Self-Supervised Learning to Guide Scientifically Relevant Categorization of Martian Terrain Images

Tejas Panambur[1,*] Deep Chakraborty[2,*] Melissa Meyer[3] Ralph Milliken[3]
Erik Learned-Miller[2] Mario Parente[1]
[1] Department of Electrical and Computer Engineering, UMass Amherst
[2] Manning College of Information and Computer Sciences, UMass Amherst
[3] Department of Earth, Environmental, and Planetary Sciences, Brown University

Abstract

Automatic terrain recognition in Mars rover images is an important problem not just for navigation, but for scientists interested in studying rock types, and by extension, conditions of the ancient Martian paleoclimate and habitability. Existing approaches to label Martian terrain either involve the use of non-expert annotators producing taxonomies of limited granularity (e.g. *soil, sand, bedrock, float rock, etc.), or rely on generic class discovery approaches that tend to produce perceptual classes such as rover parts and landscape, which are irrelevant to geologic analysis. Expert-labeled datasets containing granular geological/geomorphological terrain categories are rare or inaccessible to public, and sometimes require the extraction of relevant categorical information from complex annotations. In order to facilitate the creation of a dataset with detailed terrain categories, we present a self-supervised method that can cluster sedimentary textures in images captured from the Mast camera onboard the Curiosity rover (Mars Science Laboratory). We then present a qualitative analysis of these clusters and describe their geologic significance via the creation of a set of granular terrain categories. The precision and geologic validation of these automatically discovered clusters suggest that our methods are promising for the rapid classification of important geologic features and will therefore facilitate our long-term goal of producing a large, granular, and publicly available dataset for Mars terrain recognition. Code and datasets are available at https://github.com/TejasPanambur/mastcam.*

1. Introduction

Automatic terrain recognition has aided the navigational operations of Mars rovers by solving challenges such as traversability analysis [45, 49], slip prediction [12, 16], and minimizing driving energy [22]. A primary scientific objective of both the Mars Science Laboratory (MSL or Curiosity rover) [18] and Mars 2020 (Perseverance rover) [58] is to answer questions about water activity and the potential for past life on Mars. Toward this objective, a necessary step is to discriminate and/or correlate the various rock types and rock textures encountered along the rover traverse. The missions also seek to understand the geological history and evolution of the planet, and to prepare for future robotic and human exploration. These objectives are pursued through a plethora of geological and geochemical experiments onboard the rovers [9, 10, 35, 37]. Crucially important are investigations related to imaging devices as they provide essential geologic context and highlight the presence of morphological or sedimentological features that may be indicative of water alteration (*e.g*. the presence of fractures in-filled by veins or nodular structures that are directly responsible for chemical and mineralogical changes in the rock due to the interaction with water). The prospect of automating the difficult and somewhat subjective task of identifying and cataloging geomorphological and textural classes in Martian terrain would greatly improve the scientific return of Martian missions by allowing scientists to focus on more fundamental analyses, and speculations on formational mechanics, rather than classification. The need to provide scientists with relevant and granular terrain classes clearly arises. Deep learning approaches have made great strides towards solving terrain classification tasks [4,16,41,44,45,51,52,56], but the datasets and approaches available for rock classification are still very limited.

In this paper, we develop an approach that can automatically assemble morphological and textural categories from the large amounts of unlabeled images available. An assessment from planetary scientists is used to validate our discoveries and their ability to represent clear scientific phe-

[*] Authors contributed equally. Correspondence to: {tpanambur, dchakraborty}@umass.edu

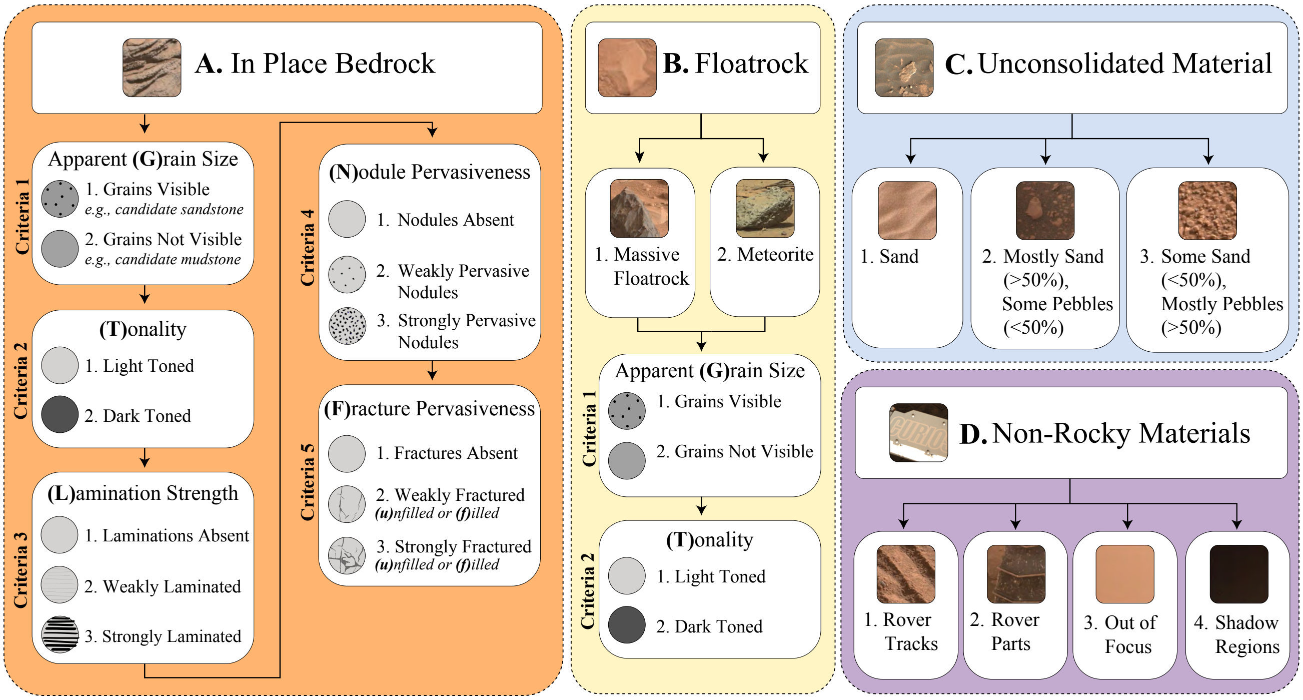

Figure 1. Taxonomy of geological classes (as found in our dataset). Note that a given class can be coded into any of the following categories using a combination of alphabets and numbers as shown in bold. For instance, weakly laminated red mudstone, which is non-nodular, and has calcium sulfate-filled hairline planar fractures can be represented as `A-G2-T1-L2-N1-F2f`.

nomena relevant to current scientific research.

