AI4MARS: A Dataset for Terrain-Aware Autonomous Driving on Mars

R. Michael Swan, Deegan Atha, Henry A. Leopold, Matthew Gildner, Stephanie Oij
Cindy Chiu, and Masahiro Ono
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, CA 91109, USA
robert.m.swan@jpl.nasa.gov; deegan.j.atha@jpl.nasa.gov; henry.leopold@jpl.nasa.gov;
stephanie.l.oij@jpl.nasa.gov; matthew.gildner@jpl.nasa.gov; cindy.h.chiu@jpl.nasa.gov;
masahiro.ono@jpl.nasa.gov

Abstract

Deep learning has quickly become a necessity for self-driving vehicles on Earth. In contrast, the self-driving vehicles on Mars, including NASA's latest rover, Perseverance, which is planned to land on Mars in February 2021, are still driven by classical machine vision systems. Deep learning capabilities, such as semantic segmentation and object recognition, would substantially benefit the safety and productivity of ongoing and future missions to the red planet. To this end, we created the first large-scale dataset, AI4Mars, for training and validating terrain classification models for Mars, consisting of ∼326K semantic segmentation full image labels on 35K images from Curiosity, Opportunity, and Spirit rovers, collected through crowdsourcing. Each image was labeled by ∼10 people to ensure greater quality and agreement of the crowdsourced labels. It also includes ∼1.5K validation labels annotated by the rover planners and scientists from NASA's MSL (Mars Science Laboratory) mission, which operates the Curiosity rover, and MER (Mars Exploration Rovers) mission, which operated the Spirit and Opportunity rovers. We trained a DeepLabv3 model on the AI4Mars training dataset and achieved over 96% overall classification accuracy on the test set. The dataset is made publicly available.[1] [2]

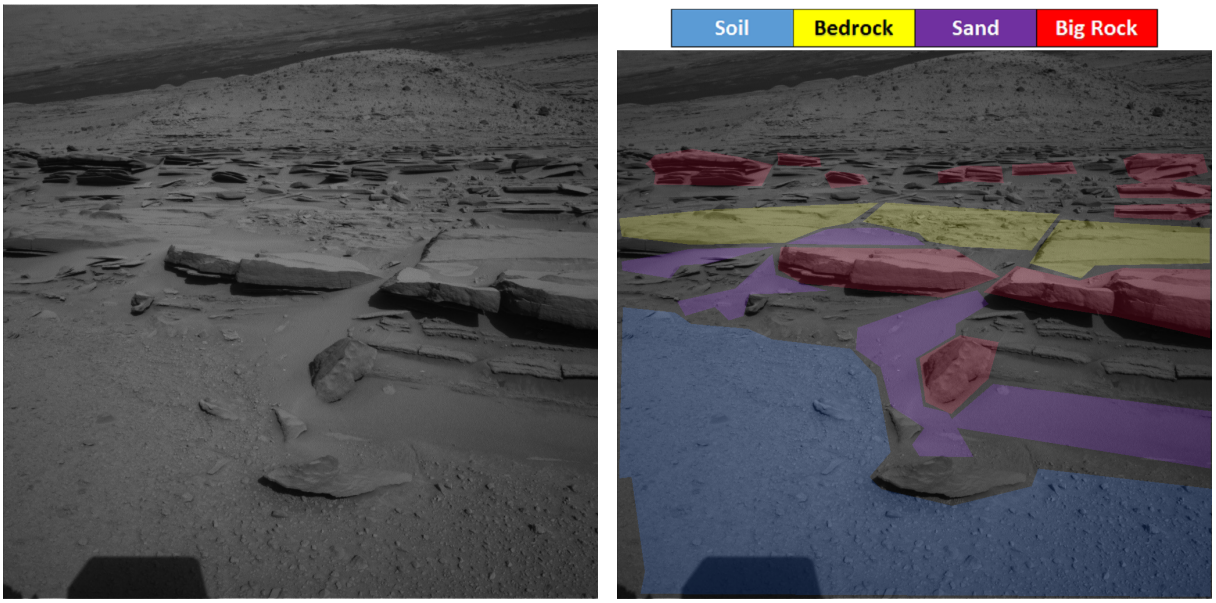
1. Introduction

NASA's Mars rovers have possessed an autonomous driving capability, called *AutoNav*, for more than a decade [3] on the *Spirit* and *Opportunity* rovers, which started exploration of the red world in 2004, and later on the *Curiosity* rover, which landed in 2012. Substantial enhancements of AutoNav were implemented for the latest rover, *Perseverance*, which is on the way to Mars at the time of writing [18, 26]. Still, its perception system remains purely based on classical machine vision algorithms, consisting of stereo matching for 3D reconstruction and obstacle detection, and visual odometry for state estimation [26]. This means that AutoNav assesses the traversability of the terrains solely based on geometric information. However, like vehicles on Earth, we empirically know that terrain types have substantial implications for traversability. For example, the Spirit rover was immobilized by a sand trap; the Curiosity rover, too, was nearly embedded when driving on a sandy surface at Hidden Valley; the aluminum wheels of Curiosity were punctured when it drove on pointy rocks on hard surfaces.

