

# Vulcan Centaur: towards end-to-end real-time perception in lunar rovers

J. de Curtó and R. Duvall.

Iris Lunar Rover. Carnegie Mellon.

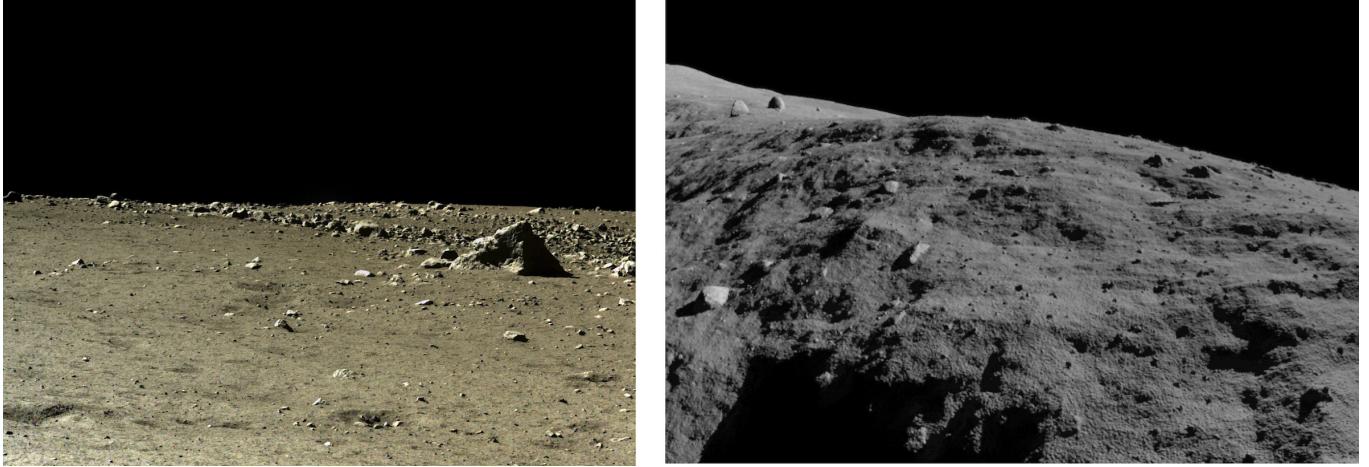


Figure 1. Left: Real image from the Moon. Right: Synthetic Moon.

We introduce a new real-time pipeline for Simultaneous Localization and Mapping (SLAM) and Visual Inertial Odometry (VIO) in the context of planetary rovers. We leverage prior information of the location of the lander to propose an object-level SLAM approach that optimizes pose and shape of the lander together with camera trajectories of the rover. As a further refinement step, we propose to use techniques of interpolation between adjacent temporal samples; videlicet synthesizing non-existing images to improve the overall accuracy of the system. The experiments are conducted in the context of the Iris Lunar Rover, a nano-rover that will be deployed in lunar terrain in 2021 as the flagship of Carnegie Mellon, being the first unmanned rover of America to be on the Moon.

CCS Concepts: • Robotics → Perception.

Additional Key Words and Phrases: Robotics, Lunar Rover, Perception, SLAM, VIO, Segmentation.

## 1 Introduction

Our aim is to present a novel pipeline to deploy state-of-the-art DL techniques in planetary rovers. With the advent of a new wave of planetary exploration missions, the need to call on generalizable perception and control systems that can operate autonomously in other worlds will become ubiquitous in the coming years.

## 2 Overall System

Following the design principles and the perception pipeline proposed in [Allan et al. 2019] in the context of the NASA Mission Resource Prospector, we put forward an improved technique for Visual Odometry (VIO) that could be exploited in a rover of the same characteristics. Although at the present time data from the

Moon is scarce, there are already some open datasets available in analogue environments such as the POLAR Stereo Dataset [Wong et al. 2017] that includes stereo pairs and LiDAR information or [Vayugundla et al. 2018], that contains IMU, stereo pairs and odometry plus some additional localization data, all obtained on Mount Etna. Specifically for the task of semantic segmentation, Kaggle provides images from a rendered environment of the Moon and masks. More recently, as a benchmark for tasks of Computer Vision in the context of space exploration, a dataset containing PNG images and positioning information from the mission Chang'E-4 to the Moon has been released [de Curtó and Duvall 2020], the data from CE4 consists on post-processed original files from the mission Chang'E<sup>1</sup>. Our specific sensor suite, that will be on-board the Iris Lunar Rover [de Curtó and Duvall 2020], a project led by Carnegie Mellon that will deploy a four pound rover into the surface of the Moon by 2021 and that will be the first unmanned rover of America to explore the surface of the Moon, consists on IMU, two high-fidelity cameras and odometry sensors. Furthermore, it also has a UWB module [Alarifi et al. 2016; Ledegerber et al. 2015; Mueller et al. 2015; Xu et al. 2020] on-board to localize the rover with respect to the lander.

