Segment Anything Model with LiDAR based images for building segmentation in forest areas

Emilia Hattula [1], Jere Raninen [1], Lingli Zhu [1]

[1] National Land Survey of Finland (NLS) firstname.lastname@nls.fi

Keywords: SAM, U-Net, Building Segmentation, LiDAR, Digital Surface Model, Remote Sensing

Abstract

The Segment Anything Model (SAM) represents a significant advancement in image segmentation, with growing applications for LiDAR-based data alongside traditional RGB imagery. Recent work, such as Ošep et al. (2024) on Segment Anything in LiDAR (SAL) and Yarroudh (2023) on automatic unsupervised LiDAR segmentation with SAM, highlights its potential for enhancing segmentation accuracy in complex environments. Building segmentation, especially in forested areas, poses unique challenges due to difficulties in distinguishing structures from dense vegetation. Prior research indicates that utilizing height information from LiDAR digital surface models (DSM) and digital elevation models (DEM) is beneficial, suggesting SAM could improve forest building segmentation accuracy with LiDAR-based images.

This study explores SAM's application for building segmentation using true orthophotos and LiDAR-derived DSMs and DEMs. Its performance is compared against the U-Net neural network (Ronneberger et al. 2015), which utilizes the same multi-modal data. While existing SAM studies often focus on RGB imagery or point clouds, this research specifically investigates its capabilities within challenging forest environments.

A 72km² rural forested area, covering mapsheet L4211D near Karkkila and N3244E near Närpiö, Finland, was selected for testing. Both models were trained using datasets from multiple Finnish cities. Their performance was evaluated using F1-scores during training. For the test areas, which had true orthophotos, LiDAR DSMs, and DEMs from 2024, the number of correctly identified buildings was analyzed against the topographic database of Finland (1,380 buildings in Karkkila, 1,020 in Närpiö). Additionally, the shape and accuracy of segmented buildings were visually compared. This evaluation of SAM's effectiveness aims to advance methodologies for building extraction in forested landscapes, ultimately seeking to reduce manual labor in future mapping tasks.

1. Introduction

Accurate building segmentation is critical for applications such as urban planning, disaster management, and environmental monitoring (Zhou et al., 2018). However, in forested landscapes, challenges such as occlusions from dense vegetation, challenges in differentiating tree canopies from rooftops, and complex terrain hinder the performance of traditional segmentation methods (Awrangjeb et al., 2013). Recent advances in remote sensing, particularly Light Detection and Ranging (LiDAR), have enabled the generation of high-resolution Digital Surface Models (DSMs) and Digital Elevation Models (DEMs), which capture 3D structural information that can be utilized as such or to complement 2D imagery like true orthophotos (Miliaresis & Kokkas, 2007). These datasets provide opportunities to overcome forest-related challenges by distinguishing buildings from vegetation through elevation analysis.

Deep learning models, such as U-Net—a convolutional neural network (CNN) architecture with an encoder-decoder structure and skip connections—have demonstrated success in building segmentation by learning hierarchical representations from labeled datasets (Ronneberger et al., 2015). However, their reliance on extensive labeled training data limits scalability in diverse environments. The emergence of foundation models like the Segment Anything Model (SAM) (Kirillov et al., 2023), a vision transformer (ViT)-based architecture, offers transformative potential for automating geospatial tasks. Unlike CNNs, which prioritize local spatial patterns through convolutional filters, SAM employs self-attention mechanisms to model global contextual relationships across images, enabling generalization across domains with minimal input prompts. Pre-trained on a massive corpus of 11 million images, SAM achieves robust zero-shot segmentation of generic objects. However, its performance in specialized contexts—particularly forested environments where spectral ambiguity between buildings and vegetation challenges ViT's reliance on texture and color cues—remains underexplored.

