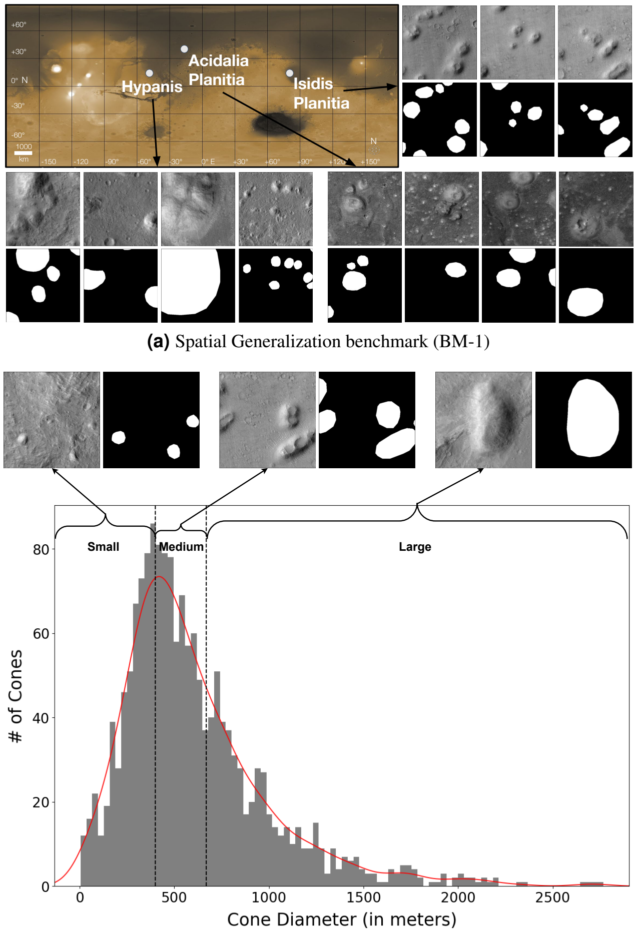
ConeQuest: A Benchmark for Cone Segmentation on Mars

Mirali Purohit Jacob Adler Hannah Kerner
Arizona State University
{mpurohi3, jbadler2, hkerner}@asu.edu

Abstract

Over the years, space scientists have collected terabytes of Mars data from satellites and rovers. One important set of features identified in Mars orbital images is pitted cones, which are interpreted to be mud volcanoes believed to form in regions that were once saturated in water (i.e., a lake or ocean). Identifying pitted cones globally on Mars would be of great importance, but expert geologists are unable to sort through the massive orbital image archives to identify all examples. However, this task is well suited for computer vision. Although several computer vision datasets exist for various Mars-related tasks, there is currently no open-source dataset available for cone detection/segmentation. Furthermore, previous studies trained models using data from a single region, which limits their applicability for global detection and mapping. Motivated by this, we introduce ConeQuest, the first expert-annotated public dataset to identify cones on Mars. ConeQuest consists of >13k samples from 3 different regions of Mars. We propose two benchmark tasks using ConeQuest: (i) Spatial Generalization and (ii) Cone-size Generalization. We finetune and evaluate widely-used segmentation models on both benchmark tasks. Results indicate that cone segmentation is a challenging open problem not solved by existing segmentation models, which achieve an average IoU of 52.52% and 42.55% on in-distribution data for tasks (i) and (ii), respectively. We believe this new benchmark dataset will facilitate the development of more accurate and robust models for cone segmentation. Data and code are available at https://github.com/kerner-lab/ConeQuest.



(a) Spatial Generalization benchmark (BM-1)

(b) Cone-size Generalization benchmark (BM-2)

Figure 1. Illustrative examples corresponding to both benchmarks

1. Introduction

With the advancement of camera technology and as data downlinking rates have improved, the entirety of the Martian surface has been imaged by multiple instruments that have acquired terabytes of data. Analyzing the data returned by spacecraft instruments is the only current way to gain insights into Mars surface processes as humans have yet to land on a planetary body other than the Moon. Finding evidence of past water is a top goal of the Mars science community [1, 3, 35]. Mapping the presence of water-related features improves scientists' understanding of the planet's past climate and its potential to have oceans or localized habitable environments, and helps identify key sites to send a future rover or human mission. One important set of features on the martian surface is pitted *cones* that are believed to be mud volcanoes [2, 6, 10, 17, 19, 21, 34, 36, 43]. These rounded mound-shaped features range in diameter from meter-sized to a few kilometers and are believed to form in regions that were once saturated in water (i.e., lakes or oceans). Three billion years ago, this water receded underground and turned to ice, or was buried by other deposits, only to emerge later as a mud volcano when tectonic compression, impact events, and other processes squeezed buried mud back to the surface along cracks and faults [43].

Manually reviewing the growing volumes of high-resolution remote sensing data of Mars to identify and characterize cones is prohibitively time-consuming and labor-intensive. Hence, an automated process is necessary to detect/segment and analyze the characteristics of cones. Machine learning methods provide a promising solution for developing an automation pipeline. However, training these models requires a substantial amount of data. Despite the availability of numerous Mars-related datasets, there is currently no open-source dataset for cone detection/segmentation.

Our study introduces a novel dataset called ConeQuest, which has been annotated by experts enabling ML methods to identify cones on Mars. This dataset comprises more than $13k$ samples from the Isidis Planitia, Acidalia Planitia, and Hypanis regions of the martian surface. Additionally, we provide metadata for every data sample, such as latitude-longitude, area, and bounding box. We formulated cone detection as a binary segmentation problem and developed two benchmarks (BMs) tasks based on ConeQuest: Spatial Generalization (BM-1) and Cone-size Generalization (BM-2). Figure 1 shows the overview of both BMs. The evaluation of spatial generalization examines the model's performance across different regions, in which test data is from a region not used for training. Similarly, cone-size generalization assesses the model's performance based on variations in cone size within the data, involving training on specific size ranges and combining different size ranges.

