Journal of Image Processing Theory and Applications (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2026.090104
ISSN 2560-6239 Vol. 9 Num. 1

Stratified Evaluation of SAM 2 for Zero-Shot Building Segmentation in Aerial Imagery

Bingning Xiong[1, a,*], Mingyu Ou[1, b]

[1]Beijing Technology and Business University, Beijing, China
[a]2642443369@qq.com, [b]179437891@qq.com
**Corresponding author*

Keywords: SAM 2, Building Segmentation, Zero-Shot Learning, Aerial Imagery

Abstract: Building footprint extraction from remote sensing imagery underpins urban planning, population estimation, and disaster damage assessment. Deep learning methods have achieved high accuracy for this task, but their dependence on large-scale pixel-level annotations creates a severe bottleneck: annotating a city-scale dataset demands hundreds of hours of manual labor, limiting rapid deployment to new regions. The Segment Anything Model 2 (SAM 2), a foundation model with zero-shot segmentation capability, offers a potential solution by eliminating the need for task-specific annotations entirely. Yet SAM 2 was trained exclusively on natural scene images and videos, raising a critical question: can it generalize to the fundamentally different visual characteristics of aerial remote sensing imagery? This paper presents the first systematic zero-shot evaluation of SAM 2 for aerial building segmentation. We conduct three groups of experiments: (1) benchmarking four SAM 2 model variants to identify the optimal accuracy-efficiency trade-off; (2) stratified evaluation across dense, sparse, large-scale, and small-scale building morphologies to reveal which architectural characteristics challenge SAM 2 most; and (3) comparison of single-point, multi-point, and bounding-box prompting strategies to derive practical guidelines. Results demonstrate that SAM 2 Base+ achieves an IoU of 0.7738 without any training data, while oracle bounding-box prompting reaches 0.8755. SAM 2 excels on dense and large-scale buildings but struggles with sparse scenes. These findings establish SAM 2 as a viable tool for rapid building mapping while highlighting where domain adaptation remains necessary.

1. Introduction

Building footprint extraction from remote sensing imagery is a foundational task in urban computing, supporting urban planning [1], population estimation [2], disaster damage assessment [3], and land-use mapping [4]. Over the past decade, deep learning methods—most notably U-Net [5], DeepLabV3+ [6], and HRNet [7]—have pushed segmentation accuracy to impressive levels on benchmark datasets. However, these supervised approaches carry a fundamental limitation: they rely on large quantities of pixel-level annotated training data. In the remote sensing domain, producing such annotations is extraordinarily costly. A single 512×512 tile can take 5–10 minutes for a professional annotator, and city-scale datasets can demand hundreds of hours of manual labor

[1]. This annotation bottleneck severely constrains the rapid deployment of building extraction models to new geographic regions.

A promising pathway to circumvent this bottleneck lies in foundation models with zero-shot capability. The Segment Anything Model 2 (SAM 2) [8], released by Meta AI in 2024, is trained on 51,000 natural-scene videos and over 600,000 masklet annotations. It can generate object masks from minimal prompts (points, boxes, or masks) without any task-specific training. If SAM 2 can directly segment buildings from aerial images, the entire annotation stage could be bypassed, enabling rapid building footprint mapping for any city worldwide. This prospect is particularly compelling given that SAM 2 achieves inference speeds 6 times faster than its predecessor SAM [9] while maintaining state-of-the-art zero-shot generalization.

Yet a critical barrier stands in the way: domain mismatch. SAM 2 was trained exclusively on natural-scene images and videos—indoor scenes, outdoor environments, portraits, and everyday objects. It has never been exposed to the distinct visual characteristics of aerial remote sensing imagery: overhead perspective, geometric building footprints, shadow patterns, and spectral properties of man-made structures. Whether a model trained on ground-level photography can generalize to airborne imagery is far from guaranteed. Prior studies have evaluated SAM (v1) on remote sensing tasks [10-11], but SAM 2 differs substantially in architecture (Hiera hierarchical encoder [12], memory attention mechanism, redesigned mask decoder) and training data. The applicability of SAM 2 to aerial building extraction—arguably one of the most economically significant tasks in geospatial analysis—remains an open question.

Moreover, existing evaluations of SAM/SAM 2 on remote sensing imagery rarely control for building morphology. Building footprints are not homogeneous: dense urban clusters, sparse suburban developments, large industrial complexes, and small residential structures each present fundamentally different visual patterns. Adjacent buildings in dense areas merge together; isolated buildings in sparse regions are easily missed; expansive structures fragment into pieces; compact units vanish against textureless backgrounds. Without a stratified evaluation across these morphological categories, it is impossible to determine where SAM 2 succeeds and where it fails, limiting practical deployment guidance.