This study evaluates SAM's capability for building segmentation in forested landscapes using multi-modal LiDAR data (DSMs, DEMs) and true orthophotos, comparing its results against a U-Net baseline. By addressing SAM's adaptability to elevation data and occlusion challenges, this work aims to advance automated mapping methodologies, reduce manual labour, and provide insights into the integration of foundation models in remote sensing workflows.

2. Background

2.1 Deep Learning and SAM in Remote Sensing

Accurate building segmentation in forested landscapes is complicated by occlusions from dense vegetation and spectral similarities between tree canopies and rooftops. Traditional methods, such as rule-based Object-Based Image Analysis (OBIA), have relied on LiDAR-derived DSMs and DEMs to mitigate these challenges by incorporating elevation and texture data (Miliaresis & Kokkas, 2007). Modern deep learning approaches, such as U-Net—a CNN with an encoder-decoder architecture—have improved segmentation accuracy by learning

hierarchical features from labeled datasets (Ronneberger et al., 2015). However, U-Net's dependency on large-scale annotated training data limits its adaptability to diverse or understudied environments.

SAM, a ViT-based foundation model pre-trained on 11 million images, addresses this limitation through zero-shot generalization (Kirillov et al., 2023). SAM employs self-attention mechanisms to model global contextual relationships, enabling segmentation of unseen objects with minimal prompts. Recent studies highlight SAM's potential in remote sensing: for example, Zhang et al. (2024) fused SAM with LiDAR-derived DSMs and true orthophotos to extract building footprints in residential areas, demonstrating improved performance over RGB-only inputs. However, their work also revealed SAM's struggles in forested regions, where spectral ambiguity between vegetation and buildings persists even with elevation data. In 2024, Ravi et al. published a new, more advanced version of SAM, SAM2 (Ravi et al., 2024). SAM2 builds upon the groundbreaking foundation of SAM by refining its performance on images, making it a more powerful tool for segmentation tasks.

2.2 SAM and LiDAR Integration

LiDAR data provides critical 3D structural information to disentangle overlapping features in complex landscapes. While CNNs like U-Net process LiDAR-optical data through multi-channel inputs (Xu et al., 2021), SAM's ViT architecture lacks native support for 3D point clouds. To bridge this gap, Yarroudh et al. (2023) developed an open-source framework (segment-lidar) that projects LiDAR point clouds into 2D depth maps compatible with SAM, enabling unsupervised segmentation of urban buildings. However, their method achieved only modest accuracy in densely vegetated areas, underscoring SAM's limitations in non-urban contexts.

Efforts to enhance SAM's elevation-awareness include encoding DSM height bands as auxiliary input channels (Zhang et al., 2024) and fusing SAM with synthetic aperture radar (SAR) data for cloud-penetrating capabilities (Liu et al., 2024). While these adaptations improve urban building detection, their performance in forested environments—where irregular building layouts and occlusions dominate—remains untested.

2.3 Challenges in Forested Landscapes

Forested environments pose unique challenges for SAM due to its reliance on RGB spectral cues and limited elevation-awareness. For instance, Zhang et al. (2024) demonstrated that SAM struggles with detection between buildings and vegetation even when fused with LiDAR-derived DSMs, particularly in occlusion-heavy forested regions. Similarly, Yarroudh et al. (2023) observed that SAM's unsupervised LiDAR-based segmentation framework achieved only modest accuracy in densely vegetated areas, where irregular building layouts and overlapping tree canopies dominate. These studies underscore the need for multi-modal adaptations of SAM that explicitly leverage LiDAR's structural data to resolve ambiguities, such as encoding height thresholds or incorporating DEM-derived slope information as segmentation prompts. Studying the capabilities of SAM in the forested areas with 5-channel data, like RGB and height information from both DSMs and DEMs, remains unstudied.

