OTAS: Open-vocabulary Token Alignment for Outdoor Segmentation

Simon Schwaiger[1,2] Stefan Thalhammer[2] Wilfried Wöber[2] Gerald Steinbauer-Wagner[1]

[1]Graz University of Technology, Faculty of Computer Science and Biomedical Engineering, Institute of Software Engineering and Artificial Intelligence, 8010 Graz, Austria

[2]University of Applied Sciences Technikum Wien, Faculty of Industrial Engineering, Research Group Digital Manufacturing, Automation and Robotics, 1200 Vienna, Austria

Abstract: Understanding open-world semantics is critical for robotic planning and control, particularly in unstructured outdoor environments. Current vision-language mapping approaches rely on object-centric segmentation priors, which often fail outdoors due to semantic ambiguities and indistinct semantic class boundaries. We propose OTAS—an Open-vocabulary Token Alignment method for Outdoor Segmentation. OTAS overcomes the limitations of open-vocabulary segmentation models by extracting semantic structure directly from the output tokens of pretrained vision models. By clustering semantically similar structures across single and multiple views and grounding them in language, OTAS reconstructs a geometrically consistent feature field that supports open-vocabulary segmentation queries. Our method operates zero-shot, without scene-specific fine-tuning, and runs at up to ≈17 fps. OTAS provides a minor IoU improvement over fine-tuned and open-vocabulary 2D segmentation methods on the Off-Road Freespace Detection dataset. Our model achieves up to a 151% IoU improvement over open-vocabulary mapping methods in 3D segmentation on TartanAir. Real-world reconstructions demonstrate OTAS' applicability to robotic applications. The code and ROS node will be made publicly available upon paper acceptance.

Keywords: semantic segmentation, zero-shot learning, vision-language model

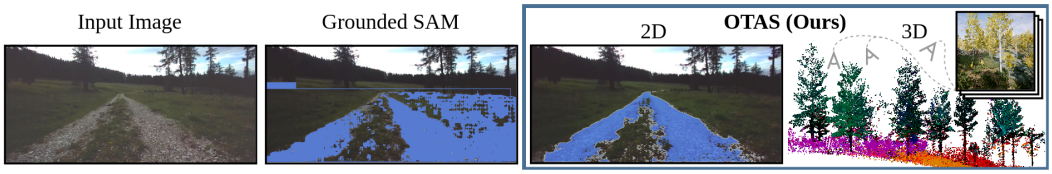

Figure 1: **OTAS** is an unsupervised segmentation method that aligns tokens from vision and language foundation models for robotic outdoor tasks. It operates zero-shot on single (2D) or multi-view (3D) inputs and achieves real-time operation. For 2D the prompt *"gravel road"* was used, 3D visualises *"trees"* in green, *"shrubbery"* in purple, *"grass"* in orange, and *"stone"* in red.

1 Introduction

Understanding the open world through semantics is a key challenge for robotics. Vision-Language Models (VLMs), that ground vision with language, have recently been shown to effectively provide semantics for mapping to facilitate task planning and navigation [1, 2]. However, open-vocabulary semantic mapping methods [3, 4, 5] rely on segmentation priors from general-purpose models to reason about the environment. These models are trained for object-centric knowledge retrieval, therefore, they are effective for segmenting structured settings with salient objects. However, segmentation fails in unstructured outdoor environments, such as forests or off-road paths, see Figure 1. Unstructured, texture-rich classes relevant to outdoor robotics, such as roads and grass, are underrepresented in typical open-vocabulary image-text pair-based datasets and are often inconsistently

labelled. Visual ambiguities and indistinct boundaries, such as overlaps between gravel and grass, further complicate the task for segmentation models, which leads to imprecise segmentation masks.

In order to obtain robust semantic segmentation in unstructured outdoor environments, we introduce **OTAS**, an **O**pen-Language **T**oken **A**lignment Method for Outdoor **S**egmentation. Instead of relying on language semantics for segmentation, we cluster tokens based on visual prototypes derived from self-supervised pretrained vision models. Language grounding is obtained through semantic and spatial alignment over token clusters, alleviating the need for linear probing or rendering. Optionally, multiple observations can be aligned to obtain a language-embedded reconstruction with geometric consistency. Hence, OTAS is not subject to the object-centric bias learned by general-purpose segmentation models, despite also performing zero-shot inference. We demonstrate real-time inference on GPU, while improving the state of the art for segmentation on Off-Road Freespace Detection (ORFD) [6] and TartanAir [7]. Additional experiments on robot data demonstrate the advantage of OTAS for language-embedded reconstruction of unstructured outdoors in comparison to volumetric rendering with LERF [8] and Feature Splatting [5].

2 Related Work

Table 1: **Comparison of Semantic Reconstruction Methods.** Assuming 15 fps as real-time—typical for low-dynamic settings like forests and agriculture—only OpenFusion and OTAS meet this threshold. Only LERF and OTAS use non-object-centric language maps. OTAS uniquely supports semantic segmentation in both 2D and 3D natively.

Method	Foundation Model	Real-Time	Zero-Shot	3D	2D	Repr.	Not Object-Centric
LERF [8]	OpenCLIP [9] DINOv2 [10]	✗	✗	✓	✗	NeRF	✓
Feature Splatting[5]	CLIP [11] DINOv2 [10] SAM [12]	✗	✗	✓	✗	Gaussian Splatting	✗
ConceptGraphs [3]	OpenCLIP [9] SAM [12]	✗	✓	✓	✗	Points	✗
OpenFusion [4]	SEEM [13]	✓	✓	✓	✗	TSDF	✗
OTAS (ours)	CLIP [11] DINOv2 [10] SAM2 [14]	✓	✓	✓	✓	Points	✓

VLMs ground vision in language by encoding a joint feature space, typically extracting one feature per image or patch [11, 9]. Many robotic tasks, however, require fine-grained spatial relationships. This motivates mapping VLM features to queryable semantic maps [4, 3, 8, 5, 15, 16].

