

# ShelfOcc: Native 3D Supervision beyond LiDAR for Vision-Based Occupancy Estimation

Simon Boeder<sup>1,2</sup>

simon.boeder@de.bosch.com fabian.gigengack@de.bosch.com simon.roesler@de.bosch.com

Holger Caesar<sup>3</sup>

h.Caesar@tudelft.nl b.risse@uni-muenster.de

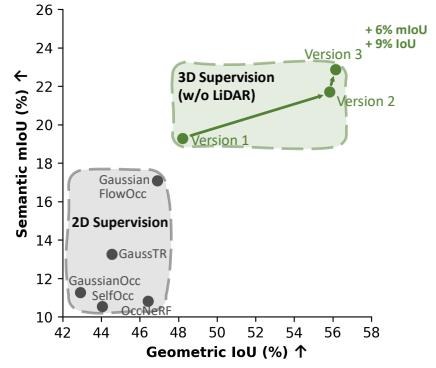
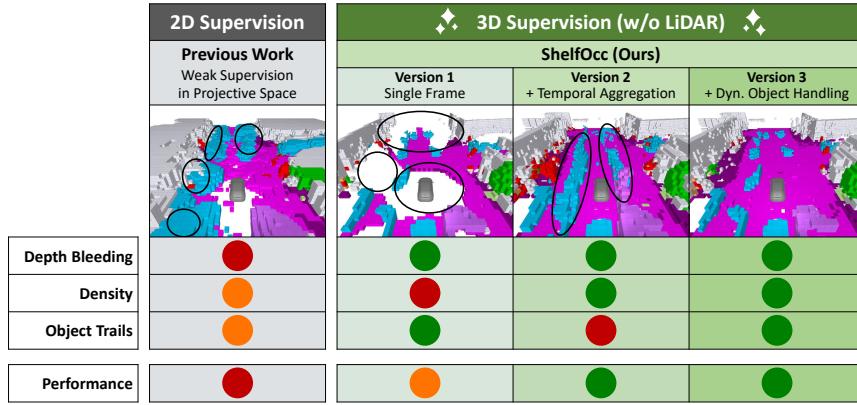
Fabian Gigengack<sup>1</sup>Simon Roesler<sup>1</sup><sup>1</sup>Bosch Research    <sup>2</sup>University of Münster    <sup>3</sup>TU Delft

Figure 1. **Contributions of ShelfOcc.** We propose a shift in supervision strategy for weakly/shelf-supervised occupancy estimation. Unlike prior 2D rendering-based approaches, which are prone to depth bleeding, *ShelfOcc* trains occupancy networks directly in native 3D voxel space with pseudo-labels generated using a combination of geometric and semantic FMs. By accumulating and filtering static geometry while handling dynamic objects separately, our approach yields clean and consistent 3D supervision relying only on images, without LiDAR. This shift in supervision leads to a significant performance gain over previous methods, as illustrated on the right.

## Abstract

Recent progress in self- and weakly supervised occupancy estimation has largely relied on 2D projection or rendering-based supervision, which suffers from geometric inconsistencies and severe depth bleeding. We thus introduce *ShelfOcc*, a vision-only method that overcomes these limitations without relying on LiDAR. *ShelfOcc* brings supervision into native 3D space by generating metrically consistent semantic voxel labels from video, enabling true 3D supervision without any additional sensors or manual 3D annotations. While recent vision-based 3D geometry foundation models provide a promising source of prior knowledge, they do not work out of the box as a prediction due to sparse or noisy and inconsistent geometry, especially in dynamic driving scenes. Our method introduces a dedicated framework that mitigates these issues by filtering and accumulating static geometry consistently across frames, handling

dynamic content and propagating semantic information into a stable voxel representation. This data-centric shift in supervision for weakly/shelf-supervised occupancy estimation allows the use of essentially any SOTA occupancy model architecture without relying on LiDAR data. We argue that such high-quality supervision is essential for robust occupancy learning and constitutes an important complementary avenue to architectural innovation. On the Occ3D-nuScenes benchmark, *ShelfOcc* substantially outperforms all previous weakly/shelf-supervised methods (up to a 34% relative improvement), establishing a new data-driven direction for LiDAR-free 3D scene understanding.

## 1. Introduction

Accurate and efficient 3D occupancy estimation is fundamental for safe and reliable autonomous driving, providing a dense understanding of the environment which is crucial for planning and navigation [46, 63]. While significant ad-

vancements have been made, many state-of-the-art methods still depend heavily on dense 3D ground truth annotations derived from LiDAR sensors. This dependency is a major bottleneck for scalability and real-world application, as manual dense 3D annotation is extremely costly and labor-intensive [54]. Moreover, fleet vehicles are rarely equipped with such reference sensors, preventing their data from being utilized for supervised training.

To overcome the challenges of 3D label acquisition, researchers have explored weakly/shelf-supervised approaches, often relying on 2D annotations or photometric losses through differentiable rendering. We use the term *shelf-supervised* [27, 66] to denote a form of self-supervision that relies on off-the-shelf foundation models as sources of geometric and/or semantic priors. Methods like SelfOcc [20], OccNeRF [71], and GaussianFlowOcc [2] utilize techniques such as Neural Radiance Fields (NeRF) [41] or 3D Gaussian Splatting (3DGS) [26] to render 3D scene representations back to 2D image space. This enables supervision from easily obtainable 2D cues, such as semantic segmentation masks produced by models like GroundedSAM [45] and monocular depth maps from, e.g., Metric3D [68]. However, a critical limitation persists: learning complex 3D geometry solely from 2D image-based losses is inherently difficult. This frequently results in artifacts such as depth bleeding, where models fail to precisely capture the volumetric extent of objects along viewing rays, since 2D signals primarily provide information about the visible object boundary. To provide more complete 3D supervision, rendering-based approaches rely on temporal consistency, which requires handling dynamic objects and further complicates the training while reducing but not removing the issue of depth bleeding. These ambiguities limit performance compared to methods with direct 3D supervision.

The growing availability of off-the-shelf 3D vision foundation models (FMs) has opened new opportunities for leveraging pretrained geometric priors in downstream perception tasks. Models such as the Visual Geometry Grounded Transformer (VGGT) [59] and MapAnything [25] can infer detailed 3D scene attributes from images, making them appealing sources of supervision for 3D occupancy learning. Trained on vast amounts of 3D-annotated data, these FMs can infer camera parameters, depth maps, and dense 3D point clouds from multiple images in a single forward pass. However, as illustrated in Fig. 1, directly applying such FMs, which typically assume static scenes and consistent camera parameters, to dynamic multi-camera driving sequences poses several challenges. These include handling non-static elements and integrating semantic information into 3D space. A naive frame-wise application (Fig. 1 Version 1) of FMs produces sparse and incomplete labels, while simple temporal aggregation (Fig. 1 Version 2) leads to model violations which manifest

themselves in dynamic object trails and ghosting artifacts.

In this paper, we propose *ShelfOcc*, a shelf-supervised learning framework that leverages 3D geometry FMs to generate high-quality 3D pseudo-labels and train high-performing occupancy networks. Our proposed 3D pseudo-label generation pipeline separates dynamic and static scene parts, filters noisy and wrong predictions, accumulates the static scene across the sequence and re-introduces dynamic objects per frame. *ShelfOcc* offers a plug-and-play solution for supervising any occupancy network with 3D labels, solely from camera images. Our central hypothesis is that direct 3D supervision, even when derived from FM-generated pseudo-labels, can substantially enhance the geometric understanding of occupancy networks, yielding superior performance compared to 2D-supervised counterparts, without requiring costly 3D annotations. Owing to its modular design, *ShelfOcc* can be seamlessly integrated into existing occupancy prediction pipelines, providing a scalable framework for 3D supervision and enabling LiDAR-free training of state-of-the-art occupancy architectures. We demonstrate that enhancing the supervision signal can lead to substantial gains in occupancy estimation without modifying the network design itself.

In summary our contributions are as follows:

- **Paradigm shift for shelf-supervised occupancy estimation.** We introduce *ShelfOcc*, a novel framework that enables direct and native 3D shelf-supervision for occupancy estimation by leveraging the emergent capabilities of off-the-shelf geometric and semantic foundation models, eliminating the need for LiDAR, manual 3D annotations or 2D rendering supervision.
- **Vision-only 3D pseudo-label generation.** We develop a sophisticated, semantics-aware pipeline that separates static and dynamic scene components, accumulates static geometry across time, reintroduces dynamic objects per frame, and filters low-confidence predictions to produce clean and consistent 3D voxel pseudo-labels.
- **Data-centric performance gains.** We empirically show that our enhanced 3D supervision signal yields substantial performance improvements, surpassing all prior shelf-supervised methods on the Occ3D-nuScenes benchmark (up to a 34% relative improvement). Importantly, these gains are consistently observed across multiple plug-and-play occupancy network architectures.

