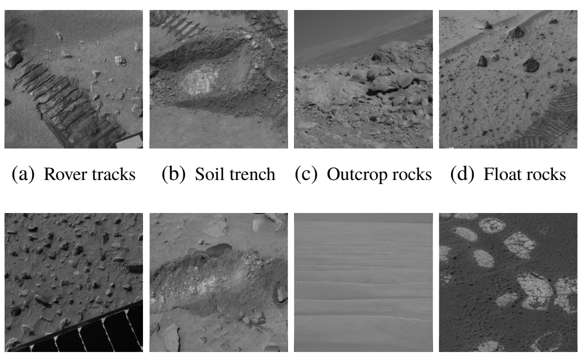
CONTENT-BASED CLASSIFICATION OF MARS EXPLORATION ROVER PANCAM IMAGES. S. Lu[1], B. Zhao[2], K. L. Wagstaff[1], S. B. Cole[3], K. Grimes[1] [1]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA, 91109, USA (you.lu@jpl.nasa.gov), [2]Duke University, Durham, NC, 27708, USA, [3]Space Science Institute, Boulder, CO, 80301, USA

Introduction: Automated classification of planetary science image data enables content-based search of large archives such as those held by the NASA Planetary Data System (PDS). For example, the PDS Image Atlas uses machine learning to help users quickly find images of interest from the HiRISE instrument on the Mars Reconnaissance Orbiter or the Mastcam and MAHLI instruments on the Mars Science Laboratory rover [1] [2].

Building on these successes, we have trained the first classifier for images collected by the Pancam instruments on the Mars Exploration Rovers (MER) and used it to classify 15,601 images in the PDS archive. These classifications can be used to search and filter Pancam images via the Atlas at https://pds-imaging.jpl.nasa.gov/search/. Unlike the classifiers employed for HiRISE, Mastcam, and MAHLI instruments, MERNet is a multi-label classifier that can detect the presence of more than one category within an image. For example, an image may be labeled as containing both "Soil" and "Clasts".



(a) Rover tracks (b) Soil trench (c) Outcrop rocks (d) Float rocks

(e) Clasts (f) Bright soil (g) Dunes/ripples (h) Spherules

Figure 1: Examples of selected classes in the MER dataset.

MER Pancam Images: The Pancam instrument is a stereo pair of science cameras, and each camera has an 8-position filter wheel. In this work, we used images taken by the second filter of the left-eye ("L2") camera, and we used only the highest resolution images (512×512 and $1{,}024 \times 1{,}024$). To generate a representative dataset, we randomly sampled 20% of the high-resolution images at each rover investigation site, yielding 3,004 images [3].

We identified 25 classes based on the MER Data Catalog User Survey with participants including scientists and engineers from MER teams, faculty members and students from research universities, and general public [4]. We also performed a class discovery process using the DEMUD [5] novelty detection algorithm which iteratively selects the most "interesting" or unusual images in a collection. Example images are shown in Figure 1, and statistics of each class are shown in Table 1. The dataset of 3,004 images was labeled using Zooniverse.org.

MERNet Classifier: The MERNet classifier is a multi-label convolution neural network (CNN) that classifies MER Pancam images into 25 classes. The MERNet classifier was built using transfer learning techniques, in which the knowledge learned from images on Earth was transferred to classify images on Mars. We used the VGG-16 CNN architecture [6] and retrained the network to output 25 Mars-relevant classes.

We employed a "classifier chain" [7] to allow the MERNet classifier to check individually for the presence of each possible class within an image. The order of classes in the chain is shown in Table 1. In contrast to the multi-label approach employed for the classification of Cassini Imaging Science Subsystem (ISS) images, in which 19 individual binary classifiers were trained to detect the presence of classes such as "craters", "horizon", "noise", "artifact", etc. [8], the classifier chain allows the explicit modeling of dependencies between classes. For example, "Bright soil" often co-occurs with "Soil trench", as in Figures 1(b) and (f).

To improve the performance of the classifier on new images, we expanded the dataset to include "augmented" variants of each image. These are generated through geometric transformation such as rotating, skewing, or shearing an image to yield a different perspective that could appear in other Pancam images. The dataset was split to use 60% for training, 15% for validation, and 25% for testing. Augmentation was applied only to the training and validation sets. The combined dataset contains 70,864 labeled images. The parameter values of the three augmentation methods were constrained such that the resultant augmented images do not crop out important features at the edges of the original images.

We fine-tuned the MERNet classifier using the training data for 34,995 iterations with a batch size of 8 images, a learning rate of 0.0001 for the of convolution layers, and a learning rate of 0.001 for the final fully connected layer. The MERNet classifier uses a confidence threshold of 0.9 to determine which images will be shown to users. To ensure that the classifier self-reported posterior probabilities are well calibrated, we employed Platt scaling [9] to adjust the output for each class. Platt scaling selects a temperature T and bias b to transform the logits Z_i output by the classifier for image x_i prior to the conversion

Table 1: Performance results for the MERNet classifier before and after calibration. Precision (0.9) and recall (0.9) indicate precision and recall given a confidence threshold of 0.9; colors indicate scores $\geq$0.80 or $\leq$0.50. Majority, minority, and extreme minority classes are marked with $*$, †, and ‡. The lowest ECE and highest precision and recall scores per row are shown in bold.

