3D Occupancy Prediction with Low-Resolution Queries via Prototype-aware View Transformation

Gyeongrok Oh[1,*], Sungjune Kim[1,*], Heeju Ko[1], Hyung-gun Chi[2], Jinkyu Kim[1], Dongwook Lee[3],
Daehyun Ji[3], Sungjoon Choi[1], Sujin Jang[3,†], Sangpil Kim[1,†]

[1]Korea University   [2]Purdue University   [3]AI Center, DS Division, Samsung Electronics

Abstract

The resolution of voxel queries significantly influences the quality of view transformation in camera-based 3D occupancy prediction. However, computational constraints and the practical necessity for real-time deployment require smaller query resolutions, which inevitably leads to an information loss. Therefore, it is essential to encode and preserve rich visual details within limited query sizes while ensuring a comprehensive representation of 3D occupancy. To this end, we introduce ProtoOcc, a novel occupancy network that leverages prototypes of clustered image segments in view transformation to enhance low-resolution context. In particular, the mapping of 2D prototypes onto 3D voxel queries encodes high-level visual geometries and complements the loss of spatial information from reduced query resolutions. Additionally, we design a multi-perspective decoding strategy to efficiently disentangle the densely compressed visual cues into a high-dimensional 3D occupancy scene. Experimental results on both Occ3D and SemanticKITTI benchmarks demonstrate the effectiveness of the proposed method, showing clear improvements over the baselines. More importantly, ProtoOcc achieves competitive performance against the baselines even with 75% reduced voxel resolution. Project page: https://kuai-lab.github.io/cvpr2025protoocc.

1. Introduction

3D occupancy prediction (*3DOP*) is the task of determining which parts of a 3D space are occupied by objects and identifying their class categories. In particular, view transformation from 2D to 3D space is an essential step for successful camera-based occupancy prediction. Prior works employ various 3D space representation strategies, such as bird's-eye-view (BEV) plane, tri-perspective-view (TPV) planes, and 3D voxel cube, which aim to map multi-view camera in-

---

*Equal contributions
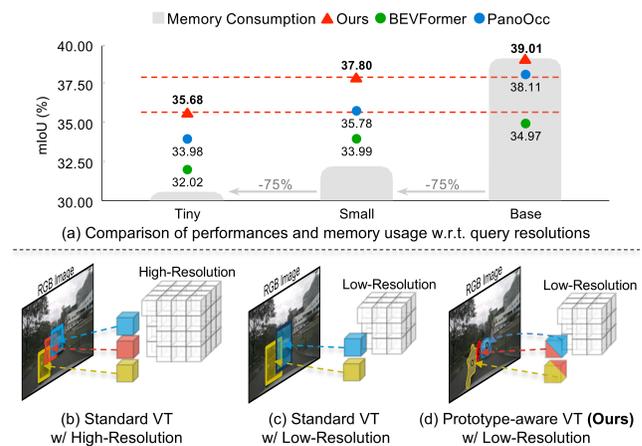†Corresponding author (s.steve.jang@samsung.com, spk7@korea.ac.kr)

Figure 1. (a) Our *ProtoOcc* can perform comparably to higher-resolution counterparts while using 75% less memory. (b-c) Reducing query resolutions in standard view transformation (VT) is required for faster inference, but brings geometrical ambiguity. (d) Our prototype-aware VT can capture high-level geometric details while preserving computational efficiency.

formation into corresponding grid cells in a unified 3D space. Among these, 3D voxels are the most common representation strategy in the 3DOP task [11, 12, 40, 43, 48, 50, 53, 55], as they naturally encode visual information into a structured semantic 3D space. Therefore, the effectiveness of 3D voxelization (*i.e.*, voxel queries) plays a crucial role in determining the prediction performance in camera-based 3DOP.

Trade-offs between computation and performance have long been a challenge in various deep learning applications [6, 8, 17, 26]. Learning voxel queries also encounters this dilemma, especially in real-time and safety-critical vision systems such as autonomous driving and robotics. As illustrated in Figure 1, utilizing voxel queries with high-resolution may produce reliable performances, but requires intensive computation. Accelerating the inference inevitably necessitates smaller query sizes. However, it results in performance degradation primarily due to its inability to capture precise high-level details within limited spatial storage,

which is crucial for 3DOP. Hence, incorporating comprehensive visual contexts into low-resolution voxel queries is essential for ensuring accurate 3DOP results while maintaining computational efficiency.

Despite its importance, existing approaches overlook the constraints posed by low-resolution queries in view transformation. They typically utilize a standard 2D-to-3D cross attention mechanism using low-level image features [12, 22, 28, 29, 47]. However, with reduced query resolutions, it becomes insufficient to solely rely on these low-level features to precisely reconstruct a 3D scene. Specifically, since images are packed into a smaller space, the encoded features lose spatial distinctiveness, thereby requiring an alternative method to preserve the necessary contextual information. In addition, it is also challenging to decode the densely compressed and low-resolution voxel queries into a sparse and high-resolution 3D occupancy scene.

To this end, we propose *ProtoOcc*, a novel occupancy network focusing on context enhancement in low-resolution view transformation for 3DOP. In particular, ProtoOcc features two main learning strategies: Prototype-aware View Transformation and Multi-perspective Occupancy Decoding.

Prototype-aware View Transformation. Image prototypes are representations of clustered image segments that provide a high-level understanding of visual structures [23, 44] (e.g. layouts and boundaries). Therefore, mapping these structural elements of 2D onto 3D voxel queries effectively alleviates the loss of spatial cues in reduced query resolutions. Furthermore, we optimize the prototype features to gain distinctiveness by leveraging pseudo-mask-based contrastive learning. As a result, each prototype becomes more exclusive and informative, enabling the voxel queries to encode more precise spatial details even with reduced resolutions.

Multi-perspective Occupancy Decoding. Reconstructing 3D occupancy from low-resolution queries is an ill-posed problem, requiring precise disentanglement of densely encoded visual details. Therefore, we propose a multi-perspective decoding strategy using voxel augmentations. Each augmentation is upsampled as a unique view of the scene, collectively forming a comprehensive representation of the 3D structure. Then, we apply a consistency regularization on different perspectives to ensure that they represent the same 3D scene, which brings robustness in training.

Through extensive experiments on the Occ3D-nuScenes [39] and SemanticKITTI [2] benchmarks, we validate that the proposed ProtoOcc can be an effective solution for ensuring accurate 3DOP results while maintaining computational efficiency. Most importantly, ProtoOcc achieves competitive performance with even 75% smaller voxel queries against models with larger-sized ones.