This paper addresses the central research question: Can SAM 2 perform zero-shot building segmentation from aerial imagery, and under what conditions does it succeed or fail? We present a systematic benchmark with three components: (1) We evaluate all four SAM 2 model sizes (Tiny, Small, Base+, Large) to establish the accuracy-efficiency frontier and identify the best model for practical deployment. (2) We conduct a stratified analysis across dense, sparse, large-scale, and small-scale building types to pinpoint which morphological characteristics challenge SAM 2 and why. (3) We compare three oracle prompting strategies (single-point, multi-point, bounding-box) to determine how much performance can be gained when approximate building locations are known. Our findings establish SAM 2 as a viable tool for rapid building mapping while revealing critical scenarios where domain adaptation remains necessary.

2. Related Work.

2.1. Building Extraction from Remote Sensing Imagery

Building extraction has evolved through several paradigms. Early methods relied on morphological operations and edge detection combined with handcrafted rules [13]. The introduction of fully convolutional networks (FCNs) [14] enabled end-to-end semantic segmentation, with U-Net [5] emerging as a dominant architecture due to its skip connections that preserve fine spatial details. Subsequent developments include DeepLabV3+ [6] with atrous separable convolutions for multi-scale feature extraction, HRNet [7] for high-resolution

representation learning, and transformer-based methods that capture long-range dependencies [15]. The WHU Building Dataset [1], derived from aerial imagery of Christchurch, New Zealand at 0.075m resolution, has become one of the most widely used benchmarks for building extraction evaluation.

2.2. The Segment Anything Model Family

SAM [9] introduced a flexible segmentation framework comprising a Vision Transformer (ViT) image encoder, a prompt encoder supporting points, boxes, and masks, and a lightweight mask decoder. Trained on the SA-1B dataset of 11 million images, SAM demonstrated strong zero-shot transfer to diverse segmentation tasks. SAM 2 [8] builds upon this foundation with several architectural innovations: a hierarchical vision encoder [12] replaces the original ViT backbone, a memory attention mechanism enables video consistency, and an improved mask decoder produces higher-quality boundaries. SAM 2 is available in four model sizes: Tiny (38.9M parameters), Small (46M), Base+ (80.8M), and Large (224.4M), enabling deployment across diverse computational constraints. Figure 1 illustrates the SAM 2 architecture and our experimental pipeline.

