

# UniOcc: A Unified Benchmark for Occupancy Forecasting and Prediction in Autonomous Driving

Yuping Wang<sup>1,2,4,\*</sup> Xiangyu Huang<sup>3,†</sup> Xiaokang Sun<sup>1,†</sup> Mingxuan Yan<sup>1</sup> Shuo Xing<sup>4</sup>  
Zhengzhong Tu<sup>4</sup> Jiachen Li<sup>1,‡</sup>

<sup>1</sup>University of California, Riverside

<sup>2</sup>University of Michigan

<sup>3</sup>University of Wisconsin, Madison

<sup>4</sup>Texas A&M University

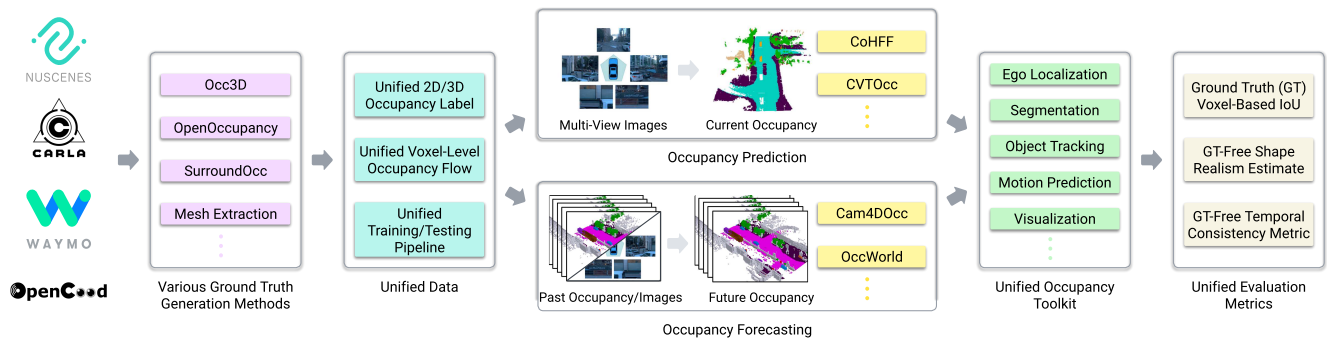


Figure 1. Our UniOcc framework incorporates various occupancy label generation methods from multiple data sources, provides the training/testing pipeline & toolkit for a variety of occupancy tasks, and supports comprehensive evaluation metrics.

## Abstract

We introduce *UniOcc*, a comprehensive, unified benchmark and toolkit for occupancy forecasting (i.e., predicting future occupancies based on historical information) and occupancy prediction (i.e., predicting current-frame occupancy from camera images). *UniOcc* unifies the data from multiple real-world datasets (i.e., *nuScenes*, *Waymo*) and high-fidelity driving simulators (i.e., *CARLA*, *OpenCOOD*), providing 2D/3D occupancy labels and annotating innovative per-voxel flows. Unlike existing studies that rely on suboptimal pseudo labels for evaluation, *UniOcc* incorporates novel evaluation metrics that do not depend on ground-truth labels, enabling robust assessment on additional aspects of occupancy quality. Through extensive experiments on state-of-the-art models, we demonstrate that large-scale, diverse training data and explicit flow information significantly enhance occupancy prediction and forecasting performance. Our data and code are available at <https://uniocc.github.io/>.

## 1. Introduction

Occupancy grid map (OGM) has been an effective representation of traffic scenes. It provides a rasterized view of the environment by discretizing the space into a grid of 2D or

3D cells, each indicating the presence or absence of objects such as vehicles, pedestrians, and static obstacles [53]. Obtaining robust occupancy representations of the dynamic environments is essential for safe motion planning, end-to-end driving systems, and various off-board applications (e.g., data generation for model development). There are two representative tasks in the context of autonomous driving: a) Occupancy forecasting [2, 24, 35, 55] aims to predict future occupancies based on historical occupancy/image observations, which enables autonomous systems to anticipate dynamic changes in the environment; b) Occupancy prediction [1, 51] focuses on estimating the current occupancy grid map from raw sensor data, which reconstructs the surrounding scene in a structured and interpretable format. While recent work has made significant progress, several critical issues have to be addressed.

**Suboptimal Occupancy Labels and Metrics.** Widely used driving datasets (e.g., *nuScenes* [3] and *Waymo* [30]) lack official occupancy annotations. Existing research thus relies on pseudo ground truth labels derived from heuristics or manual labeling [31, 36, 42]. These pseudo labels often capture only the reflective surfaces (e.g., car sides hit by LiDAR), failing to represent the true 3D occupancy of the scene. Models trained on these suboptimal labels inevitably produce suboptimal results. Even worse, standard metrics like Intersection-over-Union (IoU) cannot reveal such quality issues because they compare predictions solely against

\*ypw@umich.edu, †Equal contribution

‡jiachen.li@ucr.edu, Corresponding author

the flawed pseudo labels. To mitigate these pitfalls, we propose novel evaluation metrics that do not rely on pseudo ground truth labels. These metrics provide additional aspects for occupancy quality evaluation.

**Domain Constraints and Fragmented Data.** Existing occupancy forecasting and prediction methods are mostly restricted to a single dataset. For example, models trained on the nuScenes dataset [3] often are not directly applicable to the Waymo dataset [30] due to the differences in sensor configurations, data formats, sampling rates, and annotation types. Furthermore, each dataset typically requires its own dedicated tools and data loaders. Inspired by efforts in unified trajectory prediction (*e.g.*, UniTraj [6]), we introduce a unified occupancy dataset and framework that standardizes these discrepancies, which enables cross-dataset training with a single command. Our framework also leverages CARLA simulations to provide virtually unlimited, diverse training data. Furthermore, our unification enables the cross-domain evaluation of occupancy methods and allows us to analyze their out-of-distribution generalization performance, which is critical for safe autonomous driving.

**Lack of Per-Voxel Flows.** Current 3D occupancy labels generally lack motion flow information within each voxel, which limits the ability of models to exploit dynamic scene cues. While flows may not be critical in camera-to-occupancy prediction, they are crucial for occupancy forecasting tasks that must capture object and agent movement over time. By including forward and reverse flows for each voxel, our unified dataset facilitates more robust forecasting and simplifies downstream tasks such as object tracking. Furthermore, to our knowledge, we are the first to use per-voxel flows for 3D occupancy forecasting.

**Lack of Support for Cooperative Occupancy Forecasting.** Cooperative driving is a growing area, with research in cooperative perception and prediction [29, 40, 50], but there has been no dataset available for cooperative occupancy forecasting. Building on OpenCood [49], our framework and dataset extend to multi-agent scenarios, serving as the first to support cooperative occupancy forecasting.

To address these issues, we present UniOcc, a comprehensive, open-source benchmark unifying 2D/3D occupancy labels, per-voxel flow annotations, and multi-agent support across multiple real-world and synthetic datasets. We hope that UniOcc will catalyze occupancy-centric research, streamlining development, benchmarking, and fostering innovations in autonomous driving. The summary of our contributions is as follows:

- We introduce UniOcc, the first-of-its-kind unified 2D/3D occupancy forecasting and prediction benchmark with flow information for both conventional and cooperative driving by unifying real data from nuScenes and Waymo and synthetic data from CARLA and OpenCOOD.
- We develop a user-friendly platform for current-frame oc-

cupancy prediction and multi-frame occupancy forecasting, which enables easy setup, cross-dataset augmentation, and comprehensive occupancy evaluation with or without reference to ground-truth labels.