We conduct training on commonly used segmentation-based models which include U-Net [41], FPN [29], DeepLab [8], and MA-Net [16] for both BMs. We evaluate the model on in-distribution (id) as well as out-of-distribution (ood) data (i.e., data from region/size model has seen and unseen during training, respectively) to assess the efficiency of models. The average IoU for the id category is 52.52% and 42.55% for BM-1 and BM-2, respectively. Additionally, the average IoU on ood data is 15.04% (BM-1) and 26.92% (BM-2). The results obtained from the evaluation of ood data suggest that the model struggles to generalize on ood data and the evaluation of id data indicates that the model performs poorly in segmenting cones. These outcomes show that the cone segmentation task is not solved by existing segmentation models and there is a need for new solutions to segment cones accurately in future work. In summary, our contributions are as follows:

1. We introduce ConeQuest, the first expert-annotated publicly available dataset for cone segmentation across three different regions on Mars, along with metadata for each sample.
2. We designed two benchmarks based on ConeQuest: (i) Spatial Generalization and (ii) Cone-size Generalization, and assessed the effectiveness of various segmentation-based models in segmenting cones.
3. Evaluation of models indicates that existing models struggle to perform well on ConeQuest, highlighting the need for specialized models that can effectively capture the unique characteristics of cones.

2. Related Work

Deep learning has enabled researchers to develop models for a wide range of tasks in order to gain insights into data properties, improve labeling processes, and facilitate annotation. However, the success of these models heavily relies on the availability of large training datasets. The following sections provide an overview of the existing datasets for various Mars-related tasks created for training deep learning models (§2.1) and past research on cone detection on Mars (§2.2).

2.1. Mars Datasets

In Mars research, recognition of geological landforms and terrain classification are commonly explored tasks. Among these tasks, crater detection has been the most prominent. A few widely used datasets for Mars crater detection are [28, 31, 39, 40, 45, 46, 48]. In addition to crater detection, researchers have also generated datasets for other geological features, such as dunes, streaks, and ridges. For instance, [48] introduced a dataset encompassing 15 classes and $\sim 1k$ data samples per class across five distinct landform categories: aeolian bedforms, topographic landforms, slope feature landforms, impact landforms, and basic terrain landforms. Further, Wagstaff et al. have contributed landform datasets comprising $\sim 3k$ and $\sim 10k$ data samples from six classes [45, 46]. These datasets also include around $\sim 2.9k$ and $\sim 7k$ images representing over 20 classes of rover parts. Datasets are also available for terrain segmentation, encompassing classes such as rock, soil, sand, bedrock, and more. Two notable datasets in this domain are AI4MARS ($\sim 326k$ samples from 5 classes) [44] and S^5Mars ($\sim 5k$ samples from 9 classes) [49]. Another dataset focuses on martian frost, classifying images as either containing frost or representing background scenery [13]. Additionally, there are datasets tailored for change detection, comprising pairs of images that capture changes or the absence of changes over the same location at different times [24]. Furthermore, novelty detection and outlier detection datasets have been formulated to identify novel and anomalous samples within Mars datasets [23, 25]. Notably, there exists a dataset designed for classifying dusty versus non-dusty images [14].

Region	CTX Mosaic Folder (4°× 4°)	CTX Mosaic Tile ID (2°× 2°)	Area of Masked Region (N°× E°)	Resolution	Latitude	Longitude
Isidis Planitia	beta01_E084_N12	beta01_E084_N12.tif	Partial (0.5° × 0.5°)	5927 × 5927	13.5°	85.5°
Acidalia Planitia	beta01_E-016_N36	beta01_E-014_N38.tif	Partial (1° × 0.5°)	11855 × 5927	39.5°	-13°
Hypanis	beta01_E-044_N08	beta01_E-044_N10.tif	Full (2° × 2°)	23710 × 23710	10°	-44°
	beta01_E-044_N12	beta01_E-044_N12.tif	Full (2° × 2°)	23710 × 23710	12°	-44°
	beta01_E-048_N08	beta01_E-046_N10.tif	Full (2° × 2°)	23710 × 23710	10°	-46°
	beta01_E-048_N08	beta01_E-048_N10.tif	Full (2° × 2°)	23710 × 23710	10°	-48°
	beta01_E-048_N12	beta01_E-046_N12.tif	Full (2° × 2°)	23710 × 23710	12°	-46°
	beta01_E-048_N12	beta01_E-048_N12.tif	Full (2° × 2°)	23710 × 23710	12°	-48°

Table 1. Metadata of each CTX tile across three regions used in creation of ConeQuest

2.2. Dataset on Cone Detection

Despite the wide range of crater and other feature databases for Mars, there is a scarcity of cone detection/segmentation studies. Palafox et al. introduced MarsNet, a CNN-based classifier, for the identification of volcanic rootless cones and transverse aeolian ridges [37]. Pieterek et al. proposed a short study on pitted cones and crater detection by comparing a CNN with SVM [38]. Both of these studies lack detailed information about the annotation process, including whether the data were annotated by experts. Jiang et al. proposed a Single Shot MultiBox Detector model for cone and crater detection on Mars [22]. One limitation of their work is that they did not use experts for annotation but instead relied on an unspecified Wikipedia definition of cones and example HiRISE images found online. Furthermore, their annotations were performed at the box level, rather than accurately masking the region of interest, which does not fully align with the expectations of planetary scientists. One common weakness among all previous studies is that they train models using data from a single region, which may limit the ability of the model to generalize on another region and hinder global mapping and detection, as studies have shown that cones have unique characteristics specific to their region on Mars [43].

3. ConeQuest

This section provides information about the data source, annotation process, and overview of ConeQuest in detail.

