GeoAI Framework for Evaluation of Deep Learning Models in Water Body Segmentation

Amirhossein Mahmoudnia, Mohammad Kordmafi

Dept. of Surveying and Geomatic Eng., College of Eng., University of Tehran, Tehran, Iran
(amirh.mahmoudnia@gmail.com; mohammad.kordmafi@ut.ac.ir)

Keywords: GeoAI, Semantic Segmentation, Deep Learning, Water Body Mapping, U-Net

Abstract

Accurate delineation of surface water bodies from satellite imagery is fundamental for hydrological monitoring, disaster management, and environmental assessment. Deep learning–based semantic segmentation has shown great promise for this task, yet the relative performance of widely used encoder–decoder architectures remain insufficiently explored under consistent experimental settings. This study presents a GeoAI-based evaluation of four state-of-the-art segmentation architectures: U-Net, U-Net++, DeepLabV3+, and Feature Pyramid Network (FPN), all employing a ResNet-34 encoder pretrained on ImageNet. Using the Earth Surface Water Dataset, we applied an identical tiling strategy, preprocessing pipeline, data augmentation, optimizer configuration, and training schedule to ensure a fair comparison. Our results show that DeepLabV3+ achieved the highest accuracy with an IoU of 0.955 and Dice of 0.973, followed by U-Net++ (IoU 0.945, Dice 0.969) and U-Net (IoU 0.9352, Dice 0.9598). FPN performed comparatively lower with an IoU of 0.9218 and Dice of 0.9514. These findings demonstrate that architectural choices substantially influence segmentation quality even with identical encoders and training protocols. The study provides benchmark results and practical insights for selecting deep learning architectures for operational water body mapping in remote sensing applications.

1. Introduction

Surface water is a critical resource that underpins human well-being, providing essential services for drinking supplies, sanitation, and irrigation (Markert et al., 2020). Beyond its ecological roles, surface water is central to the hydrological cycle and enables hydropower production, navigation, and numerous industrial uses (Aekakkararungroj et al., 2020). Beyond its provisioning services, surface water contributes to climate modulation, supports ecological diversity, and shapes land-use conservation policies (Valentin et al., 2008). Consequently, up-to-date knowledge of the spatial distribution and areal extent of surface water is fundamental to sustainable water management. Equally important is the capacity to rapidly detect, map, and communicate affected areas, supporting national and local governments, NGOs, and emergency services with synoptic, decision-ready information for disaster preparedness and planned response (Nemni et al., 2020).

Satellite-based remote sensing remains a standard means of delineating surface water, with common techniques ranging from index-based detection (Gao, 1996) to machine-learning classification (Huang et al., 2018) and dynamic, context-aware thresholding (Nasirzadehdizaji et al., 2019). Furthermore, investigations of surface-water occurrence employ both passive and active sensors. Passive instruments draw primarily on measurements in the visible and infrared bands, whereas active systems operate in the microwave domain (Flores-Anderson et al., 2019). Optical satellite imagery is the dominant input for monitoring surface water at broad spatial scales and on time-critical schedules (Li et al., 2025).

Because the Earth's surface includes a mix of water and non-water features, classification is challenging: rivers, lakes, reservoirs, and seas vary in morphology and spectra, while shadows, asphalt, and buildings can appear spectrally and spatially similar to water. These overlaps commonly reduce the reliability of surface-water maps. In addition, clouds and haze commonly contaminate optical observations, thereby hindering the production of accurate, high-resolution maps of surface water. In recent years, active spaceborne microwave sensors have been deployed, providing higher-resolution data that remain unaffected by cloud cover (Oddo & Bolten, 2019).

The integration of artificial intelligence techniques into geospatial analysis, often termed GeoAI has transformed the landscape of environmental and hydrological monitoring. GeoAI brings together spatial data science, machine learning, and big geospatial data to enable spatial reasoning, pattern discovery, and predictive modelling over large and complex domains (e.g., remote sensing, GIS) (Boutayeb et al., 2024). In hydrology and water resources research, GeoAI methods have been increasingly adopted for tasks such as flood forecasting, streamflow prediction, water quality estimation, and fluvial geomorphology mapping (Gonzales-Inca et al., 2022). These methods are especially attractive because they can capture non-linear dependencies, fuse heterogeneous data sources, and scale to high spatio-temporal resolutions that traditional physical models struggle to handle (Iyer et al., 2025).

Despite their promise, GeoAI models for hydrological systems must confront challenges: the need for large labelled datasets, limited interpretability, potential overfitting in space or time, and integration with domain knowledge (Gonzales-Inca et al., 2022). Researchers have begun exploring hybrid approaches, where physical models inform or constrain data-driven ones, to improve generalizability and interpretability. In this context, water body segmentation from remote sensing imagery represents a core subtask: accurate delineation of rivers, lakes, wetlands, or reservoirs is often a prerequisite to hydrological modeling, change-detection, flood mapping, and ecosystem monitoring.

