Dr.Occ: Depth- and Region-Guided 3D Occupancy from Surround-View Cameras for Autonomous Driving

Supplementary Material

1. Implementation Details

1.1. Deformable Cross-Attention Module Architecture

We provide comprehensive architectural specifications of the Deformable Cross-Attention (DCA) module, which serves as the core building block of D^2-VFormer. As illustrated in Fig. 1, each DCA module employs a two-stage design: (1) a transformer block for geometric-aware feature learning, and (2) a feature deepening module for progressive refinement.

Transformer Block. Given input voxel features $\mathbf{F}_{in} \in \mathbb{R}^{N\times C}$, we first apply self-attention to model intra-voxel dependencies. Following standard practice, residual connection and layer normalization stabilize training:

$$\mathbf{F}_{self} = \text{LayerNorm}(\mathbf{F}_{in} + \text{SelfAttn}(\mathbf{F}_{in})). \quad (1)$$

Subsequently, we establish geometric correspondence between 3D voxel space and 2D image space through deformable cross-attention, where the normalized features $\mathbf{F}_{self}$ query corresponding image features $\mathbf{F}_{img}$:

$$\mathbf{F}_{cross} = \text{DeformCrossAttn}(\mathbf{Q} = \mathbf{F}_{self}, \mathbf{K}, \mathbf{V} = \mathbf{F}_{img}). \quad (2)$$

Finally, an FFN with a residual connection further refines the attended features:

$$\mathbf{F}_{trans} = \text{LayerNorm}(\mathbf{F}_{cross} + \text{FFN}(\mathbf{F}_{cross})). \quad (3)$$

Feature Deepening Module. To preserve low-level geometric cues while incorporating high-level semantic information from the transformer, we concatenate the original input $\mathbf{F}_{in}$ with the transformer output $\mathbf{F}_{trans}$:

$$\mathbf{F}_{cat} = [\mathbf{F}_{in}, \mathbf{F}_{trans}] \in \mathbb{R}^{N\times 2C}, \quad (4)$$

where $[\cdot,\cdot]$ denotes channel-wise concatenation. The concatenated features are then processed through a progressive refinement pathway consisting of three 3D convolutional layers ($3\times3\times3$ kernel) with batch normalization and ReLU activation:

$$\mathbf{F}_{out} = \text{Conv}_3^C \circ \text{ReLU} \circ \text{BN} \circ \text{Conv}_2^{2C} \circ \text{ReLU} \circ \text{BN} \circ \text{Conv}_1^{2C}(\mathbf{F}_{cat}), \quad (5)$$

where superscripts denote output channel dimensions. The first two convolutions maintain the expanded dimensionality ($2C$) for sufficient representational capacity, while the final convolution projects back to C dimensions to enable residual connection with subsequent layers. This bottleneck-like design balances feature expressiveness and computational efficiency.

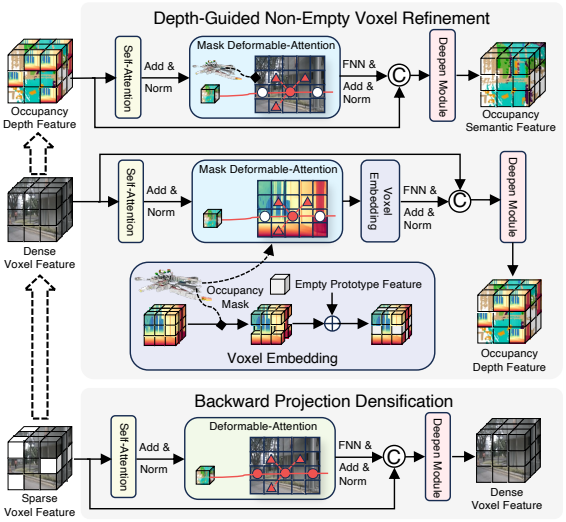

Figure 1. **Depth-guided Dual Projection View Transformer.**

1.2. Loss Functions

During training, we employ a combination of loss functions to supervise both semantic and geometric learning:

$$\mathcal{L} = \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}}, \quad (6)$$

