ECAI 2025
I. Lynce et al. (Eds.)

© 2025 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA250844

GeoSAM: Fine-Tuning SAM with Multi-Modal Prompts for Mobility Infrastructure Segmentation

Rafi Ibn Sultan[a], Chengyin Li[b], Hui Zhu[a], Prashant Khanduri[a], Marco Brocanelli[c] and Dongxiao Zhu[a,*]

[a]Department of Computer Science, Wayne State University, Detroit, MI, USA 48202
[b]Department of Radiation Oncology, Henry Ford Health, Detroit, MI, USA 48202
[c]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, Ohio, USA 43210

Abstract. In geographical image segmentation, performance is often constrained by the limited availability of training data and a lack of generalizability, particularly for segmenting mobility infrastructure such as roads, sidewalks, and crosswalks. Vision foundation models like the Segment Anything Model (SAM), pre-trained on millions of natural images, have demonstrated impressive zero-shot segmentation performance, providing a potential solution. However, SAM struggles with geographical images, such as aerial and satellite imagery, due to its training being confined to natural images and the narrow features and textures of these objects blending into their surroundings. To address these challenges, we propose Geographical SAM (GeoSAM), a SAM-based framework that fine-tunes SAM using automatically generated multi-modal prompts. Specifically, GeoSAM integrates point prompts from a pre-trained task-specific model as primary visual guidance, and text prompts generated by a large language model as secondary semantic guidance, enabling the model to better capture both spatial structure and contextual meaning. GeoSAM outperforms existing approaches for mobility infrastructure segmentation in both familiar and completely unseen regions by at least 5% in mIoU, representing a significant leap in leveraging foundation models to segment mobility infrastructure, including both road and pedestrian infrastructure in geographical images. The source code is publicly available.

1 Introduction

While a substantial amount of research [7, 43, 17, 28, 13] has focused on road infrastructure segmentation from geographical and remote sensing imagery like aerial and satellite images, pedestrian infrastructure, such as sidewalks or crosswalks, has received comparatively little attention, despite its importance in daily life. Historically, research efforts have predominantly focused on assisting drivers in navigation rather than pedestrians [18]. Existing accessibility studies often use simplified road data, but accurate segmentation of pedestrian infrastructure can better reveal accessible routes and destinations, especially for people with disabilities.

Rooted in historical context, mobility infrastructure segmentation has predominantly relied on traditional models, including Convolutional Neural Networks (CNNs) [31, 45, 20, 27] and Vision Transformer (ViT) models [14, 11]. These models typically require large collections of human-labeled data for task-specific training [18, 3], something that is oftentimes a luxury for these tasks, and are often too sensitive to changes in data. However, the scarcity of high-quality labeled datasets remains a major challenge, especially in the context of mobility infrastructure, limiting scalability and adaptability to diverse tasks.

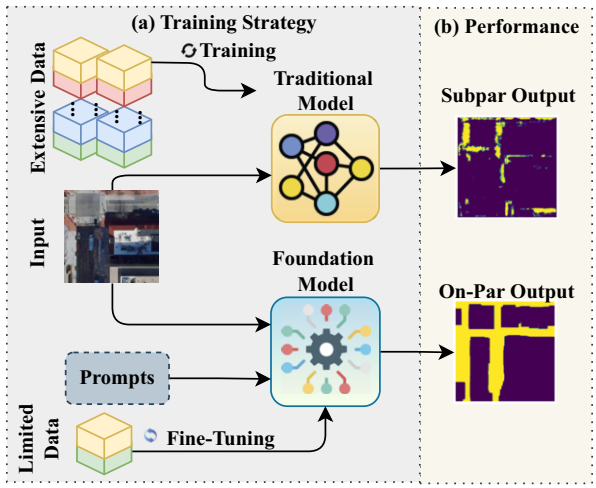

Figure 1. Mobility infrastructure segmentation: (a) Traditional models need large task-specific datasets, (b) struggle with narrow, texture-similar objects, yielding subpar results. Fine-tuning a promptable foundation model with limited data and prompts achieves on-par performance.

Traditional models, when trained on limited and homogeneous datasets(Figure 1a), often fail to distinguish fine-grained classes such as sidewalks and roads, which exhibit subtle visual differences like thin boundaries, similar textures, and frequent occlusions (Figure 1b). Moreover, their learned representations are typically domain-specific, resulting in poor generalization when deployed in unseen regions or datasets with different visual characteristics. Even minor shifts in data distribution, such as moving from one geographic region to another, often lead to significant performance degradation. In contrast, vision foundation models, pre-trained task-agnostically on large-scale and diverse image distributions [21, 26], offer a promising alternative with superior generalization ability across varying domains. These models adapt to new downstream tasks without re-training, relying on user-provided prompts for contextual guidance. In this work, we leverage the Segment Anything Model (SAM) [21], a promptable vision foundation model, to overcome the limitations of traditional approaches and enable effective

* Corresponding Author. Email: dzhu@wayne.edu.

segmentation of mobility infrastructure, even with limited labeled data and across geographically diverse regions.

However, unlike compact objects in natural images, where a single point (e.g., placed on a dog's body) often suffices for segmentation, spatially extensive structures like roads and sidewalks usually require multiple iterative prompts to capture their full extent. This process is often exhaustive and error-prone, and even with multiple prompts, zero-shot SAM struggles in remote sensing tasks due to its pre-training on natural images, which lack the large, texture-similar structures common in geographical data [21]. Nonetheless, SAM's general segmentation capability can be adapted to geographical imagery via fine-tuning on limited data (bottom of Figure 1), allowing it to learn domain-specific patterns and remain effective under regional distribution shifts. Capitalizing on this strength, we introduce Geographical SAM (GeoSAM), an end-to-end model tailored for segmenting mobility infrastructure through multi-class segmentation of road and pedestrian infrastructure.

To address these challenges, we propose Geo-Point Generation (GPG), an automated prompt generation technique that generates point prompts for geographical images from a domain-specific pretrained model for precise spatial guidance. It is complemented by text prompts for semantic clarity to resolve ambiguities inherent in point-based guidance. Point prompts focus the model on specific pixels, but a single pixel can often belong to multiple objects. Text prompts, containing semantic information about the class, clarify the object of interest and provide a broader understanding [37]. This complementary design ensures precise geometrical guidance from point prompts and broader contextual understanding from text prompts, enhancing segmentation accuracy.