In summary, the contributions of this work include:

- Introducing ProtoOcc as an exemplar in 3DOP by using computationally efficient low-resolution voxel queries.

- A prototype-aware view transformation and multi-perspective decoding strategy for enhancing the representations of low-resolution voxel queries.
- Demonstrating clear improvements over previous state-of-the-art methods in two major benchmarks, along with detailed analyses.

2. Related Work

2.1. Camera-based 3D Occupancy Prediction

The task of camera-based 3D occupancy prediction (3DOP) originated from Semantic Scene Completion (SSC) [3, 14, 19, 38, 54], which leverages a single monocular image or sparse LiDAR supervision of the SemanticKITTI benchmark [2] to reconstruct a static 3D scene. The monocular camera-based methods MonoScene [3] and VoxFormer [21] serve as the foundation for further research expansion into multi-view images. Following the previous advances, Occ3D [39] recently released a benchmark for dynamic 3DOP using multi-view camera images, triggering the popularity of camera-based 3DOP [11, 12, 40, 43, 48, 50, 55].

Despite their potential, the tremendous computation load of intermediate voxel representations poses a significant challenge. Several approaches [25, 32, 39] sparsify the voxel queries based on occupancy scores. A recent work COTR [30] first embeds 2D observations into a larger-sized voxel queries and further downsamples them into a compact representation. However, our proposed ProtoOcc directly exploits a smaller-sized query for view transformation. Our prototype-aware view transformation and multi-perspective voxel decoding strategies on low-resolution queries efficiently improve both computation and performance.

2.2. 2D-to-3D View Transformation

View transformation is an indispensable process in camera-based perception models, bridging the cross-view environment. LSS [33] achieves this by utilizing camera parameters to estimate the depth distribution from an image. The strategy is adopted from diverse 3D downstream tasks, demonstrating its effectiveness [9, 10, 20, 30]. However, the reliance on depth estimation causes geometric uncertainty, thus presenting a fundamental challenge. The attention-based approach is another line of view transformation that effectively aggregates 2D spatio-temporal features [15, 22, 36]. Here, learnable queries with diverse configurations [12, 22, 28] attend on specific regions in an image and transfers 2D semantics into a unified 3D space. In 3DOP, 3D voxel queries are routinely exploited for their ability to preserve as much spatial information of a 3D scene [39, 48].

ProtoOcc aims at providing high-level 2D contextual guidance in attention-based view transformation, specifically targeted at enhancing the representations of low-resolution voxel queries for efficient 3DOP. Several attempts have been
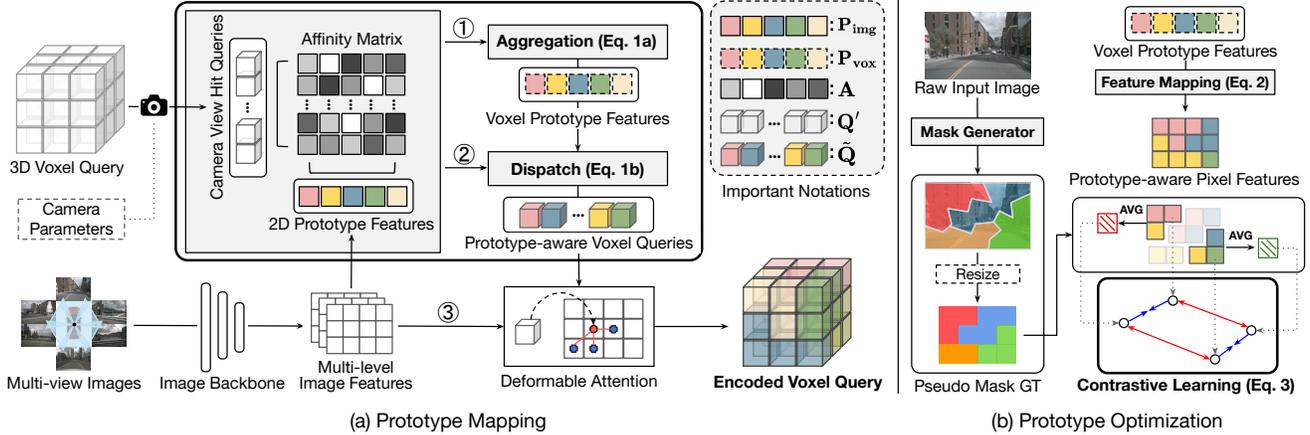
Figure 2. **Prototype-aware View Transformation.** (a) In the Prototype Mapping stage, we fully exploit the hierarchies of 2D image features via a clustering method to map 2D prototype representations onto 3D voxel query. (b) Contrastive learning on the prototype features based on the pseudo ground truth masks enhances the discrimination between the prototypes for better feature learning. Best viewed in color.

previously made in 3D object detection task, to improve representations in view transformation [18, 49]. However, these methods mainly focus on enhancing object-wise representations, which makes them unsuitable for 3DOP, which requires holistic understanding of a scene.

3. Method

ProtoOcc tackles the information loss problem occurring from reduced query resolutions through two main strategies: Prototype-aware View Transformation and Multi-perspective Occupancy Decoding. In this section, we first go over the overview of ProtoOcc (Sec. 3.1), then explain the technical details of each proposed method (Sec. 3.2 and 3.3).

3.1. Overview

Given N multi-view images $\{\mathcal{I}_i\}_{i=1}^N$, camera-based 3D occupancy prediction aims to reconstruct a 3D occupancy scene $\mathcal{O} \in \mathbb{R}^{L \times H \times W \times Z}$, where H, W, and Z represent the spatial dimensions and L is the semantic label. Specifically, the image backbone network (e.g. ResNet50 [7]) first extracts multi-scale image features $\{\{\mathcal{F}_i^{(t)}\}_{i=1}^N\}_{t=1}^T$ using Feature Pyramid Network (FPN) [24]. Here, T denotes the total number of different feature scales, and each feature has a channel size of d. Then, these 2D features across different views are converted into a unified 3D representation by employing a learnable 3D voxel query $\mathbf{Q} \in \mathbb{R}^{d \times h \times w \times z}$. $\mathbf{Q}' \in \mathbb{R}^{N \times K \times d}$ denotes a subset of $\mathbf{Q}$, where K represents the maximum number of hit queries on each N camera views ($0 \leq K \leq h \times w \times z$). Following the attention-based view transformation [22, 39, 48], the queries aggregate spatial and temporal cue of the surrounding 3D scene via deformable attention [56]. Finally, the voxel upsampling network restores high-resolution voxel volume from the encoded voxel query and predicts an occupancy state and

semantic label for each voxel.

