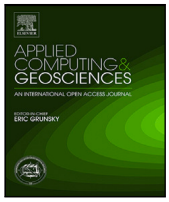
Applied Computing and Geosciences 21 (2024) 100150

Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.sciencedirect.com/journal/applied-computing-and-geosciences



Comparative performance analysis of simple U-Net, residual attention U-Net, and VGG16-U-Net for inventory inland water bodies

Ali Ghaznavi [a,b,c], Mohammadmehdi Saberioon [c,*], Jakub Brom [d], Sibylle Itzerott [c]

[a] Institute of Energy and Climate Research (IEK9), Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Straße, 52425 Jülich, Germany
[b] Faculty of Fisheries and Protection of Waters, South Bohemian Research Center of Aquaculture and Biodiversity of Hydrocenoses, Institute of Complex Systems, University of South Bohemia in České Budějovice, Zámek 136, 373 33 Nové Hrady, Czech Republic
[c] Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Section 1.4 Remote Sensing and Geoinformatics, Telegrafenberg, Potsdam 14473, Germany
[d] Department of Applied Ecology, Faculty of Agriculture and Technology, University of South Bohemia in České Budějovice, Studentská 1668, České Budějovice 37005, Czech Republic

ARTICLE INFO

Keywords:
Automated mapping
Deep learning
Land cover
Satellite imagery
Segmentation
Water bodies

ABSTRACT

Inland water bodies play a vital role at all scales in the terrestrial water balance and Earth's climate variability. Thus, an inventory of inland waters is crucially important for hydrologic and ecological studies and management. Therefore, the main aim of this study was to develop a deep learning-based method for inventorying and mapping inland water bodies using the RGB band of high-resolution satellite imagery automatically and accurately.

The Sentinel-2 Harmonized dataset, together with ZABAGED-validated ground truth, was used as the main dataset for the model training step. Three different deep learning algorithms based on U-Net architecture were employed to segment inland waters, including a simple U-Net, Residual Attention U-Net, and VGG16-U-Net. All three algorithms were trained using a combination of Sentinel-2 visible bands (Red [B04; 665nm], Green [B03; 560nm], and Blue [B02; 490 nm]) at a 10-meter spatial resolution.

The Residual Attention U-Net achieved the highest computational cost due to the increased number of trainable parameters. The VGG16-U-Net had the shortest run time and the lowest number of trainable parameters, attributed to its architecture compared to the simple and Residual Attention U-Net architectures, respectively. As a result, the VGG16-U-Net provided the best segmentation results with a mean-IoU score of 0.9850, a slight improvement compared to other proposed U-Net-based architectures.

Although the accuracy of the model based on VGG16-U-Net does not make a difference from Residual Attention U-Net, the computation costs for training VGG16-U-Net were dramatically lower than Residual Attention U-Net.

1. Introduction

Inland waters (i.e., rivers, streams, lakes, reservoirs, wetlands, and flood plains) significantly impact hydrological and biogeochemical cycles. They play a vital role at all scales in the terrestrial water balance and Earth's climate variability (Zhang et al., 2021a; Cooley et al., 2021). Furthermore, inland waters provide vital resources for humans and are the sole habitat for an extraordinarily rich, endemic, and sensitive biota. However, like many other ecosystems over the past century, humans' high demands on freshwater, continuous demographic pressure, and climate change have threatened the existence of inland water resources and biodiversity around the world (Dudgeon et al., 2006). Consequently, tracking and quantifying human and climate change influence on global inland water is essential, particularly for small water bodies, and delineating them is a prerequisite for further monitoring, modeling, and management.

Since the 1970s, remote sensing techniques have become increasingly popular for detecting and mapping inland waters regionally and globally (Bukata, 2013; Palmer et al., 2015). Since the launch of Sentinel-2, this trend has increased as Sentinel-2 is continuously acquiring high-resolution images from the land surface. Therefore, the scientific community and public and private sectors have used Sentinel-2 data extensively for land cover/use monitoring, including water bodies detection (Xu et al., 2019; Phiri et al., 2020). Many former studies using methods like spectral indices (Feyisa et al., 2014; Zou

* Corresponding author.
E-mail address: saberioon@gfz-potsdam.de (M. Saberioon).

https://doi.org/10.1016/j.acags.2023.100150
Received 6 July 2023; Received in revised form 14 December 2023; Accepted 18 December 2023
Available online 19 December 2023

et al., 2017), single band density slicing (Worden et al., 2021), or supervised classification (Bangira et al., 2019; Ghasemigoudarzi et al., 2020) for detecting and mapping water bodies as water bodies appear dark in optical remote sensing due to high absorbance of irradiance in the near-infrared (NIR) spectrum. Nevertheless, these methods exhibit limitations and can be challenging when attempting to inventory inland waters with satisfactory accuracy. For example, the determination of a consistent threshold value is frequently hindered by fluctuations in the physical environment across both space and time, as highlighted in prior research (Worden and de Beurs, 2020). While some approaches involve the use of multiple threshold values to identify an optimal threshold (Ji et al., 2009), this may not be universally applicable, particularly for water bodies with complex shapes, sizes, and spectral characteristics. Additionally, such methods may lack robustness in handling variations in image quality, resolution, and acquisition conditions (Sekertekin, 2021; Kavats et al., 2022). In water body classification, shadows produced by mountains, trees, buildings, and river banks can contaminate satellite imagery classification of water bodies (Pan et al., 2020). Therefore, a new method is still desirable for detecting and mapping inland waters where high-resolution orbital remote sensing data automatically and accurately.

Recent advancements in deep learning, particularly the use of semantic segmentation algorithms, play a pivotal role in classifying remote sensing images (Zhao et al., 2017; Lv et al., 2022). Among various semantic segmentation approaches, the U-Net architecture has gained attention for achieving excellent recognition of fine objects in complex scenes with relatively small amounts of training data (Abdi et al., 2018; Li et al., 2019; He et al., 2023). Importantly, our research seeks to investigate the robustness of U-Net architectures in handling diverse geographical and environmental conditions, including different terrains, climates, and characteristics of inland water bodies. Despite the growing use of U-Net-based algorithms, there remains a need for exploring diverse architectures tailored to specific challenges and scenarios in remote sensing applications. Previous studies have highlighted the superiority of U-Net in classifying land covers from medium-resolution remote sensing data (Zhang et al., 2021b). Additionally, innovative modifications, such as replacing the convolution layer with a bottleneck structure, have demonstrated impressive accuracy in segmenting water bodies while reducing model size and prediction time (An and Rui, 2022). U-Net has also proven effective in addressing shadow contamination issues in complex geographical scenes (Wang et al., 2021). Continued exploration of U-Net-based models with diverse architectures is crucial for segmenting various remote sensing scenarios and feature types. The adaptability and customizability of U-Net architectures make them versatile for addressing specific challenges, providing potential solutions, and demonstrating applicability across diverse remote sensing scenarios. Given the wealth of remote sensing data available for applications such as land cover mapping and environmental monitoring, U-Net-based models can be optimized for specific tasks, such as segmenting small objects, handling multi-spectral data, and processing high-resolution images.

This study stands out as a pioneering effort in its domain, as a comprehensive review of the literature underscores a notable gap—no similar research has been conducted to specifically address the challenges of detecting and mapping inland waters in diverse geographical and environmental conditions using U-Net architectures. The current investigation aims to bridge this void and contribute novel insights to the field, making it a unique and valuable addition to the current body of knowledge. The primary objective of this research is to develop, implement, and test an accurate deep learning segmentation method with reasonable computational cost for detecting and segmenting inland water bodies from high spatial resolution (10 m) remote sensing images. The choice of the U-Net is motivated by its solid performance in semantic segmentation tasks. Additionally, two other U-Net architectures, Residual Attention U-Net and VGG16-U-Net, are explored to identify the best architecture for automated inland water detection based on accuracy and computational cost.

