Semantic Causality-Aware Vision-Based 3D Occupancy Prediction

Dubing Chen[1], Huan Zheng[1], Yucheng Zhou[1],
Xianfei Li[2], Wenlong Liao[2], Tao He[2], Pai Peng[2], Jianbing Shen[1]✉
[1]SKL-IOTSC, CIS, University of Macau [2]COWAROBOT Co. Ltd.
https://github.com/cdb342/CausalOcc

Abstract

Vision-based 3D semantic occupancy prediction is a critical task in 3D vision that integrates volumetric 3D reconstruction with semantic understanding. Existing methods, however, often rely on modular pipelines. These modules are typically optimized independently or use pre-configured inputs, leading to cascading errors. In this paper, we address this limitation by designing a novel causal loss that enables holistic, end-to-end supervision of the modular 2D-to-3D transformation pipeline. Grounded in the principle of 2D-to-3D semantic causality, this loss regulates the gradient flow from 3D voxel representations back to the 2D features. Consequently, it renders the entire pipeline differentiable, unifying the learning process and making previously non-trainable components fully learnable. Building on this principle, we propose the Semantic Causality-Aware 2D-to-3D Transformation, which comprises three components guided by our causal loss: Channel-Grouped Lifting for adaptive semantic mapping, Learnable Camera Offsets for enhanced robustness against camera perturbations, and Normalized Convolution for effective feature propagation. Extensive experiments demonstrate that our method achieves state-of-the-art performance on the Occ3D benchmark, demonstrating significant robustness to camera perturbations and improved 2D-to-3D semantic consistency.

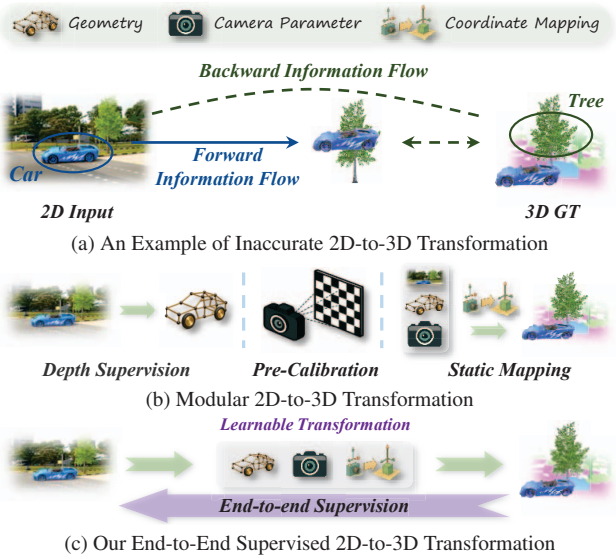


Figure 1. **(a) Illustrates a Visual Analysis of Semantic Ambiguity in VisionOcc.** Inaccurate 2D-to-3D transformations may lead to positional shifts, misaligning supervision signals and resulting in semantic ambiguity. **(b) Depicts the Conventional Modular 2D-to-3D Transformation Paradigm** [7, 13, 27, 35], which employs depth supervision for geometry estimation, pre-calibrated camera parameters, and fixed mapping for lifting. **(c) Presents Our Holistic, End-to-End Supervised 2D-to-3D Transformation Paradigm**, which eliminates the need for separate modular supervision or pre-calibration, enabling unified error propagation to supervise all components.

1. Introduction

Predicting dense 3D semantic occupancy is a fundamental task in 3D vision, providing a fine-grained voxel representation of scene geometry and semantics [39, 40, 43, 47]. The challenge of performing this prediction from vision alone, an approach known as vision-based 3D semantic occupancy prediction (VisionOcc), has recently become a focal point of research. By leveraging only commodity cameras, VisionOcc is pivotal for a wide range of 3D applications, serving as a comprehensive digital replica of the environment for tasks like analysis, simulation, and interactive visualization [7, 9, 14, 27, 39, 40, 44].

VisionOcc unifies the challenges of feed-forward 3D reconstruction and dense semantic understanding. Existing pipelines typically decompose this task into two metaphases [7, 14, 27, 30, 35]. The initial 2D-to-3D transfor-

✉ Corresponding author: *Jianbing Shen*. This work was supported in part by the Science and Technology Development Fund of Macau SAR (FDCT) under grants 0102/2023/RIA2 and 0154/2022/A3 and 001/2024/SKL and CG2025-IOTSC, the University of Macau SRG2022-00023-IOTSC grant, and the Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC).

mation uses large receptive field operators like view lifting [7, 37] or cross attention [14, 27] to construct an initial 3D feature volume. Subsequently, a 3D representation learning phase employs operators with a small receptive field (*e.g.*, 3D convolutions or local self-attention) to refine 3D features and produce the final prediction. Our work targets the 2D-to-3D transformation phase, which is more critical and error-prone. Fig. 1a showcases a primary failure mode where features of one class (*e.g.*, a 2D 'car') are erroneously transformed to the 3D location of another (*e.g.*, a 'tree'). This creates a flawed learning objective, forcing the model to learn a spurious association between 'car' features and a 'tree' label. Such *semantic ambiguity* is a principal obstacle to achieving high performance.

The prevailing VisionOcc methods, typically based on Lift-Splat-Shoot (LSS), employ a modular approach for the 2D-to-3D transformation (Fig. 1b) [7, 35, 37]. This involves supervising geometry with a proxy depth loss while relying on fixed, pre-calibrated camera parameters and a static lifting map. However, this modularity raises critical questions about robustness and optimality. First, it is susceptible to compounding errors; for example, the reliance on fixed camera parameters makes the system vulnerable to real-world perturbations like camera jitter during motion. More fundamentally, the optimality of such proxy supervision is questionable. An intermediate representation ideal for depth estimation may not be optimal for the final semantic occupancy task, inherently limiting the transformation's expressive power due to this objective misalignment. This motivates our central research question: **Can we devise an end-to-end supervision framework**[1] that holistically optimizes the entire 2D-to-3D transformation, enabling unified semantic-aware error backpropagation and allowing traditionally fixed modules to become fully learnable?

We approach this problem from a causal perspective. In VisionOcc, the 2D image semantics are the "cause" of the final 3D semantic "effect". Semantic misalignment arises from disrupted information flow from cause to effect (Fig. 1a). Therefore, instead of correcting the erroneous output, we propose to directly regularize the information flow itself. We posit that a 3D prediction for a given class should be influenced predominantly by 2D image regions of that same class. To enforce this information flow, we leverage gradients as a proxy, inspired by prior work [18, 42, 53]. For each semantic class, the gradient of its aggregated 3D features is computed *w.r.t.* the 2D feature map, producing a saliency-like map of 2D influence. This map is then directly supervised with the ground truth 2D segmentation mask. As shown in Fig. 1c, this establishes a principled, end-to-end supervision signal for the 2D-to-3D transformation, enabling holistic optimization of all its components.

