Comparative Analysis of Vision Foundation Models For Building Segmentation in Aerial Imagery

Zeynep Akbulut [1], Samed Özdemir [1], Fevzi Karslı [2]

[1] Gumushane University, Department of Geomatics Engineering, Gumushane, Turkey – (zeynepakbulut, samed.ozdemir)@gumushane.edu.tr
[2] Karadeniz Technical University, Department of Geomatics Engineering, Trabzon, Turkey – fkarsli@ktu.edu.tr

Keywords: Aerial Imagery, Building Segmentation, Grounded-SAM, Segment Anything Model, Vision Foundation Models.

Abstract

Visual Foundation Models (VFMs) demonstrate impressive generalization capabilities for image segmentation and classification tasks, leading to their increasing adoption in the remote sensing field. This study investigates the performance of VFMs in zero-shot building segmentation from aerial imagery using two model pipelines: Grounded-SAM and SAM+CLIP. Grounded-SAM integrates the Grounding DINO backbone with a Segment Anything Model (SAM) while SAM+CLIP first employs SAM for generating masks followed by Contrastive Language Image Pretraining (CLIP) for classification. The evaluation, performed on the WHU building dataset using Precision, Recall, F1 score, and intersection over union (IoU) metrics, revealed that Grounded-SAM achieved F1-score of 0.83 and IoU of 0.71. SAM+CLIP achieved F1-score of 0.65 and IoU of 0.49. While Grounded-SAM excelled at accurately delineating partially occluded and irregularly shaped buildings, SAM+CLIP was able to segment larger buildings but struggled with delineating smaller ones. Given the impressive performance of VFMs in zero-shot building segmentation, future efforts aimed at refining these models through fine-tuning or few-shot learning could significantly expand their application in remote sensing.

1. Introduction

Accurate building segmentation is essential for various applications, including urban planning, monitoring, and mapping (Hajjar et al., 2024). In the field of remote sensing (RS), many researchers (Erdem and Avdan, 2020; Wang et al., 2022; Chang et al., 2024; Yildirim et al., 2024) have contributed to studies on building segmentation using deep learning (DL). Over the past few years, DL has emerged as a widely used method for automatic feature learning, driving significant advancements in computer vision (Wang et al., 2022). One of the key challenges in deploying deep neural networks in real-world applications is their dependence on large amounts of precisely annotated training data, particularly for dense prediction tasks like semantic segmentation and change detection (Ding et al., 2024). Another point noted by Li et al. (2021) is that each sensor requires its own training data, and single-sensor models cannot be effectively transferred to other sensors. Lately, Vision Foundation Models (VFMs) have emerged and attracted considerable attention in computer vision research (Ding et al., 2024).

In recent years, a variety of foundational models have been developed (Liu et al., 2024a), and interest in these models has surged due to their extensive pre-training on web-scale datasets, which grants them a remarkable ability to generalize across various downstream tasks (Ji et al., 2024). Liu et al. (2024a) provided examples of foundation models for computer vision, including SimCLR (Chen et al., 2020), Masked AutoEncoder (MAE) (He et al., 2022), and Segment Anything Model (SAM) (Kirillov et al., 2023). Moreover, recent innovations in VFMs have focused on integrating multiple foundational models into unified pipelines to leverage their complementary strengths. A notable example for these hybrid frameworks is Grounded-SAM, introduced by Ren et al. (2024), which combines Grounding DINO (Liu et al., 2024b), an open-set object detector, with SAM. Similarly, OV-SAM (Yuan et al., 2025) integrates SAM with open-vocabulary detector Contrastive Language Image Pretraining (CLIP) (Radford et al., 2021), allowing for segmentation based on textual descriptions.

In the remote sensing field there are several new foundational models that employs VFMs. Liu et al. (2024a) introduce RemoteCLIP, marking it as the first vision-language foundation model tailored for RS. RSPrompter (Chen et al., 2024) which is based on SAM, is another adaptation of VFMs for RS applications. SAM-RSIS (Luo et al., 2024) is based on fine-tuned SAM and has automatic box prompting for remote sensing instance segmentation applications. RingMo-SAM (Yan et al., 2023), a multimodal image segmentation model based on SAM which can segment and identify categories of optical imagery and synthetic aperture radar (SAR) images.

