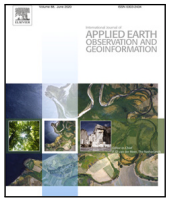
International Journal of Applied Earth Observation and Geoinformation 124 (2023) 103540

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



The Segment Anything Model (SAM) for remote sensing applications: From zero to one shot

Lucas Prado Osco [a,*], Qiusheng Wu [b], Eduardo Lopes de Lemos [c], Wesley Nunes Gonçalves [c], Ana Paula Marques Ramos [d], Jonathan Li [e], José Marcato Junior [c]

[a] University of Western São Paulo (UNOESTE), Rod. Raposo Tavares, km 572, Limoeiro, Presidente Prudente, 19067-175, Brazil
[b] University of Tennessee (UT), 1331 Circle Park Drive, Knoxville, 37996-0925, United States
[c] Federal University of Mato Grosso do Sul (UFMS), Av. Costa e Silva-Pioneiros, Cidade Universitária, Campo Grande, 79070-900, Brazil
[d] São Paulo State University (UNESP), Centro Educacional, R. Roberto Simonsen, 305, Presidente Prudente, 19060-900, Brazil
[e] University of Waterloo (UW), 200 University Avenue West, Waterloo, N2L 3G1, Canada

ARTICLE INFO

Dataset link: GitHub: AI-RemoteSensing, GitHub: Segment-Geospatial

Keywords:
Artificial intelligence
Image segmentation
Multi-scale datasets
Text-prompt technique

ABSTRACT

Segmentation is an essential step for remote sensing image processing. This study aims to advance the application of the Segment Anything Model (SAM), an innovative image segmentation model by Meta AI, in the field of remote sensing image analysis. SAM is known for its exceptional generalization capabilities and zero-shot learning, making it a promising approach to processing aerial and orbital images from diverse geographical contexts. Our exploration involved testing SAM across multi-scale datasets using various input prompts, such as bounding boxes, individual points, and text descriptors. To enhance the model's performance, we implemented a novel automated technique that combines a text-prompt-derived general example with one-shot training. This adjustment resulted in an improvement in accuracy, underscoring SAM's potential for deployment in remote sensing imagery and reducing the need for manual annotation. Despite the limitations, encountered with lower spatial resolution images, SAM exhibits promising adaptability to remote sensing data analysis. We recommend future research to enhance the model's proficiency through integration with supplementary fine-tuning techniques and other networks. Furthermore, we provide the open-source code of our modifications on online repositories, encouraging further and broader adaptations of SAM to the remote sensing domain.

1. Introduction

The field of remote sensing deals with capturing images of the Earth's surface from airborne or satellite sensors. Analyzing these images allows us to monitor environmental changes, manage disasters, and plan urban areas efficiently (Gómez et al., 2016; Song et al., 2023; Yuan et al., 2020). A critical part of this analysis is the ability to accurately identify and segment various objects or regions within these images, a process known as image segmentation. Segmentation allows us to isolate specific objects or areas within an image for further study or monitoring (Kotaridis and Lazaridou, 2021). Traditional segmentation techniques often require extensive human input and intervention for accurate results. However, with the advent of advanced artificial intelligence (AI) and deep learning methods (Bai et al., 2022; Aleissaee et al., 2023), the segmentation process has become more automated, albeit still facing challenges, particularly in the effective segmentation of images with minimal human input.

The Segment Anything Model (SAM), developed by Meta AI, is a groundbreaking approach to image segmentation that has demonstrated exceptional generalization capabilities across a diverse range of image datasets, requiring no additional training for unfamiliar objects (Kirillov et al., 2023). This approach enables it to make accurate predictions with little to no training data. However, its potential can be limited when facing specific domain conditions. To overcome this limitation, SAM can be modified by a re-learning approach (Zhang et al., 2023b), feeding it with a single example of a new class or object for better results.

Zero-shot learning pertains to a model's capability to accurately process and act upon input data that it has not explicitly encountered during training (Alayrac et al., 2022; Sun et al., 2021). This ability is derived from gaining a generalized understanding of the data rather than specific instances. Zero-shot learning systems can recognize objects or understand tasks they have never seen before based on learning

* Corresponding author.
E-mail address: lucasosco@unoeste.br (L.P. Osco).

https://doi.org/10.1016/j.jag.2023.103540
Received 5 July 2023; Received in revised form 19 October 2023; Accepted 26 October 2023
Available online 1 November 2023

underlying concepts or relationships. In contrast, one-shot learning denotes a model's ability to interpret and make accurate inferences from just a single example of a new class (Zhang et al., 2023b). By feeding SAM with a single example (or 'shot') of this new class, we can potentially enhance its performance, as it has more specific information to work with.

The best-known one-shot methods for SAM are named PerSAM and PerSAM-F, both being training-free personalization approaches (Zhang et al., 2023b). Given a single image with a reference mask, PerSAM localizes the target concept using a location prior to an initial estimate of where the object of interest is likely to be. The second method is PerSAM-F, a variant of PerSAM that uses one-shot fine-tuning to reduce mask ambiguity. In this case, the entire SAM is frozen (i.e., its parameters are not updated during the fine-tuning process), and two learnable weights are introduced for multi-scale masks. This one-shot fine-tuning variant requires training only two parameters and can be done in as little as ten seconds to enhance performance (Zhang et al., 2023b). Both are capable of improving SAM, making it a flexible model.

Another important aspect relates to SAM's ability to perform segmentation with minimal input, requiring only a bounding box or a single point as a reference, or even a prompt text as guidance (Kirillov et al., 2023). This capability has the potential to reduce human labor during the annotation process. Many existing techniques require intensive annotations for each new object of interest, resulting in significant computational overhead and potential delays in time-sensitive applications. SAM, on the other hand, presents an opportunity to alleviate this time-intensive task.

Since SAM's release in April 2023, the geospatial community has shown strong interest in adapting SAM for remote sensing image segmentation. However, a more in-depth investigation is needed. In this context, we present a first-of-its-kind evaluation of SAM, developing both its zero and one-shot learning performance on segmenting remote sensing imagery. We adapted SAM to our data structure, benchmarked it against multiple datasets, and assessed its potential to segment multiscale images. We then evolved SAM's zero-shot characteristic to a one-shot approach and demonstrated that with only one example of a new class, SAM's segmentation performance can be significantly improved.

Our proposal's innovation is within the one-shot technique, which involves using a prompt-text-based segmentation as a training sample (instead of a human-labeled sample), making it an automated process for refining SAM on remote sensing imagery. In this study, we also discuss the implications, limitations, and potential future directions of our findings. Understanding the effectiveness of SAM in this domain is of paramount importance for novel development. In short, with its promise of zero-shot and one-shot learning, SAM has the potential to transform current practices by significantly reducing the time and resources needed for training and annotating data, thereby enabling a quicker, more efficient approach.

2. Remote sensing image segmentation: A brief summary

The remote sensing field has experienced impressive advancements in recent years, largely driven by improvements in aerial and orbital platform technologies, sensor capabilities, and computational resources (Toth and Jóźków, 2016; Osco et al., 2021a). One of the most critical tasks in remote sensing is image segmentation, which involves partitioning images into multiple segments or regions, each, ideally, corresponding to a specific object or class (Kotaridis and Lazaridou, 2021). In this section, we focus on providing comprehensive information regarding segmentation processes, deep learning-based methods, and techniques, and explain the overall importance of conducting zero-to-one shot learning.

Traditional image segmentation techniques in remote sensing often rely on pixel-based or object-based approaches. Pixel-based methods, such as clustering and thresholding, involve grouping pixels with similar characteristics, while object-based techniques focus on segmenting images based on properties of larger regions or objects (Hossain and Chen, 2019; Wang et al., 2020b). However, these methods can be limited in their ability to handle the complexity, variability, and high spatial resolution of modern remote sensing imagery (Kotaridis and Lazaridou, 2021).

Segmentation involves various methods designed to separate or group portions of an image based on certain criteria (Zhang et al., 2021). Each method has a unique approach and application. Interactive Segmentation, for example, is a niche within image segmentation that actively incorporates user input to improve the segmentation process, making it more precise and tailored to specific requirements (Li et al., 2020; Wu et al., 2021). Different interactive segmentation methods utilize various strategies to include human intelligence in the loop. This makes interactive segmentation particularly useful in tasks where high precision is required, and generic segmentation methods may not suffice.

Super Pixelization is another method that groups pixels in an image into larger units, or "superpixels", based on shared characteristics such as color or texture (Gharibbafghi et al., 2018). This grouping can simplify the image data while preserving the essential structure of the objects. Object Proposal Generation goes a step further by suggesting potential object bounding boxes or regions within an image (Hossain and Chen, 2019; Su et al., 2019). These proposals serve as a guide for a more advanced model to identify and classify the actual objects' pixels. Foreground Segmentation, also known as background subtraction, is a technique primarily used to separate the main subjects or objects of interest (the foreground) from the backdrop (the background) in an image (Zheng et al., 2020; Ma et al., 2022).

Semantic Segmentation is a more comprehensive approach where every pixel in an image is assigned to a specific class, effectively grouping regions of the image based on semantic interest (Zhang et al., 2020; Adam et al., 2023). Instance Segmentation identifies each pixel recognizes distinct objects of the same class and recognizes the individual objects as separate entities or instances (Gao et al., 2021; Qurratulain et al., 2023). Panoptic Segmentation merges the concepts of semantic and instance segmentation, assigning every pixel in the image a class label and a unique instance identifier (Hua et al., 2021; de Carvalho et al., 2022). This method aims to give a complete understanding of the image by identifying and classifying every detail.

All these methods have been intensively studied, but one that surged in recent years, with the advancements of Visual Foundation Models (VFM) and Large Multimodal Models (LMM), is known as "Promptable Segmentation", an approach that aims to create a versatile model capable of adapting to a variety of segmentation tasks (Mialon et al., 2023; Zhang et al., 2023a). This is achieved through "prompt engineering", where prompts are carefully designed to guide the model toward generating the desired output (Lobry et al., 2020; Sun et al., 2021). This concept is a departure from traditional multi-task systems where a single model is trained to perform a fixed set of tasks. The unique feature of a promptable segmentation model is its ability to take on new tasks at the time of inference, serving as a component in a larger system (Sun et al., 2021; Mialon et al., 2023). For instance, to perform instance segmentation, a promptable segmentation model could be combined with an existing object detector.

Object detection is a crucial task in computer vision, focusing on identifying and locating objects within images. This task is foundational for various applications such as surveillance, autonomous vehicles, and many others. In the realm of object detection and image segmentation, different techniques have been employed. Traditional methods often focus on detecting objects that the model has been specifically trained on, known as closed-set detection. However, real-world applications demand more flexibility and the ability to detect and classify objects not seen during training, known as open-set detection.

One state-of-the-art open-set object detector that stands out is Grounding DINO (GroundDINO), an enhanced transformer-based object detector capable of identifying a broader range of objects based on various human inputs (Liu et al., 2023b). This system is an enhancement of the Transformer-based object detector called DINO (Zhang et al., 2022a), enriched with grounded pre-training to be able to identify a broader range of objects based on human inputs, such as category names or referring expressions. An open-set detector is meant to identify and classify objects that were not part of the model's training data, as opposed to a closed-set detector that can only recognize objects it has been specifically trained on. The information from Grounding DINO can potentially be used to guide the segmentation process, providing class labels or object boundaries that the segmentation model could use.

Most NLMs incorporate deep-learning-based networks and, with the rise of these methods, more advanced segmentation techniques have been developed for remote sensing applications. Convolutional Neural Networks (CNNs), which emerged as a popular choice due to their ability to capture local and hierarchical patterns in images (Martins et al., 2021; Bressan et al., 2022), have widely been used as the backbone for these tasks. CNNs consist of multiple convolutional layers that apply filters to learn increasingly complex features, making them well-suited for segmenting objects in many remote sensing images (Yuan et al., 2021; Bai et al., 2022). However, they are computationally intensive and may require substantial training data.

Generative Adversarial Networks (GANs) have also shown potential in the field of image processing. GANs consist of a generator and a discriminator network, where the generator tries to create synthetic data to fool the discriminator, and the discriminator aims to distinguish between real and synthetic data (Jozdani et al., 2022). For image segmentation, GANs can be used to generate realistic images and their corresponding segmentations, which can supplement the training data and improve the robustness of the segmentation models (Benjdira et al., 2019).

Vision Transformer (ViT), on the other hand, is a recent development in deep learning that has shown promise in image segmentation tasks. Unlike CNNs, which rely on convolutional operations, ViT employs self-attention mechanisms that allow it to model long-range dependencies and global context within images (Li et al., 2023b,a). This approach has demonstrated competitive performance in various computer vision tasks, including remote sensing image segmentation (Aleissaee et al., 2023), and it is currently outperforming CNNs in remote sensing data (Gonçalves et al., 2023).