[1]Here, "end to end" refers to a unified supervision scheme for a process, distinct from the concept of a monolithic network architecture.

To fully leverage our end-to-end supervision, we introduce a more expressive and learnable 2D-to-3D view transformation, termed the Semantic Causality-Aware Transformation (SCAT). A key challenge is that directly supervising gradients is inherently unstable. Therefore, the entire SCAT module is designed to constrain its gradient flow to a stable [0, 1] range. Specifically, SCAT introduces three targeted designs: *i) Channel-Grouped Lifting:* To better disentangle semantics, we move beyond LSS's uniform weighting and apply distinct learnable weights to different groups of feature channels. *ii) Learnable Camera Offsets:* To mitigate motion-induced pose errors, we introduce learnable offsets to the camera parameters, which are implicitly supervised by the 2D-3D semantic consistency enforced by our causal loss. *iii) Normalized Convolution:* Finally, we employ a normalized convolution to densify the sparse 3D features from LSS [28], ensuring this final step also adheres to our global gradient stability requirement.

Our contributions are as follows: **i)** We systematically analyze the 2D-to-3D transformation in VisionOCC, identifying a critical failure mode we term semantic ambiguity. We provide a theoretical analysis proving how the modularity of prior methods leads to error propagation, offering clear guidance for future work. **ii)** To address these problems, we propose the Causal Loss that directly regularizes the information flow of the 2D-to-3D transformation. This enables true end-to-end optimization of all constituent modules, mitigating error accumulation and making previously fixed components, such as camera parameters, fully learnable. **iii)** We instantiate our principles in the Semantic Causality-Aware Transformation, a novel 2D-to-3D transformation architecture. SCAT incorporates Channel-Grouped Lifting, Learnable Camera Offsets, and Normalized Convolution to explicitly tackle the challenges of semantic confusion, camera perturbations, and limited learnability. **iv)** Extensive experiments show our method significantly boosts existing models, achieving a 3.2% absolute mIoU gain on BEVDet. Furthermore, it demonstrates superior robustness to camera perturbations, reducing the relative performance drop on BEVDet from a severe -32.2% to a mere -7.3%.

2. Related Work

2.1. Semantic Scene Completion

Semantic Scene Completion (SSC) [10, 20, 24, 31, 51] refers to the task of simultaneously predicting both the occupancy and semantic labels of a scene. Existing methods can be classified into indoor and outdoor approaches based on the scene type, with the former focusing on occupancy and semantic label prediction in controlled environments [10, 20, 31], while the latter shifts towards more complex outdoor settings, particularly in the context of au-

tonomous driving [1, 6]. The core principle of SSC lies in its ability to infer the unseen, effectively bridging gaps in incomplete observations with accurate semantic understanding. MonoScene [6] introduces a 3D SSC framework that infers dense geometry and semantics from a single monocular RGB image. VoxFormer [24] presents a Transformer-based semantic scene completion framework that generates complete 3D volumetric semantics from 2D images by first predicting sparse visible voxel queries and then densifying them through self-attention with a masked autoencoder design. OccFormer [51] introduces a dual-path transformer network for 3D semantic occupancy prediction, efficiently processing camera-generated 3D voxel features through local and global pathways, and enhancing occupancy decoding with preserve-pooling and class-guided sampling to address sparsity and class imbalance.

2.2. Vision-based 3D Occupancy Prediction

Vision-based 3D Occupancy Prediction [9, 14–16, 34, 38–40, 43**?**] aims to predict the spatial and semantic features of 3D voxel grids surrounding an autonomous vehicle from image data. This task is closely related to SSC, emphasizing the importance of multi-perspective joint perception for effective autonomous navigation. TPVFormer [14] is prior work that lifts image features into the 3D TPV space by leveraging an attention mechanism [26, 41]. Different from TPVFormer, which relies on sparse point clouds for supervision, subsequent studies, including OccNet [40], SurroundOcc [44], Occ3D [39], and OpenOccupancy [43], have developed denser occupancy annotations by incorporating temporal information or instance-level labels. Methods such as BEVDet [13], FBOcc [27], COTR [35], and ALOcc [7] leverage depth-based LSS [22, 23, 25, 37] for explicit geometric transformation, demonstrating strong performance. Some methods [2, 4, 17, 36, 46, 49, 52] have explored rendering-based methods that utilize 2D signal supervision, thereby bypassing the need for 3D annotations. Furthermore, recent research like [3, 7, 8, 21, 29, 32, 40] introduced 3D occupancy flow prediction, which addresses the movement of foreground objects in dynamic scenes by embedding 3D flow information to capture per-voxel dynamics. Unlike the above methods, we analyze the 2D-to-3D transformation process from the perspectives of error propagation and semantic causal consistency, proposing a novel approach that enhances causal consistency.

3. Method

Preliminary. VisionOcc predicts a dense 3D semantic occupancy grid from surround-view images, modeled as a causal dependency chain as shown in Fig. 2. This process involves key variables: input image $\mathbf{I}$, estimated geometry $\mathbf{G}$, camera parameters P (intrinsic & extrinsic), potential camera parameter errors e_P, intermediate 3D feature $\mathbf{L}$, and final occupancy output $\mathbf{O}$. In LSS-based methods [7, 13, 35], the pipeline starts with the image backbone extracting features $\mathbf{f}_i = F_i(\mathbf{I}), \mathbf{f}_i \in \mathbb{R}^{U,V,C}$; geometry $\mathbf{G} \in \mathbb{R}^{U,V,D}$ is predicted as a probability distribution over discretized depth bins D, *i.e.*, $\mathbf{G} = F_g(\mathbf{f}_i)$; 2D features are then transformed to 3D via an outer product $\mathbf{f}'_L = \mathbf{G} \otimes \mathbf{f}_i, \mathbf{f}'_L \in \mathbb{R}^{U,V,D,C}$; camera parameters P map these to voxel coordinates $R_P(u,v,d) \to (h,w,z) \in [0, H-1] \times [0, W-1] \times [0, Z-1]$, yielding $\mathbf{f}_L \in \mathbb{R}^{H \times W \times Z}$, where $H \times W \times Z$ defines the occupancy grid resolution; finally, $\mathbf{L}$ is decoded to produce the semantic occupancy output: $\mathbf{O} = F_o(\mathbf{f}_L), \mathbf{O} \in \mathbb{R}^{H \times W \times Z \times S}$, where F_o is the decoding function and S is the number of semantic classes. The prediction $\mathbf{O}$ is supervised by the ground-truth $\tilde{\mathbf{O}}$.