3.1. Source Imagery

The Mars Reconnaissance Orbiter (MRO) Context Camera (CTX) acquires high-resolution images of the martian surface and has been operational since 2006 [4]. To build ConeQuest, we used open-source CTX data from the Murray Lab [7]. The dataset is a seam-corrected global image mosaic of Mars rendered at $\sim$ 5.0 meters/pixel [11, 33]. Data covers the entirety of the martian surface (> 99.5%). The global image data is divided into 3960 tiles (4° × 4°) from 88°S to 88°N [11, 12]. Each tile is subdivided into 4 subtiles (2° × 2°). This is the highest resolution complete-coverage global image data for Mars and is freely accessible at [7].

3.2. Data Annotation

There have been at least six fields of pitted cones on Mars identified so far (where a leading hypotheses is formation by mud volcanism), with each field containing hundreds to tens of thousands of cones [43]. However, a labeled dataset of cone annotations did not exist before our study. To create the labeled dataset, a planetary geologist annotated eight CTX subtiles spanning three different regions of Mars with known pitted cone fields (Table 1). Using the CTX mosaic subtiles as a basemap, a shapefile was created with polygons outlining the shape of each cone. The shapefile was then converted to a bitmask of the same resolution and dimensions as the CTX basemap (where pixels inside a cone shape were 1, and those not on a cone were set to 0). The labeled dataset thus contained eight CTX subtile images and eight corresponding bitmasks. The total number of annotated cones was 163 in Acidalia Planitia, 325 in Isidis Planitia, and 1691 in the Hypanis region of Southern Chryse Planitia. For Isidis Planitia and Acidalia Planitia there was a high density of pitted cones in the subtile, so only part of the CTX subtile was mapped ('Partial' indicated in Table 1), and the CTX image and bitmask were cropped to the mapped extent. The latitude and longitude in Table 1 correspond to the bottom-left corner of each annotated tile.

A planetary geologist with expertise in the morphology of cones in the dataset regions (co-author Adler) created all annotations of Isidis Planitia (IP) and Acidalia Planitia (AP) for this work. The Hypanis annotation shapes were sourced from a previously published peer-reviewed journal article [2] and thus have gone through quality control by other planetary mapping experts. While IP and AP regions have not been peer-reviewed, we deem such review unnecessary because 1) the same cone fields have been previously published in figures [10, 17, 19, 34, 36] (but were not digitized at the individual cone level), 2) the geologic setting within a uniform background unit makes identification obvious to experts, and 3) annotation was performed by a Mars mud volcano expert trained in geologic mapping. While we cannot be 100% certain about the annotations without ground-truth confirmation, we think it is unlikely there are false positives or false negatives at the object-level. In the Hypanis region, erroneous labels may be more likely because this region is more varied and geologically complex with compound features (many that erode into a rounded shape) that could be misinterpreted [2]. We estimate that no more than 5-20% of the annotations may be erroneous.

3.3. Data Preparation

The original data provided by Murray Lab are large CTX subtiles of size $23,710 \times 23,710$ pixels ($\sim$ 300 MB). To prepare these images for deep learning models, it is necessary to create smaller patches of the data, that can be compatible with the DL-based models and the system on which the model will be trained. We generated input samples by dividing each subtile into chunks measuring 512×512 pixels (which covers $\sim 2.5km^2$ area). The patches are generated in a column-wise manner. All the generated patches are distinct, ensuring that there is no overlap. This approach ensures there is no data leakage between training and test partitions. The resulting ConeQuest dataset has a total of 13,686 patches from 8 different subtiles across 3 regions.

3.4. Data Overview

The ConeQuest dataset includes input image and target segmentation mask pairs as well as metadata for every sample pair. Figure 1 shows the example of input data and their corresponding ground truth mask from each region. Each input-output pair has a unique name prefixed by its CTX Mosaic Tile ID. Generated input-output pairs can be used to train and evaluate segmentation or object detection models.

Additionally, ConeQuest provides the metadata of every CTX Mosaic Tile used in its creation (as displayed in Table 1), and a set of attributes about each patch, which can be further used for model training and evaluation, and mapping cones on Mars. These attributes are crucial as they record the specifics of *every cone* found across all the patches. The description of each attribute is as follows:

- **Patch Id:** Unique name of each input-output pair (e.g., *E-044_N10_00516.tif*). For Isidis Planitia and Acidalia Planitia name follows *_P* which indicates a patch from a partially annotated tile (e.g., *E084_N12_00068_P.tif*).
- **Region:** Denotes the region name to which each patch belongs (e.g., Isidis Planitia).
- **CTX Mosaic Folder:** CTX Mosaic folder from where subtile of CTX Mosaic Tile is taken (e.g., *beta01_E-044_N08*).
- **CTX Mosaic Tile ID:** CTX Mosaic Tile ID which was annotated and patch created from (e.g., *beta01_E-044_N10.tif*).
- **Latitude-Longitude Bounding Box:** This attribute provides the latitude-longitude coordinates of the bounding box for each cone present in the patch in the below format:

  [Polygon ((left-top, right-top, left-bottom, right-bottom)), Polygon ((left-top, right-top, left-bottom, right-bottom)), ...]

