environmental sciences proceedings

MDPI

Article

Zero-Shot Refinement of Buildings' Segmentation Models using SAM†

Ali Mayladan[1,2], Hasan Nasrallah[2], Hasan Moughnieh[2], Mustafa Shukor[3] and Ali J. Ghandour[2]*

[1] Lebanese University, Lebanon;
[2] National Center for Remote Sensing - CNRS, Lebanon;
[3] Sorbonne University, France;
* Corresponding author: aghandour@cnrs.edu.lb;
† Presented at the 5th International Electronic Conference on Remote Sensing, 721 November 2023.

Abstract: Foundation models have excelled in various tasks but are often evaluated on general benchmarks. The adaptation of these models for specific domains, such as remote sensing imagery, remains an underexplored area. In remote sensing, precise building instance segmentation is vital for applications like urban planning. While Convolutional Neural Networks (CNNs) perform well, their generalization can be limited. For this aim, we present a novel approach to adapt foundation models to address existing models' generalization dropback. Among several models, our focus centers on the Segment Anything Model (SAM), a potent foundation model renowned for its prowess in class-agnostic image segmentation capabilities. We start by identifying the limitations of SAM, revealing its suboptimal performance when applied to remote sensing imagery. Moreover, SAM does not offer recognition abilities and thus fails to classify and tag localized objects. To address these limitations, we introduce different prompting strategies, including integrating a pre-trained CNN as a prompt generator. This novel approach augments SAM with recognition abilities, a first of its kind. We evaluated our method on three remote sensing datasets, including the WHU Buildings dataset, the Massachusetts Buildings dataset, and the AICrowd Mapping Challenge. For out-of-distribution performance on the WHU dataset, we achieve a 5.47% increase in IoU and a 4.81% improvement in F1-score. For in-distribution performance on the WHU dataset, we observe a 2.72% and 1.58% increase in True-Positive-IoU and True-Positive-F1 score, respectively. Our code is publicly available at this Repo, hoping to inspire further exploration of foundation models for domain-specific tasks within the remote sensing community.

Keywords: foundation models; buildings footprint; instance segmentation; Segment Anything Model; prompt engineering

Citation: Mayladan, A.; Nasrallah, H.; Moughnieh, H.; Shukor M.; Ghandour A. J. Zero-Shot Refinement of Buildings' Segmentation Models using SAM . *Environ. Sci. Proc.* **2023**, *1*, 0. https://doi.org/

Published:

1. Introduction

Most current state-of-the-art remote sensing models are CNN-based [1] that struggle with out-of-distribution generalization. This challenge is mainly due to the significant variations in imagery when observed in various regions, seasons, and periods. This demonstrates the need for more robust and adaptive techniques to manage these variances efficiently.

Foundation models [2] have demonstrated unparalleled proficiency in a wide range of tasks, from high-resolution image interpretation to multi-modal data analysis. These models not only have equaled, but frequently outperformed, the performance of previous task-focused algorithms, particularly in complex tasks such as dense prediction and spatial pattern recognition [3]. However, many of these models have been trained and benchmarked primarily against generic datasets and usually underperform on domain-specific tasks like remote sensing segmentation.

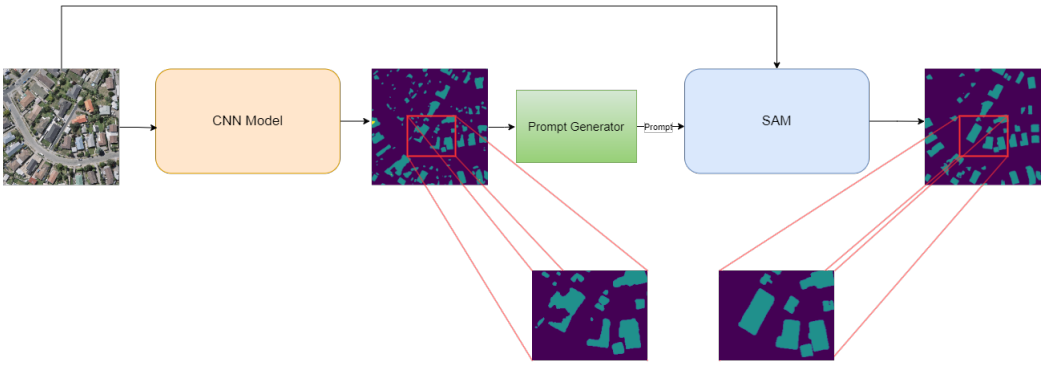

Figure 1. Input RGB image undergoes rooftop instance segmentation via the CNN model. Segmentation masks are passed to the Prompt Generator used to prompt SAM. This approach would equip SAM with recognition abilities and generate precise buildings output masks.

