One Prompt Fits All: Visual Prompt-Tuning for Remote Sensing Segmentation

ICLR

Climate Change AI

Spotlight

Marshall Wang,[1] John Willes,[1] Deval Pandya[1]

[1]Vector Institute

VECTOR INSTITUTE | INSTITUT VECTEUR

Background

Image segmentation is crucial in climate change research for analyzing satellite imagery.

- **Deforestation Monitoring**: Detect changes in forest cover over time, helping to pinpoint areas of illegal logging or deforestation.
- **Urban Planning**: Identify land use patterns, infrastructure, and green spaces, enabling better decision making and sustainable development.
- **Natural Disaster Response**: Quick analysis of areas affected by natural disasters to aid in efficient allocation of resources for relief efforts.
- And much more!

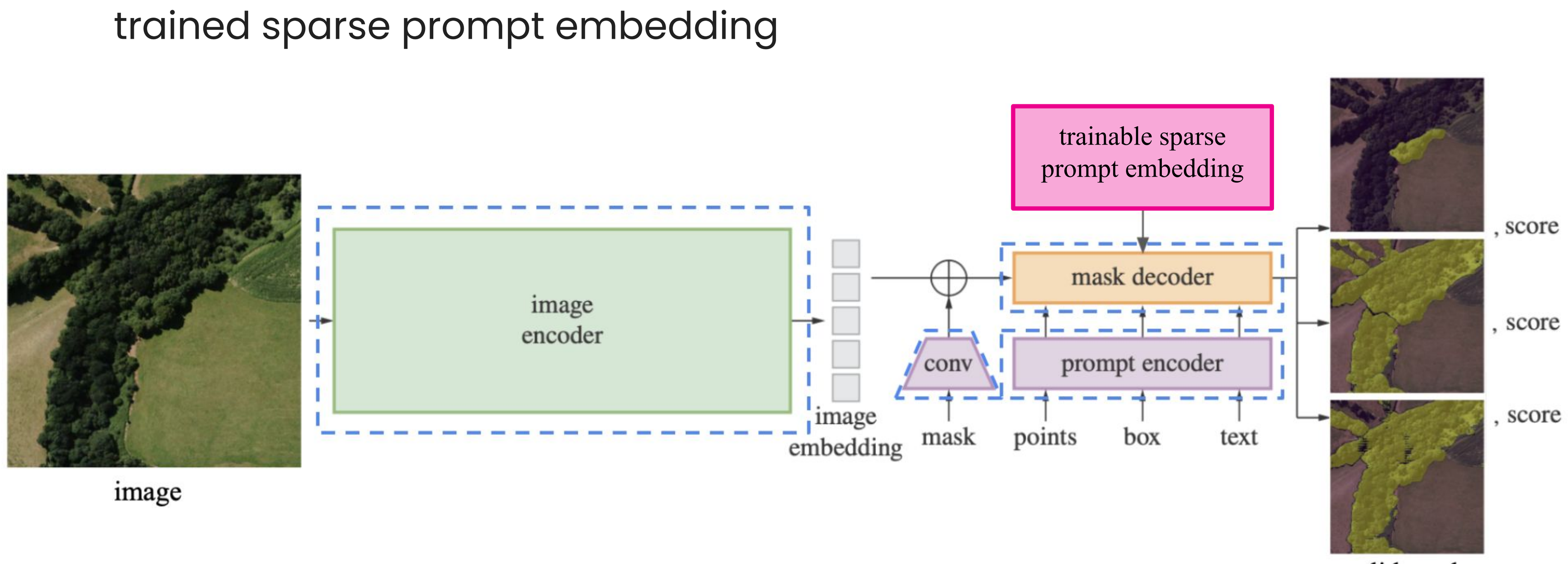
The advent of vision-based foundational models like the **Segment Anything Model (SAM)** opens new avenues in climate research and remote sensing (RS). SAM can perform zero-shot segmentation tasks from manually-crafted prompts

Motivation

- **The efficacy of SAM largely depends on the quality of input prompts.**
- This issue is particularly pronounced with RS data, which are inherently complex.
  - One would need to create **complex prompts** for each image, which typically involves selecting dozens of points or bounding boxes.
- **Hard to scale up!**

Prompt-Tuning SAM (PT-SAM)

- **A method that reduces the need for extensive manual input** by incorporating a trainable, lightweight prompt embedding to SAM
- **The embedding captures essential semantic information for specific objects**, enabling effective segmentation of unseen images without human intervention.
- Embedding training has **minimal dataset and hardware requirements**, making it accessible for widespread use → All experiments conducted on a single T4 GPU, training can also be conducted on CPU
- **Plug-and-play**: Users can easily replace the prompt encoder with various trained sparse prompt embedding