2. Materials and pre-processing

2.1. Data preparation and pre-processing

This study acquired the raw images using the sentinel-2 Harmonized dataset archived on the Google Earth Engine JavaScript platform (GEE). The southern part of the Czech Republic, including the South Bohemian region, was selected as the region of interest (Fig. 1). This part of Czech republic were considered to train the model because of the more water bodies in and artificial lakes existing in this region of the country. Including images with more related RoI regions were helpful to train more efficient models to predict the water bodies. Sentinel-2 images acquired during summer 2022 with less than 10% of cloud covering were considered as datasets for training and testing algorithms.

In this study, the combination of visible bands of sentinel-2 (Red [B04; 665 nm], Green [B03; 560 nm], and Blue [B02; 490 nm]) were considered and used to obtain true color images for segmentation purpose. The reason of considering RGB bands is because the more bands used, the more complex and computationally expensive the segmentation model. In other words, increasing model development and deploy the model requires more time and computation power. Additionally, not all bands may provide useful information for segmenting of water bodies, so it is often more efficient to select a relevant subset of bands. Therefore, using only the RGB bands, which produce true color images, was a reasonable choice, given their sufficiency in achieving good accuracy in segmenting water bodies. Using fewer bands can also help reduce overfitting, which occurs when a model becomes too complex and fits the training data too closely, resulting in poor generalization to new data. By using a simpler model with fewer input features, the risk of overfitting can be reduced and the generalization performance of the segmentation model can be improved.

To achieve RGB images and render the image as a true-color composite, The Earth Engine visualization parameters and specific bands are configured as ‘B4’ (665 nm), ‘B3’ (560 nm), and ‘B2’ (490 nm) for red, green, and blue color channels with 10-meter spatial resolution, respectively. The “min” and “max” values in visualization parameters are suitable for displaying reflectance from typical Earth surface targets. The min value was set to zero, the max value was considered equal to 4000, and the Gamma correction factor was set to 1.4. After collecting the raw images from the Google Earth Engine (GEE) JavaScript platform, Raw images were downloaded and transferred into the QGIS software for further processing.

After transferring the raw image data into the QGIS, the specific parts of the South Bohemian region (Fig. 1, the rectangle) was selected as the main dataset. On the other hand, the labeled data from Czech Republic inland waters provided by ZABAGED (Czech Geodetic and Cadastral Office, 2019) were imported into the QGIS to generate the shape file of the inland water for all parts of the Czech Republic. Then, the same specific coordination from the GEE image and the labeled data were exported as “Tiff” file with a big size of $46K \times 46K$ pixel resolution.

In the next step, the image and mask in big size were patchified into smaller parts (Fig. 2). That process generated the main dataset for further analysis. The patchifing step splits images into small patches by given patch cell size (Weiyuan et al., 2017) (ie. like cropping image in big size into the small parts). Images were patchified and masked into the 2048 × 2048 pixel resolution to achieve suitable region of interest (ROI) area and avoid pixelating and blurring problems in the smaller size of the images. The patchifying step helped us to convert the image in big size into the images in smaller size to use in training step. After patchifying the image and mask into smaller parts, we achieved 504 images as the main dataset. The main dataset was split into three parts: (1) train set by randomly considering 322 images (80% of the main dataset), (2) test set by randomly considering 101 images (20% of the main dataset), (3) for model validation progress, 20% of the train set randomly selected (81 images) to prevent over-fitting problem during training progress and reach more stable performance for generated models.

Fig. 1. The map of the study area. The red region represented the area selected for the data collection phase. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2. Neural network architecture

2.2.1. Simple U-Net

Deep neural network methods delivered promising outcomes in classification and segmentation tasks in terms of accuracy when dealing with a large dataset. One of the promising neural network architectures for semantic segmentation is U-Net. The U-Net based methods deliver promising outcome in different sensitive research fields including medical and microscopy regions (Ronneberger et al., 2015; Ghaznavi et al., 2022). The U-Net was proposed and created for semantic segmentation based on the convolutional neural network (CNN) architecture and comprised of an encoder–decoder convolutional network topology. The encoder and decoder blocked in each level were connected to each other via a bridge to combine features from the encoder part with extracted features from the decode section. The feature representation extracted by the decoder part is useful for positioning, whereas encoder part features are efficient in achieving accurate segmentation. The proposed architecture for the simple U-Net method applied in this research is displayed in Fig. 3.

The first layer of the encoder part (Fig. 3, Part *A*) accepts images with the size 512 × 512 with three color channel (RGB) mode as input. The proposed U-Net structure has five levels. Each level consists of two 3 × 3 convolutions followed by Batch normalization for each convolution layer and applying a rectified linear unit "ReLu" as activation functions. In each level of the encoder part (down-sampling), the image size was halved by applying 2 × 2 max pooling operation, and the number of feature channels was doubled using convolutions. The maximum value was selected in the 2 × 2 area with the stride of two by max pooling operation. The encoder part of the network extracts the features and learns an abstract representation of the input image through a sequence of the encoder blocks.

In the decoder or up-sampling section (Fig. 3, Part *B*), the dimension of the feature maps in each level was doubled from the layer at the bottom to the top layer till achieved the exact same size as the input images. The bridge connection combined the extracted features from the encoder part into the decoder section. As a result of the concatenation step, the channels of the output feature maps will be twice as big as the size of the input features. The Concatenation step of feature maps in U-Net gives us better localization information. The output of the last decoder layer at the top includes 1 × 1 convolution with Sigmoid activation to predict the probabilities value of pixels for classification purposes. The size of the feature map at the output layer was achieved the exactly as same size as the input layer by applying Padding in the convolution process. The decoder part of the network used extracted abstract representation from the encoder part and generated a semantic segmentation mask. The Binary Focal Loss was used as loss function of the U-Net.

2.2.2. Residual attention U-Net

The architecture of U-Net consists of encoder and decoder blocks that are connected via a bridge at each level (Fig. 3). The bridge connections are responsible for merging the down-sampling and up-sampling paths together to reach spatial information. On the other hand, the concatenation step may transfer many unimportant and useless feature representations from the encoder part during the combination process. The attention mechanism implemented based on U-Net architecture (Fig. 4, part *D*) was proposed by Oktay et al. (2018) with a promising outcome in medical imaging. The soft attention mechanism was implemented to keep and highlight the most representative features and enhance achieved segmentation results by simple U-Net. The soft attention mechanism remark the important features and represses activations in the unrelated regions. As a result, model sensitivity and performance were slightly improved by employing the attention gate without requiring complicated and heavy computational costs (Ghaznavi et al., 2022).

The employed soft attention gate (Fig. 4, part *D*) getting two inputs, x and g. The input x was achieved by the concatenation bridges from the early layers of the encoder part and includes better spatial information. Input g comes from the deeper layers of the network known as the gating signal, which includes more efficient feature representation and contextual information to identify the focus region and gives weight to the different parts of the images. The attention coefficients $\alpha \in [0, 1]$ identify, extract, and assign weights to the features belong to the important part of the image regions in our case the water bodies. The attention mechanism progress, getting the weights to the pixels according to their relevance in training steps (Oktay et al., 2018). The more relevant part of the image will get weights bigger than the less relevant parts. So, by applying the achieved weights in the training process, we trained model that is more attentive to the relevant image parts. The multiplication of the input feature maps x^l and the achieved attention coefficient α generate the output of the attention gate:

$$q_{att}^{I} = \psi^{T}(\sigma_1(W_x^{T} x_i^{I} + W_g^{T} g_i + b_g)) + b_\psi, \tag{1}$$

$$\alpha_i^{I} = \sigma_2(p_{att}^{I}(x_i^{I}, g_i; \Theta_{att})), \tag{2}$$

whereas the σ_1 and σ_2 parameters correspond to the relu and sigmoid activation functions and Θ_{att} indicate different parameters including

