OPTIMIZED FIELD SAMPLING OF MARS-ANALOG SERPENTINE ZONES VIA MACHINE LEARNING. D.R. Jayakody[1], T.D. Ambegoda[1], S. Karunatillake[2], E.B. Hughes[3], [1]Department of Computer Science and Engineering, University of Moratuwa, Moratuwa, Sri Lanka, [2]Department of Geology and Geophysics, Louisiana State University, Baton Rouge, LA 70803, [3]Earth and Atmospheric Sciences Department, Georgia Institute of Technology, Atlanta, GA 30332.

Introduction: Earth's serpentine zones have been investigated as analogs to those on Mars [1, 2], especially given geologic conditions for habitability as well as insight on paleo-magnetism and planetary evolution [3]. However, field sites are often biased by past works (e.g., mentor-mentee generational transfer) with only limited mapping independent of prior knowledge. We consider a case study of independent mapping, along Sri Lanka's Mars-analog serpentine zone, at the litho-tectonic boundary between the Highland Complex and the Vijayan Complex (HC-VC boundary) [4]. The currently identified serpentinite field sites are limited to 6 discontinuous localities [4], of which only two contain public geo-specific coordinates [5].

To support further exploration of the serpentine zones of Sri Lanka, we propose Serp-Seg, a Python-based system to automatically detect previously unidentified serpentine zones along the HC-VC boundary. The system primarily contains a machine learning model which we've calibrated using previously identified serpentine landscapes (i.e., dominated areally by serpentinites and derived soils). The scripts required to use or calibrate this model shall be packaged into a Python library known as serp-seg before archival at NASA's Github repository.

Our system is expected to significantly reduce the time spent by geologists on localizing sampling sites for serpentine-derived soil. Furthermore, the system can be used for the detection (localizing on a given object on an image [6]) and segmentation (classifying the pixels of an image [7]) of any geological zones of interest, for which the number of labeled samples that can be used for training are very limited (a.k.a. few-shot [8] segmentation). Thus, it can also have resource prospecting utility for economically valuable minerals, before in situ sampling and analysis.

Dataset: The dataset used is based on the Sentinel-2 satellite imagery. The highest resolution bands of the dataset are of a resolution of 10m per pixel.

For calibrating our system, the frames of the Sentinel-2 dataset that cover the HC-VC boundary are divided into images of 256x256 pixels. Of these, the images of regions with in-situ verified serpentine exposure are separated, labeled, and used for training the model. Accordingly, this dataset consists of 40 labeled images containing serpentine sites and 1470 unlabeled images of candidate sites (the rest of the HC-VC boundary).

Methodology: There are two main indicators that can be used to identify the presence of serpentine zones: (a) infrared reflectance spectral features and

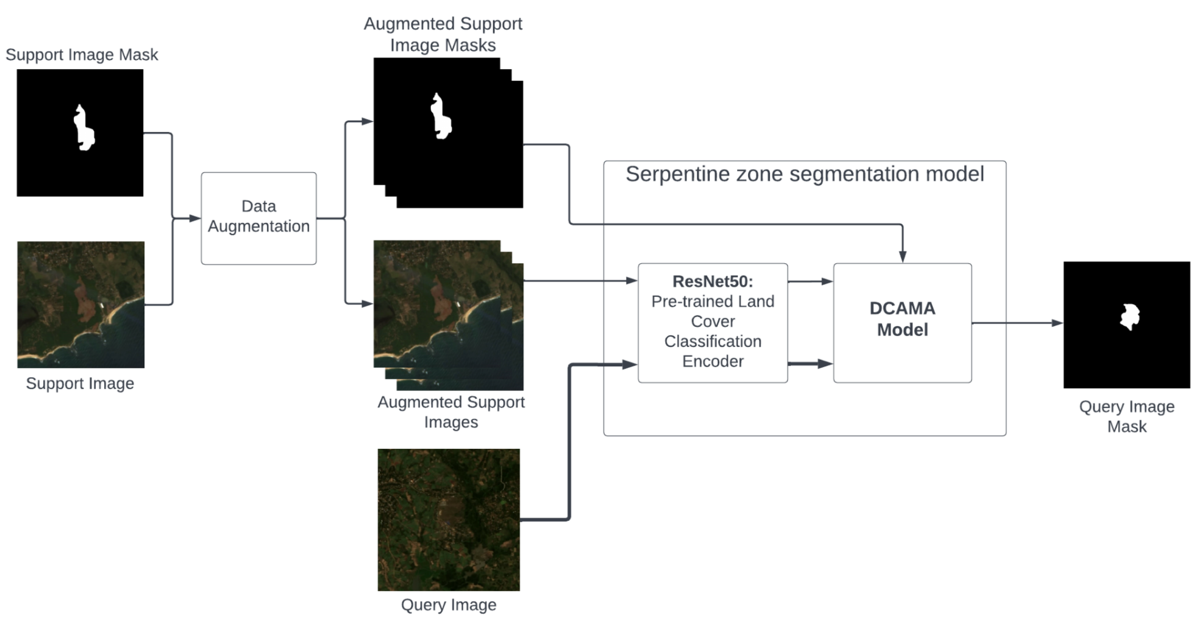

Figure 1: High-level architecture of the Serpentine Zone Segmentation System.