3. Material

Data utilized in this study was collected by the National Land Survey of Finland (NLS). True orthophotos, LiDAR DSM and LiDAR DEM used for fine-tuning U-Net and SAM had a pixel resolution of 25cm and covered 9 different areas of Finland, with high-quality labels corrected with the help of both true orthophoto and LiDAR DSM data, the LiDAR DSM data helping to distinguish also building, for example, covered by trees and not visible in the true orthophoto data. The creation method for the LiDAR DSM was the last and only pulse.

The area of data used for fine-tuning was 344.90 km ² of Finland from the 9 separate areas. The image sizes of each area and their total coverage is found from Table 1.

Name of area	Image size	Area
Lahti	11,048 x 11,632 px	8.03km ²
Närpiö	28,000 x 28,000 px	49.00km ²
Rovaniemi	46,819 x 21,108 px	61.77km ²
Uusikaarlepyy	38,773 x 38,211 px	92.60km ²
Vaala	28,066 x 48,762 px	85.53km ²
Ylitornio	15,985 x 5,584 px	5.58km ²
Oulu	24,000 x 12,000 px	18.00km ²
Heinävesi	12,583 x 11,915 px	9.37km ²
Varkaus	18,613 x 12,910 px	15.02km ²

Table 1. The areas in the training dataset used for fine-tuning, their image sizes and areas covered.

Validation set was a separate set of 512 x 512 pixel images carefully selected to include various types of buildings and environments from multiple Finnish areas to ensure the model performance with different kinds of buildings and areas. It included 752 images.

The U-Net model was pretrained with Finnish datasets collected during earlier years, having 30cm pixel resolution and DSM and DEM data produced from aerial images instead of LiDAR data. The pretraining dataset included 24 different training areas of Finland, covering an area of 167.42km ².

The versions of the original SAM used for fine-tuning were the ViT-B and ViT-L pretrained models. For SAM2.1, the newest version of SAM, baseplus model was selected to be fine-tuned.

Figure 1. A true orthophoto and LiDAR DSM image from the test area of Karkkila, Finland. Water areas have been removed from the LiDAR DSM image for reducing false detections.

Figure 2. A true orthophoto and LiDAR DSM image from the test area of Närpiö, Finland

For evaluating the performance of the models, mapsheets covering areas of 36km ² were selected from two different areas of Finland. These two areas were Karkkila and Närpiö, and they provided excellent forest environments for evaluating the performance of the models with occluded buildings. The evaluation areas are seen in Figures 1 and 2.

4. Methods

Three deep learning methods for building extraction were tested and compared; the U-Net introduced by Ronneberger et al. in 2015, SAM introduced by Kirillov et al. in 2023 and SAM2 by Ravi et al. in 2024.

Data augmentation of the training datasets included random cropping, as well as vertical and horizontal flipping. U-Net and original SAM models were fine-tuned with all available datasets corrected with RGB and LiDAR information. For SAM2, 6 of the available datasets were utilized. All the models used learning rate of 1e-4 for fine-tuning. For model training, the supercomputer Puhti of CSC, Finland, was utilized.

4.1 U-Net

The architecture of the U-Net trained is presented in Figure 3.

U-Net with an almost identical structure to the original U-Net introduced in 2015 (Ronneberger et al. 2015) was used for fine-tuning with building dataset with LiDAR DSMs and DEMs. The differences of the utilized model architecture and the original U-Net architecture laid in the use of dropout with a rate of 0.25, always after a sequence of a convolutional layer, batch normalization, and ReLU. Upsampling was also done a little differently: It consisted of a sequence of PyTorch's classes UpsamplingNearest2d, ConstantPad2d, and a 2d convolution. The model architecture utilized had only three layers in the expanding and contracting paths in addition to the bottom layer (Hattula et al., 2023).

