ReSyRIS - A Real-Synthetic Rock Instance Segmentation Dataset for Training and Benchmarking

Wout Boerdijk[1], Marcus G. Müller[1], Maximilian Durner[1], Rudolph Triebel[1]

[1]German Aerospace Center (DLR)
Institute of Robotics and Mechatronics
Münchner Str. 20, 82234 Weßling, Germany
Contact: Wout.Boerdijk@dlr.de

***Abstract*—The exploration of our solar system for understanding its creation and investigating potential chances of life on other celestial bodies is a fundamental drive of human mankind. After early telescope-based observation, Apollo 11 was the first space mission able to collect samples on the lunar surface and take them back to earth for analysis. Especially in recent years this trend accelerates again, and many successors were (or are in the process of being) launched into space for extra-terrestrial sample extraction. Yet, the abundance of potential failures makes these missions extremely challenging. For operations aimed at deeper parts of the solar system, the operational working distance extends even further, and communication delay and limited bandwidth increase complexity. Consequently, sample extraction missions are designed to be more autonomous in order to carry out large parts without human intervention. One specific sub-task particularly suitable for automation is the identification of relevant extraction candidates. While there exists several approaches for rock sample identification, there are often limiting factors in the form of applicable training data, lack of suitable annotations of the very same, and unclear performance of the algorithms in extra-terrestrial environments because of inadequate test data. To address these issues, we present *ReSyRIS* (Real-Synthetic Rock Instance Segmentation Dataset), which consists of real-world images together with their manually created synthetic counterpart. The real-world part is collected in a quasi-extra-terrestrial environment on Mt. Etna in Sicily, and focuses recordings of several rock sample sites. Every scene is re-created in OAISYS, a Blender-based data generation pipeline for unstructured outdoor environments, for which the required meshes and textures are extracted from the volcano site. This allows not only precise re-construction of the scenes in a synthetic environment, but also generation of highly realistic training data with automatic annotations in similar fashion to the real recordings. We finally investigate the generalization capability of a neural network trained on incrementally altered versions of synthetic data to explore potential sim-to-real gaps. The real-world dataset together with the OAISYS config files to create its synthetic counterpart are publicly available at `https://rm.dlr.de/resyris_en`. With this novel benchmark on extra-terrestrial stone instance segmentation we hope to further push the boundaries of autonomous rock sample extraction.**

TABLE OF CONTENTS

1. INTRODUCTION 1
2. RELATED WORK 2
3. RESYRIS 4
4. EXPERIMENTS 5
5. CONCLUSION 6
ACKNOWLEDGMENTS 7
REFERENCES 8
BIOGRAPHY 9

978-1-6654-9032-0/23/$31.00 ©2023 IEEE

Figure 1: **ReSyRIS consists of a set of real-world images from rocks in lunar-analogue environments (left), and a mimicked synthetic version (right) with corresponding instance annotations for both parts (colored overlay and green boxes). The created synthetic rock models and texture allow for infinitely many variations thereof.**

1. INTRODUCTION

The autonomous detection of rocks is an essential capability for future space exploration missions. The majority of obstacles on extra-terrestrial surfaces, like Mars or the Moon, are rocks. A collision between an autonomously exploring rover and a mis-detected rock can cause severe damage leading to mission failure. Besides safety reasons, rocks are also valuable for scientific analysis, since their shape and size contain significant geologic clues. However, the manual investigation and collection of statistics is extremely time intensive [1]. Furthermore, the collection of organic matter is an important goal of future space missions. Such complex tasks require capabilities which enhance the rover's autonomy, such as the detection of rock samples.

There readily exists a plethora of (often deep-learning based) algorithms for image-based identification of rocks. Yet, training data is usually extremely scarce and merely contains a few images. For instance, in [2] an adapted U-net is trained with solely 300 images of a Mars-like environment. Schenk *et al.* [3] manually label data of a muck pile of stones set-up in their laboratory. The works in [4] and [5] address this problem by using synthetic training data for the applied networks. While synthetic renderings represent a promising alternative, the environment and its textures need to be properly modeled to support sim-to-real transfer, thereby allowing for a better applicability of the model on real-world imagery. This results in a circularity problem, since the generalization capability of the algorithm can solely be verified by the already mentioned hardly available on-site test data. For the

©2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. DOI: 10.1109/AERO55745.2023.10115802

minority of large-scale data recorded on celestial bodies, thoroughly annotated images are generally difficult to obtain due to the laborious effort and often ambiguous interpretations. Furthermore, most of the existing datasets focus on pixel-wise semantic segmentation for rover traversability, which lacks separate annotations for each individual instance.

In this work, we propose to address these issues by presenting ReSyRIS, a real-world dataset for stone instance segmentation and its mimicked, synthetic counterpart (see Figure 1). The real-world part is collected on Mt. Etna in Sicily, Italy, which resembles the geological structure of celestial bodies. It focuses on several rock sample sites recorded with different sensors, where all suitable candidates are manually pixel-wise labeled in each frame. Every scene is re-created in OAISYS [4], a simulation environment specifically targeted at planetary robotics. To provide the required information, we generate high-quality textured meshes of the rocks by photogrammetric 3D reconstruction. Gravel texture is extracted from a collection of image patches recorded on the volcano. Given the possibility to generate large amounts of synthetic training data accompanied with the real-world pendant, ReSyRIS is an ideal dataset for benchmarking rock detection and instance segmentation algorithms.

Additionally, since OAISYS allows for arbitrary texture properties, sample placements and the like, ReSyRIS enables the generalization capabilities for incrementally altered training data given the assets at hand. In other words, we are able to investigate the sim-to-real gap in detail. To this end, we fine-tune a Mask-RCNN [6] on different sets of training data to explore the contributions for bridging the sim-to-real gap.

