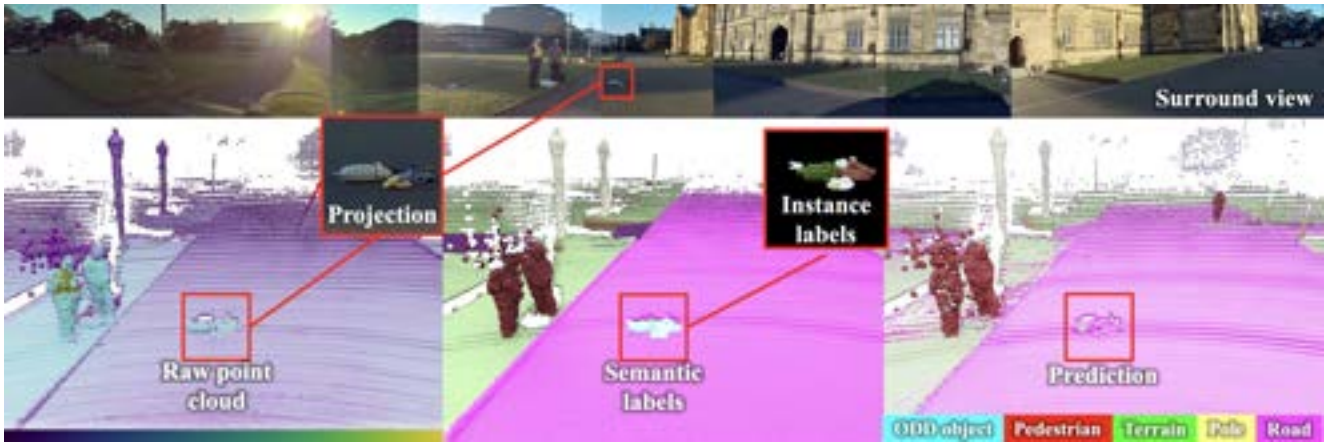# Spotting the Unexpected (STU): A 3D LiDAR Dataset for Anomaly Segmentation in Autonomous Driving

Alexey Nekrasov[1]   Malcolm Burdorf[1]   Stewart Worrall[2]   Bastian Leibe[1]
Julie Stephany Berrio Perez[2]
[1] RWTH Aachen University, [2] The University of Sydney

https://vision.rwth-aachen.de/stu-dataset

Figure 1. We present Spotting the Unexpected (STU) a novel anomaly segmentation dataset for autonomous driving. The dataset contains semantic and instance labels for out-of-distribution (OOD) objects, and includes surround-view setup with synchronized cameras.

Abstract

To operate safely, autonomous vehicles (AVs) need to detect and handle unexpected objects or anomalies on the road. While significant research exists for anomaly detection and segmentation in 2D, research progress in 3D is underexplored. Existing datasets lack high-quality multimodal data that are typically found in AVs. This paper presents a novel dataset for anomaly segmentation in driving scenarios. To the best of our knowledge, it is the first publicly available dataset focused on road anomaly segmentation with dense 3D semantic labeling, incorporating both LiDAR and camera data, as well as sequential information to enable anomaly detection across various ranges. This capability is critical for the safe navigation of autonomous vehicles. We adapted and evaluated several baseline models for 3D segmentation, highlighting the challenges of 3D anomaly detection in driving environments. Our dataset and evaluation code will be openly available, facilitating the testing and performance comparison of different approaches.

1. Introduction

Autonomous vehicles have become an increasingly important part of modern transportation systems. Despite significant advances, these vehicles continue to face challenges in accurately detecting unexpected objects on the road. One particular challenge is road debris, which refers to objects that are not typically part of the roadway environment and pose significant safety risks. According to the AAA Foundation for Traffic Safety [53], road debris was a contributing factor in more than $50,000$ crashes, $10,000$ injuries, and 125 fatalities annually between 2011 and 2014. These objects can vary widely in size and appearance, making them particularly challenging to detect. Addressing this issue is critical for ensuring the safe and reliable operation of autonomous vehicles, as it directly impacts their ability to handle anomalies in real-world driving conditions.

State-of-the-art anomaly segmentation and anomaly detection methods have received substantial attention in recent years [17, 37, 41], especially in the context of autonomous driving [11, 16, 18, 21, 31, 54]. One of the first

benchmarks for evaluating anomaly segmentation is Lost and Found [45], whose data serves as the basis for many complex benchmarks [4, 10]. These benchmarks provide a wide range of possible road anomalies and obstacles, and recent research has focused on exploring the temporal [38] and multimodal [6, 36] nature of anomaly segmentation.

While current autonomous driving datasets [1, 7, 51] use a setup with surround view cameras, radar, and high-resolution LiDAR sensors, existing anomaly segmentation and detection benchmarks focus primarily on image data [5] and mostly use Cityscapes [15] as the in-distribution dataset. Image-first benchmarks often lack the comprehensive sensor setup found in modern autonomous vehicles, and overlook the use of 3D modality for anomaly segmentation. Recent research that has begun to investigate the use of LiDAR [49] sensor for anomaly segmentation and detection, lacks high-resolution sensors found in modern autonomous vehicles. Several approaches have sought to find anomalies in validation sets of existing datasets [36, 46] or have created datasets using simulation [6] due to the difficulty of collecting and annotating such data in the real world. Despite these efforts, extending existing data or simulating anomalies only partially addresses the lack of accessible and high-quality data, shown in Table 1.

To address this issue, we analyzed the labels of two widely used in-distribution (ID) datasets [1, 7] to identify an appropriate set of anomalies that would not overlap with objects in the ID training or validation datasets. This informed our creation of a real-world dataset containing challenging anomaly objects placed on roads, with annotated sequences in 3D LiDAR point clouds. Our dataset consists of multiple high-resolution LiDAR sequences acquired in challenging environments, with point-level labels for individual anomaly objects (shown in Figure 1). We acquired data in both rural and urban environments, with objects of varying degrees of difficulty, and include a number of real-world anomalies found on public roads. In addition, to evaluate in-distribution performance, we annotated additional sequences without anomalies with inlier labels to train and evaluate models to ensure a small domain gap between our and SemanticKITTI [1] datasets.

To establish simple baselines for 3D LiDAR segmentation, we adapted the methods used in 2D anomaly segmentation [4, 24, 34, 41, 50] and applied them to the 3D domain, investigating anomaly segmentation methods for 3D LiDAR point clouds. Our models, trained jointly on the SemanticKITTI dataset and our own dataset, show strong segmentation performance for in-distribution data and novel additional sequences. However, we observed that the baseline methods struggled with out-of-distribution objects, and models tend to predict anomaly objects as inliers with a high degree of confidence. We believe that our dataset will benefit the community by allowing the evaluation of anomaly segmentation methods in the LiDAR domain and further exploration of multimodal anomaly segmentation methods.

Our contributions provide a basis for the development of more advanced anomaly segmentation methods in the 3D LiDAR domain. Specifically, we propose a new anomaly segmentation dataset, and perform an analysis of the recorded and annotated data. The purpose of the paper is to provide practitioners with the necessary tools to evaluate their anomaly segmentation models. We also provide a set of baselines that can serve as a reference for future challenges and research directions in this area.

2. Related work

2.1. Anomaly Datasets

Several multimodal datasets for autonomous driving are available, including Waymo [51], nuScenes [7], SemanticKITTI [1], among others. These datasets were collected in urban environments under typical driving conditions. Although these datasets are relevant for evaluating various tasks related to autonomous vehicle operation, they lack data for uncommon scenarios. One such scenario involves the presence of foreign objects on the road. Although this situation is rare during typical driving, autonomous vehicles must be able to recognize and respond to it safely. Here, assigning a unique label to each distinct object is impractical, since datasets have a limited set of labels, and objects on the road can vary widely. These objects are often grouped under a general category like "other object" and are excluded from the benchmarks due to their variability.

In recent years, several datasets have been introduced to assess the detection of foreign objects on the road [4, 10, 25, 36, 38, 42, 45, 59]. Some of these datasets have been developed using data augmentation techniques or simulations, but only a few are based on data collected with real-world equipment in real driving scenarios. Simulations and augmentations are great tools, but they lack realism in terms of the sensor model [4] and the sizes of the objects [25]. However, the aforementioned datasets are limited to image data only. Autonomous vehicles use LiDAR sensors because they offer greater range, robustness to lighting conditions, and more precise 3D information about the environment compared to monocular depth estimations or stereo cameras. Therefore, incorporating 3D information is essential for accurately assessing road anomalies.

