EVIOCC: PRESERVING VISUAL EVIDENCE FOR CROSS-DATASET 3D OCCUPANCY ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous vehicles need to locate unexpected obstacles even when their categories are absent from training data. Occupancy anomaly prediction addresses this need by combining scene geometry and semantics with voxel-level anomaly detection. However, the image features used to detect unfamiliar objects also change with camera geometry and scene appearance. When a model is transferred across datasets, these changes can cause normal objects to be flagged as anomalies. Addressing these errors through additional data collection, annotation, and model adaptation for each new camera configuration or environment is costly. We therefore introduce EviOcc, a source-only framework that predicts semantic occupancy and anomaly scores from a single RGB image and camera calibration. Using matching patches in resized and cropped versions of the same source image, EviOcc learns normal features expected at each patch's distance, viewing direction, and sampling scale. Deviations from these expected features provide anomaly scores for individual patches before feature pooling. Estimated depth locates each patch in 3D, where local kernels predict semantic occupancy and distribute patch anomaly scores to voxels. Directly projected scores are also retained, even at voxels predicted as free. Experiments across four training sources and three anomaly benchmarks demonstrate strong same-family and cross-dataset anomaly detection without target-domain adaptation, together with competitive geometric and semantic occupancy accuracy on source-domain and same-family evaluations.

1 INTRODUCTION

3D semantic occupancy prediction represents occupied and free space together with semantic labels, supporting motion planning in autonomous driving (Li et al., 2024). Road scenes can also contain unexpected obstacles, such as fallen cargo, whose categories are absent from training data (Chan et al., 2021). Localizing their spatial extent remains important even when their semantic identity is unknown. Occupancy anomaly prediction addresses this need by combining semantic occupancy prediction with voxel-level anomaly detection.

Camera-based methods infer dense geometry and semantics from images, including regions beyond direct observation (Cao & De Charette, 2022; Li et al., 2023; Yu et al., 2024). OccOoD combines complementary voxel and bird's-eye-view representations to improve semantic occupancy prediction and anomaly detection (Zhang et al., 2025). ProOOD further improves completion and rare-class representations through prototype-guided refinement, and combines logit coherence with prototype matching for anomaly scoring (Zhang et al., 2026). Both methods derive anomaly scores from representations and semantic predictions produced by the occupancy network. However, an unfamiliar feature response need not indicate an unfamiliar object: it may also reflect a change in how a normal object is observed.

Such observation changes arise naturally across cameras and environments. Differences in camera geometry and scene appearance can alter the features of normal objects (Li et al., 2024; Baek et al., 2024), potentially causing false anomaly responses (Yang et al., 2023; Gao et al., 2024). The challenge is therefore to detect unfamiliar objects without mistaking changes in normal observations for

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.

anomalies. Addressing these changes through additional data collection, dense 3D annotation, and model adaptation for each new camera configuration or environment is costly. We therefore study source-only cross-dataset generalization: each model is trained and selected using a single source dataset, then evaluated on other datasets without target-domain adaptation.

To distinguish unfamiliar objects from normal objects seen under different conditions, we propose **EviOcc**, a source-only framework that predicts semantic occupancy and voxel-level anomaly scores from a single RGB image and camera calibration. Our *sampling-aware evidence measure lifting* uses estimated depth to place image patch features in 3D and weights each patch by the spatial area it represents. It pools these features for occupancy prediction while keeping individual features for anomaly detection. Our *paired observation-conditioned normality* learns normal feature variations using matching patches in resized and cropped versions of the same source image. A patch receives a higher anomaly score when its feature differs from normal features expected at its distance, viewing direction, and sampling scale. In parallel, *geometry-tied local evidence completion* uses the pooled features to predict occupancy and semantic labels through local 3D kernels. These kernels also distribute patch anomaly scores to the voxels they cover. The final voxel score combines these distributed scores with directly projected patch scores, which remain even when a voxel is predicted as free.

Experiments across four training sources (Behley et al., 2019; Li et al., 2024) and three anomaly benchmarks (Zhang et al., 2025) demonstrate strong anomaly detection in both same-family and cross-dataset settings, together with competitive geometric and semantic occupancy accuracy on source-domain and same-family evaluations.

2 Related Work

Camera-Based Semantic Occupancy Prediction. Camera-based methods infer geometry and semantics using voxel or planar representations (Cao & De Charette, 2022; Huang et al., 2023; Zhang et al., 2023; Wei et al., 2023). Depth- and coordinate-aware lifting (Li et al., 2023; Mei et al., 2024; Miao et al., 2023; Yao et al., 2023), contextual and instance reasoning (Jiang et al., 2024; Yu et al., 2024; Li et al., 2026a), and distillation (Zheng et al., 2024) improve completion. Other approaches explore self-supervision (Huang et al., 2024a; Gan et al., 2025), sparse Gaussian representations (Huang et al., 2024b; 2025), and visibility-aware completion (Lu et al., 2025). EviOcc instead scores observations before pooling and couples occupancy prediction with anomaly propagation through shared local kernels, while preserving direct scores even in voxels predicted as free.

Anomaly Detection in Scene Understanding. Road-scene anomaly segmentation uses calibrated logits and energy (Jung et al., 2021; Tian et al., 2022), or mask-based scoring (Nayal et al., 2023; Rai et al., 2023). LiDAR methods identify unknown objects among observed points (Cen et al., 2022; Xu et al., 2025; Deng et al., 2025; Li et al., 2026c; Mosco et al., 2026; Li et al., 2026b); STU provides a real-world anomaly benchmark (Nekrasov et al., 2025). Occupancy approaches combine voxel–BEV refinement with entropy/similarity scoring (Zhang et al., 2025), or prototype-guided refinement and scoring (Zhang et al., 2026). Feature reconstruction learns normality while suppressing identity shortcuts (You et al., 2022), and robust segmentation distinguishes semantic from domain shifts (Gao et al., 2024). For source-only cross-dataset detection, EviOcc learns observation-conditioned normality from paired source resamplings, scoring deviations from expected normal features under each observation condition independently of predicted class probabilities.

