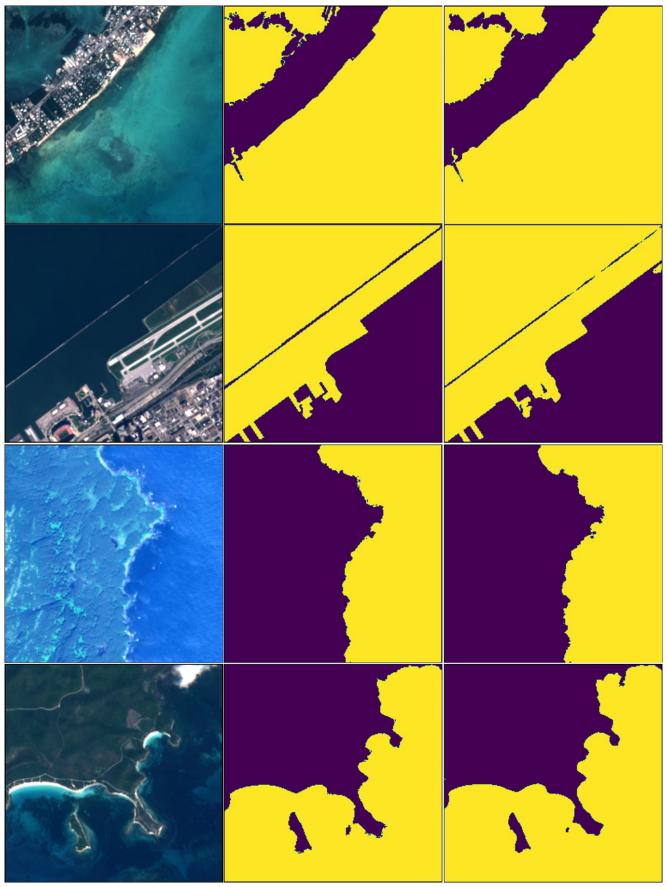

Fig. 2. Visualised segmentation results. The first and second columns show the source image and the ground-truth, respectively. Column 3 shows the segmentation results obtained using the MU-Net model [12] with configuration no. 10, based on Table I.

TABLE I

THE RESULTS OF THE MODELS THAT WERE TRAINED ACCORDING TO THE CONFIGURATIONS USED IN THE LOSS FUNCTION EXPERIMENTS, L_{CE+SD} (3) AND $L_{Seg+Edge}$ (6). THESE WERE OBTAINED USING THE SNOWED TEST SET. THE BEST RESULTS ARE SHOWN IN BOLD.

No.	Model	No. of channels in subsequent layer inputs	Loss function	F1-score	IoU$_{water}$	IoU$_{land}$	IoU$_{mean}$
1	SeNet	[4, 16, 32x2, 64x7, 32x2, 16x2]	$L_{Seg+Edge}$	73.72%	64.38%	12.56%	38.47%
2	SeNet	[4, 32, 64x2, 128x7, 64x2, 32x2]	$L_{Seg+Edge}$	81.74%	76.39%	40.2%	58.30%
3	SeNet	[4, 16, 32x2, 64x7, 32x2, 16x2]	L_{CE+SD}	95.69%	95.51%	90.82%	93.16%
4	SeNet	[4, 32, 64x2, 128x7, 64x2, 32x2]	L_{CE+SD}	96.69%	96.28%	92.61%	94.45%
5	DeepUNet	[4, (16, 32)x7, (32, 16)x7, 32]	$L_{Seg+Edge}$	96.03%	95.52%	90.82%	93.17%
6	DeepUNet	[4, (32, 64)x7, (64, 32)x7, 32]	$L_{Seg+Edge}$	96.71%	96.43%	92.66%	94.54%
7	DeepUNet	[4, (16, 32)x7, (32, 16)x7, 32]	L_{CE+SD}	95.74%	95.4%	91%	93.2%
8	DeepUNet	[4, (32, 64)x7, (64, 32)x7, 32]	L_{CE+SD}	96.55%	96.3%	92.63%	94.46%
9	MU-Net	[4, 32, 64, 128, 256, 512, 256, 128, 64, 32]	$L_{Seg+Edge}$	97.27%	96.98%	93.9%	95.44%
10	MU-Net	[4, 64, 128, 256, 512, 1024, 512, 256, 128, 64]	$L_{Seg+Edge}$	**97.71%**	**97.39%**	**94.71%**	**96.05%**
11	MU-Net	[4, 32, 64, 128, 256, 512, 256, 128, 64, 32]	L_{CE+SD}	96.29%	96.41%	92.65%	94.53%
12	MU-Net	[4, 64, 128, 256, 512, 1024, 512, 256, 128, 64]	L_{CE+SD}	97.4 %	97.01%	93.94%	95.52%

TABLE II

THE VALUES OF FOUR HYPERPARAMETERS, NAMELY THE NUMBER OF EPOCHS, BATCH SIZE, LEARNING RATE, LOSS FUNCTION AND OPTIMISER, ARE PRESENTED FOR THE SENET [5], DEEPUNET [6] AND MU-NET [12] MODELS.