In fact, human rover drivers on Earth heavily rely on semantic information from rover images to plan a path. The manually planned drives by human rover drivers are able to use this semantic information about terrain types to drive over longer distances and over challenging terrain that would otherwise cause faults when using AutoNav and endanger the rover's hardware. For this reason the vast majority of Mars rover driving on existing missions is performed through manual terrain assessment and path planning by humans, while the usage of AutoNav has been limited to situations where manual control of drive path is impossible [19].

Had there been an ability to identify terrain types onboard, the rovers would be able to predict slip [8], plan a path for minimizing driving energy [12, 17] or localization error [13] (both which depend on terrain type), and autonomously identify promising targets for scientific observations [17]. The foundation for such terrain-aware autonomy is the ability to classify terrain from on-board images. For this reason, Rothrock *et al.* developed a machine learning-based terrain classifier for Mars named SPOC (Soil Property and Object Classification)[20], which has been deployed on the ground operation system of Cu-

[2] https://data.nasa.gov/d/cykx-2qix



(a) Original MSL NAVCAM image.

(b) A high rated AI4MARS label.

Figure 1: Kimberley region, imaged by Curiosity's NAVCAM on Sol 574 (March 18, 2014). Image Credit: NASA/JPL

riosity and field-tested on Earth with a test rover. While the results from these terrestrial deployments are highly promising, to further elevate SPOC's level of reliability and meet the highly demanding standard for on-board algorithms, a large-scale, high-quality labeled dataset is necessary.

Not surprisingly, Earth is the only planet on which a variety of public large-scale datasets for deep learning are available to humans. Previously we asked in-house experts to create Mars training datasets, but the number of labels that we could collect was severely limited (up to a few thousand) due to the availability of the experts. More recently, ESA's LabelMars project successfully ran a crowdsourcing effort and collected semantic segmentation labels on 5,000 images from Spirit, Opportunity, and Curiosity [23]. Their labels involve ∼20 geological/geomorphological terrain categories, such as "concretions/nodules," "dark-toned magmatic outcrop," and "light-toned sedementary float rock." Such categorization can only be interpreted by trained experts, hence posing a challenge to scale the dataset to the volume needed for highly accurate deep learning models, whose accuracy seems to scale logarithmically with dataset size [24]. This sharply contrasts to terrestrial datasets for autonomous driving such as KITTI [10], Cityscapes [7], nuScenes [5], Berkeley DeepDrive [27], and Oxford RobotCar [14], which mostly consist of common objects that are interpretable by non-experts such as people, animals, vehicles, and road signs.

Our key observation is that, while the geomorphological terrain categorization as in LabelMars is needed for planetary science, a substantially simpler and intuitive categorization, such as "sand," "rock," and "soil," suffices for traversability assessment of rovers on Mars, including the prediction of slip, driving energy, and wheel wear. A simplified terrain categorization allows the general public to participate in crowdsourcing with minimal training, which can be provided by a short web-based tutorial. At the same time, a high-quality dataset is necessary for system validation to meet stringent requirements for space missions. For this reason, we also collected a substantially smaller (∼1.5K labels) holdout dataset for model testing, labeled by about ten domain experts consisting of experienced Mars rover drivers and project scientists on MSL and MER missions. To summarize, the curation approach of the AI4Mars dataset is to gather a large-scale, focused dataset for training through crowdsourcing, combined with a high-quality, expert-labeled dataset for testing.

2. Related Works

We are not the first to utilize crowdsourcing for collecting labels on planetary images. There are many exciting space-bound citizen science projects created to find scientific features of interest on orbital images. For example, Moon Zoo [4] mobilized citizen scientists to identify and characterize impact locations of craters and other geological features on high resolution images of the moon obtained from NASA's Lunar Reconnaissance Orbiter (LRO) spacecraft. The project successfully used an online citizen science platform called Zooniverse, which allows web-based

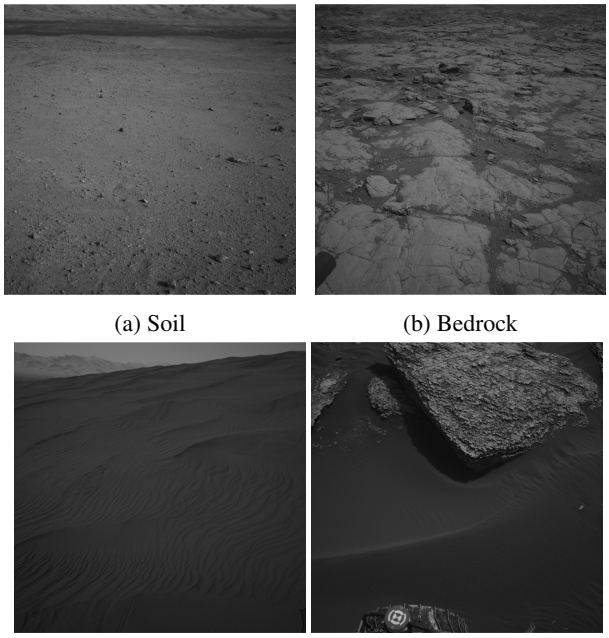
annotation of images by the general public. As for Mars, the Planet Four [2] project successfully used citizen science to identify a number of scientific features of interest on 221 high-resolution images on the southern polar region of Mars, taken by the Mars Reconnaissance Orbiter (MRO). Planet Four focused on intuitively recognizable features, such as fans and blotches, that are indicative of seasonal changes on Mars. The COSMIC project [9] developed an algorithm that detects changes on the Martian surface, such as fresh impact craters and avalanches. It also utilized Zooniverse to collect a training data set from citizen scientists. These projects were highly successful partially because macro-scale scientific features on orbital images often exhibits distinctive patterns that can be intuitively explained for non-experts, such as "Swiss cheese terrain"[25] and "spidery channels"[2].