Three datasets that integrate camera and LiDAR data are SOD [22], TOR4D [58] and CODA [36]. The CODA series is derived from validation sets of existing autonomous vehicle datasets. Still, it has the limitation that anomalous objects (*e.g.*, traffic cones, barriers) may also appear in the training set, which is acceptable for the task of open-set segmentation, but contradicts the task of anomaly segmentation. Recently, a few methods for 3D anomaly detection

Table 1. Comparison of our dataset with other community datasets [5]. SM = Semantic Mask, IM=Instance Mask, BB=Bounding Box

Dataset	Camera	LiDAR Beams	Source	#ODD	Label	Size Test/Val	Temporal	Environments Test Sequences	Related Dataset
Fishyscapes [3, 4]									
FS Lost and Found	1 RBG	✗	Staged	1	SM	275/100	✗	n/a	Lost and Found [45]
FS Static	1 RGB	✗	Augmented	1	SM	1000/30	✗	n/a	CityScapes [15]
CAOS [25]									
StreetHazards	1 RGB	✗	Simulated	1	SM	1500	✓	2 sim towns - 150 seq	
BDD-Anomaly	1 RGB	✗	Validation Set	3	SM	810	✗	n/a	BDD100K [61]
SegmentMeIfYouCan [10]									
RoadAnomaly21	1 RGB	✗	Web Sourced	1	SM	100/10	✗	n/a	
RoadObstacle21	1 RGB	✗	Staged	1	SM	382/30	✓	9 streets - 90 seq	
CODA [36]									
CODA-KITTI	1 RGB	✓ - 64	Validation Set	6	BB	309	✗	n/a	KITTI [19]
CODA-nuScenes	1 RGB	✓ - 32	Validation Set	17	BB	134	✗	n/a	NuScenes [7]
CODA-ONCE	1 RGB	✓ - 40	Validation Set	32	BB	1057	✗	n/a	ONCE [39]
CODA22-ONCE	1 RGB	✓ - 40	Validation Set	29	BB	717	✗	n/a	ONCE [39]
CODA22-SODA10M	1 RGB	✗	Validation Set	29	BB	4167	✗	n/a	SODA [23]
Wuppertal ODD [38]									
SOS	1 RGB	✗ - dist	Staged	13	IM	1129	✓	9 streets - 20 seq	
CWL	1 RGB	✗ - dist	Simulated	18	IM	1210	✓	26 streets - 26 seq	
Lost and Found [45]	1 Stereo	✗	Staged	42	SM	1068	✓	5 streets - 5 seq	
SOD [49]	1 RGB	✓ - 16	Staged	1	SM	460/530	✗	2 streets - 5 seq	
WD-Pascal [2]	1 RGB	✗	Simulated	1	SM	70	✗	n/a	
Vistas-NP [20]	1 RGB	✗	Class exclusion	4	SM	11167	✗	n/a	Mapillary Vistas [43]
STU (Ours)	**8 RGB**	**✓ - 128**	**Staged**	**1**	**IM**	**8022/1960**	**✓**	**6 streets - 51 seq**	**SemKITTI [1], P-CUDAL [55]**

have been proposed [28, 33, 46]. However, these methods use validation sequences to evaluate anomaly detection capabilities. The SOD dataset offers depth information for small road obstacles using a low-resolution 16-beam LiDAR, and that introduces a large domain gap between the training and test data. TOR4D and Rare4D [58] are the only datasets that contain real-world LiDAR data to benchmark road anomalies; however, these datasets are proprietary and not accessible to the public. We release out-of-distribution validation and in-distribution training sequences with a permissive license and open the private test set to submissions from the public.

2.2. Open-Set Segmentation in 3D

Open-set and unsupervised segmentation are closely related to the out-of-distribution detection task. Multiple methods were proposed for unsupervised segmentation [44, 48] and open-set prediction in 3D [9, 58]. Recent evaluations in 2D anomaly segmentation [18] have demonstrated that methods such as OSIS [58] and EOPSN [30] perform poorly on anomaly objects. This weakness stems from open-set methods' reliance on the presence of unknown instances inside the void regions during training, which may not generalize well to real-world scenarios. Similarly, MLUC [8] assesses open-set performance on the UDI [8] dataset, but this dataset is also not publicly accessible and makes the same assumptions as OSIS methods. In addition, open-set methods typically remove the ground plane [44] to obtain clusters of potential objects, but ground plane removal [35] methods are often sensitive to chosen hyperparameters and yield noisy results, missing many smaller objects. For the evaluation of anomaly segmentation, it is important to have a test dataset that is free of anomaly objects which might occur during training. In contrast to open-set methods, our work focuses on collecting and densely annotating real-world anomaly objects, ensuring that there is no intersection between in-distribution training sets and out-of-distribution test sets.

2.3. Anomaly Segmentation in Autonomous Driving

To tackle the problem of anomaly segmentation, multiple works [4, 5, 10] have recently been proposed for the problem of 2D anomaly segmentation. Earlier works [4, 11, 21, 24, 34, 37, 54] were applied to per-pixel segmentation models such as DeepLab [12]; however, recently many works were focused on MaskFormer-style models [13, 14]. A line of work [16, 41, 47] proposed different methods to adapt Mask2Former, exploring the mechanism of mask prediction, where multiple masks might ignore certain regions of an image. These models provide more reliable anomaly scoring mechanisms not limited to per-pixel predictions and achieve higher performance on the anomaly segmentation task. For our benchmark, we decided to evaluate methods that also use MaskFormer-style models and investigate the application of the Mask4Former [60] model for 3D anomaly segmentation. We obtain all our scores using a MaskFormer-Style model, laying the foundation for future baselines.

3. The Spotting the Unexpected (STU) Dataset

Our data collection platform consists of a rigid frame mounted on a vehicle and equipped with eight hardware-triggered cameras and a LiDAR sensor to ensure synchronized data collection. The LiDAR and the cameras were calibrated using the method described in [56], with the cal-

(b) Data collection in controlled environment.

Figure 2. Data collection conducted in a naturalistic manner (a) and controlled environment (b) with objects on the road.

ibration process repeated for each camera. This setup follows the configuration discussed in Panoptic-CUDAL [55]. We refer to the Supplementary Material for detailed information on the vehicle setup.

3.1. Data collection.

The data was collected using two sets of conditions: one in a naturalistic environment where objects were found on public roads, and one in a controlled environment. In the naturalistic scenario, several hours of driving data were recorded, with specific events marked when objects appeared on the road. Figure 2a shows a bucket on the road recorded while driving on a highway.

The controlled environment data collection protocol was rigorous for safety reasons. We employed four people: two spotters positioned at either end of a low-traffic and low-speed road, one responsible for placing objects on the road, and a driver. The spotters would signal when it was safe to place the objects, and once positioned, the driver would begin recording the scene. If one of the spotters reported an approaching vehicle, the data recording was stopped, and the object was immediately removed. This was done to avoid disrupting the normal flow of traffic. The vehicle was driven at a maximum speed of 50 km / h during data recording. Figure 2b shows a computer monitor placed on the road during sensor data logging. Some of the objects placed on the road are also shown in Figure 3.

Postprocessing. After data collection, post-processing is required to obtain vehicle poses and anonymize the images. Point cloud registration was performed using KISS ICP [57], a LiDAR odometry pipeline. KISS ICP includes point cloud motion compensation, subsampling, adaptive thresholding to determine correspondences, and LiDAR pose estimation. The calculated LiDAR pose was then exported in KITTI format as required for the labeling tool [1].

Ethical considerations for this dataset include anonymizing camera images. Identifiable data, such as faces and license plates, are processed using DashcamCleaner [52] and DeepPrivacy2 [29]. DashcamCleaner uses a license plate detector to locate and blur license plates, while DeepPrivacy2 detects facial features and generates new, unidentifiable faces to replace the original ones. Examples of anonymization can be found in the supplementary material.

3.2. Labeling

We define anomalies as objects that may pose a danger to the autonomous vehicle and its passengers, especially those present on the driving surface and not present in the training dataset. In this work, we focus on road anomalies and do not consider unusual situations such as vehicle collisions or unexpected driver behavior, and we focus on the fine-grained anomaly segmentation. In particular, we follow previous work [4, 45] and focus on objects that are not present in the training set.

We labeled the data using the SemanticKITTI labeler [1]. To generate initial labels for the scenes, we used pseudo-labeling techniques using a pre-trained SemanticKITTI model. We instruct our annotators to group all points belonging to inliers as a single semantic class, questionable points as unlabeled, and anomaly points as the outlier class. We instruct our labelers that it is very important to label inliers and outliers carefully while allowing for larger unlabeled regions where the annotator may be challenged.

Annotation process. The domain gap between urban environments and SemanticKITTI street scenes is minimal, especially in regions close to the vehicle. This similarity allows the acquisition of reliable initial estimates. The annotation process was performed by three annotators, each spending eight hours a week over several months. Each annotator spent approximately four hours on the initial labeling of each scene and an additional two hours on refining the final labels. To ensure accuracy and consistency, the annotators cross-validated each other's work. To further improve the quality of the labels, we examined the predictions of the baseline methods to see if any known objects were missed. In particular, current methods tend to be highly sensitive to objects that are present but ignored during training, such as those classified in the category "other object". We annotate these objects as unlabeled and ignore them in the evaluations, as they could alter the results of the anomaly segmentation methods.

