Kalpa Publications in Computing

Volume 22, 2025, Pages 207–221

Proceedings of The Sixth International Conference on Civil and Building Engineering Informatics

Lightweight Transfer Learning for Water Body Segmentation Using Adaptor-Based Fine-Tuning

Jiapei Zhao[1*], Nobuyoshi Yabuki[2], Tomohiro Fukuda[3]

1) Ph.D. Student, Division of Sustainable Energy and Environmental Engineering, Osaka University, Osaka, Japan. Email: u812893b@ecs.osaka-u.ac.jp

2) Ph.D., Prof., Division of Sustainable Energy and Environmental Engineering, Osaka University, Osaka, Japan. Email: yabuki@see.eng.osaka-u.ac.jp

3) Ph.D., Assoc. Prof., Division of Sustainable Energy and Environmental Engineering, Osaka University, Osaka, Japan. Email: fukuda.tomohiro.see.eng@osaka-u.ac.jp

Abstract

This paper presents a lightweight transfer learning approach for water body segmentation by applying Adaptor-based fine-tuning on general image datasets. Traditional deep learning models often require full-scale retraining for each new task, which is computationally expensive and time-consuming. In contrast, Adaptor networks—lightweight modules that selectively fine-tune task-specific layers while retaining most pre-trained model parameters—offer an efficient alternative. Water bodies present unique challenges for segmentation, such as varying lighting, reflections, and seasonal fluctuations. These factors can confuse distinguishing water from land, particularly in cases where reflections resemble adjacent features. Adaptor-based fine-tuning helps to reduce computational costs while ensuring the model captures the fine distinctions between similar regions like shallow water and land. This paper evaluated the method on the ATLANTIS dataset, which includes diverse categories of water bodies such as lakes, rivers, and wetlands. This dataset is recognized as a comprehensive collection for evaluating semantic segmentation performance in varied environmental conditions. The results indicate that Adaptor-based fine-tuning achieves comparable performance to fully fine-tuned models, with a significant reduction in computational costs and training time. The method also demonstrated high precision in segmenting water bodies under challenging conditions, such as occlusions and reflections. This study highlights the potential of lightweight transfer learning in resource-constrained environments, with applications in environmental monitoring, hydrological modeling, and geographic information systems (GIS). By demonstrating the effectiveness of Adaptor networks, this work contributes to the broader field of efficient transfer learning, showcasing how minimal adjustments to pre-trained models can yield accurate task-specific performance.

I. Hkust, J. Cheng and Y. Yantao (eds.), ICCBEI 2025 (Kalpa Publications in Computing, vol. 22), pp. 207–221

1 Introduction

Water body segmentation is a fundamental task in environmental monitoring, particularly for applications in geographic information systems (GIS), hydrological modeling, and urban planning (Kadhim & Premaratne, 2023). The increasing frequency of extreme weather events due to climate change makes accurate segmentation of water bodies crucial for disaster management, urban flood prediction, and resource allocation (Zaffaroni & Rossi, 2020). Satellite imagery provides an efficient means for this purpose, but the complexity of natural environments, including reflections, occlusions, and seasonal variations, complicated the accurate delineation of water boundaries (Saleh et al., 2018).

Traditional deep learning approaches, such as convolutional neural networks (CNNs), have been widely employed for semantic segmentation (Pinaya et al., 2020; Yuan et al., 2021). However, these methods often require a complete retraining process for each new segmentation task, making them computationally expensive and unsuitable for scenarios with limited computational resources (Chen et al., 2017; Zhao et al., 2017). Moreover, the need for vast labeled datasets and the high computational cost associated with full-scale retraining limit the scalability of these models to new environments and tasks. In contrast, recent advancements leverage Vision Transformers (ViTs), which have shown great promise for segmentation tasks due to their ability to model long-range dependencies efficiently (Dosovitskiy, 2020).

To address these challenges, we were the first to apply Adaptor-based fine-tuning with Vision Transformer (ViT) architecture to the task of waterbody segmentation, proposing a lightweight transfer learning approach. This approach leverages Vision Transformers (ViTs), specifically the SegFormer architecture (Xie et al., 2021), which has shown promise in capturing multi-scale features suitable for segmentation (Xie et al., 2021). By selectively fine-tuning task-specific layers using lightweight Adaptor modules (Liu et al., 2023), this method retains most pre-trained model parameters while adapting efficiently to the unique characteristics of water bodies, such as their dynamic boundaries and varying appearances. This strategy effectively reduces the computational load, facilitating the deployment of the model in resource-constrained environments (Dong et al., 2023). The experimental evaluation presented in this paper highlights the efficacy of our method when applied to 15 different water body categories, demonstrating its robustness and efficiency in real-world scenarios.

2 Related Work

Recent advancements in semantic segmentation have seen a shift from conventional convolutional neural networks (CNNs) to transformer-based architectures, such as Vision Transformers (ViTs) (Dosovitskiy, 2020). These models have shown notable improvements in capturing global context and multi-scale information, which are essential for accurately delineating objects. The SegFormer model, in particular, has gained prominence for its hierarchical transformer architecture, which is effective for extracting multi-scale features suitable for segmentation tasks (Xie et al., 2021).

