A BENCHMARK FOR KNOWING UNKNOWN THINGS IN MULTI-MODAL SENSOR STREAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable autonomous driving requires perception systems to recognize and consistently reason about situations that fall outside their training distribution. Such out-of-distribution (OoD) events are rare by nature, yet can be safety-critical and particularly challenging when objects are partially occluded, dynamically interacting with the scene, or observed differently across sensing modalities. Evaluation therefore requires benchmarks that extend beyond independent, single-modality frames. We introduce **Kn**owing **U**nknown **T**hings (KnUT), a multi-modal, sequence-based benchmark for evaluating OoD perception and tracking in dynamic driving scenes recorded with a cargo-bike sensor setup. KnUT provides synchronized camera and LiDAR recordings with OoD instances that are consistently identifiable across modalities and over time. The benchmark covers three complementary tasks: pixel-wise OoD detection in camera images, point-wise OoD detection in LiDAR point clouds, and temporal tracking of OoD instances. The data are recorded in live traffic such that OoD objects are routinely occluded by other road users and observed in the far distance. We release an annotated validation set and the complete evaluation code, while keeping the test annotations private for benchmark evaluation. In addition, we benchmark existing OoD methods to provide reproducible reference results. Overall, KnUT provides a unified benchmark for evaluating OoD perception across different perception modalities in both, static frames and aligned sequences.

1 INTRODUCTION

Neural networks excel in dense image recognition tasks, like semantic or instance segmentation (Hümmer et al., 2024; Cheng et al., 2022), when deployed in their training domain. Autonomous vehicles (AV), however, are deployed in the open world and can be confronted with instances from outside the semantic categories contained in the training data. To guarantee the AV's safe operation in the presence of such out-of-distribution (OoD) objects, OoD detection is a must-have. The development of reliable OoD detection methods requires data. Camera-based OoD datasets like Lost and Found (Pinggera et al., 2016), Fishyscapes (Blum et al., 2021), CAOS (Hendrycks et al., 2022) and SegmentMeIfYouCan (Chan et al., 2021a) responded to this need. Numerous OoD detection methods for semantic segmentation and instance segmentation have been developed since then, solving these benchmarks to a large extent (Delić et al., 2024).

With today's solid research landscape in OoD detection established, benchmark challenges are ready to represent more challenging and complex scenarios from the automated driving (AD) domain. First, a single frame-based OoD detection should be replaced by OoD detection on data streams. Second, AD is often based on multi-modal sensor setups, including LiDAR, RADAR and potentially further sensor modalities, instead of a single front camera optical sensor. Lastly, either existing datasets contain in-pasted objects (Blum et al., 2021; Hendrycks et al., 2022) or were recorded on country roads with few objects placed on an empty street (Pinggera et al., 2016; Chan et al., 2021a). This makes it hard to distinguish between genuine OoD detection performance, reaction to synthetic texture and foreground-background segmentation.

A number of datasets attempted to close this research gap. Maag et al. (2023); Nekrasov et al. (2025a) focus on OoD instance tracking, while Nekrasov et al. (2025a) adds LiDAR data as an additional modality. Bogdoll et al. (2024) contains dynamic scenes containing OoD objects, poses and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

actions, but is based on synthetic data generated by the CARLA AD simulator (Dosovitskiy et al., 2017), see section 2 for further discussion. None of these datasets, however, cover all shortcomings at once.

In this paper we present **Kn**owing **U**nknown **T**hings (KnUT) - a dataset covering camera, depth-camera, LiDAR and RADAR data in densely annotated data streams containing scenes in busy urban environments with traffic, occlusion and varying weather conditions, from sunny with sharp shadow contrasts, over snowy weather and a large variety of traffic participants. The data is processed with a newly designed labeling pipeline, containing in-domain pseudo labels from the state-of-the-art semantic segmentation model VLTSeg (Hümmer et al., 2024) with human revision and refinement, human-annotated OoD annotations in images and LiDAR point clouds and subsidiary depth-cam and RADAR data. Anonymization is built-in by license plate blurring and a face swap module. KnUT contains 14 sequences with $1{,}799$ labeled frames for validation and 71 sequences with $7{,}337$ frames with labels held private for independent benchmarking through an online submission system. KnUT develops several new metrics beyond frame to frame OoD detection, putting emphasis also on early detection in remote sensing tasks and consistent tracking. Our contributions can be summarized as follows:

1. We provide a new multi-modal (LiDAR, Camera with supplementary depth-cam and RADAR data) video OoD-Benchmark, called Knowing Unknown Things (KnUT), involving remote sensing scenarios, occlusion scenarios and classical OoD scenarios in busy urban environments with varying lighting conditions.
2. KnUT goes beyond single frame OoD detection and provides dense annotation in time allowing for OoD tracking tasks, supported by the development of new safety relevant metrics.
3. KnUT is based on an advanced annotation pipeline for images and LiDAR data developed specifically for this dataset, which also provides state-of-the-art pseudo labels, a human revision interface and modules guaranteeing anonymity. The annotation pipeline is published with the dataset.
4. State-of-the-art baselines are provided for camera and LiDAR standalone OoD detection and tracking.
5. KnUT comes with an online submission platform for testing on withheld private labels for independent evaluation.

2 Related Works

Table 1: Comparison of road driving OoD benchmarks. For multi-modal benchmarks we state which modality they label: camera (C) or LiDAR (L).

Sensor	Dataset	#Frames (test/val)	#Sequences (test/val)	Domain	OoD source	Complex scenes	Label	Tracking	Hidden test
Camera	Lost and Found	1068	5	Real	Staged	✗	Binary	✗	✗
	Fishyscapes L&F	275/100	–	Real	Staged	✗	Binary	✗	✓
	Fishyscapes Static	1000/30	–	Hybrid	Augmented	✗	Binary	✗	✓
	RoadAnomaly	60	–	Real	Web	✓	Binary	✗	✗
	SMIYC Anomaly	100/10	–	Real	Web	✗	Binary	✗	✓
	SMIYC Obstacle	382/30	90	Real	Staged	✗	Binary	✗	✓
	StreetHazards	1500	–	Synthetic	Simulated	✓	Binary	✗	✗
	MUAD	6501	–	Synthetic	Simulated	✓	Open-set	✗	✗
	CODA	1500	–	Real	Mined	✓	Bounding box	✗	✗
	OoDIS	787/140	–	Real	Re-annotated	✗	Binary instance	✗	✓
	ISSU static	980	–	Real	Natural	✓	Open-set	✗	✗
	PANIC	800	–	Real	Natural	✓	Open-world instance	✗	✓
	ISSU temporal	1140	–	Real	Natural	✓	Open-set	✗	✗
	SOS	1129	20	Real	Staged	✗	Binary instance	✓	✗
	WOS	938	44	Real	Staged	✗	Binary instance	✓	✗
	CWL	1210	26	Synthetic	Simulated	✗	Binary instance	✓	✗
	Tian et al.	132000	220	Synthetic	Simulated	✓	Open-set	✗	✗
	AsRO	12480	12	Real	Re-used	✓	Binary	✗	✗
Camera & LiDAR	LidarSOD	2919	15	Real	Staged	✓	Binary (C)	✗	✗
	AnoVox	conf.	–	Synthetic	Simulated	✓	Open-set instance (C + L)	✓	✗
	SDAC	–	–	Synthetic	Simulated	✓	panoptic (C) & BBox (L)	✗	✗
	STU	8022/1960	51/19	Real	Staged	✗	Binary instance (L)	✗	✓
	KnUT (ours)	7337/1799	71/14	Real	Staged	✓	Binary instance (C + L)	✓	✓

Most benchmarks for dense OoD detection in road driving evaluate single camera images. Only a few benchmarks provide labels for consecutive frames, and even fewer provide LiDAR data. The data of existing benchmarks can be real, synthetic, or a mix of both. Moreover, OoD objects can occur naturally or be staged. Often, when there is a mix of real and synthetic data, the OoD objects are simply pasted into in-distribution frames from other auxiliary datasets. The annotations differ in granularity across datasets, ranging from a binary anomaly mask to individually delineated instances, and from a single anomaly class to a full taxonomy in which unknown categories are distinguishable. Multi-modal benchmarks do not differ all modality. Lastly, only a few datasets provide instance labels across frames. We provide an overview of existing benchmarks in Table 1.

2.1 Camera-based OoD benchmarks

Lost and Found (Pinggera et al., 2016) was among the first to collect street scene sequences with small road hazards on the drivable area. **Fishyscapes** (Blum et al., 2021) builds on these recordings, adds a split with anomalies pasted into Cityscapes (Cordts et al., 2016) images, and introduces the protocol of releasing a small validation set while withholding test annotations behind a submission server, a convention we follow. **RoadAnomaly** (Lis et al., 2019) instead collects web images, covering a broader range of viewpoints, and **SegmentMeIfYouCan** (Chan et al., 2021a) splits the task into an anomaly track, covering the whole scene, and an obstacle track, restricted to the road surface. **StreetHazards** (Hendrycks et al., 2022) and **MUAD** (Franchi et al., 2022) render anomalies in simulation, which avoids the artifacts of pasted objects at the cost of a domain shift, while **CODA** (Li et al., 2022) mines real corner cases from KITTI (Geiger et al., 2013), nuScenes (Caesar et al., 2019) and ONCE (Mao et al., 2021). Later benchmarks enrich the annotation. **ISSU** (Laskar et al., 2025) labels in- and out-of-distribution classes jointly in cluttered scenes; its temporal split provides consecutive frames, but without instance identities across them. **OoDIS** (Nekrasov et al., 2025b) re-annotates Fishyscapes and SMIYC at instance level, and **PANIC** (Sodano et al., 2024) extends this to open-world panoptic segmentation.

