*DOI:*
10.1109/iSpaRo60631.2024.10687827

*Document Version*
Peer reviewed version

# Deep Learning based Semantic Segmentation for Mars Rover Terrain Classification

Fakher Mohammad[1,2], Yang Gao[1], Steven Kay[3], Robert Field[3], Matteo De Benedetti[3], Evridiki Vasileia Ntagiou[4]
[1]Centre for Robotics Research, Department of Engineering, NMES, King's College London, UK
[2]School of Mathematics and Physics, University of Surrey, UK
[3]GMV NSL Ltd, Airspeed 2, Eighth Street, Harwell Campus, Oxfordshire, UK, OX11 0RL
[4]European Space Operations Centre, ESA, Darmstadt, DE, 64293
Corresponding author (fakher.mohammad@kcl.ac.uk)

***Abstract*— Terrain classification is crucial for the successful execution of autonomous navigation and path planning during Mars rover missions. This study focuses on enhancing the rover's capability to traverse the Martian surface by investigating the integration of advanced semantic segmentation models based on deep learning. The aim is to identify the most effective deep learning model from recent advancements and establish efficient training approaches.**

**The study selected the state-of-the-art U-Net and DeepLabV3+ models for further assessment and evaluation, utilizing both the AI4Mars and ESA's LabelMars datasets. Techniques such as preprocessing, augmentation, and various loss functions were investigated to improve model performance and class imbalance issues are tackled. To mitigate overfitting, regularization techniques like weight decay and early stopping were applied, ensuring robust model training. Additionally, to further enhance the model's performance, especially in recognizing rare classes, the study investigated the use of state-of-the-art GAN models for generating new Mars rover images.**

**Our findings reveal that excluding the background class from training and testing significantly improves model performance. Using early stopping regularization reduces the training time drastically while giving high model performance. Notably, the DeepLabV3+ model surpasses the performance reported in existing literature, achieving a maximum segmentation accuracy and mIoU of 99% and 87% on the AI4Mars dataset, and 87% and 72% on the LabelMars dataset, respectively. The integration of GAN-generated images into training further improved rare class performance by up to 2%. These advancements in deep learning models for terrain classification promise to significantly enhance the capabilities of Mars rovers in autonomous navigation and path planning.**

## I. Introduction

The successful exploration of Mars depends on the ability of rovers to autonomously navigate the complex and often hazardous terrain of the Red Planet. A crucial aspect of this navigation is terrain classification, which entails identifying and categorizing various surfaces and obstacles on Mars. This classification is vital for enabling the rover to make informed decisions regarding safe paths, avoiding potential hazards, and planning its route efficiently.

In recent years, the field of deep learning has experienced significant progress in computer vision and image analysis. These advancements offer exciting prospects for applying deep learning techniques to Mars rover terrain classification. Deep learning models, known for their capacity to learn complex patterns and features from large datasets, present a promising solution for enhancing the rover's autonomy and adaptability to the challenging Martian environment. This article explores the intersection of deep learning and Mars rover terrain classification, aiming to develop and assess advanced deep learning models for efficiently classifying the geological properties of the Martian surface. The objective is to identify the best-performing model, establish effective training methodologies, and select optimal hyperparameters to achieve accurate semantic segmentation of Martian terrain. Leveraging state-of-the-art methods, evaluating preprocessing techniques, and addressing class imbalances, this study aspires to elevate the rover's navigation precision and safety on Mars.

Terrain classification is viewed as a semantic segmentation task, a more complex process than general classification tasks. Semantic segmentation involves assigning a label to every pixel in an image, with each label identifying a specific object. This technique is crucial for Mars rover terrain classification, where the model needs to accurately differentiate between various surfaces like soil, sand, bedrock, and large rocks. Thanks to recent advancements in deep learning, approaches such as U-Net and DeepLabV3 have emerged as leading methods for semantic segmentation. These models have shown great success in various fields, including Martian terrain classification. In our experiments, we employ these models with varying hyperparameters to compare their effectiveness. Additionally, we explore the use of advanced generative models, such as GauGAN and semanticStyleGAN, to create new Mars surface images to expand the datasets.

One common challenge in machine learning, including deep learning, is model overfitting. This issue arises, particularly with small datasets or complex models, when a model learns to fit the training data too precisely, struggling to generalize to new, unseen data. To mitigate overfitting, various methods such as dropouts, L1 & L2 regularizations, and early stopping are employed. Early stopping, involving monitoring the model's performance on a validation dataset during training and stopping the process when the model's performance no longer improves, proves effective and is utilized in our experiments.

## II. BACKGROUND

The attempt to use deep learning for Mars rover terrain classification has been driven by a growing body of research in the field of planetary science and artificial intelligence. Several scientific papers have laid the foundation for the application of deep learning techniques which are reviewed in the followings.