3 Method

3.1 Overview

EviOcc predicts occupancy probabilities over free space and K source classes, together with voxel anomaly scores, from one RGB image and calibration using source-only training. Fig. 1 shows two paths: image features are aggregated in 3D to predict occupancy through local kernels, while individual patches are scored against normal features expected under their observation conditions. Paired source views teach how sampling changes normal features. The kernels propagate patch scores to voxels, while directly projected scores remain even at voxels predicted as free.

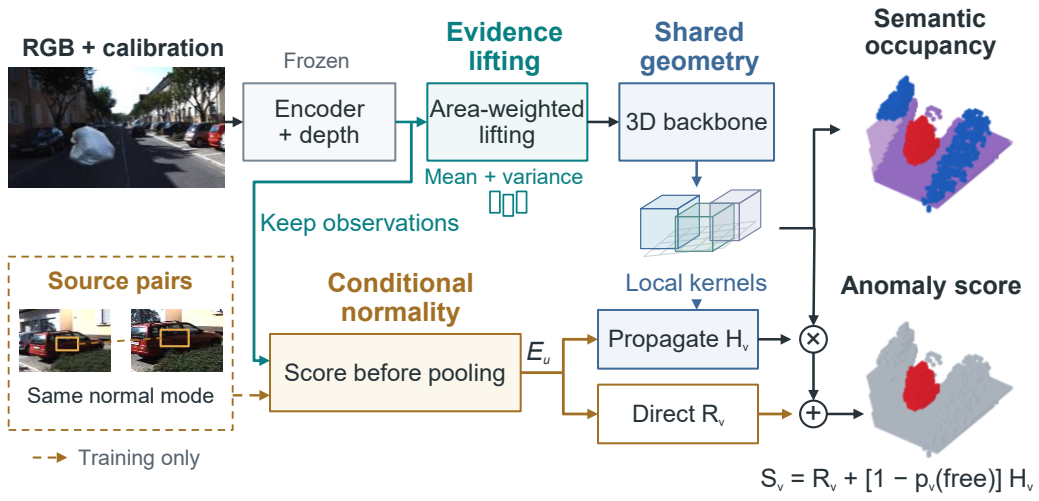

Figure 1: **Overview of EviOcc.** Frozen models extract patch features and depth from one RGB image. Using calibration, area-weighted lifting aggregates features in 3D, and a 3D backbone predicts local kernels for semantic occupancy. In parallel, retained patches are scored before pooling against normal features expected under their observation conditions. During training, matching patches from resized and cropped source images share a normal appearance mode. Patch scores reach voxels through direct projection (R_v) and kernel propagation (H_v). Only propagated scores are weighted by occupied probability, preserving direct evidence even at voxels predicted as free.

3.2 Sampling-Aware Evidence Measure Lifting

To avoid averaging away anomaly cues, we retain individual observations while aggregating features for occupancy prediction. Frozen DINOv3 (Siméoni et al., 2025) provides patch features $\mathbf{f}_u \in \mathbb{R}^d$ with norm $\sqrt{d}$. Frozen depth estimation (Lin et al., 2025) and calibration locate each patch at $\mathbf{x}_u$ in 3D.

Patches represent different spatial areas depending on depth and camera intrinsics. We weight them by the camera-normal footprint approximation $A_u = z_u^2 \Delta p_x \Delta p_y / (f_x f_y)$, where z_u is camera-axis depth, $\Delta p_x, \Delta p_y$ are patch dimensions in pixels, and f_x, f_y are focal lengths. A descriptor $\mathbf{q}_u$ records range, ray direction, footprint, and local depth variation, normalized using source statistics, to condition normality modeling.

For occupancy, a learned projection adapts patch features, and a lifting head uses image features and geometry to assign soft voxel weights along each ray, including an out-of-grid outcome. This lets occupancy supervision adjust feature placement. Multiplying assignments by A_u and normalizing per voxel produces weighted feature means and elementwise variances: the mean summarizes appearance, while variance exposes feature disagreement. Both feed a shared 3D backbone; empty voxels receive zeros.

For anomaly scoring, fixed trilinear splatting at $\mathbf{x}_u$, weighted by A_u and normalized per voxel, assigns each original $(\mathbf{f}_u, \mathbf{q}_u)$ to nearby voxels with weights b^0_{vu}. This *evidence measure* preserves individual observations independently of learned aggregation; unobserved voxels have empty measures.

3.3 Geometry-Tied Local Evidence Completion

A local kernel weights locations within a learned 3D region. Sharing these regions across occupancy prediction, normality learning, and score propagation lets occupancy labels constrain where anomaly evidence spreads.

The backbone predicts one tent kernel κ_r per coarse grid cell: a unit-integral product of axis-wise triangular functions, with its center inside the cell and positive axis-wise scales setting its nonzero support. Each kernel has nonnegative strength α_r and a joint distribution π_{rkm} over occupied classes k and normal modes m, representing feature patterns within a class. Its contribution to voxel volume $\mathcal{V}_v$ is $\lambda_{vr} = \alpha_r \int_{\mathcal{V}_v} \kappa_r(\mathbf{x})\, d\mathbf{x}$, with relative weight $a_{vr} = \lambda_{vr} / \sum_j \lambda_{vj}$. Total mass determines occupancy, while relative contributions mix class probabilities:

$$\begin{aligned} p_v(\text{free}) &= \exp\left(-\sum_r \lambda_{vr}\right), \\ p_v(k) &= \big(1 - p_v(\text{free})\big) \sum_r a_{vr} \sum_m \pi_{rkm}. \end{aligned} \tag{1}$$

Zero mass gives $a_{vr} = 0$ and $p_v(\text{free}) = 1$. Kernels need not match whole objects.

Weights $b_{ur} \propto \alpha_r \kappa_r(\mathbf{x}_u)$, normalized over kernels, connect patches to kernels. They provide mode weights during training and pool patch scores at inference; a_{vr} distributes kernel scores to voxels. Scores pass only through kernels containing the observation and intersecting the voxel. Unsupported patches have $b_{ur} = 0$.