In contrast, citizen science projects on planetary *surface* images are relatively rare, even though there are a substantial number of images taken on Lunar and Martian surfaces. This is perhaps because in-situ geology often requires analyses with expert knowledge, such as the interpretation of stratigraphy and classification of rock types. To the best of our knowledge, LabelMars [23] is the only large-scale label collection effort from citizen scientists on planetary in-situ images. Using their own website (www.labelmars.net), it accumulated terrain classification labels on five thousand images gained from Martian rovers, such as Curiosity and Spirit, with plans to expand the data set using images from the Opportunity and Perseverance rovers. Their labels involve ∼20 geological/geomorphological terrain categories, such as "concretions/nodules," "dark-toned magmatic outcrop," and "light-toned sedementary float rock." Since such classification requires expert knowledge, labels are collected from qualified experts, mostly consisting of undergraduate students in geology majors.

While AI4Mars uses the same data source as LabelMars, our data set is distinct from LabelMars in two aspects: i) the number of labels is two orders of magnitude greater than LabelMars (we have collected ∼ 326K labels at present), and ii) we employed a substantially simpler, four-way categorizaion of terrain: sand, soil, bedrock, and big rocks, which are intuitively understandable by non-experts *and* highly informative for assessing the traversability of Mars rovers. The purpose of the data set is also distinct from the data sets mentioned above in that AI4Mars is not for planetary science, geology, or geomorphology studies, but for training deep learning models to enable safe self-driving on Mars. To the best of our knowledge, this work is the first to successfully collect more than 100K semantic segmentation labels for images taken on the surface of any celestial body other than the Earth.

Dataset	Images	Train	Test
MSL NAVCAM	17K	160K	943
MSL Mastcam	9K	82K	TBD
Opportunity NAVCAM	6K	54K	573
Spirit NAVCAM	3K	30K	
Total (Combined)	35K	326K	1.5K

Table 1: Dataset summary. MER (i.e., Spirit and Opportunity) labels use a joint test set as the rovers are identical. Merged label counts as described in Section 3.3 are equal to the number of images. Production of a Mastcam test set is future work.



(a) Soil (b) Bedrock

(c) Sand (d) Big rock and sand

Figure 2: Representative examples of each class. Note that there are few if any images of big rock by itself.

3. The AI4Mars Dataset

The AI4Mars dataset includes the majority of the existing high-resolution images taken on the surface of Mars. It is comprised of ∼35K images sourced from the Planetary Data System[3] (PDS), covering the grey-scale navigation camera (NAVCAM) and color mast camera (Mastcam) images from the Curiosity (MSL), as well as the grey-scale NAVCAM images from Opportunity (MER), and Spirit (MER) Mars rovers, as noted in Table 1 (note that dataset counts pertain to the Nov. 2020 initial release). Notable exclusions from the dataset are microscopic images from Mars Hand Lens Imager (MAHLI) and telescopic images from

[3]https://pds-imaging.jpl.nasa.gov/index.html

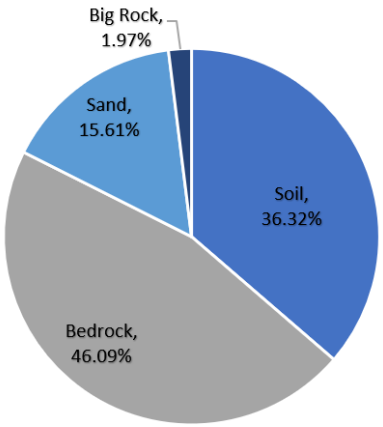


Figure 3: Composition of MSL NAVCAM labels by class.

Chemistry and Camera (ChemCham) because these images are not usually helpful for traversability assessment. The images from fish-eye Harzard Avoidance Cameras mounted on the rover's body are also not included because NAVCAM and MASTCAM images are better suited for terrain classification due to its higher mounting point and better angular resolution. The images from Panoramic Cameras (Pancam) of Spirit and Opportunity are planned to be included in the future. Since multiple labels were collected for each image, the number of labels greatly exceeds the actual number of images; this is discussed further in Section 3.1. The dataset is split into four different label types, these are: Soil, Bedrock, Sand, and Big Rock. Results relating to each of these classes is discussed in Section 4. The dataset also provides rover time and location data, camera parameters, and depth data when available.

Depth All provided MSL NAVCAM images and many MER NAVCAM images have associated stereo range data. The stereo ranging error for images increases linearly as function of distance. For distances of 1-30m, the MSL NAVCAM has an error of 0.0005-0.4m as noted in [16]. For the same range of 1-30m, the MER NAVCAM images have an error of 0.001-0.5m as noted in [15].

