Trinity: Unifying Class-Agnostic Terrain and Semantic Segmentation for Planetary Environments by Leveraging Synthetic Data

Marcus G Müller[12*], Wout Boerdijk[1], Maximilian Durner[1], Abel Gawel[3], Wolfgang Stürzl[1], Roland Siegwart[2], Rudolph Triebel[1]

***Abstract*— Terrain understanding is fundamental for mobile robots operating in unstructured outdoor environments, including planetary surfaces on Earth, Mars, the Moon, and beyond. Existing vision-based traversability estimation methods rely on robot-specific annotations or semantic class mappings, limiting transferability across platforms and requiring costly re-annotation when robot capabilities change, while standard semantic segmentation methods only focus on specific predefined classes, which do not capture the variety of terrains. In this work, we propose a transformer-based architecture that jointly performs class-specific semantic segmentation and class-agnostic terrain segmentation within a unified network, called Trinity. Terrain regions are segmented based solely on visual appearance, without predefined semantic labels or robot-dependent traversability scores. This formulation enables the learning of robot-agnostic visual terrain priors that can be combined with robot-specific experience for downstream tasks such as traversability estimation, visual odometry, and mission planning. To enable large-scale training with diverse terrain appearances, we extend the OAISYS simulator and introduce RUGDSynth, a synthetic dataset inspired by RUGD with class-agnostic terrain samples. Furthermore, we present the EXTerra Dataset, providing real-world images annotated with both class-specific and class-agnostic terrain labels. Experiments demonstrate the feasibility of the proposed task and the effectiveness of our joint segmentation approach in complex planetary outdoor environment.**

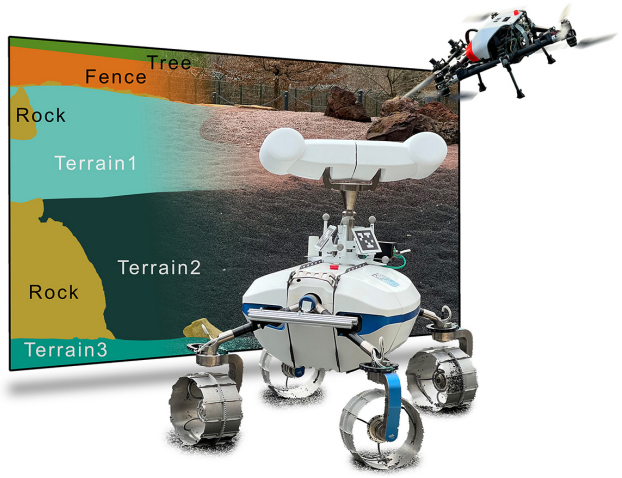

Fig. 1. Rover LRU [3] and drone ARDEA [4] observe a planetary exploration outdoor laboratory. The overlapping annotations shown on the color image illustrate the ideal segmentation results produced by the proposed method. The task is to segment different terrain types (shown in shades of cyan) in a class-agnostic manner, enabling their use across a wide range of robotic applications and platforms, while simultaneously identifying class-specific regions that are important for a mission (here: rock, fence, and tree).

I. Introduction

Semantic scene understanding is a fundamental capability for mobile robots operating in complex outdoor environments, especially for planetary space robotics. In such settings, reliable perception supports navigation, state estimation, path planning, and manipulation. It can further enable higher-level objectives such as identifying scientifically relevant areas or suitable terrain for infrastructure deployment in planetary exploration.

A central component of outdoor scene understanding is terrain segmentation, with traversability estimation as a closely related task. While supervised semantic segmentation has achieved remarkable success in structured urban environments [1], transferring these approaches to natural terrain remains challenging. In contrast to roads or buildings with well-defined geometry, outdoor environments exhibit irregular regions, gradual transitions, and visually similar surface types. Boundaries are often ambiguous, and semantic categories are less clearly separable, making also the annotation process difficult [2].

Beyond these perceptual challenges, terrain semantics are inherently context-dependent. The appearance and meaning of concepts such as soil or grass vary with environmental conditions (e.g., weather, season, illumination) and with the operational context of the robot. For instance, sand may pose no difficulty for a tracked vehicle but obstruct a small wheeled robot. Consequently, semantic labels and especially traversability assessments are not absolute properties, but depend on both environmental and robot-specific factors. This variability makes it difficult to define a universally valid, fixed taxonomy of terrain classes.

Most existing methods, however, assume precisely such predefined class sets and operate under a closed-world assumption in which all relevant categories are known during training. In traversability estimation, supervision is further tied to a specific robot platform, since annotations implicitly encode its mobility capabilities. As a result, deploying models in new environments or on different robots typically requires new data collection and retraining, limiting scalability and robustness.

* Corresponding author marcus.mueller@dlr.de

**This work was supported by the Helmholtz Association project iFOODis (contract number KA2-HSC-06)

[1]Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Weßling, Germany

[2]Federal Institute of Technology Zurich (ETH Zurich), Zurich, Switzerland

[3]Robotics and AI Institute (RAI), Zurich, Switzerland

To address these challenges, we introduce Trinity-Net, a unified transformer-based architecture for flexible scene understanding that jointly performs class-agnostic and class-specific pixel-wise segmentation without relying on fixed terrain taxonomy.

