Study the Segment Anything Model (SAM) for remote sensing image segmentation.

Group 26
Nhữ Tùng Lâm (BI12-233)
Lê Hoàng Long (BI12-251)
Phạm Thái Dương (BA12-057)
Nguyễn Vũ Thế Bình (BA12-022)
Võ Trường Giang (BI12-130)
Phan Duy Long (BI12-252)

University of Science and Technology
of Hanoi

1. Introduction

1.1 What is Segment Anything Model (SAM):

- Released by Meta in April 2023,Segment Anything Model (SAM) is a cutting-edge image segmentation model that enables promptable segmentation, offering exceptional flexibility in image analysis. It is part of the Segment Anything initiative, which introduces a new model, task, and dataset for image segmentation.
- SAM's design allows it to handle new image distributions and tasks without prior training, a feature called zero-shot transfer. It was trained on the extensive SA-1B dataset, which includes over 1 billion masks across 11 million images.

1.2 What is remote sensing:

- Remote means something that is not exactly in contact or physical contact, Sensing means getting information, data, something like temperature, pressure, photograph, etc.

- Remote sensing is the process of acquiring information, detecting, analyzing, and monitoring an area's physical characteristics by recording how it is reflected and emitted radiation energy without having any physical contact with the object under study. This is done by capturing the reflected radiation/energy.

1.3 Advantages of SAM in remote sensing applications:

- Generalization Across Image Types: SAM handles various aerial and satellite images, adapting to different environments like urban, forest, and ocean regions.
- Zero-Shot Learning: SAM performs segmentation without task-specific training, vital for remote sensing where labeled data is limited and expensive.
- Minimal Human Input: SAM uses simple prompts like bounding boxes or single points for segmentation, reducing manual annotation needs.
- Efficiency with High-Resolution Data: SAM effectively segments complex features and fine details in high-resolution remote sensing images.

2. Remote Sensing Image Segmentation Overview

2.1 About Remote Sensing Image Segmentation

- Remote sensing image segmentation refers to the process of dividing images into meaningful regions or objects, essential for analyzing geographical data. Traditional techniques, such as pixel-based or object-based approaches, often struggle to handle the complexity and high resolution of modern remote sensing imagery, requiring substantial human intervention. These limitations become more evident with the increasing detail and variability in aerial and satellite imagery.

- To address these challenges, the Segment Anything Model (SAM) leverages deep learning to effectively handle segmentation tasks, offering flexibility and adaptability. SAM stands out by integrating a variety of prompts (e.g., bounding boxes, points, or textual descriptions) to guide the segmentation process, thus making it more adaptable to diverse tasks without the need for extensive labeled data.

- Key Segmentation Techniques:
  - Interactive Segmentation: Involves user input to refine segmentation for improved precision.
  - Superpixel Segmentation: Groups pixels into larger units based on similar characteristics, simplifying data while maintaining object structure.
  - Semantic Segmentation: Classifies every pixel in the image into specific classes.
  - Instance Segmentation: Differentiates individual objects within the same class, treating them as distinct entities.
- SAM excels in "Promptable Segmentation," where prompts guide the model in performing various segmentation tasks dynamically. This approach significantly enhances its applicability to remote sensing, allowing for accurate and flexible segmentation with minimal human intervention. SAM’s ability to generalize across different objects and scenes makes it a valuable tool for efficiently processing large-scale remote sensing imagery.

2.2 Comparison of Object and Remote Sensing Imagery Using SAM

Factor	Object Imagery	Remote Sensing Imagery
Resolution & Scope	High resolution, small scope	Variable resolution, wide geographical coverage
Perspective	Direct or close-up view, minimal perspective distortion	Aerial view, geometric distortions due to angle and Earth's curvature
Object Type	Clear, distinct objects, rarely occluded	Diverse, complex objects, often occluded or blended with background
Preprocessing	Minimal, usually brightness/contrast adjustments	Complex preprocessing like geometric/radiometric corrections
Segmentation Mode	Bounding box or points suffice for accurate segmentation	Requires multiple segmentation modes for optimal results
Postprocessing	Simple, mainly focused on mask smoothing and classification	Complex, involves spatial information extraction and geographical analysis

This comparison highlights SAM's adaptability, enabling it to handle both object-level and large-scale remote sensing imagery effectively, even with the increased complexity of geographic data.