The use of these foundation models for remote sensing applications such as land cover categorization, change detection, and instance segmentation is still unexplored. Therefore, the adaptation of foundational models is crucial to address the evolving challenges in satellite and aerial data analysis.

A central question then arises: **How can foundation models be effectively leveraged for remote sensing segmentation, notably, in buildings' footprint instance segmentation?**

We mainly focus in this manuscript on Meta's newly unveiled transformer-based "Segment Anything Model" (SAM) [4], a powerful foundation model for image segmentation, promising broad applicability and high accuracy. SAM is trained on an extensive high-quality dataset (SA-1B) encompassing more than 11 million images and more than 1.1 billion masks, constituting the largest segmentation dataset, with 400x more masks than any existing segmentation dataset.

Although SAM excels in localization capabilities, it does not offer recognition abilities and thus fails to classify and tag localized objects. Therefore, the use of SAM for segmentation is not a straightforward task. Hence, we propose to leverage the SAM foundation model to improve the performance of pre-trained CNN segmentation models. Complementing CNNs with SAM might harness: (*i*) the collaboration between CNNs and transformers, on the one hand, and (*ii*) the generalization power of SAM with the domain specificity of pre-trained CNNs. Specifically, we propose to Prompt Engineer (PE) SAM to enhance its performance by integrating a pre-trained CNN as a prompt generator.

Our contributions can be summarized as follows: (*i*) Investigate various distinct single and composite prompting strategies for SAM and (*ii*) Experiment with two CNN models as prompt generators for SAM to get more accurate instance segmentation results.

2. Related Work

In remote sensing, instance segmentation is vital for precisely extracting building footprints from satellite and aerial imagery, allowing for exact land-use analysis and infrastructure planning. It provides real-time insights for various geographic tasks by differentiating between individual items of the same class.

Foundation models have emerged as a transformational force in the ever-evolving field of artificial intelligence. Among these, large language models (LLMs) [5,6] have demonstrated unparalleled capabilities in natural language processing and generation, enabling a wide range of applications ranging from chatbots to content development.

On another front, multimodality-based models [7,8] can analyze and integrate data from various modalities such as images, text, and sound. Within the visual domain, the foundation models [9–11] have set new standards for image recognition, object detection, and various other computer vision tasks. These models have served as the backbone for many applications, ranging from autonomous vehicles to healthcare diagnostics.

Leveraging and adapting foundation models to specialized tasks has been a notable trend in recent research [12,13]. SAM, specifically, has been widely deployed in a short period of time for various applications [14–17]. With a strong Zero-Shot performance, SAM has attracted attention for its outstanding capacity to generate high-quality object masks from a variety of input prompts.

3. SAM Prompt Engineering

Experiment	Precision	Recall	IoU	F1	TP-IoU	TP-F1
Single-point	69.40	63.63	47.21	51.83	81.12	89.05
Single-point + Negative-point	72.31	66.52	50.62	55.36	81.89	89.56
Skeleton Multiple-points	83.18	76.97	60.97	65.96	84.15	91.03
Random Multiple-points	84.12	78.01	61.52	66.64	83.89	90.88
Random Multiple-points + Single-point	84.09	78.04	61.86	67.00	83.92	90.89
Random Multiple-points + Negative-point	83.72	77.68	61.12	66.23	83.79	90.81
Bounding-box	84.78	78.62	63.82	68.52	**85.67**	**91.98**
Bounding-box + Single-point	**84.88**	78.72	**63.86**	**68.62**	85.53	91.90
Bounding-box + Multiple-points	84.87	**78.81**	63.54	68.43	85.10	91.65
baseline U-Net-based CNN [18]	84.76	78.68	61.79	67.34	82.95	90.40

Table 1. Comprehensive set of experiments conducted while integrating the U-Net CNN model [18] with SAM on WHU Buildings' dataset, encompassing various prompt types. These experiments are evaluated based on precision, recall, IoU, F1-score, True-Positive IoU (TP-IoU) and True-Positive F1-score (TP-F1) metrics.

Meta AI recently unveiled the Segment Anything Model (SAM) [4], a class-agnostic segmentation model that incorporates automatic mask generation and quality filters. SAM utilizes a Vision Transformer (ViT) for image encoding and employs a two-layer mask decoder with transformer-based architecture. SAM has outstnading localization capabilities but lacks any recognition abilities.

In our proposed methodology shown in Figure 1, we augmnented SAM with the capability to recognize objects, mainly buildings. The input image is fed initially to a CNN-based model pre-trained for buildings' instance segmentation. We then developed a prompt generator component capable of providing SAM with various promptst. At the core of our proposed architecture lies this prompt generator, which operates by taking the output masks of the CNN model as input and using them to generate SAM prompts of the following three different categories: (*i*) single-point prompts, where a single representative point is generated for each input mask. (*ii*) Multiple-point prompts, using either random points localized within the input mask or by extracting skeleton-shaped points from the input buildings' mask. (*iii*) Bounding box prompts for each mask, with the box coordinates serving as prompts for SAM.