where $\mathcal{L}_{\text{seg}}$ denotes the semantic supervision loss with two variants: Weight-CELoss employs manually-predefined voxel-level spatial weights that decrease with distance from the ego vehicle. Specifically, for a voxel at position $\mathbf{v} = (x, y, z)$, the weight is defined as $w(\mathbf{v}) = \exp(-\alpha \cdot d(\mathbf{v}))$, where $d(\mathbf{v}) = \sqrt{x^2 + y^2}$ is the distance from the ego vehicle and $\alpha = 0.01$ is a decay factor. Probability-CELoss adaptively generates voxel-level weights from the model itself. Following [2, 6], we introduce an auxiliary decoder that estimates the contribution of each voxel by predicting a confidence score $p(\mathbf{v}) \in [0, 1]$. This decoder shares the same feature backbone but outputs a single-channel probability map. The adaptive weight for each voxel is then computed as $w(\mathbf{v}) = p(\mathbf{v})$, enabling the model to automatically learn spatially-varying importance without manual tuning. In practice, Weight-CELoss is adopted when using the R-EFormer module, where fixed spatial priors better align with its manually defined regional structure, while Probability-CELoss is employed for the R^2-EFormer variant to leverage its adaptive recursive refinement and dynamically learn voxel-wise importance. $\mathcal{L}_{\text{depth}}$ follows

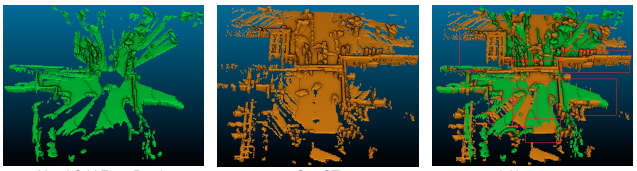

Figure 2. **Visibility-occupancy mismatch in pseudo-LiDAR projection. Left**: Voxel grid reconstructed from monocular depth. **Middle**: Occupancy ground truth. **Right**: Overlay visualization achieving only 13.54% IoU. Key discrepancies: (1) **Over-densification**—depth produces dense voxels on visible surfaces where GT is sparse, and (2) **Occlusion incompleteness**—depth misses occluded regions that GT captures through multi-view aggregation. This fundamental visibility-occupancy gap renders direct depth-to-voxel conversion inadequate for camera-only occupancy prediction.

BEVDepth [3] to supervise the forward-projected depth distribution. Following MonoScene [1], we introduce semantic affinity loss $\mathcal{L}_{\text{sem}}$ and geometric affinity loss $\mathcal{L}_{\text{geo}}$ to strengthen supervision on semantic and geometric features. The hyper-parameters are set to $\lambda_{\text{seg}} = 10$, $\lambda_{\text{depth}} = 1$, $\lambda_{\text{sem}} = 1$, and $\lambda_{\text{geo}} = 1$, following [1, 7] to maintain consistent optimization scales across losses.

2. Further Experiments

2.1. Alternative Depth Integration Strategies

To systematically validate our depth-guided dual projection mechanism, we design comprehensive ablation studies against three representative depth integration strategies (see Table 1 and Fig. 2). First, *Direct Depth Replacement* (following BEVDepth [3]) substitutes learned depth with externally predicted one-hot depth distributions from MoGe-2 [4], which causes severe feature–geometry misalignment and leads to a notable drop in occupancy accuracy. Second, *RGBD Concatenation* fuses depth and RGB as a four-channel input using the depth-aware DFormer-v2 [5] backbone; although this partially alleviates misalignment, dense depth signals tend to amplify local estimation errors in texture-less or reflective regions, thereby contaminating high-level semantics. Third, *Pseudo-LiDAR Projection* voxelizes depth for LiDAR-style occupancy prediction, as illustrated in Fig. 2; this approach suffers from visibility–occupancy gaps, resulting in missing occluded structures and over-densification of visible surfaces. Overall, these comparisons indicate that naive depth injection or direct voxelization disrupts coherent feature learning, whereas our geometry-aware dual projection leverages depth voxels as adaptive geometric priors, achieving superior geometric alignment and semantic consistency.

Method	FP Depth	mIoU(%)	IoU(%)
BEVDet4D (Baseline)	DepthNet	36.01	70.36
BEVDet4D	MoGe	18.69	60.00
RGBD-based	DepthNet	22.97	43.49
Dr.Occ (Ours)	DepthNet	43.43	72.87

Table 1. **Comparison of alternative depth integration strategies on Occ3D-nuScenes.** All methods are built upon BEVDet4D with R50 backbone and 704×256 input resolution. FP Depth denotes the depth source used in forward projection.