A smaller group treats OoD perception as a temporal problem. **Street Obstacle Sequences (SOS)**, **Wuppertal Obstacle Sequences (WOS)** and **CARLA-WildLife (CWL)** (Maag et al., 2023) are among the earliest benchmarks for OoD tracking: SOS and WOS record a vehicle approaching an object placed on the street, restricted to the drivable area, and CWL provides a synthetic counterpart. A synthetic benchmark by Tian et al. (2024) introduces latency-aware streaming metrics and a temporal consistency metric, arguing that frame-wise evaluation ignores the time a method needs to react, though it evaluates pixel-wise throughout and withholds no test split. **AsRO** (Rai et al., 2025) is curated from Apolloscape (Huang et al., 2019) and provides by far the longest real-world sequences. A related similar line of work studies video anomaly detection (Ramachandra & Jones, 2020), where anomalous events are defined by appearance or motion deviations rather than by object identity.

2.2 OoD Benchmarks with LiDAR Data

Benchmarks that move beyond the camera remain rare, and even fewer annotate both modalities. **LidarSOD** (Singh et al., 2020) releases point clouds alongside the images, but uses them only to supply confidence maps to a monocular network; annotation is pixel-wise on the camera, into road, off-road and obstacle. **AnoVox** (Bogdoll et al., 2024) and **SDAC** (Gong et al., 2024) label both modalities, but inside a simulation, where cross-modal correspondence follows from the renderer rather than from annotation. **Spotting the Unexpected (STU)** (Nekrasov et al., 2025a) is the only prior real-world benchmark for 3D OoD segmentation, combining naturally encountered anomalies with deliberately placed objects. It labels the point cloud alone as inlier, outlier or ignored under the SemanticKITTI taxonomy (Behley et al., 2019), and evaluates single scans.

KnUT instead annotates OoD instances in both the image and the point cloud, and evaluates their association over time. It is also recorded in live traffic, whereas STU halts recording when a vehicle approaches, so its scenes contain no surrounding road users. We use a sparser LiDAR, reflecting lower-cost platforms, and treat the resulting difficulty as a benchmark property rather than a limitation. Altogether, KnUT is, to the best of our knowledge, the first real-world benchmark to jointly evaluate pixel-wise OoD detection, point-wise OoD detection and OoD tracking on the same sequences.

3 KNUT: KNOWING UNKNOWN THINGS

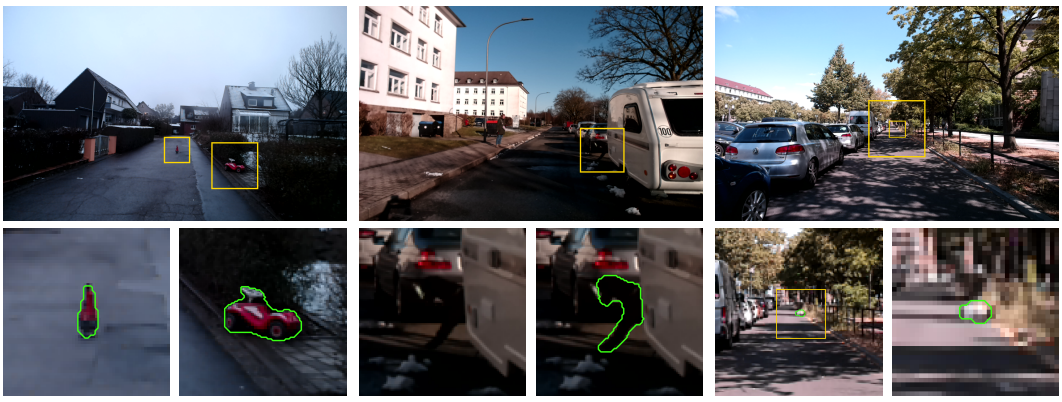
We present **Kn**owing **U**nknown **T**hings (KnUT), a real-world multi-modal benchmark for OoD detection and OoD tracking across diverse urban scenarios under varying lighting conditions. KnUT provides synchronized camera and LiDAR data recorded from a cargo bike navigating urban traffic, enabling evaluation in both 2D image space and 3D point cloud space. In this section we give an overview of the dataset, its recording and the annotation process. We further introduce new scenarios. We publish our code TBA and provide a Leaderboard TBA.

3.1 SENSOR SETUP

We use two sensors for data collection: an OAK-D LR Luxonis camera (lux), which provides RGB and stereo depth images, and a Blickfeld Cube 1 Outdoor LiDAR (bli). For reproducibility, we provide the dimensions and distances of the cargo box and the sensor setup in the Appendix in Figure 4. The cargo bike allows for recording typical urban driving scenes at speeds up to 20 km/h. The modular sensor setup on the cargo bike allows for quick, flexible deployment and recording across a wide variety of locations.

The camera has horizontal and vertical fields of view of $80°$ and $64°$, respectively; the corresponding LiDAR fields of view are $65°$ and $28°$. The LiDAR field of view therefore lies within the camera image, as illustrated in Figure 2a. The camera records RGB and stereo images at 786×480 pixels and 30 Hz, while the LiDAR records at 11.14 Hz. The LiDAR configuration can be manually adjusted and was set to achieve the highest possible frame rate while maintaining a sufficiently wide FOV. This configuration results in 48 scan lines with an angular spacing of $0.4°$. The scan lines are drawn in Figure 2b. We combine the upward- and downward-facing scan lines, showing in Figure 2c, resulting in pairs of scan lines with a small distance between them, followed by a larger gap to the next pair. Both sensors are integrated using ROS Noetic (Quigley et al., 2009) and are synchronized with a time difference of at most 20 ms. The recording is already synchronized. We calibrate the sensors using MFCalib (Ye et al., 2024).

3.2 SCENARIOS AND TASKS



(a) near OoD (classical scenario) (b) bad lighting & occluded OoD (c) distant OoD (remote sensing)

Figure 1: Added complexity for the OoD detection task in KnUT. (a) Similar to existing classical OoD datasets, KnUT contains scenes where objects are fully visible and appear at near distance. (b) Additionally, there are diverse lighting conditions, and OoD objects can overlap with other objects. (c) Also, in KnUT, OoD objects can appear at distances of up to 200m, where they are barely visible.

KnUT includes complex scenarios. In existing OoD segmentation datasets like SMIYC (Chan et al., 2021a), most OoD objects are clearly visible and non-overlapping. KnUT includes scenes with occlusion and distant, barely visible objects (also called remote sensing), which make detection and localization more difficult. Examples for these scenarios are shown in Figure 1. For all three scenarios, we provide three different types of tasks: OoD segmentation, OoD tracking and a new task called *early-recognition*. Early-recognition measures the initial frame in which an object is

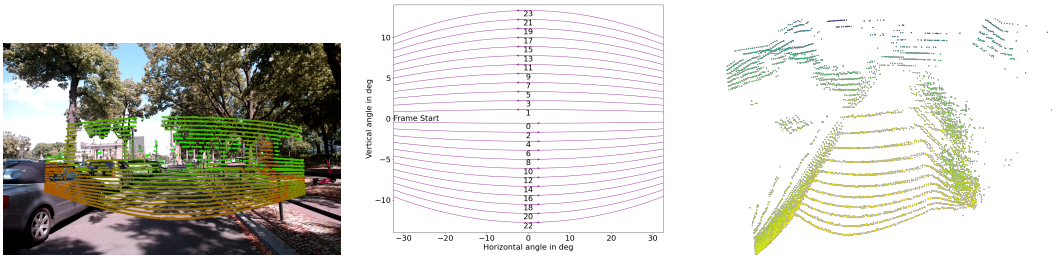
detected relative to the ground truth. For LiDAR, it additionally measures the distance at which the object is detected. From a safety-critical perspective, it is important not only to detect an object, but to detect it as early as possible, in order to react as fast as possible.

3.3 RECORDING KNUT

To obtain a diverse set of environments, recordings were conducted on 18 different roads with varying terrain, mainly asphalt and paving stones. The roads range from parking places to frequently used roads. Therefore, the road texture range from cobblestones, patched and repaired asphalt to clear asphalt. Additionally, also the environment surround of the roads ranges from vegetation, or buildings to frequently parked vehicle in a busy environment with lots of cars and vehicles. We provide a selection of roads and surrounding in Figure 10a. Most captured scenes include parked vehicles on at least one side of the street, which allows for capturing realistic and potentially hazardous scenarios, e.g., a ball rolling onto the street from behind a parked vehicle. The recordings cover different OoD situations, including remote sensing, occlusion, and classical OoD scenarios. Further details about the different scenarios are provided in section 3.2.

The recordings were performed during daytime, under varying lighting and cloud coverage, as well as on dry and snowy roads. Different lighting condition occur due to sunny or cloudy weather as well as shadows, ranging from building to movable shadows from vegetation. We show a collection of different lighting condition in Figure 10b. In total, 27 different OoD objects were used for the recordings. These objects can be divided into two categories. The first category contains objects that can appear in urban street environments but are typically not included in in-distribution road scene datasets such as Cityscapes (Cordts et al., 2016) or SemanticKITTI (Behley et al., 2019). Examples include animals, outdoor toys, or garden equipment. The second category contains objects that are typically found indoors, such as furniture and office equipment. The validation set provides 9 different OoD-categories, namely: Garden gnome, Bobby car, dog, stuffed animal, banner, gaming seat, air mattress, seat, and toolbox.

Individual recorded sequences range from 2 seconds to > 1 minute and can contain multiple OoD scenarios. During post-processing, these recordings are split into final sequences, each corresponding to a single OoD scenario.



(a) Camera-LiDAR overlay (b) Scan lines for LiDAR (c) Point Cloud of LiDAR data

Figure 2: LiDAR information. The LiDAR FOV (a) is smaller as the one for images, such that the whole LiDAR point cloud fits in the image. The scan line recording (b) show that 24 lines are used for up and 24 lines for downward resulting in 48 lines in total. The scan lines (c) are combined from upward- and downward facing resulting in two lines next to each other. The point cloud is matching with the image of (a).