3.4 Paired Observation-Conditioned Normality

To distinguish sampling changes from anomalies, we model normal features under their observation conditions. Each class k has M normal modes with density $\phi_{km}(\mathbf{f}, \mathbf{q}) = \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_{km}(\mathbf{q}), \nu(\mathbf{q})\mathbf{I})$, where $\mathbf{I}$ is the identity. The mean starts from a learned reference for each class and mode. A shared low-dimensional multilayer perceptron encodes $\mathbf{q}$; a class–mode-specific linear map converts its output into a reference adjustment. The resulting mean is rescaled to norm $\sqrt{d}$. Positive variance $\nu(\mathbf{q})$ describes expected spread and is shared across classes and modes. Both depend only on $\mathbf{q}$, preventing the tested feature from adjusting its own reference distribution.

Resizing, cropping, and resampling a source image produce paired views with updated calibration. Matching locations valid in both views, called anchors, yield $(\mathbf{f}_u^a, \mathbf{q}_u^a)$ and $(\mathbf{f}_u^b, \mathbf{q}_u^b)$. They show the same source content, so we explain both through one normal mode under their respective observation conditions. One uniformly selected view supplies the kernel field and anchor locations for b_{ur}. For an anchor labeled class k, normalizing $\sum_r b_{ur}\pi_{rkm}$ over m gives mode weights $\rho_{um|k}$, uniform outside kernel support. The paired likelihood is

$$p_u^{\text{pair}} = \sum_{m=1}^{M} \rho_{um|k}\, \phi_{km}(\mathbf{f}_u^a, \mathbf{q}_u^a)\, \phi_{km}(\mathbf{f}_u^b, \mathbf{q}_u^b). \tag{2}$$

Gradients through $\rho_{um|k}$ also train the geometry that assigns observations to kernels.

At inference, we measure variance-scaled feature mismatch D_{ukm} and combine mode matches within each class:

$$\begin{aligned} D_{ukm} &= \frac{\|\mathbf{f}_u - \boldsymbol{\mu}_{km}(\mathbf{q}_u)\|_2^2}{2\nu(\mathbf{q}_u)}, \\ E_u &= \min_k \left[-\frac{1}{d} \log \left(\frac{1}{M} \sum_{m=1}^{M} \exp(-D_{ukm}) \right) \right]. \end{aligned} \tag{3}$$

A high score means no source class explains the patch well. Omitting the Gaussian log-variance term avoids penalizing expected variability itself. Uniform mode weights and minimization over classes avoid reliance on predicted semantic labels or kernel mode weights. The nonnegative E_u is computed before pooling.

3.5 Training and Inference

We minimize $\mathcal{L} = \mathcal{L}_{\text{occ}} + \mathcal{L}_{\text{normal}}$. The first term is voxel-wise cross-entropy over free space and occupied classes. The second averages $-\log p_u^{\text{pair}}/(2d)$, including Gaussian normalization terms, within each observed source class and then equally across classes, preventing frequent classes from dominating normality learning. Anchor labels come from visible source occupancy labels projected

into the image. Both losses train the lifting, backbone, and kernel field; the paired likelihood also trains normality parameters. No anomaly labels or auxiliary losses are used. Both views are encoded, but only the selected view enters the 3D backbone. Depth is estimated once on the original image and resampled for both views.

Inference uses one image and calibration. Direct evidence R_v projects patch scores using b^0_{vu}. Propagated evidence H_v averages scores within each kernel using $A_u b_{ur}$, then mixes these averages using a_{vr}:

$$\begin{aligned} R_v &= \sum_u b^0_{vu} E_u, \\ H_v &= \sum_r a_{vr} \frac{\sum_u A_u b_{ur} E_u}{\sum_u A_u b_{ur}}, \\ S_v &= R_v + \big(1 - p_v(\text{free})\big) H_v. \end{aligned} \tag{4}$$

Empty sums and unsupported averages are zero. Only propagated evidence is weighted by occupied probability, so predicting free space cannot erase direct evidence.

4 Experiments

We evaluate whether EviOcc generalizes anomaly detection across datasets while maintaining semantic occupancy accuracy, and examine the contributions of its components. Models are trained independently on four sources and evaluated without target-domain adaptation.

4.1 Experimental Setup

Datasets and protocol. We train models independently on SemanticKITTI (Behley et al., 2019) and the KITTI-360, Waymo, and nuScenes components of SSCBench (Li et al., 2024). Following ProOOD (Zhang et al., 2026), we evaluate OOD detection on the three VAA benchmarks (Zhang et al., 2025): VAA-KITTI, VAA-KITTI-360, and VAA-STU. VAA-KITTI and VAA-KITTI-360 are constructed by adding synthetic anomalies to SemanticKITTI and SSCBench-KITTI-360, respectively, whereas VAA-STU is derived from STU (Nekrasov et al., 2025) and contains real-world anomalies. The underlying scenes used to construct VAA-KITTI and VAA-KITTI-360 are excluded from source training and model selection. Each source-trained model is evaluated on all three benchmarks. For models trained on SemanticKITTI and SSCBench-KITTI-360, evaluation on the corresponding VAA benchmark measures performance within the same dataset family. All other source–target combinations assess cross-dataset generalization. This yields two same-family and ten cross-dataset settings. Training and model selection use source data only, without target-domain adaptation. Checkpoints are selected by semantic mIoU on the corresponding source validation split. For each source, the selected checkpoint and scoring hyperparameters are fixed across all three target benchmarks. We additionally evaluate semantic occupancy on each source's native evaluation split using its own semantic label set. We also evaluate semantic occupancy on VAA-KITTI and VAA-KITTI-360 using models trained on their corresponding source datasets.

Evaluation metrics. Following OccOoD (Zhang et al., 2025) and ProOOD (Zhang et al., 2026), we report AuROC for anomaly ranking and regional AuPRC_r using ground-truth dilation with $r \in \{0.8, 1.0, 1.2\}\,\mathrm{m}$ to allow spatial tolerance. An AuROC of 50% indicates chance-level ranking. For semantic occupancy, we report geometric occupancy IoU and mIoU over occupied semantic classes, excluding free space.