BP	GR	SE	mIoU(%)	IoU(%)
			36.01	70.36
✓			39.08	70.52
✓	✓		41.16	71.07
✓	✓	✓	41.45	71.29

Table 2. **Ablation study on D^2-VFormer components.** BP: Backward Projection Densification; GR: depth-guided Geometric Refinement; SE: depth-guided Semantic Enhancement.

2.2. Component Analysis of D^2-VFormer

We conduct comprehensive ablation studies on the three key components of D^2-VFormer in Table 2. The results demonstrate that each component contributes progressively to performance improvement. Backward Projection Densification (BP) alone brings a substantial gain of +3.07% mIoU and +0.16% IoU, validating the effectiveness of dual projection. Adding depth-guided Geometric Refinement (GR) further improves performance by +2.08% mIoU and +0.55% IoU, showing that explicit geometric guidance enhances both semantic accuracy and occupancy prediction. Finally, incorporating Semantic Enhancement (SE) yields additional improvements of +0.29% mIoU and +0.22% IoU. The cumulative gains demonstrate that all three components effectively contribute to our final design, achieving +5.44% mIoU and +0.93% IoU over the baseline.

2.3. Runtime Analysis

As reported in Table 3, we analyze the runtime characteristics of each component in our framework. The **baseline (BL)** configuration represents the standard feature extraction and projection pipeline, while **Depth Cues (DC)** serve as the default geometric prior provided by MoGe-2 [4]. DC introduces only a minor runtime overhead of 37 ms, making it an efficient default setting for all subsequent modules. The **D^2-VFormer (D2V)** requires 144 ms due to dual-projection and depth-guided deformable fusion, reflecting a deliberate trade-off that substantially improves 3D geometric alignment and voxel representation quality. The **R-EFormer (RE)** exhibits the highest runtime (313 ms) as its multi-expert routing performs region-wise semantic rea-