These multi-modal prompts fine-tune SAM through its lightweight decoder. By integrating spatial precision with semantic context, we introduce **Geographical SAM (GeoSAM)**, an end-to-end SAM-based model fine-tuned for multi-class segmentation of roads and pedestrian infrastructure. GeoSAM outperforms traditional CNN-based approaches [18, 17, 43], not only improving segmentation accuracy but also demonstrating the potential of combining natural language and visual interaction within foundation models for geographical imagery. Our contributions are three-fold: (1) We pioneer the use of SAM for multi-class mobility infrastructure segmentation, integrating point and text prompts in geographical imagery. (2) We introduce fine-tuning and automated prompt generation techniques that inject domain knowledge from traditional models via multi-modal prompts. (3) We conduct extensive evaluations on datasets from two cities, demonstrating GeoSAM's strong performance and generalizability across diverse locations.

2 Related Work

2.1 Traditional Geographical Methods

Before the emergence of foundation models, traditional task-specific works, such as UNet-based approaches like [17, 28] and more advanced encoder-decoder-based works like [43, 7, 13] were developed to execute various geographical image segmentation tasks. Furthermore, CNN-based work such as [18] focuses more on pedestrian infrastructure segmentation in aerial images. Researchers have also explored machine learning techniques to enhance CNN-based segmentation for geographic objects [5, 1], along with transfer learning approaches that leverage pretrained models [41]. While these efforts improve remote sensing segmentation, they often rely on extensive supervision and retraining. Accuracy gains aside, they fall short in addressing the core challenge of generalizing to new locations.

2.2 Geographical Foundation Models

Task-agnostic vision foundation models address traditional segmentation limitations by using prompts to adapt to unseen classes across diverse tasks. While their use in geographical imagery, such as SAM, remains limited, some studies have begun exploring its potential. Works like [34, 23, 25] leverage SAM's zero-shot capabilities for tasks beyond segmentation, with [25] employing a hybrid zero-shot and one-shot learning approach for geographical imagery segmentation. However, these approaches are largely effective for objects with well-defined boundaries and distinguishable physical contexts, relying primarily on sensible prompts without requiring extensive domain-specific knowledge.

Most research focuses on manual human prompting during inference, though automated prompt generation has gained attention. Studies like [4, 35, 40] develop automated prompt-generation techniques requiring substantial training data, while others, such as [39, 25], use text queries in two-stage pipelines to generate bounding box prompts for SAM. Direct integration of natural language text prompts for improving SAM in geographical imagery remains unexplored. Methods like [22, 44] eliminate the need for prompts using additional networks, but require extensive training data, while [36] depends on auxiliary inputs like trajectory points, making it road-specific and less generalizable to other classes.

To address domain-specific challenges, some works have fine-tuned SAM using Parameter Efficient Fine-Tuning (PEFT) techniques [19]. In geographical imagery, studies like [38, 10, 4, 12] explore fine-tuning for diverse downstream tasks, yet no work, to our knowledge, focuses on fine-tuning SAM specifically for mobility tasks such as pedestrian infrastructure segmentation. This critical gap presents an opportunity for significant social impact in underperforming tasks.

2.3 Domain-Specific Geographical Foundation Models

Researchers have also explored training domain-specific foundation models on large-scale geographical imagery for targeted tasks. Similar to SAM, works like [3, 32] develop non-promotable foundation models using scaled ViT architectures, focusing on specific tasks without user interaction. In a related effort, [38] employs a SAM-like architecture trained on a massive remote sensing dataset. While many of these studies target road segmentation, they overlook the critical task of mobility infrastructure segmentation, such as sidewalks and crosswalks. Moreover, the lack of public source code makes it difficult to evaluate their effectiveness for pedestrian infrastructure.

3 Method

3.1 Problem Definition

Given a geographical or remote sensing image i.e. aerial or satellite imagery dataset D containing n sample images, where each image $I \in \mathbb{R}^{H\times W\times 3}$ represents a standard high-resolution RGB image with height H, width W, and 3 color channels. We implement GeoSAM (illustrated in Figure 2a), which produces a multi-class segmentation map $S \in \mathbb{R}^{H\times W}$ for each input image, where each pixel stores the predicted class index (e.g., background, pedestrian infrastructure, or road infrastructure). For training purposes, we convert this into a one-hot encoded multi-channel representation $\hat{S} \in \{0,1\}^{H\times W\times C}$, where C is the number of classes, and $\hat{S}_{i,j,c} = 1$ if pixel (i, j) belongs to class c, and 0 otherwise.