3.1. Collection

Training Set Our approach to label collection involved collecting 3 or more full image labels from different labelers for the same image. A single "label" will consist of all regions and classes identified within a single image. 93% of images have 9 or more associated labels. This single-image multi-label approach ensures that all images will likely include at least one quality label and extra labels may be used for some model training experiments (more discussion in Section 4.1).

Labels were primarily sourced from volunteer citizen scientists on Zooniverse [4] who were given a concise, yet complete, web-based training. When a labeler logged into the Zooniverse project page, they were presented images with a randomized order and asked to provide labels on each image. It was completely voluntary; there were no obligations or compensations. The collected labels went through algorithmic and manual acceptance review (Section 3.2), followed by label merging (Section 3.3) to enhance the overall quality of the dataset. A model trained by this dataset resulted in 96% overall accuracy, evaluated against the expert-labeled test set after label merging, as described in detail in Section 4.

Test Set A small "golden standard" test set was additionally created by a group of expert labelers in order to evaluate the model performance against a trusted reference, as well as to compare to final model performance on unseen data. Due to the highly limited availability and the high labor cost of the domain experts, the number of images in the test set needed to be small so that it could be labeled in a reasonable amount of time. We targeted a test set of approximately $\sim$1% of the images of training set. To ensure the test set properly represents the diversity of terrains contained within the training set, we sampled the images from locations distributed over all the major terrain classes mapped from orbital images, as shown in Table 2. Unlike the four-way local classification employed in the labeling, these orbital terrain classes represent large-scale geological units, each of which have unique composition and appearance for the four local terrain classes. This orbital terrain mapping had been generated manually by MSL project scientists [1] with an emphasis on geological terrain classes relevant to rover mobility near Curiosity's traverse path. Orbital terrain classes were correlated with rover locations during the mission. From these locations, approximately the same number of unique NAVCAM images were selected from the locations representing each orbital terrain class. The test set was then manually pruned of images of low quality (typically acquired near dusk), highly similar images, and those with large amounts of rover hardware. This resulted in a total 323 images for Curiosity' NAVCAM images and 205 Spirit/Opportunity's NAVCAM images (larger "Test" numbers in Table 1 are raw counts before the merging process discussed in Section 3.3).

Each orbital terrain class was assigned 3 expert labelers, with labelers being assigned to no more than 2 classes. Expert labelers were given the same general classification guidance as the Zooniverse participants with some modifications. The expert labelers only generated labels of high confidence and label coverage of all the terrain in an image

[4] https://www.zooniverse.org/projects/hiro-ono/ai4mars

Orbital Terrain Class	Quantity of NAVCAM Images
Smooth Terrain	38
Ridged Terrain	45
Pitted Terrain	57
Highly Dissected Terrain	57
Fractured Terrain	41
Sandy Pits and Ripple Fields	46
Sand Dunes	39

Table 2: Distribution of NAVCAM test set images within orbital terrain classes. Each image was labeled multiple times.

was not treated as a priority. This meant that some images had only a small fraction of the terrain labeled. Labeling was performed individually by the experts meaning that the set of labels collected for a specific image could vary from expert to expert.

3.1.1 Preprocessing

Before images were labeled, some preprocessing was applied to the images to make it easier for labelers to provide quality labels. Figure 4 shows an example of a preprocessed image. The white trapezoid-shaped marker in the center serves as a scale bar, where its width is always 50 cm in Curiosity images and 20 cm in Spirit/Opportunity images. These sizes are particularly relevant to the *Big Rock* class, which was defined as rocks which stand more than 30 cm high and are at least 50cm in width. This class definition was introduced because rocks higher than $\sim$30 cm are considered potentially risky for rovers to drive over. The tutorial stated that the labelers may ignore features smaller than the width of the scale bar. Note that the top portion of Figure 4 is darkened out; it indicates the portion of the image that is more than 30 m in distance from the cameras. Labelers were instructed to ignore features beyond 30 m. This distance was chosen to ensure only closer and clearer features needed for autonomy would be trained upon. The preprocessing was performed using the range product of the images, which were created from the stereo processing of images (all full-resolution NAVCAM in this dataset and many of the Mastcam images are stereo pairs).

3.1.2 Labeler Training

The classification categories of the AI4Mars dataset were designed so that those as young as fourth grade students can properly label with a quick, web-based tutorial. The tutorial, which is shown automatically when opening the labeling site for the first time, provides basic guidelines for labeling (e.g., no overlapping between labels) and intuitively explains the four terrain categories with examples. Before the launch of the citizen science project, we performed a beta test, participated by 138 volunteers, to get feedback on the tutorial and the user interface. For the expert labelers, in addition to the web-based tutorial, we held a series of meetings to build a consensus on our labeling policy and ensure the consistency of the test set.