## 2. Related Work

### 2.1. 3D Occupancy Estimation

The 3D semantic occupancy estimation task has garnered significant attention in autonomous driving research. Broadly, vision-based approaches can be categorized into fully supervised and shelf-supervised methods.

**Fully Supervised Methods** leverage dense 3D ground

truth annotations, typically derived from LiDAR sensors with additional manual labeling. Most approaches adapt architectures originally developed for object detection, lifting multi-view camera features into a 3D voxel grid, refining them through 3D convolutions or deformable attention, and applying losses directly in voxel space [5, 18, 21, 32, 56]. Recent advancements have improved the model efficiency [35, 38, 47, 53, 58, 69], optimized training procedures [1, 12, 16, 22, 43, 50], or improved the overall occupancy estimation performance through refined architectural designs and specific modeling strategies [7, 8, 23, 33, 34, 39, 40, 51, 67, 73, 74]. A growing research direction explores the integration of 3D occupancy with feature spaces of foundation models. Some methods align predicted voxel features with vision-language representations to enable open-vocabulary occupancy estimation [3, 31, 52, 57, 70, 75], while [24, 48] distill DINO features to obtain strong semantic priors.

**Shelf-Supervised Methods**, on the other hand, aim to mitigate the reliance on expensive 3D labels by utilizing 2D annotations or self-supervision. SelfOcc [20] and Oc-*c*NeRF [71] employ volume rendering techniques to project estimated 3D occupancy into 2D image space, where photometric and semantic losses from pretrained models [45, 72] provide supervision. GaussianOcc [13] and GaussTR [24] instead use 3D Gaussian Splatting for rendering supervision. GaussianFlowOcc [2] further models scene dynamics to mitigate temporal inconsistencies during training. While these shelf-supervised methods remove the need for 3D ground truth, they struggle with precise geometric understanding due to the inherent ambiguity of 2D signals and often face challenges in dynamic scenes. GS-Occ3D [65] recently demonstrated direct 3D Gaussian optimization for occupancy reconstruction, yet without incorporating semantic labels. LeAP [14] proposes a voxel pseudo-label pipeline for occupancy estimation using LiDAR scans as geometric priors. The concurrent work and preprint Easy-Occ [15] instead lifts 2D semantic segmentation masks into 3D using (only) monocular depth estimates, but does not achieve a notable performance improvement over previous work. It is important to note that some methods, such as AGO [31], AutoOcc [78], and VEON [75], while not using semantic 3D annotations for training, still utilize LiDAR data for geometric supervision, which sets them apart from the purely vision-based, shelf-supervised paradigm we focus on. Although these methods are not directly comparable to ours, we include a quantitative comparison in the supplement for completeness. Furthermore, some research is moving towards 4D occupancy and motion prediction, and even world models, incorporating occupancy for future state prediction [10, 39, 60, 76]. Finally, there is considerable amount of work on 4D scene reconstruction for driving scenes, enabling tasks like novel view synthesis, scene edit-

ing and simulation [19, 37, 49, 62, 64].

## 2.2. Vision Foundation Models

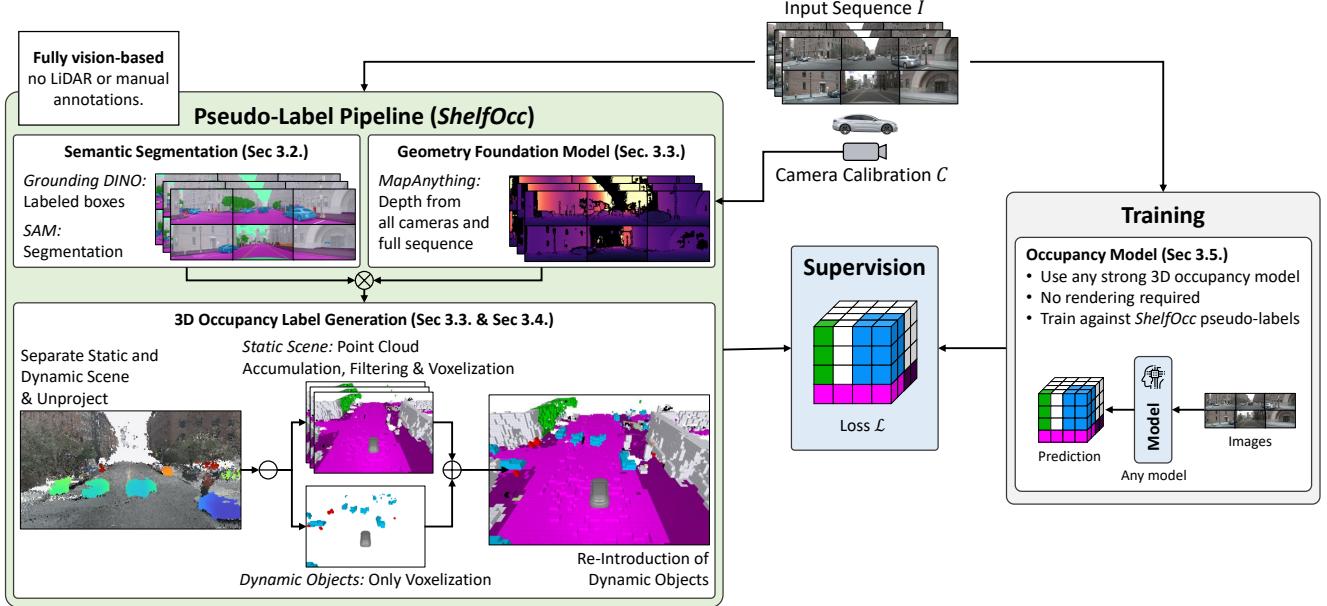
The rapid progress of deep learning has led to the emergence of powerful Vision Foundation Models (VFs), which have transformed a wide range of computer vision tasks. Models such as DINO [6, 42] and CLIP [44] are trained on large-scale image corpora to learn general-purpose visual representations. In parallel, Vision-Language Models (VLMs) combine visual perception with natural language understanding. Grounding DINO [36] and Detic [77] are prominent examples of open-vocabulary object detectors capable of localizing objects based on arbitrary text prompts. GroundedSAM [45] extends this concept by integrating Grounding DINO with the Segment Anything Model (SAM) [28], enabling open-vocabulary segmentation by generating pixel-accurate masks for any textual query without additional training. Text4Seg [29] further leverages multi-modal language models to enhance open-vocabulary segmentation capabilities.

Recently Geometry Foundation Models emerged, which specialize in inferring 3D geometric information from 2D images. Models such as DUST3R [61], MAST3R [30], VGGT [59], and MapAnything [25] are trained on large-scale datasets with 3D supervision (e.g., structure-from-motion data, synthetic scenes) and can estimate depth, camera poses, and dense 3D point clouds. MapAnything [25], in particular, is designed for large-scale scene reconstruction and can leverage known camera parameters, making it highly suitable for autonomous driving applications. These models provide a powerful source of metrically consistent 3D geometry from standard camera inputs, which is foundational to our approach for generating 3D pseudo-labels.

## 3. Methodology

Our proposed ShelfOcc framework generates high-fidelity, metrically-scaled 3D semantic voxel pseudo-labels from multi-view image sequences. These pseudo-labels act as a plug-and-play source of supervision for any occupancy estimation network that leverages 3D labels. The framework is specifically designed to overcome the limitations of directly applying 3D geometry foundation models (FMs) to dynamic scenes.

The overall framework is illustrated in Fig. 2. We leverage the 3D geometry model MapAnything [25] and the 2D semantic segmentation model GroundedSAM [45] to construct semantics-aware 3D pseudo-labels. The framework comprises six key stages: 2D semantic segmentation mask prediction, initial 3D geometry estimation and filtering, static scene accumulation, dynamic object reintroduction, voxelization with visibility mask generation, and training an occupancy network using the generated labels. We describe each stage in detail in the following subsections.



**Figure 2. Overview of the *ShelfOcc* framework.** We leverage a 3D geometry foundation model (MapAnything [25]) and a 2D semantic foundation model (GroundedSAM [45]) to construct precise 3D semantic voxel pseudo-labels. The pipeline processes image sequences, segregating static and dynamic scene elements, filtering and aggregating static elements and carefully reintroducing dynamic objects to mitigate artifacts. These generated 3D pseudo-labels serve as a plug-and-play supervision for any 3D occupancy network.

