Towards Foundation Models for Zero-Shot Semantic Tree Segmentation

Mareike Weishaupt, Guneet Mutreja, and Jiaojiao Tian

Remote Sensing Technology Institute, German Aerospace Center (DLR), Münchener Str. 20, 82234 Oberpfaffenhofen, Germany. mareike.weishaupt@dlr.de

Abstract. Tree detection and segmentation from remote sensing (RS) images is important for both forest management and urban planning. Previous methods using spectral sensors or height information to separate trees from background are often costly. Supervised tree classification models are restricted in cross-domain generalization. While foundation models (FMs) have demonstrated strong zero-shot capabilities in natural image understanding and RS, their robustness in semantic tree segmentation in RGB imagery remains largely unexplored. We evaluate recent vision and vision-language FMs on three heterogeneous RGB datasets to assess their zero-shot generalization across unseen domains. GeoVLM achieves the most robust overall performance with a mIoU of 42.9 %, whereas SAM3 performs particularly well on high-resolution single-tree segmentation. The results demonstrate that geospatial vision-language models (VLMs) can serve for semantic tree segmentation in previously unseen domains, enabling downstream ecological and forestry analyses without requiring task-specific annotated training data.

Keywords: Segmentation · VLM · Zero-Shot · Trees · RS · RGB

1 Introduction

Trees play a crucial role in urban planning, forestry, biodiversity monitoring, and environmental management [9]. Semantic tree segmentation enables applications such as canopy cover estimation, biomass assessment, tree inventory, and health monitoring. However, accurate tree segmentation from remote sensing (RS) imagery remains challenging in the absence of rich spectral and structure information [8].

Recent deep learning methods have substantially improved semantic tree segmentation from RGB imagery, but their performance depends on annotated training data that closely match the target domain [4]. Supervised learning models often struggle with cross-domain tree segmentation tasks due to the significant variations in tree species like morphology, phenological stages, illumination, or spatial resolution [17]. Creating pixel-level annotations across diverse environments, acquisition conditions, and tree species is labor-intensive and expensive, limiting scalability and cross-domain generalization.

Recently, foundation models (FMs), including SAM3 [1], RemoteSAM (RSAM) [16], Geospatial Vision Language Model (GeoVLM) [5], and EarthMind [11], have demonstrated promising zero-shot capabilities by leveraging large-scale visual or vision-language pretraining. Over the past decade, CNNs [2,10,13] have

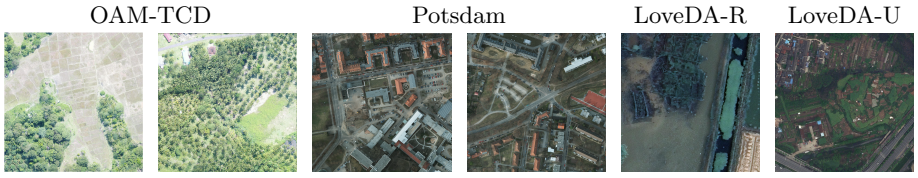


Fig. 1: Two samples from each evaluated dataset.

Table 1: Comparison of the three evaluated RGB RS datasets.

Property	**OAM-TCD [12]**	**Potsdam [6]**	**LoveDA [14]**
Product	Orthorectified UAV RGB	True RGB orthophotos	Orthorectified RGB
GSD	10 cm/px	5 cm/px	30 cm/px
Image size	2048×2048	6000×6000	1024×1024
Images (Test)	4,608 (439)	38 (14)	5,987 (1,669*; R:992, U:677)
Original class	Tree	Tree	Forest
Tree cover	25.65%	17.04%	R:4.94%; U:10.45%
Location	Worldwide	Germany	China
Year	2024	2014	2021
Tree-Specific Challenges	Global diversity; tree density and seasonality	Winter images (deciduous trees); dense urban layout	Low resolution; image quality; complex BG

Abbreviations: U/R = urban/rural; BG = Background; GSD = Ground Sampling Distance *Validation data.

dominated this field by achieving high segmentation accuracy and more recently, transformer-based architectures [3,7,15], have further improved segmentation performance. Although these approaches represent the current state of the art and continue to deliver excellent performance, they generally rely on large amounts of annotated training data and often exhibit limited generalization when applied to unseen datasets or domains [4].

Despite these advances, the effectiveness of FMs for zero-shot semantic tree segmentation in RGB imagery has not been systematically evaluated. This work presents a benchmark of recent foundation and vision-language models (VLMs) for zero-shot semantic tree segmentation across three well-selected RGB datasets covering both local and worldwide geographic regions. As shown in Figure 1 these datasets represent diverse tree distributions, species compositions, and seasonal conditions. Our objective is to evaluate their cross-domain generalization and identify the strengths and limitations of current FMs for their generalization performance across varying environments.

