FDR-Occ: Factorized Dense Routing for Full-Spectrum 3D Occupancy Prediction

Dubing Chen[1], Huan Zheng[1], Tianyi Yan[1], Yucheng Zhou[1], Runzhou Tao[2], Zhongying Qiu[2], Jianfei Yang[3], and Jianbing Shen[1,✉]

[1] SKL-IOTSC, CIS, University of Macau
[2] Chongqing Afari Intelligent Drive Co., Ltd.
[3] Geely Automobile Research Institute (Ningbo) Co., Ltd.

Abstract. Vision-based 3D occupancy prediction fundamentally relies on the 2D-to-3D view transformation. Current paradigms predominantly utilize explicit physical projection, which artificially restricts the routing matrix to strict, sparse camera rays. While computationally efficient, this imposes a severe *Locality Bottleneck*, preventing the network from constructing holistic contextual understanding and degrading sharply when camera extrinsics are unreliable or absent. To break this bottleneck, we abstract view transformation as unconstrained bipartite routing and propose **Factorized Dense Routing (FDR)**. By approximating dense 2D-to-3D mixing through hierarchical tensor contractions, FDR guarantees a fully-global receptive field with tractable, sub-quadratic complexity. Crucially, the mandatory spatial contraction in dense routing exposes a fundamental *Resolution-Context Trade-off*. To address this, we introduce a **Resolution-Context Decoupled Architecture**. We factorize the 3D space into a global macroscopic topological anchor (via FDR) and precise local geometric planes (via explicit projection). This decoupling enables global semantic inference and exact surface localization to complement each other without mutual compromise. Extensive experiments demonstrate that our framework achieves state-of-the-art performance on the Occ3D-nuScenes and Occ3D-Waymo benchmarks. More notably, in an uncalibrated setting where physical extrinsics are withheld, our global routing internalizes the implicit multi-camera rig topology and exhibits substantially stronger structural robustness than physical-projection baselines under the same protocol.

Keywords: 3D Occupancy Prediction · View Transformation · Dense Routing · Spatial Decoupling

1 Introduction

✉ Corresponding author: *Jianbing Shen*. This work was supported by the Science and Technology Development Fund of Macau SAR (FDCT) under grants 0134/2025/RIA2 and 0102/2023/RIA2, and the Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC).

Vision-based 3D occupancy prediction has emerged as a fundamental task for autonomous driving and robotic navigation [8, 10, 11, 26, 27, 29, 33, 34]. The core challenge lies in the 2D-to-3D view transformation (VT), which governs how perspective image features are lifted into a 3D volumetric space. Historically, the field has been dominated by methods relying on explicit physical ray priors. These can be broadly categorized into active forward-projection paradigms, represented by Lift-Splat-Shoot (LSS) family [4, 6, 13, 22] and reactive backward-querying mechanisms, leveraging 3D-to-2D cross-attention [15, 17].

Despite their widespread adoption, physical-prior-based methods are bottlenecked by a fundamental architectural flaw: **the strict ray-wise localization of receptive fields**. In these frameworks, the 2D-to-3D routing matrix is artificially constrained to an extreme sparsity pattern tightly aligned with physical lines-of-sight [17]. This rigid constraint causes *Restricted Reachability* during VT: a 2D pixel can only broadcast its features strictly along its physical ray (as shown in Figure 1), leaving massive 3D volumetric regions completely unreachable to direct image evidence. Consequently, the network fails to construct global topological understanding, splatting continuous structures (*e.g.*, holistic road layouts or heavily occluded cross-camera vehicles) as isolated fragments. Furthermore, it relies on an overly rigid assumption: the availability of perfectly precise camera parameters. When physical calibrations are noisy, perturbed, or entirely absent, this localized projection mathematically collapses.

To break this locality bottleneck, we argue that the VT operator must structurally evolve from a locally constrained physical ray toward fully global reachability. Mathematically, as we formally prove in Section 3.2, the function family defined by local physical routing is merely a strictly constrained subset of the unconstrained global routing family. Thus, a global routing operator possesses a strictly higher representational upper bound. However, solving the global routing involves explicitly computing a dense bipartite mixing matrix across massive spatial dimensions, which incurs an intractable quadratic complexity. By analogy with the architectural evolution from fully-connected networks to factorized convolutional fields, we propose **Factorized Dense Routing (FDR)**. FDR introduces a scalable structural inductive bias, approximating the dense lifting operator with a hierarchy of localized tensor operations. By progressively contracting the 2D spatial dimensions and expanding into the 3D volume, FDR guarantees a fully-global receptive field while reducing the computational burden by nearly two orders of magnitude.

While FDR theoretically breaks the bottleneck of localized routing, realizing this linear efficiency practically necessitates progressive spatial aggregation. This mathematical property exposes a fundamental **Resolution-Context Trade-off** in view transformation: expanding the global receptive field inherently entails aggregating fine-grained spatial details. Consequently, FDR excels at abstracting macroscopic global topology but prioritizes holistic scene layouts over microscopic geometric precision. To exploit unconstrained global context without sacrificing pixel-exact spatial precision, we propose a **Resolution-Context Decoupled Architecture**. We intentionally factorize the 3D representation

space into two orthogonal pathways. We configure FDR as the *Global Context Pathway*. Concurrently, we position the classical LSS as the *Local Resolution Pathway*; by circumventing spatial contraction entirely, it acts as an optimal extractor to carve precise local geometries. These two orthogonal capacities are unified on the shared Bird's-Eye-View (BEV) plane, resolving the trade-off between unconstrained global semantic inference and precise geometric boundary preservation.

We extensively evaluate our method on the Waymo Open [25] and the nuScenes datasets [1], establishing new state-of-the-art results for 3D occupancy prediction under the ResNet-50 backbone with minimal temporal frames. Crucially, we design a rigorous **uncalibrated setting** where explicit camera parameters are entirely withheld. In such scenarios where classical methods degrade sharply, our method learns to internalize the implicit rig topology, demonstrating robustness to unconstrained real-world conditions.