- Modules in blue dash boxes are frozen during training

Algorithm 1 Prompt-Tuning SAM

1: Pre-compute image embeddings and generate fixed dense prompt embedding
2: Initialize a random sparse prompt embedding
3: **for** each image **do**
4: Pass the trainable sparse prompt embedding, the fixed dense prompt embedding, and the pre-computed image embedding to the mask decoder to generate masks
5: Compare the loss (BCE + DICE) between the generated masks and the corresponding ground truth masks
6: Perform backpropagation on the learnable sparse prompt embedding based on the loss
7: **end for**

Results

- PT-SAM is baselined against SAM following the evaluation protocol used by the SAM authors
- The baseline method, referred to as Point-Prompted SAM (PP-SAM), uses the center point of the ground truth masks (or the closest point to the center on the masks) to prompt SAM.

Table 1: Forest Embedding Evaluation Metrics

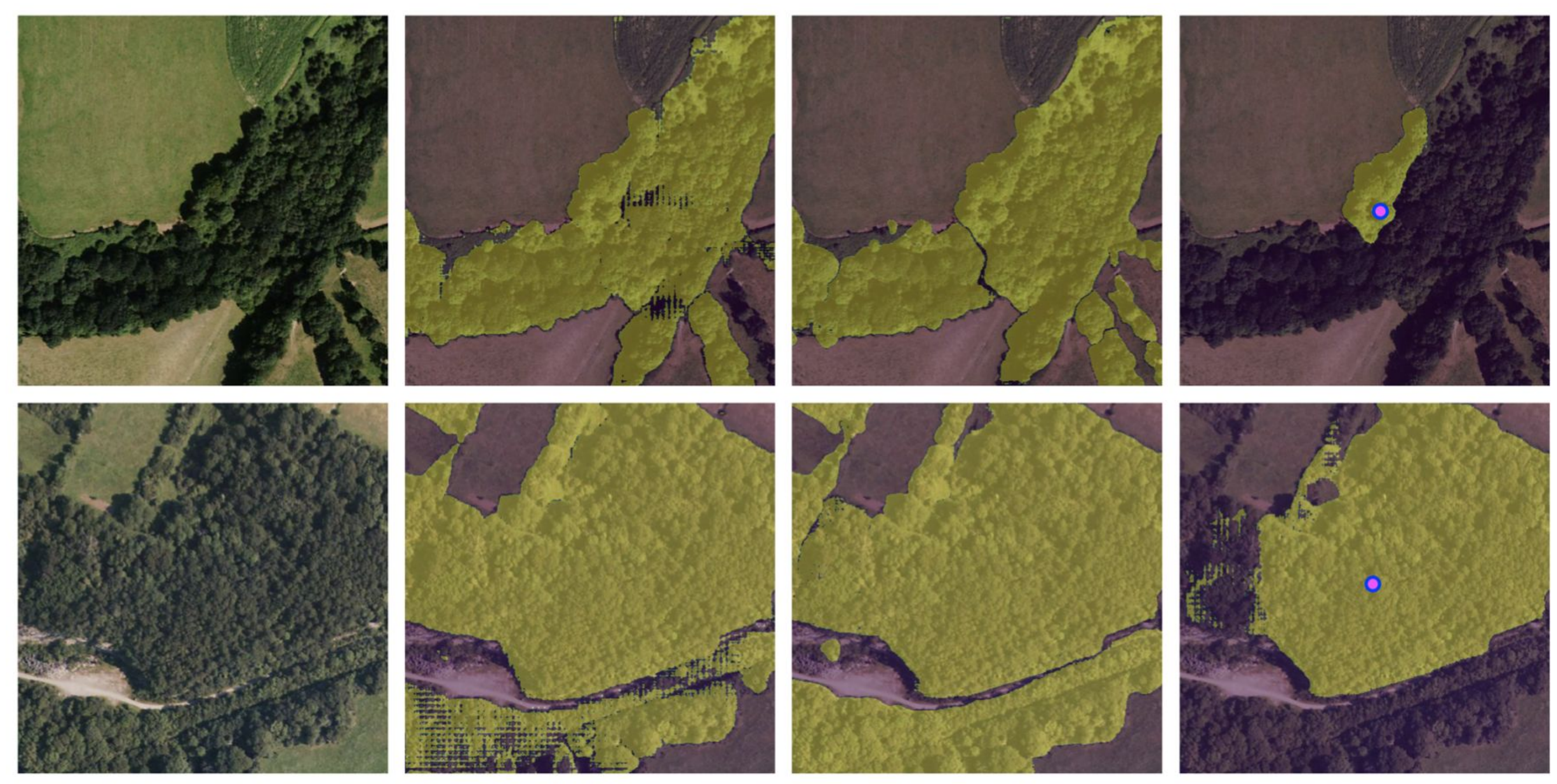
Average Scores	IoU	Accuracy	F1
PT-SAM (Ours)	**0.7639**	**0.8636**	**0.8558**
PP-SAM	0.4327	0.6220	0.4939