3. Materials and Methods

3.1. Data Description

3.1.1 General Information

- **Purpose**: This step involves gathering and organizing different types of remote sensing images that will be used in the study. The diversity of data sources helps ensure that the evaluation covers various scenarios.
- **Case Study**: Building Recognition Using the Segment Anything Model (SAM)

3.1.2 Material

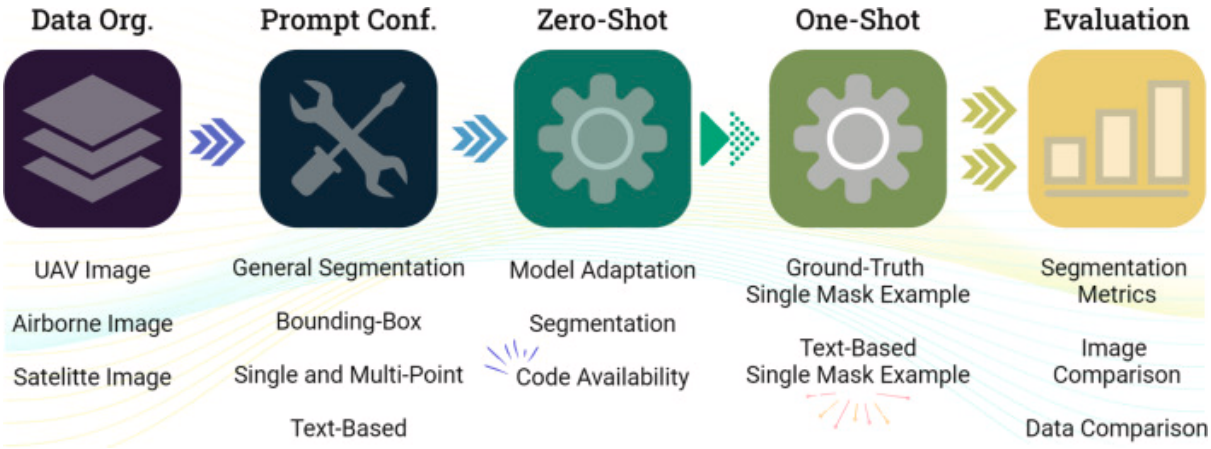
satellite.png

Image Size: 1270x768

Image Data Type: 24 bit

Resolution (dpi): 96

3.2. Method



3.2.1 Data Preparation:

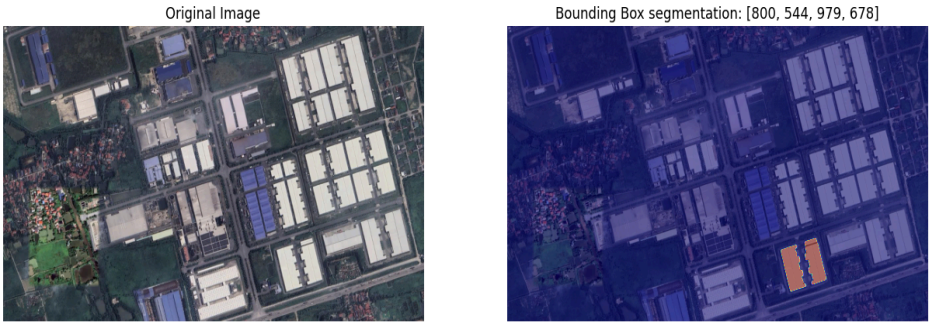
- **Image format**: SAM supports common image formats like JPG, PNG, and TIFF. Ensure your remote-sensing images are in one of these formats.
- **Image size:** SAM can handle various input sizes. However, very large images might slow down processing. You may need to resize the image to fit your computational resources.

3.2.2 Preprocessing

- Depending on the quality of the remote sensing images and your specific requirements, you might need to perform some preprocessing steps such as:
  + **Geometric correction**: Remove geometric distortions caused by the angle of capture and the Earth's curvature.
  + **Radiometric correction**: Adjust brightness and contrast to enhance image quality.
  + **Noise removal**: Minimize random or systematic noise in the images.
  + **Resolution enhancement**: Improve image sharpness using interpolation or super-resolution techniques.

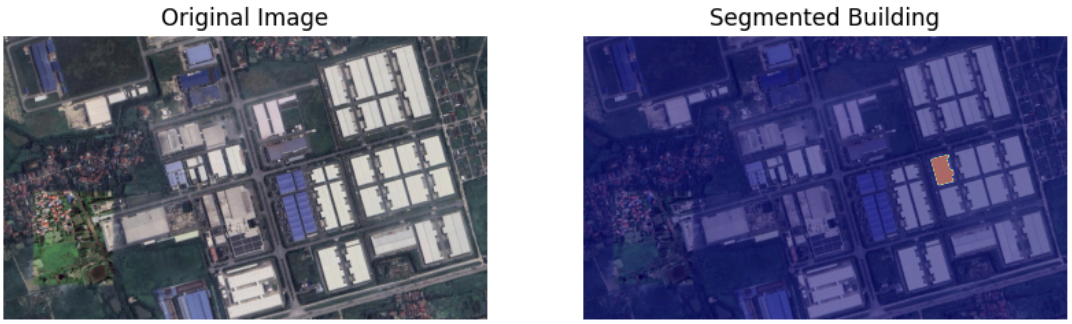
3.2.3 Using SAM to Generate Masks:

- Choose segmentation mode: SAM provides three segmentation modes:
- Bounding box prompt: You provide one or more bounding boxes around the objects you want to segment.