Our main contributions are summarized as follows:

- We formally identify the *Restricted Reachability* bottleneck inherent in explicit physical projection and characterize its structural limitation in constructing unconstrained global context.
- We propose Factorized Dense Routing, a novel global routing operator. By approximating dense 2D-to-3D mixing through hierarchical spatial tensor contractions, FDR achieves a fully-global receptive field with tractable complexity.
- To resolve the fundamental *Resolution-Context Trade-off*, we design a decoupled volumetric architecture. By combining a globally-connected Holistic Context Anchor (via FDR) with high-resolution geometric planes (via LSS), our framework captures both unconstrained macroscopic scene layouts and precise local geometric boundaries.
- Our method achieves state-of-the-art performance on the Waymo Open and the nuScenes datasets. Experimental results under a strictly uncalibrated configuration demonstrate that it can robustly infer implicit multi-camera topology even in the complete absence of explicit camera parameters.

2 Related work

2.1 Semantic Occupancy Prediction

The task of semantic occupancy prediction is fundamentally linked to Semantic Scene Completion (SSC), which focuses on generating dense 3D volumetric semantics from monocular or incomplete observations. Within the SSC domain [18,23,24,32], MonoScene [2] pioneered monocular 3D semantic completion from single RGB images. Subsequent architectures like VoxFormer [12] and OccFormer [38] then introduced mechanisms to densify sparse voxel queries and address class imbalance through dual-path transformer structures.

In modern autonomous driving research, vision-based 3D occupancy prediction has shifted toward multi-camera joint perception to predict spatial and

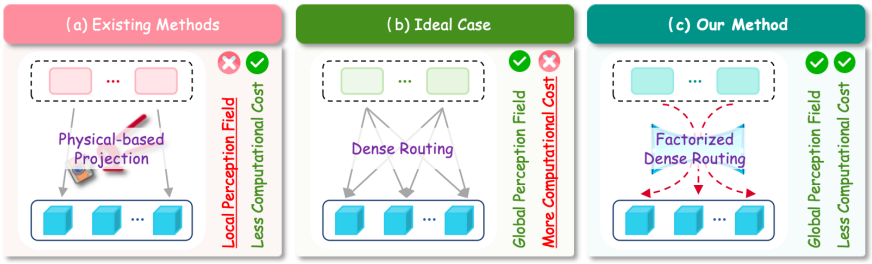

Fig. 1: Evolution of View Transformation (VT) routing. Left: Explicit physical projection [4, 22] enforces a rigid ray-wise sparsity, causing structural fragmentation. **Middle:** Ideal dense routing enables 100% reachability but suffers from an intractable quadratic complexity. **Right:** Our **Factorized Dense Routing (FDR)** structurally factorizes the dense mapping into hierarchical tensor contractions (the blue module). It achieves fully-global reachability with tractable complexity, while effectively internalizing holistic 3D topology that is unreachable via localized physical rays.

semantic features of 3D voxel grids surrounding a vehicle. Methodologies can be broadly divided into two paradigms. Backward-querying methods, exemplified by TPVFormer [8], aggregate features into a tri-perspective space via attention. Geometry-explicit methods such as BEVDet [7], FBOcc [16], COTR [19], and ALOcc [4] instead rely on the Lift-Splat-Shoot (LSS) framework for geometric transformation. To benchmark these models, datasets like Occ3D [26] provide dense, voxel-based annotations derived from temporal and instance-level data. Additionally, recent work has explored 2D rendering-based supervision [20] to bypass 3D labeling needs and occupancy flow prediction to account for per-voxel dynamics in moving foreground objects.

Despite these technical strides, traditional geometric-prior-based pipelines often encounter "semantic ambiguity" due to their reliance on modular components and fixed physical constraints. These frameworks typically employ fixed, pre-calibrated camera parameters and static mapping. Theoretical analysis indicates that inherent mapping errors in these fixed structures lead to gradient deviations that prevent the model from reaching an optimal solution. This rigid constraint results in a disruption of information flow where 2D features are erroneously projected to incorrect 3D locations. Consequently, existing methods struggle with error propagation and exhibit a significant lack of robustness when camera parameters are affected by real-world perturbations like motion jitter.

2.2 2D-to-3D View Transformation

The transformation of 2D image features into a unified 3D volumetric space is a cornerstone of vision-based 3D occupancy prediction. Current research primarily follows two distinct paradigms: bottom-up and top-down transformations.

Bottom-up methods rely on explicit depth estimation to project multi-view features into the 3D or BEV space. This paradigm, pioneered by LSS [22] and

adopted by the BEVDet [7, 13, 14] series, utilizes camera parameters to "lift" features based on estimated depth distributions. However, as noted in recent studies, the reliance on depth estimation causes inherent geometric uncertainty. Furthermore, the resulting BEV features are typically sparse, posing significant challenges for tasks that require dense spatial representations.

Top-down transformer-based methods employ learnable queries to aggregate 2D spatio-temporal features into a unified 3D space. Notable frameworks such as BEVFormer [15], VoxFormer [12], and PanoOcc [28] generate dense representations but often lack explicit geometric constraints, making them vulnerable to occlusions and depth inconsistencies. To bridge these paradigms, methods like FB-OCC [16] and COTR [19] use LSS-generated features to initialize transformer queries, among other strategies. Additionally, FlashOcc [36] explores memory-efficient occupancy prediction via a channel-to-height mechanism.

Despite these advancements, existing geometric-prior-based methods encounter specific bottlenecks in complex environments. ProtoOcc [9] identifies a Resolution-Context Trade-off: standard cross-attention degrades spatial distinctiveness when operating on low-resolution queries. It proposes a prototype-aware view transformation to provide high-level 2D contextual guidance. Concurrently, LIAR [31] highlights that most methods overlook environmental factors. Challenging nighttime conditions such as underexposure and overexposure severely degrade visual features, motivating illumination-affined representations for robust perception.

These existing frameworks, however, remain largely constrained by a ray-wise localization of receptive fields. While they attempt to enrich context within fixed projection patterns, they do not fundamentally address the Restricted Reachability inherent in explicit physical routing, where massive volumetric regions remain unreachable to direct image evidence.

