Rock Instance Segmentation from Synthetic Images for Planetary Exploration Missions

W. Boerdijk*[1,2], M. G. Müller*[1,3], M. Durner*[1,2], M. Sundermeyer[1,2],
W. Friedl[1], A. Gawel[3], W. Stürzl[1], Z.-C. Márton[4], R. Siegwart[3], R. Triebel[1,2]

***Abstract*— As the complexity and operation distance of space missions rises, the demand of highly autonomous rovers increases as well. An aspect of autonomous rovers that has been specifically attracting much attention from the space community is semi-autonomous sampling from celestrial bodies. Detecting possible samples is important for their extraction, which is challenging due to the unstructured and unknown enviroment, and the lack of suitable datasets. This work addresses the task of sample collection in an unknown and unstructured environment by presenting a module for visual stone segmentation. Due to the limited training data for such scenarios, we apply a photo-realistic simulator to optimize an unknown instance segmentation network. We evaluate various manners of fine-tuning and show the positive effect of training on data highly related to the test data.**

I. INTRODUCTION

Collecting geological samples from other planetary bodies is becoming increasingly important for the space community. As a result, several missions are mainly based on returning different kind of samples, such as soil and rocks, back to Earth. Examples include the the Hyabusa2 mission [1], that returned a sample of the asteroid *162173 Ryugu* as well as the Mars Sample Return Mission [2]. Similarly, the ARCHES mission [3] combines the exploration and sample return with a heterogeneous robotic team.

In most cases, a fully remote controlled rock sample extraction is not feasible due to the communication delay and limited bandwidth on such missions. Therefore, it is necessary that the sample extracting robot is detecting samples autonomously and sending the list of potential rocks for extraction to the ground team. Then, a team of scientist can select the desired sample and send a high-level command to the robot to execute the extraction procedure.

However, the autonomous detection of rocks is challenging. First of all, every rock has a unique and highly unstructured shape, which is not known beforehand. This renders the method of detecting arbitrary rocks difficult since it has to be applied on a diverse dataset. Another problem results from the lack of suitable datasets. While there exist a few datasets from planetary [4] and analog planetary environments [5], [6], they do not provide the necessary annotations to train an instance segmentation approach. Therefore, it is difficult to directly use them for training such methods.

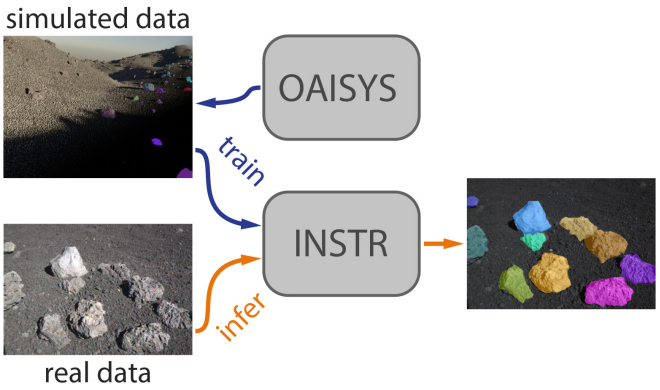

Fig. 1. Illustration of the presented pipeline. OAISYS is used to create a training dataset with which INSTR is trained to segment arbitrary rocks.

The problem of rock detection has already been addressed by several works. Di et al. [7] use a combination of classical techniques such as the mean-shift algorithm and plane fitting to detect small and large stones in 3D point cloud data. In [8] a gradient-region constrained level set method is presented. More recently, in [9] an adapted U-net is used to segment stones in a Mars-like environment. Schenk et al. [10] fine-tune a Mask R-CNN [11] with manually annotated data of a muck pile of stones set-up in the laboratory. Besides the tedious labeling effort, this might also introduce a bias due to the low variance in the recordings. To overcome the aforementioned challenges, we propose to generate photo realistic renderings of a planetary environment using OAISYS [12] and train our Instance Stereo Transformer (INSTR) [13], as illustrated in Fig. 1.

II. FINE-TUNE INSTR FOR PLANETARY USE-CASE

A. Unknown Instance Segmentation Method

Our goal is to segment rocks in an unknown environment for manipulation purposes. Therefore, as shown in [10], [12], an existing network (e.g. [11]) could be fine-tuned with one category: *rocks*. However, such network only relies on RGB cues, which might not be optimal in the underlying scenario, since stones and background have similar color and texture. Instead, more robust predictions can be obtained by incorporating additional modalities such as depth. Since high-quality depth data cannot always be ensured, we argue to use the INSTR network [13]. By taking a stereo image

*Equal Contribution
[1]Institute of Robotics and Mechatronics, German Aerospace Center (DLR), 82234 Wessling, Germany; <first>.<second>@dlr.de
[2]Department of Computer Science, Technical University of Munich (TUM), 85748 Garching, Germany
[3]Autonomous Systems Lab, Swiss Federal Institute of Technology (ETH Zürich), Switzerland
[4]Agile Robots AG, 81477 Munich, Germany

pair as input, the network implicitly fuses RGB and disparity information and avoids the necessity of depth data. Originally, the network is trained on synthetic data to segment any unknown instance on a dominant horizontal surface (e.g. tables) in an indoor environment. Since rocks on a planar surface state a similar problem, the pre-trained INSTR is able to partially segment instances. To further improve, we propose to specialize the network on the underlying use-case. Therefore, we evaluate the effect of fine-tuning with photo-realistic synthetic data of stone instances and compare it to the pre-trained version.