Adaptor-based fine-tuning methods have also been explored in natural language processing (NLP) applications, such as BERT (Kenton & Toutanova, 2019) and T5 (Raffel et al., 2020), to reduce computational costs while retaining model performance. This technique has gradually been introduced to the computer vision domain to address similar challenges (Dong et al., 2023). This approach draws upon these advancements, extending the idea of adaptor modules to the segmentation of water bodies—a task characterized by diverse and challenging environmental conditions, including fluctuating water levels, occlusions, and reflections.

Prior studies have utilized datasets such as COCO for semantic segmentation of multiple categories. For instance, the ATLANTIS dataset, proposed by Erfani et al., (2022), contains 5,195 training images and 1,296 testing images, with a wide variety of water body types. This study focuses on 15 specific categories from ATLANTIS to validate the method. Compared to state-of-the-art models like DNLNet, GCNet, and AQUANet (Cao et al., 2019; Erfani et al., 2022; Ni et al., 2022), this adaptor-based method offers improved performance by selectively fine-tuning high-frequency components and embedding features critical to water body segmentation.

3 Method

In this section, a lightweight transfer learning framework for water body segmentation by applying Adaptor-based fine-tuning is introduced. The framework utilizes Vision Transformers pre-trained on large-scale image datasets such as ImageNet (Deng et al., 2009). This method focuses on selectively fine-tuning layers critical to task-specific segmentation, which allows to avoid the computational cost of full-scale retraining. By targeting high-frequency components and image embeddings, the model effectively adapts to challenges such as varying lighting conditions, occlusions, and reflections in water body segmentation.

Figure 1 presents an overview of the Adaptor-based fine-tuning approach. The frozen pretrained Transformer model, represented with an ice symbol, indicates the components that remain unchanged during the training process. The Adaptor module, marked with a flame symbol, highlights the elements that are specifically fine-tuned to adapt to the target water body segmentation task. This Adaptor module leverages the generalization capabilities of the pretrained model while enabling efficient adaptation to the specific requirements of water body segmentation. The input image undergoes processing through the model, producing a segmented output that accurately delineates the water body.

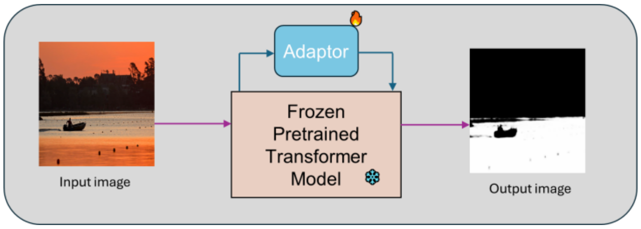

Figure 1. Overview of Adaptor-Based Fine-Tuning for Water Body Segmentation

3.1 Pre-trained Model Architecture

This approach builds upon the hierarchical structure of SegFormer (Xie et al., 2021), a Vision Transformer (ViT) model renowned for its multi-scale feature extraction. The encoder of SegFormer is composed of a series of transformer blocks that progressively capture finer image details. Each block outputs features at different resolutions, providing multi-level representations.

As Figure 1 shows, given that the pre-trained model already captures essential structural features, its backbone layers are frozen to retain general visual knowledge learned from ImageNet. Only a small number of task-specific layers are leveraged for fine-tuning. These layers learn to specialize in water body segmentation without the need to modify the majority of the pre-trained parameters, drastically reducing computational overhead. The frozen layers are shown in Equation (1):

$$\theta_f \in \mathrm{R}^d \tag{1}$$

where θ_f represents the frozen parameter set, the symbol R represents the set of real numbers, and d is the number of parameters in the backbone.

3.2 High-frequency Component Extraction

One of the core aspects of the Adaptor model is the extraction and fine-tuning of high-frequency components (HFC) (Wang et al., 2020), which capture fine structural details crucial for water body segmentation. These high-frequency components are extracted using the Fourier Transform. The process decomposes the input image into low- and high-frequency components, where the high-frequency components focus on sharp boundaries and textures, essential for delineating water from land.

$$\mathrm{f}(\tau) = \begin{cases} 1, & \text{if } \frac{4\left(\frac{H}{2}-x\right)\left(\frac{W}{2}-y\right)}{HW} \leq \tau \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

As Equation (2) shows, H and W are the height and width of the image, x and y are the coordinates of a given pixel, and τ controls the frequency threshold. After applying this mask, the high-frequency component is obtained using the inverse Fourier transform using Equation (3). The high-frequency components I_{hf} are computed by applying a binary mask M_{hf}, which retains frequencies above a threshold τ. fft is the Fourier transform and $ifft$ is its inverse, and I represents the input image.

$$I_{hf} = ifft\left(M_{hf}\right) \cdot fft(I) \tag{3}$$

3.3 Adaptor-based Fine-tuning

Figure 2 shows the structure of an Adaptor module used for fine-tuning. The tunable components (marked with flames) are updated. The Embedding Tune and HFC Tune modules adapt the embeddings and high-frequency features, respectively, and their outputs are combined. The GELU activation introduces non-linearity. The MLP_{tune}^{i} is specifically fine-tuned, while MLP_{up} layer, both shared and unshared, manage up-projection, allowing for a balance between shared learning and task-specific adaptation.