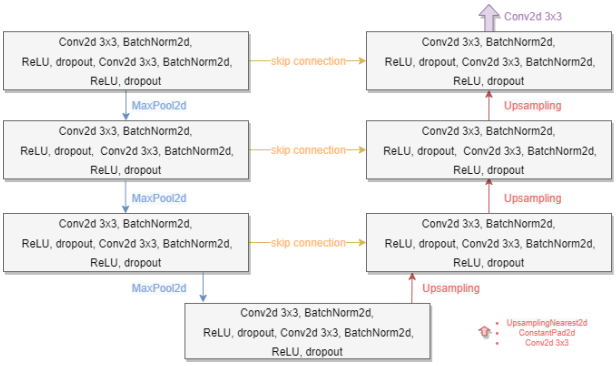

Figure 3. The architecture of UNet trained and fine-tuned (Hattula et al., 2023). Original UNet was developed by Ronneberger et al. in 2015 (Ronneberger et al., 2015).

U-Net was trained with Adam optimizer with the images being randomly cropped and augmented from the large training areas until its performance stopped increasing. Tversky loss was used to weight between precision and recall, recall was given weigh of 0.7 and precision weight of 0.3. Early stopping was utilized for saving the best model according to the validation F1-score.

4.2 SAM

4.2.1 Original SAM

To leverage the capabilities of SAM (Kirillov et al. 2023) for building detection using 5-channel data (RGB+DSM+DEM), its architecture was adapted and fine-tuned. While the core SAM structure—comprising an image encoder, a prompt encoder, and a mask decoder—is retained, key modifications and a specific training strategy were employed.

The ViT-based image encoder was modified to accept 5-channel input instead of the original 3 (RGB). This involved changing the `in_chans` parameter of the initial convolutional layer to 5. Correspondingly, 5-channel pixel mean and pixel standard deviation values, calculated from the fine-tuning dataset, were used for input normalization. As the goal was binary building segmentation, the mask decoder was configured to produce a single output mask. Components like the mask token embeddings and prediction heads were dimensioned accordingly.

The standard SAM's prompt encoder module was included in the architecture. The module contains learnable embeddings representing different prompt types (e.g., positive/negative points, box corners) and positional encoding capabilities. Crucially, it also includes a learnable embedding designed to be used as a default dense prompt when no explicit mask input is provided.

A checkpoint loading mechanism allowing model initialization from pre-trained SAM weights (typically trained on 3-channel RGB data) was implemented. The mechanism loads weights where layer names and shapes match between the checkpoint and the modified model. Layers with shape mismatches, such as the input layer or parts of the mask decoder, are skipped during loading and retain their random initialization, enabling them to be fine-tuned on the target 5-channel task.

The model was fine-tuned with the building detection dataset presented in Table 1. Binary Cross-Entropy with Logits loss was utilized for training with the AdamW optimizer.

A key aspect of the implementation relates to how prompts are handled during this fine-tuning process, diverging from the

prompt-driven training of the original SAM: During the training phase, the input dictionaries supplied contained only the 5-channel image tensor ("image") and its original spatial dimensions ("original_size"). Critically, no explicit sparse prompts (e.g., "point_coords", "point_labels", "boxes") or dense mask prompts ("mask_inputs") were provided from the dataset during training iterations.

Despite the absence of external prompts in the training data, the prompt encoder was invoked for each input image. Given `points=None`, `boxes=None`, and `masks=None`, it executed as follows:

1. It generates an empty tensor for the sparse embeddings, as no point or box coordinates are provided.

2. It utilizes its internal learnable `self.no_mask_embed` parameter. This embedding is reshaped and tiled spatially to match the dimensions of the image encoder's output feature map, effectively creating a default dense prompt embedding.

In summary, the adapted SAM model leverages the core architectural components of SAM, including the prompt encoder's mechanisms. However, the fine-tuning strategy implemented trained the model primarily as a prompt-agnostic semantic segmentation network for 5-channel building detection. It utilized the prompt encoder's default `no_mask_embed` internally but does not explicitly train the model to condition its output on user-provided sparse prompts like points or boxes. Therefore, while the architecture retains the potential for promptable segmentation, the trained weights are specialized for direct, prompt-free inference on this specific task. Early stopping was utilized based on the validation F1-score to save the best performing model during fine-tuning.