3 Methodology

3.1 Problem Formulation

The core objective of vision-based 3D occupancy prediction is to transform multi-view perspective images into a dense, semantically annotated 3D volumetric grid. Given a set of surround-view images, a 2D backbone first extracts dense image-plane features $\mathcal{F}_{2D} \in \mathbb{R}^{N\times C\times H\times W}$, where N is the number of camera views, C is the feature dimension, and H, W are the spatial resolutions. The crux of the pipeline is the View Transformation module, an operator $\mathcal{T}$ that lifts the perspective features into a 3D ego-car coordinate space to form a volumetric tensor $\mathcal{V}_{3D} = \mathcal{T}(\mathcal{F}_{2D}; \Theta) \in \mathbb{R}^{C\times X\times Y\times Z}$. Here, Θ denotes the provided camera calibration parameters (intrinsics and extrinsics), and X, Y, Z represent the spatial dimensions of the 3D grid. Following the VT module, the 3D encoder and the decoder processes $\mathcal{V}_{3D}$ to output the final voxel-wise semantic probabilities $\mathcal{O} \in \mathbb{R}^{B\times X\times Y\times Z}$, where B is the number of semantic categories. The representational capacity of the view transformation operator $\mathcal{T}$ is therefore the primary bottleneck for the entire system.

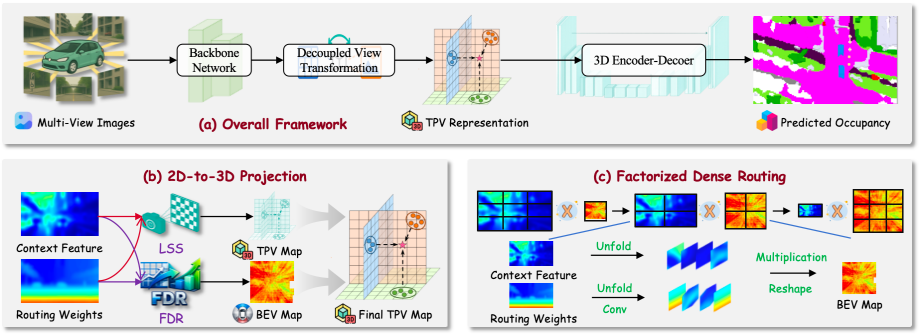

Fig. 2: Overview of our Resolution-Context Decoupled Architecture. (a) Overall Framework. (b) Decoupled 2D-to-3D Projection: To solve the spatial-contextual trade-off, the view transformation is factorized into two orthogonal pathways. Explicit LSS preserves high-frequency local geometry via orthogonal planes (TPV Map), while FDR captures unconstrained global topology via a Z-compressed BEV anchor. **(c) Factorized Dense Routing:** FDR approximates intractable dense global routing via hierarchical tensor contractions.

3.2 View Transformation as Spatial Routing: The Locality Bottleneck

To systematically analyze the representational capacity of existing VT operators, we abstract the 2D-to-3D lifting process into a generalized spatial routing framework [4]. This perspective allows us to mathematically expose the inherent locality bottlenecks of physical projection.

Let $S_0 = HW$ and $V_T = XYZ$ denote the total spatial elements in the 2D multi-view images and the 3D ego-space, respectively. The view transformation can be universally formulated as a bipartite routing:

$$\mathrm{vec}(\mathcal{V}_{3D}) = \mathbf{W}\,\mathrm{vec}(\mathcal{F}_{2D}) \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{V_T \times S_0}$ represents the routing matrix governing the information flow from perspective pixels to 3D voxels, and $\mathrm{vec}(\cdot)$ defines the flattened feature.

Explicit Physical Projection. Practically computing a dense unconstrained $\mathbf{W}$ incurs a prohibitive quadratic complexity of $\mathcal{O}(S_0 V_T)$. To make the 2D-to-3D lifting computationally tractable, classical physical-prior-based methods [22] artificially enforce extreme sparsity on the routing matrix. Specifically, LSS predicts a categorical depth distribution over D discrete bins along the camera ray of each pixel, and splats the probability-weighted features strictly into those localized bins.

We can formally characterize this computational compromise as a ray-local support constraint. Let $\mathbf{W}[:, s] \in \mathbb{R}^{V_T}$ denote the s-th column of the routing matrix, representing the projection weights from a single 2D pixel s. Mathematically, the *support set* of this vector, denoted as $\mathrm{supp}(\mathbf{W}[:, s])$, is defined as the set of 3D voxel indices where the routing weight is strictly non-zero (i.e., the specific voxels that successfully receive direct evidence from pixel s).

Given the calibration Θ, let $\mathcal{R}_\Theta(s) \subseteq \{1, \ldots, V_T\}$ denote the small subset of 3D voxels that physically intersect with the geometric ray of token s. The classical LSS explicitly limits the non-zero elements of $\mathbf{W}$ to this ray, defining a highly constrained feasible set:

$$\mathcal{S}_{ray}(\Theta) = \left\{ \mathbf{W} : \mathrm{supp}(\mathbf{W}[:, s]) \subseteq \mathcal{R}_\Theta(s),\ \forall s \right\} \tag{2}$$

The Defect: Restricted Reachability. By restricting the 2D-to-3D routing matrix to a highly sparse pattern strictly aligned with the ray support set, explicit lifting secures computational efficiency but suffers from severe locality limitations. This rigid structural constraint leaves a massive number of 3D voxels completely unreachable or only weakly reachable to direct image evidence during the VT stage. Consequently, when observing continuous macroscopic topologies (*e.g.*, cross-camera vehicles or occlusion boundaries), the network independently splats them as isolated structural fragments.

Because the necessary cross-ray geometric constraints are mathematically blocked by $\mathcal{R}_\Theta(s)$, LSS-based architectures are forced to rely on post-hoc 3D/BEV convolutional networks to progressively diffuse information and bridge these fragments. This *posterior remedy* is computationally inefficient, prone to generating fractured occupancy predictions, and structurally suboptimal.

To overcome this inherent suboptimality, the VT operator must theoretically evolve from the locally constrained space toward a fully global reachability. Let $\mathcal{S}_{global} = \mathbb{R}^{V_T \times S_0}$ denote the completely unconstrained dense routing space, where a 2D pixel can theoretically route to any 3D voxel. For any feasible set $\mathcal{S}$, let $\mathcal{F}(\mathcal{S}) = \{ f_{\mathbf{W}} : \mathcal{F}_{2D} \mapsto \mathbf{W}\mathcal{F}_{2D} \mid \mathbf{W} \in \mathcal{S} \}$ denote the induced function family. We establish the following theoretical motivation:

Proposition 1 (Expressivity Inclusion by Relaxing Support). *For any given camera calibration Θ, the ray-local routing is mathematically merely a subset constraint bounded by the physical geometry. Therefore, the feasible set satisfies $\mathcal{S}_{ray}(\Theta) \subset \mathcal{S}_{global}$, and the induced function families mathematically satisfy $\mathcal{F}(\mathcal{S}_{ray}(\Theta)) \subseteq \mathcal{F}(\mathcal{S}_{global})$, with the inclusion being strict whenever cross-ray voxel correlations carry non-trivial information.*

This proposition shows that relaxing the routing constraints yields a function family with representational capacity at least as large as, and generically strictly larger than, that of the ray-constrained family. We therefore propose to approximate global dense routing at tractable complexity.

