AN2DL -Report on Segmentation of Mars Landscape Neurova

Albin Sigurdson, Vladimir Blizniukov, Hedieh Raeisi, Homa Easy

albinsigurdson, vladimirblz, hediehraeisi, homaeasy

276619, 273495, 277545, 277875

December 14, 2024

1 Introduction

This project focuses on the semantic segmentation of Mars terrain images, which is essential for automated terrain analysis in planetary exploration and rover navigation. Each image is classified into one of five terrain classes: background, soil, bedrock, sand, and big rock. To enhance the model's capability to generalize across diverse Martian landscapes, we train convolutional neural networks (CNNs) using a labeled dataset and apply various preprocessing and augmentation techniques to diversify the data.

2 Problem Analysis

The training dataset consists of 2,615 grayscale images, each measuring 64×128 pixels. One of the main issues was the significant numerical prevalence of a test set, prompting the need for data augmentation. Additionally, there was a class imbalance, especially in the "Big Rock" class, which represented a minority and made up only about 10% of all labels.

3 Method

3.1 Preprocessing

During the data exploration, it was quickly found that many of the irrelevant images had been intentionally put in the training set to confuse the model training. Indifferent to the label, these images contained aliens, as shown in figure 1. As a result, 110 images were removed from the training set.

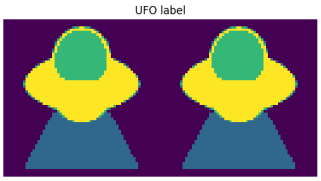

Figure 1: Label corresponding to alien images

To ensure a more robust and clean dataset, we performed a thorough cleaning process that involved two main steps: morphological cleaning and statistical cleaning.

3.1.1 Morphological Cleaning

The goal of this step is to refine the given masks by removing noise, such as small isolated regions or holes in the labeled areas. This process eliminates small disconnected regions while ensuring that valid areas are not fragmented due to minor imperfections.

3.1.2 Statistical Cleaning

This step identifies and addresses mislabeled masks by analyzing the proportion of each class within the

masks. We calculate the Z-score to measure how far a class's proportion deviates from the mean proportions of all classes. The Z-score is defined as follows:

$$Z = \frac{(X - \mu)}{\sigma}$$

where X is the proportion of the class, μ is the mean proportion across classes, and σ is the standard deviation of the proportions. If a class's Z-score exceeds a specific threshold, which we set at 2.5 standard deviations, it indicates that the class proportions are statistically outliers. The flagged masks and regions corresponding to these flagged classes are removed by setting their pixel values to 0.

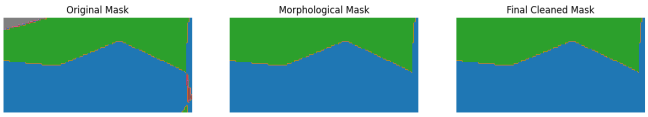

Figure 2: Example of cleaned data

3.2 Data Augmentation

To address the limitations of the dataset and improve training robustness, we implemented a data augmentation pipeline before applying any cleaning processes. This decision was made after observing that applying cleaning in the first hand *disproportionately affected classes with fewer pixels*, leading to their further reduction or misclassification as outliers.

Augmentation Pipeline

The augmentation process aimed to expand the dataset and balance class distributions while preserving semantic structures. The Augmentation increased the dataset to the size of 25,000 samples, ensuring balanced class distributions and improved generalization. Key steps included:

3.2.1 Geometric Transformations

We did random rotations (90°, 180°, 270°), horizontal flips, and transpositions to introduce spatial variability in the data.

3.2.2 Intensity Adjustments

CLAHE (Contrast Limited Adaptive Histogram Equalization) was added in the augmentation for contrast enhancement, random brightness/contrast adjustments, and Gaussian noise to mimic variations in image acquisition. CLAHE improves local contrast and highlights details in images, particularly in regions with low contrast. This is the formula for Gaussian noise, where μ is the mean and σ is the standard deviation of the noise.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

3.2.3 Elastic, Grid, and Optical Distortions

The Elastic grid applies random, non-linear deformations to simulate elastic stretching or compression. By dividing the image into a grid and adding random displacements, it mimics perspective inconsistencies caused by uneven terrains. It also simulates lens anomalies, like barrel or pincushion distortions, reflecting camera imperfections.