In this study, we explored the use of two foundational models Grounded-SAM and SAM+CLIP pipeline for zero-shot building segmentation in aerial imagery. The SAM+CLIP pipeline works in two stages: first, images are segmented using SAM, specifically the SAM Automatic Mask Generator with the pre-trained ViT-H model. Once segmentation is complete, the second stage employs the CLIP RSICD model (Arutiunian et al., 2021) for zero-shot classification. This model integrates CLIP introduced by Radford et al. (2021), with the Remote Sensing Image Caption Dataset (RSICD) developed by Lu et al. (2018). In the Grounded-SAM model (Ren et al., 2024) pipeline, input images goes through the grounding stage in which the model looks for image segments align with the given prompt. After the grounding stage, SAM model segments the detected object with the predictor method using the bounding boxes extracted in the grounding stage. For the experiments, the WHU aerial image dataset (Ji et al., 2019) was used, and precision, recall, F1 score, and Intersection over Union (IoU) metrics were employed to evaluate the performance of the models. Also, we evaluated the CPU and GPU processing times of the models. To evaluate the models' sensitivity to textual prompts, we ran experiments in which we systematically varied the input prompts and observed the resulting changes in segmentation performance.

2. Methodology

2.1 Review of Vision Foundation Models

Foundational models, a concept introduced by Bommasani et al. (2021) at Stanford’s Institute for Human-Centered AI, are primary models developed using extensive self-supervised or semi-supervised training on large datasets and can be adapted for various downstream applications (Awais et al., 2025). These applications include autonomous driving, medical diagnostics, and remote sensing image analysis (Yu et al., 2024). Training on multi-modal data enables foundation models to recognize complex patterns, ensuring strong generalization and robustness across all types of tasks (Yu et al., 2024). A notable example is the SAM, which offers class-agnostic segmentation capabilities for specialized domains such as medical imaging, robotics, or remote sensing (Awais et al., 2025). As stated by Ding et al. (2024), another significant example is CLIP (Radford et al., 2021), a model that represents visual content through textual descriptions and was trained on 400 million imagetext pairs. Its zero-shot image classification performance is on pair with fully-supervised convolutional neural network (CNN) (Ding et al., 2024).

2.1.1 Segment Anything Model: The SAM is a newly introduced image segmentation model, trained on one of the largest datasets in computer vision, featuring over one billion masks from 11 million images (Kirillov et al., 2023). Notably, SAM exhibits remarkable zero-shot transfer capabilities, frequently surpassing prior supervised methods. First proposed by Kirillov et al. (2023), SAM is specifically designed for promptable segmentation and is structured around three key components: an image encoder, a prompt encoder, and a mask decoder (Figure 1).

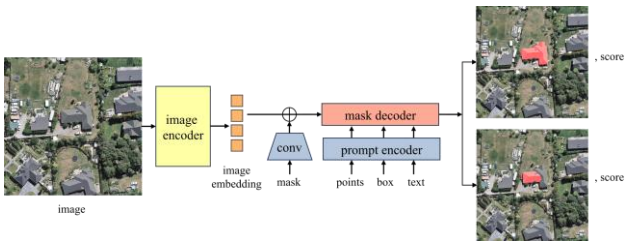

Figure 1. Segment Anything Model (adapted from Kirillov et al. (2023)).

2.1.2 Contrastive Language Image Pre-Training Model: The CLIP model (Figure 2), introduced by Radford et al. (2021), offers an innovative way to learn visual representations using natural language supervision. Rather than striving to match exact text descriptions with images, CLIP employs a contrastive learning strategy by identifying correct image-text pairs from a large batch of candidates. This is accomplished via a symmetric cross-entropy loss applied to the cosine similarity between image and text embeddings (Radford et al., 2021). It is worth noting, however, that the original CLIP model was not explicitly trained on remote sensing imagery. For this reason we choose to run CLIP RSICD Arutiunian et al. (2021) model which is CLIP model fine tuned on Remote Sensing Image Caption Dataset (RSICD) by Lu et al. (2018).