2 Methodology

Four state-of-the-art FMs are selected and investigated for zero-shot tree semantic segmentation. Figure 1 illustrates example images of each dataset. In Table 1 the datasets and their characteristics are summarized. The datasets differ in many attributes, enabling a comprehensive evaluation under diverse conditions. We further illustrate common tree-specific challenges. For LoveDA [14], validation data are used with two subsets, since no masks are available for test data. For consistent evaluation, all non-tree classes are treated as background.

Table 2: Overview of the evaluated foundation models.

Model	Type	Domain	Used prompts
SAM3 [1]	FM	General	*tree, tree crown, forest tree, tall vegetation*
EarthMind [11]	MLLM	EO	*Please segment [...]. [tree, tree crown, tall vegetation]*
RSAM [16]	FM	EO	*tree, tree crown, tall vegetation*
GeoVLM [5]	VLM	EO	*Segment all the [...] in the image. [tree, tree crown, forest tree, tall vegetation]*

Abbreviations: MLLM: Multimodal Large Language Model; EO: Earth Observation

Table 3: Tree segmentation performance in percent (%). Best results are shown in bold. R/U denote rural/urban.

	OAM-TCD [12]				Potsdam [6]				LoveDA-R [14]				LoveDA-U [14]				Mean
Model	IoU	F1	Pre.	Rec.	IoU	F1	Pre.	Rec.	IoU	F1	Pre.	Rec.	IoU	F1	Pre.	Rec.	IoU
SAM3[1]	19.1	32.1	95.0	19.3	**30.7**	**47.0**	**96.6**	**31.0**	04.4	08.5	**54.7**	04.6	37.7	54.8	70.4	44.8	23.0
EarthMind[11]	31.4	47.8	**97.0**	31.7	0.01	0.02	99.0	0.01	13.9	24.4	41.1	17.4	18.9	31.8	**72.4**	20.4	16.3
RSAM[16]	41.5	58.7	62.9	54.9	00.8	01.6	57.7	00.8	04.8	09.2	38.0	05.3	08.8	16.0	56.7	09.3	14.0
GeoVLM[5]	**71.9**	**83.7**	79.2	**88.7**	18.3	30.9	79.8	19.1	**30.9**	**47.3**	41.4	**55.0**	**50.6**	**67.2**	57.7	**80.5**	**42.9**

The selected models represent different approaches to vision-language understanding and segmentation, ranging from general-purpose FM to specialized RS models (Table 2). This diversity enables a comprehensive assessment of their suitability for zero-shot semantic tree segmentation. During inference, each model was applied to all datasets to produce binary predictions consisting of tree (1) and background (0). All models are evaluated using their default confidence thresholds. Prompts used for inference are stated in Table 2 and integrated into the vision-language models' queries for EarthMind and GeoVLM. For the Potsdam dataset, inference was performed on 1024×1024 tiles with a 896-pixel overlap because the original 6000×6000 images exceeded the supported input size. Performance was evaluated using IoU, F1-score, Recall, and Precision. Overall rankings are based on the arithmetic mean IoU across the datasets.

3 Results and Discussion

Table 3 summarizes the quantitative results. GeoVLM achieves the most robust zero-shot performance, obtaining the highest mIoU and outperforming the other evaluated models, which frequently show high precision but poor recall, especially EarthMind and RSAM. The only exception is the Potsdam dataset, where SAM3 achieves the highest IoU and F1 scores. This can likely be attributed to the ground sampling distance (GSD) of Potsdam (5 cm), whereas GeoVLM was trained for imagery with GSD between 10 and 100 cm [5]. The RS-specific FMs generally outperform SAM3 on datasets with coarser GSD, highlighting the importance of geospatial pretraining for robust cross-domain generalization. Although zero-shot performance remains below those of fully supervised CNN- and transformer-based approaches, GeoVLM demonstrates robust performance across all three datasets without requiring task-specific training.

The selected examples in Figure 2 highlight the strengths and weaknesses of the evaluated models. GeoVLM provides the most consistent outputs overall but struggles on the Potsdam dataset, likely due to its high spatial resolution.

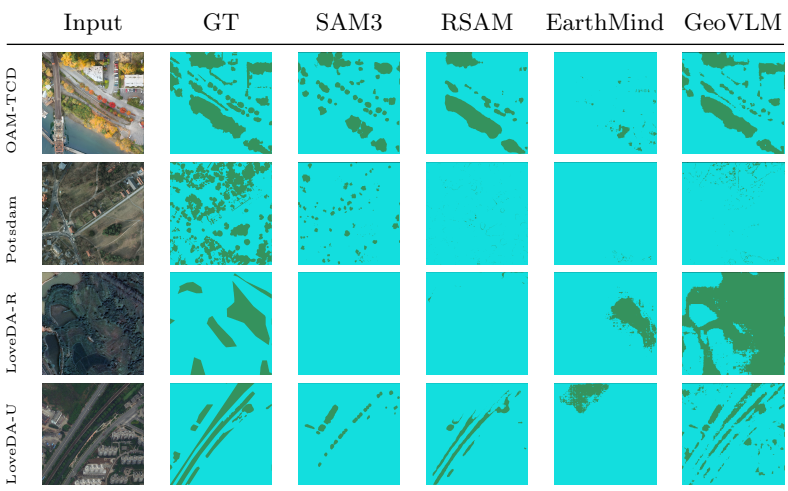


Fig. 2: Qualitative comparison of FMs with the ground truth (GT).