Model	No. of epochs	Batch size	Learning rate	Optimiser
SeNet	1000	4	0.00001	Adam
DeepUNet	100	11	0.0001	Adam
MU-Net	100	10	0.001	Adam

IV. CONCLUSION

This paper presented the results of three models based on the U-Net architecture for segmenting water and land areas from optical satellite images using the publicly available SNOWED dataset. During the research, the possibility of using different loss functions was analysed in detail, and optimal results were obtained by using the sum of two loss functions in the following configurations: (a) cross-entropy and Sørensen-Dice, (b) local regularisation and structured edge information. The most optimal results were obtained with the MU-Net model using a sum of two loss functions, namely local regularisation and structured edge information. The F1-score and IoU$_{mean}$ metrics results for this model were 97.71% and 96.05%, respectively. The research also reached the interesting conclusion that half-reduced number of filters in subsequent convolution layers in the DeepUnet and MU-Net models allows results close to the maximum values. It is therefore recommended that further research be conducted in order to optimise the models by reducing their size even further, thus facilitating their storage on devices with limited memory. This could also potentially improve the generalisation of the models, leading to higher metric values in the test set and faster inference.

REFERENCES

[1] J. E. Pardo-Pascual and et al., "Automatic extraction of shorelines from Landsat TM and ETM+ multi-temporal images with subpixel precision. Remote Sensing of Environment," *Remote Sens. Environ.*, vol. 123, pp. 1-13, 2012.

[2] C. Ieracitano and et al., " An explainable embedded neural system for on-board ship detection from optical satellite imagery, " *Eng. Appl. Artif. Intell.*, vol. 133, PaperId 108517, 2024.

[3] Homepage-Copernicus. Available online: https://www.copernicus.eu/en

[4] Landsat Science. Available online: https://landsat.gsfc.nasa.gov/

[5] D. Cheng and et al., "SeNet: Structured Edge Network for sea-land Segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 14(2), pp. 247–251, 2017.

[6] R. Li and et al., "DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11(11), pp. 3954–3962, 2018.1.

[7] Y. Li and et al., "Water body classification from high-resolution optical remote sensing imagery: Achievements and perspectives," *ISPRS J. Photogramm. Remote Sens.*, vol. 187, pp. 306-327, 2022.

[8] F. Sultana and et al., " Evolution of image segmentation using deep convolutional neural network: A survey," *Knowledge-Based Syst.*, vol. 201, Art. no. 106062, 2020.

[9] C. Subakan and et al., "Attention Is All You Need In Speech Separation," *in ICASSP*, pp. 21-25, 2021.

[10] T. Lin and et al., "A survey of transformers,". *AI open*, vol. 3, pp. 111-132, 2022.

[11] H. Thisanke and et al., "Semantic segmentation using Vision Transformers: A survey," *Eng. Appl. Artif. Intell.* Engineering Applications of Artificial Intelligence vol. 126, Art. no. 106669, 2023.

[12] Y. Zhang and et al., "MU-Net: Embedding MixFormer into Unet to Extract Water Bodies from Remote Sensing Images," *Remote Sensing*, vol. 15(14), 3559, 2023.

[13] Q. Chen and et al., "MixFormer: Mixing Features across Windows and Dimensions," *in CVPR*, pp. 5239-5249, 2022.

[14] O. Ronneberger and et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation," *in MICCAI*, pp. 234–241, 2015.

[15] G. Andria and et al., "SNOWED: Automatically Constructed Dataset of Satellite Imagery for Water Edge Measurements," *Sensors*, vol. 23(9), 4491, 2023.

[16] M. Scarpetta and et al., "Use of the SNOWED Dataset for Sentinel-2 Remote Sensing of Water Bodies: The Case of the Po River," *Sensors*, vol. 24(17), 5827, 2024.

[17] F. Milletari and et al., "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *in 3DIMPVT*, pp. 565-571, 2016.

[18] R. El Jurdi and et al., "High-level prior-based loss functions for medical image segmentation: A survey," *Comput. Vis. Image Understand.*, vol. 210, Art. no. 103248, 2021.