- We provide the Python toolkit for occupancy grid processing: localization, detection, tracking, alignment, and visualization. (See supplementary and code for details.)
- We validate our dataset-agnostic training/testing pipeline and the proposed evaluation metrics on state-of-the-art occupancy forecasting/prediction models. Our experiments show that (1) incorporating voxel-level flow yields performance gains in occupancy forecasting and (2) existing methods face challenges in cross-domain generalization, highlighting avenues for future research.

## 2. Related Work

### 2.1. Occupancy Datasets

The nuScenes [3] and Waymo [30] datasets are widely used autonomous driving datasets collected from real-world driving, which provide raw sensor data (*i.e.*, camera and LiDAR) with 3D annotations. Nevertheless, they do not provide 3D occupancy labels. As a result, existing studies often rely on automatic label generation methods introduced in Occ3D [31], SurroundOcc [42], or OpenOccupancy [36]. On the other hand, CarlaSC [43] and CoHFF [28] provide synthetic datasets collected with the CARLA simulator [5], where the ground truth occupancy can be easily obtained. A detailed comparison between existing datasets and UniOcc (ours) is shown in Table 1.

### 2.2. Occupancy Prediction

Recent studies in 3D perception have explored using only camera inputs to produce dense 2D or 3D occupancy grid maps with semantic labels. Early methods often employ a single frame of monocular or multi-camera images to estimate the 2D occupancy [11, 15, 16] and 3D occupancy [12, 29, 31, 32, 36, 42] at the current frame. Despite these advances, single-frame methods are limited by their ability to estimate depth. Recently, researchers have turned to historical camera frames for more robust geometric cues, allowing better handling of occlusion and complex scene dynamics. CVT-Occ [51], for example, refines current-frame occupancy with a *temporal cost volume* constructed from past images, thereby leveraging multi-view images across time for improved depth estimation. In the cooperative autonomous driving domain, CoHFF [28] explores cooperative prediction from a multi-connected vehicle (CAV) setting by having each CAV share its perception information.

### 2.3. Occupancy Forecasting

Beyond static occupancy reconstruction, a growing line of work tackles *temporal* occupancy prediction, inferring

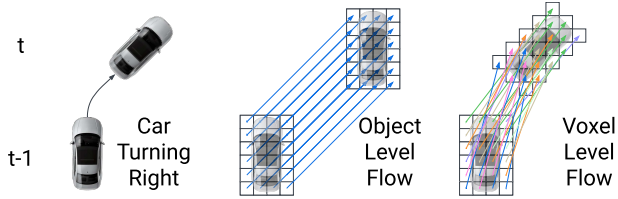


Figure 2. Comparison of object-level flow and voxel-level flow.

how 3D grids evolve over time. Several recent methods predict future occupancy either from historical occupancy grids [2, 9, 24, 35, 41, 55], often conditioned on ego trajectories or high-level navigation intents. By capturing future scene states, these models facilitate proactive planning and safer driving in dynamic environments.

## 2.4. Occupancy Flow

Early works on flow-driven occupancy forecasting primarily focused on 2D grids, using object bounding boxes and map information to predict future occupancy and flow [20, 21, 26]. While LetOccFlow [23] extends flow to 3D, it only considers horizontal directions and thus cannot capture rich object rotations. CarlaSC [43] provides per-object flow by assigning each voxel the object’s velocity, but this approach similarly neglects rotational motion. In contrast, our method annotates each voxel with its *unique* 3D displacement for the next time step, thereby preserving full rotation and translation. We show their difference in Figure 2.

Most prior approaches adopt a *forward flow* convention (a vector pointing from the current voxel to its next location). Liu *et al.* [21] introduce an alternative *reverse flow* that points backward in time to simplify multi-future training. To accommodate both conventions, we provide both forward and reverse flows in our unified dataset, enabling versatile modeling of complex, fully 3D motion dynamics.

## 3. UniOcc Framework

### 3.1. Unified Data Format and Features

Our benchmark supports a wide range of occupancy-centric tasks, including occupancy forecasting, single-frame occupancy prediction, and flow estimation. Our framework defines the following task-agnostic data formats:

**Semantic Occupancy Label.** We represent the scene as a 3D voxel grid  $G \in \{0, \dots, C\}^{L \times W \times H}$ , where  $C$  denotes the number of classes (see Table 10 in supplementary materials), and  $L, W, H$  are the grid’s dimensions along the ego vehicle’s heading, lateral, and vertical axes, respectively. This grid is centered on the ego vehicle, with the  $+x$ -axis aligned to the direction of travel,  $+y$ -axis to the left, and  $+z$ -axis upward. For certain 2D tasks (*e.g.*, motion planning), we collapse the height dimension via a priority scheme (*e.g.*,  $Pedestrian > Car > Road$ ), such that each vertical pillar adopts the label of its highest-priority

voxel. This approach prevents occlusion of essential object classes (like pedestrians) by lower-priority labels in the same grid column, ensuring meaningful representation for downstream tasks.

**Camera Images.** We store raw RGB images in a 4D tensor  $I \in \{0, \dots, 255\}^{K_{\text{cam}} \times \text{Img}_x \times \text{Img}_y \times 3}$ , where  $K_{\text{cam}}$  denotes the number of onboard cameras and each image has resolution  $\text{Img}_x \times \text{Img}_y$ .

**Camera Field-of-View (FOV) Mask.** A binary 3D tensor  $U \in \{0, 1\}^{L \times W \times H}$  indicates which voxels lie within each camera’s observable frustum ( $U = 1$  for visible voxels and  $U = 0$  otherwise). This mask is crucial for camera-based occupancy methods that require explicit delineation of occluded regions or unobserved space.

**Camera Intrinsic and Extrinsic.** We represent camera intrinsics as  $\text{Int} \in \mathbb{R}^{K_{\text{cam}} \times 3}$ , while extrinsic transformations (from each camera to the ego frame) are given by  $\text{Ext} \in \text{SE}(3)^{K_{\text{cam}}}$ , where  $\text{SE}(3)$  denotes the group of 3D homogeneous transformation. These parameters unify the projection from 3D ego coordinates onto 2D image planes.

**Ego-to-World Transformation.** A homogeneous transformation matrix  $T_e^w \in \text{SE}(3)$  denotes the pose of the ego vehicle in a global world frame, enabling precise alignment of data from multiple sensors and coordinate systems.

**Forward Occupancy Flow.** We define a 4D tensor  $F \in \mathbb{R}^{L \times W \times H \times 3}$  that records per-voxel forward-motion vectors. Unlike prior methods [43] that assign a single velocity to all voxels of an object (thus missing object rotation), our approach computes individual voxel flows capturing both translation and rotation. We separately compute the flow for dynamic foreground objects (*e.g.*, *Car*, *Pedestrian*) and static background environment (*e.g.*, *Road*, *Vegetation*) and merge dynamic and static flows into  $F$ . As illustrated in Figure 3, this voxel-level flow captures full 3D motion, including rotation. The details of our computation algorithm can be found in the supplementary materials and code.