Baselines and implementation details. For semantic occupancy, we compare against SGN (Mei et al., 2024), CGFormer (Yu et al., 2024), Symphonies (Jiang et al., 2024), VoxDet (Li et al., 2026a), OccOoD (Zhang et al., 2025), ProOOD (+SGN), and ProOOD (+VoxDet) (Zhang et al., 2026). For OOD detection, we compare SGN, VoxDet, and OccOoD using Anomaly-Sensitive Scoring (ASS) (Zhang et al., 2025), and ProOOD (+SGN) and ProOOD (+VoxDet) using EchoOOD (Zhang et al., 2026). All baselines are trained independently on each source, following the recommended configurations in their official repositories. All models are trained on 4 NVIDIA RTX A5000 GPUs. Our model uses AdamW with a learning rate of 10^{-4} and weight decay 10^{-2}.

4.2 OOD Detection across Datasets

Table 1 compares OOD detection performance across four training sources and three target benchmarks. Methods are compared within each source–target pair. The two same-family settings serve as reference evaluations, while the remaining ten settings assess cross-dataset generalization without target-domain adaptation. We report AuROC to measure how reliably a method assigns higher anomaly scores to anomalous voxels than to normal ones, with 50% corresponding to chance-level ranking. Regional AuPRC_r measures the ability to detect anomaly regions while limiting false positives. It allows spatial tolerance by expanding the ground-truth anomaly regions by radius r, so that detections near an anomaly can still be counted as correct despite imperfect localization. We evaluate three tolerances: 0.8, 1.0, and 1.2 m.

The compared methods show substantial limitations under cross-dataset transfer. In several source–target pairs, their AuROC approaches or falls below 50%, indicating that anomaly scores provide little reliable separation or even rank anomalies below normal voxels more often than above them. Low regional AuPRC_r in many transfer settings further highlights their difficulty in detecting anomaly regions while suppressing false positives. Our method achieves the highest AuROC in all twelve source–target pairs, remaining above chance in every setting. It also consistently achieves the highest regional AuPRC_r at all three tolerances in all ten cross-dataset settings. Together, these results demonstrate stronger cross-dataset generalization in both anomaly discrimination and localization, without target-domain adaptation.

Table 1: OOD detection across training sources and target benchmarks (%; higher is better). KITTI-360, Waymo, and nuScenes denote SSCBench components. Within each source block, the best, second-best, and third-best values in each column are marked in **bold**, underline, and *italics*, respectively.

Method	VAA-KITTI				VAA-KITTI-360				VAA-STU			
	AuROC	AuPRC_r			AuROC	AuPRC_r			AuROC	AuPRC_r		
		0.8 m	1.0 m	1.2 m		0.8 m	1.0 m	1.2 m		0.8 m	1.0 m	1.2 m
Source: SemanticKITTI												
SGN	60.68	8.35	14.61	22.78	43.92	5.72	12.46	23.81	*64.66*	2.87	5.81	11.26
VoxDet	59.36	4.68	8.21	14.19	46.57	4.63	10.17	20.54	59.70	1.51	3.12	5.97
OccOoD	*62.13*	10.64	16.22	23.58	43.61	6.48	14.73	27.16	65.21	*3.08*	*6.23*	*11.52*
ProOOD (+SGN)	64.10	28.04	43.32	62.87	42.76	7.91	17.26	32.68	61.94	8.12	14.38	24.46
ProOOD (+VoxDet)	61.09	*21.79*	*36.88*	*56.68*	*45.28*	*6.83*	*15.84*	*30.19*	54.95	2.64	4.92	8.68
Ours	**70.26**	**30.84**	**46.73**	**67.92**	**60.84**	**12.38**	**25.61**	**44.73**	**70.46**	**10.63**	**18.72**	**30.15**
Source: SSCBench-KITTI-360												
SGN	*46.72*	3.91	6.84	11.56	47.40	13.13	28.41	53.16	*54.23*	1.64	3.42	6.83
VoxDet	40.36	2.74	4.91	8.63	50.59	11.94	25.24	48.31	50.71	0.92	1.93	3.87
OccOoD	48.05	4.36	7.82	12.74	*47.96*	*14.34*	*29.88*	*54.18*	55.86	1.92	3.87	7.42
ProOOD (+SGN)	28.13	8.73	14.92	23.68	46.78	**14.69**	30.66	57.74	52.48	3.76	7.14	12.63
ProOOD (+VoxDet)	35.67	*6.72*	*11.84*	*19.61*	47.47	12.85	27.04	49.98	48.94	*2.45*	*4.83*	*8.94*
Ours	**66.37**	**16.82**	**27.46**	**41.73**	**57.83**	14.41	**31.87**	**59.26**	**68.32**	**6.47**	**11.83**	**20.76**
Source: SSCBench-Waymo												
SGN	*51.74*	3.82	6.95	11.38	42.38	4.63	10.27	19.84	*54.82*	1.83	3.76	7.14
VoxDet	49.28	2.63	4.87	8.76	45.16	3.82	8.91	17.32	52.39	1.17	2.42	4.93
OccOoD	53.16	4.51	8.12	13.47	43.07	5.36	11.84	22.47	56.17	2.14	4.39	8.26
ProOOD (+SGN)	50.63	9.24	16.37	26.82	41.62	*6.74*	*14.53*	*27.86*	53.61	4.28	8.17	14.63
ProOOD (+VoxDet)	49.61	*7.68*	*13.85*	*22.73*	*44.29*	7.13	15.47	29.62	48.82	*2.96*	*5.89*	*10.42*
Ours	**65.48**	**16.73**	**28.16**	**43.52**	**57.93**	**10.82**	**22.67**	**40.31**	**69.24**	**7.36**	**13.28**	**22.47**
Source: SSCBench-nuScenes												
SGN	48.63	2.94	5.37	9.18	41.26	3.76	8.43	16.72	56.38	2.16	4.27	8.13
VoxDet	46.91	2.16	3.92	6.84	44.91	4.07	9.15	17.38	53.72	1.43	2.96	5.64
OccOoD	50.27	3.48	6.42	10.73	42.17	4.32	9.68	18.91	*55.91*	2.38	4.83	9.17
ProOOD (+SGN)	*51.38*	7.16	12.83	21.46	40.73	*5.48*	*12.16*	*23.74*	54.86	4.71	8.92	15.38
ProOOD (+VoxDet)	52.06	*6.43*	*11.26*	*19.05*	*44.24*	6.01	12.78	25.17	51.09	*3.34*	*6.52*	*11.71*
Ours	**62.71**	**13.62**	**23.47**	**37.18**	**56.42**	**9.37**	**19.84**	**35.26**	**70.18**	**7.82**	**14.16**	**23.89**