Mars surface images originate from two primary sources: Mars rover missions *e.g*. MSL, Mars2020, etc., *and* Mars orbiter missions *e.g*. Mars Global Surveyor (MGS), Mars Reconnaissance Orbiter (MRO), etc. While terrain can be identified both from ground and orbiter images, orbiter images are simply not acquired at scales fine enough to discern finer textural details that are necessary to uniquely identify categories related to rock types or the presence of diagnostic small-scale alterations. Therefore datasets constructed using these images [14, 45, 57] are not useful for the task at hand. Ground images captured from a variety of cameras mounted on rovers (*e.g*. mastcam, navcam, and chemcam) on the other hand, are acquired at the necessary scale to facilitate detailed geomorphological analysis. However, efforts to label these images [49, 51, 52] usually produce coarse labels such as “sand”, “soil”, and “bedrock” as a result of non-expert annotations *or* perceptual classes such as “rover parts” and “landscape” as a result of generic class discovery approaches [53] precluding geologic analysis. Expert-labeled datasets on the other hand, exhibit more detailed terrain categories [48] but occasionally suffer from limited usability due to their complex annotation-based representations [44]. Such labeled datasets are, furthermore, inaccessible to the public.

This work is born from the desire to create a large, publicly accessible database of scientifically relevant textural/morphological terrain categories from readily available rover images, with a method that addresses the discrepancies described above. We propose a self-supervised deep clustering algorithm that can automatically group robust terrain categories by utilizing a network designed for texture recognition. A geology expert then provides a qualitative assessment of the discovered clusters and the scientific significance of such clusters. The expert also assigns a set of granular labels to images selected randomly from the discovered clusters. These labels have broad categories that can be further divided into subcategories based on different attributes (similar to LabelMars [48]). Here, top-level categories such as bedrock, floatrock, unconsolidated material, and non-rocky materials, have been further classified based on types of rock formations and other attributes such as grain size, tonality (hue), apparent lamination strength, nodule and fracture pervasiveness, as applicable (Figure 1). We don’t claim that our work is a gold standard for terrain classification, but believe it to be a step in the right direction. Finally, we use the expert-derived labels to evaluate the quality of our clustering algorithm both by showing that it can produce homogeneous and well separated clusters, as well as evaluating its precision on a test set.

Our contributions in this paper are two-fold:

1. We develop a novel synthesis of deep texture encoding techniques and self-supervised deep clustering algorithms to support rapid and robust terrain categorization.
2. We produce an exhaustive taxonomy for classifying Martian terrain as seen in curiosity mastcam images, designed by a planetary geologist and supported with a review of how our approach could help geologic exploration.

2. Related work

2.1. Efforts to label Martian terrain

Several efforts have been have been proposed for Mars terrain classification [14, 44, 45, 48, 49, 51, 52, 55, 57]. Earlier works, such as [14, 45, 57] annotated orbiter images from the MRO Context Camera (CTX) [33] or High Resolution Imaging Science Experiment (HiRISE) [36] to facilitate rover navigation. Although some of these works have defined important geomorphological categories, orbiter images with a resolution of around 25 cm/pixel are only partially suited for detailed geological analysis. This motivated the creation of labeled datasets using ground images captured from Mars rover cameras that have a much higher resolution (150 or 450 μm/pixel in the MSL right and left mastcam [1] for instance). It is the following works therefore, with which we compare and contrast in this paper. Wagstaff *et al*. curate a set of around 9000 images in total using a combination of expert annotation and automatic class discovery using the DEMUD [53] algorithm [51, 52]. These labels include a variety of rover parts and artifacts, and a small set of geological categories such as float rock, layered rock, veins, sand, etc. AI4Mars [49] is currently the largest labeled Mars terrain dataset, containing $326k$ images with semantic segmentation labels of categories such as soil, bedrock, sand, and big rock. These labels, while simple and useful for tasks such as navigation, are not granular enough for geological analyses. Inspired by the need for a content-based search system that enables scientists to interact with the rover using natural language descriptions, Qiu *et al*. propose SCOTI [44]. They create a Mars image caption dataset starting from 1250 expert captioned images, training an image caption model, and progressively growing using predictions on unlabeled images with open/expert review. The final dataset contains more than $12,500$ images with natural language descriptions which include relevant geomorphological features together with general statements of limited relevance. A scientist seeking to carefully catalog the different textural/morphological categories in an image would find it difficult to extract a large set of uniquely identifiable class labels from the long descriptions. This annotated dataset is publicly unavailable. More similar to our effort is the the LabelMars project [48] that annotated 5000 images with the help of undergraduate geology students using a labeling scheme based on hierarchical morphological categories. Coarse categories such as "sedimentary", "magmatic", or "meteoric" are further divided into sub-categories such as concretions/nodules and light/dark tonality, but this dataset is also publicly unavailable. Our proposed approach also discovers similar classes with high granularity that are labeled by an expert (see Fig. 1), albeit in an self-supervised fashion. This work pushes the labeling effort to the point that the identified categories reproduce exhaustively the set of criteria that are necessary for a complete characterization of the terrain that is possible by an expert with the imaging data available. Our dataset is described in detail in Sec. 4.