Figure 3 illustrates the integration of Adaptor modules within a Vision Transformer (ViT) for water body segmentation. It includes both frozen and tunable components: the frozen parts (marked with an ice crystal) remain unchanged during training, while the tunable components (marked with flames) are updated. The model starts with patch embedding of the input dataset, followed by a series of transformer layers that leverage frozen pre-trained features. The key innovation lies in integrating Adaptors between these transformer layers. Each Adaptor is specifically fine-tuned to handle unique aspects of water body segmentation, such as distinguishing between water and non-water regions under varying environmental conditions. This selective fine-tuning mechanism effectively balances computational efficiency with task-specific adaptability, resulting in robust segmentation outcomes suitable for diverse real-world scenarios.

The core contribution of this work is the Adaptor-based fine-tuning mechanism, which allows selective tuning of the frozen model layers. The Adaptors are lightweight modules designed to modify both the image embeddings and the high-frequency components in a task-specific manner.

Two types of tunable components are defined. One is Embedding Tune: This module fine-tunes the image embeddings E by learning a projection to a lower-dimensional space using the Equation (4):

$$F_{pe} = L_{pe}\ (E) \tag{4}$$

where L_{pe} is a linear layer that maps the embeddings to a task-specific feature space, and F_{pe} is the fine-tuned embedding output. Another one is HFC Tune: For high-frequency components I_{hf}, an additional layer of fine-tuning is applied. The high-frequency patches are mapped into a low-dimensional representation through a linear layer using the Equation (5):

$$F_{hfc} = L_{hfc}(I_{hf}) \tag{5}$$

where, L_{hfc} is a linear transformation that compresses the high-frequency information into a task-specific form, F_{hfc}, which is then combined with the embeddings.

The final feature used for segmentation is the combination of both embedding and high-frequency tuned components using the Equation (6):

$$p_i = MLP_{adaptor}\left(GELU\left(F_{pe} + F_{hfc}\right)\right) \tag{6}$$

where p_i is the prompt for the *i*-th transformer layer and $MLP_{adaptor}$ is a multi-layer perceptron tasked with merging these features. The activation function $GELU$ introduces non-linearity, and the summed features are passed to subsequent layers in the transformer.

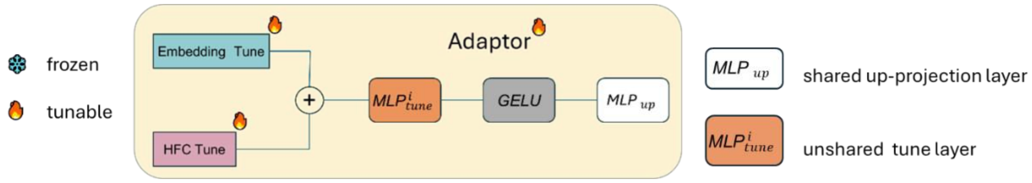

Figure 2. The Architecture of Adaptor-Based Fine-Tuning Modules

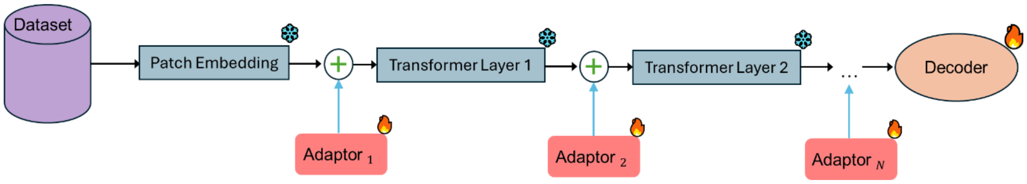

Figure 3. Adaptor Integration in Vision Transformer for Water Body Segmentation

3.4 Efficiency of the Adaptor Network

The Adaptor network works by adjusting only the critical layers responsible for handling task-specific challenges in segmentation, such as reflections or occlusions that are common in water body images. This selective fine-tuning drastically reduces the computational cost compared to traditional methods that retrain entire models.

Given the task of water body segmentation, which often involves challenging scenarios such as subtle distinctions between water and land under different environmental conditions, the fine-tuned high-frequency and embedding components enhance the model's robustness. The compact nature of the Adaptor enables efficient transfer learning without sacrificing performance.