## 3 SLAM/VIO

Simultaneous Localization and Mapping (SLAM) and Visual Inertial Odometry (VIO) are defined as a function that transform raw data from the sensors into a distribution over the states of the robot. SLAM and VIO [Schneider et al. 2018; Usenko et al. 2019] have been for decades unparalleled problems in robot perception and state

decurtoidiaz@ieee.org, rduvall@andrew.cmu.edu.

<sup>1</sup>moon.bao.ac.cn.

estimation. Although typical dense SLAM systems are not differentiable, new approaches to solve this problem propose gradient-based learning over computational graphs to go all the way from 3D maps to 2D pixels [Murthy et al. 2020].

The first task to tackle in geometric computer vision, being SLAM [Engel et al. 2017, 2014; Newcombe et al. 2011], Structure-from-Motion (SfM) [Agarwal et al. 2009; Bloesch et al. 2018; Graham and Novotny 2020; Snavely et al. 2006; Tateno et al. 2017; Z. Teed 2020], camera calibration or image matching, is to extract interest points [DeTone et al. 2018; Ono et al. 2018] from still images. We can define interest points as 2D specific locations in a given sample which can be considered stable and repeatable along different ambient conditions and viewpoints. The techniques used to traditionally attack this problem pertain to Multiple View Geometry [Hartley and Zisserman 2004], a subfield of mathematics that sets forth theorems and algorithms built on the assumption that those interest points can indeed be reliably extracted and matched across overlapping frames. Nonetheless, real-world computer vision operates on raw images that are far from the idealized conditions assumed in the proposed theory. Blending traditional modules with learning representations have lately been proven to be incredibly effective [DeTone et al. 2018; Tang and Tan 2019; Yang et al. 2020] as a way to bridge the gap between the conditions that we face in the real world and the assumptions made to design the algorithms. Plentiful of approaches also explore unsupervised learning of depth and egomotion [Godard et al. 2017; Yin and Shi 2018; Zhou et al. 2017].

State-of-the-art approaches also deal with related problems such as SLAM object-level, that is, a system capable of optimizing object poses and shapes together with camera trajectory [McCormac et al. 2018; Sucar et al. 2020; Sünderhauf et al. 2017]. Although a SLAM system capable of incrementally mapping multi-object scenes seems not related to our task, its importance is revealed when we understand the fact that in many occasions the rover will localize itself with respect to the lander, which location is known; therefore a SLAM solution capable of optimizing the pose and shape of the lander along camera trajectory of the rover, would be distinctly adequate. With respect to this, we have to bear in mind that the principal technique that the rover will be using on-board to localize itself will be the UWB module [Xu et al. 2020]; that will indeed use the lander as a way-station for data communication. The reason for this is that critical weight and power can be hugely saved using RF for communication and state estimation. Thus, SLAM and VIO computation will be done on-ground. Using the same philosophy, it seems natural also to rely on a technique that will jointly optimize pose and shape of the lander together with camera trajectories.

#### 4 Shape and Pose of the Lander

We assume here that we have a segmentation mask of the lander that in our specific case is obtained by the use of semantic segmentation [Arbeláez et al. 2011, 2014; Chen et al. 2018a, 2017, 2018b; Chen and Koltun 2017; Hariharan et al. 2015; He et al. 2017; Mostajabi et al. 2015; Yu and Koltun 2016; Zhou et al. 2017]. On some of these approaches, the segmentation process is guided by the use of