3.2. Prototype-aware View Transformation

In this stage, we encode prototype representations of 2D images into 3D voxel queries, which is an unexplored concept in the previous literature. Prototypes are grouped representations of image clusters that integrate semantically similar features [1, 13, 52]. Inspired by this principle, we enrich the 3D voxel query representations by mapping hierarchical image representation encompassing both raw and clustered features, thereby compensating for the loss of spatial cues at reduced resolutions. We provide an illustration of our method in Figure 2.

3.2.1. Prototype Mapping

Mapping image prototype representations onto 3D voxel queries starts from clustering the images. We adopt an iterative grouping strategy [1, 4, 13, 52] to segment the image feature space. Specifically, we first divide the feature map into regular grids and obtain initial prototype representation by calculating the average within each grid. Then, we iteratively update this feature by multiplying the soft-assigned pixel-prototype similarity with the feature map. The acquired 2D prototype features are denoted as $\mathbf{P}_{\text{img}} \in \mathbb{R}^{N \times M \times d}$, where M is the number of prototypes.

The next step is to lift these 2D prototype features $\mathbf{P}_{\text{img}}$ into the 3D voxel query space. In this process, we introduce an innovative application of *Feature Aggregation & Dispatch* technique [23, 31]. Originally, the idea is introduced as a feature extraction paradigm in 2D image space. In this work, we utilize this as a mapping function from a 2D prototype feature space to a 3D voxel query space. We start by projecting $\mathbf{P}_{\text{img}}$ to 3D voxel feature space using MLP layers and obtain $\hat{\mathbf{P}}_{\text{img}}$. Then, we compute the affinities $\mathbf{A} \in \mathbb{R}^{N \times M \times K}$ via pairwise cosine similarity between

$\hat{\mathbf{P}}_{\text{img}}$ and $\mathbf{Q}'$. Using the sigmoid function σ, the similarity values are further re-scaled to range between $(0, 1)$, indicating the probability of each query cell in $\mathbf{Q}'$ being assigned to a certain prototype. Multiplying these probability with the voxel query feature $\mathbf{Q}'$ produces 3D voxel prototype feature $\mathbf{P}_{\text{vox}} \in \mathbb{R}^{N \times M \times d}$ (*Aggregation*). Then, these features are redistributed to individual voxel queries to form a prototype-aware 3D voxel query $\tilde{\mathbf{Q}} \in \mathbb{R}^{N \times K \times d}$ (*Dispatch*). The overall process is formulated as follows:

$$\texttt{Aggregate: } \mathbf{P}_{\text{vox}} = \frac{1}{R}\Big(\hat{\mathbf{P}}_{\text{img}} + \sigma(\mathbf{A}) \cdot \mathbf{Q}'\Big) \quad \text{(1a)}$$

$$\texttt{Dispatch: } \tilde{\mathbf{Q}} = \mathbf{Q}' + \texttt{MLP}\Big(\sigma(\mathbf{A})^{\mathsf{T}} \cdot \mathbf{P}_{\text{vox}}\Big), \quad \text{(1b)}$$

where $\hat{\mathbf{P}}_{\text{img}}$ and $\mathbf{Q}'$ are added as residual connection for *Aggregation* and *Dispatch*, respectively. R is the total sum of the similarities $\sigma(\mathbf{A})$, which is divided to ensure a stable training process. Subsequently, the prototype-aware voxel query $\tilde{\mathbf{Q}}$ and the multi-scale image feature $\mathcal{F}$ are fed as inputs for deformable attention. Owing to this strategy, the encoded 3D voxel query carries both high-level and fine-grained visual contexts that are highly essential for predicting 3D occupancy.

3.2.2. Prototype Optimization

The prototype-aware 3D voxel query contains rich 2D contextual information, yet the quality of these features highly depends on 2D clustering quality. However, standard optimization objectives in 3DOP (e.g., $\mathcal{L}_{\text{occ}}$ and $\mathcal{L}_{\text{Lov}}$) do not provide suitable guidance for cluster learning. Therefore, we incorporate an explicit pseudo 2D supervision, by a contrastive learning between the prototype-aware pixel features in accordance with pseudo ground truth masks.

Prototype-aware Pixel Features. These features, denoted as $\mathbf{X} \in \mathbb{R}^{N \times d \times h' \times w'}$, are the mapping of 3D voxel prototype features $\mathbf{P}_{\text{vox}}$ onto an implicit 2D grid cell. For simplicity, we flatten the spatial dimension and denote it as D in the following (i.e., $D = h' \times w'$). Since $\mathbf{P}_{\text{vox}}$ itself does not contain any spatial information, a direct mapping between the two spaces is limited. Therefore, we first employ a deformable attention map [35] $\mathbf{G} \in \mathbb{R}^{N \times D}$, which highlights the salience regions on the image features for each 3D voxel queries. Then, the element-wise multiplication between $\mathbf{G}$ and the prototype-query affinity matrix $\mathbf{A}$ computes the probability of each grid cell being assigned with a certain prototype. This acts as the spatial bridge for mapping $\mathbf{P}_{\text{vox}}$ onto the 2D grid cell. Consequently, through a matrix multiplication between this feature and $\mathbf{P}_{\text{vox}}$, we obtain the prototype-aware pixel features. The explained mapping procedure can be formulated as:

$$\mathbf{X} = \big\{\mathbf{G} \odot \mathcal{H}(\mathbf{A})\big\} * \mathbf{P}_{\text{vox}}, \quad \text{(2)}$$

where $\mathcal{H}$ is a linear mapping function and $\odot$ is the element-wise multiplication. Through this proposed methodology,

each feature in $\mathbf{X}$ stores the prototype features from the 3D voxel space, yet precisely aligned within the 2D pixel grid.