  Here, each *Polygon* element represents the coordinates of the bounding box of a single cone, where each coordinate represents *(longitude, latitude)* pair.
- **Latitude-Longitude Perimeter:** This attribute provides the latitude-longitude vertices of the perimeter for each cone present in the patch in the below format:

  [Polygon ((v_0*,* v_1 v_2*,* v_3*, ...)), Polygon ((*v_0*,* v_1 v_2*,* v_3*, ...)), ...]*

  Here, each *Polygon* element represents the vertices of the perimeter of a single cone, where each vertex (v_i) represents *(longitude, latitude)* pair.
- **Bounding Box:** Lists of bounding box list (e.g., [[x_{min}, y_{min}, $width$, $height$], [x_{min}, y_{min}, $width$, $height$], ...], here, each element in the list is a bounding box of a single cone).
- **Perimeter:** Lists of polygon vertices list (e.g., [[x_1, y_1, x_2, y_2, ...], [x_1, y_1, x_2, y_2, ...], ...], here, each element in the list is a polygon of a single cone).
- **Area:** Area of every cone in the patch[1] (e.g., [A_1, A_2, ...], here, each A_i is a float value).
- **Average Cone Diameter:** This attribute shows the average cone diameter of all cones in the patch. To calculate this, we have used the following formula which assumes the cone has a round shape:

  $$D = \frac{1}{N} \sum_{i=1}^{N} 2 * \sqrt{\frac{A_i * 25}{\pi}} \tag{1}$$

  Where N is the number of cones in the patch; A_i is the area of cone i; and we multiplied the area by 25 to convert the area into $meter^2$ (as 1 pixel covers an area of 5× meters).
- **Number of Cones:** A total number of cones in the patch (N).

Latitude and Longitude for the bounding box and perimeter follow a standard format, which simplifies the process for users to plot individual cones or clusters of cones from any patch on Mars using any Geographic Information Systems (GIS) software. Bounding Box, Area, and Perimeter are given in a similar format as the COCO dataset [30].

[1] calculated using *contourarea* function from OpenCV

It is crucial to note that ConeQuest includes samples that do not contain any cones, and these are referred to as *negative samples* or *non-cone patches*. Samples that contain cone/s are referred as *positive samples* or *cone-patches*. It is essential to recognize the significance of negative samples as they help to identify characteristics in the data that do *not* represent cones. For non-cone patches, attribute values contain empty list values or 0 accordingly.

4. Benchmarks

As discussed in §3.4, the ConeQuesttask is binary segmentation where the goal is to segment the cones in the input data and to mask (segment) the particular region where the cone is present (as shown in Figure 1). Cone segmentation is a very difficult task as shadows, contrast, and lighting change due to the time of day and season the images were acquired which strongly affects whether cones stand out from the background terrain and have similar shadow angles and lengths. Also, cones present in each of the three regions exhibit variations in terms of size, shape, and other characteristics [43]. For example, the morphology of cones in the Hypanis region is highly variable (small-large, bright-dark, circular-elongated-clustered) [2]. These characteristics make global segmentation of cones challenging, similar to other remote sensing tasks with high intra-class variance (such as building damage detection [5]). Based on this, we defined two benchmark tasks using ConeQuest: (i) Spatial Generalization (BM-1) and (ii) Cone-size Generalization (BM-2).

4.1. Spatial Generalization

In this benchmark, our objective is to assess model performance across the regional variability of cones (Figure 1a). Table 2 shows the total number of patches and the cone-patches in each region. For experiments, we train the model on two configurations: (i) single-region: model is trained using data from each region individually, and (ii) multi-region: model is trained on a combination of 2 or all 3 regions. For both configurations, we evaluate the model on each region separately where we denote regions included in the training as *in-distribution*, and excluded from the training as *out-of-distribution*.

Region	# of patches created	# of cone-patches
Isidis Planitia (IP)	144	131
Acidalia Planitia (AP)	288	135
Hypanis (HP)	13,254	1,392
Total	13,686	1,658

Table 2. Number of patches and cone-patches across three regions

4.2. Cone-size Generalization

This task evaluates the capability of a model to segment cones of different sizes which can identify any model biases in terms of cone size, which would substantially impact the downstream use of the model. Figure 1b shows the histogram of cone size in terms of *Average Cone Diameter* of patches and shows the variations in cone size across different patches. The peak of the distribution is approximately 400 m. We split all cone-patches into 3 categories: i) small, ii) medium, and iii) large. Table 3 gives statistics of size range and number of samples in each category. We create almost equal splits to fairly compare a model's performance when trained on different size categories.

Category	Size range (Cone Diameter)	# of Patches
Small (S)	5m < D ≤ 400	537
Medium (M)	400m < D ≤ 670m	569
Large (L)	670m < D	550

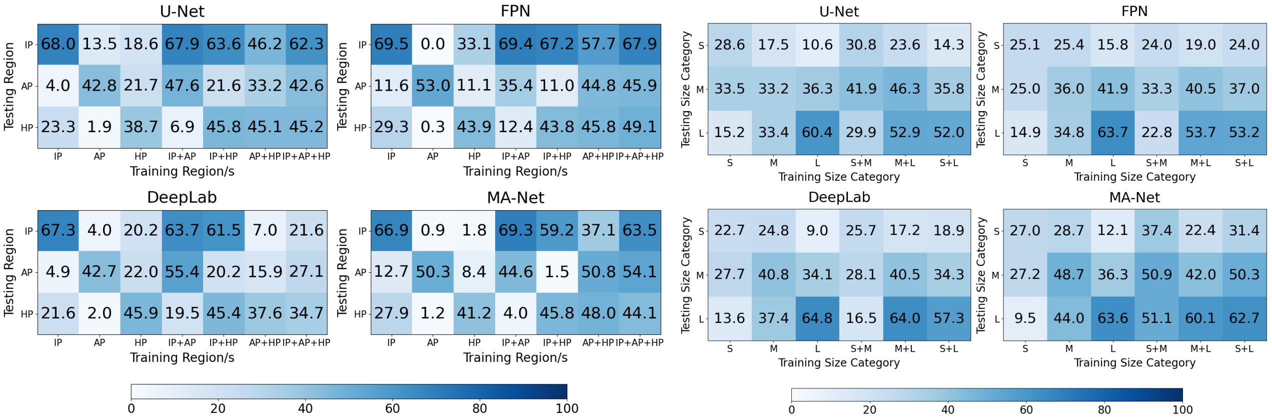
Table 3. Size range and number of patches across three categories

For experiments, we train the model on two configurations: (i) single-category (i.e., a model trained on each size category), and (ii) multi-category (i.e., a model trained on a combination of two size categories). For both configurations, we evaluate the model on each category separately where we denote the category included in the training as *in-distribution* (id), and excluded from the training as *out-of-distribution* (ood). id and ood data used for evaluation in both tasks will be denoted as $\mathcal{D}_{id}$ and $\mathcal{D}_{ood}$, respectively.