2.2. Terrain recognition

Terrain recognition is a popular area of research in computer vision due to its vast applications in autonomous driving and terrain classification. It is usually cast as a texture recognition problem, and traditional approaches used geometric features, particularly curvature, color features, lighting, illumination direction, and photometric properties [5,15,29,61], followed by feature pooling as seen in bag-of-words models [11,28]. Recently, deep learning models with end-to-end texture/terrain recognition have gained popularity. However, naive CNN architectures are inadequate for texture recognition as they aren't invariant to spatial layout, and recognizing textures typically requires preserving some orderless information. Therefore, texture recognition approaches usually have distinct architectures compared to CNNs used for object recognition, to preserve fine textural details. Zhang *et al*. introduce a CNN architecture with end-to-end dictionary learning and feature pooling for orderless texture recognition [65]. Further, Xue *et al*. hypothesize that surfaces are not completely orderless, and propose the Deep Encoding Pooling Network (DEP) [60] that incorporates both orderless texture details and local spatial information, combined using bilinear pooling [30]. Several state-of-the-art approaches have since been proposed by integrating stronger geometric priors into deep networks or by encoding features from different layers in the network [8,23,59,63,64]. However, our method is based off of Xue *et al*. [60] for its simplicity and adaptability to a self-supervised objective such as the one we use in this paper.

Self-supervised and unsupervised deep embedding learning is a relatively new and active area of research [3, 6, 7, 17, 20, 25, 27, 40, 62]. A majority of the existing methods work fairly well on object-centric [26, 32, 46, 54] or scene-centric [13, 19, 31] datasets. However, very little research exists for unsupervised texture/terrain recognition [42, 43]. Panambur and Parente [43] use an alternating clustering and classification approach similar to DeepCluster [2] to cluster mars terrain images. The resulting learned representations form homogeneous clusters of different kinds of terrain. In a followup work, Deep Clustering using Metric Learning (DCML) [42], the authors train a triplet network by iteratively clustering the features and using the cluster labels to form triplets. The features generated have high *inter*-class distance while preserving low *intra*-class distance, making it ideal for retrieval tasks. However, there is no expert evaluation of the quality of clustering, unlike our work. Moreover, these approaches still suffer from the same drawbacks of traditional CNN architectures used for texture recognition. To rectify this problem, we incorporate the texture encoding module [60] into the CNN architecture to get better terrain recognition performance, while using the same metric learning approach as [42] in order to leverage the large amounts of unlabeled Mars terrain images. We hypothesize that the learned representation can capture better the nuances between rocks that are geologically significant, while grouping similar types of terrain together.

3. Deep self-supervised texture recognition

We first present the network architecture that is used in our approach in Sec. 3.1. This architecture was specifically designed for texture classification, and modified to support self-supervised training. We then describe our self-supervised training objective in Sec. 3.2.

3.1. Deep encoding pooling network

We start from the CNN architecture described in [60]. Given an input image $\mathbf{I}$, and the backbone feature extraction function F_θ (in this case ResNet-18 [21]), we obtain features $\mathbf{X}_f = F_\theta(\mathbf{I})$. The outputs from the feature extractor feed two separate layers in the network. One of these is a texture encoding layer [65] that produces an orderless representation of the features and preserves fine textural details, the kind that can be seen in terrain images. Given a feature map $\mathbf{X}_f$ and an encoding layer F_e we obtain a texture embedding $\mathbf{X}_t = F_e(\mathbf{X}_f)$. The other layer is the usual global average pooling (GAP) layer that preserves spatial information, and the pooled embedding is defined as $\mathbf{X}_g$. The outputs of these layers are combined using a bilinear pooling layer [30] that helps in capturing the relationship between the orderless texture information and spatial information. The output embedding $\mathbf{X}_b$ from the bilinear layer combines texture features $\mathbf{X}_t$ and spatial features $\mathbf{X}_g$. Further, a fully connected layer and a linear classification layer are normally used to train the network using a supervised classification objective. We remove the final classification layer from the network that is generally used to map the embeddings to the fixed number of labeled categories in the dataset. Since we train our network on unlabeled data, we directly use the embeddings generated on the penultimate layer (fc-7 features) given by $\mathbf{X}_{emb}$.

3.2. Deep clustering using metric learning

The DCML algorithm alternatively clusters the embeddings using standard K-means clustering algorithm and uses the subsequent assignments as pseudo-labels to train a metric learning objective [42]. We cluster the features from the embedding layer $\mathbf{X}_{emb}$, and assign pseudolabels to these clusters so that they can be used to generate triplets comprising an anchor, a positive, and a negative example. Positive examples are samples from the same cluster as the anchor, and negative examples belong to a different cluster. These triplets are used to minimize a distance metric objective called triplet loss. The triplet loss minimizes the distance between the anchor and positive examples and maximizes the distance between the anchor and negative examples in the embedded space. We use the same triplet loss objective with triplet sampling strategy defined in [42] as:

$$\mathcal{L} = \sum_{i=1}^{N} ||\mathbf{X}^i_{emb_a} - \mathbf{X}^i_{emb_p}||^2_2 - ||\mathbf{X}^i_{emb_a} - \mathbf{X}^i_{emb_n}||^2_2 + \alpha \quad (1)$$

where $\mathbf{X}^i_{emb_a}$, $\mathbf{X}^i_{emb_p}$, $\mathbf{X}^i_{emb_n}$ are the embeddings of the anchor, positive, and negative examples, and α is the desired margin [42, 47]. We set $\alpha = 1$ in our experiments.

4. Dataset

4.1. Sourcing data and preprocessing

Our dataset consists of DRCL (decompressed, radiometrically calibrated, color corrected, and geometrically linearized) images [34] acquired by the MSL Mast cameras between sol 1 and sol 2800. We restrict our focus to terrain images by following the settings in [42] to eliminate the majority of images containing rover parts, sky, etc. In order to prevent scale disparities in terrain features, we limit our analysis to images in which the distance of a target to the rover is within $15m$. The resulting dataset contains $30,000$ images of size typically 1200×1600 pixels. Since an image of this size might contain several different kinds of terrain, smaller patches of size 128×128 and 256×256 were extracted from it using a sliding window with a stride 50% that of the patch size. The difference in patch size corresponds to the difference in focal lengths of the two cameras that make up the “left eye” and “right eye” of the curiosity rover, and were applied accordingly. This yields a total of $2.4M$ patches which is twice the size of the ImageNet-1k dataset [46].