Early VLM-based navigation approaches detect objects, extract VLM features per instance, and ground them on 2D occupancy grids (e.g., VLMaps [1], VLFM [17]) by interpolating features spatially. They rely on general-purpose detection or segmentation models, which introduce an object-centric prior into feature extraction [18]. This paradigm has been extended to 3D. OpenFusion [4] fuses SEEM [13] features into a 3D semantic Simultaneous Localisation and Mapping (SLAM) map. Similarly, ConceptGraphs [3] uses SAM [12] masks and OpenCLIP [9] features, projected to 3D and fused via geometric and semantic similarity. While effective indoors, all retain object-centric biases from their segmentation models. An alternative direction is to reconstruct language-grounded feature fields. Feature Splatting [5] retains object priors since it uses SAM for generating segmentation masks. LERF [8] avoids object priors by extracting multiscale OpenCLIP features, yielding dense, non-object-centric feature maps refined via neural rendering. Both rely on geometric consistency across views. Using neural scene representations, such as LERF or Feature Splatting, requires rendering, resulting in slow scene-specific training and making them neither zero-shot nor real-time capable.

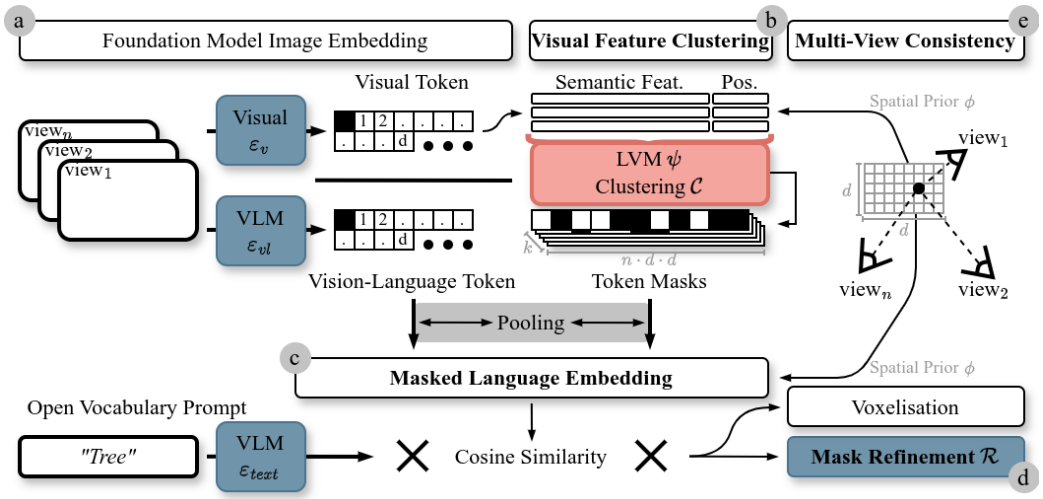

Figure 2: **Method Overview.** a) OTAS encodes input views using frozen encoders. b) Patch tokens of the visual encoder are reduced and clustered to obtain semantic masks. c) The masks are pooled with normalised patch tokens of a vision-language encoder for natural language grounding. d) A frozen mask refinement network projects semantic similarity to prompts to pixel-level. e) Clustering and pooling are optionally conditioned on environment geometry through projection.

Table 1 compares state-of-the-art semantic reconstruction methods for outdoor robot navigation relevance. Key requirements include real-time performance for robot control, zero-shot applicability to new environments, and avoidance of object-centric priors for accurate segmentation of non-salient objects. OTAS is the only method meeting all criteria.

3 Method

VLMs, such as Grounded SAM and SEEM are biased towards object-centric knowledge retrieval [19]. This becomes especially problematic in the unstructured environments of outdoor robotics, where the semantic classes of interest fall outside the a priori encoded object-centric knowledge. Examples of such classes are road, woods, and shrubbery, which, however, are highly relevant to mobile outdoor robotics.

Self-supervised pretrained vision foundation models, such as DINOv2 [10], do not have this limitation, since they are not trained directly on segmentation tasks. Their training process results in an emergent semantic organization of the feature space, where semantically similar classes are embedded adjacently. Hence, we disentangle the open-vocabulary semantic segmentation by using DINOv2 for coarse zero-shot semantic clustering, followed by natural language grounding by pooling over CLIP's vision-language embeddings. Input views are embedded by the frozen vision and vision language encoders, see Figure 2 (a). Output tokens of the vision encoder are clustered to obtain semantic structures (b), and aligned with vision language tokens to obtain language grounding (c). The language-grounded semantic clusters are used as priors for zero-shot upscaling to pixel level (d) [14]. Optional spatial regularisation of steps b and c increase geometric consistency and allow multi-view reconstruction and segmentation (e).

3.1 Visual Feature Clustering

Given a monocular input image $I \in \mathbb{R}^{H \times W \times 3}$, our goal is to generate a semantic segmentation mask guided by both vision and language. The input image is first processed by a frozen vision encoder $\mathcal{E}_v$ to produce a coarse spatial feature map $F_v = \mathcal{E}_v(I) \in \mathbb{R}^{H' \times W' \times C_v}$. To align vision with language, F_v is interpolated to a shared feature dimension d using bilinear interpolation. The interpolated features are then flattened and L2 normalised, denoted by $\mathbf{f}_v$. The flattened feature map

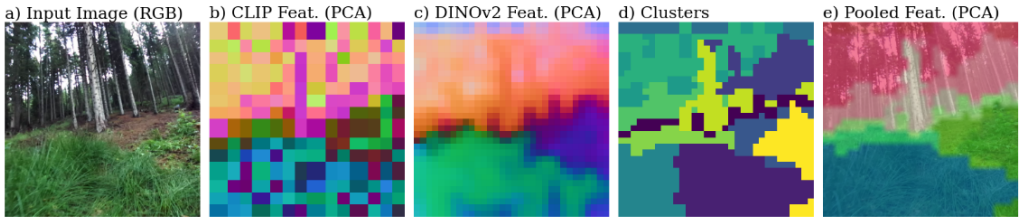

Figure 3: **Feature comparison.** CLIP (b) [11] features include view-dependent noise that is detrimental to segmentation accuracy [5]. We achieve regularisation in non-object centric environments by extracting visual prototypes from DINOv2 (c) [10], with k-Means clustering (d) and language grounding via feature pooling (e).

is decorrelated and reduced in dimensionality using a latent variable model (LVM) ψ, resulting in $\hat{\mathbf{f}}_{LVM} = \psi(\hat{\mathbf{f}}_v) \in \mathbb{R}^{d \cdot d \times C_r}$, where the reduced feature dimension C_r is a hyperparameter.