Another capability of deep learning that can enhance the segmentation process is transfer learning. With it, a model pre-trained on a large dataset is adapted for a different but related task (Tong et al., 2020). For instance, a CNN or ViTr trained on a large-scale image recognition dataset like ImageNet can be fine-tuned for the task of remote sensing image segmentation (Osco et al., 2020, 2021b). The advantage of transfer learning is that it can leverage the knowledge gained from the initial task to improve performance on the new task, especially when the amount of labeled data for the new task is limited.

One of the main challenges in applying deep learning techniques to remote sensing image segmentation is the need for large volumes of labeled ground-truth data (Chi et al., 2016). Acquiring and annotating this data can be time-consuming and labor-intensive, requiring expert knowledge and resources that may not be readily available. Furthermore, the variability and complexity of remote sensing imagery can make the labeling process even more difficult (Amani et al., 2020). As such, it becomes imperative to develop robust, efficient, and accessible solutions that can aid in the processing and analysis of such data. A model that can perform segmentation with zero domain-specific information may offer an important advantage for this process.

In this sense, the Segment Anything Model (SAM) has emerged as a potential tool for assisting in the segmentation process of remote sensing images. SAM design enables it to generalize to new image distributions and tasks effectively and already resulted in numerous applications (Kirillov et al., 2023). By using minimal human input, such as bounding boxes, reference points, or simply text-based prompts, SAM can perform segmentation tasks without requiring extensive ground-truth data. This capability can reduce the labor-intensive process of manual annotation and be incorporated into the image processing pipeline, potentially accelerating its workflow.

SAM has been trained on an enormous dataset, of 11 million images and 1.1 billion masks, and it boasts impressive zero-shot performance on already a variety of segmentation tasks (Kirillov et al., 2023). Foundation models such as this, which have shown promising advancements in NLP and, more recently, in computer vision, can carry out zero-shot learning. This means they can learn from new datasets and perform new tasks often by utilizing 'prompting' techniques, even with little to no previous exposure to these tasks. In the field of NLP, "foundation models" refer to large-scale models that are pre-trained on a vast amount of data and are then fine-tuned for specific tasks. These models serve as the "foundation" for various applications (Mai et al., 2023; Mialon et al., 2023; Wu et al., 2023).

SAM's ability to generalize across a wide range of objects and images makes it particularly appealing for remote sensing applications. That it can be retrained with a single example of each new class at the time of prediction (Zhang et al., 2023b), demonstrates the models' high flexibility and adaptability. The implementation of a one-shot approach may assist in designing models that learn useful information from a small number of examples — in contrast to traditional models which usually require large amounts of data to generalize effectively. This could potentially revolutionize how we process remote-sensing imagery. As such, by investigating SAM's innovative technology, we may be able to provide more interactive and adaptable remote sensing systems.

3. Materials and methods

In this section, we describe how we evaluated the performance of the Segment Anything Model (SAM), for both zero and one-shot approach, in the context of remote sensing imagery. The method implemented in this study is summarized in Fig. 1. The data for this study consisted of multiple aerial and satellite datasets. These datasets were selected to ensure diverse scenarios and a large range of objects and landscapes. This helped in assessing the robustness of SAM and its adaptability to different situations and geographical regions.

The study particularly investigated SAM's segmentation capacity under different prompting conditions. First, we used the general segmentation approach, in which SAM was tasked to segment objects and landscapes without any guiding prompts. This provided a baseline for SAM's inherent segmentation capabilities with zero-shot. For this, we only evaluated its visual quality, since it segments every possible object in the image, instead of just the ones with ground-truth labels. It also is not guided by any means, thus resulting in the segmentation of unknown classes, serving as just a traditional segmentation filter.

In the second scenario, bounding boxes were provided. These rectangular boxes, highlighting specific areas within the images, were used to restrict SAM's segmentation per object and see its proficiency in recognizing and segmenting them. Next, we conducted segmentation using points as prompts. In this setup, a series of specific points within the images were provided to guide SAM's processing. It allowed us to test the precision potential of SAM. Finally, we experimented with the segmentation process using only textual descriptions as prompts. This was conducted with an implementation of SAM alongside GroundingDINO's method (Liu et al., 2023b). This permitted an evaluation of these models' capabilities to understand, interpret, and transform textual inputs into precise segmentation outputs.

To measure SAM's adaptability and potential to deal with remote sensing imagery, we then devised a one-shot implementation. For each of the datasets, we presented an example of the target class to SAM.

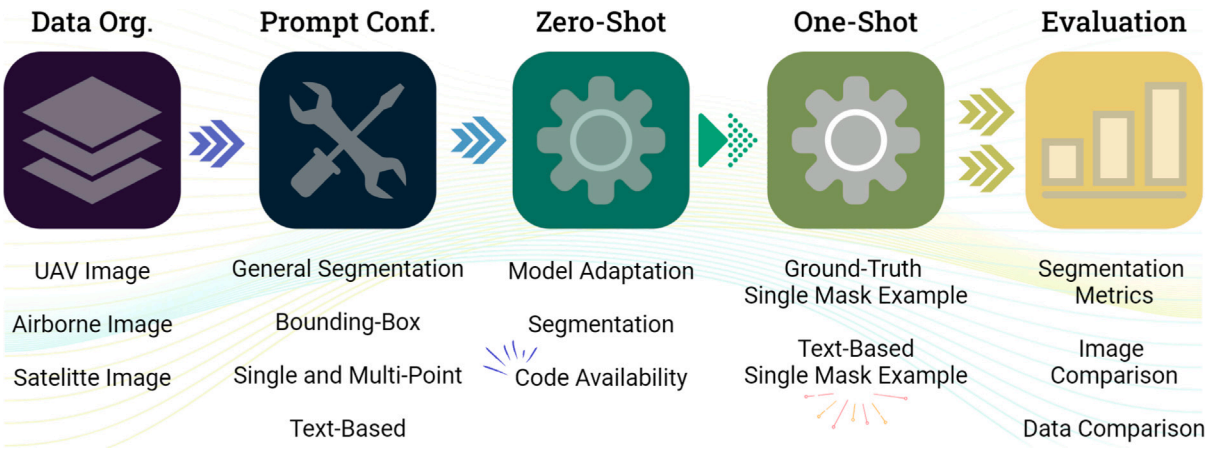


Fig. 1. Schematic representation of the step-by-step process undertaken in this study to evaluate the efficacy of SAM's approach in remote sensing image processing tasks.

Table 1
Overview of the distinct attributes and specifications of the datasets employed in this study.

#	Platform	Resolution (m)	Area (ha)	Target	General	Box	Point	Text	Reference
00	UAV	0.04	70	Tree	Yes	Yes	Centroid	Tree	
01	UAV	0.04	70	House	Yes	Yes	Centroid	House	
02	UAV	0.01	4	Plantation Crop	Yes	No	Multiple	Plantation	Osco et al. (2021a)
03	UAV	0.04	40	Plantation Crop	Yes	No	Multiple	Plantation	
04	UAV	0.09	90	Building	Yes	Yes	Centroid	Building	Gao et al. (2021)
05	UAV	0.09	90	Car	Yes	Yes	Centroid	Car	
06	Airborne	0.20	120	Tree	Yes	Yes	Centroid	Tree	
07	Airborne	0.20	120	Vehicle	Yes	Yes	Centroid	Vehicle	
08	Airborne	0.45	190	Lake	Yes	Yes	Centroid	Lake	
09	Satellite	0.30	–	Building; Road; Water; Barren; Forest; Farm	Yes	Yes	Multiple	Building; Road; Water; Barren; Forest; Farm	LoveDA
10	Satellite	0.50	480	Building; Street; Water; Vehicle; Tree	Yes	Yes	Yes	Building; Street; Water; Vehicle; Tree	SkySat ESA

For that, we adapted the model with a novel combination of the text-prompt approach and the one-shot learning method. Specifically, we selected the best possible example (highest logits) of the target object, using textual prompts to define the object for mask generation. This example was then presented to SAM as the sole representative of the class, effectively guiding its learning process. The rationale behind this combined approach was to leverage the context provided by the text prompts and the efficacy of the one-shot learning method to the adaptability of SAM to an automated enhancement process.

3.1. Description of the datasets

We begin by separating our dataset into three categories related to the platform used for capturing the images: 1. Unmanned Aerial Vehicle (UAV); 2. Airborne, and; 3. Satellite. Each of these categories provides unique advantages and challenges in terms of spatial resolution and coverage. In our study, we aim to evaluate the performance of SAM across these sources to understand its applicability and limitations in diverse contexts. Their characteristics are summarized in Table 1. We also provided illustrative examples from these datasets in Fig. 2 as in bounding boxes and point prompts.

The UAV category comprises data that have the advantage of very-high spatial resolution, returning images and targets with fine details. This makes them particularly suitable for local-scale studies and applications that require high-precision data. However, the coverage area of UAV datasets is limited compared to other data sources. The images comprised particularly single-class objects per dataset, so they were tackled in binary form. In the case of linear objects, specifically continued plantation crops cover, we used multi-points spread within its extremes, to ensure that the model was capable of understating it better. For more condensed targets such as houses and trees, we used the centered position of the object as a point prompt.

The second category is Airborne data, which includes data collected by manned aircraft. These datasets typically offer a good compromise between spatial resolution and coverage area. We processed these datasets with the same approach as with the UAV images since they also consisted of binary problems. The total quantifiable size of these datasets surpasses 90 Gigabytes and comprises more than 10,000 images and image patches. Part of the dataset, specifically the aerial one (UAV and Airborne), is currently being made public in the following link for others to use: GeomaticsandComputerVision/Datasets. These datasets cover different area sizes and their corresponding ground-truth masks were generated and validated by specialists in the field.

The third category consists of Satellite data, which provides the widest coverage and is focused on multi-class problems. The spatial resolution of satellite data is generally lower than that of UAV and Airborne data. Furthermore, the quality of the images is more affected by atmospheric conditions, with differing illumination conditions, thus providing additional challenges for the model. These datasets consist of publicly available images from the LoveDA dataset (Wang et al., 2022) and from the SkySat ESA archive (European Space Agency, 2023) and present a multi-class segmentation problem. To facilitate's SAM evaluation, specifically with the guided prompts (bounding box, point, and text), we conducted a one-against-all approach, in which we separated the classes into individual classifications ("specified class" versus "background").

3.2. Protocol for promptable image segmentation

In this section, we explain how we adapted SAM to the remote sensing domain and how we conducted the promptable image segmentation with it. All of the implemented code, specifically designed for this paper, is made publicly available in an under-construction educational repository (Osco, 2023). Also, as part of our work, we are focusing on developing the "segment-geospatial" package (Wu and Osco, 2023), which implements features that will simplify the process of using SAM models for geospatial data analysis. This is a work in progress, but it is publicly available and offers a suite of tools for performing general segmentation on remote-sensing images using SAM. The goal is to

Fig. 2. Collection of image samples utilized in our research. The top row features UAV-based imagery with bounding boxes and point labels, serving as prompts for SAM. The middle row displays airborne-captured data representing larger regions, with both points and a rectangular box provided as model inputs. The bottom row reveals satellite imagery, again with bounding boxes and points as prompt inputs, offering a trade-off between lower spatial resolution and wider area coverage.

enable users to engage with this technology with a minimum of coding effort.

Our geospatial analysis was conducted with the assistance of a custom tool, namely "SamGeo", which is a component of the original module. SAM possesses different models to be used, namely: ViT-H, ViT-L, and ViT-B (Kirillov et al., 2023). These models have different computational requirements and are distinct in their underlying architecture. In this study, we used the ViT-H SAM model, which is the most advanced and complex model currently available, bringing most of the SAM capabilities to our tests.

To perform the general prompting, we used the generate method of the SamGeo instance. This operation is simple enough since it segments the entire image and stores it as an image mask file, which contained the segmentation masks. Each mask delineates the foreground of the image, with each distinct mask allocated a unique value. This allowed us to segment different geospatial features. The result is a non-classified segmented image that can also be converted into a vector shape. As

mentioned, we only evaluated this approach visually, since it was not possible to appropriately assign the segmented regions outside of our reference class.

For the bounding box prompt, we used the SamGeo instance in conjunction with the objects' shapefile. Bounding boxes are extracted from any multipart polygon geometry returning a, which returned a list of geometric boundaries for our image data based on its coordinates. To efficiently process these boundaries, we initialized the predictor instance. In this process, the image was segmented and passed through the predictor along with a designated model checkpoint. Once established, the predictor processed each clip box, creating the masks for the segmented regions. This process enabled each bounding box's contents to be individually examined as instance segmentation masks. These binary masks were then merged and saved as a single mosaic raster to create a comprehensive visual representation of the segmented regions. Although not focused on remote sensing data, the official implementation is named Grounded-SAM (IDEA-Research, 2023).

The single-point feature prompt was implemented similarly to the bounding-box method. For that, we first defined functions to convert the geodata frame into a list of coordinates [x, y] instead of the previous [x1, y1, x2, y2] ones. We utilized SamGeo again for model prediction but with the distinction of setting its automatic parameter to 'False' and applying the predictor to individual coordinates instead of the bounding boxes. This approach was conducted by iterating through each point, predicting its features in instances, and saving the resulting mask into a unique file per point (also resulting in instance segmentation masks). After the mask files were generated, we proceeded to merge these masks into a single mosaic raster file, giving us a complete representation of all the segmented regions from the single-point feature prompt.