**Backward Occupancy Flow.** Analogous to the forward flow, we define a 4D tensor  $B \in \mathbb{R}^{L \times W \times H \times 3}$  to capture *backward* motion vectors. Instead of computing each voxel’s displacement from  $t$  to  $t+1$ , we evaluate the motion from  $t$  to  $t-1$ . This backward flow is particularly useful for models that benefit from reverse-time supervision or multi-future training strategies [21].

**Object Annotations.** We also provide object-level annotations as a list of dictionaries, each containing: ① **Agent-to-Ego Transformation.** A transformation matrix  $T_a^e \in \text{SE}(3)$  that maps the agent’s local coordinate system into the ego frame. This captures both the agent’s position and orientation relative to the ego vehicle. ② **Size.** A 3D vector  $d \in \mathbb{R}^3$  describing the bounding box dimensions of the agent (*length*, *width*, *height*). ③ **Category.** The object’s semantic class label, following the definitions in Table 10.

Table 1. Comparison of popular occupancy datasets. **Length** is the total time this dataset covers. **Scenarios** are the number of scenarios, usually a proxy for data diversity. **Voxel Range** is the range of the occupancy grid. **Resolution** is the per-voxel size. **Flow** is whether this dataset provides occupancy flow. **Obj Categories** are the number of category labels provided in the dataset.

Dataset	Data Source	Length	Scenarios	Sampling Rate	Voxel Range (m)	Resolution (m)	Flow	Obj Categories
Occ3D nuScenes [31]	nuScenes	9.5 hrs	1110	2 Hz	$[\pm 40, \pm 40, -1 \sim 5.4]$	0.2 / 0.4	-	17
Occ3D Waymo [31]	Waymo	4.0 hrs	998	10 Hz	$[\pm 40, \pm 40, -1 \sim 5.4]$	0.2 / 0.4	-	15
SurroundOcc [42]	nuScenes	9.5 hrs	1110	2 Hz	$[\pm 40, \pm 40, -1 \sim 5.4]$	0.5	-	17
OpenOccupancy [36]	nuScenes	9.5 hrs	1110	2 Hz	$[\pm 51.2, \pm 51.2, -5 \sim 3]$	0.1	-	17
CoHFF [28]	OpenCOOD	0.69 hrs	44	10 Hz	$[\pm 51.2, \pm 51.2, -5 \sim 3]$	1.0	-	10
UniOcc (Ours)	nuScenes, Waymo CARLA, OpenCOOD	14.2 hrs	2152	2 Hz / 10 Hz	$[\pm 40, \pm 40, -1 \sim 5.4]$	0.2 / 0.4	Voxel Level	10, 15, 17

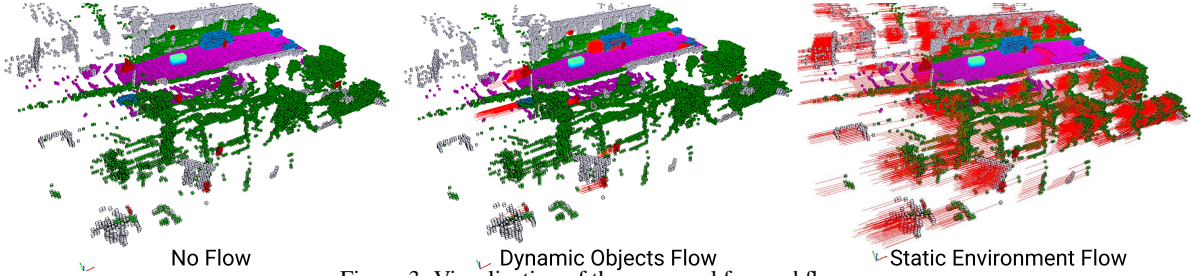


Figure 3. Visualization of the per-voxel forward flows.

### 3.2. Task Categories

Our unified inputs enable a broad range of occupancy-centric tasks, spanning both static prediction and dynamic forecasting. By employing a unified representation across multiple domains, we simplify cross-dataset training and allow fair comparisons of methods that tackle different sub-problems. Below, we outline three representative tasks:

**Occupancy Prediction.** Here, the model consumes the past  $W_{\text{obs}}$  camera frames  $\{I^{t-W_{\text{obs}}}, \dots, t\}$ , together with their FOV masks  $\{U^{t-W_{\text{obs}}}, \dots, t\}$  and camera parameters (*intrinsics* Int, *extrinsics* Ext). The output is the current 3D occupancy grid  $G^t$ , which captures the scene at time  $t$ .

**Occupancy Forecasting with Optional Flow.** In the forecasting setting, the input is the historical data over  $W_{\text{obs}}$  frames—either voxel grids  $\{G^{t-W_{\text{obs}}}, \dots, t\}$  or camera images  $\{I^{t-W_{\text{obs}}}, \dots, t\}$ . The model predicts future occupancies  $\{G^{t, \dots, t+W_{\text{fut}}}\}$ , optionally conditioned on fine-grained ego trajectories  $T_e^{w, t: t+W_{\text{fut}}}$  or high-level driving intentions (e.g., *Turn Right*). For certain use cases, forecasting methods may also produce the future flow  $F^{t: t+W_{\text{fut}}}$  or future ego movement  $T_e^{w, t: t+W_{\text{fut}}}$ . As discussed in Section 4.2, this joint occupancy-and-flow forecasting scheme can help capture complex motion patterns over time.

**Cooperative Occupancy Prediction and Forecasting with Optional Flow.** Under cooperative settings, multiple connected vehicles (CAVs) collaborate by sharing either image or occupancy data. From the ego vehicle’s perspective, it receives the shared historical observations  $\{I_{\text{CAV}}^{t-W_{\text{obs}}}, \dots, t\}$  or  $\{G_{\text{CAV}}^{t-W_{\text{obs}}}, \dots, t\}$  alongside transformations mapping CAV frames to the ego frame. The output remains the same (i.e., single-ego occupancy or forecast), but the increased viewpoint coverage can mitigate occlusions and improve overall scene understanding.

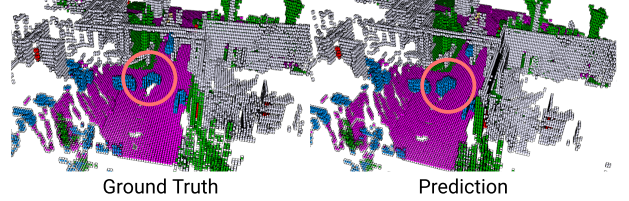


Figure 4. Illustration of imperfect ground-truth labels. Left: partial car shape from Occ3D [31]. Right: a more complete shape predicted by OccWorld [55]. Standard IoU may penalize the model for producing a fuller shape, despite it being more realistic.