	Class distribution		Test set (before calibration)			Test set (after calibration)		
Class Name	Count	Percent	ECE	Precision (0.9)	Recall (0.9)	ECE	Precision (0.9)	Recall(0.9)
Rover deck†	222	7.39%	0.026	0.80	**0.62**	**0.012**	**0.91**	0.53
Pancam calibration target‡	14	0.45%	0.016	0.00	0.00	**0.011**	0.00	0.00
Arm hardware‡	4	0.13%	0.009	0.00	0.00	**0.006**	0.00	0.00
Other hardware‡	116	3.86%	0.044	0.00	0.00	**0.017**	0.00	0.00
Rover tracks$*$	301	10.02%	0.073	**1.00**	**0.24**	**0.020**	**1.00**	0.18
Soil trench‡	34	1.13%	0.012	0.00	0.00	**0.007**	0.00	0.00
RAT brushed target‡	17	0.57%	0.019	0.00	0.00	**0.011**	0.00	0.00
RAT hole‡	30	1.00%	0.042	0.00	0.00	**0.028**	0.00	0.00
Outcrop rocks$*$	1,915	63.75%	**0.011**	0.95	**0.57**	**0.011**	**0.96**	0.47
Float rocks$*$	860	28.63%	0.021	**0.90**	**0.40**	**0.012**	0.87	0.29
Clasts$*$	1,676	55.79%	0.033	0.94	**0.64**	**0.010**	**0.95**	0.52
Misc. rocks†	249	8.29%	0.033	0.00	0.00	**0.017**	0.00	0.00
Bright soil‡	122	4.06%	0.033	0.68	**0.34**	**0.017**	**0.92**	0.32
Dunes/ripples$*$	1,000	33.29%	0.041	0.83	**0.48**	**0.017**	**0.90**	0.33
Rock (linear feature)$*$	943	31.39%	0.016	0.84	**0.43**	**0.007**	**0.88**	0.28
Rock (round feature)†	219	7.29%	0.017	0.67	0.05	**0.009**	**1.00**	**0.08**
Soil$*$	2,891	96.24%	0.027	**0.99**	**0.96**	**0.014**	**0.99**	0.93
Astronomy‡	12	0.40%	0.010	0.00	0.00	**0.008**	0.00	0.00
Spherules$*$	868	28.89%	0.016	0.88	**0.25**	**0.012**	**0.90**	0.12
Distant vista$*$	903	30.23%	0.049	**0.96**	**0.69**	**0.023**	**0.96**	0.60
Sky$*$	954	31.76%	0.030	**0.97**	**0.86**	**0.017**	**0.97**	0.83
Close-up rock‡	23	0.77%	0.042	0.00	0.00	**0.017**	0.00	0.00
Nearby surface$*$	2,006	66.78%	0.013	0.97	**0.90**	**0.012**	**0.99**	0.83
Rover parts$*$	301	10.02%	0.060	**1.00**	**0.53**	**0.025**	**1.00**	0.39
Artifacts‡	28	0.93%	0.032	0.00	0.00	**0.012**	0.00	0.00

of Z_i into a posterior probability $p = \frac{1}{1+e^{-Z_{i,k}/T_k+b_k}}$, where $Z_{i,k}$ is the logit for image i and class k. The parameters T and b are optimized using L-BFGS algorithm with respect to the binary cross entropy loss on the validation set. Expected calibration error (ECE) [10] is used to quantitatively measure classifier calibration.

Results: The performance of the MERNet classifier is measured in terms of precision and recall per class on the test set (Table 1). We designate the classes as majority (>10%), minority (5–10%), or extreme minority (<5%) based on their representation. Majority classes (e.g., "Outcrop rocks", "Soil") achieve good performance. Minority class (e.g., "Miscellaneous rocks", "Rock (round feature)") performance is generally lower. Extreme minority classes (e.g., "Pancam calibration target", "Arm hardware") often exhibit precision and recall of zero due to the very small number of labeled examples. Up-sampling minority and extreme minority classes may be helpful for improving their performance, and this is an area of future work. Applying the Platt scaling method consistently improved the classifier calibration (see Table 1 ECE columns; lower is better) and generally increased the thresholded precision scores at the cost of reducing recall scores (by abstaining), thereby yielding more reliable predictions. The only class to exhibit lower precision was "Float rocks" (declined from 0.90 to 0.87).

Conclusions and Next Steps: In this work, we demonstrated the use of deep learning fine-tuning techniques to create the first machine learning application that annotates MER Pancam image content and enables users to quickly find images of interest. Future work will focus on improving performance for the minority classes. In addition, we aim to expand these capabilities to images of other bodies such as the Moon.

Acknowledgments: Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References: [1] Wagstaff, K.L. *et al.* (2018) *Innovative Applications of AI*. [2] Wagstaff, K.L. *et al.* (2021) *Innovative Applications of AI*. [3] Zhao, B. *et al.* (2020) *Zenodo*, http://doi.org/10.5281/zenodo.4302760. [4] Cole, S. B. *et al.* (2020) *LPSC*. [5] Wagstaff, K. *et al.* (2013) *AAAI*. [6] Simoyan, K. *et al.* (2015) *ICLR* . [7] Read, J. *et al.* (2011) *Machine Learning*. [8] Stanboli, A. *et al.* (2017) *3rd Planetary Data Workshop*. [9] Platt, J. (1999) *Advances in Large Margin Classifiers*. [10] Guo, C. *et al.* (2017) *ICML*.