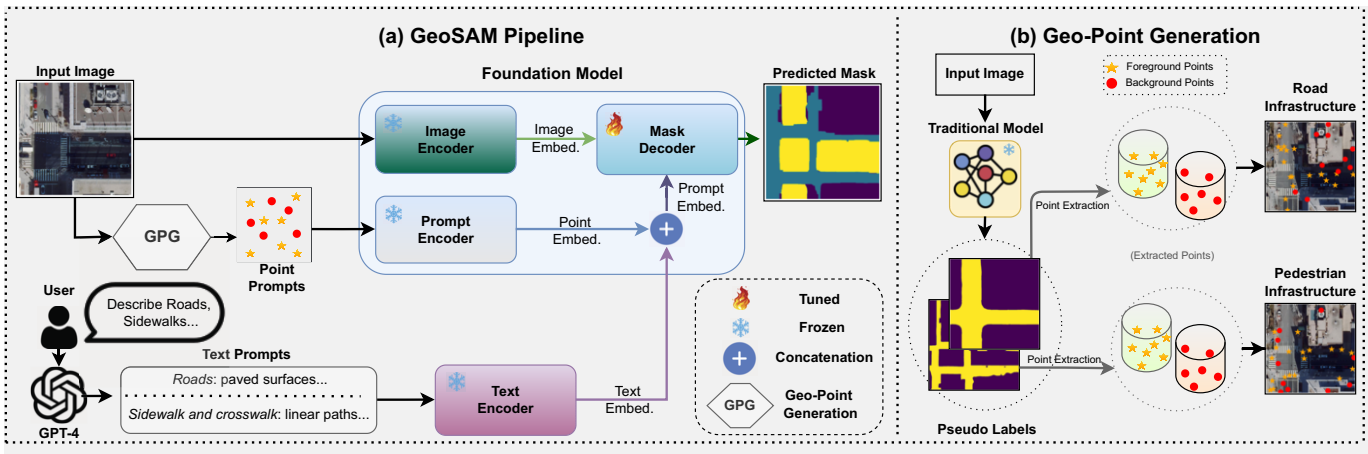

Figure 2. (a) **GeoSAM**: Training pipeline for mobility infrastructure segmentation. Text prompts generated by GPT-4 and point prompts generated from a pre-trained traditional model are utilized to tune a foundation model decoder in producing segmentation masks (yellow = road, blue = pedestrian). (b) **Geo-Point Generation**: Point prompts are generated from pseudo labels created by a pre-trained traditional model; stars and circles represent foreground and background points, respectively.

3.2 SAM: Background

Segment Anything Model (SAM) consists of an image encoder ($\mathrm{Enc_I}$), a prompt encoder ($\mathrm{Enc_P}$), and a mask decoder ($\mathrm{Dec_M}$). Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a set of prompts P, SAM encodes the image as $F_I = \mathrm{Enc_I}(I)$ and the prompts as $T_P = \mathrm{Enc_P}(P)$. These embeddings are passed to the decoder, which performs attention-based interactions and predicts the segmentation mask:

$$S = \mathrm{Dec_M}(F_I, T_P).$$

Prompt embeddings T_P guide the decoder using both spatial and semantic information.

3.3 Multi-Modal Prompt Generation

Point Prompts To segment sparse and spatially extensive structures like roads and sidewalks in remote sensing images, GeoSAM leverages point prompts as its primary guidance mechanism. Unlike bounding boxes, which are often impractical for such objects due to their large spatial extent, point prompts enable precise localization with minimal input. We introduce GPG, an automated approach (Figure 2b) that generates multiple foreground and background point prompts from a pre-trained traditional model f_{pre}. Each point is a 2D spatial coordinate on the image, serving as either a foreground cue to guide the model's attention or a background cue to indicate regions to avoid. Using multiple points helps reduce ambiguity and enables the model to more accurately localize the target object, especially in complex or overlapping areas.

The input image $I \in \mathbb{R}^{H\times W\times 3}$ is processed by the pre-trained model f_{pre}, which outputs a pseudo-label segmentation map of the same size:

$$M_{\text{pseudo}} = f_{\text{pre}}(I), \quad M_{\text{pseudo}}[i,j] \in \{0, 1, \ldots, C_{\text{pre}} - 1\}, \tag{1}$$

where C_{pre} is the number of semantic classes in the pre-trained model, and each pixel in M_{pseudo} is assigned one of these class labels.

We decompose the pseudo-label map M_{pseudo} into multiple binary masks, one for each semantic class, as illustrated in Figure 2b. For each class-specific mask, we randomly sample a set of point prompts $x = \{x_i\}_{i=1}^{k}$, where each point x_i is a 2D coordinate in the image domain $\Omega_I \subset \mathbb{R}^2$. The total number of sampled points per class is denoted by k, and these points are used as point prompts to guide the segmentation model. Then the set x is partitioned into foreground and background points:

$$\begin{aligned} x^{\text{fg}} &= \{x_i \in \Omega_I \mid M_{\text{pseudo}}(x_i) \in \mathcal{C}_{\text{fg}}\}, \qquad |x^{\text{fg}}| = k_1, \\ x^{\text{bg}} &= \{x_i \in \Omega_I \mid M_{\text{pseudo}}(x_i) \in \mathcal{C}_{\text{bg}}\}, \qquad |x^{\text{bg}}| = k_2, \\ x &= x^{\text{fg}} \cup x^{\text{bg}}, \end{aligned} \tag{2}$$

where $k = k_1 + k_2$ denotes the total number of sampled point prompts and $\mathcal{C}_{\text{fg}}$ and $\mathcal{C}_{\text{bg}}$ are the sets of class labels corresponding to foreground and background, respectively. These point prompts are transformed into high-dimensional embeddings of dimension C using the frozen prompt encoder $\mathrm{Enc_P}$.

$$T_x = \mathrm{Enc_P}(x) \in \mathbb{R}^{k\times C}. \tag{3}$$

The pre-learned position embeddings from SAM's pre-training (indicating whether a point is in the foreground or background) are appended with the T_x. Here, the accuracy of M_{pseudo} is not critical; as long as the generated points are approximately within the foreground or background regions of the class, they can effectively guide the focus of the segmentation process. For our experiments, we adopt a standard semantic segmentation model f_{pre} as the pre-trained traditional model, which produces a multi-class segmentation map containing class sets of: $\mathcal{C}_{\text{fg}}^{\text{road}} = \{\text{road}\}$, $\mathcal{C}_{\text{fg}}^{\text{pedestrian}} = \{\text{sidewalk}, \text{crosswalk}\}$, $\mathcal{C}_{\text{bg}} = \{\text{background}\}$. Each binary segmentation task, either between $\mathcal{C}_{\text{fg}}^{\text{road}}$ and $\mathcal{C}_{\text{bg}}$, or between $\mathcal{C}_{\text{fg}}^{\text{pedestrian}}$ and $\mathcal{C}_{\text{bg}}$, uses its corresponding foreground and background point prompts to provide GeoSAM with class-specific spatial guidance.

Text Prompts In addition to geometric guidance from point prompts, GeoSAM incorporates semantic context through text prompts t, enhancing the model's ability to distinguish between overlapping objects. While point prompts indicate specific pixel locations, a single pixel may belong to multiple classes (e.g., both road and crosswalk at the same time). To resolve such ambiguities, text prompts provide class-specific descriptions of the target object.