The work [1] introduces Soil Property and Object Classification (SPOC), a novel software capability that utilizes DeepLab FCNNs implementation and machine learning to visually identify terrain types and features on planetary surfaces, working with both orbital and ground-based images. SPOC utilizes a machine learning strategy, learning from a limited set of examples provided by human experts and then effectively applying this learned model to process a substantial amount of data. As a result, it efficiently offers crucial terrain data for rover assessments, streamlining the labor-intensive manual classification procedure. The paper details the technology behind SPOC and its successful applications in Mars rover missions, including terrain classification of 17 terrain object classes for the Mars 2020 Rover's landing site selection and slip prediction for the Mars Science Laboratory (MSL) mission.

A subsequent paper [2] introduces practical enhancements to the SPOC method. Initially, the approach undergoes pretraining using data sourced from the publicly available AI4Mars dataset and is subsequently fine-tuned for application to the Mars 2020 Rover (M2020). This fine-tuning involves feeding a small volume of data between different Sols, resulting in an overall pixel accuracy of 84.2% and a recall rate of 93.4% for the identification of sand, a critical class affecting the rover's traversability. Furthermore, the SPOC model pretrained on ImageNet, leading to a significant reduction in the decline of accuracy over time. Lastly, to enable deployment on mobile devices like rovers, the SPOC method is reimplemented with the lightweight CNN model MobileNetV2.

The NOAH-H project [3] aimed to create a comprehensive set of ontological classes for diverse surface textures and aeolian bedforms in Oxia Planum and Mawrth Vallis on Mars. Following this, a deep learning-based system for terrain classification was applied to categorize the diverse terrain types. The paper discusses the process of selecting these ontological classes, evaluates the model's pixel-scale accuracy, and explores how qualitative factors can impact its reliability and practicality. The Google DeepLab model was used for the semantic segmentation of the collected terrain images, yielding 74.15% mIoU.

A large-scale dataset AI4Mars created for training and validating terrain classification models for Mars rovers is introduced by NASA [4]. The dataset is used to train a DeepLabv3 model with a ResNet-101 backend pretrained on ImageNet, which achieved over 96% overall classification accuracy on the testing set. The testing is conducted on a gold standard testing set where each image was labelled by three expert labellers. However, other important semantic segmentation performance metrics like mIoU or IoU for individual classes are not provided.

Liu et al. [5] proposes a semantic segmentation method for the Chinese Zhurong Mars rover's terrain classification. By combining historical mission data from AI4Mars dataset and simulation rover data with semantic segmentation labels generated using Unreal Engine 4 from Epic Games. Contributions include a knowledge transfer-based segmentation strategy, integrating historical and simulation data to distinguish Martian landforms, and creating a virtual Mars scene. Evaluation with real Zhurong rover images shows an overall accuracy of 98.33%.

In [6], an semantic segmentation method with a hybrid attention-based approach is presented. This method utilizes a dual-branch network to effectively merge both the broader global context and finer local context information for unstructured terrains. This integration is facilitated by a merging module and a newly crafted loss function. The method's performance is assessed on two datasets: MarsScapes, which is a recently collected panorama dataset of Martian landforms, and AI4Mars, a publicly available dataset. The results show a 60% mIoU on MarsScapes and a high performance on AI4Mars with a 91% mIoU and 97% accuracy. Nevertheless, it's important to highlight that the computational performance, specifically the inference speed, has not been confirmed and could potentially be slower due to the inclusion of the dual network branch. Additionally, there is a lack of specific information regarding the AI4Mars testing set and the number of classes used for training and performance evaluation.

A semi-supervised learning framework is proposed for Mars terrain classification [7] where the deep segmentation network is trained in an unsupervised manner on unlabeled images and then transferred to the task of training terrain segmentation model using a small number of labeled images. The evaluation utilizes the AI4Mars dataset, which originally comprises four classes: soil, bedrock, sand, and big rock. However, the dataset is expanded by adding two additional classes that represent the rover itself and areas that are beyond 30 meters. The findings demonstrate a remarkable pixel-level accuracy of 97.5% when evaluated on the M3 testing set. This outperforms the accuracy achieved through standard supervised learning, which reached 95% on the same dataset. Nevertheless, the paper fails to provide IoU and mIoU values for the proposed approach, crucial metrics for evaluating the effectiveness of any semantic segmentation technique. In our experiments, we combined three categories – rover hardware, distances exceeding 30 meters, and unlabeled pixels – into one single class. This approach resulted in a segmentation model that distinguishes four distinct classes, while the combined fifth class was not included in the training process.

Zhang et al. [8] introduced a semi-supervised learning (SSL) framework designed for semantic segmentation of Mars images, employing a two-branch teacher-student architecture. The student model is built upon DeepLabV3+ with a ResNet50 backbone pretrained on ImageNet. Additionally, they proposed two augmentation methods: AugIN, which generates new images by altering the statistics of two distinct images, and SAM-Mix, which utilizes out-of-shelf segmentation (SAM) to duplicate an object from one image and paste it onto another, thereby enhancing the SSL framework's performance. The evaluation of this framework was conducted on the

AI4Mars and S5Mars datasets, resulting in an impressive 75% mIoU and 80% mean Accuracy (MACC) on the AI4Mars dataset.