3.3 Factorized Dense Routing (FDR)

While Proposition 1 establishes the theoretical necessity of $\mathcal{S}_{global}$, explicitly parameterizing a fully unconstrained routing matrix incurs an intractable computational complexity of $\mathcal{O}(S_0 V_T C)$. To practically instantiate unconstrained global reachability, we propose **Factorized Dense Routing (FDR)**, as shown in Figure 2 (c). Motivated by the architectural evolution from fully-connected dense networks to hierarchical convolutional fields, FDR approximates the intractable

Table 1: Progression of spatial dimensions across FDR stages. To maintain linear complexity, the 2D dimensions progressively contract via Δ_h, Δ_w, while the 3D dimensions exponentially expand via $\Delta_x, \Delta_y, \Delta_z$.

Stage t	**2D Resolution**	**3D Volume**	**Patch** P_t	**Expansion** K_t
0 (Input)	$H \times W$	$1 \times 1 \times 1$	-	-
1	$\frac{H}{\Delta_h^{(1)}} \times \frac{W}{\Delta_w^{(1)}}$	$\Delta_x^{(1)} \times \Delta_y^{(1)} \times \Delta_z^{(1)}$	$\Delta_h^{(1)}\Delta_w^{(1)}$	$\Delta_x^{(1)}\Delta_y^{(1)}\Delta_z^{(1)}$
...	...	...	...	...
T (Output)	1×1	$X \times Y \times Z$	-	-

dense matrix $\mathbf{W}$ as a product of T localized, stage-wise tensor operations:

$$\mathbf{W} \approx \mathbf{M}^{(T)}\mathbf{M}^{(T-1)} \dots \mathbf{M}^{(1)} \tag{3}$$

where each $\mathbf{M}^{(t)}$ acts as a sparse block-matrix operator performing local dense routing between a contracted 2D patch and an expanded 3D micro-volume.

Stage-wise Spatial Contraction and Expansion. Let $\mathcal{X}^{(t)} \in \mathbb{R}^{S_t \times V_t \times C}$ denote the flattened spatial tensor at stage t. The algorithm is initialized with $\mathcal{X}^{(0)} = \mathcal{F}_{2D} \in \mathbb{R}^{S_0 \times 1 \times C}$, where the initial 3D volume dimension is degenerated at $V_0 = 1$. Our objective is to progressively contract the 2D spatial size $S_t \to 1$ while exponentially expanding the 3D volume size V_t toward the macroscopic 3D dimensions. Between adjacent stages, standard 2D convolutions on the 2D space are applied to refine the spatial features.

At each stage t, we partition the 2D spatial grid into non-overlapping patches with adaptive sizes $P_t = \Delta_h^{(t)} \times \Delta_w^{(t)}$. We unfold the features within each patch, yielding:

$$\tilde{\mathcal{X}}^{(t-1)} = \text{Unfold}_{P_t}(\mathcal{X}^{(t-1)}) \in \mathbb{R}^{S_t \times V_{t-1} \times C \times P_t} \tag{4}$$

where $S_t = S_{t-1}/P_t$, with adaptive zero-padding applied to ensure exact divisibility. Subsequently, each 2D patch is dense-routed to an expanded 3D micro-volume containing $K_t = \Delta_x^{(t)} \times \Delta_y^{(t)} \times \Delta_z^{(t)}$ tokens via a dynamically generated routing matrix $\mathbf{A}^{(t)} \in \mathbb{R}^{S_t \times V_{t-1} \times P_t \times K_t}$. This local dense routing is executed as a tensor contraction over the patch dimension P_t:

$$\mathcal{Y}^{(t)} = \tilde{\mathcal{X}}^{(t-1)} \otimes \mathbf{A}^{(t)}, \quad \mathcal{Y}^{(t)} \in \mathbb{R}^{S_t \times V_{t-1} \times C \times K_t}, \tag{5}$$

where $\otimes$ denotes batched matrix multiplication along the P_t axis. Finally, a Fold operation merges the newly expanded K_t dimension into the 3D volume axis, yielding the updated spatial tensor $\mathcal{X}^{(t)} \in \mathbb{R}^{S_t \times V_t \times C}$, with $V_t = V_{t-1} \times K_t$. The exact progression of these spatial dimensions is detailed in Table 1.

Adaptive Geometric Context Routing. The dynamic weights $\mathbf{A}^{(t)}$ dictate the critical probability mapping. Instead of utilizing rigid static parameters, we dynamically generate them. Specifically, we fuse the intermediate spatial features with an aligned geometric context $\mathcal{E}_{geo}^{(t)}$, process them through a lightweight mapping network $g_\theta^{(t)}$ (comprising standard convolutions), and subsequently unfold

the resulting pre-routing logits:

$$\mathbf{A}^{(t)} = \sigma\left(\text{Unfold}_{P_t}\left(g_\theta^{(t)}\left(\mathcal{X}^{(t-1)} \parallel \mathcal{E}_{geo}^{(t)}\right)\right)\right) \tag{6}$$

where $\parallel$ denotes channel concatenation, and σ is the Softmax function normalized over the expanded K_t micro-volume dimension.

Crucially, $\mathcal{E}_{geo}^{(t)}$ comprises explicit local depth features (derived from a lightweight depth head) and Plücker ray coordinates $\mathcal{E}_{ray} = \text{MLP}\left(\mathbf{d} \parallel (\mathbf{o} \times \mathbf{d})\right)$. Conditioning the routing weights on $\mathcal{E}_{geo}^{(t)}$ treats physical geometry as a soft inductive bias rather than a rigid structural constraint. Unlike LSS, which hard-zeros routing weights outside the physical ray, this context-aware conditioning allows FDR to discover implicit 3D correlations.