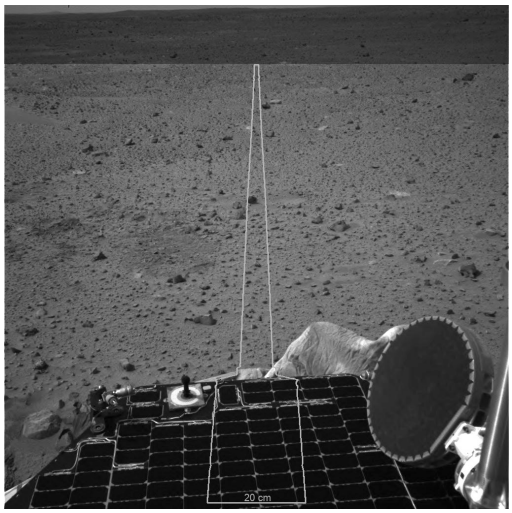


Figure 4: An example of the preprocessed image presented to labelers on Zooniverse; the image is taken by the Spirit rover on Sol 9 (9th Martian day after the landing).

3.1.3 Post-processing

After images were labeled, masks were applied to remove any annotations which overlapped with the rover itself (if it was in the image) and annotations which covered regions further than 30 m.

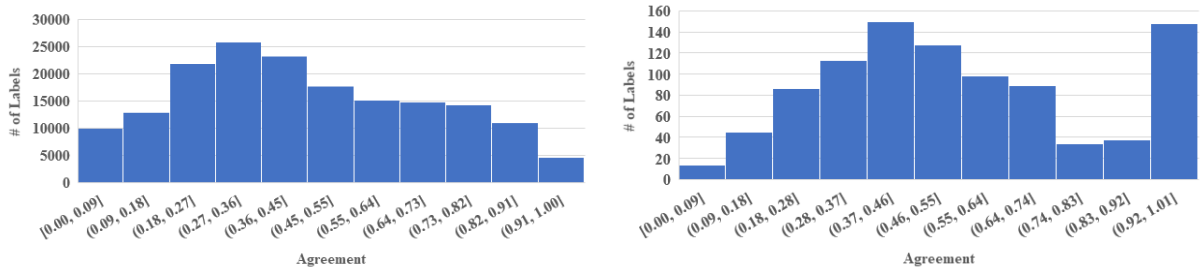
3.2. Cleaning & Label Acceptance

To enhance the overall quality of the training set, we performed algorithmic and manual review of the labels submitted by the citizen scientists. Labels which contained no valid annotations after post-processing were thrown out. Each label was compared with other labels for the same image to provide an "agreement score" that is based on the well known Jaccard Index/mIoU noted in Equation 1.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

We define the set of all labels for an image as L and all pairwise combinations of those labels (without replacement) as C per Equation 2.

$$C = \binom{L}{2} \tag{2}$$



(a) MSL NAVCAM Training Set, labeled by citizen scientsts

(b) MSL NAVCAM Test Set, labeled by domain experts

Figure 5: A histogram comparing agreement scores of the coarse train set and fine-grained test set.

The agreement score of some label $i \in L$ is then defined as the mean of the Jaccard Index for all pairs in C containing label i and any other label per Equations 3 and 4. For example, given 3 labels where we want to find the agreement score of label 1, we find all combination pairs containing label 1: (1,2) (1,3). We then find the Jaccard Index for pair (1,2) and for pair (1,3). These numbers are summed and then divided by the total number of pairs (2).

$$C_i = \{(x, y) \in C | x = i\} \tag{3}$$

$$A(C_i) = \overline{J(C_i)} \tag{4}$$

Labels with low agreement scores (bottom 20% of distribution of all labels) were then preferentially reviewed by our team; preliminary investigation suggests that low agreement scores are correlated with poor reviews. The team reviewed labels by providing a rating from 1-5 regarding the quality of each label. Labels given a rating of 1 are thrown out. Possible usage of the remaining ratings is discussed in Section 5.

The histogram of agreement scores for the test set and training set shown in Figure 5 provides a quantitative review of the agreement between labels in the test set. Based on notable shift between the coarse training labels and fine test labels, we hypothesize that higher agreement scores are strongly correlated with higher quality labels. It is possible that other projects with the unique problem of having more labels than data may be able to make use of agreement as a quality metric.

3.3. Label Merging

Given an image still has multiple quality labels after data cleaning and acceptance, there is some question as to how the remaining labels should be used, much of which is discussed in Sections 4.1 and 5. One approach considers the idea of merging existing labels where they agree and where other labels were left blank (most labels are sparsely annotated on purpose). To that end, a merged label dataset is provided which uses a majority rule to determine which annotations or parts of annotations to include.

To merge multiple labels for a single image into a singular label, two criteria were used. The first was that for an individual pixel to be accepted, the most commonly labeled class for that pixel had to be labeled by at least three different labelers. The second was that for each pixel, the accepted class had to have over 65% agreement out of the total number of labels for the pixel. Note that the pixel-wise "agreement" noted here refers to pixel-wise label overlap, which is not the same as the "agreement score" defined in Section 3.2. Unlabeled pixels are ignored for this calculation. As an example of this, Pixel X was within an image that was labeled by 10 labelers. The breakdown was: 2 unlabeled, 6 soil, 1 sand, and 1 bedrock label. The merged class for Pixel X will be soil as it received 6 labels and a 75% agreement. As another example, Pixel Y was within an image annotated by 9 labelers. The breakdown was: 3 unlabeled, 3 soil, 2, sand, and one bedrock label. There are 3 soil labels but it only has a 50% agreement, so Pixel Y will be unlabeled within the merged annotation.

