Towards Robust Perception of Unknown Objects in the Wild

Wout Boerdijk[1], Maximilian Durner[1], Martin Sundermeyer[1] and Rudolph Triebel[1,2]

***Abstract*— To be able to interact in dynamic and cluttered environments, detection and instance segmentation of only known objects is often not sufficient. Our recently proposed *Instance Stereo Transformer* (INSTR) [1] addresses this problem by yielding pixel-wise instance masks of unknown items on dominant horizontal surfaces without requiring potentially noisy depth maps. To further boost the application of INSTR in a robotic domain, we propose two improvements: First, we extend the network to semantically label all non-object pixels, and experimentally validate that the additional explicit semantic information further enhances the object instance predictions. Second, knowledge about *some* detected objects might often readily be available, and we utilize Dropout as approximation of Bayesian inference to robustly classify the detected instances into known and unknown categories. The overall framework is well suited for various robotic applications, e.g. stone segmentation in planetary environments or in an unknown object grasping setting.**

I. INTRODUCTION

INSTR [1] is a transformer-based network that is able to predict instance-level binary masks for unknown objects on dominant horizontal surfaces. In contrast to existing methods relying on depth or RGB-D inputs (e.g. [2], [3]), INSTR processes a pair of stereo images and is guided to implicitly reason about geometric information by an auxiliary disparity loss. This circumvents operating on potentially incomplete and noisy depth maps, and is well suited to be employed in dynamic, real-world scenarios [4].

The network poses no assumptions on the environment regarding object shape, texture or the like, and solely requires a planar surface as indication of object presence. Yet, in many real-world robotic settings, further semantic information might be required for successful navigation and manipulation aside the detection of unknown instances. En plus, one might readily have knowledge about a particular subset of objects in the scene. And even if all instances are perfectly known and categorized, an unknown distraction object could be present - let it be someone's forgotten coffee mug. While INSTR is well suited to detect the cup on a plane, it treats *every* object instance as unknown. Addressing the separation into known and unknown objects is therefore an important consequential step, and also offers application of INSTR to anomaly detection, bin sorting or the like.

In this work we propose two modifications to INSTR that allow to reason about semantics in the scene, and to separate detected instances into known and unknown objects.

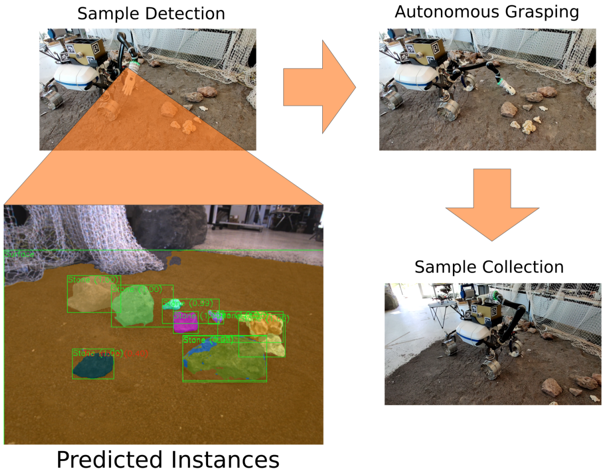

Fig. 1: Application of INSTR for autonomous rock segmentation and sample extraction on a *Lightweight Rover Unit* (LRU) (best viewed magnified and in color).

Regarding the former, jointly predicting semantic and instance level information is usually referred to as *Panoptic Segmentation* (PS) [5]. Concretely, the task is to assign a semantic label to each pixel as well as to detect and segment each individual instance. While early work often focuses on proposal-based segmentation (most prominently [6]), after the success of DETR [7] many transformer-based architectures have been introduced (e.g. [8], [9]). In our case, we extend INSTR with a segmentation head and fuse the output with the instance predictions.

Out-of-Distribution (OOD) detection is an important research area on its own, and for brevity we refer the reader to e.g. [10] for an extensive overview. To robustly separate detected instances into known and unknown categories, we follow previous work on (the approximation of) Bayesian inference, specifically by utilizing Dropout [11], [12], in the following referred to as *Bayesian Neural Network* (BNN).

In summary, our contributions are threefold:

- We extend INSTR to PS and show that explicitly modeling semantic information aids the object instance segmentation objective.
- We examine the task of separating the detected instances into known and unknown categories, and investigate performance on false-positive detections of INSTR.
- We demonstrate the applicability of INSTR for stone segmentation in extra-terrestrial environments and grasping for table clearing, where the BNN well alleviates autonomous object interaction (see also Figure 1).