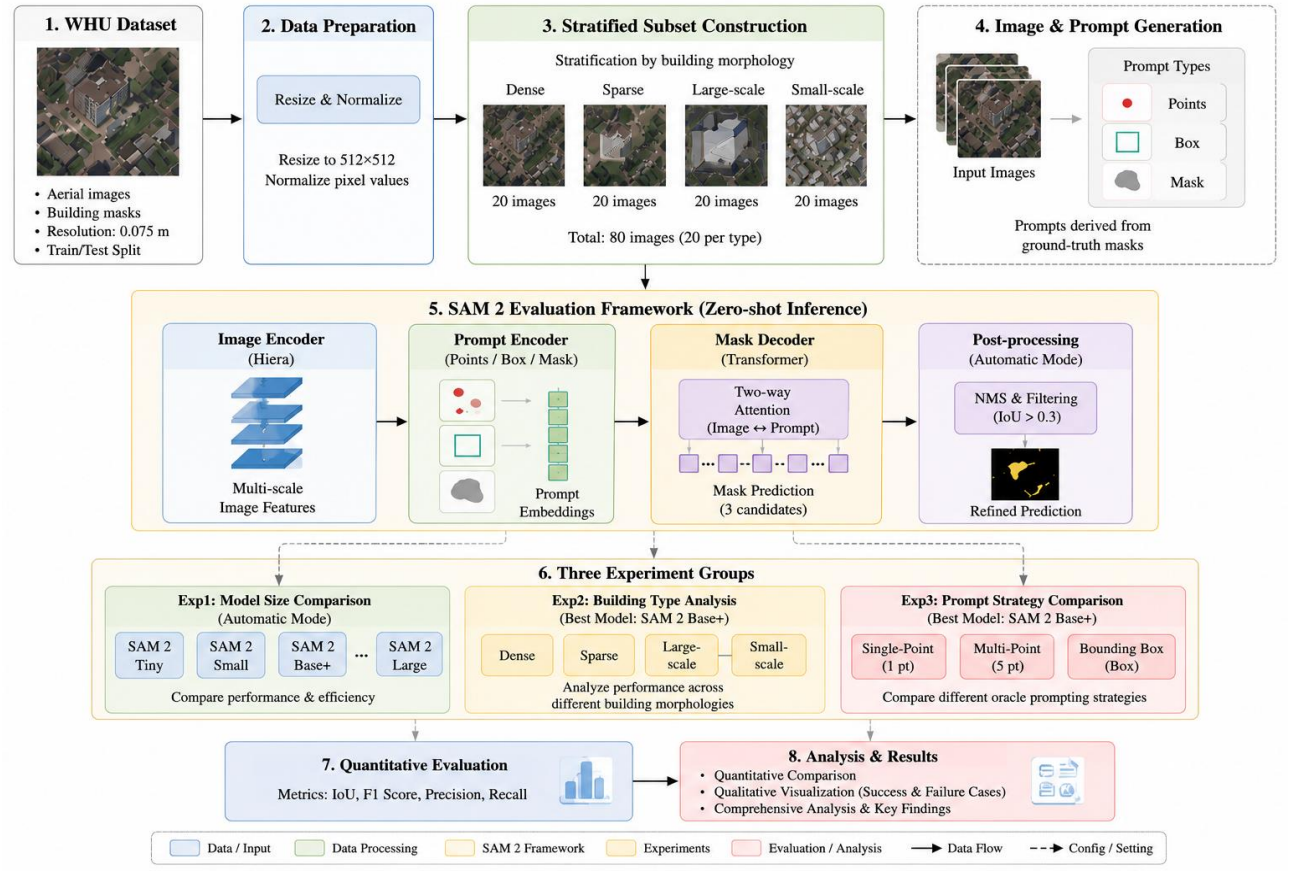

Figure 1. Overview of the proposed framework. Top: Data preparation and stratified subset construction (80 images across 4 building types). Middle: SAM 2 zero-shot inference pipeline with image encoder, prompt encoder, mask decoder, and post-processing (NMS and filtering with IoU > 0.3). Bottom: Three experiment groups.

2.3. SAM for Remote Sensing Applications

The adaptation of SAM to remote sensing has garnered significant research interest. Osco et al.

[10] evaluated SAM across multiple remote sensing tasks including building extraction, finding that performance varies substantially with prompting strategy and scene complexity. Ren et al. [11] conducted the first comprehensive evaluation of SAM on overhead imagery, identifying key challenges including the domain gap between natural and aerial images. Chen et al. [16] proposed SAM-Adapter, which introduces lightweight adapter modules to fine-tune SAM for underperforming scenes. Zhang et al. [17] combined SAM with LiDAR and VHR imagery for residential building extraction. SAM has also been applied to agricultural field delineation from Sentinel-2 imagery [18] and forest canopy segmentation from LiDAR [19], and automatic unsupervised LiDAR segmentation [23]. However, these works primarily target SAM v1, and a systematic evaluation of SAM 2 on aerial building segmentation across diverse morphologies remains lacking.

3. Methodology.

3.1. SAM 2 Architecture and Experimental Pipeline

Figure 1 presents the SAM 2 architecture alongside our experimental pipeline. SAM 2 [8] comprises three core components: (1) an image encoder [12] that processes input images into hierarchical feature representations; (2) a prompt encoder that converts user inputs (points, boxes, or masks) into embedding vectors; and (3) a mask decoder that fuses image and prompt embeddings to predict segmentation masks. For single-image segmentation, the memory bank is empty. In automatic mask generation mode, SAM 2 samples point prompts on a 32×32 grid, generates masks, and applies non-maximum suppression.

3.2. Experiment Design

3.2.1. Experiment 1: Model Size

We evaluate all four SAM 2.1 checkpoints released by Meta AI [20]: sam2.1-hiera-tiny (38.9M parameters), sam2.1-hiera-small (46M), sam2.1-hiera-base-plus (80.8M), and sam2.1-hiera-large (224.4M). All models operate in automatic mask generation mode. Following Ren et al. [11], we apply an oracle filtering strategy: we retain masks with IoU > 0.3 with any ground-truth building region, then merge the retained masks to form the final prediction.

3.2.2. Experiment 2: Building Morphology Analysis

We define four building morphology types: Dense—closely packed structures with minimal separation; Sparse—isolated structures with significant inter-building spacing; Large-scale—industrial complexes or large residential blocks; and Small-scale—compact residential units. From the WHU test set, we manually select 20 representative images for each type (80 images total). We use the best-performing model from Experiment 1.

3.2.3. Experiment 3: Prompt Strategy Comparison

We implement three oracle prompting strategies derived from ground-truth annotations: Single-point prompting—a single point at the centroid; Multi-point prompting—five randomly sampled interior points; and Bounding-box prompting—the tight axis-aligned bounding box. We use the Base+ model across all strategies.