A problem of class balancing still remained as a large number of patches contain mostly unconsolidated terrain and classes of geologic relevance are less frequent (long-tailed distribution). Therefore, in order to avoid any distribution mismatch between the training and test sets, we do not select patches randomly for each set (as is common practice). Instead, we follow a careful patch extraction strategy where patches in the training set are selected from the left 60% portion of a given image, and patches from the remaining 40% of the image are reserved for the test set. Criticism then arises that such an approach could lead to different views of the same geographical area appearing in both training and test sets due to panning of the camera. Though this is a rare occurrence (mastcam has a small field of view), two patches imaging the same area have different viewpoints and are as similar or different as two samples drawn from the same distribution albeit strongly correlated. Therefore, in addition to ensuring that the same "patch" is never used for both training and testing, patches captured from the same *site* and *drive* of the rover are removed from the results before evaluation.

4.2. Grouping data and discovering classes

Once the model described in Sec. 3 is trained on the unlabeled data obtained as above, we collect expert feedback on the model performance through a web-based user interface. We think that a natural way to do this is using query images to retrieve the top-K most similar images from the dataset (K-Nearest Neighbors). The query images are randomly sampled from the clusters found by the model. The nearest neighbors are the top-K images in the dataset whose embeddings (output of the penultimate fully connected layer of the trained network or fc-7 features) are closest (using a distance measure such as euclidean distance) to the embedding of the query image. We then display the query images along with top-K nearest neighbors on a webpage which supports image annotation. The interface is shown in the appendix (Fig. A.1). We ask an expert in planetary geology to do the following:

1. Characterize and label the query image using detailed geological/geomorphological criteria.
2. Assess how many of the top-K neighbors belong to the same category as the query image and identify the mistakes.
3. Evaluate the homogeneity of a randomly sampled subset of clusters produced by the network and comment on their geological relevance.

Several geological categories naturally emerge from this process (See Sec. 5.2 for the analysis). The top level categories include "bedrock", "floatrock", "unconsolidated material", and "non-rocky materials". These are further subdivided based on rock type, grain size, tonality (hue), lamination strength, fracture and nodule pervasiveness, and preponderance of unconsolidated materials as apparent from the image. The full taxonomy is shown in Fig. 1.

We can traverse this hierarchy to generate detailed and uniquely-defined classes and also assign them a taxonomy code. For instance, a particular traversal could be represented as `A-G2-T1-L2-N1-F2f`. Here `A-G2` denotes bedrock with no visible grain (*e.g.* mudstone), `T1` is light-toned (in this case red colored), `L2` indicates weakly laminated, `N1` encodes the absence of nodules on this rock, and `F2f` indicates that the rock is commonly (lightly) fractured with calcium sulfate filled veins. Note that since the above classifications were formulated by the expert after looking at the clusters produced by our model, our model reflects the kind of granular observations of terrain features a geologist would have to make as part of their scientific study (see discussion in Sec. 5.2). Our taxonomy provides an exhaustive set of criteria that could be used to classify terrain, and leaves room for the addition of even finer categories still (which may not be present in our dataset), depending on the application, without any need to alter the hierarchy.

This level of granularity in categories in our dataset, while complex, is unprecedented for any terrain recognition dataset on Mars (and possibly Earth) and demands a high level of sophistication from automatic terrain classifiers. Note however that expert review is a long process, and due to the time available and also potential undersampling of certain categories, only a finite number of labels (25) could actually be identified in the sampled images. The full list of class descriptions along with their taxonomies is presented in the appendix (Tab. A.1). Most of these classes are from the *bedrock* category, as it is the focus of MSL missions and also a central category for geologic analysis. We plan to expand the number of classifications available for floatrocks and unconsolidated materials in future work.

5. Experiments

5.1. Implementation details

For the proposed method we use the DEP architecture [60] with ImageNet pretrained 18-layer ResNet [21] as the backbone. The dimensionality of the embedding layer is set to 512. For the clustering, the embeddings are PCA-reduced to 256 dimensions, whitened and ℓ^2-normalized. We use Faiss K-means clustering algorithm [24] with K set to 150 as determined by visual inspection of clusters. We use the SGD optimizer to train our network with a learning rate of $1\text{e}{-4}$ and weight decay of $1\text{e}{-5}$. The size of a minibatch corresponds to the number of samples per cluster times the number of clusters. We set number of samples per cluster as 4, and the resulting batch size is 600. We train on 4 Nvidia M40 GPUs, and training takes around 3 days. The criteria for convergence is the cluster stability over epoch. This is measured using Normalized Mutual Information (NMI) by calculating it between cluster assignments at epoch t and $t-1$ [2]. We find that the training saturates at epoch 40 with NMI$= 0.756$. We resize the image patches into 224×224. Since our model is trained to identify terrain features, it