Subsequently, a clustering model $\mathcal{C}$ is applied to the flattened feature map $\hat{\mathbf{f}}_v$ to derive k clusters, that constitute mixtures of visual tokens, referred to as visual prototypes. The affiliation of each data point to a cluster is denoted by $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_{d \cdot d}\}, C_j \in \{1, ..., k\} \forall j$, representing the assignment of the latent representations $\hat{\mathbf{f}}_{LVM}$ to a visual prototype. The clusters are interpreted as a set of k binary masks $\mathcal{M} = \{\mathcal{M}_1, ..., \mathcal{M}_k\}$, where each mask $\mathcal{M}_i \in \{0, 1\}^{n \times d \times d}$ corresponds to the shared feature dimension d across n input images.

3.2 Masked Language Embedding

DINOv2 embeddings are not correlated with semantics such as language. An intuitive way to retrieve semantic categories is linear probing. This, however, requires annotated data in the target domain. Instead, we use a vision-language encoder $\mathcal{E}_{vl}$ to produce language-grounded tokens and align them with the visual tokens, resulting in $F_{vl} = \mathcal{E}_{vl}(I) \in \mathbb{R}^{H_{vl} \times W_{vl} \times C_{vl}}$. These tokens are subsequently interpolated to match d using nearest neighbour interpolation ($\mathcal{U}_{nn}$): $F_{vl}^{shared} = \mathcal{U}_{nn}(F_{vl}) \in \mathbb{R}^{d \times d \times C_{vl}}$. We adopt masked average pooling (MAP) to address token alignment, following [5], who showed its regularising effect on VLMs. Unlike prior work, we apply MAP over coarse feature maps in the shared embedding space rather than at pixel level. MAP computes the mean language feature vector for each mask. This is done per image, also in the case of multi-view inputs.

$$F_{pooled}(x, y) = \frac{1}{|M_c|} \sum_{(x,y) \in M_c} F_{vl}^{shared}(x, y) \tag{1}$$

Since each patch is only assigned to a single mask in $\mathcal{M}$, the resulting F_{pooled} is a feature map of shape $d \times d \times C_{vl}$. F_{pooled} represents a language-grounded image embedding, regularised by the token mask structure (see Figure 3). Ultimately, pooled features are normalised using the L2 norm.

A frozen text encoder $\mathcal{E}_{text}$ maps text prompts to the vision-language feature dimension $F_{text} = \mathcal{E}_{text}(t) \in \mathbb{R}^{C_{vl}}$. Cosine similarity between F_{text} and each feature in F_{pooled} produces a similarity map of shape $d \times d$. As done by [8, 5], $\mathcal{E}_{text}$ and the similarity computation are applied to a set of positive prompts t^+ and negative prompts t^-, indicating target and undesired concepts, respectively, resulting in the combined similarity map $\mathcal{S}_{combined}$.

$$\mathcal{S}_{combined} = \sum_{t \in t^+} \mathcal{S}(t, F_{pooled}) - \sum_{t \in t^-} \mathcal{S}(t, F_{pooled}) \tag{2}$$

3.3 Mask Refinement

We use the similarity map as a language-grounded prior to obtain a binary pixel-level segmentation mask M. Depending on the used encoders and interpolation to the shared feature resolution d, the similarity map resolution will be lower than the input image resolution. Typically, the similarity map

is at 1/8th or 1/16th of the input image resolution. In order to refine the coarse mask we employ a frozen mask refinement network $\mathcal{R}$ that takes the image I and the similarity map $\mathcal{S}_{combined}$ as input and outputs the final high-resolution segmentation mask.

$$M = \mathcal{R}(I, \mathcal{U}_{bl}(\mathcal{S}_{combined})) \in \{0,1\}^{H \times W} \tag{3}$$

3.4 Multi-View Consistency

To expand OTAS to the multi-view case, information is aggregated over multiple views using the depth map $D \in \mathbb{R}^{H \times W}$ and camera pose $T \in SE(3)$ associated with each frame. During image embedding, D is projected to 3D points $P \in \mathbb{R}^{N \times 3}$. Median depth $\tilde{D}$ is sampled in each grid of size $d \times d$ to align the 3D points with the vision and vision language features. Using camera intrinsics K, 3D points P are projected to the image plane via $P = \pi(\tilde{D}, K) \in \mathbb{R}^{d \cdot d \times 3}$. A mapping ϕ_i tracks the relationship between 3D points P_i and patch indices (i, j). The points are transformed to a global coordinate frame using camera poses $\{T_1, ..., T_n\}$ to construct $P_{global} = \bigcup_{i=1}^{n} T_i P_i$.

Spatially Conditioned Clustering. The global point positions and relationship ϕ allow conditioning the visual feature clustering by concatenating semantic features F_v^{shared} with 3D coordinates P_{global}. This yields a combined feature map $F_{spatial} \in \mathbb{R}^{d \cdot d \times (C_v + 3)}$ that replaces F_v^{shared} as input for the LVM, where each feature vector $F_{spatial}(i, j)$ contains both semantic and spatial information for the corresponding point p.

Spatially Conditioned Pooling. After pooling the visual and vision-language features for each input view separately, each F_{pooled} is projected on the global point cloud P_{global} using the relationship ϕ, resulting in a spatial 3D feature volume $P_{semantic} \in \mathbb{R}^{d \cdot d \times (C_{vl} + 3)}$ where $P_{semantic} = \text{concat}(F_{pooled}(i, j), P_{global}(p)) \mid p \in P_{global}, (i, j) = \phi_i(p)$. The feature volume consists of keypoint position and language-grounded feature embedding pairs. Knowing the keypoint position, the feature volume is downsampled using a configurable voxel-size v. During downsampling, all pooled features in a voxel are linearly interpolated to further condition the language-embeddings with spatial context, where $\hat{P}_{semantic} = (\frac{1}{|V_k|} \sum_{(f,p) \in V_k} f)$ with $V_k = \{(f, p) \in P_{semantic} \mid \lfloor \frac{p}{v} \rfloor = k\}$. $\hat{P}_{semantic}$ describes a language-queryable 3D occupancy grid directly usable for robotic applications such as obstacle avoidance and goal-based navigation.

