RadOcc: Learning Cross-Modality Occupancy Knowledge through Rendering Assisted Distillation

Haiming Zhang[1,2*], **Xu Yan**[3†], **Dongfeng Bai**[3], **Jiantao Gao**[3], **Pan Wang**[3], **Bingbing Liu**[3], **Shuguang Cui**[2,1], **Zhen Li**[2,1†]

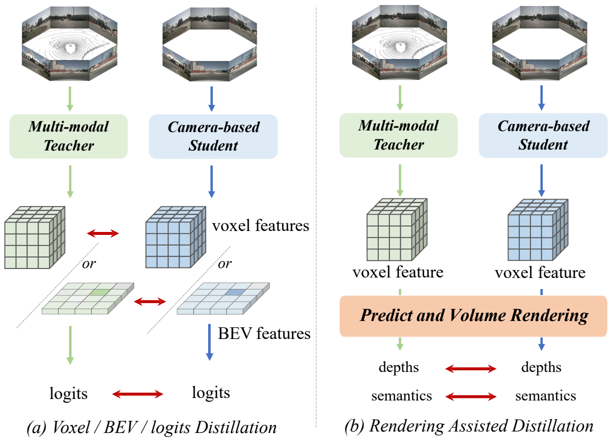
[1]FNii, CUHK-Shenzhen, Shenzhen, China
[2]SSE, CUHK-Shenzhen, Shenzhen, China
[3]Huawei Noah's Ark Lab
{haimingzhang@link., xuyan1@link., lizhen@}cuhk.edu.cn

Abstract

3D occupancy prediction is an emerging task that aims to estimate the occupancy states and semantics of 3D scenes using multi-view images. However, image-based scene perception encounters significant challenges in achieving accurate prediction due to the absence of geometric priors. In this paper, we address this issue by exploring cross-modal knowledge distillation in this task, *i.e.,* we leverage a stronger multi-modal model to guide the visual model during training. In practice, we observe that directly applying features or logits alignment, proposed and widely used in bird's-eye-view (BEV) perception, does not yield satisfactory results. To overcome this problem, we introduce **RadOcc**, a **R**endering **a**ssisted **d**istillation paradigm for 3D **Occ**upancy prediction. By employing differentiable volume rendering, we generate depth and semantic maps in perspective views and propose two novel consistency criteria between the rendered outputs of teacher and student models. Specifically, the depth consistency loss aligns the termination distributions of the rendered rays, while the semantic consistency loss mimics the intra-segment similarity guided by vision foundation models (VLMs). Experimental results on the nuScenes dataset demonstrate the effectiveness of our proposed method in improving various 3D occupancy prediction approaches, *e.g.,* our proposed methodology enhances our baseline by **2.2%** in the metric of mIoU and achieves **50%** in Occ3D benchmark.

Introduction

3D occupancy prediction (3D-OP) is a crucial task within the field of 3D scene understanding, which has garnered considerable attention, particularly in the field of autonomous driving (Wang et al. 2023b; Tong et al. 2023; Tian et al. 2023). In contrast to other 3D perception tasks, such as object detection using bounding box representations, 3D-OP involves the simultaneous estimation of both the occupancy state and semantics in the 3D space using multi-view images (Tian et al. 2023). This is achieved by leveraging geometry-aware cubes to represent a wide range of objects and background shapes.



(a) Voxel / BEV / logits Distillation

(b) Rendering Assisted Distillation

Figure 1: Rendering Assisted Distillation. (a) Existing methods conduct alignment on features or logits. (b) Our proposed RadOcc method constrains the rendered depth maps and semantics simultaneously.

In the realm of 3D occupancy prediction, remarkable advancements have been achieved thus far. These advancements have been made possible by adopting a pipeline inspired by Bird's Eye View (BEV) perception, which utilizes either forward projection (Huang et al. 2021) or backward projection (Li et al. 2022b) techniques for view transformation. This process generates 3D volume features that capture the spatial information of the scene, which are then fed into the prediction head for occupancy predictions. However, relying solely on camera modality poses challenges in accurate prediction due to the lack of geometric perception. To overcome this bottleneck, two mainstream solutions have emerged in the field of BEV perception: **1)** integrating geometric-aware LiDAR input and fusing the complementary information of the two modalities (Liu et al. 2023), and **2)** conducting knowledge distillation to transfer the complementary knowledge from other modalities to a single-modality model (Zhou et al. 2023a). As the first solution introduces additional network designs and computational overhead, recent works have increasingly focused on the second solution, aiming to develop stronger single-modal

[*] Work done during an internship at Huawei Noah's Ark Lab.

[†] Corresponding authors: Xu Yan and Zhen Li.

models through distilling multi-modal knowledge.

In this paper, we present the first investigation into cross-modal knowledge distillation for the task of 3D occupancy prediction. Building upon existing methods in the field of BEV perception that leverage BEV or logits consistency for knowledge transfer, we extend these distillation techniques to aligning voxel features and voxel logits in the task of 3D occupancy prediction, as depicted in Figure 1(a). However, our preliminary experiments reveal that these alignment techniques face significant challenges in achieving satisfactory results in the task of 3D-OP, particularly the former approach introduces negative transfer. This challenge may stem from the fundamental disparity between 3D object detection and occupancy prediction, where the latter is a more fine-grained perception task that requires capturing geometric details as well as background objects.