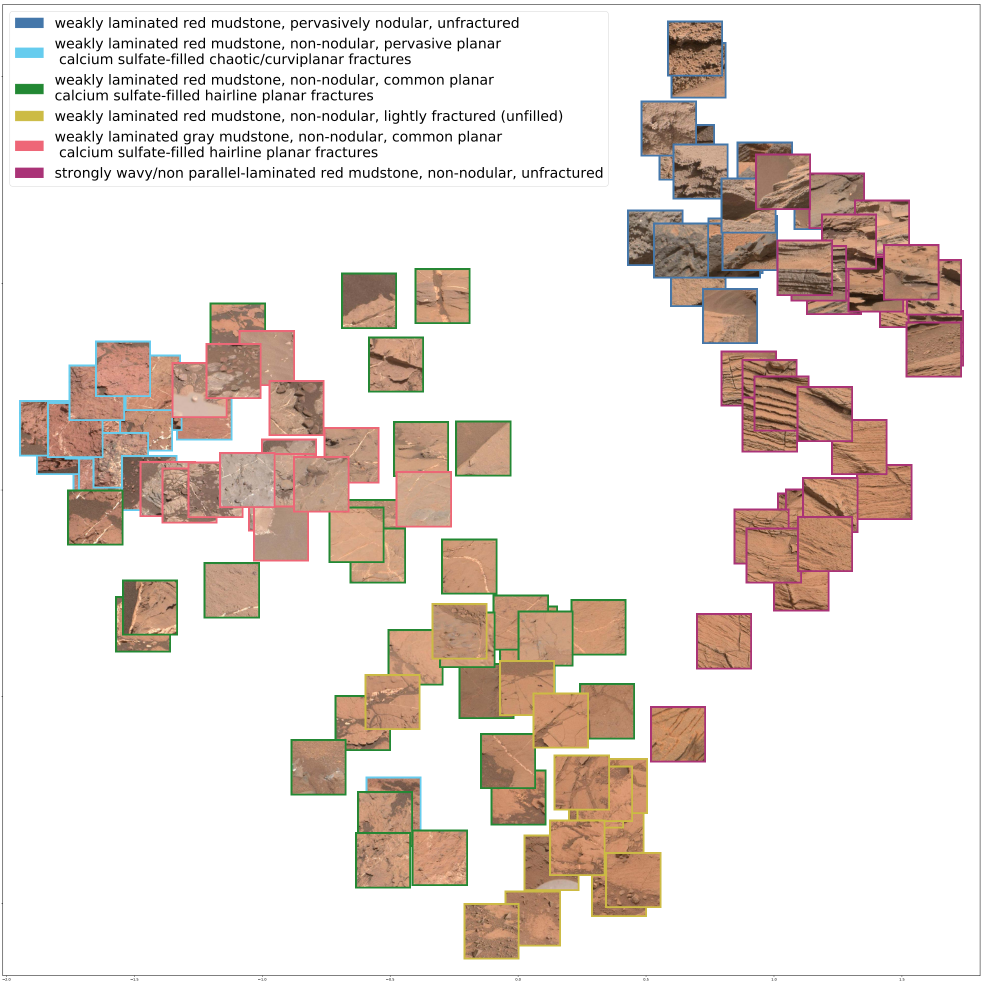

Figure 2. t-SNE [50] visualization of 6 clusters from the model. The clusters show good separation, and overlap only when an increasing number of features are common (*e.g*. green and yellow). Best viewed in color (accessible) on a pdf processor with zoom.

can be very sensitive to scale and orientation of the images. Therefore, we avoid using augmentations such as random resized crops and horizontal flips in order to avoid training on instances that might represent unrealistic viewpoints or non-existent geological formations.

5.2. Evaluation

Our clustering performance is evaluated in two ways: using a visual inspection of homogeneity and cluster separation from t-SNE [50] plots of learned embeddings, and computing the precision of a retrieval task from a test set given a query image. The precision of our model for retrieval tasks demonstrates the usefulness of our approach by allowing scientists to automate the process of finding visually similar terrain. Additionally, an expert opinion on the scientific significance of the clusters found by our model adds depth to the analysis of our model's performance and shows how our approach could be used to support geologic exploration.

Qualitative. Figure 2 shows a plot of 6 clusters obtained by projecting the embedding vectors for the points (512-dimensions) onto 2-dimensions. Notice how the clusters are homogeneous and well separated in most cases, with occasional overlap in the case of highly similar terrain. We asked the expert to review a small subset of clusters produced by the model by looking at randomly sampled images from these clusters, and explain if our clusters could be useful for geological analysis the details of which follow.

Here, the clusters in green and yellow represent two classes of weakly laminated red mudstone, that are non-nodular and fractured. The sole difference is that the fractures are either filled with calcium sulfate or unfilled, respectively. Using automation to make these sorts of subtle distinctions could be useful for scientists to rapidly make interpretations with important science and operational impli-