The text-based prompt differentiates from the previous approach since it required additional steps to be implemented. This method combines GroundingDINO's (Liu et al., 2023b) capabilities for zero-shot visual grounding with SAM's object segmentation functionality for retrieving the pre-trained models. For instance, once Grounding DINO has detected and classified an object, SAM is used to isolate that object from the rest. As a result, we have been able to identify and segment objects within our images based on a specified textual prompt. This procedure opens up a new paradigm in geospatial analysis, harnessing the power of state-of-the-art models to extract image features based only on natural language input.

Since remote sensing imagery often contained multiple instances of the same object (e.g., several 'houses', 'cars', 'trees', etc.), we have added a looping procedure. The loop identifies the object with the highest probability in the image (i.e. logits), creates a mask for it, removes it from the image, and then restarts the process to identify the next highest probable object. This process continues until the model reaches a defined minimum threshold for both detection, based on a box threshold, and text prompt association, also based on an specific threshold. The precise balancing of these thresholds (ranging from 0 to 1) is crucial, with implications for the accuracy of the model, so we manually set them for each dataset based on trial and error tentatively:

- Box Threshold: Utilized for object detection in images. A higher value augments model selectivity, isolating only those instances the model identifies with high confidence. A lower value, conversely, expands model tolerance, enhancing overall detections but possibly including less certain ones.
- Text Threshold: Utilized for associating detected objects with provided text prompts. An elevated value mandates a robust association between the object and text, ensuring precision but potentially limiting associations. A diminished value permits broader associations, potentially boosting the number of associations but potentially compromising precision.

These thresholds are critical for ensuring the balance between precision and recall based on specific data and user requirements. The optimal values may diverge depending on the nature and quality of the images and the specificity of text prompts, warranting user experimentation for optimal performance. The segmented individual images and their corresponding boxes are subsequently generated, while the resulting segmentation mask is saved and mosaicked.

3.3. One-shot text-based approach

The one-shot training was conducted following the recommendation in Zhang et al. (2023b) by using its PerSAM and PerSAM-F approaches. We begin by adapting the text-based approach of the combination of the GroundDINO (Liu et al., 2023b) and SAM (Kirillov et al., 2023) methods to return the overall most probable object belonging to the specified class in its description. By doing so, we enable an automated process of identifying a single object and including it on a personalized pipeline for training SAM with this novel knowledge. In this section, we describe the procedures involved in the one-shot training mechanism as well as the methods used for object identification and personalization. To summarize the whole process, we illustrate the main phases in Fig. 3.

Following Fig. 3, the initial phase of the one-shot training mechanism involves the model derived from the object with the highest logits calculated from the text-based segmentation. This ensures the object is accurately recognized and selected for further steps. It is this aspect of the process that the text-based approach starts, capitalizing on GroundDINO's capabilities for zero-shot visual grounding combined with SAM's object segmentation for pre-trained model retrieval. As such, the selected object becomes the "sample" of the one-shot training process due to its high probability of belonging to the specified class by the text.

Once the object has been identified through this method, the next phase involves creating a single-segmented object mask. This mask is used for the retraining of SAM in a one-shot manner. The text-based approach adds value by helping SAM distinguish between the different object instances present in the remote sensing imagery, such as multiple "houses", "cars", or "trees", for example. Each object is identified based on its individual likelihood, leading to the creation of a unique mask for retraining SAM. The third phase starts once the object with the highest probability has been identified and its mask has been used for SAM's one-shot training. The selected input object is removed from the original image, making the remaining objects ready for further segmentation.

The final phase involves a dynamic, interactive loop, where the remaining objects are continuously segmented until no more objects are detectable by the PerSAM approach (Zhang et al., 2023b). This phase is critical as it ensures that every potential object within the image is identified and segmented. Here again, the loop approach aids the process, using a procedure that identifies the next highest probable object, as it creates a mask, removes it from the image, and repeats. This cycle continues until a breakpoint is reached, where it detects the previous position again.

Another important aspect of the one-shot approach regards the choice of the method for its training. An early exploration of both PerSAM and PerSAM-F methods (Zhang et al., 2023b) was conducted to assess their utility in the context of remote sensing imagery. Our investigations have shown that PerSAM-F emerges as a more suitable choice for this specific domain. PerSAM, in its original formulation, leverages one-shot data through a series of techniques such as target-guided attention, target-semantic prompting, and cascaded post-refinement, delivering favorable personalized segmentation performance for subjects in a variety of poses or contexts. However, there were occasional failure cases, notably where the subjects comprised hierarchical structures to be segmented.

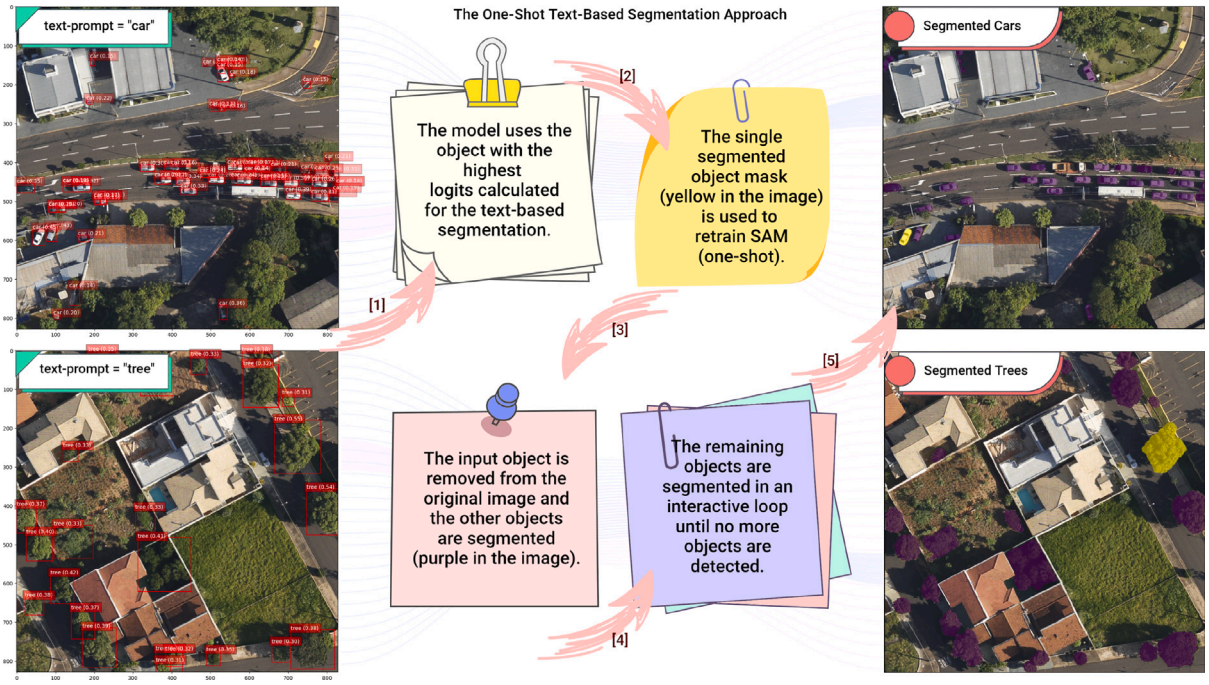


Fig. 3. Visual representation of the one-shot-based text segmentation process in action. The figure provides a step-by-step illustration of how the model identifies and segments the most probable object based on a text prompt with "car" and "tree" as examples.

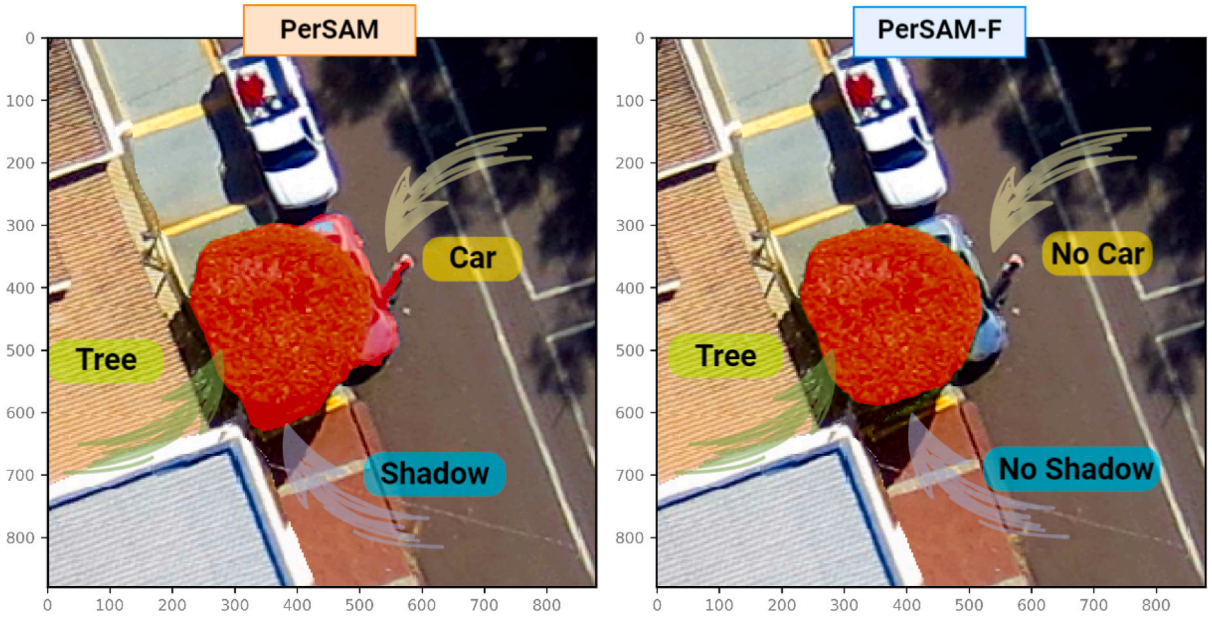


Fig. 4. Comparative illustration of tree segmentation using PerSAM and PerSAM-F. On the left, the PerSAM model segments not only the tree but also its shadow and a part of the car underneath it. On the right, the PerSAM-F model, fine-tuned for hierarchical structures and varying scales, accurately segments only the tree, demonstrating its improved ability to discern and isolate the target object in remote sensing imagery.

Examples of such cases in traditional images are discussed in Zhang et al. (2023b), where ambiguity provides a challenge for PerSAM in determining the scale of the mask as output (e.g. a "dog wearing a hat" may be segmented entirely, instead of just the "dog"). In the context of remote sensing imagery, such hierarchical structures are commonly encountered. An image may contain a tree over a house, a car near a building, a river flowing through a forest, and so forth. These hierarchical structures pose a challenge to the PerSAM method, as it struggles to determine the appropriate scale of the mask for the segmentation output. An example of such a case, where a tree covers a car, can be seen in Fig. 4.

To address this challenge, we used PerSAM-F, the fine-tuning variant of PerSAM. As previously mentioned, PerSAM-F freezes the entire SAM to preserve its pre-trained knowledge and only fine-tunes two parameters within a ten seconds training window (Zhang et al., 2023b). Crucially, it enables SAM to produce multiple segmentation results with different mask scales, thereby allowing for a more accurate representation of hierarchical structures commonly found in remote sensing imagery. PerSAM-F employs learnable relative weights for each scale, which adaptively select the best scale for varying objects. This strategy offers an efficient way to handle the complexity of segmentation tasks in remote sensing imagery, particularly when dealing with objects that

exhibit a range of scales within a single image. This, in turn, preserves the characteristics of the segmented objects more faithfully.

As such, PerSAM-F exhibited better segmentation accuracy in our early experiments, thus being the chosen method to be incorporated with the text-based approach. In our training phase with PerSAM-F, the DICE loss and Sigmoid Focal Loss are computed, and their summation forms the final loss that is backpropagated to update the model weights. The learning rate is scheduled using the Cosine Annealing method (Loshchilov and Hutter, 2017), and the model is trained for 1000 epochs. With hardware acceleration incorporated, the model can be trained within a reasonable time frame without requiring excessive computational resources. This careful setup ensures the extraction of meaningful features from the reference image, contributing to the effectiveness of our one-shot text-based approach.

To evaluate the performance and utility of the text-based one-shot learning method, we conduct a comparative analysis against a traditional one-shot learning approach. The traditional method used for comparison follows the typical approach of one-shot learning, providing the model with a single example from the ground-truth mask, manually labeled by human experts. To ensure fairness, we provided the model with multiple random samples from each dataset, and mimic the image inputs to return a direct comparison for both approaches. We calculated the evaluation metrics from each input and returned its average value alongside with its standard deviation. Since the text approach always uses the same input (i.e. the highest logits object), we were able to return a single measurement of their accuracies.

3.4. Model evaluation

The performance of both zero-shot and one-shot models was measured by evaluating their prediction accuracy on a ground-truth mask. For that, we used metrics like Intersection over Union (IoU), Pixel Accuracy, and Dice Coefficient. These metrics are commonly used in evaluating imaging segmentation, as they provide a more nuanced understanding of model performance. For that, we compared pairs of predicted and ground-truth masks.