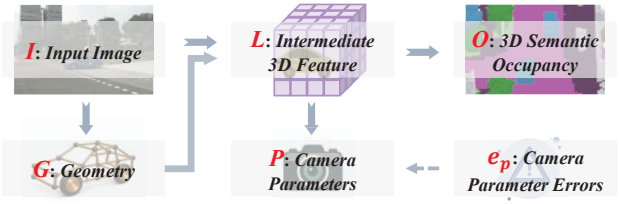


Figure 2. **The Causal Structure of VisionOcc.** It illustrates the dependency chain from the image input $\mathbf{I}$ to the semantic occupancy output $\mathbf{O}$. $\mathbf{G}$: geometry for 2D-to-3D transformation. P: camera intrinsic and extrinsic. $\mathbf{L}$: intermediate 3D feature. e_P: errors in camera parameters.

3.1. Error Propagation in Depth-Based LSS

Theorem 1. *In Depth-Based LSS methods with a fixed 2D-to-3D mapping M_{fixed}, inherent mapping error δM leads to gradient deviations, preventing convergence to an ϵ-optimal solution. This is formalized as:*

$$M_{fixed} = M_{ideal} + \delta M \implies \nabla_\theta L_{LSS} \neq \nabla_\theta L_{ideal}, \quad (1)$$

where M_{ideal} is the ideal mapping and L_{ideal} is the loss function using M_{ideal}.

Proof. **Mapping Error Quantification:** Let $\mathbf{x} \in \mathbb{R}^2$ be 2D pixel coordinates and $d(\mathbf{x})$ be the ground truth depth. Estimated depth is $\hat{d}(\mathbf{x}) = d(\mathbf{x}) + \epsilon_d(\mathbf{x})$, with $\epsilon_d(\mathbf{x})$ as depth error. Let $\mathbf{K}_{ideal} \in \mathbb{R}^{3\times3}$ be the ideal camera intrinsics, and $\mathbf{K} = \mathbf{K}_{ideal} + \epsilon_K$ be the estimated camera intrinsics with error ϵ_K. The fixed mapping is $M_{fixed} = M_{ideal} + \delta M$, where δM encompasses errors from various sources, including depth estimation error $\epsilon_d(\mathbf{x})$ and camera extrinsic error ϵ_K. We assume the total mapping error is bounded $\|\delta M\|_F \leq \Delta_M < \infty$. The 3D coordinates are:

$$\mathbf{X} = M_{fixed}(\mathbf{x}, \hat{d}(\mathbf{x}), \mathbf{K}) \quad (2)$$

$$\mathbf{X}_{ideal} = M_{ideal}(\mathbf{x}, d(\mathbf{x}), \mathbf{K}_{ideal}) \quad (3)$$

$$\Delta\mathbf{X} = \mathbf{X} - \mathbf{X}_{ideal} = \delta M(\mathbf{x}, \hat{d}(\mathbf{x}), \mathbf{K}), \quad (4)$$

where $\|\Delta\mathbf{X}\|_2 \leq C \cdot \Delta_M$ for bounded inputs.

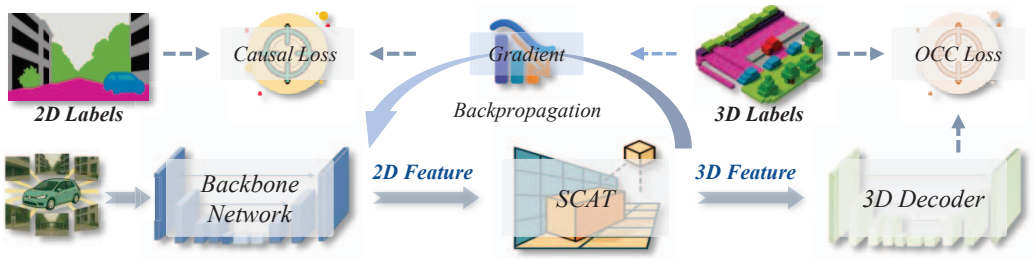


Figure 3. **The Overall Framework of the Proposed Semantic Causality-Aware VisionOcc.** The proposed framework consists of three primary components: a backbone network for extracting 2D features, an SCAT module for transforming these features into 3D space, and an Encoder-Decoder network for learning 3D semantics. The SCAT module is supervised by our causal loss.

Method	mIoU	mIoU_D	IoU
Depth-Based LSS	44.5	40.4	78.9
SCL-Aware LSS	50.5↑6.0	46.9↑6.5	85.7↑6.8

Table 1. **Performance of Depth-Based LSS *vs*. SCL-Aware LSS on Occ3D (in Ideal Conditions).** BEVDetOcc is the baseline.

Feature Space Deviation and Loss: Let $F_{2D}(\mathbf{x})$ be 2D features and $F_{3D}(\cdot) = Lift(F_{2D}(\mathbf{x}), \cdot)$. Assuming $Lift$ is L_{Lift}-Lipschitz continuous, the 3D feature deviation is:

$$\begin{aligned}\|F_{3D}(\mathbf{X}) - F_{3D}(\mathbf{X}_{ideal})\|_F &\leq L_{Lift}\|\mathbf{X} - \mathbf{X}_{ideal}\|_2 \\ &\leq L_{Lift}\|\Delta\mathbf{X}\|_2 \leq \Delta_{F_{3D}}.\end{aligned} \quad (5)$$

The loss function is $L_{LSS} = \mathcal{L}(P_{3D}(\mathbf{X}), GT_{3D})$, where $P_{3D}(\mathbf{X}) = Seg_{3D}(F_{3D}(\mathbf{X}))$.