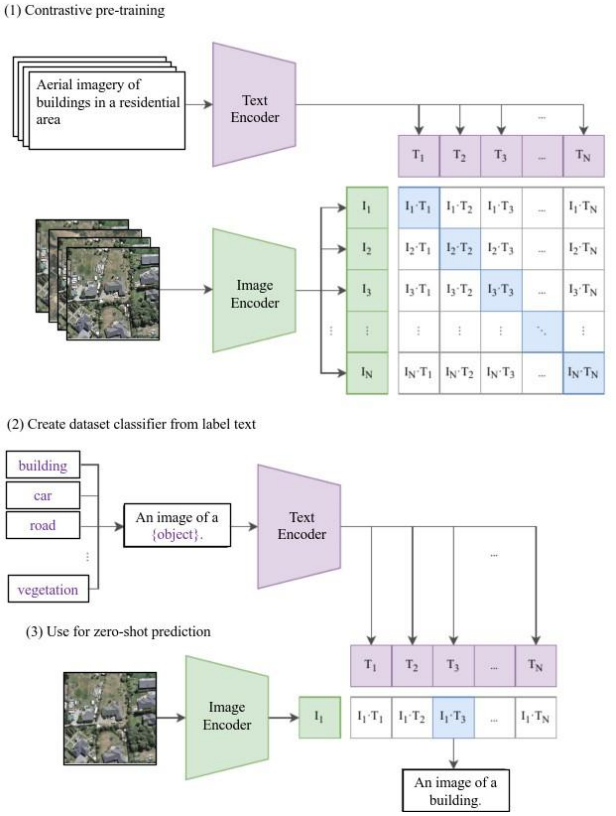

Figure 2. CLIP model architecture (adapted from Radford et al. (2021)).

2.1.3 Grounded-SAM: Ren et al. (2024) presented Grounded-SAM, which leverages Grounding DINO (Liu et al., 2024b) as an open-set object detector in conjunction with the SAM. When provided with an input image and a text prompt, Grounded-SAM initially utilizes Grounding DINO to create accurate bounding boxes for objects or regions by conditioning on the textual information. These annotated boxes then serve as prompts for SAM, which generates precise mask annotations (Ren et al., 2024). Grounding DINO (Liu et al., 2024b) is a dual-encoder-single-decoder framework designed to detect and localize objects in an image while associating them with corresponding textual prompts. The model uses dedicated backbones to extract vanilla image and text features, which are then merged through a feature enhancer. A language-guided query selection module further refines these fused features, generating cross-modality queries that the cross-modality decoder uses to produce bounding boxes and extract matching phrases. As a result, Grounding DINO can perform both object detection and referring expression comprehension by aligning object proposals with the relevant text inputs (Liu et al., 2024b).

2.2 Dataset

In this study, we employed the WHU Building Aerial Imagery Dataset (Ji et al., 2019) to evaluate the performance of the models. The dataset consists of 8,189 aerial images, each image has 512 × 512 pixels with a ground resolution of 0.3 meters. Since the models do not require any training, we proceeded directly to the testing phase. The WHU Building Dataset includes 3,810 test images, from which we randomly selected 10% (381 images) to conduct our experiments.

2.3 Experiments

In this study, both Grounded-SAM and SAM+CLIP pipelines were utilized to segment building boundaries in a variety of urban environments. Experiments were conducted on a workstation equipped with an Intel i5-11400F processor, 32 GB of RAM, and an NVIDIA RTX 3060 GPU with 12 GB of dedicated memory. For the Grounded-SAM pipeline (Figure 3), we employed the pre-trained groundingDINO swinb cogcoor checkpoint, which integrates a SwinB backbone with coordinate-based enhancements for improved spatial alignment. In the Grounded-SAM pipeline, input images goes through the grounding stage in which the model looks for image segments align with the given prompt which is building. After the grounding stage, SAM segments the detected object, which is defined by a bounding box, utilizing predictor method of SAM with default parameters.