In summary, we propose the following contributions:

- A novel real-world dataset for rock sample segmentation recorded in a lunar-analogue environment with corresponding pixel-wise annotations.
- Highly realistic synthetic assets comprised of textured rock meshes and gravel texture.
- A synthetically mimicked version of the real-world dataset, and configuration files to create (almost) infinite variations thereof.
- A novel benchmark for extra-terrestrial rock instance segmentation including both synthetic and real data.
- An elaborative study on the generalization capabilities of neural networks for the task of rock instance segmentation.

This paper is organized as follows: In the next section, we give an overview on existing rock datasets within our scope of work. Then, we go in detail on the recording procedure of the real-world data on Mt. Etna, and explain the creation of the synthetic assets. Last but not least, we explore the generalization capability of a neural network on different variations of synthetic data trained on the test scenes.

2. Related Work

A major part of literature on extracting information of rocks from 2D images focuses on *classification* of stones and minerals on earth (e.g. [16], [17]) as well as on celestial bodies [18]. The *detection* of the respective substances is a necessary prerequisite, yet often limited by sufficient data. Importantly, together with successive instance segmentation, it is a substantial step for autonomous sample extraction. The following section attempts to give an overview on existing datasets from (quasi-) extra-terrestrial domains which are in scope of our work, and discuss their applicability for rock sample extraction. A tabulated overview of datasets particularly relevant to our approach is listed in Table 1.

For clarification, in the following we will refer to the task of *(semantic) segmentation* for classifying pixels into semantic categories (e.g. sky, rocks, sand). *Detection* refers to localizing specific instances in an image, usually by drawing a bounding box around them, and classifying them (e.g. *rock1*, *rock2*, ...), while *instance segmentation* additionally provides a binary mask for every detected item.

ADE20K [10] is initially proposed for Panoptic Segmentation (the combination of semantic and instance segmentation), and also includes approximately 1,150 images depicting various kinds of rocks. All images are recorded on earth, and many are insufficient for planetary rock segmentation since they depict stones in arbitrary environments (see Figure 2a).

While intended for localization and mapping in GNSS-denied environments, the *Katwijk Beach Planetary Rover Dataset* [7] additionally includes a collection of approximately 20,000 stereo images. For navigation purposes, artificial rock models in three different sizes (0.7 - 1.9 m diameter) were placed before recording; yet, their given poses are limited to 2D locations, making automatic labeling infeasible. In an attempt to produce large-scale annotations from this data, Kuang *et al.* [19] artificially extend a small set of images by pasting manually annotated rocks on background area. A mathematical model is established to finally determine the grayscale value of the added stone, which is comprised of a set of averaged pixel values of the target area, the approximate density of rays, and a correction constant. Different from our proposed approach, the algorithm is limited to generating morphological altered, grayscale versions from a set of labeled source images, and its incremental nature allows to solely *add* rock samples to the scene.

The *Artificial Lunar Landscape Dataset* [11] is a collection of renderings from a synthetic lunar environment created with the software Terragen. The dataset consists of semantic annotations for all 9,766 images into four classes (*sky*, *ground*, *small* and *large* rocks). Bounding boxes derived by blob detection are provided for large rocks above a certain pixel area. Aside the synthetic renderings, a small annotated test set of 36 real moon imagery is provided (Figure 2b).

Between 2012 and 2018, the *Curiosity* rover took a plethora of images on Mars, and a collection of its Mastcam recordings is formed by the *mars32k*[1] dataset (Figure 2c). To further artificially increase the amount of labeled data, a selection of *mars32k* is first manually classified into five categories in [20], and then employed to train a Style-GAN to create novel images. While the algorithm provides means to synthesize different shape and texture properties, the overall system behaves like a black-box where camera angle, light conditions and the like cannot be controlled arbitrarily, and pixel-wise instance labels cannot be derived. Xiao *et al.* [12] develop a kernel principal component - based rock detection framework, and in scope of their work annotate 405 images of *mars32k*. The annotations of the around 20,000 stones are available as semantic data (see Figure 2d).

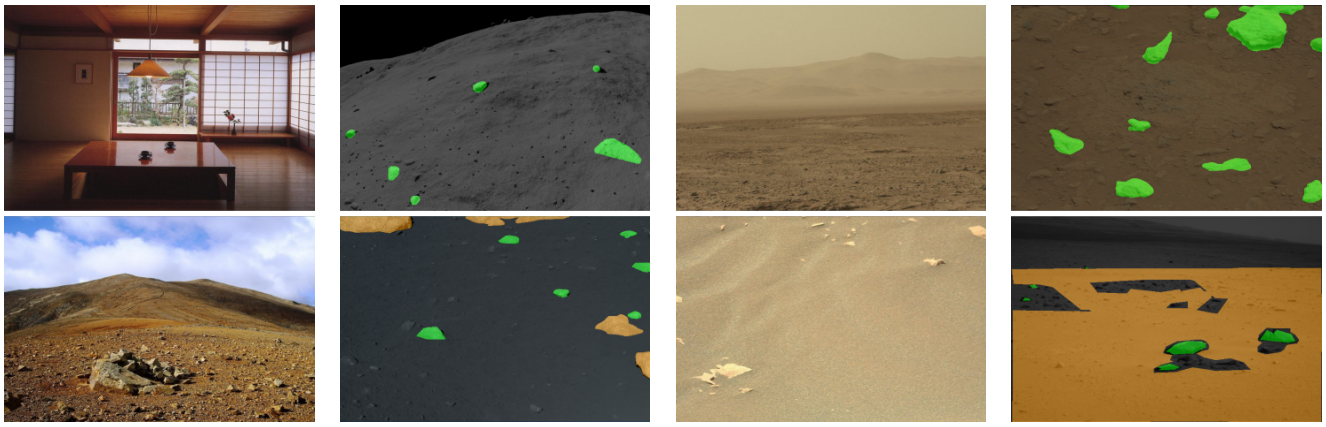
SPOC [13] presents an approach for slip prediction based on terrain classification of Marsian images for predicting traversability classes. The respective data is a collection

