

TriO: Tri-Modal Unsupervised Occupancy World Model for Anything Perception

Quinlan Sykora^{*1,2}, Sourav Biswas^{*1,2}, Christopher Diehl^{*1},
Andrew Cunningham¹, Thomas Gilles¹, and Raquel Urtasun¹

¹ Waabi

² University of Toronto,

{qsykora,sbiswas,cdiehl,acunningham,tgilles,urtasun}@waabi.ai

Abstract. We present TriO, a multi-modal unsupervised world model that predicts 4D occupancy, obstacle segmentation, flow and LiDAR. In contrast to prior work, TriO utilizes three distinct sensor modalities (camera, LiDAR, and RADAR) as both inputs and sources of self-supervision, eliminating the need for additional human annotations. Thanks to its novel supervision, the model is able to segment any occupancy from the drivable surface, overcoming the limitations of existing open-set methods in handling long-tail objects. TriO achieves state-of-the-art results in multiple 3D and 4D tasks, including occupancy, flow, and LiDAR prediction, as well as zero-shot road obstacle segmentation across multiple datasets such as Argoverse 2, and Spotting the Unexpected.

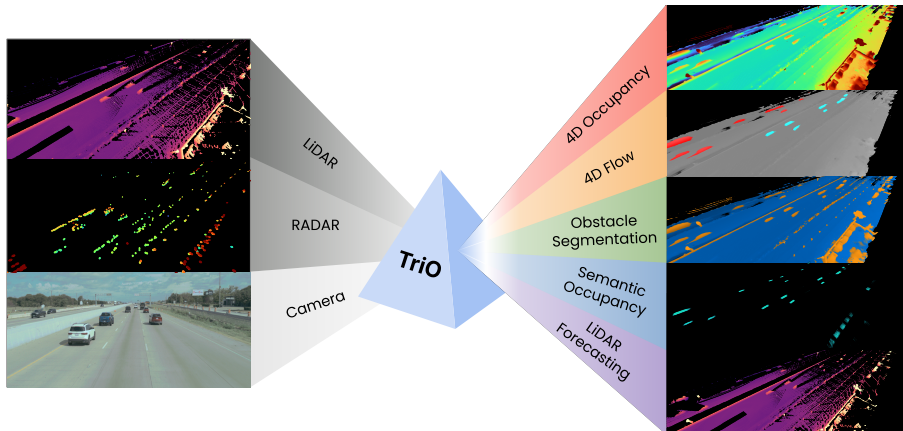


Fig. 1: We present TriO, an unsupervised tri-modal occupancy world model designed to perceive and forecast anything that could be considered path-blocking.

* denotes equal contribution

1 Introduction

In order for a self-driving vehicle (SDV) to safely navigate through complex environments, it must first *perceive anything*, i.e. segment above ground obstacles from drivable areas at the current time, then forecast the evolution of the scene to plan its future motion. This problem has been predominantly tackled with object-based approaches [9, 10, 12, 15, 27, 28, 41, 64, 73, 74], which first detect objects and then forecast their future trajectories. In contrast, object-free occupancy approaches [2, 11, 16, 24, 47, 57, 63] overcome the downsides of thresholding detections and the limited representation capability of a finite set of trajectories.

However, most 3D/4D occupancy approaches [21, 22, 46, 82, 84] rely on dense voxel grid labels [67, 72] which are computationally expensive at high resolutions, and are constructed with human annotations (e.g., semantic LiDAR point annotations or bounding box labels). Due to the cost of the latter, and their difficulty to generate at scale, the volume of the labeled data is commonly orders of magnitude smaller than the amount of unlabeled data [1]. Moreover, the mentioned approaches are usually supervised with a finite set of semantic classes, which do not encapsulate safety-critical rare events.

Conversely, recent occupancy approaches [1, 17, 34] leverage LiDAR self-supervision, pre-training on vast amounts of sensor data and fine-tuning with a small amount of labeled data for downstream applications. However, LiDAR is susceptible to blooming [69], multipath returns [61], and limited range. Additional sensor modalities can be of benefit for self-supervised training, and multi-sensor fusion can increase robustness to sensor failures and adverse conditions.

Other methods have leveraged RGB cameras, either as inputs [55], distilling foundation model features [45], or predicting semantic pseudo-labels [6, 7, 19, 26, 38, 42] from camera-based models. Furthermore, many prior works [31, 32, 59, 77, 79] generate semantic labels with open-vocabulary segmentation models queried with text-prompts. However, common failure modes have been shown in open-set detectors [3] and segmentors [70], challenging their overall robustness to rare classes. Finally, the segmentation models require image-text pair annotations.

One generally overlooked modality is RADAR, which can enhance training by introducing radial velocity measurements, improve robustness under adverse weather conditions [39], and increase the observation range. While some works investigate RADAR-based 3D occupancy models using human annotations [18, 56], and others focus on RADAR-only input with self-supervision [36] or cross-modal LiDAR supervision [43], none integrate all aforementioned modalities in a unsupervised tri-modal 4D world model.

Motivated by these downsides, our goal is to develop a model that can leverage all three common sensor modalities (LiDAR, camera, and RADAR) as both input and supervision for occupancy in a self-supervised manner. Towards this goal, we propose TRIO: a multi-modal world model that can predict and forecast 4D continuous occupancy, obstacle segmentation, and flow at any continuous (x, y, z, t) coordinate from tri-modal observations. To achieve this, we leverage each sensor as both an input and as a unique form of supervision, ultimately combining the strengths of each modality to produce the best possible represen-

tation of the world. We further introduce a novel multi-sensor fusion architecture leading to better forecasting capabilities. Notably, being pre-trained on a vast amount of raw sensor data, TriO can be effectively leveraged for multiple downstream applications.

Historically, to determine which entities in the scene are relevant to the self-driving task, prior work uses semantic supervision from open-vocabulary models prompted with a closed set of multiple classes [31, 32, 59, 77, 79]. These however do not generalize to long tail events with rare obstacles, and hence do not cover all safety-critical scenarios. We argue the most general and robust requirement for occupancy is to semantically distinguish between what is *occupied but not path-blocking* (e.g., a flat surface that is traversable by the SDV) vs. *occupied and path-blocking* (e.g., a rare object on the road). Towards this end, we define the task of **anything-perception** as the ability to segment any object that exists above the road from the ground surface.

Despite demonstrating impressive open-set detection capabilities, the segmentation models of prior work used for $2D \rightarrow 3D$ label distillation struggle to segment rare objects [70] (the long tail), even when prompted with the ground truth class, highlighting a lack of semantic understanding and generalization in current VLMs. To circumvent this, we propose a simple yet effective unsupervised¹ *geometric image/LiDAR segmentation* method that leverages representations from image depth model Metric3Dv2 [25], and the LiDAR segmentation heuristic from Patchwork++ [37]. Our method detects any obstruction, leading to significant improvements in generalizable unknown object segmentation. Notably, in contrast to other vision-language based approaches which require at least textual image descriptions [4, 32, 59, 68, 77] or (semantic) mask labels [31, 35, 77, 79], our obstacle segmentation model relies solely on geometric properties rather than semantic annotations.

Contributions: **1)** TriO uniquely uses all three common sensor modalities (camera, LiDAR, RADAR), both as input and for self-supervision. **2)** We introduce a novel image/LiDAR-based obstacle segmentation technique that generates pseudo-labels for distillation into TriO, without requiring additional human labels in our training pipeline, while maintaining robustness and generalizability. **3)** We propose a novel architecture fusing all modalities leading to better long-term forecasting. **4)** TriO achieves state-of-the-art performance on a variety of tasks including point cloud forecasting, obstacle segmentation, and both 4D/3D occupancy and flow forecasting, validated on three real-world autonomous driving datasets (*Argoverse2*, *Spotting the Unexpected*, *Long Range RADAR*). Overall, the novelty of our contribution is the unique integration of all three sensor modalities as inputs, and as source of self-supervision for our output representations.

¹ We note that Metric3Dv2 [25] was partially trained on datasets with semi-supervised human annotations. In this context, we define *unsupervised* to mean not requiring direct human annotations. For more discussion refer to the supplementary.