In [9], an obstacle mapping technique is introduced for enabling autonomous navigation in robotic platforms such as planetary rovers. While the traditional approach of using LiDAR sensors for occupancy grid mapping is widespread in obstacle detection, it encounters difficulties in recognizing flat obstacles like sandy or rocky terrains. To overcome this challenge, the proposed method employs DeepLabV3+ for semantic segmentation, enabling the identification of theses flat obstacles within planetary environments. These obstacles are then integrated with depth data from a stereo camera to construct a laser scan like model using ORB-SLAM. The method's performance is assessed using images from the ESA Katwijk Beach Planetary Rover Dataset, with a comparison made between the resulting occupancy map and a manually segmented orthomosaic map obtained through drone surveys.

The work in [10] introduces a lightweight ViT-based terrain segmentation method named SegMarsViT. The proposed approach utilizes a mobile vision transformer (MViT) block in the encoder to extract local–global spatial information and capture multiscale contextual details. Additionally, cross-scale feature fusion modules (CFF) in the decoder integrate hierarchical context information, while the compact feature aggregation module (CFA) combines multi-level feature representation. The method is evaluated on three public datasets—AI4Mars, MSL-Seg, and S5Mars—achieving mIoU scores of 68.4%, 78.22%, and 67.28%, respectively, at a speed of 69.52 frames per second (FPS). The results demonstrate the efficiency and effectiveness of SegMarsViT for on-board satellite deployment in Martian terrain segmentation.

A rock detection in a Mars-like environment is proposed using a modified U-net-based model to segment images into rock and background [11]. The U-Net has fewer parameters to enhance inference speed. The methodology's effectiveness is demonstrated on Devon Island dataset comprising Mars-like environment images, achieving an impressive F-score of 78.5%.

## III. Model Architectures

In the following, both U-Net [12] and DeeplabV3Plus [13] models for semantic segmentations used in this work are briefly described:

### A. U-Net Model

U-Net, initially designed for medical/biomedical image segmentation, features an encoder-decoder architecture. The U shape of the model is attributed to its encoding (contracting or downsampling) and decoding (expanding or upsampling) paths, as illustrated in Figure 1. To facilitate the reuse of features acquired during downsampling, the feature maps from the downsampling path are concatenated with their mirrored counterparts along the upsampling path. This integration, denoted by the grey arrows in Figure 1, allows the model to capture diverse levels of abstractions.

### B. DeepLabV3+ Model

This model represents the latest model within the well-known DeepLab family, encompassing earlier versions such as DeepLabV1, DeepLabV2, DeepLabV3, and now DeepLabV3+. This model adopts an encoder-decoder architecture and employs atrous/dilated convolutions, effectively expanding the field of view without a proportional increase in parameters, as depicted in Fig 2. This approach enhances the receptive field of convolutions without imposing additional computational costs. DeepLabV3+ utilizes filters at various sampling rates to capture diverse objects and multiscale image contexts. The model integrates cascaded and parallel modules of dilated convolutions, providing a comprehensive strategy for effective feature extraction and semantic segmentation.

Figure 1. U-Net Architecture [12]

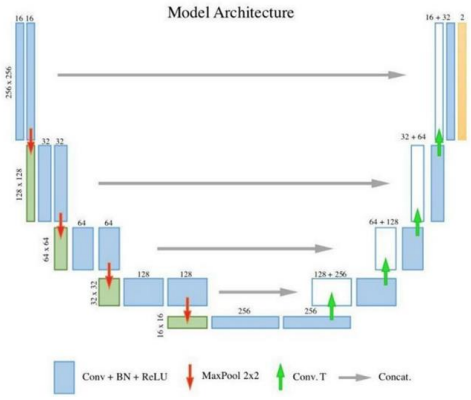
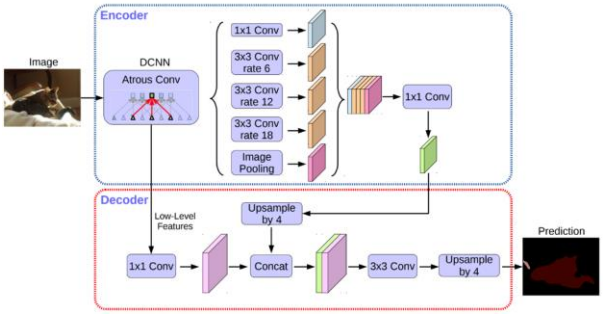



Figure 2. DeepLabV3+ Architecture [13]




Both models leverage encoder-decoder architectures. U-Net employs a distinctive approach by copying uncompressed activations, often referred to as skip connections, from encoding blocks to their mirrored counterparts within the decoding blocks. In contrast, DeepLabV3+ utilizes filters at multiple sampling rates to capture multiscale image contexts. Additionally, it incorporates atrous convolutions to expand the field of view.