4 Experiments

Datasets and Metrics. Monocular semantic segmentation is evaluated on the Off-Road Freespace Detection Dataset (ORFD) [6]. ORFD aims to identify traversable road types in the outdoors, such as gravel, dirt and sand. 3D feature reconstruction is evaluated on TartanAir [7], a large-scale, photorealistic synthetic dataset for visual SLAM and robot navigation. Since TartanAir does not provide 3D ground truth labels, 2D labels are projected onto the point clouds via majority vote over its 5 nearest neighbours [20]. In order to evaluate unstructured outdoor segmentation, we evaluate segmenting vegetation, labels 152 and 109. Following previous work [21], Intersection over Union (IoU), F-score (Fsc), Precision (Pre), and Recall (Rec) are evaluated for all quantitative experiments. Qualitative results are presented on real-world teleoperated robotic trials in the alps [22].

Implementation Details. OTAS is provided in three configurations. All models use CLIP ViT-B-16 [11, 23] and DINOv2 ViT-S-14 with 4 registers [10, 24]. *OTAS Small* uses a shared feature dimension of $d = 16$ and SAM2.1 Hiera-T [14] for mask refinement. *OTAS Large* uses $d = 32$ and SAM2.1 Hiera-L. *OTAS Spatial* uses $d = 64$, a voxel-size of $v = 0.5m$, and no mask refinement, as segmentations are regularised geometrically. All models use GPU-accelerated Principal Component Analysis (PCA) for ψ and k-Means for C. Evaluations are done on an Intel i7-12700 CPU and NVIDIA 4070 TI Super GPU.

Table 2: **Semantic Segmentation on ORFD.** We include the current state of the art in fine-tuned off-road segmentation methods as well as other zero-shot segmentation methods that serve as the baseline for language-grounded semantic scene representations. The † indicates results optioned from the reimplementation by [25].

	Method	IoU(%)	Fsc(%)	Pre(%)	Rec(%)
Fine-tuned	OFF-Net [6]	82.30	90.30	86.60	94.30
	RTFNet† [26]	90.70	95.10	93.80	<u>96.50</u>
	RoadFormer [25]	92.51	96.11	95.08	**97.17**
	M2F2-Net [27]	93.10	96.40	97.30	95.50
	NAIFNet [28]	<u>94.10</u>	<u>97.00</u>	97.50	96.40
Zero-Shot	SEEM [13]	51.31	59.12	61.44	60.93
	Grounded SAM [29]	90.49	94.13	95.12	93.32
	Grounded SAM-2 [30]	93.32	96.38	<u>97.73</u>	95.38
	OTAS Small (**Ours**)	91.72	95.59	96.93	94.58
	OTAS Large (**Ours**)	**94.34**	**97.05**	**97.83**	96.39

4.1 2D Semantic Segmentation on ORFD

This section compares OTAS to the state of the art for fine-tuned and open-vocabulary 2D semantic segmentation. For open-vocabulary, we report Grounded SAM [29] and SEEM [13], since these are the models used by Concept Graphs [3] and OpenFusion [4] respectively. Table 2 presents results on ORFD. OTAS achieves the highest IoU, Fsc and precision among fine-tuned and zero-shot methods. OTAS reports the highest recall among zero-shot methods. Yet, the segmentation recall of the fine-tuned RoadFormer marginally improves over OTAS. Interestingly, this phenomenon can be observed for all zero-shot methods. They exhibit lower recall as compared to fine-tuned methods. This is a consequence of the lack of dense supervision for specific classes and the necessity to generalise over broad, noisy semantics, whereas fine-tuned models directly optimise for segmenting the specific classes, including dataset characteristics like annotation errors and noise.

4.2 3D Semantic Mapping on TartanAir

Semantic mapping is evaluated against the state of the art for scene reconstruction: Concept Graphs [3] and OpenFusion [4]. Since both methods do not directly provide semantic labels, but rather language-grounded point clouds, we threshold using the same language queries as for OTAS.

Table 3 presents 3D segmentation results on outdoor scenes of TartanAir using the first annotated trajectory. OTAS improves all evaluated metrics over OpenFusion and ConceptGraphs on Amusement, Gascola, and Seasonsforest. Especially in environments with barely any discrete objects, such as Gascola, the margin for improvement is huge, reaching up to 151% on IoU. The lower contrast reduces segmentation quality of object-centric open-vocabulary segmentation, highlighting the advantages of OTAS for outdoor robotics. We observe that ConceptGraphs performs better in snowy scenes of Seasonsforest Winter. This is likely due to the high contrast between objects and the uniform snow, which enhances object boundaries and thus benefits object-centric methods.

4.3 Feature Reconstruction in Alpine Environment

This section directly compares OTAS to LERF [8] and Feature Splatting [5] for semantic reconstruction in the foothills of the Alps. While neither zero-shot nor real-time, these methods excel at language-embedded reconstruction for robotics. We use a ROS bagfile of RoboNav[22]. This allows for reproducible testing on real sensor data since the bagfile captures the full sensor and actuation context of the robot in representative environments.

Figure 4 shows language-embedded reconstructions of a challenging forest scene featuring dense vegetation and different ground types, such as grass, dirt and puddles. LERF and Feature Splatting require highly accurate camera poses for reconstructing scenes with differential rendering. Usually,

Table 3: **Semantic Mapping on TartanAir.** All methods reconstruct a language-grounded point cloud given known camera poses. Sec denotes total reconstruction time excluding evaluation. We compare per point segmentation performance in identifying vegetation (labels 152 and 109).

	Amusement					Gascola				
	IoU	Fsc	Pre	Rec	Sec↓	IoU	Fsc	Pre	Rec	Sec↓
OpenFusion [4]	23.13	37.09	39.17	37.86	55	10.24	18.37	18.23	20.36	52
ConceptGraphs [3]	34.86	46.15	47.00	48.17	2201	30.68	38.03	30.68	50.00	333
OTAS Spatial (**Ours**)	**47.11**	**64.04**	**65.16**	**65.48**	**22**	**67.87**	**80.27**	**79.23**	**81.73**	**12**
	Seasonsforest					**Seasonsforest Winter**				
	IoU	Fsc	Pre	Rec	Sec↓	IoU	Fsc	Pre	Rec	Sec↓
OpenFusion [4]	25.09	35.18	47.38	39.07	53	22.16	36.01	39.37	40.84	103
ConceptGraphs [3]	17.39	28.96	51.06	52.25	151	36.48	53.33	54.26	54.49	479
OTAS Spatial (**Ours**)	**43.63**	**57.23**	**57.09**	**57.42**	**10**	**39.61**	**55.13**	**56.22**	**55.33**	**18**