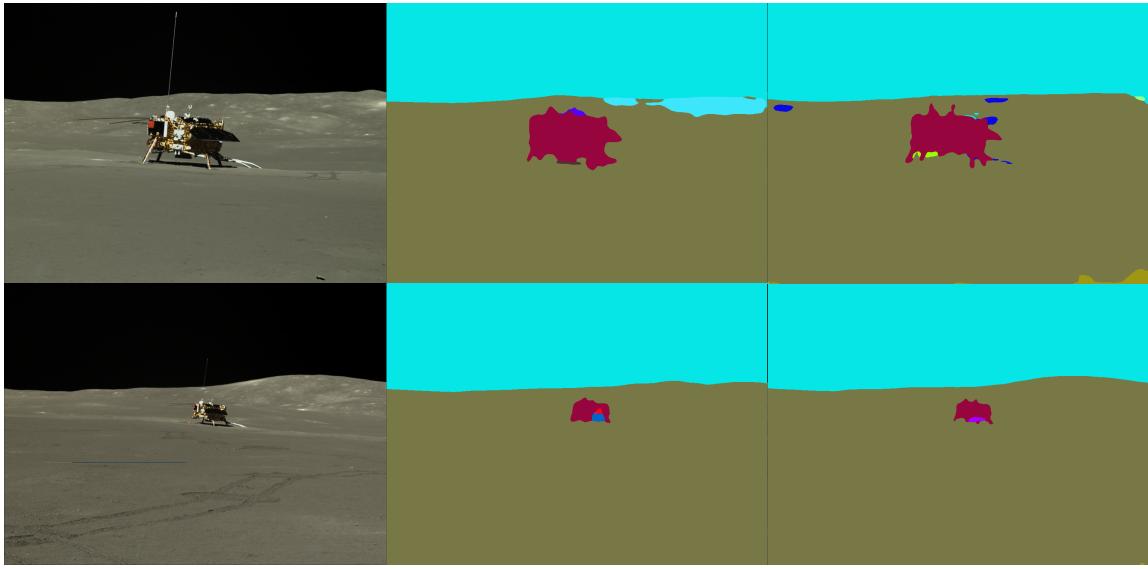
a prior object detector [Girshick 2015; Huang et al. 2017; Liu et al. 2016; Redmon et al. 2016; Redmon and Farhadi 2017; Ren et al. 2015]. Specifically, we finetune our model building on DilatedResnet-101 [Zhao et al. 2017; Zhou et al. 2018] and UperNet-101 [Lin et al. 2017; Xiao et al. 2018] trained on ADE20K [Zhou et al. 2017]. Some examples of the mask given by our segmenter can be observed in Figure 2. To infer the shape and pose we will leverage existing techniques [Sucar et al. 2020] that given a depth image, full shape and pose is determined. These techniques normally address multi-object categories; where a previous classification step and object observation is necessary, however our approach is somewhat simpler in the sense that the only object under consideration will be the lander per se.

#### 5 Temporal Interpolation between Subsequent Samples

In the absence of continuous data between adjacent temporal samples given by the camera and to mitigate the effects that this will incur in the algorithms used to localize the rover, we propose to adopt techniques from video frame interpolation. Although signaling breakthroughs have been achieved by the use of recent deep convolutional neural networks, the quality of the resulting samples is often dubious due to object motion or occlusions. The main aim here is to synthesize non-existent frames in-between original samples to improve accuracy in the proposed VIO/SLAM approaches. Specifically for this purpose, we build on a recent depth-aware flow projection layer that achieves compelling upshots to synthesize intermediate sequences [Bao et al. 2019].