The class-agnostic part separates scenes into visually coherent terrain regions without predefined semantic labels. By focusing purely on appearance, it provides a robot-agnostic terrain representation that can serve as a visual prior for (self-supervised) downstream tasks [5], [6]. Unlike local region-proposal methods [7], Trinity-Net is able to enforce global visual consistency across spatially disjoint terrain regions.

The class-specific branch predicts platform- and context-independent or mission critical semantic categories (e.g., sky, rocks, robotic infrastructure). This separation allows the model to retain stable semantic anchors while avoiding an overconstrained terrain taxonomy. Fig. 1 gives an illustration of the overall task.

Training such a system requires substantial terrain diversity to ensure robust generalization and genuinely class-agnostic behavior. However, annotating large-scale datasets with terrain labels is costly and time-consuming. To address this limitation, we additionally leverage synthetic data. Unfortunately, many existing outdoor datasets do not provide terrain-specific annotations. Moreover, most simulation environments are unable to output such information, as their semantic labels are typically object-based rather than texture- or material-based. In contrast, the OAISYS simulator [8] provides semantic annotations using a material-based labeling approach. This enables the extraction of terrain annotations corresponding to different surface textures.

To this end, we leverage and extend the OAISYS simulator to generate a large-scale synthetic dataset for joint training. For class-specific categories we are following the taxonomy of the RUGD dataset [9]. That makes it possible to also use our approach for other field robotic tasks on Earth. Our approach achieves strong class-agnostic segmentation results on a real-world dataset recorded for planetary exploration scenarios while being able to segment class-specific classes as well.

To summarize, our contributions are: (i) **Trinity-Net, a unified transformer architecture** for joint class-specific semantic and class-agnostic segmentation, trained on synthetic data to learn transferable appearance-based terrain representations; (ii) an **extension of the OAISYS simulator** enabling large-scale synthetic data generation for this task, resulting in **RUGDSynth**, a synthetic counterpart to the field-robotics RUGD benchmark, and (iii) the **EXTerra Dataset**, a real-world planetary exploration dataset from an analog mission site.

II. Related Work

Our work builds on semantic segmentation [10], [11], with emphasis on terrain segmentation, traversability estimation, and open-world segmentation.

Terrain Segmentation is commonly formulated as supervised semantic segmentation with predefined terrain taxonomies, requiring extensive manual annotation and fixed class definitions. These limitations become particularly evident in planetary exploration, where annotated data is scarce and highly variable terrain appearance challenges rigid semantic categories. Existing approaches range from rock-focused obstacle classification [12], [13] to broader supervised semantic labeling efforts [2], [14], [15]. To improve robustness multimodal perception is additionally explored, by fusing visual and thermal imagery [16]. Several works also investigate reduced-supervision, including sparse labeling [17], semi-supervised learning [18], and text-guided foundation-model-based segmentation [19]. Despite mitigating annotation scarcity, these methods still operate within predefined terrain categories. Other approaches reduce semantic labels to navigability-oriented categories. In [20] a multi-scale transformer to aggregate contextual features and cluster terrains into roughness-based groups for navigation is employed. Similarly, Li *et al.* [21] propose a contextual-aware segmentation network that models local and global dependencies while operating within the same roughness-based terrain taxonomy. Closely related to our motivation, Ellis *et al.* [22] introduce a temporally consistent unsupervised approach for class-agnostic terrain segmentation. While they only address class-agnostic region discovery, we additionally support class-specific semantic segmentation within a single framework.

Traversability Estimation. While terrain segmentation assigns semantic labels to surface regions, traversability estimation models navigability, commonly formulated as a binary decision or continuous feasibility score. Schilling *et al.* [23] combine visual and geometric features using random forest. Other methods use elevation maps to predict traversability via learning-based approaches [24], [25]. In [26] dense traversability maps from sparse point clouds for off-road navigation are learned. Several approaches exploit self-supervision from robot interaction. Kahn *et al.* [27] cast traversability as an end-to-end reinforcement learning problem. Other methods leverage platform-dependent (weak) supervision, such as foothold reprojection for legged robots [28], MPC trajectories [29], or proprioceptive signals including odometry and IMU signals [30]. Building on these ideas combined with foundation-model representations, Frey *et al.* [5] propose an online self-supervised system that adapts their model using visual embeddings and short human demonstrations, while Jung *et al.* [6] incorporate segmentation priors using Segment Anything Model (SAM) [7] to improve fine-grained traversability prediction.