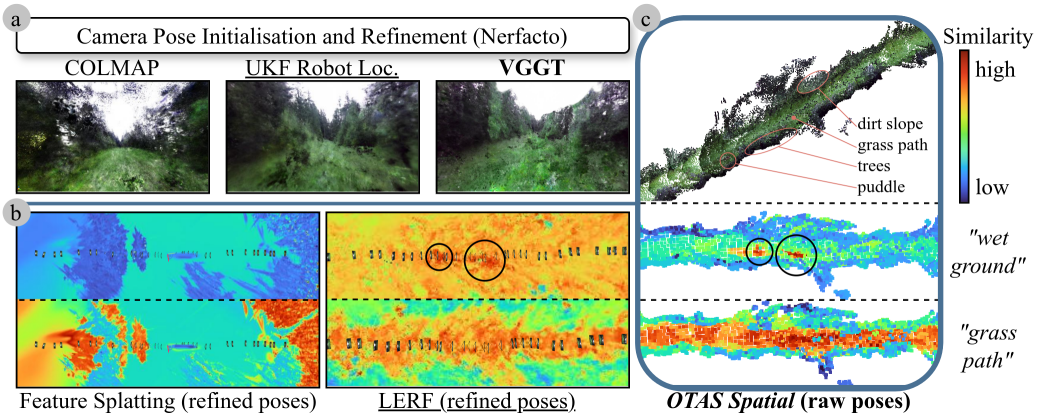

Figure 4: **Ground Comprehension in Alpine Environment.** Language-embedded reconstruction requires accurate camera poses. a) Reconstruction obtained using COLMAP, UKF Robot Localisation, and VGGT. All poses are refined using Nerfacto. b) Semantic similarity of Feature Splatting and LERF to prompts. c) Semantic reconstruction and prompt similarity of *OTAS Spatial*.

Structure from Motion, like COLMAP [31], is used for camera pose initialisation. However, due to the cluttered, highly-textured scene, COLMAP, UKF Robot Localisation [32], and VGGT [33] fail to provide poses with sufficient accuracy, see Figure 4 a). Hence, camera poses are initialised using VGGT, scaled using metric depth estimation [34], and refined using Nerfacto [35]. Even with pose refinement, Feature Splatting fails to properly reconstruct the ground. LERF correctly locates the grass-path itself and puddles (black circles) thanks to non-object-centric language grounding, Figure 4 b). However, it is computationally intensive with ≈40 minutes for this scene. OTAS shows a geometrically accurate reconstruction with detailed language similarity at ≈1.3 seconds (without pose initialisation), Figure 4 c). Additional experiments can be found in the supplementary material.

4.4 Ablations

Model Size and Inference Time (2D). We provide multiple model configurations for different compute capabilities. Table 4 presents their speed-accuracy trade-off on GPU and CPU. No mask refinement refers to normalising the similarity map $\mathcal{S}_{combined}$ to $[0, 1]$ and binary thresholding. No alignment with mask refinement represents $\mathcal{R}(I, \mathcal{U}_{bl}(\mathcal{S}(F_{text}, F_{vl})))$. Small and Large model configurations are outlined in Section 4. Mask refinement adds ≈50% to runtime on GPU. *OTAS Small* without refinement runs at real-time (assuming 15 fps). While the performance difference between of the refined *OTAS Small* and *Large* is insignificant, the larger version still runs at 5 fps on GPU.

Table 4: **Influence of Model Size.** Comparison of accuracy, memory and fps of OTAS on ORFD.

Model	Ref.	IoU (%)	Fsc (%)	Pre (%)	Rec (%)	Mem. (GB)	fps (s^{-1})
No align. (GPU)	no	68.25	80.46	79.57	82.48	1.6	≈25
	yes	75.48	84.54	92.90	82.03	2.4	≈13
Small (GPU)	no	84.71	91.35	91.12	92.84	1.6	≈17
	yes	91.72	95.59	96.93	94.58	2.4	≈11
Small (CPU)	no	84.80	91.41	91.20	92.87	-	≈1.6
	yes	91.71	95.58	96.93	94.57	-	≈0.38
Large (GPU)	yes	94.34	97.05	97.83	96.39	3.5	≈5

Reduction and Clustering Methods. Table 5 examines the choice of LVM (ψ) and clustering model (C). This comparison shows that k-Means clustering leads to the cleanest segmentation results, with PCA and Kernel-PCA being equally suitable for dimensionality reduction.

Number of Clusters and Components. The top of Figure 5 shows an ablation of PCA components (C_r) and k-Means clusters (k) on a 20% split of ORFD. Positive prompts are *gravel, road, dirt* and negative prompts are *sky, grass, forest*. The denoted score is an average of the IoU, F1 score, precision and recall. The highest score is achieved with $C_r = 4$ and $k = 4$.

Table 5: **Dimensionality Reduction and Clustering Algorithms.** Score refers to an average of IoU, Fsc, Pre, and Rec of *OTAS Small* with mask refinement on ORFD.

Clustering (C)	**Reduction** (ψ)	Score
GMM	PCA (GPU)	0.9390
	KPCA	0.9395
HDBSCAN	PCA	0.9394
	KPCA	0.9394
	PCA (GPU)	0.9394
k-Means (GPU)	PCA	0.9427
	PCA (GPU)	0.9424
k-Means	PCA	0.9466
	KPCA	**0.9467**
	PCA (GPU)	**0.9467**

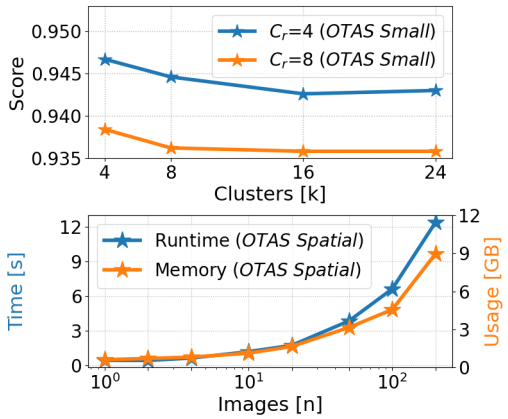

Figure 5: **Clusters, Runtime and Memory.** Top presents the number of k-Means clusters (k). Bottom the runtime and the memory usage.

Model Scalability (3D). Bottom of Figure 5 shows the time requirements for reconstruction with *OTAS Spatial* on TartanAir Seasonsforest Winter in blue, and the memory usage in orange. Both time and space complexity are comparably low to the state of the art. At 10 views, both time and memory usage are marginally above 1 second and Gigabyte respectively. Using 200 views takes 12.42 seconds and requires 9.97 Gigabytes of GPU memory.