### C. GAN Model

Generative Adversarial Networks (GANs) [14] are common deep learning methods for generating synthetic data that closely resembles original training data. They operate through two contesting deep learning networks, hence the term 'adversarial.' Among the various models, such as DCGAN [15], CycleGAN [16], and SAGAN [17], Conditional GANs are particularly noteworthy. These type of GANs generate images based on additional information like class labels, data

from other modalities, or a semantic mask, which aligns well with our current semantic segmentation task. These models include GauGAN [18] and SemanticStyleGAN [19].

GauGAN, for instance, creates new images conditioned on input semantic masks, using either original training masks or manually created ones. In contrast, SemanticStyleGAN simultaneously generates images and their corresponding semantic segmentation masks, eliminating the need for additional labeling. In our work, we utilize SemanticStyleGAN to generate synthetic images, thus expanding our dataset for training terrain classification models. To evaluate the quality of these generated images, we employ standard metrics such as the Inception Score (IS) [20] and Fréchet Inception Distance (FID) [21].

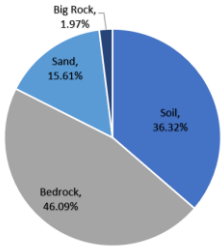
## IV. Datasets

Two datasets AI4Mars [4] and LabelMars [22] are used in the experiments conducted in this study. In the following, we briefly describe each of the datasets.

### A. *Ai4Mars*

AI4Mars, a publicly available comprehensive dataset for Mars terrain classification, comprises 35,000 high-resolution images from the Curiosity, Opportunity, and Spirit rovers. It features semantic segmentation labels in four primary categories: Soil, Bedrock, Sand, and Big Rock, with Bedrock being the most common and Big Rock the rarest, as depicted in Figure 3. Unlabeled pixels and elements like the sky, distances beyond 30 meters, and rover hardware are categorized as background class, which can be ignored during training and testing. The dataset also includes three distinct testing sets of 255 images (M1, M2, M3) with labels verified by up to three specialists, enhancing the dataset's reliability for model evaluation.

Figure 3. AI4Mars MSL class composition [4]




### B. *LabelMars*

The LabelMars dataset, curated by the European Space Agency (ESA) under the NOAH project, comprises 5000 images without a designated testing set. It is organized into five main categories: Artificial, Float Rock, Outcrop, Unconsolidated, and Sky, further divided into 25 sub-categories. Similar to AI4Mars, LabelMars employs a 'Don't Know' class for unassigned pixels, which can be excluded in training and testing.

## V. Experimental Results

### A. *Experimental Setup*

U-Net and DeepLabV3+ models are developed using PyTorch, deployed locally on Anaconda/Spyder on a Linux machine utilizing Quadro RTX 4000 GPU with 8GB RAM for training. Testing sets for AI4Mars were provided, and the training set was randomly split into 90% training and 10% validation sets. The LabelMars dataset underwent a similar random split into 80% training, 10% validation, and 10% testing sets. On both datasets the background class is excluded from training and testing processes.

In all experiments, models were trained and validated on designated training and validation sets and subsequently tested on the testing sets. Performance metrics such as pixel-level accuracy (Acc), class Intersection over Union (IoU), and mean IoU (mIoU) over all classes were recorded for the testing sets. During the training process, the best model is identified as the one that achieves the highest mIoU on the validation set within the final five epochs before the early stopping mechanism is triggered. The models trained using two distinct backbones, namely ResNet50 and ResNet101, with the inclusion of two different input image sizes, (256x256) and (512x512), to assess their impact on the model's performance.

Key parameters for the experiments included a base learning rate of 0.001 with a polynomial learning rate decay. Although the number of epochs was set at 100 for all experiments, training concluded at varying epochs due to early stopping regularization. Stochastic gradient descent (SGD) was the chosen optimizer, as experiments with the Adam optimizer did not exhibit performance differences. The cross-entropy loss function consistently outperformed other alternatives such as focal and dice losses, and thus was adopted in all experiments. A weight decay (L2 regularization) of 0.0001 was implemented, and batch sizes were optimized to the highest feasible values before GPU memory limitations were reached.

### B. *Using semanticStyleGAN*

In our experiments with semanticStyleGAN, we addressed the lower IoU performance of the Big Rock class in the AI4Mars dataset and the Float Rock class in the LabelMars dataset. The GAN model trained on the original images, specifically those containing the rare class.