Intersection over Union (IoU) is a common evaluation metric for object detection and segmentation problems. It measures the overlap between the predicted segmentation and the ground truth (Rahman and Wang, 2016). The IoU is the area of overlap divided by the area of the union of the predicted and ground truth segmentation. A higher IoU means a more accurate segmentation. The equation to achieve it is presented as:

$$IoU = \frac{TP}{TP + FP + FN} \tag{1}$$

Here, TP represents True Positives (the correctly identified positives), FP represents False Positives (the incorrectly identified positives), and FN represents False Negatives (the positives that were missed).

Pixel Accuracy is the simplest used metric and it measures the percentage of pixels that were accurately classified (Minaee et al., 2021). It is calculated by dividing the number of correctly classified pixels by the total number of pixels. This metric can be misleading if the classes are imbalanced. The following equation returns it:

$$Pixel\ Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

Here, TN represents True Negatives (the correctly identified negatives).

Dice Coefficient (also known as the Sørensen–Dice index) is another metric used to gauge the performance of image segmentation methods. It is particularly useful for comparing the similarity of two samples. The Dice Coefficient is twice the area of overlap of the two segmentations divided by the total number of pixels in both images (the sum of the areas of both segmentations) (Minaee et al., 2021). The Dice Coefficient ranges from 0 (no overlap) to 1 (perfect overlap). The equation to perform it is given as follows:

$$Dice\ Coefficient = 2 * \frac{TP}{2 * TP + FP + FN} \tag{3}$$

We also utilized other metrics, particularly, True Positive Rate (TPR) and False Positive Rate (FPR) to measure the effectiveness of SAM, juxtaposed with the accurately labeled class from each dataset. The interpretation of these metrics as per (Powers, 2020) is: The True Positive Rate (TPR) denotes the fraction of TP cases among all actual positive instances, while the False Positive Rate (FPR) signifies the fraction of FP instances out of all negative instances. A model with a higher TPR is proficient at correctly pinpointing lines and edges and performs better at avoiding incorrect detections of lines and edges when the FPR is lower. Both metrics are calculated as:

$$\text{TPR} = \frac{TP}{(TP + FN)} \tag{4}$$

$$\text{FPR} = \frac{FP}{(FP + TN)} \tag{5}$$

In alignment with the inherent structure of SAM, a transformer network, our objective was to maintain the comprehensive context of our images to fully harness the model's attention mechanism. This consideration led to our decision to process larger image crops or entire orthomosaics as a single unit, rather than fragmenting them into fixed-sized smaller patches. While this approach enhances the model's contextual understanding, it understandably augments the computational time.

For most larger patches or quartered orthomosaics, the inference duration on a GPU was kept under 10 min, providing a balance between computational load and contextual analysis. When processing entire datasets as a whole, the time requirement extended to approximately 1 to 2 h. Despite the augmented processing time for larger datasets, the assurance of comprehensive contextual analysis justifies this computational investment. Still, in fixed-sized patches such as the ones from the publicly available datasets, the inference time was under a second for each patch. These inferences were executed on an NVIDIA RTX 3090 equipped with 24 GB GDDR6X video memory and 10,496 CUDA cores, operating on Ubuntu 22.04.

4. Results and discussion

4.1. General segmentation

Our exploration of SAM for remote sensing tasks involved an evaluation of its performance across various datasets and scenarios. This section presents the results and discusses their implications for SAM's role in remote sensing image analysis. This process commenced with an investigation of SAM's general segmentation approach, which requires no prompts. By merely feeding SAM with remote sensing images, we aimed to observe its inherent ability to detect and distinguish objects on the surface. Examples of different scales are illustrated in Fig. 5, where we converted the individual regions to vector format. This approach demonstrates its adaptability and suitability for various applications. However, as this method is not guided by a prompt, it is not returning specific segmentation classes, making it difficult to measure its accuracy based on our available labels.

As depicted in Fig. 5, the higher the spatial resolution of an image, the more accurately SAM segmented the objects. An interesting observation pertained to the processing of satellite images where SAM encountered difficulties in demarcating the boundaries between contiguous objects (like large fragments of trees or roads). Despite this limitation, SAM exhibited an ability to distinguish between different regions when considering very-high spatial resolution imagery, indicative of an effective segmentation capability that does not rely on any prompts. This approach offers value for additional applications that are based on object regions, such as classification algorithms. Moreover, SAM can expedite the process of object labeling for refining other models, thereby significantly reducing the time and manual effort required for this purpose.

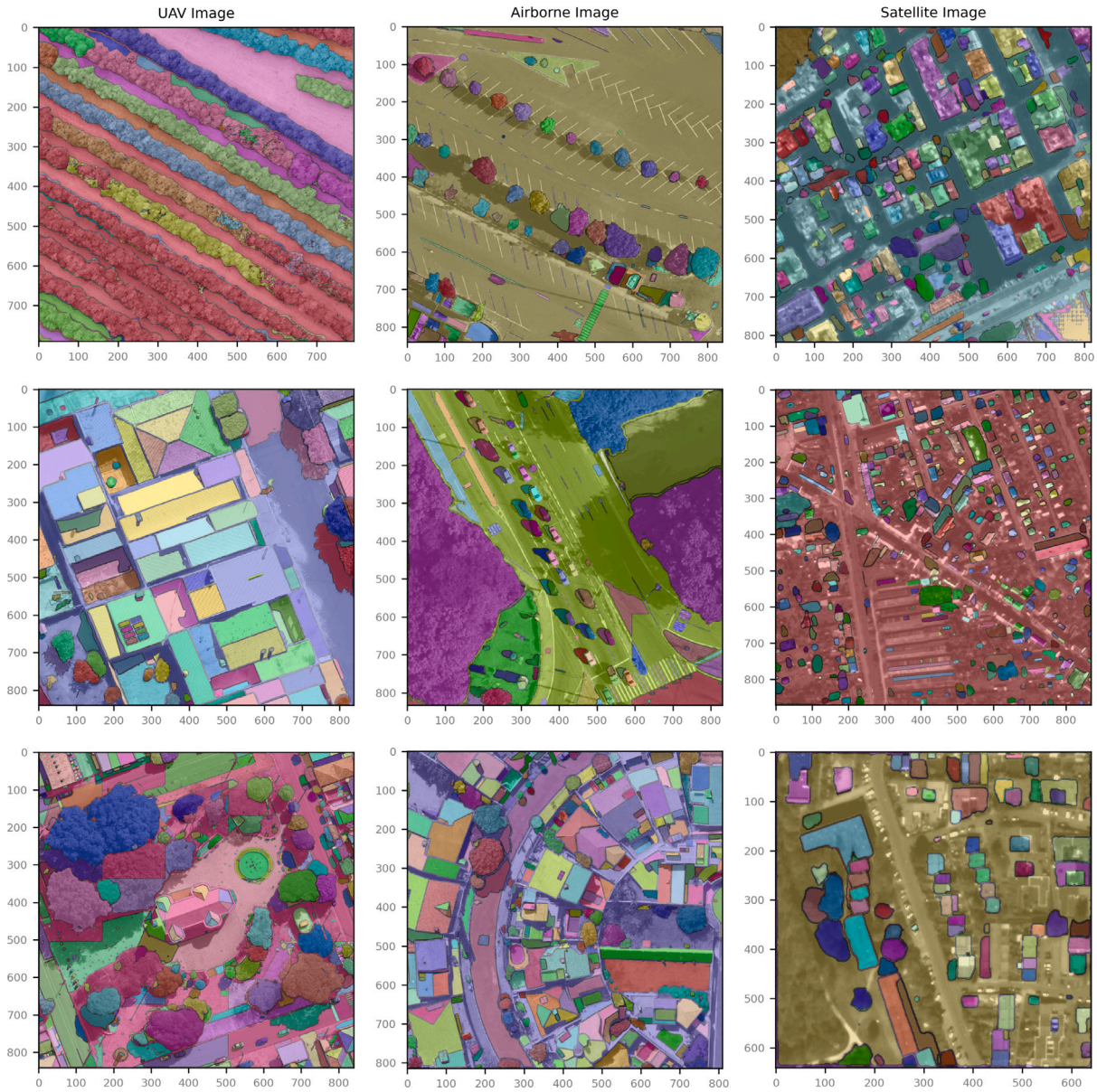


Fig. 5. Examples of segmented objects using SAM's general segmentation method, drawn from diverse datasets based on their platforms. Objects are represented in random colors. As the model operates without any external inputs, it deduces object boundaries leveraging its zero-shot learning capabilities.

4.2. Zero-shot segmentation

Following this initial evaluation, we proceeded to test SAM's promptable segmentation abilities using bounding boxes, points, and text features. The resulting metrics for each dataset are summarized in Table 2. Having compiled a dataset across diverse platforms, including UAVs, aircraft devices, and satellites with varying pixel sizes, we noted that SAM's segmentation efficacy is also quantitatively influenced by the image's spatial resolution. These findings underscore the significant influence of spatial resolution on the effectiveness of different prompt types.

For instance, on the UAV platform, text prompts showed superior performance for object segmentation tasks such as trees, with higher Dice and IoU values. However, bounding box prompts were more effective for delineating geometrically well-defined and larger objects like houses and buildings. The segmentation of plantation crops was a unique case. Point prompts performed well at a finer 0.01 m resolution for individual plants. However, as the resolution coarsened to 0.04 m and the plantation types changed, becoming denser with the plant canopy covering entire rows, bounding box prompts outperformed the others. This outcome suggests that, for certain objects, the type of input prompt can greatly influence detection and segmentation in the zero-shot approach.

With the airborne platform, point prompts were highly effective at segmenting trees and vehicles at a 0.20 m resolution. This trend continued for the segmentation of lakes at a 0.45 m resolution. It raises the question of whether the robust performance of point prompts in these scenarios is a testament to their adaptability to very high-resolution imagery or a reflection of the target object's specific characteristics. These objects primarily consist of very defined features (like cars and vehicles) or share similar characteristics (as in bodies of water).

In the context of satellite-based remote sensing imagery, point prompts proved most efficient for multi-class segmentation at the examined resolutions of 0.30 m and 0.50 m. This can be attributed to the fact that bounding box prompts tend to overshoot object boundaries, producing more false positives compared to point prompts. This finding indicates the strong ability of point prompts to manage a diverse set of objects and categories at coarser resolutions, making them a

Table 2
Summary of metrics for the image segmentation task across different platforms, targets, and resolutions, and using different prompts for SAM in zero-shot mode. The values in red indicate the best performance for a particular target under specific conditions.

#	Platform	Target	Resolution (m)	Prompt	Dice (%)	IoU (%)	Pixel Acc. (%)	TPR (%)	FPR (%)
00	UAV	Tree	0.04	Box	88.8	79.9	96.0	94.2	3.6
				Point	91.8	84.8	97.6	91.6	1.4
				Text	92.2	85.2	98.1	92.1	1.2
01	UAV	House	0.04	Box	92.7	86.3	98.4	97.4	1.5
				Point	70.8	54.8	84.0	96.6	19.2
				Text	89.2	79.8	95.6	97.1	10.1
02	UAV	Plantation	0.01	Box	86.2	82.8	85.5	88.2	11.1
				Point	95.8	92.0	95.0	98.0	9.2
				Text	67.1	64.4	66.5	68.6	12.0
03	UAV	Plantation	0.04	Box	80.1	68.9	95.2	94.4	10.4
				Point	72.7	57.1	93.5	93.4	6.5
				Text	44.1	32.8	49.9	45.0	6.1
04	UAV	Building	0.09	Box	69.7	53.5	81.3	95.5	22.8
				Point	69.1	52.8	84.2	91.1	17.5
				Text	66.3	50.9	77.2	90.7	24.0
05	UAV	Car	0.09	Box	78.8	65.0	97.0	66.0	0.2
				Point	90.0	81.9	99.1	86.7	0.3
				Text	92.7	84.3	97.3	89.3	0.1
06	Airborne	Tree	0.20	Box	68.8	52.4	91.2	84.4	7.9
				Point	91.7	84.7	93.5	88.3	2.9
				Text	89.0	82.2	90.7	85.6	3.7
07	Airborne	Vehicle	0.20	Box	86.1	75.6	99.5	86.9	0.3
				Point	86.3	75.9	99.1	78.5	0.1
				Text	84.6	74.4	97.1	76.9	0.2
08	Airborne	Lake	0.45	Box	57.4	40.3	98.3	98.8	1.7
				Point	97.2	94.5	99.9	99.1	0.1
				Text	89.4	86.9	91.9	91.2	0.8
09	Satellite	Multiclass	0.30	Box	39.1	22.5	94.5	22.6	0.4
				Point	82.3	56.7	87.8	67.8	3.7
				Text	74.0	51.0	79.1	61.0	3.9
10	Satellite	Multiclass	0.50	Box	26.1	15.0	93.6	15.1	0.5
				Point	54.9	37.8	87.0	45.2	4.2
				Text	49.4	34.0	78.3	40.7	4.4

promising tool for satellite remote sensing applications. The text-based approach was found to be the least effective, primarily due to the model's difficulty in associating low-resolution objects with words. Still, it is important to notice that, from all the datasets, the satellite multiclass problem proved to be the most difficult task for the model, with generally lower metrics than the others.