[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

II. Proposed Extensions

A. Panoptic Segmentation

To further facilitate object instance segmentation we add a second decoder head which directly upsamples *Transformer Encoder* (TF_{Enc}) features into semantic categories (Figure 2 top right). Since INSTR only considers unknown objects as instances (or *things*, as commonly referred to in PS), we refrain from upsampling semantic object information as this is implicitly defined by the object query outputs. In other words, the semantic head only upsamples *stuff* classes, while the mask head of INSTR predicts *things*. As an example, let us consider background and object as semantic regions. In addition to the predicted instance tensor $\mathbf{I} \in \mathbb{R}^{b \times n_q \times h \times w}$ with n_q object queries and b, h and w being batch size, height, and width, let $\mathbf{S} \in \mathbb{R}^{b \times n_s \times h \times w}$ be a semantic output tensor, where n_s defines the number of semantic classes (in case of only predicting the background as semantic class $n_s = 1$). A complete panoptic prediction tensor $\hat{\mathbf{P}}$ is then derived with

$$\hat{\mathbf{P}} = \underset{dim=1}{softmax}(\underset{dim=1}{concatenate}(\mathbf{S}, \mathbf{I})). \quad (1)$$

Notably, softmax-scores are calculated across semantic and instance information. We further match instance predictions to their ground truth labels and apply the modified Dice loss as in [1]. This is easily extendable to multiple semantic (*stuff*) categories, yet the prediction of *things* of semantic categories other than unknown objects would require additional means of differentiation, e.g. in the form of a classification head (as for example in [7]).

B. Separating the Known from the Unknown

We explore two different settings for known / unknown classification (Figure 2 bottom right) with a ResNet50 [13] classifier, once with *Monte Carlo Dropout* (MCD) [11] and a second model with trainable *Concrete Dropout* (CD) [12]. It was shown that dropout can be interpreted as sampling from an approximate of the posterior distribution $p(\boldsymbol{\omega}|D)$ (with $\boldsymbol{\omega}$ being a set of learnable weights and $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^{N}\}$ being the training data set). In contrast to using the likelihood $p(\boldsymbol{y}^*|\boldsymbol{x}^*, \boldsymbol{\omega})$, the approximation of the posterior has the advantage of incorporating epistemic uncertainty (stemming from the model's parameters), thus providing more interpretable uncertainty estimates. In practice, MCD inference is done by averaging multiple stochastic forward passes to obtain the model's predictive mean.

Yet, for well-calibrated uncertainty estimates, the dropout probability has to be determined well, which can be a computationally expensive search since it has to be set before training. To mitigate this, Gal *et al.* [12] propose CD as extension, which allows optimizing the dropout probability directly, and we refer the reader to the original paper for further details. Implementation-wise, for the MCD case we insert a dropout layer before the final classifier layer; for CD we follow the authors' implementation and append the original multi-layer CD perceptron to the ResNet.

TABLE I: Mean IoU [%] on object instance masks across all scenes of *Stereo Instances On Surfaces* (STIOS) for both cameras (*rc_visard*, *Zed*). Values in **bold** denote the best results.

Semantic classes	rc_visard			Zed		
	mIoU	F1	PQ	mIoU	F1	PQ
Background	77.43	86.17	72.23	75.34	84.74	73.06
Background + Table	**77.72**	**86.49**	70.35	**75.48**	**85.02**	68.86
None (*base version*)	74.93	84.50	**73.27**	74.06	83.80	**73.52**

III. Experiments

A. Metrics

For PS we calculate the binary *Intersection over Union* (IoU) and F1 score on the matched pairs in a *size-sensitive* way - i.e., we summarize binary TP, FP and FN scores over all objects in a scene before deriving IoU and F1 scores to be consistent with [1]. Additionally, *Panoptic Quality* (PQ) [5] is listed which instead averages across detected instances:

$$PQ = \frac{\sum_{(p,g) \in TP} IoU(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}. \quad (2)$$

For the classifier we depict standard ROC and Precision/Recall curves, and list the ratio for OOD detection.