The efficiency of the model can be quantified by comparing the total number of parameters trained in traditional full-scale fine-tuning versus the parameters trained in the Adaptor-based method. θ_t denotes the parameters trained in a traditional method and θ_a denotes the parameters trained in the Adaptor-based method, achieving a significant reduction in trainable parameters, as Equation (7) shows:

$$\frac{\theta_a}{\theta_t} \ll 1 \tag{7}$$

where θ_a is typically an order of magnitude smaller than θ_t. The computational complexity is further reduced due to the focused fine-tuning of high-frequency components. Instead of performing a full forward and backward pass for all model layers, the complexity is concentrated in the computation of the high-frequency components and their subsequent fine-tuning, as Equation (8) shows.

$$O(Adaptor) = O(L_{pe}) + O(L_{hfc}) \tag{8}$$

$O(Adaptor)$ represents the overall computational complexity of the Adaptor-based method. $O(L_{pe})$ represents the computational complexity of the linear layer L_{pe} used for fine-tuning the embeddings. $O(L_{hfc})$ represents the computational complexity of the linear layer L_{hfc} used for fine-tuning the high-frequency components. This complexity remains manageable, even for large images, as only the most important task-specific features are adjusted, making the method suitable for real-world applications in resource-constrained environments.

4 Experiments

The experiments was conducted using Adaptor-based fine-tuning on the SegFormer model as backbone. The training was performed on a single NVIDIA GeForce RTX 3080 GPU for 50 epochs, using a batch size of 4 and an initial learning rate of 0.0002, with the Adam optimizer.

4.1 Dataset

The ATLANTIS dataset (Erfani et al., 2022) is recognized as the largest annotated collection for semantic segmentation of water bodies and their related facilities, consisting of 56 categories of 5,195 training images and 1,296 testing images as Table 1 shows. This study specifically focused on 15 distinct water body categories without their related facilities from the ATLANTIS dataset, as Table 2 shows. This subset includes 1,609 training images, 260 validation images, and 662 testing images. The images are resized to 352 × 352 for consistent evaluation. For performance assessment, the mean Intersection over Union (mIoU) is utilized as the evaluation metric.

Table 1. Complete List of Categories from the ATLANTIS Dataset

Number	Label	Number	Label	Number	Label	Number	Label
0	background	1	bicycle	2	boat	3	Breakwater
4	bridge	5	building	6	bus	7	canal
8	car	9	cliff	10	culvert	11	cypress tree
12	dam	13	ditch	14	fence	15	hydrant
16	fjord	17	flood	18	glaciers	19	hot spring
20	lake	21	levee	22	lighthouse	23	mangrove
24	marsh	25	motorcycle	26	offshore	27	parking
28	person	29	pier	30	pipeline	31	pole
32	puddle	33	rapids	34	reservoir	35	river
36	river delta	37	road	38	sea	39	ship
40	shoreline	41	sidewalk	42	sky	43	snow
44	spillway	45	swimming pool	46	terrain	47	traffic sign
48	train	49	truck	50	umbrella	51	vegetation

52	wall	53	water tower	54	water well	55	waterfall
56	wetland						

Table 2. Selected Water Body Categories Used for Segmentation in This Study

Number	Label	Number	Label	Number	Label	Number	Label
0	background	7	canal	13	ditch	16	fjord
17	flood	19	hot spring	20	lake	32	puddle
33	rapids	34	reservoir	35	river	36	River delta
38	sea	45	swimming pool	55	waterfall	56	wetland

4.2 Data Preprocessing

In the data preprocessing stage, the raw images were transformed to facilitate effective segmentation. As illustrated in Figure 4, the original images were annotated with multi-class masks, distinguishing between various elements within the scene, including different types of water bodies and non-water features. To enhance the focus of water body segmentation and simplify the model's learning process, these original multi-class masks were converted into binary masks. This conversion produced processed masks, where the water bodies were distinguished from the background, thereby framing the segmentation task as a foreground-background classification problem.

This approach reduced the complexity of the segmentation model by focusing solely on the target water bodies, thus allowing the model to concentrate on differentiating water from non-water areas without the need to classify multiple object categories. The preprocessing pipeline ensured that each image was represented consistently, facilitating robust training of the segmentation network across diverse types of water body environments.

4.3 Comparison with Existing Models

The results of the introduced Adaptor-based fine-tuning method for water body segmentation are presented in Table 3 and Table 4. The evaluation metric used is Intersection over Union (IoU), which measures the accuracy of segmentation by comparing the predicted output with the ground truth data.

Table 3 presents the performance comparison of this method, Adaptor, against various state-of-the-art models such as DNLNet, GCNet, DeepLabv3, and AQUANet across 15 water body categories. The Adaptor method demonstrates a clear improvement in the mean IoU (mIoU) with 79.38%, outperforming all the baseline models significantly. For example, fjord: The Adaptor method achieves an IoU of 91.7%, which is much higher compared to DNLNet (48.8%) and GCNet (44.7%). River delta: Adaptor achieves 88.8% IoU, surpassing all other methods, including PSPNet (65.5%) and OCNet (65.9%). Flood: The performance of the Adaptor method is 84.6%, whereas models like CCNet and EMANet achieve only 26.9% and 23.0%, respectively.

These results indicate that the selective fine-tuning of task-specific layers using Adaptor-based networks effectively captures the unique characteristics of water bodies under varying conditions, such as occlusions and reflections, leading to superior segmentation accuracy.