[1] https://dominikschmidt.xyz/mars32k/

Table 1: **Comparison of existing Rock Datasets for Segmentation Tasks.**

Dataset	Modality	No. Images	Real / Synthetic	Annotation
Katwijk Beach Planetary Rover Dataset [7]	Stereo-RGB	20,000	real	-
↳ [2]	RGB	331	real	instance-level (detection)
Devon Island Rover Navigation Dataset [8]	Stereo-RGB	49,410	real	-
↳ [9]	RGB	400	real	semantic
↳ [2]	RGB	1,600	real	instance-level (detection)
ADE20K [10] *(subset)*	RGB	∼1,150	real	instance-level
Lunar Landscape Dataset [11]	RGB	9,766	synthetic	semantic†
mars32k	RGB	32,368	real	-
↳ [12]	RGB	405	real	semantic
SPOC [13]	Grayscale	700	real	semantic
AI4Mars [14]	Grayscale	35,000	real	semantic‡
S^5Mars [15]	RGB	6,000	real	semantic‡
ReSyRIS	Stereo-RGB		both	instance-level

† Additional bounding box annotations are provided for large stones
‡ Only sparsely annotated



(a) ADE20K (b) Lunar Landscape Dataset (c) mars32k (d) mars32k labeled

Figure 2: **Exemplary images of relevant datasets; *green* overlay denote semantic masks for (small) rocks and *orange* color other semantic classes. ADE20K (a) depicts stones in arbitrary environments. (b) shows a rendering (top) and real image (bottom) of the lunar surface with corresponding semantic annotations. For mars32k (c), there exist various annotations (d), but regarding stones these are limited to semantics.**

of around 700 semantically labeled images from Curiosity's Navcam, including small and large rock classes.

Going beyond the scope of Curiosity's recordings, the publicly released *AI4Mars* [14] forms a collection of 35K images from Curiosity, Opportunity and Spirit rovers. Being particularly targeted for traversability of Mars rovers, the dataset is sparsely semantically annotated into sand, soil, bedrock and big rocks regions. To this end, every image is independently labeled by several people through crowd-sourcing, and the resulting masks are merged by accordance criteria. As a downside, this process results in further unlabeled regions due to inconsistencies across the annotators. [21] make use of *AI4Mars* in a semi-supervised setting for terrain segmentation. Specifically, a semantic segmentation network is first pre-trained in self-supervised fashion using classical contrastive loss settings. A subsequent supervised fine-tuning on 161 images (∼1% of the data) leads to improved performance than plain supervised learning.

S^5Mars [15] deals with a similar issue for semantic labeling of 6,000 Curiosity' Mastcam images, and proposes to solely annotate regions with high human confidence. In a first step, a segmentation network is pre-trained on restoring masked images in RGB color space. Since objects in images from extra-terrestrial domains usually have similar colors and indistinguishable contours, reconstruction is additionally constrained by a texture-aware operator. Finally, by using task uncertainty, confident pseudo-labels can be used to incorporate unlabeled image regions into the training.

The *Devon Island Rover Navigation Dataset* [8] is recorded at a Moon/Mars analogue site on Devon Island located in the Canadian High Arctic, and intended for localization in GPS-denied environments. The various data logged during the 10 km rover traverse includes almost 50K color stereo image pairs, captured approximately every 20cm. A selection of 400 images is hand-labeled in [9] for semantic segmentation of rocks with the goal of aiding navigation with obstacle

detection. In a subsequent work [2], a larger selection (1,600 images) is labeled for rock detection, consisting of more than 8,000 candidates. A SSD (Single-shot Multibox Detector) is then trained on the data at hand, and authors also showcase generalization on the *Katwijk Beach Planetary Rover Dataset*.

In summary, two important finds can be drawn from current literature: First, data from environments relevant for execution is often scarce. Annotations thereof are difficult to obtain due to unclear object boundaries, labeling ambiguity and the like. Second, while all of the aforementioned datasets depict rocks in (analogous) planetary environments, their applicability for autonomous rock sample extraction is impractical, if not impossible: Most of them concentrate on terrain segmentation for traversability prediction and obstacle avoidance, providing annotations for semantic use-cases. While some works provide pixel-wise instance-level annotations of rocks, information on the rock size in order to determine its suitability for grasping is not available.

To fill this gap, we present a real-world dataset recorded in a lunar-analogous environment depicting stones in arbitrary configurations, and corresponding instance annotations. Furthermore, we meticulously re-create the natural scene in OAISYS while carefully mimicking geometric and texture properties. The required textured rock meshes as well as the gravel texture allow for the creation of infinitely many, arbitrarily varied versions thereof and can well serve as training data.

3. ReSyRIS

In the following section, we will introduce ReSyRIS, explain the recording setup of the real-world data as well as the annotation format, and in detail explain the creation of its synthetic counterpart for rendering in OAISYS.