Gradient Deviation and Optimization Limit: The gradient of L_{LSS} *w.r.t.* parameters θ is given by chain rule. However, since M_{fixed} is fixed, $\frac{\partial \mathbf{X}}{\partial \theta} = 0$. Thus, the gradient becomes:

$$\nabla_\theta L_{LSS} = \frac{\partial L_{LSS}}{\partial P_{3D}} \frac{\partial P_{3D}}{\partial F_{3D}} \left(\frac{\partial F_{3D}}{\partial F_{2D}} \frac{\partial F_{2D}}{\partial \theta} \right). \quad (6)$$

Due to $\Delta_{F_{3D}} > 0$, the computed gradient $\nabla_\theta L_{LSS}$ is based on the deviated feature space, i.e.,

$$\begin{aligned}\nabla_\theta L_{LSS}(\theta) &= \nabla_\theta \mathcal{L}(P_{3D}(\mathbf{X}), GT_{3D}) \\ &\neq \nabla_\theta \mathcal{L}(P_{3D}(\mathbf{X}_{ideal}), GT_{3D}) = \nabla_\theta L_{ideal}(\theta).\end{aligned} \quad (7)$$

The gradient deviation prevents mapping's direct optimization and limits convergence to an ϵ-optimal solution. □

The theoretical analysis reveals that the inherent error in the fixed 2D-to-3D mapping of Depth-Based LSS methods fundamentally hinders gradient-based optimization from achieving optimal performance.

3.2. Semantic Causal Locality in VisionOcc

As revealed in our theoretical analysis (Sec. 3.1), Depth-Based LSS methods suffer from inherent error propagation due to their fixed 2D-to-3D mapping, which limits optimization efficacy and potential performance. To overcome these limitations, we particularly focus on strengthening the semantic causality of the 2D-to-3D transformation. We argue that VisionOcc's 2D-to-3D semantic occupancy prediction should exhibit *semantic causal locality (SCL)* for robust perception in autonomous driving. Ideally, 2D causes should drive 3D semantic effects. For instance, a predicted "car" at 3D location (h, w, z) should originate from a matching 2D image region of "car". Per the causal chain, camera parameters P and estimated geometry $\mathbf{G}$ enable this dependency, with $\mathbf{G}$ being crucial to maintain SCL.

Next, we formulate the ideal SCL condition. For a 2D pixel (u, v) with semantic label s, the projection probability p_d (the value of $\mathbf{G}$ at this coordinate and depth $d \in D$) should be high if its corresponding 3D ground-truth semantic is s, and low otherwise:

$$p_d \propto \mathbb{1}(\tilde{\mathbf{O}}(R_P(u, v, d) + e_P) = s),$$

where $\mathbb{1}$ is the indicator function, and e_P represents the potential coordinate transformation error caused by factors such as camera pose error. In practice, p_d acts as a weight multiplied by 2D features (Eq. (8)), enabling a probabilistic mapping that supports differentiable backpropagation.

Limitations of Depth-Based LSS in SCL. Depth-Based LSS does not fully account for semantic causal consistency. It only preserves semantic causal locality intuitively under ideal conditions where $e_P = 0$ and depth estimation is perfectly accurate. However, with coordinate transformation errors e_P, even a high p_d for the ideal depth may project 2D semantics to incorrect 3D locations, *i.e.*,

$$\tilde{\mathbf{O}}(R_P(u, v, d) + 0) = s, \text{ but } \tilde{\mathbf{O}}(R_P(u, v, d) + e_P) \neq s.$$

This misalignment causes 2D semantics s (*e.g.*, "car") to link with wrong 3D semantics (*e.g.*, "tree"), causing semantic ambiguity and hindering training. Moreover, depth-based LSS often propagates semantics to surface points, weakening semantic propagation to occluded regions [7].

Empirical Validation. We conduct an empirical study to validate our analysis of SCL, comparing two ideal geometric transformations. Using BEVDetOcc [13] as the baseline, we replace its estimated depth-based geometry for LSS

with: *i)* ground-truth LiDAR depths; *ii)* SCL-Aware geometry, which computes p_d from 2D and 3D semantic ground truths, where for (u, v, d, s), $p_d = 1$ if $\tilde{\mathbf{O}}(R_P(u, v, d)) = s$, else $p_d = 0$. We evaluate their performance in semantic occupancy prediction. Tab. 1 demonstrates that 2D-to-3D Transformation achieves significant performance improvements over depth-based LSS in ideal conditions, validating the benefits of semantic causal locality.

Summary. Building on the above analysis, we propose our solution to enforce semantic causality constraints during training. The overall framework is shown in Fig. 3.

3.3. Semantic Causality-Aware Causal Loss

For lifting methods like LSS, we could directly supervise the transformation geometry $\mathbf{G}$ using the causal semantic geometry derived from ground-truth labels (as described in the previous section). However, we aim to enhance the lifting method in the next section, rendering direct supervision impractical. Thus, we design a gradient-based approach to enforce semantic causality.

We begin within the LSS framework. For a 2D pixel feature at location (u, v), LSS multiplies it by the depth-related transformation probability p_d and projects it to the 3D coordinate corresponding to depth d:

$$\mathbf{f}_L(R_P(u, v, d)) = p_d(u, v, d) \cdot \mathbf{f}_i(u, v, d). \tag{8}$$

Here, $\mathbf{f}_L(R_P(u, v, d)) \in \mathbb{R}^C$ represents the 3D voxel feature at the projected location, with e_P omitted for notational simplicity. $\mathbf{f}_i(u, v, d) \in \mathbb{R}^C$ denotes the 2D image feature at location (u, v). We backpropagate gradients from $\mathbf{f}_L$ to $\mathbf{f}_i$, *i.e.*,

$$\frac{\partial \sum \mathbf{f}_L(R_P(u, v, d))}{\partial \mathbf{f}_i(u, v, d)} = p_d \cdot \mathbf{I}, \tag{9}$$

where $\mathbf{I}$ is the all-ones vector.

For each semantic class s, we aggregate the features $\mathbf{f}_L$ at all 3D positions where the ground truth class equals s. This aggregation is backpropagated to the 2D features $\mathbf{f}_i$, yielding a gradient map $\nabla_s \in \mathbb{R}^{U \times V \times C}$ for class s:

$$\nabla_s(u, v, c) = \sum_{(h', w', z') \in \Omega_s} \frac{\partial \sum \mathbf{f}_L(h', w', z', c)}{\partial \mathbf{f}_i(u, v, c)}, \tag{10}$$

where $\Omega_s = \{(h', w', z') \mid O(h', w', z') = s\}$ is the set of 3D positions with semantic label s, and c indexes the feature channels. Averaging over channel C produces an attention map $A_s \in \mathbb{R}^{U \times V}$:

$$A_s(u, v) = \frac{1}{C} \sum_{c=1}^{C} \nabla_s(u, v, c). \tag{11}$$

Finally, we enforce per-pixel constraints using 2D ground-truth labels with a binary cross-entropy loss:

$$\begin{aligned} L^s_{bce} = -\frac{1}{U \cdot V} \sum_{u,v} \big[& Y_s(u, v) \log A_s(u, v) \\ & + (1 - Y_s(u, v)) \log(1 - A_s(u, v)) \big], \end{aligned} \tag{12}$$

where $Y_s(u, v) \in \{0, 1\}$ is the 2D ground-truth label for semantic class s at pixel (u, v).