Qualitatively, our observations also revealed that bounding boxes were particularly effective for larger objects (Fig. 6). However, for smaller objects, SAM tended to overestimate the object size by including shadows in the segmented regions. Despite this overestimation, the bounding box approach still offers a useful solution for applications where an approximate estimate of such larger objects suffices. For these types of objects, a single point or central location does not suffice, they are defined by a combination of features within a particular area. Bounding boxes provide a more spatially comprehensive prompt, encapsulating the entire object, which makes them more efficient in these instances.

The point-based approach outperformed the others across our dataset, specifically for distinct objects. By focusing on a singular point, SAM was able to provide precise segmentation results, thus proving its capability to work in detail (Fig. 7). In the plantation dataset with 0.01 m resolution, for instance, when considering individual small plants, the point approach returned better results than bounding boxes. This approach may hold particular relevance for applications requiring precise identification and segmentation of individual objects in an image. Also, when isolating entities like single trees and vehicles, these precise spatial hints might suffice for the model to accurately identify and segment the object.

The textual prompt approach also yielded promising results, particularly with very high-resolution images (Fig. 8). While it was found to be relatively comparable in performance with the point and bounding box prompts for the aerial datasets, the text prompt approach had notable limitations when used with lower spatial resolution images. The text-based approach also returned worse predictions on the plantation with 0.04 m. This may be associated with the models' limitation on understanding the characteristics of specific targets, especially when considering the bird's eye view of remote sensing images. Since it relies on GroundDINO to interpret the text, it may be more of a limitation on it than on SAM, mostly because, when applying the general segmentation, the results visually returned overall better segmentation on these datasets (Fig. 5).

Text prompts, though generally trailing behind in performance, still demonstrated commendable results, often closely following the top-performing prompt type. Text prompts offer ease of implementation as their primary advantage. They do not necessitate specific spatial annotations, which are often time-consuming and resource-intensive to produce, especially for extensive remote sensing datasets. However, their effectiveness hinges on the model's ability to translate text to image information. Currently, their key limitation is that they are typically not trained specifically on remote sensing images, leading to potential inaccuracies when encountering remote sensing-specific terms or concepts. Improving the effectiveness of text prompts can be achieved through fine tuning models on remote sensing-specific datasets and terminologies. This could enable them to better interpret the nuances of remote sensing imagery, potentially enhancing their performance to match or even surpass spatial prompts like boxes and points.

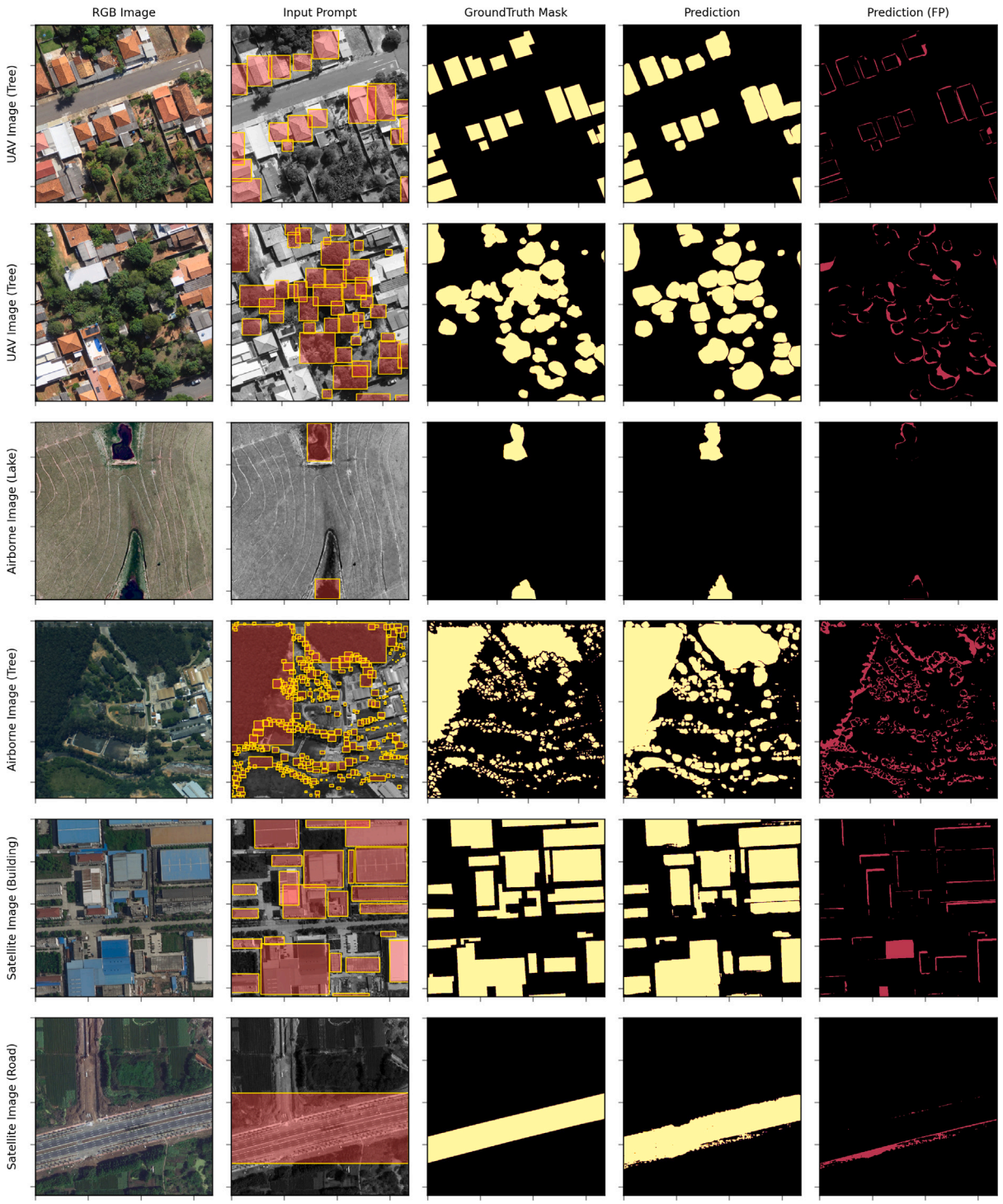


Fig. 6. Illustrations of images processed using bounding-box prompts. The first column consists of the RGB image, while the second column demonstrates how the prompt was handled. The ground-truth mask is presented in the third column and the prediction result from SAM in the fourth. The last column indicates the false positive (FP) pixels from the prediction.

4.3. One-shot segmentation

Regarding our one-shot approach, we noticed that the models' performance is improved in most cases, as evidenced by the segmentation metrics calculated on each dataset. Table 3 presents a detailed comparison of the different models' performance providing a summary of the segmentation results. Fig. 9 offers a visual illustration of example results obtained from both approaches, particularly highlighting the performance of the model. The metrics indicate that, while the PerSAM approach with a human-sampled example may be more appropriate than the proposed text-based approach, this may not always be the case when considering the metric's standard deviation. This opens up the potential for adopting the automated process instead. However, in some instances, specifically where GroundDINO's not capable of identifying the object, to begin with, the human-labeling provides a more appropriate result.

In its zero-shot form, SAM tends to favor selecting shadows in some instances alongside its target, which can lower its performance in tasks

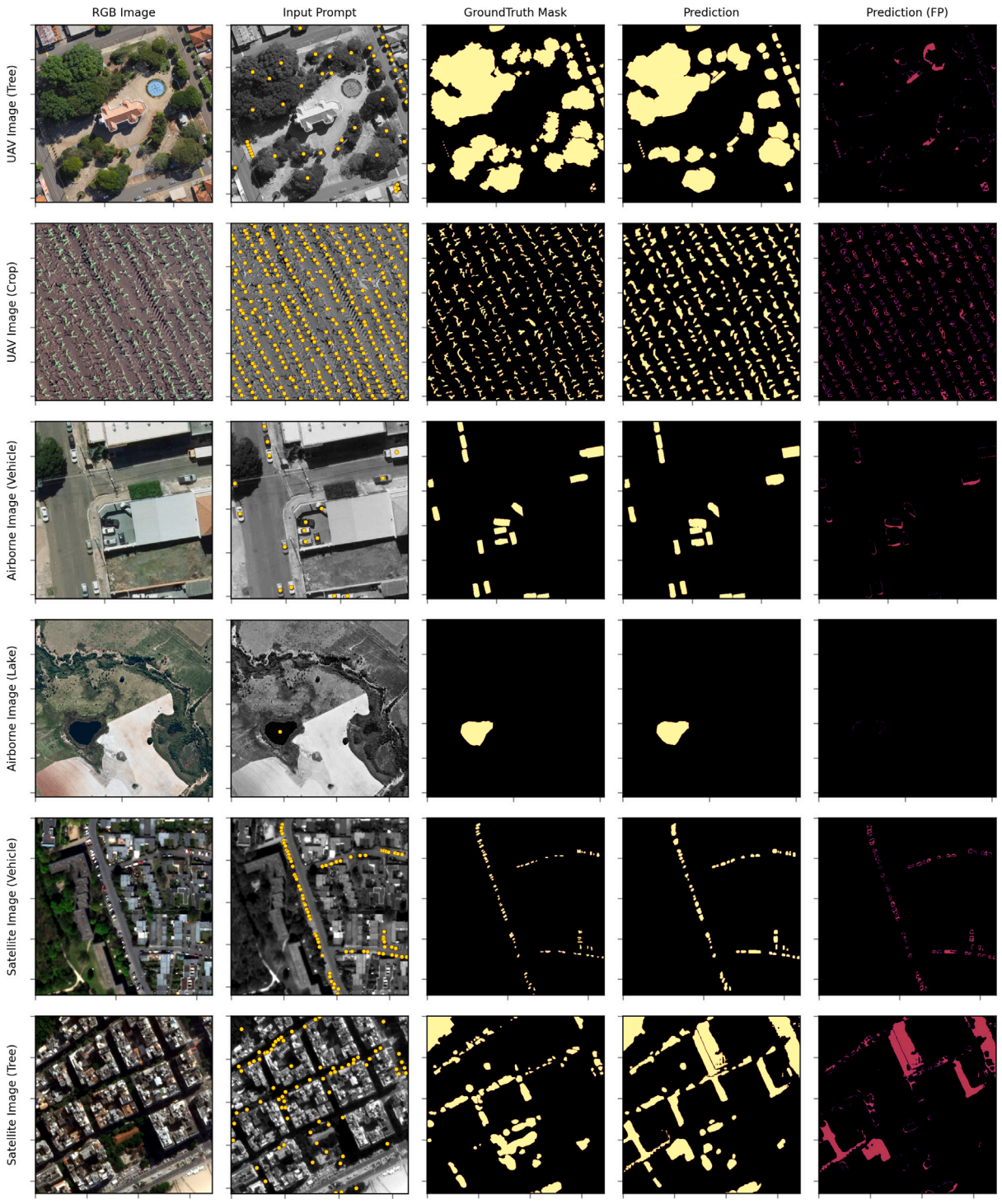


Fig. 7. Illustrations of images processed using point prompts. The first column presents the RGB image, while the second column demonstrates the handling of the point prompt. The third column showcases the ground-truth mask, and the fourth column shows the prediction result from SAM. The final column highlights the false positive (FP) pixels from the prediction.

like tree detection. Segmenting objects with similar surrounding elements, especially when dealing with construction materials like streets and sidewalks, can be challenging for SAM, as noticed in our multi-class problem. Moreover, its performance with larger grouped instances, particularly when using the single-point mode, can be unsatisfactory. Also, the segmentation of smaller and irregular objects poses difficulties for SAM independently from the given prompt. SAM may generate disconnected components that do not correspond to actual features, specifically in satellite imagery where the spatial resolution is lower.