2 Related Work

Semantic Occupancy. The task of *scene completion* [33, 50, 78, 83] involves reconstructing a complete scene from sparse sensor measurements. *Semantic scene completion* [5, 13, 62, 76] further extends this by aiming to infer semantic classes for each completed region, a task closely related to *semantic occupancy prediction*. Contemporary methods in this domain [21, 22, 46, 82, 84] typically focus on predicting 3D or 4D occupancy voxel grids with their associated semantic labels. These approaches generally require dense voxel-level supervision [67, 72], which is often derived from costly human-annotated labels, such as 3D bounding boxes or semantic point clouds. More recently, several approaches [6, 8, 42, 80] have proposed distilling 2D pseudo-semantic labels from VLMs. However, these VLMs themselves rely on large-scale human annotations in the form of image-text pairs, and supervising with a closed set of classes inherently restricts generalization in open-world scenarios.

Anything Perception. Open-vocabulary segmentation methods perform classification via text prompts at inference time, overcoming the limitations of fixed, closed-set label spaces. In the 2D domain, two primary paradigms have emerged: one leverages open-set detectors like GroundingDINO [44] or OWLv2 [49] to prompt mask predictors such as SAM [35], while another [31, 77, 79] predicts mask proposals and classifies them by computing the similarity between mask embeddings and CLIP-style [58] text embeddings. These paradigms have been extended to the 3D domain through methods like SAL [53], which distills CLIP features for LiDAR segmentation. Similarly, recent open-vocabulary occupancy methods [30, 38, 71, 81] distill these features into voxel representations to enable open-world prompting. Despite their flexibility, these methods face two critical challenges: (1) they remain fundamentally dependent on models pre-trained on vast image-text and per-pixel mask annotations, and (2) they exhibit significant performance degradation for rare classes [3, 70]. In contrast, TRIOs supervision relies on geometric cues from different sensors, enabling a more reliable segmentation of any object relative to the drivable surface. While ground segmentation has been extensively studied for LiDAR point clouds [20, 37, 54, 65], its application to occupancy remains limited. For example, [60] utilizes surface normals from stereo cameras and LiDAR to estimate ground regions *in current time BEV (2D)* occupancy maps, yet is constrained by the range of current measurements and segmentation networks that require pixel-level human annotations. In comparison, we distill fused obstacle segmentation estimates from LiDAR and monocular camera data and TRIO has a unique ability of *forecasting these segmentations in 4D*.

Unsupervised Occupancy. Early non-learning-based methods [14, 52, 66] employed classical filtering techniques to estimate occupancy and flow without human annotations. With the advent of deep learning, 4D-Occ [34] proposed a LiDAR-only self-supervised approach to forecast 4D occupancy. However, its explicit 4D voxel grid representation suffers from significant memory and run-time inefficiencies, often forcing a trade-off between spatial resolution and reconstruction quality. To address these limitations, recent implicit occupancy

methods [1, 17, 42, 45, 55] represent scenes in continuous space-time, leveraging self-supervision from LiDAR or camera inputs. Departing from implicit representations, SceneDino [29] predicts 3D geometry and feature fields from monocular input, achieving semantic scene completion through clustering and unsupervised distillation. Other recent works, such as [43], adopt cross-modal strategies using RADAR as input with LiDAR-based supervision. In contrast to all prior work, TriO utilizes camera, LiDAR, and RADAR data for both input and self-supervision, exploiting the inherent physical synergies between sensors to perform 4D forecasting of occupancy, segmentation, and flow. We show an extensive tabular comparison and further related works in the supplementary.

3 Unsupervised Multi-Modal Occupancy-Flow Method

To learn a robust representation of the world, we leverage the complementary physical properties of the standard SDV sensor suite: active sensors (LiDAR and RADAR) operating on Time-of-Flight principles, and passive cameras. While LiDAR provides high-precision 3D geometry capable of capturing free-space [1] and object surfaces [37], its sparse returns degrade at distance and in adverse weather. In contrast, RADAR uses longer wavelengths that penetrate fog and rain, and provide instantaneous radial velocity via the Doppler effect. Passive RGB cameras offer dense, high-resolution semantic and textural information, but lack native depth. To overcome this limitation, we extract dense depth and surface normal estimates using recent vision foundation models [25], enabling the reliable segmentation of obstacles. In summary, we combine the complementary synergies of each sensor modality and propose TriO, a 4D world model that learns geometry, dynamics, and object segmentation while remaining unsupervised.

3.1 Unsupervised Task

To define our training objective, let $T \in \mathbb{N}$ denote the future time-horizon and $H \in \mathbb{N}$ the historical observation time steps. For notational clarity, in the following we define $\mathbf{S}_H = \mathbf{S}_{-H, \dots, 0}$ as the historical sensor observations at time $-H, \dots, 0$ and $\mathbf{S}_T$ as the future observations at time $0, \dots, T$. Here, $\mathbf{S}$ is a placeholder for either multi-view camera images $\mathbf{C} \in \mathbb{R}^{N_C \times W \times H \times 3}$, LiDAR points $\mathbf{L} \in \mathbb{R}^{N_L \times d_L}$, or RADAR targets $\mathbf{R} \in \mathbb{R}^{N_R \times d_R}$. N_C denotes the number of camera images with width W and height H . Further, N_L , N_R , d_L , d_R are the number of LiDAR/RADAR points and their feature dimensionalities, respectively. Finally, we define the combined historical and future sensor observations as $\mathbf{X}_H = (\mathbf{C}_H, \mathbf{L}_H, \mathbf{R}_H)$ and $\mathbf{X}_T = (\mathbf{C}_T, \mathbf{L}_T, \mathbf{R}_T)$, respectively. Our goal is to learn a model:

$$\mathbf{y} = (o, s, \mathbf{f}) = f_\theta(\mathbf{X}_H, \mathbf{q}), \quad (1)$$

parameterized by θ , that can predict occupancy probability $o \in [0, 1]$, obstacle segmentation probability $s \in [0, 1]$ and 3D flow $\mathbf{f} \in \mathbb{R}^3$ at *query point* $\mathbf{q} = (q_x, q_y, q_z, q_t)$ in *continuous* 4D space-time (x, y, z, t) .

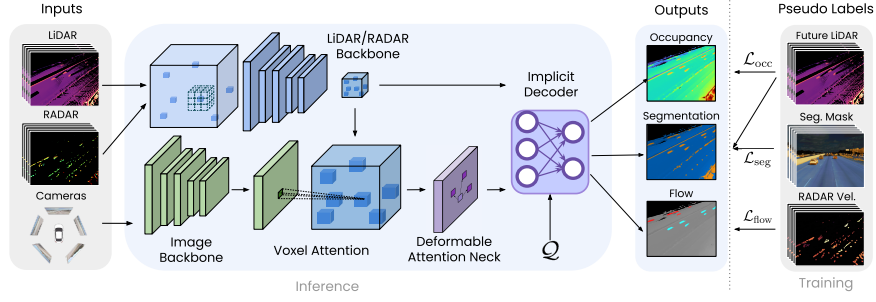


Fig. 2: TRIO’s multi-modal architecture fusing LiDAR, RADAR, and camera inputs via voxel-attention. An implicit decoder generates occupancy, segmentation, and flow for a set of query points $\mathcal{Q}$ supervised by pseudo-labels derived from sensor data.

Geometric occupancy and 3D flow offer a simple representation that captures important characteristics of real-world environments (being occupied by a physical object and its velocity). However, a motion planner must distinguish between what is path-blocking (e.g., a child sitting on the street) and what is traversable (e.g., a drivable slope), motivating the prediction of obstacle segmentation s . To train our method, we require labels for each query point $\mathbf{q}$. We construct supervisory signals (pseudo-labels) as $\bar{\mathbf{y}} = g(\mathbf{X}_T, \mathbf{q})$, which are used in the loss function $\mathcal{L}(\mathbf{y}, \bar{\mathbf{y}})$.

In the remainder of this section, we first detail the architecture of f_θ in Sec. 3.2. We then detail the function g by introducing our segmentation method in Sec. 3.3 and how pseudo-labels are generated from all three sensor modalities in Sec. 3.4. The training procedure is described in Sec. 3.5.