Complexity and Theoretical Advantage. While fully unconstrained global routing dictates an intractable complexity of $\mathcal{C}_{global} = \mathcal{O}(S_0 V_T C)$, FDR factorizes this process to reduce the relative computational cost to a strict ratio of $\mathcal{C}_{FDR}/\mathcal{C}_{global} = \mathcal{O}\big(\sum_{t=1}^{T} \frac{S_{t-1}}{S_0}\frac{V_t}{V_T}\big)$. Physically, early stages are computationally trivialized by the unexpanded 3D volume ($V_t \ll V_T$), while late stages are exponentially accelerated by the highly contracted 2D grid ($S_{t-1} \ll S_0$). As mathematically detailed in the Supplementary Material, under our practical settings ($T = 3$), FDR achieves fully-global reachability while requiring merely $\sim 1.4\%$ of the dense theoretical computation.

FDR also subsumes classical LSS as a special case. As proved in the Supplementary Material, *the classical LSS operator is strictly a degenerate instance of FDR* (obtained when $T = 1$, $P_1 = 1 \times 1$, and weights are heavily masked). This inclusion ($\mathcal{F}_{LSS} \subseteq \mathcal{F}_{FDR}$) guarantees that FDR's hypothesis space strictly contains that of explicit physical priors, allowing it to internalize multi-camera rig topologies even when physical extrinsics are uncalibrated or absent.

3.4 Resolution-Context Decoupled Unification

As established above, circumventing the intractable quadratic complexity of dense global routing intrinsically requires progressively contracting the spatial dimensions via $P_t > 1$. This mathematical necessity reveals a fundamental **Resolution-Context Trade-off**: significantly reducing the computational cost necessitates spatial coarsening. Therefore, a single monolithic routing operator cannot simultaneously maintain pixel-exact 2D spatial precision and achieve fully-global contextual reachability.

Rather than forcing a single operator to violate this structural bound, we propose a decoupled architecture that factorizes the 3D representation space into two orthogonal pathways (Figure 2 (b)):

- **Global Context Pathway ($P_t > 1$):** The mandatory patch-wise aggregation in FDR explicitly expands the contextual footprint. It excels at gathering wide-range global semantics (*e.g.*, continuous road layouts and holistic occlusion relationships) to achieve the unconstrained reachability of $\mathcal{S}_{global}$.

– **Local Resolution Pathway ($P_t = 1 \times 1$):** Explicit LSS strictly avoids any cross-patch aggregation. While this Kronecker-delta-like routing bounds its receptive field to zero ($\mathcal{S}_{ray}$), it perfectly preserves exact 2D spatial precision, acting as an optimal pathway to carve sharp local geometric boundaries.

To unify these orthogonal capacities without the memory overhead of dense 3D fusion, we design a volumetric decomposition. We instantiate the local pathway via classical LSS to yield a high-resolution precise surface volume $\mathcal{V}_{LSS} \in \mathbb{R}^{C \times X \times Y \times Z}$. Concurrently, rather than forcing the resolution-coarsened FDR to reconstruct sharp 3D surfaces, we configure its final stage to compress the Z-axis. This outputs a globally connected, gravity-aligned *Holistic Context Anchor* $\mathcal{V}_{global} \in \mathbb{R}^{C \times X \times Y \times 1}$.

We then orthogonally pool the high-resolution $\mathcal{V}_{LSS}$ into three axis-aligned planes ($\mathcal{P}_{xy}, \mathcal{P}_{xz}, \mathcal{P}_{yz}$). Because global driving topologies inherently reside in the BEV plane, the global-local unification is performed via addition on the shared XY plane:

$$\tilde{\mathcal{P}}_{xy} = \mathcal{P}_{xy} + \mathcal{V}_{global} \tag{7}$$

This formulation combines the unbounded contextual reachability of dense routing with the local geometric precision of sparse physical priors.

3.5 Occupancy Decoding and Training

The unified plane $\tilde{\mathcal{P}}_{xy}$ is processed by a dedicated 2D convolutional encoder, while the remaining geometric planes $\mathcal{P}_{xz}$ and $\mathcal{P}_{yz}$ are processed by a secondary shared 2D encoder. To reconstruct the dense 3D volume, $\tilde{\mathcal{P}}_{xy}$ is first passed through a linear projector to recover the Z-axis dimension from its channels. The encoded $\mathcal{P}_{xz}$ and $\mathcal{P}_{yz}$ planes are then spatially broadcasted along their respective missing axes. Finally, these three expanded tensors are element-wise summed and fed into an MLP decoder to predict the voxel-wise semantic probabilities. The framework is optimized end-to-end using a joint multi-task objective:

$$\mathcal{L} = \mathcal{L}_{BCE} + \mathcal{L}_{dice} + \mathcal{L}_{depth} + \mathcal{L}_{sem}, \tag{8}$$

where the 3D occupancy loss combines Binary Cross-Entropy (BCE) and Dice loss to address voxel-class imbalance. For auxiliary supervision, $\mathcal{L}_{depth}$ explicitly guides the depth distribution in the LSS branch, and $\mathcal{L}_{sem}$ enforces 2D perspective-view segmentation following ALOcc [4].

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the proposed decoupled architecture on two standard autonomous driving benchmarks: Occ3D-nuScenes [26] and Occ3D-Waymo [26]. The Occ3D-nuScenes dataset, derived from the nuScenes dataset [1], provides dense 3D occupancy annotations encompassing 17 foreground semantic classes. It contains 600 training sequences and 150 validation sequences, with a predefined

voxel resolution of $0.4\text{m} \times 0.4\text{m} \times 0.4\text{m}$ covering the spatial range of $[-40\text{m}, 40\text{m}] \times [-40\text{m}, 40\text{m}] \times [-1\text{m}, 5.4\text{m}]$. Conversely, the Occ3D-Waymo dataset is constructed from the Waymo Open Dataset [25]. It offers an expanded scale of 798 training and 202 validation sequences with 14 foreground semantic categories, presenting a more complex and expansive urban driving environment.

Evaluation Metrics. Following the established evaluation protocols [26], we adopt the mean Intersection over Union (mIoU) metric to quantify the geometric and semantic accuracy of the 3D occupancy predictions. We also report the geometric-only Intersection over Union (IoU) to isolate and assess pure structural completeness.