### 3.1. Notation

We denote the dataset as a collection of driving sequences  $\mathcal{S} = \{S_1, S_2, \dots, S_N\}$ , where each sequence  $S_k$  represents a short temporal segment composed of multiple consecutive frames. Formally, a sequence  $S_k$  contains  $\mathcal{F}_k = \{f_1, f_2, \dots, f_{T_k}\}$ , where  $T_k$  is the total number of frames in sequence  $S_k$ . Pseudo-label generation is performed independently for each sequence. At a given time step  $t$ , the frame  $f_t$  consists of a set of multi-view images  $\mathcal{I}_t = \{I_{1,t}, I_{2,t}, \dots, I_{C,t}\}$ , where  $C$  denotes the total number of cameras. Each image  $I_{i,t}$  is associated with its intrinsic calibration matrix  $\mathbf{K}_{i,t}$  and extrinsic camera pose  $\mathbf{T}_{i,t}$ , which together define the projection from 3D world coordinates to the 2D image plane and vice versa.

### 3.2. 2D Pseudo-Semantic Masks

The initial step in the ShelfOcc pipeline involves generating dense 2D pseudo-semantics segmentation masks for all input images  $I_i$  within the sequence. We employ GroundedSAM [45], which combines an open-vocabulary object detector and a highly capable segmentation model. The process unfolds in two distinct stages. Firstly, each image is fed into Grounding DINO [36], an open-vocabulary object detection method, along with a set of text prompts corresponding to the target semantic classes. The predicted bounding boxes then serve as input prompts for SAM (Segment Anything Model) [28], which then generates precise 2D masks, inheriting the label from its corresponding predicted box.

**Mitigating False Negatives / Positives via sky grounding.** We observed that querying Grounding DINO for all target classes simultaneously often results in too few detections, causing many objects in the image to be missed. However, querying the model for each class individually can lead to a bias towards detecting objects even when they are not present, resulting in high-confidence but incorrect boxes and frequent class confusions. To address this issue, we input each class prompt individually into Grounding DINO, together with a generic background label (e.g., ‘sky’). This strategy greatly reduces false positives by providing Grounding DINO an alternative high-confidence prediction when the query object is absent. Any boxes predicted with the background label are discarded before the second stage with SAM. The full vocabulary used for generating the semantic segmentation masks can be found in the supplementary material. These dense 2D masks are crucial for assigning semantic information to the 3D points and for dynamically identifying objects within the scene, a critical step detailed in Sec. 3.3.

### 3.3. 3D Geometry Estimation and Static/Dynamic Separation

We process the entire multi-view image sequence (across all cameras  $C$  and time steps  $T$ ) through a 3D geometry foundation model. For this work, we specifically leverage MapAnything [25]. MapAnything is particularly suitable because, unlike some other FMs, it can optionally take avail-

able camera poses (intrinsics  $\mathbf{K}_i$  and extrinsics  $\mathbf{T}_i$ ) as input, which are readily available from autonomous driving datasets like nuScenes [4]. This enables MapAnything to directly predict metrically-scaled depth maps  $\mathbf{D}_i$  for each input image  $I_i$ . The depth maps can be unprojected using the camera poses and intrinsics to create a 3D point cloud representing the scene geometry. One of the primary challenges in generating accurate 3D pseudo-labels for dynamic driving scenes is the handling of moving objects. Naively accumulating points from all time steps of the sequence into a single 3D scene would result in moving objects appearing multiple times along their trajectory, polluting the scene representation. To prevent this, we separate the scene into static and a dynamic components.

**Static Scene Construction.** To construct the static part of the scene, which remains consistent across all time steps within a sequence, we first use the 2D pseudo-semantics masks from Sec. 3.2 to determine which pixels belong to dynamic objects. We then unproject only those pixels that are *not* identified as dynamic into 3D points. This selective unprojection prevents motion artifacts, which are common because FMs like MapAnything are primarily trained on static scenarios. The 3D coordinate  $\mathbf{P}(u, v)$  for a pixel  $(u, v)$  in image  $\mathbf{I}_i$  are obtained by:

$$\mathbf{P}(u, v) = \mathbf{T}_i \cdot \left( \mathbf{K}_i^{-1} \cdot \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \cdot \mathbf{D}_i(u, v) \right). \quad (1)$$

Here,  $\mathbf{D}_i(u, v)$  is the depth predicted by MapAnything,  $\mathbf{K}_i$  represents the camera intrinsics, and  $\mathbf{T}_i$  are the camera extrinsics. All static points, collected from all cameras and time steps within a given sequence, are then aggregated into a single global static point cloud  $\mathcal{P}_{\text{static}}$ .

**Confidence Filtering.** To mitigate noise in the accumulated static point cloud and suppress spurious occupied voxels in the final 3D volume, we apply two confidence filtering strategies. First, we remove points likely resulting from erroneous depth predictions by retaining only those that are confirmed by multiple frames within the sequence. Concretely, for each pixel ray, we record how often it traverses a given voxel cell and how often it terminates within that cell based on the unprojected depth points. Points located in cells that are intersected more frequently than they are confirmed are discarded, as they likely correspond to incorrect depth estimates. Second, we prune voxels with insufficient point density (fewer than four points in our case), as such sparsely populated regions tend to represent unreliable or noisy predictions. Together, these filtering steps substantially improve the consistency and reliability of the generated point cloud, preserving only high-confidence geometric information for subsequent voxelization.

**Dynamic Scene Construction.** For each time step  $t$  in the sequence, we generate a dedicated dynamic point cloud  $\mathcal{P}_{\text{dynamic}, t}$ . This is achieved by unprojecting all pixels that were identified as dynamic by the 2D semantic segmentation in all cameras  $\mathbf{I}_{i,t}$  at that time step using Eq. (1). This ensures that dynamic objects are captured precisely at their positions in each frame.

### 3.4. Voxelization and Visibility Mask Generation

The final 3D semantic point cloud for any given frame  $t$  is then constructed by combining the global static point cloud with the dynamic point cloud specific to that time step. The final step in the ShelfOcc pipeline converts the 3D semantic point clouds  $\mathcal{P}_t$  into high-fidelity, metrically-scaled 3D semantic voxel labels, which are ready for direct supervision.

**Voxelization.** A target voxel grid is first defined with specific dimensions and resolution (e.g.,  $[-40m, 40m]$  in X/Y and  $[-1m, 5.4m]$  in Z with a  $0.4m$  resolution for nuScenes). For each voxel  $\mathbf{v}$  in this grid, we aggregate all 3D points from  $\mathcal{P}_t$  that fall within its spatial bounds. The semantic label of  $\mathbf{v}$  is then determined by a majority voting scheme among these collected points. To mitigate class imbalance effects near the ground where frequent classes (*road* or *terrain*) tend to dominate, we prioritize minority object classes (e.g., *traffic cones*). If no points fall within a voxel, it is marked as *empty*. This process yields a dense 3D semantic voxel grid  $\mathbf{V}_t$  for each frame in the sequence  $S_t$ .

**Camera Visibility Mask.** To effectively train an occupancy network, it is essential to distinguish between truly empty space and regions that are merely unobserved by the cameras. For each frame  $t$ , we generate a camera visibility mask  $\mathbf{M}_{\text{vis}, t}$ . This is achieved by casting rays from each ground truth camera position through the scene. For each ray, we identify the first occupied voxel. All voxels along the ray *before* this first occupied voxel are marked as 'visible free space'. Conversely, voxels that lie behind occupied voxels or are entirely outside the frustum of any camera are marked as 'unobserved'. This visibility mask  $\mathbf{M}_{\text{vis}, t}$  is critical during network training, as it guides the loss computation, ensuring that the model is only penalized for incorrect predictions within regions theoretically observable by the cameras. This accurate voxelization process ultimately provides high-quality, dense 3D semantic voxel grids  $\mathbf{V}_t$  and corresponding visibility masks  $\mathbf{M}_{\text{vis}, t}$ .

### 3.5. Training with ShelfOcc Labels

The central advantage of ShelfOcc is its modular design, enabling seamless integration across different occupancy prediction algorithms. Any existing occupancy network architecture that uses 3D voxel labels for supervision can be directly trained using ShelfOcc labels. This circumvents the

Table 1. **Occupancy estimation performance on the Occ3D-nuScenes validation set.** Best performing per column in **bold**, second best in *italics*. All methods ignore the *others* and *other flat* classes. *IoU* represents the geometric performance independent of the semantic label, while *mIoU* is the mean IoU over all classes.

