GeoSAM: Fine-tuning SAM with Sparse and Dense Visual Prompting for Automated Segmentation of Mobility Infrastructure

Rafi Ibn Sultan[1], Chengyin Li[1], Hui Zhu[1], Prashant Khanduri[1], Marco Brocanelli[2], Dongxiao Zhu[1]
[1]Wayne State University, [2]Ohio State University
[1]{hm4013, gv2145, hq2197, dzhu}@wayne.edu [2]brocanelli.1@osu.edu

Abstract

The Segment Anything Model (SAM) has shown impressive performance when applied to natural image segmentation. However, it struggles with geographical images like aerial and satellite imagery, especially when segmenting mobility infrastructure including roads, sidewalks, and crosswalks. This inferior performance stems from the narrow features of these objects, their textures blending into the surroundings, and interference from objects like trees, buildings, vehicles, and pedestrians - all of which can disorient the model to produce inaccurate segmentation maps. To address these challenges, we propose Geographical SAM (GeoSAM), a novel SAM-based framework that implements a fine-tuning strategy using the dense visual prompt from zero-shot learning, and the sparse visual prompt from a pre-trained CNN segmentation model. The proposed GeoSAM outperforms existing approaches for geographical image segmentation, specifically by 20%, 14.29%, and 17.65% for road infrastructure, pedestrian infrastructure, and on average, respectively, representing a momentous leap in leveraging foundation models to segment mobility infrastructure including both road and pedestrian infrastructure in geographical images.

1. Introduction

While a substantial amount of research has focused on road infrastructure segmentation from geographical imagery like aerial and satellite images, pedestrian infrastructure such as sidewalks and 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 [24, 32]. Even existing pedestrian route accessibility studies rely on simplified road data rather than actual pedestrian infrastructure segmentation [37]. This overlooks critical needs, especially for people with disabilities, who require detailed accessibility information to navigate safely [51]. Therefore accurately segmenting mobility infrastructure including both road and pedestrian infrastructure could provide invaluable information about accessible pedestrian routes and trip locations.

Current approaches to mobility infrastructure segmentation are developed under the traditional Convolution Neural Network (CNN) [41, 49, 67] or Vision Transformer (ViT) models [45, 46] that require human-labeled roads, sidewalks, and crosswalks as the training set [23, 24, 51, 52]. For example, the project sidewalk [51] employs crowd workers to label and validate the geographical objects manually onsite in the city of Seattle. Projects like this have contributed high-quality annotated geographical imagery data sets and achieved impressive segmentation performance in several urban cities. Yet, similar projects with a comparable segmentation performance have not been implemented in the rest of the country, particularly in the under-resourced rural areas owing to the lack of quality training data. The limitations of traditional segmentation models are rooted in their dependence on large volumes of high-quality labeled image datasets, which can hinder scalability and adaptability to diverse tasks.

The rise of vision foundation models [29, 47, 48] represents a big leap in scaling up segmentation models, allowing for powerful zero-shot or few-shot learning capabilities and flexible prompting. Without the need for re-training, these models can quickly adapt to a new downstream task. To tackle this problem, we turn to the Segment Anything Model (SAM) [29], one of the first vision foundation models for image segmentation. With the introduction of SAM, designed with the ambition of segmenting virtually anything in images, the field of image segmentation is in a fast pace of transition from traditional CNN or ViT models to pre-trained foundation models.

Zero or few-shot learning and fine-tuning using Parameter Efficient Fine-Tuning (PEFT) are the two primary approaches for leveraging the capabilities of foundation models. Zero-shot learning sets the initial groundwork for a downstream task, utilizing the model without specific contextual information [29, 47]. While this basic utilization has shown early success, zero-shot SAM often struggles to generalize effectively across various downstream tasks [26, 33]. To ad-

dress this limitation, few-shot learning and PEFT have been introduced. In few-shot learning, one or more input images along with their corresponding ground truth labels are provided as the context for adapting to a specific task [3, 64]. Whereas, PEFT is another strategy that fine-tunes a subset (typically a lightweight module) of a large foundation model with billions of parameters while keeping the majority of parameters frozen to optimize task performance without full retraining [7, 13, 15, 25, 30].

Considering the task of mobility infrastructure segmentation employed in this work, the following challenges exist in utilizing zero-shot SAM: i) sidewalks tend to have very thin borders alongside the road borders from an aerial view, and ii) sidewalks have very similar texture (sometimes the same) to roads which makes it difficult for the zero-shot SAM to distinguish between them. To address these challenges, we introduce Geographical SAM (GeoSAM), an end-to-end model tailored for segmenting pedestrian infrastructure via binary segmentation of road and pedestrian infrastructure.

As illustrated in Figure 1, our approach incorporates an automated prompt generation process. We employ sparse prompts generated by a domain-specific CNN encoder and complement them with dense prompts produced by zero-shot SAM to perform additional fine-tuning of SAM using PEFT techniques. Sparse prompts, which are essentially clicks on the image, provide the model with a context for segmentation by indicating where to focus. These sparse prompts are complemented by dense prompts (which are low-quality segmentation maps), which offer the model additional context about the objects to be segmented. Instead of relying on human intervention to provide these prompts for contextual information, we have devised an automated system that operates independently.