3.2 Unsupervised Multi-Modal Occupancy-Flow Model

Our model follows an encoder-decoder structure and the overall architecture is shown in Fig. 2. First, we encode the measurements of each sensor with two encoders, one for LiDAR-RADAR (f_{lr}) and one for camera (f_c):

$$\mathbf{Z}_c = f_c(\mathbf{C}_H), \quad \mathbf{Z}_{lr} = f_{lr}(\mathbf{L}_H, \mathbf{R}_H). \quad (2)$$

The camera encoder f_c is a pretrained version of ResNet50 [23], which only takes in the most recent camera frame. The LiDAR-RADAR encoder discretizes locations at the voxel level of multiple historical frames, and encodes them as features in a 3D sparse feature map. The features from the LiDAR $\mathbf{l} = [x, y, z, i, \Delta t]^\top \in \mathbb{R}^5$ (including intensity i and relative timestep Δt) and RADAR points $\mathbf{r} = [x, y, z, v_r, \sigma, \Delta t]^\top \in \mathbb{R}^6$ (including range rate v_r , RADAR cross section σ , and Δt) are each encoded through separate shallow MLPs.

The latent representations $\mathbf{Z}_c$ and $\mathbf{Z}_{lr}$ are then fused by allowing the sparse LiDAR features at the coarsest level to attend to the image features through a voxel attention mechanism (f_{voxelatt}), similar to what is done in [40], giving $\mathbf{Z}_0 =$

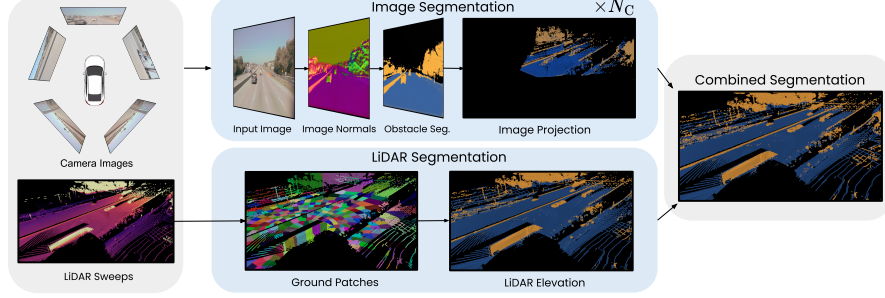


Fig. 3: **Obstacle** and **ground** pseudo-label generation process.

$[\mathbf{V}_{2x}, \mathbf{V}_{4x}, \mathbf{V}_{8x}, \mathbf{M}_{16x}]$, with feature volumes $\mathbf{V}_{2x}, \mathbf{V}_{4x}, \mathbf{V}_{8x}$ and $\mathbf{M}_{16x}$ having been downsampled by 2x, 4x, 8x, and 16x respectively. We then densify the coarsest feature map $\mathbf{M}_{16x}$ and pass it through several deformable attention layers (f_{defatt}) to increase the effective receptive field of the model. Finally, we use the same deformable attention multi-resolution header proposed by [17] for each of the model’s predictions. We describe the specific details in the supplementary. The process is summarized as:

$$\mathcal{Z}_0 = f_{\text{voxelatt}}(\mathbf{Z}_c, \mathbf{Z}_l) \rightarrow \mathcal{Z}_1 = f_{\text{defatt}}(\mathcal{Z}_0) \rightarrow (o, \mathbf{f}, s) = f_{\text{dec}}(\mathcal{Z}_1, \mathbf{Z}_l, \mathbf{q}). \quad (3)$$

3.3 Obstacle Segmentation Model

To provide supervision for obstacle segmentation s , we propose a novel method for extracting obstacle/ground segmentation labels from image and LiDAR features in a probabilistic manner, combining the strength of each sensor modality. This approach ensures that the objects which are distinctly above ground are captured by the LiDAR segmentation, which is robust due to the 3D information it receives, while also detecting smaller close-to-the-ground obstacles that may only be observable in image space. The overall process is illustrated in Fig. 3.

Image Obstacle Segmentation. We propose using an image foundation model to produce estimates of normals and depths of the image for each camera. Specifically, given a camera image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ we apply Metric3Dv2 [25] to obtain per-pixel metric depth $d \in \mathbb{R}^{H \times W}$, depth confidence $c \in \mathbb{R}^{H \times W}$, and surface normals $\hat{\mathbf{n}} \in \mathbb{R}^{H \times W \times 3}$ with associated confidence $c_n \in \mathbb{R}^{H \times W}$. We then follow a similar process to [48] to produce a segmentation of the obstacles relative to the rest of the scene based on the normals and depth information. In contrast to prior work, which produce discrete labels, we output continuous probabilities for obstacles/ground ($p_{\text{obst}}^{\text{image}}$), in addition to probabilities of the validity of the estimation ($p_{\text{valid}}^{\text{image}}$). Producing these continuous image segmentation labels (ISL) enables the downstream model to learn a more calibrated representation.

LiDAR Obstacle Segmentation. We leverage ideas from Patchwork++ [37] for LiDAR segmentation to discretize the point cloud into polar patches, where

for each patch j , a ground plane is characterized by its centroid $\mathbf{c}_j \in \mathbb{R}^3$ and surface normal $\hat{\mathbf{n}}_j \in \mathbb{R}^3$. For a LiDAR return $\mathbf{l} \in \mathbb{R}^3$, we identify its corresponding patch in the xy -plane. We then get the elevation $e \in \mathbb{R}$ as the point to plane distance using the patch normal, and map this into continuous probabilities for the obstacle $p_{\text{obst}}^{\text{LiDAR}}$ and the valid distribution $p_{\text{valid}}^{\text{LiDAR}}$.

Combined Segmentation. To leverage the complementary strengths of both modalities, we employ a confidence-driven fusion using the per-point validity and obstacle probabilities p_{valid} , p_{obst} from both image (projected onto LiDAR) and LiDAR sources. Formally, the combined probability is

$$\mathbf{p}^{\text{comb}} = \begin{cases} \mathbf{p}^{\text{image}}, & \text{if } (p_{\text{valid}}^{\text{LiDAR}} < 0.5) \wedge (p_{\text{valid}}^{\text{image}} > p_{\text{valid}}^{\text{LiDAR}}) \\ \mathbf{p}^{\text{LiDAR}}, & \text{otherwise} \end{cases} \quad (4)$$

with $\mathbf{p}^{\text{SOURCE}} = [p_{\text{valid}}^{\text{SOURCE}}, p_{\text{obst}}^{\text{SOURCE}}]^\top$, $\text{SOURCE} \in \{\text{comb}, \text{image}, \text{LiDAR}\}$. The probability vector $\mathbf{p}^{\text{comb}}$ describes the combined soft labels (CSL) to supervise s . By prioritizing image labels for low-confidence LiDAR segmentation results, this approach captures small visual details while maintaining LiDAR’s robustness to points that are clearly distinct from the ground plane. Additional details of our segmentation method are in the supplementary.

3.4 Multi-Modal Self-Supervision

We will now outline the pseudo-labels used to supervise TRIO. These can all be represented by a 4D query point $\mathbf{q}$, and a label $\bar{y}$, summarized as $\bar{\mathbf{y}} = (\mathbf{q}, \bar{y})$.