B. Panoptic Segmentation

We train two INSTR networks on panoptic segmentation from scratch following the original training schedule of [1], and list results in Table I. While mean IoU and F1 scores increase in contrast to the baseline INSTR (bottom row of Table I), the PQ decreases. Potentially, the *softmax* removes low-probable instances which would otherwise be kept by evaluating them individually (and independently) with a *sigmoid*. Note that the comparison to the plain INSTR is drawn to emphasize the overall benefit of the panoptic head, and we leave further comparison with similar frameworks (e.g. [14]) for future work. Exemplary qualitative results are depicted in Figure 3.

C. Classification of Known vs. Unknown

To simulate known and unknown objects, we randomly separate the 15 YCB objects of the STIOS dataset [1] into ten known and five unknown bins. The model from Section II-B is trained for 30 epochs on the first 10,000 images from the BOP challenge [2], where an input is the cropped, masked RGB detection. We use standard cross entropy loss and AdamW optimizer, and repeat the training ten times on different random object separations to avoid any bias on specific items. The accuracy-wise best performing model on the validation set is taken for inference. As comparison we also list the performance without MCD.

[1] Available at `https://www.dlr.de/rm/en/desktopdefault.aspx/tabid-17628#gallery/36367`

[2] Available at `https://bop.felk.cvut.cz/challenges/#datasets`.

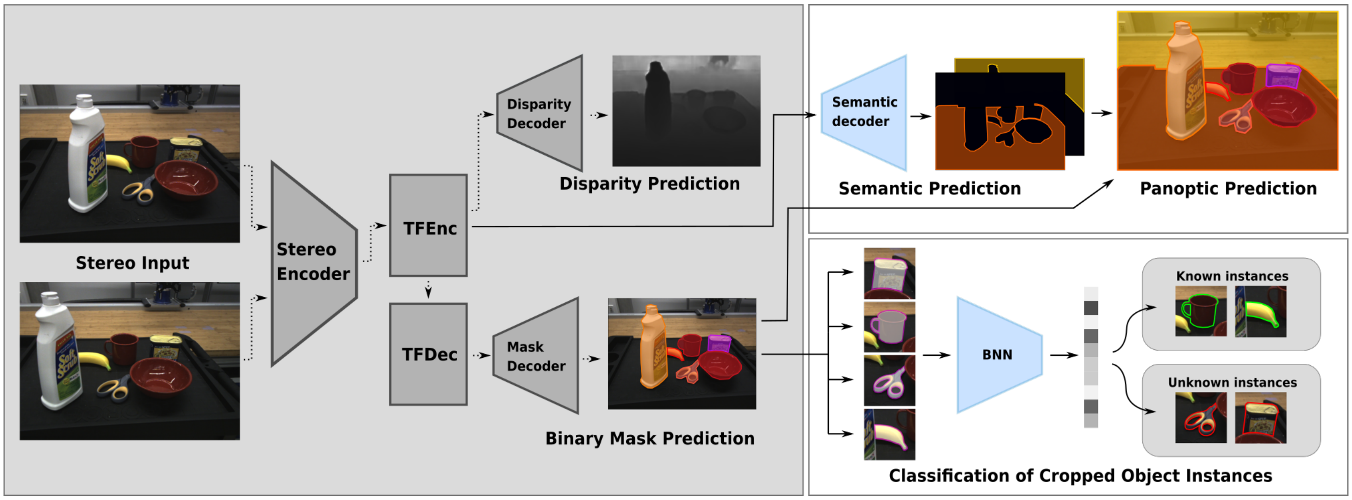

Fig. 2: Proposed extensions to INSTR (left, gray background): A semantic decoder upsamples TF_{Enc} features which are fused with object instance predictions for PS (top right). Object instance predictions are also forwarded to a BNN, here trained on *mug* and *banana*, but not on *scissors* and *potted meat can* (bottom right). This allows to further separate objects into known and unknown instances.