4.2.2 SAM2

SAM2.1 baseplus model was studied for fine-tuning with the 5-channel data (RGB+DSM+DEM).

The architecture of SAM2 was modified to handle 5-channel inputs and to produce segmentation of buildings. The main changes made included adding `sam2_image_predictor` a [1.0, 1.0] standard deviation list for accepting 5 channels during preprocessing and not changing the values of DSM and DEM channels. For required asserts, 3 channels were modified into 5. In addition, for `SAM2Transforms` suitable mean and standard deviation values were added. SAM2's input channel was modified for 5 channels and original weights were copied to it, in addition to adding the mean of the weights for the new fourth and fifth channels. Similarly to the original SAM, SAM2 was trained without prompts.

For fine-tuning, AdamW optimizer was utilized together with Dice loss.

4.3 Evaluation

During training, the performance of the models was evaluated with F1-score.

For the two evaluation mapsheets from Karkkila and Närpiö, the reference data was gotten from the topographic database. The Karkkila mapsheet had 1,380 buildings and the Närpiö mapsheet had 1,020 buildings. The number of correctly detected buildings by the models were inspected.

5. Results

5.1 U-Net results

To adapt the U-Net to the forest environment as well as possible, precision and recall and their relation was inspected. The model achieved a recall of 0.96 on the validation set while the F1-score was 0.89. The model found 1,158 buildings from Karkkila test tile, meaning 84% of all buildings from the area, and 939 buildings from Närpiö, mearning 92% of all buildings from the area. The model produced some false detections due to the high weight given for recall instead of precision. Model predictions with the evaluation areas can be seen in Figure 4.

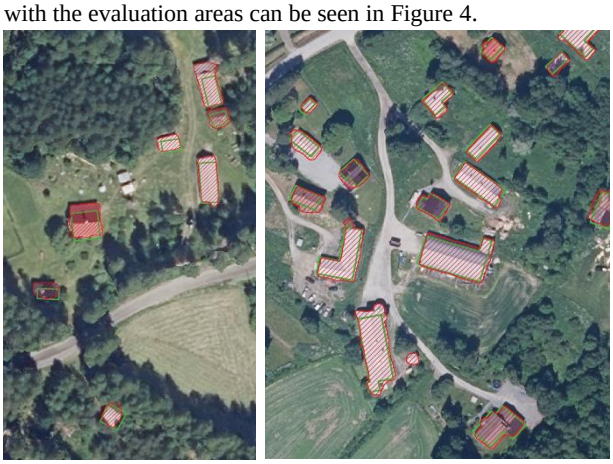

Figure 4. Results from the two test areas with the finetuned U-Net: Karkkila on the left and Närpiö on the right.

5.2 SAM results

5.2.1 Original SAM

Two different pretrained versions of the original SAM were tested to be fine-tuned with the Finnish datasets, both the smallest ViT-B model and the larger ViT-L model. After fine-tuning both models, the ViT-L model achieved slightly higher validation F1-score and was selected to be evaluated with the Karkkila and Närpiö evaluation areas. The ViT-B pretrained model achieved a validation F1-score of 0.85 after training it for 15 epochs, the ViT-L model seemed to achieve higher performance while the training continued after that and finally achieved an F1-score of 0.87. The model found 1,074 buildings from Karkkila test tile, meaning 78% of all buildings from the area, and 893 buildings from Närpiö, mearning 88% of all buildings from the area. Model predictions with the evaluation areas can be seen in Figure 5.

Figure 5. Results from the two test areas with the fine-tuned SAM, ViT-L model: Karkkila on the left and Närpiö on the right.