3.4 DATA ANONYMIZATION

After recording, we anonymize the image data by replacing faces and blurring license plates. We use YOLO11 (Jocher & Qiu, 2024) to detect and track pedestrians and their faces. The detected faces are replaced using InsightFace (Ren et al., 2023; Guo et al., 2021; Gecer et al., 2021; An et al., 2022; 2021; Deng et al., 2020a;b; Guo et al., 2018; Deng et al., 2018; 2019) and faces from the public CelebA dataset (Liu et al., 2015). License plates are detected using SAM3 (Carion et al., 2025) and anonymized using Gaussian blurring. LiDAR data do not contain personally identifiable information and therefore do not require anonymization.

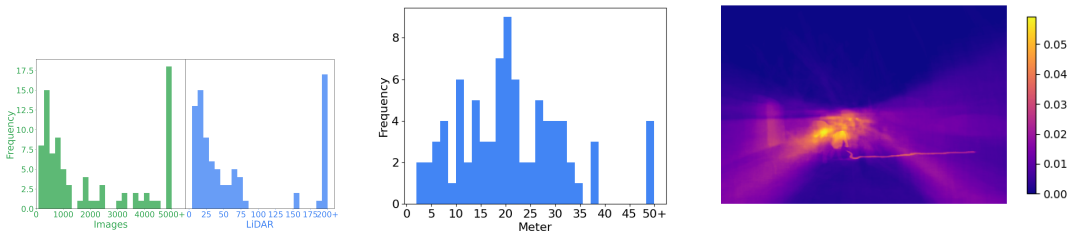
3.5 LABELING PIPELINE

We annotate the images using both semantic and instance segmentation. The annotation process is split into five steps: (1) pre-labeling images, (2) rectification of pre-labeling images in semantic segmentation and OoD instance annotation, (3) removing artifacts from pre-labeling, (4) label LiDAR with instance annotation, and (5) ap instance annotations to OoD annotations for LiDAR. For the pre-labeling VLTSeg (Hümmer et al., 2024) is used. We obtain a semantic segmentation of the KnUT images in the 19-class Cityscapes taxonomy. Misclassifications are re-annotated in the six most important category groups: road, sidewalk, terrain, bicycles, persons (+ pedestrian, rider) and vehicles (+ car, truck, caravan, bus, motorcycle). To replace those misclassifications we use SAM3 (Carion et al., 2025) and in failure cases Labelme (Wada et al., 2021) as a manual annotation tool. For instance segmentation of the OoD objects, we also use SAM3 or manual labeling backup. SAM3 also provides tracking capabilities, which helps to maintain consistent instance identities across frames. OoD objects can be distinguished by their individually assigned object IDs. Objects that are physically fixed together and form a single object, such as a stack of boxes, are annotated as one instance. Since the VLTSeg pre-annotations are not designed to represent OoD objects, they can contain incorrect regions around the OoD objects. We therefore apply a manually post-processing step, as the third step, that detects manually implausible class labels surrounding an OoD object and automatically reassigns these regions to the appropriate class. This removes artifacts from the final instance annotations.

After annotating the images, the labels were used as a pre-annotation for LiDAR data. Specifically we used the pre-annotations of the OoD objects, identify OoD instances with camera pixels to obtain an initial location of the OoD instances inside the LiDAR point cloud. To correct the asynchronization map error, we apply SegmentsAI (Segments.ai) to manually re-label the objects in a fourth step. The last step contains the mapping of instance LiDAR segmentation to OoD segmentation.

The labeling for both types of data were reviewed for a second time to verify that the labels contain no errors. A detailed description of the annotation pipeline is provided in Appendix A.2.

3.6 DATASET STATISTICS



(a) Average size of all OoD objects over all scenarios.

(b) Maximal distance in meters of all OoD objects over all scenarios.

(c) Distribution of OoD objects in the scenarios for image data.

Figure 3: Data statistic for LiDAR and image data. For image data the counting is in pixel, while for LiDAR data it is in points. The distribution is given for the image data, while the distance data are measured with the LiDAR.

KnUT contains a validation and a test set with LiDAR and camera data. The test data contains 7,337 frames in 71 sequences and the validation data 1,799 frames in 14 sequences. The test set will be kept private to ensure a consistent and fair evaluation for all participants. The leaderboard can be found at TBA. The validation set can be downloaded from TBA. We provide some examples of our KnUT dataset with annotations in Appendix C.1. KnUT contains sequences of a length between 22 and 332 frames. Furthermore, the OoD objects cover 0.735% of all pixels and 1.598% of all LiDAR point clouds. The amount of OoD instances per clip ranges from one to seven.

Figure 3c shows the distribution of the OoD objects over all data. The shift to the left is due to the position of the camera and LiDAR which is not exactly in the middle of the recording vehicle and the placement of the obstacles. The size of the objects differs on the one hand due to the object itself and on the other hand due to the distance to the object. The average object size is 1,791 pixels.The OoD

objects contain a size up to $49{,}703$ pixels which is around 12% of the whole frame. A distribution for images and LiDAR data are shown in Figure 3a. Most objects are relatively small, however there are a few large outliers, such that the distribution is bounded for the size. Furthermore, in Figure 3b we show the maximal distance over all OoD objects over all scenarios. The greatest distance is 200 meters. Additional, we provide in Figure 6 a size-distance ratio overview, showing that the size reduce with the distance.

4 Evaluation Metrics

We evaluate OoD perception on KnUT along two axes: frame-wise detection and temporal instance tracking which we extend with metrics for how early and how far away each instance is first recognized.

4.1 OoD Detection Metrics

For comparability with existing OoD segmentation benchmarks, we follow the evaluation protocol of ISSU (Laskar et al., 2025), which builds on the SegmentMeIfYouCan (SMIYC) benchmark (Chan et al., 2021a). At pixel level, OoD detection is a binary classification problem over all pixels, in-distribution against out-of-distribution. Since OoD pixels are heavily underrepresented, receiver-operating-characteristic metrics are uninformative under this imbalance. We therefore use Average Precision (AP) as the primary ranking metric. As complementary metrics, we report the False Positive Rate at 95% recall (FPR_{95}), since false positives in the high-recall regime matter for safety-critical applications, and the True Positive Rate at 5% false positive rate (TPR_5), which measures detection under a strict false-positive budget. At component level, we evaluate individual OoD segments. The adjusted segment-wise Intersection-over-Union (sIoU) measures how well a ground-truth segment is covered by predictions, the Positive Predictive Value (PPV) how well predicted segments match any ground-truth OoD region, and the component-wise F_1-score aggregates true positives, false positives, and false negatives. We report each averaged over thresholds $\{0.25, 0.30, \ldots, 0.75\}$, denoted $\overline{\mathrm{sIoU}}$, $\overline{\mathrm{PPV}}$, and $\overline{F_1}$.

For LiDAR, we report detection metrics at both point and object level. At point level, recall denotes the fraction of ground-truth OoD points correctly identified, and precision (denoted as tp/pp) the fraction of correctly identified OoD points among all points predicted as OoD. At object level, we use sIoU to quantify the overlap between predicted and ground-truth OoD objects.

4.2 OoD Instance Tracking Metrics

We evaluate the temporal consistency of predicted tracks using the multiple-object-tracking metrics MOTA (accuracy) and MOTP (precision) (Bernardin & Stiefelhagen, 2008). MOTA summarizes false positives, false negatives, and identity switches relative to the total number of ground-truth objects across the sequence. MOTP measures localization accuracy as the average distance between the geometric centers of matched predicted and ground-truth objects. We additionally report the proportions of mostly tracked (MT), partially tracked (PT), and mostly lost (ML) objects, where an object is mostly tracked if recovered in at least 80% of its frames, mostly lost if recovered in fewer than 20%, and partially tracked otherwise.

4.3 Early-recognition metrics

Standard detection and tracking metrics measure whether an OoD instance is recognized, but not how early or from what distance. In safety-critical settings such as autonomous driving, early recognition of an approaching unknown is as important as eventual detection. We therefore introduce two metrics that quantify the timeliness of OoD recognition along a sequence. **First view (FV)** measures how early an instance is first detected, defined as $\mathrm{FV} = 1 - (i - N_0)/(N_f - N_0)$, where N_0 and N_f are the first and last frame, respectively in which the instance is visible and i is the frame of first prediction. An instance counts as predicted once its IoU with the ground-truth exceeds 50%. **Highest distance (HD)** measures how far away an instance is first recognized, defined as $\mathrm{HD} = (d - D_f)/(D_0 - D_f)$, where D_0 and D_f are the minimum instance distances in the first and last visible frames and d is the distance at first prediction. We measure d over the intersection of

prediction and ground-truth, taking the minimum distance among those points, and assume distance decreases over time. Where it does not, we swap D_0 and D_f. Since distance is available only in 3D, we report HD for LiDAR alone. Both metrics are bounded in $[-1, 1]$, with higher values indicating earlier recognition, and we penalize an instance that is never detected with -1. For both, we distinguish small, middle, and large instances by the terciles of the validation split, averaged over each sequence.

5 EXPERIMENTS

We benchmark existing OoD methods on KnUT to establish reproducible reference points and to characterize the difficulty of the benchmark. We organize baselines by sensor modality, and evaluate each on the detection and tracking tasks.

5.1 BASELINE METHODS

We evaluate three methods introduced on SMIYC, each representing a different source of OoD signal: uncertainty estimation, outlier exposure, and image resynthesis. **Maximum Softmax Probability** (Hendrycks & Gimpel, 2017) derives a pixel-wise anomaly score from the confidence of a closed-set Cityscapes classifier (Cordts et al., 2016), without additional training or auxiliary data. **Entropy Maximization** (Chan et al., 2021b) fine-tunes the segmentation network to produce high-entropy predictions on auxiliary OoD data sampled from COCO (Lin et al., 2014). **SynBoost** (Di Biase et al., 2021) resynthesizes an image from the predicted segmentation and flags high-discrepancy regions, combining the resynthesis signal with uncertainty scores. All three are built on a DeepLabv3+ backbone (Chen et al., 2018). We additionally include **UNO** (Delić et al., 2024), which combines per-pixel uncertainty with a mask-level negative-objectness signal from Mask2Former (Cheng et al., 2022) and ranks among the strongest methods on the SMIYC leaderboard. We include the SMIYC baselines to test whether they transfer to KnUT, and UNO to test whether a method close to saturating SMIYC still faces substantial headroom in our setting.