Occupancy. We generate dense occupancy pseudo-labels without any human annotation, following the methodology of UnO [1], DIO [17], and GASP [45]. For a LiDAR return $\mathbf{l}$ we assume the continuous ray segment between the sensor position and $\mathbf{l}$ as the unoccupied region $\mathcal{R}^-$, and a short segment immediately behind $\mathbf{l}$ along the ray direction as the occupied region $\mathcal{R}^+$. The sets of occupied and unoccupied points in the scene are $\mathcal{O}^+ = \bigcup_{\mathbf{l} \in \mathbf{L}} \mathcal{R}^+$ and $\mathcal{O}^- = \bigcup_{\mathbf{l} \in \mathbf{L}} \mathcal{R}^-$, respectively. The set of $2N$ sampled points with occupancy pseudo-labels is

$$\mathcal{Y}_{\text{occ}} = \{(\mathbf{q}, 1) \mid \mathbf{q} \in \mathcal{O}^+\} \cup \{(\mathbf{q}, 0) \mid \mathbf{q} \in \mathcal{O}^-\}. \quad (5)$$

Obstacle Segmentation. In order to train for the segmentation class, we revisit the pseudo-labels generated in Sec. 3.3. Recall that this formulation provides a combined probability vector $\mathbf{p}^{\text{comb}}$ for each point $\mathbf{r} \in \mathcal{P}_{\text{sweep}}$ in LiDAR sweep $\mathcal{P}_{\text{sweep}}$, containing both the validity and obstacle probabilities. We therefore use the locations of these LiDAR points with their time of emission (x, y, z, t) as coordinates for our query $\mathbf{q}$, and their corresponding combined probabilities as the soft labels. The set of points with pseudo-labels $\mathcal{Y}_{\text{seg}}$ is

$$\mathcal{Y}_{\text{seg}} = \{(\mathbf{q}, \mathbf{p}^{\text{comb}}(\mathbf{q})) \mid \mathbf{q} \in \mathcal{P}_{\text{sweep}}\}. \quad (6)$$

Instantaneous Flow. In order to train for unsupervised instantaneous flow, we use the Doppler readings from the RADAR sensor as our supervision source.

As RADAR points can be noisy, we first define a refined set of RADAR points $\mathcal{R}_{\text{clean}}$ where the measurement uncertainty is below a certain threshold. Due to the inherent characteristic that RADAR only captures velocity in the radial direction, our pseudo-labels must include the sensor velocity at the time of measurement $\mathbf{v}_{\text{sensor}}$, the unit radial direction of the RADAR point $\mathbf{d}_{\text{radar}}$, and the scalar Doppler velocity v_{doppler} . We define the flow pseudo-label set $\mathcal{Y}_{\text{flow}}$ as:

$$\mathcal{Y}_{\text{flow}} = \{(\mathbf{q}, (v_{\text{doppler}}, \mathbf{v}_{\text{sensor}}, \mathbf{d}_{\text{radar}})) \mid \mathbf{q} \in \mathcal{R}_{\text{clean}}\}. \quad (7)$$

3.5 Training

We train our model with a multi-task loss for occupancy ($\mathcal{L}_{\text{occ}}$), flow ($\mathcal{L}_{\text{flow}}$), and obstacle probability ($\mathcal{L}_{\text{seg}}$)

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{occ}} + \lambda_2 \mathcal{L}_{\text{flow}} + \lambda_3 \mathcal{L}_{\text{seg}} \quad (8)$$

using the pseudo-labels defined by $\mathcal{Y}_{\text{occ}}$, $\mathcal{Y}_{\text{seg}}$, and $\mathcal{Y}_{\text{flow}}$. Let TRIO’s occupancy, obstacle probability, and flow output at a query point $\mathbf{q}$ be denoted as $f_{\theta, \text{occ}}(\mathbf{q}) \in [0, 1]$, $f_{\theta, \text{seg}}(\mathbf{q}) \in [0, 1]$, and $f_{\theta, \text{flow}}(\mathbf{q}) \in \mathbb{R}^3$ respectively. For notational simplicity, we drop the conditioning on the input $\mathbf{X}_H$. The occupancy loss is the binary cross-entropy loss over all query points:

$$\mathcal{L}_{\text{occ}} = \frac{1}{|\mathcal{Y}_{\text{occ}}|} \sum_{(\mathbf{q}, \bar{y}_{\text{occ}}) \in \mathcal{Y}_{\text{occ}}} \text{BCE}(f_{\theta, \text{occ}}(\mathbf{q}), \bar{y}_{\text{occ}}). \quad (9)$$

The obstacle probability loss is a soft focal loss (SFocalLoss):

$$\mathcal{L}_{\text{seg}} = \frac{1}{|\mathcal{Y}_{\text{seg}}|} \sum_{(\mathbf{q}, p^{\text{comb}}(\mathbf{q})) \in \mathcal{Y}_{\text{seg}}} \text{SFocalLoss}(f_{\theta, \text{seg}}(\mathbf{q}), p_{\text{obst}}^{\text{comb}}(\mathbf{q}), p_{\text{valid}}^{\text{comb}}(\mathbf{q})). \quad (10)$$

In contrast to $\mathcal{L}_{\text{occ}}$, which uses hard labels $\bar{y}_{\text{occ}} \in \{0, 1\}$, $\mathcal{L}_{\text{seg}}$ operates on soft targets $p_{\text{obst}}^{\text{comb}} \in [0, 1]$, with each point’s contribution scaled by its label confidence $p_{\text{valid}}^{\text{comb}}$. We find this soft parameterization to be important for generating a well-calibrated model and elaborate further in Sec. 4. Full loss definitions are provided in the supplementary material. Finally, the flow supervision is determined as follows. We begin by moving the model’s instantaneous 3D flow prediction $f_{\theta, \text{flow}}(\mathbf{q})$ into the sensor’s frame of reference by subtracting the sensor’s velocity $\mathbf{v}_{\text{sensor}}$, yielding the relative flow $\mathbf{f}_{\text{rel}}(\mathbf{q}) = f_{\theta, \text{flow}}(\mathbf{q}) - \mathbf{v}_{\text{sensor}}$. We then project this relative flow onto the radial direction vector $\mathbf{d}_{\text{radar}}$ to obtain the predicted radial velocity: $f_{\text{rad}}(\mathbf{q}) = \mathbf{d}_{\text{radar}} \cdot \mathbf{f}_{\text{rel}}(\mathbf{q})$. Finally, we supervise this value using the Doppler measurements with an L_1 loss over the cleaned RADAR points $\mathcal{R}_{\text{clean}}$:

$$\mathcal{L}_{\text{flow}} = \frac{1}{|\mathcal{R}_{\text{clean}}|} \sum_{\mathbf{q} \in \mathcal{R}_{\text{clean}}} |f_{\text{rad}}(\mathbf{q}) - v_{\text{doppler}}|. \quad (11)$$

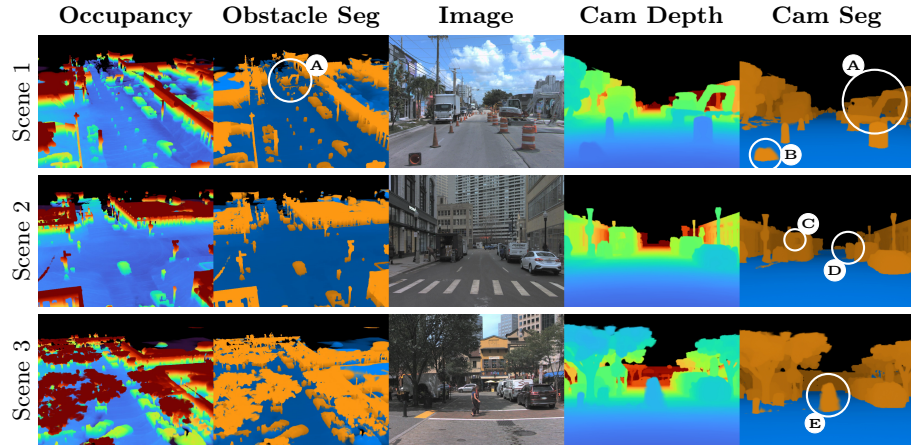


Fig. 4: Visualization of TRIO on AV2. Each row shows (left to right): perspective-view occupancy, perspective-view obstacle segmentation, camera image, rendered depth, and rendered obstacle segmentation. TRIO captures key elements of the scenes: **A)** excavator, **B)** fallen cone, **C)** inside of truck, **D)** open car door, **E)** pedestrians.

4 Experiments

In this section, we investigate the following research questions while comparing against the state-of-the-art: **Q1:** Can TRIO reliably perceive unknown objects in a zero-shot, open-set manner? **Q2:** Does TRIO’s architecture lead to a better 4D geometric representation? **Q3:** Does unsupervised multi-modal pre-training improve downstream task performance via fine-tuning?