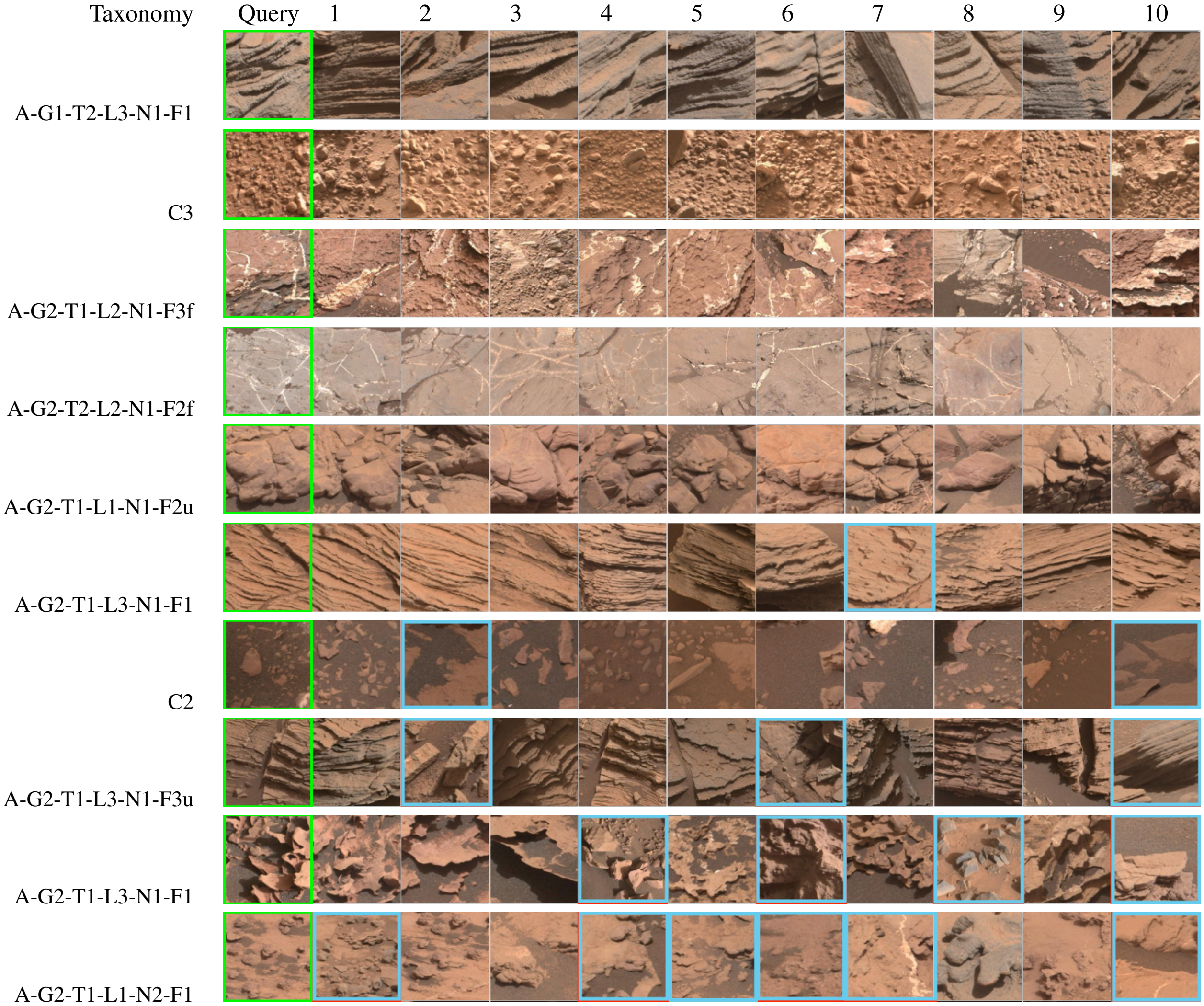

Figure 3. Top-10 nearest neighbors retrieved for 10 query images (green). Nearest neighbors are from the test set, and restricted to be from different sites along the rover traverse. The rows are arranged in ascending order of mistakes (cyan) made by the model while retrieving patches. Only 10 out of 25 classes have been shown here along with their taxonomies. Best viewed in color (accessible).

cations. In this example scenario, one could plausibly identify two generations of fracturing events within the dataset: one in which water was available to fill the fractures with calcium sulfate, and one in which it was not. Note how these clusters still have some overlap in the t-SNE plot which is consistent with their highly similar nature.

Another similar type of rock is seen in the red cluster, where the rocks appear less red. This could either be because dust cover is minimized here, or because there is an inherent change in bedrock composition. This could be verified by spatially mapping these images and comparing whether these regions overlap with spectrally bland regions, which are often attributed to dust cover in CRISM data [38]. In yet another well separated cluster (purple), we can see instances of well laminated, coherent bedrock from which sedimentary geologists typically infer ancient depositional processes. This cluster may be useful to this community as a way to rapidly identify exposure to facilitate more detailed sedimentological analysis.

Moreover, geologists are often interested in finding and mapping peculiarities in terrain, such as an increased amount of nodule formations in the bedrock as seen in the blue cluster, or those which appear unusual such as the inordinately red rocks seen in the cyan cluster. The large number of images present make this task impossible to accomplish manually. Our model can quickly retrieve all images that are similar to a provided example, which could then be used to generate viewsheds on top of Mars orbital views [39] allowing geologists to corroborate their findings using global context.

Table 1. Precision@10 obtained by expert review of each category present in our dataset. Only taxonomies are shown here to preserve space, the full description of classes is available in Tab. A.1.

ID	Taxonomy	Precision@10
1	A-G1-T2-L3-N1-F1	1.0
2	A-G2-T1-L3-N1-F1	0.9
3	A-G2-T1-L3-N1-F1	0.9
4	A-G2-T1-L3-N1-F3u	0.7
5	A-G2-T1-L3-N1-F1	0.6
6	A-G2-T1-L2-N1-F1	0.8
7	A-G2-T1-L2-N1-F2u	1.0
8	A-G2-T1-L2-N1-F3u	0.9
9	A-G2-T1-L2-N1-F2f	1.0
10	A-G2-T2-L2-N1-F2f	1.0
11	A-G2-T1-L2-N1-F3f	1.0
12	A-G2-T1-L2-N3-F1	0.9
13	A-G2-T1-L2-N3-F3f	1.0
14	A-G2-T1-L1-N1-F2	1.0
15	A-G2-T1-L1-N3-F1	1.0
16	A-G2-T1-L1-N2-F1	0.4
17	B1-G2-T1	0.7
18	B1-G2-T2	0.3
19	C1	1.0
20	C2	0.8
21	C3	1.0
22	D1	0.1
23	D2	0.9
24	D3	1.0
25	D4	1.0
	Avg.	**0.836**

Quantitative. Due to the lack of supervised annotations available for our data, we measure the performance of our model using a retrieval task. Given a query image $\mathbf{I}$ with known category $c_{\mathbf{I}}$, we poll the model to retrieve the top-K images $\mathbf{I}'_k$ from the test set whose embeddings $F_\theta(\mathbf{I}'_k)$ have the least distance to the embedding of the query image $F_\theta(\mathbf{I})$. Here F_θ represents our deep network parameterized by θ, and the distance measure selected is the euclidean distance between two vectors, $d(\mathbf{u}, \mathbf{v}) = ||\mathbf{u}-\mathbf{v}||_2$. We use the metric Precision@K to evaluate the quality of the retrieved images. Precision@K is defined as follows:

$$\text{Precision@}K \triangleq \frac{1}{K} \sum_{k \in 1..K} \mathbb{I}(c_{\mathbf{I}} = c_{\mathbf{I}'_{\mathbf{k}}}) \qquad (2)$$

where $\mathbb{I}$ is the indicator function. Here, we use $K = 10$. Since we don't have actual labels available for the retrieved images, we once again seek expert review to correctly identify the classes of all retrieved images. The results for each category are shown in Tab. 1. Figure 3 shows a subset of labeled query images from our dataset and the retrieved nearest neighbors. The nearest neighbors are restricted to be from different sites along the rover traverse in the Gale crater. This demonstrates our model's generalization performance and ability to retrieve interesting terrain images from different locations using a query image. This mechanism could potentially be used to propagate the label of a query image to similar *unlabeled* images and overlaid on a map to study terrain changes along the rover traverse. Our model obtains an overall Precision@10 of 83.6%. Queries from 12 (out of 25) classes in our data have 100% precision of retrieval (subset shown in first 5 rows in Fig. 3) and the performance degrades slightly for other classes. Common failure modes include bedrock with nodules confused with pebbles on the surface of the bedrock (Row 10 in Fig. 3) and rover tracks mistaken for fractured or strongly laminated bedrock (not shown). The former is a result of natural scale variations in the dataset due to the focus distance and the model's inability to capture them, whereas the latter happens in part because of the high visual similarity of sand tracks to laminated rocks and in part because of limited data available for such classes.