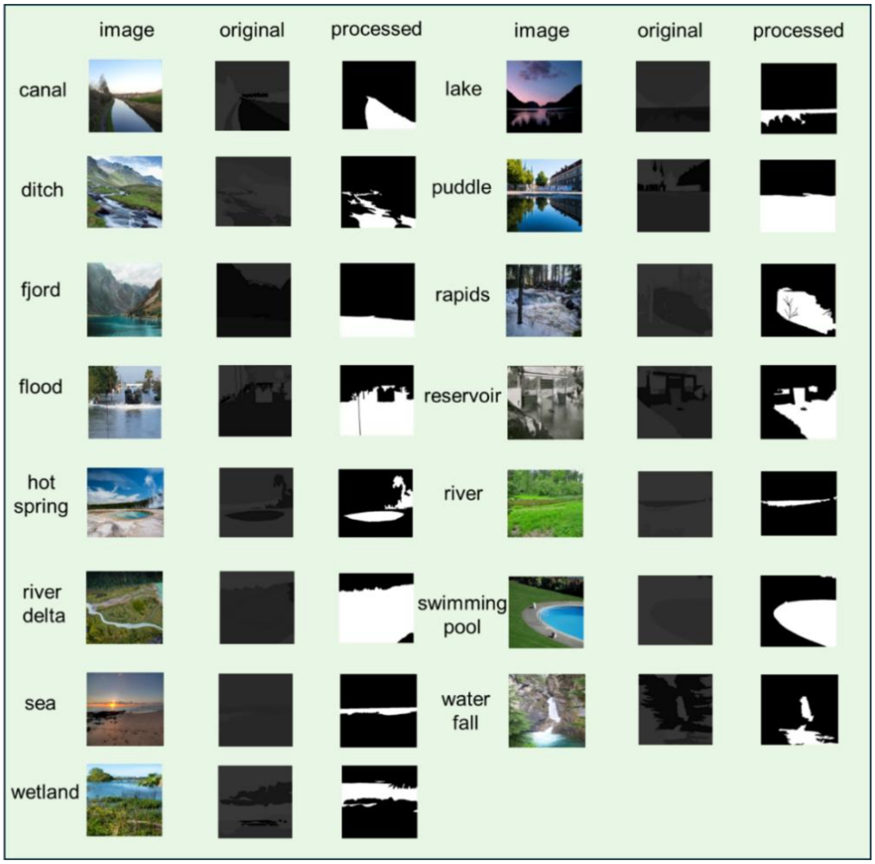

Figure 4. Transformation of Original Dataset Labels into Processed Masks for Model Training and Testing

4.4 Comparison with Different Fine-Tuning Methods

Figures 5, 6, and 7 demonstrate the performance of different fine-tuning methods, namely AdaptFormer, Linear, and Adaptor, on water body segmentation across various types of water bodies. Each figure presents a comparison of segmentation results for distinct water body categories, highlighting the Intersection over Union (IoU) values achieved by each method.

Table 4 illustrates the performance comparison between the Adaptor method and other related approaches, Adaptformer, and the Linear fine-tuning method, all based on the same backbone model, Segformer. The experimental results show that the Adaptor method consistently outperforms both Adaptformer and the Linear approach across various water body categories. Specifically, in the canal category, the Adaptor method achieved an Intersection over Union (IoU) score of 83.6%, significantly surpassing Adaptformer (26.0%). In the lake category, the Adaptor method reached an IoU of 90.9%, compared to Adaptformer at 37.1% and Linear at 85.1%. For the reservoir category, the Adaptor method achieved an IoU of 82.7%, while the Linear method performed slightly lower, with an IoU of 75.2%. The results demonstrate that the Adaptor-based fine-tuning strategy is more effective than both Adaptformer and Linear, in challenging water body categories. The method's ability to tune high-frequency components and embedding features enables it to handle variations in water appearance, such as seasonal changes and reflective surfaces.

The experimental results clearly highlight the strength of the Adaptor-based fine-tuning method in achieving high segmentation accuracy across diverse water body types. By focusing on selective tuning, the approach significantly reduces computational costs compared to full-scale retraining while maintaining a high level of accuracy. These results demonstrate the feasibility of using Adaptor-based models for real-world water body segmentation tasks, offering both efficiency and robustness.

Table 3. Performance Comparison of State-of-the-Art Models for Water Body Segmentation, the best result is bold