### 3.3. Unified Datasets

We build our unified dataset from the following sources:

- **nuScenes [3] and Waymo [30].** Both datasets provide camera images, LiDAR sweeps, and object-level annotations. As neither directly includes 3D occupancy labels, we synthesize occupancy ground truth via three pseudo-labeling pipelines: Occ3D [31], OpenOccupancy [36], and SurroundOcc [42]. This multi-tool approach increases robustness and diversity in labeled outputs.
- **CARLA [5].** We use CARLA’s simulation engine to generate an unlimited variety of virtual driving scenes, from which we can extract “perfect” 3D occupancy labels (meshes, object states, etc.). These realistic yet controllable scenarios are publicly released, enabling straight-forward large-scale training. Our framework offers the option to generate an arbitrary length of data.
- **OpenCOOD [49].** Built on CARLA, OpenCOOD offers multi-vehicle cooperation scenarios. We extend its data-generation scripts to export 3D occupancy from simulated meshes, thus expanding our dataset with collaborative driving examples.

### 3.4. Unified Occupancy Processing Toolkit

Most occupancy-based approaches focus solely on generating an occupancy grid but provide limited tools for downstream processing or motion analysis. To address this gap, our framework includes a toolkit for object segmentation and tracking directly within the voxel space, enabling more advanced tasks such as shape analysis or motion planning. (Some details, *e.g.*, localization, tracking, are deferred to the supplementary and the code).

#### 3.4.1. Object Identification

Given an occupancy grid  $G \in \{0, \dots, C\}^{L \times W \times H}$ , we identify and segment relevant objects in the following steps, which are shown in Figure 5.

1. **Object Segmentation.** We extract voxels by category (*e.g.*, *Car*, *Pedestrian*), then run 6-connected component labeling (CCL) implemented via Breadth-First Search:

$$L = \text{CCL}(G), \quad t \in \{0, 1, \dots, T\}, \quad (1)$$

where  $L \in \{1, \dots, N\}^{L \times W \times H}$  assigns each connected component a unique object ID, and  $N$  is the total number of objects.

2. **Voxel Extraction.** For each object ID  $n$ , we gather its voxel coordinates  $V_n$ :

$$V_n = \{ \langle x, y, z \rangle \mid L(x, y, z) = n \} \quad (2)$$

3. **Horizontal Axis Bounding Box.** Voxel predictions can be partial (see Fig. 4), making direct bounding-box measurement (length, width, height) unreliable. We therefore fit a bounding rectangle in the horizontal plane using a rotating-calipers method [33], which is  $O(n^2)$  in the number of object voxels, under the assumption that each object moves parallel to the ground. This yields a 2D minimum bounding rectangle, from which we recover heading and planar extents.
4. **Dimension Extraction.** We take the rectangle’s length and width as the object’s planar dimensions, then compute height from the vertical extent of the voxels. All dimensions are scaled by the voxel resolution  $\epsilon$  to convert to metric units.

#### 3.4.2. Object Tracking

Leveraging the forward occupancy flow introduced in Section 3.1 predicted for each voxel, we also provide a simple occupancy-based object tracking algorithm:

1. **Object Voxel Extraction.** For each identified object in the occupancy grid at frame  $t$ , we retrieve its voxel coordinates  $V_n^t$  (Eq. 2) and corresponding flow vectors  $F_n^t \in \mathbb{R}^{n \times 3}$ .
2. **Step Prediction.** We estimate the next-frame voxel positions  $\widetilde{V}_n^{t+1}$  by adding the flow:

$$\widetilde{V}_n^{t+1} = V_n^t + F_n^t. \quad (3)$$

3. **Centroid Extraction.** Let  $\widetilde{c}_n^{t+1}$  be the centroid of the predicted voxel set  $\widetilde{V}_n^{t+1}$ . We also compute the true object voxels at frame  $t + 1$ ,  $V_n^{t+1}$ , and its centroid  $c_n^{t+1}$ :

$$\widetilde{c}_n^{t+1} = \frac{1}{|\widetilde{V}_n^{t+1}|} \sum_{(i,j,k) \in \widetilde{V}_n^{t+1}} \langle i, j, k \rangle, \quad (4)$$

$$c_n^{t+1} = \frac{1}{|V_n^{t+1}|} \sum_{(i,j,k) \in V_n^{t+1}} \langle i, j, k \rangle. \quad (5)$$

4. **Bipartite Association.** We match predicted centroids  $\{\widetilde{c}_p^t\}$  with observed centroids  $\{c_q^{t+1}\}$  using the Hungarian algorithm to minimize pairwise distances:

$$P^* = \underset{P}{\operatorname{argmin}} \sum_{p,q} \|\widetilde{c}_p^t - c_q^{t+1}\|_2 P_{pq}, \quad P \in \{0, 1\}^{n_s^t \times n_s^{t+1}}. \quad (6)$$

where  $P^*$  is the matching matrix and  $n_s^t, n_s^{t+1}$  represents the number of objects in the consecutive frames. We assign the same ID for matched objects between frames and assign new IDs for newly appeared objects. We discard objects where the matched distance is greater than a threshold of  $\gamma$ , which we empirically choose as 0.5 m for cars and cyclists, 0.2 m for pedestrians.

The above process yields cross-frame associations that unify object identities over time, enabling motion interpretation and analysis directly in the voxel space.

#### 3.4.3. Object Alignment

Finally, we align the voxel sets of tracked objects for shape analysis or consistency checks:

1. **Translation Alignment.** We translate each object’s voxel coordinates to center them at the origin as  $\bar{V}_n^t$ :

$$\bar{V}_n^t = V_n^t - \frac{1}{|V_n^t|} \sum_{\mathbf{v} \in V_n^t} \mathbf{v}. \quad (7)$$

2. **Rotation Alignment.** We apply Principal Component Analysis (PCA) to each frame’s voxel set to resolve a canonical orientation. For consistency, we adjust the sign of the new principal axes to align with the previous frame’s orientation (see supplementary materials). The final rotated voxel coordinates are denoted as  $\hat{V}_n^t$ .

With these steps, we facilitate object-centric analyses (*e.g.*, measuring shape changes or rotation consistency) entirely in the occupancy grid domain without the need for a reference ground truth label or annotation.

### 3.5. Unified Evaluation Metrics

Our benchmark includes multiple metrics for assessing the quality of generated or predicted occupancy grids. Section 3.5.1 describes the widely adopted *voxel-based metrics*, while Section 3.5.2 proposes *ground-truth-free* methods that address two major issues: the imperfect nature of

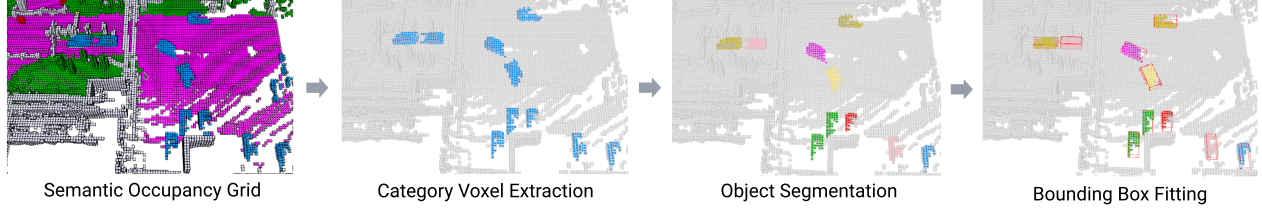


Figure 5. An example showing our pipeline for voxel extraction, connected-component segmentation, and bounding-box fitting.

real-world labels (Fig. 4) and the inherent multi-modality of some forecasting tasks (where only a single future is recorded but we expect model to produce multiple futures).