Datasets. We use the *Argoverse 2* (AV2) [75] *Sensor* split (700 train/ 150 val sequences) and evaluate on the AV2 Occupancy Forecasting leaderboard. The dataset *Spotting the Unexpected* (STU) [51] uniquely provides 3D LiDAR point labels for rarely seen objects on the road, aiming to test models on anything perception. We utilize all 22 available sequences with labels from STU (train and val) for evaluation. Note, STU is only for *zero-shot evaluation* – i.e. we train TRIO and baselines on AV2 and apply it without any fine-tuning on STU, containing different sensor characteristics. Given the limited label range and no RADAR in previous datasets, we use an unreleased dataset *Long Range RADAR* (LRR) where we label bounding boxes up to 350m in front of the SDV. LRR has 2280 sequences (highway and urban). See supplementary for details.

4.1 Zero-Shot Anything Perception (Q1)

We evaluate our method’s ability to perceive unusual, path-blocking objects on STU. First, we assess pseudo-label quality by comparing our ISL and CSL against STU ground truth. Using the correspondence between image and LiDAR labels in STU, we benchmark our soft pseudo-labels against open-set 2D and 3D

Table 1: Pseudo-label model segmentation performance on STU. MCPP refers to MaskCLIP++.

Method	mIoU $\uparrow$	F1 $\uparrow$	Recall $\uparrow$
MCPP + SAM	22.2	38.5	51.8
MCPP + SAM (Oracle)	22.2	38.3	51.9
MCPP + FC-CLIP	24.8	30.3	43.9
MCPP + FC-CLIP (Oracle)	27.4	35.1	52.4
MCPP + MAFT+	25.6	34.0	51.6
MCPP + MAFT+ (Oracle)	28.0	37.8	59.5
GroundedSAM	22.0	27.9	35.3
GroundedSAM (Oracle)	26.8	18.7	55.2
OWLv2 + SAM	20.2	21.9	23.4
OWLv2 + SAM (Oracle)	28.6	29.5	57.2
Patchwork++	42.1	51.4	62.8
ISL (Ours)	38.6	35.0	76.3
CSL (Ours)	44.6	57.8	79.9

Table 2: Semantic occupancy prediction performance of distilled methods. $\dagger$ is QueryOcc+ with LiDAR input.

Method	mIoU $\uparrow$	AP $\uparrow$	F1 $\uparrow$	Rec. $\uparrow$
UnO (ISL)	40.1	7.4	11.6	23.5
DiO (ISL)	49.7	29.3	34.3	50.2
UnO (CSL)	40.6	23.1	29.8	25.5
DiO (CSL)	49.8	74.8	72.9	65.6
TriO (MCPP + SAM)	27.9	25.7	27.5	24.0
TriO (MCPP + FC-CLIP)	23.1	2.2	5.2	56.7
TriO (MCPP + MAFT+)	20.6	10.2	14.8	10.0
TriO (GroundedSAM) $\dagger$	50.4	50.1	52.1	54.8
TriO (OWLv2 + SAM)	50.1	72.3	70.1	68.8
TriO (Patchwork++)	39.4	74.8	70.8	59.5
TriO (ISL) (Ours)	41.8	57.1	53.7	59.9
TriO (CSL Hard) (Ours)	50.3	76.1	71.7	64.5
TriO (CSL Soft) (Ours)	54.9	81.6	77.7	71.3

baselines on equal footing. We also provide qualitative results on AV2 in Fig. 4.

Anything Perception Baselines. We benchmark against state-of-the-art open-vocabulary methods, including *detect-then-segment* (GroundedSAM [59], OWLv2 [49]+SAM [35]) and *segment-then-classify* (MaskCLIP++ [79] with FC-CLIP [77], MAFT+ [31], or SAM) approaches. Following [53], we use standard prompts for ground ($\mathcal{P}_{\text{ground}}$) and non-ground ($\mathcal{P}_{\text{obstacle}}$) classes; see supplementary material for details. To establish an upper performance bound, we include an (Oracle) variant where scenes are manually inspected and specific unknown object classes are appended to $\mathcal{P}_{\text{obstacle}}$. Finally, we also include Patchwork++ [37] as a LiDAR-based baseline. For the second set of experiments, we train TriO’s segmentation objective ($\mathcal{L}_{\text{seg}}$) using each of the pseudo-label from the baselines. This allows us to measure how effectively open-world knowledge can be distilled into the downstream world model, and in turn how well this model generalizes when evaluated on a different domain. Notably, our model under GroundedSAM supervision serves as a LiDAR-enhanced version of QueryOcc+ [42]. Finally, we compare against LiDAR-only models (UnO, DiO) trained with ISL and CSL.

Anything Perception Metrics. We report *mIoU* (mean Intersection over Union) [67, 72] for semantic predictions and *F1* score to evaluate pseudo-label quality at a fixed operating point without unfairly advantaging soft probability methods. We emphasize *Recall* to capture safety-critical missed detections, while Average Precision (*AP*) evaluates continuous probabilities in distilled occupancy models. See supplementary for further details.

Comparison of Pseudo-Label Methods. Our generated pseudo-labels (CSL) achieve the best balance of detecting the obstacles (measured by recall) on the road without too many false positive detections (F1 and mIoU) as shown in Tab. 1. Notably, this strong performance is not solely reliant on multi-modal fusion; even our image-only pseudo-labels (ISL) surpass all evaluated baselines in terms of recall. We also find that text-based segmentation baselines, even when prompted with the exact text description of the objects in the road, struggle to

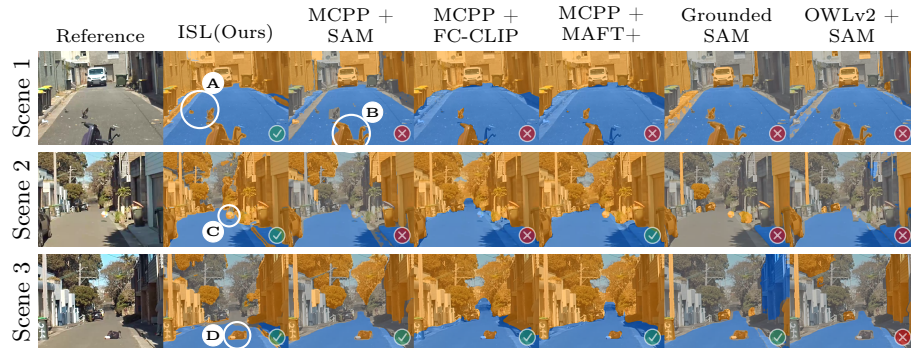


Fig. 5: Segmentation results on the STU validation set. Each scene contains an unexpected obstacle: **A**) two pigeons and **B**) chair, **C**) ball, and **D**) bookbag. An obstacle segmentation is correct if the road is labeled traversable (**blue**) and the obstacle is fully segmented as an obstacle (**orange**). ISL robustly segments long-tail obstacles.

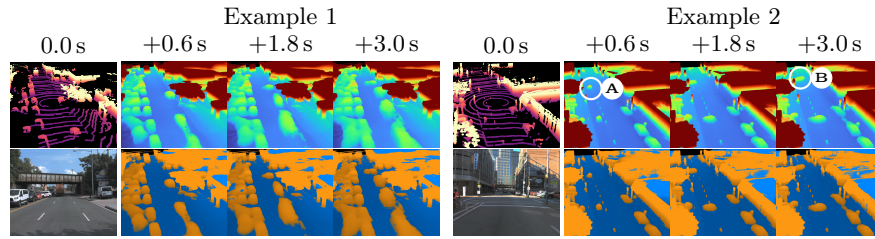


Fig. 6: Scene forecasting results of TRIO on AV2. LiDAR and camera sensor inputs are displayed at present time for reference. TRIO’s forecasts maintain coherent obstacle segmentation throughout time, distinguishing occupied from free space even as agents move. TRIO captures a bicyclist **A**) and predicts its future multimodality **B**).

generalize to these rarely seen classes, reinforcing the findings of [70]. We provide visual evidence for that claim in Fig. 5 and the supplementary. In contrast, our model, relying on geometry, is able to generalize in a more robust manner.

Semantic Occupancy Evaluation. As shown in Tab. 2 TRIO performs the best after the obstacle labels have been distilled into our occupancy model. We also find that using soft probabilistic labels (CSL Soft) tends to result in better performance than when trained with thresholded discretized labels (CSL Hard) due to better calibration (see AP, mIoU). TRIO improves also over LiDAR-only methods highlighting the benefit of using multi-modal sensor information.