The fine-tuning process plays a crucial role in imparting domain-specific knowledge to SAM. Our tailored strategy to use zero-shot SAM to generate dense prompts is inspired by a key technique from the training strategy of SAM itself [29], which involves supplying the mask prediction as dense prompts from the previous iteration as additional prompts to the model. Conversely, we employ a domain-specific CNN encoder for the segmentation of geographical objects, enabling us to capture precise location information for the generation of sparse prompts specific to mobility infrastructure in the geographical imagery domain.

GeoSAM stands out by utilizing an improved vision foundation model where both categories (sparse and dense) of prompts are employed for mobility segmentation, setting itself apart from previous CNN-based approaches [2, 22, 24, 66]. Hence, GeoSAM not only excels in the segmentation of mobility infrastructure by improving on other contemporary works in accuracy but also showcases the capabilities of foundation models, expanding the horizons of geographical image analysis. Our contributions are summarized below:

- We pioneer the adaptation of the foundation model, SAM, for mobility infrastructure segmentation using geographical imagery, without any human intervention, overcoming the limitations of zero-shot SAM.
- We develop the fine-tuning and prompting of SAM for geographical imagery, empowering SAM with domain-specific knowledge drawn from the utilization of both sparse and dense prompts.
- We design and implement a novel automated pipeline to generate both dense prompts from zero-shot learning and sparse prompts from a pre-trained CNN encoder to enhance SAM's effectiveness and efficiency on the underperforming mobility infrastructure segmentation task.

2. Related Work

2.1. CNN-based Geographical Image Segmentation

Prior to the emergence of foundation models, CNN-based models like UNet [49] and vision transformer-based [14] works, which follow an encoder-decoder architecture for semantic segmentation, were the standard choice for various geographical segmentation tasks. Simple UNet-based approaches like [22, 50] and more advanced encoder-decoder-based works like [2, 8, 18, 66] are developed to execute various geographical image segmentation tasks. Additionally, there are endeavors like [1, 6, 19], where researchers exploit multiple machine learning techniques to enhance the performance of these CNN-based segmentation models for geographical object segmentation.

Furthermore, Tile2Net [24], also based on CNN, is one of the most established works in mobility infrastructure segmentation in geographical imagery. Their primary focus is mapping sidewalk networks using aerial and satellite imagery, involving segmenting mobility infrastructure elements like roads, sidewalks, and crosswalks. For the semantic segmentation part of their network creation, they train a hierarchical multi-scale attention model [56] and HRNet-W48 [55] from scratch; [60] with object-contextual representations [63] as the backbone. While these diverse efforts contribute significantly to the field of geographical image segmentation, many of them share similar limitations, as they typically require a lot of supervised data for each different task and necessitate retraining. While accuracy improvements are valuable, they do not necessarily address some of the more fundamental issues inherent in geographical image segmentation, i.e., generalizability to new locations.

In addition to conventional CNN models trained from scratch, there have been mentionable transfer learning-based efforts, exemplified in works like [28, 65], where a model trained from the source task is used to reduce the computational demands for various related downstream tasks. However, it's worth noting that in practice, these researchers often

encounter the need to conduct further fine-tuning or retraining of these models to align them with the precise objectives of their respective tasks. This lack of generalizability leads to these source models being re-trained to achieve competitive performance in the downstream task. These transfer learning-based approaches remain task-specific, in contrast to foundation models, which are designed to be more general and not tied to specific tasks.

2.2. SAM-based Geographical Image Segmentation

Vision foundation models, in their essence, aim to address the shortcomings of CNN-based segmentation models by being readily adaptable to segment previously unknown classes for various downstream tasks. SAM [29], a foundation model dedicated to segmentation tasks has three main components: (i) a ViT-based [14] image encoder that has been trained with over 1 billion masks on 11 million images to compute the image embeddings, (ii) a prompt encoder that takes prompts from users (guiding the model for the context where to focus at) and encodes the embeddings, and, (iii) a lightweight mask decoder to generate segmentation map based on the received image embedding, and prompts embedding. These prompts can take the form of sparse prompts (such as clicks, bounding boxes, or texts) or dense prompts (mask inputs).

While the application of SAM in the field of geographical imagery is not as extensive as in other domains, it has been utilized in several research endeavors. Considering the zero-shot SAM approaches, several works, including [16, 59, 61, 62], have employed SAM's zero-shot capabilities for a range of downstream tasks, extending beyond segmentation. Additionally, in [42], a hybrid approach combining both zero-shot and one-shot learning is applied to SAM for segmenting geographical imagery. However, it's crucial to note that these zero-shot-based approaches primarily target objects with well-defined boundaries and distinguishable physical contexts in their surroundings. In such cases, SAM doesn't necessitate extensive domain-specific knowledge for accurate segmentation.

When zero-shot SAM encounters difficulties in specific domains, there have been attempts to fine-tune SAM using PEFT techniques. Mixed works in geographical imagery like [11, 27] delve into the exploration of fine-tuning using diverse PEFT techniques for a range of downstream tasks such as geo-localization and mapping. Beyond geographical imagery, [7, 12, 43] fine-tune a small number of parameters exploiting various PEFT techniques across a variety of natural imagery. However, as far as our knowledge extends, no fine-tuning work specifically tailored to geographical images for mobility infrastructure segmentation has been done, a significant under-performing task that may generate tremendous social impact.

While most of these works are based on manual human prompting in the inference stage, there are also works focusing on automating prompt generation, mainly in the medical imaging segmentation domain. Works such as [33, 53] report about developing auto prompts generation techniques by replacing the SAM prompt encoder with a trainable network, a process that demands substantial training data. However, auto-prompting in geographical image segmentation represents a non-trivial task due to the lack of curated geographic infrastructure data for prompt generation.