### 3.5.1. Voxel-Based Evaluation

Following prior occupancy prediction [51] and forecasting [2, 35, 41, 55] studies, we employ two standard metrics: *geometric IoU* (or simply  $\text{IoU}_{\text{geo}}$ ) and *mIoU* (mean intersection over union across semantic classes). Concretely, for a predicted occupancy grid  $G_{\text{pred}}$  and ground-truth grid  $G_{\text{gt}}$ ,

$$\text{IoU}_{\text{geo}} = \frac{|G_{\text{pred}} \cap G_{\text{gt}}|}{|G_{\text{pred}} \cup G_{\text{gt}}|}, \quad (8)$$

where  $|G_{\text{pred}} \cap G_{\text{gt}}|$  is the number of voxels occupied in both the prediction and ground truth, and  $|G_{\text{pred}} \cup G_{\text{gt}}|$  is the total occupied voxels in either grid. For multi-class occupancy ( $C$  total classes), the *mIoU* is computed via:

$$\text{mIoU}_{\text{geo}} = \frac{1}{C} \sum_{c=1}^C \text{IoU}_{\text{geo},c}, \quad (9)$$

where  $\text{IoU}_{\text{geo},c}$  is computed from Eq. (8) with restricting the voxels to class  $c$ .

While these voxel-based metrics are straightforward, they can penalize predictions that exceed the pseudo-ground truth (Fig. 4). Additionally, tasks like multi-modal forecasting may produce many plausible futures not captured by a single reference label. For these reasons, we propose evaluation strategies that do *not* require perfect ground truth.

### 3.5.2. Ground-Truth-Free Evaluation

Complementary to label-dependent IoU, we propose metrics that assess geometric plausibility without referencing a single ground-truth scene. These metrics are particularly useful for multi-modal generation or cases where ground-truth labels are incomplete.

**Key Object Dimension Probability.** Given a predicted object’s bounding box  $\langle l, w, h \rangle$  for a category  $c$ , we evaluate its plausibility by computing a *Gaussian Mixture Model* (GMM) likelihood. Specifically, each category  $c$  has a pre-trained GMM, denoted  $\text{GMM}_c$ , learned from real or synthetic data in our unified dataset (Appendix 7.2). At inference time, we query  $\text{GMM}_c$  with the object’s dimensions:

$$P_n = \max_k p(\langle l, w, h \rangle | \text{GMM}_c^k), \quad (10)$$

This probability  $P_n$  gives a heuristic for whether the object has realistic dimensions for its reported category. We use an

empirical value  $\rho = 0.5$  as the threshold to determine if the shape is likely real or not.

**Temporal Foreground Object Shape Consistency.** For dynamic objects forecasted across multiple frames, we measure shape consistency by aligning each object’s voxels over time (see Section 3.4.3) and computing a voxel-wise intersection over union:

$$\text{IoU}_{\text{object}} = \frac{|\hat{V}^t \cap \hat{V}^{t+1}|}{|\hat{V}^t \cup \hat{V}^{t+1}|}. \quad (11)$$

A high  $\text{IoU}_{\text{object}}$  implies stable shape geometry from frame  $t$  to  $t+1$ . We then average these IoUs within each category to assess overall consistency across time.

**Temporal Background Environment Consistency.** For static background regions, we expect persistent occupancy between consecutive frames within the overlapping field of view. Let  $V_e^t$  be the environment voxels at time  $t$ , and  $\tilde{V}_e^{t+1}$  be their projected coordinates at  $t+1$  (using known ego-motion, see Section 3.2). We discard out-of-bound voxels and compute the binary IoU of the overlap:

$$\text{IoU}_{\text{bg}} = \frac{|\tilde{V}_e^{t+1} \cap V_e^{t+1}|}{|\tilde{V}_e^{t+1} \cup V_e^{t+1}|}. \quad (12)$$

Higher  $\text{IoU}_{\text{bg}}$  indicates a consistent static background across frames, even without a perfect ground truth label.

Overall, these ground-truth-free metrics complement standard IoU by providing deeper insights into scene realism and temporal coherence, especially valuable for generative or multi-modal occupancy tasks.

## 4. Experiments

### 4.1. Experimental Settings

In all experiments, we use a voxel size of  $[200, 200, 16]$  and grid resolution of 0.4 m. For the occupancy forecasting task, the model takes in the 3 s historical occupancy and forecasts the future 3 s. For the occupancy prediction task, the models take in the 3 s historical camera images and predict the occupancy at the current frame. For nuScenes and Waymo data sources, we leverage the pseudo occupancy labels from Occ3D [31]. For the CARLA data source, we generate simulated driving data in 16 diverse scenarios.

### 4.2. Occupancy Forecasting with Flows

To investigate the impact of explicitly modeling flow in occupancy forecasting, we augment OccWorld [55] to consume both the dynamic and static voxel flows provided by

Table 2. Occupancy forecasting performance of OccWorld [55] on nuScenes with different types of flow. (See Supplementary for Waymo.)

Train and Test Source	Flow Type	mIoU <sub>geo</sub> ↑				IoU <sub>geo</sub> ↑				IoU <sub>bg</sub> ↑	IoU <sub>car</sub> ↑	P <sub>car</sub> ↑
		0s	1s	2s	3s	0s	1s	2s	3s			
nuScenes	None	66.79	30.23	21.67	18.13	60.66	33.33	24.93	20.67	53.27	78.39	77.83
nuScenes	Object	65.41	31.09	20.97	18.40	60.72	33.28	24.57	20.55	54.15	77.36	76.25
nuScenes	Voxel	<b>70.64</b>	<b>32.13</b>	<b>22.50</b>	<b>19.06</b>	<b>62.62</b>	<b>35.93</b>	<b>26.03</b>	<b>21.04</b>	<b>59.56</b>	<b>81.50</b>	<b>82.57</b>

Table 3. Occupancy forecasting performance of OccWorld [55] on various train/test data source combinations.