5 Conclusion

This work addressed open-vocabulary segmentation in unstructured outdoor environments. We introduce OTAS, an open-vocabulary segmentation method that aligns semantic tokens across single and multiple views to reconstruct a geometrically consistent feature field. It aligns the output tokens of a pretrained vision model to a language embedding by clustering semantically similar tokens through unsupervised learning and pooling. Results show a minor improvement over open-vocabulary and fine-tuned baselines on the ORDF dataset, a significant improvement over the state of the art on TartanAir, and robust applicability to real-world robotic tasks. Future work will investigate employing our semantic maps for outdoor navigation, e.g., through costmap modification [16] or with learned policies [36].

6 Limitations

Feature Map Resolution. Although OTAS demonstrates strong segmentation capabilities in unstructured outdoor environments, it can fail to detect all objects in geometrically complex and highly textured scenes. This is because clustering and therefore VLM model regularisation are limited by DINOv2's patch grid size. This limits the density of features and masks that can be extracted from each image. As a result, the model may miss small objects or fine-grained details. While mask refinement in 2D and spatial pooling in 3D partially mitigate this issue, increasing the number of patch tokens per image would improve the detection of small objects and fine structures. Exploring foundation models with finer patch granularity may be a promising direction for future work.

Spatial Consistency and Depth Maps. To achieve real-time performance, OTAS does not perform explicit feature matching across multiple views. Instead, it relies on tracking spatial relationships between tokens with a projection function, making it highly dependent on the quality of camera poses and depth maps to align features in 3D and across multiple views. While this is often available in robotic applications, for example from RGB-D sensors or SLAM, it may not always be accessible or may suffer from limited accuracy. If depth or pose information is unavailable, initialisation using structure-from-motion is possible, but the resulting depth maps may be less accurate. Camera pose refinement and rejecting outliers of the depth map could improve alignment. However, these operations are computationally expensive and may compromise the model's real-time capability.

Baseline Comparison. Comparing OTAS to RayFronts [20] on TartanAir2 would be highly interesting and informative, however, the code and the dataset are not publicly available at the time of writing.

Acknowledgments

This work was supported by the city of Vienna (MA23 – Economic Affairs, Labour and Statistics) through the project *Stadt Wien Kompetenzteam für Drohnentechnik in der Fachhochschulausbildung* (DrohnFH, MA23 project 35-02).

References

[1] C. Huang, O. Mees, A. Zeng, and W. Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615, 2023.

[2] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler. Open-vocabulary queryable scene representations for real world planning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11509–11522, 2023.

[3] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024.

[4] K. Yamazaki, T. Hanyu, K. Vo, T. Pham, M. Tran, G. Doretto, A. Nguyen, and N. Le. Openfusion: Real-time open-vocabulary 3d mapping and queryable scene representation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9411–9417. IEEE, 2024.

[5] R.-Z. Qiu, G. Yang, W. Zeng, and X. Wang. Language-driven physics-based scene synthesis and editing via feature splatting. In *European Conference on Computer Vision (ECCV)*, pages 368–383, 2024.

[6] C. Min, W. Jiang, D. Zhao, J. Xu, L. Xiao, Y. Nie, and B. Dai. Orfd: A dataset and benchmark for off-road freespace detection. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2532–2538, 2022.

[7] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916, 2020.

[8] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik. Lerf: Language embedded radiance fields. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19672–19682, 2023.

[9] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, June 2023.

[10] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

[13] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee. Segment everything everywhere all at once. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 19769–19782. Curran Associates, Inc., 2023.

[14] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[15] N. M. M. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv preprint arXiv:2210.05663*, 2022.

[16] R.-Z. Qiu, Y. Hu, Y. Song, G. Yang, Y. Fu, J. Ye, J. Mu, R. Yang, N. Atanasov, S. Scherer, and X. Wang. Learning generalizable feature fields for mobile manipulation. *arXiv preprint arXiv:2403.07563*, 2024.

[17] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 42–48, 2024.

[18] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[19] Y. Zhang, N. Konz, K. Kramer, and M. A. Mazurowski. Quantifying the limits of segmentation foundation models: Modeling challenges in segmenting tree-like and low-contrast objects. *arXiv preprint arXiv:2412.04243*, 2025.

[20] O. Alama, A. Bhattacharya, H. He, S. Kim, Y. Qiu, W. Wang, C. Ho, N. Keetha, and S. Scherer. Rayfronts: Open-set semantic ray frontiers for online scene understanding and exploration. *arXiv preprint arXiv:2504.06994*, 2025.

[21] C. Min, S. Si, X. Wang, H. Xue, W. Jiang, Y. Liu, J. Wang, Q. Zhu, Q. Zhu, L. Luo, F. Kong, J. Miao, X. Cai, S. An, W. Li, J. Mei, T. Sun, H. Zhai, Q. Liu, F. Zhao, L. Chen, S. Wang, E. Shang, L. Shang, K. Zhao, F. Li, H. Fu, L. Jin, J. Zhao, F. Mao, Z. Xiao, C. Li, B. Dai, D. Zhao, L. Xiao, Y. Nie, Y. Hu, and X. Li. Autonomous driving in unstructured environments: How far have we come?, radiological, and nuclear disaster response. *arXiv preprint arXiv:2410.07701*, 2024.

[22] M. Eder, R. Prinz, F. Schöggl, and G. Steinbauer-Wagner. Traversability analysis for off-road environments using locomotion experiments and earth observation data. *Robotics and Autonomous Systems*, 168:104494, 2023.

[23] C. Zhou, C. C. Loy, and B. Dai. Extract free dense labels from clip. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, pages 696–712, Cham, 2022. Springer Nature Switzerland.

[24] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

[25] J. Li, Y. Zhang, P. Yun, G. Zhou, Q. Chen, and R. Fan. Roadformer: Duplex transformer for rgb-normal semantic road scene parsing. *IEEE Transactions on Intelligent Vehicles*, 9(7): 5163–5172, 2024.

[26] Y. Sun, W. Zuo, and M. Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019.

[27] H. Ye, J. Mei, and Y. Hu. M2f2-net: Multi-modal feature fusion for unstructured off-road freespace detection. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–7, 2023.