Method	IoU (%)															
	canal	ditch	fjord	flood	hot spring	lake	puddle	rapids	reservoir	river	river delta	sea	swimming pool	waterfall	wetland	mIoU (%)
DNLNet	54.4	26.3	48.8	36.3	55.3	35.5	52.3	40.4	32.1	31.3	37.1	61.7	52.4	48.7	54.6	44.48
GCNet	56.6	19.0	44.7	34.8	36.1	35.8	39.4	39.9	41.6	32.4	67.0	62.2	42.9	50.7	59.7	44.19
OCRNet	52.4	19.4	46.9	34.9	58.8	30.4	39.7	42.5	29.8	31.9	55.5	55.4	43.6	56.8	51.5	43.30
CCNet	41.1	17.4	35.2	26.9	47.9	18.6	43.8	29.9	16.6	23.7	48.3	53.3	38.4	51.1	34.1	35.09
EMANet	46.1	16.6	27.1	23.0	63.7	17.2	43.6	42.2	17.2	21.0	68.6	53.5	36.1	52.1	36.2	37.61
ANNet	50.9	22.8	31.6	32.0	58.1	25.6	52.9	48.4	20.8	28.6	56.8	60.4	43.9	57.9	51.4	42.81
DANet	50.5	34.1	37.1	37.0	61.6	23.8	51.5	42.8	30.2	31.5	63.5	60.4	43.1	55.2	54.6	45.13
DeepLab v3	52.5	27.2	52.3	43.8	42.5	31.1	54.2	46.0	34.4	27.1	51.1	61.5	53.6	52.8	52.9	45.40
PSPNet	53.8	29.0	42.9	46.5	53.9	29.7	54.7	38.2	29.8	28.8	65.5	63.5	47.7	48.4	47.5	45.33
OCNet	56.4	33.6	48.0	37.3	55.2	29.2	50.6	43.8	35.1	35.6	65.9	62.7	47.9	53.1	54.9	47.29
AQUANet	55.0	27.7	53.4	47.0	60.5	33.2	54.4	46.3	39.0	34.7	63.2	64.2	44.9	53.0	66.1	49.51
Adaptor	**83.6**	**58.1**	**91.7**	**84.6**	**74.1**	**90.9**	**74.9**	**86.1**	**82.7**	**84.2**	**88.8**	**86.7**	**61.5**	**71.9**	**72.4**	**79.38**

Table 4. Performance Comparison of Adaptor-Based Fine-Tuning with Other Methods, the best result is bold

Method	IoU (%)															mIoU (%)
	canal	ditch	fjord	flood	hot spring	lake	puddle	rapids	reservoir	river	river delta	sea	swimming pool	waterfall	wetland	
adaptformer	26.0	12.6	30.4	19.1	25.4	37.1	13.3	33.0	32.9	31.5	24.8	35.1	26.6	31.0	24.6	26.68
linear	79.3	50.2	83.2	76.9	68.7	85.1	59.2	79.0	75.2	78.2	78.4	82.1	59.1	68.7	65.0	72.65
Adaptor	**83.6**	**58.1**	**91.7**	**84.6**	**74.1**	**90.9**	**74.9**	**86.1**	**82.7**	**84.2**	**88.8**	**86.7**	**61.5**	**71.9**	**72.4**	**79.38**

5. Discussion

The experimental results demonstrate that the Adaptor-based fine-tuning approach provides superior segmentation performance across diverse water body categories compared to conventional full-scale retraining methods. This approach, which utilizes Adaptors, significantly reduces computational costs without sacrificing accuracy. The improvement in mean Intersection over Union (mIoU), highlights the efficacy of leveraging high-frequency component tuning alongside embedding-based adjustments. The significant gains observed for categories such as fjord, river delta, and flood indicate that the Adaptor-based method effectively handles challenges like occlusions, reflections, and seasonal variations. Unlike traditional models, which often struggle to distinguish between visually similar regions, this approach's ability to retain pre-trained knowledge while adapting to specific features of water bodies leads to a more robust segmentation output.

Another key observation is the effectiveness of Adaptor networks when compared to recent related approaches such as adaptformer and linear fine-tuning. The Adaptor method consistently outperforms these methods, particularly in challenging categories such as "canal" and "lake," which require precise boundary detection. The combination of embedding and high-frequency component tuning allows the ViT model to handle complex water body characteristics, ensuring that subtle differences between water and non-water regions are accurately captured. The reduction in trainable parameters, quantified in equation (7), further validates the efficiency of the method, making it suitable for deployment in real-world, resource-constrained environments. This efficiency is critical for applications in environmental monitoring, where computational resources are often limited, and rapid processing is required for decision-making.

Compared to existing segmentation models, the adaptor-based method significantly reduces the number of parameters. Models such as ANNet (63.1M), GCNet (28.1M), and DeepLabv3 (15.4M) exhibit substantially higher parameter counts, whereas the adaptor-based method achieves remarkable efficiency with only 0.55M parameters. This parameter efficiency makes it highly suitable for resource-constrained environments, enabling a substantial reduction in computational demands while maintaining competitive performance. Notably, training this model required only 3 hours. Furthermore, the existing segmentation models used for comparison were trained from scratch on the dataset in this study, rather than leveraging pre-trained models, with training times ranging from 7 hours to 2 days. The adaptor-based method demonstrates a clear advantage in terms of speed and efficiency.