3.3. Evaluation Metrics

We evaluate segmentation quality using four standard metrics. Let P denote the predicted building region and G the ground truth. Intersection over Union (IoU) = $|P \cap G| / |P \cup G|$. Precision = $|P \cap G| / |P|$. Recall = $|P \cap G| / |G|$. F1 Score = $2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$.

4. Experiments and Results

4.1. Dataset and Implementation Details

The WHU Building Dataset [1] consists of aerial imagery of Christchurch, New Zealand, captured at 0.075m ground sample distance. The dataset provides 8,188 training images and 2,416 test images, each resized to 512 ×512 pixels at 0.3m resampled resolution. We construct a stratified subset of 80 images (20 per building type). All experiments are conducted on Google Colab Pro [21] with an NVIDIA Tesla T4 GPU (16GB VRAM). We use the official SAM 2.1 checkpoints from Meta AI [20] with default parameters: points_per_side=32, pred_iou_thresh=0.86.

4.2. Experiment 1: Effect of Model Size

Table 1 presents the performance of all four SAM 2 variants. The Base+ model achieves the highest IoU of 0.7738 and F1 of 0.8630, outperforming the Tiny model by 17.3 percentage points. Surprisingly, the Large model achieves a lower IoU (0.7386) than Base+, suggesting potential overfitting to the natural image training distribution. Inference time increases modestly with model size, from 4.78s for Tiny to 5.49s for Large. GPU memory remains stable (3.3–4.0 GB). The Base+ model offers the best accuracy-efficiency trade-off.

Table 1. Performance comparison of four SAM 2 model variants

Model	Params	IoU↑	F1↑	Prec.	Rec.	Time(ms)	GPU(MB)
Tiny	38.9M	0.6004	0.7240	0.8942	0.6518	4781.5	3636
Small	46M	0.6881	0.7999	0.8934	0.7478	4827.9	3366
Base+	80.8M	0.7738	0.8630	0.9168	0.8387	5108.2	3593
Large	224.4M	0.7386	0.8335	0.9004	0.8032	5489.8	4029

4.3. Experiment 2: Building Morphology Analysis

Table 2. Performance of SAM 2 (Base+) across building morphology types

Building Type	Images	IoU↑	F1↑	Prec.	Rec.
Dense	20	0.8150	0.8979	0.8934	0.9030
Sparse	20	0.7181	0.8068	0.9074	0.7922
Large-scale	20	0.7904	0.8779	0.9741	0.8086
Small-scale	20	0.7718	0.8696	0.8924	0.8510

Table 2 reports SAM 2 (Base+) performance across four building types. Dense buildings achieve the highest IoU (0.8150), benefiting from strong visual contrast between adjacent structures. Large-scale buildings follow closely (IoU 0.7904, precision 0.9741). Small-scale buildings achieve IoU 0.7718. Sparse buildings exhibit the lowest performance (IoU 0.7181) due to lower recall (0.7922)—isolated small buildings in large backgrounds are missed by the automatic point grid.

Figure 2 visualizes representative success cases for each building type. For large-scale buildings (Row 3, IoU 0.968), SAM 2 accurately captures industrial warehouse structures with minimal

boundary error. Dense buildings (Row 1, IoU 0.853) show successful delineation despite close adjacency. Small-scale buildings (Row 4, IoU 0.930) demonstrate precise segmentation of isolated compact structures. Sparse buildings (Row 2, IoU 0.839) illustrate effective detection of suburban residential buildings.