The gradient computation can be performed using the automatic differentiation calculator like *torch.autograd*, requiring S backward passes to iterate over the S semantic classes. This incurs a computational overhead scaling linearly with the class amount S. To mitigate this overhead, we reformulate the loss computations in terms of the expectation of an unbiased estimator [11]. We define the expected BCE loss across all semantic classes $s \in \{1, \ldots, S\}$ as:

$$\mathbb{E}[L^s_{bce}] = \frac{1}{S} \sum_{s=1}^{S} L^s_{bce}. \tag{13}$$

Based on this relationship, we uniformly sample a single semantic class s during training:

$$L_{causal} = L^s_{bce}, \quad s \sim \text{Uniform}(1, S). \tag{14}$$

This sampling preserves the unbiased nature of the gradient and loss estimates and reduces the computational cost to $\frac{1}{S}$. L_{causal} focuses on enhancing the geometric transformation, serving as an auxiliary term to complement the occupancy loss (case-by-case, *e.g.*, cross-entropy in BEVDet [13]).

3.4. Semantic Causality-Aware Transformation

We propose semantic causality-aware 2D-to-3D transformation to enhance 2D-to-3D lifting, as shown in Fig. 4. Eq. (14) constrains the geometry $\mathbf{G}$ using 2D and 3D semantics. This overcomes the rigid, per-location hard alignment of geometric probabilities (*e.g.*, using LiDAR depth supervision) in prior methods [7, 13, 22, 23, 27]. It enables advanced lifting designs, addressing errors from camera pose and other distortions.

3.4.1. Channel-Grouped Lifting

Vanilla LSS applies uniform weights to all 2D feature channels. We argue this is trivial as 2D and 3D features have distinct locality biases. For instance, a 2D "car" edge may capture "tree" semantics via convolution, but in 3D, these objects are distant. Uniform weighting both semantics causes ambiguity. Since different channels typically encode distinct semantics, we group the feature channels and learn unique weights for each group:

$$\mathbf{f}_{L,g}(R_P(u, v, d)) = \omega_{g,d} \cdot \mathbf{f}_{i,g}(u, v, d),\ g \in \{1, \ldots, N_g\}, \tag{15}$$

where $\mathbf{f}_{L,g} \in \mathbb{R}^{C/N_g}$ and $\mathbf{f}_{i,g} \in \mathbb{R}^{C/N_g}$ are the 3D and 2D features for group g. $\omega_{g,d}$ is the learned weight for group

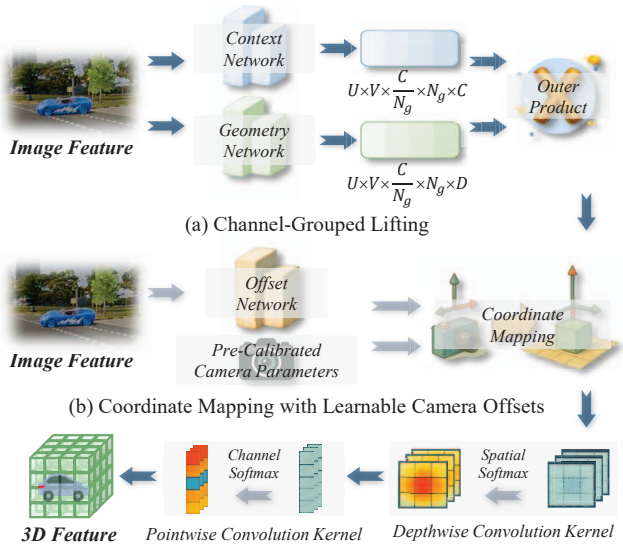


Figure 4. **The Detailed Structure of the Proposed Modules in Semantic Causality-Aware 2D-to-3D Transformation.** There are three key components, including (a): Channel-Grouped Lifting, (b): Learnable Camera Offsets, and (c): Normalized Convolution, enabling accurate and robust 2D-to-3D transformation.

g, replacing p_d which uniformly lifts all channels. N_g is the number of groups. This preserves semantic distinction, ensuring channel-specific causal alignment.

3.4.2. Learnable Camera Offsets

To address camera parameter errors, especially pose inaccuracies, we introduce learnable offsets into camera parameters. First, we ensure the lifting process is coordinate-differentiable. This is crucial for the offsets to receive gradients and adapt during training. The transformation of 2D image coordinates (u, v) and depth d to 3D voxel coordinates can be represented as matrix multiplication:

$$[h, w, z]^T = P \cdot [u \cdot d, v \cdot d, d, 1]^T, \tag{16}$$

where $P \in \mathbb{R}^{3\times4}$ is the camera projection matrix combining intrinsics and extrinsics. LSS typically rounds floating-point coordinates to integers, rendering them non-differentiable. Following ALOcc [7], we use the soft filling to enable differentiability *w.r.t.* position. This method calculates distances between floating-point 3D coordinates and their eight surrounding integer coordinates. These distances serve as weights to distribute a 2D feature at (u, v, d) across multiple 3D locations. Lifting can be rewritten to

$$\begin{aligned}\mathbf{f}_{L,g}(h', w', z') = \omega_{g,d} \cdot \omega_{h',w',z'} \cdot \mathbf{f}_{i,g}(u, v, d),\\ \forall (h', w', z') \in \text{neighbors},\end{aligned} \tag{17}$$

where $\omega_{h',w',z'}$ are trilinear interpolation weights.

Next, we propose learning two offsets. First, we directly predict an offset applied to the camera parameters:

$$P := P + \Delta P, \quad \Delta P = F_{offset1}(\mathbf{f}_i, P), \tag{18}$$

where ΔP is the predicted parameter offset, and $F_{offset1}$ denotes the network. Second, we estimate per-position offsets for each (u, v, d) in the image coordinate system:

$$\begin{aligned}(u, v, d) := (u + \Delta u, v + \Delta v, d + \Delta d),\\ (\Delta u, \Delta v, \Delta d) = F_{offset2}(\mathbf{f}_i(u, v, d)),\end{aligned} \tag{19}$$

where $F_{offset2}$ is another network. These offsets enable the model to adaptively compensate for camera parameter errors e_P, improving geometric accuracy while preserving semantic causal locality under such errors.