#### References

- S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. 2009. Building Rome on a Day. *IEEE International Conference on Computer Vision* (2009).
- A. Alarifi, A. Al-Salmán, M. Alsaleh, A. Alnafessah, S. Al-Hadrami, M. A. Al-Ammar, and H. S. Al-Khalifa. 2016. Ultra Wideband Indoor Positioning Technologies: Analysis and Recent Advances. *Sensors* (2016).
- M. Allan, U. Wong, P. M. Furlong, A. Rogg, S. McMichael, T. Welsh, I. Chen, S. Peters, B. Gerkey, M. Quigley, M. Shirley, M. Deans, H. Cannon, and T. Fong. 2019. Planetary Rover Simulation for Lunar Exploration Missions. *IEEE Aerospace Conference* (2019).
- P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. 2011. Contour Detection and Hierarchical Image Segmentation. *IEEE Transactions (TPAMI)* 33 (2011).
- P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. 2014. Multiscale Combinatorial Grouping. *IEEE Conference on Computer Vision and Pattern Recognition* (2014).
- W. Bao, W. Lai, C. Ma, X. Zhang, Z. Gao, and M. Yang. 2019. Depth-aware Video Frame Interpolation. *IEEE Conference on Computer Vision and Pattern Recognition* (2019).
- M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. Davison. 2018. CODESLAM - Learning a Compact, Optimisable Representation for Dense Visual SLAM. *IEEE Conference on Computer Vision and Pattern Recognition* (2018).
- L. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam. 2018a. MaskLab: Instance Segmentation by Refining Object Detection with Semantic and Direction Features. *IEEE Conference on Computer Vision and Pattern Recognition* (2018).
- L. Chen, G. Papandreou, F. Schroff, and H. Adam. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv:1706.05587* (2017).
- L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018b. Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation. *EUROPEAN Conference on Computer Vision* (2018).
- Q. Chen and V. Koltun. 2017. Photographic Image Synthesis with Cascaded Refinement Networks. *IEEE International Conference on Computer Vision* (2017).
- J. de Curtó and R. Duvall. 2020. Cycle-consistent Generative Adversarial Networks for Neural Style Transfer Using Data from Chang'E-4. *arXiv:2011.11627* (2020).
- D. DeTone, T. Malisiewicz, and A. Rabinovich. 2018. SuperPoint: Self-supervised Interest Point Detection and Description. *IEEE Conference on Computer Vision and Pattern Recognition* (2018).
- J. Engel, V. Koltun, and D. Cremers. 2017. Direct Sparse Odometry. *T-PAMI* (2017).
- J. Engel, T. Schöps, and D. Cremers. 2014. LSD-SLAM: Large-scale Direct Monocular SLAM. *EUROPEAN Conference on Computer Vision* (2014).
- R. Girshick. 2015. FAST R-CNN. *IEEE International Conference on Computer Vision* (2015).



**Figure 2. Segmentation of the Lander.** Left: Image from CE4 [de Curtó and Duvall 2020]. In particular we are using color images from the panoramic camera of the rover of the mission to the Moon Chang'E-4. Middle: Generated mask given by model DilatedResnet-101 [Zhao et al. 2017; Zhou et al. 2018]. Right: Generated mask given by model UperNet-101 [Lin et al. 2017; Xiao et al. 2018].