Contrastive Learning with Pseudo Masks. The absence of explicit ground truth to contrast the prototype-aware pixel features remains a challenge. Therefore, we generate S pseudo ground truth masks by applying either a classical clustering algorithm [41] or visual foundation model [16] on the image. This is further resized to match the spatial dimension of $\mathbf{X}$ to provide distinct spatial boundaries. Then, we compute S mask centroids by averaging the features of $\mathbf{X}$ that fall within each boundary. These mask centroids, denoted as $\{\mathbf{M}_i\}_{i=1}^{S} \in \mathbb{R}^{d}$, serve as the anchor for our contrastive learning. We calculate the contrastive loss [42, 45] for the feature in the (x, y) position of $\mathbf{X}$ as:

$$\mathcal{L}_{\text{cls}}^{(x,y)} = -\log \frac{\sum_{s=1}^{S} m_s \exp\left(\langle \mathbf{M}_s, \mathbf{X}_{(x,y)}\rangle / \tau_{\text{cls}}\right)}{\sum_{s=1}^{S} \exp\left(\langle \mathbf{M}_s, \mathbf{X}_{(x,y)}\rangle / \tau_{\text{cls}}\right)}, \quad \text{(3)}$$

where $\langle \cdot, \cdot \rangle$ represents the cosine similarity between the two elements, and τ_{cls} is the temperature hyperparameter. The binary value $m_s \in \{0, 1\}$ is set as 1 if the position (x, y) lies within the boundary of sth mask. By summing up the losses across all grid cells, we obtain the final contrastive loss (i.e. $\mathcal{L}_{\text{cls}} = \sum_{x=1}^{h'} \sum_{y=1}^{w'} \mathcal{L}_{\text{cls}}^{(x,y)}$), which enhances the distinctiveness of the prototype features when minimized.

3.3. Multi-Perspective Occupancy Decoding

When reconstructing 3D occupancy from low-resolution queries, the lack of resolution causes inevitable geometrical ambiguity. That being said, an encoded query has the potential to be decoded in diverse 3D perspectives. We address this ill-posed property by enhancing contextual diversity in occupancy decoding through two essential techniques: 1) Multi-perspective View Generation and 2) Scene Consistency Regularization.

3.3.1. Multi-perspective View Generation

We generate multi-perspective contexts from voxel queries through augmentations. Since augmentation on voxel representations is not widely explored, we establish two categories of voxel augmentation: 1) a feature-level augmentation (e.g. Random Dropout and Gaussian Noise), and 2) a spatial-level augmentation (e.g. Transpose and Flips). Combining the augmentations from these categories, we obtain a query set $\mathbb{Q} = \{\mathbf{Q}^{(0)}, \mathbf{Q}^{(1)}, \dots, \mathbf{Q}^{(P-1)}, \mathbf{Q}^{(P)}\}$, where $\mathbf{Q}^{(0)}$ denotes the original voxel query derived from the encoder, and P is the number of voxel augmentations. Then, these queries pass through a transposed 3D convolution layer with shared weights, generating an upsampled query set $\mathbb{V} = \{\mathbf{V}^{(0)}, \mathbf{V}^{(1)}, \dots, \mathbf{V}^{(P-1)}, \mathbf{V}^{(P)}\}$, where $\mathbf{V} \in \mathbb{R}^{L \times H \times W \times Z}, \forall \mathbf{V} \in \mathbb{V}$. For each grid cell position, the shared kernel convolves with diverse features of local voxel neighbors. As a result, the upsampled voxels can be inter-

preted as unique perspective of a scene, collectively forming a holistic representation of the 3D structure.

3.3.2. Scene Consistency Regularization

Although the upsampled query set $\mathbb{V}$ contains varying contextual perspectives, they should all depict the same semantic occupancy. Therefore, we couple with the augmentations a regularization term to maintain a semantically consistent prediction. Specifically, we adopt a simple yet effective regularization technique of GRAND [5], and minimize the L_2 distances between the predicted label distributions and their average distribution for each grid cell. For example, the average of the predicted label distribution in the (i, j, k) position of the voxels is calculated as $\hat{\mathbf{V}}_{(i,j,k)} = \frac{1}{P+1} \sum_{p=0}^{P} \mathcal{G}(\mathbf{V}_{(i,j,k)}^{(p)})$, where $\mathcal{G}(\cdot)$ is the label classifier network. Then, this distribution is sharpened as:

$$\tilde{\mathbf{V}}_{(i,j,k)}[c] = \hat{\mathbf{V}}_{(i,j,k)}^{\frac{1}{\tau_{\text{cons}}}}[c] \bigg/ \sum_{l=1}^{L} \hat{\mathbf{V}}_{(i,j,k)}^{\frac{1}{\tau_{\text{cons}}}}[l], (1 \le c \le L),$$

$$(4)$$

which denotes the guessed probability on the cth class, with τ_{cons} as the temperature hyperparmeter that controls the sharpness of the categorical distribution. Accordingly, the final consistency regularization loss $\mathcal{L}_{\text{cons}}$ is obtained by taking the average of the distances between the sharpened average and each prediction across all grid cells and augmentations:

$$\mathcal{L}_{\text{cons}} = \frac{\sum_{p=0}^{P} \sum_{k=1}^{Z} \sum_{j=1}^{W} \sum_{i=1}^{H} \left\| \tilde{\mathbf{V}}_{(i,j,k)} - \mathcal{G}(\mathbf{V}_{(i,j,k)}^{(p)}) \right\|_2^2}{H \cdot W \cdot Z \cdot (P+1)}. \quad (5)$$

3.4. Optimization

ProtoOcc minimizes four loss functions for each objectives: 1) $\mathcal{L}_{\text{occ}}$ for accurate semantic label prediction, 2) $\mathcal{L}_{\text{Lov}}$ as an complementary loss for maximizing the mean Intersection over Union (mIoU), 3) $\mathcal{L}_{\text{cls}}$ for enhancing the prototype feature quality, and lastly 4) $\mathcal{L}_{\text{cons}}$ for regularizing the outputs from diverse contextual perspectives. The final objective function of the model is defined as:

$$\mathcal{L}_{\text{total}} = \sum_{p=0}^{P} (\lambda_1 \mathcal{L}_{\text{occ}}^{(p)} + \lambda_2 \mathcal{L}_{\text{Lov}}^{(p)}) + \lambda_3 \mathcal{L}_{\text{cls}} + \lambda_4 \mathcal{L}_{\text{cons}}, \quad (6)$$

where λ_1, λ_2, λ_3, and λ_4 are coefficient hyperparameters that determine the weights of four losses. The values we set for each coefficient are specified in Appendix B.1.