For LiDAR, we evaluate **RbA** (Nayal et al., 2023), **Maximum Softmax Probability** (Hendrycks & Gimpel, 2017), and a naive entropy baseline computing the entropy from softmax values. The Maximum Softmax Probability is transferred directly from the camera setting to 3D. All three build on a Mask4Former3D backbone (Yilmaz et al., 2024). We follow the protocol of Nekrasov et al. (2025a) but replace their baselines, retaining only RbA. RbA treats each class-wise prediction as an independent binary classifier and assigns outliers to an additional class that captures objects fitting none of the known classes.

For both modalities, we adopt a lightweight, training-free tracking heuristic (Maag et al., 2020) that maps predicted OoD segments across consecutive frames by spatial proximity and segment overlap. For LiDAR, we cluster points using DBSCAN (Ester et al., 1996) and adapt the matching from 2D to 3D. We release all baseline code in the supplementary material.

5.2 NUMERICAL RESULTS

Table 2 reports baseline performance on the camera data. UNO detects OoD regions reliably, while the three SMIYC baselines perform poorly, trailing by at least 66 pp AP. We attribute this gap to complexity that KnUT introduces and SMIYC does not capture, including heavy occlusion, remote sensing, and dynamic scene interaction. Notably, even UNO leaves substantial headroom, which indicates that a method close to saturating SMIYC remains far from reliable OoD detection in our setting. This holds on the test split, where UNO leads detection by roughly 60 pp AP in comparison th the other baselines, but still reaches an AP of only 80.61.

Tracking exposes a trade-off between detection quality and track completeness. UNO tracks most reliably, with the lowest MOTP and the only non-negative MOTA, yet mostly tracks only 3 of the 18 objects. Maximum Softmax and SynBoost invert this, mostly tracking 11 and 12 objects but reaching strongly negative MOTA through frequent ID switches and spurious tracks. The pattern persists on the test split: UNO retains the lowest MOTP and the only positive MOTA, while the best MOTA on either split remains only marginally positive. We therefore find that OoD tracking on KnUT is essentially unsolved. For early-recognition, small objects are frequently never detected,

yielding negative sFV for three of the four baselines on validation and for all four on test. We attribute this to object size, since small objects occupy at most 0.04% of the image and provide little signal for early detection.

Table 2: Benchmark results on the camera data of KnUT.

Split	Method	Pixel-level OoD detection						OoD instance tracking							
		$\mathrm{FPR}_{95}\downarrow$	$\mathrm{TPR}_5\uparrow$	AP $\uparrow$	$\overline{\mathrm{F_1}}\uparrow$	$\overline{\mathrm{sIoU}}\uparrow$	$\overline{\mathrm{PPV}}\uparrow$	MOTP $\downarrow$	MOTA $\uparrow$	MT	PT	ML	$\overline{\mathrm{FV}}\uparrow$	hFv	miFV	sFV
Validation	Max. entropy	24.49	73.61	9.45	6.46	22.24	12.37	16.72	−4.71	7	6	5	−0.11	0.80	0.06	−0.54
	Max. softmax	49.74	27.31	2.43	0.90	1.10	4.29	32.37	−35.88	12	4	2	−0.62	−0.05	−1.00	−0.69
	SynBoost	35.58	66.63	7.84	5.80	44.38	5.31	18.78	−14.72	11	7	0	0.67	0.95	0.91	0.46
	UNO	3.35	96.30	75.47	47.44	54.02	40.11	2.57	0.21	3	8	7	0.14	0.29	0.41	−0.03
Test	Max. entropy	52.06	68.14	19.15	6.87	16.27	12.72	29.14	−3.34	24	65	34	−0.20	0.38	−0.30	−0.42
	Max. softmax	44.64	36.87	7.77	0.93	8.59	41.67	41.67	−23.23	74	37	12	−0.67	−0.39	−0.73	−0.75
	SynBoost	39.54	64.81	19.95	6.16	23.96	10.85	22.46	−8.31	78	35	10	0.10	0.38	0.10	−0.05
	UNO	8.02	94.16	80.61	50.87	50.35	48.17	7.11	0.35	24	47	52	0.03	0.06	0.09	−0.06

KnUT LiDAR data. Table 3 reports baseline performance on the LiDAR data. RbA is the strongest baseline, with an $\overline{\mathrm{sIoU}}$ of 60.63 and a recall of 88.71%. The low $\mathrm{tp/pp}$ ratios across all methods reveal a large gap between predicted-positive and true-positive points. Tracking fails across all methods: MOTP exceeds 180 throughout, and only RbA tracks any objects reliably (5 of 18). For early-recognition, RbA again leads on every metric. These conclusions carry over to the test split, where RbA remains the strongest detector, tracking again fails for every method with MOTP now above 260, and RbA continues to lead on early-recognition. Qualitative examples are provided in Appendix C.2.

Table 3: Benchmark results on the LiDAR data of KnUT.

Split	Method	Point-level OoD detection			OoD instance tracking										
		tp/pp $\uparrow$	recall $\uparrow$	$\overline{\mathrm{sIoU}}\uparrow$	MOTP $\downarrow$	MT	PT	ML	$\overline{\mathrm{FV}}\uparrow$	hFV	miFV	sFV	$\overline{\mathrm{HD}}\uparrow$	hHD	miHD	sHD
Validation	RbA	3.55	88.71	60.63	182.80	5	12	1	0.53	0.92	0.33	0.48	0.59	0.65	0.65	0.28
	Entropy	6.16	48.38	45.27	201.23	0	8	10	0.06	0.56	0.01	−0.05	0.11	0.43	−0.05	0.13
	Max. softmax	5.31	42.86	38.24	215.47	0	9	9	0.16	0.51	0.10	0.08	0.25	0.17	0.33	0.08
Test	RbA	4.37	82.88	48.90	264.74	41	74	6	0.23	0.66	0.34	0.03	0.31	0.80	0.18	0.04
	Entropy	6.46	44.38	30.13	266.89	0	51	70	−0.02	0.19	0.06	−0.13	0.02	0.07	0.00	0.00
	Max. softmax	5.59	40.84	27.32	271.89	1	72	48	0.08	0.24	0.29	−0.08	0.14	0.25	0.10	0.09

6 Conclusion and Outlook

We introduce a new multi-modal, sequence-based real-world OoD dataset, called Knowing Unknown Things (KnUT). From this results a semantic and instance annotated benchmark, containing 14 validation sequences and 71 test sequences. KnUT is recorded in the real world, and contains scenarios such as occlusion or remote sensing scenarios. We provide semantic annotations for images and OoD instance masks for both, images and LiDAR data. Furthermore, the KnUT benchmark involves OoD detection and tracking baselines accompanied by novel early-recognition metrics measuring first detection frame and distance. For LiDAR data, we include three different standard methods for reference performance. For images, our benchmark consists of four different methods, containing top-performing methods from other OoD benchmarks.

We measured poor performance for all included baselines in at least one evaluation attesting the increased complexity of KnUT. With KnUT, we reach a new level of real-world OoD data containing a variety of real world scenarios. We encourage other researchers to submit their methods and improve the OoD detection and tracking to a new level helping to make autonomous driving safer. In the future we want to improve end-to-end models to get awareness of those scenarios, which can be tested on our dataset. Toward the end-to-end applications in real-world it becomes a higher importance of having also datasets with actions labels and time awareness. We serve the early-recognition metric as a first step towards the time awareness.

AI USE STATEMENT

We used generative AI tools for grammar and spelling corrections, as well as for literature search and discovery during the preparation of the related work. All information obtained with AI assistance was independently verified by the authors against the original sources. Generative AI tools were also used for code assistance, specifically to improve the LiDAR tracking and adapt the existing image-tracking code to LiDAR data. All other code was written manually or reused from publicly available sources, which are referenced in the respective scripts. Generative AI was not used for the recording setup, data collection, or data annotation, these processes were conducted entirely by the authors. The authors reviewed and verified all AI-assisted contributions and take full responsibility for the final content of this work.

ETHICS STATEMENT

We introduce a dataset which has contained during the recording personal information, like license plates or human faces. As we stated in section 3.1 we blurred the license plates and swap the faces by using classical machine learning models. Afterwards, a person reviewed the whole process and anonymized manually all missed faces or license plates. We can guarantee to the best of our knowledge, that it is with classical methods at the moment not possible to recover personal information from our data. The recording vehicle, visible in Figure 4, contains signs to inform that we record data as additional a QR-Code and written information, where persons they did not wanted to be recorded can contradict against their filming and we can directly delete corresponding scenes. As we stated, we were always aware of other traffic users and never endanger them. We highly encourage that with our dataset we want to make autonomous driving safety.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the evaluation code and implementations of the evaluated baselines. The supplementary material further includes the annotation code and annotation guidelines used to ensure a consistent labeling process, as well as the anonymization code. Example validation data are included in the supplementary material. Following the review process, we will release the complete validation data together with the corresponding annotations. The test data will remain private to ensure an independent evaluation, and a public leaderboard will be provided after the review process. Details of the recording setup and sensor configuration are described in the paper to facilitate the construction of comparable recording systems.

REFERENCES

`https://www.blickfeld.com/wp-content/uploads/2022/10/blickfeld_Datasheet_Cube1-Outdoor_v1.1.pdf`. [Online, accessed 18-Aug-2026].

`https://docs.luxonis.com/hardware/products/OAK-D%20LR`. [Online, accessed 18-Aug-2026].

Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. Partial fc: Training 10 million identities on a single machine. In *ICCVW*, 2021.

Xiang An, Jiangkang Deng, Jia Guo, Ziyong Feng, Xuhan Zhu, Yang Jing, and Liu Tongliang. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *CVPR*, 2022.

J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008.

Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, 129(11):3119–3135, Nov 2021. ISSN 1573-1405. doi: 10.1007/s11263-021-01511-6. URL https://doi.org/10.1007/s11263-021-01511-6.