5.2.2 SAM2

Fine-tuned SAM2.1 baseplus model was evaluated. The fine-tuned model found 1,086 buildings from Karkkila test tile, meaning 79% of all buildings from the area, and 927 buildings from Närpiö, mearning 91% of all buildings from the area. Model predictions with the evaluation areas can be seen in Figure 6.

Figure 6. Results from the two test areas with the fine-tuned SAM2.1 baseplus model: Karkkila on the left and Närpiö on the right.

5.3 Comparison

The evaluation across the Karkkila and Närpiö test areas revealed distinct performance characteristics for each model (Table 2). The U-Net model, fine-tuned with a focus on recall (achieving 0.96 on the validation set alongside an F1-score of 0.89), demonstrated the highest building detection rates, identifying 84% and 92% of reference buildings in Karkkila and Närpiö, respectively. However, this high recall came at the cost of generating a significant number of false positive detections (81 in Karkkila, 332 in Närpiö).

Model	Karkkila (1,380 buildings)	Närpiö (1,020 buildings)
U-Net	1,158	939
SAM, ViT-L	1,074	893
SAM2	1,086	927

Table 2. Test areas, how many buildings they included in the forested areas and how many of the buildings each model found.

The fine-tuned original SAM ViT-L model achieved a balance between detection and precision. While its detection rates were slightly lower than U-Net (78% in Karkkila, 88% in Närpiö), it achieved a competitive validation F1-score (0.87) and, notably, produced considerably fewer false detections than the U-Net (46 in Karkkila, 286 in Närpiö). This suggests the ViT-L architecture, even when fine-tuned in a prompt-agnostic manner on 5-channel data, may possess good generalization capabilities or inherent regularization against spurious detections compared to the recall-weighted U-Net in this setup. The smaller SAM ViT-B variant achieved a lower validation F1-score (0.85) and was thus not selected for the final test area evaluation.

The SAM2.1 baseplus model, fine-tuned also with 5 input channels, yielded similar detection rates as the fine-tuned SAM ViT-L (79% Karkkila, 91% Närpiö). While it detected more buildings in comparison to the SAM ViT-L, it produced the most false detections (1,208 on Närpiö and 1297 on Karkkila), as the model tended to segment other objects from images near buildings. In Närpiö test area its performance was close to the U-Net's performance.

Qualitatively, the building polygons generated by the SAM ViT-L model tended to exhibit more rounded shapes compared to the U-Net outputs and were a bit more regular in comparison to SAM2.1 baseplus model's outputs. All models struggled with detecting smaller, heavily occluded buildings, highlighting the persistent challenge of segmentation under dense forest canopies.

Furthermore, a significant practical difference lies in the models' computational requirements. The U-Net model is lightweight (7.76MB), facilitating faster training and inference. In contrast, the SAM ViT-L model is substantially larger (1.25GB), demanding more computational resources and time for both training and deployment. The SAM2 model, while based on SAM, also involves a multi-step inference process that adds complexity while at the same time it offers faster image segmentation in comparison to SAM. These factors are crucial considerations for practical applications and large-scale mapping efforts.

6. Discussion

This study evaluated the effectiveness of U-Net, a fine-tuned original SAM (ViT-L), and a fine-tuned SAM2 for building segmentation in challenging forested environments using multi-modal aerial imagery and LiDAR-derived elevation data. The findings indicate that while the established U-Net architecture, particularly when fine-tuned for high recall, achieved the highest raw building detection count, the foundation model SAM, specifically the ViT-L variant adapted for 5-channel input, presented a compelling alternative by offering a better balance between detection and precision with fewer false positives. The SAM2 offered a bit higher detection rate in comparison to the ViT-L variant with the cost of higher amount of false detections but shows potential especially for tasks like detection of demolished buildings.