Real-world Data

Real-world imagery is recorded in the valley between *Cratere del Laghetto* and *La Cisternazza* on Mt. Etna in Sicily, Italy. A collection of 36 stones is iteratively placed in small sets (6-9 per image) in *close* and *spread out* configurations on the floor. For each set around 20 images are recorded while moving a camera stack circularly around the scene of interest. The stack consists of a *rc_visard 65* and two *Manta G-201* monochrome cameras with 8mm F1.8 lenses, and is depicted in Figure 3. Data for intrinsic and extrinsic calibration of the cameras is recorded on-site. Finally, every left color image of the rc_visard is manually annotated to yield pixel-wise masks for every rock. An additional selection of recorded images can be found in Figure 4.

Synthetic Data Creation

As readily explained we aim for two sets of synthetic data: (1) a mimicked test set with equivalent synthetic meshes and texture, and (2) assets to create synthetic training data. Both require precise textured mesh models of the rock samples from Mt. Etna, and corresponding gravel texture for the scene. In the following, the proposed pipeline to obtain the required assets for OAISYS is described. The workflow is depicted in Figure 5.

3D Stone Reconstruction—For every rock a synthetic counterpart is generated by photogrammetric 3D reconstruction with

Figure 3: **Camera stack used to record the real-world data, consisting of a *rc_visard 65* (bottom) and two Manta cameras (top).**

the publicly available software *Meshroom*[2]. A Sony ILCE-7M4 with a 35mm F1.8 lens is used to capture images of the object slowly rotating on a white table in a well-lit environment. The camera height is adjusted several times during recording to capture various viewing angles. In total, around 200 images per stone are aggregated. The collection of images is first processed in *Darktable*[3], where exposure and white balance are corrected according to a gray card placed aside the rock sample, and auto-corrections such as filmic rgb and highlight reconstruction are disabled. Every image is then cropped to the region of interest, and masked such that all background pixels are set to a uniform (white) value. We observe that this further facilitates the matching process in Meshroom, since feature extraction should primarily focus on the rock. Meshroom's photogrammetry pipeline is finally used to generate a high-quality textured mesh. Exemplary, this process is visualized in Figure 6, and all reconstructed rocks are depicted in Figure 7.

Texturing—Aside the comparably low resolution, the varying lighting conditions affecting the auto exposure and white balancing of the cameras used for recording on Mt. Etna prohibit the creation of a high-quality texture. Instead, the previously mentioned Sony camera is employed to record an approximately 25m^2 gravel area in small patches from above ($\sim$50cm distance), resulting in a collection of more than 150 images of volcano gravel. To almost nullify sunlight influence, a shading plane is held above the captured area. Given enough overlap between the images, we can re-use the same 3D reconstruction pipeline as explained above, with one minor difference: Prior to reconstruction, a second set of images is exported in Darktable, with particular emphasis on contrast and sharpness in order to boost fine detail and aid the reconstruction. The first image set is solely used at the very end for texturing. Note that the effect of a separate texture set was not necessary for rock reconstruction since cameras were matched successfully with the default texture images.

For a realistic texture it is desirable to not only have RGB information available, but also capture the geometric structure of the surface. To this end, we also extract a displacement and normal map from the mesh by baking in Blender. Precisely, we iteratively decimate the mesh up to retrieving an almost flat surface, followed by a re-calculation of UV-maps. Then, the mesh is subdivided again alongside a *shrinkwrap*

[2] https://github.com/alicevision/Meshroom.
[3] https://github.com/darktable-org/darktable.

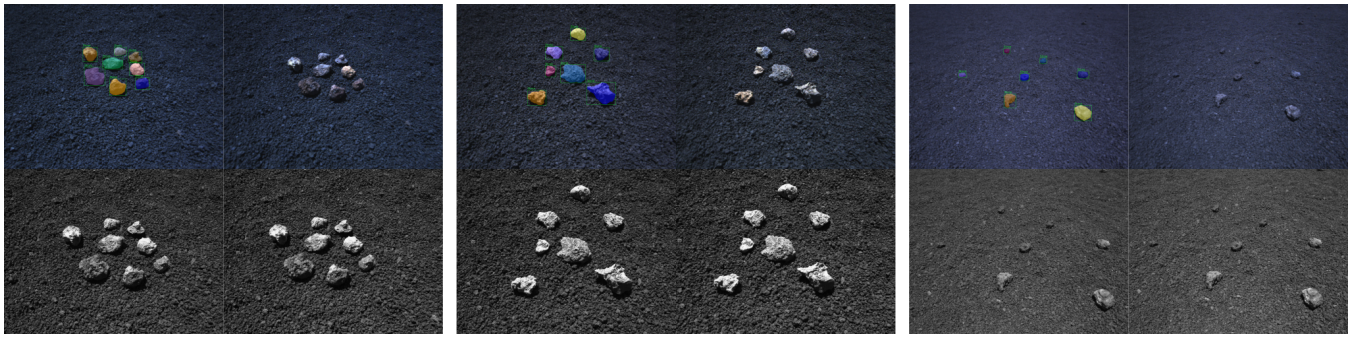


Figure 4: **Two *close* and one *spread out* examples of the recorded test data (best viewed magnified and in color). Every set of tiles depicts the rc_visard stereo images in the top, and the grayscale images of the two Mantas in the bottom. Every left rc_visard colour image has been annotated manually. Bounding boxes have been added to increase visibility.**