Daniel Bogdoll, Iramm Hamdard, Lukas Namgyu Rößler, Felix Geisler, Muhammed Bayram, Felix Wang, Jan Imhof, Miguel De Campos, Anushervon Tabarov, Yitian Yang, et al. Anovox: A benchmark for multimodal anomaly detection in autonomous driving. In *European Conference on Computer Vision*, pp. 206–223. Springer, 2024.

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. *arXiv preprint arXiv:1903.11027*, 2019.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021a.

Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5108–5117. IEEE, 2021b.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299, 2022.

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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Anja Delić, Matej Grcic, and Siniša Šegvić. Outlier detection by ensembling uncertainty with negative objectness. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024. URL https://papers.bmvc2024.org/0779.pdf.

Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, 2018.

Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*, 2020a.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020b.

Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16918–16927, 2021. doi: 10.1109/CVPR46437.2021.01664.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg (eds.), *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 1–16. PMLR, 13–15 Nov 2017. URL https://proceedings.mlr.press/v78/dosovitskiy17a.html.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231. AAAI Press, 1996.

Gianni Franchi, Xuanlong Yu, Andrei Bursuc, Ángel Tena, Rémi Kazmierczak, Séverine Dubuisson, Emanuel Aldea, and David Filliat. Muad: Multiple uncertainties for autonomous driving, a benchmark for multiple uncertainty types and tasks. In *British Machine Vision Conference*, 2022.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets, 2021. URL https://arxiv.org/abs/1803.09010.

Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Ostec: One-shot texture completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

Lei Gong, Yu Zhang, Yingqing Xia, Yanyong Zhang, and Jianmin Ji. Sdac: A multimodal synthetic dataset for anomaly and corner case detection in autonomous driving. In *AAAI Conference on Artificial Intelligence*, 2024.

Jia Guo, Jiankang Deng, Niannan Xue, and Stefanos Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. In *BMVC*, 2018.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

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)*, 2017.

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.

Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2702–2719, 2019.

Christoph Hümmer, Manuel Schwonberg, Liangwei Zhou, Hu Cao, Alois Knoll, and Hanno Gottschalk. Strong but simple: A baseline for domain generalized dense perception by clip-based transfer learning. In *ACCV*, 2024.

Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. URL https://github.com/ultralytics/ultralytics.

Zakaria Laskar, Tomáš Vojíř, Matej Grcic, Iaroslav Melekhov, Shankar Gangisetty, Juho Kannala, Jiri Matas, Giorgos Tolias, and C.V. Jawahar. A Dataset for Semantic Segmentation in the Presence of Unknowns. In *CVPR*, 2025.

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*, pp. 406–423. Springer, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2152–2161, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Kira Maag, Matthias Rottmann, and Hanno Gottschalk. Time-dynamic estimates of the reliability of deep semantic segmentation networks. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 502–509, 2020. doi: 10.1109/ICTAI50040.2020.00084.

Kira Maag, Robin Chan, Svenja Uhlemeyer, Kamil Kowol, and Hanno Gottschalk. *Two Video Data Sets for Tracking and Retrieval of Out of Distribution Objects*, pp. 476–494. Springer Nature Switzerland, 2023. ISBN 9783031263484. doi: 10.1007/978-3-031-26348-4_28. URL http://dx.doi.org/10.1007/978-3-031-26348-4_28.

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, 2021. URL https://arxiv.org/abs/2106.11037.

Nazir Nayal, Mısra Yavuz, João F. Henriques, and Fatma Güney. Rba: Segmenting unknown regions rejected by all, 2023. URL https://arxiv.org/abs/2211.14293.

Alexey Nekrasov, Malcolm Burdorf, Stewart Worrall, Bastian Leibe, and Julie Stephany Berrio Perez. Spotting the Unexpected (STU): A 3D LiDAR Dataset for Anomaly Segmentation in Autonomous Driving. In *"Conference on Computer Vision and Pattern Recognition (CVPR)"*, 2025a.

Alexey Nekrasov, Rui Zhou, Miriam Ackermann, Alexander Hermans, Bastian Leibe, and Matthias Rottmann. Oodis: Anomaly instance segmentation and detection benchmark. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2764–2771, 2025b. doi: 10.1109/ICRA55743.2025.11128111.

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 *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1099–1106. IEEE, 2016.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Ng. Ros: an open-source robot operating system. volume 3, 01 2009.

Shyam Nandan Rai, Shyamgopal Karthik, Mariana-Iuliana Georgescu, Barbara Caputo, Carlo Masone, and Zeynep Akata. Road obstacle video segmentation. In *DAGM German Conference on Pattern Recognition*, pp. 186–201. Springer, 2025.

Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2569–2578, 2020.

Xingyu Ren, Alexandros Lattas, Baris Gecer, Jiankang Deng, Chao Ma, and Xiaokang Yang. Facial geometric detail recovery via implicit representation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 2023.

Segments.ai. Segments.ai: Multi-sensor labeling platform for robotics and autonomous vehicles. URL https://segments.ai.

Aasheesh Singh, Aditya Kamireddypalli, Vineet Gandhi, and K Madhava Krishna. Lidar guided small obstacle segmentation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8513–8520. IEEE, 2020.

M. Sodano, F. Magistri, J. Behley, and C. Stachniss. Open-World Panoptic Segmentation. arXiv:2412.12740, 2024.

Beiwen Tian, Huan-ang Gao, Leiyao Cui, Yupeng Zheng, Lan Luo, Baofeng Wang, Rong Zhi, Guyue Zhou, and Hao Zhao. Latency-aware road anomaly segmentation in videos: A photorealistic dataset and new metrics. *arXiv preprint arXiv:2401.04942*, 2024.

Kentaro Wada, Mpitid, Martijn Buijs, N. Zhang Ch., , Bc. Martin Kubovčík, Alex Myczko, Latentix, Lingjie Zhu, Naoya Yamaguchi, Shohei Fujii, Iamgd67, IlyaOvodov, Akshar Patel, Christian Clauss, Eisoku Kuroiwa, Roger Iyengar, Sergei Shilin, Tanya Malygina, Kento Kawaharazuka, Jonne Engelberts, Aleksi J, AlexMa, Changwoo Song, , Charlie, Daniel Rose, Douglas Livingstone, , Doug, , Erik, and Henrik Toft. wkentaro/labelme: v4.6.0, 2021. URL https://zenodo.org/record/5711226.

Tianyong Ye, Wei Xu, Chunran Zheng, and Yukang Cui. Mfcalib: Single-shot and automatic extrinsic calibration for lidar and camera in targetless environments based on multi-feature edge. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 864–871, 2024. doi: 10.1109/IROS58592.2024.10802395.

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.

Appendix

This appendix provides additional information about dataset acquisition and annotation, split construction and composition, qualitative benchmark results, the evaluation-server protocol, auxiliary sensor streams, and the KnUT dataset card.

A Dataset Construction and Reproducibility

A.1 Sensor rig

Figure 4 documents the acquisition platform and the relative placement of the camera and LiDAR. KnUT is recorded with a bicycle instead of a car, since it is cheaper, has less emissions while able to drive at the same speed as cars in most cities drive.

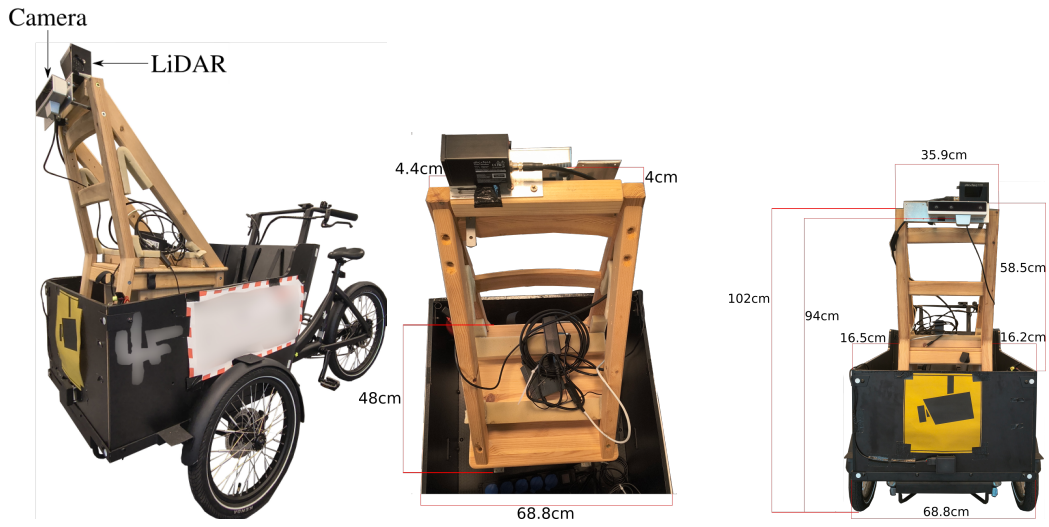


Figure 4: Acquisition platform and sensor mounting geometry. (a) Cargo bicycle with the removable sensor rig mounted in the cargo box. (b) Top view showing the camera and LiDAR origins and their lateral and longitudinal displacement. (c) Front view showing sensor heights and lateral offsets.

A.2 Annotation pipeline

We provide a detailed description of our annotation procedure, as an extension of Section 3.5. Figure 5 shows a diagram of our annotation pipeline, described below.

Pre-labeling We fine-tune VLTSeg (Hümmer et al., 2024) on Cityscapes images with reduced sizes of 819×409. We use this VLTSeg for pre-labeling to the Cityscapes (Cordts et al., 2016) taxonomy since it has shown good generalization under domain shifts. However, by manual inspection we observed miss-classification errors in the predictions of VLTSeg. For reducing the annotation effort and focus on the main points of OoD and autonomous driving we refrain from re-annotating everything. Instead, we focus on six main class groups relevant for the driving domain: road, sidewalk, terrain, vehicles, persons and bicycles. Inside of vehicles, we group car, truck, trailer, caravans and motorcycles and for persons we include pedestrians as well as riders. In the following step, we relabel in the area of that six classes and the OoD area.