Implementation Details. For a fair comparison with existing baselines, we utilize ResNet-50 [5] as the 2D image backbone. The input image resolutions are set to 256×704 for nuScenes and 640×960 for Waymo. For our Factorized Dense Routing (FDR), we employ a 3-stage configuration ($T = 3$) where, since the Z-axis is compressed to 1, the BEV micro-volumes are expanded by area factors of 10×10, 5×5, and 4×4, respectively. Correspondingly, the 2D spatial dimensions are contracted with patch sizes P_t of 4×4, 2×4, and 2×2. The model is optimized end-to-end using AdamW [39] with a learning rate of 2×10^{-4} and a global batch size of 16 for a total of 24 training epochs. Our experiments on Occ3D-nuScenes focus on settings with minimal temporal frames to explicitly emphasize the structural improvements derived from our 2D-to-3D View Transformation architecture. Due to the scarcity of single-frame baselines on Occ3D-Waymo, we follow ALOcc [4] to fuse 16 temporal frames when evaluating on this dataset.

4.2 State-of-the-Art Comparison

Results on Occ3D-nuScenes. We present a comprehensive quantitative comparison on the Occ3D-nuScenes validation set in Table 2, which is organized into two settings: a single-frame setting without temporal voxel feature fusion (upper block) and a setting with one historical frame (+**1f**, lower block). Our framework achieves the best mIoU in both settings, reaching 41.2% among single-frame methods and 42.2% in the +1f setting, surpassing all competing methods under each respective protocol. By replacing strict ray constraints with FDR, our model gains consistent improvements on large-scale classes such as driveable surface, which require extensive cross-ray contextual reasoning. The Resolution-Context Decoupled Unification preserves local geometric fidelity, yielding competitive results on tightly localized objects such as *car* and *motor*.

Results on Occ3D-Waymo. The results on Occ3D-Waymo are summarized in Table 3. Our method achieves a state-of-the-art mIoU of **31.25**%. Specifically, it outperforms ALOcc-3D [4] by a margin of +**1.22**% mIoU. The gains on spatially continuous classes, notably *Road* (81.21%) and *Walkable* (64.46%), confirm that FDR builds global contextual priors and reduces the spatial isolation inherent to pure physical-projection methods.

Table 2: 3D Semantic Occupancy Prediction Results on Occ3D-nuScenes validation set. The upper block reports the single-frame setting (no temporal voxel feature fusion); the lower block reports the setting with one historical frame, where the suffix **+1f** denotes the incorporation of one historical frame. The top performance within each block is highlighted in **bold**.

Method	Backbone	Size	**mIoU**	**IoU**	others	barrier	bicycle	bus	car	cons. veh.	motor.	pedes.	tfc. cone	trailer	truck	drv. surf.	other flat	sidewalk	terrain	manmade	vegetation
FB-Occ [16]	ResNet-50	256 × 704	35.7	66.5	11.2	40.0	23.2	43.1	46.6	21.8	23.3	25.9	23.5	29.5	33.7	77.6	39.0	47.3	51.0	37.5	33.0
DHD-S [30]	ResNet-50	256 × 704	36.5	-	10.6	43.2	23.0	40.6	47.3	21.7	23.3	23.9	23.4	31.8	34.2	80.2	41.3	50.0	54.1	38.7	33.5
COTR [19]	ResNet-50	256 × 704	39.1	69.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BEVDetOcc [6]	ResNet-50	256 × 704	37.1	70.4	9.2	45.6	15.9	42.1	50.0	23.0	21.0	21.9	22.7	31.1	36.8	80.2	38.7	51.6	55.2	45.4	40.2
LightOcc-S [37]	ResNet-50	256 × 704	37.9	-	11.7	45.6	25.4	43.1	48.7	21.4	25.6	26.6	29.2	33.2	35.1	80.0	41.8	50.4	53.9	39.4	34.0
ProtoOcc [9]	ResNet-50	256 × 704	39.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ALOcc [4]	ResNet-50	256 × 704	40.1	70.2	12.4	46.4	26.2	42.6	52.1	25.6	27.5	28.1	30.2	32.9	39.5	81.2	42.5	52.6	56.0	46.1	39.7
Ours	ResNet-50	256 × 704	**41.2**	**71.0**	13.6	46.5	28.3	43.4	53.0	26.8	29.1	29.6	31.6	36.4	39.9	81.6	44.3	53.0	55.5	47.2	40.6
COTR [19]	ResNet-50	256 × 704	41.4	-	12.2	48.5	29.1	44.7	53.3	27.0	29.2	28.9	31.0	35.0	39.5	81.8	42.5	53.7	56.9	48.2	42.1
DHD-M [30]	ResNet-50	256 × 704	41.5	-	12.7	48.7	26.3	43.2	52.9	27.3	28.5	28.5	30.0	35.8	40.2	83.1	44.7	54.7	57.7	48.9	42.1
SA-Occ [3]	ResNet-50	256 × 704	40.7	-	10.9	48.5	23.5	45.8	52.8	24.5	23.6	24.2	22.7	35.0	40.3	83.5	44.1	55.5	60.0	51.2	45.3
Ours (+1f)	ResNet-50	256 × 704	**42.2**	**72.1**	14.2	48.7	28.8	44.8	53.8	26.8	29.9	30.3	33.0	35.4	41.3	82.6	45.2	54.2	57.3	49.2	42.2

Table 3: 3D Semantic Occupancy Prediction Results on Occ3D-Waymo validation set. The best results are highlighted in **bold**.

Method	**mIoU**	Gen. Obj.	Vehicle	Pedestrian	Sign	Bicyclist	Traffic Light	Pole	Cons. Cone	Bicycle	Building	Vegetation	Tree Trunk	Road	Walkable
BEVFormer-w/o TSA	23.87	7.50	34.54	21.07	9.69	20.96	11.48	11.48	14.06	14.51	23.14	21.82	8.57	78.45	56.89
BEVFormer [15]	24.58	7.18	36.06	21.00	9.76	20.23	12.61	14.52	14.70	16.06	23.98	22.50	9.39	79.11	57.04
SOLOFusion [21]	24.73	4.97	32.45	18.28	10.33	17.14	8.07	17.83	16.23	19.30	31.49	28.98	16.93	70.95	53.28
BEVFormer-WrapConcat	25.07	6.20	36.17	20.95	9.56	20.58	12.82	16.24	14.31	16.78	25.14	23.56	12.81	79.04	56.83
CVT-Occ [35]	27.37	7.44	41.00	23.93	11.92	20.81	12.07	18.03	16.88	**21.37**	29.40	27.42	14.67	79.12	59.09
ALOcc-3D [4]	30.03	6.51	39.61	24.14	20.84	20.56	20.56	24.28	17.95	12.22	35.67	37.25	22.45	78.42	59.91
Ours	**31.25**	**9.03**	**42.74**	**27.52**	**22.91**	**25.38**	**21.47**	**27.84**	**22.65**	20.69	**39.98**	**38.89**	**24.02**	**81.21**	**64.46**

