POLITECNICO
MILANO 1863

AN2DL - Second Homework Report
riconvoluzione informatica

Valeria de Gennaro, Donato Fiore, Lorenzo Fonnesu, Gabriele Lorenzetti

valeriadegennaro, donatofiore01, LolloFonne, lorentz18

251450, 248399, 249663, 259410

February 10, 2026

1 Introduction

In this project, we developed a deep learning model to address the challenge of **image segmentation** of the Martian terrain. The main objective was semantic segmentation, i.e., assigning a class label to each pixel in the images. The classes represent five different terrain types, each associated with specific features of the Martian terrain.

To achieve this, we designed a custom model based on a U-Net++ architecture integrated with the Path Aggregation Network (PAN). This combination improved the model's ability to capture both local and global details and context in images. In addition, to improve performance and reduce the risk of overfitting, we applied data augmentation strategies and introduced dropout during training.

The dataset used was subjected to a careful preliminary analysis and cleaning phase, where misleading were removed. This process ensured a more consistent and reliable database for training the model.

2 Problem Analysis

2.1 Dataset characteristics

The problem was presented with an initial dataset as described above. The first step was the analysis of the latter so as to have a deeper understanding of the problem. The analysis showed:

- **Distribution** The entire dataset presented two fields: the training set, composed by 2615 images with 64x128 shape, and the test set, composed by 10022 unlabelled images
- **Disruptive Elements** The data set included 110 out-of-context images characterized by their own out-of-context label.

2.2 Main challenges

The analysis of the dataset brought the following conclusions:

- **pre-processing** with the goal of cleaning the dataset of disturbing elements needed;
- addition of **image generality** needed, due to the limited number of images, in order to increase the generality of the model out of the training phase;
- select and customise an **appropriate model** that can learn efficiently from the resulting dataset.

2.3 Initial assumptions

In our work, we assumed the label 0, i.e. the background, is not considered in the final score, as requested.

3 Method

The following section describe all the techniques and approaches used to obtain the final model.

3.1 Pre-Processing

The pre-processing phase of the dataset plays a crucial role in achieving optimal results, not only during training but also in real-world applications. This process consists of three key steps:

- **Cleaning**: Out-of-context images were removed to improve the stability and effectiveness of training, enabling the model to achieve the target **Mean IOU** in less time.
- **Augmentation Techniques**: Various transformations were applied to the images to increase their generalization capacity and enhance the model's robustness to real-world variations. The augmentation techniques implemented include `rotation`, `width_shift`, `height_shift`, `shear`, `zoom`, and `horizontal_flip`, all utilizing a `nearest fill mode`.

For our purpose, we adopted the following approach: the dataset was initially split, with 300 images allocated to the **VAL_TEST** set. The remaining dataset was then duplicated, and the two parts were concatenated. Data augmentation was applied exclusively to the second part, and the resulting dataset was shuffled to ensure diversity.

3.2 Model development

The development started with a basic U-Net model for image segmentation, which delivered decent results. Building on this foundation, a more advanced **U-Net++** model was adopted, since it uses a nested skip connection network to aggregate the features while decoding the encoded data. By aggregating these features from different paths in the network, the model is able to improve the accuracy of the segmentation mask, also using or not deep supervision to enhance the model performance by providing a regularization to the network while training; like this work [5] suggests.

To further enhance performance, the model was integrated with a Feature Pyramid Network (FPN), which was later replaced by a **Pyramid Attention Network (PAN)** to exploit the impact of global contextual information, combining attention mechanism and spatial pyramid to extract precise dense features for pixel labelling instead of complicated dilated convolution and artificially designed decoder networks, like this work [2] suggests.

Both frameworks were then combined with **Squeeze-and-Excitation (SE) blocks**; this blocks adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels, like this work [1] suggests.

Simultaneously, various loss functions were tested and analyzed, including Focal Loss, Dice Loss, and Intersection over Union (IoU).

At the end of the training, the pixels that predicted 0 were changed to the second most probable value, i.e. 1, for the reason stated in the assumptions.