The text-based one-shot learning approach, on the other hand, automates the process of selecting the example. It uses the text-based prompt to choose the object with the highest probability (highest logits) from the image as the training example. This not only reduces the need for manual input but also ensures that the selected object is highly representative of the specified class due to its high probability. Additionally, while the text-based approach is capable of handling multiple instances of the same object class in a more streamlined manner, thanks to the looping mechanism that iteratively identifies and

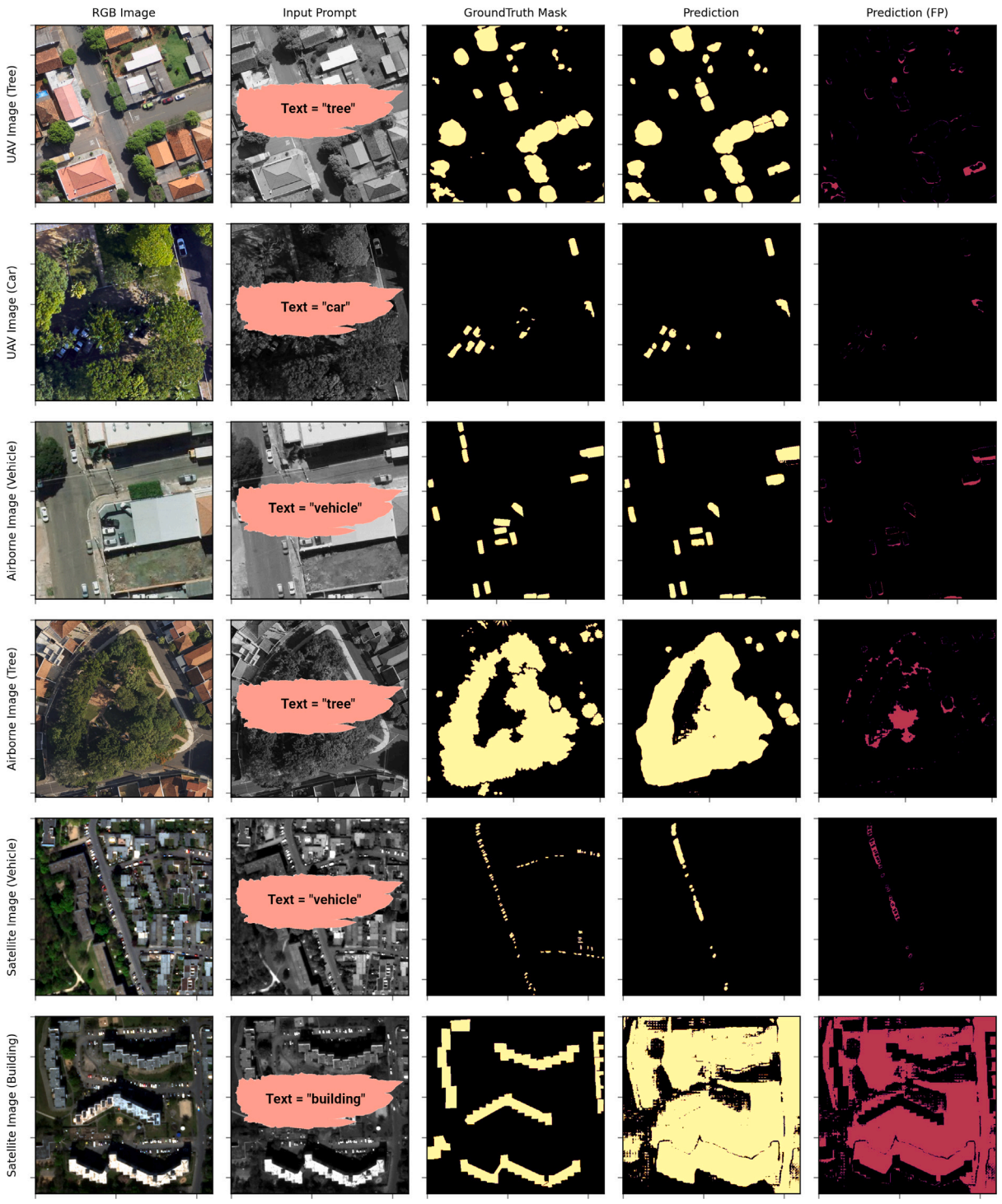


Fig. 8. Examples of images processed through text-based prompts. The first column contains the RGB image, while the second column indicates the text prompt used for the model. The ground-truth mask is shown in the third column, with the prediction result from SAM in the fourth. The last column indicates the false positive (FP) pixels from the prediction.

segments objects based on their probabilities. The one-example policy, however, excluded some of the objects in the image to favoring only the objects similar to the given sample.

In summary, upon comparing these two methods, we found that the traditional one-shot learning approach outperforms the zero-shot learning approach in all datasets. Additionally, the combination of text-based with one-shot learning also, even when not improving on it, gets close enough in most cases. This comparison underscores the benefits and potential of integrating state-of-the-art models with natural language processing capabilities for efficient and accurate geospatial analysis. Nevertheless, it is important to remember that the optimal choice between these methods may vary depending on the specific context and requirements of a given task.

5. Future perspectives on SAM for remote sensing

SAM has several advantages that make it an attractive option for remote sensing applications. First, it offers zero-shot generalization

Table 3
Comparison of segmentation results on different platforms and targets when considering both the one-shot and the text-based one-shot approaches. The baseline values are referent to the best metric obtained by the previous zero-shot investigation, be it from a bounding box, a point, or a text prompt. The red colors indicate the best result for each scenario.

#	Platform	Target	Resolution (m)	Sample	Dice (%)	IoU (%)	Pixel Acc. (%)	TPR (%)	FPR (%)
00	UAV	Tree	0.04	Baseline	92.2	85.2	98.1	92.1	1.2
				PerSAM-F	94.5 ± 4.2	87.4	98.8	94.4	1.1
				Text PerSAM-F	95.0 ± 4.9	87.8	99.3	96.3	0.9
01	UAV	House	0.04	Baseline	92.7	86.3	98.4	97.4	1.5
				PerSAM-F	95.4 ± 2.1	88.9	99.3	98.1	1.1
				Text PerSAM-F	95.0 ± 2.7	88.5	98.8	99.8	1.4
02	UAV	Plantation Crop	0.01	Baseline	80.1	68.9	95.2	94.4	10.4
				PerSAM-F	82.1 ± 6.4	70.6	98.8	96.8	9.6
				Text PerSAM-F	64.1 ± 7.2	55.1	76.2	75.5	15.6
03	UAV	Plantation Crop	0.04	Baseline	95.8	92.0	95.0	98.0	9.2
				PerSAM-F	98.2 ± 1.1	94.3	98.8	100.4	8.5
				Text PerSAM-F	76.7 ± 1.3	73.6	76.0	78.4	13.8
04	UAV	Building	0.09	Baseline	69.7	53.5	81.3	95.5	22.8
				PerSAM-F	87.2 ± 6.2	66.9	98.0	96.6	21.0
				Text PerSAM-F	73.2 ± 6.7	54.9	94.3	97.9	21.1
05	UAV	Car	0.09	Baseline	92.7	84.3	97.3	89.3	0.1
				PerSAM-F	95.0 ± 2.4	86.4	98.8	91.5	0.1
				Text PerSAM-F	95.5 ± 3.0	86.9	99.3	93.3	0.1
06	Airborne	Tree	0.20	Baseline	91.7	84.7	93.5	88.3	2.9
				PerSAM-F	94.0 ± 1.3	86.8	98.8	90.5	2.7
				Text PerSAM-F	94.5 ± 1.5	87.3	99.3	92.3	2.1
07	Airborne	Vehicle	0.20	Baseline	86.3	75.9	99.1	78.5	0.1
				PerSAM-F	88.4 ± 5.6	77.8	99.8	80.4	0.2
				Text PerSAM-F	86.7 ± 6.5	76.3	99.6	78.9	0.1
08	Airborne	Lake	0.45	Baseline	97.2	94.5	99.9	99.1	0.1
				PerSAM-F	97.6 ± 1.5	94.9	99.9	99.5	0.1
				Text PerSAM-F	97.3 ± 1.3	94.6	99.8	99.2	0.1
09	Satellite	Multiclass	0.30	Baseline	82.3	56.7	87.8	67.8	3.7
				PerSAM-F	90.5 ± 5.2	68.0	96.6	74.5	3.5
				Text PerSAM-F	89.7 ± 5.3	61.8	95.8	73.9	3.5
10	Satellite	Multiclass	0.50	Baseline	54.9	37.8	87.0	45.2	4.2
				PerSAM-F	60.3 ± 10.4	45.3	95.7	49.7	3.9
				Text PerSAM-F	59.8 ± 12.3	41.2	94.8	49.2	4.0

to unfamiliar objects and images without requiring additional training (Kirillov et al., 2023). This capability allows SAM to adapt to the diverse and dynamic nature of remote sensing data, which often consists of varying land cover types, resolutions, and imaging conditions. Second, SAM's interactive input process can significantly reduce the time and labor required for manual image segmentation. The model's ability to generate segmentation masks with minimal input, such as a text prompt, a single point, or a bounding box, accelerates the annotation process and improves the overall efficiency of remote sensing data analysis. Lastly, the decoupled architecture of SAM, comprising a one-time image encoder and a lightweight mask decoder, makes it computationally efficient. This efficiency is crucial for large-scale remote sensing applications, where processing vast amounts of data on time is of utmost importance.

However, our study consists of an initial exploration of this model, where there is still much to be investigated. In this section, we discuss future perspectives on SAM and how it can be improved upon. Despite its potential, SAM has some limitations when applied to remote sensing imagery. One challenge is that remote sensing data often come in different formats, resolutions, and spectral bands. SAM, which has been trained primarily on RGB images, may not perform optimally with multispectral or hyperspectral data, which are common in remote sensing applications. A possible approach to this issue consists of either adapting SAM to read in multiple bands by performing rotated 3-band combinations or performing a fine-tuning to domain adaption. In our early experiments, a simple example run on different multispectral datasets demonstrated that, although the model has the potential to segment different regions or features, it still needs further exploration. This is something that we intend to explore in future research, but expect that others may look into it as well.

Regardless, the current model can be effectively used in various remote sensing applications. For instance, we verified that SAM can be easily employed for land cover mapping, where it can segment forests, urban areas, and agricultural fields. It can also be used for monitoring urban growth and land use changes, enabling policymakers and urban planners to make informed decisions based on accurate and up-to-date information. Furthermore, SAM can be applied in a pipeline process to monitor and manage natural resources. Its efficiency and speed make it suitable for real-time monitoring, providing valuable information to decision-makers. This is also a feature that could be potentially explored by research going forward with its implementation.

Nevertheless, it is crucial to underscore a significant limitation concerning the complexity of our data. While our primary objective was to analyze results across varying spatial resolutions and broad remote sensing segmentation tasks, the limited regional diversity of our data may not fully capture the range of object characteristics encountered worldwide. Future research, therefore, could emphasize utilizing and adapting to a more diverse array of the same object, thereby bolstering the robustness and applicability of the model or its adaptations. For instance, in the detection of buildings and water bodies, exploration of publicly available datasets from diverse regions (Boguszewski et al., 2022; Zhang et al., 2023c) could provide a more comprehensive understanding of these objects' varied characteristics, and contribute to the enhancement of algorithmic performance across varied geographical contexts.

For the one-shot technique based on SAM, which is the capacity to generate accurate segmentation from a single example (Zhang et al., 2023b). Our experimental results indicate an improvement in performance across most investigated datasets, especially considering the

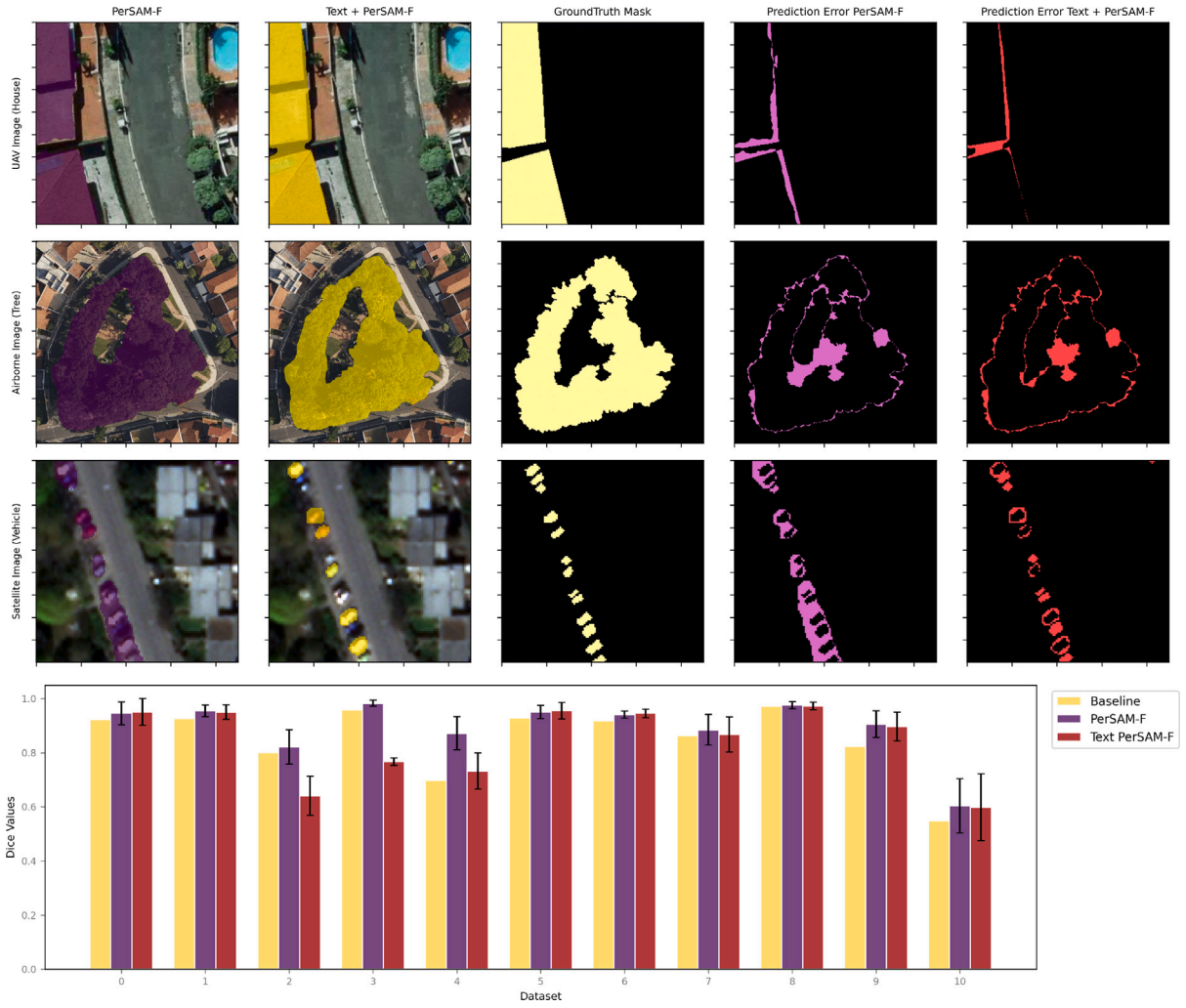


Fig. 9. Visual illustration of the segmentation results using PerSAM and text-based PerSAM. from The last two columns highlights the difference in pixels the PerSAM prediction and the text-based PerSAM prediction to its ground truth. The graphic compares the range from the Dice values of both PerSAM and text-based PerSAM, illustrating how the proposed approach remains similar to the traditional PerSAM approach, underscoring the potential of most practices to adopt the automated process in such cases.

border of the objects. However, it is essential to note that one-shot learning may pose challenges to the generalization capability of the model. This may be an issue of remote sensing data that often exhibit a high degree of heterogeneity and diversity (Zia et al., 2022). For instance, a "healthy" tree can be a good sample for the model, but it can bias it to ignore "unhealthy" trees or canopies with different structures.