4. Experiments

4.1. Experiment Setup

Datasets and Evaluation Metrics. We train and evaluate our method on two representative tasks in 3D scene understanding: 3D occupancy prediction (3DOP) and 3D semantic scene completion (3DSSC). For 3DOP, we utilize the

Occ3D-nuScenes [39] benchmark. The qualitative result is measured with a mean Intersection-over-Union (mIoU) metric. For 3DSSC, we use the SemanticKITTI [2] benchmark, and measure IoU and mIoU for evaluation. More details of the datasets can be found in Appendix A.

Implementation Details. As our primary goal is to observe the influences of voxel query resolutions on view transformation, categorize the experimental setting based on query resolutions. For Occ3D-nuScenes, we make three variants: `Base`, `Small` and `Tiny`. `Base` follows the standard baseline settings: a query size of 200×200 for BEVFormer and $100 \times 100 \times 16$ for PanoOcc and our ProtoOcc. For `Small`, we use a query size of 100×100 for BEVFormer, and $50 \times 50 \times 16$ for PanoOcc and ProtoOcc. Lastly for `Tiny`, we further reduce the resolution into 50×50 for BEVFormer, and $50 \times 50 \times 4$ for PanoOcc and ProtoOcc. All three categories are trained with 432×800 sized images for 12 epochs. For SemanticKITTI, we make two variants: `Base` and `Small`. Here, we observe how our proposed methods can be applied and improve existing baselines: VoxFormer [21] and Symphonies [14]. A query size of $128 \times 128 \times 16$ is used for `Base`, and $64 \times 64 \times 8$ is used for `Small` for all baselines. More implementation details are reported in Appendix B.1.

4.2. Quantitative Analysis

3D Occupancy Prediction. Table 1 presents a quantitative comparison of 3D occupancy prediction between baselines and ProtoOcc. Here, we observe two key insights. First, ProtoOcc achieves the highest mIoU in all query settings. This strongly demonstrates the effectiveness of voxel representation learning through our proposed methods. Specifically, ProtoOcc excels in predicting important road agents (e.g. pedestrian, car, and bus). Furthermore, by comparing the results within the same color marks (● and ●), we notice that ProtoOcc achieves competitive results against the baselines with higher-resolution queries. For example, with 75% fewer parameters in view transformation, ProtoOcc achieves performance comparable to PanoOcc [48] in the larger resolution category. The gaps in mIoU are marginal (0.31 in `Small`-`Base` and 0.1 in `Tiny`-`Small`), and our ProtoOcc even outperforms predictions in a number of semantic classes.

3D Semantic Scene Completion. Table 2 reports the results of 3D semantic scene completion with SemanticKITTI benchmark. As clearly seen from the table, our method can also bring benefits in scene understanding as a plug-and-play. Applying our method enhances the performance of both baselines in all metrics. It is important to note that, except for IoU in VoxFormer, all the smaller-resolution models surpass its larger variant when our ProtoOcc is combined. This implies that ProtoOcc overcomes 87.5% deficiency of spatial storage through its prototype-aware view transformation and multi-perspective occupancy decoding.

Table 1. **Quantitative results on the Occ3D-nuScenes [39] validation set.** We highlight the best and runner-up results for each category in **bold** and plain, respectively. Not only does ProtoOcc stand out in its category, but also by comparing the results within the same color marks (● and ●), it is apparent that ProtoOcc can overcome query deficiencies, performing on par even with higher-resolution counterparts.

Query Size	Model	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	mammade	vegetation
Base	CTF-Occ [39]	28.50	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.00
	TPVFormer [12]	34.20	7.68	44.01	17.66	40.88	46.98	15.06	20.54	24.69	24.66	24.26	29.28	79.27	40.65	48.49	49.44	32.63	29.82
	SurroundOcc [50]	34.60	9.51	38.50	22.08	39.82	47.04	20.45	22.48	23.78	23.00	27.29	34.27	78.32	36.99	46.27	49.71	35.93	32.06
	OccFormer [55]	37.04	9.15	45.84	18.20	42.80	50.27	24.00	24.00	22.86	20.98	31.94	38.13	80.13	38.24	50.83	54.3	**46.41**	**40.15**
	BEVFormer● [22]	34.97	7.53	41.77	16.39	44.06	48.48	17.27	20.01	23.36	21.16	28.88	35.59	80.12	35.35	47.65	51.89	40.68	34.28
	PanoOcc● [48]	38.11	9.75	45.31	22.45	43.13	50.19	22.25	27.35	24.49	**25.17**	31.74	37.95	81.74	42.29	50.82	54.80	40.81	37.14
	ProtoOcc (Ours)	**39.01**	9.75	**46.08**	**24.34**	**46.09**	**52.45**	24.21	**28.11**	24.72	19.79	**32.90**	**40.50**	**82.29**	**43.02**	**52.47**	**55.94**	42.46	38.13
Small	BEVFormer● [22]	33.98	6.75	41.67	13.91	41.97	48.49	17.83	18.01	22.19	19.08	29.64	33.23	79.42	36.48	46.82	49.26	39.04	33.91
	PanoOcc● [48]	35.78	8.18	41.60	20.79	41.25	47.78	21.87	23.42	21.03	19.29	29.71	36.10	81.20	40.00	49.22	53.94	38.09	34.83
	ProtoOcc (Ours)●	**37.80**	9.28	43.64	22.30	44.72	50.07	23.68	25.23	22.77	19.66	30.43	38.73	82.05	42.61	51.68	55.84	41.91	38.05
Tiny	BEVFormer [22]	32.02	4.86	39.79	7.17	42.46	47.10	18.46	13.18	17.76	12.46	28.74	33.19	78.64	35.36	45.27	47.29	38.93	33.61
	PanoOcc [48]	33.99	6.97	39.60	18.80	40.67	45.63	18.19	21.43	19.10	16.53	25.99	35.15	80.60	38.44	49.02	52.11	36.81	32.87
	ProtoOcc (Ours)●	**35.68**	8.33	40.55	19.84	42.95	48.08	20.31	22.78	21.21	17.00	28.22	36.60	81.42	41.32	50.16	53.82	38.63	35.35

Table 2. **Quantitative results on the SemanticKITTI [2] validation set.** Results of each semantic class are provided in Appendix D. Performance increases through our method are denoted in blue.