Differences to SemanticKITTI. In contrast to SemanticKITTI, which provides labels up to a range of 50 meters, we labeled all points visible to the LiDAR sensor. This was facilitated by using a LiDAR with a higher vertical resolution. The maximum distance resulting from an anomaly in our dataset is 150 meters, allowing for comprehensive evaluations. However, since we train on the Se-

Figure 3. Different anomalies in the STU dataset. Different objects on the road used for staged data collection. We pick objects such that have no intersection with the inlier dataset and place them on roads in different locations and illumination conditions. Objects might touch each other, be very small, as large as a chair or a surf board, and could cause an accident if a car would drive over them.

manticKITTI dataset, which contains only labels up to 50 meters, we limit our evaluations to the 50 meter range. The sensor setup has 128 LiDAR beams, compared to 64 beams in SemanticKITTI.

3.3. Dataset analysis

We recorded 70 sequences with anomalies, which we fully annotated for anomaly segmentation. Two of these labeled sequences were captured in a naturalistic manner, while the remaining 68 were staged to include various anomalies. We designated 19 sequences as a validation set and 51 as a closed test set for benchmarking purposes. In addition, we recorded and annotated two sequences with no anomalies for training and validation of the inlier performance "STU-inlier", to reduce the domain gap between our dataset and SemanticKITTI.

Our dataset defines two primary classes: *inliers* and *outliers*. *Inliers* encompasses classes that are used to train a model, and contains objects expected on the road, such as vehicles, pedestrians, and road infrastructure. *Outliers*, on the other hand, refer to out-of-distribution elements or anomalies that are atypical or unexpected in the driving scene, shown in Figure 3. One additional class, *unlabeled* is ignored in the evaluation and typically includes classes that the model has seen during training, but was not supervised in, such as parking meters, utility boxes, and lamps. We provide anomaly annotations at both the instance and semantic levels, enabling a comprehensive evaluation of models for semantic and instance segmentation in 3D environments.
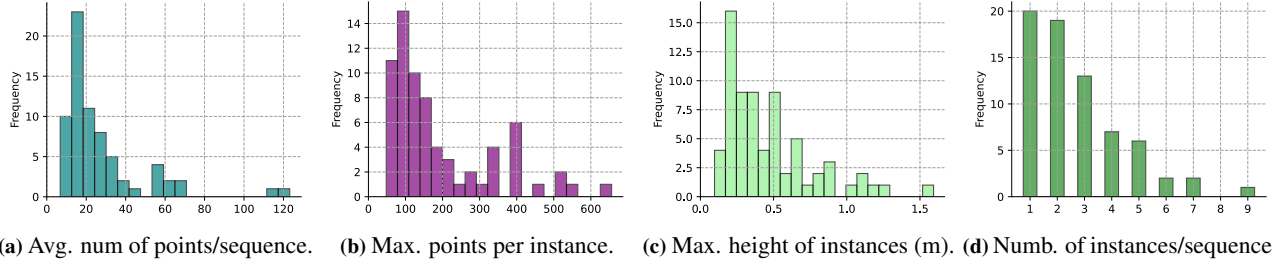
Each outlier object in our dataset has an instance la-

bel unique for a sequence. Figure 4 presents histograms that display the distribution of key properties for these instances, including the average and maximum number of LiDAR points per instance, instance heights, and the number of instances per scan. This analysis provides insights into the structural characteristics and variability of data within point clouds. Specifically, Figure 4 indicates that most instances have very few LiDAR points on average, but a sequence may have multiple instances per sequence.

Furthermore, Figure 5 presents a scatter plot of detected anomalies within the vehicle frame, highlighting a higher density of anomalies along the longitudinal (X) axis both in front and behind the vehicle. The accompanying histograms depict the distribution of the anomalies along each axis, noting that the anomalies are identified at distances ranging up to 150 meters in front of the vehicle. Figure 6 explores the relationship between the distance to an anomaly and its size, showing that the proportion of LiDAR point cloud obstacles decreases as the distance increases. This trend is due to the radial nature of LiDAR, which has lower vertical resolution at greater ranges. For benchmarking purposes, we focus exclusively on instances with more than five LiDAR points, highlighted in blue, ensuring that the evaluation targets detectable anomalies.

3.4. Evaluated Baselines

The formulation of the anomaly segmentation task restricts training to avoid direct supervision of OOD objects. Generally, OOD objects are not present in the training set, but could be present during inference [18]. Since there are no methods that could be applied directly for our setup, we

(a) Avg. num of points/sequence. (b) Max. points per instance. (c) Max. height of instances (m). (d) Numb. of instances/sequence.

Figure 4. Anomaly Instance Properties. A typical recorded anomaly has on average less then 50 points per sequence (a), with less then 300 (b) points at maximum, and a maximum height below one meter (c). We record up to nine individual anomaly instances in the same sequence (d).
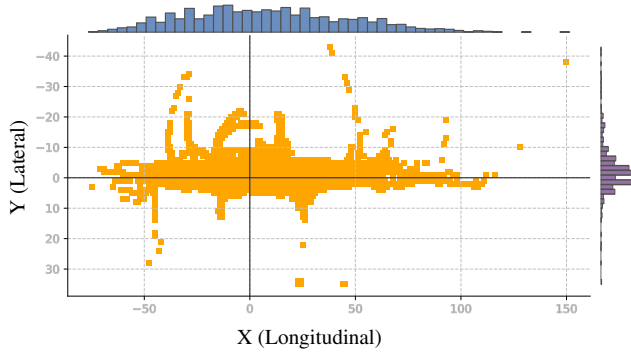
Figure 5. Distribution of anomalies along the vehicle's reference frame. Most of the points appear around the vehicle.
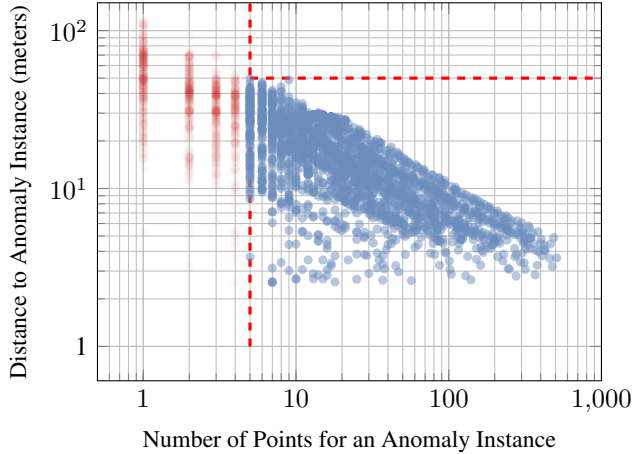
Figure 6. We follow SemanticKitti's maximum distance threshold and evaluate instances that are within a 50 meter radius of the vehicle. In addition, we constrain evaluation of point and object level metrics to a minimum of 5 points.

adopt several baseline methods from the 2D anomaly segmentation to 3D LiDAR segmentation. First, we evaluate simple baselines, such as Max-Logit [24], Monte-Carlo Dropout [50], and Deep Ensembles [34]. Secondly, we train one of the models to predict ignored regions during train-

ing, which could be used to classify "void" points during testing [4]. We also adapt RbA [41], one of the state-of-the-art methods for image anomaly segmentation, to the 3D LiDAR segmentation method Mask4Former-3D [60].

Deep Ensembles, MC Dropout, and MaxLogit. Deep Ensembles [34] and MC Dropout [50] are very common methods that practitioners use due to ease of implementation and application. For Deep Ensembles, we train three Mask4Former-3D [60] models. For MC dropout, we train one Mask4Former-3D model with enabled dropout in the transformer decoder for self- and cross-attention operations, and during inference we repeat the transformer decoder operation 50 times. As a simple baseline that uses only one model and a single forward pass, we use the MaxLogit [24] method.

RbA [41] is a method that uses the ability of MaskFormer-Style models [13, 14] to predict multiple overlapping segmentation masks. MaskFormer models have an option to disable queries and might not predict any masks if no objects are present. Points for which no masks were predicted will have the highest anomaly score. The Mask4Former model follows the same architectural paradigm and has a query deactivation mechanism that allows for no prediction in uncertain regions. This allows us to use the RbA method to generate anomaly scores within our setup. Unlike the original RbA model setup, we did not modify the transformer decoder.

Void Classifier We follow a common baseline [4] by treating the unlabeled regions as an additional semantic class during training.

4. Baseline Evaluation

Training data. We use SemanticKITTI [1], Panoptic-CUDAL [55], and an additional recorded sequence from to jointly train our models. The Panoptic-CUDAL dataset, contains six sequences for training and one for validation, and was mostly recorded in rural locations, while our dataset consists mainly of urban scenes. To reduce the do-

Table 2. Anomaly Segmentation Performance of the selected baseline methods.