2.3. Pre-training Geographical Image Segmentation Foundation Model

Very recently, researchers have also attempted to train domain-specific foundation models using large geographical imagery data sets. [58] uses plain ViT models with about 0.1 billion parameters to train large vision models tailored to remote sensing tasks and investigate how these large ViT models perform on object detection, scene classification, and semantic segmentation tasks. The work in [40] compares different visions of foundation models with CNN-based fine-tuned models for geographical images and they conclude that the foundation models fall short compared to the CNN-based fine-tuned models. Similar to SAM, [5, 36] develop their own foundation models for geographical imagery based on scaled versions of ViT and CLIP models [47] respectively. Many of these works focus on road infrastructure segmentation tasks, which, unfortunately, do not specifically address the key issue of pedestrian infrastructure segmentation tasks, i.e., sidewalk/ crosswalk segmentation. Additionally, given that they do not provide the source code, assessing their model's effectiveness in the context of pedestrian infrastructure presents challenges.

3. Method

This section is organized to provide an overview of SAM at first followed by a discussion of the training and inference phases of GeoSAM.

3.1. SAM: Background

SAM comprises three elements: an image encoder (referred to as $\mathrm{Enc_I}$), a prompt encoder (referred to as $\mathrm{Enc_P}$), and a mask decoder (referred to as $\mathrm{Dec_M}$). SAM, designed as a model that can be prompted, accepts an image, denoted as I, and a collection of prompts, known as P. These prompts can represent a point, a box, or a dense mask. In its operation, SAM initially employs $\mathrm{Enc_I}$ to extract features from the input image. It then utilizes $\mathrm{Enc_P}$ to transform the human-provided prompts, which have a length of k, into prompt tokens. Specifically:

$$F_I = \mathrm{Enc_I}(I), \quad T_P = \mathrm{Enc_P}(P), \tag{1}$$

In equation 1, F_I is the feature embedding of the input image where $F_I \in \mathbb{R}^{h \times w \times c}$, h and w represent the resolution of

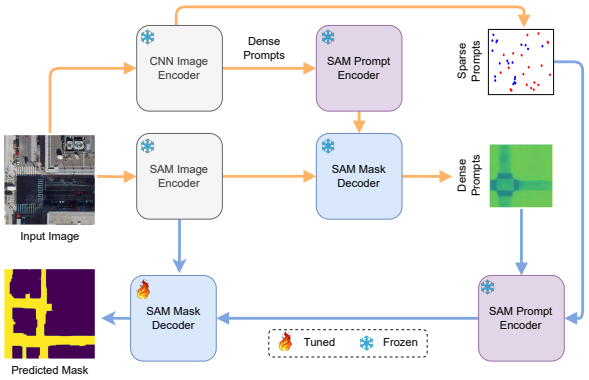

Figure 1. Training GeoSAM, an automated mobility infrastructure segmentation pipeline. In **Prompts Generation** (orange arrows), the model generates the sparse and dense prompts with the help of a secondary CNN-based geographical image encoder. Sparse prompts are generated automatically from the output of the secondary model and the soft mask is generated by the SAM mask decoder which is used as the dense prompts. In **Fine-tuning** (blue arrows), the prompts generated from the previous stage are used in the tunable decoder to produce the mask.

Figure 2. Sparse prompts generated based on segmentation maps created by the pre-trained CNN image encoder. Here, the foreground class is the sidewalk/crosswalk, blue and red circles represent foreground and background clicks respectively.

the image feature map, and c denotes the feature dimension. Likewise, T_P is the feature embedding of the prompts where $T_P \in \mathbb{R}^{k \times c}$, k is the length of the prompts.

Following this, the encoded image and prompts are supplied to the decoder, called $\mathrm{Dec_M}$, which employs attention-based mechanisms for feature interaction. SAM prepares the input tokens for the decoder by merging several mask tokens, denoted as T_M, with the prompt tokens T_P. These mask tokens play a crucial role in generating the mask output, which is defined as:

$$S = \mathrm{Dec_M}\left(F_I, \mathrm{Concat}(T_M, T_P)\right), \tag{2}$$

where S in equation 2 represents the output segmentation mask predicted by SAM.

3.2. GeoSAM: Training Strategy

3.2.1 Prompts Generation

In the proposed framework, we provide the model with both sparse and dense prompts to offer contextual information for the segmentation task.

Sparse Prompts Generation In the sparse prompts generation process, we implement a system that creates random click points referred to as sparse prompts, on both the foreground and background of the object of interest. We use a pre-trained model, Tile2Net [24], to classify pixels into foreground and background, selecting points for foreground and background clicks as sparse prompts.

Tile2Net generates a multi-class segmentation map containing four distinct labels, including three foreground classes and one background class. For the sake of clarity, we refer to this segmentation map produced by Tile2Net as the "pseudo label". However, this pseudo label differs from our binary class segmentation map, where there are only two classes: foreground and background. To this end, we make specific adjustments to adapt this pseudo label for our task. In both road and pedestrian segmentation tasks, we classify specific pixel values as foreground (road infrastructure for roads, and sidewalk and crosswalk for pedestrians), while the rest are considered background, effectively turning the pseudo labels into binary segmentation maps.