4.2 Geometric Occupancy/ LiDAR Point Cloud Forecasting (Q2)

We evaluate TRIO on the AV2 Occupancy Forecasting benchmark; the goal of this benchmark is to determine how TRIO performs on occupancy prediction (visualized in Fig. 6), evaluated via LiDAR point cloud forecasting. Following

Table 3: LiDAR forecasting on the Argoverse 2 leaderboard.

	L1 (m) ↓	AbsRel (%) ↓	NFCD (m ²) ↓	CD (m ²) ↓
RayTracing	4.88	35.00	3.62	17.03
occformer_ep15	3.57	22.00	3.35	91.61
4D-Occ	3.22	19.00	2.45	72.74
occformer	2.97	17.00	2.03	71.33
Progressive ARM	2.32	13.00	1.81	71.41
UnO	2.24	12.00	0.86	8.10
DIO	2.11	11.00	0.85	13.96
TriO	1.82	10.63	0.93	10.90

prior work [1, 17, 34], after training the world model f_θ , we freeze all parameters θ and introduce a LiDAR forecasting header f_ψ to predict LiDAR depth. We provide more details in the supplementary.

LiDAR Forecasting Baselines. We compare our method against the state-of-the-art from the Argoverse 2 leaderboard. Notably, 4D-Occ [34] serves as a representative unsupervised baseline for explicit occupancy prediction via voxel grid forecasting. In contrast, UnO [1] and DiO [17], also perform implicit occupancy prediction. However, unlike TriO, these methods do not incorporate camera input, representing architectural differences in sensor modality.

LiDAR Forecasting Metrics: For point cloud forecasting we use well-established metrics of the AV2 Occupancy Forecasting leaderboard, namely Chamfer distance (CD), Near Field Chamfer Distance (*NFCD*), depth L1 error (*L1*), depth relative *L1* error (AbsRel) as our metrics.

Comparison against state-of-the-art. Tab. 3 compares TriO against all the AV2 leaderboard baselines. TriO is best in the L1 and AbsRel metrics, while competitive with baselines in other metrics. This underlines TriO’s ability to accurately forecast occupancy and its effectiveness as a world model.

4.3 Fine-tuning for Downstream Occupancy Tasks (Q3)

We show the versatility of the TriO pre-trained feature extractor by evaluating its performance on various downstream tasks. We follow [1, 45] and finetune the model to forecast semantic 3D (AV2) and semantic 2D BEV (LRR) occupancy and flow over time supervised with bounding box labels for the vehicle class.

Downstream Occupancy Baselines. For fine-tuning experiments, we pre-train UnO and DiO baselines via LiDAR self-supervision then fine-tune on 2D or 3D bounding box labels. We also compare against ImplicitO [2] which does not utilize unsupervised pre-training.

Downstream Occupancy Metrics. We use well-established metrics to evaluate the performance of the different methods on these downstream tasks by following prior work [1, 2, 17, 47]. We report mAP, SoftIoU [47] for occupancy and End Point Error (EPE) for flow. For LRR, we also evaluate each metric at over 200m in front of the SDV. We hypothesize that RADAR specifically having a longer effective range is able to improve our model performance at these dis-

Table 4: Fine-tuning downstream performance on AV2 and LRR.

Method	AV2			LRR					
	mAP $\uparrow$	Soft-IoU $\uparrow$	EPE $\downarrow$	All			200m+		
				mAP $\uparrow$	Soft-IoU $\uparrow$	EPE $\downarrow$	mAP $\uparrow$	Soft-IoU $\uparrow$	EPE $\downarrow$
Implicit-O	45.3	18.9	3.04	49.1	22.6	5.69	29.5	10.7	10.2
UnO	59.0	29.2	2.04	51.6	21.2	4.92	27.2	8.30	8.29
DIO	57.6	26.2	1.43	58.0	31.5	3.01	36.2	13.8	4.40
TRIO	63.3	31.9	1.97	60.5	35.7	2.86	42.6	21.4	4.01

tances, where other sensor readings may be too sparse to be consistently useful. This emphasizes the use case of LRR as only it is able to provide validation labels at a distance of 350m in front of the SDV.

Semantic Occupancy and Flow Forecasting. Tab. 4 and Tab. 5 show the BEV and 3D occupancy and flow forecasting results on AV2 and LRR. Our method outperforms all the baselines when finetuned to the supervised occupancy task. The drop in EPE performance in Tab. 4 compared to DIO is explained as TriO on AV2 has no flow supervision (due to the absence of RADAR data) during pre-training, whereas DIO does. Notably, our model outperforms the baselines significantly at range. The results show an improvement of 6.4 AP over the state-of-the-art showcasing the benefits of using RADAR at long-range.

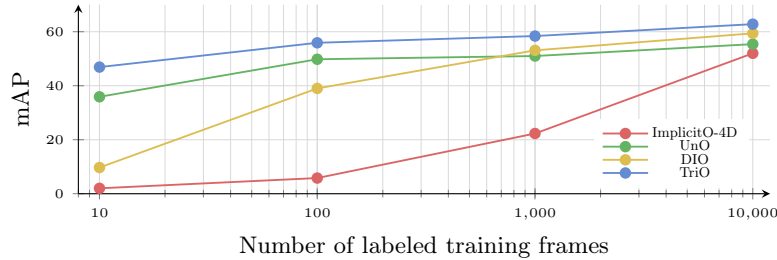
Ablation Study. We evaluate the contributions of pre-training, sensor modalities, and loss objectives in Tab. 5. Unsupervised pre-training provides consistent gains across all metrics, validating the world model objectives. Omitting the radar loss leads to degradation in flow metrics across all ranges. Conversely, adding camera inputs improves far range occupancy performance at the expense of EPE; this suggests that single-frame camera features, which lack temporal context, may introduce noise into the flow estimation process. Removing RADAR as an input modality significantly degrades performance on long range metrics, supporting our hypothesis that this sensor modality directly lends itself to long range tasks.

Scaling Labeled Data. In Fig. 7 we scale the number of labeled training samples when fine-tuning pre-trained models (TRIO, UnO, DIO) and models trained from scratch (ImplicitO-4D). TRIO outperforms at all levels of supervision, and with 10^2 labeled training examples, reaches similar performance to the strongest baselines with 10^4 examples. TRIO is *two-orders of magnitude* more sample efficient, underscoring the generalizability of our method. Due to its novel obstacle segmentation supervision TRIO better understands obstacles and free space. This underlines the strong representation learned during pre-training and TRIO’s ability to encode the geometry, semantics, and dynamics of the scene.

Limitations, Discussion, and Future Work. Currently, TRIO is limited to using RADAR for flow self-supervision, but future work could add occupancy supervision from RADAR. While our model uses RADAR to generalize occupancy

Table 5: Ablation studies of TriO on LRR.

Method	All			200m+		
	mAP $\uparrow$	Soft-IoU $\uparrow$	EPE $\downarrow$	mAP $\uparrow$	Soft-IoU $\uparrow$	EPE $\downarrow$
TriO (no pretraining)	59.7	33.9	4.09	41.0	19.8	4.70
TriO (no radar input)	57.7	31.9	3.04	34.5	14.3	4.59
TriO (no camera input)	60.5	35.7	2.68	42.2	20.9	3.64
TriO (no radar loss)	60.5	35.6	4.35	42.6	21.3	6.02
TriO	60.5	35.7	2.86	42.6	21.4	4.01

**Fig. 7:** Occupancy fine-tuning under a varying number of labelled frames.

predictions beyond the LiDAR range, RADAR can also be used as additional occupancy supervision in adverse weather to robustify predictions. This requires techniques beyond the line-of-sight assumption done in [36] due to nature of RADAR returns. Furthermore, the ability to perceive and forecast the movement of rare objects is an understudied problem in the SDV space. We see a general need for additional larger scale benchmarks focusing on the task of *Anything Perception* to ensure the overall robustness of many pre-existing systems.

5 Conclusion

We present TriO, a novel unsupervised tri-modal occupancy model that learns to predict 4D occupancy, obstacle semantics, and flow by leveraging the complementary physical synergies of camera, LiDAR, and RADAR data. It achieves state-of-the-art performance across multiple tasks, including zero-shot anything perception, LiDAR forecasting, and 3D/4D occupancy prediction, while being unsupervised and independent of additional costly human annotations. We observe failure-cases of open-vocabulary methods for rare classes, and show that TriO outperforms these methods in open-world scenarios. Notably, it achieves robust performance in perceiving anything relative to the drivable surface, including rare classes, showcasing its potential for a scalable deployment of SDVs.