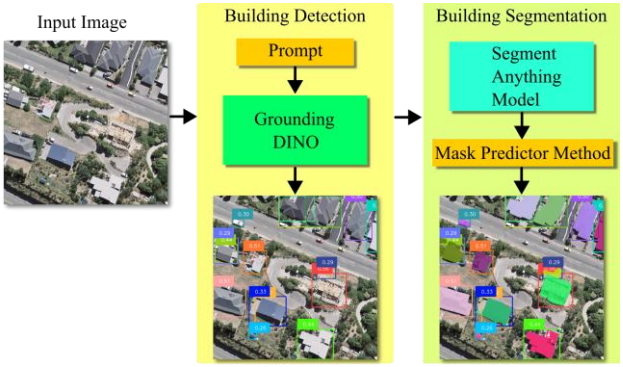

Figure 3. Grounded-SAM pipeline.

For the SAM+CLIP pipeline (Figure 4), the pre-trained *CLIP RSICD v4* model utilized. For the SAM model the pre-trained ViT-H backbone was employed. In this pipeline, the SAM model's automatic mask generator method was used with the default parameters. This method prompts the SAM model with a predefined number of points spread equally inside the image boundary in a grid pattern. SAM then identifies and generates segments within the image. Following the segmentation step, each extracted image segment is fed into the CLIP model to determine semantic categories. The desired class names are merged into a prompt, which guides the classification process. The classes used in the input prompt include: building, roof, parking, house, commercial, center, medium residential, square, industrial, dense residential, bare land, sparse residential, grassland, meadow, forest, and park. Image segments with a cumulative probability of 85% for the building, roof, and house classes are marked as buildings, while segments below this threshold are discarded.

2.4 Evaluation Metrics

In order to evaluate the performance of the Grounded-SAM and the SAM+CLIP, precision, recall, F1 score and IoU metrics were employed. The formulas are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

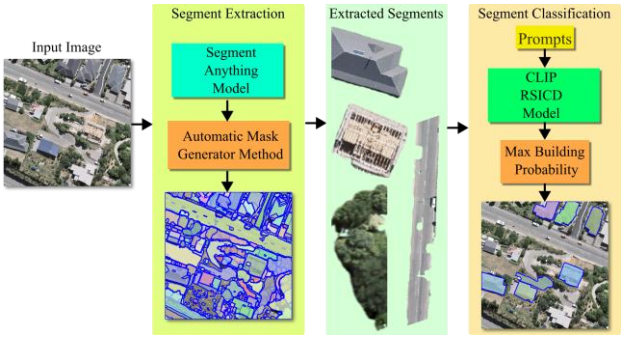

Figure 4. SAM+CLIP pipeline.

$$\text{Precision} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{4}$$

where, TP is true positive, TN is true negative, FP is false positive, and finally FN is false negative.

3. Results

3.1 Building Extraction Performance

In this section, we present the results of the Grounded-SAM and SAM+CLIP pipelines. Both the Grounded-SAM and SAM+CLIP pipelines accurately segment building boundaries across diverse urban environments, including industrial complexes and densely populated residential districts. This is further supported by the accuracy metrics shown in Table 1. Grounded-SAM outperforms SAM+CLIP across all evaluation metrics. Specifically, Grounded-SAM achieves precision of 0.86, recall of 0.80, F1-score of 0.83, and IoU of 0.71. In contrast, SAM+CLIP achieves slightly lower values: 0.69 for precision, 0.62 for recall, 0.65 for F1-score, and IoU of 0.49. The obtained results are presented through six examples from the study area, as shown in Figure 5.

MODEL	Precision	Recall	F1-Score	IoU
SAM+CLIP	0.69	0.62	0.65	0.49
Grounded-SAM	0.86	0.80	0.83	0.71

Table 1. Building extraction results for Grounded-SAM and SAM+CLIP models.

Despite roof shape and colour variations or partial occlusions from surrounding vegetation (Figure 6), the segmentation masks produced by both methods align closely with actual building boundaries. Notably, Grounded-SAM demonstrates enhanced boundary precision in challenging areas such as roofs partially occluded by trees which indicates the grounding mechanism contributes additional contextual cues for more accurate building extraction.