To address the aforementioned challenges, we propose **RadOcc**, a novel approach that leverages differentiable volume rendering for cross-modal knowledge distillation. The key idea of RadOcc is conducting alignment between rendered results generated by teacher and student models, as Figure 1(b). Specifically, we employ volume rendering (Mildenhall et al. 2021) on voxel features using the camera's intrinsic and extrinsic parameters, which enables us to obtain corresponding depth maps and semantic maps from different viewpoints. To achieve better alignment between the rendered outputs, we introduce the novel Rendered Depth Consistency (**RDC**) and Rendered Semantic Consistency (**RSC**) losses. On the one hand, the RDC loss enforces consistency of ray distribution, which enables the student model to capture the underlying structure of the data. On the other hand, the RSC loss capitalizes on the strengths of vision foundation models (Kirillov et al. 2023), and leverages pre-extracted segments to conduct an affinity distillation. This criterion allows the model to learn and compare semantic representations of different image regions, enhancing its ability to capture fine-grained details. By combining the above constraints, our proposed method effectively harnesses the cross-modal knowledge distillation, leading to improved performance and better optimization for the student model. We demonstrate the effectiveness of our approach on both dense and sparse occupancy prediction and achieve state-of-the-art results on both tasks.

In summary, our main contributions are threefold:

- We propose a rendering assisted distillation paradigm for 3D occupancy prediction, named **RadOcc**. Our paper is the first to explore cross-modality knowledge distillation in 3D-OP and provides valuable insights into the application of existing BEV distillation techniques for this task.
- Two novel distillation constraints, *i.e.,* rendered depth and semantic consistency (**RDC & RSC**), are proposed, which effectively enhance the knowledge transfer process through aligning ray distribution and affinity matrices guided by vision foundation models.
- Equipped with the proposed methodology, RadOcc achieves state-of-the-art performance on the Occ3D and nuScenes benchmarks for dense and sparse occupancy prediction. Furthermore, we verify that our proposed distillation approach can effectively boost the performance of several baseline models.

Related Work

Camera-based 3D Perception

Camera-based 3D perception has emerged as a significant research focus in the field of autonomous driving, owing to its cost-effectiveness and rich visual attributes. Recent advancements have aimed to integrate multiple tasks into a unified framework by transforming image-based features into a Bird's Eye View (BEV) space. One mainstream follows the forward projection paradigm proposed in LSS (Philion and Fidler 2020), where multi-view image features are projected onto the BEV plane through predicted depth maps (Huang et al. 2021; Li et al. 2023, 2022a). Another mainstream (*i.e.,* backward projection) draws inspiration from DETR3D (Wang et al. 2022b), which involves using learnable queries and a cross-attention mechanism to extract information from image features (Li et al. 2022b; Lu et al. 2022; Jiang et al. 2023). Although these methods effectively compress information onto the BEV plane, they may lose some of the essential structural details inherent in 3D space. Introducing LiDAR priors through cross-modal knowledge distillation makes them ideal for understanding the structure of 3D scenes while keeping efficiency.

3D Occupancy Prediction

The field of 3D occupancy prediction (3D-OP) has garnered significant attention in recent years, with the aim of reconstructing the 3D volumetric scene structure from multi-view images. This area can be broadly classified into two categories based on the type of supervision: *sparse prediction* and *dense prediction*. On the one hand, sparse prediction methods utilize LiDAR points as supervision and are evaluated on LiDAR semantic segmentation benchmarks. For instance, TPVFormer (Huang et al. 2023) proposes a tri-perspective view method for predicting 3D occupancy, while PanoOcc (Wang et al. 2023b) unifies the task of occupancy prediction with panoptic segmentation in a coarse-to-fine scheme. On the other hand, dense prediction methods are more akin to the Semantic Scene Completion (SSC) task (Song et al. 2017; Yan et al. 2021), with the core difference being whether to consider the area that the camera cannot capture. Recently, several studies focus on the task of dense occupancy prediction and introduce new benchmarks using nuScenes dataset (Caesar et al. 2020) at the same period, such as OpenOccupancy (Wang et al. 2023a), OpenOcc (Tong et al. 2023), SurroundOcc (Wei et al. 2023) and Occ3D (Tian et al. 2023). These works mainly adopt the architecture from BEV perception and use 3D convolution to construct an extra head for occupancy prediction. We find out that concurrent work (Gan et al. 2023) also utilizes volume rendering technique, however, they naively apply rendered results as auxiliary supervision. Still, we first time investigate cross-modal knowledge distillation in this field, and our proposed method can be integrated into arbitrary previous works.

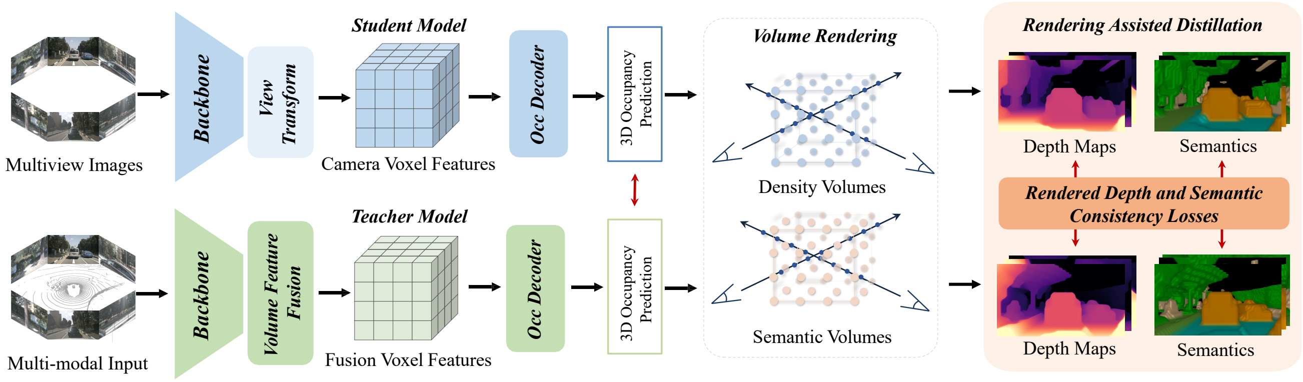


Figure 2: Overall framework of RadOcc. It adopts a teacher-student architecture, where the teacher network is a multi-modal model while the student network only takes camera inputs. The predictions of two networks will be utilized to generate rendered depth and semantics through differentiable volume rendering. The newly proposed rendered depth and semantic consistency losses are adopted between the rendered results.