Method	OoD Data	Point-Level OOD			Object-Level OOD				
		AUROC ↑	FPR@95 ↓	AP ↑	RecallQ	SQ	RQ	UQ	PQ
Deep Ensemble [34]	✗	**86.74**	**58.05**	**5.17**	16.75	84.49	**10.43**	14.16	**8.81**
MC Dropout [50]	✗	61.51	82.37	0.11	2.25	**86.72**	1.95	2.14	1.86
Max Logit [24]	✗	84.53	81.49	0.95	**26.14**	83.06	2.13	**21.71**	1.77
Void Classifier [4]	✓	85.99	78.60	3.92	17.64	84.40	8.19	14.89	6.91
RbA [41]	✗	66.38	100.0	0.81	24.04	83.28	3.23	20.02	2.69

Table 3. Closed-set performance of the methods on validation sets of SemanticKITTI [1] and STU-inlier dataset. * denotes a model trained only on SemanticKITTI dataset.

Method	SemanticKITTI			STU-inlier Dataset		
	PQ	SQ	RQ	PQ	SQ	RQ
Mask-PLS* [40]	59.9	76.4	69.1	39.6	67.0	50.5
Mask-PLS [40]	55.6	76.0	64.5	50.9	69.4	62.2
Mask4Former-3D* [60]	**61.9**	81.3	71.6	42.8	70.6	53.2
Mask4Former-3D [60]	60.7	81.0	70.4	**52.7**	73.8	64.7

main gap, we have recorded and annotated two additional inlier scenes from an urban setting "STU-inlier", one for training and one for evaluation. These training sequences closely resemble the environments in which anomaly scenes were recorded. We train models with additional data recorded from the same vehicle setup as Panoptic-CUDAL and labeled the data with 19 SemanticKITTI labels. However, SemanticKITTI and our recorded dataset contain "other-object" that acts as a miscellaneous class for all objects that do not fit other-class definitions. A typical object is a bench, a garbage bin, or an advertisement post. To strictly define anomalies, we analyze all other-objects present in our training set, as well as the "Movable_Object.Debris" and "Pushable.Pullable" class objects in the NuScenes training set. We specifically design our dataset such that we do not have an intersection between anomalous objects and the aforementioned classes. Our Void Classifier training procedure consists of training on this unlabeled class that is present in the training data.

Models. We train MaskPLS [40] and Mask4Former [60] models jointly on both SemanticKitti and our training sequences. These models achieve state-of-the-art results on the SemanticKITTI [1] dataset for single-scan and multi-scan settings. However, while our dataset supports evaluation of temporal sequences, we performed our evaluation and training on single scans only. In our experiments, we refer to Mask4Former trained on single scans as Mask4Former-3D. All anomaly segmentation methods use Mask4Former-3D as the main model.

Metrics. To evaluate the performance of our methods, we use common metrics in the anomaly segmentation commu-

nity in 2D and panoptic segmentation in 3D LiDAR. Following SemanticKitti, we evaluated both inlier and outlier performance on points within 50 meters of the vehicle, and on objects with at least five points.

Point-Level Metrics. To evaluate the performance of the anomaly segmentation at the point level, we use the common metrics of Average Precision (AP), False-Positive Rate at 95% True-Positive Rate (FPR95), and Area Under Receiver Operating characteristic Curve (AUROC). Commonly used in the anomaly segmentation community [4].

Object-Level Metrics. To be consistent with the panoptic segmentation setup, we use panoptic metrics to evaluate the model performance on inlier classes. For the inlier data, we compute the Panoptic Quality (PQ) on the SemanticKitti classes. We report detailed class scores in the Supplementary. In contrast to anomaly segmentation, the open task formulation emphasizes the recall of all objects. Formally, the Unknown Quality (UQ) metric is used in such a setting [58]. However, in the case of anomaly segmentation, since anomalies can pose serious threats to the vehicle, it is important to predict as few false positives as possible, while also producing accurate segmentation labels for anomaly objects. Panoptic Quality uses an F-score and penalizes the prediction of multiple false positives. Thus, for anomaly segmentation baselines, we report both Panoptic Quality and Unknown Quality on a single class. For inlier performance, we report the Panoptic Quality score on the SemanticKITTI classes.

Panoptic Quality [32] is the geometric mean between Recognition Quality (RQ) and Segmentation Quality (SQ) for a class c:

$$\text{PQ}_c = \underbrace{\frac{\sum_{(p,g)\in TP_c} \text{IoU}(p,g)}{|TP_c|}}_{\text{Segmentation Quality (SQ)}} \times \underbrace{\frac{|TP_c|}{|TP_c| + \frac{1}{2}|FP_c| + \frac{1}{2}|FN_c|}}_{\text{Recognition Quality (RQ)}}$$

where TP, FP, and FN are True Positives, False Positives, and False Negatives for objects with Intersection over Union (IoU) higher than 50%. The final PQ metric is calculated as an average over all classes c. Ignore points are removed from the scene prior to evaluation and erroneous predictions in the ignore region are not penalized.

To evaluate the recall of anomaly objects, we use the Unknown Quality (UQ) metric [58]:

$$\text{UQ} = \underbrace{\frac{\sum_{(p,g)\in TP}\text{IoU}(p,g)}{|TP|}}_{\text{Segmentation Quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP|+|FN|}}_{\text{Recall Quality (RecallQ)}} .$$

The metric does not penalize FP and allows for the evaluation of object recall. However, the metric counts objects as TP only if they have an IoU of at least 50% with ground truth.

4.1. Experiments

Inlier Performance on SemanticKitti and STU-inlier dataset. We train models jointly with SemanticKITTI [1], Panoptic-CUDAL [55] and our additional sequence, and evaluate the SemanticKITTI and STU-inlier validation sets. The jointly trained model demonstrates improved performance on our STU-inlier dataset compared to models trained solely on the SemanticKITTI dataset, suggesting a small domain gap between our dataset and SemanticKITTI. Evaluation on the proposed STU-inlier sequence shows good segmentation performance for inlier classes that is critical for anomaly segmentation. The metrics for the performance of the baseline method in the validation sets are provided in Tab. 3.

Out-of-Distribution (OOD) Performance. We evaluate baseline methods for anomaly segmentation at both the point and object levels. Our findings indicate that directly transferring methods from 2D to 3D does not yield similar performance. Our point clouds typically contain around $100,000$ points, with only a few dozen being anomalous, as shown in Fig. 6. Due to the significant class imbalance, point-level scores are generally low, highlighting the difficulty of the task. A high FPR@95, as noted in ODIN [27], suggests that models often misclassify road objects as inlier classes. Since LiDAR point clouds have lower resolution at a distance, predicting a small anomaly object consisting of a few points is a hard task. It is common for large anomalous objects to be predicted as pedestrians or other vehicles, as depicted in Fig. 7. Anomalous objects frequently occur in familiar contexts, which leads the model to confidently assign inlier labels to such anomalies in many cases. This behavior contrasts significantly with models trained on 2D data, where a simple max logit approach [26] outperforms Deep Ensembles [10]. However, the Deep Ensemble model stands out by producing fewer false positives at the point level, resulting in a lower FPR@95 and a higher PQ metric. Despite this reduction in false positives, the average precision (AP) remains low compared to 2D datasets. We attribute this to the high confidence with which individual models predict inlier objects, with disagreements between models contributing to the improved FPR@95. We refer

Figure 7. Example of a failure case anomaly segmentation. For the chair on the road, labeled as anomaly in upper left, model predicts "other-vehicle" ■ class, in lower left, with a high certainty, as indicated by MaxLogit scores on the bottom right.

readers to the supplementary material for a more detailed investigation of the results.

5. Conclusions and Future Work

In this paper, we introduced the STU dataset, designed for anomaly segmentation in LiDAR point clouds. The dataset was collected using eight cameras and a LiDAR sensor, which provides a surround view of environments through images and high-resolution LiDAR data. We extensively annotated point clouds with "unlabeled", "anomaly", and "inlier" classes, as well as introduced additional sequences without anomalies for training and evaluation. The results of the baseline methods indicate that there is a significant performance gap between 2D and 3D anomaly segmentation, which underscores numerous ongoing research opportunities in the field. Our objective with this paper is to provide researchers with annotated data and standardized benchmarks to facilitate the comparison of their anomaly segmentation approaches in 3D. We believe the dataset could be especially useful for the evaluation of methods that use temporal LiDAR data and multi-modal data in the future, and can drive innovation in developing more robust anomaly detection systems that can better handle complex real-world scenarios.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 4, 6, 7, 8

[2] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Simultaneous Semantic Segmentation and Outlier Detection in Presence of Domain Shift. In *German Conference on Pattern Recognition (GCPR)*, 2019. 3

[3] Hermann Blum, Paul-Edouard Sarlin, Juan I. Nieto, Roland Y. Siegwart, and César Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. *International Conference on Computer Vision Workshop (ICCV'W)*, 2019. 3

[4] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The Fishyscapes Benchmark: Measuring Blind Spots in Semantic Segmentation. *International Journal on Computer Vision (IJCV)*, 2021. 2, 3, 4, 6, 7