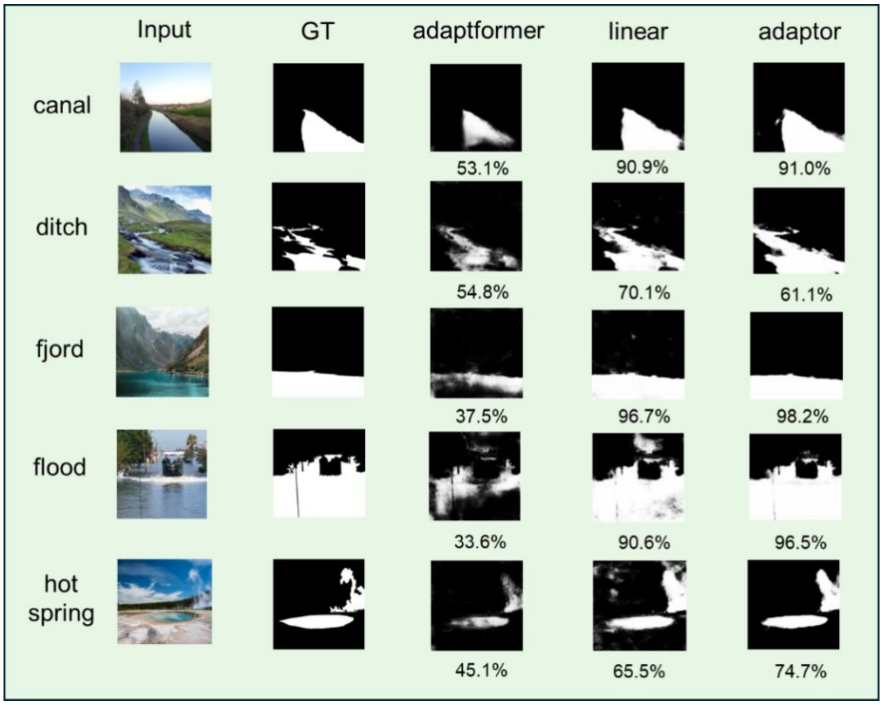

Figure 5. IoU for Various Water Body Segmentation using Different Light-Weight Fine-Tuning Methods

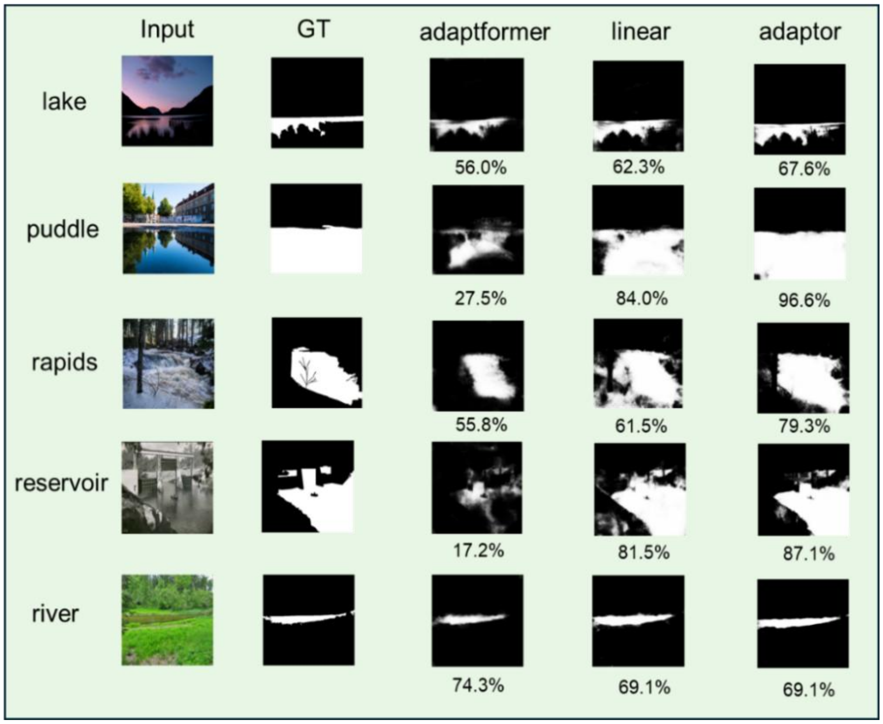

Figure 6. IoU for Various Water Body Segmentation using Different Light-Weight Fine-Tuning Methods