Labeling We use SAM3 (Carion et al., 2025) as our labeling tool, since it works well for object segmentation and most foreground detection. Additionally, sequences can be used as input and

therefore, tracking objects is possible. We use SAM3 in a clicking mode. For reproducibility we also add the code for it in the supplementary material. SAM3 replaced misclassified classes, especially in the foreground well. However, for background object or for difficult labeled objects, due to lighting condition as well as for worse distinguishable objects SAM3 fails. Therefore, we used Labelme (Wada et al., 2021) for manual corrections.

Removing artifacts In cases where an object is re-classified or new annotated via SAM3 or manual corrections, failures on the border of those re-labeled object can occur. Therefore, we use a manual pipeline, producing a mask of miss-classification and insert the correct mask via nearest-neighbor approach.

LiDAR annotations The instance segmentation for images is used as a pre-labeling for our LiDAR data. We map the LiDAR point clouds into the image space. Therefore, we obtain a point-wise instance mapping for the LiDAR points. However, the synchronization is only up to 20 ms. It is possible, that due to the moving of the recording vehicle, mapped class points can mapped to wrong instances. Especially on the border or on fare away objects this happens. We use SegmentsAI (Segments.ai) for manual labeling, since there is no similar method like SAM3 for 3D point clouds. After the manual labeling for LiDAR point clouds, the instance annotation is finished.

As a last step, we remap all instances to the OoD class, obtaining a semantic version for OoD objects.

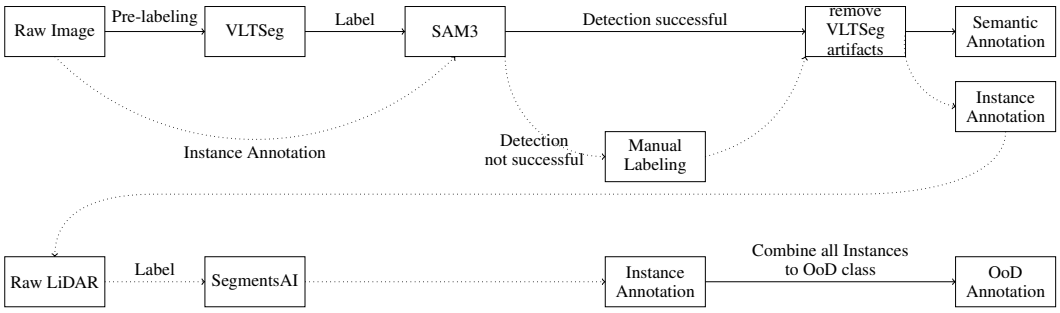


Figure 5: Labeling pipeline for Image and LiDAR data. Raw images are pre-labeled by VLTSeg(Hümmer et al., 2024). Wrong segments or OoD objects are re-annotated by SAM3 (Carion et al., 2025). If SAM3 already shows good results, OoD segments are removed onside their shaped by artifacts from VLTSeg. If SAM3 produces errors, or did not detect all segments, manual labeling by (Wada et al., 2021) is needed. After removing artifacts semantic annotations and instance annotations for OoD objects for Image data are finished. The instance annotations serves as an input for LiDAR data. Those instance as re-labeled manually by SegmentsAI (Segments.ai). For OoD semantic annotation all instances get the same class.

B Dataset Analysis

Figure 6 shows the ratio between a size of an object, here in points, and its distance. Most objects have a small size and a small distance, up to 30 meters. However, the dataset contains also large objects with a small distance, as well as small objects with a large distance.

C Qualitative Examples and Baselines

C.1 OoD Scenarios

We introduce three types of OoD scenarios: Occlusion, remote sensing, and classical OoD scenes, define as, an object is only partially covered from a known class, and the distance of the OoD instances is less than 20 meters. Examples of those tasks are shown in Figure 7. In contrast to Figure 1, Figure 7 shows additionally the semantic annotation as well as the instance annotation for the OoD objects.

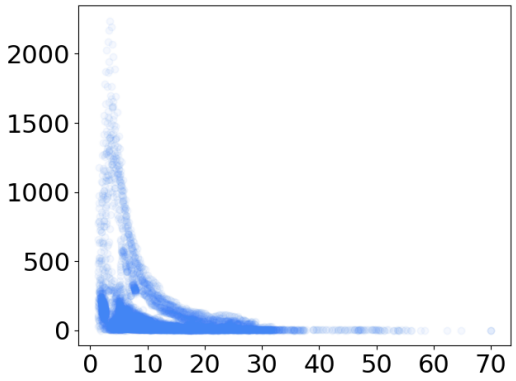


Figure 6: Ratio between OoD object size and distance, showing for LiDAR point clouds.

C.2 Visualization of Baseline Results

Figures 8 and 9 show an exemplary output of all baseline methods. As brighter the color in the heatmaps, as higher is the uncertainty score. The LiDAR data are mapped into the image for better visibility only. For images, maximized softmax predicts more the shapes as uncertain as the object itself. For maximized entropy the surrounding of the OoD object as well as the OoD object itself has a high uncertainty score. From this it follows, that maximized entropy predict the OoD objects, however it also predict a lot of non-OoD objects as OoD objects. For SynBoost, less surrounding is predicted as OoD. UNO shows the strongest performance and show that the object can be predicted as OoD without having a strong uncertainty prediction in surrounding. However, UNO shows also that the surrounding is much more complex. For LiDAR RbA shows that all objects except the road are detected as OoD. However for maximized softmax and entropy only less points provide a high uncertainty score. This indicates that maximized softmax and entropy overlook the true OoD object, while RbA predict almost anything as OoD. The example shows also the different and difficult lighting condition which occur normally in the real-world. It complicates the task, also for a human to detect the object.

C.3 Diversity of KnUT Data

Figure 10 provides an overview of the diversity of our data. We only publish the validation data, however the test data contains further different lighting conditions and road diversity.

C.4 Depth Images and RADAR point clouds

We provide further unlabeled data containing depth-images from the camera and RADAR point clouds. Figures 11 provide the output of the camera used for the recordings. RGB-image and depth-image are synchronized. The depth-image is computed over a stereo camera, containing two monocular cameras. Over disparity the depth-estimation is computed and saved. For further information we refer to the documentation of Luxonis (lux). The RADAR provides 40 horizontal points, providing information about its position, motion (only if the object is moving or not) and velocity. The radar data are saved in a pickle format. We highly encourage researchers to use this data and test how its performance are. We discard from using RADAR, since it is difficult to detect if an object is OoD or not, from obtaining only few information, or at most one point of the object. Furthermore, the distance information is already provided by the LiDAR image. However, those data can be used for a lot of different other research fields in the OoD detection domain.

D Evaluation Server and Auxiliary Modalities

After the review process, we will provide a Leaderboard to test methods on the withhold test data. Depending on the metric the methods can be adjust in their order, by clicking to any metric. The

team name of the submission team and the submission date is announced. We provide separate metrics for LiDAR and images, however it is possible to use both data for a fusion.

E Dataset Card

We follow (Gebru et al., 2021) for answer questions regarding the dataset recording and conditions.

E.0.1 Motivation

- **For what purpose was the dataset created?** KnUT provides an anomaly dataset and is multi-modal and sequence-based. The dataset can be used to evaluate new methods in more realistic scenes for autonomous driving.
- **Who created the dataset and on behalf of which entitiy?** Will be answered after reviewing process.
- **Who funded the creation of the dataset** Will be answered after the reviewing process.
- **Any other comments** The dataset provides complex scenarios and evaluates detection and tracking tasks.

E.0.2 Compositions

- **What do the instances that comprise the dataset represent?** Real-world scenarios in sequence-based format.
- **How many instances are there in total?** We provide 14 validation sequences with 1799 frames and 71 test sequences with 7337 frames for evaluation.
- **Does the dataset contain all possible instances or is it a sample of instances from a larger set?** It contains all possible instances, however not all will be published. The test data will be withhold for evaluation.
- **What data does each instance consist of?** Sequence-based image and LiDAR point clouds.
- **Is there a label or target associated with each instance** For images we provide semantic and instance OoD annotations. For LiDAR point clouds we provide instance OoD annotations.
- **Is any information missing from individual instances?** No.
- **Are relationships between individual instances made explicit?** Yes. OoD instances are assigned consistent instance IDs across frames, making the temporal correspondence of individual objects explicit. For the multi-modal data, corresponding OoD instances are also consistently identified across camera and LiDAR data.
- **Are there recommended data splits?** Yes, we provide a public validation set and a private test set.
- **Are there any errors, sources of noise, or redundancies in the dataset?** Minor annotation errors may remain despite manual verification and correction. LiDAR measurements can also contain noise or sparse observations due to the sensor characteristics and object distance. No intentional redundancies are included in the dataset.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources?** Yes, it is self-contained.
- **Does the dataset contain data that might be considered confidential?** No.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** No.
- **Does the dataset identify any subpopulations?** No.
- **Is it possible to identify individuals, either directly or indirectly from the dataset?** No, we anonymized all persons and personal information such as license plates.
- **Does the dataset contain data that might be considered sensitive in any way?** No, we anonymized all persons and personal information such as license plates.
- **Any other comments?** No.

E.0.3 COLLECTION PROCESS

- **How was the data associated with each instance acquired?** Directly observable.
- **What mechanisms or procedures were used to collect the data?** We used a camera and LiDAR sensor. Additionally we have a RADAR sensor producing RADAR points.
- **If the dataset is a sample from a larger set, what was the sampling strategy?** We have no larger set.
- **Who was involved in the data collection process and how were they compensated?** Will be answered after reviewing process.
- **Over what timeframe was the data collected?** Over more than 9 months.
- **Were any ethical review processes conducted?** No.
- **Did you collect the data from the individuals in question directly, or obtain it via third paries or other sources?** The data were collected directly by the dataset authors through our own recording sessions. No data were obtained from third-party datasets or external sources.
- **Were the individuals in question notified about the data collection?** Yes, data collection was indicated by signs placed on the recording vehicle, informing people in the surrounding area that recording was taking place.
- **Did the individuals in question consent to the collection and us of their data?** The consent is silent agreement of the roads, and can be contradict via informing the authors. We provide a contradiction website which can be directly loaded during the recording. Those how are directly involved in the recording e.g. by placing object on the road sign an agreement.
- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain use?** Yes, we provide a contradiction website and informed the persons about it.
- **Has an analysis of the potential impact of the dataset and its use on data subject been conducted?** No.