Cross-Modal Knowledge Distillation

Knowledge distillation has been a popular technique in the field of computer vision since its introduction in (Hinton, Vinyals, and Dean 2015). This technique initially involves compressing a large network (teacher) into a more compact and efficient one (student), while simultaneously improving the performance of the student. Over the years, the effectiveness of knowledge distillation has led to its widespread exploration in various computer vision tasks, including object detection (Dai et al. 2021; Guo et al. 2021; Zhang and Ma 2020), semantic segmentation (Hou et al. 2020; Liu et al. 2019) and other tasks (Yan et al. 2022b; Zhao et al. 2023; Yuan et al. 2022; Zhou et al. 2023b). Recently, knowledge distillation has been introduced into 3D perception tasks for knowledge transfer between models using different modalities. For instance, (Chong et al. 2022) transfers depth knowledge of LiDAR points to a camera-based student detector by training another camera-based teacher with LiDAR projected to perspective view. 2DPASS (Yan et al. 2022a) utilizes multiscale fusion-to-single knowledge distillation to enhance the LiDAR model with image priors. In the field of BEV perception, CMKD (Hong, Dai, and Ding 2022), BEVDistill (Chen et al. 2022) and UniDistill (Zhou et al. 2023a) perform cross-modality distillation in BEV space. Specifically, these methods transform prior knowledge through distillation in feature, relation, and output levels. Although these efforts have greatly enhanced the performance of student models, they cannot achieve satisfactory performance gains when directly applied to the task of 3D occupancy prediction.

Methodology

Problem Setup

3D occupancy prediction leverages multiview images as input to predict a semantic volume surrounding the ego-vehicle. Specifically, it takes into account the current multiview images denoted as $\mathcal{I}_t = \{\mathcal{I}_1^t, ..., \mathcal{I}_n^t\}$, as well as the previous frames $\mathcal{I}^{t-1}, ..., \mathcal{I}^{t-k}$, where k represents the number of history frames and n denotes the camera view index. By incorporating this temporal information, the model finally predicts the semantic voxel volume $\mathcal{Y}_t \in \{w_1, ..., w_{C+1}\}^{H\times W\times Z}$ for the current frame. Here, $C+1$ includes C semantic classes with an occupancy state in the scene, while $w_{(\cdot)}$ represents the voxel grid.

Distillation Architecture

Framework overview. The overall architecture is illustrated in Figure 2, consisting of teacher and student networks. The teacher network takes both LiDAR and multi-view images as input, while the student network solely utilizes multi-view images. Both branches are supervised by ground truth occupancy, and the distillation constraints are applied between 3D occupancy predictions and rendered results.

Camera-based student. Our student network takes multi-frame multi-view images as input and first extracts the feature using an image backbone. To leverage the benefits of Bird's Eye View (BEV) perception, we apply pixel-wise depth estimation on image features and then project them from the perspective view into a 3D volume via the view-transform operation proposed in (Huang et al. 2021), forming a low-level volume feature. Moreover, to introduce the temporal information in our model, we adopt the technique proposed in (Li et al. 2022a), which dynamically warps and fuses the historical volume feature and produces a fused feature. To obtain more fine-grained predicted shapes, the volume feature is fed into an occupancy decoder to generate the prediction.

Multi-modal teacher. Inspired by LiDAR-based detectors presented in (Shi et al. 2020), the unstructured point clouds are scattered into pillars (Lang et al. 2019). Subsequently, the volume features are extracted by SECOND and SECOND-FPN (Yan, Mao, and Li 2018). Building upon the success of LiDAR-camera-based BEV detectors, as presented in (Liu et al. 2023), we further concatenate features from two modalities and process the result with a fully convolutional network to produce the fused features. Finally, a

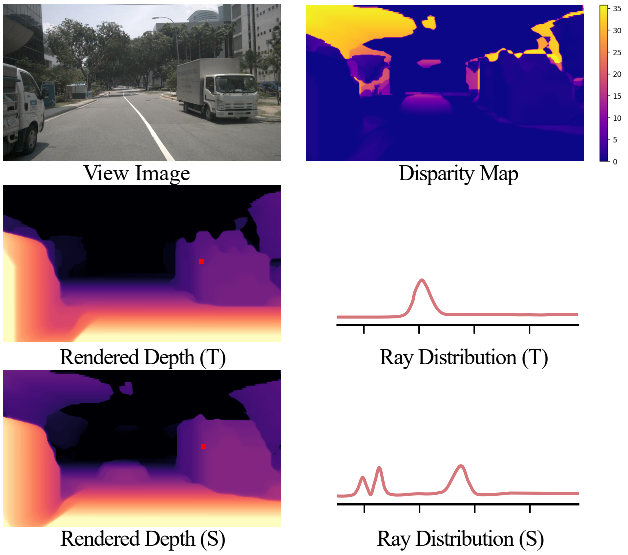


Figure 3: The analysis of rendered depths. Although the rendered depths of teacher (T) and student (S) are similar, especially for the foreground objects, their ray termination distribution shows a great disparity.