3.4.3. Normalized Convolution

Prior work [28] notes that the direct mapping in LSS yields sparse 3D features. To address this, an intuitive solution is to use local feature propagation operators (*e.g.*, convolutions) with causal loss supervision for preserving semantic causality. However, vanilla convolutions lack gradient constraints during causal loss computation, producing poor results. We address this by normalizing convolution weights to keep gradient map values in [0, 1]. Given the difficulty of normalizing standard convolution weights, we follow MobileNet and ConvNext to adopt depthwise (spatial) and pointwise (channel) decomposition. For the depthwise kernel $W_{\text{spatial}} \in \mathbb{R}^{3\times3\times3\times C}$, we apply softmax across spatial dimensions (h, w, z) per channel c:

$$W'_{\text{spatial}}[h,w,z,c] = \frac{\exp(W_{\text{spatial}}[h,w,z,c])}{\sum_{h',w',z'} \exp(W_{\text{spatial}}[h',w',z',c])}. \tag{20}$$

For the pointwise kernel $W_{\text{channel}} \in \mathbb{R}^{C\times C}$, we apply softmax across input channels per output channel:

$$W'_{\text{channel}}[c_{\text{in}}, c_{\text{out}}] = \frac{\exp(W_{\text{channel}}[c_{\text{in}}, c_{\text{out}}])}{\sum_{c'_{\text{out}}} \exp(W_{\text{channel}}[c_{\text{in}}, c'_{\text{out}}])}. \tag{21}$$

Specifically, we use transposed convolution for depthwise convolutions. It diffuses features from non-zero to zero positions, but the zero position does not affect others. We prove in the *supplement* that the derived gradient mask remains within [0, 1] even with our semantic causality-aware 2D-to-3D transformation.

3.5. Validation of Gradient Error Mitigation

Fig. 5 shows the occupancy loss curves during training for BEVDet, comparing performance with and without our method. The results show that BEVDet integrated with our approach (blue line) achieves a significantly faster and steeper loss reduction compared to the original BEVDet (red line). This empirical evidence validates our theoretical analysis of gradient error mitigation. As Theorem Theorem 1 formalizes, Depth-Based LSS methods are inherently limited by gradient deviations due to fixed 2D-to-3D

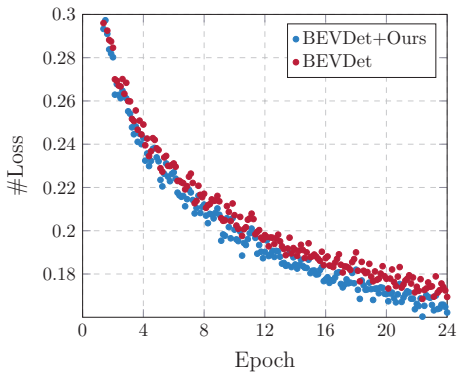


Figure 5. **Occupancy Loss Curves with/without Our Method.** Our method reduces training loss through semantic causal 2D-to-3D geometry transformation, as shown in the comparison before **(red)** and after **(blue)** its application.

Method	mIoU ↑	Drop	mIoU_D↑	Drop	IoU↑	Drop
BEVDetOcc [12]	37.1	-32.3%	30.2	-49.0%	70.4	-8.5%
BEVDetOcc *w/ Noise*	25.1		15.4		64.4	
BEVDetOcc+Ours	38.3	**-7.3%**	31.5	**-10.8%**	71.2	**-1.4%**
BEVDetOcc+Ours *w/ Noise*	35.5		28.1		70.2	
ALOcc [7]	40.1	-21.9%	34.3	-28.6%	70.2	-8.4%
ALOcc *w/ Noise*	31.3		24.5		64.3	
ALOcc+Ours	40.9	**-3.3%**	35.5	**-4.8%**	70.7	**-1.0%**
ALOcc+Ours *w/ Noise*	39.6		33.8		70.0	

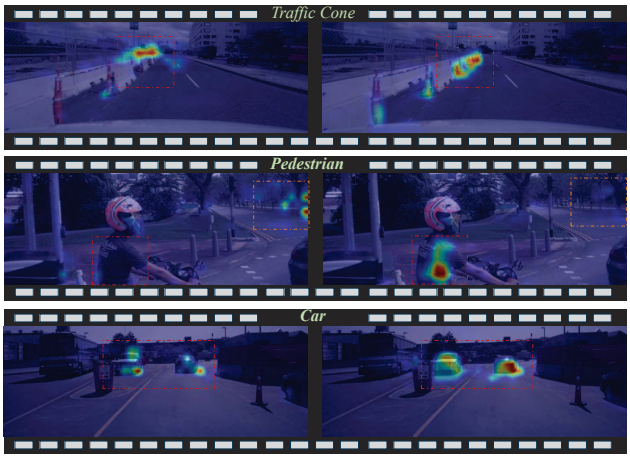
Table 2. **Performance Comparison on the Occ3D Dataset with Gaussian Noise Added to Camera Parameters.** The "Drop (%)" columns show degradation, with our methods (BEVDetOcc+Ours, ALOcc+Ours) achieving much smaller drops (*e.g.*, -7.3% mIoU vs. -32.4% for BEVDetOcc).

mapping errors. In contrast, our method aims to alleviate this by enabling a learnable mapping and incorporating a causal loss. It indicates that by mitigating gradient error through semantic causality-aware 2D-to-3D transformation, our approach facilitates more efficient gradient-based learning, leading to faster convergence and a lower loss curve.

4. Experiment

4.1. Experimental Setup

Dataset. In this study, we leverage the Occ3D-nuScenes dataset [5, 39], a comprehensive dataset with diverse scenes for autonomous driving research. It encompasses 700 scenes for training, 150 for validation, and 150 for testing. Each scene integrates a 32-beam LiDAR point cloud alongside 6 RGB images, captured from multiple perspectives encircling the ego vehicle. Occ3D [39] introduces voxel-based annotations, covering a spatial extent of -40 m to 40 m along the X and Y axes, and -1 m to 5.4 m along the Z axis, with a consistent voxel size of 0.4 m across all dimensions. Occ3D delineates 18 semantic categories, comprising 17 distinct object classes plus an *empty* class to signify unoccupied regions. Following [7, 27, 35, 39], we evaluate the occupancy prediction performance with mIoU across 17 semantic object categories, mIoU_D for 8 dynamic categories, and occupied/unoccupied IoU for scene geometry.



(a) BEVDet (b) BEVDet + Ours

Figure 6. **Visualization of 2D-to-3D Semantic Causal Consistency Using LayerCAM [18].** We enhance BEVDet with our method for comparison. Attention maps are computed for critical traffic classes: "traffic cone", "pedestrian", and "car". Areas of greatest difference are marked with boxes. Each class-specific localization highlights our method's precise focus over vanilla BEVDet, showing improved semantic alignment.