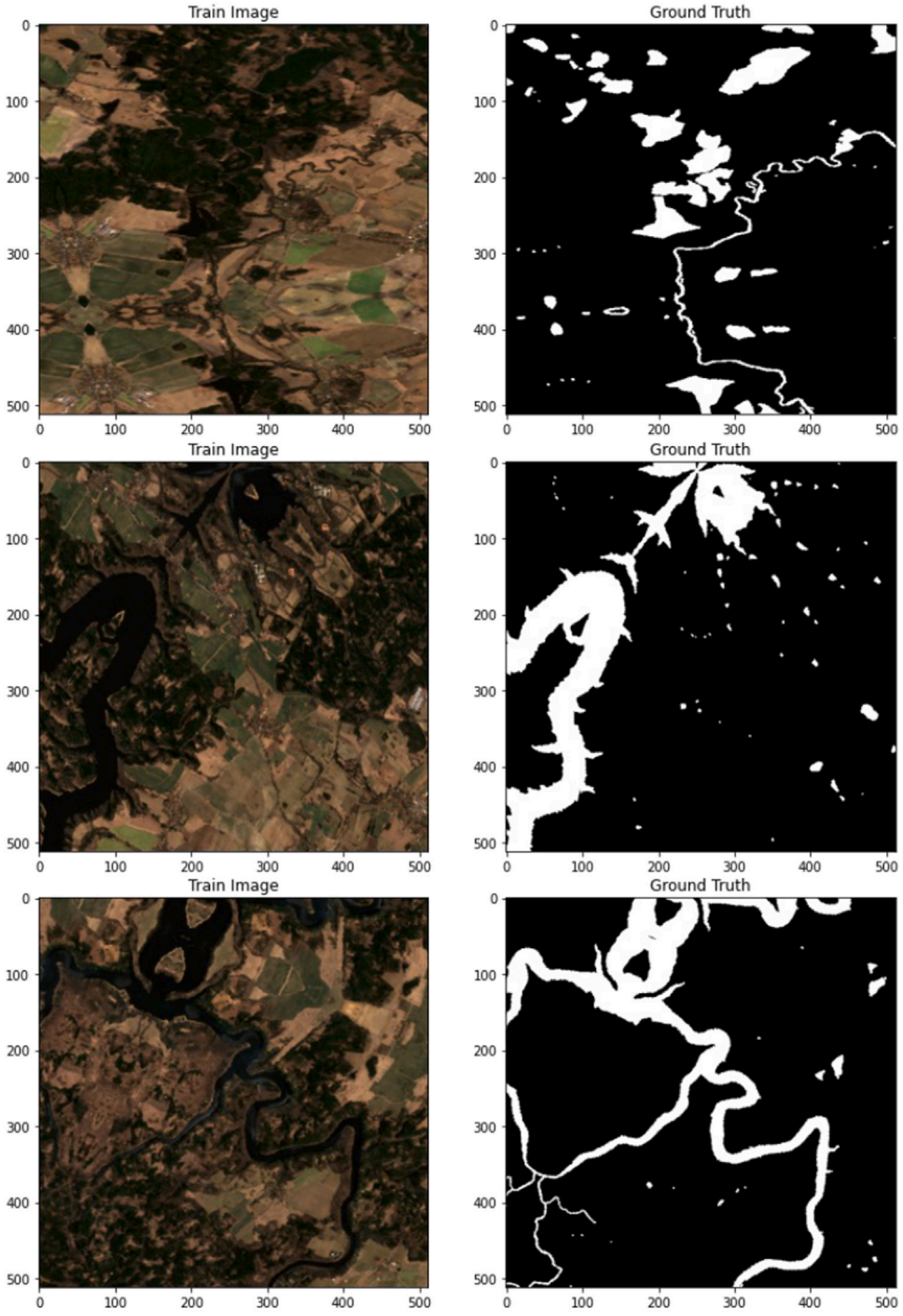


Fig. 2. Train set images and corresponded ground truth images. The size of image is 512 × 512.

linear transformations W_x and W_g, function ψ and bias terms b_ψ and b_g (Oktay et al., 2018).

Deeper neural networks deliver more effective performance in complex classification and segmentation tasks (Nishimura et al., 2021). Each level of the proposed U-Net-based architectures consists of many convolutional blocks (Fig. 4). The input value enters into the Convolutional blocks, the convolution operation, and the activation function applied in the input value and generates the output. In neural networks, the output of each convolutional block is the input of the next convolutional block. So, by making the neural network architecture deeper, the calculated gradient value from one block to another will be smaller because of the gradient vanishing effect, and the accuracy of the trained model will degrade rapidly instead of improving. The gradient vanishing problem appeared during the training procedure and affected the model's generalization ability. To mitigate this problem, the residual mechanism was implemented and applied to the proposed method to continuously update the calculated gradient values in each convolutional block and improve the performance of trained models (Ni et al., 2019). The proposed residual blocks, known as skip connections, will bypass one or more layers and update the gradient values from one or more previous layers into the layer step ahead. By combining the soft attention mechanism with the residual mechanism, we will get the weights into the important part of the image and overcome the gradient vanishing problem during training progress.

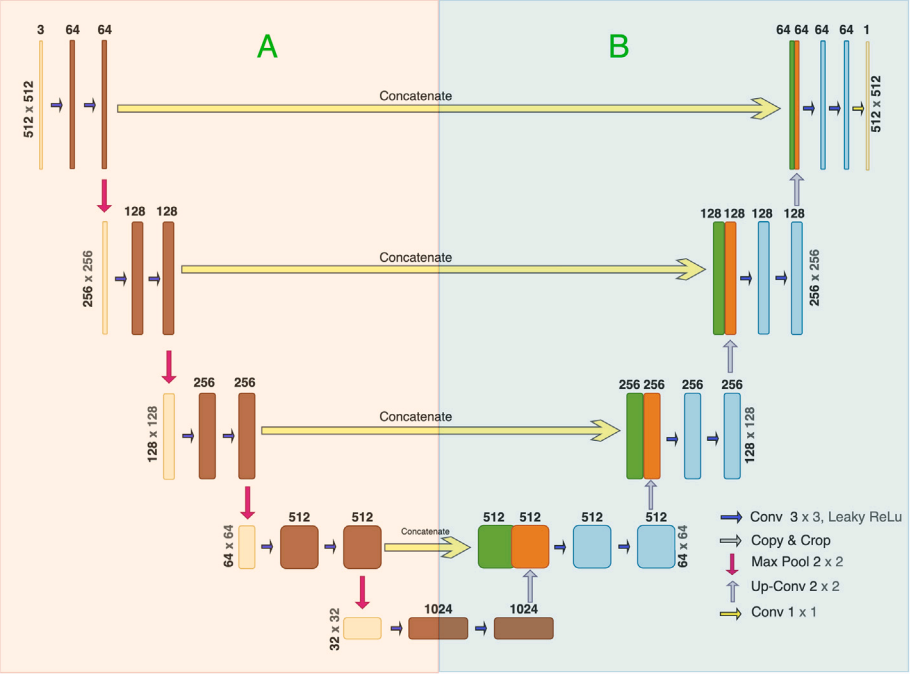


Fig. 3. The simple U-Net Architecture. Part *A* represent the encoder section and part *B* represent decoder section.

2.2.3. VGG16-U-Net

Different CNN architectures have been proposed to be combined with the U-Net architecture for improving the trained model accuracy and computational cost of the U-Net and reducing the number of trainable parameters in comparison to the simple U-Net. The VGG is the basis of CNN architecture proposed by Simonyan and Zisserman (2015) and developed by the Visual Geometry Group from Oxford University. The VGG was developed and proposed to reduce the number of trainable parameters in the Convolutional layers and improve the training time because of the structure of the developed architecture proposed by Simonyan and Zisserman (2015). The VGG architecture has many different variants depending on the number of layers from VGG11 to VGG19. The VGG16 efficiently performed many object detection and image classification tasks (Hamwi and Almustafa, 2022; Wahyuni et al., 2021). Due to this, in this research, the hybrid VGG16-U-Net architecture was chosen and implemented to compare with two other methods and improve the semantic segmentation results in term of performance and computational costs. To implement the proposed hybrid network, the encoder part of the U-Net, which is responsible for extracting the feature representation, was completely replaced with the VGG16 structure (Fig. 5, part *B*). The VGG16 architecture at the encoder part (Fig. 5, part *A*) consists of sixteen layers, including thirteen convolutional layers and three dense layers. The 3 fully connected layers of Vgg16 (Fig. 5, part *A*, green rectangles) were replaced with architecture that resembled the decoding part of U-Net, which formed the expanding path with convolution layers and upsampling layers (Fig. 5, part *B*). Hence, the VGG16 without the final 3 fully connected layers was retained as the contracting path (Balakrishna et al., 2018).