Table 2: Building Embedding Evaluation Metrics

Average Scores	IoU	Accuracy	F1
PT-SAM (Ours)	**0.6686**	**0.9209**	**0.7990**
PP-SAM	0.3256	0.4399	0.4747

Forest Embedding

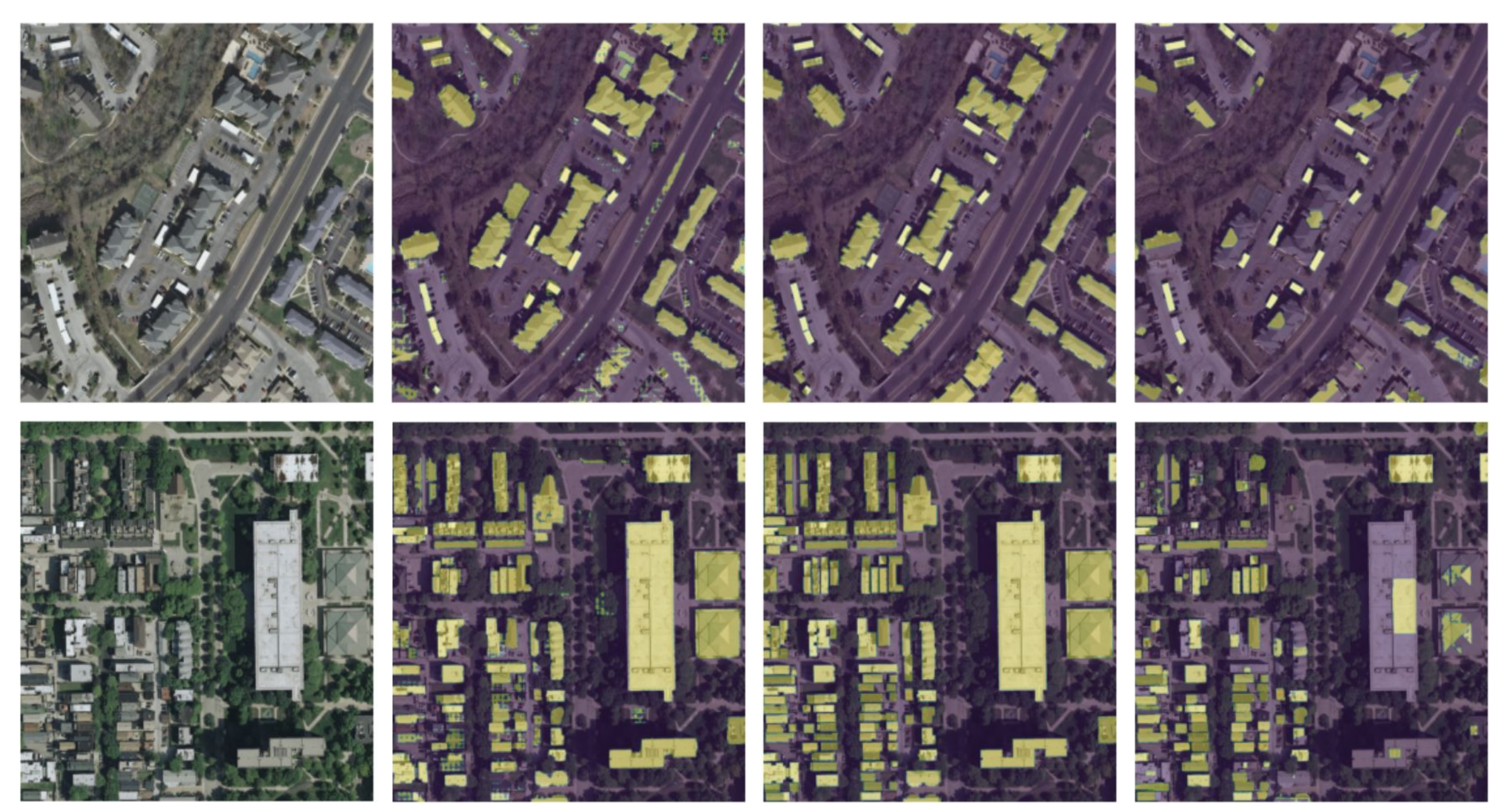
Trained on 100 images (500 x 500 pixels)



Original Image | PT-SAM | Ground Truth | PP-SAM

Building Embedding

Trained on 1000 images (1000 x 1000 pixels)



Original Image | PT-SAM | Ground Truth | PP-SAM

PT-SAM in Action - Deforestation Analysis