E.0.4 PREPROCESSING/CLEANING/LABELING

- **Was any preprocessing/cleaning/labeling of the data done?** We label the data for semantic segmentation (images) and instance segmentation for OoD objects (images and LiDAR). Detailed descriptions are done in section 3.5
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?** Yes, we provide it after review process.
- **Is the software that was used to preprocess/clean/label the data available?** Yes, we provide it in the supplementary material.
- **Any other comments?** No.

E.0.5 USES

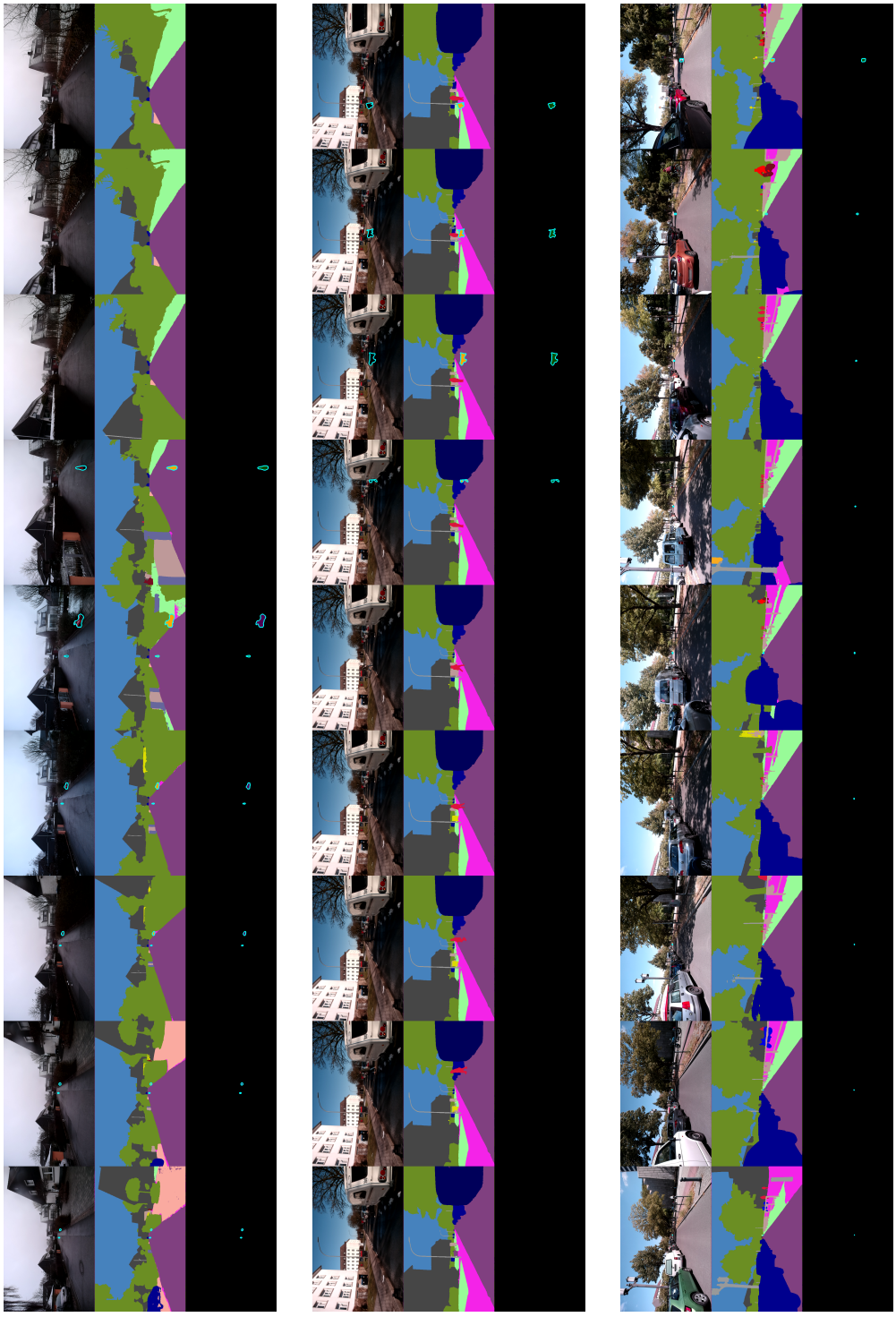
- **Has the dataset been used for any tasks already?** We provide a baseline, however no other task were applied to this dataset.
- **Is there a repository that links to any or all papers or systems that use the dataset?** We provide a Leaderboard and repository after the reviewing process.
- **What (other) tasks could the dataset be used for?** All tasks, where LiDAR or image data are needed, e.g. Gaussian splatting.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** No.
- **Are there tasks for which the dataset should not be used?** No.
- **Any other comments?** No.

E.0.6 DISTRIBUTION

- **Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?** Yes, we will publish it for the research community.
- **How will the dataset will be distributed?** Over common dataset platforms.
- **When will the dataset be distributed?** After the reviewing process.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** By a CC BY 4.0 license.
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** No.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** No.
- **Any other comments?** No.

E.0.7 MAINTENANCE

- **Who will be supporting/hosting/maintaining the dataset?** We will answer it after reviewing process.
- **How can the owner/curator/manager of the dataset be contacted?** We will provide a contact formula in the Leaderboard and on the download website.
- **Is there an erratum?** No.
- **Will the dataset be updated?** Not yet planned.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?** The dataset does not relate to people.
- **Will older versions of the dataset continue to be supported/hosted/maintained?** There are no older versions, a updated version is not planned.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** The validation data are public and can be used for it. The test data will be not published at all.
- **Any other comments?** No.



(a) Classical OoD (b) Occlusion (c) Remote sensing

Figure 7: Three different kinds of OoD scenarios: Classical OoD, where the distance of each object to the recording vehicle is smaller than 20 meters, Occlusion where at least one OoD object is covered by another (not necessarily OoD) object, Remote sensing, where an OoD object is fare away from the recording vehicle and it drives towards the object.



(a) RGB images

(b) Annotated OoD object for image data

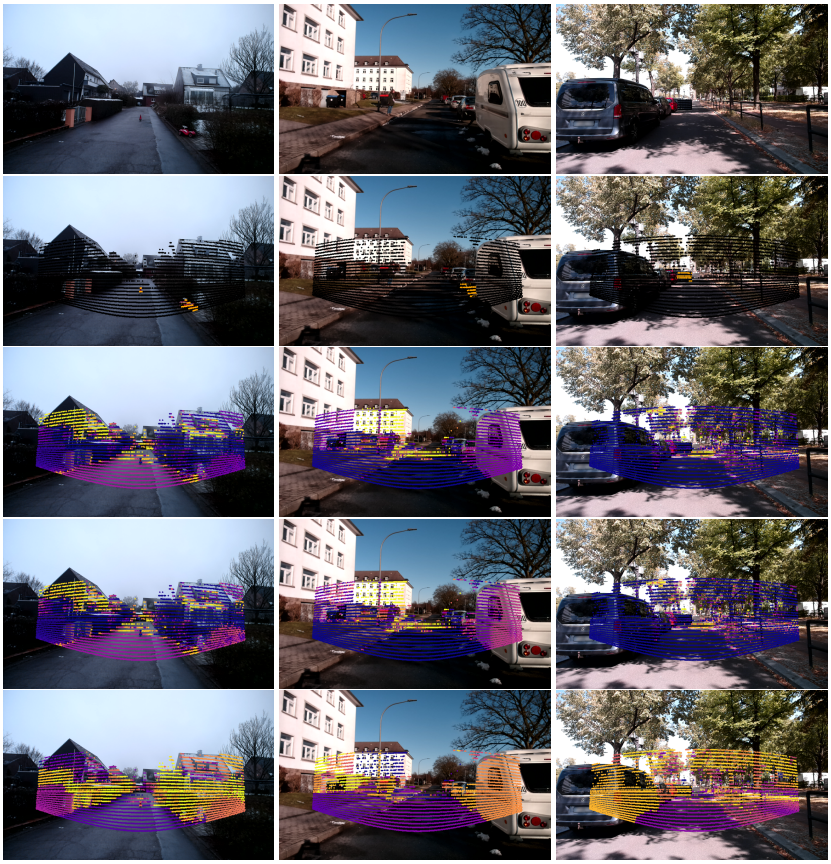
(c) Maximized Entropy

(d) Maximized Softmax

(e) Synboost

(f) UNO

Figure 8: Random picked frame with OoD-object showing baseline methods results. As brighter the color, as higher the uncertainty value.



(a) RGB images

(b) Annotated OoD object for LiDAR data

(c) Entropy

(d) Maximized Softmax

(e) RbA

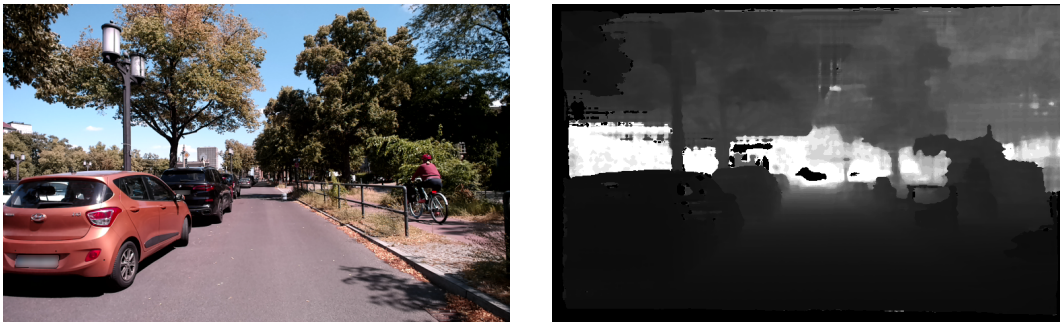
Figure 9: Random picked frame with OoD-object showing baseline methods results. As brighter the color, as higher the uncertainty value. For LiDAR data, we map the 3D point clouds into the image space, to obtain an image connection.