- C. Godard, O. Mac Aodha, and G. Brostow. 2017. Unsupervised Monocular Depth Estimation with Left-right Consistency. *IEEE Conference on Computer Vision and Pattern Recognition* (2017).
- B. Graham and D. Novotny. 2020. RidgeSfM: Structure from Motion via Robust Pairwise Matching under Depth Uncertainty. *3DV* (2020).
- B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. 2015. Hypercolumns for Object Segmentation and Fine-grained Localization. *IEEE Conference on Computer Vision and Pattern Recognition* (2015).
- R. Hartley and A. Zisserman. 2004. *Multiple View Geometry in Computer Vision* (2nd edition ed.). CAMBRIDGE University Press.
- K. He, G. Gkioxari, P. Dollár, and R. Girshick. 2017. MASK R-CNN. *IEEE International Conference on Computer Vision* (2017).
- J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. 2017. Speed/Accuracy Trade-offs for Modern Convolutional Object Detectors. *IEEE Conference on Computer Vision and Pattern Recognition* (2017).
- A. Ledergerber, M. Hamer, and R. D'Andrea. 2015. A Robot Self-localization System Using One-way Ultra-wideband Communication. *IEEE International Conference on Intelligent Robots and Systems (IROS)* (2015).
- T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. 2017. Feature Pyramid Networks for Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition* (2017).
- W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg. 2016. SSD: Single Shot Multibox Detector. *EUROPEAN Conference on Computer Vision* (2016).
- J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger. 2018. Fusion++: Volumetric Object-level SLAM. *3DV* (2018).
- M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. 2015. Feedforward Semantic Segmentation with Zoom-out Features. *IEEE Conference on Computer Vision and Pattern Recognition* (2015).
- M. W. Mueller, M. Hamer, and R. D'Andrea. 2015. Fusing Ultra-wideband Range Measurements with Accelerometers and Rate Gyroscopes for Quadcopter State Estimation. *ICRA* (2015).
- K. Murthy, S. Saryazdi, G. Iyer, and L. Paull. 2020. gradslam: Dense SLAM Meets Automatic Differentiation. *ICRA* (2020).
- R. Newcombe, S. Lovegrove, and A. Davison. 2011. DTAM: Dense Tracking and Mapping in Real-time. *IEEE International Conference on Computer Vision* (2011).
- Y. Ono, E. Trulls, P. Fua, and K. M. Yi. 2018. LF-Net: Learning Local Features from Images. *NIPS* (2018).
- J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. 2016. You Only Look Once: Unified, Real-time Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition* (2016).
- J. Redmon and A. Farhadi. 2017. YOLO9000: Better, Faster, Stronger. *IEEE Conference on Computer Vision and Pattern Recognition* (2017).
- S. Ren, K. He, R. Girshick, and J. Sun. 2015. FASTER R-CNN: Towards Real-time Object Detection with Region Proposal Networks. *NIPS* 28 (2015).
- T. Schneider, M. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart. 2018. Maplab: an Open Framework for Research in Visual-inertial Mapping and Localization. *IEEE Robotics and Automation Letters* (2018).
- N. Snavely, S. M. Seitz, and R. Szeliski. 2006. Photo Tourism: Exploring Photo Collections in 3D. *ACM Transactions on Graphics (SIGGRAPH)* 25 (2006).
- E. Sucar, K. Wada, and A. Davison. 2020. NODESLAM: Neural Object Descriptors for Multi-view Shape Reconstruction. *3DV* (2020).
- N. Sünderhauf, T. T. Pham, Y. Latif, M. Milford, and I. Reid. 2017. Meaningful Maps with Object-oriented Semantic Mapping. *IEEE International Conference on Intelligent Robots and Systems (IROS)* (2017).
- C. Tang and P. Tan. 2019. BA-Net: Dense Bundle Adjustment Network. *ICLR* (2019).
- K. Tateno, F. Tomboli, I. Laina, and N. Navab. 2017. CNN-SLAM: Real-time Dense Monocular SLAM with Learned Depth Prediction. *IEEE Conference on Computer Vision and Pattern Recognition* (2017).
- V. Usenko, N. Demmel, D. Schubert, J. Stückler, and D. Cremers. 2019. Visual-inertial Mapping with Non-linear Factor Recovery. *IEEE Robotics and Automation Letters* (2019).
- M. Vayugundla, F. Steidle, M. Smisek, M. J. Schuster, K. Bussmann, and A. Wedler. 2018. Datasets of Long Range Navigation Experiments in a Moon Analogue Environment on Mount Etna. *International Symposium on Robotics* (2018).
- U. Wong, A. Nefian, L. Edwards, X. Buoyssounoue, P. M. Furlong, M. Deans, and T. Fong. 2017. POLAR (Polar Optical Lunar Analog Reconstruction) Stereo Dataset. *NASA Ames Research Center* (2017).
- T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. 2018. Unified Perceptual Parsing for Scene Understanding. *EUROPEAN Conference on Computer Vision* (2018).
- H. Xu, L. Wang, Y. Zhang, K. Qiu, and S. Shen. 2020. Decentralized Visual-Inertial-UWB Fusion for Relative State Estimation of Aerial Swarm. *ICRA* (2020).
- N. Yang, L. von Stumberg, R. Wang, and D. Cremers. 2020. D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry. *IEEE Conference on Computer Vision and Pattern Recognition* (2020).
- Z. Yin and J. Shi. 2018. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. *IEEE Conference on Computer Vision and Pattern Recognition* (2018).
- F. Yu and V. Koltun. 2016. Multi-scale Context Aggregation by Dilated Convolutions. *ICLR* (2016).
- J. Deng Z. Teed. 2020. DeepV2D: Video to Depth with Differentiable Structure from Motion. *ICLR* (2020).
- J. Zhao, M. Mathieu, and Y. LeCun. 2017. Energy-based Generative Adversarial Networks. *ICLR* (2017).
- B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. 2018. Semantic Understanding of Scenes through the ADE20K Dataset. *IJCV* (2018).

- T. Zhou, M. Brown, N. Snavely, and D. Lowe. 2017. Unsupervised Learning of Depth and Ego-motion from Video. *IEEE Conference on Computer Vision and Pattern Recognition* (2017).