From these selected sets of pixels, we randomly select 2000 foreground and 1000 background pixels and obtain their corresponding pixel coordinates, which are subsequently utilized as sparse prompts for the prompt encoder. We experimented with different numbers of points and found that a ratio of 2:1 for foreground and background points works particularly well (refer to the ablation study in Table 1). Figure 2 visually outlines the random sparse prompts generation process using the pseudo labels generated by Tile2Net (for illustration purposes, we reduce the number of points).

Dense Prompts Generation In addition to the sparse prompts, we have also designed another automated system to supply the model with dense prompts, serving as an additional means to define the context effectively. The dense prompts can be regarded as a soft mask i.e. an unthresholded prediction map of lower quality. The soft mask produced by zero-shot SAM is unthresholded, meaning it provides continuous values representing the model's confidence in pixel assignments. This unthresholded mask retains more detailed semantic information compared to a simplified binary mask that arises from applying a threshold.

The process of creating dense prompts starts by providing the model with the input image to be segmented and its corresponding feature embeddings (generated from the pre-trained CNN image encoder), which the decoder in zero-shot SAM in turn uses to generate dense prompts. Primarily, these embeddings act as the initial dense prompts for the SAM's prompt encoder (illustrated in Figure 1). Normally, SAM's prompt encoder takes dense prompts in the shape of (64×64×256), where 256 represents the channel dimension and (64×64) represents the height and width, respectively. To emulate the extracted feature embeddings as dense prompts, we primitively resize them to this format. We do this by applying a 1×1 convolution to map the channel

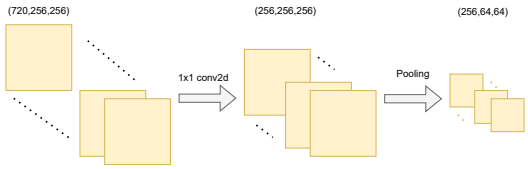

Figure 3. Resizing the feature embeddings of the pre-trained CNN encoder to generate the dense prompts for SAM.

dimensions of feature embeddings to 256-dimensional channels and resizing the height and width using average pooling to 64×64 (this resizing process is illustrated in Figure 3). The obtained embeddings can now be used as inputs into the prompt encoder. Finally, using the outputs produced by the image encoder and prompt encoder, the decoder generates the output i.e. soft mask, effectively creating self-generated dense prompts for the model.

3.2.2 Fine-Tuning

In this part of the training, we concurrently provide the model with both dense prompts and sparse prompts to the prompt encoder that were generated earlier. Subsequently, the SAM mask decoder generates a segmentation map based on the prompts and the input image. At this stage, we compute the loss between the output segmentation map and the ground truth. We then begin updating the decoder's parameters while keeping the other parameters fixed, starting the fine-tuning process using PEFT. It's important to highlight that we didn't make any modifications to the model's parameters before this point, following the PEFT strategy. The fine-tuning process occurs here, where we adjust the decoder's parameters.

For the loss function to fine-tune the decoder, we opt for a combination of Dice Loss [35] and Focal Loss [54]. The Dice Loss is based on the Dice Similarity Coefficient (DSC), a popular metric using the overlap between two regions for evaluating the accuracy of a segmentation algorithm. The Dice Loss can be represented as the complement of the dice coefficient metric, therefore, minimizing the Dice Loss during training is equivalent to maximizing the Dice Coefficient. The equation can be expressed as

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{c=1}^{C}\sum_{i=1}^{N} s_i^c g_i^c}{\sum_{c=1}^{C}\sum_{i=1}^{N} s_i^c + \sum_{c=1}^{C}\sum_{i=1}^{N} g_i^c}, \quad (3)$$

where g_i^c represents the ground truth binary indicator of class label c for the pixel i, and s_i^c denotes the corresponding predicted segmentation probability.

Focal Loss is a weighted Cross Entropy Loss designed to focus on hard-to-classify examples while down-weighting easy-to-classify examples. In addition to the previous notations, we also introduce α and γ as the balancing and focusing parameters, respectively. The balancing factor α assigns different weights to different classes to provide more importance to the minority class whereas the focusing parameter γ affects how much the loss is focused on hard-to-classify examples [34]. Parameters α and γ are combined with the basic Cross Entropy Loss to get the equation of Focal Loss:

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \alpha\left[(1-s_i^c)^{\gamma} g_i^c \log(s_i^c)\right]. \quad (4)$$

Finally, we combine both losses to get the overall Dice Focal Loss used for fine-tuning the GeoSAM.

$$\mathcal{L}_{\text{DiceFocal}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{Focal}}. \quad (5)$$

To assess the performance of geographical object segmentation, we employ the Intersection over Union (IoU), also known as the Jaccard index. This widely used region-wise evaluation metric considers the regions of overlap and union between predicted and ground truth segmentations. The IoU score is a ratio of the area of overlap between the predicted and ground truth regions to the area of union. The value 0 indicates no overlap, while a value of 1 indicates perfect overlap between the predicted and ground truth regions.

3.3. The End-to-End Inference Pipeline

During the inference stage, we utilize the fine-tuned decoder obtained during training, along with sparse prompts and dense prompts, to automatically generate segmentation maps from input geographical images. This end-to-end pipeline mirrors the approach used during training, where the pre-trained CNN encoder provides sparse prompts, and zero-shot SAM contributes dense prompts. With the fine-tuned decoder, GeoSAM processes these prompts to generate the final segmentation map output. The model also undergoes postprocessing on these generated maps to refine them for more practical use (more on section 4.3.1). This inference pipeline ensures that the model is capable of segmenting pedestrian infrastructure in geographical images without any human intervention.