Expanding the one-shot learning to a few-shot scenario could potentially improve the model's adaptability to different environments or tasks by enabling it to learn from more than one example (2 to 10) instead of a single one. This would involve using a small set of labeled objects for each land cover type during the training process (Sun et al., 2021; Li et al., 2022a). A more robust learning approach, which uses a larger number of examples for each class, could further enhance the model's ability to capture the nuances and variations within each class. This approach, however, may require more computational resources and training data, and thus may not be suitable for all applications.

Additionally, While SAM is a powerful tool for image segmentation, its effectiveness can be boosted when combined with other techniques. For example, integrating SAM into another ViT framework in a weakly-supervised manner could potentially improve the segmentation result, better handling the spatial-contextual information. However, it is worth noting that integrating it might also bring new challenges (Wang et al., 2020a). One potential issue could be the increased model complexity and computational requirements, which might limit its feasibility. But, as the training of transformers typically requires large amounts of data, SAM can provide fast and relatively accurate labeled regions for it.

Furthermore, one of the key challenges to tackle would be improving SAM's performance when applied to low spatial resolution imagery. Thus, as the original training data of SAM primarily consisted of high-resolution images, it is inherently more suitable for similar high-resolution conditions, even in the remote sensing domain. The noticeable decrease in accuracy at resolutions above 30 cm, noted in our tests, further substantiates this observation. This shortcoming can be further explored by coupling SAM with a Super-Resolution (SR) technique (Yang et al., 2015), for instance, creating a two-step process, where the first step involves using an SR model to increase the spatial resolution of the imagery, and the second step involves using the enhanced resolution image as an input to SAM. It is acknowledged that while this method can theoretically enhance the performance of SAM with low-resolution images, the Super-Resolution techniques themselves can introduce errors, potentially offsetting the benefits (Yang et al., 2015). Therefore, the proposed two-step process should be approached with caution, ensuring meticulous testing and validation. A dedicated exploration into refining and optimizing SAM for lower-resolution images, possibly involving adaptation and training of the model on lower-resolution data, will be integral to ensuring its effective and reliable application in diverse remote sensing scenarios.

As we explored the integration of SAM with other types of methods, such as GroundDINO (Liu et al., 2023b), we noticed both strengths and

limitations that were already discussed in the previous section. This combination demonstrates a high degree of versatility and accuracy in tasks such as instance segmentation, where GroundDINO's object detection and classification guided SAM's segmentation process. However, the flexibility of this approach extends beyond these specific models. Any similar models could be swapped in as required, expanding the applications and robustness of the system. Alternatives such as GLIP (Li et al., 2022b) or CLIP (Liu et al., 2023a) may replace GroundDINO, allowing for further experimentation and optimization (Zhang et al., 2022b). Furthermore, integrating language models like ChatGPT (OpenAI, 2023) could offer additional layers of interaction and nuances of understanding, demonstrating the far-reaching potential of combining these expert models. This modular approach underpins a potent and adaptable workflow that could reshape our capabilities in handling remote sensing tasks.

The integration of Geographical Information Systems (GIS) with models like SAM holds significant promise for enhancing the annotation process for training specific segmentation and change detection models. A fundamental challenge often lies in the discrepancy between training data and the image data employed due to different acquisition times and since the data used could be marred with annotator errors, leading to a compatibility issue with the used image. The integration with SAM could help users optimize the creation of annotations and, when suitable, improve its results with editing, thus creating a quicker and more robust dataset. Lastly, a topic which is not discussed in this paper, but which is an important issue for applications particularly in the area of geospatial intelligence is AI security. A recent survey paper on this topic is Xu et al. (2023). It discusses issues such as that it can be unclear based on which data a (foundation) model has been trained and what deficits may arise from this. Particularly, an adversary might have contaminated the training data.

In short, our study focused on demonstrating the potential of SAM adaptability for the remote sensing domain, as well as presenting a novel, automated approach, to retrain the model with one example from the text-based approach. While there is much to be explored, it is important to understand how the model works and how it could be improved upon. To summarize this discussion, there are many potential research directions and applications for SAM in remote sensing applications, which can be condensed as follows:

- Examining the most effective approaches and techniques for adapting SAM to cater to a variety of remote sensing data, including multispectral and hyperspectral data.
- Analyzing the potential of coupling SAM with few-shot or multi-shot learning, to enhance its adaptability and generalization capability across diverse remote sensing scenarios.
- Investigating potential ways to integrate SAM with prevalent remote sensing tools and platforms, such as Geographic Information Systems (GIS), to augment the versatility and utility of these systems.
- An issue particularly important for applications in the area of geospatial intelligence is AI security, where an adversary might, e.g., contaminate the training data for a (foundation) model.
- Assessing the performance and efficiency of SAM in real-time or near-real-time remote sensing applications to understand its capabilities for timely data processing and analysis.
- Exploring how domain-specific knowledge and expertise can be integrated into SAM to enhance its ability to understand and interpret remote sensing data.
- Evaluating the potential use of SAM as an alternative to traditional labeling processes and its integration with other image classification and segmentation techniques in a weakly-supervised manner to boost its accuracy and reliability.
- Integrating SAM with super resolution approach to enhance its capability to handle low-resolution imagery, thereby expanding the range of remote sensing imagery it can effectively analyze.

6. Conclusions

In this study, we conducted a comprehensive analysis of both the zero and one-shot capabilities of the Segment Anything Model (SAM) in the domain of remote sensing imagery processing, benchmarking it against aerial and satellite datasets. Our analysis provided insights into the operational performance and efficacy of SAM in the sphere of remote sensing segmentation tasks. We concluded that, while SAM exhibits notable promise, there is a tangible scope for improvement, specifically in managing its limitations and refining its performance for task-specific implementations.

In summary, our data indicated that SAM delivers notable performance when contrasted with the ground-truth masks, thereby underscoring its potential efficacy as a significant resource for remote sensing applications. Our evaluation reveals that the prompt capabilities of SAM (text, point, box, and general), combined with its ability to perform object segmentation with minimal human supervision, can also contribute to a significant reduction in annotation workload. This decrease in human input during the labeling phase may lead to expedited training schedules for other methods, thus promoting more streamlined and cost-effective workflows.

The chosen datasets were also selected with the express purpose of representing a broad and diverse context at varying scales, rather than exemplifying complex or challenging scenarios. By focusing on more straightforward datasets, the study went in on the fundamental aspects of segmentation tasks, without the additional noise of overly complicated or intricate scenarios. In this sense, future research should be oriented towards improving SAM's capabilities and exploring its potential integration with other methods to address more complex and challenging remote sensing scenarios.

Nevertheless, despite the demonstrated generalization, there are certain limitations to be addressed. Under complex scenarios, the model faces challenges, leading to less optimal segmentation outputs, by overestimating most of the objects' boundaries. Additionally, SAM's performance metrics display variability contingent on the spatial resolution of the input imagery (i.e., being prone to increase mistakes as the spatial resolution of the imagery is lowered). Consequently, identifying and rectifying these constraints is essential for further enhancing SAM's applicability within the remote sensing domain.

CRediT authorship contribution statement

Lucas Prado Osco: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **Qiusheng Wu:** Methodology, Software, Writing – review & editing. **Eduardo Lopes de Lemos:** Data curation, Methodology. **Wesley Nunes Gonçalves:** Methodology, Writing – review & editing. **Ana Paula Marques Ramos:** Validation, Visualization, Writing – review & editing. **Jonathan Li:** Validation, Visualization, Writing – review & editing. **José Marcato Junior:** Validation, Supervision, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Here, we provide an open-access repository designed to facilitate the application of the Segment Anything Model (SAM) within the domain of remote sensing imagery. The incorporated codes and packages provide users the means to implement point and bounding box-based shapefiles in combination with the SAM. The repositories also include notebooks

that demonstrate how to apply the text-based prompt approach, alongside one-shot modifications of SAM. These resources aim to bolster the usability of the SAM approach in diverse remote sensing contexts, and can be accessed via the following online repositories: GitHub:AI-RemoteSensing (Osco, 2023) and; GitHub:Segment-Geospatial (Wu and Osco, 2023).

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil - Finance Code 001. The authors are funded by the Support Foundation for the Development of Education, Science, Technology of the State of Mato Grosso do Sul, Brazil (FUNDECT; 71/009.436/2022), the Brazilian National Council for Scientific and Technological Development (CNPq; 433783/2018-4, 310517/2020-6; 405997/2021-3; 308481/2022-4; 305296/2022-1), and CAPES Print, Brazil (88881.311850/2018-01).

References

Adam, J.M., Liu, W., Zang, Y., Afzal, M.K., Bello, S.A., Muhammad, A.U., Wang, C., Li, J., 2023. Deep learning-based semantic segmentation of urban-scale 3D meshes in remote sensing: A survey. Int. J. Appl. Earth Obs. Geoinf. 121, 103365. http://dx.doi.org/10.1016/j.jag.2023.103365.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K., 2022. Flamingo: a visual language model for few-shot learning. arXiv:2204.14198.

Aleissaee, A.A., Kumar, A., Anwer, R.M., Khan, S., Cholakkal, H., Xia, G.-S., Khan, F.S., 2023. Transformers in remote sensing: A survey. Remote Sens. 15 (7), 1860. http://dx.doi.org/10.3390/rs15071860.

Amani, M., Ghorbanian, A., Ahmadi, S.A., Kakooei, M., Moghimi, A., Mirmazloumi, S.M., Moghaddam, S.H.A., Mahdavi, S., Ghahremanloo, M., Parsian, S., Wu, Q., Brisco, B., 2020. Google earth engine cloud computing platform for remote sensing big data applications: A comprehensive review. IEEE J. Select. Top. Appl. Earth Observations Remote Sens. 13, 5326–5350. http://dx.doi.org/10.1109/jstars.2020.3021052.

Bai, Y., Zhao, Y., Shao, Y., Zhang, X., Yuan, X., 2022. Deep learning in different remote sensing image categories and applications: status and prospects. Int. J. Remote Sens. 43 (5), 1800–1847. http://dx.doi.org/10.1080/01431161.2022.2048319.

Benjdira, B., Bazi, Y., Koubaa, A., Ouni, K., 2019. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. Remote Sens. 11 (11), 1369. http://dx.doi.org/10.3390/rs11111369.

Boguszewski, A., Batorski, D., Ziemba-Jankowska, N., Dziedzic, T., Zambrzycka, A., 2022. LandCover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. arXiv:2005.02264.

Bressan, P.O., Junior, J.M., Martins, J.A.C., de Melo, M.J., Gonçalves, D.N., Freitas, D.M., Ramos, A.P.M., Furuya, M.T.G., Osco, L.P., de Andrade Silva, J., Luo, Z., Garcia, R.C., Ma, L., Li, J., Gonçalves, W.N., 2022. Semantic segmentation with labeling uncertainty and class imbalance applied to vegetation mapping. Int. J. Appl. Earth Obs. Geoinf. 108, 102690. http://dx.doi.org/10.1016/j.jag.2022.102690.