3.2 Prompt Selection

In this section, we evaluate the prompt sensitivity of the Grounded-SAM and the SAM+CLIP pipelines. For the Grounded-SAM, we used three distinct prompts: structure, building, and roof. During the building detection phase, the

Figure 5. Grounded-SAM and SAM+CLIP results.

model responded similarly to each prompt (Figure 7). Focusing on the Test1 3213 area, which is a medium-density residential zone with similarly sized houses, we observed that the building prompt enabled the model to detect an additional small building located at the eastern corner of the image.

In an industrial scene such as Test3 101, the Grounded-SAM demonstrates behavior consistent with previous observations. In this scene, all of the prompts yields slightly different outcomes for building detection with the Grounding DINO component. The building detection results for the different prompts are shown in Figure 7. As shown in Figure 8, in Test3 101, the SAM model struggles to accurately extract large roofs, particularly for buildings with white roofs that feature repeating textures.

To evaluate the CLIP model’s sensitivity to different input prompts, we compiled three prompt sets (Figure 9). Prompt 1 included classes such as building, roof, parking, house, commercial, center, medium residential, square, industrial, dense

Figure 6. Buildings extracted with Grounded-SAM.

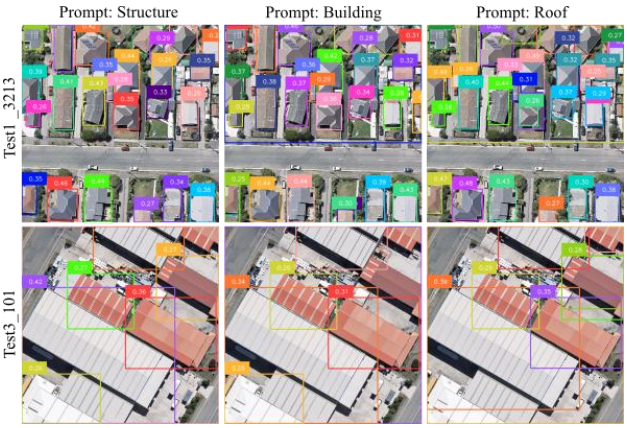

Figure 7. Building detection results for different prompts with Grounded-SAM model.

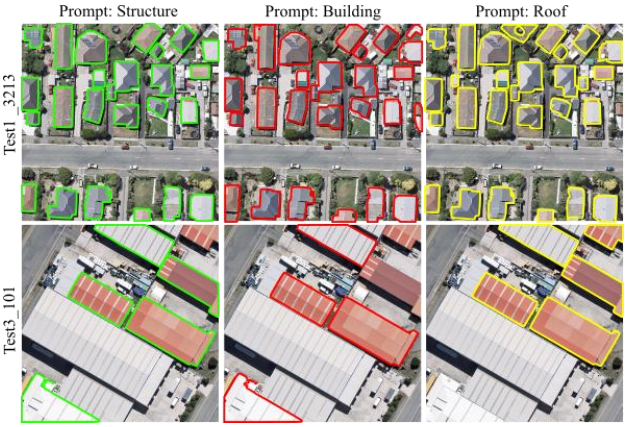

Figure 8. Building extraction results for different prompts with the Grounded-SAM.

residential, bare land, sparse residential, grassland, meadow, forest, and park. Prompt 2 contains building, roof, house, square, bare land, grassland, meadow, and forest. Prompt 3 combined classes from the first two prompts, building, roof, parking, house, commercial, center, medium residential, industrial, dense residential, sparse residential, and park. These prompts were deliberately chosen to establish contrasting classes, thereby enhancing CLIP’s capacity to differentiate between them.