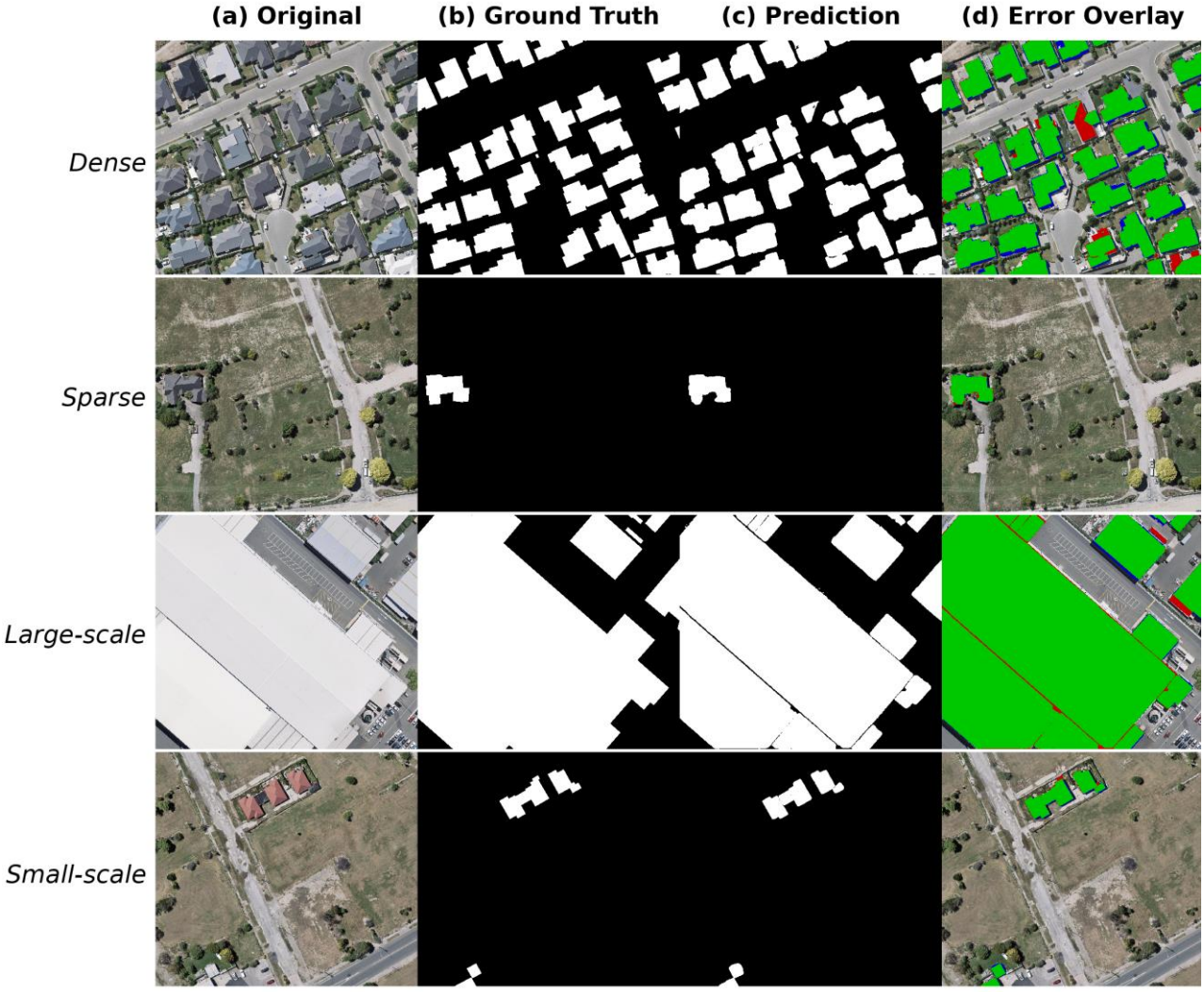

Figure 2. Success case visualization across building morphology types. Each row shows: Dense (IoU 0.853), Sparse (IoU 0.839), Large-scale (IoU 0.968), Small-scale (IoU 0.930). Columns: input image, ground truth, prediction, error overlay (G=TP, R=FN, B=FP).

Figure 3 presents corresponding failure cases. For dense buildings (Row 1, IoU 0.752), SAM 2 merges adjacent structures into larger blobs. Sparse buildings (Row 2, IoU 0.622) suffer from missed detections. Large-scale buildings (Row 3, IoU 0.563) exhibit fragmentation. Small-scale buildings (Row 4, IoU 0.128) show the most severe failure with complete omission.