6. Conclusion

We presented a framework that would allow the creation of a large database of robust, readily available, and geologically relevant terrain categories of the Martian surface based on mastcam images. Our self-supervised deep clustering algorithm can automatically identify nuanced terrain categories by utilizing a network designed for texture recognition. The automatically discovered clusters enabled the creation of a robust taxonomy of scientifically relevant terrain categories through expert assessment, that can be used to rapidly label images. The granularity and homogeneity of the discovered clusters were evaluated using such labels qualitatively and quantitatively. The agreement between the membership identity of the clusters and the expert terrain categories show promise for extensive, automated analysis of geologic features on the Martian surface pertaining to a better understanding of depositional processes and interpreting the paleoclimate to ultimately answer the question of whether life once existed on Mars.

Acknowledgements

This work was sponsored in part by the AFRL and DARPA under agreement FA8750-18-2-0126, and by NASA under grant 80NSSC19K1227. Part of this work was conducted on HPC equipment managed by the MassTech Collaborative. The authors would like to thank the anonymous reviewers for their invaluable comments, and Debasmita Ghose, Akanksha Atrey, and Karan Rakesh for proofreading the manuscript.

References

[1] James F Bell III, A Godber, S McNair, MA Caplinger, JN Maki, MT Lemmon, J Van Beek, MC Malin, D Wellington, KM Kinch, et al. The mars science laboratory curiosity rover mastcam instruments: Preflight and in-flight calibration, validation, and data archiving. *Earth and Space Science*, 4(7):396–452, 2017. 3

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 4, 5

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 3

[4] Anirudh S Chakravarthy, Roshan Roy, and Praveen Ravirathinam. Mrscatt: A spatio-channel attention-guided network for mars rover image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1961–1970, 2021. 1

[5] Michael John Chantler, G McGunnigle, A Penirschke, and M Petrou. Estimating lighting direction and classifying textures. 2002. Eleventh British Machine Vision Conference, BMVC 2000 ; Conference date: 11-09-2000 Through 14-09-2000. 3

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 3

[8] Zhile Chen, Feng Li, Yuhui Quan, Yong Xu, and Hui Ji. Deep texture recognition via exploiting cross-layer statistical self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5231–5240, 2021. 3

[9] Agnes Cousin, Violaine Sautter, Valérie Payré, Olivier Forni, Nicolas Mangold, Olivier Gasnault, Laetitia Le Deit, Jeff Johnson, Sylvestre Maurice, Mark Salvatore, et al. Classification of igneous rocks analyzed by chemcam at gale crater, mars. *Icarus*, 288:265–283, 2017. 1

[10] A Cousin, V Sautter, V Payré, O Forni, N Mangold, O Gasnault, L Le Deit, PY Meslin, J Johnson, S Maurice, et al. Classification of 59 igneous rocks analyzed by chemcam at gale crater, mars. In *Ninth International Conference on Mars*, volume 2089, page 6075, 2019. 1

[11] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004. 3

[12] Chris Cunningham, Masahiro Ono, Issa Nesnas, Jeng Yen, and William 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. IEEE, 2017. 1

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 3

[14] Soumya Ghosh, Tomasz F Stepinski, and Ricardo Vilalta. Automatic annotation of planetary surfaces with geomorphic labels. *IEEE Transactions on Geoscience and Remote Sensing*, 48(1):175–185, 2009. 2, 3

[15] D. B. Goldof and T. S. Huang. A curvature-based approach to terrain recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(11):1213–1217, nov 1989. 3

[16] Ramon Gonzalez and Karl Iagnemma. Deepterramechanics: Terrain classification and slip estimation for ground robots via deep learning. *arXiv preprint arXiv:1806.07379*, 2018. 1

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 3

[18] John P Grotzinger, Joy Crisp, Ashwin R Vasavada, Robert C Anderson, Charles J Baker, Robert Barry, David F Blake, Pamela Conrad, Kenneth S Edgett, Bobak Ferdowski, et al. Mars science laboratory mission and science investigation. *Space science reviews*, 170(1):5–56, 2012. 1

[19] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[21] 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. 4, 5

[22] Shoya Higa, Yumi Iwashita, Kyohei Otsu, Masahiro Ono, Olivier Lamarre, Annie Didier, and Mark 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. 1

[23] Y. Hu, Z. Long, and G. AlRegib. Multi-level texture encoding and representation (multer) based on deep neural networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019. 3

[24] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 5

[25] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 3

[27] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6874–6883, 2017. 3

[28] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE, 2006. 3

[29] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision (IJCV)*, 43(1):29–44, 2001. 3

[30] Tsung-Yu Lin and Subhransu Maji. Improved bilinear pooling with cnns. *CoRR*, abs/1707.06772, 2017. 3, 4

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[32] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 3

[33] Michael C Malin, James F Bell III, Bruce A Cantor, Michael A Caplinger, Wendy M Calvin, R Todd Clancy, Kenneth S Edgett, Lawrence Edwards, Robert M Haberle, Philip B James, et al. Context camera investigation on board the mars reconnaissance orbiter. *Journal of Geophysical Research: Planets*, 112(E5), 2007. 3

[34] Michal C Malin, Michael A Ravine, Michael A Caplinger, F Tony Ghaemi, Jacob A Schaffner, Justin N Maki, James F Bell III, James F Cameron, William E Dietrich, Kenneth S Edgett, et al. The mars science laboratory (msl) mast cameras and descent imager: Investigation and instrument descriptions. *Earth and Space Science*, 4(8):506–539, 2017. 4