The superior detection rate of the U-Net can likely be attributed to two key factors: its pretraining on a large, geographically similar (though lower resolution) Finnish dataset, and the explicit fine-tuning towards high recall using Tversky loss. This pretraining likely provided a strong initialization advantage for recognizing building features common in the Finnish landscape. However, the high recall objective inevitably led to a higher rate of false positives, classifying non-building features as buildings. Conversely, the SAM ViT-L model, despite being initialized from a general-purpose checkpoint (not specifically pretrained on Finnish remote sensing data) and fine-tuned using a standard BCE loss without explicit recall weighting, demonstrated robust performance. Its ability to generate fewer false positives suggests that the ViT architecture's global attention mechanisms might capture contextual information more effectively, aiding in distinguishing buildings from spectrally or structurally similar forest elements, even when trained without prompts. The observed rounder polygon shapes from SAM could stem from the patch-based nature of ViTs or the interpolation during mask upscaling.

A critical limitation acknowledged is the difference in pretraining histories. The U-Net benefited from pretraining on relevant Finnish data, whereas the SAM models were initialized from

checkpoints trained on general natural images. Fine-tuning SAM ViT-L and SAM2.1 baseplus models on the 30cm dataset prior to the 25cm data could potentially bridge this gap and further improve its performance. Furthermore, the prompt-agnostic fine-tuning strategy employed for the SAM models, while successful for direct segmentation, did not leverage the model's inherent promptable capabilities. Exploring prompt-based fine-tuning, perhaps using initial U-Net predictions or height thresholds as prompts, could unlock additional performance gains.

The tendency of the SAM models (and to a lesser extent, the recall-focused U-Net) to produce false detections underscores the difficulty of this task. While the evaluation focused on correctly detected building counts—valuable for applications like change detection (for example, identifying demolished buildings from different areas) where false negatives are critical—reducing false positives remains crucial for clean map generation. The lower false positive rate of SAM ViT-L is promising in this regard, while also the higher detection rate of SAM2.1 baseplus model shows promise. The significant difference in model size and computational cost between U-Net and SAM is a major practical consideration, potentially favouring U-Net for resource-constrained applications or rapid large-area processing, while original SAM might be preferred where higher precision (fewer false positives) is paramount, despite the computational overhead.

Especially the SAM models trained in the experiments tended to produce false detections. In the results the number of correctly detected buildings were focused on, as it has multiple promising applications where the false detections can be ignored, for example, studying the number of demolished buildings in the areas, where false negative building detection results are more dire. The U-Net produced 332 false detections on the Närpiö test area and 81 false detections on the Karkkila test area. SAM ViT-L model produced 286 false detections on the Närpiö test area and 46 false detections on the Karkkila test area.

Future research should prioritize several avenues. Firstly, investigating the impact of pretraining the SAM models on the larger 30cm dataset before fine-tuning could clarify the influence of domain-specific pretraining. Secondly, exploring different loss functions (e.g., Focal Loss, Lovász-Softmax) and learning rate schedulers could further optimize model convergence and performance. Incorporating additional LiDAR-derived features, such as slope or vegetation indices derived from NIR (which was available but excluded), could provide richer inputs. Refining prompt-engineering techniques for both original SAM and SAM2, potentially integrating geometric priors or height information directly into the prompting mechanism, holds significant potential for improving segmentation accuracy and robustness in complex forested landscapes. For example, exploring the usage of RGB remote sensing data together with other LiDAR-based products and incorporating DEM-derived slope information as segmentation prompts for SAM could offer new perspectives.

3D point cloud data has been investigated together with SAM-based approach (Yarroudh et al., 2023), but in the future, new SAM-based methods and even more accurate point cloud data could enhance the accuracy of building detection in the forested areas.

7. Conclusions

This study compared the performance of a U-Net, a 5-channel adapted Segment Anything Model (SAM ViT-L), and a 5-channel adapted SAM2 for the challenging task of building segmentation in forested areas using true orthophotos and LiDAR-derived DSM/DEM data.