The first layer of the encoder section takes the input image with the size of 512 × 512 in RGB color mode and has 64 channels. Each convolutional blocks in each level have max pooling progress with the size of 2 × 2 and a stride of two to extract the maximal value. In each level of the encoder section, the size of the image was half, and the size of feature channels was doubled from 64 to a maximum of 512. The right side of the network (Fig. 6, Part *B*) represents the decoder part with five levels. The structure of the decoder section remained the same as we applied in the simple U-Net method. Each level of the encoder and decoder parts was connected via a concatenation bridge. The concatenation step combines features extracted from the encoder section with the decoder section, and this concatenation step is important for achieving localization information. The last encoder layer has 1 × 1 convolutional size to predict the probability value of each pixel and generate the semantic segmentation by applying the "Sigmoid" activation function.

2.3. Training models

The computational platform used for implementing all methods is Python 3.9. All deep learning frameworks were implemented using Keras with the backend of Tensorflow (Abadi et al., 2016) to train the best stable models. After developing methods and completing of implementation phase for all CNN architectures, the complete method was transferred and compiled on the Google Collab Pro + cluster account. The google clusters are equipped with two vCPU as processors, 24 Gb of RAM as memory, and P100 and T4 graphical processor unit (GPU)(Google LLC, California, USA). By the completion of the data pre-processing step (Section 2), 80% of the main dataset was chosen randomly as a train set (322 images), and the rest of 20% was considered randomly as a test set (101 images) for testing and evaluating the generated models' performance. Meanwhile, 20% of the training set was chosen randomly as the validation set (81 images) to validate the model and prevent over-fitting problems during the training process.

The input image size used in proposed CNN architectures was 512 × 512 px. All dataset images were resized from 2048 × 2048 px into 512 × 512 px as proper and specific input image size for proposed CNN's. We employed data augmentation variables during model training for all three CNN methods. The best-achieved values for each hyperparameter were reported in Table 1. The early stopping parameters are useful to prevent the over-fitting problem in the training phase. The threshold for patient value is set equal to 20. The "Relu" was selected as an activation function, and the Batch size value was considered 8. As a description of data Augmentation parameters, the "rotation range" means randomly rotating images between [−90, 90] degrees. The "width shift range" shift the image to the left or right (horizontal shifts), and the "height shift range" parameter shifts the

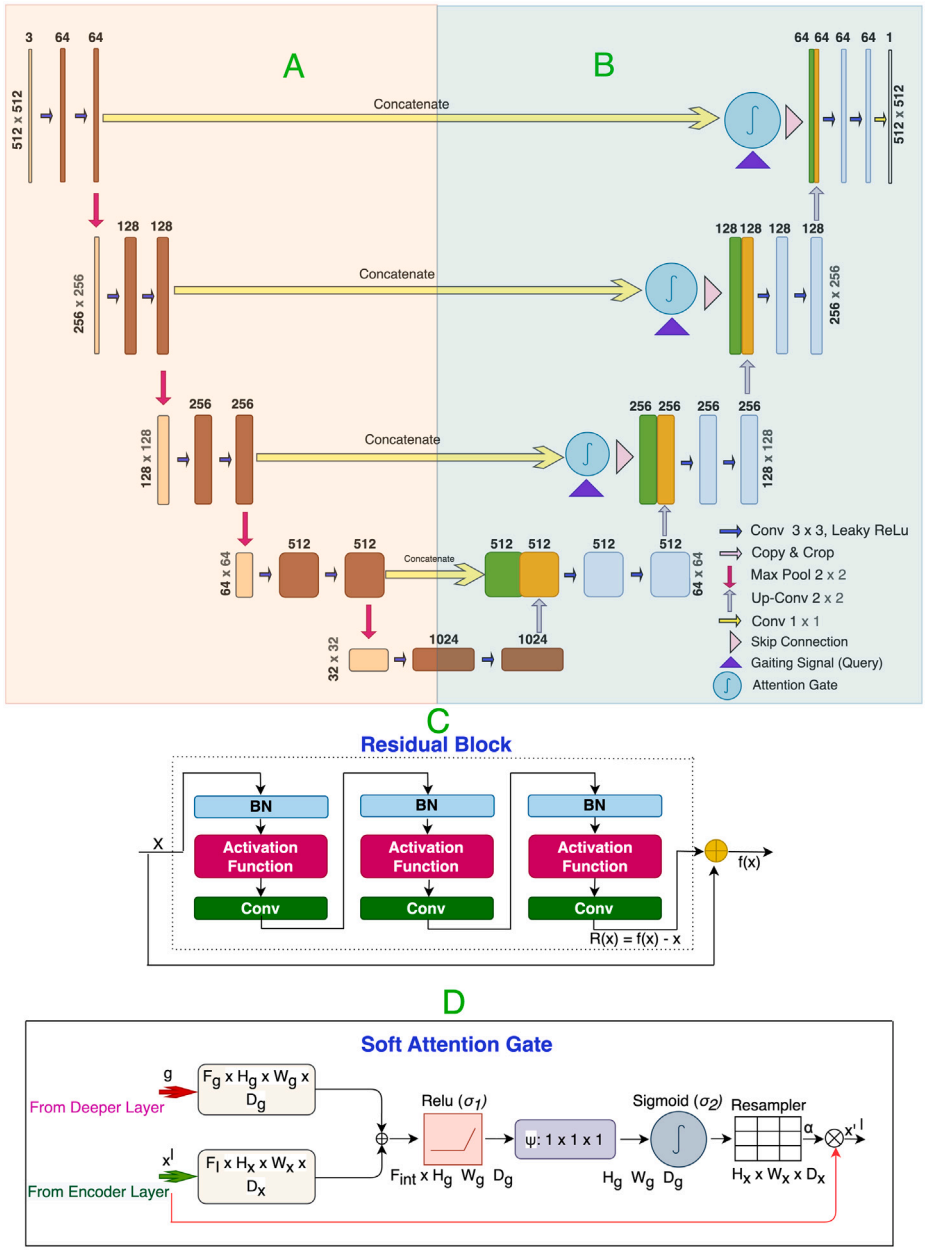


Fig. 4. The proposed architecture for Residual attention U-Net. Part *A* represents the encoder section, and part *B* represents the decoder section. Part *C* represents the residual mechanism. Part *D* represent the soft Attention mechanism. Each feature map has size as $H \times W \times D$, which H, W, and D represent height, width, and number of channels.

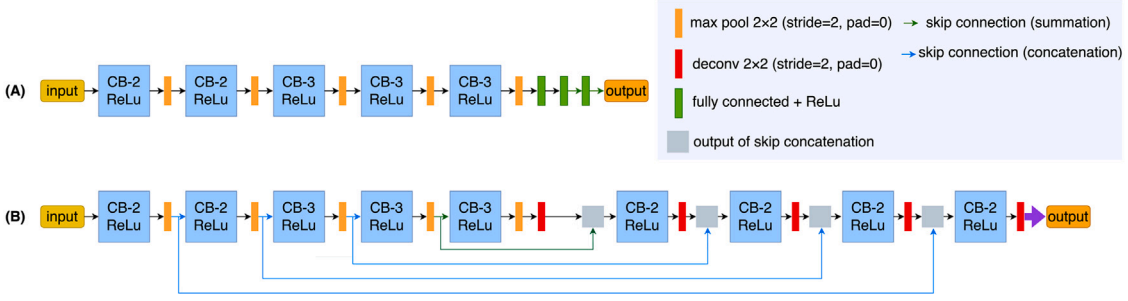


Fig. 5. Architecture of the VGG16 and its variants. *A*) represent the VGG16 network architecture. *B*) represent VGG16–U–Net architecture.

