Segment Any Stream: Scalable Water Extent Detection with the Segment Anything Model

Haozhen Zheng[1,3*], **Chenhui Zhang**[3,4*†], **Kaiyu Guan**[1,2,3], **Yawen Deng**[1,2]
Sherrie Wang[4], **Bruce L Rhoads**[5], **Andrew J. Margenot**[1,2], **Shengnan Zhou**[2], **Sheng Wang**[1,2†]
[1] Agroecosystem Sustainability Center,
Institute for Sustainability, Energy, and Environment,
University of Illinois Urbana-Champaign
[2] College of Agricultural, Consumer and Environmental Sciences,
University of Illinois Urbana-Champaign
[3] National Center for Supercomputing Applications,
University of Illinois Urbana-Champaign
[4] MIT Institute for Data, Systems, and Society
[5] Department of Geography & Geographic Information Science,
University of Illinois Urbana-Champaign

Abstract

The accurate detection of water extent in streams and rivers is pivotal to understanding inland water hydrodynamics and terrestrial-aquatic interactions of biogeochemical cycles, in particular bank erosion and the resulting transfer of nutrient elements such as phosphorus (P). Prior studies have employed a variety of computational methods, ranging from hand-crafted decision rules based on spectral indices to advanced image segmentation techniques. However, these methods are limited in their generalizability when implemented in new regions. Furthermore, the recent development of vision foundation models such as the Segment Anything Model (SAM) has brought about opportunities for water extent detection due to their exceptional generalization capabilities. Nevertheless, the adaptation of these models remains challenging due to the computational overhead of fully fine-tuning the entire model. Taking these desiderata into account, this work proposes **S**egment **A**ny **S**tream (SAS), which employs the Low-Rank Adaptation (LoRA) method to perform low-rank updates on a pretrained SAM with a small amount of curated high-resolution aerial imagery to map the water extents in the Mackinaw watershed, a HUC-8 watershed in central Illinois. Through our experiments, we show that SAS is lightweight yet highly effective: it enables efficient fine-tuning on a single consumer-grade GPU while achieving a high IoU of 0.76. This research highlights a generalizable framework for repurposing foundation models to support river/stream segmentation. We believe this framework can benefit the accurate and scalable quantification of streambank erosion as assessed by bank migration and width changes over time, a significant source of sediment and nutrient losses in agricultural landscapes. Code and data are released at `https://github.com/zoezheng126/SAMed-river/tree/development`.

*Equal Contribution
†Correspondence to: Chenhui Zhang, Sheng Wang

Computational Sustainability Workshop at NeurIPS 2023.

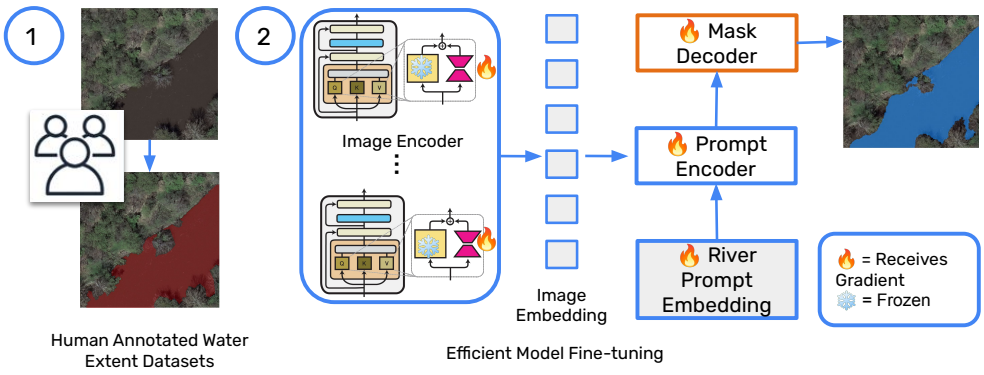

Figure 1: Segment Any Stream (SAS) implements an efficient and effective fine-tuning pipeline for adapting the Segment Anything Model (SAM) to water extent segmentation tasks.