Chi, M., Plaza, A., Benediktsson, J.A., Sun, Z., Shen, J., Zhu, Y., 2016. Big data for remote sensing: Challenges and opportunities. Proce. IEEE 104 (11), 2207–2219. http://dx.doi.org/10.1109/jproc.2016.2598228.

de Carvalho, O.L.F., Júnior, O.A.d., e Silva, C.R., de Albuquerque, A.O., Santana, N.C., Borges, D.L., Gomes, R.A.T., Guimarães, R.F., 2022. Panoptic segmentation meets remote sensing. Remote Sens. 14 (4), 965. http://dx.doi.org/10.3390/rs14040965.

European Space Agency, 2023. SkySat - EOGateway. URL https://earth.esa.int/eogateway/missions/SkySat.

Gao, K., Chen, M., Narges Fatholahi, S., He, H., Xu, H., Marcato Junior, J., Nunes Gonçalves, W., Chapman, M.A., Li, J., 2021. A region-based deep learning approach to instance segmentation of aerial orthoimagery for building rooftop extraction. Geomatica 75 (3), 148–164. http://dx.doi.org/10.1139/geomat-2021-0009, arXiv:https://cdnsciencepub.com/doi/pdf/10.1139/geomat-2021-0009.

Gharibbafghi, Z., Tian, J., Reinartz, P., 2018. Modified superpixel segmentation for digital surface model refinement and building extraction from satellite stereo imagery. Remote Sens. 10 (11), 1824. http://dx.doi.org/10.3390/rs10111824.

Gómez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: A review. ISPRS J. Photogram. Remote Sens. 116, 55–72. http://dx.doi.org/10.1016/j.isprsjprs.2016.03.008.

Gonçalves, D.N., Marcato, J., Carrilho, A.C., Acosta, P.R., Ramos, A.P.M., Gomes, F.D.G., Osco, L.P., da Rosa Oliveira, M., Martins, J.A.C., Damasceno, G.A., de Araújo, M.S., Li, J., Roque, F., de Faria Peres, L., Gonçalves, W.N., Libonati, R., 2023. Transformers for mapping burned areas in Brazilian pantanal and amazon with PlanetScope imagery. Int. J. Appl. Earth Obs. Geoinf. 116, 103151. http://dx.doi.org/10.1016/j.jag.2022.103151.

Hossain, M.D., Chen, D., 2019. Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. ISPRS J. Photogramm. Remote Sens. 150, 115–134. http://dx.doi.org/10.1016/j.isprsjprs.2019.02.009.

Hua, X., Wang, X., Rui, T., Shao, F., Wang, D., 2021. Cascaded panoptic segmentation method for high resolution remote sensing image. Appl. Soft Comput. 109, 107515. http://dx.doi.org/10.1016/j.asoc.2021.107515.

IDEA-Research, 2023. Grounded-segment-anything. URL https://github.com/IDEA-Research/Grounded-Segment-Anything.

Jozdani, S., Chen, D., Pouliot, D., Johnson, B.A., 2022. A review and meta-analysis of generative adversarial networks and their applications in remote sensing. Int. J. Appl. Earth Obs. Geoinf. 108, 102734. http://dx.doi.org/10.1016/j.jag.2022.102734.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment anything. arXiv:2304.02643.

Kotaridis, I., Lazaridou, M., 2021. Remote sensing image segmentation advances: A meta-analysis. ISPRS J. Photogramm. Remote Sens. 173, 309–322. http://dx.doi.org/10.1016/j.isprsjprs.2021.01.020.

Li, X., Deng, J., Fang, Y., 2022a. Few-shot object detection on remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–14. http://dx.doi.org/10.1109/tgrs.2021.3051383.

Li, X., Ding, H., Zhang, W., Yuan, H., Pang, J., Cheng, G., Chen, K., Liu, Z., Loy, C.C., 2023a. Transformer-based visual segmentation: A survey. arXiv:2304.09854.

Li, K., Hu, X., Jiang, H., Shu, Z., Zhang, M., 2020. Attention-guided multi-scale segmentation neural network for interactive extraction of region objects from high-resolution satellite imagery. Remote Sens. 12 (5), 789. http://dx.doi.org/10.3390/rs12050789.

Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y., 2023b. UniFormer: Unifying convolution and self-attention for visual recognition. arXiv:2201.09450.

Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.-W., Gao, J., 2022b. Grounded language-image pre-training. arXiv:2112.03857.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Zhou, J., 2023a. RemoteCLIP:A vision language foundation model for remote sensing. arXiv:2306.11029.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L., 2023b. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv:2303.05499.

Lobry, S., Marcos, D., Murray, J., Tuia, D., 2020. RSVQA: Visual question answering for remote sensing data. IEEE Trans. Geosc. Remote Sens. 58 (12), 8555–8566. http://dx.doi.org/10.1109/tgrs.2020.2988782.

Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic gradient descent with warm restarts. arXiv:1608.03983.

Ma, A., Wang, J., Zhong, Y., Zheng, Z., 2022. FactSeg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 60, 1–16. http://dx.doi.org/10.1109/tgrs.2021.3097148.

Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., Cundy, C., Li, Z., Zhu, R., Lao, N., 2023. On the opportunities and challenges of foundation models for geospatial artificial intelligence. arXiv:2304.06798.

Martins, J.A.C., Nogueira, K., Osco, L.P., Gomes, F.D.G., Furuya, D.E.G., Gonçalves, W.N., Sant'Ana, D.A., Ramos, A.P.M., Liesenberg, V., dos Santos, J.A., de Oliveira, P.T.S., Junior, J.M., 2021. Semantic segmentation of tree-canopy in urban environment with pixel-wise deep learning. Remote Sens. 13 (16), 3054. http://dx.doi.org/10.3390/rs13163054.

Mialon, G., Dessí, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., Scialom, T., 2023. Augmented language models: a survey. arXiv:2302.07842.

Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N., Terzopoulos, D., 2021. Image segmentation using deep learning: A survey. IEEE Trans. Pattern Anal. Machine Intell. 1. http://dx.doi.org/10.1109/tpami.2021.3059968.

OpenAI, 2023. GPT-4 technical report. arXiv:2303.08774.

Osco, L., 2023. AI-RemoteSensing: a collection of jupyter and google colaboratory notebooks dedicated to leveraging artificial intelligence (AI) in remote sensing applications. http://dx.doi.org/10.5281/zenodo.8092269.

Osco, L.P., dos Santos de Arruda, M., Junior, J.M., da Silva, N.B., Ramos, A.P.M., Moryia, É.A.S., Imai, N.N., Pereira, D.R., Creste, J.E., Matsubara, E.T., Li, J., Gonçalves, W.N., 2020. A convolutional neural network approach for counting and geolocating citrus-trees in UAV multispectral imagery. ISPRS J. Photogramm. Remote Sens. 160, 97–106. http://dx.doi.org/10.1016/j.isprsjprs.2019.12.010.

Osco, L.P., Junior, J.M., Ramos, A.P.M., de Castro Jorge, L.A., Fatholahi, S.N., de Andrade Silva, J., Matsubara, E.T., Pistori, H., Gonçalves, W.N., Li, J., 2021a. A review on deep learning in UAV remote sensing. Int. J. Appl. Earth Obs. Geoinf. 102, 102456. http://dx.doi.org/10.1016/j.jag.2021.102456.

Osco, L.P., Nogueira, K., Ramos, A.P.M., Pinheiro, M.M.F., Furuya, D.E.G., Gonçalves, W.N., de Castro Jorge, L.A., Junior, J.M., dos Santos, J.A., 2021b. Semantic segmentation of citrus-orchard using deep neural networks and multispectral UAV-based imagery. Precis. Agricul. 22 (4), 1171–1188. http://dx.doi.org/10.1007/s11119-020-09777-5.

Powers, D.M.W., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv:2010.16061.

Qurratulain, S., Zheng, Z., Xia, J., Ma, Y., Zhou, F., 2023. Deep learning instance segmentation framework for burnt area instances characterization. Int. J. Appl. Earth Obs. Geoinf. 116, 103146. http://dx.doi.org/10.1016/j.jag.2022.103146.

Rahman, M.A., Wang, Y., 2016. Optimizing intersection-over-union in deep neural networks for image segmentation. In: Advances in Visual Computing. Springer International Publishing, pp. 234–244. http://dx.doi.org/10.1007/978-3-319-50835-1_22.

Song, Y., Kalacska, M., Gašparović, M., Yao, J., Najibi, N., 2023. Advances in geocomputation and geospatial artificial intelligence (GeoAI) for mapping. Int. J. Appl. Earth Obs. Geoinf. 120, 103300. http://dx.doi.org/10.1016/j.jag.2023.103300.

Su, H., Wei, S., Yan, M., Wang, C., Shi, J., Zhang, X., 2019. Object detection and instance segmentation in remote sensing imagery based on precise mask R-CNN. In: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1454–1457. http://dx.doi.org/10.1109/igarss.2019.8898573.

Sun, X., Wang, B., Wang, Z., Li, H., Li, H., Fu, K., 2021. Research progress on few-shot learning for remote sensing image interpretation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 2387–2402. http://dx.doi.org/10.1109/jstars.2021.3052869.

Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237, 111322. http://dx.doi.org/10.1016/j.rse.2019.111322.

Toth, C., Jóźków, G., 2016. Remote sensing platforms and sensors: A survey. ISPRS J. Photogramm. Remote Sens. 115, 22–36. http://dx.doi.org/10.1016/j.isprsjprs.2015.10.004.

Wang, S., Chen, W., Xie, S.M., Azzari, G., Lobell, D.B., 2020a. Weakly supervised deep learning for segmentation of remote sensing imagery. Remote Sens. 12 (2), 207. http://dx.doi.org/10.3390/rs12020207.

Wang, Y., Lv, H., Deng, R., Zhuang, S., 2020b. A comprehensive survey of optical remote sensing image segmentation methods. Can. J. Remote Sens. 46 (5), 501–531. http://dx.doi.org/10.1080/07038992.2020.1805729.

Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y., 2022. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. arXiv:2110.08733.

Wu, Z., Hou, B., Ren, B., Ren, Z., Wang, S., Jiao, L., 2021. A deep detection network based on interaction of instance segmentation and object detection for SAR images. Remote Sens. 13 (13), 2582. http://dx.doi.org/10.3390/rs13132582.

Wu, Q., Osco, L.P., 2023. samgeo: A Python package for segmenting geospatial data with the Segment Anything Model (SAM). Zenodo, http://dx.doi.org/10.5281/ZENODO.7966658.

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N., 2023. Visual ChatGPT: Talking, drawing and editing with visual foundation models. arXiv:2303.04671.

Xu, Y., Bai, T., Yu, W., Chang, S., Atkinson, P.M., Ghamisi, P., 2023. AI security for geoscience and remote sensing: Challenges and future trends. IEEE Geosci. Remote Sens. Mag. 11 (2), 60–85. http://dx.doi.org/10.1109/mgrs.2023.3272825.

Yang, D., Li, Z., Xia, Y., Chen, Z., 2015. Remote sensing image super-resolution: Challenges and approaches. In: 2015 IEEE International Conference on Digital Signal Processing (DSP). IEEE, pp. 196–200. http://dx.doi.org/10.1109/icdsp.2015.7251858.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 111716. http://dx.doi.org/10.1016/j.rse.2020.111716.

Yuan, X., Shi, J., Gu, L., 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Syst. Appl. 169, 114417. http://dx.doi.org/10.1016/j.eswa.2020.114417.

Zhang, J., Huang, J., Jin, S., Lu, S., 2023a. Vision-language models for vision tasks: A survey. arXiv:2304.00685.

Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Gao, P., Li, H., 2023b. Personalize segment anything model with one shot. arXiv:2305.03048.

Zhang, X., Jin, J., Lan, Z., Li, C., Fan, M., Wang, Y., Yu, X., Zhang, Y., 2020. ICENET: A semantic segmentation deep network for river ice by fusing positional and channel-wise attentive features. Remote Sens. 12 (2), 221. http://dx.doi.org/10.3390/rs12020221.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.-Y., 2022a. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection. arXiv:2203.03605.

Zhang, R., Li, G., Wunderlich, T., Wang, L., 2021. A survey on deep learning-based precise boundary recovery of semantic segmentation for images and point clouds. Int. J. Appl. Earth Obs. Geoinf. 102, 102411. http://dx.doi.org/10.1016/j.jag.2021.102411.

Zhang, H., Zhang, P., Hu, X., Chen, Y.-C., Li, L.H., Dai, X., Wang, L., Yuan, L., Hwang, J.-N., Gao, J., 2022b. GLIPv2: Unifying localization and vision-language understanding. arXiv:2206.05836.

Zhang, Z., Zhang, Q., Hu, X., Zhang, M., Zhu, D., 2023c. On the automatic quality assessment of annotated sample data for object extraction from remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 201, 153–173. http://dx.doi.org/10.1016/j.isprsjprs.2023.05.026.

Zheng, Z., Zhong, Y., Wang, J., Ma, A., 2020. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. arXiv:2011.09766.

Zia, U., Riaz, M.M., Ghafoor, A., 2022. Transforming remote sensing images to textual descriptions. Int. J. Appl. Earth Obs. Geoinf. 108, 102741. http://dx.doi.org/10.1016/j.jag.2022.102741.