4.3 SEMANTIC OCCUPANCY PREDICTION

Tables 2–5 report occupancy IoU, semantic mIoU, and per-class IoU on each source's evaluation split. Following OccOoD (Zhang et al., 2025) and ProOOD (Zhang et al., 2026), we also evaluate SemanticKITTI- and SSCBench-KITTI-360-trained models on VAA-KITTI and VAA-KITTI-360, respectively, to assess semantic prediction in scenes containing anomalies. We follow each dataset's standard semantic label set and compute mIoU over non-empty classes.

Our method achieves the highest occupancy IoU and semantic mIoU among the compared methods across all four source datasets, demonstrating strong performance in both occupancy estimation and semantic classification. Per-class improvements span multiple categories, showing that the gains extend beyond a few classes. On VAA-KITTI and VAA-KITTI-360, our method also delivers strong occupancy and semantic accuracy, supporting reliable prediction of existing classes in scenes containing anomalies. Together with Sec. 4.2, these results show that stronger cross-dataset OOD detection is achieved while maintaining competitive semantic occupancy performance.

Table 2: Semantic occupancy results on the SemanticKITTI validation split and VAA-KITTI (%; higher is better).

Method	IoU	mIoU	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
SemanticKITTI																					
SGN	43.52	14.52	31.86	0.60	0.18	13.33	5.67	1.24	1.40	0.00	59.16	18.56	30.47	0.46	21.63	9.75	25.79	7.26	34.27	9.68	4.66
Symphonies	41.95	14.90	28.79	2.74	2.73	20.24	*13.84*	3.45	2.41	0.00	56.45	15.41	27.72	1.00	21.54	8.29	25.70	6.78	30.78	9.48	5.73
CGFormer	46.07	16.88	34.46	4.62	2.67	19.47	7.65	2.51	4.17	0.00	*65.50*	20.81	32.44	0.17	23.62	9.01	26.75	8.86	*39.44*	10.60	*7.95*
VoxDet	47.35	*18.74*	**34.57**	*4.15*	5.05	**30.98**	14.51	**4.47**	3.30	0.00	65.49	20.86	*34.28*	0.03	*25.60*	11.68	*28.90*	10.40	**41.32**	12.06	8.47
OccOoD	*46.73*	16.81	33.61	0.58	1.13	*25.33*	11.79	0.42	0.15	0.00	60.95	*23.50*	32.26	0.05	26.40	9.78	28.58	9.06	36.86	**13.02**	5.96
ProOOD (+SGN)	44.07	14.95	31.68	0.87	0.90	14.57	5.57	2.99	2.18	0.00	59.47	20.35	27.81	0.09	21.56	9.74	26.03	8.51	33.99	11.04	6.73
ProOOD (+VoxDet)	46.66	18.86	*34.08*	3.79	**5.79**	30.42	**15.58**	*3.82*	*4.12*	0.00	**67.06**	23.76	34.83	*0.66*	24.55	*10.78*	29.05	*9.79*	41.09	*11.67*	7.50
Ours	**48.80**	**19.83**	29.77	**5.66**	*4.88*	7.53	6.17	3.95	**4.89**	**1.74**	66.23	**37.04**	**37.69**	**15.35**	**29.98**	**23.02**	**30.37**	**15.22**	34.32	11.40	**11.56**
VAA-KITTI																					
SGN	39.05	11.13	20.54	0.45	0.35	0.62	1.21	0.47	1.32	0.06	46.61	15.96	26.70	0.12	18.26	12.25	24.89	*6.49*	21.86	8.77	4.46
Symphonies	39.80	11.80	19.89	0.90	0.77	2.41	2.70	1.13	1.07	**0.51**	49.67	17.79	27.44	*0.18*	19.51	13.17	25.48	5.86	22.47	8.25	5.04
CGFormer	40.72	12.71	21.55	1.11	0.69	3.79	2.14	0.85	1.43	0.12	53.34	19.97	28.15	0.05	*21.65*	14.55	26.61	6.63	24.35	8.98	5.45
VoxDet	41.58	*13.14*	20.27	3.20	*1.10*	*5.65*	4.10	1.10	1.66	*0.29*	*55.88*	*21.30*	27.15	0.02	22.78	*15.57*	27.46	5.88	23.04	8.69	4.61
OccOoD	39.99	11.48	20.18	0.27	0.57	0.75	1.06	0.66	1.99	0.09	50.05	18.22	*27.46*	0.01	17.25	11.34	25.09	5.58	*23.57*	8.27	*5.80*
ProOOD (+SGN)	40.37	11.86	20.10	0.42	0.96	3.34	1.31	*1.54*	1.18	0.08	49.41	16.61	27.13	0.15	19.98	12.75	25.94	6.37	23.24	9.09	5.76
ProOOD (+VoxDet)	*40.97*	13.64	*21.40*	*2.91*	**3.70**	**7.03**	*3.50*	**2.57**	*1.94*	0.16	56.47	**23.71**	28.15	**0.25**	*21.65*	15.89	*27.01*	6.45	21.25	*9.03*	6.02
Ours	**43.14**	**15.00**	**23.70**	**3.92**	1.67	6.76	**5.52**	2.13	**2.56**	0.47	**58.69**	23.32	**31.16**	0.23	**25.66**	**17.57**	**30.33**	**7.81**	**26.14**	**10.34**	**7.11**

Table 3: Semantic occupancy results on SSCBench-KITTI-360 and VAA-KITTI-360 (%; higher is better).