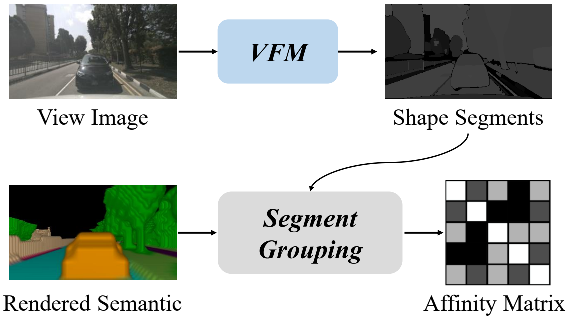


Figure 4: The generation of affinity matrix. We first adopt a visual foundation model (VFM), *i.e.,* SAM, to extract segments into the original image. After that, we conduct segment grouping in rendered semantic features in each segment, obtaining the affinity matrix.

similar occupancy decoder is applied to the fused feature, resulting in the prediction of occupancy.

Rendering Assisted Distillation

Volume rendering. In this paper, we adopt the volume rendering technique as proposed in NeRF (Mildenhall et al. 2021) to obtain depth and semantic maps for knowledge distillation. By incorporating camera intrinsic and external parameters, we are able to compute the corresponding 3D ray for each pixel in the 2D image. After that, we employ the volume rendering technique to perform a weighted sum on the sampled points along the ray, thereby calculating the predicted depths and semantics in perspective views. Given N_p sampled points $\{p_i = (x_i, y_i, z_i)\}_{i=1}^{N_p}$ along the ray in pixel (u, v), the rendered depth $\hat{d}$ and semantic logits $\hat{s}$ at this pixel can be calculated via

$$T_i = \exp(\sum\nolimits_{j=1}^{i-1} \sigma(p_j)\delta_j), \tag{1}$$

$$\hat{d}(u,v) = \sum\nolimits_{i=1}^{N_p} T_i(1 - \exp(-\sigma(p_i)\delta_i))d(p_i), \tag{2}$$

$$\hat{s}(u,v) = \sum\nolimits_{i=1}^{N_p} T_i(1 - \exp(-\sigma(p_i)\delta_i))s(p_i), \tag{3}$$

where $d(\cdot)$, $\sigma(\cdot)$ and $s(\cdot)$ are distance, volume density and semantic of the sampled point, respectively. Since the occupancy network will predict the occupancy probability and semantics, we can easily obtain $\sigma(p_i)$ and $s(p_i)$ by scattering the voxel predictions into the corresponding sampled point p_i. Moreover, $\delta_i = d(p_{i+1}) - d(p_i)$ is the distance between two adjacent sampled points. Finally, we obtain depth and semantic maps in i-th perspective view through collecting results from all pixels, *i.e.,* $\mathcal{S}_i = \{\hat{s}(u,v) \mid u \in [1,H], v \in [1,W]\}$ and $\mathcal{D}_i = \{\hat{d}(u,v) \mid u \in [1,H], v \in [1,W]\}$, where (H,W) is the size of view image. To facilitate the definition, we respectively denote rendered depth and semantics results from teacher and student as $\mathcal{D}^{T/S} = \{\mathcal{D}_1^{T/S}, ..., \mathcal{D}_n^{T/S}\}$ and $\mathcal{S}^{T/S} = \{\mathcal{S}_1^{T/S}, ..., \mathcal{S}_n^{T/S}\}$, where n is the number of views.

Rendered depth consistency. After acquiring the rendered depth, a simplistic approach involves directly imposing constraints between the output of teacher and student models. However, this approach is a hard constraint, and the differences in rendered depths between the teacher and student models are typically within a narrow range. To address this issue, we propose an innovative approach that aligns the ray termination distribution during the volume rendering process. As shown in Figure 3, we plot ray distribution over the distance traveled by the ray. Although the rendered depths of the two models are quite similar, their ray distribution shows a great discrepancy. When a ray traverses through single objects (the red point), we find that the ray termination distribution of the teacher model is typically unimodal, while that of the student exists multiple peaks. Aligning this distribution makes the student model tend to predict a similar latent distribution as the teacher model. Finally, rendered depth consistency (RDC) loss $\mathcal{L}_{rdc}$ is formulated as

$$\mathcal{R}_{(u,v)}^{(\cdot)} = \{T_i(1 - \exp(-\sigma(p_i)\delta_i))\}_{i=1}^{N_p}, \tag{4}$$

$$\mathcal{L}_{rdc} = \frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} \mathcal{D}_{KL}(\mathcal{R}_{(u,v)}^{\text{teacher}} || \mathcal{R}_{(u,v)}^{\text{student}}). \tag{5}$$

Here, T_i is calculated as Eqn. (1). The notation $\mathcal{R}^{\text{teacher}}$ and $\mathcal{R}^{\text{student}}$ respectively denote the ray distribution of the teacher and student networks, which are aligned through KL divergence $\mathcal{D}_{KL}(\cdot||\cdot)$.

Rendered semantic consistency. Besides simply using KL divergence to align the semantic logits, we also leverage the strengths of vision foundation models (VFMs) (Kirillov et al. 2023) to perform a segment-guided affinity distilla-