Figure 4. Synthetic image generation using semanticStyleGAN trained on AI4Mars dataset

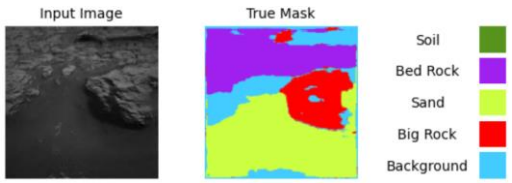
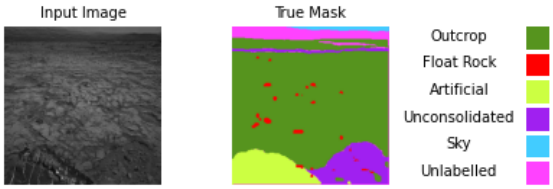



Figure 5. Synthetic image generation using semanticStyleGAN trained on LabelMars dataset



We generated synthetic images and corresponding masks matching the quantity in the original training set to duplicate the size of the training set. To evaluate the quality of the generated images, we used the Inception Score (IS) and Fréchet Inception Distance (FID) metrics. Figure 4 and 5 illustrate generated sample images along with its corresponding semantic masks.

### *C. Experiments on AI4Mars Dataset*

In each experiment the IoU of the four classes and mIoU over all classes are recorded for each of the three testing sets. Additionally, the Acc of the validation set and the testing sets were recorded (see Table I).

It's noteworthy that all models exhibit highest performance on the M3 testing set, where consensus among three specialists is required for labelling, establishing it as a gold standard for evaluation. Across all models, pixel-level testing accuracies (Acc) remain similar. However, notable variations emerge in terms of mIoU and the rare class's IoU. Interestingly, utilizing a larger image size (512x512) consistently yields better overall performance for both U-Net and DeepLabV3+. DeepLabV3+ outperforms U-Net, achieving the highest mIoU of 87% and the highest IoU for "Big Rock" at 59% on the M3 testing set. The performance is slightly better on the other two testing sets too, establishing DeepLabV3+ as the best performing model. The final row in the table shows results from a repeated experiment using the best model configuration with expanded training set using GAN-generated images enhancing the rare class's IoU by approximately 1-2%.

### *D. Experiments on LabelMars Dataset*

Here the same experiments in the last section are repeated on the LabelMars dataset, reporting only the results using 5 main class categories. Initial trials involving 26 sub-class categories yielded lower results, with a recorded mIoU of approximately 20%. This performance drop could be attributed to the heigh complexity of the dataset, possibly due to the increased number of classes.

Similar to the experiments on AI4Mars, the adoption of a larger input image size of 512x512 proved advantageous for both U-Net and DeepLabV3+. Unlike the previous experiment, both the DeepLabV3+ and U-Net models exhibit comparable performance, attaining the highest mIoU of 72% and a testing accuracy of 87% establishing both models as equally good. Notably, the "Float Rock" class exhibits the lowest IoU at a maximum of 21%, echoing a similar behavior observed in the case of the "Big Rock" class in the AI4Mars dataset. An observation emerges regarding the performance of ResNet50 and ResNet101 as backbones. Despite their equivalence in performance, ResNet50 stands out as the preferred choice due to its efficiency – requiring fewer parameters and less memory, thereby accelerating the training and inference process. Doubling the size of the training set with images generated by GAN model, trained on original images including the rare class 'Float Rock,' resulted in a modest increase of about 1% in both testing accuracy and the IoU for the rare class.

## VI. Results Comparison & Discussion

Our model shows notable improvements on the AI4Mars dataset, with a 2% increase in accuracy and a 5% rise in mIoU compared to existing literature (see Table III). These gains are achieved through the exclusion of the background class during training and testing and implementing early stopping regularization in addition to expanding the dataset using GAN model. It's important to note that there is a lack of comparable works using the LabelMars dataset, preventing direct comparison.

TABLE I. Acc, IoU, and mIoU of the semantic segmentation models using AI4Mars dataset.