In recent years, deep learning has emerged as the dominant paradigm for semantic segmentation of remote sensing and Earth observation imagery. Earlier approaches relying on spectral indices, thresholding, and hand-crafted features are being superseded by convolutional neural networks that learn hierarchical representations directly from raw pixel data. Surveys show that modern segmentation architectures such as U-Net variants, DeepLab, and transformer-based models—consistently outperform traditional baselines on remote sensing benchmarks (Maniyar et al., 2025).

In light of this, a comparative and systematic evaluation of different segmentation architectures becomes not only timely but necessary. By situating such a comparison within a consistent GeoAI framework, one can reveal the relative merits, limitations, and trade-offs of each model combination specifically for water body mapping. In doing so, we bridge the methodological advances of deep learning with the applied demands of hydrological monitoring and environmental mapping.

In this study, we present a systematic comparative evaluation of four widely used semantic segmentation architectures: U-Net, U-Net++, DeepLabV3+, and Feature Pyramid Network (FPN) for the task of water body segmentation using satellite imagery. While all models are implemented with standardized encoder backbones (ResNet-34) initialized with ImageNet weights, the primary focus lies in assessing the architectural differences among the segmentation models themselves. Utilizing the Earth Surface Water Dataset (Luo et al., 2021) as a benchmark, we investigate both quantitative metrics and qualitative outputs to determine model suitability for operational water mapping applications. The findings aim to guide the remote sensing and GeoAI communities in selecting appropriate segmentation frameworks for hydrological analysis and environmental monitoring tasks.

2. Related Work

In remote sensing, the extraction of surface water has long been addressed via spectral indices (e.g., NDWI, MNDWI) and thresholding, but these methods often fail under heterogeneous landscapes, mixed pixels, or shadow interference. Recent advances in deep learning—particularly convolutional neural networks (CNNs) and their segmentation derivatives—have outperformed classical techniques by learning hierarchical, contextual features automatically (Luo et al., 2022).

Architectures adapted from generic segmentation tasks such as U-Net and DeepLab, are frequently repurposed for water delineation, often augmented with multi-scale fusion, attention modules, or residual connections to boost boundary precision in complex settings (Sun et al., 2024). For wetlands and transitional zones, fusion of optical and SAR modalities has also been proposed to reduce ambiguity and expand generalization (Peña et al., 2024).

Some specialized models have been designed for water mapping. The study by (Cao et al., 2024) presents EU-Net, which builds on a U-Net backbone but incorporates enhanced residual connections, attention modules, and multi-scale dilated convolutions specifically tailored for extracting water bodies from high-resolution imagery. Other studies have modified DeepLab's ASPP module or introduced strip pooling layers to better capture linear water features (Luo et al., 2022).

Beyond purely supervised approaches, there have been efforts in weakly supervised segmentation (e.g. NFANet) and self-supervised models (e.g. DeepAqua) that aim to relax dependence on pixel-level annotation. NFANet, for instance, uses point-level supervision combined with neighbour feature aggregation to infer water boundaries (Lu et al., 2022).
DeepAqua employs knowledge distillation from index-based methods (e.g. NDWI) as pseudo-labelling to train a student CNN, notably in wetland settings (Peña et al., 2023).

The Earth Surface Water Dataset is increasingly adopted as a benchmarking resource for segmentation tasks on Sentinel-2 imagery. It provides large-scale annotated tiles of surface water features, enabling standardized assessment of model performance (Luo et al., 2021). While some works reference this dataset in passing, few studies conduct a controlled comparative evaluation exclusively on it. This underutilization suggests a latent opportunity to establish more rigorous baselines for water segmentation across architectures.

The remote sensing field frequently compares segmentation models like U-Net, DeepLab, and PSPNet. These benchmarks typically measure their success in tasks such as land cover classification or urban feature extraction. Researchers use multispectral or hyperspectral imagery to weigh segmentation accuracy against computational expense. A common difficulty in these studies is that they alter several design elements at once. Changes to preprocessing steps, backbone networks, and data augmentation are often made concurrently. This convoluted approach makes it hard to pin performance gains or losses directly on the core architecture itself.

Deep learning shows great potential for mapping surface water. Yet current research contains significant gaps that hinder a clear understanding of the results. A primary issue is the lack of consistent experimental conditions across studies. Researchers usually change multiple variables at once including the core segmentation model along with different encoder backbones training methods and data preparation techniques. This variability makes it difficult to determine which component is truly responsible for a model's performance. This inconsistent methodology makes it hard to link performance directly to a model's design. Another problem is the underutilization of major public datasets. Even with resources like the Earth Surface Water Dataset available few studies use them in a way that allows for true replication. The scope of comparison is also often narrow. Many analyses only pit a couple of architectures against each other which provides a limited view of their relative strengths and weaknesses. Finally, the practical computational needs for applying these models are frequently ignored. Critical factors like inference speed, model size, and the ability to scale up are just as important as accuracy for real-world use. These limitations collectively underscore the need for a controlled, architecture-focused benchmarking study that evaluates state-of-the-art segmentation models using a standardized framework and dataset. The present work addresses this gap by offering a comparative analysis of four prominent semantic segmentation architectures; U-Net, U-Net++, DeepLabV3+, and FPN, under uniform experimental conditions, thereby contributing a clearer understanding of architectural efficacy in the context of water body mapping.

https://doi.org/10.5194/isprs-annals-X-4-W8-2025-465-2026 | © Author(s) 2026. CC BY 4.0 License.