1 Introduction

Water extent detection plays a crucial role in our understanding of the temporal change in inland water hydrodynamics and bank erosion. A conventional approach to map the water extent is to calculate spectral indices such as NDWI shown in Equation (2), and then perform thresholding to obtain a binary water extent mask [McFeeters, 1996, 2013, Xu, 2006]. In addition, deep-learning-based image segmentation models such as U-Net and VGG-Net are also utilized to map lake dynamics [Pi et al., 2022, Nyberg et al., 2023, Moortgat et al., 2022]. More recently, computer vision foundation models such as CLIP [Radford et al., 2021], Segment Anything [Kirillov et al., 2023] and DINOv2 [Oquab et al., 2023] have demonstrated exceptional performance on a variety of downstream computer vision tasks such as image classification, semantic segmentation, depth estimation, etc, in a few-shot or even zero-shot manner, for which we discuss in detail in Section 2. However, despite having the emergent capabilities of existing foundation models, the adaptation to water extent mapping is still non-trivial due to the high computational overhead of fully fine-tuning the model and the out-of-distribution nature of overhead aerial imagery. In light of the aforementioned challenges and the imperatives of water extent mapping, this work aims to ask: *Can we repurpose an existing vision foundation model such as Segment Anything to achieve better and more efficient water extent mapping? How does the adapted model perform on out-of-distribution aerial images?*

To address these challenges, this work develops *Segment Any Stream* (SAS), a regional watershed dataset and an efficient fine-tuning strategy to adapt SAM to our new dataset, which is illustrated in Figure 1. This work is highly relevant to national and regional water quality efforts, notably the Mississippi River Basin nutrient reduction targets set by U.S. Environmental Protection Agency (EPA). Our contributions can be summarized as follows:

- We propose to employ the Low-Rank Adaptation (LoRA) method to fine-tune a pretrained SAM in an efficient manner. LoRA constrains the weight updates to a low-rank subspace, facilitating a lightweight yet effective adaptation of SAM, circumventing the computational exigencies typically associated with full-scale fine-tuning.
- We curated a high-quality regional river extent dataset (as visualized in Figure 2) with high-resolution aerial imagery from the United States Department of Agriculture (USDA).

2 Related Works

Foundation Models in Computer Vision. In the quest to develop a foundation model that can be easily adapted for computer vision tasks, a variety of works try to tackle this problem from the perspectives of pretraining methods, model architecture, and downstream applications. Radford et al. proposed CLIP, an image-text pertaining framework that learns a joint vision-language embedding useful for a variety of downstream tasks such as zero-shot classification and image retrieval. In addition, He et al. [2022] proposes MAE, a scalable self-supervised pretraining method for vision transformer (ViT) models by reconstructing randomly masked image patches. In order to overcome the limitations of image-text pretraining, Oquab et al. [2023] proposes DINOv2, a data curation

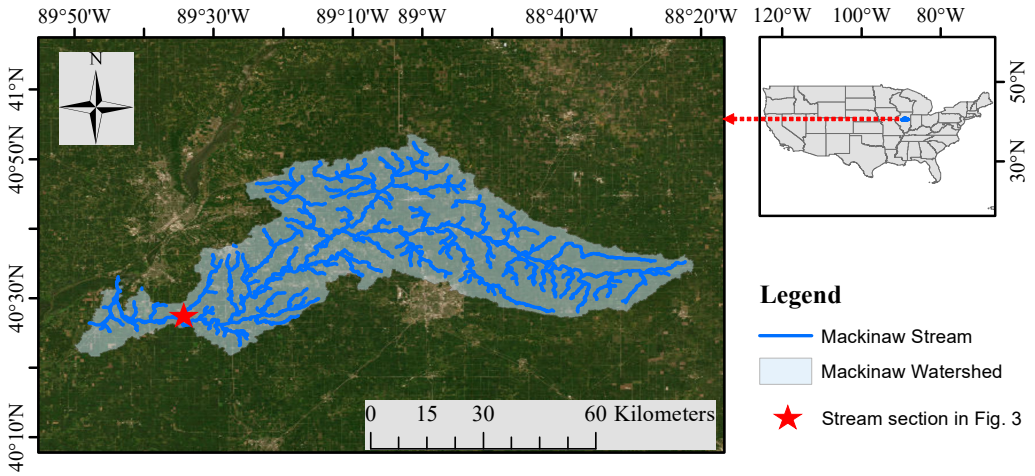

Figure 2: Mackinaw River Watershed in Illinois. The red star refers to the location of the image visualized in Figure 3.