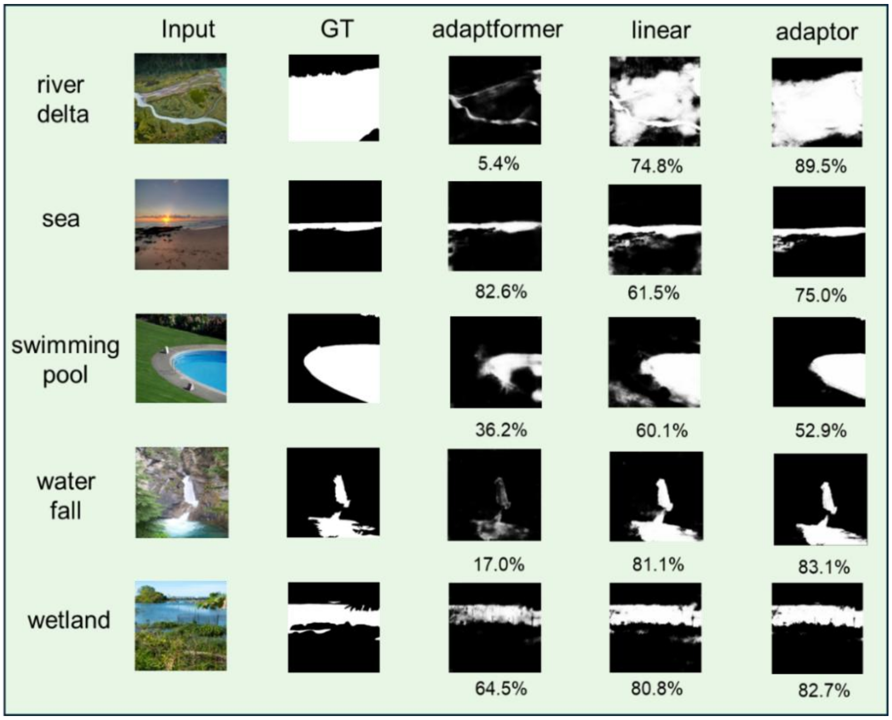

Figure 7. IoU for Various Water Body Segmentation using Different Light-Weight Fine-Tuning Methods

6. Conclusions

This paper presents a lightweight, Adaptor-based fine-tuning framework for water body segmentation that significantly improves both computational efficiency and segmentation accuracy. By adding Adaptors in Vision Transformers, the approach avoids the need for full-scale retraining, thereby reducing the computational load and training time. The experimental results indicate that this Adaptor-based method outperforms existing state-of-the-art models and related fine-tuning approaches, achieving a mean IoU of 79.38% across 15 water body categories. The ability to effectively capture high-frequency details and adapt embeddings to specific segmentation tasks allows for robust performance, even under challenging conditions such as occlusions, reflections, and seasonal variations. This research's contributions include demonstrating the feasibility of using lightweight Adaptor networks for effective transfer learning, particularly in the context of water body segmentation. Future work will explore extending this methodology to other environmental monitoring tasks, leveraging the flexibility and efficiency of Adaptor-based models to address various segmentation challenges in remote sensing and geographic information systems (GIS).

ACKNOWLEDGMENTS

This work was supported by JST SPRING, Grant Number JPMJSP2138.

References

Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. (2019). GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. 0–0. https://openaccess.thecvf.com/content_ICCVW_2019/html/NeurArch/Cao_GCNet_Non-Local_Networks_Meet_Squeeze-Excitation_Networks_and_Beyond_ICCVW_2019_paper.html.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://ieeexplore.ieee.org/abstract/document/5206848/

Dong, W., Yan, D., Lin, Z., & Wang, P. (2023). Efficient Adaptation of Large Vision Transformer via Adapter Re-Composing. *Advances in Neural Information Processing Systems*, *36*, 52548–52567.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv Preprint arXiv:2010.11929*.

Erfani, S. M. H., Wu, Z., Wu, X., Wang, S., & Goharian, E. (2022). ATLANTIS: A benchmark for semantic segmentation of waterbody images. *Environmental Modelling & Software*, *149*, 105333.

Kadhim, I. J., & Premaratne, P. (2023). A Novel Deep Learning Framework for Water Body Segmentation from Satellite Images. *Arabian Journal for Science and Engineering*, *48*(8), 10429–10440. https://doi.org/10.1007/s13369-023-07680-5

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of naacL-HLT*, *1*, 2. https://www.waqasrana.me/assets/papers/N19-1423.pdf

Liu, W., Shen, X., Pun, C.-M., & Cun, X. (2023). Explicit visual prompting for low-level structure segmentations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19434–19445. http://openaccess.thecvf.com/content/CVPR2023/html/Liu_Explicit_Visual_Prompting_for_Low-Level_Structure_Segmentations_CVPR_2023_paper.html

Ni, J., Wu, J., Elazab, A., Tong, J., & Chen, Z. (2022). DNL-Net: Deformed non-local neural network for blood vessel segmentation. *BMC Medical Imaging*, *22*(1), 109. https://doi.org/10.1186/s12880-022-00836-z

Pinaya, W. H. L., Vieira, S., Garcia-Dias, R., & Mechelli, A. (2020). Convolutional neural networks. In *Machine learning* (pp. 173–191). Elsevier. https://www.sciencedirect.com/science/article/pii/B9780128157398000109

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Saleh, F. S., Aliakbarian, M. S., Salzmann, M., Petersson, L., & Alvarez, J. M. (2018). Effective use of synthetic data for urban scene semantic segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 84–100. http://openaccess.thecvf.com/content_ECCV_2018/html/Fatemeh_Sadat_Saleh_Effective_Use_of_ECCV_2018_paper.html

Wang, H., Wu, X., Huang, Z., & Xing, E. P. (2020). High-frequency component helps explain the generalization of convolutional neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8684–8694. http://openaccess.thecvf.com/content_CVPR_2020/html/Wang_High-requency_Component_Helps_Explain_the_Generalization_of_Convolutional_Neural_Networks_CVPR_2020_paper.html

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, *34*, 12077–12090.

Yuan, K., Zhuang, X., Schaefer, G., Feng, J., Guan, L., & Fang, H. (2021). Deep-learning-based multispectral satellite image segmentation for water body detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 7422–7434.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2881–2890. http://openaccess.thecvf.com/content_cvpr_2017/html/Zhao_Pyramid_Scene_Parsing_CVPR_2017_paper.html.