Open-World Segmentation moves beyond fixed categories and closed-set assumptions to handle previously unseen categories during execution. Some works estimate predictive uncertainty to detect unknown regions, through Bayesian inference [31], [32]. Further, auxiliary datasets are leveraged to explicitly expose models to unknown samples during training [33], [34]. Generative [35], [36] and reconstruction-based [37], [38] methods detect novel regions by identifying discrepancies between input images and reconstructed outputs, as unseen patterns typically yield

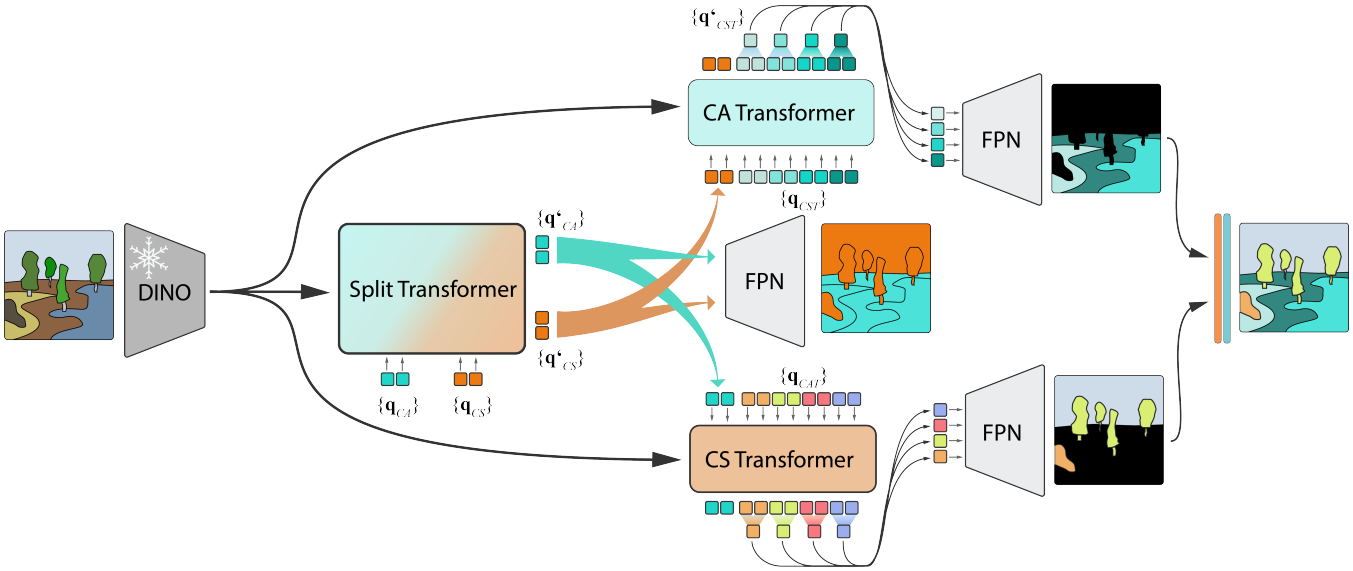

Fig. 2. The figure provides an overview of Trinity-Net. The model consists of an image encoder and three main components: the Split Transformer, the CS Transformer, and the CA Transformer. The input image is first processed by a DINOv2 image encoder to obtain a feature map. Queries in the Split Transformer attend to the image features and implicitly divide the feature space into a CS and a CA component. The resulting queries that attend to the CS component are extended to form the queries of the CA Transformer, while the queries attending to the CA component are expanded to form the queries of the CS Transformer. The queries of both the CS and CA Transformers attend to the feature map. The resulting output queries are then aggregated and upsampled. Finally, the class-specific output is concatenated with the class-agnostic output.

higher reconstruction errors. In [39] distance-based unsupervised anomaly detection is introduced. Finally, feature-space approaches aim to separate known from unknown classes by measuring distances to learned class clusters within the feature space [40], [41].

III. Class-Specific and Class-Agnostic Ground Segmentation

We introduce Trinity-Net, a novel unified framework for class-agnostic terrain and class-specific semantic segmentation, enabled by a large synthetic dataset spanning diverse terrains. It is composed of three main parts, hence the name of the model. We first present the architecture, followed by the simulation and dataset generation pipeline.

A. Architecture - Trinity-Net

Our method addresses two related but distinct sub-tasks: *class-specific (CS) semantic segmentation* of predefined categories known a priori, capturing platform-independent and mission-relevant semantics, and *class-agnostic (CA) terrain segmentation* that groups visually coherent ground regions in a robot- and context-agnostic manner. Each pixel p is assigned exclusively to either a specific class C_{CS} or to a non-specified ground class C_{CA}:

$$p \in C_{CS} \cup C_{CA} \quad \text{where} \quad C_{CS} \cap C_{CA} = \varnothing. \tag{1}$$

The overall architecture is illustrated in Fig. 2. We first encode the input RGB image using a frozen DINOv2 backbone [42]. The extracted image features are passed to a DETR-like *Split-Transformer* with learnable query sets, $\{\mathbf{q}_{CS}\}$ and $\{\mathbf{q}_{CA}\}$, which partition the feature space into class-specific and class-agnostic regions. To explicitly encourage the intended partitioning of the feature space, we apply an auxiliary loss to the updated query embeddings. For this purpose, we map the ground-truth annotations to a binary representation distinguishing CS from CA regions.