4. Experiments

4.1. Datasets

For fine-tuning and inference purposes, we create the training and testing dataset of geographical imagery from separate regions of Washington DC. We denote the training dataset as $D_{\text{train}} = \{S, G_{\text{road}}, G_{\text{ped}}\}$, where $S = \{s_1, ..., s_n\}$, $G_{\text{road}} = \{g_1^{\text{road}}, ..., g_n^{\text{road}}\}$ and $G_{\text{ped}} = \{g_1^{\text{ped}}, ..., g_n^{\text{ped}}\}$ correspond to the n sample of images and segmentation ground truth masks of road infrastructure and pedestrian infrastructure respectively. Similarly, we denote the test dataset as $D_{test} = \{\hat{S}, \hat{G}_{\text{road}}, \hat{G}_{\text{ped}}\}$.

We choose Washington DC as our primary location due to the easy accessibility of GIS data sources[9, 10], as it is the

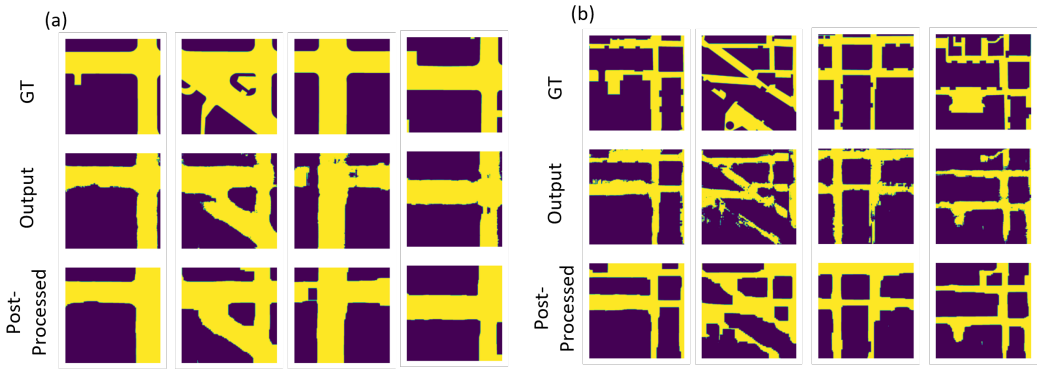

Figure 4. Postprocessing operations on the two tasks. Each of the columns represents a single randomly picked image from the testing dataset with two tasks: (a) road infrastructure segmentation, and (b) pedestrian infrastructure segmentation. Here, GT = ground truth.

only accessible source in the USA. For our geographical image analysis, we utilize orthorectified imagery sourced from aerial images. At first, we obtain high-resolution orthorectified imagery (corrected aerial imagery) and subsequently generate annotation labels or mask images using the GIS data. These orthorectified images can be obtained through the U.S. Geographical Survey (USGS) [57], an agency studying and mapping the Earth's natural resources and geological features.

We employ the pipeline from [24] for downloading orthorectified image tiles based on specific geographical coordinates. We utilize Google Earth [17] to define the region bounded by coordinates 38°53'48.6"N 77°00'26.7"W and 38°54'25.1"N 76°59'20.2"W, approximately encompassing an area of approximately 2.31 square kilometers. We use a zoom level of 20 for downloading the image tiles, producing 256×256 pixel tiles (note that at zoom level 0, the entire world is represented as a single tile). A total of 2240 image tiles are selected representing the whole region. Following the procedure detailed in [24], we stitch these tiles into larger 1024×1024 resolution tiles, reducing the number of training image tiles to 560. To generate the corresponding ground truth images, we operate on the same coordinates and its affiliated GIS data to create equivalent two sets of 560 ground truth images, one containing road infrastructure and the other containing pedestrian (sidewalk/crosswalk) infrastructure.

We follow the same methodology as used for the training dataset to construct the test dataset, which covers a non-overlapping region to the training dataset (coordinates 38°54'34.9"N 77°01'10.9"W and 38°54'20.8"N 77°02'42.1"W) in Washington DC. Employing identical settings as before, we acquired 296 stitched tiles and corresponding masks.

4.2. Implementation Details

As mentioned earlier we treat our objective as two distinct single-task binary class segmentation tasks, road infrastructure and pedestrian infrastructure segmentation, where the objects of interest are roads and sidewalks/crosswalks respectively. We adopted ViT-H [14] 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 [29], the choice of optimizer was the AdamW ($\beta 1$ =0.9, $\beta 2$ = 0.999) optimizer [39], with an initial learning rate set at e^{-5} and weight decay of 0.1 and no data augmentation techniques were applied. To have an adaptable learning rate, we also employed a cosine annealing learning rate scheduler with a maximum learning rate decaying smoothly to a minimum value (e^{-7}) over the course of training. As the pre-trained image encoder, we opt for Tile2Net's semantic segmentation component, specifically the Hierarchical Multi-Scale Attention model, and initialized the model with publicly available pre-trained weights released from the Tile2Net team [24].

All the experiments are done on an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory with the Python version for the project is 3.10.9. We use a total of 100 epochs to train our model as well as the other baseline models. While doing the training for GeoSAM, we follow PEFT techniques where we keep the parameters of the encoder part (both image and prompt encoder) frozen and only update the gradients of the decoder. The source code will be released upon publication.

