POLITECNICO
MILANO 1863

AN2DL - Second Homework Report
SoFarSoGood

Federico de Introna, Francesco Di Giore, Davide Corradina

fdeintrona, digioref, corrapiano

251274, 252126, 252117

July 22, 2025

1 Introduction

This project addresses the challenge of **semantic segmentation** of Mars terrain images [12]. The primary goal is to assign a class label to each pixel, categorizing the surface into one of five distinct terrain types.

The overarching objective is to develop an accurate and efficient model for pixel-wise classification. A custom U-Net-inspired architecture was developed, and the project focuses on creating a robust segmentation model, evaluating adaptive weighting schemes and diverse loss functions, and refining results through post-processing. The approach includes data analysis, iterative model training, and post-processing to enhance performance, aiming to improve Mars surface understanding.

2 Problem Analysis

The dataset comprises 64x128 grayscale images, with each pixel categorized into one of five terrain classes representing distinct geological features. The classes are: Background, Soil, Bedrock, Sand and Big Rock.

2.1 Initial assumptions

The analysis begins with several foundational assumptions about the dataset and problem domain. It is presumed that the dataset provides adequate variability to train a model capable of generalizing to unseen images. Each terrain class is assumed to exhibit distinctive textural and intensity characteristics, which the model can leverage to differentiate between classes. While external factors such as noise or image distortions may be present, these are assumed to be minimal or manageable through preprocessing steps. Additionally, the application of **data augmentation** is anticipated to improve the dataset's effective diversity, enhancing the model's training process.

2.2 Main challenges

The dataset presents significant challenges. **High intra-class variability** is a significant issue, as terrain types can display a wide range of appearances within the same class. Conversely, **inter-class similarities**, where distinct terrain types share overlapping features, add further complexity to the segmentation process. **Class imbalance** poses another challenge, as certain terrain types are underrepresented, necessitating the use of adaptive weighting techniques to ensure fair learning and accurate segmentation.

A particularly notable challenge is the nature of the provided masks, which segmented the dataset in a manner that often diverges from intuitive human perception. Lastly, the low resolution of 64x128 im-

ages limits the level of detail that can be captured, increasing the difficulty of distinguishing features.

Overcoming these challenges requires the model to generalize effectively to unseen terrain variations. Insights from the problem analysis shape the development and training strategies, providing a foundation for experimentation and refinement.

3 Method

The main aspects of our project are: a **U-Net** [10, 4] structured model, with some additions to improve and refine the general structure of the simple U-Net, use of **adaptive weights** to deal with class imbalance and **TTA** [15] to improve the predictions.

3.1 Architecture

The model we have produced is reported in the following table.

Table 1: Our Unet architecture

Layer (type)	Output Shape
InputLayer	(None, 64, 128, 1)
Encoder 1	(None, 32, 64, 32)
Encoder 2	(None, 16, 32, 64)
Encoder 3	(None, 8, 16, 128)
Encoder 4	(None, 4, 8, 256)
Bridge	(None, 4, 8, 512)
Decoder 1	(None, 4, 8, 256)
Decoder 2	(None, 8, 16, 128)
Decoder 3	(None, 16, 32, 64)
Decoder 4	(None, 32, 64, 32)
Concatenate	(None, 64, 128, 480)
OutputLayer	(None, 64, 128, 5)

This architecture was chosen for its ability to capture spatial hierarchies through a combination of down-sampling and up-sampling pathways, facilitating precise pixel-wise segmentation. Each encoder is made up of a **residual block**, combined with **squeeze-and-excite** mechanism, to improve the flow of the gradient, and to enhance feature representation by adapting the information flow. The bridge is composed of a **squeeze-and-excite** [3] block, a module that combines multiple **ASPP** [13, 8] blocks and finally the concatenation of the last two outputs from the encoders. ASPP module is used for multi-scale fusion of representations at different scales using dilated convolutions. Finally, the decoders are composed of **transpose convolutions**, that are learnable with respect to up-sampling, and **weighted fusion** of the output of the decoder with the skip connection from the corresponding encoder. The skip connection is obtained by using an ASPP module followed by an **attention gate** mechanism [2, 7, 9, 6, 14], used to properly fuse features from encoders and decoders.

3.2 Loss functions

Loss functions are central in guiding the model to learn effectively. Several loss functions [1] were evaluated to determine their impact on segmentation performance. These included **Categorical Cross-Entropy** (CE), Tversky loss, Dice loss, Jaccard loss, SSIM, focal variants of CE, Tversky and Dice.

In addition to testing these loss functions individually, combinations of multiple losses were experimented with. However, these combinations did not yield significant improvements over single-loss functions.

3.3 Adaptive weights

To mitigate class imbalance, adaptive weights were incorporated during training. New weights are computed each epoch based on the following formula as presented in [5]:

$$w_k = 1/(IoU_k + \alpha)$$

using the validation IoU for each class k and a parameter alpha for regularization purposes. The weights were then normalized to keep them between 0 and 1.

4 Experiments

The experiments aimed to evaluate the performance of the proposed model under different configurations and training strategies. The training process was carried out using various loss functions, including cross-entropy and Dice loss, with and without adaptive class weighting. The dataset was split into training and validation sets, and metrics such as mean Intersection over Union (mIoU) and pixel-wise accuracy were used for evaluation.

Quantitative results are summarized in Table 2.

Table 2: Performance comparison for different configurations

Loss function	mIoU	Accuracy
Dice Loss	64.67	65.30
Tversky Loss	65.55	63.54
Focal Loss	68.20	67.40
Categorical Cross-Entropy	**72.95**	**69.59**

4.1 Post-processing

Qualitative results, illustrated in Figure 1, demonstrate the post-processing technique's ability to redefine the borders of the different labels.