The updated querys $\{\mathbf{q}'_{CS}\}$ and $\{\mathbf{q}'_{CA}\}$ are then separated and passed to two subsequent transformer modules: the *CS-Transformer (CST)*, responsible for segmenting class-specific semantics, and the *CA-Transformer (CAT)*, which handles class-agnostic regions. Each transformer operates with its own set of learnable queries, $\{\mathbf{q}_{CST}\}$ and $\{\mathbf{q}_{CAT}\}$, representing prototypes for the respective regions. Since a single query cannot capture the full complexity of a CA region or CS category in the real-world, we employ multiple queries per region to represent its variability and increase robustness.

To guide each transformer toward its relevant regions, we expand the task-specific queries with the updated queries: $\{\mathbf{q}'_{CA}\} \cup \{\mathbf{q}_{CST}\}$ and $\{\mathbf{q}'_{CS}\} \cup \{\mathbf{q}_{CAT}\}$. Intuitively, the Split-Transformer attends to complementary regions of the scene for each task, thereby providing contextual information that facilitates separation.

We average the resulting task-specific queries, $\{\mathbf{q}'_{CST}\}$ and $\{\mathbf{q}'_{CAT}\}$, that belong to the same region. This aggregation step reduces the queries to match the final CS and CA output channels and is performed prior to upsampling for memory efficiency. Note, the number of CA prototypes is chosen to be sufficiently large to accommodate diverse ground configurations. The amount of final CS output channels depends on the mission-relevant semantic classes. Finally, the aggregated queries are subsequently upsampled and

concatenated to form a unified prediction tensor containing both class-specific and class-agnostic segmentation maps.

Since the CAT does not enforce a fixed alignment between queries and regions, the same terrain may be represented by different queries across samples. To resolve this ambiguity, predictions are matched to ground-truth masks using a Hungarian matcher to determine the optimal alignment between predicted class-agnostic regions and ground-truth masks. After this matching step, a cross-entropy loss is applied to the full network output.

B. Data Generation - RUGDSynth

To train the class-agnostic component of our method, we require a large and diverse set of terrain samples. In addition to covering a wide range of ground types, the dataset must include the mission-relevant, class-specific categories. Since collecting such data in the real world is costly and time-consuming, we rely on synthetic data to efficiently meet the requirements. Therefore, we extend OAISYS [8], an open-source simulator pipeline based on Blender, that enables the creation of outdoor environments. It relies on material-based semantics, ideally suitable for our use-case to simulate different terrains and their corresponding annotations. Additionally, OAISYS provides the *MeshParticleScatter* option, which enables to scatter meshes using the internal particle system of Blender. This is particularly useful for simulating large environments with numerous objects (e.g., rocks, trees). On the other hand, this concept is less suitable when simulating a smaller amount of objects (e.g., cars, rovers) that have to be place more deliberately. To this end, we extend the simulator with a physics-based placement module called *MeshPhysicsScatter*. Objects can be spawned at a predefined height, where they fall under physical simulation onto a specified surface. Spawning locations can also be defined relative to the camera. Given the large simulated environments, this gives a higher chance that the objects are visible in the camera frame. Additionally, objects can be optionally respawned within a batch. Originally, each object had to be listed in the config file, which is cumbersome, dealing with many objects. We now allow a directory of assets, from which objects are randomly spawned. We also extended the modules with a probabilistic activation parameter, controlling whether a module runs in a given batch. This prevents all assets from spawning simultaneously, increasing scene variability while keeping computation manageable.

OAISYS is further enhanced with a custom material module, *MaterialCSCATerrain*, to simulate diverse ground surfaces. This module randomly selects textures from a predefined set and supports three configuration modes: class-specific assignment, class-agnostic assignment, or a mixture of both. For each set, a minimum and maximum number of textures can be specified. The maximum number is particularly important to ensure that the number of terrain variations does not exceed the downstream network's capabilities. Additionally, this constraint helps maintain GPU memory consumption within reasonable limits. Furthermore, the developed module is not only applying selected textures using the default scales but is additionally randomizing the sizes of each texture to further increase visual variability.

TABLE I
USED 3D ASSETS FOR RUGDSYNTH (TOTAL: 547)

3D Asset Type	No. Assets	Scatter Type	Activation Prob. [%]
Grass	210	Particle	40.0
Tree	96	Particle	60.0
Pole	6	Physical	40.0
Vehicle	41	Physical	99.2
Generic-Objects	10	Physical	40.0
Building	12	Physical	99.9
Log	40	Physical	40.0
Bicycle	5	Physical	30.0
Person	26	Physical	40.0
Fence	10	Physical	30.0
Bush	23	Particle	70.0
Traffic Sign	25	Physical	30.0
Rock	41	Particle	27.0
Picnic-Table	2	Physical	30.0

Based on the OAISYS extensions, we created a synthetic version of RUGD [9], called *RUGDSynth*. By mimicking RUGD, the synthetic dataset and trained model can be applied not only to planetary tasks but also to a broader range of field robotics applications. The dataset features diverse ground types as well as the class-specific categories found in RUGD. Trees, grass, and rocks are distributed across the scene using a particle system, with objects (e.g., vehicles, logs) spawned at the initial sensor location. Since the sensor is not moved far from its initial location, respawning within a batch is unnecessary, reducing computational overhead. Table I gives an overview over the used mesh assets. The sky is simulated using 68 randomly selected HDR images with randomized emission intensity. In total, 204 unique ground textures are used.