Method	Image Backbone	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
Performances on Validation Set																			
MonoScene	R101-DCN	6.06	1.75	7.23	4.26	4.93	9.38	5.67	3.98	3.01	5.90	4.45	7.17	14.91	6.32	7.92	7.43	1.01	7.65
CTF-Occ	R101-DCN	28.53	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
BEVFormer	R101-DCN	39.24	10.13	47.91	24.90	47.57	54.52	20.23	28.85	28.02	25.73	33.03	38.56	81.98	40.65	50.93	53.02	43.86	37.15
PanoOcc	R101-DCN	42.13	11.67	50.48	29.64	49.44	55.52	23.29	**33.26**	30.55	30.99	34.43	42.57	**83.31**	44.23	54.40	56.04	45.94	40.40
BEVDet†	Swin-B	42.02	12.15	49.63	25.10	52.02	54.46	27.87	27.99	28.94	27.23	36.43	42.22	82.31	43.29	54.62	57.90	48.61	43.55
Baseline (ours)	Swin-B	44.14	**13.39**	52.20	**31.43**	52.01	56.70	**30.66**	32.95	31.56	**31.31**	39.87	44.64	82.98	**44.97**	**55.43**	**58.90**	48.43	42.99
RadOcc (ours)	Swin-B	**46.06**	9.78	**54.93**	20.44	**55.24**	**59.62**	30.48	28.94	**44.66**	28.04	**45.69**	**48.05**	81.41	39.80	52.78	56.16	**64.45**	**62.64**
Teacher (ours)	Swin-B	49.38	10.93	58.23	25.01	57.89	62.85	34.04	33.45	50.07	32.05	48.87	52.11	82.9	42.73	55.27	58.34	68.64	66.01
Performances on 3D Occupancy Prediction Challenge																			
BEVFormer	R101-DCN	23.70	10.24	36.77	11.70	29.87	38.92	10.29	22.05	16.21	14.69	27.44	33.13	48.19	33.10	29.80	17.64	19.01	13.75
SurroundOcc†	R101-DCN	42.26	11.7	50.55	32.09	41.59	57.38	27.93	38.08	30.56	29.32	48.29	38.72	80.21	48.56	53.20	47.56	46.55	36.14
BEVDet†	Swin-B	42.83	18.66	49.82	31.79	41.90	56.52	26.74	37.31	30.01	31.33	48.18	38.59	80.95	50.59	53.87	49.67	46.62	35.62
PanoOcc-T⋆	Intern-XL	47.16	**23.37**	50.28	36.02	47.32	59.61	31.58	39.59	34.58	33.83	52.25	43.29	**83.82**	**55.81**	**59.41**	53.81	53.48	43.61
Baseline-T (ours)	Swin-B	47.74	22.88	50.74	**41.02**	**49.39**	55.40	33.41	45.71	38.57	35.79	48.94	44.40	83.19	52.26	59.09	**55.83**	51.35	43.54
RadOcc-T (ours)	Swin-B	**49.98**	21.13	**55.17**	39.31	48.99	**59.92**	**33.99**	**46.31**	**43.26**	**39.29**	**52.88**	**44.85**	83.72	53.93	59.17	55.62	**60.53**	**51.55**
Teacher-T (ours)	Swin-B	55.09	25.94	59.04	44.93	57.95	63.70	38.89	52.03	53.21	42.16	59.90	50.45	84.79	55.70	60.83	58.02	67.66	61.40

Table 1: 3D occupancy prediction performance on the Occ3D. † denotes the performance reproduced by official codes. ⋆ means the results provided by authors. '-T' represents results through test-time augmentation (TTA). Please note that our visual model achieves a benchmark ranking of Top-4 on 16/08/2023, outperforming all previously published methods.

tion (SAD). Specifically, we first employ the VFM to over-segment patches using the original view images as input, as illustrated in Figure 4. With the rendered semantic features from both the teacher and student networks, *i.e.,* $\mathcal{S}^T$, $\mathcal{S}^S \in \mathbb{R}^{H\times W\times C}$, we can divide the rendered semantics into several groups based on the indices of aforementioned patches. After that, an average pooling function is applied within each group, extracting multiple teacher and student semantic embedding, *i.e.,* $\mathcal{E}^T \in \mathbb{R}^{M\times C}$ and $\mathcal{E}^S \in \mathbb{R}^{M\times C}$. Here, M is the number of patches generated by the VFM. Inspired but different from the previous work (Hou et al. 2022), we calculate an affinity matrix $\mathcal{C}^{(\cdot)}$ according to the above segments for the further distillation:

$$\mathcal{C}_{i,j,r} = \frac{\mathcal{E}(i,r), \mathcal{E}(j,r)}{||\mathcal{E}(i)||_2||\mathcal{E}(j)||_2}. \tag{6}$$

The affinity score captures the similarity of each segment of semantic embedding and it can be taken as the high-level structural knowledge to be learned by the student. After that, the final RSC loss is a linear combination of affinity distillation loss and KL divergence between rendered semantics:

$$\mathcal{L}_{sad} = \sum_{r=1}^{C}\sum_{i=1}^{M}\sum_{j=1}^{M} ||C^T_{i,j,r} - C^S_{i,j,r}||_2^2, \tag{7}$$

$$\mathcal{L}_{rsc} = \mathcal{L}_{sad}/CM^2 + \omega\mathcal{D}_{KL}(\mathcal{S}^T||\mathcal{S}^S), \tag{8}$$

where C^T and C^S are affinity matrices of teacher and student networks, and ω is a hyperparameter in our experiment.

Experiments

Dataset ane Metric

Dataset. We evaluate our proposed method on nuScenes (Caesar et al. 2020) for sparse prediction and Occ3D (Tian et al. 2023) for dense prediction. The data descriptions are provided in supplementary material.