Method	IoU	mIoU	car	bicycle	motorcycle	truck	other-vehicle	person	road	parking	sidewalk	other-ground	building	fence	vegetation	terrain	pole	traffic-sign	other-structure	other-object
SSCBench-KITTI-360																				
SGN	46.18	17.72	28.20	2.12	3.01	12.01	3.74	4.20	59.47	14.52	36.56	4.27	39.79	7.08	36.61	*23.13*	14.81	16.08	8.30	5.00
Symphonies	44.16	18.58	30.03	1.88	5.89	**25.10**	**11.99**	**8.24**	54.90	13.84	32.75	6.92	35.08	8.57	38.36	11.57	14.08	9.62	**14.38**	11.27
CGFormer	48.06	*20.04*	*29.90*	3.36	3.96	17.58	*6.79*	*6.69*	63.91	17.11	40.72	*5.59*	42.78	8.16	38.74	24.92	16.19	17.43	10.24	6.73
VoxDet	48.19	19.97	28.99	3.87	*5.21*	16.15	6.62	6.11	61.98	*16.30*	39.70	5.32	*43.31*	*9.15*	38.72	22.78	17.19	**19.80**	10.33	7.94
OccOoD	46.92	18.38	28.46	2.99	0.95	9.49	4.59	3.04	58.95	16.23	36.97	5.36	43.05	7.81	37.35	23.00	17.51	*18.94*	9.01	7.17
ProOOD (+SGN)	46.62	17.98	27.26	2.79	4.27	12.03	5.69	4.89	60.02	14.09	36.64	4.96	40.86	7.18	36.91	22.78	14.10	15.44	8.86	4.88
ProOOD (+VoxDet)	*48.18*	20.27	28.85	*3.81*	**7.32**	*18.75*	6.55	5.87	*62.26*	16.26	*40.13*	5.43	43.69	9.21	*38.73*	22.59	*17.24*	19.39	*10.61*	*8.17*
Ours	**49.95**	**22.69**	**32.72**	**5.76**	5.02	22.00	9.76	6.81	**66.47**	**19.65**	**42.56**	**8.77**	**45.73**	**10.69**	**41.09**	**27.06**	**19.73**	18.85	12.85	**12.90**
VAA-KITTI-360																				
SGN	39.63	12.03	20.65	0.06	0.14	0.43	*2.96*	0.19	47.56	14.70	24.65	2.56	32.82	3.58	28.64	11.65	7.41	8.71	4.35	5.51
Symphonies	40.63	13.08	22.99	0.17	0.38	**1.36**	2.11	1.90	48.32	13.71	26.43	3.17	33.86	*3.92*	29.19	11.95	8.99	13.24	*6.47*	7.30
CGFormer	42.07	*14.08*	23.14	*0.42*	0.29	0.92	1.23	2.33	52.09	*15.33*	29.45	3.17	*36.59*	3.89	30.77	13.63	*10.33*	15.78	6.16	7.92
VoxDet	*42.26*	13.49	21.36	**0.99**	*0.47*	0.81	0.38	2.64	51.40	12.13	*28.85*	2.62	36.13	3.55	*29.93*	12.54	9.31	*15.85*	5.46	8.32
OccOoD	39.30	12.57	22.82	0.03	0.08	0.32	**3.57**	0.57	49.15	14.67	27.90	3.36	32.56	1.55	27.40	11.44	8.48	12.24	3.63	6.47
ProOOD (+SGN)	39.39	12.56	22.19	0.05	0.26	0.56	0.04	2.05	48.22	**16.51**	25.54	3.60	33.69	2.92	28.11	*12.93*	8.72	10.34	5.60	4.71
ProOOD (+VoxDet)	42.63	14.26	*23.07*	0.11	**0.63**	*1.07*	0.02	**3.52**	*51.51*	14.46	28.48	*3.54*	37.26	**4.10**	29.72	12.53	10.60	**18.95**	7.25	**9.80**
Ours	**44.02**	**15.51**	**25.14**	0.86	0.57	1.28	3.44	*2.45*	**54.44**	15.65	**32.14**	**5.64**	**38.65**	4.02	**33.25**	**15.72**	**11.81**	17.95	**8.14**	*8.08*

Table 4: Semantic occupancy results on SSCBench-Waymo (%; higher is better).

Method	IoU	mIoU	car	bicycle	motorcycle	person	bicyclist	road	sidewalk	other-ground	building	vegetation	trunk	pole	traffic-sign	other-object
SGN	41.56	14.19	21.36	3.02	0.34	7.68	5.21	55.84	28.63	23.17	18.72	14.65	4.23	5.79	5.38	4.61
Symphonies	40.83	15.17	26.48	4.56	0.91	**10.72**	**8.94**	53.76	27.38	22.94	17.63	16.27	4.96	5.46	4.89	7.52
CGFormer	44.62	*17.33*	*27.08*	5.37	0.83	9.48	7.91	60.42	33.56	28.87	23.12	18.97	*6.59*	7.21	6.84	6.41
VoxDet	*45.17*	17.18	26.91	**6.38**	*1.26*	10.14	7.82	58.46	31.92	27.61	22.38	18.26	6.47	*7.58*	**7.46**	*7.93*
OccOoD	42.39	14.82	22.08	3.41	0.42	7.96	5.06	56.32	29.14	23.86	20.34	15.27	4.58	6.31	6.94	5.82
ProOOD (+SGN)	42.04	14.72	21.74	3.79	0.58	8.53	6.12	55.17	28.48	23.52	19.87	15.68	4.96	5.94	6.41	5.26
ProOOD (+VoxDet)	45.49	17.51	27.32	*6.11*	**1.47**	*10.29*	*8.03*	*59.18*	*32.74*	*28.32*	*22.85*	*18.54*	6.91	7.81	7.38	8.22
Ours	**47.34**	**18.85**	**29.39**	6.21	1.31	10.64	8.74	**62.68**	**35.51**	**30.57**	**24.98**	**20.39**	**7.94**	**8.88**	*7.31*	**9.35**

Table 5: Semantic occupancy results on SSCBench-nuScenes (%; higher is better).