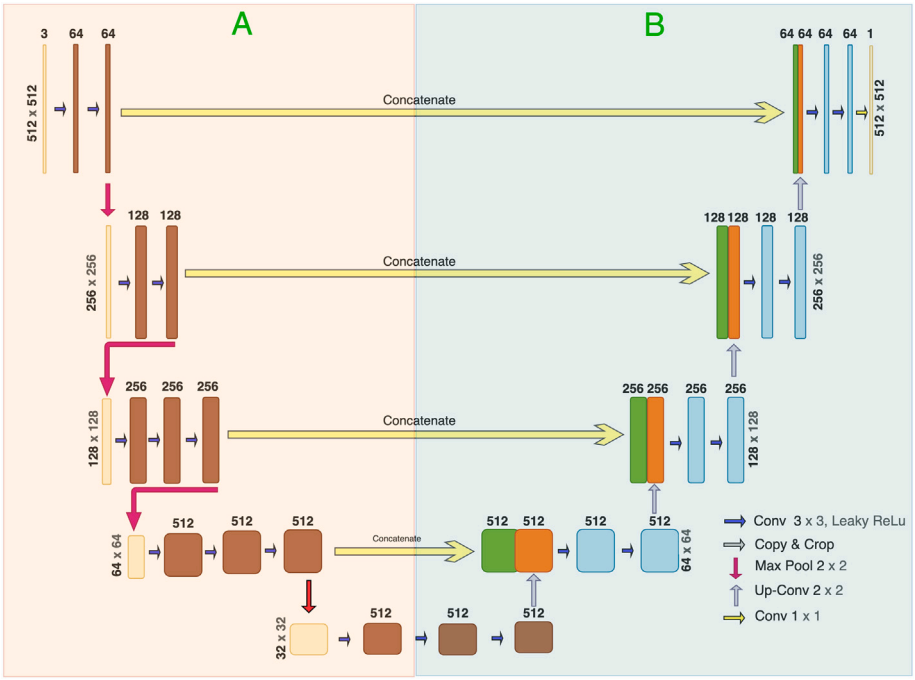


Fig. 6. Architecture of the proposed Hybrid VGG16-U-Net model. *A*) represent the encoder part of VGG16 architecture, *B*) represent the decoder part of U-Net respectively.

Table 1
The value of Hyperparameters used for all CNN models.

Hyperparameter	Value
Activation function	Relu
Learning rate	10^{-3}
Size of the Bach	8
Number of the Epochs	70
Early stopping	20
Number of steps in each epochs	100
Rotation range	90
Width shift	0.3
Height shift	0.3
Shear range	0.5
Zoom range	0.3

image vertically (up or down). The "shear range" parameter shows a distorted image along an axis to create or rectify the perception angle. The random zoom for the training images was obtained by the "zoom range" parameter. For optimizing the network, we choose the 'Adam' optimizer. The learning rate value was considered to 10^{-3}.

Semantic segmentation progress could be defined as a classification task at the pixel level to classify those pixels into water bodies or other classes. The segmented water bodies' images with the ground truth (GT) were compared to minimize the difference between them during the training using the Dice loss. The Binary Focal Loss was used as a loss function for semantic segmentation (Eq. (3)) (Lin et al., 2020):

$$\text{Focal Loss} = -\alpha_t(1-p_t)^{\gamma}\log(p_t), \tag{3}$$

Which $p_t \in [0,1]$ represents the predicted probability value achieved by the model for the ground truth class with label $y = 1$; $\alpha_t \in [0,1]$ corresponding to the weighting factor for class 1 and $1-\alpha_t$ for class 0; and $\gamma \geq 0$ representing tunable focusing parameter. Applying focal loss efficiently achieved better segmentation performance in regions of images that are challenging to segment (e.g., narrow inland water bodies or inland bodies with a similar texture to forest) and separate sensitive inland water bodies from the background. On the other hand, the focal loss as loss function manages and reduces the participation of the pixels belonging to the specific region that can be segmented easier (e.g., big and visible inland waters) over the image region in the model training progress. The model has the responsibility of updating the gradient direction. This progress depends on the loss of the model.

2.4. Evaluation metrics

To evaluate segmentation models generated by CNN's, different evaluation metrics were used (Eqs. (4)–(8)). The TP represents a true positive, FP indicates a false positive, FN corresponds to a false negative, and TN represents true negative values, respectively (Pan et al., 2017). The generated models were evaluated with the test sets using described metrics, and mean values of each metric were reported in Table 3.

The accuracy (Acc) metric indicates the percentage of the pixels which segmented correctly from water bodies. The Precision (Pre) metric represents a ratio of the pixels segmented as water bodies that exactly match the masks (GT). The Recall metric indicates the ratio of pixels belonging to the water bodies in the mask (GT), which is detected properly over the segmentation process. The Dice coefficient, known as F1-score, indicates if the segmented area is equal to the mask of the image (GT) in terms of location and level of detail. The F1-score represents ascertaining how accurate is the segmentation result in boundary regions (Csurka et al., 2013) and is more important than the ACC metric for evaluating model performance. The most important metric for segmentation model evaluation is Intersection over Union (IoU), also known as the Jaccard similarity index. The mentioned

Table 2
CNN's architecture trainable parameters and runtimes.

Network name	Training time	Trainable parameters
U-Net	3:01':47''	31,402,501
Residual Attention U-Net	4:17':23''	39,090,377
VGG16-U-Net	2:53':19''	25,862,337

metric represents the correlation between the prediction of the model and mask (GT) (Long et al., 2015; Vijay et al., 2015), and indicates the overlap and union area proportion for the model predicted and mask (GT).

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{4}$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recl} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{Dice} = \frac{2 \times \text{Pre} \times \text{Recl}}{\text{Pre} + \text{Recl}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{IoU} = \frac{|\, y_t \cap y_p \,|}{|\, y_t \,| + |\, y_p \,| - |\, y_t \cap y_p \,|} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{8}$$

3. Results and discussion

The proposed neural network models were well trained by processing 70 epochs according to the training/validation loss and accuracy plots (Fig. 7). To achieve the best training performance and stability, we assume all models were trained well according to the best-optimized hyperparameter values listed in Table 1. The best hyperparameter values were achieved by training several models based on different values of hyperparameters to achieve the best model performance and training stability. The trained models were evaluated using a test dataset to assess the performance of the proposed models based on the metrics written in Eqs. (4)–(8).

The simple U-Net model had an average computational cost in comparison with the Residual attention and VGG16-U-Net architecture. However, the number of the trainable parameters in the Residual attention U-net increased dramatically because of soft attention and residual mechanism, which cause the highest computational cost by this architecture. On the other hand, VGG16-U-Net had the lowest number of trainable parameters and, as a result, the shortest run time because of the structure of this architecture and achieved the best performance compared with the other two proposed methods (Table 2).

Fig. 8 shows the segmentation results achieved by different proposed CNN architectures. The result of segmentation accomplished by U-Net did not manage to segment all the water bodies over the test set image and suffered from a miss segmentation problem (Fig. 8, red circle). The Residual Attention U-Net segmented the borders of water bodies in complete shape, and the segmentation result was improved in comparison with the simple U-Net. Nevertheless, the result achieved by Residual Attention U-Net faced the under-segmentation problems in some water bodies regions to detect and segment some edges as visualized in Fig. 8 (green circle). The best performance of the segmentation was achieved by the VGG16-U-Net method. The result represents a more precise and accurate segmentation of the water bodies' borders, especially in the edge region and sensitive areas (Fig. 8, light blue circle).

Table 3 displays the evaluation of different U-Net-based proposed models with different evaluation metrics using (Eqs. (4)–(8)) as the mean value for all the metrics. The simple U-Net achieved the lowest segmentation performance according to the value of Mean-IoU and other evaluation metrics. The Residual Attention U-Net model represents a more improved segmentation result in comparison with the U-Net model in terms of the same test set image and evaluation metric values. In one more step, the segmentation result was further improved after applying the VGG16 encoder architecture with U-Net as a hybrid VGG16-U-Net method.