RUGDSynth contains 42,672 samples across more than 4,500 worlds (batches), each with roughly 10 samples. The dataset is split into 42,672 training, 1,000 validation, and 1,924 test samples. Figure 3 shows example scenes.

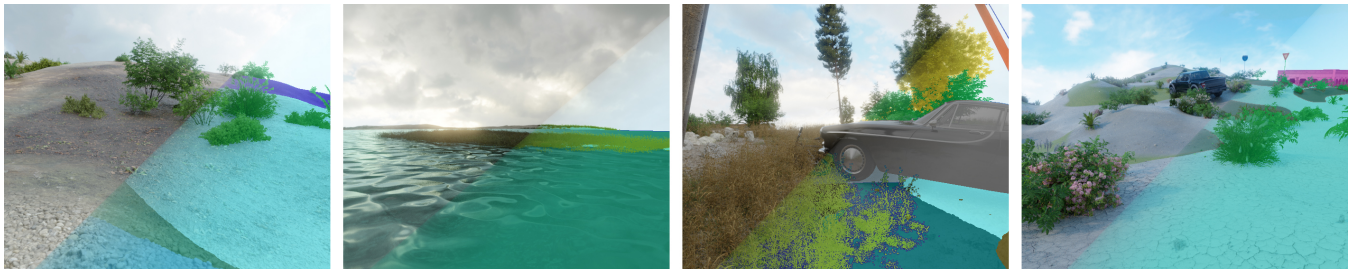

Fig. 3. Sample images from RUGDSynth overlaid with the corresponding annotations. Terrain regions are randomly colored in shades of cyan.

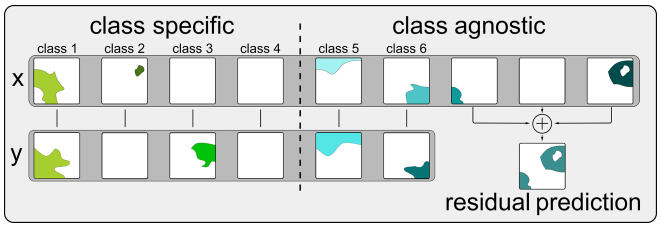

Fig. 4. Illustration of the residual prediction calculation. All region proposal masks that are not matched to a ground-truth mask are summed to produce the residual prediction mask. In the ideal case, this mask is empty.

IV. Experiments

In this section, we present the performance of our method. We introduce the EXTerra (Plan**e**tary E**x**ploration **Terr**ain) dataset, we collected and annotated data from a planetary exploration outdoor lab. This dataset will be used for evaluation in this section. The data were acquired using the LRU rover system [3]. The semantic categories are labeled according to the RUGD [9] taxonomy. All ground surfaces are treated as class-agnostic, while mission-relevant, platform- and context-independent categories are labeled as class-specific. Consequently, terrains such as sand, gravel, and mud, including their sub-classes, are assigned to the class-agnostic set, while trees, vegetation, vehicles, and buildings are treated as class-specific. The set of specific classes consists of the classes: tree, pole, sky, person, vehicle, grass, generic-object, building, log, fence, bush, sign, rock, and picnic-table. In principle, any separation into different subsets should be possible and is independent of the method presented here. The dataset comprises 14 additional distinct terrain types and consists of 124 samples. Figure 5 gives an overview over the outdoor lab and examples from the EXTerra Dataset.

Our network features 16 semantic classes, and provides 20 possible class-agnostic prediction slots. If not mentioned differently, we train Trinity-Net with the training set of RUGD and RUGDSynth.

A. Evaluation Metric

In semantic segmentation, the number of classes in both the annotations and the predictions is predefined and consistent. This assumption is not given for the class-agnostic task. Although the number of annotations is fixed, the number of potential region predictions is arbitrary and determined by the method. Relying solely on the standard Intersection over Union (IoU) values would fail to account for errors arising from predicting more class-agnostic regions than exist. Consequently, the IoU as standard semantic segmentation metric is insufficient for our case, therefore, we introduce an additional metric. Specifically, we aggregate all residual class-agnostic prediction masks that were not matched with any ground truth annotation mask into a single mask. Ideally, the consquent residual mask should contain no entries, whereas in the worst case, it has entries at every pixel. Based on the residual mask, we calculate the recall value, comparing it to a ground truth mask where all entries are set to one. The resulting Residual Recall (resR) value quantifies the ratio of predicted to actual class-agnostic terrain regions in the scene. Larger values indicate an increasing surplus of predicted regions. Figure 4 illustrates the metric calculation.

B. Class-Specific and Class-Agnostic Evaluation

Quantitative results on the EXTerra Dataset are provided in Table II, while qualitative results are found in Fig. 6. Trinity-Net outperforms SAM in the class-agnostic case by a high margin. In Fig. 6 we can observe that SAM is preforming well when boundaries are visually clear. If boundaries are fuzzy and intersect, like it is often the case for terrains, the mask predictions quality is decreasing as can be seen in the example of second column of Fig. 6.