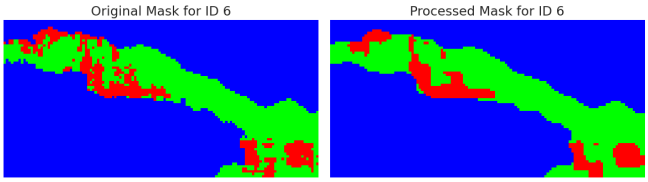

Figure 1: Example figure showing the elaboration of the image using post-processing techniques, such as library cv2

5 Results

The performance of the model was evaluated using the mean Intersection over Union (mIoU) metric, excluding class 0, along with the IoU of each terrain class. The results demonstrate consistent segmentation performance across most classes, with no unexpected outcomes. The final IoU scores for all classes are summarized in Table 3.

Table 3: Final Results

IoU_1 (%)	IoU_2 (%)	IoU_3 (%)	IoU_4 (%)	mIoU (%)
89.09	80.16	86.15	36.39	**72.95**

These results validate the model's ability to segment Mars terrain with reasonable accuracy, despite the challenges posed by class imbalance and low-resolution input images.

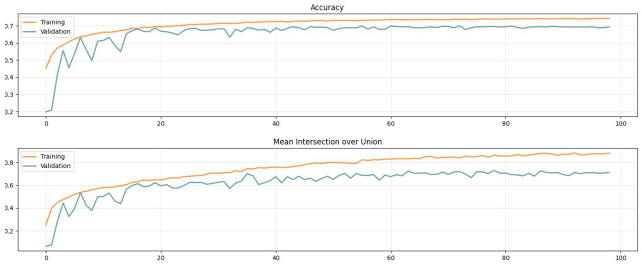

Figure 2: Accuracy and Mean IoU over the course of training.

6 Discussion

The model exhibits notable strengths and weaknesses. A significant challenge was segmenting terrain labeled as class 4, which was underrepresented in the dataset. The scarcity of samples for this class led to limited learning and lower performance compared to other classes, even with the use of adaptive weights. This highlights the inherent difficulty of addressing extreme class imbalances in semantic segmentation tasks.

Despite this, the model demonstrates strong overall performance. The use of a relatively simple U-Net-inspired architecture proves effective in achieving reliable segmentation across most terrain types. The results show that the architecture can generalize well to the dataset, leveraging adaptive weighting and optimized loss functions to balance disparities among classes.

7 Conclusions

The results indicate that a straightforward architecture, combined with adaptive weights and tailored loss functions, can produce a robust segmentation model.

Future work could focus on several areas for improvement. Finding an effective balance for stepwise training and combining multiple loss functions may enhance the model's performance further. Additionally, exploring the use of deep supervision [11], where auxiliary losses are applied to intermediate layers of the network, could help address challenges like class imbalance and low-resolution details. These enhancements have the potential to improve the model's ability to learn from limited or imbalanced datasets and achieve even more accurate segmentation of Mars terrain.

8 Contributors

In this section, we list the contributions of every member to the various tasks.

Member	Contributions
Davide Corradina	Data Analysis, Report, Hyperparameter Tuning
Federico de Introna	Adaptive Weights , Model
Francesco Di Giore	Post Processing, Model

References

[1] R. Azad. Loss functions in the era of semantic segmentation: A survey and outlook. *arXiv:2312.05391v1*, 2023.

[2] Y. Cai and Y. Wang. Ma-unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation. *arXiv:2012.10952*, 2020.

[3] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *arXiv:1709.01507v4*, 2019.

[4] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu. Unet 3+: A full-scale connected unet for medical image segmentation. *arXiv:2004.08790v1*.

[5] J. Li, K. Chen, G. Tian, L. Li, and Z. Shi. Marsseg: Mars surface semantic segmentation with multi-level extractor and connector. *arXiv:2404.04155*, 2024.

[6] H. Liu, F. Liu, X. Fan, and D. Huang. Polarized self-attention: Towards high-quality pixel-wise regression. *arXiv:2107.00782v2*, 2021.

[7] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention u-net: Learning where to look for the pancreas. *arXiv:1804.03999v3*, 2018.

[8] X. Qiu. U-net-aspp: U-net based on atrous spatial pyramid pooling model for medical image segmentation in covid-19. *Journal of Applied Science and Engineering*, 25:1167–1176, febbraio 2022.

[9] M. Rahman, S. Shokouhmand, S. Bhatt, and M. Faezipour. Mist: Medical image segmentation transformer with convolutional attention mixing (cam) decoder. *arXiv:2310.19898*.

[10] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv:1505.04597v1*, 2015.

[11] M. Sheng, W. Xu, J. Yang, and Z. Chen. Cross-attention and deep supervision unet for lesion segmentation of chronic stroke. *Front. Neurosci. 16:836412.doi: 10.3389/fnins.2022.836412*, 2022.

[12] R. M. Swan, D. Atha, H. A. Leopold, M. Gildner, and S. Oij. Ai4mars: Adataset for terrain-aware autonomous driving on mars. 2021.

[13] Z. Wang, Y. Chen, F. Wang, and Q. Bao. Improved unet model for brain tumor image segmentation based on aspp-coordinate attention mechanism. *arXiv:2409.08588v1*, 2024.

[14] S.-K. Yeom and J. von Klitzing. U-mixformer: Unet-like transformer with mix-attention for efficient semantic segmentation. *arXiv:2312.06272v1*, 2023.

[15] R. Zhou, Z. Yuan, W. Sun, Y. Ye, K. Zhang, X. Li, Z. Yan, Y. Li, L. He, and L. Sun. Ttt-unet: Enhancing u-net with test-time training layers for biomedical image segmentation. *arXiv:2409.11299v3*, 2024.