3.3 Loss Functions

To optimize model training, we developed a loss function by combining already known loss functions, such as **Focal Loss** 1 that optimizes the extreme imbalance between foreground and background classes found of dense detectors and prevents the vast number of easy negatives from overwhelming the detector during training [3]; **Dice Loss** 2 suitable for imbalanced data optimizing the overlap between predicted and ground truth masks [4]. This combination proved to be a robust approach to improve model performance in complex scenarios, ensuring more accurate segmented masks and effective handling of unbalanced classes.

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \tag{1}$$

$$\mathcal{L}_{Dice} = 1 - 2\frac{\sum(y_{true} \cdot y_{pred}) + smooth}{\sum y_{true} + \sum y_{pred} + smooth} \tag{2}$$

3.4 Final Model

The final solution is based on **U-Net++** model, integrated with **Path Aggregation Network (PAN)**, and **Squeeze-and-Excitation (SE)** to optimize performance in semantic segmentation tasks. Each component contributes uniquely to the overall architecture, enhancing its capacity to process and classify segmented images with high accuracy. Below, we detail the role and functionality of each component.

3.4.1 Unetpp

U-Net++ forms the backbone of the model, providing a robust and flexible architecture for semantic segmentation. Its defining feature is the dense skip connections, which enhance the flow of information between the encoder and decoder. These connections improve the model's ability to capture fine-grained details and ensure better convergence during training. Additionally, U-Net++ employs a nested structure that helps in generating more precise segmentation masks, particularly in complex regions of the image. Finally, in our model we opted to a non deep-supervised implementation, since it allows us to reach better results.

3.4.2 PAN

The Path Aggregation Network (PAN) complements the U-Net++ architecture by improving the integration of local and global contextual information. PAN utilizes adaptive feature pooling and hierarchical feature fusion to ensure that the model captures essential details at multiple scales. This capability is particularly valuable for processing segmented images where both fine local structures and broader spatial patterns play a critical role. By incorporating PAN, the model gains better robustness in handling variations in terrain and other complexities of the Martian dataset.

3.4.3 Squeeze-and-Excitation - SE

The Squeeze-and-Excitation (SE) module is integrated into the final model to enhance its ability to selectively emphasize the most informative features. SE achieves this by dynamically recalibrating channel-wise feature responses, allowing the model to focus on the most relevant patterns while suppressing noise. This mechanism is particularly effective in boosting the model's discriminative power and contributes to improved accuracy across all segmentation classes. By embedding SE, the model achieves a more efficient utilization of features, leading to better performance overall.

4 Experiments

We tried different combinations of strategies in order to achieve better scoring. In particular, as it can be seen in 1, our main experiments involved the use of Unet++, more or less deep, together with PAN/FPN. Moreover, we also combine different loss functions together, such as the sparse categorical crossentropy, the dice loss, the focal loss or the tversky loss.

Model	Score Kaggle	val MIOU	val accuracy	accuracy	val loss
1	0.58680	0.510	0.801	0.940	0.196
2	0.60415	0.470	0.786	0.940	0.215
3	**0.63753**	**0.500**	**0.7512**	**0.9198**	**0.2321**

Table 1: Progression of model performance through architectures and techniques: 1-unetpp+fpn+se, 2-unetpp+pan+se, 3-unetpp+pan+se+augmentation

5 Results

After different tests, a maximum score of 0.63753 (MIOU) was achieved, as shown in the table above 1.

An unexpected result was the lack of performance increase after the application of augmentation techniques. The augmentation was intentionally applied in a subtle manner for this very reason.

6 Discussion

The final model presents a decent capacity for the recognition of the various labels, but deteriorates in accuracy at the edges, probably due to the absence of a boundary loss function. The model also occasionally presents various difficulties in the recognition of certain specific classes.

7 Conclusions

During our work, the different tasks (model selection, augmentation, etc.) were equally divided among the group members, so as to balance the workload of each member.

In terms of future work, integrating transformer techniques would be an interesting avenue to explore.

Another possible strategy could involve implementing Attention Mechanisms, which allow the model to focus on the most relevant parts of the input data while minimizing distractions from less informative regions.

References

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

[2] H. Li, P. Xiong, J. An, and L. Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:*, 1805.10180, 2018.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *arXiv preprint arXiv:*, 1708.02002, 2018.

[4] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *arXiv preprint arXiv:*, 1707.03237, 2017.

[5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. *arXiv preprint arXiv:*, 1807.10165, 2018.