SAM is performing slightly better in the class-specific case. We hypothesise that most appearance of such classes are in the far field of the robot. Since SAM is having an uniformly grid of point prompts, it does not differentiate too much between fore- and background.

Next, we focus on the relevance of synthetic data for the generalization capabilities of our model. We train Trinity-Net with exclusively either the RUGD data or RUGDSynth, and show the results in Table II and Fig. 6. The model trained on only synthetic data shows stronger metrics than the model only trained on RUGD, indicating the beneficial value of RUGDSynth. The union of both datasets shows the best results, which demonstrates that both can be used well together to fill each others shortcomings.

V. Conclusions

We presented Trinity-Net, a unified transformer-based architecture enabling flexible scene understanding without relying on fixed terrain taxonomy. By introducing a novel split-transformer design with interacting query sets, our approach decouples robot-agnostic terrain representation from mission-relevant semantic prediction. To support training at scale, we extended the OAISYS simulator and introduced RUGDSynth, a large synthetic dataset targeted for the proposed task. Furthermore, we introduced EXTerra, a real-world planetary exploration dataset, showcasing the domain-shift capabilities of Trinity-Net. We hope that the proposed task formulation, datasets, and methodology will stimulate further research in this field and beyond and contribute to advancing the level of autonomy in future robotic systems.

TABLE II
Evaluation on EXTerra Dataset

Model	cs	ca		
	mIoU	mIoU	mPre	mRec
SAM[7]	**35.06**	10.06	**51.73**	9.33
Trinity (ours)				
⊢ RUGDSynth	29.78	34.93	48.67	28.04
⊢ RUGD	23.36	27.00	41.44	22.54
⊢ RUGDSynth+RUGD	34.82	**41.84**	51.57	**34.82**

Fig. 5. Image of the planetary exploration outdoor laboratory and example samples from the EXTerra dataset with partially overlaid annotations and the approximate capture locations indicated by orange lines.