Train Sources	Test Source	mIoU <sub>geo</sub> ↑				IoU <sub>geo</sub> ↑				IoU <sub>bg</sub> ↑	IoU <sub>car</sub> ↑	P <sub>car</sub> ↑
		0s	1s	2s	3s	0s	1s	2s	3s			
nuScenes	nuScenes	70.64	32.13	22.50	19.06	62.62	35.93	26.03	21.04	59.56	81.50	82.57
Waymo	nuScenes	63.22	23.47	18.11	15.80	60.42	27.35	20.86	17.63	49.90	79.41	72.54
CARLA	nuScenes	29.93	11.94	10.85	10.54	49.32	13.51	11.24	10.82	22.39	59.64	78.99
nuScenes	Waymo	64.37	31.08	23.48	20.90	65.38	39.38	30.94	27.40	61.25	81.34	72.69
Waymo	Waymo	71.35	32.04	25.77	23.76	72.69	36.04	30.48	27.96	58.26	89.30	86.68
CARLA	Waymo	30.64	12.09	11.13	10.79	55.99	16.16	13.84	13.07	23.05	57.05	77.18
nuScenes	CARLA	79.62	49.70	49.25	48.72	68.93	17.62	16.45	15.50	86.74	97.52	71.02
Waymo	CARLA	80.32	48.54	48.06	47.60	71.38	15.72	14.17	12.99	86.79	91.37	80.38
CARLA	CARLA	79.66	48.87	47.28	46.69	69.67	20.05	15.34	12.78	24.34	59.39	80.92
nuScenes + Waymo	nuScenes	71.80	32.57	22.87	20.15	63.12	36.24	27.93	21.48	60.23	86.08	83.14
nuScenes + Waymo	Waymo	71.23	33.42	26.52	24.99	73.13	36.30	31.61	28.22	59.23	88.72	87.01
nuScenes + CARLA	nuScenes	71.47	31.70	22.69	18.11	62.94	35.83	28.29	21.03	55.07	77.87	82.97
Waymo + CARLA	nuScenes	64.99	24.88	19.65	15.55	61.33	27.50	20.38	18.94	49.20	81.13	82.50
nuScenes + CARLA	Waymo	65.71	31.48	23.84	21.25	66.95	40.37	31.47	27.50	57.34	83.80	81.98
Waymo + CARLA	Waymo	71.66	37.05	29.50	26.16	72.94	42.54	35.46	31.52	56.32	86.67	81.29
nuScenes + CARLA	CARLA	84.88	49.25	48.69	47.88	74.15	17.02	15.81	14.41	86.54	98.48	81.91
Waymo + CARLA	CARLA	83.81	54.31	52.94	52.13	74.34	27.42	24.60	23.04	73.89	90.47	81.48
nuScenes + Waymo + CARLA	nuScenes	72.53	33.98	22.76	20.18	63.32	36.31	27.83	21.89	57.51	80.51	83.01
nuScenes + Waymo + CARLA	Waymo	74.49	34.32	28.28	24.61	73.58	43.42	32.46	27.44	62.20	87.54	80.93
nuScenes + Waymo + CARLA	CARLA	85.26	55.19	52.58	50.96	74.63	28.33	22.31	19.35	74.15	88.61	82.35

our unified dataset. Specifically, we introduce an additional flow encoder that processes the per-voxel flow, followed by cross-attention [34] to fuse the encoded flow features with the scene tokens in the Spatial-Temporal Generative Transformer of OccWorld. We further append a flow decoder to predict next-step voxel flows, supervised via an  $L_2$  loss against our ground-truth flow annotations.

As shown in Table 2 and 8, flow consistently improves forecasting performance and enhances temporal consistency for both nuScenes and Waymo datasets. In particular, we observe larger gains in the category-averaged mIoU, implying that using flow information helps the network better capture object-level motion, thereby improving predictions for dynamic classes (e.g., moving vehicles and pedestrians).

### 4.3. Cross Data Source Training and Evaluation for Occupancy Forecasting

A key advantage of our unified dataset is the ability to train and evaluate models across multiple data sources, thereby measuring out-of-distribution (OOD) performance. We use our flow-augmented OccWorld [55] to illustrate this cross-domain generalization, as it can be trained on large datasets with minimal computational overhead. Our results are shown in Table 3. The key insights are summarized below.

**Diverse Data Improves OOD Generalization in Real-World Driving.** Models trained on a single data source

(e.g., only nuScenes) tend to perform well on in-domain test data but exhibit weaker transfer to unseen domains (e.g., Waymo). In contrast, the model trained on our unified occupancy dataset (combining nuScenes and Waymo) consistently achieves higher mIoU and IoU scores over a range of prediction horizons. This implies the importance of multi-domain coverage: exposure to a broader set of scenes and motion patterns reduces the severity of domain shift and improves OOD performance. Consequently, our unified dataset marks a substantial step toward more robust occupancy forecasting in real-world driving.

**Diverse Data Improves Long Term Accuracy.** As expected, forecast accuracy degrades with increased time horizons (1s-3s), highlighting the challenge of long-horizon occupancy prediction. Yet, this degradation is consistently less severe for the models that are trained from multi-domain data (nuScenes and Waymo), which indicates that diverse training data helps improve forecasting accuracy over time.

**Simulation Data Enhances Object Shape Learning.** As is shown in Table 3, incorporating CARLA data alongside real-world datasets increases the likelihood of accurate object predictions (noted by higher  $P_{car}$ ), especially in scenarios where object shapes are imperfectly captured by LiDAR or pseudo-labeling in nuScenes and Waymo datasets (see Figure 4). The simulation data, by providing “perfect”

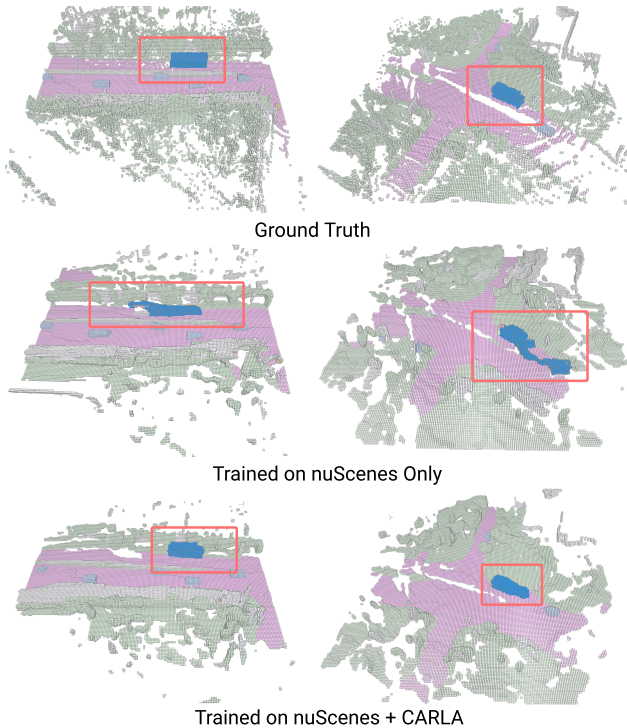


Figure 6. Comparisons of the predicted car shapes between the models trained on nuScenes-only and nuScenes + CARLA data sources. In both cases, we observe a vehicle with a strange shape in the nuScenes-only forecasting. In contrast, the model trained on both data sources generates more reasonable predictions.

shapes for both static and dynamic elements, enables better learning of the geometry of objects. Figure 6 provides a qualitative illustration of this improvement.

#### Sim-Real Compatibility Remains an Open Challenge.

As is shown in Table 3, incorporating CARLA data alongside real-world datasets increases the likelihood of accurate object predictions (noted by higher  $\mathbf{P}_{\text{car}}$ ), especially in scenarios where object shapes are imperfectly captured by LiDAR or pseudo-labeling in nuScenes and Waymo datasets (see Figure 4). The simulation data, by providing “perfect” shapes for both static and dynamic elements, enables better learning of the geometry of objects. Figure 6 provides a qualitative illustration of this improvement. However, we also note that a naive mix of simulation data and real driving degrades temporal consistency (after adding CARLA to nuScenes,  $\text{IoU}_{\text{bg}}$  degrades by 7.5%). We found that one potential cause is the mismatch in speed range between CARLA and real-world datasets, since the former tends to be slower on average. We explore this perspective further in Section 8 in the supplementary materials.