4.3. Results

4.3.1 Qualitative Results

We implement postprocessing techniques after the initial segmentation maps are generated from the SAM mask decoder. To further improve results, techniques such as morphological erosion and dilation [44] are utilized to enhance performance. In our task, the priority is on path creation rather than the classification of individual pixels. We place greater emphasis on identifying and delineating all available paths.

In the generated segmentation maps, which represent real-world pedestrian paths, the existence of scattered isolated

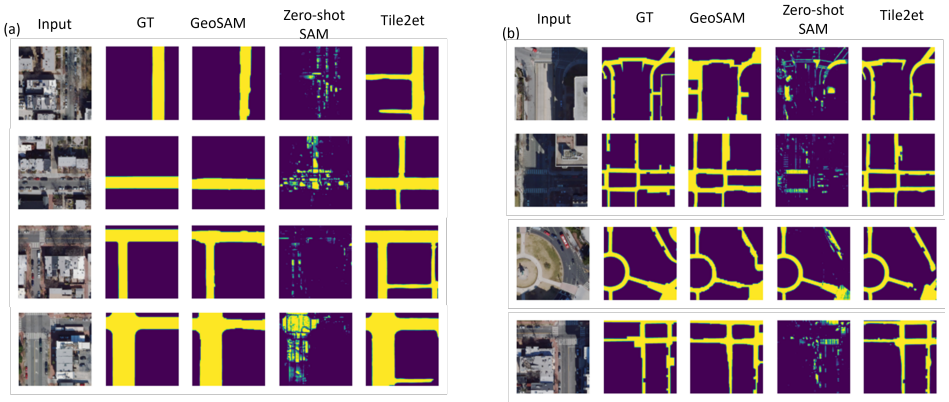

Figure 5. Qualitative segmentation results of GeoSAM on the two tasks compared to Tile2Net and zero-shot SAM. Each of the rows represents a single randomly picked image from the testing dataset with two tasks: (a) road infrastructure segmentation, and (b) pedestrian infrastructure segmentation. Here, GT = ground truth.

Method	Road Infras.	Pedestrian Infras.	Mean
GeoSAM (Ours)	**0.72**	**0.48**	**0.60**
Zero-shot SAM [29]	0.25	0.16	0.20
Tile2Net [24]	0.60	0.42	0.51
UNet [49]	0.32	0.15	0.23
AttUNet [41]	0.28	0.18	0.23
UNet++ [67]	0.63	0.35	0.49
UNETR [21]	0.48	0.20	0.34
Swin UNETR [20]	0.45	0.26	0.35

Table 1. Evaluation results of GeoSAM in IoU compared to other works, Infras. = infrastructure. Results with the best performance are boldfaced.

regions indicates inaccuracies in the model's performance. To rectify this issue, we employ an erosion technique to eliminate such isolated regions. Additionally, when the segmentation map displays abrupt interruptions or gaps in connected paths, it indicates failure in accurately segmenting the entire path. Therefore, we utilize dilation to address these issues by establishing connections between disjointed regions, bringing the map closer to the ground truth representation.

The equations for erosion and dilation are described below where (x, y) is a pixel in the image, $B(i, j)$ is the structuring element or the mask to do the operation, (i, j) are the coordinates within the structuring element, $\bigcap$ is the intersection and $\bigcup$ is the union operation:

$$E(x,y) = \bigcap_{(i,j)\in B} I(x+i, y+j), \tag{6}$$

$$D(x,y) = \bigcup_{(i,j)\in B} I(x+i, y+j). \tag{7}$$

To do these operations, we select a (1010) filter in a (1024×1024) resolution segmentation map. This filter is passed over the whole segmentation map and performs erosion and dilation respectively in the regions it covers. We run these operations for a total of 10 iterations to get a better-refined map. Figure 4 illustrates how these postprocessing techniques have been performed to improve performance. For instance, you can observe the removal of isolated regions and a connection has been established where a path should be. We plot both the initial output and the postprocessing output to distinguish the difference in quality.

Figure 5 provides a visual representation of the qualitative outcomes achieved by GeoSAM on several images from the test dataset. Given the objective of performing two distinct binary class segmentation tasks, we have included visualizations for both tasks in the figure. Upon examination of the figure, it becomes evident that GeoSAM's performance is on par with Tile2Net. Additionally, GeoSAM significantly outperforms zero-shot SAM when compared to zero-shot SAM's capacity to handle similar tasks. This observation underscores the earlier assertion in Section 1 that zero-shot SAM does not exhibit strong performance for the thin boundary objects when applied to geographical images.

4.3.2 Quantitative Results

Next, we have some quantitative results based on the evaluation metric we discussed in Section 3.2.2. In Table 1 we demonstrate and compare the detailed results obtained by using GeoSAM with Tile2Net, zero-shot SAM, and some of the more popular semantic segmentation models (both CNN-based and ViT-based as benchmarks). We pick UNet [49], AttUNet [41], and UNet++ [67] as CNN-based benchmark models to compare with. For ViT-based models, among popular benchmark models such as [20, 21, 31, 38], we have selected [20, 21] for comparison.