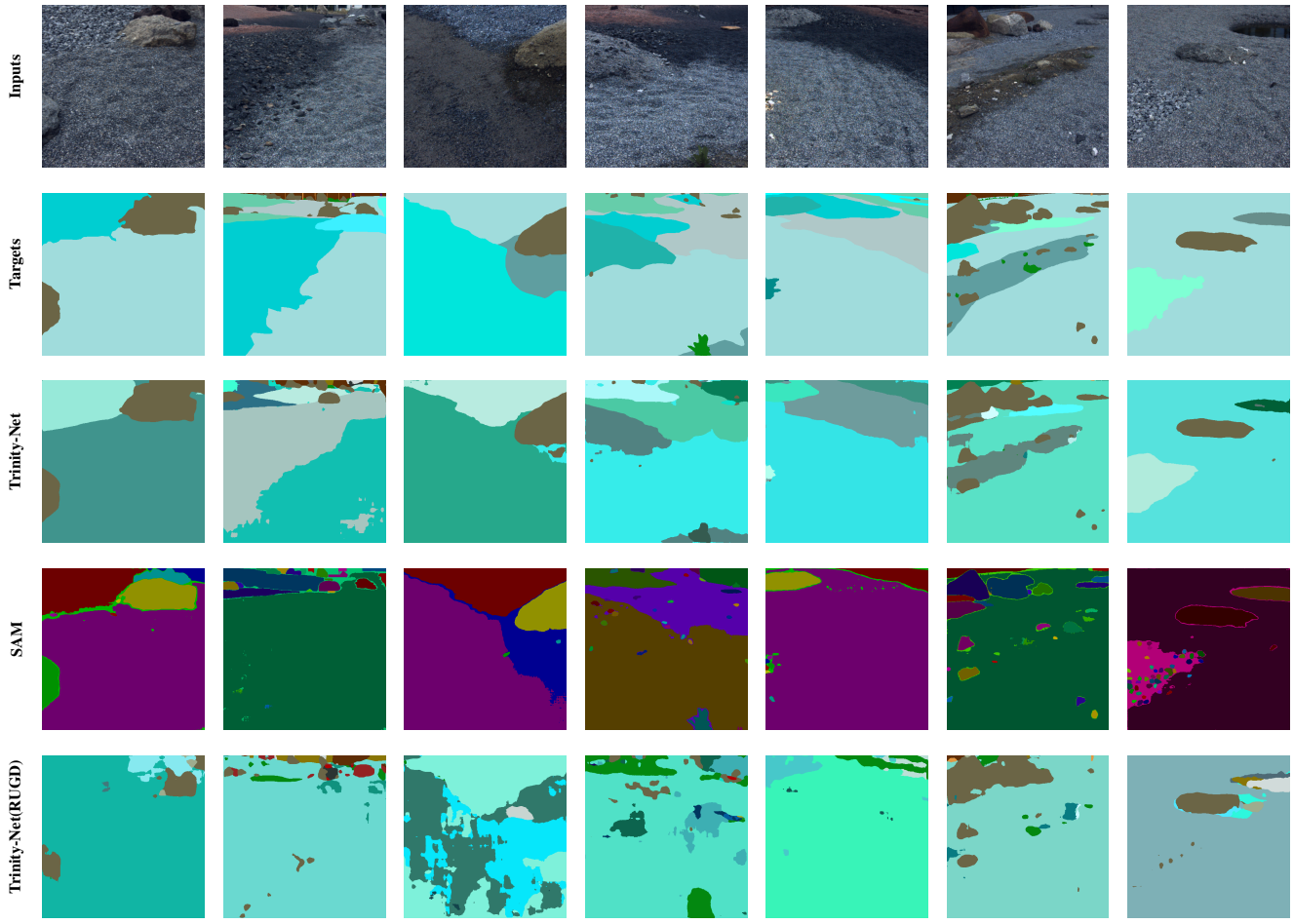

Fig. 6. Selected qualitative results for the EXTerra dataset. Class-agnostic regions are colored in shades of cyan. Note that the cyan colors used for the class-agnostic predictions are randomly assigned for each sample and do not necessarily match the ground-truth colors. For visualization clarity, the region predictions of SAM are randomly selected.

REFERENCES

[1] J. Van Brummelen, M. O'Brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," *Transportation Research Part C: Emerging Technologies*, 2018.

[2] 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 *IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.

[3] A. Wedler, B. Rebele, J. Reill, M. Suppa, H. Hirschmüller, C. Brand, M. Schuster, B. Vodermayer, H. Gmeiner, A. Maier, B. Willberg, K. Bussmann, F. Wappler, and M. Hellerer, "Lru – lightweight rover unit," in *Symposium on Advanced Space Technologies in Robotics and Automation (ASTRA)*, 2015.

[4] P. Lutz, M. G. Müller, M. Maier, S. Stoneman, T. Tomić, I. von Bargen, M. J. Schuster, F. Steidle, A. Wedler, W. Stürzl, and R. Triebel, "Ardea—an mav with skills for future planetary missions," *Journal of Field Robotics*, vol. 37, no. 4, pp. 515–551, 2020.

[5] J. Frey, M. Mattamala, N. Chebrolu, C. Cadena, M. Fallon, and M. Hutter, "Fast traversability estimation for wild visual navigation," *Proc. of Robotics: Science and Systems (RSS)*, 2023.

[6] S. Jung, J. Lee, X. Meng, B. Boots, and A. Lambert, "V-STRONG: Visual Self-Supervised Traversability Learning for Off-road Navigation," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2024.

[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2023.

[8] M. G. Müller, M. Durner, A. Gawel, W. Stürzl, R. Triebel, and R. Siegwart, "A Photorealistic Terrain Simulation Pipeline for Unstructured Outdoor Environments," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Sep. 2021.

[9] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A RUGD Dataset for Autonomous Navigation and Visual Perception in Unstructured Outdoor Environments," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2019.

[10] G. Csurka, R. Volpi, and B. Chidlovskii, "Semantic Image Segmentation: Two Decades of Research," *Foundations and Trends in Computer Graphics and Vision*, 2022.

[11] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic segmentation using Vision Transformers: A survey," *Engineering Applications of Artificial Intelligence*, Nov. 2023.

[12] M. Durner, W. Boerdijk, Y. Fanger, R. Sakagami, D. L. Risch, R. Triebel, and A. Wedler, "Autonomous Rock Instance Segmentation for Extra-Terrestrial Robotic Missions," in *Proc. of the IEEE Aerospace Conf.*, 2023.

[13] H. Liu, M. Yao, X. Xiao, and Y. Xiong, "RockFormer: A U-Shaped Transformer Network for Martian Rock Segmentation," *IEEE Trans. on Geoscience and Remote Sensing*, 2023.

[14] R. Gonzalez and K. Iagnemma, "Deepterramechanics: Terrain classification and slip estimation for ground robots via deep learning," *arXiv preprint arXiv:1806.07379*, 2018.

[15] M. G. Müller, M. Durner, W. Boerdijk, H. Blum, A. Gawel, W. Stürzl, R. Siegwart, and R. Triebel, "Uncertainty estimation for planetary robotic terrain segmentation," in *2023 IEEE Aerospace Conference*, 2023, pp. 1–8.

[16] R. Castilla-Arquillo, C. Perez-del Pulgar, L. Gerdes, A. Garcia-Cerezo, and M. A. Olivares-Mendez, "OmniUnet: A Multimodal Network for Unstructured Terrain Segmentation on Planetary Rovers Using RGB, Depth, and Thermal Imagery," *arXiv preprint arXiv:2508.00580*, 2025.

[17] E. Goh, J. Chen, and B. Wilson, "Mars Terrain Segmentation with Less Labels," in *Proc. of the IEEE Aerospace Conf.*, 2022.

[18] J. Zhang, L. Lin, Z. Fan, W. Wang, and J. Liu, "S5Mars: Semi-Supervised Learning for Mars Semantic Segmentation," *IEEE Trans. on Geoscience and Remote Sensing*, 2024.