Each text prompt follows the format: "[class]: Description.", where [class] denotes the target class (e.g., roads or sidewalks/crosswalks). For example, a generated prompt might be "Roads: paved surfaces, vehicle lanes" (as can be seen in Figure 2a). To improve robustness and avoid overfitting to static descriptions, we dynamically generate diverse class-specific text prompts during training using OpenAI's GPT-4 [24]. These prompts are generated solely based on the class name, without any access to image content, ensuring variability [9] in phrasing while maintaining class relevance. The following instruction is provided to GPT-4 to create these class definitions:

> "role: system, content: You are a creative assistant, skilled in providing detailed visual descriptions of objects as seen in aerial imagery."
> "role: user, content: Print out a visual description (don't mention their names) that can be seen from aerial images of [CLS] (in one line, 4 to 5 words, not more, not less)."

For each class, we generate a set of t_n text prompts $t = \{t_1, t_2, \ldots, t_{t_n}\}$, which are encoded using CLIP's text encoder [26] to produce text embeddings $T_t \in \mathbb{R}^{t_n \times C}$. CLIP (Contrastive Language–Image Pretraining) is a vision-language model trained to align image and text pairs in a shared embedding space. We leverage CLIP's inherent ability to project text and image inputs into a shared embedding space for effective cross-modal alignment. To enhance class discrimination, we append a learnable class-specific embedding $E_{\text{cls}} \in \mathbb{R}^{1\times C}$ to each text embedding, where $C = 512$ is the embedding dimension of CLIP. This mitigates the variability introduced by natural language descriptions (e.g., from GPT-4) by allowing the model to learn a consistent, discriminative representation for each class. The resulting embeddings are L2-normalized along the feature dimension and subsequently projected to match SAM's embedding dimension of 256 using a trainable linear projection layer $W_t \in \mathbb{R}^{C\times 256}$:

$$T_t = \text{NORM}[f_{\text{clip}}(t)]W_t \in \mathbb{R}^{n\times 256}. \tag{4}$$

Joint Multi-Modal Prompts As text prompt embeddings T_t encode the semantic representation of the target class, they naturally complement the foreground point embeddings T_x, which capture precise spatial localization. The two types of prompt embeddings are then concatenated along the batch (prompt) dimension to form the joint prompt embedding:

$$T_P = \begin{bmatrix} T_x \\ T_t \end{bmatrix} \in \mathbb{R}^{(k+n)\times C}.$$

This design allows T_x to provide geometric position cues, while T_t enriches the representation with high-level semantic information, enabling the model to reason more effectively about the target object. Then, T_P along with F_I are concatenated and supplied to the decoder.

3.4 Fine-Tuning the Decoder

Decoder Architecture The decoder utilizes a combination of bidirectional transformers, where image embeddings (F_I) are updated through repeated Self Attention (SA) and Cross Attention (CA) with prompts. The self-attention operation on T_P enables interaction among different prompts, allowing them to exchange information and refine their representations before attending to the image features.

$$\begin{aligned} T'_P &= \mathbf{SA}(T_P), \\ \hat{T_P} &= T'_P + \mathbf{MLP}_P(\mathbf{CA}(T'_P, F_I)), \\ \hat{F_I} &= F_I + \mathbf{MLP}_I(\mathbf{CA}(F_I, \hat{T_P})), \end{aligned} \tag{5}$$

where $\hat{T_P}$ represents the refined set of prompt embeddings, and $\hat{F_I}$ denotes the updated set of visual embeddings after refining the embeddings by attending to the positional and semantic information of the prompts, enabling context-aware representations that ultimately produce the segmentation map. GeoSAM fine-tunes only the decoder while keeping the rest of the model frozen; a common strategy in foundation model adaptation [19]. This PEFT approach leverages the encoder's general representation capabilities while reducing the computational overhead of a large foundation model by restricting updates to the task-specific decoder.

Segmentation Map Adaptation SAM is originally a binary-class segmentation model, producing a map that distinguishes only foreground from background for a single class. GeoSAM extends this by generating a multi-channel segmentation map, where each channel corresponds to a target class, such as road or pedestrian infrastructure. Both classes are processed jointly throughout the pipeline by a single shared decoder, and the model outputs all channels simultaneously. This design makes the framework easily extensible to additional classes, and the loss is computed by comparing the resulting multi-channel outputs with one-hot encoded ground truth maps.

Loss Function We employ Dice Focal Loss, a synergistic combination of Dice Loss and Focal Loss, to address the challenges inherent in segmenting high-resolution remote sensing images. Given that mobility infrastructure occupies only a small fraction of these images, Dice Focal Loss effectively balances the need for precise overlap accuracy while mitigating the impact of severe class imbalance. Let $S^c \in [0,1]^N$ and $G^c \in \{0,1\}^N$ denote the predicted and ground-truth binary masks for class c, flattened over all N pixels. Dice Loss is defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{c=1}^{C}\langle S^c, G^c\rangle}{\sum_{c=1}^{C}(\|S^c\|_1 + \|G^c\|_1)}, \tag{6}$$

where $\langle\cdot,\cdot\rangle$ denotes the dot product and $\|\cdot\|_1$ the ℓ_1-norm (sum of elements).

Focal Loss applies a balancing factor α_c and focusing parameter γ to emphasize hard examples:

$$\mathcal{L}_{\text{Focal}} = -\sum_{c=1}^{C}\alpha_c\,(1-S^c)^{\gamma}\,G^c\log(S^c), \tag{7}$$

The total loss is computed as:

$$\mathcal{L}_{\text{DiceFocal}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{Focal}}. \tag{8}$$

4 Experiments

Our objective is to confirm the efficacy of the newly proposed GeoSAM in enhancing segmentation performance across various metrics. This will be accomplished by conducting a comprehensive set of experiments designed to answer critical research inquiries. ***Q1***: *Does GeoSAM outperform the current state-of-the-art (SOTA) methods in terms of performance in mobility infrastructure segmentation?* ***Q2***: *Can GeoSAM demonstrate superior generalizability by performing effectively on previously unseen datasets?* ***Q3***: *Is automated prompt generation necessary in the case of mobility infrastructure segmentation?*

Table 1. Details of the datasets used in this study. D_{train} and D_{test} are collected from Washington, D.C., while D_{gen} is used to evaluate generalization performance on Cambridge, MA.