B. Generating Training Data

Since datasets for the described use-case are scarce (or not available), we are using OAISYS [12] to synthesize a dataset. It is a simulator which can auto generate photo-realistic outdoor environments and is specialized for planetary use-cases. One can provide textures for the underlying terrain and a set of objects, which are scattered on the surface. In order to create a useful dataset, the simulated environment is supposed to look similar to our target environment, Mt. Etna. Therefore, we use three gravel textures as terrains and 14 different kind of rocks as mesh assets. To distribute the rocks over the surface, we apply the particle system option of OAISYS. To create a realistic composition, we previously adjusted the color of all assets to be similar. In the simulator arbitrary number of sensors can be simulated. Here, a stereo set-up is configured. For each sensor, the activated render passes can be configured. By default the following rendering passes exist: color, depth, instance, and semantic segmentation. The created dataset consists of $\sim$1800 stereo images with the additional meta data. Fig. 2 illustrates example images of the training data.

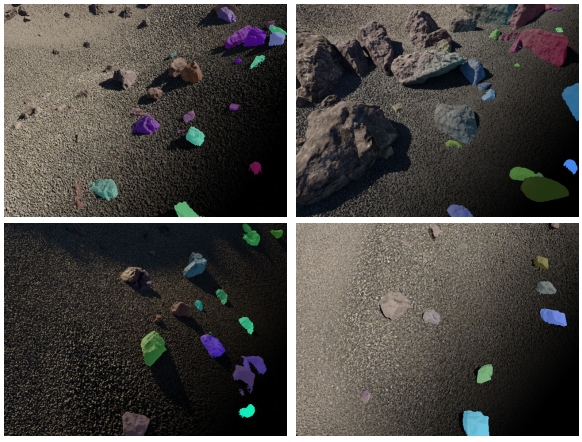

Fig. 2. Example images of the training dataset created by OAISYS. Color images partly overlayed by ground truth instance segmentation map.

III. Evaluation

The network consists of various modules which can all be fine-tuned individually (see Fig. 2 in [13] for details). Besides the fine-tuning of the whole network (a), we additionally evaluate fine-tuning of: (b) only the disparity and segmentation decoder; (c) transformer + (b); (d) only ResNet encoder. Both networks (pre-trained and fine-tuned) are evaluated on a real dataset recorded on a site on the volcano Mt. Etna in Sicily, Italy. The affinity of texture and color between the floor and the rocks as well as the variety of lightning conditions make this dataset quite challenging. The data was manually annotated and has a total of 26 test images. While (a) results in the best performance (Tab. I), (b) and (c) vastly degrade. We hypothesize that this is due to the distance change between the terrain dataset and the original indoor dataset, which makes an adaption of the encoder weights inevitable. This is confirmed by (d), where we freeze everything except the siamese ResNet encoder. Specifically, while the correlation computation itself does not have to be adjusted due to its inherent adaptability to varying camera intrinsics even in inference mode, weights of e.g. the downsampling layer after the second correlation layer - which directly operates on the correlation result and encodes distance information for future channels - have to be re-configured to provide plausible information to subsequent layers and the transformer. Qualitative results can be found in Fig. 3.

TABLE I

mIOU[%] of fine-tuning approaches on Mt. Etna test data

pre-trained	(a)	(b)	(c)	(d)
53.55	**63.88**	36.55	36.60	62.25

Fig. 3. Qualitative results of INSTR pre-trained (left) and fine-tuned with the synthetic data (right). After fine-tuning INSTR is well aware of objects in greater proximity to the camera.

IV. Conclusion

In this work, the specialisation of the INSTR network on the task of stone segmentation in a planetary scenario is presented. We analyse multiple fine-tuning manners on real test images. The experiments show the positive effect of fine-tuning with photo-realistic, synthetic images generated by OAISYS.

Acknowledgment

This work was supported by the Helmholtz Association, project ARCHES (`www.arches-projekt.de/en/`, contract number ZT-0033).