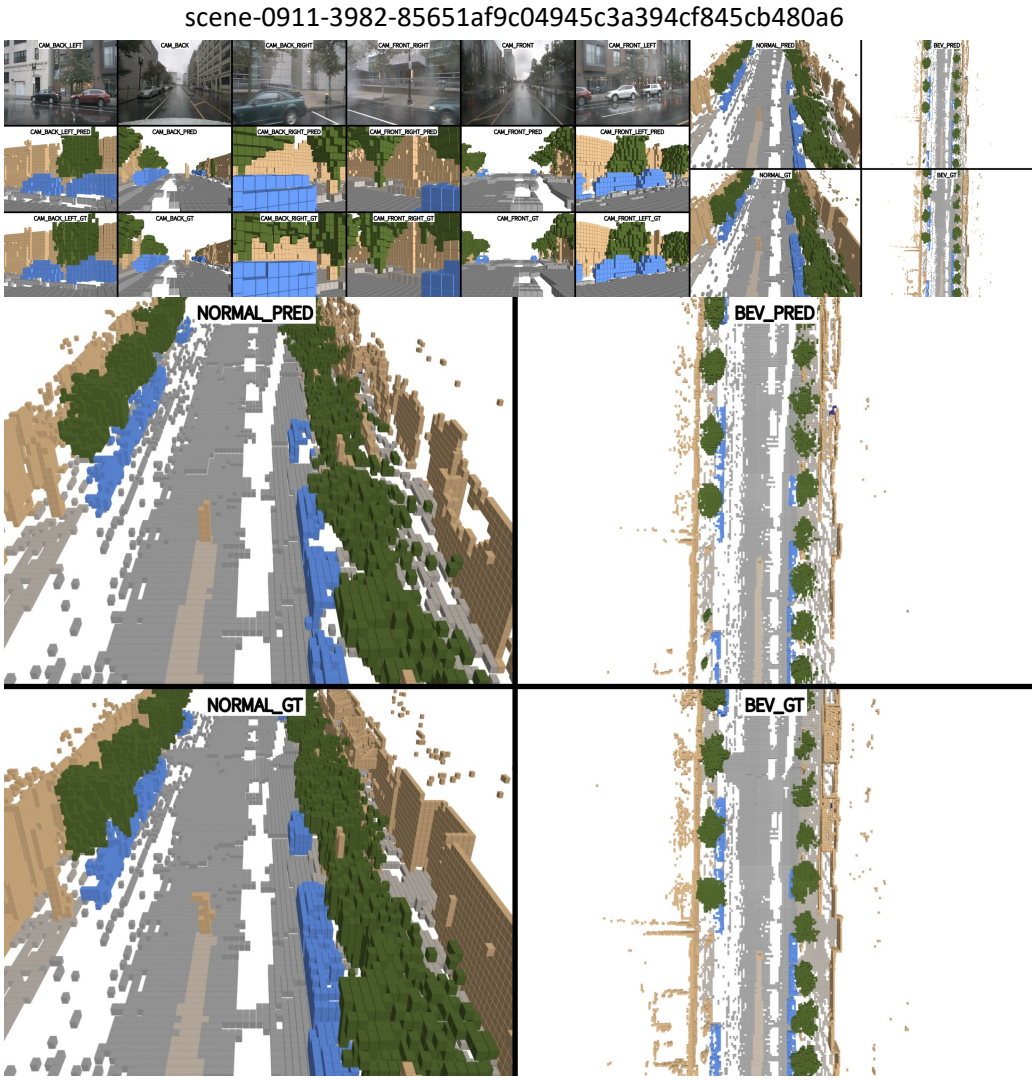

Figure 3. **Typical case.** Qualitative results show that Dr.Occ recovers road markings, vehicles, buildings, and vegetation with high fidelity, closely matching the ground truth. Readers are encouraged to zoom in to inspect fine-grained geometric and semantic details.

soning across the entire volume, leading to significant accuracy gains but a large computational burden. To address this limitation, the recursive variant **R^2-EFormer (R2E)** replaces multiple experts with a single recursively applied one, reducing runtime to 72 ms while preserving the semantic discrimination. This analysis highlights that although modules such as D2V and RE are computationally demanding, each substantially contributes to representation quality, and the recursive refinement in R2E achieves an effective balance between accuracy and efficiency.

	BL	DC	D2V	RE	R2E
Time [ms]	84	37	144	313	72

Table 3. **Runtime Performance of Each Component.** BL: Baseline; DC: Depth Cues; D2V: D^2-VFormer; RE: R-EFormer; R2E: R^2-EFormer.

Although our latency (337ms vs. 84ms) and memory usage (5.5GB vs. 2.9GB) increase, these costs remain well within the operational envelope of modern automotive SoCs (e.g., NVIDIA Orin). The substantial +7.4% mIoU gain represents a significant leap in perception quality that simple scaling the baseline cannot achieve, as efficiency-centric models often hit a performance ceiling due to geometric loss. Our model is also practical to train, requiring 38GB VRAM per GPU (BS=2). Finally, the proposed decoupled design allows heavy attention components to be replaced with hardware-friendly operators for reduced latency.