Implementation Details. We integrate our approach into BEVDet [12] and ALOcc [7] for performance evaluation. The model parameters, image, and BEV augmentation strategies are retrained as the original. For ALOcc, we remove its ground-truth depth denoising module to adapt to our method. We use multiple convolutional layers to predict the two camera parameter offsets. The loss weight of L_{causal} is set to 0.02 in the main experiments. We optimize using AdamW [33] with a learning rate of 2×10^{-4}, a global batch size of 16, and for 24 epochs. Experiments use single-frame surrounding images, emphasizing improvements from enhanced 2D-to-3D transformations.

4.2. Evaluation of Camera Perturbation Robustness

To assess our method's robustness under extreme noise, we add Gaussian noise with (0.1 variances) to camera parameters in training and testing, as shown in Tab. 2. Compared to vanilla BEVDetOcc and ALOcc, our methods show smaller performance drops. BEVDetOcc+Ours has a mIoU drop of -7.3% *vs.*-32.3% for BEVDetOcc, and ALOcc+Ours drops -3.3% *vs.*-21.9% for ALOcc, demonstrating enhanced resilience to noisy parameters. This effectively counters motion-induced errors, benefiting self-driving tasks.

4.3. Semantic Causality Visualization

As shown in the Fig. 6, we visualize 3D-to-2D semantic causal consistency using LayerCAM [18]. We backpropagate the final 3D semantic occupancy predictions of dis-

Method	Backbone	Input Size	mIoU	$mIoU_D$	IoU
MonoScene [6]	ResNet-101	928 × 1600	6.1	5.4	-
CTF-Occ [39]	ResNet-101	928 × 1600	28.5	27.4	-
TPVFormer [14]	ResNet-101	928 × 1600	27.8	27.2	-
COTR [35]	ResNet-50	256 × 704	39.1	33.8	69.6
ProtoOcc [19]	ResNet-50	256 × 704	39.6	34.3	-
LightOcc-S [50]	ResNet-50	256 × 704	37.9	32.4	-
DHD-S [45]	ResNet-50	256 × 704	36.5	30.7	-
FlashOCC [48]	ResNet-50	256 × 704	32.0	24.7	65.3
FB-Occ [27]	ResNet-50	256 × 704	35.7	30.9	66.5
BEVDetOcc [12]	ResNet-50	256 × 704	37.1	30.2	70.4
BEVDetOcc+Ours	ResNet-50	256 × 704	**38.3** ↑1.2	**31.5** ↑1.3	**71.2** ↑1.2
ALOcc [7]	ResNet-50	256 × 704	40.1	34.3	70.2
ALOcc+Ours	ResNet-50	256 × 704	**40.9** ↑0.8	**35.5** ↑1.1	**70.7** ↑0.5

Table 3. **Comparison of 3D Semantic Occupancy Prediction Using Single Frame on the Occ3D Dataset, Evaluated mIoU, $mIoU_D$, and IoU Metrics.** Performance gains are indicated by red arrows ↑. Our proposed approach (**+Ours**) consistently demonstrates superior enhancement over existing methods.

tinct classes to the 2D feature maps fed into SCAT. Notably, we are the first to apply LayerCAM for cross-dimensional analysis. Fig. 6 (b) shows our method precisely focuses on class-specific locations. This confirms LayerCAM's cross-dimensional effectiveness. Our approach surpasses vanilla BEVDet (Fig. 6 (a)) in targeting class-associated objects. This proves improved semantic localization.

4.4. Benchmarking with Previous Methods

As shown in Tab. 3, we compare our method with leading 3D semantic occupancy prediction approaches on Occ3D [39]. Specifically, compared to baseline models BEVDet and ALOcc, our method achieves significant improvements in mIoU, $mIoU_D$, and IoU. For instance, the **BEVDetOcc+Ours** variant achieves an mIoU of **38.3**, surpassing BEVDetOcc [12] by **1.2**, while improving $mIoU_D$ by **1.3** and IoU by **1.2**. Similarly, **ALOcc+Ours** shows gains of **0.8** in mIoU, **1.1** in $mIoU_D$, and **0.5** in IoU over ALOcc [7]. These results validate the superiority of our semantic causality-aware 2D-to-3D transformation.

4.5. Ablation Study

Effect of Causal Loss. We first investigate the effectiveness of the proposed Causal Loss in enhancing the occupancy prediction performance. As demonstrated in Tab. 4, the impact of the proposed Causal Loss is evaluated through a series of experiments. Specifically, using BEVDetOcc as the baseline (**Exp. 0**), we conducted two ablation studies: one removing the depth supervision loss (**Exp. 1**), and another incorporating the proposed Causal Loss (**Exps. 2, 3**). The results reveal that the removal of the depth supervision loss leads to marginal decrease in performance, whereas the addition of the proposed Causal Loss yields a significant improvement. This suggests that Causal Loss facilitates superior 2D-to-3D transformation and ultimately enhances the

Exp.	Method	mIoU	Diff.	$mIoU_D$	IoU	Latency
0	Baseline (BEVDetOcc) [12]	37.1	-	30.2	70.4	416/125
1	*w/o* Depth Sup	36.8	-0.3	29.6	70.3	414/125
2	+ ***Causal Loss***	37.6	+0.8	31.0	70.1	450/125
3	+ ***Unbiased Estimator***	37.5	-0.1	30.7	70.5	417/125
4	*w/o* Post Conv	37.3	-0.2	30.7	70.2	379/122
5	+ ***Channel-Grouped Lifting***	37.6	+0.3	30.7	70.7	419/128
6	+ Soft Filling	37.6	-	30.6	70.7	434/149
7	+ ***Learnable Camera Offset***	37.9	+0.3	31.1	71.0	446/150
8	+ ***Normalized Convolution***	**38.3**	+0.4	**31.5**	**71.2**	466/159

Table 4. **Ablation Study of 3D Semantic Occupancy Prediction on Occ3D.** We comprehensively evaluated the impact of the individual strategy (**bolded rows**) proposed in our paper, with BEVDetOcc as the baseline. The final column reports single-frame training/inference latency (ms) on an RTX 4090 GPU.

precision of 3D semantic occupancy prediction. Comparing **Exp. 3** to **Exp. 2**, the Unbiased Estimator simplifies the Causal Loss computation, reducing training overhead.