Model	Pub.	IoU	mIoU
MonoScene [3]	CVPR 22	36.86	11.08
TPVFormer [12]	CVPR 23	35.61	11.36
OccFormer [55]	ICCV 23	36.50	13.46
HASSC [46]	CVPR 24	44.82	13.48
VoxFormer-S [21]	CVPR 23	43.10	11.51
+ ProtoOcc	-	43.55 (+0.35)	12.39 (+0.88)
VoxFormer-B [21]	CVPR 23	44.02	12.35
+ ProtoOcc	-	**44.90** (+0.85)	13.57 (+1.22)
Symphonies-S [14]	CVPR 24	41.67	13.64
+ ProtoOcc	-	43.02 (+1.35)	14.50 (+0.86)
Symphonies-B [14]	CVPR 24	41.85	14.38
+ ProtoOcc	-	42.12 (+0.27)	**14.83** (+0.45)

Table 3. Comparison of Computational Efficiency.

Model	Inf. Time	FLOPs (G)	Params	mIoU
PanoOcc-B [48]	266 ms	1,310	46.24M	38.11
PanoOcc-S	66 ms	330	15.52M	35.78
+ DFA3D [18]	153 ms	360	22.33M	36.27
ProtoOcc–S (Ours)	105 ms	378	16.11M	37.80

View Transformation Efficiency. Table 3 compares the cost of inference time, FLOPs, and the number of parameters during the view transformation stage. Initially, by comparing PanoOcc-Base and PanoOcc-Small, we observe huge savings from a reduced query resolution. Next, we compare two view transformation methods introduced from DFA3D [18] and ProtoOcc on the Small query setting. Although additional processing brings inevitable increases in computations, our method shows significant computational efficiency compared to DFA3D, as it produces better results with less costs in inference time and number of parameters. Above all, our ProtoOcc-Small produces comparable mIoU with PanoOcc-Base while requiring 60.53%, 71.15%, and 65.16% less computational costs in each metric, respectively.

4.3. Qualitative Analysis

Predictions on Challenging Scenarios. In this section, we visualize the prediction results on some challenging sce-narios that the numerical results were unable to highlight. In the top example of Figure 3, the labels for approaching vehicles in the distance are absent in the ground truth. Accordingly, the baselines fail in predicting the occupancy of the corresponding region. However, our high-level 2D prototype-aware encoding enables ProtoOcc to capture these long-range objects, even without ground-truth guidance. Furthermore, the bottom example emphasizes the superiority of ProtoOcc in occluded regions. We consider that this ability is developed through the multi-perspective learning of 3D space. More examples are provided in Appendix D.2.

Attention Map Visualization. Attention map is a crucial indicator of whether the 2D-to-3D view transformation has been successful. Figure 4 compares how different qualities of attention maps can lead to substantial differences in predicted 3D occupancies. The high-level 2D prototypes guide ProtoOcc to distinguish the scene components with clear boundaries. This naturally enables the model to identify important parts in the image to focus on with fine granularity. For example, ProtoOcc captures small yet crucial scene components (e.g. pedestrian and motorcycle in Figure 4) and assigns high attention weights, which leads to accurate predictions. However, the baseline fails to capture the details and result in prediction misses that jeopardize the safety of the driving environment.

Feature Map Visualization. The prototype-aware view transformation forms a cluster per each significant region in

FRONT LEFT **FRONT** **FRONT RIGHT** **BACK LEFT** **BACK** **BACK RIGHT**

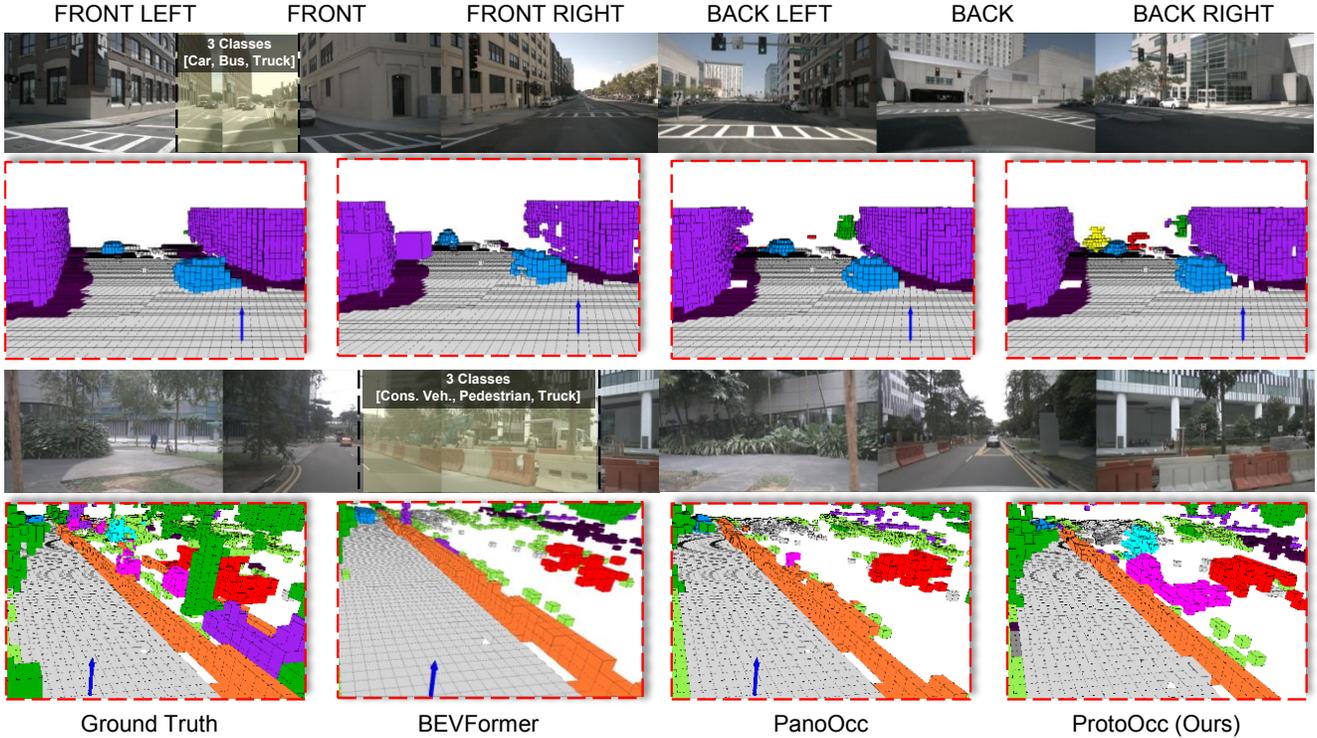Ground Truth · BEVFormer · PanoOcc · ProtoOcc (Ours)

Figure 3. **Predictions in challenging scenarios.** We visualize the prediction comparisons between the baseline and our ProtoOcc. The corresponding camera view is highlighted in yellow, and we label important semantic classes on top. Best viewed in color.