For the gold standard test set, each image was annotated by three different expert labelers. Due to the reduced number of labelers and increased confidence of the labels, different merge parameters were used. Three different gold standard test sets were generated. All three required 100% agreement for a label of a specific pixel, but the sets varied based on the minimum number of labeled pixels required to accept a label using 1, 2, and 3 total labels per pixel. The gold test set that contains a merged label with a minimum of 3 labels per pixel and 100% accuracy is the most confident set and the accuracy of a model should be the best. However, this will also be the sparsest test set and therefore the other sets are useful to gain a better understanding of performance in the less confident regions of the image.

A summary of pixel-wise agreement between experts on each image in the MSL test set is shown in Figure 6. Classes like *Big Rock* and *Bedrock* which have a higher proportion of pixels where there was only agreement between two labelers or one labeler (no agreement) suggest that the definition of the classes or the method of labeling them has some

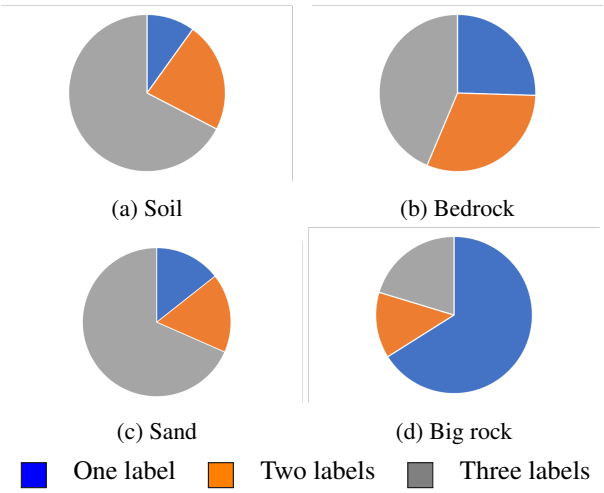


Figure 6: Breakdown of proportions of how many expert labelers labeled a pixel a specific class for the MSL test set.

inherent ambiguity as compared to other classes.

4. Experiments

For all of our experiments we made use of DeepLabv3+ with a ResNet-101 backend pretrained on ImageNet [6, 11, 21]. DeepLab was selected due to the maturity of its codebase and the state of the art semantic segmentation performance it maintains.Training was done on machines with either two NVIDIA GeForce GTX TITAN X or two NVIDIA Tesla P100 GPUs. Images were resized from 1024x1024 pixels to 513x513 in order to match the settings of the pretrained model. Batch size was chosen to be as large as possible before running into GPU memory issues as recommended by DeepLab documentation.

All experiments shown here were done using MSL data. This dataset is the largest and was completed first, so there was more time for analysis. Hyperparameters were determined experimentally; the same hyperparameters were used for all experiments with the exception of label weights. Label weighting was chosen to be $1 - composition$ where composition refers to the percent taken up by each class in the training set. Composition numbers for all MSL NAVCAM labels are shown in Figure 3. In our experiments, it was found that using no label weight or incorrect label weights resulted in slower model convergence, but increasing the number of steps in these situations resulted in similar mIoU scores and class accuracy as models trained using the weighting approach mentioned previously.

Better performance on all metrics is likely possible given further testing. The experiments noted here are intended to serve as a baseline to improve upon. A number of possible variations on this approach which could provide improved performance are discussed further in Section 5.

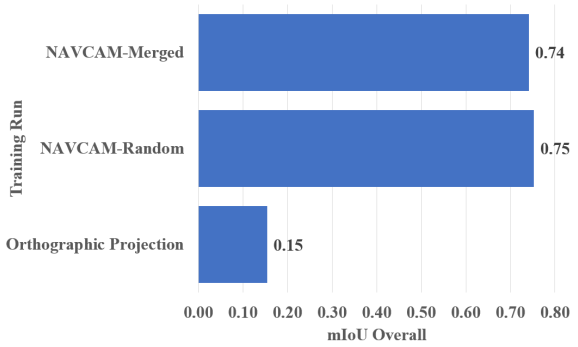


Figure 7: Overall mIoU scores calculated against random validation sets for variants of MSL data. The NAVCAM variants are defined in Section 4.1

		Predicted			
		Soil	Bedrock	Sand	Big Rock
Actual	Soil	96.00	0.31	3.69	0
	Bedrock	6.15	90.87	2.54	0.44
	Sand	0.25	3.23	96.51	0.01
	Big Rock	11.67	0.03	5.48	82.83

Table 3: MSL *NAVCAM-Random* confusion matrix percentages calculated with respect to the 3 label agreement test set. Overall accuracy is 94.97%.