The proposed component can also generate hybrid prompts of various categories such as a "single-point and bounding-box" prompt. We also used the concept of negative points that can be located either in the image background or inside the bounding box, but not within the building's mask. More details about the three prompt categories are presented in Table 1.

We experimented with two different CNN models trained for buildings' footprints instance segmentation: (*i*) a U-Net-based CNN model [18] that employs Efficient-Net-B3 backbone for feature extraction, ensuring accuracy and precision in mask generation, and (*i*) D-LinkNet [19] that builds upon LinkNet and utilizes ResNet34 as encoder. The encoder includes dilated convolution layers for context capture, and the decoder efficiently restores feature map resolution through transposed convolution layers. More details of these experiments are elaborated on in the next section.

Experiment	Precision	Recall	IoU	F1	TP-IoU	TP-F1
Bounding-box	**43.25**	**52.16**	**29.96**	**32.44**	**83.72**	**90.64**
baseline DCNN [19]	39.89	47.59	24.49	27.63	76.16	85.93

Table 2. SAM Bouding-box prompt results using the D-LinkNet mode [19] out-of-distribution on the WHU Buildings dataset.

4. Experimental Results

In this section, we provide insights into our dual-architecture shown in Figure 1 that consists of a CNN prompt generator and SAM for Zero-Shot mask refinement.

We conducted a series of experiments to assess the performance of SAM under various types of prompts in three remote sensing datasets: the WHU Buildings dataset [20] and the Massachusetts Buildings dataset [21], in addition to the AICrowd Mapping Challenge dataset [22]. We consistently used prediction masks generated by one of the two CNN models as input to the prompt generator.

As detailed in Table 1, we initially used single-point prompts for buildings' rooftop instance segmentation. We replace each CNN-predicted building's mask by a single-representative point, which is then provided as input to SAM alongside the original RGB image. The representative point, by definition, is guaranteed to be within the building, irrespective of the building's shape. The SAM output in Figure 2(a) reveals that the representative single-point sometimes fails to accurately segment the target object due to irregular shapes (e.g., L-shaped or U-shaped buildings). Single-point prompt scores 47.21%, 51.83%, 81.12% and 89.05%, in terms of IoU, F1, TP-IoU, and TP-F1 scores, respectively, on the WHU dataset. Using one Negative-point along the single-point improved IoU and F1-score with more than 3%.

We explored the use of multiple-points prompts, to ensure comprehensive coverage of the building area, and improve segmentation accuracy, particularly for larger buildings where a single-point prompt might be insufficient to encompass the entire structure. We distribute 5 points within each mask following two approaches: (*i*) Random distribution as shown in Figure 2(d) and (*ii*) Skeleton form depicted in Figure 2(c) where one point is the building centroid and the others along the edges. Surprisingly, Table 1 reveals that both Random and SkeletonMultiple-points prompt almost exhibit the same performance. Future research is needed to investigate why the Skeleton approach did not outperform the random scheme. We also conducted additional pairs of experiments using "RandomMultiple-points + Single-point" and "RandomMultiple-points + Negative-point" where both did not provide substantial improvement.

Among all the experiments depicted in Table 1, the Bounding-box prompts exhibited the most promising results. We explored three difference version including: (*i*) Bounding-box, (*ii*) Bounding-box + Single-point and (*ii*) Bounding-box +Multiple-points. The prompt of SAM with Bounding-box led to 2.03%, 1.18%, 2.72% and 1.58% improvement in terms of IoU, F1, TP-IoU, and TP-F1 scores, respectively, on the WHU dataset.

We also performed prompt engineering experiments with bounding-boxes using D-LinkNet CNN [19] out-of-distribution on the WHU dataset. Using bounding-box prompt showed substantial improvement with a **5.47%**, **4.81%**, **7.56%** and **4.71%** increase in IoU, F1-score, TP-IoU and TP-F1-score, respectively, on the WHU dataset as shown in Table 2.

Additionally, we expanded our experiments to include the Massachusetts buildings and the AICrowd Mapping Challenges datasets using bounding boxes as prompts. On the AICrowd dataset, SAM proficiently predicts building segments, even those obscured by trees, while ground truth designates tree-covered sections as integral building parts. Similarly, on the Massachusetts Buildings dataset, we noticed improvements in terms of TP-IoU and TP-F1. Detailed results over these two datasets are omitted for space limitations.