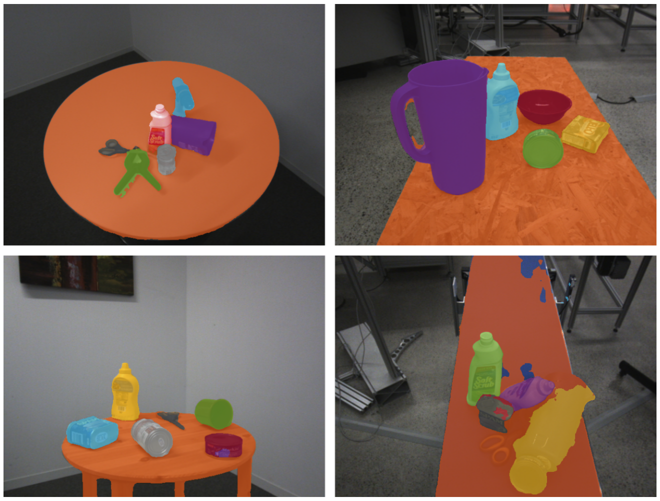

Fig. 3: Exemplary panoptic-like results on STIOS (best viewed magnified and in color). Semantic classes are *background* and *table*; the orange color denotes the table class, the background is not colored. All other colors are assigned randomly. The bottom right image depicts a failure case for heavily stacked objects.

Figure 4 depicts ROC and PR curves for in-distribution items as well as number of correctly and mis-classified instances, and the ratio of OOD samples including false-positive predictions of INSTR for different confidence levels. Results are averaged across all ten training runs.

Overall, CD performs slightly better on known instances (ROC / PR curve (top) and green / orange bars (bottom)). While the MCD model excels at identifying OOD samples (orange-red bars are comparably lower at higher confidences), the baseline model is spuriously more confident, which matches the theoretical foundation of the MCD.

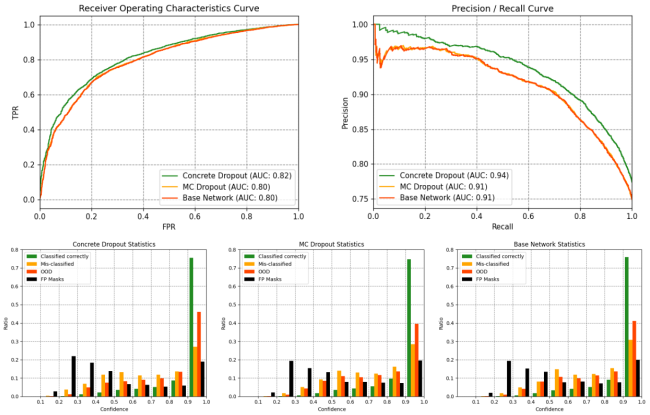

Fig. 4: Top: ROC and PR curve for in-distribution samples averaged across ten runs on INSTR predictions of the STIOS dataset. Bottom: Rate of (mis-)classification, OOD detection and the performance of false-positive detections of INSTR. Overall, CD marginally outperforms MCD and the plain ResNet in our setting.

D. Applications in the Wild

INSTR will be employed for stone segmentation as part of the ARCHES mission [15], where its task is to identify rocks in extra-terrestrial environments to either provide instance masks to a scientist in the ground station for further selection, or to directly grasp them in an autonomous top-down manner. Especially for completely autonomous grasp selection it is vital to only segment stones, and the LRU's sample extraction box or its battery packs pose challenging anomalies. To this end, we train a BNN with CD as in Section II-B on renderings of OAISYS [16] and a collection of other objects formed by the 15 YCB objects in STIOS from the aforementioned BOP dataset. We display quantitative results

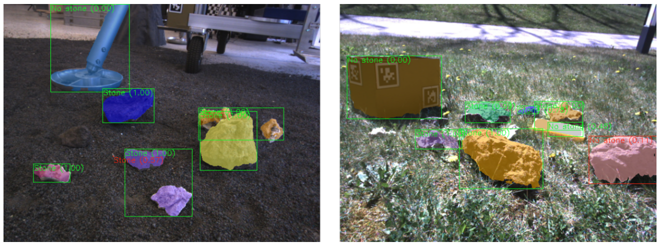

Fig. 5: Exemplary BNN predictions for stone segmentation on the LRU2 (best viewed magnified and in color). Green and red boxes denote correct and wrong classified instances, respectively. The number in brackets lists the confidence for the class *stone*. For a detailed explanation we refer to Section III-D

in Figure 5 and proceed by shortly discussing the findings: In the left image, the foot of the lander is identified as unknown object, but correctly labeled as *no stone*. The pixel-wise, artifact-like prediction (in the center, in red) is labeled as stone, but with a comparably low confidence. In the middle, the container box (with AprilTags, on the left) and the yellow battery pack (middle right) are successfully identified as *no stone*. based The instance on the right is falsely labeled as stone, but with lower confidence. The BNN is similarly applicable to INSTR with PS (Figure 1), where we hardcode semantic labels (in this case the surface). A particular failure case is shown in Figure 7 (left).