Method	IoU	mIoU	car	bicycle	motorcycle	truck	other-vehicle	person	road	sidewalk	other-ground	building	vegetation	other-object
SGN	31.47	10.56	12.68	1.26	3.38	9.47	10.16	4.12	40.83	15.94	13.26	8.65	6.87	0.06
Symphonies	31.02	11.68	14.23	2.41	*5.26*	**13.84**	13.47	5.73	39.64	15.36	12.89	9.48	7.64	*0.17*
CGFormer	34.27	12.36	14.86	2.13	4.71	11.92	11.84	5.19	*43.76*	18.27	15.32	11.26	8.94	0.14
VoxDet	*34.82*	*13.05*	*15.72*	*3.28*	**6.13**	12.67	12.38	*5.86*	43.19	*18.64*	*16.17*	*12.49*	*9.78*	**0.31**
OccOoD	32.16	10.96	13.14	1.47	3.69	9.12	10.63	4.76	41.27	16.48	13.92	9.31	7.58	0.09
ProOOD (+SGN)	31.93	10.88	12.46	1.91	4.23	10.16	10.37	5.08	40.49	15.78	13.54	9.07	7.36	0.08
ProOOD (+VoxDet)	35.26	13.41	16.08	**3.46**	6.02	*13.19*	*12.86*	6.14	44.02	19.15	16.64	12.97	10.16	0.28
Ours	**36.81**	**14.65**	**17.92**	3.32	6.02	13.73	**14.91**	**7.08**	**46.96**	**21.09**	**18.45**	**14.57**	**11.47**	0.28

4.4 Ablation Study

We evaluate four cumulative variants: Base, +A, +A+B, and Full. Base uses fixed mean lifting, a dense class/mode head, and static normal components, with anomaly scores computed from pooled features. A introduces learned footprint-weighted lifting and scores individual observations before aggregation. B replaces the dense head with a shared local kernel field for occupancy prediction, normal-mode routing, and anomaly evidence transport. C introduces observation-conditioned normal components and a shared-mode paired likelihood, replacing static components and independent per-view mixture likelihoods.

All variants are trained separately on SemanticKITTI using the same frozen encoders, 3D backbone, source view pairs, training schedule, and source-only checkpoint selection rule. Both occupancy supervision and normality-likelihood training are retained throughout. Each selected checkpoint is evaluated on SemanticKITTI validation and all three VAA benchmarks without target-domain adaptation. Table 6 reports source-domain occupancy performance and same-family and cross-dataset anomaly detection. Consecutive rows assess incremental effects in the chosen order, rather than isolated component effects.

Table 6: Cumulative ablation trained on SemanticKITTI. Occupancy IoU and mIoU are evaluated on SemanticKITTI validation; anomaly detection is evaluated on three VAA benchmarks. Regional AuPRC uses $r = 1.0\,\mathrm{m}$. All values are percentages; higher is better.

	SemanticKITTI		VAA-KITTI		VAA-KITTI-360		VAA-STU	
Variant	IoU	mIoU	AuROC	AuPRC_r	AuROC	AuPRC_r	AuROC	AuPRC_r
Base	39.86	12.74	57.42	20.64	43.16	7.12	54.18	5.24
+A	44.63	16.52	64.83	31.46	49.37	13.52	60.26	9.73
+A+B	47.72	18.91	67.58	43.82	53.86	22.34	64.18	15.61
Full (+A+B+C)	**48.80**	**19.83**	**70.26**	**46.73**	**60.84**	**25.61**	**70.46**	**18.72**

Adding A improves both occupancy prediction and anomaly detection. Adding B further improves regional AuPRC_r on all three benchmarks, consistent with the role of local geometry in propagating

anomaly evidence. Adding C yields larger AuROC gains on the two cross-dataset benchmarks than on VAA-KITTI, supporting observation-conditioned normality for cross-dataset transfer.

4.5 Qualitative Results

Fig. 2(a,b) shows cross-dataset anomaly detection from SSCBench-nuScenes. EviOcc concentrates anomaly responses near the anomalous objects, with less response on surrounding normal regions than ProOOD (+VoxDet) and OccOoD in these examples. Panels (c,d) show comparable or better recovery of scene geometry and known-class regions.

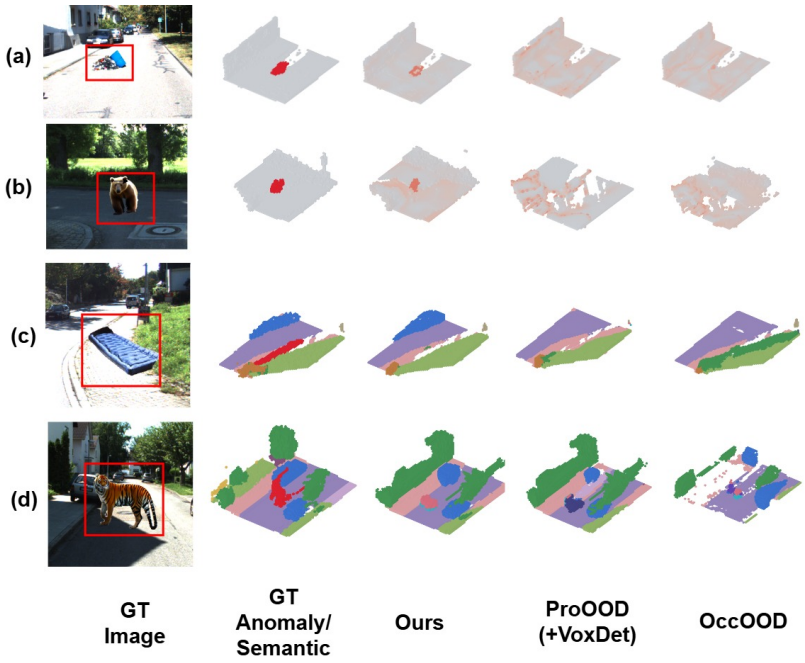

Figure 2: **Qualitative comparisons.** (a,b) Anomaly detection from SSCBench-nuScenes to VAA-KITTI and VAA-KITTI-360, respectively. (c,d) Semantic occupancy from SemanticKITTI to VAA-KITTI and from SSCBench-KITTI-360 to VAA-KITTI-360, respectively. Columns show the input image, ground truth, EviOcc, ProOOD (+VoxDet), and OccOoD. Red boxes mark anomalous objects in the input images.