Also classifying additional vegetation as trees is a weakness of the model. In contrast, SAM3 performs best on Potsdam, especially for individual tree crowns. R-SAM occasionally captures the overall structure of tree canopies but produces incomplete segmentations, whereas EarthMind rarely identifies meaningful tree regions across the evaluated datasets. The selected examples further illustrate the challenges posed by the benchmark datasets. OAM-TCD demonstrates that the models adapt reasonably well to seasonal color variations during autumn. In contrast, the leaf-off conditions in Potsdam considerably reduce segmentation quality. LoveDA-R represents the most challenging dataset due to dense forest cover, lower image quality, and the low proportion of tree pixels (approximately 5 %). Conversely, the models generally perform better in urban environments, suggesting that large-scale pretraining enables robust recognition of common urban structures such as roads, buildings, and vegetation.

4 Conclusion

This work investigates recent foundation and vision-language models for zero-shot semantic tree segmentation across three heterogeneous RGB datasets. Among the evaluated models, GeoVLM demonstrated the most reliable cross-domain generalization, outperforming the other FMs on unseen datasets, whereas SAM3 performed strongest on the Potsdam data. The results further indicate that geospatial pretraining is essential for robust zero-shot semantic tree segmentation in RS, while substantial challenges remain for dense forests, leaf-off conditions, and differing resolutions. Overall, this work provides a reference for evaluating FMs for semantic tree segmentation and demonstrates their potential as an annotation-efficient starting point for downstream ecological and forestry applications, although robust generalization across diverse conditions remains a significant challenge.

References

1. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), https://arxiv.org/abs/2511.16719
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation (2018), https://arxiv.org/abs/1802.02611
3. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation (2022), https://arxiv.org/abs/2112.01527
4. Cheng, J., Deng, C., Su, Y., An, Z., Wang, Q.: Methods and datasets on semantic segmentation for unmanned aerial vehicle remote sensing images: A review. ISPRS Journal of Photogrammetry and Remote Sensing **211**, 1–34 (2024). https://doi.org/https://doi.org/10.1016/j.isprsjprs.2024.03.012, https://www.sciencedirect.com/science/article/pii/S0924271624000844
5. Esri: Introduction to the geospatial vision language model (2026), https://doc.arcgis.com/en/pretrained-models/latest/imagery/introduction-to-geospatial-vision-language-model.htm, accessed: 2026-07-22
6. ISPRS: 2d semantic labeling contest – potsdam. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx (2014)
7. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021), https://arxiv.org/abs/2103.14030
8. Onishi, M., Ise, T.: Explainable identification and mapping of trees using uav rgb image and deep learning. Scientific reports **11**(1), 903 (2021)
9. Pretzsch, H., Moser-Reischl, A., Rahman, M., Pauleit, S., Rötzer, T.: Towards sustainable management of the stock and ecosystem services of urban trees. from theory to model and application. Trees - Structure and Function **37**(1), 177–196 (Feb 2023). https://doi.org/10.1007/s00468-021-02100-3, publisher Copyright: © 2021, The Author(s).
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015), https://arxiv.org/abs/1505.04597
11. Shu, Y., Ren, B., Xiong, Z., Paudel, D.P., Gool, L.V., Demir, B., Sebe, N., Rota, P.: Earthmind: Leveraging cross-sensor data for advanced earth observation interpretation with a unified multimodal llm (2025), https://arxiv.org/abs/2506.01667
12. Veitch-Michaelis, J., Cottam, A., Schweizer, D.S., Broadbent, E.N., Dao, D., Zhang, C., Almeyda Zambrano, A., Max, S.: Oam-tcd: A globally diverse dataset of high-resolution tree cover maps. In: Advances in Neural Information Processing Systems. vol. 37, pp. 1–12. Curran Associates, Inc. (2024)
13. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition (2020), https://arxiv.org/abs/1908.07919

14. Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y.: LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation (Oct 2021). https://doi.org/10.5281/zenodo.5706578, https://doi.org/10.5281/zenodo.5706578
15. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers (2021), https://arxiv.org/abs/2105.15203
16. Yao, L., Liu, F., Chen, D., Zhang, C., Wang, Y., Chen, Z., Xu, W., Di, S., Zheng, Y.: Remotesam: Towards segment anything for earth observation (2025), https://arxiv.org/abs/2505.18022
17. Yun, T., Li, J., Ma, L., Zhou, J., Wang, R., Eichhorn, M.P., Zhang, H.: Status, advancements and prospects of deep learning methods applied in forest studies. International Journal of Applied Earth Observation and Geoinformation **131**, 103938 (2024)