5. Experiments Setup

Data split: As stated in §3.4, ConeQuest has negative samples, and it is necessary to include those in the training. Table 2 shows that for BM-1 in Acidalia Planitia and Hypanis, there are equal and fewer negative samples compared to positive samples, respectively. Hence, for BM-1, we use balanced data of positive and negative samples to ensure that the model does not overfit the majority class except for Isidis Planitia (IP) region. Also, we have stratified positive and negative samples across dataset splits. In BM-2, we added negative samples to balance the data for each category. To evaluate the effectiveness of training with negative patches, we performed an ablation experiment in which models were trained only on positive samples.

For all experiments, we split the data into training, validation, and testing sets with a ratio of 7:1:2. To do a fair comparison across experiments, the same dataset splits for all regions and size categories are used for single and multi configurations training. These splits are also provided as



(a) Pixel IoU for BM-1

(b) Pixel IoU for BM-2

Figure 2. Results on $\mathcal{D}_{id}$ and $\mathcal{D}_{ood}$ for both BMs based on Pixel IoU (not a confusion matrix). IP: Isidis Planitia, AP: Acidalia Planitia, HP: Hypanis, S: Small, M: Medium, L: Large.

part of our public benchmark dataset. As mentioned in §3.3, non-overlapping samples were created with a stride of 512, which is equivalent to the patch height/width. This ensures that all samples in the data are distinct from each other.

Models: For training and evaluation, we selected commonly-used state-of-the-art segmentation-based models: U-Net [41], FPN [29], DeepLab [8], and MA-Net [16]. For the encoder of each model, we used ResNet-101 [18] as the backbone pre-trained on ImageNet [42].

Training Configuration: All models were trained for 200 epochs with a batch size of 8. Soft binary cross-entropy with logits was used as the loss function with the Adam optimizer [26]. We used early stopping to avoid overfitting and all models were evaluated on the model from the epoch with the lowest validation loss. All the experiments were conducted on Tesla V100-SXM2 with 16 GB GPU RAM.

Metrics: We report standard pixel-wise segmentation metrics and object-wise metrics for evaluation. We use pixel-Intersection over Union (IoU) [15] and mask IoU [30]. Pixel accuracy, pixel precision, and pixel recall are calculated by using all 4 quadrants of the confusion matrix as defined in [20]. Models are also evaluated on mean Average Precision (mAP) [30] and Panoptic Quality [27]. Since planetary scientists are also interested in instance-level metrics, we have shown evaluation based on object-wise metrics. For object IoU, we ran Hungarian matching [47] between all ground truth and predicted bounding boxes with the threshold of 0.5. To calculate object-wise accuracy, precision, and recall, the object is considered as True Positive (TP) if IoU is above 0.5.

As discussed in §3.4, ConeQuest includes non-cone patches and it is important to analyze the model's performance when it incorrectly tries to segment the object. Most previous research on segmentation does not consider this scenario. To incorporate this, we evaluated the model's performance by computing pixel-wise area segmented as a cone which is defined as follows:

$$A_{FP} = \frac{\text{FP} \quad (\text{\# pixels segmented as cone})}{\text{FP + TN} \quad (\text{total \# pixels in image})} * 100 \qquad (2)$$

In Equation 2, *lower* A_{FP} indicates better model performance, i.e., lower false positive area in negative patches.

6. Analysis and Discussion

In the following two sections, we describe quantitative and qualitative analysis of the results for both BMs.

6.1. Quantitative Analysis

Figure 2a and 2b show pixel IoU for all 4 models on $\mathcal{D}_{id}$ and $\mathcal{D}_{ood}$ for BM-1 and BM-2, respectively. For BM-1, the average IoU across all models on $\mathcal{D}_{id}$ is 52.52% for single-region and 49.03% for multi-region training. For BM-2, the average IoU across all models on single-category and multi-category is 42.88% and 41.08%, respectively. Table 4 and 5 report results on $\mathcal{D}_{id}$ for all evaluation metrics for BM-1 and BM-2, respectively. Results for each model on $\mathcal{D}_{ood}$ are reported in the Appendix. From Table 4 and 5, it can be observed that the maximum mAP is 33.16% for BM-1 and 22.54% for BM-2. These results indicate that cone segmentation is an open challenging problem, even on $\mathcal{D}_{id}$, for both BMs. From Figure 2, it can be observed that an average pixel IoU is 8.3% higher for BM-1 compared to BM-2. This indicates that current models generalize better across different regions than different size categories of cones. Comparing model-wise performance, MA-Net outperformed all 3 models for single and multi-group training for both benchmarks. Moreover, other metrics shown in Table 4 and 5 show similar observations.

Training Region	Testing Region	Cone											Non-Cone
		Mask IoU	Pixel IoU	Pixel Accuracy	Pixel Precision	Pixel Recall	Panoptic Quality	mAP	Object IoU	Object Accuracy	Object Precision	Object Recall	A_{FP}
IP	IP	64.81	67.92	96.66	83.54	79.09	54.87	33.15	74.29	60.71	71.77	75.51	0.00
AP	AP	45.13	45.83	96.21	82.97	52.77	38.91	19.44	49.30	44.45	52.25	50.87	0.23
HP	HP	41.39	42.43	92.09	82.71	49.99	31.41	14.39	43.41	36.97	43.72	44.31	0.50
IP + AP	IP	64.19	67.58	96.68	85.24	76.54	54.96	34.15	74.81	60.89	72.06	72.68	0.00
	AP	45.34	45.76	96.36	90.16	50.35	36.10	20.92	50.52	39.67	47.15	50.42	0.31
IP + HP	IP	61.25	62.82	96.24	83.32	74.58	50.14	28.69	71.68	54.70	69.81	63.05	0.00
	HP	42.82	43.95	92.28	79.47	53.16	32.20	14.26	45.14	37.34	44.29	45.63	0.66
AP + HP	AP	36.53	36.17	95.52	89.52	40.14	28.32	11.55	37.71	32.88	39.80	38.03	1.46
	HP	42.66	44.12	92.25	81.40	52.92	32.51	14.01	47.10	37.35	44.78	46.48	1.08
IP + AP + HP	IP	55.85	56.36	95.76	89.19	63.82	45.55	26.91	63.78	51.40	65.94	56.95	0.00
	AP	45.95	45.76	96.40	85.09	50.85	40.53	16.36	53.68	45.48	53.84	52.50	2.01
	HP	43.01	44.58	92.30	81.60	53.56	32.85	13.46	45.99	38.50	44.94	47.60	0.79