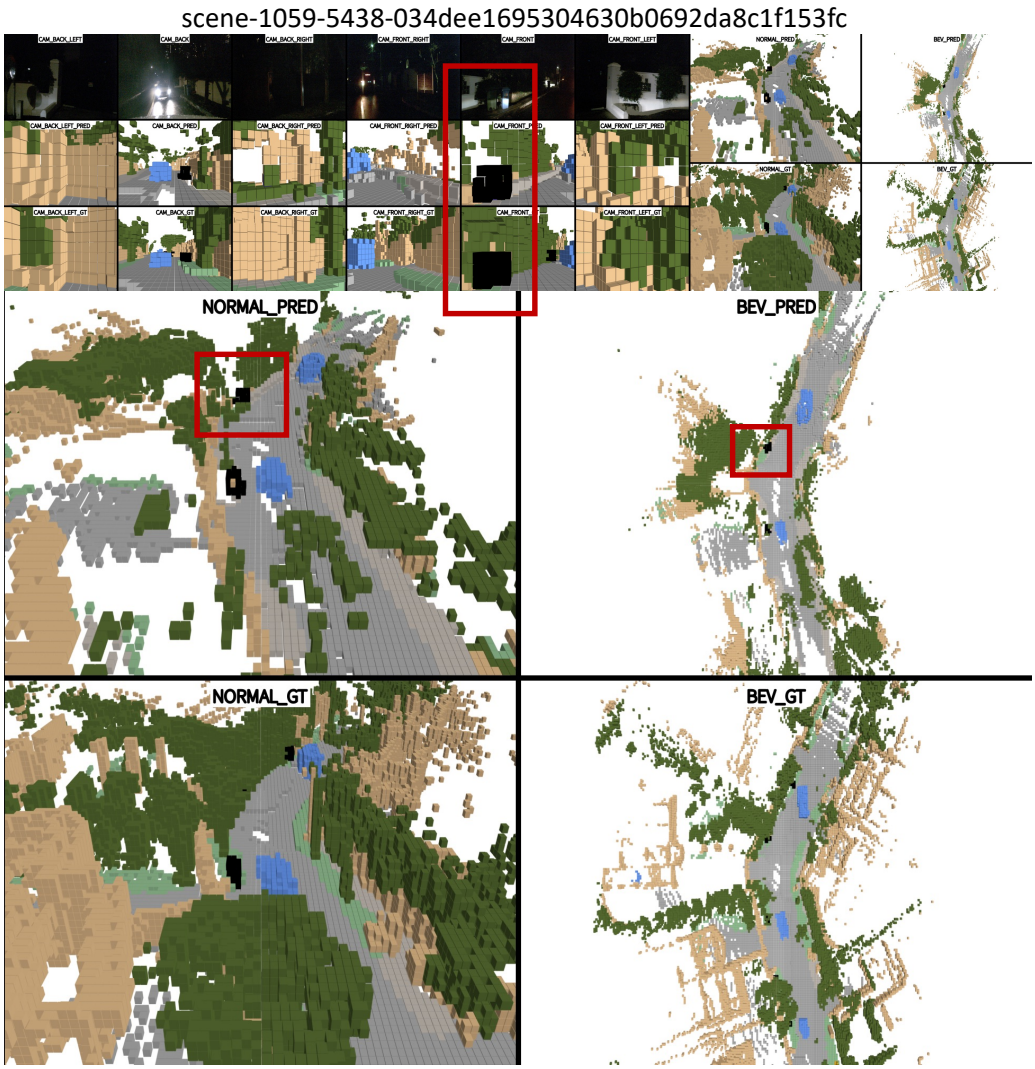

Figure 4. **Low-light failure case.** An extremely challenging nighttime scene in which the red box highlights the rare "others" class. Dr.Occ maintains reasonable predictions despite the severe illumination drop.

3. Additional Qualitative Results

In this section, we provide qualitative visualizations to illustrate the prediction quality and robustness of Dr.Occ. The first example shows a typical scene to familiarize the reader with our visualization layout, while the latter two examples highlight challenging conditions, including low-light and heavy occlusion (see Fig. 3, Fig. 4, and Fig. 5).

Typical Scene. Figure 3 presents a representative daytime scene. The top three rows show the six synchronized RGB camera views, the predicted occupancy (`Occ pred`), and the ground-truth occupancy (`Occ GT`). On the right, we visualize the scene in a BEV view and a 3D top-down perspective view, with enlarged crops displayed below for clarity. These visualizations demonstrate that Dr.Occ can produce occupancy maps that align well with the ground truth and capture fine-grained geometric and semantic details across multiple perspectives.

Low-light Scenes. We then examine an extremely challenging nighttime example (scene–1059, Fig. 4). Specifically, the red bounding box denotes the rare "others" category; despite the severely degraded illumination, Dr.Occ successfully detects this low-frequency class and generates a reasonable occupancy map. However, in completely dark image regions, the network tends to predict empty space, suggesting that additional sensing modalities (e.g., LiDAR or radar) are needed to recover missing occupancy information.