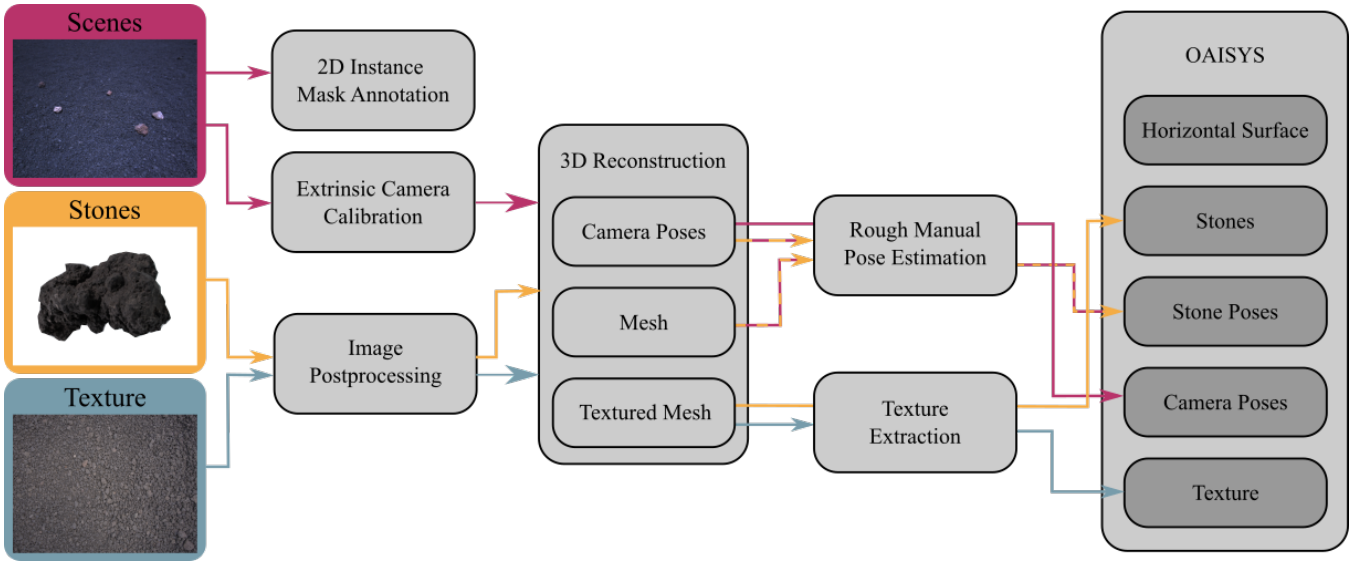


Figure 5: **Our suggested pipeline to create a synthetic, mimicked version of the recorded test scenes. The five main input types to OAISYS (*surface*, *stones*, *stone poses*, *camera poses* and *texture*) can be varied arbitrarily to generate synthetic training data.**

modifier, which constrains the surface to the original, undecimated mesh. This displacement from a (relatively) low-face-count mesh to the high-polynomial mesh achieved by the constrained subdivision can then be captured into a displacement, normal, and color map. While one advantage is the almost identical visual appearance, the rendered images can now be used as material in OAISYS to create realistic looking terrains on a large scale.

Synthetic Test Set Reconstruction—One aim of this work is the creation of a synthetic equivalent of the real-world recordings on Mt. Etna. The processing of the rock meshes as well as the gravel has readily been explained in the previous sections, and the camera poses can be obtained from the photogrammetric pipeline. While of insufficient quality for rendering synthetic data, the derived mesh of the test scene from the reconstruction pipeline is utilized to loosely estimate a rock's pose. Afterwards, the scenes can be rendered in OAISYS to obtain the synthetic test counterpart. A comparison between the two domains is shown in Figure 1 and in Figure 8.

4. Experiments

Last but not least, we would like to explore the possibilities of synthetic data creation with the generated real-world assets and OAISYS. Specifically, we are interested in the performance of neural networks for the task of stone instance segmentation trained on synthetic data with varying quality, and would like to investigate (1) *How important are geometrically representative, textured stone meshes?*, and (2) *How important is the realism of the texture in the scene?* To elaborate on these questions, we generate several small sets of training data for fine-tuning (each around 1000 samples), where we vary between (a) our 3D reconstructed rocks and more simpler, less realistic stone meshes; (b) a default texture from OAISYS and our volcano texture; and (c) between a bright and dark light setting. For all generated data, camera poses are sampled in similar height and angle as in the real-world imagery. Then, a Mask-RCNN is trained on the data at hand, following default configurations as presented in [22]. Results on the validation set (90/10 split), the synthetic test set and its real-world counterpart are listed in Table 2.

Overall, the network performs considerably well on the synthetic validation data, reaching average precision values

Table 2: **Quantitative results on diverse variations of synthetic training data. Results in *bold* denote the best performer; *underlined* values the second-best. All values are in [%].**

Synth. Training Setting			Synth. Val. Set			Synth. Test Set			Real Test Set		
Rocks	Texture	Light	AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
✗	✗	1	**81.04**	98.93	95.25	48.84	76.59	58.04	28.44	58.54	25.58
✗	✗	2	80.07	98.83	93.33	50.12	80.4	58.74	15.91	37.61	10.02
✓	✗	1	68.78	95.48	79.19	66.73	96.71	82.53	44.77	77.97	48.62
✓	✗	2	72.18	96.54	85.63	**67.75**	**96.79**	**84.0**	**50.54**	78.37	**59.7**
✗	✓	1	80.87	98.79	**96.27**	45.65	77.22	50.65	31.35	65.55	25.74
✗	✓	2	78.67	**99.86**	94.08	46.78	80.3	52.96	36.07	70.73	34.46
✓	✓	1	72.81	97.94	86.62	58.69	87.61	71.89	40.64	72.29	42.52
✓	✓	2	64.78	95.72	73.7	61.9	94.26	74.78	47.5	**78.67**	51.12