and self-supervised training recipe that better learns the local information in an image. Kirillov et al. [2023], on the other hand, proposes the Segment Anything Model (SAM), a data curation and large-scale supervised training pipeline to tackle the problem of interactive image segmentation. Due to the superior performance of these models, the earth observation (EO) community has gained substantial interest in their applications. In particular, Cong et al. [2022] proposes SATMAE, a variant of MAE that can pretrain ViTs on multitemporal and multispectral satellite imageries. Zhang et al. [2023] proposes TEXT2SEG, a pipeline to perform text-conditioned segmentation for EO data. Taking advantage of the generalization capability of SAM, Wang et al. proposes an annotation pipeline for EO data to obtain a diverse set of remote sensing segmentation datasets.

Parameter-Efficient Adaption of Foundation Models. The efficient adaption of foundation models towards specific downstream tasks has been of growing interest to the research community due to the need to leverage their capabilities under computational and memory constraints. Houlsby et al. [2019] introduces a parameter-efficient transfer learning method for NLP tasks using adapter modules, allowing task-specific training with fewer additional parameters, without altering the original network's parameters. Furthermore, motivated by the observation that over-parametrized models lie in a low intrinsic dimension, Hu et al. [2022] proposes a finetune technique to adapt large language models (LLMs) with a few low-rank matrices inserted into self-attention layers, significantly reducing computational and memory overhead compared with full finetuning. During inference time, LoRA can be merged with main weights, posing no additional overhead during inference. Similarly, Zaken et al. [2022] introduces a straightforward approach that only finetunes the bias terms of a model. Extending into the vision domain, He et al. [2023] presents a method for adapting Vision Transformers (ViTs), which first selects adaptation candidates by measuring their local intrinsic dimensions and then projects them into subspace for further decomposition via a novel Kronecker Adaptation method.

3 Method

Parameter-Efficient Adaptation of the Image Encoder. In order to specialize SAM into the river segmentation task under data and computation constraints, we use LoRA Hu et al. [2022] to constrain the weight updates during fine-tuning to a low-rank subspace. LoRA contributes to our goal in the following ways: **1)** it is parameter-efficient, which allows us to adapt SAM on watershed segmentation tasks by tuning less than 5% of the original parameters with limited computation and memory; **2)** it poses no additional overhead during inference, allowing us to merge the LoRA modules seamlessly back to the main model for future inference; **3)** it can preserve the utility of the original model. Intuitively, since we are fine-tuning the model on relatively out-of-distribution (OoD) data

compared to the original training dataset, limiting the model update to a low-rank subspace can prevent the drastic change of the model weight due to certain data points that incur high losses.

Concretely, given a weight matrix $W_0 \in \mathbb{R}^{d\times k}$ of the projection layers in the transformer blocks of a pretrained SAM, we insert a low-rank matrix $\Delta W = BA$ where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ with $r \ll \min(d, k)$. Here r is a tunable hyper-parameter during our fine-tuning process. During fine-tuning, we completely freeze W_0 and only let ΔW receive gradients with a reparametrized forward pass as below.

$$(W_0 + \Delta W)x = W_0 x + \Delta W x \tag{1}$$

During fine-tuning, we insert LoRA modules into the query and key projection layers of the self-attention blocks of the image encoder as a reparametrization shown on the right-hand side of Equation 1. This is simply a residual connection over the selected weight matrices. After training, we seamlessly merge the learned ΔW with the original weight W_0 as the left-hand side of Equation 1 so that we can perform inference with the modified weights without additional overheads.

Adaptation of Prompting Techniques. To make the downstream performance of SAS less dependent on the prompting techniques, we initialize an additional learnable prompt embedding vector and fine-tune it together with the prompt encoder. During inference, we simply use this learned embedding vector associated with rivers as the input to the prompt encoder. By doing so, we eliminate the need for the point and box prompts from the user.