Table 4. Results for BM-1 for all metrics on $\mathcal{D}_{id}$. Here, the results in each row are the average across 4 models. See Appendix B (Table 1 and 2) for individual model results.

6.2. Analysis

Models fail to generalize on $\mathcal{D}_{ood}$: Figure 2a shows that all models do not generalize well to $\mathcal{D}_{ood}$ for single-region or multi-region training. For example, U-Net trained on HP achieves 38.7-pixel IoU on HP ($\mathcal{D}_{id}$), but drops substantially to 18.6 for IP and 21.7 for AP ($\mathcal{D}_{ood}$). This discrepancy in performance could be attributed to the variations in cone characteristics among the three regions. Models show the same generalization gap on $\mathcal{D}_{ood}$ for BM-2.

Segmenting all cones: Table 4 and 5 show that pixel precision is higher compared to pixel recall. This indicates that the models are better at reducing false positives than false negatives, which means all true cone pixels are not accurately captured. Object precision and object recall exhibit similar patterns across most cases. Moreover, in BM-1 for IP, the disparity between pixel precision and pixel recall is minimal, however, AP and HP show a larger discrepancy. Figure 3 shows similar observations for AP and HP[2]. Identical trend for BM-2, the small and medium categories have higher discrepancies between pixel precision and pixel recall compared to the large cone category.

Non-cone patches: To assess the significance of negative samples, we trained all models for both BMs only on positive samples. Results indicate that models perform poorly for non-cone patches when *trained solely on positive samples* ($\mathcal{T}_p$) compared to those *trained on both positive and negative samples* ($\mathcal{T}_{pn}$). Results show that $\mathcal{T}_p$ performs relatively lower compared to $\mathcal{T}_{pn}$ by 3.94% (for BM-1) and 2.14% (for BM-2) in terms of A_{FP} for non-cone patches. Additionally, performance for $\mathcal{T}_p$ does not show any significant improvement for positive samples, showing relative improvement of 0.09% (for BM-1) and 0.15% (for BM-2) on A_{FP}. Hence, we can conclude that negative samples help in training to improve the performance of non-cone patches while maintaining the performance of cone-patches. Detailed results of $\mathcal{T}_p$ are shown in Appendix C (Tables 5, 6, 7, and 8).

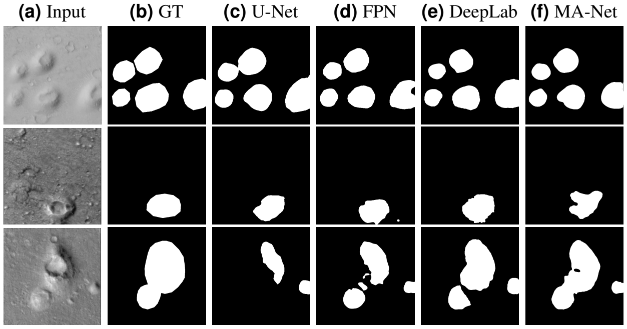


Figure 3. Illustration of predictions of in-distribution evaluation for single-region training (BM-1). Row 1, row 2, and row 3 represent test data samples from IP, AP, and HP, respectively. Columns c, d, e, and f show predictions from models with their corresponding Input (column a) and Ground Truth (GT) (column b).

6.2.1 Benchmark - 1

Multi-Region training: In natural language processing and general-domain computer vision, it has been established that models outperform multi-domain learning compared to single-task learning [9, 32]. However, comparing results on $\mathcal{D}_{id}$ for single-region and multi-region training (from Figure 2a) shows lower performance for multi-region. This suggests that tested models struggle to learn the multiple data distributions in multi-region scenarios, making global mapping a challenging task.

Performance differences across regions: From Figure 2a and Table 4, it can be observed that test metrics are higher for

[2] Detailed results of BM-1 and BM-2 are shown in Appendix A.

Training Size Category	Testing Size Category	Cone											Non-Cone
		Mask IoU	Pixel IoU	Pixel Accuracy	Pixel Precision	Pixel Recall	Panoptic Quality	mAP	Object IoU	Object Accuracy	Object Precision	Object Recall	A_{FP}
S	S	26.65	25.85	97.61	76.50	32.40	16.64	10.26	25.21	19.34	25.20	24.73	0.12
M	M	39.84	39.66	92.90	81.51	45.40	27.67	12.65	43.15	32.65	42.42	39.00	0.35
L	L	59.40	63.13	90.68	84.50	72.65	47.89	22.55	62.32	55.03	62.36	66.80	0.88
S + M	S	30.40	29.47	97.67	77.65	37.56	20.56	11.86	30.41	24.02	31.81	29.30	0.21
	M	40.31	38.54	92.84	85.79	44.01	29.72	15.57	43.12	35.64	44.69	42.09	0.35
M + L	M	42.78	42.31	92.69	79.04	49.82	29.38	13.90	45.12	34.50	46.72	39.94	0.51
	L	54.91	57.67	89.90	90.19	62.63	45.05	20.58	59.02	52.29	59.08	63.00	2.38
S + L	S	22.58	22.19	97.17	78.60	29.71	13.66	6.22	21.43	16.00	21.45	20.63	0.89
	L	52.08	56.31	89.48	87.50	63.24	41.37	17.83	54.19	47.37	52.86	59.5	2.12

Table 5. Results for BM-2 for all metrics on $\mathcal{D}_{id}$. Here, the results in each row are the average across 4 models. See Appendix B (Table 3 and 4) for individual model results.

the IP region compared to AP and HP. The average number of cones per patch for IP, AP, and HP are 3.52, 1.92, and 2.08, respectively. The higher cone density within IP compared to the other two regions, even though the number of training patches is smaller in IP, may explain the higher performance.