(b) the change in vegetation density when moving across a serpentine zone boundary [4], due in part to the common toxicity of serpentine soils to vegetation [9]. Details on both these indicators can be approximately inferred from the bands of the Sentinel-2 dataset (e.g.: bands 4 and 8 for obtaining the NDVI, bands 4 and 3 for Ferric ions, and bands 3, 4, 8, and 12 for Ferrous ions).

Our task is to identify whether a given image contains serpentine exposures and if so, their spatial distribution. To do so, we use a machine learning model known as DCAMA [10] (Figure 1). This specific model was chosen due to its strength in working with very few labeled images. Due to having just two geo-specific coordinates and six geological locations, the number of images that can be confidently labeled is highly limited for Lanka's serpentine zone.

DCAMA works by mapping the visual characteristics of the serpentine-rich landscapes in the labeled images to those of an unlabeled image (if any). To identify these visual characteristics, the image is passed through a secondary model known as an encoder [11]. The encoder is responsible for creating a numerical representation of the images, which are then used by DCAMA for segmenting.

To further improve the performance of the model, we replace the encoder of DCAMA with a model (ResNet50 [12]) which has been pretrained on the So2Sat [13, 14] landcover classification dataset.

(a) (b)

Figure 2: (a) Satellite Image of Ussangoda (b) Predicted serpentine zone

Calibration: To calibrate the system, we feed the model with 6 labeled images along with labels of 5 of the images (5-shot segmentation). The model then predicts a label for the remaining image. It is then calibrated against the actual label (Figure 2).

We calibrate the model for 1000 iterations over the entire dataset. This process takes approximately one hour on a computer with 16 GB of RAM and a GTX 1060 6GB GPU.

Evaluation: To evaluate the performance improvement of the DCAMA model, we compare it against a U-Net [15] model with the previously mentioned ResNet50 model as the encoder. We test the models both with and without the pretrained backbone to evaluate their contributions.

Both the models were evaluated using the Jaccard similarity index (a.k.a. intersection over union / IoU) and were cross-validated using 5 folds. The Jaccard similarity index is calculated by dividing the total number of correctly identified serpentine zone pixels by the total number of pixels in the union of the predicted serpentine-dominated pixels and the actual serpentine-dominated pixels (i.e., on an areal basis).

Comparison against regular approaches: It can be seen by the below metrics that the proposed approach improves over the regular segmentation approaches.

Table 1. Summary of the IoU scores of the different experiments, with higher score indicating greater accuracy.

ResNet50	ResNet50 with pretrained encoder	DCAMA	DCAMA with pretrained encoder
41.4	50.2	44.14	61.2

Outcome: Finally, we ran the model on the 1470 unlabeled images. At the specified resolution, this covers approximately 9630 km^2 of land along the HC-VC boundary. Running our program on this entire dataset (using a computer with 16 GB of RAM and a GTX 1060 6GB GPU) takes 4m 45s in the 5-shot scenario (feeding 5 labeled images with the unlabeled image) and 12m 55s in the 10-shot scenario (feeding 10 labeled images with the unlabeled image; more accurate). In other words, it takes 11.6 seconds per image in the 5-shot scenario and 31.6 seconds per image in the 10-shot scenario.

References: [1] Kumarathilaka P. et al. Appl. Geochem., 2016, [2] Vithanage M. et al. *Environ. Monit. Assess.*, 2014 [3] Karunatillake S. et al., *Terrestrial Analogs,* 2021. [4] Hewawasam et al., *JES*, 2014. [5] He X.F. et al. *Precambrian Res.,* 2016. [6] Licheng J., et al. *IEEE Access 7*, 2019. [7] Shervin M., et al. IEEE TPAMI, 2021. [8] Wang, Yaqing, et al. *ACM csur* 2020. [9] Kruckeberg A.R., Annu. Rev. Ecol. Evol. Syst., 1985. [10] Shi X. et al., *ECCV,* 2022 [11] Garcia-Garcia A., *arXiv preprint*, 2017. [12] He K. et al., *CVPR*, 2016. [13] Neumann M. *arXiv preprint*, 2019. [14] Zhu X. X. et al., *IEEE GRSM,* 2020. [15] Ronneberger O. et al, *MICCAI* 2015.