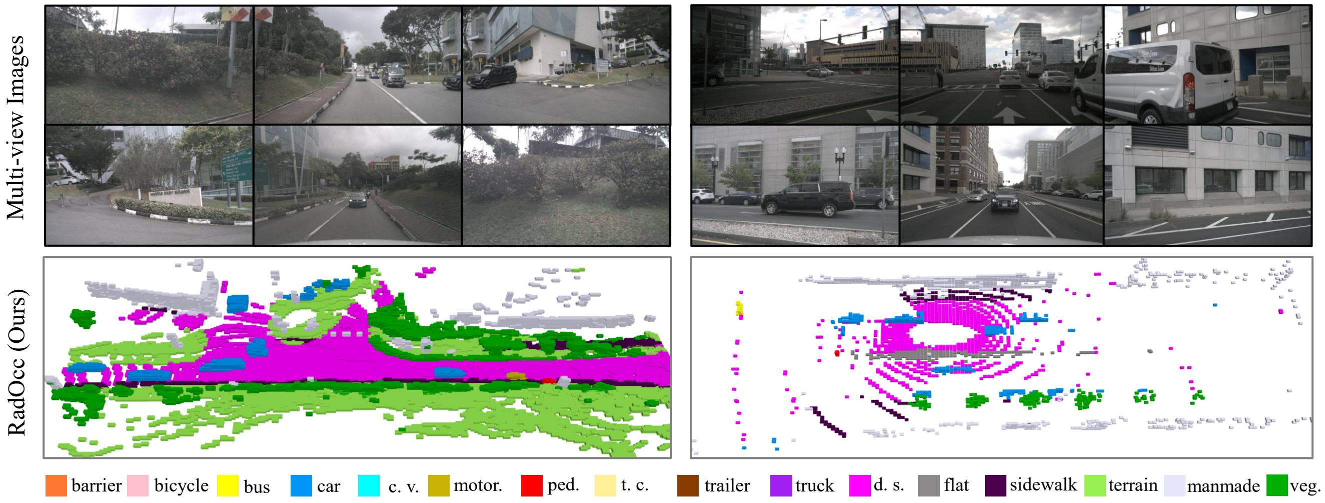
Evaluation metrics. Our study presents an independent evaluation of the model's performance in both dense and sparse prediction tasks. Specifically, for dense prediction, we conduct experiments on the Occ3D dataset, which quantifies the mean Intersection over Union (mIoU) for 17 semantic categories within the camera's visible region. On the other hand, for sparse prediction, we train the model with single-sweep LiDAR and assess the model's performance on the nuScenes-lidarseg benchmark, which measures the mIoU for 16 semantic categories, with the 'others' category being treated as 'ignored'.

Experimental Settings

Implementation. For the dense prediction, we follow the setting of BEVDet (Huang et al. 2021) and use Swin Transformer (Liu et al. 2021) as the image backbone. We adopt the semantic scene completion module proposed in (Yan et al. 2021) as our occupancy decoder, which contains several 3D convolutional blocks to learn a local geometry representation. Afterward, the features from different blocks are concatenated to aggregate information. Finally, a linear projection is utilized to map the feature into $C + 1$ dimensions. Since the challenging nature of the Occ3D test benchmark, we utilize 8 historical frames for temporal encoding and use

Method	Input Modality	Image Backbone	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
PolarNet	LiDAR	-	69.4	72.2	16.8	77.0	86.5	51.1	69.7	64.8	54.1	69.7	63.5	96.6	67.1	77.7	72.1	87.1	84.5
Cylinder3D	LiDAR	-	77.2	82.8	29.8	84.3	89.4	63.0	79.3	77.2	73.4	84.6	69.1	97.7	70.2	80.3	75.5	90.4	87.6
2DPASS	LiDAR	-	80.8	81.7	55.3	92.0	91.8	73.3	86.5	78.5	72.5	84.7	75.5	97.6	69.1	79.9	75.5	90.2	88.0
TPVFormer	Camera	R50-DCN	59.2	65.6	15.7	75.1	80.0	48.8	43.1	44.3	26.8	72.8	55.9	92.3	53.7	61.0	59.2	79.7	75.6
BEVDet†	Camera	Swin-B	65.2	31.3	**63.9**	74.6	79.1	51.5	59.8	63.4	56.2	74.7	59.8	92.8	61.4	69.5	65.7	84.1	82.9
TPVFormer (BL)	Camera	R101-DCN	69.4	**74.0**	27.5	**86.3**	85.5	**60.7**	68.0	62.1	49.1	**81.9**	68.4	94.1	59.5	66.5	63.5	83.8	79.9
RadOcc (ours)	Camera	R101-DCN	**71.8**	49.1	34.2	84.5	**85.8**	59.2	**70.3**	**71.4**	**62.5**	79.7	**69.0**	**95.4**	**66.2**	**75.1**	**72.0**	**87.4**	**86.0**
Teacher (ours)	Cam+Li	R101-DCN	75.2	62.7	33.2	88.7	88.8	64.6	78.1	74.1	65.0	83.1	72.2	96.5	68.3	77.6	74.4	88.7	87.1

Table 2: LiDAR semantic segmentation results on nuScenes test benchmark. † denotes the performance is reproduced by official codes. Our method achieves state-of-the-art performance in camera-based methods. BL denotes the baseline method.



(a) Dense 3D Occupancy Prediction

(b) Sparse 3D Occupancy Prediction

Figure 5: Qualitative results on Occ3D and nuScenes validation sets. RadOcc takes multi-view images as input and produces voxel predictions. More visualization comparisons can be found in the supplementary materials.

3 frames on the validation set. For the sparse prediction, we use previous art TPVFormer (Huang et al. 2023) as our baseline. The rendered size of the network is configured to 384 × 704. To speed up the rendering and reduce memory usage, we randomly sample 80,000 rays during each step.

Results and Analysis

Dense Prediction. To evaluate the performance of dense 3D occupancy prediction, we compare our proposed method with current state-of-the-art approaches on the Occ3D dataset (Tian et al. 2023), including the validation set and online benchmark. The upper part of Table 1 presents the validation set results, where all methods are trained for 24 epochs. Specifically, we compare our approach with MonoScene (Cao and de Charette 2022), BEVFormer (Li et al. 2022b), CTF-Occ (Tian et al. 2023) and PanoOcc (Wang et al. 2023b), which all employ the ResNet101-DCN (Dai et al. 2017) initialized from FCOS3D (Wang et al. 2021) checkpoint as the image backbone. Additionally, we report the results of BEVDet (Huang et al. 2021) that uses the same image backbone as ours. Our baseline model, trained from scratch, already outperforms prior state-of-the-art methods. However, by leveraging our proposed distillation strategy, we achieve significantly better occupancy results in terms of mIoU.

