Comparison of Boulder Detection Performance on Asteroids Using Fine-Tuned Foundation Model and CNN Model

*Junho Hur[1,2], Toru Kouyama[2], Xuanchao Fu[1,2], Ichiro Yoshikawa[1], Chikatoshi Honda[3]

1. The University of Tokyo, 2. National Institute of Advanced Industrial Science and Technology, 3. University of Aizu

In asteroid exploration, measuring the size and investigating the distribution of boulders play a crucial role in selecting landing sites for spacecraft and understanding the processes involved in the formation and evolution of asteroids. At the individual asteroid level, it is known that examining the size and shape distribution of boulders can help verify whether the asteroid was formed through the catastrophic collision of its parent body (cf. Michikami et al., 2021). Such verification requires counting and contour measurement of tens of thousands to hundreds of thousands of boulders.
However, since the task of annotating asteroid rocks has traditionally been conducted primarily through human visual inspection, the use of image analysis techniques that can match human visual capabilities has become increasingly important for asteroid exploration. In previous studies, deep learning models such as Mask R-CNN, based on Convolutional Neural Networks (CNN), have been used for detecting rocks in asteroid image data.
On the other hand, a new machine learning technology called Transformer, which has improved performance by learning the entire context of data has appeared. In image learning, it is called as Vision Transformer (ViT, Dosovitskiy et al., 2021). Foundation models that contain numerous Transformers and are pre-trained on massive datasets with billions of samples have been reported to achieve image recognition capabilities comparable to or surpassing humans (Bommasani, et al., 2021).
This study focuses on investigating whether foundation model incorporating multiple Transformers can achieve scientifically reliable precision and be applied to planetary exploration, as well as identifying challenges in such adaptation. The study specifically explores the application of Meta's foundation model "Segment Anything (Kirillov et al., 2023)" to asteroid images, with a particular focus on the concrete application of "boulder size and shape measurement," which currently remains a prominent analytical method conducted through human visual inspection.
One notable advantage of foundation models is their capability for zero-shot learning. Despite encountering images for the first time, the model successfully detected numerous boulders on asteroids. In contrast, traditional deep learning models require at least a few target images and fine-tuning to achieve such detection capabilities.
We used custom Ryugu boulder dataset for this research. It consists of annotated data where boulders were manually detected by human inspection from close-up optical images of Ryugu captured by the ONC-T camera. The annotation process has been completed for over 228 close-up images, focusing on boulders larger than 10 pixels.
Based on these factors, the study compares the boulder detection performance of fine-tuned pre-trained foundation models against traditional deep learning models (detection accuracy: 68%, Seki, 2023) and Transformer-embedded deep learning models (detection accuracy: 72%, Hur, 2024). This comparison aims to explore the adaptability of foundation models to asteroid images.

Keywords: Deep Learning, Foundation Model, Instance Segmentation, Asteroid