To compare GeoSAM with the popular benchmark semantic segmentation models (both CNN-based and ViT-based) in Table 1, we train each of the models from scratch in their de-

Techniques				Sparse Prompt			IOU	
Zero-shot	DP	PE	FT	For. Points	Back. Points	Ratio	Road Infras.	Pedestrian Infras.
✓	✓	✓	✓	0	0	-	0.01	0.01
✓	✓	✓	✓	100	50	2:1	0.69	0.44
✓	✓	✓	✓	2000	2000	1:1	0.66	0.46
✓	✓	✓	✓	**2000**	**1000**	**2:1**	**0.72**	**0.48**
✓	✓	✓	✓	2000	4000	1:2	0.57	0.43
✓	×	×	×	2000	1000	2:1	0.06	0.04
✓	×	×	✓	2000	1000	2:1	0.68	0.41
✓	✓	✓	×	2000	1000	2:1	0.35	0.22
✓	✓	×	✓	2000	1000	2:1	0.70	0.44

Table 2. An ablation study. The segmentation performance is examined across the varying number of sparse point prompts and various model components. DP = dense prompt, PE = pre-trained encoder, FT = fine-tuning. The white and gray sections represent two distinct types of ablation studies.

fault settings using the training dataset (described in section 4.1). Following the settings we train the GeoSAM model using Monai, an open-source framework [4] (without any preprocessing and post-processing for a fair comparison). In the inference stage, we assess these models using the dedicated test dataset (described in section 4.1) to obtain the IoU values.

The results in Table 1 demonstrate that GeoSAM outperforms Tile2Net by a margin of 17% in terms of mean IoU. Compared to zero-shot SAM (aided with only automated sparse prompts), GeoSAM performs much better across both tasks (a 200% increase in mean IoU). In comparison with the other benchmark segmentation models, GeoSAM has demonstrated its superiority by significantly outperforming these models across both tasks. We observe that GeoSAM surpasses the top-performing CNN-based model (UNet++) in mIoU by a significant margin of 22% and outperforms the best ViT-based model (UNETR) by approximately 71%. Although [5] is relevant to our work, especially in using foundation models for geographical images, the unavailability of their source code hinders a direct comparison. Instead, we reference the highest IoU value for road segmentation from their results (0.59), which is surpassed by GeoSAM with an IoU of 0.72. Importantly, [5] lacks results on sidewalk/crosswalk segmentation, our primary focus is on mobility infrastructure segmentation. Additional results, including some other evaluation metrics, are provided in the appendix.

Table 2 reports some of the ablation studies we performed for our approach. We split the ablation study into two parts. For the first part of the ablation study, we examine the effect of the number of sparse prompts on the performance of GeoSAM. We tried different ratios of foreground and background points and also we ran GeoSAM without any points as well to compare the performance. The results strongly support the assertion that sparse prompts play a critical role, as demonstrated by GeoSAM's significant drop in performance when sparse prompts are omitted. Further, from analyzing the data, it can be inferred that the initial assumption, which involved employing a ratio of 2:1 and selecting 2000:1000 foreground-to-background points, results in the best performance for this specific case. As a result, this configuration becomes the default setting.

In the second part of the ablation study, as shown in Table 2, we conducted an ablation study to assess the importance of several key techniques, as outlined in Table 2. These key components include the significance of the CNN Encoder for generating dense prompts, the role of dense prompts in enhancing sparse prompts, and the importance of fine-tuning the decoder over the original decoder. Utilizing the optimal settings established in the previous section, we assessed the significance of the individual techniques opted for within GeoSAM. The results from Table 2 indicate that each component integrated into GeoSAM contributes uniquely to enhancing the overall performance. Significantly, the exclusion of FT (fine-tuned decoder) in the last row results in the most significant drop in performance, underscoring that the original SAM decoder is not well-suited for this specific type of task. Additionally, without the other techniques, only automated sparse prompts given to zero-shot SAM (first row) resulted in very poor performance, highlighting the importance of each of the techniques we utilized.

5. Conclusion

GeoSAM is a pioneering work that uses SAM's capabilities for pedestrian infrastructure within geographical images, without human intervention. We introduce an innovative architecture of utilizing both sparse prompts and dense prompts with the addition of fine-tuning for the foundation segmentation model SAM. In addition, different from other existing work, the training and end-to-end inference pipeline for mobility infrastructure segmentation that we developed here are reproducible can be adapted for various segmentation tasks, and are transferable to other geo-locations by using a different domain-specific encoder and fine-tuning the decoder with a different type of dataset.

References

[1] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021.

[2] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*, 2015.

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[4] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[5] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*, 2023.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[7] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 2023.

[8] Xin Chen, Qun Sun, Wenyue Guo, Chunping Qiu, and Anzhu Yu. Ga-net: A geometry prior assisted neural network for road extraction. *International Journal of Applied Earth Observation and Geoinformation*, 114:103004, 2022.

[9] DC GIS. Roads 2019. https://opendata.dc.gov/datasets/DCGIS::roads-2019, 2019.

[10] DC GIS. Sidewalks 2019. https://opendata.dc.gov/datasets/sidewalks-2019, 2019.

[11] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. *arXiv preprint arXiv:2303.11851*, 2023.

[12] Lei Ding, Kun Zhu, Daifeng Peng, Hao Tang, and Haitao Guo. Adapting segment anything model for change detection in hr remote sensing images. *arXiv preprint arXiv:2309.01429*, 2023.

[13] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12799–12807, 2023.

[16] Iraklis Giannakis, Anshuman Bhardwaj, Lydia Sam, and Georgios Leontidis. Deep learning universal crater detection using segment anything model (sam). *arXiv preprint arXiv:2304.07764*, 2023.