The U-Net architecture is one of the promising semantic segmentation methods which have been used in different research fields. The simple U-Net have been selected as first method to implement and apply in our study. As next phase, we slightly improved the obtained result by modifying the simple U-Net architecture by adding the residual mechanism together with soft attention mechanism as extension into the simple U-Net. At the last step, we replaced the encoder (feature extraction) part of the U-Net with more powerful VGG16 architecture to build hybrid CNN architecture with more efficient feature extraction section and compare the obtained result with previous methods in term of performance and computational costs.

To the best knowledge, there is no similar research that has been done before based on the proposed methods for detecting and segmenting inland water. However, some researchers applied different deep learning algorithms to detect and segment the inland waters. Table 4 represent the comparison of the similar literature with the proposed methods in this study. Zhong et al. (2022) proposed a noise-cancelling transformer network (NT-Net) for the automatic extraction of lake water bodies from remote sensing images and resolve the over-segmentation problem obtained by other literature. The proposed method obtained a 0.862 accuracy value in terms of the IoU metric. Zhang and Wang (2019) proposed a modified feature extraction network and a modified encoder–decoder network based on depth-wise separable convolution for segmenting the water bodies. The proposed method achieved 0.984 IoU metric accuracy. The authors in Xiang et al. (2023) proposed a dense pyramid pooling module (DensePPM) to extract global prior knowledge with a dense scale distribution for Segmenting Water Bodies From Aerial Images. The proposed method obtained a 0.842 metric value in terms of the IoU metric. Chang et al. (2022) proposed modified U-Net with residual mechanism and attention mechanism in encoder section based on PMS1 remote sensing data of GF2 satellite. The authors achieved good result (i.e., IoU = 0.9270). Ch et al. (2022) used Sentinel-2 image with two Band3 (Sentinel-2 Green Channel) and Band8 (Sentinel-2 Infrared Channel) and combined these two channel by following "NWDI" formula (as described in original paper) to achieve dataset images and then applied simple U-Net architecture to analyze them. The authors achieved 0.89 of Mean IoU score based on suggested method.

4. Conclusions

The efficiency and quality of the segmentation of orbital remote sensing images are the fundamental elements influencing the application of remote sensing for land cover/use mapping. Image semantic segmentation methods based on deep learning remarkably eliminated conventional segmentation methods' shortcomings (e.g., no distinct segmentation due to complex image background or many target instances in one image). This paper analyzed and compared three different deep learning, U-Net-based methods, including simple U-Net, Residual Attention U-Net, and VGG16-U-Net, to detect and segment inland water bodies using high-resolution satellite images. The results of this study indicate that the U-Net-based algorithms can be employed to inventory inland water bodies fast, accurately, and inexpensively in terms of computation cost. The results of this study can pave the way for implementing precision land cover mapping based on high-resolution satellite imagery by providing an objective, fast, accurate algorithm for inventorying land covers globally. Therefore, this study can be extended further to investigate other state-of-the-art deep learning algorithms also to evaluate them for other types of land cover/use mapping. The code used in this study is publicly available on our Gitlab repository (https://git.gfz-potsdam.de/ali/remotesensing-hida).

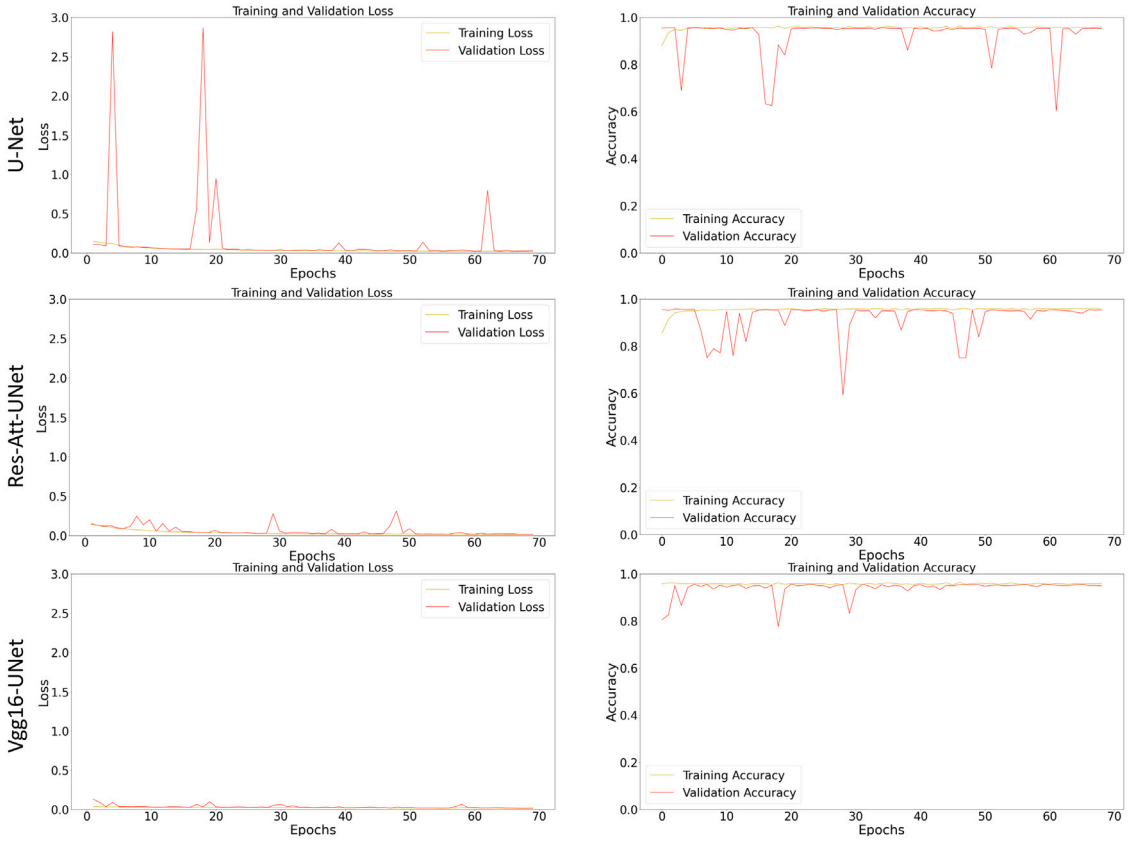


Fig. 7. The training loss and accuracy plots for U-Net (first raw), Residual Attention U-Net (second raw), and VGG16-U-Net (third raw).