The lower part of Table 1 presents the results on the 3D occupancy prediction challenge, where our proposed method achieves state-of-the-art performance and outperforms all previously published approaches by a large margin. Note that though PanoOcc (Wang et al. 2023b) adopts a stronger image backbone, *i.e.,* InternImage-XL (Wang et al. 2022a), the results of them are still lower than ours, especially for the foreground objects with challenge nature. The visualization results for both dense and sparse prediction are

Method	Consistency	mIoU	Gains
BEVDet (baseline)	-	36.10	-
Hinton *et al.*	Prob.	37.00	+0.90
Hinton *et al.*	Feature	35.89	-0.21
BEVDistill	Prob. + Feature	35.95	-0.15
RadOcc (ours)	Render	37.98	+1.88
RadOcc (ours)	Prob. + Render	**38.53**	+2.43

Table 3: Comparison for knowledge distillation. The results are obtained on Occ3D. To speed up the evaluation, we take BEVDet (Huang et al. 2021) with ResNet50 image backbone as our baseline. †: Since there is no object-level prediction, we replace the sparse distillation of BEVDistill (Chen et al. 2022) with logits distillation.

shown in Figure 5. More visualization results can be found in the supplementary material.

Sparse Prediction. To evaluate the effectiveness of model using sparse LiDAR supervision, we evaluate the performance of our proposed RadOcc model on the nuScenes LiDAR semantic segmentation benchmark. Our results, as shown in Table 2, demonstrate a significant improvement over the baseline TPVFormer (Huang et al. 2023) and outperform previous camera-based occupancy networks such as BEVDet (Huang et al. 2021). Surprisingly, our method even achieves comparable performance with some LiDAR-based semantic segmentation methods (Zhang et al. 2020; Zhou et al. 2020). It should be noted that since we use voxelized single-sweep LiDAR as supervision, where the geometric details in data may be lost during the voxelization, the results of a multi-modal teacher network may not achieve comparable performance with state-of-the-art LiDAR-based methods (Yan et al. 2022a).

Comparison for knowledge distillation. To further validate the efficacy of our proposed methodology upon previous teacher-student architectures, we conduct a comparative analysis of RadOcc with conventional knowledge transfer techniques as presented in Table 3. To facilitate the experimentation process, we choose BEVDet (Huang et al. 2021) with ResNet50 image backbone as our baseline, and all methods are trained with the same strategies for a fair comparison. The results in the table indicate that direct application of feature and logits alignment (Hinton, Vinyals, and Dean 2015; Chen et al. 2022) fails to achieve a significant boost on the baseline model, particularly for the former, which results in negative transfer. Notably, leveraging rendering-assisted distillation leads to a substantial improvement of 2% on mIoU. Furthermore, even when applying logit distillation, the model can still enhance the mIoU by 0.6%.

Ablation study. We conduct an ablation study of rendering distillation in Table 4. Here, BEVDet with ResNet50 image backbone is selected as our baseline model. *Model A* directly conducts alignment through Scale-Invariant Logarithmic (Eigen, Puhrsch, and Fergus 2014) on rendered depth maps but fails to improve the performance. In contrast, *Model B* aligns the latent distribution of depth rendering and achieves an improvement of 0.7% in mIoU. On the other hand, *Model C* demonstrates the results sorely using segment-guided affinity distillation (SAD) on rendered semantics, which increases the mIoU by 1.0%. Applying additional KL divergence between two rendered semantics can boost the performance to 37.42%. Finally, when we combine RDC and RSC losses, the model achieves the best result.

Method	RDC(-)	RDC	SAD	RSC	mIoU
BEVDet					36.10
Model A	✓				35.08
Model B		✓			36.76
Model C			✓		37.13
Model D				✓	37.42
RadOcc (ours)		✓		✓	**37.98**

Table 4: Ablation study on Occ3D. We use BEVDet with ResNet50 image backbone as our baseline. Here, RDC and RSC are rendered depth and semantic consistency losses. RDC (-) denotes directly aligning the rendered depth map with Scale-Invariant Logarithmic loss.

Method	Segment	mIoU	Gains
BEVDet w/ RSC	SAM	**37.42**	-
Model E	Super Pixel	37.05	-0.37

Table 5: Design analysis of SAD. We replace the segment extraction strategy with other designs.

In Table 5, we analyze the design of SAD by replacing its segment with other implementations in *Model E*. Specifically, when we use super-pixel (Achanta et al. 2012), the performance will decrease by about 0.37%.

Conclusion

In this paper, we present RadOcc, a novel cross-modal knowledge distillation paradigm for 3D occupancy prediction. It leverages a multi-modal teacher model to provide geometric and semantic guidance to a visual student model via differentiable volume rendering. Moreover, we propose two new consistency criteria, depth consistency loss and semantic consistency loss, to align the ray distribution and affinity matrices between the teacher and student models. Extensive experiments on the Occ3D and nuScenes datasets show RadOcc can significantly improve the performance of various 3D occupancy prediction methods. Our method achieves state-of-the-art results on the Occ3D challenge benchmark and outperforms existing published methods by a large margin. We believe that our work opens up new possibilities for cross-modal learning in scene understanding.