Method			Semantic Classes														
	mIoU	IoU	barrier	bicycle	bus	car	cons. vehicle	motorcycle	pedestrian	traffic cone	trailer	truck	drv. surf.	sidewalk	terrain	mannade	vegetation
SelfOcc [20]	10.54	44.05	0.15	0.66	5.46	12.54	0.00	0.80	2.10	0.00	0.00	8.25	55.49	26.30	26.54	14.22	5.60
OccNeRF [71]	10.81	46.43	0.83	0.82	5.13	12.49	3.50	0.23	3.10	1.84	0.52	3.90	52.62	20.81	24.75	18.45	13.19
GaussianOcc [13]	11.26	42.91	1.79	5.82	14.58	13.55	1.30	2.82	7.95	9.76	0.56	9.61	44.59	20.10	17.58	8.61	10.29
GaussTR [24]	13.26	44.54	2.09	5.22	14.07	20.43	5.70	7.08	5.12	3.93	0.92	13.36	39.44	15.68	22.89	21.17	21.87
EasyOcc [15]	15.96	38.86	1.85	8.18	16.66	22.12	1.01	7.74	<b>14.74</b>	<i>12.84</i>	0.98	13.76	55.91	27.96	22.73	15.77	17.20
GaussianFlowOcc [2]	17.08	46.91	6.75	9.68	18.98	17.15	4.19	11.78	9.27	10.30	1.83	12.33	61.03	31.17	34.78	14.66	12.40
Ours: ShelfOcc + COTR [40]	18.65	<i>53.71</i>	9.10	6.20	22.92	22.08	1.66	5.94	9.92	8.55	0.0	<i>15.32</i>	67.93	31.13	38.76	<i>23.11</i>	17.15
Ours: ShelfOcc + CVT-Occ [67]	19.21	52.72	<i>11.53</i>	6.38	20.39	21.92	4.20	10.18	9.02	10.67	0.89	13.08	<b>68.42</b>	31.23	41.42	22.74	16.15
Ours: ShelfOcc + STCOcc [34]	<b>22.87</b>	<b>56.14</b>	<b>13.98</b>	<b>11.36</b>	<b>25.27</b>	<b>25.80</b>	<b>7.25</b>	<b>16.61</b>	<b>12.91</b>	<b>13.42</b>	<b>5.37</b>	<b>17.15</b>	<b>68.01</b>	<b>34.66</b>	<b>42.73</b>	<b>25.63</b>	<b>22.89</b>

need for complex differentiable rendering pipelines that are typically employed in 2D-supervised methods.

An occupancy network takes multi-view images as input and predicts a dense 3D semantic voxel grid  $\hat{\mathbf{V}}_t$ , representing semantic probabilities for each voxel. This output is directly compared with the generated ShelfOcc labels  $\mathbf{V}_t$  using loss functions like the cross-entropy loss

$$\mathcal{L} = \sum_t \sum_{\mathbf{v} \in \mathbf{V}_t} \mathbf{M}_{\text{vis},t}(\mathbf{v}) \cdot \mathcal{L}_{\text{CE}}(\hat{\mathbf{V}}_t(\mathbf{v}), \mathbf{V}_t(\mathbf{v})), \quad (2)$$

however the exact loss functions used depends on the model of choice. The camera visibility mask  $\mathbf{M}_{\text{vis},t}(\mathbf{v})$  is usually applied as a weighting factor in the loss computation. This ensures that the loss is computed only for voxels within observable regions of the scene.

This direct 3D supervision offers two key advantages: First, it significantly enhances 3D geometric understanding by allowing the network to learn directly from explicit 3D targets, effectively mitigating issues such as depth bleeding and promoting a more complete and accurate geometric representation. Second, it simplifies the training pipeline by eliminating the need for complex differentiable rendering mechanisms, thereby reducing memory consumption and computational overhead.

## 4. Experiments

### 4.1. Dataset

We conduct our experiments on the Occ3D-nuScenes benchmark [55], which builds upon the widely used nuScenes dataset [4, 11] and provides 3D semantic occupancy ground truth for all scenes in the dataset (that we only use during validation). Following standard protocol, we evaluate model performance using the Intersection-over-Union (IoU) across a predefined set of semantic classes, with the mIoU being the mean over all classes. In addition, we report the geometric IoU, which measures the voxel-wise accuracy of occupied versus free space regardless of

the underlying semantic category, providing a measure of geometric fidelity. To further assess consistency along the depth axis, we also report the RayIoU metric [35], a ray-based evaluation that mitigates inconsistencies in traditional voxel-level IoU by comparing occupancy predictions along camera rays.

### 4.2. Experimental Setup

We focus on the shelf-supervised setting, where models are trained using only multi-view camera inputs without access to LiDAR data or any 3D annotations from the target dataset. To assess the quality and generality of our generated 3D pseudo-labels, we train several state-of-the-art occupancy networks originally designed for fully supervised training, namely COTR [40], CVT-Occ [67], and STCOcc [34], directly on our generated pseudo-labels. This enables direct comparison to previous shelf-supervised methods while leveraging the advantages of established 3D architectures. All models are trained with input images at a resolution of  $256 \times 704$ , using a ResNet-50 [17] backbone pretrained on ImageNet [9]. Source code will be made available after publication.

### 4.3. Main Results

As summarized in Tab. 1, our method establishes new state-of-the-art results across all evaluated architectures in the shelf-supervised occupancy estimation setting. Models trained on our ShelfOcc pseudo-labels consistently outperform prior approaches, achieving improvements of up to +5.79 mIoU and +9.23 geometric IoU over the previous best-performing method, GaussianFlowOcc [2]. This corresponds to relative gains of 34% in mIoU and 20% in geometric IoU. We believe the strong performance stems from our direct 3D supervision strategy. High-quality pseudo-labels enable occupancy networks to be trained natively in 3D space, avoiding the instability and ambiguity associated with 2D rendering-based supervision. The inherently multi-view consistent geometric predictions from MapAny-

**Table 2. Occupancy estimation performance in terms of RayIoU [35].** We compare the performance of STCOcc trained with our *ShelfOcc* labels against previous works.

Method	mRayIoU	RayIoU@1	RayIoU@2	RayIoU@4
GaussianOcc [13]	11.85	8.69	11.90	14.95
GaussianFlowOcc [2]	16.47	11.81	16.58	20.98
ShelfOcc + COTR [40]	17.24	12.22	17.33	22.18
ShelfOcc + CVT-Occ [67]	18.28	13.11	18.29	23.45
ShelfOcc + STCOcc [34]	<b>19.97</b>	<b>14.38</b>	<b>20.07</b>	<b>25.47</b>

thing [25] provide metrically accurate 3D supervision, facilitating more stable optimization and improved volumetric reasoning compared to methods relying on 2D depth rendering. The comparison in Fig. 1 further illustrates the clear performance gap between 2D rendering-based methods and our 3D shelf-supervised approach. A similar trend is observed on the RayIoU metric [35], which provides a more depth-consistent evaluation of volumetric predictions. The results in Tab. 2 show that across all evaluated ranges (1m, 2m, and 4m), our pseudo-labels lead to consistent improvements in RayIoU when training COTR, CVT-Occ and STCOcc, highlighting the generality and robustness of the generated supervision signal. Furthermore, the qualitative results in Fig. 3 demonstrate several key advantages of our approach. The pseudo-labels generated by our framework are noticeably denser than the LiDAR-based ground truth, providing more comprehensive spatial coverage of the scene. When training on these labels, the networks learn to effectively suppress noise and artifacts present in the raw pseudo-labels, yielding smooth and coherent occupancy maps. Moreover, the models demonstrate strong completion capabilities, successfully reconstructing the full extent of partially visible objects and improving overall scene consistency. We provide further results and comparisons with LiDAR-based and supervised approaches in the supplementary material.

#### 4.4. Ablation Study

**Label Quality and Semantic Segmentation.** We first evaluate the quality of the generated pseudo-labels by directly comparing them against the Occ3D-nuScenes benchmark, as shown in Tab. 3. This experiment assesses the intrinsic accuracy of the pseudo-labels independent of any model training. Without sky grounding, the pseudo-labels already capture the overall scene structure but contain substantial noise and misclassifications. Introducing the proposed sky grounding technique leads to a notable improvement of +3.41 mIoU (a 55% relative increase) and +8.79 geometric IoU (a 51% relative increase), indicating that better sky segmentation substantially enhances 2D semantic mask quality and reduces false positives and negatives in the projected 3D labels. While the pseudo-labels capture the general scene structure, their performance remains considerably below that of models trained on them, confirming

that off-the-shelf foundation models alone are insufficient for high-quality occupancy prediction. However, when used as supervision, the models trained on these pseudo-labels exhibit strong generalization and completion capabilities, filling in missing geometry and correcting inconsistencies. The direct gain from sky grounding persists but is smaller after training (+0.39 mIoU and +1.8geometric IoU), indicating that the model can effectively learn to correct residual noise and resolve label ambiguities.