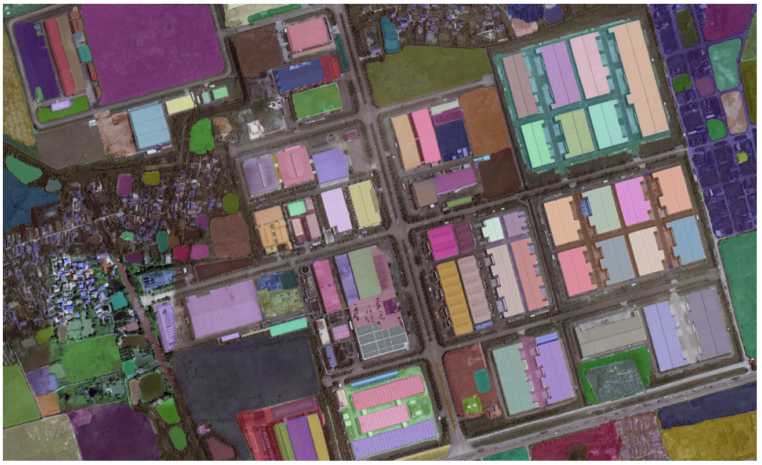
Segment two objects buildings by a given bounding box

- Point prompt: You specify foreground and background points on the image.

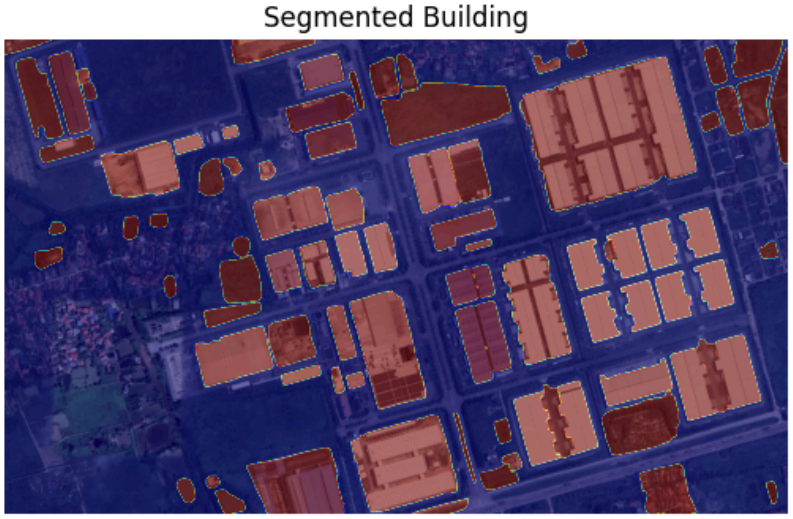


Segment an object building by a specific point

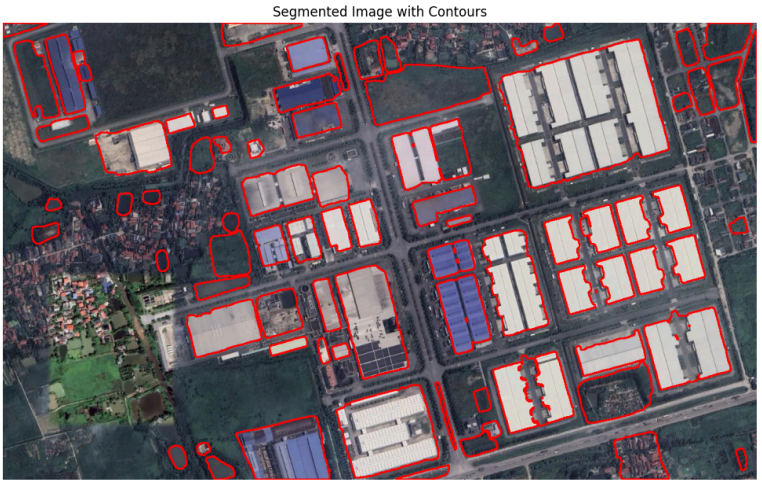
- Automatic mode (everything): SAM automatically segments all objects in the image.