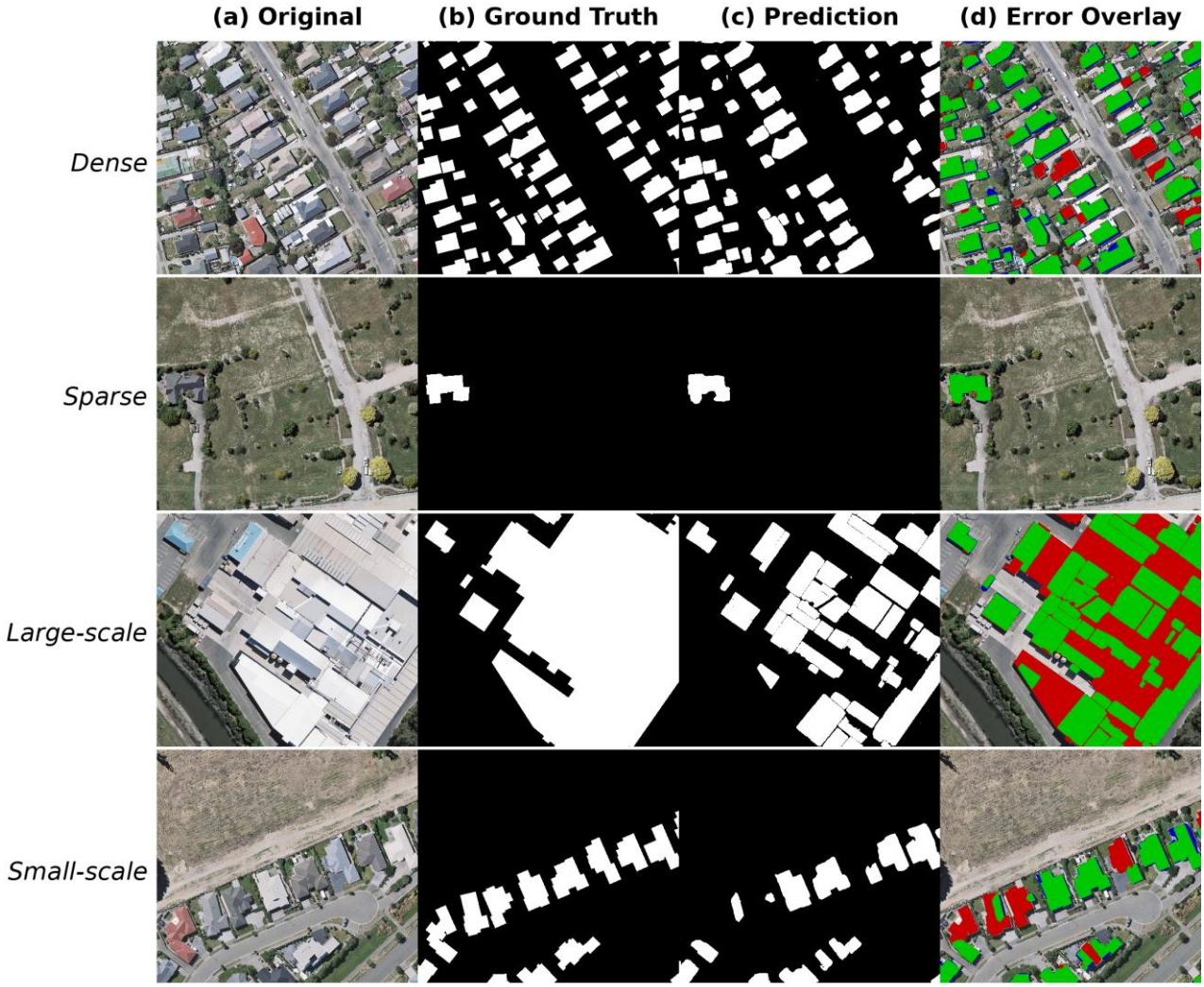

Figure 3. Failure case visualization across building morphology types. Each row shows: Dense (IoU 0.752), Sparse (IoU 0.622), Large-scale (IoU 0.563), Small-scale (IoU 0.128). Columns: input image, ground truth, prediction, error overlay.

4.4. Experiment 3: Effect of Prompt Strategy

Table 3 compares the three oracle prompting strategies. Bounding-box prompting achieves the highest IoU of 0.8755 and F1 (0.9330), significantly outperforming point-based strategies. This advantage stems from the explicit spatial constraint provided by bounding boxes. Multi-point prompting achieves IoU 0.8179, a 16.4 percentage point improvement over single-point (0.6537). The additional points provide richer spatial information about building interior regions [22].

Table 3. Comparison of three oracle prompting strategies using SAM 2 (Base+).

Prompt Strategy	IoU↑	F1↑	Prec.	Rec.
Single-point	0.6537	0.7639	0.9211	0.7063
Multi-point (5)	0.8179	0.8962	0.9128	0.8918
Bounding-box	0.8755	0.9330	0.9278	0.9408

Figure 4 presents a visual comparison of the three prompting strategies across four representative scenes. Single-point prompting produces severe under-segmentation for complex scenes (IoU as

low as 0.201), while multi-point prompting substantially improves coverage. Bounding-box prompting produces the most complete and accurate masks across all scenes, as the explicit spatial boundary prevents both over-segmentation and under-segmentation.

