AN2DL - Second Homework Report
Team: Machine Dumbening

Seyedehyekta Kamaneh, Prachi Kedar, Francesco Frontera

yektakamaneh, prachikedar27, frankman110

243997, 245581, 247174

December 14, 2024

1 Introduction

This report, prepared for the Artificial Neural Networks course, addresses the semantic segmentation of Mars terrain images. Each image is accompanied by a mask that assigns a class label to every pixel. The segmentation labels include the following categories: 1) background, 2) soil, 3) bedrock, 4) sand, and 5) large rocks. Examples of these images and their corresponding masks are shown in Figure 1.

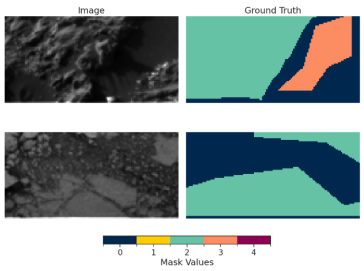

Figure 1: Example image

2 Data Inspection

After a random inspection of the dataset, we identified recurring outlier images with varying backgrounds and orientations. Cleaning these outliers effectively proved challenging using basic methods like plain similarity searches. As a result, we opted to manually identify and remove them from the dataset.

- **Original dataset size:** 2,615
- **Outlier instances removed:** 106
- **Cleaned dataset size:** 2,509

Figure 2 shows some of the detected outliers.

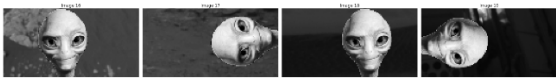

Figure 2: Data contamination samples

3 Data Augmentation

To enhance the model's robustness and improve feature learning, we are performing data augmentation by applying random flips along both the x and y axes. Figure 3 illustrates the effect of this transformation on a sample image and its corresponding label.

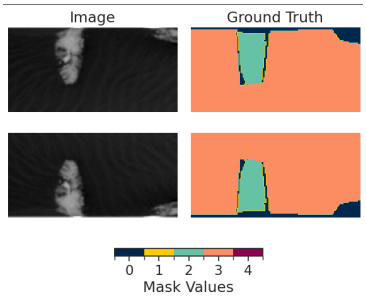

Figure 3: Synthetic data samples

4 Model Architecture

For the semantic segmentation task, we employed a modified **U-Net** architecture, a popular deep learning model designed for pixel-wise predictions. The U-Net model is particularly well-suited for image segmentation tasks due to its encoder-decoder structure, which captures both global context and fine-grained details.

The model comprises the following main components:

Encoder Path (Contraction Path):

- The encoder consists of four levels of convolutional layers with increasing filter sizes: 16, 32, 64, and 128.
- Each level employs two 3×3 convolutional layers with ReLU activation, followed by max pooling for down-sampling.
- Dropout layers are included to mitigate overfitting during training.
- This path progressively reduces the spatial dimensions while capturing high-level features.

Bottleneck:

- At the bottleneck, the model utilizes two convolutional layers with 256 filters, increasing the receptive field and capturing the most abstract features of the image.

Decoder Path (Expansive Path):

- The decoder mirrors the encoder, consisting of up-sampling layers implemented via transposed convolutions.
- At each up-sampling step, the corresponding features from the encoder are concatenated with the up-sampled output using skip connections. This enables the model to recover fine-grained spatial details.
- The decoder path gradually reduces the number of filters back to 128, 64, 32, and 16.

Output Layer:

- A 1×1 convolutional layer with a softmax activation function produces the final pixel-wise class predictions. The number of output channels corresponds to the number of segmentation classes.

Overall, the model contains nine levels with convolutional layers, each featuring He initialization and ReLU activation. Dropout is applied at various stages to improve generalization. This U-Net variant effectively balances computational efficiency and accuracy, ensuring robust segmentation of Mars terrain images.

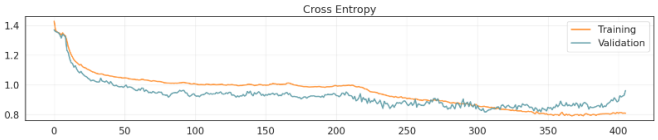

Figure 4: Cross Entropy

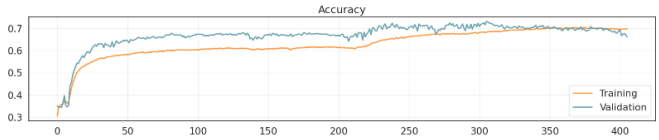

Figure 5: Accuracy

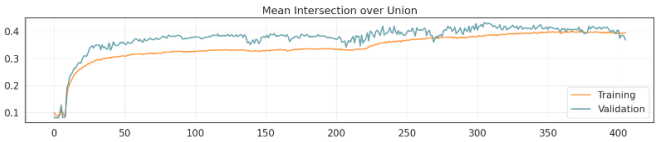

Figure 6: Mean Intersection Over Union

5 GAN-Based Segmentation

For a generative adversarial network (GAN)-based segmentation approach, we employed the Pix2Pix framework, which is well-suited for paired image-to-image translation tasks. In this setup, the network learns to predict segmentation masks from Mars terrain images by training two core components:

1. **Generator:** The generator is a modified U-Net trained to predict multi-class segmentation masks from input images. Its encoder-decoder structure with skip connections allows the model to capture both global context and fine-grained details. The downsampling layers extract increasingly abstract features, while the upsampling layers restore full resolution by integrating low-level features, thereby preserving spatial accuracy. A final softmax layer produces probability distributions for each pixel-class combination, ensuring robust and detailed segmentations.

2. **Discriminator:** The discriminator, inspired by PatchGAN, evaluates whether an (image, mask) pair is real or generated. By operating on image patches, it provides local feedback to the generator, pushing it to produce segmentation masks indistinguishable from the real ground truth. This local scrutiny improves the visual fidelity and contextual coherence of the generated masks.

Training Process: During training, the generator and discriminator engage in a minimax game: the generator strives to fool the discriminator by producing increasingly realistic masks, while the discriminator attempts to distinguish them from real ones. In addition to the adversarial loss (often binary cross-entropy), a reconstruction loss (such as categorical cross-entropy or MAE) ensures that the generated masks closely match the ground truth at a pixel level. This balanced combination of losses encourages the network to produce outputs that are both statistically accurate and perceptually convincing.

Inference:

After training, only the generator is retained and used for segmentation. It directly predicts segmentation masks for unseen terrain images. Indeed, the generator is also used for data augmentation since training data were not enough to be able to build a pretty good model.

By leveraging this adversarial framework, Pix2Pix facilitates the creation of high-quality segmentation masks that balance global coherence with local detail. This approach is particularly advantageous when pixel-wise accuracy and the ability to capture fine-grained patterns are critical.

6 Results

By the end of the training (for example, around epoch 189 out of 200), the reported loss values indicate a balanced performance. The **discriminator loss (D_loss)**, approximately 1.24, shows that the discriminator is neither completely fooled nor overly rejecting all generated masks. The **generator loss (G_loss)**, about 2.85, combines a very low **segmentation loss (Seg_loss)** of 0.0183—indicating a close match to the ground truth—and an **adversarial loss (Adv_loss)** near 1.02, which ensures sufficient realism. Taken together, these values suggest that the model is producing accurate and perceptually plausible segmentation masks, maintaining a well-rounded equilibrium between the adversarial and reconstruction objectives.

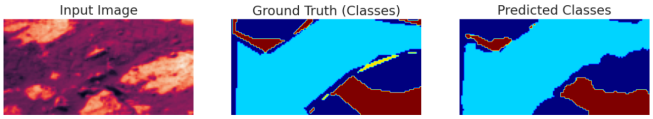

Figure 7: GAN Results Example 1

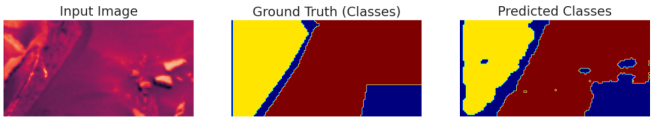

Figure 8: GAN Results Example 2

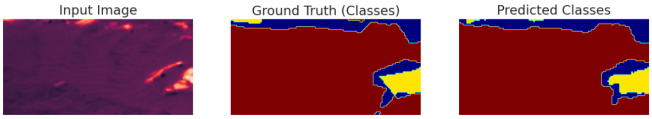

Figure 9: GAN Results Example 3

7 Hyperparameter Fine-Tuning

To optimize performance of the selected model, we applied hyperparameter tuning using random search, grid search, and Bayesian optimization techniques to explore different configurations. The results for the best parameters found are listed here:

- Number of filters per level: 16, 32, 64, 128
- Activation Function: relu
- Kernel size: 3
- Number of levels: 4
- Dropout rate of each layer: 0.2

* *This project reflects a collaborative effort, combining the contributions of all team members. Using a shared IPython notebook, we worked together through a mix of in-person and online sessions, exchanging ideas and integrating our work to achieve the final result.*