**Table 3. Evaluation of pseudo-label quality and sky grounding.** We report the occupancy estimation performance of the generated *ShelfOcc* pseudo-labels when directly evaluated on the Occ3D-nuScenes benchmark, treating them as model predictions.

w/ STCOcc [34]	Sky Grounding	mIoU	IoU
✗	✗	6.21	17.21
✗	✓	9.62	26.00
✓	✗	22.48	54.26
✓	✓	22.87	56.14

**Visibility Mask.** We analyze the influence of using the generated camera visibility mask, which restricts supervision to regions observable by the cameras, cf. Tab. 4. Applying the mask leads to a significant improvement in both IoU and mIoU compared to training without it, highlighting its importance for 3D supervision in this setting.

**Confidence Filter.** We further ablate the effect of the proposed confidence filtering techniques in Sec. 3.3. As shown in Tab. 5, disabling the confidence filter during pseudo-label generation leads to a decrease of –1.88 in mIoU and –1.63 in IoU, showing that our proposed pipeline produces cleaner pseudo-labels.

**Table 4. Ablation of camera mask.** Training STCOcc w/o mask, evaluated w/ mask.

Camera Mask	mIoU	IoU
✗	13.41	26.21
✓	22.87	56.14

**Table 5. Ablation of conf. filter.** Training STCOcc on pseudo-labels w/o conf. filters.

Confidence Filter	mIoU	IoU
✗	20.99	54.51
✓	22.87	56.14

**Aggregation.** We further ablate the geometric aggregation pipeline to isolate the benefits of each version (see Fig. 1). We compare three configurations: (1) Pseudo-label generation on a per-frame basis, (2) Aggregation over time without explicit handling of dynamic objects, (3) Full pipeline including dynamic object processing. We generate pseudo-labels with each configuration and train STCOcc on each, cf. Tab. 6. Even the single-frame version surpasses previous shelf-supervised approaches in both IoU and mIoU, underscoring the inherent strength of our 3D supervision. Temporal aggregation yields a substantial boost,

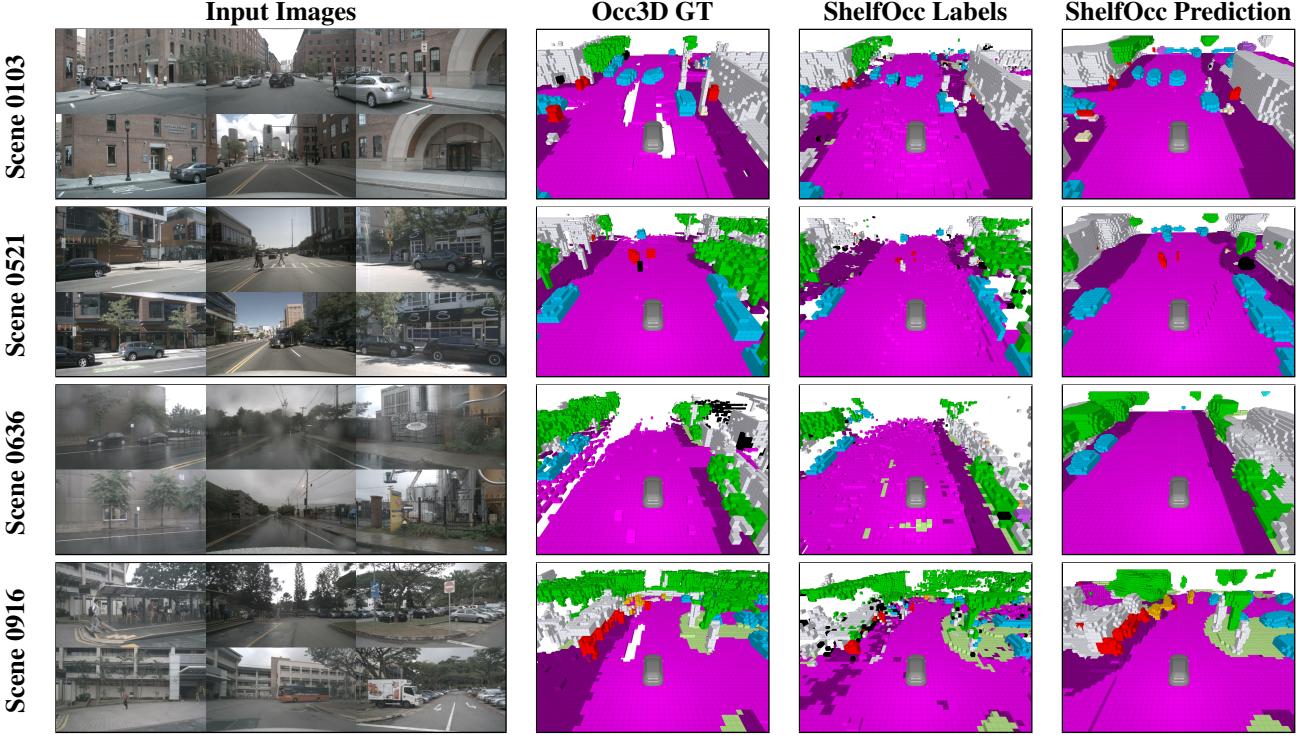


Figure 3. **Qualitative results on the Occ3D-nuScenes dataset.** We show the images, ground truth occupancy, the *ShelfOcc* pseudo-labels and the predictions of *ShelfOcc* + STCOcc [34]. Best viewed when zoomed in.

**Table 6. Effect of scene accumulation and dynamic object handling.** We show results when training STCOcc on *ShelfOcc* pseudo-labels and not aggregating the point clouds across multiple frames or without handling dynamic objects during accumulation.

Temporal Aggregation	Dynamic object handling	mIoU	IoU
✗	✗	19.29	48.22
✓	✗	21.70	55.84
✓	✓	22.87	56.14

particularly in geometric IoU, by reinforcing consistent static scene structures. However, this includes practically unacceptable model violations for dynamic objects. Finally, incorporating explicit dynamic object handling achieves an additional relative gain of approximately 5% in mIoU, indicating that accurate modeling of dynamic objects, despite their rarity, is essential for autonomous driving.

## 5. Conclusion and Future Work

In this paper, we introduced ShelfOcc, a novel training framework for shelf-supervised occupancy estimation. We presented a comprehensive framework to generate high-quality, metrically-scaled 3D semantic voxel pseudo-labels by leveraging a 3D geometry foundation model (MapAny-

thing [25]) in conjunction with a 2D semantic foundation model (GroundedSAM [45]). Our generation pipeline addresses challenges posed by large dynamic scenes, enabling the creation of precise 3D supervisory signals. By training different occupancy networks with these 3D pseudo-labels, we demonstrated substantial performance improvements over previous shelf-supervised methods, notably closing the gap to methods trained with full LiDAR supervision. This work marks a significant milestone in making high-performance 3D occupancy estimation more scalable and accessible for real-world autonomous driving applications.

Future work could focus on further enhancing the quality of the generated pseudo-labels, particularly for rare or fine-grained categories. Current models struggle with long-tail classes that are underrepresented or visually ambiguous, which can lead to incomplete or noisy semantic occupancy. Another important avenue is the improved modeling of dynamic objects. In the current framework, dynamic elements are reconstructed based on single-frame observations, which limits the completeness of their 3D shapes. Integrating motion cues such as optical flow, scene flow, or multi-frame object tracking could enable temporal alignment and accumulation of dynamic object geometry across time. This would not only produce more complete dynamic reconstructions but also extend shelf-supervised 3D labeling toward temporally consistent 4D scene understanding.