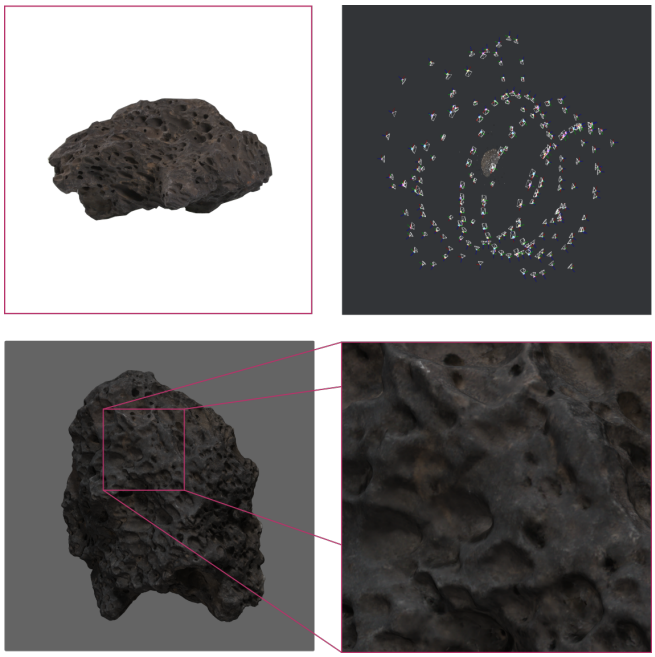


Figure 6: **Process of reconstructing a high-quality 3D textured mesh. Top left: exemplary cropped and masked stone image; top right: reconstructed cameras in Meshroom; bottom: fully textured mesh rendering in Blender with close-up.**

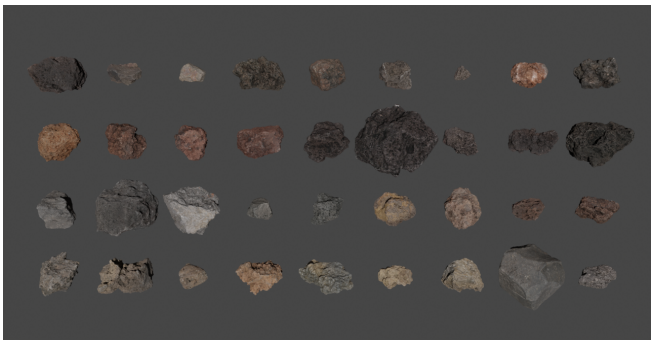


Figure 7: **A rendering of all 36 reconstructed stones scaled randomly (best viewed magnified and in color).**

of above 80%. While the plainest version (first two rows, referred to as *baseline* in the following) performs comparatively best on the validation set, the generalization onto the test set variants is seemingly difficult. By replacing the more basic rocks with our scanned assets (third and fourth row), the average precision on the test sets increases by 18 and 16 percent points, respectively, when compared to the baseline. The darker renderings (fourth row) further increase performance - a plausible trend that can be observed for most variants, since the test set features darker light conditions. Albeit not as strong, similar can be observed for replacing the scene texture with the real-world counterpart from the volcano (fifth and sixth row). From these findings, a logical conclusion would be that replacing both rocks and scene texture with our synthetic scans (last two rows). Interestingly, this is not completely evident given the results: While the models perform superior to the baseline and replacing scene texture only (row five and six), it merely comes close to or performs en par with the version with rock-replacement only (fourth row). Undeniably, there are an abundance of other crucial aspects influencing a model's performance on data of a different domain, and we believe that our experiments support this claim. Qualitatively, similar findings can be seen in Figure 8: For relatively easy scenes (left column), most models perform well. As rock texture blends in with the floor, the baseline model fails to identify most instances (middle column). For difficult settings (right column), only the best performing setting seems to reliably perform the task at hand.

5. Conclusion

In this work, we present ReSyRIS, a collection of real-world scans from lunar-analogue environments with a synthetic counterpart for the task of rock instance segmentation. The laboriously created high-quality textured meshes and terrain texture are particularly suitable for synthetic data generation, and will be made publicly available. In an additional experimental evaluation we investigated performance on different data altered in various dimensions, where we could confirm the contributions of our real-world scanned assets for aiding the generalization capabilities of neural networks on the task of extra-terrestrial stone segmentation. Given the undeniable existence of other important contributions for reducing the sim-to-real gap beyond our exploration, we think that further investigation in this direction is an interesting area of future work. Altogether, we believe that the introduced rock meshes alongside the volcano texture provide a sufficient starting point for pushing the boundaries of autonomous robotic systems in extra-terrestrial environments.