[35] Nicolas Mangold, Mariek E Schmidt, Martin R Fisk, Olivier Forni, Scott M McLennan, Doug W Ming, Violaine Sautter, Dawn Sumner, Amy J Williams, Samuel M Clegg, et al. Classification scheme for sedimentary and igneous rocks in gale crater, mars. *Icarus*, 284:1–17, 2017. 1

[36] Alfred S McEwen, Eric M Eliason, James W Bergstrom, Nathan T Bridges, Candice J Hansen, W Alan Delamere, John A Grant, Virginia C Gulick, Kenneth E Herkenhoff, Laszlo Keszthelyi, et al. Mars reconnaissance orbiter's high resolution imaging science experiment (hirise). *Journal of Geophysical Research: Planets*, 112(E5), 2007. 3

[37] MJ Meyer, RE Milliken, KM Robertson, and JA Hurowitz. Microscale chemical and spectral characterization of clay-bearing evaporites and implications for the mars 2020 rover. In *Lunar and Planetary Science Conference*, number 2326, page 1705, 2020. 1

[38] Scott Murchie, R Arvidson, Peter Bedini, K Beisser, J-P Bibring, J Bishop, J Boldt, P Cavender, T Choo, RT Clancy, et al. Compact reconnaissance imaging spectrometer for mars (crism) on mars reconnaissance orbiter (mro). *Journal of Geophysical Research: Planets*, 112(E5), 2007. 7

[39] Marion Nachon, Schuyler Borges, RC Ewing, Frances Rivera-Hernández, Nathan Stein, and JK Van Beek. Coupling mars ground and orbital views: Generate viewsheds of mastcam images from the curiosity rover, using arcgis® and public datasets. *Earth and Space Science*, 7(9):e2020EA001247, 2020. 7

[40] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 3

[41] Leon F. Palafox, Christopher W. Hamilton, Stephen P. Scheidt, and Alexander M. Alvarez. Automated detection of geological landforms on mars using convolutional neural networks. *Computers & Geosciences*, 101:48–56, 2017. 1

[42] Tejas Panambur and Mario Parente. Improved deep clustering of mastcam images using metric learning. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 2859–2862. IEEE, 2021. 3, 4

[43] Mario Parente and Tejas Panambur. Classification of martian terrains via deep clustering of mastcam images. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 1054–1057. IEEE, 2020. 3

[44] Dicong Qiu, Brandon Rothrock, Tanvir Islam, Annie K. Didier, Vivian Z. Sun, Chris A. Mattmann, and Masahiro Ono. Scoti: Science captioning of terrain images for data prioritization and local image search. *Planetary and Space Science*, 188:104943, 2020. 1, 2, 3

[45] 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*, page 5539. 2016. 1, 2, 3

[46] 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. 3, 4

[47] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015. 4

[48] SP Schwenzer, M Woods, S Karachalios, N Phan, and L Joudrier. Labelmars: Creating an extremely large martian image dataset through machine learning. In *50th Annual Lunar and Planetary Science Conference*, number 2132, page 1970, 2019. 2, 3

[49] R Michael Swan, Deegan Atha, Henry A Leopold, Matthew Gildner, Stephanie Oij, Cindy Chiu, and Masahiro Ono. Ai4mars: A dataset for terrain-aware autonomous driving on mars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2021. 1, 2, 3

[50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[51] Kiri Wagstaff, Steven Lu, Emily Dunkel, Kevin Grimes, Brandon Zhao, Jesse Cai, Shoshanna B Cole, Gary Doran, Raymond Francis, Jake Lee, et al. Mars image content classification: Three years of nasa deployment and recent advances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15204–15213, 2021. 1, 2, 3

[52] Kiri Wagstaff, You Lu, Alice Stanboli, Kevin Grimes, Thamme Gowda, and Jordan Padams. Deep mars: Cnn classification of mars imagery for the pds imaging atlas. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 2, 3

[53] Kiri L Wagstaff, Nina L Lanza, David R Thompson, Thomas G Dietterich, and Martha S Gilmore. Guiding scientific discovery with explanations using demud. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013. 2, 3

[54] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 3

[55] Cong Wang, Zian Zhang, Yongqiang Zhang, Rui Tian, and Mingli Ding. Gmsri: A texture-based martian surface rock image dataset. *Sensors*, 21(16), 2021. 3

[56] Wenjing Wang, Lilang Lin, Zejia Fan, and Jiaying Liu. Semi-supervised learning for mars imagery classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 499–503. IEEE, 2021. 1

[57] Thorsten Wilhelm, Melina Geis, Jens Püttschneider, Timo Sievernich, Tobias Weber, Kay Wohlfarth, and Christian Wöhler. Domars16k: A diverse dataset for weakly supervised geomorphologic analysis on mars. *Remote Sensing*, 12(23), 2020. 2, 3

[58] Kenneth H Williford, Kenneth A Farley, Kathryn M Stack, Abigail C Allwood, David Beaty, Luther W Beegle, Rohit Bhartia, Adrian J Brown, Manuel de la Torre Juarez, Svein-Erik Hamran, et al. The nasa mars 2020 rover mission and the search for extraterrestrial life. In *From habitability to life on Mars*, pages 275–308. Elsevier, 2018. 1

[59] Yong Xu, Feng Li, Zhile Chen, Jinxiu Liang, and Yuhui Quan. Encoding spatial distribution of convolutional features for texture representation. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[60] Jia Xue, Hang Zhang, and Kristin Dana. Deep texture manifold for ground terrain recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2018. 3, 4, 5

[61] Cheng-peng Yu and Xia Yuan. Terrain classification for autonomous navigation using ladar sensing. In *2009 First International Conference on Information Science and Engineering*, pages 1467–1470, 2009. 3

[62] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 3

[63] Wei Zhai, Yang Cao, Zheng-Jun Zha, HaiYong Xie, and Feng Wu. Deep structure-revealed network for texture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11010–11019, 2020. 3

[64] Wei Zhai, Yang Cao, Jing Zhang, and Zheng-Jun Zha. Deep multiple-attribute-perceived network for real-world texture recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3613–3622, 2019. 3

[65] Hang Zhang, Jia Xue, and Kristin J. Dana. Deep TEN: texture encoding network. *CoRR*, abs/1612.02844, 2016. 3, 4