[5] Daniel Bogdoll, Svenja Uhlemeyer, Kamil Kowol, and J Marius Zöllner. Perception datasets for anomaly detection in autonomous driving: A survey. In *Intelligent Vehicles Symposium (IV)*, 2023. 2, 3

[6] Daniel Bogdoll, Iramm Hamdard, Lukas Namgyu Rößler, Felix Geisler, Muhammed Bayram, Felix Wang, Jan Imhof, Miguel de Campos, Anushervon Tabarov, Yitian Yang, Hanno Gottschalk, and J. Marius Zöllner. AnoVox: A Benchmark for Multimodal Anomaly Detection in Autonomous Driving. In *ECCV 2024 W-CODA workshop*, 2024. 2

[7] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[8] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Open-set 3D Object Detection. In *International Conference on 3D Vision (3DV)*, 2021. 3

[9] Anirudh S. Chakravarthy, Meghana Reddy Ganesina, Peiyun Hu, Laura Leal-Taixe, Shu Kong, Deva Ramanan, and Aljosa Osep. Lidar Panoptic Segmentation in an Open World. *International Journal on Computer Vision (IJCV)*, 2024. 3

[10] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 2, 3, 8

[11] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 3

[12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[13] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 6

[14] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Neural Information Processing Systems (NeurIPS)*, 2021. 3, 6

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3

[16] Anja Delić, Matej Grcic, and Siniša Šegvić. Outlier detection by ensembling uncertainty with negative objectness. In *British Machine Vision Conference (BMVC)*, 2024. 1, 3

[17] kuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[18] Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Nassir Navab, Benjamin Busam, and Federico Tombari. Segmenting Known Objects and Unseen Unknowns without Prior Knowledge. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 3, 5

[19] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. In *The International Journal of Robotics Research*, 2013. 3

[20] Matej Grcic, Petra Bevandic, and Sinisa Segvic. Dense open-set recognition with synthetic outliers generated by real nvp. *ArXiv*, abs/2011.11094, 2020. 3

[21] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 3

[22] Krishnam Gupta, Syed Ashar Javed, Vineet Gandhi, and K. Madhava Krishna. MergeNet: A Deep Net Architecture for Small Obstacle Discovery. In *International Conference on Robotics and Automation (ICRA)*, 2018. 2

[23] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Jiageng Mao, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. SODA10M: A Large-Scale 2D Self/Semi-Supervised Object Detection Dataset for Autonomous Driving. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 3

[24] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 3, 6, 7, 4

[25] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling Out-of-Distribution Detection for Real-

World Settings. In *International Conference on Machine Learning (ICML)*, 2022. 2, 3

[26] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Xiaodong Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, 2022. 8

[27] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data . In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[28] Chengjie Huang, Van Duong Nguyen, Vahdat Abdelzad, Christopher Gus Mannes, Luke Rowe, Benjamin Therien, Rick Salay, and K. Czarnecki. Out-of-distribution detection for lidar-based 3d object detection. *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2022. 3

[29] Håkon Hukkelås and Frank Lindseth. Deepprivacy2: Towards realistic full-body anonymization. In *Winter Conference on Applications of Computer Vision (WACV)*, 2023. 4, 2

[30] Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-Based Open-Set Panoptic Segmentation Network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[31] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized Max Logits: A Simple yet Effective Approach for Identifying Unexpected Road Obstacles in Urban-Scene Segmentation. In *International Conference on Computer Vision (ICCV)*, 2021. 1

[32] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[33] Michael Kösel, Marcel Schreiber, Michael Ulrich, Claudius Gläser, and Klaus Dietmayer. Revisiting Out-of-Distribution Detection in LiDAR-based 3D Object Detection. In *Intelligent Vehicles Symposium (IV)*, 2024. 3

[34] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Neural Information Processing Systems (NeurIPS)*, 2017. 2, 3, 6, 7, 4

[35] Seungjae Lee, Hyungtae Lim, and Hyun Myung. Patchwork++: Fast and robust ground segmentation solving partial under-segmentation using 3D point cloud. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022. 3, 2

[36] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[37] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. GMMSeg: Gaussian Mixture based Generative Semantic Segmentation Models. In *Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3

[38] Kira Maag, Robin Chan, Svenja Uhlemeyer, Kamil Kowol, and Hanno Gottschalk. Two Video Data Sets for Tracking

and Retrieval of Out of Distribution Objects. In *Asian Conference on Computer Vision (ACCV)*, 2022. 2, 3

[39] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Hang Xu, and Chunjing Xu. One Million Scenes for Autonomous Driving: ONCE Dataset. In *Neural Information Processing Systems (NeurIPS)*, 2021. 3

[40] Rodrigo Marcuzzi, Lucas Nunes, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. Mask-Based Panoptic LiDAR Segmentation for Autonomous Driving. In *IEEE Robotics And Automation Letters (RAL)*, 2023. 7, 4

[41] Nazir Nayal, Mısra Yavuz, João F. Henriques, and Fatma Güney. RbA: Segmenting Unknown Regions Rejected by All. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 6, 7, 4

[42] Alexey Nekrasov, Rui Zhou, Miriam Ackermann, Alexander Hermans, and Matthias Rottmann Bastian Leibe. OoDIS: Anomaly Instance Segmentation Benchmark. *arXiv preprint arXiv:2406.11835*, 2024. 2

[43] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *International Conference on Computer Vision (ICCV)*, 2017. 3

[44] Lucas Nunes, Xieyuanli Chen, Rodrigo Marcuzzi, Aljosa Osep, Laura Leal-Taixé, Cyrill Stachniss, and Jens Behley. Unsupervised Class-Agnostic Instance Segmentation of 3D LiDAR Data for Autonomous Vehicles. *IEEE Robotics And Automation Letters (RAL)*, 2022. 3

[45] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and Found: Detecting Small Road Hazards for Self-Driving Vehicles. In *International Conference on Intelligent Robots and Systems (IROS)*, 2016. 2, 3, 4

[46] Aldi Piroli, Vinzenz Dallabetta, Johannes Kopp, Marc Walessa, Daniel Meissner, and Klaus Dietmayer. LS-VOS: Identifying Outliers in 3D Object Detections Using Latent Space Virtual Outlier Synthesis. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2023. 2, 3

[47] Shyam Nandan Rai, Fabio Cermelli, Dario Fontanel, Carlo Masone, and Barbara Caputo. Unmasking Anomalies in Road-Scene Segmentation. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[48] Jenny Seidenschwarz, Aljoša Ošep, Francesco Ferroni, Simon Lucey, and Laura Leal-Taixé. SeMoLi: What Moves Together Belongs Together. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[49] Aasheesh Singh, Aditya Kamireddypalli, Vineet Gandhi, and K. Madhava Krishna. LiDAR guided Small obstacle Segmentation. In *International Conference on Intelligent Robots and Systems (IROS)*, 2020. 2, 3

[50] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. In *Neural Information Processing Systems (NeurIPS)*, 2014. 2, 6, 7, 3, 4

[51] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou,

Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[52] DashcamCleaner Development Team. Dashcamcleaner: Censor identifiable information in videos from dashcam recordings. https://github.com/tfaehse/DashcamCleaner, 2024. 4, 2

[53] Tefft, Brian C. . The Prevalence of Motor Vehicle Crashes Involving Road Debris, United States, 2011-2014. https://aaafoundation.org/wp-content/uploads/2017/12/RoadDebris_REPORT_2015.pdf, 2015. [Online]. 1

[54] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise Energy-biased Abstention Learning for Anomaly Segmentation on Complex Urban Driving Scenes. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 3

[55] Tzu-Yun Tseng, Alexey Nekrasov, Malcolm Burdorf, Bastian Leibe, Julie Stephany Berrio Perez, Mao Shan, and Stewart Worrall. Panoptic-CUDAL Technical Report: Rural Australia Point Cloud Dataset in Rainy Conditions. *arXiv preprint arXiv:2503.16378*, 2025. 3, 4, 6, 8

[56] S. Verma, J. S. Berrio, S. Worrall, and E. Nebot. Automatic extrinsic calibration between a camera and a 3d lidar using 3d point and plane correspondences. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019. 3

[57] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. KISS-ICP: In Defense of Point-to-Point ICP – Simple, Accurate, and Robust Registration If Done the Right Way. In *IEEE Robotics And Automation Letters (RAL)*, 2023. 4, 2

[58] Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying Unknown Instances for Autonomous Driving. In *Conference on Robot Learning (CoRL)*, 2019. 2, 3, 7, 8

[59] Yajun Xu, Huan Hu, Xiaoya Zhu, Yibing Nan, Kai Wang, ZhaoXiang Liu, and Shiguo Lian. Raod: A benchmark for road abandoned object detection from video surveillance. *IEEE Access*, 2024. 2

[60] Kadir Yilmaz, Jonas Schult, Alexey Nekrasov, and Bastian Leibe. Mask4Former: Mask Transformer for 4D Panoptic Segmentation. In *International Conference on Robotics and Automation (ICRA)*, 2024. 3, 6, 7, 4

[61] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

Spotting the Unexpected (STU): A 3D LiDAR Dataset for Anomaly Segmentation in Autonomous Driving

Supplementary Material

6. Hardware Setup

The sensors and hardware included in the data collection platform are as follows:

- 5 SF3325 automotive GMSL cameras (ONSEMI CMOS image sensor AR0231), SEKONIX ultra high-resolution lens with 60 horizontal and 38 vertical FOV, images captured at a resolution of 1928 × 1208 (2.3M pixel) at 30 Hz.
- 3 SF3324 automotive GMSL cameras, (ONSEMI CMOS image sensor AR0231), SEKONIX ultra high-resolution lens with 120 horizontal and 73 vertical FOV, images captured at a resolution of 1928 × 1208 (2.3M pixel) at 30 Hz.
- 1 OS1-128 Ouster Lidar, with a vertical resolution of 128 beams within a 45 FOV and range of 200 meters, point cloud captured at 10 Hz.
- 1 NVIDIA DRIVE Pegasus, with two NVIDIA Xavier™ SoCs.