As illustrated in Figure 9, CLIP effectively identified the roof, building, and house classes while accurately discerning unrelated

classes such as park and forest. However, the 9 also reveals that CLIP responds differently to various roof types. Specifically, for larger roofs and those with diverse styles as shown in the last two rows of Figure 9, CLIP assigns higher scores to the roof class. In a residential setting (first and second rows of 9), building and roof classes have similar probability scores across all of the prompts. The observed behavior of the CLIP model can be attributed to several factors. Roofs, especially those with differing shapes or colors, possess distinct visual features that CLIP can identify and prioritize. Larger or more varied roof types further emphasize these characteristics. This may confuse CLIP and cause it to weigh these classes differently based on the prominence and distinctiveness of the roof features in the image. Moreover, the specific composition of the prompts influences how CLIP interprets and prioritizes classes.

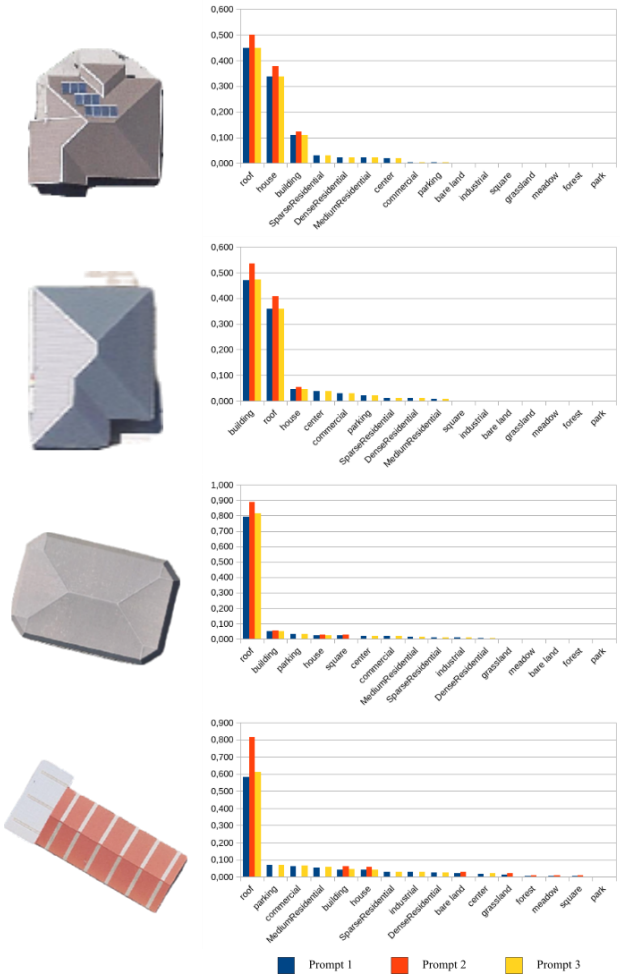

Figure 9. CLIP prompt selection results.

3.3 Time Complexity

Regarding the time complexity of the models, we performed a series of tests. For the Grounded-SAM, processing a single image requires about 0.3059 seconds of wall clock time on the GPU (0.3255 seconds of CPU time) for the DINO component, and 1.2960 seconds of wall clock time on the GPU (1.3629 seconds of CPU time) for the SAM component. The total GPU time measured for SAM is 1.258 seconds, while Grounding DINO’s total GPU time is about 0.256 seconds processing a single image. For the SAM+CLIP, running SAM alone takes approximately 4.724 seconds of CPU time and 3.752 seconds of GPU time to process a single image. The CLIP model, which was executed on the CPU only, requires about 1.2015 seconds of CPU time. However, it should be noted that these times refer to processing a single mask. If there are multiple masks, the total time will scale linearly with the number of masks. In terms of wall clock time, SAM completes the segmentation of the entire image in about 4.2818 seconds, while CLIP takes about 0.4772 seconds to process a single mask.The difference in processing times within the SAM model is attributable to the use of the predictor method in Grounded-SAM, while the automatic mask extractor method is employed in the SAM+CLIP pipeline. It has been observed that the Grounded-SAM is more efficient than SAM+CLIP in terms of time complexity.