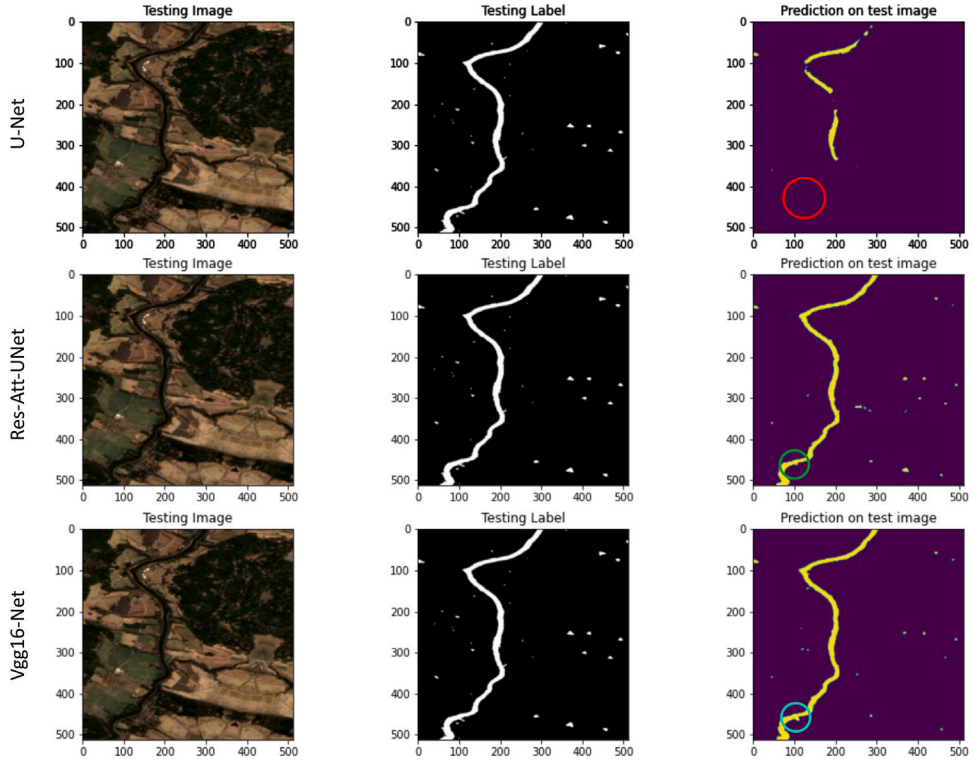


Fig. 8. Result of Segmentation for the U-Net (the red circle visualizes the miss-segmentation of water bodies), Residual Attention U-Net (the green circle visualizes the under-segmentation issue), and the VGG16-U-Net (light blue circle visualizes the accurate segmentation of the water bodies. The size of images is 512 × 512. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
The performance of the CNN Models evaluated by the different metrics. Green highlighted values indicate the best performance of segmentation according to the reported metrics.

Network	Accuracy	Precision	Recall	m-IoU	m-Dice
U-Net	0.9710	0.9997	0.9709	0.9707	0.9849
Residual Attention U-Net	0.9852	0.9986	0.9861	0.9848	0.9923
VGG16-U-Net	0.9855	0.9981	0.9869	0.9850	0.9924

Table 4
Comparison of the proposed CNNs with other similar literature. The highlighted Green value represent the highest segmentation accuracy achieved by proposed methods.

Models	IoU	Dice	Acc
prop. U-Net	0.9707	0.9849	0.9710
prop. Residual Attention-U-Net	0.9848	0.9923	0.9852
prop. VGG16-U-Net	0.9850	0.9924	0.9855
NT-U-Net (Zhong et al., 2022)	0.862	–	–
Modified Encoder-Decoder (Zhang and Wang, 2019)	0.984	–	–
DensePPM (Xiang et al., 2023)	0.842	–	–
Res2U-Net (Chang et al., 2022)	0.9270	–	–
ResNet50 (An and Rui, 2022)	0.9781	–	–
U-Net (Ch et al., 2022)	0.89	–	–

Funding

The authors would like to thank the EU, German's Federal ministry of education and research (BMBF), and The Technology Agency of the Czech Republic (TAČR) for funding in the frame of the collaborative international consortium AIHABs financed under the ERA-NET AquaticPollutants Joint Transnational Call (GA N° 869178). This ERA-NET is an integral part of the activities developed by the Water, Oceans, and AMR Joint Programming Initiatives. Furthermore, the authors appreciate the Helmholtz Information and data science academy (HiDA) funding in the frame of the Helmholtz Visiting Researcher Grant. The authors would like to thank the European Regional Development Fund in the frame of the project ImageHeadstart (ATCZ215) in the Interreg V-A Austria–Czech Republic programme and the project GAJU 114/2022/Z.

Code availability section

UNet-based methods for Inventory Inland Water Bodies using remote sensing

Contact: saberioon@gfz-potsdam.de, +49(33)1626427539

Hardware requirements: For training the model, developed methods require Graphical computation resources (GPU) with high level of memory (e.g., Google Colab with 16 GB of GPU memory) depending on the size of the dataset.

Program language: python

Program size: 241.1 MB

The source codes are available for downloading at the link: https://git.gfz-potsdam.de/ali/remotesensing-hida

CRediT authorship contribution statement

Ali Ghaznavi: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Mohammadmehdi Saberioon:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jakub Brom:** Writing – review & editing, Data curation. **Sibylle Itzerott:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Mohammadmehdi Saberioon reports financial support was provided by Federal Ministry of Education and Research Bonn Office. Jakub Brom reports financial support was provided by TACR.

Data availability

Data will be made available on request.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., 2016. TensorFlow: large-scale machine learning on heterogeneous distributed systems. In: OSDI'16: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. pp. 265–283.

Abdi, G., Samadzadegan, F., Reinartz, P., 2018. Deep learning decision fusion for the classification of urban remote sensing data. J. Appl. Remote Sens. http://dx.doi.org/10.1117/1.jrs.12.016038.

An, S., Rui, X., 2022. A high-precision water body extraction method based on improved lightweight U-net. Remote Sens. 14 (17), http://dx.doi.org/10.3390/rs14174127.

Balakrishna, C., Dadashzadeh, S., Soltaninejad, S., 2018. Automatic detection of lumen and media in the IVUS images using U-net with VGG16 encoder. http://dx.doi.org/10.48550/arXiv.1806.07554, arxiv.

Bangira, T., Alfieri, S.M., Menenti, M., van Niekerk, A., 2019. Comparing thresholding with machine learning classifiers for mapping complex water. Remote Sens. 11 (11), http://dx.doi.org/10.3390/rs11111351.

Bukata, R.P., 2013. Retrospection and introspection on remote sensing of inland water quality: "Like Déjà Vu All Over Again". J. Gt. Lakes Res. 39, 2–5. http://dx.doi.org/10.1016/j.jglr.2013.04.001, Remote Sensing of the Great Lakes and Other Inland Waters.

Ch, A., Ch, R., Gadamsetty, S., Iwendi, C., Gadekallu, T., Dhaou, I., 2022. ECDSA-based water bodies prediction from satellite images with UNet. Water http://dx.doi.org/10.3390/w14142234.

Chang, X., Fei, Y., Bao, Z., Deng, B., Yuan, F., 2022. High-resolution remote sensing water extraction based on improved U-net. In: ISCTT 2022; 7th International Conference on Information Science, Computer Technology and Transportation. pp. 1–5.

Cooley, S.W., Ryan, J.C., Smith, L.C., 2021. Human alteration of global surface water storage variability. Nature 591 (7848), 78–81. http://dx.doi.org/10.1038/s41586-021-03262-3.

Csurka, G., Larlus, D., Perronnin, F., 2013. What is a good evaluation measure for semantic segmentation? In: Proceedings of the British Machine Vision Conference. BMVA Press, pp. 32.1–32.11. http://dx.doi.org/10.5244/C.27.32.

Czech Geodetic and Cadastral Office, 2019. The basic database of geographic data of the Czech Republic (ZABAGED®) map. Retrieved September 15, 2022, [Dataset] URL http://data.europa.eu/88u/dataset/https-atom-cuzk-cz-api-responses-cz-00025712-cuzk_zabaged-jsonld.

Dudgeon, D., Arthington, A.H., Gessner, M.O., Kawabata, Z.-I., Knowler, D.J., Lévêque, C., Naiman, R.J., Prieur-Richard, A.-H., Soto, D., Stiassny, M.L.J., Sullivan, C.A., 2006. Freshwater biodiversity: importance, threats, status and conservation challenges. Biol. Rev. 81 (2), 163–182. http://dx.doi.org/10.1017/s1464793105006950.

Feyisa, G.L., Meilby, H., Fensholt, R., Proud, S.R., 2014. Automated water extraction index: A new technique for surface water mapping using landsat imagery. Remote Sens. Environ. 140, 23–35. http://dx.doi.org/10.1016/j.rse.2013.08.029.

Ghasemigoudarzi, P., Huang, W., Silva, O.D., Yan, Q., Power, D., 2020. A machine learning method for inland water detection using CYGNSS data. IEEE Geosci. Remote Sens. Lett. 19, 1–5. http://dx.doi.org/10.1109/lgrs.2020.3020223.

Ghaznavi, A., Rychtáriková, R., Saberioon, M., Štys, D., 2022. Cell segmentation from telecentric bright-field transmitted light microscopy images using a residual attention U-net: A case study on hela line. Comput. Biol. Med. 147, http://dx.doi.org/10.1016/j.compbiomed.2022.105805.