[28] Y. Lv, Z. Liu, G. Li, and X. Chang. Noise-aware intermediary fusion network for off-road freespace detection. *IEEE Transactions on Intelligent Vehicles*, pages 1–11, 2024.

[29] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[30] IDEA-Research. Grounded-sam-2: Ground and track anything in videos with grounding dino, florence-2, and sam 2. https://github.com/IDEA-Research/Grounded-SAM-2, 2025. Accessed: 2025-04-30.

[31] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016.

[32] T. Moore and D. Stouch. A generalized extended kalman filter implementation for the robot operating system. In *Intelligent Autonomous Systems 13: Proceedings of the 13th International Conference IAS-13*, pages 335–348. Springer, 2016.

[33] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[34] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. . *arXiv preprint arXiv:2302.12288*, 2023.

[35] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, 2023.

[36] P. Maheshwari, W. Wang, S. Triest, M. Sivaprakasam, S. Aich, J. G. R. III, J. M. Gregory, and S. Scherer. Piaug – physics informed augmentation for learning vehicle dynamics for off-road navigation. *arXiv preprint arXiv:2311.00815*, 2023.

Supplementary Material for OTAS: Open-vocabulary Token Alignment for Outdoor Segmentation

A Supplementary Feature Reconstruction Results

Figure 6 presents *OTAS Spatial* reconstructions on RoboNav [22]. The leftmost column shows an image of the scenes. The second column to the left shows the RGB reconstruction of P_{global}. Language-grounded semantic information is visualised using PCA over $P_{Semantic}$ in the third column, and a 3D segmentation using common labels in outdoor robotics (rightmost column). Segmentations are obtained by thresholding the similarity between $\hat{P}_{Semantic}$ and a set of text prompts for each scene. *Grass* is depicted in light green, *trees* and *shrubbery* in dark green, the *duff layer* (comprising dead leaves and small twigs) in brown, *gravel* in light grey, *road* in dark grey, and *puddle* and *water* in blue.

Figure 6a) depicts a clearing in a forest. The primary challenge in this scene is to differentiate between tall grass, vegetation flattened by repeated vehicular traffic, and shrubbery located in the centre of the scene. The geometric reconstruction shows high visual similarity of tall grass and shrubbery, yet they have distinct implications for traversability. In the PCA visualisation, trees and shrubbery are indicated as semantically similar (red), and are clearly separated from grass (green). At the boundaries between semantic classes, the dark black regions indicate ambiguous semantic associations between tall grass and shrubbery in the raw feature reconstruction. Nevertheless, when features are queried using open-vocabulary prompts, OTAS successfully segments the narrow grass path for traversing the shrubbery.

Figure 6b) illustrates a crossroad of a gravel road and a field track. The primary challenge is in correctly identifying both roads for obtaining information about traversability. The road on the right is a forest road composed of gravel, while the left path is a field track. The PCA visualisation shows that the road and the path are semantically similar (pink and orange) in OTAS' feature reconstruction. Through the open-vocabulary prompt *"road"*, OTAS correctly identifies both as traversable areas (grey).

Figure 6c) presents a forest road under varying lighting conditions, a common scenario in outdoor robotics. OTAS clearly distinguishes the road (grey) from vegetation (dark green) and grass (light green).

Figure 6d) shows a steep, highly unstructured area containing grass, trees, bushes and duff. The primary challenge in this scene is to correctly distinguish between different ground types, namely tall grass, shrubbery, tree stumps, and duff. This is particularly challenging as these ground types blend into each other, leading to fuzzy semantic boundaries. Furthermore, the ground is cluttered. The PCA reconstruction demonstrates that in the language-grounded embeddings, the ground types can be clearly distinguished from each other. Shrubbery and tree stumps are dark green, tall grass is red, and duff is yellow. The PCA visualisation also illustrates how these ground types mix with each other in the raw language-grounded embeddings, depicting how they blend into each other in reality. When prompted, the resulting segmentation distinguishes between *grass* (light green), *shrubs* and *trees* (dark green), and *duff* (brown).

Figure 6e) shows a grass path with puddles. The primary challenge is to correctly distinguish the wet ground and puddles from the traversable grass. In the PCA visualisation, the *puddles* (blue) are clearly separated from the *grass* (red). However, in the prompt-based segmentation, the *puddles* (blue) are not as clearly defined, indicating the need for a smaller voxel size.

Figure 6f) shows an asphalt road going downhill. The street smoothly transitions to the gravel strip on the left, which itself blends into grass and vegetation. This scene presents very unclear semantic boundaries. OTAS' PCA visualisation shows that the ground types are clearly distinguishable in the semantics. Furthermore, the blending of ground types is also apparent in the raw language-grounded embeddings. *Asphalt* (purple) mixes with *gravel* (red) and *grass* (dark red). In the prompt-based segmentation, OTAS determines clear boundaries between these ground types.