Figure 2a and Table 4 show performance is worst in HP for the single and multi-region training even on $\mathcal{D}_{id}$. Figure 1a shows data samples from all 3 regions. This may be due to the greater variability in cone appearance in HP compared to the other regions, which can be seen in the example images in Figure 1a. As discussed in §4 and §3.2, cones in HP have greater diversity in terms of size, shape (circular, elongated, and clustered cones), and brightness/appearance, and it is possible a small percentage of annotations have errors. This makes HP a challenging region for cone segmentation.

6.2.2 Benchmark - 2

Model behavior *vs.* cone size: From Figure 2b, we can observe that performance for all 4 models increases as cone size increases for single-category and multi-category training. This effect is likely a result of the larger size category providing the model with more information about the cone morphology compared to other smaller categories.

Multi-category Training: Similar to BM-1, in BM-2 multi-category performs worse than single-category training, except for the combination of small and medium categories. When training with small and medium combined, there is a significant improvement in object-wise metrics for small cones, while pixel-wise metrics show slight improvement. This indicates that the inclusion of the medium category has a positive impact on the performance of the small category. Although the model may not be able to precisely mask the cone at the pixel level, it is able to accurately localize the cone compared to training with only the small category.

Results on Non-cone patches: From Table 5, we can observe that A_{FP} increases for non-cone patches, as we go from a small to large category. This is due to the fact that when cone size in the patches increases across all training data, the model encounters more samples with larger cone areas. For instance, a model trained on the small category exhibits a lower false positive area (0.12) in non-cone patches compared to the one trained on the large category (0.88).

7. Conclusion

Despite the importance of cone segmentation in planetary science and the potential for computer vision techniques to facilitate this task, cone segmentation is under-explored and no publicly available dataset exists. In this research, we introduced ConeQuest, a benchmark for cone segmentation in Mars orbital images. We proposed two benchmark tasks based on ConeQuest: (i) Spatial Generalization and (ii) Cone-size Generalization. We evaluated four widely-used segmentation-based models for these tasks. Results show that for both benchmark tasks, existing models struggle in segmenting cones accurately for both in-distribution and out-of-distribution sub-groups. Furthermore, the evaluation of multi-region and multi-category training shows that models do not generalize for multi-domain learning. To enhance the model's performance, various techniques can be employed: (1) Employing a pixel-wise ensemble by combining the outputs of multiple models can benefit. (2) Histogram Matching can be used to improve the results of multi-region training as different lightning and brightness across regions can affect the model's performance. In the future, we plan to expand ConeQuest to include additional regions on Mars and evaluate model performance in these additional geologic settings. The dataset and metadata provided in ConeQuest enable researchers to develop customized models that may also incorporate context from metadata in model learning. We hope that ConeQuest will facilitate the development of models for cone segmentation and global mapping of cones and other important features on Mars and ultimately improve scientists' understanding of the Red Planet.

References

[1] *Vision and voyages for planetary science in the decade 2013-2022*. National Academies Press, 2012. i

[2] Jacob B Adler, James F Bell, Nicholas H Warner, Eldar Noe Dobrea, and Tanya N Harrison. Regional geology of the hypanis valles system, mars. *Journal of Geophysical Research: Planets*, 127(3):e2021JE006994, 2022. i, iii, iv, v

[3] Jeffrey L Bada, Andrew D Aubrey, Frank J Grunthaner, Michael Hecht, Richard Quinn, Richard Mathies, Aaron Zent, and John H Chalmers. Seeking signs of life on mars: In situ investigations as prerequisites to a sample return mission. *Planetary science decadal survey White Paper, Scripps Institution of Oceanograph, USA*, 2009. i

[4] JF Bell III, MC Malin, MA Caplinger, J Fahle, MJ Wolff, BA Cantor, PB James, T Ghaemi, LV Posiolova, MA Ravine, et al. Calibration and performance of the mars reconnaissance orbiter context camera (ctx). *International Journal of Mars Science and Exploration*, 8:1–14, 2013. iii

[5] Vitus Benson and Alexander Ecker. Assessing out-of-domain generalization for robust building damage detection. *arXiv preprint arXiv:2011.10328*, 2020. v

[6] Petr Brož, Ernst Hauber, Ilse Van de Burgt, V Špillar, and Gregory Michael. Subsurface sediment mobilization in the southern chryse planitia on mars. *Journal of Geophysical Research: Planets*, 124(3):703–720, 2019. i

[7] California Institute of Technology - Division of Geological and Planetary Sciences. The Bruce Murray Laboratory for Planetary Visualization. `http://murray-lab.caltech.edu/CTX/`. iii

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. ii, vi

[9] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020. vii

[10] PA Davis and KL Tanaka. Curvilinear ridges in isidis planitia, mars–the result of mud volcanism. In *Lunar and Planetary Science Conference*, volume 26, 1995. i, iii

[11] JL Dickson, BL Ehlmann, LH Kerber, and CI Fassett. Release of the global ctx mosaic of mars: An experiment in informationpreserving image data processing. In *54th Lunar and Planetary Science Conference*, pages 1–2, 2023. iii