Dataset	Region	Geographical Coordinates	Base Image Size	#Base Images	#Stitched Images
D_{train}	Washington DC	38.905788, -77.045019 38.90968, -77.019694	(512, 512)	2240	560
D_{test}	Washington DC	38.8968333, -77.0074118 38.906958, -76.988948	(512, 512)	1184	296
D_{gen}	Cambridge	42.360067, -71.144373 42.395258, -71.051704	(256, 256)	38080	2380

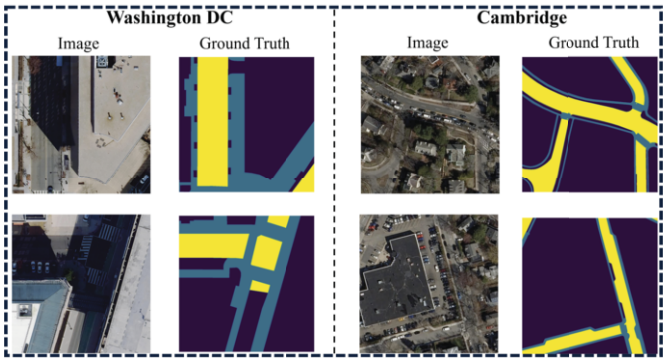

Figure 3. Randomly picked examples of Washington, D.C. and Cambridge from the datasets (yellow = road, blue = pedestrian infrastructure).

4.1 Datasets

We define the training dataset as $D_{\text{train}} = \{I_{\text{train}}, G_{\text{train}}\}$, where I_{train} and G_{train} represent n images and corresponding segmentation ground truths (masks for roads and pedestrian infrastructure). Similarly, the test dataset is denoted as $D_{\text{test}} = \{I_{\text{test}}, G_{\text{test}}\}$, and the generalization dataset as $D_{\text{gen}} = \{I_{\text{gen}}, G_{\text{gen}}\}$.

These datasets are constructed from high-resolution orthorectified aerial images and publicly available GIS data [8, 2], following the methodology in [18]. Orthorectified tiles are aerial images that have been geometrically corrected to ensure uniform scale and true top-down perspective, enabling accurate spatial measurements. These tiles [33] are downloaded using geographical bounding boxes and appropriate zoom levels (e.g., zoom level 0 spans the entire world). The GIS data, provided by respective local government authorities, contains accurate coordinate information on urban infrastructure such as roads and sidewalks, enabling reliable mask generation for the two infrastructure classes. We additionally perform manual inspection and correction on the generated masks to fix any potential inconsistencies or missing annotations. GeoSAM is trained and tested in separate regions of Washington, D.C., with an additional test conducted in Cambridge, MA, to evaluate generalization. We first download the base images at their native resolutions (Table 1) and stitch adjacent base image tiles within each region to form input images of size 1024×1024, using zoom level 20. This resolution corresponds to high-detail aerial imagery, capturing fine-grained urban structures like lanes and sidewalks, consistent with standard geographical mapping scales. Figure 3 provides a couple of examples, and the dataset preparation is explained further in Appendix A.3 [30].

4.2 Implementation Details

Experiments Setup We adopted ViT-H [11] as the encoder version of SAM and initialized the model with pre-trained weights from SAM's ViT-H version. Following the original SAM paper settings [21], the choice of optimizer was the AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$), with an initial learning rate set at 10^{-5} and weight decay of 0.1, and no data augmentation techniques were applied. Following our experimentation on various values, we chose 0.8 for the balancing factor (α) and 2 for the focusing parameter (γ) loss function. To have an adaptable learning rate, a cosine annealing learning rate scheduler was employed with a maximum learning rate decaying smoothly to a minimum value (10^{-7}) over the course of training. A pre-trained nnU-Net [20] model (trained on the training dataset) has been selected as the pre-trained traditional model for point prompt generation. We adopted CLIP's ViT-B version [26] as the text encoder. Finally, for point prompt generation, we selected 2000 foreground and 1000 background points. All the experiments were conducted on an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory and Python 3.10.9. We use a total of 100 epochs to train GeoSAM as well as the other baseline models. The source code can be found in this publicly accessible GitHub repository [29].

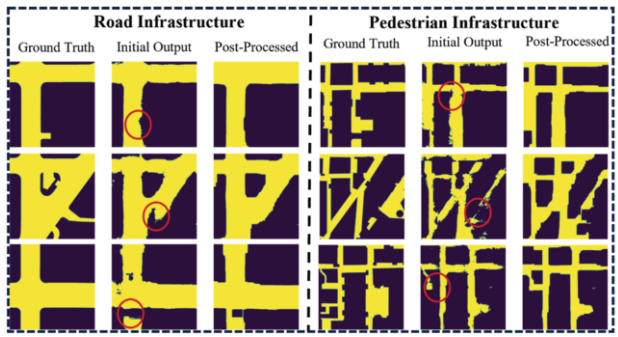

Figure 4. Postprocessing on two classes, with each row showing a randomly selected test image for (a) road infrastructure and (b) pedestrian infrastructure segmentation. Circles indicate cosmetic improvements.

Postprocessing We apply postprocessing uniformly to all model outputs, aiming to improve the structural coherence of segmentation maps, with a particular emphasis on ensuring path connectivity over precise pixel-wise correctness. Morphological operations like erosion and dilation address common segmentation issues. Erosion removes isolated regions, ensuring cleaner segmentation of pedestrian paths, while dilation connects disjointed paths to improve route continuity. These operations are performed with a (10×10) filter over a (1024×1024) resolution map and iterated 10 times for effective refinement. Figure 4 illustrates these techniques, showing improved connectivity and alignment with the ground truth. Comparisons between initial and postprocessed outputs highlight the removal of isolated regions and enhanced path continuity.