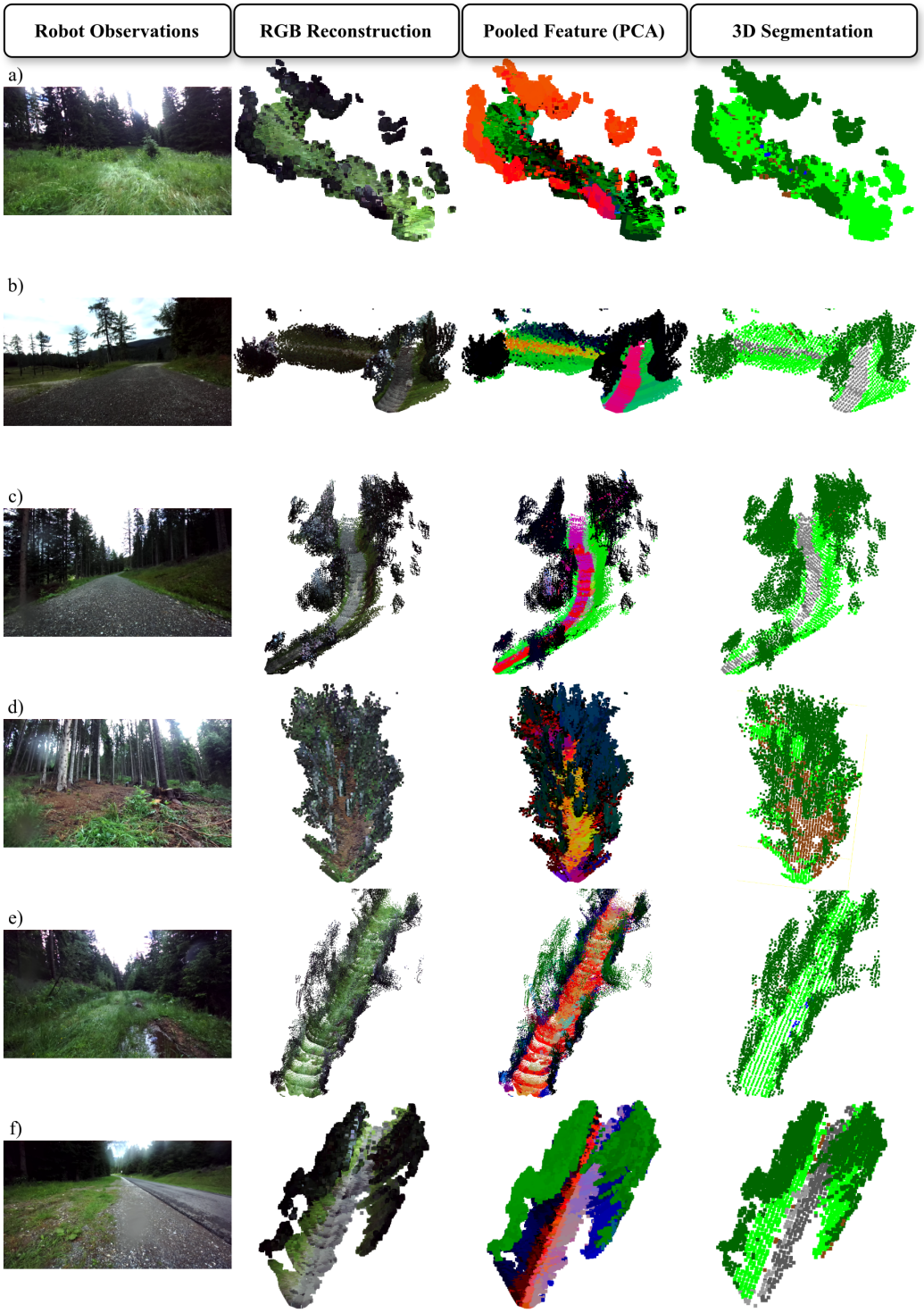

Figure 6: **Alpine 3D Segmentation.** LTR: Input image (of n), RGB point cloud P_{global}, PCA over $P_{Semantic}$ and segmentation over $\hat{P}_{Semantic}$. Scenes are a) open field with shrubbery, b) cross-road between a road and field track, c) forest road, d) steep area including trees, bushes and duff, e) grass path with puddles, and f) asphalt street. Semantic classes *trees*, *shrubs*, and *bushes* are dark green, *grass* is light green, *road* and *gravel* are dark and light grey, *duff* is brown, and *puddles* are blue.

B Supplementary Experimental Detail

B.1 Datasets

Off-Road Freespace Detection (ORFD) [6] ORFD experiments use $k = 4$ clusters and $C_r = 4$ components, with positive prompts *"gravel", "road", "dirt"* and negative prompts *"sky", "grass", "forest"*. For OTAS without mask refinement, a similarity threshold of 0.5 is considered a positive label. SEEM [13] and Grounded-SAM [29] use the same positive prompts as OTAS. Negative prompts are not supported by them. Grounded-SAM 2 [30] neither supports negative nor multiple prompt input. Therefore, *"road."* was used as it achieved the best results.

TartanAir [7] The TartanAir experiments use $k = 30$ clusters and $C_r = 30$ components, with positive prompts *"tree", "bush", "vegetation"*, and negative prompts *"sky", "stone", "object"*, again with a similarity threshold of 0.5. Invalid, infinite, and depth values over 150 metres are discarded. For sequences too large to fit into 16GB of GPU VRAM, only every third image is used for reconstruction. This was necessary for OpenFusion on all sequences, and for *OTAS Spatial* on Amusement and Gascola. The baselines OpenFusion [4] and ConceptGraphs [3] are configured to output point cloud reconstructions with per-point CLIP [11] features. These point clouds are prompted identically to the OTAS reconstructions. For ConceptGraphs, we used the configuration of the original paper based on SAM.

RoboNav [22] COLMAP [31] with the unaltered camera stream and default matcher parameters failed to provide camera poses with sufficient accuracy for obtaining reconstructions. Hence, input images were pre-processed with a sharpening kernel, brightness and contrast adjustment, and fast non-local means denoising. This enabled pose initialisation using COLMAP's sequential matcher with relaxed parameters: overlap (10), quadratic overlap disabled, loop detection every 10 frames, a reduced loop detection window (50 images), fewer nearest neighbours (5), and a reduced number of checks (256). UKF Robot localisation fuses wheel odometry with GNSS using an Unscented Kalman Filter.

B.2 Feature Reconstruction Baselines

VGGT [33] does not provide metric depth. Consequently images are scaled by the inverse ratio between VGGT's and metric depth. Scaled camera extrinsics and depth maps are then used for reconstruction. Areas depicting the sky (obtained using single-view segmentation on *OTAS Small* with positive *"sky", "clouds"* and negative *"ground", "object"* prompts) and overexposed areas (obtained using a threshold of 0.75) are excluded from scale estimation. For *OTAS Spatial*, these areas and depth values over 15 metres are also excluded from the reconstruction.

Nerfacto [35], Feature Splatting [5], and LERF [8] are trained using default settings and for the default number of iterations. *OTAS Spatial* uses $k = 12$ clusters and $C_r = 24$ components. Qualitative results use the prompts and segmentation thresholds *"grass"* (0.5), *"gravel"* (0.8), *"road"* (0.65), *"tree"* (0.5), *"shrubbery"* (0.6), *"tree stump"* (0.8), *"duff layer"* (0.7), and *"water"* (0.9). Each similarity query also includes the positive prompt of *"ground"* and a negative prompt of *"object"* to combat noisy depth predictions.

Visualised point clouds have their outliers removed using radius outlier rejection (removing points with fewer than 9 neighbours within a 4 metre radius) followed by statistical outlier rejection (removing points with distances to their 30 nearest neighbours larger than 1.8 standard deviations from the mean distance). Points with more than 0.75 brightness are removed as they are likely overexposed. For semantic visualisation, the point cloud has all points further than 0.5 metres from the geometric point cloud removed. These steps are exclusively cosmetic to improve the clarity of the visualisations.