| ***Model (Backbone) (Image size, Batch size)*** | ***Test.*** | ***Acc*** | ***mIoU*** | ***IoU*** | | | |
|---|---|---|---|---|---|---|---|
| | | | | ***Soil*** | ***Bedrock*** | ***Sand*** | ***Big Rock*** |
| **U-Net** (Resnet101) (256, 16) | M1 | 0.92 | 0.66 | 0.91 | 0.80 | 0.84 | 0.11 |
| | M2 | 0.96 | 0.72 | 0.95 | 0.87 | 0.91 | 0.13 |
| | M3 | 0.98 | 0.81 | 0.97 | 0.93 | 0.95 | 0.38 |
| **U-Net** (Resnet50) (256, 16) | M1 | 0.92 | 0.66 | 0.91 | 0.80 | 0.83 | 0.11 |
| | M2 | 0.96 | 0.71 | 0.95 | 0.87 | 0.90 | 0.12 |
| | M3 | 0.98 | 0.82 | 0.97 | 0.93 | 0.94 | 0.42 |
| **U-Net** (Resnet50) (512, 4) | M1 | 0.92 | 0.66 | 0.90 | 0.80 | 0.84 | 0.10 |
| | M2 | 0.96 | 0.71 | 0.95 | 0.88 | 0.91 | 0.09 |
| | M3 | 0.98 | 0.81 | 0.98 | 0.94 | 0.95 | 0.39 |
| **U-Net** (Resnet101) (512, 4) | M1 | 0.93 | 0.67 | 0.92 | 0.81 | 0.86 | 0.12 |
| | M2 | 0.96 | 0.72 | 0.96 | 0.89 | 0.93 | 0.12 |
| | M3 | 0.99 | 0.84 | 0.98 | 0.95 | 0.97 | 0.44 |
| **DeepLabV3+** (Resnet50) (256, 8) | M1 | 0.92 | 0.66 | 0.91 | 0.80 | 0.84 | 0.10 |
| | M2 | 0.96 | 0.71 | 0.95 | 0.88 | 0.91 | 0.10 |
| | M3 | 0.98 | 0.82 | 0.97 | 0.94 | 0.94 | 0.42 |
| **DeepLabV3+** (Resnet101) (256, 8) | M1 | 0.92 | 0.67 | 0.92 | 0.81 | 0.85 | 0.10 |
| | M2 | 0.96 | 0.72 | 0.96 | 0.88 | 0.92 | 0.11 |
| | M3 | 0.98 | 0.79 | 0.98 | 0.94 | 0.96 | 0.29 |
| **DeepLabV3+** (Resnet50) (512, 4) | M1 | 0.92 | 0.67 | 0.86 | 0.80 | 0.79 | 0.11 |
| | M2 | 0.96 | 0.71 | 0.96 | 0.88 | 0.92 | 0.10 |
| | M3 | 0.98 | 0.79 | 0.98 | 0.93 | 0.96 | 0.30 |
| **DeepLabV3+** (Resnet101) (512, 4) | M1 | 0.93 | **0.68** | 0.91 | 0.81 | 0.86 | **0.12** |
| | M2 | 0.96 | **0.73** | 0.96 | 0.89 | 0.93 | **0.13** |
| | M3 | 0.99 | **0.87** | 0.99 | 0.94 | 0.97 | **0.59** |
| **DeepLabV3+ & GAN** (Resnet101) (512, 4) | M1 | 0.93 | **0.68** | 0.92 | 0.82 | 0.85 | **0.13** |
| | M2 | 0.97 | **0.73** | 0.97 | 0.89 | 0.93 | **0.13** |
| | M3 | 0.99 | **0.88** | 0.99 | 0.95 | 0.97 | **0.61** |

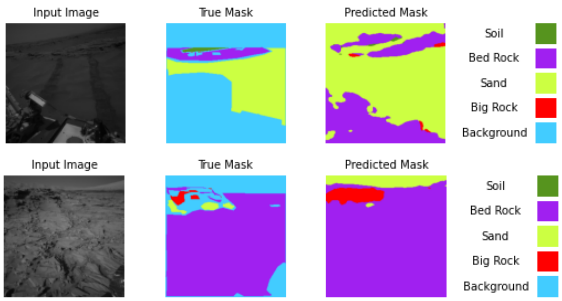
While excluding the background class improves our model's performance in training and testing for Mars rover terrain classification, its application needs careful consideration. The AI4Mars dataset's background class includes varied elements like the sky, rover hardware, distant beyond 30 meters, and unlabelled pixels. Notably, our model often misclassifies the sky as sand and rover hardware as bedrock (refer to Fig 6). However, these misclassifications may not critically impact rover navigation. The rover doesn't need to avoid the sky or its hardware, and the hardware is

usually outside the front camera's view, reducing the need for avoidance. Thus, these misclassifications are unlikely to significantly affect the rover's operational decision-making during exploration. In contrast, the LabelMars dataset's background class contains only unlabeled pixels, presenting different considerations than the AI4Mars dataset.

TABLE II. ACC, IOU, AND MIOU OF THE SEMANTIC SEGMENTATION MODELS USING LABELMARS DATASET.