### 4.4. Occupancy Prediction

Although our UniOcc framework primarily targets 3D occupancy forecasting from unified voxel grids, it also facil-

Table 4. Occupancy prediction performance of Cam4DOcc [24] and CVTOcc [51] on the nuScenes data source.

Model	mIoU <sub>geo</sub> ↑	IoU <sub>geo</sub> ↑	IoU <sub>bg</sub> ↑	IoU <sub>car</sub> ↑	P <sub>car</sub> ↑
CVTOcc [51]	31.57	81.20	48.93	80.60	74.91
Cam4DOcc [24]	13.59	13.33	52.46	56.13	73.28

Table 5. Cooperative occupancy prediction performance of CoHFF [28] on the OpenCOOD [49] data source.

Model	mIoU <sub>geo</sub> ↑	IoU <sub>geo</sub> ↑	IoU <sub>bg</sub> ↑	IoU <sub>car</sub> ↑	P <sub>car</sub> ↑
CoHFF [28]	34.16	50.46	51.90	87.22	66.19

itates camera-based models, provided that domain-specific calibration and pre-processing pipelines are carefully integrated. We incorporate two open-source camera-based occupancy prediction approaches, Cam4DOcc [24] and CVTOcc [51], by aligning their input requirements and output grids with our unified evaluation protocol. Notably, we evaluate each model only on the domain where its official weights were originally trained (*e.g.*, Cam4DOcc on nuScenes), which enables a fair comparison of methods on a consistent voxel labeling and metric setup. Table 4 illustrates that CVTOcc achieves notably higher object-centric IoU and geometry-aware mIoU than Cam4DOcc due to its more flexible cost-volume fusion mechanism.

### 4.5. Cooperative Occupancy Prediction

While most existing occupancy methods focus on single-ego perception, multi-vehicle collaboration offers a promising avenue for enhanced scene understanding. To highlight this, we integrate and evaluate CoHFF [28], a cooperative occupancy prediction approach, within our framework. By sharing sensor observations and intermediate features across multiple agents, CoHFF mitigates occlusions and extends coverage in complex driving scenarios. Table 5 reports the performance of CoHFF on the OpenCOOD [49] data source, showing that multi-agent fusion yields reasonably high IoU for car instances (87.22) and background occupancy (51.90). These results demonstrate the potential benefits of cooperative perception and underscore our framework’s flexibility in accommodating multi-agent settings with standardized occupancy representations.

## 5. Conclusion

We present UniOcc, a unified benchmark for occupancy forecasting and prediction in autonomous driving. By integrating diverse real-world and synthetic data sources, our approach enables cross-dataset training and evaluation on occupancy tasks ranging from single-ego to cooperative multi-vehicle settings. Beyond occupancy grid labels, we provide comprehensive occupancy flow annotations (forward and backward), voxel-based segmentation and tracking tools, and ground-truth-free evaluation metrics. We release our benchmark to foster new opportunities in the exploration of occupancy-based autonomous driving.

## References

- [1] Ben Agro, Quinlan Sykora, Sergio Casas, Thomas Gilles, and Raquel Urtasun. Uno: Unsupervised occupancy fields for perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14487–14496, 2024. 1
- [2] Hengwei Bian, Lingdong Kong, Haozhe Xie, Liang Pan, Yu Qiao, and Ziwei Liu. Dynamiccity: Large-scale occupancy generation from dynamic scenes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. 1, 3, 6
- [3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 1, 2, 4
- [4] Xiaoyu Deng, Zhengjian Kang, Xintao Li, Yongzhe Zhang, and Tianmin Guo. Covis: A collaborative framework for fine-grained graphic visual understanding. *arXiv preprint arXiv:2411.18764*, 2024. 1
- [5] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2, 4
- [6] Lan Feng, Mohammadhossein Bahari, Kaouther Mersaoud Ben Amor, Éloi Zablocki, Matthieu Cord, and Alexandre Alahi. Unitraj: A unified framework for scalable vehicle trajectory prediction. In *European Conference on Computer Vision*, pages 106–123. Springer, 2025. 2
- [7] Xiangbo Gao, Yuheng Wu, Xuewen Luo, Keshu Wu, Xinghao Chen, Yuping Wang, Chenxi Liu, Yang Zhou, and Zhengzhong Tu. Airv2x: Unified air-ground vehicle-to-everything collaboration. *arXiv preprint arXiv:2506.19283*, 2025. 1
- [8] Xiangbo Gao, Yuheng Wu, Rujia Wang, Chenxi Liu, Yang Zhou, and Zhengzhong Tu. Langcoop: Collaborative driving with language. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4226–4237, 2025. 1
- [9] Songen Gu, Wei Yin, Bu Jin, Xiaoyang Guo, Junming Wang, Haodong Li, Qian Zhang, and Xiaoxiao Long. Dome: Taming diffusion model into high-fidelity controllable occupancy world model. *arXiv preprint arXiv:2410.10429*, 2024. 3
- [10] Congrui Hetang and Yuping Wang. Novel view synthesis from a single rgbd image for indoor scenes. In *2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, pages 447–450. IEEE, 2023. 1
- [11] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2
- [12] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023. 2
- [13] Zhengjian Kang, Ye Zhang, Xiaoyu Deng, Xintao Li, and Yongzhe Zhang. Lp-detr: Layer-wise progressive relation for object detection. In *International Conference on Intelligent Computing*, pages 144–156. Springer, 2025. 1
- [14] Peiran Li, Xinkai Zou, Zhuohang Wu, Ruifeng Li, Shuo Xing, Hanwen Zheng, Zhikai Hu, Yuping Wang, Haoxi Li, Qin Yuan, et al. Safeflow: A principled protocol for trustworthy and transactional autonomous agent systems. *arXiv preprint arXiv:2506.07564*, 2025. 1
- [15] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1477–1485, 2023. 2
- [16] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 2
- [17] Fangzhou Lin, Yun Yue, Songlin Hou, Xuechu Yu, Yajun Xu, Kazunori D Yamada, and Ziming Zhang. Hyperbolic chamfer distance for point cloud completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14595–14606, 2023. 1
- [18] Fangzhou Lin, Yun Yue, Ziming Zhang, Songlin Hou, Kazunori Yamada, Vijaya Kolachalama, and Venkatesh Saligrama. Infocd: A contrastive chamfer distance loss for point cloud completion. *Advances in Neural Information Processing Systems*, 36:76960–76973, 2023.
- [19] Fangzhou Lin, Zilin Dai, Rigved Sanku, Songlin Hou, Kazunori D Yamada, Haichong K Zhang, and Ziming Zhang. A strong view-free baseline approach for single-view image guided point cloud completion. *arXiv preprint arXiv:2506.15747*, 2025. 1
- [20] Haochen Liu, Zhiyu Huang, and Chen Lv. Strajnet: Occupancy flow prediction via multi-modal swin transformer. *arXiv preprint*, 2022. 3
- [21] Haochen Liu, Zhiyu Huang, and Chen Lv. Multi-modal hierarchical transformer for occupancy flow field prediction in autonomous driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1449–1455. IEEE, 2023. 3
- [22] Xu Liu, Tong Zhou, Chong Wang, Yuping Wang, Yuanxin Wang, Qinjingwen Cao, Weizhi Du, Yonghuan Yang, Junjun He, Yu Qiao, et al. Toward the unification of generative and discriminative visual foundation model: A survey. *The Visual Computer*, pages 1–42, 2024. 1
- [23] Yili Liu, Linzhan Mou, Xuan Yu, Chenrui Han, Sitong Mao, Rong Xiong, and Yue Wang. Let occ flow: Self-supervised 3d occupancy flow prediction. In *Proceedings of The 8th Conference on Robot Learning*, pages 2895–2912. PMLR, 2025. 3
- [24] Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao Gu, Rui Ai, and Hesheng Wang. Cam4docc: Benchmark for camera-only 4d occupancy forecasting in autonomous driving applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21486–21495, 2024. 1, 3, 8