Data Curation. To obtain our training and testing dataset, we downloaded the high-resolution aerial images of Illinois taken during the leaf-off season of 2011 by the National Agriculture Imagery Program (NAIP) [Earth Resources Observation and Science (EROS) Center, 2018], with a spatial resolution of 0.3 m and RGB and NIR bands. Based on the previous records of river centerlines, we cropped the NAIP imageries for a human annotator with basic knowledge of hydrology to annotate a polygon of the current extent of the Mackinaw watershed. Finally, the annotated polygon is overlaid with the raster image to generate ground truth masks. We randomly split the resulting dataset to 512×512 chips. Finally, we visualize our study region in Figure 2.

4 Experiments

Baselines. To evaluate SAS, We mainly consider spectral indices thresholding, U-Net, and untuned SAM as our baselines. Concretely, we first calculate the NDWI index based on the reflectance and set the NDWI threshold for each image based on the Otsu's algorithm [Otsu, 1979] or as the corresponding NDWI at the local minimum of the Kernel Density Estimation (KDE) function.

$$NDWI = \frac{X_{green} - X_{nir}}{X_{green} + X_{nir}} \tag{2}$$

We also consider U-Net [Ronneberger et al., 2015] as another baseline for which we train the model from scratch on the same training dataset for 60 epochs with AdamW optimizer and learning rate decay from $1e-3$. In addition, we consider the vanilla SAM without fine-tuning as another baseline.

Setup. In order to fine-tune the LoRA layers in SAS, we follow the recipe in Zhang and Liu [2023] to use learning rate warmup and a weighted combination of cross entropy and dice loss. We also fine-tune for 60 epochs for a fair comparison.

4.1 Segment Any Stream

In Table 1, we compare SAS to the aforementioned baselines in terms of IoU, precision, recall, accuracy, f-1 score, and kappa score. In terms of all metrics, SAS demonstrates superior performance compared with the baseline methods. Most notably, SAS demonstrates a substantial performance increase compared with SAM with default prompting methods. In Figure 3, we further provide a qualitative comparison between different methods. Compared with NDWI thresholding and U-Net, SAS produces smoother segmentation results without fuzzy edges and holes in the middle.

Table 1: Test performance of spectral-index thresholding (NDWI), U-Net, SAM, and SAS in selected study regions. SAM-D uses the default prompt and untuned SAM. SAM-P uses random point prompts sampled from the NDWI mask and untuned SAM.

Method	Input Bands	IoU	Precision	Recall	Accuracy	f-1	Kappa
NDWI (KDE)	G, NIR	0.35	0.55	0.42	0.86	0.78	0.68
NDWI (Otsu)	G, NIR	0.35	0.55	0.42	0.83	0.78	0.64
U-Net	RGB	0.71	0.87	0.79	0.89	0.83	0.75
SAM-D	RGB	0.17	0.33	0.26	0.57	0.29	-0.01
SAM-P	RGB	0.54	0.66	0.75	0.78	0.70	0.53
SAS (ours)	RGB	**0.76**	**0.90**	**0.83**	**0.90**	**0.86**	**0.79**

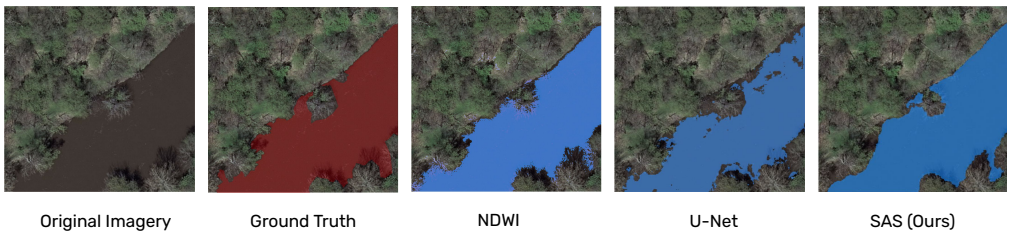

Figure 3: Visualization of the Comparison Among Different Methods

Table 2: Test Performance of SAS in Selected Study Regions. The best result regarding each metric is highlighted in **bold**.