REFERENCES

[1] T. Morota, S. Sugita, Y. Cho, M. Kanamaru, E. Tatsumi, N. Sakatani, R. Honda, N. Hirata, H. Kikuchi, M. Yamada, Y. Yokota, S. Kameda, M. Matsuoka, H. Sawada, C. Honda, T. Kouyama, K. Ogawa, H. Suzuki, K. Yoshioka, M. Hayakawa, N. Hirata, M. Hirabayashi, H. Miyamoto, T. Michikami, T. Hiroi, R. Hemmi, O. S. Barnouin, C. M. Ernst, K. Kitazato, T. Nakamura, L. Riu, H. Senshu, H. Kobayashi, S. Sasaki, G. Komatsu, N. Tanabe, Y. Fujii, T. Irie, M. Suemitsu, N. Takaki, C. Sugimoto, K. Yumoto, M. Ishida, H. Kato, K. Moroi, D. Domingue, P. Michel, C. Pilorget, T. Iwata, M. Abe, M. Ohtake, Y. Nakauchi, K. Tsumura, H. Yabuta, Y. Ishihara, R. Noguchi, K. Matsumoto, A. Miura, N. Namiki, S. Tachibana, M. Arakawa, H. Ikeda, K. Wada, T. Mizuno, C. Hirose, S. Hosoda, O. Mori, T. Shimada, S. Soldini, R. Tsukizaki, H. Yano, M. Ozaki, H. Takeuchi, Y. Yamamoto, T. Okada, Y. Shimaki, K. Shirai, Y. Iijima, H. Noda, S. Kikuchi, T. Yamaguchi, N. Ogawa, G. Ono, Y. Mimasu, K. Yoshikawa, T. Takahashi, Y. Takei, A. Fujii, S. Nakazawa, F. Terui, S. Tanaka, M. Yoshikawa, T. Saiki, S. Watanabe, and Y. Tsuda, "Sample collection from asteroid (162173) ryugu by hayabusa2: Implications for surface evolution," *Science*, 2020.

[2] B. K. Muirhead, A. Nicholas, and J. Umland, "Mars sample return mission concept status," in *2020 IEEE Aerospace Conference*, 2020.

[3] M. J. Schuster, M. G. Müller, S. G. Brunner, H. Lehner, P. Lehner, R. Sakagami, A. Dömel, L. Meyer, B. Vodermayer, R. Giubilato, M. Vayugundla, J. Reill, F. Steidle, I. von Bargen, K. Bussmann, R. Belder, P. Lutz, W. Stürzl, M. Smek, M. Maier, S. Stoneman, A. F. Prince, B. Rebele, M. Durner, E. Staudinger, S. Zhang, R. Phlmann, E. Bischoff, C. Braun, S. Schröder, E. Dietz, S. Frohmann, A. Börner, H. Hübers, B. Foing, R. Triebel, A. O. Albu-Schäffer, and A. Wedler, "The arches space-analogue demonstration mission: Towards heterogeneous teams of autonomous robots for collaborative scientific sampling in planetary exploration," *IEEE Robotics and Automation Letters*, 2020.

[4] K. Wagstaff, Y. Lu, A. Stanboli, K. Grimes, T. Gowda, and J. Padams, "Deep mars: Cnn classification of mars imagery for the pds imaging atlas," *AAAI Conference on Artificial Intelligence*, 2018.

[5] L. Meyer, M. Smíšek, A. F. Villacampa, L. O. Maza, D. Medina, M. J. Schuster, F. Steidle, M. Vayugundla, M. G. Müller, B. Rebele, A. Wedler, and R. Triebel, "The MADMAX dataset for visual-inertial rover navigation on Mars," *Journal of Field Robotics*, 2021.

[6] M. Vayugundla, F. Steidle, M. Smisek, M. Schuster, K. Bussmann, and A. Wedler, "Datasets of long range navigation experiments in a moon analogue environment on mount etna," in *International Symposium on Robotics*, 2018.

[7] K. Di, Z. Yue, Z. Liu, and S. Wang, "Automated rock detection and shape analysis from mars rover imagery and 3d point cloud data," *Journal of Earth Science*, 2013.

[8] J. Yang and Z. Kang, "A gradient-region constrained level set method for autonomous rock detection from mars rover image," in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2019.

[9] F. Furlán, E. Rubio, H. Sossa, and V. Ponce, "Rock Detection in a Mars-Like Environment Using a CNN," in *Pattern Recognition*, 2019.

[10] F. Schenk, A. Tscharf, G. Mayer, and F. Fraundorfer, "Automatic much pile characterization from UAV images," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2019.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *International Conference on Computer Vision*, 2017.

[12] M. G. Müller, M. Durner, A. Gawel, W. Stürzl, R. Triebel, and R. Siegwart, "A Photorealistic Terrain Simulation Pipeline for Unstructured Outdoor Environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.

[13] M. Durner, W. Boerdijk, M. Sundermeyer, W. Friedl, Z.-C. Márton, and R. Triebel, "Unknown Object Segmentation from Stereo Images," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.