- [25] Yunsheng Ma, Wenqian Ye, Can Cui, Haiming Zhang, Shuo Xing, Fucai Ke, Jinhong Wang, Chenglin Miao, Jintai Chen, Hamid Rezaatofghi, et al. Position: Prospective of autonomous driving-multimodal llms world models embodied intelligence ai alignment and mamba. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1010–1026, 2025. 1
- [26] Reza Mahjourian, Jinkyu Kim, Yuning Chai, Mingxing Tan, Ben Sapp, and Dragomir Anguelov. Occupancy flow fields for motion forecasting in autonomous driving. *IEEE Robotics and Automation Letters*, 7(2):5639–5646, 2022. 3
- [27] Rui Pan, Shuo Xing, Shizhe Diao, Wenhe Sun, Xiang Liu, Kashun Shum, Jipeng Zhang, Renjie Pi, and Tong Zhang. Plum: Prompt learning using metaheuristics. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2177–2197, 2024. 1
- [28] Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *2024 IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, 2024. 2, 4, 8
- [29] Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17996–18006, 2024. 2
- [30] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 4
- [31] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 4, 6
- [32] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8406–8415, 2023. 2
- [33] Godfried T Toussaint. Solving geometric problems with the rotating calipers. In *Proc. IEEE Melecon*, page A10, 1983. 5
- [34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 7
- [35] Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *arXiv preprint arXiv:2405.20337*, 2024. 1, 3, 6
- [36] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17850–17859, 2023. 1, 2, 4
- [37] Yuping Wang and Jier Chen. Eqdrive: Efficient equivariant motion forecasting with multi-modality for autonomous driving. In *2023 8th International Conference on Robotics and Automation Engineering (ICRAE)*, pages 224–229. IEEE, 2023. 1
- [38] Yuping Wang and Jier Chen. Equivariant map and agent geometry for autonomous driving motion prediction. In *2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–6. IEEE, 2023. 1
- [39] Yuping Wang, Shuo Xing, Cui Can, Renjie Li, Hongyuan Hua, Kexin Tian, Zhaobin Mo, Xiangbo Gao, Keshu Wu, Sulong Zhou, et al. Generative ai for autonomous driving: Frontiers and opportunities. *arXiv preprint arXiv:2505.08854*, 2025. 1
- [40] Zehao Wang, Yuping Wang, Zhuoyuan Wu, Hengbo Ma, Zhaowei Li, Hang Qiu, and Jiachen Li. Cmp: Cooperative motion prediction with multi-agent communication. *IEEE Robotics and Automation Letters*, pages 1–8, 2025. 2
- [41] Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. Occllama: An occupancy-language-action generative world model for autonomous driving. *arXiv preprint arXiv:2409.03272*, 2024. 3, 6
- [42] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 1, 2, 4
- [43] Joey Wilson, Jingyu Song, Yuewei Fu, Arthur Zhang, Andrew Capodiceci, Paramsothy Jayakumar, Kira Barton, and Maani Ghaffari. Motionsc: Data set and network for real-time semantic mapping in dynamic environments. *IEEE Robotics and Automation Letters*, 7(3):8439–8446, 2022. 2, 3
- [44] Shuo Xing, Hongyuan Hua, Xiangbo Gao, Shenzhe Zhu, Renjie Li, Kexin Tian, Xiaopeng Li, Heng Huang, Tianbao Yang, Zhangyang Wang, et al. Autotrust: Benchmarking trustworthiness in large vision language models for autonomous driving. *arXiv preprint arXiv:2412.15206*, 2024. 1
- [45] Shuo Xing, Lanqing Guo, Hongyuan Hua, Seoyoung Lee, Peiran Li, Yufei Wang, Zhangyang Wang, and Zhengzhong Tu. Demystifying the visual quality paradox in multimodal large language models. *arXiv preprint arXiv:2506.15645*, 2025. 1
- [46] Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. Openemmas: Open-source multimodal model for end-to-end autonomous

- driving. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1001–1009, 2025. [1](#)
- [47] Shuo Xing, Zezhou Sun, Shuangyu Xie, Kaiyuan Chen, Yanjia Huang, Yuping Wang, Jiachen Li, Dezhen Song, and Zhengzhong Tu. Can large vision language models read maps like a human? *arXiv preprint arXiv:2503.14607*, 2025. [1](#)
- [48] Shuo Xing, Yuping Wang, Peiran Li, Ruizheng Bai, Yueqi Wang, Chengxuan Qian, Huaxiu Yao, and Zhengzhong Tu. Re-align: Aligning vision language models via retrieval-augmented direct preference optimization. *arXiv preprint arXiv:2502.13146*, 2025. [1](#)
- [49] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. [2](#), [4](#), [8](#)
- [50] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird’s eye view semantic segmentation with sparse transformers. In *Conference on Robot Learning (CoRL)*, 2022. [2](#)
- [51] Zhangchen Ye, Tao Jiang, Chenfeng Xu, Yiming Li, and Hang Zhao. Cvt-occ: Cost volume temporal fusion for 3d occupancy prediction. In *European Conference on Computer Vision*, pages 381–397. Springer, 2024. [1](#), [2](#), [6](#), [8](#)
- [52] Junming Zhang, Weijia Chen, Yuping Wang, Ram Vasudevan, and Matthew Johnson-Roberson. Point set voting for partial point cloud analysis. *IEEE Robotics and Automation Letters*, 6(2):596–603, 2021. [1](#)
- [53] Yanan Zhang, Jinqing Zhang, Zengran Wang, Junhao Xu, and Di Huang. Vision-based 3d occupancy prediction in autonomous driving: a review and outlook. *arXiv preprint arXiv:2405.02595*, 2024. [1](#)
- [54] Ziming Zhang, Fangzhou Lin, Haotian Liu, Jose Morales, Haichong Zhang, Kazunori Yamada, Vijaya B Kolachalama, and Venkatesh Saligrama. Gps: A probabilistic distributional similarity with gumbel priors for set-to-set matching. In *The Thirteenth International Conference on Learning Representations*, 2025. [1](#)
- [55] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2025. [1](#), [3](#), [4](#), [6](#), [7](#), [5](#)
- [56] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022. [1](#)
- [57] Jun Zhuang, Haibo Jin, Ye Zhang, Zhengjian Kang, Wenbin Zhang, Gaby G Dagher, and Haohan Wang. Exploring the vulnerability of the content moderation guardrail in large language models via intent manipulation. *arXiv preprint arXiv:2505.18556*, 2025. [1](#)