## References

- [1] Simon Boeder and Benjamin Risse. Occflownet: Occupancy estimation via differentiable rendering and occupancy flow. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 306–316. IEEE, 2025. 3
- [2] Simon Boeder, Fabian Gigengack, and Benjamin Risse. 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)*, pages 24943–24954, 2025. 2, 3, 6, 7, 1
- [3] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Lanocc: Open vocabulary occupancy estimation via volume rendering. In *2025 International Conference on 3D Vision (3DV)*, pages 200–210. IEEE, 2025. 3
- [4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liang, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 5, 6
- [5] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 3
- [6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 3
- [7] Dubing Chen, Jin Fang, Wencheng Han, Xinjing Cheng, Junbo Yin, Chengzhong Xu, Fahad Shahbaz Khan, and Jianbing Shen. Alocc: Adaptive lifting-based 3d semantic occupancy and cost volume-based flow predictions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4156–4166, 2025. 3
- [8] Dubing Chen, Huan Zheng, Yucheng Zhou, Xianfei Li, Wenlong Liao, Tao He, Pai Peng, and Jianbing Shen. Semantic causality-aware vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24878–24888, 2025. 3
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6
- [10] Tuo Feng, Wenguan Wang, and Yi Yang. Gaussian-based world model: Gaussian priors for voxel-based occupancy prediction and future motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 25239–25249, 2025. 3
- [11] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuScenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2):3795–3802, 2022. 6
- [12] Wanshui Gan, Ningkai Mo, Hongbin Xu, and Naoto Yokoya. A simple attempt for 3d occupancy estimation in autonomous driving. *arXiv preprint arXiv:2303.10076*, 2023. 3
- [13] Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. *arXiv preprint arXiv:2408.11447*, 2024. 3, 6, 7
- [14] Simon Gebräad, Andras Palffy, and Holger Caesar. Leap: Consistent multi-domain 3d labeling using foundation models. *arXiv preprint arXiv:2502.03901*, 2025. 3
- [15] Seamie Hayes, Ganesh Sistu, and Ciarán Eising. Easyocc: 3d pseudo-label supervision for fully self-supervised semantic occupancy prediction models, 2025. 3, 6
- [16] Adrian Hayler, Felix Wimbauer, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. S4c: Self-supervised semantic scene completion with neural fields. In *2024 International Conference on 3D Vision (3DV)*, pages 409–420. IEEE, 2024. 3
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [18] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 3
- [19] Nan Huang, Xiaobao Wei, Wenzhao Zheng, Pengju An, Ming Lu, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. S3gaussian: Self-supervised street gaussians for autonomous driving. *arXiv preprint arXiv:2405.20323*, 2024. 3
- [20] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. *arXiv preprint arXiv:2311.12754*, 2023. 2, 3, 6
- [21] 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. 3
- [22] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2405.17429*, 2024. 3
- [23] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. *arXiv preprint arXiv:2306.15670*, 2023. 3
- [24] Haoyi Jiang, Liu Liu, Tianheng Cheng, Xinjie Wang, Tianwei Lin, Zhizhong Su, Wenyu Liu, and Xinggang Wang. Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding. *arXiv preprint arXiv:2412.13193*, 2024. 3, 6
- [25] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 2, 3, 4, 7, 8

- [26] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2
- [27] Mehar Khurana, Neehar Peri, James Hays, and Deva Ramanan. Shelf-supervised cross-modal pre-training for 3d object detection. *arXiv preprint arXiv:2406.10115*, 2024. 2
- [28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3, 4
- [29] Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*, 2024. 3
- [30] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 3
- [31] Peizheng Li, Shuxiao Ding, You Zhou, Qingwen Zhang, Onat Inak, Larissa Triess, Niklas Hanselmann, Marius Cordts, and Andreas Zell. Ago: Adaptive grounding for open world 3d occupancy prediction. *arXiv preprint arXiv:2504.10117*, 2025. 3, 1
- [32] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9087–9098, 2023. 3
- [33] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 3
- [34] Zhimin Liao, Ping Wei, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. Stcocc: Sparse spatial-temporal cascade renovation for 3d occupancy and scene flow prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1516–1526, 2025. 3, 6, 7, 8, 1, 2
- [35] Haisong Liu, Yang Chen, Haiguang Wang, Zetong Yang, Tianyu Li, Jia Zeng, Li Chen, Hongyang Li, and Limin Wang. Fully sparse 3d occupancy prediction, 2024. 3, 6, 7
- [36] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 3, 4
- [37] Hao Lu, Tianshuo Xu, Wenzhao Zheng, Yunpeng Zhang, Wei Zhan, Dalong Du, Masayoshi Tomizuka, Kurt Keutzer, and Yingcong Chen. Drivingrecon: Large 4d gaussian reconstruction model for autonomous driving. *arXiv preprint arXiv:2412.09043*, 2024. 3
- [38] Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. Octreeocc: Efficient and multi-granularity occupancy prediction using octree queries. *arXiv preprint arXiv:2312.03774*, 2023. 3
- [39] Junyi Ma, Xieyanli 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. 3
- [40] Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19936–19945, 2024. 3, 6, 7
- [41] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2
- [42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3
- [43] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. *arXiv preprint arXiv:2309.09502*, 2023. 3
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [45] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2, 3, 4, 8
- [46] Yining Shi, Kun Jiang, Jiusi Li, Zelin Qian, Junze Wen, Mengmeng Yang, Ke Wang, and Diange Yang. Grid-centric traffic scenario perception for autonomous driving: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 1
- [47] Yiang Shi, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Xinggang Wang. Occupancy as set of points. In *European Conference on Computer Vision*, pages 72–87. Springer, 2025. 3
- [48] Sophia Sirkó-Galouchenko, Alexandre Boulch, Spyros Gidaris, Andrei Bursuc, Antonin Vobecky, Patrick Pérez, and Renaud Marlet. Occfeat: Self-supervised occupancy feature prediction for pretraining bev segmentation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4493–4503, 2024. 3
- [49] Rui Song, Chenwei Liang, Yan Xia, Walter Zimmer, Hu Cao, Holger Caesar, Andreas Festag, and Alois Knoll. Coda4dgs: Dynamic gaussian splatting with context and deformation awareness for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28031–28041, 2025. 3

- [50] Qianpu Sun, Changyong Shu, Sifan Zhou, Zichen Yu, Yan Chen, Dawei Yang, and Yuan Chun. Gsrender: Deduplicated occupancy prediction via weakly supervised 3d gaussian splatting. *arXiv preprint arXiv:2412.14579*, 2024. 3
- [51] Xin Tan, Wenbin Wu, Zhiwei Zhang, Chaojie Fan, Yong Peng, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Geocc: Geometrically enhanced 3d occupancy network with implicit-explicit depth fusion and contextual self-supervision. *arXiv preprint arXiv:2405.10591*, 2024. 3
- [52] Zhiyu Tan, Zichao Dong, Cheng Zhang, Weikun Zhang, Hang Ji, and Hao Li. Ovo: Open-vocabulary occupancy. *arXiv preprint arXiv:2305.16133*, 2023. 3
- [53] Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15035–15044, 2024. 3
- [54] Levente Tempfli, Esteban Rivera, and Markus Lienkamp. Vespa: Towards un (human) supervised open-world point-cloud labeling for autonomous driving. *arXiv preprint arXiv:2507.20397*, 2025. 2
- [55] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*, 2023. 6, 1
- [56] 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. 3
- [57] Antonin Vobecky, Oriane Siméoni, David Hurých, Spyridon Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Pop-3d: Open-vocabulary 3d occupancy prediction from images. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [58] Jiabao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, Jie Yang, Wei Liu, Qibin Hou, and Mingming Cheng. Opus: occupancy prediction using a sparse set. In *Advances in Neural Information Processing Systems*, 2024. 3
- [59] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual semantic grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 3
- [60] 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. 3
- [61] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3
- [62] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 3
- [63] Huaiyuan Xu, Junliang Chen, Shiyu Meng, Yi Wang, and Lap-Pui Chau. A survey on occupancy perception for autonomous driving: The information fusion perspective. *Information Fusion*, 114:102671, 2025. 1
- [64] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision*, pages 156–173. Springer, 2024. 3
- [65] Baijun Ye, Minghui Qin, Saining Zhang, Moonjun Gong, Shaoting Zhu, Hao Zhao, and Hang Zhao. Gs-occ3d: Scaling vision-only occupancy reconstruction with gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 25925–25937, 2025.
- [66] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8839–8848, 2021. 2
- [67] Zhangchen Ye, Tao Jiang, Chenfeng Xu, Yiming Li, and Hang Zhao. Cvt-occ: Cost volume temporal fusion for 3d occupancy prediction. *arXiv preprint arXiv:2409.13430*, 2024. 3, 6, 7
- [68] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 2
- [69] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023. 3
- [70] Zhu Yu, Bowen Pang, Lizhe Liu, Runmin Zhang, Qihao Peng, Maochun Luo, Sheng Yang, Mingxia Chen, Si-Yuan Cao, and Hui-Liang Shen. Language driven occupancy prediction. *arXiv preprint arXiv:2411.16072*, 2024. 3
- [71] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *arXiv preprint arXiv:2312.09243*, 2023. 2, 3, 6
- [72] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 3
- [73] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2304.05316*, 2023. 3
- [74] Linqing Zhao, Xiuwei Xu, Ziwei Wang, Yunpeng Zhang, Borui Zhang, Wenzhao Zheng, Dalong Du, Jie Zhou, and