Rank	% of Parameters Trained	IoU	Precision	Recall	Accuracy	f-1	Kappa
1	4.19	0.75	**0.89**	**0.82**	**0.91**	**0.84**	0.77
2	4.23	0.74	**0.89**	**0.82**	0.90	0.83	0.76
4	4.31	0.75	0.88	**0.82**	0.88	0.82	**0.78**
6	4.39	0.74	0.88	**0.82**	**0.91**	**0.84**	**0.78**
8	4.79	**0.75**	**0.89**	**0.82**	**0.91**	**0.84**	0.77

4.2 Ablation Studies

In order to gauge the influence of the choice of LoRA rank on the segmentation results, we perform an ablation study with different LoRA ranks ranging from 1 to 8. In Table 2, we report the corresponding testing results along with the number of trained parameters during fine-tuning. As detailed in Table 2, we do not observe a significant change in testing results when varying the rank.

5 Conclusions

In this work, we propose SAS, a parameter-efficient fine-tuning framework that enables the adaptation of a state-of-the-art SAM model on a consumer-grade GPU within two hours. We also curate a regional watershed extent dataset in Illinois based on our annotations of the Mackinaw watershed. Through our experiments, we show that SAS demonstrates superior performance despite low computational overhead.

References

Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. B. Lobell, and S. Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In *NeurIPS*, 2022. URL `http://papers.nips.cc/paper_files/paper/2022/hash/01c561df365429f33fcd7a7faa44c985-Abstract-Conference.html`.

Earth Resources Observation and Science (EROS) Center. Usgs eros archive - aerial photography - national agriculture imagery program (naip), 2018. URL https://www.usgs.gov/centers/eros/science/usgs-eros-archive-aerial-photography-national-agriculture-imagery-program-naip. Accessed: October 1, 2023.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://doi.org/10.1109/CVPR52688.2022.01553.

X. He, C. Li, P. Zhang, J. Yang, and X. E. Wang. Parameter-efficient model adaptation for vision transformers. In B. Williams, Y. Chen, and J. Neville, editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 817–825. AAAI Press, 2023. doi: 10.1609/aaai.v37i1.25160. URL https://doi.org/10.1609/aaai.v37i1.25160.

N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for NLP. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 2019. URL http://proceedings.mlr.press/v97/houlsby19a.html.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. B. Girshick. Segment anything. *CoRR*, abs/2304.02643, 2023. doi: 10.48550/arXiv.2304.02643. URL https://doi.org/10.48550/arXiv.2304.02643.

S. K. McFeeters. The use of the normalized difference water index (ndwi) in the delineation of open water features. *International journal of remote sensing*, 17(7):1425–1432, 1996.

S. K. McFeeters. Using the normalized difference water index (ndwi) within a geographic information system to detect swimming pools for mosquito abatement: A practical approach. *Remote Sensing*, 5(7):3544–3561, 2013. ISSN 2072-4292. doi: 10.3390/rs5073544. URL https://www.mdpi.com/2072-4292/5/7/3544.

J. Moortgat, Z. Li, M. Durand, I. Howat, B. Yadav, and C. Dai. Deep learning models for river classification at sub-meter resolutions from multispectral and panchromatic commercial satellite imagery. *Remote Sensing of Environment*, 282:113279, 2022.

B. Nyberg, G. Henstra, R. L. Gawthorpe, R. Ravnås, and J. Ahokas. Global scale analysis on the extent of river channel belts. *Nature Communications*, 14(1):2163, 2023.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

N. Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

X. Pi, Q. Luo, L. Feng, Y. Xu, J. Tang, X. Liang, E. Ma, R. Cheng, R. Fensholt, M. Brandt, et al. Mapping global lake dynamics reveals the emerging roles of small lakes. *nature communications*, 13(1):5777, 2022.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference*

on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

D. Wang, J. Zhang, B. Du, D. Tao, and L. Zhang. Scaling-up remote sensing segmentation dataset with segment anything model. *arXiv preprint arXiv:2305.02034*, 2023.

H. Xu. Modification of normalised difference water index (ndwi) to enhance open water features in remotely sensed imagery. *International journal of remote sensing*, 27(14):3025–3033, 2006.

E. B. Zaken, Y. Goldberg, and S. Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1–9. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-short.1. URL https://doi.org/10.18653/v1/2022.acl-short.1.

J. Zhang, Z. Zhou, G. Mai, L. Mu, M. Hu, and S. Li. Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models. *CoRR*, abs/2304.10597, 2023. doi: 10.48550/arXiv.2304.10597. URL https://doi.org/10.48550/arXiv.2304.10597.

K. Zhang and D. Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.