3. Methodology

3.1 Dataset Description

The Earth Surface Water Dataset is a publicly available benchmark specifically constructed for deep learning–based surface water segmentation on Sentinel-2 imagery (Luo et al., 2021). An example of the data is provided in Figure 1 and Figure 2.

The dataset comprises 95 scenes globally distributed, sampling diverse climatic zones, land cover types, and water dynamics. The global spread helps ensure that models evaluated on it are exposed to a wide range of water–land interaction scenarios. The imagery is drawn from Sentinel-2 multispectral data (with nominal 10 m, 20 m, and 60 m bands), which is well suited for medium-resolution surface water mapping. Each scene is accompanied by a binary water mask annotation (i.e. water vs non-water). The creators assert that the masks are carefully quality-checked to reflect realistic water boundaries across varied settings. However, like many large-scale remote sensing masks, minor boundary inaccuracies or mixed pixels are inevitable, particularly in transition zones or in areas of turbid water, shadowing, or adjacency to vegetation or built surfaces. The authors constructed this dataset to facilitate training and evaluation of deep learning models for surface water mapping worldwide; they also used it to validate their proposed WatNet architecture. Because the scenes sample heterogeneous backgrounds (e.g., urban, rural, mountainous, arid), the dataset is well-suited to benchmark generalization performance across environmental settings. Table 1 summarizes the key attributes of this dataset.

Attribute	Description
Number of Scenes	95
Sensor	Sentinel-2 MSI
Spatial Resolution	10 m (selected bands), 20 m, 60 m (multispectral inputs)
Spectral Bands	Primarily B2 (Blue), B3 (Green), B4 (Red), B8 (NIR), B11, B12
Label Type	Binary masks (water = 1, non-water = 0)
Annotation Method	Semi-automatic method verified against expert knowledge
Coverage	Global (scenes from Asia, Africa, Europe, Americas, and Oceania)
Environmental Diversity	Includes urban, rural, mountainous, arid, tropical, and temperate zones

Table 1. Summary of Earth Surface Water Dataset Attributes.

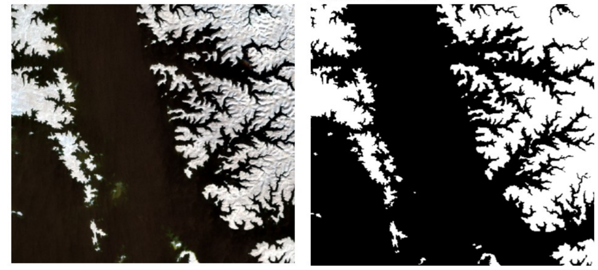

Figure 1. Example input scene (RGB composite) and its corresponding ground truth water mask.

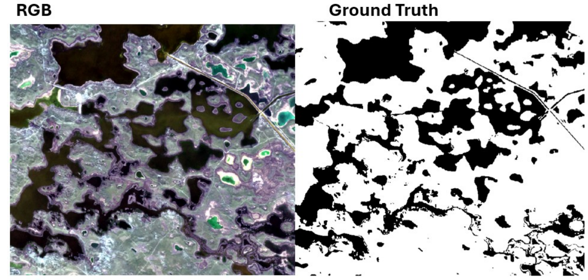

Figure 2. An example scene against its ground truth.

This dataset was selected for our benchmarking study due to its public availability, global representativeness, inclusion of challenging scenarios, and, most importantly, its high-quality manual annotations which serve as an objective ground truth.

3.2 Deep Learning Model Architectures

In this study, we use four architectures: U-Net, U-Net++, DeepLabV3+, and FPN, under the same encoder.

U-Net's symmetric encoder–decoder with skip connections is a strong baseline for geospatial segmentation because it couples broad contextual encoding with precise, high-resolution decoding. The direct skips mitigate information loss from pooling and help delineate thin, sinuous water features (e.g., narrow channels, oxbows) that are easily missed by purely downsampling encoders. Its data efficiency—originally demonstrated on small biomedical datasets—makes it attractive when dense annotations are limited, as is typical in remote sensing (Ronneberger et al., 2015). Figure 3 illustrates the U-Net architecture