[17] GoogleEarth contributors. Google Earth. https://earth.google.com/web/, 2017.

[18] Povilas Gudžius, Olga Kurasova, Vytenis Darulis, and Ernestas Filatovas. Deep learning-based object recognition in multispectral satellite imagery for real-time applications. *Machine Vision and Applications*, 32(4):98, 2021.

[19] Povilas Gudzius, Olga Kurasova, Vytenis Darulis, and Ernestas Filatovas. Automl-based neural architecture search for object recognition in satellite imagery. *Remote Sensing*, 15 (1):91, 2022.

[20] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[21] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang 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.

[22] Corentin Henry, Seyed Majid Azimi, and Nina Merkle. Road segmentation in sar satellite images with deep fully convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(12):1867–1871, 2018.

[23] Maryam Hosseini, Mikey Saugstad, Fabio Miranda, Andres Sevtsuk, Claudio T Silva, and Jon E Froehlich. Towards global-scale crowd+ ai techniques to map and assess sidewalks for people with disabilities. *arXiv preprint arXiv:2206.13677*, 2022.

[24] Maryam Hosseini, Andres Sevtsuk, Fabio Miranda, Roberto M Cesar Jr, and Claudio 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.

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[26] Xinrong Hu, Xiaowei Xu, and Yiyu Shi. How to efficiently adapt large segmentation model (sam) to medical images. *arXiv preprint arXiv:2306.13731*, 2023.

[27] Sahib Julka and Michael Granitzer. Knowledge distillation with segment anything (sam) model for planetary geological mapping. *arXiv preprint arXiv:2305.07586*, 2023.

[28] Jae Hong Kim, Sugie Lee, John R Hipp, and Donghwan Ki. Decoding urban landscapes: Google street view and measurement sensitivity. *Computers, Environment and Urban Systems*, 88:101626, 2021.

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[30] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[31] Chengyin Li, Hassan Bagher-Ebadian, Vikram Goddla, Indrin J Chetty, and Dongxiao Zhu. Focalunetr: A focal transformer for boundary-aware segmentation of ct images. *arXiv preprint arXiv:2210.03189*, 2022.

[32] Chengyin Li, Zheng Dong, Nathan Fisher, and Dongxiao Zhu. Coupling user preference with external rewards to enable driver-centered and resource-aware ev charging recommendation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 3–19. Springer, 2022.

[33] Chengyin Li, Prashant Khanduri, Yao Qiang, Rafi Ibn Sultan, Indrin Chetty, and Dongxiao Zhu. Auto-prompting sam for mobile friendly 3d medical image segmentation. *arXiv preprint arXiv:2308.14936*, 2023.

[34] Xiangrui Li, Xin Li, Deng Pan, and Dongxiao Zhu. On the learning property of logistic and softmax losses for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4739–4746, 2020.

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[36] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *arXiv preprint arXiv:2306.11029*, 2023.

[37] Shiqin Liu, Carl Higgs, Jonathan Arundel, Geoff Boeing, Nicholas Cerdera, David Moctezuma, Ester Cerin, Deepti Adlakha, Melanie Lowe, and Billie Giles-Corti. A generalized framework for measuring pedestrian accessibility around the world using open data. *Geographical Analysis*, 54(3):559–582, 2022.

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[40] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.

[41] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[42] Lucas Prado Osco, Qiusheng Wu, Eduardo Lopes de Lemos, Wesley Nunes Gonçalves, Ana Paula Marques Ramos, Jonathan Li, and José Marcato Junior. The segment anything model (sam) for remote sensing applications: From zero to one shot. *arXiv preprint arXiv:2306.16623*, 2023.

[43] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Xiaokang Yang, and Wei Shen. Sam-parser: Fine-tuning sam efficiently by parameter space reconstruction. *arXiv preprint arXiv:2308.14604*, 2023.

[44] Martino Pesaresi and Jon Atli Benediktsson. A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE transactions on Geoscience and Remote Sensing*, 39(2):309–320, 2001.

[45] Yao Qiang, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu. Fairness-aware vision transformer via debiased self-attention. *arXiv preprint arXiv:2301.13803*, 2023.

[46] Yao Qiang, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu. Interpretability-aware vision transformer. *arXiv preprint arXiv:2309.08035*, 2023.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

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

[51] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, et al. Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[52] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1422–1431, 2022.

[53] Tal Shaharabany, Aviad Dahan, Raja Giryes, and Lior Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder. *arXiv preprint arXiv:2306.06370*, 2023.

[54] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.

[55] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

[56] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

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

[58] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022.

[59] Di Wang, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Scaling-up remote sensing segmentation dataset with segment anything model. *arXiv preprint arXiv:2305.02034*, 2023.

[60] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[61] Zhe Wang, Shoukun Sun, Xiang Que, and Xiaogang Ma. Interactive segmentation in aerial images: a new benchmark and an open access web-based tool. *arXiv preprint arXiv:2308.13174*, 2023.

[62] Liangliang Yao, Haobo Zuo, Guangze Zheng, Changhong Fu, and Jia Pan. Sam-da: Uav tracks anything at night with sam-powered domain adaptation. *arXiv preprint arXiv:2307.01024*, 2023.

[63] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020.

[64] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023.

[65] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[66] Lichen Zhou, Chuang Zhang, and Ming 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.

[67] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming 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: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.