4.3 Robustness in Uncalibrated Scenarios

As theoretically analyzed in Section 3, explicit physical projection methods are mathematically bound to strict geometric constraints. To empirically validate the resilience of FDR's global hypothesis space, we design an uncalibrated scenario on nuScenes. We completely withhold the dataset-provided camera parameters and instead rely on a lightweight convolutional head that blindly predicts rough camera poses from images. As shown in Table 5, explicit physical methods suffer catastrophic performance degradation. Our decoupled framework maintains a stable 28.0% mIoU. This confirms that FDR's unconstrained global capacity lets the network internalize multi-camera rig topology from visual context alone, without relying on physical calibration.

Table 4: Ablation on Model Components and Unification Strategies. Evaluated under a lightweight single-frame setting on Occ3D-nuScenes.

Component Strategy	mIoU	IoU
Baseline (Our Default)	**38.1**	**67.1**
w/o Global Pathway (FDR)	37.0	66.0
w/o Soft Geometric Hint	37.7	66.7
Only BEV Representation	37.7	67.9
Adaptive Weighting	38.0	67.2
Late 3D Fusion	37.7	66.9

Table 5: Robustness without Camera Parameters. Evaluated on nuScenes using implicitly predicted poses.

Method	mIoU	IoU
BEVDetOcc [6]	4.7	33.1
ALOcc [4]	9.0	43.5
FB-Occ [16]	12.0	54.5
Ours	**28.0**	**63.2**

Table 6: Ablation on Stage Configurations. V_t denotes volumetric expansion factor at each stage.

Stage Schedule (V_t)	mIoU	IoU
3 Stages (10→5→4)	**38.1**	**67.1**
2 Stages ($20 \rightarrow 10$)	37.4	66.5
3 Stages ($4 \rightarrow 5 \rightarrow 10$)	37.7	66.8
3 Stages ($10 \rightarrow 10 \rightarrow 2$)	37.9	66.9

4.4 Ablation Studies and Architectural Analysis

As shown in Tables 4 and 6, we conduct extensive ablation studies on the Occ3D-nuScenes dataset. For computational efficiency, all ablations are trained under a lightweight single-frame configuration with the channel dimension halved.

Model Design. We first evaluate the necessity of the proposed dual-pathway unification in Table 4. Removing the global pathway (FDR) entirely degrades the performance of 1.1%. This confirms that introducing fully-global contextual reachability is beneficial for overcoming the receptive field isolation of strict geometric priors. Additionally, omitting the soft geometric context ($\mathcal{E}_{geo}$) during the dynamic generation of FDR's routing weights yields a suboptimal 37.7% mIoU, indicating that physical hints effectively guide the unconstrained dense mapping. We also observe that implementing the unification late in the network (*e.g.*, after the 3D encoder, 37.7%), or substituting the tri-perspective representation with BEV representation (37.7%) or adaptive spatial weighting (38.0%), offers no discernible advantage. This affirms our design choice: because FDR and LSS occupy orthogonal structural regimes, simple element-wise addition on the gravity-aligned BEV plane is the natural and principled choice, avoiding additional engineering complexity.

Stage-wise Contraction Configurations. We ablate the factorized spatial contraction across different T-stage settings in Table 6. We observe that compressing the dense routing into a shallower 2-stage network ($T = 2$) degrades the mIoU to 37.4%, suggesting that an abrupt spatial aggregation hinders smooth contextual feature learning. Conversely, heavily delaying the volumetric expansion (*e.g.*, 4-5-10 or 10-10-2 configurations) yields slightly lower results (37.7% and 37.9%) compared to our default progressive schedule (10-5-4, 38.1%). This shows that a progressive spatial abstraction schedule is crucial for building robust macroscopic context.

4.5 Qualitative Results

Figure 3 compares the occupancy predictions of our framework against ALOcc [4] in complex scenes with heavy occlusions. As shown, existing ray-based methods

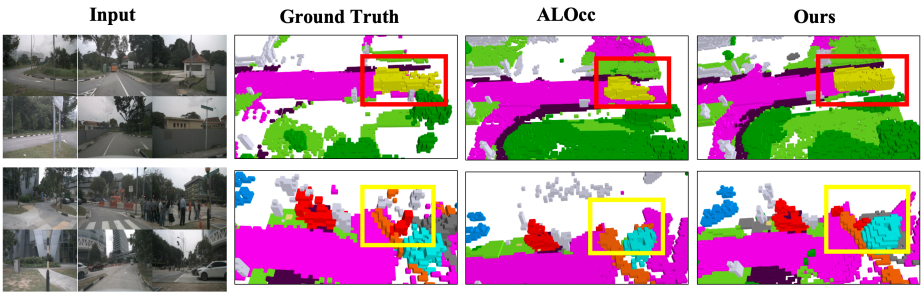

Fig. 3: Qualitative comparison of 3D semantic occupancy prediction. We highlight the regions with complex structures and heavy occlusions in red and yellow boxes. While existing methods ALOcc often produce fragmented or incomplete occupancy topologies due to the strict ray-wise locality bottleneck, our proposed framework successfully recovers continuous and holistic structures by effectively internalizing global contextual relationships.

often produce fragmented or incomplete topologies, particularly for large objects spanning multiple cameras. This is a direct consequence of *Receptive Field Isolation*: since individual 3D voxels can only aggregate evidence from strictly aligned rays, they fail to associate disjointed spatial clues into a coherent whole. In contrast, our framework recovers continuous, holistic structures in occluded regions (highlighted in red and yellow boxes). By internalizing global context through *Factorized Dense Routing*, our model establishes structural continuity where direct image evidence is absent. This empirical result verifies that our global routing operator effectively bridges the structural gaps that remain unreachable to sparse physical-masking baselines.