Acknowledgments

This work was supported by NSFC with Grant No. 62293482, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen HK S&T Cooperation Zone, by Shenzhen General Program No. JCYJ20220530143600001, by Shenzhen-Hong Kong Joint

Funding No. SGDX20211123112401002, by Shenzhen Outstanding Talents Training Fund, by Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, by the NSFC 61931024&81922046&61902335, by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055), and the Key Area R&D Program of Guangdong Province with grant No. 2018B03033800, by Tencent&Huawei Open Fund.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11): 2274–2282.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Cao, A.-Q.; and de Charette, R. 2022. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3991–4001.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; and Zhao, F. 2022. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*.

Chong, Z.; Ma, X.; Zhang, H.; Yue, Y.; Li, H.; Wang, Z.; and Ouyang, W. 2022. Monodistill: Learning spatial features for monocular 3d object detection. *arXiv preprint arXiv:2201.10830*.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Dai, X.; Jiang, Z.; Wu, Z.; Bao, Y.; Wang, Z.; Liu, S.; and Zhou, E. 2021. General instance distillation for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7842–7851.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*.

Gan, W.; Mo, N.; Xu, H.; and Yokoya, N. 2023. A Simple Attempt for 3D Occupancy Estimation in Autonomous Driving. *arXiv preprint arXiv:2303.10076*.

Guo, J.; Han, K.; Wang, Y.; Wu, H.; Chen, X.; Xu, C.; and Xu, C. 2021. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2154–2164.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hong, Y.; Dai, H.; and Ding, Y. 2022. Cross-modality knowledge distillation network for monocular 3d object detection. In *European Conference on Computer Vision*, 87–104. Springer.

Hou, Y.; Ma, Z.; Liu, C.; Hui, T.-W.; and Loy, C. C. 2020. Inter-region affinity distillation for road marking segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12486–12495.

Hou, Y.; Zhu, X.; Ma, Y.; Loy, C. C.; and Li, Y. 2022. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8479–8488.

Huang, J.; Huang, G.; Zhu, Z.; Ye, Y.; and Du, D. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9223–9232.

Jiang, Y.; Zhang, L.; Miao, Z.; Zhu, X.; Gao, J.; Hu, W.; and Jiang, Y.-G. 2023. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1042–1050.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

Li, Y.; Bao, H.; Ge, Z.; Yang, J.; Sun, J.; and Li, Z. 2022a. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *arXiv preprint arXiv:2209.10248*.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2023. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1477–1485.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022b. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; and Wang, J. 2019. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2604–2613.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2774–2781. IEEE.

Lu, J.; Zhou, Z.; Zhu, X.; Xu, H.; and Zhang, L. 2022. Learning ego 3d representation as ray tracing. In *European Conference on Computer Vision*, 129–144. Springer.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Philion, J.; and Fidler, S. 2020. 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*, 194–210. Springer.

Shi, S.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE transactions on pattern analysis and machine intelligence*, 43(8): 2647–2664.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1746–1754.

Tian, X.; Jiang, T.; Yun, L.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*.

Tong, W.; Sima, C.; Wang, T.; Wu, S.; Deng, H.; Chen, L.; Gu, Y.; Lu, L.; Luo, P.; Lin, D.; et al. 2023. Scene as Occupancy. *arXiv preprint arXiv:2306.02851*.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 913–922.

Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H.; et al. 2022a. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. *arXiv preprint arXiv:2211.05778*.

Wang, X.; Zhu, Z.; Xu, W.; Zhang, Y.; Wei, Y.; Chi, X.; Ye, Y.; Du, D.; Lu, J.; and Wang, X. 2023a. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. *arXiv preprint arXiv:2303.03991*.

Wang, Y.; Chen, Y.; Liao, X.; Fan, L.; and Zhang, Z. 2023b. PanoOcc: Unified Occupancy Representation for Camera-based 3D Panoptic Segmentation. *arXiv preprint arXiv:2306.10013*.

Wang, Y.; Guizilini, V. C.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2022b. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, 180–191. PMLR.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. *arXiv preprint arXiv:2303.09551*.

Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; and Cui, S. 2021. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3101–3109.

Yan, X.; Gao, J.; Zheng, C.; Zheng, C.; Zhang, R.; Cui, S.; and Li, Z. 2022a. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European Conference on Computer Vision*, 677–695. Springer.

Yan, X.; Zhan, H.; Zheng, C.; Gao, J.; Zhang, R.; Cui, S.; and Li, Z. 2022b. Let images give you more: Point cloud cross-modal training for shape analysis. *Advances in Neural Information Processing Systems*, 35: 32398–32411.

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.

Yuan, Z.; Yan, X.; Liao, Y.; Guo, Y.; Li, G.; Cui, S.; and Li, Z. 2022. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8563–8573.

Zhang, L.; and Ma, K. 2020. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*.

Zhang, Y.; Zhou, Z.; David, P.; Yue, X.; Xi, Z.; Gong, B.; and Foroosh, H. 2020. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9601–9610.

Zhao, W.; Zhang, H.; Zheng, C.; Yan, X.; Cui, S.; and Li, Z. 2023. CPU: Codebook Lookup Transformer with Knowledge Distillation for Point Cloud Upsampling. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3917–3925.

Zhou, H.; Zhu, X.; Song, X.; Ma, Y.; Wang, Z.; Li, H.; and Lin, D. 2020. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. *arXiv preprint arXiv:2008.01550*.

Zhou, S.; Liu, W.; Hu, C.; Zhou, S.; and Ma, C. 2023a. UniDistill: A Universal Cross-Modality Knowledge Distillation Framework for 3D Object Detection in Bird's-Eye View. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5116–5125.

Zhou, W.; Yan, X.; Liao, Y.; Lin, Y.; Huang, J.; Zhao, G.; Cui, S.; and Li, Z. 2023b. BEV@ DC: Bird's-Eye View Assisted Training for Depth Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9233–9242.