		Predicted			
		Soil	Bedrock	Sand	Big Rock
Actual	Soil	99.10	0.32	0.57	0.01
	Bedrock	3.64	94.90	0.37	1.09
	Sand	0.88	5.62	93.45	0.05
	Big Rock	6.76	0	0	93.24

Table 4: MSL *NAVCAM-Merged* confusion matrix percentages calculated with respect to the 3 label agreement test set. Overall accuracy is 96.67%

4.1. Label Merging Versus Random Selection

To examine the effectiveness of label merging versus random selection of available labels, a number of experiments were done with data that was merged as described in Section 3.3 and randomly selected data (e.g. given an image has 10 labels, we randomly pick one of those to use). The best results we were able to achieve for each approach are noted in Figure 7 and Tables 4 and 3. Initial results indicated a clear benefit to label merging, but amended results after fixing implementation errors are inconclusive. We provide them for completeness and discuss possible improvements in Section 5.

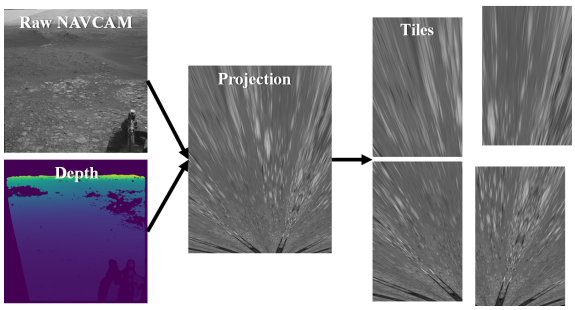


Figure 8: The process required to project and tile images. Depth is used both for projection and to mask distant areas out. In this case the top half of the raw image is removed before projection. Tiling shown is exaggerated; many images require as many as 100 tiles.

4.2. Tiling & Orthographic Projection

One idea for integrating depth data into training a model and improving our performance on the *Big Rock* class is to make use of orthographic projection, which is a well established technique in the domain of computer graphics [22]. This approach was used in [20] and replicated in this work using a nearly identical implementation for projection and tiling. This process is summarized in Figure 8.

The mIoU scores we were able to achieve as noted in Figure 7 are objectively poor; it is unclear whether the cause of this poor performance was an issue with our implementation or the approach itself. We hypothesize that the method used for tiling causes too much image context to be lost on average, such that the model is not able to consistently determine differences between similar classes.

5. Discussion and Future Work

There is a plethora of future work which can be done to improve upon experiments done with this dataset as well as autonomy efforts on other planets. This paper provides a snapshot of the dataset as it currently stands. We have plans to continue making improvements and additions to the dataset, including the images from the new Mars rover *Perseverance*.

We believe performance on *Big Rock* class included in the dataset can be greatly increased by making use of existing depth data (when available) or by using a separate rock instance classifier. Other work is ongoing which identifies rock faces via an instance detector, and then uses stereo data to estimate the rock height. Using this model approach in the future, it would be possible to train a three class semantic segmentation approach with the rock detector filling in the big rock class. Additionally, a panoptic segmentation model could be explored to combine the two networks.

The label merging approach we used in Sections 3.3 and 4.1 is not the only way of handling extra labels, and it is currently unclear what the precise benefit of label merging is. The use of label metrics could be used to create a confidence score for labels which neural networks could integrate in order to improve training accuracy. A couple of metrics which might be used are the label ratings and computed agreement scores mentioned in Section 3.2.

Another avenue of future work is the utilization of the automated terrain classifier, trained by this dataset. For example, we empirically know that slip and driving energy is highly correlated with terrain type [8]; rovers could choose energy-optimal paths by knowing the terrain type [17]; while the AI4Mars dataset does not employ geological terrain categories, it could help rovers to support scientific exploration because most scientific observations focus on evidences found in bedrocks, where geological contexts are much better preserved than sand, soil, or float rocks. Highly accurate Martian terrain classification, enabled by AI4Mars, would be a foundation for these advanced applications.

6. Conclusions

We have presented AI4MARS, a large dataset for terrain-aware autonomy on Mars. We provided an extensive overview of our process for collecting and handling the data, statistics on its composition, and approaches for using data which has more labels than images. Our experiments provide baseline results for this data and ideas for how to handle it in order to achieve the best possible model performance. We hope this dataset will foster future work on extraterrestrial autonomy and look forward to reading studies that make use of it.

7. Acknowledgements

We extend our thanks to the Zooniverse staff, in particular Cliff Johnson, for their assistance with collection of the dataset, the annotation tool, and support. We would also like to thank Brandon Rothrock, Ryan Kennedy, and Jeremie Papon for their guidance in reproducing and extending the results of SPOC [20] in addition to Xiangyu Hong who aided work on label acceptance. We also thank Keri Bean, Fred Calef, Doug Ellison, Abigail Fraeman, Sharon Laubach, Camden Miller, Tyler del Sesto, Nathaniel Stein, Ashley Stroupe, and Nathan Williams for providing expert labels. Most importantly, we would like to thank the thousands of citizen scientists that participated in the AI4Mars effort; we would not have been able to achieve these results without your help. The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

References

[1] RE Arvidson, P DeGrosse Jr, JP Grotzinger, MC Heverly, J Shechet, SJ Moreland, MA Newby, N Stein, AC Steffy, F Zhou, et al. Relating geologic units and mobility system kinematics contributing to curiosity wheel damage at gale crater, mars. *Journal of Terramechanics*, 73:73–93, 2017.