5 Conclusion

We address the *Locality Bottleneck* of classical ray-based lifting, formally showing that physical projection is a degenerate, constrained subset of a fully-global routing space. Our **Factorized Dense Routing (FDR)** approximates dense 2D-to-3D mapping at linear complexity via hierarchical spatial aggregations. We then address the resulting *Resolution-Context Trade-off* with a macro-micro decoupled architecture, combining the global topological anchor from FDR with the pixel-exact geometric precision of sparse physical priors. Empirically, our unified framework establishes new state-of-the-art results across large-scale benchmarks. We further demonstrate that FDR's unconstrained capacity enables the network to internalize multi-camera topology from visual context alone, remaining functional even when physical calibration is entirely absent. We hope this work encourages further research into geometry-free 3D representation learning for autonomous driving.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
2. Cao, A.Q., de Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3991–4001 (2022)
3. Chen, C., Wang, Z., Sheng, T., Jiang, Y., Li, Y., Cheng, P., Zhang, L., Chen, K., Hu, Y., Yang, X., et al.: Sa-occ: Satellite-assisted 3d occupancy prediction in real world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27021–27030 (2025)
4. Chen, D., Fang, J., Han, W., Cheng, X., Yin, J., Xu, C., Khan, F.S., Shen, J.: Alocc: adaptive lifting-based 3d semantic occupancy and cost volume-based flow prediction. arXiv preprint arXiv:2411.07725 (2024)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
6. Huang, J., Huang, G.: Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054 (2022)
7. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
8. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. arXiv preprint arXiv:2302.07817 (2023)
9. Kim, J., Kang, C., Lee, D., Choi, S., Choi, J.W.: Protoocc: Accurate, efficient 3d occupancy prediction using dual branch encoder-prototype query decoder. arXiv preprint arXiv:2412.08774 (2024)
10. Li, B., Guo, J., Liu, H., Zou, Y., Ding, Y., Chen, X., Zhu, H., Tan, F., Zhang, C., Wang, T., et al.: Uniscene: Unified occupancy-centric driving scene generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11971–11981 (2025)
11. Li, B., Ma, Z., Du, D., Peng, B., Liang, Z., Liu, Z., Ma, C., Jin, Y., Zhao, H., Zeng, W., et al.: Omninwm: Omniscient driving navigation world models. arXiv preprint arXiv:2510.18313 (2025)
12. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9087–9098 (2023)
13. Li, Y., Bao, H., Ge, Z., Yang, J., Sun, J., Li, Z.: Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1486–1494 (2023)
14. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. arXiv preprint arXiv:2206.10092 (2022)
15. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. arXiv preprint arXiv:2203.17270 (2022)

16. Li, Z., Yu, Z., Austin, D., Fang, M., Lan, S., Kautz, J., Alvarez, J.M.: Fb-occ: 3d occupancy prediction based on forward-backward view transformation. arXiv preprint arXiv:2307.01492 (2023)
17. Li, Z., Yu, Z., Wang, W., Anandkumar, A., Lu, T., Alvarez, J.M.: Fb-bev: Bev representation from forward-backward view transformations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6919–6928 (2023)
18. Liu, S., Hu, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., Li, X.: See and think: Disentangling semantic scene completion. Advances in Neural Information Processing Systems **31** (2018)
19. Ma, Q., Tan, X., Qu, Y., Ma, L., Zhang, Z., Xie, Y.: Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. arXiv preprint arXiv:2312.01919 (2023)
20. Pan, M., Liu, J., Zhang, R., Huang, P., Li, X., Xie, H., Wang, B., Liu, L., Zhang, S.: Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 12404–12411. IEEE (2024)
21. Park, J., Xu, C., Yang, S., Keutzer, K., Kitani, K., Tomizuka, M., Zhan, W.: Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. arXiv preprint arXiv:2210.02443 (2022)
22. Philion, J., Fidler, S.: 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. pp. 194–210. Springer (2020)
23. Roldao, L., De Charette, R., Verroust-Blondet, A.: 3d semantic scene completion: A survey. International Journal of Computer Vision **130**(8), 1978–2005 (2022)
24. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1746–1754 (2017)
25. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
26. Tian, X., Jiang, T., Yun, L., Mao, Y., Yang, H., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. Advances in Neural Information Processing Systems **36** (2024)
27. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. arXiv preprint arXiv:2303.03991 (2023)
28. Wang, Y., Chen, Y., Liao, X., Fan, L., Zhang, Z.: Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
29. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. arXiv preprint arXiv:2303.09551 (2023)
30. Wu, Y., Yan, Z., Wang, Z., Li, X., Hui, L., Yang, J.: Deep height decoupling for precise vision-based 3d occupancy prediction. arXiv preprint arXiv:2409.07972 (2024)
31. Wu, Y., Yan, Z., Zhang, Y., Li, X., Yang, J.: See through the dark: Learning illumination-affined representations for nighttime occupancy prediction. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)

32. Xia, Z., Liu, Y., Li, X., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y.: Scpnet: Semantic scene completion on point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17642–17651 (2023)
33. Yan, T., Wu, D., Han, W., Jiang, J., Zhou, X., Zhan, K., Xu, C.z., Shen, J.: Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27531–27541 (2025)
34. Yang, Y., Liang, A., Mei, J., Ma, Y., Liu, Y., Lee, G.H.: X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. arXiv preprint arXiv:2506.13558 (2025)
35. Ye, Z., Jiang, T., Xu, C., Li, Y., Zhao, H.: Cvt-occ: Cost volume temporal fusion for 3d occupancy prediction. In: European Conference on Computer Vision. pp. 381–397. Springer (2024)
36. Yu, Z., Shu, C., Deng, J., Lu, K., Liu, Z., Yu, J., Yang, D., Li, H., Chen, Y.: Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. arXiv preprint arXiv:2311.12058 (2023)
37. Zhang, J., Zhang, Y., Liu, Q., Wang, Y.: Lightweight spatial embedding for vision-based 3d occupancy prediction. arXiv preprint arXiv:2412.05976 (2024)
38. Zhang, Y., Zhu, Z., Du, D.: Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. arXiv preprint arXiv:2304.05316 (2023)
39. Zhou, P., Xie, X., Lin, Z., Yan, S.: Towards understanding convergence and generalization of adamw. IEEE transactions on pattern analysis and machine intelligence **46**(9), 6486–6493 (2024)