Unknown object instance segmentation is often a prerequisite for autonomous grasping, for instance to clear a table, and INSTR can be paired well with grasping architectures like Contact-GraspNet [17]. Here, instance masks are used to filter out invalid grasps on e.g. the surface area, and to reduce inference speed. In Figure 6 the overall framework is employed on a LWR3 robot arm, and we highlight how the proposed extensions can be incorporated. Note that both the panoptic segmentation and the BNN classifier have been added after runtime. Yet, the qualitative results indicate that our presented extensions allow INSTR to be robustly applied for sorting known objects into correct bins, and leaving the remaining objects for further inspection with e.g. [18]. Finally, a failure case is depicted in Figure 7 (right).

IV. CONCLUSIONS

Robustly perceiving and interacting with unknown objects in real-world environments is a key challenge in robotic applications. Detecting and segmenting all available instances is an important requirement. En plus, systems could benefit from further semantic information in the environment, or would like to incorporate readily available information on (some) of the objects in the scene. In this work, we extended INSTR to Panoptic Segmentation, providing semantic labels to all pixels and thereby simultaneously increasing object instance mask quality. Additionally, we showed that a simple BNN is well suited to classify known items while importantly being less confident for unknown encounters. We highlighted the applicability of the overall pipeline for grasping stones in an extra-terrestrial sample collection setting, and for table clearing in industrial / house-hold applications.

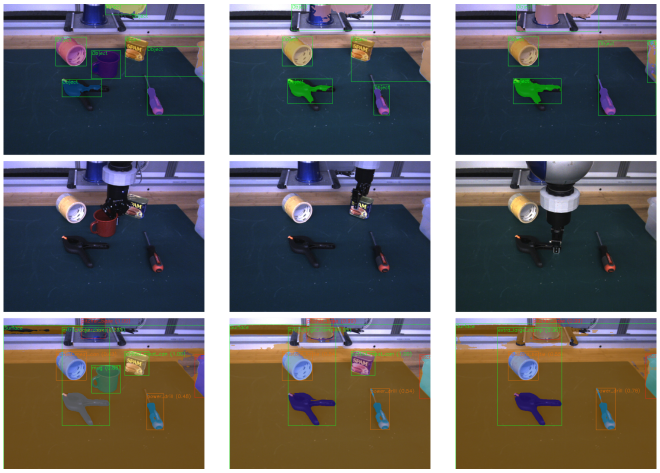

Fig. 6: INSTR together with Contact-GraspNet [17] for autonomous table clearing. Initial segmentation masks (top row) are used to successively (left to right) grasp unknown objects (middle row). PS increases mask quality (bottom row), and a classifier supports sorting into known categories and unknown items. Here, orange colors denote confidences smaller than 0.8, and red colors denote wrongly labeled objects. Note that the robot segmentation could be filtered out by the available hand-eye calibration, and the object on the right side is the drop-off bin which was not placed outside the camera view.

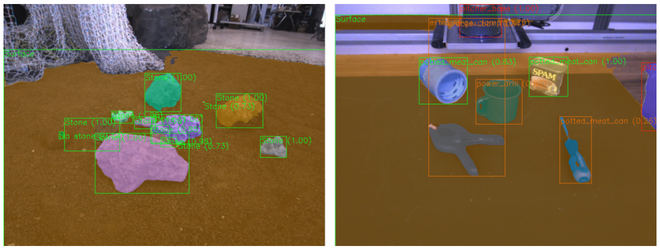

Fig. 7: Two failure cases. Left: Multiple falsely detected items for the task of stone segmentation with the panoptic head. Fine-tuning INSTR on the subset of items (here *stones*) would increase segmentation performance, as shown in [4]. Right: The BNN trained on YCB objects (here *extra_large_clamp*, *mug* and *pottet_meat_can*) mislabels the *mug*, and places low confidence values on the *extra_large_clamp* (predictions with confidence larger than 0.8 are colored green, otherwise orange. Red colors denote false predictions). We believe that the performance can be enhanced by improving the OOD dataset and further tuning of the BNN.