References

1. Agro, B., Sykora, Q., Casas, S., Gilles, T., Urtasun, R.: Uno: Unsupervised occupancy fields for perception and forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14487–14496 (June 2024)
2. Agro, B., Sykora, Q., Casas, S., Urtasun, R.: Implicit occupancy flow fields for perception and prediction in self-driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1379–1388 (2023)
3. Anonymous: Fantastic tractor-dogs and how not to find them with open-vocabulary detectors. In: The Fourteenth International Conference on Learning Representations (2026), <https://openreview.net/forum?id=jUuXNrG7wh>
4. Barsellotti, L., Bianchi, L., Messina, N., Carrara, F., Cornia, M., Baraldi, L., Falchi, F., Cucchiara, R.: Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22025–22035 (October 2025)
5. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9296–9306 (2019)
6. Boeder, S., Gigengack, F., Risse, B.: Gaussianflowocc: Sparse and weakly supervised occupancy estimation using gaussian splatting and temporal flow. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 24943–24954 (October 2025)
7. Boeder, S., Gigengack, F., Roesler, S., Caesar, H., Risse, B.: Shelfocc: Native 3d supervision beyond lidar for vision-based occupancy estimation (2025), <https://arxiv.org/abs/2511.15396>
8. Cao, A.Q., de Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3981–3991 (2022). <https://doi.org/10.1109/CVPR52688.2022.00396>
9. Casas, S., Agro, B., Mao, J., Gilles, T., Cui, A., Li, T., Urtasun, R.: Detra: A unified model for object detection and trajectory forecasting. In: ECCV (2024)
10. Casas, S., Luo, W., Urtasun, R.: Intentnet: Learning to predict intention from raw sensor data. In: CoRL (2018)
11. Casas, S., Sadat, A., Urtasun, R.: Mp3: A unified model to map, perceive, predict and plan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14403–14412 (2021)
12. Cui, A., Casas, S., Sadat, A., Liao, R., Urtasun, R.: Lookout: Diverse multi-future prediction and planning for self-driving. In: ICCV (2021)
13. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2432–2443 (2017)
14. Diehl, C., Feicho, E., Schwambach, A., Dammeier, T., Mares, E., Bertram, T.: Radar-based dynamic occupancy grid mapping and object detection. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). pp. 1–6 (2020)
15. Diehl, C., Klosek, T., Krueger, M., Murzyn, N., Osterburg, T., Bertram, T.: Energy-based potential games for joint motion forecasting and control. In: Proceedings of The 7th Conference on Robot Learning. pp. 3112–3141 (2023)

16. Diehl, C., Sievernich, T.S., Krüger, M., Hoffmann, F., Bertram, T.: Uncertainty-aware model-based offline reinforcement learning for automated driving. *IEEE Robotics and Automation Letters* pp. 1167–1174 (2023)
17. Diehl, C., Sykora, Q., Agro, B., Gilles, T., Casas, S., Urtasun, R.: Dio: Decomposable implicit 4d occupancy-flow world model. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27456–27466 (2025). <https://doi.org/10.1109/CVPR52734.2025.02557>
18. Ding, F., Wen, X., Zhu, Y., Li, Y., Lu, C.X.: Radarocc: Robust 3d occupancy prediction with 4d imaging radar. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) *Advances in Neural Information Processing Systems*. vol. 37, pp. 101589–101617. Curran Associates, Inc. (2024). <https://doi.org/10.52202/079017-3222>, https://proceedings.neurips.cc/paper_files/paper/2024/file/b81d83165e3145a2e7d33bb5e33ea913-Paper-Conference.pdf
19. Gan, W., Liu, F., Xu, H., Mo, N., Yokoya, N.: Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 28980–28990 (October 2025)
20. Gomes, T., Matias, D., Campos, A., Cunha, L., Roriz, R.: A survey on ground segmentation methods for automotive lidar sensors. *Sensors* **23**(2) (2023). <https://doi.org/10.3390/s23020601>
21. Gu, S., Yin, W., Jin, B., Guo, X., Wang, J., Li, H., Zhang, Q., Long, X.: Dome: Taming diffusion model into high-fidelity controllable occupancy world model (2024), <https://arxiv.org/abs/2410.10429>
22. Guo, E., An, P., Yang, Y., Liu, Q., Liu, A.A.: Fsf-net: Enhance 4d occupancy forecasting with coarse bev scene flow for autonomous driving (2024), <https://arxiv.org/abs/2409.15841>
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
24. Hu, A., Murez, Z., Mohan, N., Dudas, S., Hawke, J., Badrinarayanan, V., Cipolla, R., Kendall, A.: Fiery: Future instance prediction in bird’s-eye view from surround monocular cameras. In: *ICCV* (2021)
25. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
26. Huang, Y., Zheng, W., Zhang, B., Zhou, J., Lu, J.: Selfocc: Self-supervised vision-based 3d occupancy prediction. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19946–19956 (2024). <https://doi.org/10.1109/CVPR52733.2024.01885>
27. Ivanovic, B., Leung, K., Schmerling, E., Pavone, M.: Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach. *RA-L* (2020)
28. Ivanovic, B., Pavone, M.: The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In: *ICCV* (2019)
29. Jevtić, A., Reich, C., Wimbauer, F., Hahn, O., Rupprecht, C., Roth, S., Cremers, D.: Feed-forward scenedino for unsupervised semantic scene completion. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 6784–6796 (October 2025)

30. Jiang, H., Liu, L., Cheng, T., Wang, X., Lin, T., Su, Z., Liu, W., Wang, X.: Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11960–11970 (June 2025)
31. Jiao, S., Zhu, H., Huang, J., Zhao, Y., Wei, Y., Humphrey, S.: Collaborative vision-text representation optimizing for open-vocabulary segmentation. In: European Conference on Computer Vision (2024)
32. Jose, C., Moutakanni, T., Kang, D., Baldassarre, F., Darcet, T., Xu, H., Li, D., Szafraniec, M., Ramamonjisoa, M., Oquab, M., Siméoni, O., Vo, H.V., Labatut, P., Bojanowski, P.: Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24905–24916 (June 2025)
33. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the Fourth Eurographics Symposium on Geometry Processing. p. 61–70. Eurographics Association, Goslar, DEU (2006)
34. Khurana, T., Hu, P., Held, D., Ramanan, D.: Point cloud forecasting as a proxy for 4d occupancy forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1116–1124 (2023)
35. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
36. Kung, P.C., Harisha, S., Vasudevan, R., Eid, A., Skinner, K.A.: Radarsplat: Radar gaussian splatting for high-fidelity data synthesis and 3d reconstruction of autonomous driving scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 27596–27606 (October 2025)
37. Lee, S., Lim, H., Myung, H.: Patchwork++: Fast and robust ground segmentation solving partial under-segmentation using 3d point cloud. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 13276–13283 (2022). <https://doi.org/10.1109/IROS47612.2022.9981561>
38. Li, P., Ding, S., Zhou, Y., Zhang, Q., Inak, O., Triess, L., Hanselmann, N., Cordts, M., Zell, A.: Ago: Adaptive grounding for open world 3d occupancy prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8645–8655 (October 2025)
39. Li, X., Cheng, X., Ju, X., Peng, Y., Hu, J., Li, J.: Optimization on the polarization and waveform of radar for better target detection performance under rainy condition. *Remote Sensing* **16**(14) (2024). <https://doi.org/10.3390/rs16142557>, <https://www.mdpi.com/2072-4292/16/14/2557>
40. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9087–9098 (June 2023)
41. Liang, M., Yang, B., Zeng, W., Chen, Y., Hu, R., Casas, S., Urtasun, R.: Pnpnet: End-to-end perception and prediction with tracking in the loop. In: CVPR (2020)
42. Lilja, A., Lan, J., Fu, J., Hammarstrand, L.: Queryocc: Query-based self-supervision for 3d semantic occupancy (2025), <https://arxiv.org/abs/2511.17221>
43. Liu, R., Wu, X., Chen, X., Hu, L., Lou, Y.: 4d-rolls: 4d radar occupancy learning via lidar supervision. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 13602–13609 (2025). <https://doi.org/10.1109/IROS60139.2025.11246471>

44. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision (ECCV) (2024)
45. Ljungbergh, W., Lilja, A., Ling, A.T.A.L., Lindström, C., Verbeke, W., Fu, J., Petersson, C., Hammarstrand, L., Felsberg, M.: Gasp: Unifying geometric and semantic self-supervised pre-training for autonomous driving (2025), <https://arxiv.org/abs/2503.15672>
46. Ma, J., Chen, X., Huang, J., Xu, J., Luo, Z., Xu, J., Gu, W., Ai, R., Wang, H.: Cam4DOcc: Benchmark for Camera-Only 4D Occupancy Forecasting in Autonomous Driving Applications. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2024)
47. Mahjourian, R., Kim, J., Chai, Y., Tan, M., Sapp, B., Anguelov, D.: Occupancy flow fields for motion forecasting in autonomous driving. *IEEE Robotics and Automation Letters* **7**(2), 5639–5646 (2022)
48. Man, Y., Weng, X., Kitani, K.M.: Groundnet: Segmentation-aware monocular ground plane estimation with geometric consistency. *CoRR* **abs/1811.07222** (2018), <http://arxiv.org/abs/1811.07222>
49. Minderer, M., Gritsenko, A., Hounsby, N.: Scaling open-vocabulary object detection. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*. vol. 36, pp. 72983–73007. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/e6d58fc68c0f3c36ae6e0e64478a69c0-Paper-Conference.pdf
50. Nealen, A., Igarashi, T., Sorkine, O., Alexa, M.: Laplacian mesh optimization. In: *Proceedings of the 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*. p. 381–389 (2006)
51. Nekrasov, A., Burdorf, M., Worrall, S., Leibe, B., Berrio Perez, J.S.: 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 (2025). <https://doi.org/10.1109/CVPR52734.2025.01109>
52. Nuss, D., Reuter, S., Thom, M., Yuan, T., Krehl, G., Maile, M., Gern, A., Dietmayer, K.: A random finite set approach for dynamic occupancy grid maps with real-time application. *The International Journal of Robotics Research* **37**(8), 841–866 (2018). <https://doi.org/10.1177/0278364918775523>, <https://doi.org/10.1177/0278364918775523>
53. Osep, A., Meinhardt, T., Ferroni, F., Peri, N., Ramanan, D., Leal-Taixá, L.: Better call sal: Towards learning to segment anything in lidar. In: European Conference on Computer Vision (ECCV) (2024)
54. Paigwar, A., Erkent, O., Sierra-Gonzalez, D., Laugier, C.: Gndnet: Fast ground plane estimation and point cloud segmentation for autonomous vehicles. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2150–2156 (2020). <https://doi.org/10.1109/IROS45743.2020.9340979>
55. Palladin, E., Brucker, S., Ghilotti, F., Narayanan, P., Bijelic, M., Heide, F.: Self-supervised sparse sensor fusion for long range perception. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 27498–27509 (October 2025)
56. Pavković, T., Mahani, M.A.N., Niedermayer, J., Betz, J.: Gaussianfusionocc: A seamless sensor fusion approach for 3d occupancy prediction using 3d gaussians (2025), <https://arxiv.org/abs/2507.18522>

57. Phillion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: ECCV (2020)
58. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), <https://arxiv.org/abs/2103.00020>
59. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024), <https://arxiv.org/abs/2401.14159>
60. Richter, S., Bieder, F., Wirges, S., Stiller, C.: A dual evidential top-view representation to model the semantic environment of automated vehicles. *IEEE Transactions on Intelligent Vehicles* **9**(1), 2688–2700 (2024). <https://doi.org/10.1109/TIV.2023.3284400>
61. Rife, J.H., Li, Y.: Characterizing lidar range-measurement ambiguity due to multiple returns. In: Proceedings of the 38th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2025). pp. 1949–1963. Baltimore, Maryland (Sep 2025). <https://doi.org/10.33012/2025.20444>, <https://doi.org/10.33012/2025.20444>
62. Roldao, L., de Charette, R., Verroust-Blondet, A.: 3d semantic scene completion: a survey pp. 1978—2005 (2022)
63. Sadat, A., Casas, S., Ren, M., Wu, X., Dhawan, P., Urtasun, R.: Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In: ECCV (2020)
64. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: ECCV (2020)
65. Steinke, N., Goehring, D., Rojas, R.: Groundgrid: Lidar point cloud ground segmentation and terrain estimation. *IEEE Robotics and Automation Letters* **9**(1), 420–426 (2024). <https://doi.org/10.1109/LRA.2023.3333233>
66. Tanzmeister, G., Thomas, J., Wollherr, D., Buss, M.: Grid-based mapping and tracking in dynamic environments using a uniform evidential environment representation. In: 2014 IEEE International Conference on Robotics and Automation (ICRA). pp. 6090–6095 (2014). <https://doi.org/10.1109/ICRA.2014.6907756>
67. Tian, X., Jiang, T., Yun, L., Mao, Y., Yang, H., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*. vol. 36, pp. 64318–64330 (2023)
68. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmesen, J., Steiner, A., Zhai, X.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features (2025), <https://arxiv.org/abs/2502.14786>
69. Uttarkabat, S., Appukuttan, S., Gupta, K., Nayak, S., Palo, P.: Bloomnet: Perception of blooming effect in adas using synthetic lidar point cloud data. In: 2024 IEEE Intelligent Vehicles Symposium (IV). pp. 1886–1892 (2024). <https://doi.org/10.1109/IV55156.2024.10588461>
70. Šarić, J., Martinović, I., Kristan, M., Šegvić, S.: What holds back open-vocabulary segmentation? In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 4256–4266 (October 2025)

71. Wang, L., Kim, S.W., Yang, J., Yu, C., Ivanovic, B., Waslander, S.L., Wang, Y., Fidler, S., Pavone, M., Karkus, P.: Distillnerf: Perceiving 3d scenes from single-glance images by distilling neural fields and foundation model features. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
72. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17850–17859 (October 2023)
73. Weng, X., Wang, J., Held, D., Kitani, K.: 3d multi-object tracking: A baseline and new evaluation metrics. In: IROS (2020)
74. Weng, X., Yuan, Y., Kitani, K.: Ptp: Parallelized tracking and prediction with graph neural networks and diversity sampling. RA-L (2021)
75. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv:2301.00493 (2023)
76. Yan, X., Gao, J., Li, J., Zhang, R., Li, Z., Huang, R., Cui, S.: Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3101–3109 (2021)
77. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 32215–32234. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/661caac7729aa7d8c6b8ac0d39ccbc6a-Paper-Conference.pdf
78. Yuan, W., Khot, T., Held, D., Mertz, C., Hebert, M.: Pcn: Point completion network. In: 2018 International Conference on 3D Vision (3DV). pp. 728–737 (2018)
79. Zeng, Q.S., Li, Y., Zhou, D., Li, G., Hou, Q., Cheng, M.M.: High-quality mask tuning matters for open-vocabulary segmentation (2025), <https://arxiv.org/abs/2412.11464>
80. Zhang, C., Yan, J., Wei, Y., Li, J., Liu, L., Tang, Y., Duan, Y., Lu, J.: Occnerf: Advancing 3d occupancy prediction in lidar-free environments (2024), <https://arxiv.org/abs/2312.09243>
81. Zhao, L., Luo, Y., Hay, J., Gan, L.: Shelfgaussian: Shelf-supervised open-vocabulary gaussian-based 3d scene understanding (2025), <https://arxiv.org/abs/2512.03370>
82. Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., Lu, J.: Occworld: Learning a 3d occupancy world model for autonomous driving. In: ECCV (2024)
83. Zimmermann, K., Petricek, T., Šalanský, V., Svoboda, T.: Learning for active 3d mapping. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 1548–1556 (2017)
84. Zuo, S., Zheng, W., Huang, Y., Zhou, J., Lu, J.: Gaussianworld: Gaussian world model for streaming 3d occupancy prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6772–6781 (June 2025)