(a) Different texture of road and surroundings for validation data.

(b) Different lighting condition for validation data.

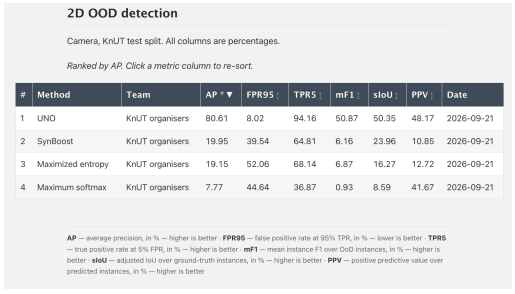
Figure 10: Diversity of KnUT data. The examples only providing the validation data. However, in the test data different roads and more lighting condition are used. Diversity of surrounding and roads are shown in (a). Different lighting condition, depending on weather as well as shadows are shown in (b).



(a) RGB-image from camera

(b) Depth-image from camera

Figure 11: Images recorded by the camera. (a) Single camera recording RGB images. (b) Depth-image recorded by stereo monocular camera, synchronized and calibrated automatically to the RGB images.



2D OOD detection

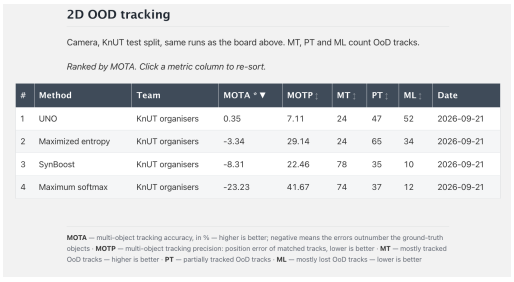
Camera, KnUT test split. All columns are percentages.

Ranked by AP. Click a metric column to re-sort.

#	Method	Team	AP * ▼	FPR95	TPR5	mF1	sIoU	PPV	Date
1	UNO	KnUT organisers	80.61	8.02	94.16	50.87	50.35	48.17	2026-09-21
2	SynBoost	KnUT organisers	19.95	39.54	64.81	6.16	23.96	10.85	2026-09-21
3	Maximized entropy	KnUT organisers	19.15	52.06	68.14	6.87	16.27	12.72	2026-09-21
4	Maximum softmax	KnUT organisers	7.77	44.64	36.87	0.93	8.59	41.67	2026-09-21

AP — average precision, in % — higher is better · **FPR95** — false positive rate at 95% TPR, in % — lower is better · **TPR5** — true positive rate at 5% FPR, in % — higher is better · **mF1** — mean instance F1 over OoD instances, in % — higher is better · **sIoU** — adjusted IoU over ground-truth instances, in % — higher is better · **PPV** — positive predictive value over predicted instances, in % — higher is better

(a) Image detection



2D OOD tracking

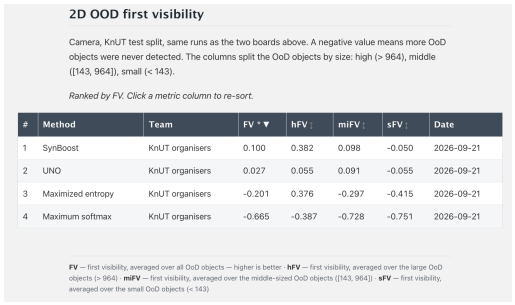
Camera, KnUT test split, same runs as the board above. MT, PT and ML count OoD tracks.

Ranked by MOTA. Click a metric column to re-sort.

#	Method	Team	MOTA * ▼	MOTP	MT	PT	ML	Date
1	UNO	KnUT organisers	0.35	7.11	24	47	52	2026-09-21
2	Maximized entropy	KnUT organisers	-3.34	29.14	24	65	34	2026-09-21
3	SynBoost	KnUT organisers	-8.31	22.46	78	35	10	2026-09-21
4	Maximum softmax	KnUT organisers	-23.23	41.67	74	37	12	2026-09-21

MOTA — multi-object tracking accuracy, in % — higher is better; negative means the errors outnumber the ground-truth objects · **MOTP** — multi-object tracking precision: position error of matched tracks, lower is better · **MT** — mostly tracked OoD tracks — higher is better · **PT** — partially tracked OoD tracks · **ML** — mostly lost OoD tracks — lower is better

(b) Image tracking



2D OOD first visibility

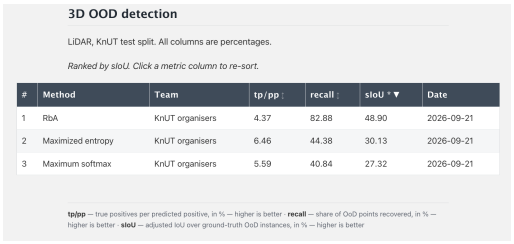
Camera, KnUT test split, same runs as the two boards above. A negative value means more OoD objects were never detected. The columns split the OoD objects by size: high (> 964), middle ([143, 964]), small (< 143).

Ranked by FV. Click a metric column to re-sort.

#	Method	Team	FV * ▼	hFV	miFV	sFV	Date
1	SynBoost	KnUT organisers	0.100	0.382	0.098	-0.050	2026-09-21
2	UNO	KnUT organisers	0.027	0.055	0.091	-0.055	2026-09-21
3	Maximized entropy	KnUT organisers	-0.201	0.376	-0.297	-0.415	2026-09-21
4	Maximum softmax	KnUT organisers	-0.665	-0.387	-0.728	-0.751	2026-09-21

FV — first visibility, averaged over all OoD objects — higher is better · **hFV** — first visibility, averaged over the large OoD objects (> 964) · **miFV** — first visibility, averaged over the middle-sized OoD objects ([143, 964]) · **sFV** — first visibility, averaged over the small OoD objects (< 143)

(c) Image early-recognition



3D OOD detection

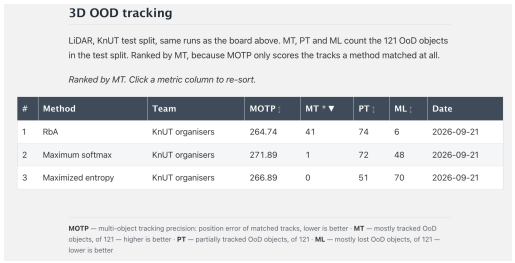
LiDAR, KnUT test split. All columns are percentages.

Ranked by sIoU. Click a metric column to re-sort.

#	Method	Team	tp/pp	recall	sIoU * ▼	Date
1	RbA	KnUT organisers	4.37	82.88	48.90	2026-09-21
2	Maximized entropy	KnUT organisers	6.46	44.38	30.13	2026-09-21
3	Maximum softmax	KnUT organisers	5.59	40.84	27.32	2026-09-21

tp/pp — true positives per predicted positive, in % — higher is better · **recall** — share of OoD points recovered, in % — higher is better · **sIoU** — adjusted IoU over ground-truth OoD instances, in % — higher is better

(d) LiDAR detection



3D OOD tracking

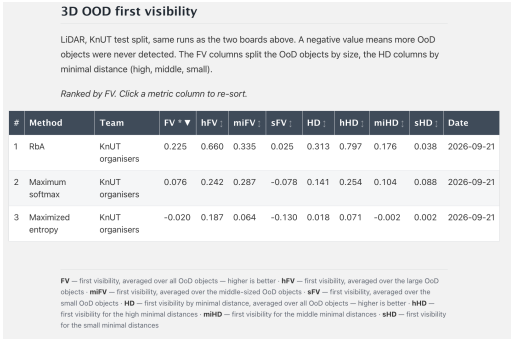
LiDAR, KnUT test split, same runs as the board above. MT, PT and ML count the 121 OoD objects in the test split. Ranked by MT, because MOTP only scores the tracks a method matched at all.

Ranked by MT. Click a metric column to re-sort.

#	Method	Team	MOTP	MT * ▼	PT	ML	Date
1	RbA	KnUT organisers	264.74	41	74	6	2026-09-21
2	Maximum softmax	KnUT organisers	271.89	1	72	48	2026-09-21
3	Maximized entropy	KnUT organisers	266.89	0	51	70	2026-09-21

MOTP — multi-object tracking precision: position error of matched tracks, lower is better · **MT** — mostly tracked OoD objects, of 121 — higher is better · **PT** — partially tracked OoD objects, of 121 · **ML** — mostly lost OoD objects, of 121 — lower is better

(e) LiDAR tracking



3D OOD first visibility

LiDAR, KnUT test split, same runs as the two boards above. A negative value means more OoD objects were never detected. The FV columns split the OoD objects by size, the HD columns by minimal distance (high, middle, small).

Ranked by FV. Click a metric column to re-sort.

#	Method	Team	FV * ▼	hFV	miFV	sFV	HD	hHD	miHD	sHD	Date
1	RbA	KnUT organisers	0.225	0.660	0.335	0.025	0.313	0.797	0.176	0.038	2026-09-21
2	Maximum softmax	KnUT organisers	0.076	0.242	0.287	-0.078	0.141	0.254	0.104	0.088	2026-09-21
3	Maximized entropy	KnUT organisers	-0.020	0.187	0.064	-0.130	0.018	0.071	-0.002	0.002	2026-09-21

FV — first visibility, averaged over all OoD objects — higher is better · **hFV** — first visibility, averaged over the large OoD objects · **miFV** — first visibility, averaged over the middle-sized OoD objects · **sFV** — first visibility, averaged over the small OoD objects · **HD** — first visibility by minimal distance, averaged over all OoD objects — higher is better · **hHD** — first visibility for the high minimal distances · **miHD** — first visibility for the middle minimal distances · **sHD** — first visibility for the small minimal distances

(f) LiDAR early-recognition

Figure 12: Leaderboard screenshots of all metrics. We provide single leaderboard baselines for LiDAR and images.