5 Conclusion

We presented EviOcc, a source-only framework for monocular semantic occupancy prediction and voxel-level anomaly detection. Sampling-aware lifting preserves observation-level visual evidence, while a shared local kernel field couples occupancy prediction with spatial anomaly evidence propagation. Paired observation-conditioned normality accounts for benign feature variations under changing observation conditions. Experiments across four training sources and three anomaly benchmarks demonstrate improved cross-dataset anomaly discrimination and localization, together with strong geometric and semantic occupancy accuracy, without target-domain adaptation. Ablations further support the complementary contributions of evidence preservation, geometry-tied completion, and observation-conditioned normality.

AI USE STATEMENT

We used ChatGPT (OpenAI) for grammar correction and language polishing. The authors reviewed all suggested edits and take responsibility for the final manuscript.

REFERENCES

Eunsu Baek, Keondo Park, Jiyoon Kim, and Hyung-Sin Kim. Unexplored faces of robustness and out-of-distribution: Covariate shifts in environment and sensor domains. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22294–22303. IEEE, 2024.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9297–9307, 2019.

Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3981–3991. IEEE, 2022.

Jun Cen, Peng Yun, Shiwei Zhang, Junhao Cai, Di Luan, Mingqian Tang, Ming Liu, and Michael Yu Wang. Open-world semantic segmentation for lidar point clouds. In *European Conference on Computer Vision*, pp. 318–334. Springer, 2022.

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. *arXiv preprint arXiv:2104.14812*, 2021.

Wenbang Deng, Xieyuanli Chen, Qinghua Yu, Yunze He, Junhao Xiao, and Huimin Lu. A novel decomposed feature-oriented framework for open-set semantic segmentation on lidar data. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11126–11132. IEEE, 2025.

Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28980–28990. IEEE, 2025.

Zhitong Gao, Bingnan Li, Mathieu Salzmann, and Xuming He. Generalize or detect? towards robust semantic segmentation under multiple distribution shifts. *Advances in Neural Information Processing Systems*, 37:52014–52039, 2024.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9223–9232. IEEE, 2023.

Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19946–19956. IEEE, 2024a.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *European Conference on Computer Vision*, pp. 376–393. Springer, 2024b.

Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27477–27486. IEEE, 2025.

Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20258–20267. IEEE, 2024.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15405–15414. IEEE, 2021.

Wuyang Li, Zhu Yu, and Alexandre Alahi. Voxdet: Rethinking 3d semantic scene completion as dense object detection. *Advances in Neural Information Processing Systems*, 38:81004–81038, 2026a.

Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9087–9098. IEEE, 2023.

Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, et al. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13333–13340. IEEE, 2024.

Zizhao Li, Zhengkang Xiang, Jiayang Ao, Feng Liu, Joseph West, and Kourosh Khoshelham. Neural distribution prior for lidar out-of-distribution detection. *arXiv preprint arXiv:2604.09232*, 2026b.

Zizhao Li, Zhengkang Xiang, Jiayang Ao, Joseph West, and Kourosh Khoshelham. Relative energy learning for lidar out-of-distribution detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 239:334–346, 2026c.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Haoang Lu, Yuanqi Su, Xiaoning Zhang, Longjun Gao, Yu Xue, and Le Wang. Vishall3d: Monocular semantic scene completion from reconstructing the visible regions to hallucinating the invisible regions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28674–28684. IEEE, 2025.

Jianbiao Mei, Yu Yang, Mengmeng Wang, Junyu Zhu, Jongwon Ra, Yukai Ma, Laijian Li, and Yong Liu. Camera-based 3d semantic scene completion with sparse guidance network. *IEEE Transactions on Image Processing*, 33:5468–5481, 2024.

Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*, 2023.

Simone Mosco, Daniel Fusaro, and Alberto Pretto. Learning to identify out-of-distribution objects for 3d lidar anomaly segmentation. *arXiv preprint arXiv:2604.23604*, 2026.

Nazir Nayal, Misra Yavuz, Joao F Henriques, and Fatma Güney. Rba: Segmenting unknown regions rejected by all. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 711–722, 2023.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11875–11885. IEEE, 2025.

Shyam Nandan Rai, Fabio Cermelli, Dario Fontanel, Carlo Masone, and Barbara Caputo. Unmasking anomalies in road-scene segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4014–4023. IEEE, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

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*, pp. 246–263. Springer, 2022.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21672–21683. IEEE, 2023.

Shaocong Xu, Pengfei Li, Qianpu Sun, Xinyu Liu, Yang Li, Shihui Guo, Zhen Wang, Bo Jiang, Rui Wang, Kehua Sheng, et al. Lion: Learning point-wise abstaining penalty for lidar outlier detection using diverse synthetic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8951–8959, 2025.

Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131(10):2607–2622, 2023.

Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9421–9431. IEEE, 2023.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022.

Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. *Advances in Neural Information Processing Systems*, 37:1531–1555, 2024.

Yuheng Zhang, Mengfei Duan, Kunyu Peng, Yuhang Wang, Ruiping Liu, Fei Teng, Kai Luo, Zhiyong Li, and Kailun Yang. Out-of-distribution semantic occupancy prediction. *arXiv preprint arXiv:2506.21185*, 2025.

Yuheng Zhang, Mengfei Duan, Kunyu Peng, Yuhang Wang, Di Wen, Danda Pani Paudel, Luc Van Gool, and Kailun Yang. Proood: Prototype-guided out-of-distribution 3d occupancy prediction. *arXiv preprint arXiv:2604.01081*, 2026.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9399–9409. IEEE, 2023.

Yupeng Zheng, Xiang Li, Pengfei Li, Yuhang Zheng, Bu Jin, Chengliang Zhong, Xiaoxiao Long, Hao Zhao, and Qichao Zhang. Monoocc: Digging into monocular semantic occupancy prediction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 18398–18405. IEEE, 2024.