ACKNOWLEDGMENT

The authors would like to thank Ryo Sakagami for assisting in data collection and stone grasping on the LRU. Finally, the authors thank the reviewers for helpful comments and feedback.

References

[1] M. Durner, W. Boerdijk, M. Sundermeyer, W. Friedl, Z.-C. Marton, and R. Triebel, "Unknown Object Segmentation from Stereo Images," *arXiv:2103.06796 [cs]*, Mar. 2021, arXiv: 2103.06796. [Online]. Available: http://arxiv.org/abs/2103.06796

[2] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "Unseen Object Instance Segmentation for Robotic Environments," *arXiv:2007.08073 [cs]*, Jul. 2020, arXiv: 2007.08073. [Online]. Available: http://arxiv.org/abs/2007.08073

[3] Y. Xiang, C. Xie, A. Mousavian, and D. Fox, "Learning RGB-D Feature Embeddings for Unseen Object Instance Segmentation," *arXiv:2007.15157 [cs]*, Mar. 2021, arXiv: 2007.15157. [Online]. Available: http://arxiv.org/abs/2007.15157

[4] W. Boerdijk, M. G. Müller, M. Durner, M. Sumdermeyer, W. Friedl, A. Gawel, W. Stürzl, Z. C. Márton, R. Siegwart, and R. Triebel, "Rock Instance Segmentation from Synthetic Images for Planetary Exploration Missions," 2021.

[5] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, "Panoptic Segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 9396–9405. [Online]. Available: https://ieeexplore.ieee.org/document/8953237/

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *arXiv:1703.06870 [cs]*, Jan. 2018, arXiv: 1703.06870. [Online]. Available: http://arxiv.org/abs/1703.06870

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," *arXiv:2005.12872 [cs]*, May 2020, arXiv: 2005.12872. [Online]. Available: http://arxiv.org/abs/2005.12872

[8] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention Mask Transformer for Universal Image Segmentation," *arXiv:2112.01527 [cs]*, Dec. 2021, arXiv: 2112.01527. [Online]. Available: http://arxiv.org/abs/2112.01527

[9] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 5459–5470. [Online]. Available: https://ieeexplore.ieee.org/document/9578908/

[10] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized Out-of-Distribution Detection: A Survey," *arXiv:2110.11334 [cs]*, Oct. 2021, arXiv: 2110.11334. [Online]. Available: http://arxiv.org/abs/2110.11334

[11] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," p. 10.

[12] Y. Gal, J. Hron, and A. Kendall, "Concrete Dropout," *arXiv:1705.07832 [stat]*, May 2017, arXiv: 1705.07832. [Online]. Available: http://arxiv.org/abs/1705.07832

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385

[14] J. Hwang, S. W. Oh, J.-Y. Lee, and B. Han, "Exemplar-Based Open-Set Panoptic Segmentation Network," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 1175–1184. [Online]. Available: https://ieeexplore.ieee.org/document/9578145/

[15] M. J. Schuster, B. Rebele, M. G. Müller, S. G. Brunner, A. Dömel, B. Vodermayer, R. Giubilato, M. Vayugundla, H. Lehner, P. Lehner, F. Steidle, L. Meyer, K. Bussmann, J. Reill, W. Stürzl, I. von Bargen, R. Sakagami, M. Smisek, M. Durner, E. Staudinger, R. Pöhlmann, S. Zhang, C. Braun, E. Dietz, S. Frohmann, S. Schröder, A. Börner, H.-W. Hübers, R. Triebel, B. Foing, A. O. Albu-Schäffer, and A. Wedler, "The arches moon-analogue demonstration mission: Towards teams of autonomous robots for collaborative scientific sampling in lunar environments," in *European Lunar Symposium (ELS)*, 2020. [Online]. Available: https://elib.dlr.de/139810/

[16] 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.

[17] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes," *arXiv:2103.14127 [cs]*, Mar. 2021, arXiv: 2103.14127. [Online]. Available: http://arxiv.org/abs/2103.14127

[18] W. Boerdijk, M. Sundermeyer, M. Durner, and R. Triebel, ""What's

This?" – Learning to Segment Unknown Objects from Manipulation Sequences," *arXiv:2011.03279 [cs]*, Nov. 2020, arXiv: 2011.03279. [Online]. Available: http://arxiv.org/abs/2011.03279