The placement and reference frames of the sensors on the vehicle are shown in Figure 8. The arrangement of the cameras and LiDAR sensors allows the vehicle to achieve a full 360° field of view (FOV) of its surroundings.
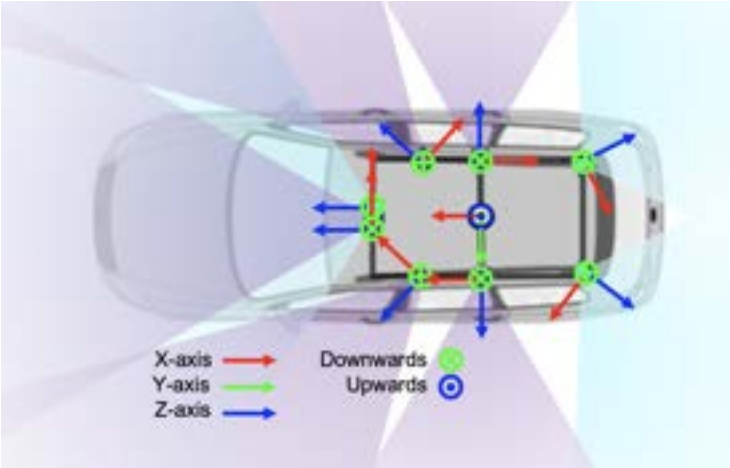

Figure 8. Sensor setup of the data collection vehicle. The field of view for the 60-degree and 120-degree cameras is represented in purple and blue, respectively.

6.1. Extrinsic Calibration

The camera positions on the vehicle were determined through a LiDAR-camera calibration process in which we computed the homogeneous transformation matrices from the LiDAR to each camera. Without ground truth for these transformations, their accuracy is typically validated visually by examining the correspondence between objects in the camera images and the LiDAR point cloud. In this case, Figure 9 shows the alignment between the LiDAR point cloud and an image captured by the front camera, illustrating the accuracy of the calibration process.
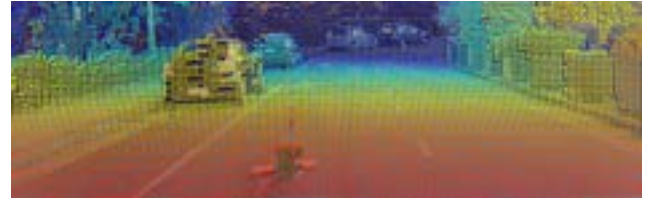

Figure 9. Lidar-Camera projection. The point cloud is colored by distance to the camera.

6.2. Software

A Robotic Operation System (ROS) framework manages the LiDAR point-cloud acquisition pipeline. Camera data is captured using the NVIDIA DRIVE Pegasus video capture and image compression pipeline. All images are encoded and stored as H.264 video, with associated metadata stored in a custom ROS message.

7. Data Collection

For staged data collection, we used a diverse collection of objects, including buckets, indoor garbage bins, brooms, chairs, pots, stuffed animals, balloons, balls, backpacks, bags, pillows, shoes, umbrellas, hats, yoga mats, helmets, swimming noodles, tissue boxes, ladders, car seats, sleeping bags, and bottles. The point cloud captures the three-dimensional structures of objects at varying distances. Figure 10 shows the point clouds of a chair, an indoor garbage

| (a) Chair. | (b) Garbage bin. | (c) Ladder. |
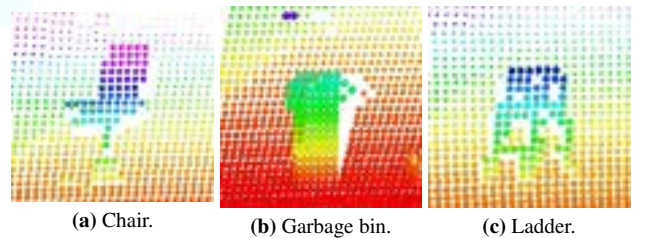
Figure 10. Point cloud of some objects on the road colored by height.

can and a ladder, each color-coded according to height to highlight their spatial dimensions.

7.1. Postprocessing

After data collection, the raw information is post-processed to estimate vehicle poses and to anonymize the image data. Point cloud registration was performed using KISS ICP [57], a lidar odometry pipeline. It includes point cloud motion compensation, subsampling, adaptive thresholding to determine correspondences, and lidar pose estimation. The calculated LiDAR pose was then exported in SemanticKITTI format as required for the labeling tool [1].

Ethical considerations for this dataset include anonymization of camera images to protect individual privacy. Identifiable information, such as faces and license plates, is processed using DashcamCleaner [52] and DeepPrivacy2 [29]. DashcamCleaner uses a license plate detector to locate and blur license plates, while DeepPrivacy2 identifies facial features and generates new unidentifiable photorealistic faces to replace the original ones. Figure 11 illustrates the anonymization process using a publicly available image from the internet.

(a) Original image (source: https://tinyurl.com/3s66za36)


(b) Anonymized image.

Figure 11. Anonymization of camera images.

7.2. Ground Plane Segmentation

One of the popular approaches for anomaly detection in the point-cloud domain involves applying ground-plane removal algorithms to reduce the search space. We used Patchwork++ [35] to remove the ground plane from the point-cloud data. While ground plane removal is effective at short ranges, its performance is reduced at longer distances and on roads with varying geometries. Under these conditions, Patchwork++ often results in many false positives or incorrectly segments objects as part of the ground plane. In addition, manually fine-tuning the parameters of such methods to adapt to different topographies is a very challenging task.
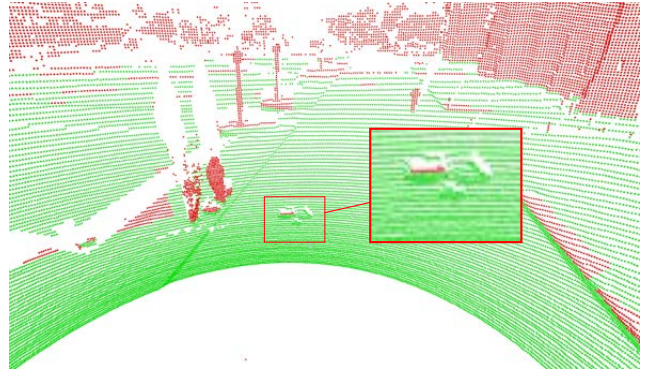
Figure 12. Patchwork++ performance in a wide environment.
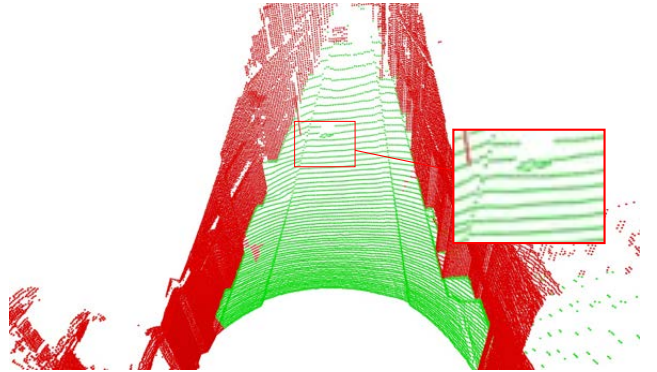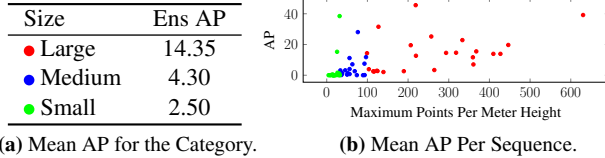

Figure 13. Patchwork++ performance in a narrow urban street.

8. Low Performance of the 3D Models

8.1. Relation of Performance to Distance and Size

We calculated the AP metric for different distance thresholds, as shown in Table 4. In the lower ranges, from 0 to 10, and from 10 to 20 range models perform better, then at other distances. Note, that methods evaluate on points

Size	Ens AP
● Large	14.35
● Medium	4.30
● Small	2.50

(a) Mean AP for the Category.