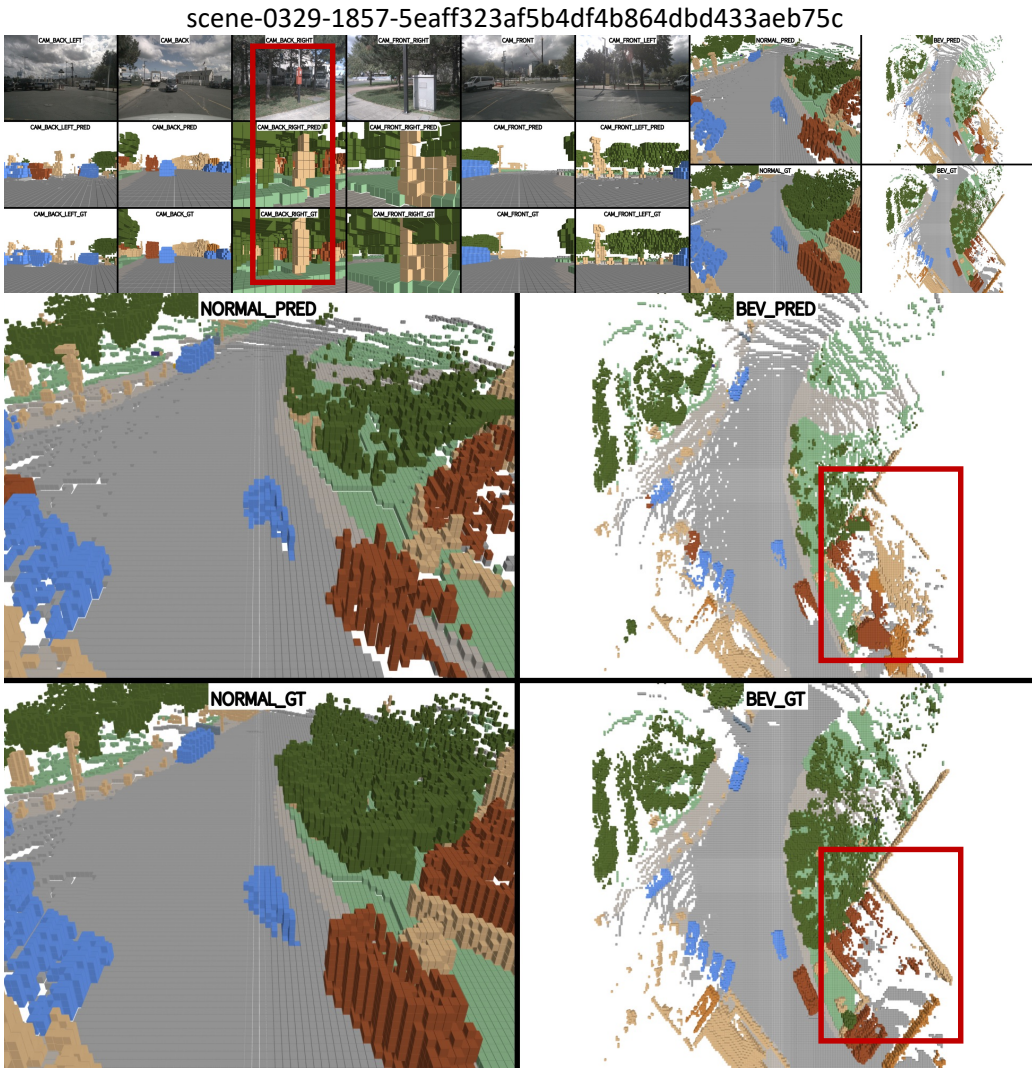

Figure 5. **Occlusion failure case.** A scene heavily blocked by trees; the model fills in sparse voxels for the visible parts but cannot recover the geometry hidden behind the foliage.

Occluded Scenes. Finally, we probe the limits of the framework under significant vegetation occlusion (scene–0329, Fig. 5). Dr.Occ reconstructs sparse voxels corresponding to the visible portions of the scene, but the details hidden behind the canopy remain unrecoverable due to the lack of visual cues. As with the low-light failure case, complete recovery would require complementary sensors such as LiDAR or radar.

Summary. Together with the results presented in the main paper, these visualizations confirm that Dr.Occ can generate reliable occupancy predictions in both typical and adverse conditions, while also making its current limitations explicit. Future work will explore multimodal fusion to address these failure scenarios.

References

[1] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 3991–4001, 2022. 2

[2] Shu Han, Xubo Zhu, Ji Wu, Ximeng Cai, Wen Yang, Huai Yu, and Gui-Song Xia. Unicalib: Targetless lidar-camera calibration via probabilistic flow on unified depth representations. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2025. 1

[3] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, pages 1477–1485, 2023. 2

[4] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv:2507.02546*, 2025. 2

[5] Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, and Qibin Hou. Dformerv2: Geometry self-attention for rgbd semantic segmentation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 19345–19355, 2025. 2

[6] Huai Yu, Xubo Zhu, Shu Han, Wen Yang, and Gui-Song Xia. I2d-locx: An efficient, precise and robust method for camera localization in lidar maps. *IEEE Robot. Autom. Lett.*, 10(8): 7899–7906, 2025. 1

[7] 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:2311.12058*, 2023. 2