4. Discussion

Grounded-SAM and SAM+CLIP demonstrated promising performance on zero-shot building segmentation from aerial imagery across complex scenarios. Nevertheless, there are certain limitations that can be broadly grouped into two categories. First, the object detection model in the case of Grounded-SAM or image captioning in the case of the SAM+CLIP pipeline can suffer from domain biases, leading to errors. Secondly, the SAM may produce less precise boundaries when dealing with objects of very irregular shapes or when presented with significant color and texture variation which leads to either oversegmentation or under-segmentation.

Although the Grounding DINO backbone was not explicitly trained on remote sensing imagery, it still achieved better building segmentation performance than the CLIP-RSICD model. As illustrated in Figure 10a-1, Grounding DINO successfully detected one large structure and two smaller structures. However, the SAM failed to segment the large structure, though it did accurately segment the smaller ones (Figure 10a-2). A similar pattern appears in Figure 10b-1, where Grounding DINO detected both large structures, but the SAM only segmented the northernmost structure (Figure 10b-2), despite both roofs having similar characteristics.

Regarding the SAM+CLIP, it notably excelled in areas where the Grounded-SAM struggled with SAM’s segmentation (Figures 10a-4 and 10b-4). The SAM+CLIP successfully segmented the large buildings (Figures 10a-3 and 10b-3) that were not detected by Grounded-SAM. However, CLIP either failed to label the smaller structures (Figure 10a-4) or mislabelled a parking lot as a building (Figure 10b-4), indicating that it is generally more effective with large-scale objects. It is worth noting that CLIP’s primary purpose is image captioning, but it can reliably caption segments extracted from images with reasonable accuracy.

Our findings indicate that prompt selection has a limited impact on the Grounded-SAM. In contrast, within the SAM+CLIP, the choice of prompts significantly affects the performance of the CLIP model. Additionally, it has been observed that the Grounded-SAM is more efficient than SAM+CLIP in terms of time complexity.

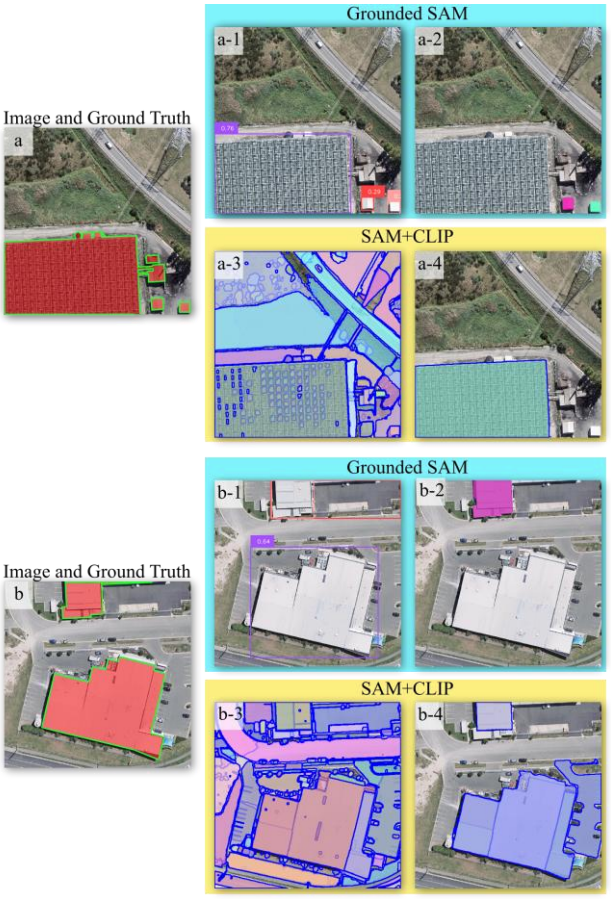

Figure 10. SAM image segmentation performance.

5. Conclusion

In this study, we employed two foundational models, Grounded-SAM and SAM+CLIP, to perform zero-shot building segmentation on WHU aerial imagery dataset. Despite employing zero-shot methods, the pipelines achieved promising segmentation accuracy. Grounded-SAM obtained an F1-score of 0.83 and an IoU of 0.71, outperforming SAM+CLIP, which achieved an F1-score of 0.65 and an IoU of 0.49. Notably, Grounded-SAM demonstrated enhanced boundary precision for partially occluded or complex roofs, whereas SAM+CLIP excelled in segmenting and classifying large-scale structures.