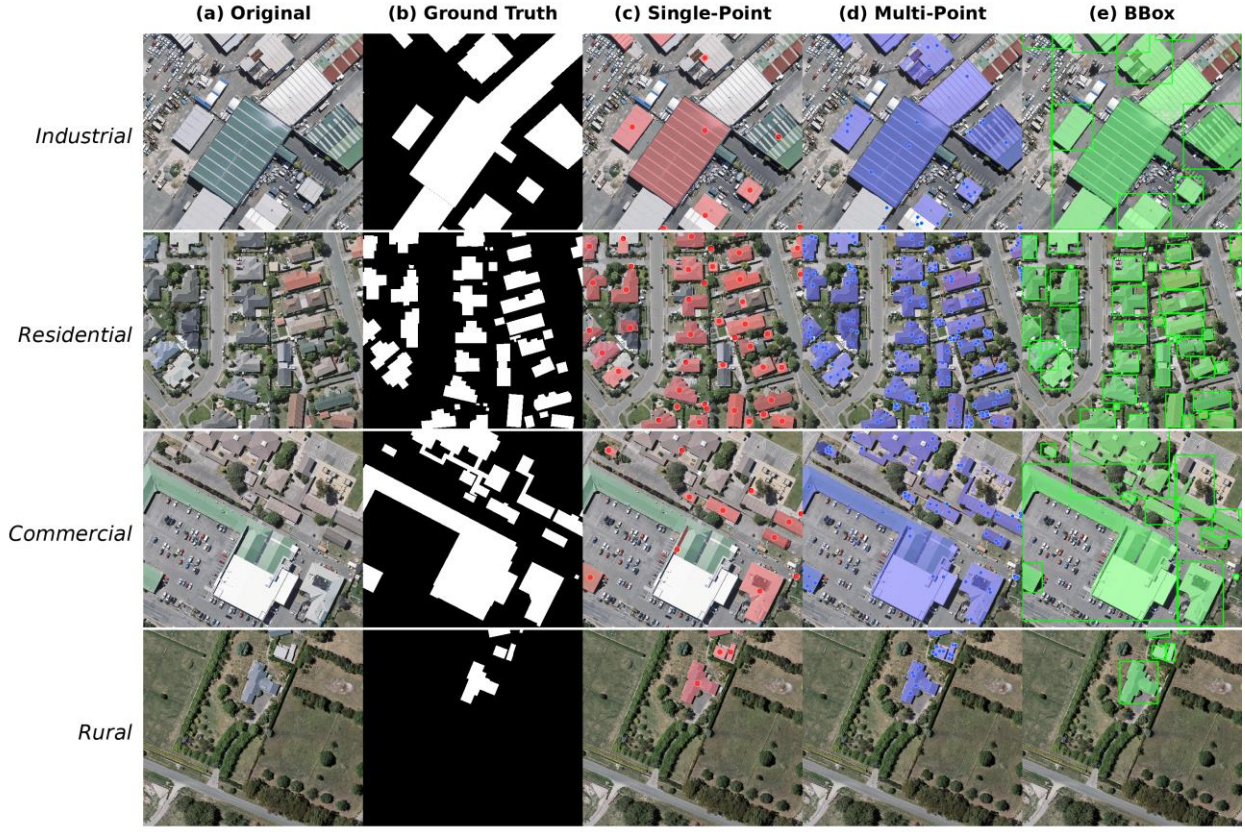

Figure 4. Visual comparison of three prompting strategies across four representative scenes. (a) Dense urban: single-point severely under-segments (IoU 0.201), box prompting achieves 0.858. (b) Suburban residential: all strategies perform well, box prompting reaches 0.900. (c) Large industrial: single-point fails (IoU 0.302), box prompting succeeds (0.879). (d) Sparse rural: single-point achieves 0.870, box prompting at 0.913.

Table 4 presents the cross-analysis. Bounding-box prompting consistently achieves the highest IoU across all types, with particularly strong performance on large-scale buildings (IoU 0.8957). Multi-point prompting substantially outperforms single-point for large-scale buildings (0.7967 vs. 0.4247), where multiple interior points help capture the full spatial extent.

Table 4. Cross-analysis: IoU by building type and prompting strategy

Building Type	Single-point	Multi-point	Bounding-box
Dense	0.7646	0.8380	0.8656
Sparse	0.7167	0.8314	0.8756
Large-scale	0.4247	0.7967	0.8957
Small-scale	0.7086	0.8053	0.8653

4.5. Summary of Key Findings

Our experiments reveal three key findings: (1) The Base+ model achieves the best accuracy-efficiency trade-off; the Large model surprisingly underperforms, suggesting diminishing returns

from increased capacity on aerial imagery. (2) Dense and large-scale buildings are easiest to segment (IoU 0.79–0.82), while sparse scenes are most challenging (IoU 0.72) due to detection difficulties. (3) Bounding-box prompting outperforms point-based methods by a substantial margin (IoU 0.8755 vs. 0.8179), establishing it as the preferred strategy when approximate building locations are available.

5. Discussion

5.1. Strengths of SAM 2 for Building Segmentation

Our evaluation demonstrates that SAM 2 possesses several notable strengths for aerial building segmentation. First, its zero-shot capability eliminates the need for task-specific training data—the Base+ model achieves an IoU of 0.7738 without any fine-tuning on remote sensing imagery, competitive with early supervised methods. Second, with bounding-box prompting, SAM 2 achieves an IoU of 0.8755, approaching state-of-the-art supervised performance. Third, the hierarchical encoder effectively captures multi-scale features, enabling consistent performance across building sizes from small residential units to large industrial complexes.