Effect of Each Module. In Tab. 4 (**Exps. 5-8**), we systematically validate the effectiveness of the three proposed modules. We first remove the two post-lifting convolutional layers from the original BEVDet (**Exp. 4**), as they serve a similar role to our proposed modules in refining volume features. Subsequently, we incrementally integrate the three proposed modules (Channel-Grouped Lifting, Learnable Camera Offsets, and Normalized Convolution) into the baseline model. To ensure gradient flow for Learnable Camera Offsets, we introduce the Soft Filling strategy from ALOcc [7] in **Exp. 6**, enabling effective training of the camera parameter offsets in **Exp. 7**. The results show progressive performance improvements with each added component (**Exps. 5, 7, 8**), confirming the efficacy of the proposed SCAT method. Additionally, the proposed modules incur acceptable computational overhead.

Please refer to the supplementary material for more comprehensive experimental results.

5. Conclusion

In this paper, we introduced a novel approach leveraging causal principles to address the limitation that ignored the reliability and interpretability by existing methods. By exploring the causal foundations of 3D semantic occupancy prediction, we propose a causal loss that enhances semantic causal consistency. In addition, we develop the SCAT module with three main components: Channel-Grouped Lifting, Learnable Camera Offsets and Normalized Convolution. This approach effectively mitigates transformation inaccuracies arising from uniform mapping weights, camera perturbations, and sparse mappings. Experiments demonstrate that our approach achieves significant improvements in accuracy, robustness to camera perturbations, and semantic causal consistency in 2D-to-3D transformations.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 3

[2] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Langocc: Self-supervised open vocabulary occupancy estimation via volume rendering. *arXiv preprint arXiv:2407.17310*, 2024. 3

[3] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Occflownet: Towards self-supervised occupancy estimation via differentiable rendering and occupancy flow. *arXiv preprint arXiv:2402.12792*, 2024. 3

[4] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Gaussianflowocc: Sparse and weakly supervised occupancy estimation using gaussian splatting and temporal flow. *arXiv preprint arXiv:2502.17288*, 2025. 3

[5] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 7

[6] 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, 8

[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 prediction. *arXiv preprint arXiv:2411.07725*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[8] Dubing Chen, Wencheng Han, Jin Fang, and Jianbing Shen. Adaocc: Adaptive forward view transformation and flow modeling for 3d occupancy and flow prediction. *arXiv preprint arXiv:2407.01436*, 2024. 3

[9] Dubing Chen, Huan Zheng, Jin Fang, Xingping Dong, Xianfei Li, Wenlong Liao, Tao He, Pai Peng, and Jianbing Shen. Rethinking temporal fusion with a unified gradient descent view for 3d semantic occupancy prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1505–1515, 2025. 1, 3

[10] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4193–4202, 2020. 2

[11] Judy Hoffman, Daniel A Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019. 5

[12] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 7, 8

[13] 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. 1, 3, 4, 5

[14] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2302.07817*, 2023. 1, 2, 3, 8

[15] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19946–19956, 2024.

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

[17] 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

[18] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 2, 7

[19] Jungho Kim, Changwon Kang, Dongyoung Lee, Sehwan Choi, and Jun Won Choi. Protoocc: Accurate, efficient 3d occupancy prediction using dual branch encoder-prototype query decoder. *arXiv preprint arXiv:2412.08774*, 2024. 8

[20] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian Reid. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7693–7702, 2019. 2

[21] Jinke Li, Xiao He, Chonghua Zhou, Xiaoqiang Cheng, Yang Wen, and Dan Zhang. Viewformer: Exploring spatiotemporal modeling for multi-view 3d occupancy perception via view-guided transformers. In *Computer Vision–ECCV 2024: 18th European Conference*, 2024. 3

[22] 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. *arXiv preprint arXiv:2206.10092*, 2022. 3, 5

[23] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1486–1494, 2023. 3, 5

[24] 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. 2, 3

[25] Yangguang Li, Bin Huang, Zeren Chen, Yufeng Cui, Feng Liang, Mingzhu Shen, Fenggang Liu, Enze Xie, Lu Sheng, Wanli Ouyang, et al. Fast-bev: A fast and strong bird's-

eye view perception baseline. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[26] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 3

[27] 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. 1, 2, 3, 5, 7, 8

[28] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6919–6928, 2023. 2, 6

[29] Zhimin Liao and Ping Wei. Cascadeflow: 3d occupancy and flow prediction with cascaded sparsity sampling refinement framework. *CVPR 2024 Autonomous Grand Challenge Track On Occupancy and Flow*, 2024. 3

[30] Lizhe Liu, Bohua Wang, Hongwei Xie, Daqi Liu, Li Liu, Zhiqiang Tian, Kuiyuan Yang, and Bing Wang. Surroundsdf: Implicit 3d scene understanding based on signed distance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[31] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[32] Yili Liu, Linzhan Mou, Xuan Yu, Chenrui Han, Sitong Mao, Rong Xiong, and Yue Wang. Let occ flow: Self-supervised 3d occupancy flow prediction. *arXiv preprint arXiv:2407.07587*, 2024. 3

[33] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 2017. 7

[34] 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

[35] Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. *arXiv preprint arXiv:2312.01919*, 2023. 1, 2, 3, 7, 8

[36] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12404–12411. IEEE, 2024. 3

[37] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2, 3

[38] Yiang Shi, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Xinggang Wang. Occupancy as set of points. In *Computer Vision–ECCV 2024: 18th European Conference*, 2024. 3

[39] 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, 3, 7, 8

[40] 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. 1, 3

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[42] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*, 2023. 2

[43] 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. *arXiv preprint arXiv:2303.03991*, 2023. 1, 3

[44] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. *arXiv preprint arXiv:2303.09551*, 2023. 1, 3

[45] Yuan Wu, Zhiqiang Yan, Zhengxue Wang, Xiang Li, Le Hui, and Jian Yang. Deep height decoupling for precise vision-based 3d occupancy prediction. *arXiv preprint arXiv:2409.07972*, 2024. 8

[46] Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Liu Haiyang, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, et al. Renderworld: World model with self-supervised 3d label. *arXiv preprint arXiv:2409.11356*, 2024. 3

[47] Shichao Yang, Yulan Huang, and Sebastian Scherer. Semantic 3d occupancy mapping through efficient high order crfs. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 590–597. IEEE, 2017. 1

[48] 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. 8

[49] 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. 3

[50] Jinqing Zhang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Lightweight spatial embedding for vision-based 3d occupancy prediction. *arXiv preprint arXiv:2412.05976*, 2024. 8

[51] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2304.05316*, 2023. 2, 3

[52] 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, 2024. 3

[53] Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15890–15902, 2024. 2