Both of the pipelines have certain limitations. The SAM can produce inaccurate boundaries for irregularly shaped or visually ambiguous targets in both pipelines. Another limitation is that the CLIP model showed occasional mislabelling of smaller buildings or non-building features, suggesting that domain biases. Considering the success of VFMs in zero-shot building segmentation in this study, future research efforts to fine-tune these models or apply few-shot approaches will pave the way for their more widespread use in the field of remote sensing.

References

Arutiunian, A., Vidhani, D., Venkatesh, G., Bhaskar, M., Ghosh, R., Pal, S., 2021. Clip-rsicd. [GitHub Repository]. https://github.com/arampacha/CLIP-rsicd.

Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.-H., Khan, F.S., 2025. Foundational Models Defining a New Era in Vision: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S. v., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., ..., Liang, P., 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

Chang, J., Cen, Y., Cen, G., 2024. Asymmetric Network Combining CNN and Transformer for Building Extraction from Remote Sensing Images. *Sensors*, 24(19), 6198.

Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z., 2024. RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-17.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 119, PMLR, 1597–1607.

Ding, L., Zhu, K., Peng, D., Tang, H., Yang, K., Bruzzone, L., 2024. Adapting Segment Anything Model for Change Detection in HR Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–11.

Erdem, F., Avdan, U., 2020. Comparison of Different U-Net Models for Building Extraction from High-Resolution Aerial Imagery. *International Journal of Environment and Geoinformatics*, 7(3), 221–227.

Hajjar, S.E., Kassem, H., Abdallah, F., Omrani, H., 2024. Enhancing building segmentation by deep multiview classification for advancing sustainable urban development. *Journal of Building Engineering*, 83, 108421.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

Ji, S., Wei, S., Lu, M., 2019. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1), 574–586.

Ji, W., Li, J., Bi, Q., Liu, T., Li, W., Cheng, L., 2024. Segment Anything Is Not Always Perfect: An Investigation of SAM on Different Real-world Applications. *Machine Intelligence Research*, 21(4), 617–630.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollar, P., Girshick, R., 2023. Segment Anything.′ *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Paris, France, 3992–4003.

Li, M., Wu, P., Wang, B., Park, H., Hui, Y., Yanlan, W., 2021. A Deep Learning Method of Water Body Extraction From High Resolution Remote Sensing Images With Multisensors. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 3120–3132.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J., 2024a. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–16.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H. et al., 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *European Conference on Computer Vision*, Springer, 38–55.

Lu, X., Wang, B., Zheng, X., Li, X., 2018. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4), 2183–2195.

Luo, M., Zhang, T., Wei, S., Ji, S., 2024. SAM-RSIS: Progressively Adapting SAM With Box Prompting to Remote Sensing Image Instance Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-14.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning Transferable Visual Models From Natural Language Supervision. *International conference on machine learning*, PmLR, 8748–8763.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F. et al., 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Wang, L., Fang, S., Meng, X., Li, R., 2022. Building Extraction With Vision Transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–11.

Yan, Z., Li, J., Li, X., Zhou, R., Zhang, W., Feng, Y., Diao, W., Fu, K., Sun, X., 2023. RingMo-SAM: A Foundation Model for Segment Anything in Multimodal Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 116.

Yildirim, F.S., Karsli, F., Bahadir, M., Yildirim, M., 2024. FwSVM-Net: A novel deep learning-based automatic building extraction from aerial images. *Journal of Building Engineering*, 96, 110473.

Yu, Z., Li, T., Zhu, Y., Pan, R., 2024. Exploring Foundation Models in Remote Sensing Image Change Detection: A Comprehensive Survey. *arXiv preprint arXiv:2410.07824*. https://arxiv.org/abs/2410.07824.

Yuan, H., Li, X., Zhou, C., Li, Y., Chen, K., Loy, C.C., 2025. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. *Computer Vision - ECCV 2024*, Springer Nature Switzerland, 419–437.