3.2.4 Synthetic Data Generation

The data generation was added with the Cut-and-Paste method. In this method, patches of specific regions (e.g., "Big Rock" class) were extracted from images and pasted onto other images. The pasted patches retained their class labels and spatial relationships, enriching the dataset.

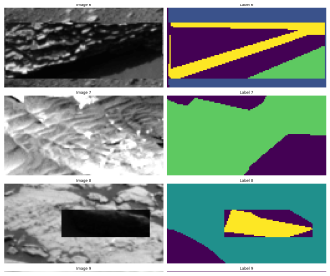

Figure 3: Example of Augmented data

3.3 Models

The primary model implemented was U-Net, known for its simplicity and accurate segmentation using an encoder-decoder structure with skip connections. The architecture consists of an encoder-decoder structure: the encoder extracts features by progressively down-sampling the input, while the decoder reconstructs the original resolution using up-sampling and concatenation with encoder layers to preserve spatial details. SegNet was tested for its computational efficiency, but lacked detailed

segmentation. Attention U-Net, with its attention gates, aimed to enhance focus on important regions, but overfitted due to increased complexity. U-Net outperformed both models, offering the best balance of accuracy and robustness for the dataset.

3.4 Metrics

The primary evaluation metric used was the mean Intersection over Union (mIoU), calculated as:

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c},$$

where C is the number of classes, TP_c is true positives for class c, FP_c is false positives, and FN_c is false negatives.

3.5 Loss

The target loss function used was categorical cross entropy with the background layer ignored. The Dice loss function was also tested as a target function, however, it led to inferior results.

4 Experiments

The original U-Net had the most amount of channels in the "bottleneck" of the architecture, right before upsampling. Extensive experimentation was done with an increasing number of channels even after the bottleneck. Results were similar, yet, the increased channels led to a greater architecture complexity and training time. The Attention U-Net demonstrated an improvement in detecting smaller regions and rare classes, with an average mIoU improvement of 2% compared to the baseline U-Net. The best model had the following characteristics: 477,141 trainable parameters, learning rate = 0.001, patience = 30, and Adam optimizer. It was trained for 1000 epochs and evaluated using K-fold (k=5) validation.

5 Results

Table 1: Comparison of Models on Test mIoU

Model	**Test mIoU (%)**
U-Net	58.24
Attention U-Net	41.82
SegNet	11.75

Qualitatively, the predictions seemed quite haphazard, sometimes accurate and sometimes far off, as seen in figure 4.

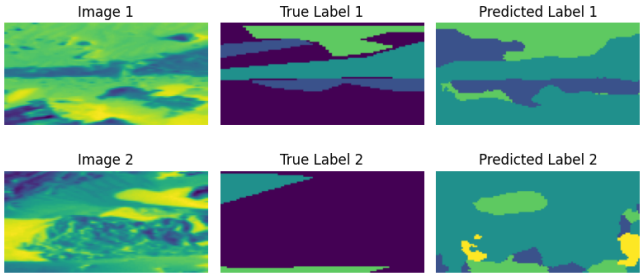

Figure 4: Predictions of two images

6 Conclusions

In conclusion, the original U-Net achieved the highest mean Intersection over Union (mIoU), while the Attention U-Net closely followed with an mIoU of 41.82, proving effective in detecting smaller classes with its attention gates. In contrast, SegNet had the lowest mIoU, likely due to its larger parameter size. Overall, U-Net is the top performer, but the Attention U-Net is a strong alternative for complex segmentation tasks.

7 Contributions

Albin worked on the architecture of the models and developed the best model. He also removed UFOs from the data.Vladimir worked on the report. Hedieh focused on data augmentation and cleaning, contributed to the report, and collaborated with Albin to develop the optimal U-Net model, also experimenting with SegNet. Homa worked on the Attention Unet.

References

[1] Ozan Oktay, Jo Schlemper, Louis Le Folgoc, et al., "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[2] Badrinarayanan, V., et al., *SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481–2495, 2017.