(b) Mean AP Per Sequence.

Figure 14. Deep Ensembles AP for differently sized objects over validation and test datasets.

Method	0–10m	10–20m	20–30m	30–40m	40–50m
Deep Ensemble [34]	7.63	8.49	3.42	0.38	0.03
MC Dropout [50]	0.16	0.53	0.06	0.04	0.01
Max Logit [24]	2.25	1.53	1.20	0.27	0.01
Void Classifier [4]	2.95	1.78	1.98	0.28	0.03
RbA [41]	1.85	1.28	0.73	0.15	0.01

Table 4. Anomaly segmentation performance per distance measured by AP.

within this range, treating points outside of the range as unlabeled, *i.e.* evaluation on 10–20 meters means that points at least as far as 10 meters are considered for evaluation. Expectedly, we see a decrease in performance as the distance to the anomaly increases. In addition, we looked at the relation between the size of the object and anomaly segmentation performance in a sequence. We observe a drop in performance for small objects in Figure 14b.

8.2. Number of Foreground Points

The class imbalance remains a challenge for Point-Level evaluation, as it is more pronounced in terms of the occupied space and number of points. If we compare to a SegmentMeIfYouCan setup (see Table 1 from [10]), our dataset has 0.03% of anomaly and 36.9% inlier points; that is twice as few anomaly points. In addition, we evaluate objects with at least 5 anomaly points (instead of 50 or more [10]). However, for some sequences, we observe performance similar to 2D methods, especially for large objects. On Figure 14a we show performance of the Ensemble method on a combined validation and test dataset for better illustration. Here, we split our data into sequences base on the effective size of an object. We divide the maximum number of points for an instance by the maximum height of an instance in a sequence, and separate sequences into three categories: sequence with an object that has $0-33$ points per meter, $33-99$ and $999+$ points per meter. We observe that deep ensembles perform better on sequences with larger objects.

9. Results on Validation Datasets

We show results for the SemanticKITTI [1] validation set in Table 6 and our dataset in Table 7. For the OOD validation set, we evaluate in three sequences and provide scores in Table 8.

Method	Aux Data	AUROC↑	FPR@95 ↓	AP↑
DenseHybrid [21]	✗	87.09	76.36	26.63
RbA [41]	✗	89.58	75.21	37.12
UNO [16]	✓	89.52	62.29	37.10
Mask2Anomaly [47]	✓	90.54	78.09	36.38

Table 5. Evaluation of 2D methods on our the validation set using only a front-view camera.

9.1. Evaluation of 2D Methods

We focus on automotive applications, where it is common practice to evaluate multimodal methods on 3D LiDAR annotations, since 3D distances to objects are important for driving. As a control experiment, we applied 2D-only methods to the frontal camera and evaluated only within the corresponding frustum of LiDAR points (see Table 5). Methods that use only RGB images have a higher false-positive rate in this setup compared to 2D benchmarks. We attribute this to a domain gap of the Cityscapes and SemanticKitti, partially because of label conventions, *i.e.*, parking lots or backs of traffic signs are "unlabeled" in Cityscapes but are "inlier" in SemanticKitti and these regions contribute to higher FPR. As well as to the LiDAR-Camera points misalignment, *i.e.* at image boundaries. However, in our dataset, anomalies appear in images from cameras with other perspectives, and evaluating a full 360-degree view would be more difficult.

10. Annotation and Qualitative Examples

We visualize the annotation interface with an example of a correctly annotated scene in the figure 15. Anomaly points are cyan, unlabeled regions are black, and inliers are purple. We provide further visualizations of the dataset and the predictions shown in the Figure 16.

11. SemanticKITTI Other-object Examples

Several examples of the other-object class in the SemanticKITTI dataset can be seen in Figure 17, Figure 18, and Figure 19. The other-object class consists of many miscellaneous items, including trash bins, advertisement posts, and small pots. We superimposed the class label on the front-facing camera, with light blue denoting the other object class.

12. Note on Training

Initially, jointly training with both SemanticKITTI and Panoptic-CUDAL led to diverging losses for Mask4Former-3D. This also occurred during training runs solely on Panoptic-CUDAL. Lowering the preset learning rate from 0.0004 to 0.0002 was enough to mitigate the loss divergence in both cases.

**Table 6.** Class-wise PQ scores on SemanticKITTI validation set.

Method	void	car	truck	bicycle	motorcycle	other vehicle	person	bicyclist	motorcyclist	road	sidewalk	parking	other ground	building	vegetation	trunk	terrain	fence	pole	traffic sign	PQ
Mask-PLS* [40]	–	94.12	83.19	44.55	61.44	59.36	77.08	91.64	0.0	93.94	78.67	37.19	0.0	86.99	84.85	53.28	57.06	21.35	59.85	53.62	59.9
Mask-PLS [40]	–	91.90	77.93	16.96	51.28	45.66	65.85	83.36	0.0	93.86	77.69	31.39	0.0	86.97	87.61	50.40	59.10	22.79	60.77	53.34	55.62
Mask4Former-3D* [60]	–	93.81	70.95	62.92	68.97	56.79	81.98	87.35	24.06	93.94	78.05	27.33	0.0	88.39	88.65	50.93	60.82	25.38	57.89	58.59	61.94
Mask4Former-3D	–	93.53	59.39	62.55	64.82	54.36	79.61	89.16	25.01	93.24	77.90	28.79	0.0	87.27	87.28	51.08	59.92	24.85	56.76	58.14	60.72
Mask4Former-3D-void	6.08	74.36	47.00	32.19	43.34	33.30	42.90	68.75	00.33	93.35	77.07	19.01	0.0	82.77	81.34	47.56	56.94	19.98	54.48	36.82	47.97

Table 7. Class-wise PQ scores on STU-inlier Validation Set.

| Method	void	car	truck	bicycle	person	road	sidewalk	parking	building	vegetation	trunk	terrain	fence	pole	traffic sign	PQ
Mask-PLS* [40]	–	78.88	1.89	0.0	70.92	56.34	26.18	0.0	54.96	74.69	1.36	55.20	48.00	41.51	44.51	39.60
Mask-PLS	–	78.66	22.16	0.0	75.33	81.77	41.46	0.0	78.79	89.16	49.92	25.53	46.53	56.79	66.94	50.93
Mask4Former-3D* [60]	–	78.45	11.29	10.59	69.44	59.07	41.99	0.27	84.70	88.46	0.0	0.0	36.51	25.92	57.50	42.80
Mask4Former-3D	–	80.99	37.28	47.65	80.99	71.46	17.74	0.0	84.08	89.73	29.34	30.79	47.6	59.62	60.96	52.73
Mask4Former-3D-void	0.07	23.88	20.78	1.01	43.30	38.24	20.03	11.11	48.45	43.09	20.20	17.31	30.80	27.26	33.16	26.96