| ***Model (Backbone) (Image size, Batch size)*** | ***Test Acc*** | ***mIoU*** | ***IoU*** | | | | |
|---|---|---|---|---|---|---|---|
| | | | ***Outcrop*** | ***Float Rock*** | ***Artificial*** | ***Uncons olidated*** | ***Sky*** |
| **U-Net** (Resnet50) (256, 8) | 0.86 | 0.70 | 0.70 | 0.18 | 0.83 | 0.77 | 0.99 |
| **U-Net** (Resnet50) (512, 4) | 0.87 | **0.72** | 0.72 | 0.21 | 0.87 | 0.78 | 0.99 |
| **U-Net** (Resnet101) (256, 8) | 0.85 | 0.69 | 0.69 | 0.17 | 0.83 | 0.77 | 0.99 |
| **U-Net** (Resnet101) (512, 4) | 0.87 | **0.72** | 0.72 | 0.21 | 0.88 | 0.79 | 0.99 |
| **DeepLabv3+** (Resnet50) (256, 8) | 0.86 | 0.69 | 0.71 | 0.16 | 0.82 | 0.77 | 0.99 |
| **DeepLabv3+** (Resnet50) (512, 4) | 0.87 | **0.72** | 0.72 | 0.21 | 0.87 | 0.79 | 0.99 |
| **DeepLabv3+** (Resnet101) (256, 8) | 0.86 | 0.68 | 0.72 | 0.14 | 0.80 | 0.77 | 0.98 |
| **DeepLabv3+** (Resnet101) (512, 4) | 0.87 | **0.72** | 0.72 | 0.21 | 0.87 | 0.79 | 0.99 |
| **DeepLabv3+ & GAN** (Resnet101) (512, 4) | 0.88 | **0.72** | 0.73 | **0.22** | 0.87 | 0.79 | 0.99 |

Figure 6. Original images and their true and pridicted masks using DeepLabV3+ model on AI4Mars dataset.




## VII. CONCLUSION

This study investigates advanced deep learning models for terrain classification, aimed at enhancing autonomous navigation and path planning in Mars rover missions. The primary objective was to identify the most effective deep learning model and establish efficient training approaches.

After evaluating the current state-of-the-art, we chose U-Net and DeepLabV3+ for further analysis. Using the AI4Mars dataset and the LabelMars dataset, we explored various preprocessing and augmentation techniques to improve model performance.

TABLE III. COMPARISON OF THE RESULTS OBTAINED USING OUR MODEL WITH EXISTING WORK. OUR BEST PERFROMING DEEPLABV3+ WITH GAN IMAGES IS USED WHICH TESTED ON AI4MARS M3 TESTING SET.

| ***Paper*** | ***Model*** | ***Input Size*** | ***Acc*** | ***mIoU*** |
|---|---|---|---|---|
| Swan et al [4] | DeepLabv3+ (ResNet101) | 513x513 | 0.96 | 0.75 |
| Atha et al [2] | DeepLabv3+ (mobileNetV2) | 513x513 | 0.97 | 0.83 |
| Zhang et al [8] | Self-Supervised model based on Teacher student (DeepLabv3+ with ResNet50) segmentation framework. | 512x512 | -- | 0.75 |
| Dai et al [10] | SegMarsViT (vision transformer ViT) Encoder (MobileViT-s) | 512x512 | 0.92 | 0.68 |
| Our Method | Based on DeepLabv3+ (ResNet101) | 512x512 | **0.99** | **0.88** |

Addressing class imbalance, we experimented with different loss functions and implemented regularization techniques such as weight decay and early stopping to mitigate overfitting. Moreover, we explored advanced generative models for generating images with rare classes, thereby improving model robustness.

A key finding of this research is the significant improvement in performance by excluding the background class during both training and testing. Early stopping regularization significantly reduced training time while maintaining high model performance. The DeepLabV3+ model exhibited highest accuracy of up to 99% and highest mIoU of 87% on the AI4Mars dataset—surpassing existing literature—and 87% and 72% on the LabelMars dataset, respectively. The use of semanticStyleGAN to augment datasets led to a 2% increase in IoU for rare classes in the AI4Mars dataset and a 1% increase in the LabelMars dataset. In future research, exploring a wider range of semantic segmentation models with different hyperparameters is feasible. Additionally, generative models could be employed to create higher resolution images, utilizing GPUs with larger memory capacities.


## ACKNOWLEDGMENT

Acknowledgments are given to open access of LabelMars dataset, and funding body European Space Agency under Contract No 4000137729/22/NL/AT “Vision Based Knowledge Extraction using Artificial Intelligence” that has led to the reported technical work and scientific findings.

## REFERENCES


[1] B. Rothrock, R. Kennedy, C. Cunningham, J. Papon, M. Heverly, and M. Ono, 'SPOC: Deep Learning-based Terrain Classification for Mars Rover Missions', in *AIAA SPACE 2016*, Long Beach, California: American Institute of Aeronautics and Astronautics, Sep. 2016. doi: 10.2514/6.2016-5539.

[2] D. Atha, R. M. Swan, A. Didier, Z. Hasnain, and M. Ono, 'Multi-mission Terrain Classifier for Safe Rover Navigation and Automated Science', in *2022 IEEE Aerospace Conference (AERO)*, Big Sky, MT, USA: IEEE, Mar. 2022, pp. 1–13. doi: 10.1109/AERO53065.2022.9843615.

[3] A. M. Barrett *et al.*, 'NOAH-H, a deep-learning, terrain classification system for Mars: Results for the ExoMars Rover candidate landing sites', *Icarus*, vol. 371, p. 114701, Jan. 2022, doi: 10.1016/j.icarus.2021.114701.