Hamwi, W.A., Almustafa, M.M., 2022. Development and integration of VGG and dense transfer-learning systems supported with diverse lung images for discovery of the coronavirus identity. Inform. Med. Unlocked 32, 101004. http://dx.doi.org/10.1016/j.imu.2022.101004.

He, J., Lin, Y.-N., Shi, F., Fu, J., Chen, B., 2023. Sentinel-2 research on the detection and classification methods of maritime ship targets from remote sensing images. J. Phys. Conf. Ser. http://dx.doi.org/10.1088/1742-6596/2425/1/012014.

Ji, L., Zhang, L., Wylie, B.K., 2009. Analysis of dynamic thresholds for the normalized difference water index. Photogramm. Eng. Remote Sens. 75 (11), 1307–1317. http://dx.doi.org/10.14358/PERS.75.11.1307.

Kavats, O., Khramov, D., Sergieieva, K., 2022. Surface water mapping from SAR images using optimal threshold selection method and reference water mask. Water 14 (24), http://dx.doi.org/10.3390/w14244030.

Li, Z., Wang, R., Zhang, W., Hu, F., Meng, L., 2019. Multiscale features supported DeepLabV3+ optimization scheme for accurate water semantic segmentation. Ieee Access http://dx.doi.org/10.1109/access.2019.2949635.

Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P., 2020. Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 318–327. http://dx.doi.org/10.1109/TPAMI.2018.2858826.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440. http://dx.doi.org/10.1109/CVPR.2015.7298965.

Lv, Z., Liu, T., Benediktsson, J.A., Falco, N., 2022. Land cover change detection techniques: Very-high-resolution optical images: A review. IEEE Geosci. Remote Sens. Mag. 10 (1), 44–63. http://dx.doi.org/10.1109/MGRS.2021.3088865.

Ni, Z.-L., Bian, G.-B., Zhou, X.-H., Hou, Z.-G., Xie, X.-L., Wang, C., Zhou, Y.-J., Li, R.-Q., Li, Z., 2019. RAUNet: Residual attention U-net for semantic segmentation of cataract surgical instruments. In: IEEE Conference on Computer Vision and Pattern Recognition. http://dx.doi.org/10.48550/arXiv.1909.10360.

Nishimura, K., Wang, C., Watanabe, K., Ker, D.F.E., Bise, R., 2021. Weakly supervised cell instance segmentation under various conditions. Med. Image Anal. 73, http://dx.doi.org/10.1016/j.media.2021.102182.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning where to look for the pancreas. In: 1st Conference on Medical Imaging with Deep Learning (MIDL 2018).

Palmer, S.C., Kutser, T., Hunter, P.D., 2015. Remote sensing of inland waters: Challenges, progress and future directions. Remote Sens. Environ. 157, 1–8. http://dx.doi.org/10.1016/j.rse.2014.09.021, Special Issue: Remote Sensing of Inland Waters.

Pan, X., Li, L., Yang, H., Liu, Z., Yang, J., Fan, Y., 2017. Accurate segmentation of nuclei in pathological images via sparse reconstruction and deep convolutional networks. Neurocomputing 229, 88–99. http://dx.doi.org/10.1016/j.neucom.2016.08.103.

Pan, F., Xi, X., Wang, C., 2020. A comparative study of water indices and image classification algorithms for mapping inland surface water bodies using landsat imagery. Remote Sens. 12 (10), http://dx.doi.org/10.3390/rs12101611.

Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V.R., Murayama, Y., Ranagalage, M., 2020. Sentinel-2 data for land cover/use mapping: A review. Remote Sens. 12 (14), http://dx.doi.org/10.3390/rs12142291.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science. Vol. 9321, Springer, Cham, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Sekertekin, A., 2021. A survey on global thresholding methods for mapping open water body using sentinel-2 satellite imagery and normalized difference water index. Arch. Comput. Methods Eng. 28 (3), 1335–1347. http://dx.doi.org/10.1007/s11831-020-09416-2.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. ICLR Conference http://dx.doi.org/10.48550/arXiv.1409.1556.

Vijay, B., Kendall, A., Cipolla, R., 2015. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 228–233. http://dx.doi.org/10.1109/TPAMI.2016.2644615.

Wahyuni, I., Wang, W.J., Liang, D., Chang, C.C., 2021. Rice semantic segmentation using unet-VGG16: A case study in yunlin, Taiwan. In: IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Hualien City, Taiwan. http://dx.doi.org/10.1109/ISPACS51563.2021.9651038.

Wang, Y., Li, S., Lin, Y., Wang, M., 2021. Lightweight deep neural network method for water body extraction from high-resolution remote sensing images with multisensors. Sensors http://dx.doi.org/10.3390/s21217397.

Weiyuan, W., Divakar, V., Wangyin, Y., 2017. patchify: A python library to split images into small patches and merge them back. https://pypi.org/project/patchify/, (Accessed: 2021-10-28).

Worden, J., de Beurs, K.M., Koch, J., Owsley, B.C., 2021. Application of spectral index-based logistic regression to detect inland water in the south caucasus. Remote Sens. 13 (24), http://dx.doi.org/10.3390/rs13245099.

Worden, J., de Beurs, K.M., 2020. Surface water detection in the caucasus. Int. J. Appl. Earth Obs. Geoinf. 91, 102159. http://dx.doi.org/10.1016/j.jag.2020.102159.

Xiang, D., Zhang, X., Wu, W., Liu, H., 2023. DensePPMUNet-a: A robust deep learning network for segmenting water bodies from aerial images. IEEE Trans. Geosci. Remote Sens. 61, 1–11. http://dx.doi.org/10.1109/TGRS.2023.3251659.

Xu, Y., Yu, L., Feng, D., Peng, D., Li, C., Huang, X., Lu, H., Gong, P., 2019. Comparisons of three recent moderate resolution African land cover datasets: CGLS-LC100, ESA-S2-LC20, and FROM-GLC-africa30. Int. J. Remote Sens. 40 (16), 6185–6202. http://dx.doi.org/10.1080/01431161.2019.1587207.

Zhang, S., Foerster, S., Medeiros, P., Araújo, J.C.d., Duan, Z., Bronstert, A., Waske, B., 2021a. Mapping regional surface water volume variation in reservoirs in north-eastern Brazil during 2009–2017 using high-resolution satellite images. Sci. Total Environ. 789, 147711. http://dx.doi.org/10.1016/j.scitotenv.2021.147711.

Zhang, W., Tang, P., Zhao, L., 2021b. Fast and accurate land-cover classification on medium-resolution remote-sensing images using segmentation models. Int. J. Remote Sens. 42 (9), 3277–3301. http://dx.doi.org/10.1080/01431161.2020.1871094.

Zhang, P., Wang, G., 2019. The modified encoder-decoder network based on depthwise separable convolution for water segmentation of real sar imagery. In: 2019 International Applied Computational Electromagnetics Society Symposium. Vol. 60, pp. 1–2. http://dx.doi.org/10.23919/ACES48530.2019.9060500.

Zhao, W., Du, S., Wang, Q., Emery, W.J., 2017. Contextually guided very-high-resolution imagery classification with semantic segments. ISPRS J. Photogramm. Remote Sens. 132, 48–60. http://dx.doi.org/10.1016/j.isprsjprs.2017.08.011.

Zhong, H.-F., Sun, Q., Sun, H.-M., Jia, R.-S., 2022. NT-Net: A semantic segmentation network for extracting lake water bodies from optical remote sensing images based on transformer. IEEE Trans. Geosci. Remote Sens. 60, 1–13. http://dx.doi.org/10.1109/TGRS.2022.3197402.

Zou, Z., Dong, J., Menarguez, M.A., Xiao, X., Qin, Y., Doughty, R.B., Hooker, K.V., David Hambright, K., 2017. Continued decrease of open surface water body area in Oklahoma during 1984–2015. Sci. Total Environ. 595, 451–460. http://dx.doi.org/10.1016/j.scitotenv.2017.03.259.