[19] Y. Fang, X. Rao, X. Gao, W. Li, and Z. Min, "MTSNet: Joint Feature Adaptation and Enhancement for Text-Guided Multi-view Martian Terrain Segmentation," in *Proc. of the ACM Int. Conf. on Multimedia*, Melbourne VIC Australia, 2024.

[20] T. Guan, D. Kothandaraman, R. Chandra, A. J. Sathyamoorthy, K. Weerakoon, and D. Manocha, "GA-Nav: Efficient Terrain Segmentation for Robot Navigation in Unstructured Outdoor Environments," *IEEE Robotics and Automation Letters*, 2022.

[21] W. Li, M. Liao, and W. Zou, "Contextual-aware terrain segmentation network for navigable areas with triple aggregation," *Expert Systems with Applications*, 2025.

[22] C. Ellis, M. Wigness, C. Lennon, and L. Fiondella, "Temporally Consistent Unsupervised Segmentation for Mobile Robot Perception," *arXiv preprint arXiv:2507.22194*, 2025.

[23] F. Schilling, X. Chen, J. Folkesson, and P. Jensfelt, "Geometric and visual terrain classification for autonomous mobile navigation," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2017.

[24] R. O. Chavez-Garcia, J. Guzzi, L. M. Gambardella, and A. Giusti, "Learning Ground Traversability From Simulations," *IEEE Robotics and Automation Letters*, 2018.

[25] B. Yang, L. Wellhausen, T. Miki, M. Liu, and M. Hutter, "Real-time Optimal Navigation Planning Using Learned Motion Costs," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2021.

[26] A. Shaban, X. Meng, J. Lee, B. Boots, and D. Fox, "Semantic Terrain Classification for Off-Road Autonomous Driving," in *Proc. of the Conf. on Robot Learning (CORL)*, 2022.

[27] G. Kahn, P. Abbeel, and S. Levine, "BADGR: An Autonomous Self-Supervised Learning-Based Navigation System," *IEEE Robotics and Automation Letters*, 2021.

[28] L. Wellhausen, A. Dosovitskiy, R. Ranftl, K. Walas, C. Cadena, and M. Hutter, "Where Should I Walk? Predicting Terrain Properties From Images Via Self-Supervised Learning," *IEEE Robotics and Automation Letters*, 2019.

[29] M. V. Gasparino, A. N. Sivakumar, Y. Liu, A. E. B. Velasquez, V. A. H. Higuti, J. Rogers, H. Tran, and G. Chowdhary, "WayFAST: Navigation With Predictive Traversability in the Field," *IEEE Robotics and Automation Letters*, 2022.

[30] A. J. Sathyamoorthy, K. Weerakoon, T. Guan, J. Liang, and D. Manocha, "TerraPN: Unstructured Terrain Navigation using Online Self-Supervised Learning," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2022.

[31] H. Sapkota and Q. Yu, "Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection," in *Conf. on Computer Vision and Pattern Recognition*, 2022.

[32] M. G. Müller, M. Durner, W. Boerdijk, H. Blum, A. Gawel, W. Stürzl, R. Siegwart, and R. Triebel, "Uncertainty Estimation for Planetary Robotic Terrain Segmentation," in *Proc. of the IEEE Aerospace Conf.*, 2023.

[33] R. Chan, M. Rottmann, and H. Gottschalk, "Entropy Maximization and Meta Classification for Out-of-Distribution Detection in Semantic Segmentation," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2021.

[34] P. Bevandić, I. Krešo, M. Oršić, and S. Šegvić, "Simultaneous Semantic Segmentation and Outlier Detection in Presence of Domain Shift," in *Pattern Recognition*, 2019.

[35] K. Lis, S. Honari, P. Fua, and M. Salzmann, "Detecting Road Obstacles by Erasing Them," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2024.

[36] Y. Zhao, "OmniAL: A Unified CNN Framework for Unsupervised Anomaly Localization," in *Conf. on Computer Vision and Pattern Recognition*, 2023.

[37] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed Students: Student-Teacher Anomaly Detection With Discriminative Latent Embeddings," in *Conf. on Computer Vision and Pattern Recognition*, 2020.

[38] X. Zhang, S. Li, X. Li, P. Huang, J. Shan, and T. Chen, "DeSTSeg: Segmentation Guided Denoising Student-Teacher for Anomaly Detection," in *Conf. on Computer Vision and Pattern Recognition*, 2023.

[39] C.-C. Tsai, T.-H. Wu, and S.-H. Lai, "Multi-Scale Patch-Based Representation Learning for Image Anomaly Detection and Segmentation," in *IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, 2022.

[40] M. Sodano, F. Magistri, L. Nunes, J. Behley, and C. Stachniss, "Open-World Semantic Segmentation Including Class Similarity," in *Conf. on Computer Vision and Pattern Recognition*, 2024.

[41] H. Blum, M. G. Müller, A. Gawel, R. Siegwart, and C. Cadena, "SCIM: Simultaneous Clustering, Inference, and Mapping for Open-World Semantic Scene Understanding," in *Robotics Research*, 2023.

[42] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.