Benchmark Models We compare GeoSAM against several popular semantic segmentation models from both CNN- and ViT-based, and SAM-based approaches. All benchmarks follow GeoSAM's training setup, using the same postprocessing and no data augmentation for fair comparison. CNN-based baselines include UNet [27], nnU-Net [20], UNet++ [45], DeepLabv3+ [6], and HRNet [31]. ViT-based models include UNETR [15], Swin UNETR [14], SwinUNETR-V2 [16], and nnFormer [42]. Additionally, we compare GeoSAM with zero-shot SAM initialized with pre-trained weights (supplemented with point prompts created) and two notable SAM-based geographical segmentation works, such as RSPrompter [4] and UV-SAM [40]. We train each of the models from scratch using their default settings on D_{train} described in Section 4.1. The summary of each of the models can be found in Appendix A.4 [30]. During inference, we evaluate these models on both D_{test} and D_{gen} datasets. For each class, we compute the Intersection over Union (IoU) using a fixed threshold to binarize predictions, and the Average Precision (AP) by integrating over all possible thresholds.

Table 2. GeoSAM evaluation results against benchmark models ("Ped." for Pedestrian, "Infras." for Infrastructure). Washington, D.C. used for Testing, and Cambridge, MA used for evaluating generalizability. The best results are in **bold**, and the second-best results are in underlined.

Method	D_{test} (Washington, D.C.)						D_{gen} (Cambridge, MA)					
	IoU		mIoU	AP		mAP	IoU		mIoU	AP		mAP
	Road Infras.	Ped. Infras.		Road Infras.	Ped. Infras.		Road Infras.	Ped. Infras.		Road Infras.	Ped. Infras.	
UNet [27]	0.45	0.17	0.31	0.44	0.22	0.33	0.24	0.12	0.18	0.11	0.05	0.08
nnU-Net [20]	0.65	0.32	0.49	0.56	0.32	0.45	0.25	0.15	0.21	0.05	0.09	0.07
UNet++ [45]	0.61	0.30	0.45	0.54	0.34	0.43	0.24	0.08	0.16	0.12	0.06	0.09
DeepLabv3+ [6]	0.47	0.18	0.32	0.46	0.22	0.34	0.10	0.06	0.08	0.05	0.04	0.04
HRNet [31]	0.50	0.19	0.34	0.49	0.23	0.36	0.13	0.08	0.10	0.08	0.06	0.06
UNETR [15]	0.48	0.20	0.34	0.50	0.27	0.38	0.27	0.11	0.18	0.12	0.05	0.08
Swin UNETR [14]	0.63	0.26	0.44	0.57	0.29	0.43	0.13	0.09	0.11	0.09	0.04	0.06
SwinUNETR-V2 [16]	0.66	0.22	0.43	0.54	0.26	0.40	0.15	0.08	0.12	0.09	0.04	0.06
nnFormer [42]	0.60	0.21	0.41	0.52	0.30	0.41	0.16	0.09	0.13	0.11	0.05	0.08
Zero-shot SAM [21]	0.30	0.18	0.24	0.34	0.23	0.27	0.25	0.12	0.18	0.11	0.06	0.08
RSPrompter [4]	0.46	0.20	0.33	0.49	0.25	0.37	0.09	0.07	0.08	0.08	0.04	0.06
UV-SAM [40]	0.57	0.21	0.39	0.55	0.26	0.40	0.11	0.07	0.09	0.08	0.05	0.06
GeoSAM (Ours)	**0.70**	**0.39**	**0.54**	**0.61**	**0.42**	**0.51**	**0.34**	**0.18**	**0.26**	**0.20**	**0.16**	**0.18**

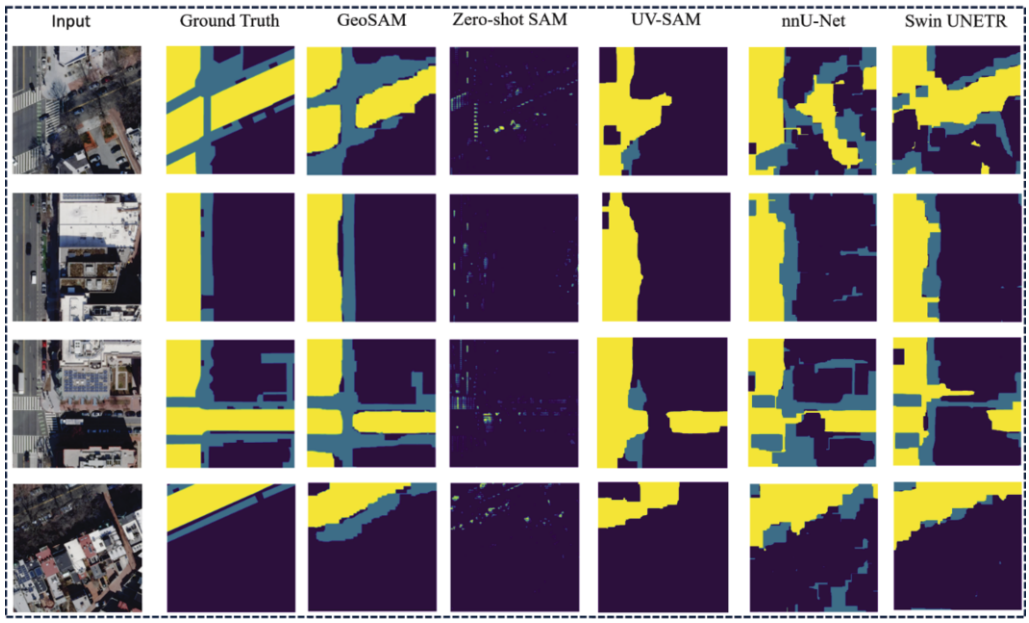

Figure 5. Comparative qualitative segmentation results: GeoSAM vs. other benchmark models. Different colors (blue=pedestrian infrastructure, yellow=road infrastructure) indicate distinct classes in the multi-class output. Each row displays a randomly selected image from the test dataset.

4.3 Results and Discussion

GeoSAM's Superiority Over Other Methods (*Q1*) Figure 5 presents a qualitative comparison of GeoSAM with zero-shot SAM, UV-SAM, nnU-Net, and Swin UNETR on randomly selected test images. The results highlight the limitations of zero-shot SAM on geographical images outside its training domain, particularly with thin boundary objects, as noted in Section 1. In contrast, GeoSAM demonstrates significantly improved segmentation accuracy, closely matching the ground truth. GeoSAM also outperforms UV-SAM, nnU-Net, and Swin UNETR, particularly in handling intricate boundaries by achieving results closest to the ground truth, underscoring its superiority for mobility infrastructure segmentation.