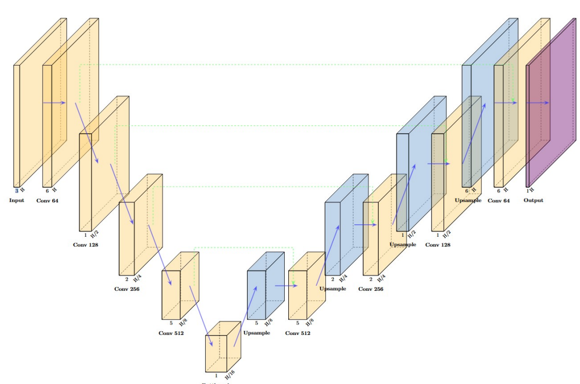

Figure 3. U-Net schematic showing encoder–decoder symmetry with skip connections.

https://doi.org/10.5194/isprs-annals-X-4-W8-2025-465-2026

U-Net++ refines U-Net's skip connections with nested, dense skip pathways and deep supervision to narrow the semantic gap between encoder and decoder representations. In practice, this yields smoother boundaries and fewer "stair-step" artifacts at land–water interfaces, a recurring challenge in heterogeneous shorelines and wetlands. The nested design also encourages multi-scale feature fusion without a large computational penalty relative to substantially deeper backbones (Zhou et al., 2018). Figure 4 illustrates the U-Net++ model.

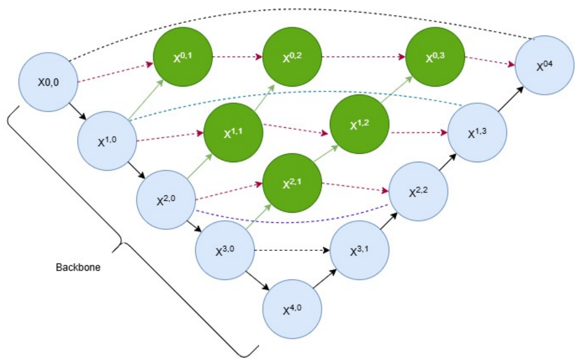

Figure 4. U-Net++ with nested dense skip pathways.

DeepLabV3+ combines Atrous Spatial Pyramid Pooling (ASPP) for rich multi-scale context with a lightweight decoder that sharpens edges as shown in Figure 5. For water mapping—where objects span scales from small irrigation ponds to wide reservoirs, ASPP's dilation rates broaden the effective receptive field, while the decoder improves shoreline crispness. Depthwise separable convolutions reduce cost, enabling higher input resolution or larger batches under fixed compute (Chen et al., 2018).

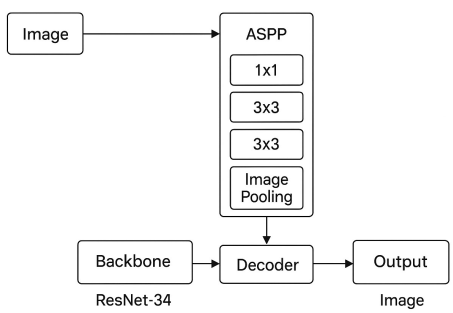

Figure 5. DeepLabV3+ highlighting ASPP for multi-scale context and a lightweight decoder (encoder: ResNet-34).

FPN constructs a top-down, lateral-connected feature pyramid so that all scales carry strong semantics. Although introduced for detection, FPN has become a general-purpose multi-scale backbone/head and underpins many segmentation systems (e.g., Mask R-CNN), making it a pertinent comparator when spatial scale variation is pronounced, as in river networks and floodplains (Lin et al., 2017). An example of simple FPN is shown in Figure 6.

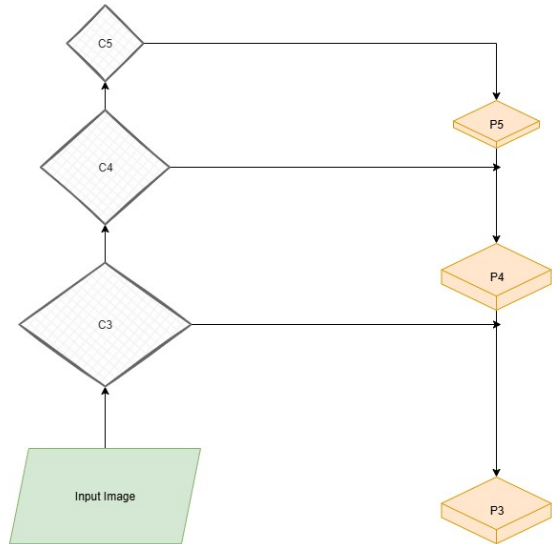

Figure 6. Feature Pyramid Network (FPN) with top-down/lateral fusion producing.

To control model capacity and isolate head-architecture effects, our primary experiments fix ResNet-34 as the encoder across all four heads. ResNet introduces residual mappings that ease optimization of deeper networks; the 34-layer variant offers a balanced trade-off between representational power and computational footprint, and it has been widely used as a solid, ImageNet-initialized backbone in segmentation pipelines. Residual connections improve gradient flow and stabilize training, which is valuable when fitting on geographically diverse yet label-scarce remote sensing data (He et al., 2016).