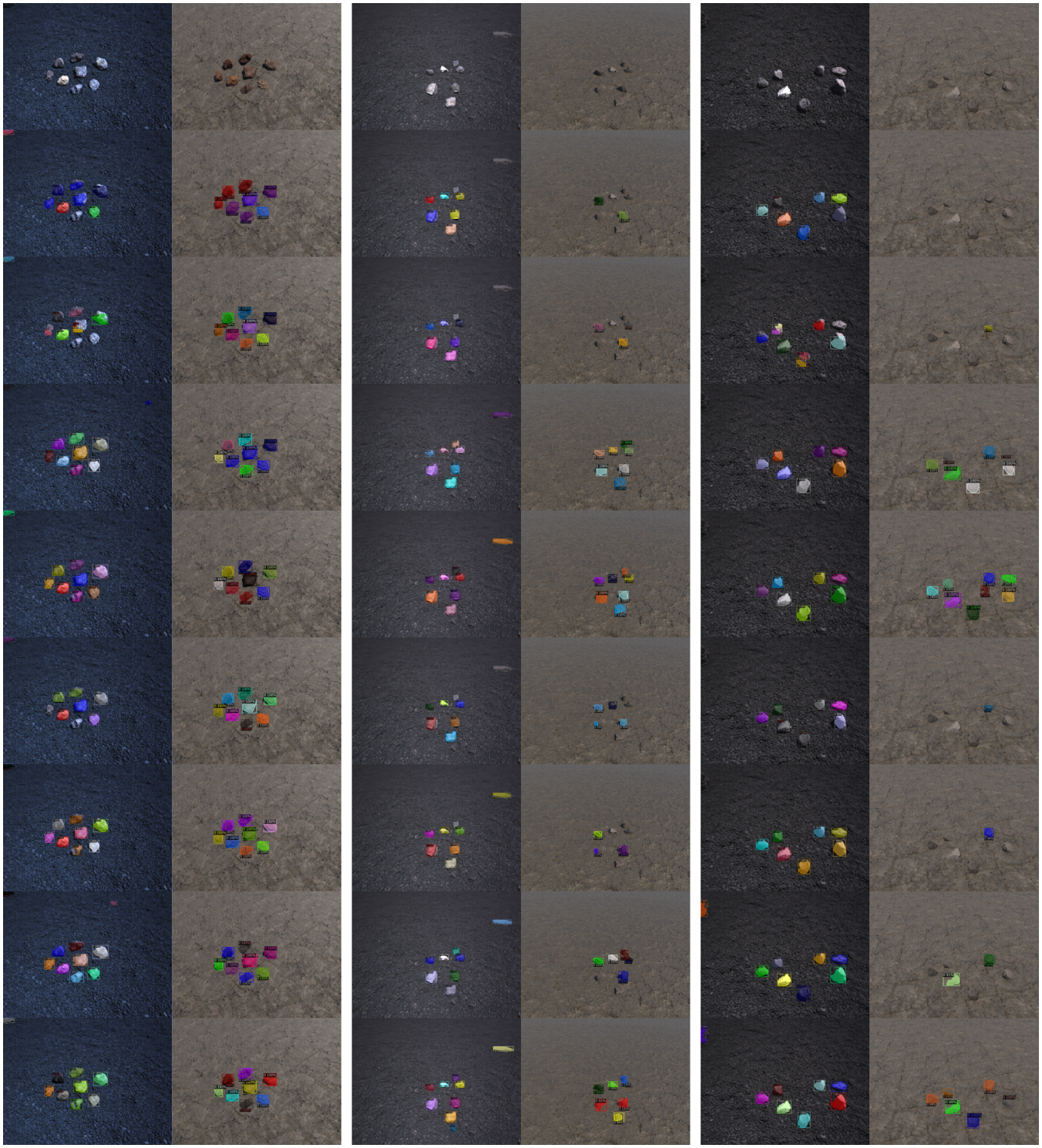


Figure 8: **Qualitative results on the two test sets (left columns: real-world data; right columns: synthetic counterpart). After the plain rgb images, the rows from top to bottom are predictions in the same order as in Table 2.**

ACKNOWLEDGMENTS

We would like to thank the Helmholtz association for understanding and supporting the postponement of the project twice in a row. Furthermore, we would also like to thank the local entities in Italy for understanding our proponent and supporting the missions with logistics, accommodations and licences. The "Parco di Etna", "INGV Catania", "Funevia di Etna", "Giro Calabria", "Glasser Sondertransporte" and the "Baia Verde Hotel" are given credits as well. This work was supported by the Helmholtz Association, project alliance ROBEX (contract no. HA-304), project ARCHES (contract no. ZT-0033), and project iFOODis (contract no. KA-HSC-06_iFOODis).

REFERENCES

[1] D. Thompson and R. Castaño, "Performance comparison of rock detection algorithms for autonomous planetary geology," in *IEEE Aerospace Conference Proceedings*, Jan. 2007, pp. 1–9. [Online]. Available: https://ieeexplore.ieee.org/document/4161563

[2] F. Furlán, E. Rubio, H. Sossa, and V. Ponce, "CNN Based Detectors on Planetary Environments: A Performance Evaluation," *Frontiers in Neurorobotics*, vol. 14, p. 590371, Oct. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7661793/

[3] F. Schenk, A. Tscharf, G. Mayer, and F. Fraundorfer, "Automatic Muck Pile Characterization From UAV Images," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W5, pp. 163–170, May 2019. [Online]. Available: https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/IV-2-W5/163/2019/

[4] M. G. Müller, M. Durner, A. Gawel, W. Stürzl, R. Triebel, and R. Siegwart, "A Photorealistic Terrain Simulation Pipeline for Unstructured Outdoor Environments," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2021, pp. 9765–9772, iSSN: 2153-0866. [Online]. Available: https://ieeexplore.ieee.org/document/9636644

[5] W. Boerdijk, M. G. Müller, M. Durner, M. Sundermeyer, W. Friedl, A. Gawel, W. Stürzl, Z.-C. Marton, R. Siegwart, and R. Triebel, "Rock Instance Segmentation from Synthetic Images for Planetary Exploration Missions," in *Advances in Space Robotics and Back to Earth (IROS 2021 Workshop)*, October 2021. [Online]. Available: https://elib.dlr.de/144626/

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2017/papers/He_Mask_R-CNN_ICCV_2017_paper.pdf

[7] R. A. Hewitt, E. Boukas, M. Azkarate, M. Pagnamenta, J. A. Marshall, A. Gasteratos, and G. Visentin, "The Katwijk beach planetary rover dataset," *The International Journal of Robotics Research*, vol. 37, no. 1, pp. 3–12, Jan. 2018, publisher: SAGE Publications Ltd STM. [Online]. Available: https://doi.org/10.1177/0278364917737153

[8] P. Furgale, P. Carle, J. Enright, and T. D. Barfoot, "The Devon Island rover navigation dataset," *The International Journal of Robotics Research*, p. 7.