5. Conclusion

In this paper, we propose to leverage SAM model in the domain of building segmentation for remote sensing applications. Our approach introduces a novel adaptation paradigm based on prompting, where we exploit the power of a pre-trained CNN as a prompt generator. We conduct an extensive evaluation of our approach on the WHU dataset, yielding remarkable improvements in SAM's building segmentation accuracy. In the context of out-of-distribution performance, our results demonstrated an impressive boost, with a notable 5.47% enhancement in IoU and a substantial 4.81% improvement in F1-score. Moreover, our evaluation also revealed noteworthy enhancements for in-distribution performance on the WHU dataset, showcasing a 2.72% increase in True-Positive-IoU and a significant 1.58% enhancement in True-Positive-F1-score. These results underline the effectiveness of our method in diverse scenarios. We hope this work will inspire the broader academic community to explore the potential of foundation models for domain-specific tasks.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. O'Shea, K.; Nash, R. An Introduction to Convolutional Neural Networks, 2015, [arXiv:cs.NE/1511.08458].
2. Awais, M.; Naseer, M.; Khan, S.; Anwer, R.M.; Cholakkal, H.; Shah, M.; Yang, M.H.; Khan, F.S. Foundational Models Defining a New Era in Vision: A Survey and Outlook, 2023, [arXiv:cs.CV/2307.13721].
3. Zuo, S.; Xiao, Y.; Chang, X.; Wang, X. Vision transformers for dense prediction: A survey. *Knowledge-Based Systems* **2022**, *253*, 109552. https://doi.org/https://doi.org/10.1016/j.knosys.2022.109552.
4. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything, 2023, [arXiv:cs.CV/2304.02643].
5. OpenAI. GPT-4 Technical Report, 2023, [arXiv:cs.CL/2303.08774].
6. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023, [arXiv:cs.CL/2307.09288].
7. Shukor, M.; Dancette, C.; Rame, A.; Cord, M. Unified Model for Image, Video, Audio and Language Tasks. *arXiv preprint arXiv:2307.16184* **2023**.
8. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, 2023, [arXiv:cs.CV/2303.05499].
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition, 2015, [arXiv:cs.CV/1512.03385].
10. Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; Huang, T. SegGPT: Segmenting Everything In Context, 2023, [arXiv:cs.CV/2304.03284].
11. 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, [arXiv:cs.CV/2103.14030].
12. Shukor, M.; Dancette, C.; Cord, M. eP-ALM: Efficient Perceptual Augmentation of Language Models. *arXiv preprint arXiv:2303.11403* **2023**.
13. Zhang, Y.; Gao, J.; Zhou, M.; Wang, X.; Qiao, Y.; Zhang, S.; Wang, D. Text-guided Foundation Model Adaptation for Pathological Image Classification, 2023, [arXiv:cs.CV/2307.14901].
14. Wu, J.; Zhang, Y.; Fu, R.; Fang, H.; Liu, Y.; Wang, Z.; Xu, Y.; Jin, Y. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation, 2023, [arXiv:cs.CV/2304.12620].
15. Osco, L.P.; Wu, Q.; de Lemos, E.L.; Gonçalves, W.N.; Ramos, A.P.M.; Li, J.; Junior, J.M. The Segment Anything Model (SAM) for Remote Sensing Applications: From Zero to One Shot, 2023, [arXiv:cs.CV/2306.16623].
16. Ding, L.; Zhu, K.; Peng, D.; Tang, H.; Guo, H. Adapting Segment Anything Model for Change Detection in HR Remote Sensing Images, 2023, [arXiv:cs.CV/2309.01429].
17. Zhang, R.; Jiang, Z.; Guo, Z.; Yan, S.; Pan, J.; Dong, H.; Gao, P.; Li, H. Personalize Segment Anything Model with One Shot, 2023, [arXiv:cs.CV/2305.03048].
18. Nasrallah, H.; Samhat, A.E.; Shi, Y.; Zhu, X.X.; Faour, G.; Ghandour, A.J. Lebanon Solar Rooftop Potential Assessment Using Buildings Segmentation From Aerial Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2022**, *15*, 4909–4918.
19. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet With Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018.
20. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing* **2019**, *57*, 574–586. https://doi.org/10.1109/TGRS.2018.2858817.

21. Mnih, V. Machine Learning for Aerial Image Labeling. PhD thesis, University of Toronto, 2013.
22. Mohanty, S.P.; Czakon, J.; Kaczmarek, K.A.; Pyskir, A.; Tarasiewicz, P.; Kunwar, S.; Rohrbach, J.; Luo, D.; Prasad, M.; Fleer, S.; et al. Deep Learning for Understanding Satellite Imagery: An Experimental Survey. *Frontiers in Artificial Intelligence* **2020**, *3*.

(a) Single-point

(b) Single + Negative-points

(c) Skeleton Multiple-points

(d) Random Multiple-points

(e) Bounding-box

Figure 2. Visualizations, over three different images from WHU dataset, of prompt engineering experiments including Single-point, Single-point + Negative-points (in red), Skeleton Multiple-points, Random Multiple-points and Bounding-box.