[2] K.-Michael Aye, Megan E. Schwamb, Ganna Portyankina, Candice J. Hansen, Adam McMaster, Grant R.M. Miller, Brian Carstensen, Christopher Snyder, Michael Parrish, Stuart Lynn, and et al. Planet four: Probing springtime winds on mars by mapping the southern polar co2 jet deposits. *Icarus*, 319:558–598, Feb 2019.

[3] Jeffrey J. Biesiadecki and MarkW. Maimone. The mars exploration rover surface mobility flight software: Driving ambition. In *2006 IEEE Aerospace Conference Proceedings*, 2006.

[4] Roberto Bugiolacchi, Steven Bamford, Paul Tar, Neil Thacker, Ian A Crawford, Katherine H Joy, Peter M Grindrod, and Chris Lintott. The moon zoo citizen science project: Preliminary results for the apollo 17 landing site. *Icarus*, 271:30–48, 2016.

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[8] C. Cunningham, M. Ono, I. Nesnas, J. Yen, and W. L. Whittaker. Locally-adaptive slip prediction for planetary rovers using gaussian processes. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5487–5494, 2017.

[9] G. Doran, S. Lu, M. Liukis, L. Mandrake, U. Rebbapragada, K. L. Wagstaff, J. Young, E. Langert, A. Braunegg, P. Horton, D. Jeong, and A. Trockman. Cosmic: Content-based onboard summarization to monitor infrequent change. In *2020 IEEE Aerospace Conference*, pages 1–12, 2020.

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] S. Higa, Y. Iwashita, K. Otsu, M. Ono, O. Lamarre, A. Didier, and M. Hoffmann. Vision-based estimation of driving energy for planetary rovers using deep learning and terramechanics. *IEEE Robotics and Automation Letters*, 4(4):3876–3883, 2019.

[13] Hiroka Inoue, Masahiro Ono, Sakurako Tamaki, and Shuichi Adachi. Active localization for planetary rovers. In *2016 IEEE Aerospace Conference Proceedings*, 2016.

[14] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.

[15] JN Maki, JF Bell, Kenneth E Herkenhoff, SW Squyres, A Kiely, M Klimesh, M Schwochert, T Litwin, R Willson, A Johnson, et al. Mars exploration rover engineering cameras. *Journal of Geophysical Research: Planets*, 108(E12), 2003.

[16] J Maki, D Thiessen, A Pourangi, P Kobzeff, T Litwin, L Scherr, S Elliott, A Dingizian, and M Maimone. The mars science laboratory engineering cameras. *Space science reviews*, 170(1-4):77–93, 2012.

[17] Masahiro Ono, Brandon Rothrock, Kyohei Otsu, Shoya Higa, Yumi Iwashita, Annie Didier, Tanvir Islam, Christopher Laporte, Vivian Sun, Kathryn Stack, Jacek Sawoniewicz, Shreyansh Daftry, Virisha Timmaraju, Sami Sahnoune, Chris A. Mattmann, Olivier Lamarre, Sourish Ghosh, Dicong Qiu, Shunichiro Nomura, Hiya Roy, Hemanth Sarabu, Gabrielle Hedrick, Larkin Folsom, Sean Suehr, and Hyoshin Park. Maars: Machine learning-based analytics for automated rover systems. In *2020 IEEE Aerospace Conference Proceedings*, 2020.

[18] Kyohei Otsu, Guillaume Matheron, Sourish Ghosh, Olivier Toupet, and Masahiro Ono. Fast approximate clearance evaluation for rovers with articulated suspension systems. *Journal of Field Robotics*, 37(5):768–785, 2020.

[19] Arturo Rankin, Mark Maimone, Jeffrey Biesiadecki, Nikunj Patel, Dan Levine, and Olivier Toupet. Driving curiosity: Mars rover mobility trends during the first seven years. In *2020 IEEE Aerospace Conference Proceedings*, 2020.

[20] Brandon Rothrock, Ryan Kennedy, Chris Cunningham, Jeremie Papon, Matthew Heverly, and Masahiro Ono. SPOC: Deep learning-based terrain classification for mars rover missions. In *AIAA SPACE 2016*, pages 1–12, September 2016.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[22] David Salomon. *Transformations and projections in computer graphics*. Springer Science & Business Media, 2007.

[23] S. P. Schwenzer, M. Woods, S. Karachalios, N. Phan, and L. Joudrier. Labelmars: Creating an extremely large martian image dataset through machine learning. In *50th Lunar and Planetary Science Conference*, 2019.

[24] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[25] R. L. Tokar, R. C. Elphic, W. C. Feldman, H. O. Funsten, K. R. Moore, T. H. Prettyman, and R. C. Wiens. Mars

odyssey neutron sensing of the south residual polar cap. *Geophysical Research Letters*, 30(13), 2003.

[26] Olivier Toupet, Tyler Del Sesto, Masahiro Ono, Steven Myint, Joshua Vander Hook, and Michael McHenry. A ros-based simulator for testing the enhanced autonomous navigation of the mars 2020 rover. In *2020 IEEE Aerospace Conference Proceedings*, 2020.

[27] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2020.