[9] F. Furlán, E. Rubio, H. Sossa, and V. Ponce, "Rock detection in a mars-like environment using a cnn," in *Pattern Recognition*, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, J. A. Olvera-López, and J. Salas, Eds. Cham: Springer International Publishing, 2019, pp. 149–158.

[10] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene Parsing through ADE20K Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 5122–5130. [Online]. Available: http://ieeexplore.ieee.org/document/8100027/

[11] R. Pessia and G. Ishigami. Artificial Lunar Landscape Dataset. [Online]. Available: https://www.kaggle.com/datasets/romainpessia/artificial-lunar-rocky-landscape-dataset

[12] X. Xiao, M. Yao, H. Liu, J. Wang, L. Zhang, and Y. Fu, "A Kernel-Based Multi-Featured Rock Modeling and Detection Framework for a Mars Rover," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2021, conference Name: IEEE Transactions on Neural Networks and Learning Systems. [Online]. Available: https://ieeexplore.ieee.org/document/9646256

[13] B. Rothrock, R. Kennedy, C. Cunningham, J. Papon, M. Heverly, and M. Ono, *SPOC: Deep Learning-based Terrain Classification for Mars Rover Missions*. Pasadena, CA, 2016. [Online]. Available: https://arc.aiaa.org/doi/abs/10.2514/6.2016-5539

[14] R. M. Swan, D. Atha, H. A. Leopold, M. Gildner, S. Oij, C. Chiu, and M. Ono, "AI4MARS: A Dataset for Terrain-Aware Autonomous Driving on Mars," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2021, pp. 1982–1991, iSSN: 2160-7516. [Online]. Available: https://ieeexplore.ieee.org/document/9523149

[15] J. Zhang, L. Lin, Z. Fan, W. Wang, and J. Liu, "S^{5}Mars: Self-Supervised and Semi-Supervised Learning for Mars Segmentation," Jul. 2022, arXiv:2207.01200 [cs]. [Online]. Available: http://arxiv.org/abs/2207.01200

[16] X. Ran, L. Xue, Y. Zhang, Z. Liu, X. Sang, and J. He, "Rock Classification from Field Image Patches Analyzed Using a Deep Convolutional Neural Network," *Mathematics*, vol. 7, no. 8, p. 755, Aug. 2019, number: 8 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2227-7390/7/8/755

[17] A. Klu. Minerals Identification Dataset. [Online]. Available: https://www.kaggle.com/datasets/asiedubrempong/minerals-identification-dataset

[18] S. Hossain, J. Uddin, and R. A. Nahin, "Rock classification dataset," 2021. [Online]. Available: https://www.kaggle.com/ds/1293628

[19] B. Kuang, M. Wisniewski, Z. A. Rana, and Y. Zhao, "Rock Segmentation in the Navigation Vision of the Planetary Rovers," *Mathematics*, vol. 9, no. 23, p. 3048, Jan. 2021, number: 23 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2227-7390/9/23/3048

[20] C. Wang, Z. Zhang, Y. Zhang, R. Tian, and M. Ding, "GMSRI: A Texture-Based Martian Surface Rock Image Dataset," *Sensors*, vol. 21, no. 16, p. 5410, Jan. 2021, number: 16 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/21/16/5410

[21] E. Goh, J. Chen, and B. Wilson, "Mars Terrain Segmentation with Less Labels," Feb. 2022, arXiv:2202.00791 [cs]. [Online]. Available: http://arxiv.org/abs/2202.00791

[22] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

BIOGRAPHY

Wout Boerdijk *is a PhD student at the Technical University of Munich and a research scientist at the German Aerospace Center, where he is part of the Perception and Cognition department in the Institute of Robotics and Mechatronics. His research interests include computer vision methods for learning of and interacting with objects.*

Marcus G. Müller *is a researcher in the department of "Perception and Cognition" at the German Aerospace Center (DLR) since 2016 and Ph.D. student at ETH Zurich. He is the leader of the MAV Exploration Team at the Institute of Robotics and Mechatronics (DLR-RM), where he is working on autonomous navigation algorithms for MAVs. Before joining DLR he conducted research at the Jet Propulsion Laboratory (JPL) of NASA in Pasadena, USA, where he worked on visual inertial navigation for MAVs and on radar signal processing. Marcus received his Master's and Bachelor's degree in Electrical Engineering from the University of Siegen, Germany.*

Maximilian Durner *received his B.Sc. and M.Sc. degree in Electrical Engineering from the Technical University of Munich in 2014 and 2016, partially studying at the Politecnico di Torino, Italy and the Universidad Nacional de Bogota, Colombia. Since then he is a researcher at the Institute of Robotics and Mechatronics, German Aerospace Center (DLR). He is the leader of the research group on semantic scene analysis, where he focuses on object-centric perception for mobile manipulation.*

Rudolph Triebel *received his PhD in 2007 from the University of Freiburg in Germany. The title of his PhD thesis is "Three-dimensional Perception for Mobile Robots". From 2007 to 2011, he was a postdoctoral researcher at ETH Zurich, where he worked on machine learning algorithms for robot perception within several EU-funded projects. Then, from 2011 to 2013 he worked in the Mobile Robotics Group at the University of Oxford, where he developed unsupervised and online learning techniques for detection and classification applications in mobile robotics and autonomous driving. Since 2013, Rudolph works as a lecturer at TU Munich, where he teaches master level courses in the area of Machine Learning for Computer Vision. In 2015, he was appointed as leader of the Department of Perception and Cognition at the Robotics Institute of DLR.*