In Table 2, we compare GeoSAM's performance with established semantic segmentation models (CNN- and ViT-based) and SAM-based models. On the Washington, D.C. test set, GeoSAM outperforms SOTA models across both classes, surpassing the second-best model, nnU-Net, by 5% in mIoU and 6% in mAP. Compared to Zero-shot SAM, GeoSAM achieves a remarkable improvement of 30% in mIoU and 24% in mAP, highlighting SAM's limitations with geographical images. GeoSAM significantly outperforms UV-SAM and RSPrompter, the leading SAM-based models in this domain. *These results confirm GeoSAM's effectiveness, even when trained in limited data scenarios.*

Generalization Performance of GeoSAM (*Q2*) GeoSAM's performance on the generalization dataset from Cambridge, MA, reveals a decline across all models due to data shifts between regions. However, GeoSAM consistently outperforms SOTA models, achieving at least double the performance of the second-best model. This highlights the limitations of traditional models, whose generalization is constrained by training data, particularly in visually distinct regions like Cambridge. Foundation models like SAM exhibit better adaptability to diverse scenarios, and GeoSAM, enhanced by automated guided prompts, further improves this adaptability, achieving 5% and 9% higher mIoU and mAP than the second-best model. *These results validate GeoSAM's scalability and superior generalization capabilities, leveraging the strengths of a foundation model.*

Necessity of Auto Prompt Generation (*Q3*) Table 3 shows the effect of varying the number and ratio of point prompts on GeoSAM's

Table 3. Segmentation performance of GeoSAM using different numbers and foreground-to-background ratios of point prompts.

Point Prompts			IoU	
Foreground Points	Background Points	Ratio	Road Infrastructure	Pedestrian Infrastructure
100	50	2:1	0.64	0.31
1000	500	2:1	0.65	0.33
2000	2000	1:1	0.67	0.34
2000	**1000**	**2:1**	**0.70**	**0.39**
2000	4000	1:2	0.66	0.27

performance. Since these prompts simulate user input, we identify the optimal configuration generated by our automated system. A 2:1 foreground-to-background ratio with 2000:1000 points performs best, likely due to the large image resolution (1024×1024), where extensive foreground coverage helps segment large structures like roads and sidewalks while fewer background points reduce ambiguity. We use this as the default in GeoSAM. Even with 150 total points (first row), pedestrian infrastructure performance drops by 8%, highlighting the model's sensitivity to the number of prompts. Our chosen 2:1 ratio is further supported by the characteristics of mobility infrastructure: roads and sidewalks are spatially extensive yet sparse, requiring dense sampling across fragmented regions, while the background is semantically redundant. Over-sampling background reduces the learning signal, but increasing foreground beyond this ratio can lead to over-segmentation, where disconnected regions may be incorrectly predicted as continuous infrastructure. *These findings underscore the necessity of automated prompt generation, as manually crafting such a large number of prompts is practically infeasible.*

Table 4. Components ablation study: examining the effects on performance based on various model components.

Components			IoU	
Point Prompts	Text Prompts	Fine-tuning Decoder	Road Infras.	Pedestrian Infras.
✓	✓	✓	**0.70**	**0.39**
✓	×	✓	0.66	0.31
×	✓	✓	0.31	0.17
✓	×	×	0.24	0.13

Ablation Study Table 4 evaluates the impact of key components on GeoSAM, including point prompts, text prompts, and a fine-tuned decoder. Using only point prompts results in an 8% decrease in pedestrian infrastructure segmentation, emphasizing the critical role of text prompts in providing semantic understanding to resolve ambiguities. However, text prompts alone lead to a 22% performance drop, demonstrating their insufficiency for nuanced segmentation tasks due to the text encoder's semantic limitations. Instead, text prompts serve as effective secondary prompts, adding context to the decoder. As secondary prompts, they effectively enhance the model's focus by providing additional context to the decoder during segmentation. Removing the fine-tuned decoder further degrades performance, with zero-shot SAM showing a 26% drop in pedestrian infrastructure segmentation, highlighting the original SAM decoder's inadequacy for geographical images. Fine-tuning adapts the decoder to the unique challenges of this domain.

Table 5 demonstrates that the choice of backbone used to generate point prompts is not overly critical. We observe only minor performance drops when replacing nnU-Net [20] (our default) with UNet [27] or Swin UNETR [14]. While the overall framework remains robust, we note that pedestrian infrastructure shows slightly higher sensitivity to backbone changes than roads. Because of its structural variability and weaker visual cues, it can become more reliant on accurate spatial guidance. Nonetheless, as long as the backbone provides reasonably well-positioned foreground and background points, GeoSAM maintains strong performance. Even with different backbones, GeoSAM consistently outperforms all state-of-the-art models listed in Table 2.

Table 5. Backbone ablation study: performance comparison with different backbones as the traditional pre-trained model to generate automated point prompts for GeoSAM.

Pre-trained Backbone	IoU		AP	
	Road Infras.	Pedestrian Infras.	Road Infras.	Pedestrian Infras.
UNet [27]	0.67	0.35	0.59	0.38
nnU-Net [20]	**0.70**	**0.39**	0.61	**0.42**
Swin UNETR [14]	0.69	0.37	**0.62**	0.39

Table 6. Average inference time of different models on D_{test} with 1024×1024 input images. Inference times are reported based on the implementation and settings described in this work; results may vary under different configurations.

Model	Tuned Params. (M)	Inference Time (sec./image)
nnUNet [20]	7.8	2.01
DeepLabV3+ [6]	5.4	2.86
Swin UNETR [16]	6.3	3.44
GeoSAM (ours)	4.2	3.06

Further, as shown in Table 6, GeoSAM attains competitive inference speed, being only marginally slower than nnUNet [20] and DeepLabV3+ [6], despite leveraging a foundation model-based encoder. This efficiency largely stems from our design choice to precompute pseudo labels offline and load them from disk during inference, enabling fast generation of point prompts without additional runtime overhead. Notably, GeoSAM outperforms Swin UNETR [16] in inference time, underscoring its efficiency among recent SOTA methods. Moreover, GeoSAM requires the fewest tunable parameters during training, highlighting its suitability for resource-constrained settings and real-world deployment.