[4] R. M. Swan *et al.*, 'AI4MARS: A Dataset for Terrain-Aware Autonomous Driving on Mars', in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 1982–1991. doi: 10.1109/CVPRW53098.2021.00226.

[5] J. Liu, S. Liu, Y. Shao, X. Wan, and H. Zhao, 'Mars Terrain Semantic Segmentation using Zhurong Rover Imagery Based on Transfer Learning of Historical Mission Data', in *2022 International Conference on Service Robotics (ICoSR)*, Chengdu, China: IEEE, Jun. 2022, pp. 139–144. doi: 10.1109/ICoSR57188.2022.00034.

[6] H. Liu, M. Yao, X. Xiao, and H. Cui, 'A hybrid attention semantic segmentation network for unstructured terrain on Mars', *Acta Astronaut.*, vol. 204, pp. 492–499, Mar. 2023, doi: 10.1016/j.actaastro.2022.08.002.

[7] E. Goh, J. Chen, and B. Wilson, 'Mars Terrain Segmentation with Less Labels', in *2022 IEEE Aerospace Conference (AERO)*, Big Sky, MT, USA: IEEE, Mar. 2022, pp. 1–10. doi: 10.1109/AERO53065.2022.9843245.

[8] J. Zhang, L. Lin, Z. Fan, W. Wang, and J. Liu, 'S$^{5}$Mars: Semi-Supervised Learning for Mars Semantic Segmentation'. arXiv, Sep. 24, 2023. Accessed: Nov. 22, 2023. [Online]. Available: http://arxiv.org/abs/2207.01200

[9] S. Chiodini, M. Pertile, and S. Debei, 'Occupancy grid mapping for rover navigation based on semantic segmentation', *ACTA IMEKO*, vol. 10, no. 4, p. 155, Dec. 2021, doi: 10.21014/acta_imeko.v10i4.1144.

[10] Y. Dai, T. Zheng, C. Xue, and L. Zhou, 'SegMarsViT: Lightweight Mars Terrain Segmentation Network for Autonomous Driving in Planetary Exploration', *Remote Sens.*, vol. 14, no. 24, p. 6297, Dec. 2022, doi: 10.3390/rs14246297.

[11] F. Furlán, E. Rubio, H. Sossa, and V. Ponce, 'Rock Detection in a Mars-Like Environment Using a CNN', in *Pattern Recognition*, vol. 11524, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, J. A. Olvera-López, and J. Salas, Eds., in Lecture Notes in Computer Science, vol. 11524. , Cham: Springer International Publishing, 2019, pp. 149–158. doi: 10.1007/978-3-030-21077-9_14.

[12] O. Ronneberger, P. Fischer, and T. Brox, 'U-Net: Convolutional Networks for Biomedical Image Segmentation', in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture Notes in Computer Science, vol. 9351. , Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, 'Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation', in *Computer Vision – ECCV 2018*, vol. 11211, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science, vol. 11211. , Cham: Springer International Publishing, 2018, pp. 833–851. doi: 10.1007/978-3-030-01234-2_49.

[14] I. J. Goodfellow, 'Generative Adversarial Nets', *Proc Adv Neural Inf Process Syst*, 2014, [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[15] A. Radford, L. Metz, and S. Chintala, 'Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks'. arXiv, Jan. 07, 2016. Accessed: Dec. 29, 2023. [Online]. Available: http://arxiv.org/abs/1511.06434

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, 'Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks'. arXiv, Aug. 24, 2020. Accessed: Dec. 29, 2023. [Online]. Available: http://arxiv.org/abs/1703.10593

[17] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, 'Self-Attention Generative Adversarial Networks'. arXiv, Jun. 14, 2019. Accessed: Dec. 29, 2023. [Online]. Available: http://arxiv.org/abs/1805.08318

[18] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, 'GauGAN: semantic image synthesis with spatially adaptive normalization', in *ACM SIGGRAPH 2019 Real-Time Live!*, Los Angeles California: ACM, Jul. 2019, pp. 1–1. doi: 10.1145/3306305.3332370.

[19] Y. Shi, X. Yang, Y. Wan, and X. Shen, 'SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing', 2021, doi: 10.48550/ARXIV.2112.02236.

[20] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, 'Improved Techniques for Training GANs'. arXiv, Jun. 10, 2016. Accessed: Jun. 10, 2023. [Online]. Available: http://arxiv.org/abs/1606.03498

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, 'GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium'. arXiv, Jan. 12, 2018. Accessed: Jun. 10, 2023. [Online]. Available: http://arxiv.org/abs/1706.08500

[22] S. P. Schwenzer, M. Woods, S. Karachalios, N. Phan, and L. Joudrier, 'LabelMars: Creating an Extremely Large Martian Image Dataset Through Machine Learning', p. 1970, Mar. 2019.