Objects Recognition with random colors using the Segment Anything Model



The model segments object large buildings in red color. Note that the model is not working well yet so it misrecognized with some large ground surround



The model segments object large buildings with red contours.

- Refine masks: SAM can generate multiple masks for different objects in the image. You can select the most appropriate mask or combine masks for more accurate segmentation results.

3.2.4 Post-processing

- Smooth masks: Remove small isolated regions or jagged edges on the mask for a smoother segmentation result.
- Object classification: Use image classification techniques to assign labels to the segmented objects.
- Extract information: Extract useful information from the segmentation results, such as the area, perimeter, and shape of the objects.

3.2.5 Zero shot learning

- **Definition:** 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. Essentially, SAM can generalize from its training data to handle new, unseen classes or objects at inference time, without needing additional labeled examples of those classes.
- **How does it work?**
  + **Learning on Seen Classes:** During training, the model learns to recognize and classify a set of "seen" classes (i.e., categories it has been explicitly exposed to). However, instead of focusing solely on memorizing examples, it also learns semantic representations or attributes associated with each class. These could be high-level descriptions like color, shape, or textual information.
  + **Defining Unseen Classes:** The unseen classes are defined by semantic descriptors or attributes that are shared with seen classes. This allows the model to infer knowledge about these unseen categories.
  + **Inference on Unseen Classes:** When the model is presented with new data from unseen classes during testing or deployment, it matches the new input with the semantic attributes or relationships it learned during training. The key is that the unseen class shares similarities with seen classes through these attributes or semantics.

3.2.6 One Shot learning

- **Definition:** One-shot learning denotes a model's ability to interpret and make accurate inferences from just a single example of a new class. 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.
- **How does it work?**
  + **Feature Extraction:** The model doesn't directly learn from the raw input (e.g., images). Instead, it extracts high-level features that are crucial for distinguishing different classes or objects. These features could be things like edges, textures, shapes, or color patterns in images, or meaningful phrases and relationships in text data.

+ **Metric Learning:** Metric learning is a key technique where the model learns to measure the distance or similarity between examples in a feature space. The idea is that examples from the same class should be close to each other in this space, while examples from different classes should be far apart. Instead of learning to classify an image directly, the model learns to determine whether two images (or instances) are the same or different. This allows it to generalize to new classes by comparing features, even with minimal data.
+ **Siamese Networks:** Siamese networks are a popular architecture for one-shot learning. They consist of two identical neural networks that share the same weights and parameters. These twin networks process two inputs simultaneously, comparing their feature representations and calculating the similarity between them. The network is trained to minimize the difference between features of the same class and maximize the difference for different classes. After training, the model can compare any new example with a reference example to determine whether they belong to the same class, based on the similarity score.

3.3 Evaluation

- **Segmentation Metrics**: The segmentation output from SAM is evaluated using specific metrics to measure its accuracy and performance.
- **Image Comparison**: The results are visually compared with ground truth or other benchmarks to assess the quality of the segmentation.
- **Data Comparison**: The outputs are compared across different datasets or segmentation methods to see how SAM performs under various conditions.
- 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:**

+ **Intersection over Union (IoU)** is a common evaluation metric for object detection and segmentation problems. The IoU is the area of

overlap divided by the area of the union of the predicted and ground truth segmentation. The equation to achieve it is presented as:

$$\text{IoU} = \frac{TP}{TP+FP+FN} \quad (1)$$

+ **Pixel Accuracy** is the simplest used metric and it measures the percentage of pixels that were accurately classified. 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:

$$\textbf{Pixel Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

+ **Dice Coefficient** is another metric used to gauge the performance of image segmentation methods. The Dice Coefficient ranges from 0 (no overlap) to 1 (perfect overlap). The equation to perform it is given as follows:

$$\textbf{Dice Coefficient} = 2 * \frac{TP}{2*TP+FP+FN} \quad (3)$$

+ 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. Both metrics are calculated as:

$$\textbf{TPR} = \frac{TP}{(TP+FN)} \quad (4)$$

$$\textbf{FPR} = \frac{FP}{(FP+FN)} \quad (5)$$

***Here**, TP represents True Positives (the correctly identified positives)*

FP represents False Positives (the incorrectly identified positives)

FN represents False Negatives (the positives that were missed)

TN represents True Negatives (the correctly identified negatives)

4. Discussion and application

4.1 Discussion

- **Success with High-Resolution Images:**
  - SAM demonstrates the ability to accurately separate objects in sharp images.
  - With high-quality images, SAM easily recognizes and distinguishes object details.
  - The model has demonstrated the ability to accurately assess image features in clear visual contexts.
- **Challenges with Low-Resolution Images:**
  - When faced with blurry images, SAM has difficulty determining the precise boundaries of objects.
  - With low resolution, image details are lost, making it difficult for the model to distinguish objects.
  - SAM's segmentation ability is limited when applied to low-resolution satellite images.