Table 4. Ablation on Model Design.

	Method			
Exp. #	Proto. Mapping	Proto. Optimization	MOD	mIoU
1	-	-	-	35.78
2	✓	-	-	35.80
3	✓	✓	-	36.55
4	-	-	✓	37.25
5	✓	✓	✓	37.80

an image. Owing to this strategy, the encoding stage gains flexibility, by not being overwhelmed by visually dominant regions. Figure 5 visualizes the feature map of voxel queries, which are flattened in the channel and the z-axis. As clearly seen in the illustration, the features of the baseline are biased towards visually dominant elements in the scene, such as the yard or building. In contrast, applying our prototype-aware view transformation on the baseline effectively enhances the recognition of smaller, yet crucial components for safe driving, such as vehicles.

4.4. Ablation Study

Model Design. Table 4 shows the effects of each method proposed in our ProtoOcc. Exp. 1 is a baseline model without any contribution of our method. When our Prototype Mapping technique is applied (Exp. 2), we observe a little increase in the performance. However, coupling with Proto-

type Optimization brings considerable improvements (Exp. 3). This implies that the explicit objective to refine the 2D prototype quality is highly beneficial for 3D voxel clustering, and as a result helps to predict accurate semantic occupancy. In Exp. 4, we apply the Multi-perspective Occupancy Decoding (MOD) method on the baseline, which shows great enhancements just by itself. This demonstrates that the contextual diversity is crucial for predicting higher-dimension 3D occupancy with lower-dimension 2D images. Lastly, the full ProtoOcc model (Exp. 5) leverages the synergies of each contribution, resulting in superior performance.

Prototype Analysis. In this section, we deeply analyze the Prototype-aware View Transformation through three experiments: 1) sensitivity on number of 2D prototypes M, 2) choice for pseudo mask generator, and 3) impact of granularity level. In the first experiment, we fix the pseudo mask generator as SEEDS [41] and observe the performance sensitivity on M, which is determined by the pre-defined downsampling ratio r. As shown in Table 6, we discover that overly fine ($r=2$) or coarse ($r=8$) clustering leads to suboptimal performance and that $M=350$ with $r=4$ is the adequate number of prototypes. Second, Table 7 shows that the pseudo mask generator SAM [16], which generates more semantically consistent clusters, outperforms the traditional superpixel algorithm SEEDS under consistent number of pseudo masks. Lastly, however, we observe in Table 8 that

Table 5. Effects of different augmentations and their consistency regularization (C.R.).

	Exp. #	1	2	3	4	5	6	7	8	9	10	11
Augmentation	Random Dropout	-	✓	-	-	-	✓	✓	-	-	✓	-
	Gaussian Noise	-	-	✓	-	-	-	-	✓	✓	✓	-
	Transpose	-	-	-	✓	-	✓	-	✓	-	-	✓
	Flips	-	-	-	-	✓	-	✓	-	✓	-	✓
mIoU	w/o C.R.	35.78	36.99	36.46	36.49	36.70	36.95	36.78	36.47	36.72	36.66	36.59
	w/ C.R.	N/A	**37.19**	**36.80**	**36.86**	**37.18**	**37.21**	**37.16**	**36.90**	**36.82**	**37.25**	**36.78**

Table 6. Sensitivity on the number of 2D prototypes M.

2D prototype M	1440 ($r = 2$)	350 ($r = 4$)	144 ($r = 6$)	91 ($r = 8$)
# Pseudo Mask		299		
mIoU	36.90	37.80	36.95	37.07

Table 7. Effect of pseudo mask quality.

2D prototype M	Mask Generator	# Pseudo Mask	mIoU
350	SEEDS [41]	91	37.14
	Segment Anything [16]	92	37.29

Table 8. Impact of the granularity level.

2D prototype M	Mask Generator	# Pseudo Mask	mIoU
91	SEEDS [41]	299	37.07
	Segment Anything [16]	92	37.21
350	SEEDS [41]	299	37.80
	Segment Anything [16]	92	37.29

aligning the consistent number of 2D prototypes M with the number of pseudo mask is more crucial for facilitating the Prototype Optimization.

Augmentations and Consistency Regularization. ProtoOcc introduces both feature- and spatial-level voxel augmentation methods for diversifying 3D contexts. Table 5 shows the influences of each augmentation and their combinations, with and without scene consistency regularization. Exp. 1 is the baseline model just as in Table 4. The results of the experiment highlight three key insights. First, the augmentation itself brings a considerable amount of benefits, demonstrating the importance of contextual diversity for occupancy decoding. Secondly, further regularization on their output consistency guarantees they represent the same scene, thereby naturally boosting robustness. Lastly, incorporating numerous augmentations for diversity can occasionally hinder robust feature learning and doesn't consistently lead to proportional performance improvements. For this reason, we conduct empirical experiments with no more than two combinations of augmentations.

5. Conclusion

This paper introduces ProtoOcc, an innovative method addressing the challenge of enriching contextual represen-
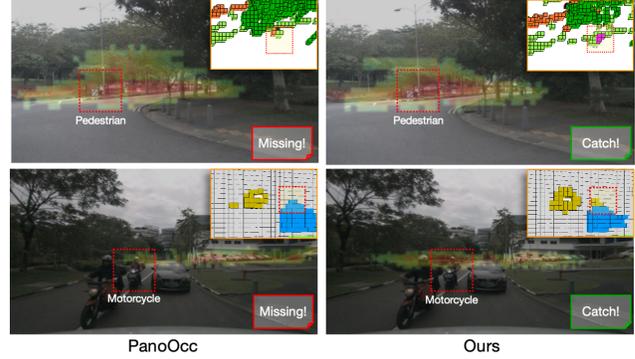

Figure 4. **Attention map visualization.** Compared to the baseline, ProtoOcc can attend to more important details in the image (e.g. red dashed boxes), which is crucial for safe driving systems.
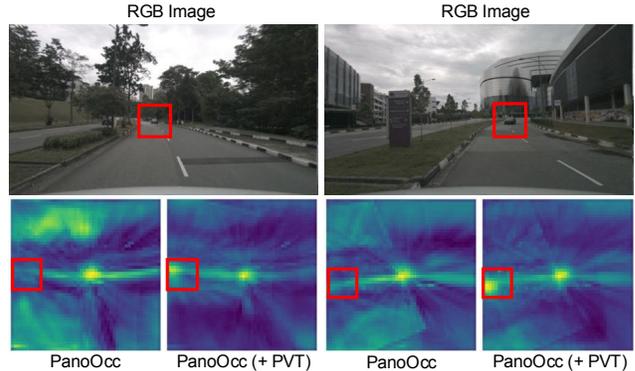