The selected models represent three distinct design approaches: the encoder-decoder framework with direct skip connections seen in U-Net and UNet++. The approach relying on dilated convolutions for context and a decoder for refining edges like DeepLabV3+ and the top-down pyramid to merge multi-scale features as in FPN. By evaluating them with identical data preparation and training procedures we can connect performance more directly to their architecture. This controlled comparison highlights how each design handles challenges specific to mapping water. We can see which best preserves sharp boundaries, remains robust against mixed pixels and shadows and detects small or narrow water channels. This controlled setup which is paired with a consistent ResNet-34 encoder—provides a fair, practically informative benchmark for GeoAI practitioners. Table 2 illustrates these models' mechanism and advantages.

Architecture	Core mechanism	strengths
U-Net	Symmetric encoder–decoder with one-to-one skip connections	May capture less global context than architectures with dilated convolutions or pyramid designs.
U-Net++	Nested, dense skip pathways; optional deep supervision	Robust delineation in heterogeneous coasts

https://doi.org/10.5194/isprs-annals-X-4-W8-2025-465-2026

DeepLabV3+	Atrous Spatial Pyramid Pooling (ASPP) + lightweight decoder	Captures both small ponds and large reservoirs
FPN	Top-down pathway with lateral 1×1 connections	Efficient scale handling for elongated river networks

Table 2. Comparative summary of segmentation architectures.

4. Implementation

We implemented a standardized GeoAI pipeline comprising four stages: dataset acquisition, tiling, model training, prediction and scoring. This way, architectural effects (U-Net, U-Net++, DeepLabV3+, FPN) can be isolated under identical conditions with a fixed encoder (ResNet-34) and identical optimization settings.

We used the Earth Surface Water (ESW) Dataset distributed via Zenodo. The collection includes 95 globally distributed Sentinel-2 scenes with binary water masks and is released for research and benchmarking. Following the accompanying WatNet resources, inputs are 6 Sentinel-2 bands (B2, B3, B4, B8, B11, B12), which capture visible, NIR, and SWIR information known to be discriminative for water mapping.

To handle large scenes efficiently and increase the number of training samples, we tile each training scene into 512×512-pixel patches with stride 128 pixel, generating paired image–mask crops. The dataset-provided validation split is held out for final model selection and reporting.

We evaluate four segmentation heads: U-Net, U-Net++, DeepLabV3+ and FPN (top-down/lateral feature pyramid)—all instantiated from segmentation-models-pytorch (SMP) and paired with a ResNet-34 encoder initialized from ImageNet. SMP supports arbitrary input channels; when using ImageNet weights with 6-channel inputs, the first convolution is initialized by reusing/replicating pretrained kernels in a documented manner, a strategy that stabilizes optimization while accommodating the multispectral stack.

We optimize with Adam at an initial learning rate of 1e-3, batch size 8, for 20 epochs. For training control, we used an internal 20% validation split of the training tiles; the held-out val_scene split remains untouched until final scoring. Final evaluation reports IoU and Dice per scene (macro-averaged across scenes) against val_truth, along with TP/FP/FN counts to characterize typical failure modes (e.g., turbid water, shoreline ambiguity, shadows). All preprocessing, tiling, optimizer settings, initialization, and seeds are held constant across architectures. The complete set of hyperparameters used for model training is provided in Table 3.

To create training samples, we tiled the training scenes into 512×512 image chips with stride = 128 pixels (i.e., 75% overlap). For inference on full validation scenes, we used a sliding window of 512 with overlap = 256 (50%). We matched image/mask pairs from the dataset's training and validation folders and only tiled training scenes; validation scenes were never included in training tiles, preventing tile-level and scene-level leakage.

Component	**Setting**
Inputs	6 Sentinel-2 bands: B2, B3, B4, B8, B11, B12
Architectures	U-Net, U-Net++, DeepLabV3+, FPN
Encoder	ResNet-34, ImageNet initialization
Optimizer	Adam, lr = 1e-3 (stable baseline for segmentation with moderate batch size).
Batch Size	8
Epochs	20
Loss	Binary cross-entropy with logits
Metrics	IoU, Dice (macro-average per scene); TP/FP/FN counts for error profiling.
Framework / Library	PyTorch + segmentation-models-pytorch (SMP)

Table 3. Hyperparameters and configuration used for training the models.

We evaluate performance using IoU and Dice (macro-averaged per scene), along with TP, FP, and FN counts for error profiling. The formulas for these metrics and their intermediate terms are provided in Equations (1) – (2).

$$IoU = \frac{TP}{TP + FP + FN}, \tag{1}$$

$$Dice = \frac{2TP}{2TP + FP + FN}, \tag{2}$$

In these equations, TP (true positive) denotes instances where the model correctly identifies the positive class. FP (false positive) refers to cases where a negative instance is incorrectly classified as positive. TN (true negative) represents correct identification of the negative class, whereas FN (false negative) corresponds to cases where the model fails to detect the positive class.

https://doi.org/10.5194/isprs-annals-X-4-W8-2025-465-2026 | © Author(s) 2026. CC BY 4.0 License.