Table 8. Anomaly Segmentation Performance on the Validation Set with Anomalies

Method	Point-Level OOD			Object-Level OOD				
	AUROC ↑	FPR@95 ↓	AP ↑	RecallQ	SQ	RQ	UQ	PQ
Deep Ensemble [34]	90.93	37.34	6.94	17.70	79.96	9.10	14.15	7.27
MC Dropout [50]	65.76	79.82	0.17	3.54	74.36	3.48	2.63	2.59
Max Logit [24]	87.27	68.76	2.02	26.64	79.26	2.06	21.12	1.63
Void Classifier [4]	89.77	79.50	2.62	17.35	81.27	8.98	14.10	7.30
RbA [41]	73.00	100.0	1.64	21.84	78.58	2.75	17.16	2.16
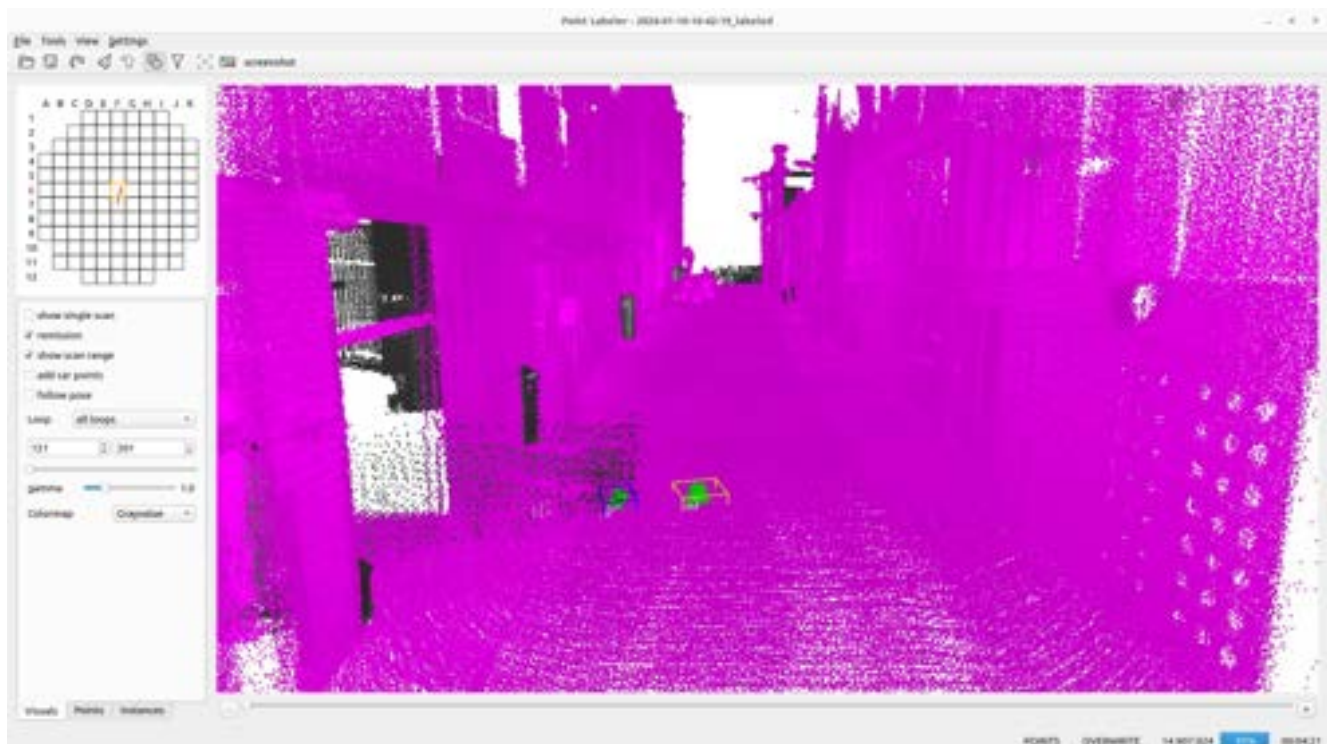
Figure 15. Data Annotation Example: Each color represents a specific label — Purple for inlier, Green for anomaly, and Black for void. Boxes represent instance boundaries.
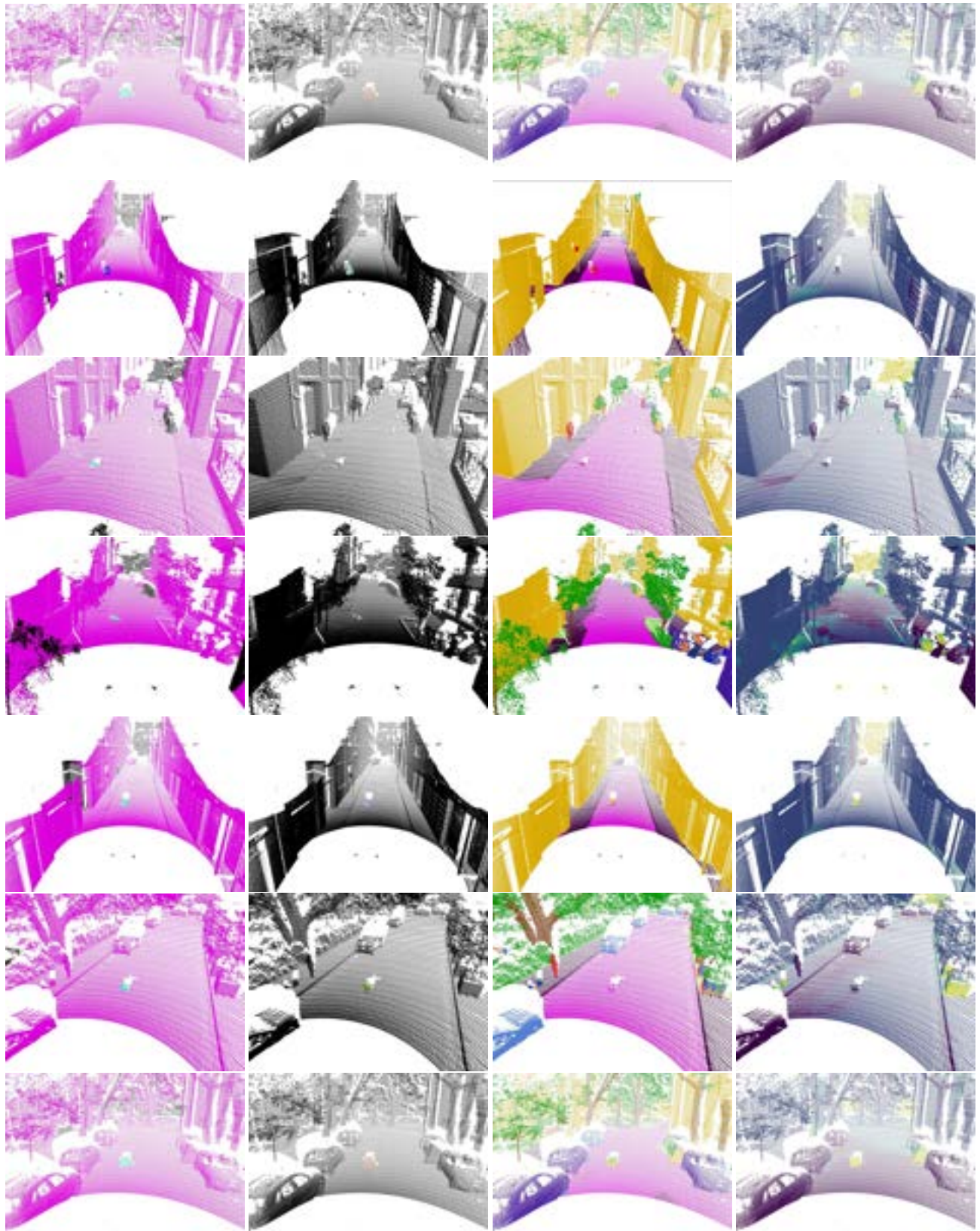
**(a)** Anomaly Label.   **(b)** Instance Label.   **(c)** Inlier Prediction.   **(d)** RBA.

Figure 16. Visualization of the proposed dataset with anomaly labels, instance labels, inlier class predictions, and anomaly scores of the selected anomaly methods.

Figure 17. Example of the other-object class: a billboard, a smaller billboard, a phone booth, and a small table, all of which belong to the other-object class.
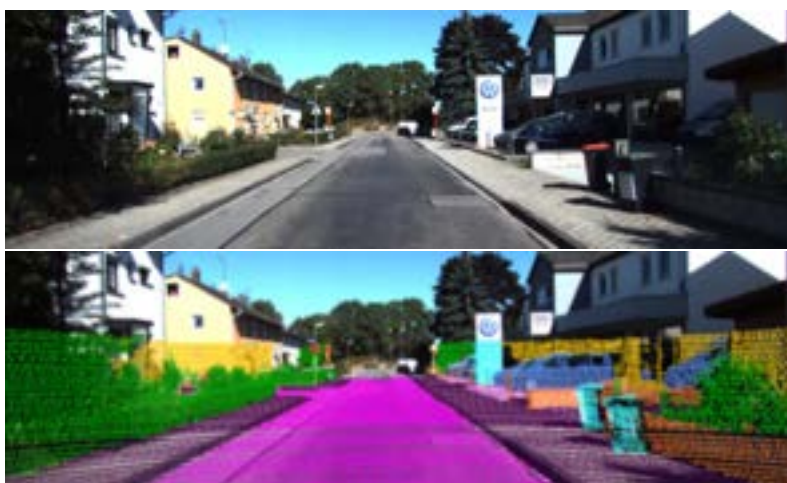
Figure 18. Example of the other-object class: a car dealership sign and two garbage cans, all belonging to the other-object class.
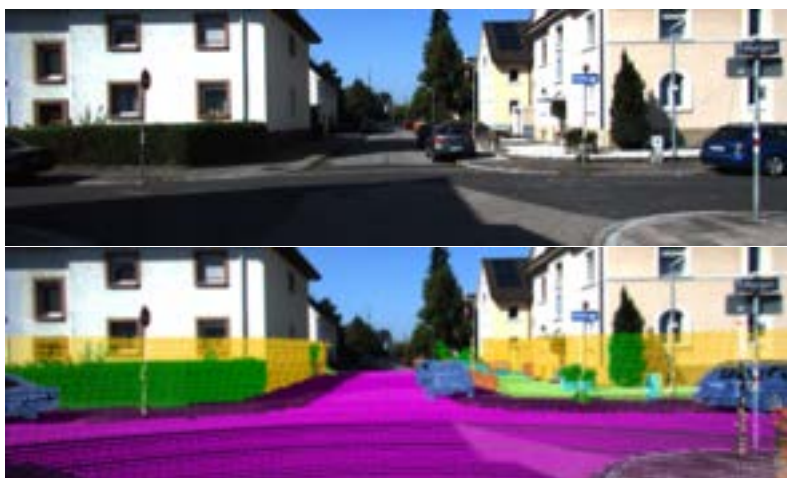
Figure 19. Example of the other-object class: a potted plant and a power adapter, all of which belong to the other-object class.