- Jiwen Lu. Lowrankocc: Tensor decomposition and low-rank recovery for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9806–9815, 2024. 3
- [75] Jilai Zheng, Pin Tang, Zhongdao Wang, Guoqing Wang, Xiangxuan Ren, Bailan Feng, and Chao Ma. Veon: Vocabulary-enhanced occupancy prediction. In *European Conference on Computer Vision*, pages 92–108. Springer, 2025. 3, 1
- [76] 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, 2024. 3
- [77] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 3
- [78] Xiaoyu Zhou, Jingqi Wang, Yongtao Wang, Yufei Wei, Nan Dong, and Ming-Hsuan Yang. Autoocc: Automatic open-ended semantic occupancy annotation via vision-language guided gaussian splatting. *arXiv preprint arXiv:2502.04981*, 2025. 3, 1

# ShelfOcc: Native 3D Supervision beyond LiDAR for Vision-Based Occupancy Estimation

## Supplementary Material

### A. Additional Qualitative Results

We provide further qualitative comparisons to complement the results in the main paper. First, we compare STCOcc [34] trained on our pseudo-labels against the previous state-of-the-art method GaussianFlowOcc [2], as well as the Occ3D-nuScenes ground truth [55] in Fig. A.1. For each scene, we show the three front-facing camera images alongside 3D predictions rendered from an elevated third-person viewpoint behind the ego vehicle for a single frame. STCOcc produces clean, dense, and well-regularized occupancy predictions with minimal noise or depth bleeding, whereas GaussianFlowOcc exhibits pronounced artifacts stemming from its 2D supervision pipeline. It is clearly visible that the model trained with our framework can correctly estimate the 3D shape of objects, while previous work suffers from depth bleeding.

We also visualize the effect of the different *versions* of our proposed pipeline introduced in Fig. 1, rendered from a top-down viewpoint in Fig. A.2. The naïve single-frame variant (version 1) yields sparse and incomplete geometry, missing large portions of the scene. Aggregating all points across the sequence without distinguishing motion (version 2) introduces object trails and leads to missing objects when low-confidence points are filtered out, both of which degrade the supervision quality. In contrast, our final design (version 3), which explicitly separates static and dynamic content and applies confidence filtering only to the static scene, produces a dense, coherent scene without trails or missing objects.

### B. Additional Quantitative Results

#### B.1. Comparison to LiDAR-Supervised Methods

Table B.1 provides an extended comparison between our approach and methods that rely on LiDAR for supervision, including weakly supervised approaches AGO [31] and VEON [75] using LiDAR data, as well as fully supervised methods trained with semantic 3D ground truth. Interestingly, both COTR and CVT-Occ, when trained solely on our pseudo-labels, achieve competitive performance relative to AGO and even surpass VEON, despite both of the latter relying on LiDAR for geometric supervision. STCOcc surpasses them even more clearly in terms of mIoU. Unfortunately, the authors of these methods do not report geometric IoU, where we would expect them to perform more strongly due to their access to LiDAR depth. These findings highlight that our LiDAR-free, shelf-supervised framework

can match or even outperform prior methods that depend on LiDAR supervision for geometry. At the same time, there remains a performance gap compared to fully supervised methods trained directly on densely annotated LiDAR voxel labels. While LiDAR-based occupancy estimation is not the focus of this work, we provide these numbers to contextualize the remaining room for improvement relative to full 3D supervision. We omit AutoOcc [78] from this comparison, as we were unable to retrace their differing evaluation protocol.

#### B.2. Per-Class Semantic Segmentation Ablation

We additionally report per-class performance for the semantic segmentation ablation (cf. Tab. 3) in Tab. B.2. The results confirm that the proposed sky grounding technique substantially improves the quality of the pseudo-labels across almost all classes. By reducing spurious false positives from the open-vocabulary detector the resulting labels become significantly cleaner and more stable. Notably, improvements are pronounced for low-frequency classes such as *bus*, *traffic cone*, and *truck*. For these categories, sky grounding prevents the detector from erroneously predicting object boxes in every frame, enabling more accurate class assignments and reducing confusion with the background.

#### B.3. Ablation on Resolution and Backbone

We further investigate the impact of input image resolution and backbone capacity on models trained with our ShelfOcc pseudo-labels in Tab. B.3. For this study, we use CVT-Occ as the representative architecture. Our results show that CVT-Occ benefits noticeably from scaling both the backbone and the input resolution. Doubling the resolution to  $512 \times 1408$  and replacing the ResNet-50 encoder with a larger ResNet-101 already yields a clear improvement in semantic mIoU. Increasing the resolution further to the full nuScenes input size leads to additional gains, improving not only mIoU but also geometric IoU. We also experimented with VoVNet-99, which has a parameter count comparable to ResNet-101. While its semantic mIoU is similar to that of ResNet-101, it achieves slightly lower geometric IoU, suggesting that encoder architecture plays a nontrivial role in exploiting the pseudo-label supervision. Overall, these findings indicate that our 3D supervision pipeline can effectively leverage higher-resolution inputs and stronger backbones, offering additional room for performance scaling.

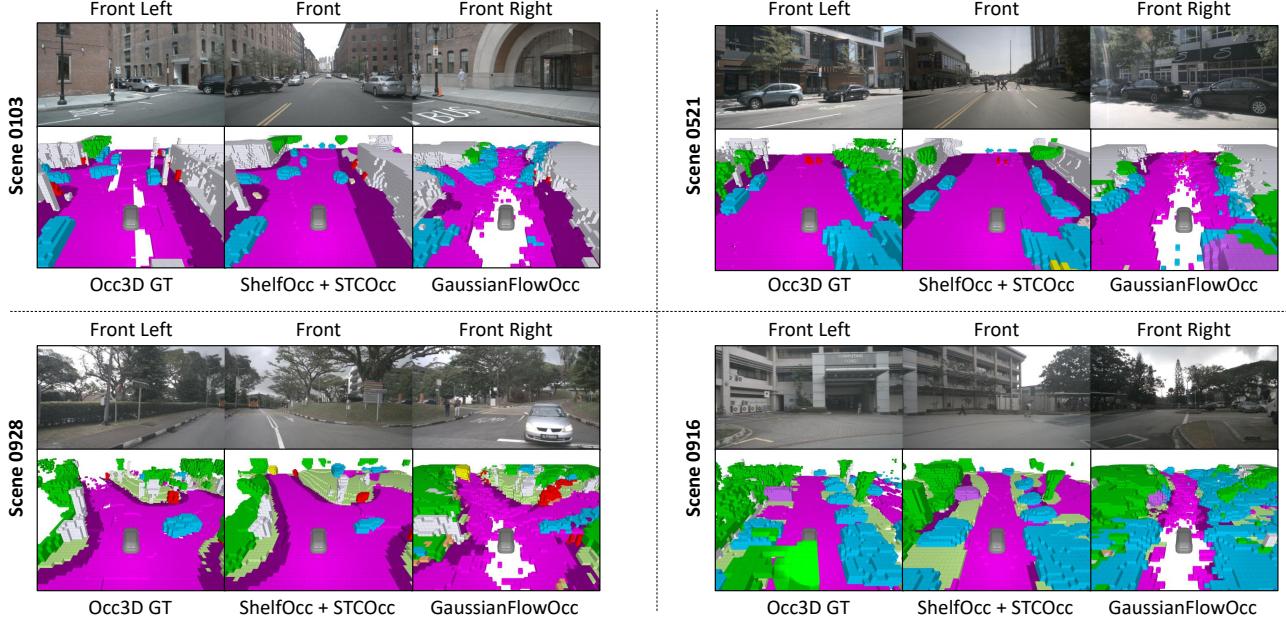


Figure A.1. **Qualitative comparison with previous state-of-the-art.** We show predictions from STCOcc [34] trained on our ShelfOcc pseudo-labels, compared against GaussianFlowOcc [2] and the Occ3D-nuScenes ground truth. STCOcc produces cleaner and more geometrically consistent occupancy predictions, demonstrating the benefits of our 3D supervision.