[12] JL Dickson, LA Kerber, CI Fassett, and BL Ehlmann. A global, blended ctx mosaic of mars with vectorized seam mapping: A new mosaicking pipeline using principles of non-destructive image editing. In *Lunar and planetary science conference*, volume 49, pages 1–2. Lunar and Planetary Institute The Woodlands, TX, USA, 2018. iii

[13] Serina Diniega, Gary Doran, Steven Lu, Kiri L. Wagstaff, Jake Widmer, and Mark Wronkiewicz. Martian frost in hirise observations of northern mid-latitude craters (1.1.0) [data set]. zenodo. `https://doi.org/10.5281/zenodo.6561242`, 2022. ii

[14] Gary Doran. Hirise image patches obscured by atmospheric dust (1.0.0) [data set]. zenodo. `https://doi.org/10.5281/zenodo.3495068`, 2019. ii

[15] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. vi

[16] Tongle Fan, Guanglei Wang, Yan Li, and Hongrui Wang. Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access*, 8:179656–179665, 2020. ii, vi

[17] William H Farrand, Lisa R Gaddis, and Laszlo Keszthelyi. Pitted cones and domes on mars: Observations in acidalia planitia and cydonia mensae using moc, themis, and tes data. *Journal of Geophysical Research: Planets*, 110(E5), 2005. i, iii

[18] 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. vi

[19] Ryodo Hemmi and Hideaki Miyamoto. High-resolution topographic analyses of mounds in southern acidalia planitia, mars: Implications for possible mud volcanism in submarine and subaerial environments. *Geosciences*, 8(5):152, 2018. i, iii

[20] Pavel Iakubovskii. Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`, 2019. vi

[21] Mikhail A Ivanov, H Hiesinger, G Erkeling, and D Reiss. Mud volcanism and morphology of impact craters in utopia planitia on mars: Evidence for the ancient ocean. *Icarus*, 228:121–140, 2014. i

[22] Shancheng Jiang, Kai Leung Yung, WH WH Ip, Zongkai Lian, and Ming Gao. Automated detection of multi-type landforms on mars using a light-weight deep learning-based detector. *IEEE Transactions on Aerospace and Electronic Systems*, 2022. iii

[23] Hannah Kerner, Umaa Rebbapragada, Kiri Wagstaff, Steven Lu, Eric Huff, Bryce Dubayah, Vinay Raman, and Sakshum Kulshrestha. Domain-agnostic outlier ranking algorithms (dora): A configurable pipeline for outlier detection in scientific datasets. In *AGU Fall Meeting Abstracts*, volume 2021, pages IN11B–02, 2021. ii

[24] Hannah Rae Kerner, Kiri L Wagstaff, Brian D Bue, Patrick C Gray, James F Bell, and Heni Ben Amor. Toward generalized change detection on planetary surfaces with convolutional autoencoders and transfer learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(10):3900–3918, 2019. ii

[25] Hannah R Kerner, Kiri L Wagstaff, Brian D Bue, Danika F Wellington, Samantha Jacob, Paul Horton, James F Bell, Chiman Kwan, and Heni Ben Amor. Comparison of novelty detection methods for multispectral images in rover-based planetary exploration missions. *Data Mining and Knowledge Discovery*, 34:1642–1675, 2020. ii

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. vi

[27] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Pro-*

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9404–9413, 2019. vi

[28] Christopher Lee. Automated crater detection on mars using deep learning. *Planetary and Space Science*, 170:16–28, 2019. ii

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. ii, vi

[30] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. iv, vi

[31] Zhou Lincoln. Mars/lunar crater dataset [data set]. roboflow. https://universe.roboflow.com/lincoln-zhou/mars-lunar-crater, 2022. ii

[32] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019. vii

[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. iii

[34] Eileen M McGowan. The utopia/isidis overlap: Possible conduit for mud volcanism on mars. *Icarus*, 212(2):622–628, 2011. i, iii

[35] Engineering National Academies of Sciences, Medicine, et al. Origins, worlds, and life: A decadal strategy for planetary science and astrobiology 2023-2032. 2022. i

[36] Dorothy Z Oehler and Carlton C Allen. Evidence for pervasive mud volcanism in acidalia planitia, mars. *Icarus*, 208(2):636–657, 2010. i, iii

[37] 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. iii

[38] B Pieterek, M Grochowski, J Ciążela, O Sokolov, and M Józefowicz. Automated detection of pitted cones and impact craters: Deep-learning approach for searching potential hydrothermal activity and related ore deposits on mars. *LPI Contributions*, 2678:1328, 2022. iii

[39] Stuart J Robbins and Brian M Hynek. A new global database of mars impact craters $\geq$ 1 km: 1. database creation, properties, and parameters. *Journal of Geophysical Research: Planets*, 117(E5), 2012. ii

[40] Stuart J Robbins and Brian M Hynek. A new global database of mars impact craters $\geq$ 1 km: 2. global crater properties and regional variations of the simple-to-complex transition diameter. *Journal of Geophysical Research: Planets*, 117(E6), 2012. ii

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. ii, vi

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. vi

[43] James A Skinner Jr and Adriano Mazzini. Martian mud volcanism: Terrestrial analogs and implications for formational scenarios. *Marine and Petroleum Geology*, 26(9):1866–1878, 2009. i, ii, iii, v

[44] 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. ii

[45] 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. ii

[46] 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. ii

[47] Jianguo Wang, Peikun He, and Wei Coo. Study on the hungarian algorithm for the maximum likelihood data association problem. *Journal of Systems Engineering and Electronics*, 18(1):27–32, 2007. vi

[48] 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):3981, 2020. ii

[49] Jiahang Zhang, Lilang Lin, Zejia Fan, Wenjing Wang, and Jiaying Liu. s^5 mars: Self-supervised and semi-supervised learning for mars segmentation. *arXiv preprint arXiv:2207.01200*, 2022. ii