- **Zero-Shot Segmentation**
  - **Impact of Prompts:**
    + The type of prompts provided to the model directly affects the segmentation results.
    + Points are an effective choice for identifying small objects, while bounding boxes are more suitable for large objects.
    + The variety of prompts allows users to customize the segmentation process.
  - **Limitations of text prompts:**
    + Text descriptions are not enough for the model to accurately understand user intent, especially with low-quality images.
    + The model's ability to transform text into images is limited, leading to inaccurate segmentation results.

- **One-shot segmentation**
  - **Improved performance:**
    + Fine-tuning the model with a single example significantly improves segmentation performance, especially for complex objects.
    + The fast learning ability of SAM is clearly demonstrated by the one-shot segmentation method.

  - **PerSAM-F effectiveness:**
    + PerSAM-F is an effective method for handling objects with complex structures and diverse scales.
    + This method helps the model better capture the relationships between the constituent parts of the object.

- **Real-time performance:**
    + SAM has been shown to be capable of performing image segmentation quite quickly, opening up potential applications in real-time systems such as video tracking, robotics.
    + However, when processing high-resolution images or complex scenes, the processing speed of SAM can be significantly reduced. Further optimizations in network architecture and algorithms are needed to improve real-time performance.

- **Flexibility in selecting objects:**
  + SAM allows users to interact directly with the model to select objects to be segmented through the use of different prompts (points, bounding boxes, text).
  + Selecting the appropriate prompt to achieve the best results requires the user to have certain knowledge of the model and data. In addition, building effective text prompts remains a challenge.

4.2 Application

- **Urban Planning and Development**
  - Urban Growth Analysis: SAM can help analyze the expansion or contraction of built-up areas over time, thereby providing data for urban development and planning forecasting. From satellite images, SAM determines the location and area of newly built buildings, residential areas, and public infrastructure.
  - Detecting illegal construction areas: Monitoring and detecting construction works that are not permitted or not planned.

- **Post-disaster damage assessment**
  - Analyzing the extent of damage: SAM can be used to analyze satellite images before and after natural disasters (hurricanes, floods, earthquakes) to assess the extent of damage to buildings and infrastructure. This is very useful for planning rescue and reconstruction after natural disasters.
  - Estimating restoration costs: Based on satellite image analysis, it is possible to estimate the cost to restore damaged infrastructure.

- **Resource and environmental management**
  - Monitoring infrastructure development in sensitive areas: SAM can be used to detect construction works in areas that need to be preserved or in ecologically sensitive areas, such as near rivers, lakes, forests or nature reserves.
  - Monitoring forest encroachment: This application is very important in detecting and monitoring construction works that encroach on forest

land, thereby taking measures to protect forests and natural resources.

- **Infrastructure analysis and management**
  - Monitoring public infrastructure: SAM can help monitor the distribution and condition of public buildings such as schools, hospitals, bridges, roads, and other essential infrastructure.
  - Detecting deterioration of old buildings: Analyze satellite images periodically to detect signs of deterioration of old buildings and infrastructure that need to be maintained or replaced.

- **Real Estate Management**
  - Analyze land development potential: SAM helps identify vacant land or less developed areas so that real estate development can be planned or investment in potential areas can be made.
  - From satellite data, it is possible to analyze construction density and surrounding infrastructure to support real estate valuation.

5. Conclusion

SAM has demonstrated good adaptability in remote sensing image segmentation, including zero-shot and one-shot learning techniques. This flexibility allows SAM to be applied flexibly to different data types and scales, opening up new prospects for remote sensing workflows.SAM shows strong potential in segmentation tasks and reduces manual annotation workload, speeding up training and streamlining workflows. However, challenges arise in complex scenarios, where SAM tends to

overestimate object boundaries and its performance varies with image resolution. Future research should aim to improve SAM's capabilities and explore its integration with other methods to handle more complex remote sensing tasks.

References

https://www.sciencedirect.com/science/article/pii/S1569843223003643?ref=pdf_download&fr=RR-2&rr=8c97db3f9d43ddc8#fig5

https://github.com/facebookresearch/segment-anything?tab=readme-ov-file#model-checkpoints

https://docs.ultralytics.com/models/sam/

https://www.v7labs.com/blog/segment-anything-model-sam#how-does-sam-support-real-life-use-cases