5.2. Limitations and Failure Analysis

Despite its strengths, SAM 2 exhibits several limitations. The domain gap between natural images (training data) and aerial imagery remains a fundamental challenge. SAM 2 was trained on natural images and videos; the overhead perspective, building textures, shadow patterns, and spectral characteristics differ substantially. The class-agnostic nature of automatic mode means SAM 2 cannot distinguish buildings from non-building structures without additional filtering. Dense urban areas cause mask merging and boundary ambiguity, while small isolated buildings are frequently missed.

5.3. Practical Recommendations

Based on our findings, we offer three recommendations: (1) Model selection: The Base+ model provides the optimal balance; the Large model offers no improvement while requiring 49% more parameters. (2) Prompt strategy: When building locations are approximately known, bounding-box prompting should be prioritized. (3) Deployment scenario: SAM 2 is best suited as a semi-automatic annotation tool or in a detection-segmentation pipeline, rather than as a fully automatic system.

6. Conclusion

This paper presented a systematic evaluation of SAM 2 for zero-shot building segmentation in high-resolution aerial imagery. Through three groups of experiments, we established that the Base+ model achieves the best accuracy-efficiency trade-off (IoU 0.7738), building morphology significantly affects performance with dense and large-scale buildings being easiest (IoU 0.79–0.82), and bounding-box prompting substantially outperforms point-based strategies (IoU 0.8755).

Our findings provide practical guidelines for leveraging foundation models for remote sensing applications. Future directions include domain adaptation through adapter modules, integration with building detection systems, and extension to additional datasets.

References

[1] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multi-source building extraction from an open aerial and satellite imagery dataset," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 1, pp. 574-586, 2018.
[2] T. Esch et al., "Delineation of urban footprints from TerraSAR-X data," IEEE Trans. Geosci. Remote Sens., vol. 48, no. 2, pp. 905-916, 2009.
[3] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review," IEEE Geosci. Remote Sens. Mag., vol. 5, no. 4, pp. 8-36, 2017.
[4] G. S. Xia et al., "AID: A benchmark for aerial scene classification," IEEE Trans. Geosci. Remote Sens., vol. 55, no. 7, pp. 3965-3981, 2017.
[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, 2015, pp. 234-241.
[6] L. Chen et al., "Encoder-decoder with atrous separable convolution," in Proc. ECCV, 2018, pp. 801-818.
[7] J. Wang et al., "Deep high-resolution representation learning," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3349-3364, 2020.
[8] N. Ravi et al., "SAM 2: Segment anything in images and videos," arXiv preprint arXiv:2408.00714, 2024.
[9] A. Kirillov et al., "Segment anything," in Proc. ICCV, 2023, pp. 4015-4026.
[10] L. P. Osco et al., "The segment anything model (SAM) for remote sensing applications," Int. J. Appl. Earth Obs. Geoinf., vol. 124, p. 103,540, 2023.
[11] S. Ren et al., "Segment anything, from space?" in Proc. WACV, 2024, pp. 3107-3117.
[12] C. Ryali et al., "Hiera: A hierarchical vision transformer," in Proc. ICML, 2023, pp. 29,461-29,472.
[13] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," Int. J. Comput. Vis., vol. 1, no. 4, pp. 321-331, 1988.
[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. CVPR, 2015, pp. 3431-3440.
[15] A. Dosovitskiy et al., "An image is worth 16x16 words," in Proc. ICLR, 2021.
[16] T. Chen et al., "SAM-Adapter: Adapting segment anything," in Proc. ICCV Workshop, 2023, pp. 3367-3377.
[17] Y. Zhang et al., "Segment anything model-based building footprint extraction," Remote Sens., vol. 16, no. 14, p. 2661, 2024.
[18] F. Panangian and K. Bittner, "Segment anything for satellite imagery," arXiv preprint arXiv:2506.16318, 2025.
[19] M. Illarionova et al., "Segmenting forest canopies with SAM and LiDAR," Int. J. Appl. Earth Obs. Geoinf., 2024.
[20] Meta AI, "SAM 2.1 checkpoints," 2024. [Online]. Available: https://github.com/facebookresearch/segment-anything-2
[21] Google Research, "Google Colaboratory," 2024. [Online]. Available: https://colab.research.google.com
[22] Z. Wang et al., "Deep learning-based interactive segmentation in remote sensing," arXiv preprint arXiv:2308.13174, 2023.
[23] A. Yarroudh, "LiDAR automatic unsupervised segmentation using SAM," GitHub repository, 2023.