5. Results

In this study, the models were trained for 20 epochs and their performance was evaluated using Intersection over Union (IoU) and Sørensen–Dice (Dice), reported per scene and macro-averaged across scenes to avoid size bias. Predictions are produced tile-wise and stitched; logits are converted to binary masks with a fixed threshold (default 0.5). Table 4 lists the models and their evaluation metrics.

Architecture	**IoU**	**Dice**
U-Net	0.9352	0.9598
U-Net++	0.945	0.969
DeepLabV3+	0.955	0.973
FPN	0.9218	0.9514

Table 4. Quantitative summary.

DeepLabV3+ attains the highest IoU/Dice, consistent with the benefit of ASPP for multi-scale context, followed by U-Net++. U-Net provides a strong baseline, while FPN trails slightly in IoU but remains competitive in Dice, indicating reasonable boundary recovery with somewhat elevated FP/FN in challenging shorelines.

To further elucidate performances, a side-by-side of original RGB scene and ground truth with the predictions are presented in Figures 7 to 10.

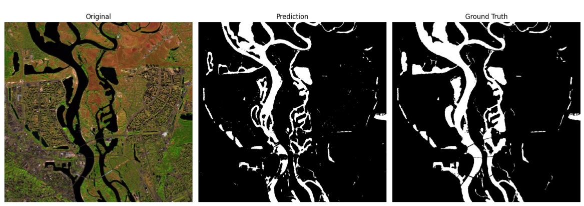

Figure 7. Result comparison for U-Net.

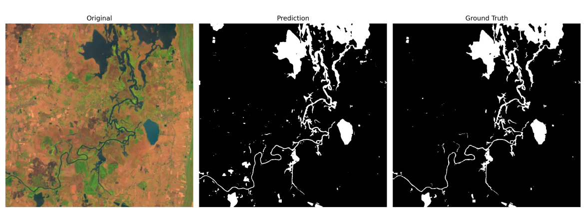

Figure 8. Result comparison for U-Net++.

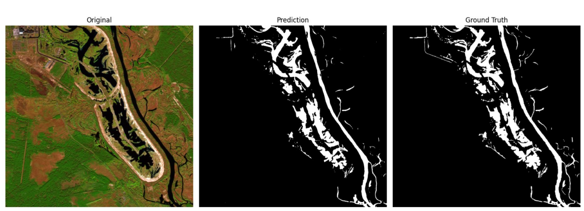

Figure 9. Result comparison for FPN.

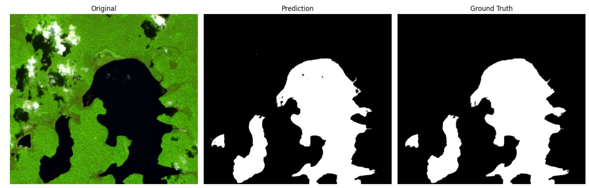

Figure 10. Result comparison for DeepLabV3+.

6. Discussion

The purpose of this study was to evaluate the performance of 4 different deep learning model architectures in water body segmentation. The study results show that after 20 epochs the U-Net model reached 0.9352 IoU, U-Net++ reached 0.945 IoU, DeepLabV3+ reached highest IoU which is 0.955 and the FPN reached 0.9218 IoU

With a fixed ResNet-34 encoder and uniform optimization, DeepLabV3+ achieved the strongest macro-IoU/Dice, followed by U-Net++, then U-Net, with FPN close but generally lower in IoU. This ordering is consistent with the architectural priors each head imposes: DeepLabV3+ aggregates multi-scale context via atrous spatial pyramid pooling (ASPP) and then sharpens boundaries with a light decoder, a combination repeatedly found to improve class separation while preserving edges in dense prediction tasks. These results are therefore in line with (Chen et al., 2018) who report boundary gains when pairing ASPP with an explicit decoder.

All four heads perform well; DeepLabV3+ is typically best because wide water bodies benefit from large effective receptive fields and robust context aggregation, which mitigate confusion from surrounding shadows or dark surfaces.

U-Net++ (and, to a lesser extent, U-Net) often shows superior boundary fidelity and recall for thin, sinuous features because dense/nested skip pathways (U-Net++) or direct skip connections (U-Net) transmit high-resolution detail to the decoder. This observation is in line with the evidence from (Zhou et al., 2018) that U-Net++ reduces the encoder–decoder semantic gap and improves delineation of small structures over U-Net.

FPN's top-down pyramid provides efficient scale handling, but without a stronger decoder its masks can be smoother and slightly under-segmented along intricate shorelines; this aligns with FPN's original design as a feature pyramid later paired with task-specific heads rather than a purpose-built semantic decoder.

Residual errors concentrate in (i) turbid or low-contrast water (false negatives), (ii) shoreline adjacency and shadows (false positives), and (iii) channels near the 10 m Sentinel-2 resolution, where sub-pixel widths limit recoverable detail. Similar challenges are reported in recent water-mapping studies and reviews, and several works explicitly note the difficulty of detecting very narrow rivers at 10 m GSD. Our qualitative failure cases therefore echo broader findings in the literature (Billson et al., 2023).