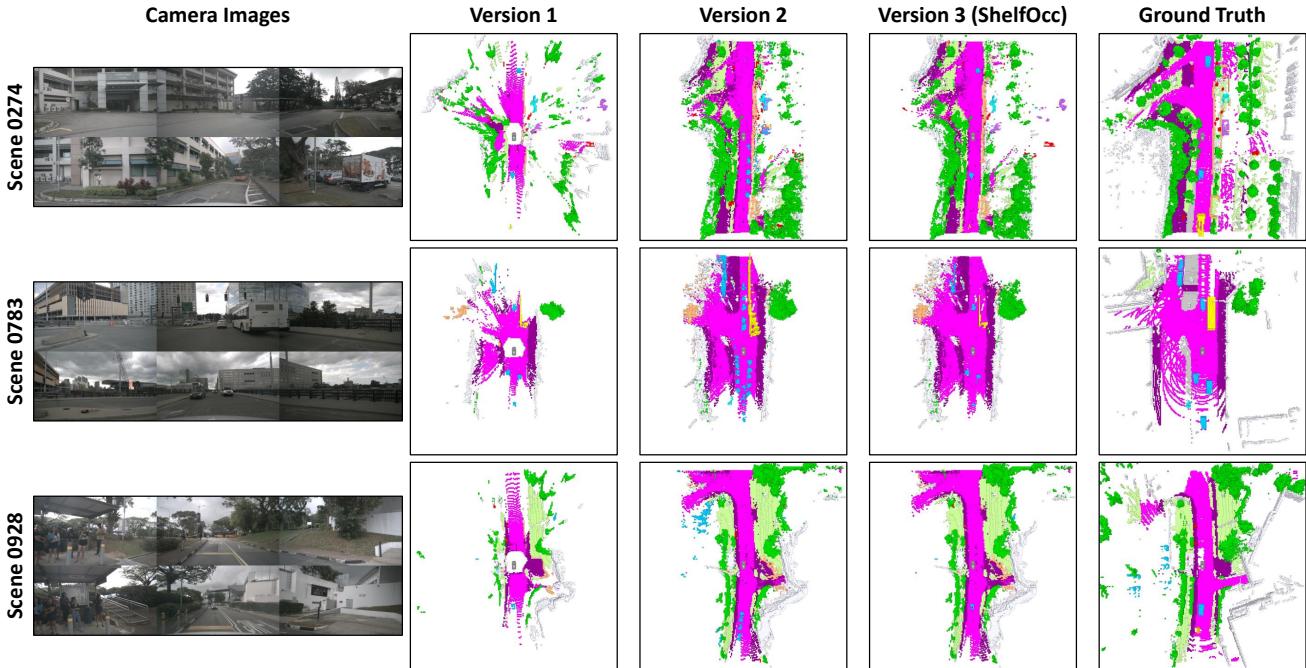


Figure A.2. **Qualitative comparison of the different versions of our proposed pipeline.** We visualize pseudo-labels produced by the three pipeline variants introduced in the main paper: (1) the naïve single-frame approach, (2) full temporal aggregation without handling motion, and (3) our final design, which aggregates static geometry while treating dynamic objects separately. The comparison highlights how version 3 avoids sparsity, object trails, and missing objects, resulting in clean and coherent 3D supervision.

**Table B.1. Performance on the Occ3D-nuScenes validation set compared to methods trained with LiDAR data.** The *LiDAR* column indicates whether a method uses raw 3D LiDAR points during training, while the *Annotations* column denotes methods that rely on semantically annotated LiDAR ground truth (e.g., voxel labels from Occ3D-nuScenes). Despite using only camera images for supervision, our shelf-supervised pipeline outperforms prior methods that depend on LiDAR-based geometric supervision.

Method	LiDAR	Annotations			barrier	bicycle	bus	car	cons. vehicle	motorcycle	pedestrian	traffic cone	trailer	truck	driv. surf.	sidewalk	terrain	manmade	vegetation
			mIoU	IoU															
AGO [31]	✓	✗	21.39	-	6.75	6.43	14.00	22.82	5.57	16.66	13.20	6.80	10.53	15.89	71.48	34.48	41.37	29.33	25.66
VEON [31]	✓	✗	17.07	-	10.40	6.20	17.70	12.70	8.50	7.60	6.50	5.50	8.20	11.80	54.50	25.50	30.20	25.40	25.40
CVT-Occ [67]	✓	✓	42.36	-	49.46	23.57	49.18	55.63	23.10	27.85	28.88	29.07	34.97	40.98	81.44	51.37	54.25	45.94	39.71
COTR [40]	✓	✓	46.41	75.01	52.11	31.95	46.03	55.63	32.57	32.78	30.35	34.09	37.72	41.84	84.48	57.55	60.67	51.99	46.33
STCOcc [34]	✓	✓	46.83	-	52.3	32.2	50.5	56.5	31.7	33.9	33.4	33.8	38.9	44.9	83.9	57.1	60.1	50.6	42.7
Ours: ShelfOcc + COTR [40]	✗	✗	18.65	53.71	9.10	6.20	22.92	22.08	1.66	5.94	9.92	8.55	0.0	15.32	67.93	31.13	38.76	23.11	17.15
Ours: ShelfOcc + CVT-Occ [67]	✗	✗	19.21	52.72	11.53	6.38	20.39	21.92	4.20	10.18	9.02	10.67	0.89	13.08	68.42	31.23	41.42	22.74	16.15
Ours: ShelfOcc + STCOcc [34]	✗	✗	22.87	56.14	13.98	11.36	25.27	25.80	7.25	16.61	12.91	13.42	5.37	17.15	68.01	34.66	42.73	25.63	22.89

**Table B.2. Effect of improved semantic segmentation.** We train STCOcc [34] on our pseudo-labels with and without using the sky grounding technique. Using the improved semantic segmentation also improves downstream occupancy estimation performance.

Method	Sky Grounding			barrier	bicycle	bus	car	cons. vehicle	motorcycle	pedestrian	traffic cone	trailer	truck	driv. surf.	sidewalk	terrain	manmade	vegetation
		mIoU	IoU															
Ours: <i>ShelfOcc</i>	✗	6.21	17.21	3.7	1.93	3.81	5.57	1.55	3.27	3.95	5.39	0.10	2.87	20.17	9.27	10.46	8.21	12.91
Ours: <i>ShelfOcc</i>	✓	9.62	26.00	6.58	3.28	7.02	8.81	2.57	4.74	4.97	8.1	0.12	5.41	34.59	14.58	18.16	10.91	14.45
Ours: <i>ShelfOcc</i> + STCOcc [34]	✗	22.48	54.26	11.96	9.89	23.64	25.88	7.89	15.24	13.31	14.43	7.73	16.84	67.64	35.51	40.13	24.43	22.67
Ours: <i>ShelfOcc</i> + STCOcc [34]	✓	22.87	56.14	13.98	11.36	25.27	25.80	7.25	16.61	12.91	13.42	5.37	17.15	68.01	34.66	42.73	25.63	22.89

**Table B.3. Ablation of Image Backbones for CVT-Occ.** Comparison of ResNet variants and VoVNet while simultaneously scaling image resolution along with backbone size. We observe further performance increases when scaling up the model size and image resolution.

Backbone	Image Size	mIoU	IoU
ResNet-50	256 x 704	19.21	52.72
ResNet-101	512 x 1408	19.78	52.77
ResNet-101	928 x 1600	20.24	53.02
VoVNet-99	928 x 1600	20.23	51.84

## C. Details on Semantic Segmentation

We provide the full vocabulary used to query the open-vocabulary detector Grounding DINO [36] in Tab. C.4. As described in the main paper, each category is queried individually by forwarding a prompt of the form “*QUERY* . *sky*” through the model. Including the background token *sky* encourages the detector to identify sky regions explicitly, which in turn reduces false positives for the target query. For each forward pass, we discard all predicted boxes corresponding to *sky* and retain only the boxes associated with the target query. To ensure high-quality detections, we further filter out any box whose predicted logit falls below 0.2. All remaining boxes across all query categories are aggregated and passed to the SAM segmentation model, which generates a mask for each box. The logit of the originating Grounding DINO detection is assigned to every pixel within the corresponding SAM mask. To construct the final semantic segmentation result, we overlay all predicted masks and perform per-pixel selection based on the highest associated logit. This produces a dense, open-vocabulary segmentation result that serves as the semantic input to our pseudo-label generation pipeline.

Table C.4. **Vocabulary used for querying Grounding DINO [36] for open-vocabulary object detection.** The left column lists the class labels, and the right column contains the prompts used during mask generation.

Class	Prompts
'barrier'	'barricade', 'barrier'
'bicycle'	'bicycle'
'bus'	'bus'
'car'	'car', 'sedan', 'van'
'construction_vehicle'	'excavator', 'crane'
'motorcycle'	'motorcycle', 'scooter'
'pedestrian'	'person', 'pedestrian'
'traffic_cone'	'traffic-cone'
'trailer'	'trailer'
'truck'	'lorry', 'truck'
'driveable_surface'	'highway', 'street', 'roadmarking'
'sidewalk'	'sidewalk', 'walkway'
'terrain'	'turf', 'grass', 'sand'
'manmade'	'building', 'wall', 'fence', 'pole', 'sign', 'light', 'bridge', 'billboard'
'vegetation'	'bush', 'plants', 'tree'