5 Conclusion

GeoSAM adapts SAM for mobility infrastructure segmentation in geographical images, with a strong social impact, particularly for pedestrian safety. It integrates multi-modal prompts (point and text) and fine-tunes SAM's decoder. Unlike existing methods, our training and end-to-end inference pipeline is transferable across locations and classes using any pre-trained traditional model. The approach is generic, reproducible, and adaptable to various domain-specific segmentation tasks.

Limitation and Future Work We aim to extend the application of GeoSAM to a wider range of geographical regions by incorporating datasets from various other cities, enabling a more comprehensive analysis of its generalizability across diverse urban layouts and visual conditions. In addition, we plan to expand support for additional object types such as stairs, islands/bridges, and potholes, as well as explore its applicability to other imaging modalities.

Acknowledgements

This paper was supported by the U.S. National Science Foundation (NSF) under Award Number 2235225. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021.

[2] Cambridge GIS (2018). Cambridge sidewalk. https://www.cambridgema.gov/GIS/gisdatadictionary, 2018.

[3] K. Cha, J. Seo, and T. Lee. A billion-scale foundation model for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[4] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi. Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[7] X. Chen, Q. Sun, W. Guo, C. Qiu, and A. Yu. Ga-net: A geometry prior assisted neural network for road extraction. *International Journal of Applied Earth Observation and Geoinformation*, 114:103004, 2022.

[8] DC GIS (2019). Roads 2019. https://opendata.dc.gov/datasets/, 2019.

[9] M. M. Derakhshani et al. Variational prompt tuning improves generalization of vision-language foundation models. In *ICLR Workshop on Foundation Models*, 2023.

[10] L. Ding, K. Zhu, D. Peng, H. Tang, K. Yang, and L. Bruzzone. Adapting segment anything model for change detection in vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[11] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[12] W. Feng, F. Guan, C. Sun, and W. Xu. Road-sam: Adapting the segment anything model to road extraction from large very-high-resolution optical remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024.

[13] P. Gudžius, O. Kurasova, V. Darulis, and E. Filatovas. Deep learning-based object recognition in multispectral satellite imagery for real-time applications. *Machine Vision and Applications*, 32(4):98, 2021.

[14] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[15] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[16] Y. He, V. Nath, D. Yang, Y. Tang, A. Myronenko, and D. Xu. Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023.

[17] C. Henry, S. M. Azimi, and N. Merkle. Road segmentation in sar satellite images with deep fully convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(12):1867–1871, 2018.

[18] M. Hosseini, A. Sevtsuk, F. Miranda, R. M. Cesar Jr, and C. T. Silva. Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery. *Computers, Environment and Urban Systems*, 101:101950, 2023.

[19] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[20] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[22] Z. Liu, Z. Li, Y. Liang, C. Persello, B. Sun, G. He, and L. Ma. Rsps-sam: A remote sensing image panoptic segmentation method based on sam. *Remote Sensing*, 16(21):4002, 2024.

[23] X. Ma, Q. Wu, X. Zhao, X. Zhang, M.-O. Pun, and B. Huang. Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[24] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[25] L. P. Osco, Q. Wu, E. L. de Lemos, W. N. Gonçalves, A. P. M. Ramos, J. Li, and J. M. Junior. The segment anything model (sam) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124:103540, 2023.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[27] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Springer, 2015.

[28] A. Saha. Conducting semantic segmentation on landcover satellite imagery through u-net architectures. In *Proceedings of the Future Technologies Conference*, pages 758–764. Springer, 2022.

[29] R. I. Sultan. *Source Code* for "GeoSAM". https://github.com/rafiibnsultan/GeoSAM, 2025. GitHub repository.

[30] R. I. Sultan, C. Li, H. Zhu, P. Khanduri, M. Brocanelli, and D. Zhu. Geosam: Fine-tuning sam with sparse and dense visual prompting for automated segmentation of mobility infrastructure. *arXiv preprint arXiv:2311.11319*, 2023. Full version of this paper.

[31] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[32] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[33] US Geological Survey (2018). USGS EROS Archive - Aerial Photography - High Resolution Orthoimagery (HRO). https://doi.org/10.5066/F73X84W6, 2018.

[34] D. Wang, J. Zhang, B. Du, M. Xu, L. Liu, D. Tao, and L. Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36, 2024.

[35] L. Wang, M. Zhang, and W. Shi. Cs-wscdnet: Class activation mapping and segment anything model-based framework for weakly supervised change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[36] T. Wu, Y. Hu, J. Qin, X. Lin, and Y. Wan. Tpp-sam: A trajectory points prompting segment anything model for zero-shot road extraction from high-resolution remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[37] Y. Yan, H. Wen, S. Zhong, W. Chen, H. Chen, Q. Wen, R. Zimmermann, and Y. Liang. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, 2024.

[38] Z. Yan, J. Li, X. Li, R. Zhou, W. Zhang, Y. Feng, W. Diao, K. Fu, and X. Sun. Ringmo-sam: A foundation model for segment anything in multimodal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

[39] J. Zhang, Z. Zhou, G. Mai, M. Hu, Z. Guan, S. Li, and L. Mu. Text2seg: zero-shot remote sensing image semantic segmentation via text-guided visual foundation models. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 2024.

[40] X. Zhang, Y. Liu, Y. Lin, Q. Liao, and Y. Li. Uv-sam: Adapting segment anything model for urban village identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38(20), 2024.

[41] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[42] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE transactions on image processing*, 32:4036–4045, 2023.

[43] L. Zhou, C. Zhang, and M. Wu. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 182–186, 2018.

[44] X. Zhou, F. Liang, L. Chen, H. Liu, Q. Song, G. Vivone, and J. Chanussot. Mesam: Multiscale enhanced segment anything model for optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[45] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.