The results demonstrate that while the U-Net, benefiting from relevant pretraining and recall-focused fine-tuning, achieved the highest building detection rate, the fine-tuned SAM models offered a strong alternative, achieving competitive detection rates. This suggests that foundation models like SAM and SAM2.1, when appropriately adapted for multi-modal remote sensing data, can effectively leverage combined image and elevation information for complex segmentation tasks, potentially offering better precision than traditional architectures fine-tuned solely for recall.

The study highlights the persistent difficulties in segmenting small and occluded buildings under dense forest canopies for all tested architectures. While height information from LiDAR aids segmentation, occlusion remains a significant hurdle. The trade-off between detection rate (recall) and precision, as well as the substantial differences in model size and computational requirements, are critical factors for selecting appropriate models for practical applications.

Overall, while U-Net currently provides the highest detection count in our specific setup, the adapted SAM ViT-L demonstrates significant promise for balancing detection performance with reduced false positives in forested building segmentation and SAM2.1 baseplus model regarding the higher detection rate in comparison to the original fine-tuned SAM. Future work focusing on domain-specific pretraining, advanced prompting strategies, and exploration of alternative loss functions is warranted to further unlock the potential of foundation models for automated and accurate mapping in complex remote sensing environments.

8. Acknowledgements

The authors wish to thank CSC - IT Center for Science, Finland (urn:nbn:fi:research-infras-2016072531) and the Open Geospatial Information Infrastructure for Research (Geoportti, urn:nbn:fi:research-infras-2016072513) for computational resources and support.

References

Awrangjeb, M., Lu, G., & Fraser, C. (2014). Automatic Building Extraction From LIDAR Data Covering Complex Urban Scenes. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XL-3. 25-32. 10.5194/isprsarchives-XL-3-25-2014.

Guan, H., Li, J., Chapman, M., Deng, F., Ji, Z., Yang, X., 2013. Integration of orthoimagery and lidar data for object-based urban thematic mapping using random forests. International Journal of Remote Sensing, 34(14), 5166-5186.

Hattula, E., Zhu, L., Raninen, J., Oksanen, J., Hyyppä, J., 2023. Advantages of Using Transfer Learning Technology with a Quantative Measurement. Remote Sensing, 15(17). https://www.mdpi.com/2072-4292/15/17/4278.

Hattula, E., Zhu, L., & Raninen, J. (2024). Building extraction in urban and rural areas with aerial and Lidar DSM. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 10, 73-79.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. arXiv preprint arXiv:2304.02643.

Miliaresis, G., & Kokkas, N. (2007). Segmentation and object-based classification for the extraction of the building class from LIDAR DEMs. Computers & Geosciences. 33. 1076-1087. 10.1016/j.cageo.2006.11.012.

Ošep, A., Meinhardt, T., Ferroni, F., Peri, N., Ramanan, D., & Leal-Taixé, L. (2024). "Better Call SAL: Towards Learning to Segment Anything in Lidar." arXiv preprint arXiv:2403.13129.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., & Feichtenhofer, C. (2024). SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 9351, 234-241. arXiv:1505.04597.

Yarroudh, A. (2023). LiDAR Automatic Unsupervised Segmentation using Segment-Anything Model (SAM) from Meta AI [GitHub repository].

Zhang, Y., Wang, R., Wu, Y., Chu, G., & Wu, X. (2024). Segment Anything Model-Based Building Footprint Extraction for Residential Complex Spatial Assessment Using LiDAR Data and Very High-Resolution Imagery. Remote Sens., 16(14), 2661.

Zhang, Y., et al. (2023). *Evaluating SAM on Satellite Imagery: Strengths and Limitations for Geospatial Applications*. Remote Sensing, 15(10).

García, M., et al. (2024). *Segmenting Forest Canopies with SAM and LiDAR: A Comparative Study*. International Journal of Applied Earth Observation.

Xu, Y., et al. (2021). Building Extraction from Very High Resolution Aerial Imagery Using Joint Attention Deep Neural Network. *Remote Sensing, 13*(1), 109.