Figure 5. **Feature map visualization.** We visualize the learned voxel queries, which are average-pooled on the channel and the z-axis. The red box indicates the region we focus on.

tation in low-resolution voxel queries for camera-based 3DOP. ProtoOcc enhances the interpretability of smaller-sized queries while mitigating information loss through two context-enriching strategies: 1) a prototype-aware view transformation and 2) a multi-perspective occupancy decoding. Experimental evaluations on both Occ3D and SemanticKITTI benchmark underscore the significance of ProtoOcc, revealing substantial enhancements over previous approaches. Notably, ProtoOcc achieves competitive performance even with smaller-sized queries, offering a promising solution for real-time deployment without compromising predictive accuracy.

Acknowledgement

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11): 2274–2282, 2012. 3, 2

[2] 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. 2, 5, 6, 1

[3] 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. 2, 6, 1, 5

[4] Jian Ding, Nan Xue, Gui-Song Xia, Bernt Schiele, and Dengxin Dai. Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15413–15423, 2023. 3, 2

[5] Wenzheng Feng, Jie Zhang, Yuxiao Dong, Yu Han, Huanbo Luan, Qian Xu, Qiang Yang, Evgeny Kharlamov, and Jie Tang. Graph random neural networks for semi-supervised learning on graphs. *Advances in neural information processing systems*, 33:22092–22103, 2020. 5

[6] Pedro J Freire, Yevhenii Osadchuk, Bernhard Spinnler, Antonio Napoli, Wolfgang Schairer, Nelson Costa, Jaroslaw E Prilepsky, and Sergei K Turitsyn. Performance versus complexity study of neural network equalizers in coherent optical systems. *Journal of Lightwave Technology*, 39(19):6085–6096, 2021. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[8] Weixiang Hong, Qingpei Guo, Wei Zhang, Jingdong Chen, and Wei Chu. Lpsnet: A lightweight solution for fast panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 16746–16754, 2021. 1

[9] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 2

[10] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2

[11] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. *arXiv preprint arXiv:2311.12754*, 2023. 1, 2

[12] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023. 1, 2, 6, 5

[13] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 352–368, 2018. 3, 2

[14] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20258–20267, 2024. 2, 5, 6, 1, 3, 4

[15] Sungjune Kim, Hadam Baek, Seunggwan Lee, Hyung-gun Chi, Hyerin Lim, Jinkyu Kim, and Sangpil Kim. Enhanced motion forecasting with visual relation reasoning. In *European Conference on Computer Vision*, pages 311–328. Springer, 2024. 2

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4, 7, 8

[17] Dmitrii Kochkov, Jamie A Smith, Ayya Alieva, Qing Wang, Michael P Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021. 1

[18] Hongyang Li, Hao Zhang, Zhaoyang Zeng, Shilong Liu, Feng Li, Tianhe Ren, and Lei Zhang. Dfa3d: 3d deformable attention for 2d-to-3d feature lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6684–6693, 2023. 3, 6

[19] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3351–3359, 2020. 2, 5

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

[21] 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, 5, 6, 1, 3, 4

[22] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 2, 3, 6, 1, 4

[23] James Liang, Yiming Cui, Qifan Wang, Tong Geng, Wenguan Wang, and Dongfang Liu. Clusterfomer: Clustering as a universal visual learner. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[25] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction. *arXiv preprint arXiv:2312.17118*, 2023. 2

[26] Lanlan Liu and Jia Deng. Dynamic deep neural networks: Optimizing accuracy-efficiency trade-offs by selective execution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[28] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 2

[29] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 2

[30] 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. 2

[31] Xu Ma, Yuqian Zhou, Huan Wang, Can Qin, Bin Sun, Chang Liu, and Yun Fu. Image as set of points. *arXiv preprint arXiv:2303.01494*, 2023. 3

[32] Jianbiao Mei, Yu Yang, Mengmeng Wang, Junyu Zhu, Xiangrui Zhao, Jongwon Ra, Laijian Li, and Yong Liu. Camera-based 3d semantic scene completion with sparse guidance network. *arXiv preprint arXiv:2312.05752*, 2023. 2

[33] 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

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[35] Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2021. 4

[36] Wonseok Roh, Gyusam Chang, Seokha Moon, Giljoo Nam, Chanyoung Kim, Younghyun Kim, Jinkyu Kim, and Sangpil Kim. Ora3d: Overlap region aware multi-view 3d object detection. *arXiv preprint arXiv:2207.00865*, 2022. 2

[37] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 5

[38] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 2

[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. 2, 3, 5, 6, 1, 4

[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, 2

[41] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin De Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12*, pages 13–26. Springer, 2012. 4, 7, 8

[42] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10052–10062, 2021. 4

[43] Antonin Vobecky, Oriane Siméoni, David Hurych, Spyridon Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Pop-3d: Open-vocabulary 3d occupancy prediction from images. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[44] Chong Wang, Yuyuan Liu, Yuanhong Chen, Fengbei Liu, Yu Tian, Davis McCarthy, Helen Frazer, and Gustavo Carneiro. Learning support and trivial prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Inter-*

national Conference on Computer Vision, pages 2062–2072, 2023. 2

[45] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5463–5474, 2021. 4

[46] Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14792–14801, 2024. 6, 1, 5

[47] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2

[48] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. *arXiv preprint arXiv:2306.10013*, 2023. 1, 2, 3, 5, 6, 4

[49] Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Frustum-former: Adaptive instance-aware resampling for multi-view 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5096–5105, 2023. 3

[50] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 1, 2, 6

[51] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3101–3109, 2021. 5

[52] Fengting Yang, Qian Sun, Hailin Jin, and Zihan Zhou. Superpixel segmentation with fully convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13964–13973, 2020. 3, 2

[53] Haiming Zhang, Xu Yan, Dongfeng Bai, Jiantao Gao, Pan Wang, Bingbing Liu, Shuguang Cui, and Zhen Li. Radocc: Learning cross-modality occupancy knowledge through rendering assisted distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7060–7068, 2024. 1

[54] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 733–749, 2018. 2

[55] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 1, 2, 6, 5

[56] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3