For large open-water targets or heterogeneous landscapes where context is critical, DeepLabV3+ is recommended. For fine hydrography (small ponds, narrow or braided channels) where boundary sharpness is paramount, U-Net++ is often preferable; U-Net remains a strong, compute-efficient baseline. FPN is attractive when a unified multi-scale backbone is desired across tasks, but pairing it with a more expressive decoder may be warranted for sharper shorelines.

Our comparison isolates head architecture by fixing the encoder (ResNet-34) and training recipe; while this strengthens attribution, it may cap absolute performance relative to larger or transformer encoders and omits task-specific tuning. Future work could (i) assess robustness across sensors, (ii) evaluate super-resolution or multi-temporal cues for sub-pixel channels, and (iii) test architecture-agnostic post-processing to further reduce shoreline errors.

7. Conclusion

This work introduced a GeoAI framework to comparatively evaluate four semantic-segmentation architectures: U-Net, U-Net++, DeepLabV3+, and FPN; for water-body mapping on the Earth Surface Water dataset (six-band Sentinel-2). Holding encoder (ResNet-34), initialization (ImageNet), tiling, and optimization constant, DeepLabV3+ achieved the highest macro-averaged performance (IoU ≈ 0.955, Dice ≈ 0.973), followed by U-Net++ (IoU ≈ 0.945, Dice ≈ 0.969), U-Net (IoU ≈ 0.9352, Dice ≈ 0.9598), and FPN (IoU ≈ 0.9218, Dice ≈ 0.9514). Qualitative analyses showed that all models accurately segment large lakes and reservoirs, while errors concentrate along complex shorelines, turbid waters, shadows, and sub-pixel channels.

Multi-scale context aggregation with a decoder (DeepLabV3+) confers a consistent advantage on heterogeneous scenes; nested skip pathways (U-Net++) improve boundary fidelity for narrow rivers and small ponds; plain skip-based encoder–decoder (U-Net) remains a strong, compute-efficient baseline; and top-down pyramids (FPN) are competitive in Dice but can under-segment intricate shorelines without a stronger decoder. Per-scene macro-averaging provides a more equitable assessment than pixel-weighted averages on geographically varied tiles.

Error typology is repeatable across heads, suggesting benefits from boundary-aware objectives and modest post-processing rather than wholesale architectural changes for certain regimes.

The study fixes a single encoder (ResNet-34) and uses one dataset, which strengthens internal validity but may limit external generalization. Future work should (i) extend comparisons to larger CNN/Transformer encoders and multi-temporal inputs; (ii) evaluate sensor fusion (e.g., SAR–optical) for turbid or shadowed waters; (iii) incorporate boundary-aware or topology-preserving losses, uncertainty estimation, and light post-processing (e.g., CRFs, contour refiners); (iv) assess domain adaptation across regions/climates; and (v) report efficiency metrics alongside accuracy to guide operational deployment. Collectively, these directions will deepen understanding of architecture–data interactions and advance robust, scalable water-mapping in Earth observation workflows.

References

Aekakkararungroj, A., Chishtie, F., Poortinga, A., Mehmood, H., Anderson, E., Munroe, T., Cutter, P., Loketkawee, N., Tondapu, G., Towashiraporn, P., Saah, D., 2020. A publicly available GIS-based web platform for reservoir inundation mapping in the lower Mekong region. Environmental Modelling & Software Abbreviation, 123(104552. doi.org/https://doi.org/10.1016/j.envsoft.2019.104552.

Billson, J., Islam, M. S., Sun, X., Cheng, I., 2023. Water body extraction from sentinel-2 imagery with deep convolutional networks and pixelwise category transplantation. Remote Sensing Abbreviation, 15(5), 1253.

Boutayeb, A., Lahsen-cherif, I., Khadimi, A. E., 2024. A comprehensive GeoAI review: Progress, Challenges and Outlooks. arXiv preprint arXiv:2412.11643 Abbreviation,

Cao, H., Tian, Y., Liu, Y., Wang, R., 2024. Water body extraction from high spatial resolution remote sensing images based on enhanced U-Net and multi-scale information fusion. Scientific Reports Abbreviation, 14(1), 16132.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV),

Flores-Anderson, A. I., Herndon, K. E., Thapa, R. B., Cherrington, E., 2019. *The SAR handbook: Comprehensive methodologies for forest monitoring and biomass estimation*.

Gao, B.-c., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment Abbreviation, 58(3), 257-266. doi.org/https://doi.org/10.1016/S0034-4257(96)00067-3.

Gonzales-Inca, C., Calle, M., Croghan, D., Torabi Haghighi, A., Marttila, H., Silander, J., & Alho, P., 2022. Geospatial artificial intelligence (GeoAI) in the integrated hydrological and fluvial systems modeling: Review of current applications and trends. Water Abbreviation, 14(14), 2211.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition,

Huang, W., DeVries, B., Huang, C., Lang, M. W., Jones, J. W., Creed, I. F., & Carroll, M. L., 2018. Automated Extraction of Surface Water Extent from Sentinel-1 Data. Remote Sensing Abbreviation, 10(5), 797. https://www.mdpi.com/2072-4292/10/5/797

Iyer, H. S., Karasaki, S., Yi, L., Hswen, Y., James, P., VoPham, T., 2025. Harnessing Geospatial Artificial Intelligence (GeoAI) for Environmental Epidemiology: A Narrative Review. Curr Environ Health Rep Abbreviation, 12(1), 34. doi.org/10.1007/s40572-025-00497-4.

Li, L., Long, D., Wang, Y., Woolway, R. I., 2025. Global dominance of seasonality in shaping lake-surface-extent dynamics. nature Abbreviation, 1-8.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,

Lu, M., Fang, L., Li, M., Zhang, B., Zhang, Y., Ghamisi, P., 2022. NFANet: A novel method for weakly supervised water extraction from high-resolution remote-sensing imagery. IEEE transactions on geoscience and remote sensing Abbreviation, 60(1-14.

Luo, X., Tong, X., Hu, Z., 2021. An applicable and automatic method for earth surface water mapping based on multispectral images. International Journal of Applied Earth Observation and Geoinformation Abbreviation, 103(102472).

Luo, Y., Feng, A., Li, H., Li, D., Wu, X., Liao, J., Zhang, C., Zheng, X., Pu, H., 2022. New deep learning method for efficient extraction of small water from remote sensing images. Plos one Abbreviation, 17(8), e0272317.

Maniyar, C. B., Yan, X., Mai, G., Srivastava, D., Samiappan, S., Oliazadeh, A., Kumar, A., Kumar, M., Mishra, D. R., 2025. Artificial intelligence in environmental remote sensing: Progress, way forward and key considerations. Progress in Environmental Geography Abbreviation, 27539687251357020.

Markert, K. N., Markert, A. M., Mayer, T., Nauman, C., Haag, A., Poortinga, A., Bhandari, B., Thwal, N. S., Kunlamai, T., Chishtie, F., Kwant, M., Phongsapan, K., Clinton, N., Towashiraporn, P., Saah, D., 2020. Comparing Sentinel-1 Surface Water Mapping Algorithms and Radiometric Terrain Correction Processing in Southeast Asia Utilizing Google Earth Engine. Remote Sensing Abbreviation, 12(15), 2469. https://www.mdpi.com/2072-4292/12/15/2469

Nasirzadehdizaji, R., Balik Sanli, F., Abdikan, S., Cakir, Z., Sekertekin, A., Ustuner, M., 2019. Sensitivity Analysis of Multi-Temporal Sentinel-1 SAR Parameters to Crop Height and Canopy Coverage. Applied Sciences Abbreviation, 9(4), 655. https://www.mdpi.com/2076-3417/9/4/655

Nemni, E., Bullock, J., Belabbes, S., Bromley, L., 2020. Fully Convolutional Neural Network for Rapid Flood Segmentation in Synthetic Aperture Radar Imagery. Remote Sensing Abbreviation, 12(16), 2532. https://www.mdpi.com/2072-4292/12/16/2532

Oddo, P. C., Bolten, J. D., 2019. The value of near real-time earth observations for improved flood disaster response. Frontiers in Environmental Science Abbreviation, 7(127.

Peña, F. J., Hübinger, C., Payberah, A. H., Jaramillo, F., 2023. Deepaqua: Self-supervised semantic segmentation of wetland surface water extent with sar images using knowledge distillation. arXiv preprint arXiv:2305.01698 Abbreviation,

Peña, F. J., Hübinger, C., Payberah, A. H., Jaramillo, F., 2024. DeepAqua: Semantic segmentation of wetland water surfaces with SAR imagery using deep neural networks without manually annotated data. International Journal of Applied Earth Observation and Geoinformation Abbreviation, 126(103624. doi.org/https://doi.org/10.1016/j.jag.2023.103624.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention,

Sun, D., Gao, G., Huang, L., Liu, Y., Liu, D., 2024. Extraction of water bodies from high-resolution remote sensing imagery based on a deep semantic segmentation network. Scientific Reports Abbreviation, 14(1), 14604.

Valentin, C., Agus, F., Alamban, R., Boosaner, A., Bricquet, J. P., Chaplot, V., de Guzman, T., de Rouw, A., Janeau, J. L., Orange, D., Phachomphonh, K., Do Duy, P., Podwojewski, P., Ribolzi, O., Silvera, N., Subagyono, K., Thiébaux, J. P., Tran Duc, T., Vadari, T., 2008. Runoff and sediment losses from 27 upland catchments in Southeast Asia: Impact of rapid land use changes and conservation practices. Agriculture, Ecosystems & Environment Abbreviation, 128(4), 225-238. doi.org/https://doi.org/10.1016/j.agee.2008.06.004.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. International workshop on deep learning in medical image analysis,