

Enhancing Indoor Occupancy Prediction via Sparse Query-Based Multi-Level Consistent Knowledge Distillation

Xiang Li, Yupeng Zheng*, Pengfei Li, Yilun Chen, Ya-Qin Zhang, Wenchao Ding†

Abstract—Occupancy prediction provides critical geometric and semantic understanding for robotics but faces efficiency-accuracy trade-offs. Current dense methods suffer computational waste on empty voxels, while sparse query-based approaches lack robustness in diverse and complex indoor scenes. In this paper, we propose DiScene, a novel sparse query-based framework that leverages multi-level distillation to achieve efficient and robust occupancy prediction. In particular, our method incorporates two key innovations: (1) a Multi-level Consistent Knowledge Distillation strategy, which transfers hierarchical representations from large teacher models to lightweight students through coordinated alignment across four levels, including encoder-level feature alignment, query-level feature matching, prior-level spatial guidance, and anchor-level high-confidence knowledge transfer and (2) a Teacher-Guided Initialization policy, employing optimized parameter warm-up to accelerate model convergence. Validated on the Occ-Scannet benchmark, DiScene achieves 23.2 FPS without depth priors while outperforming our baseline method, OPUS, by 36.1% and even better than the depth-enhanced version, OPUS†, by 21.07%. With depth integration, DiScene† attains new SOTA performance, surpassing EmbodiedOcc by 3.7% with 1.62× faster inference speed. Furthermore, experiments on the Occ3D-nuScenes benchmark and in-the-wild scenarios demonstrate the versatility of our approach in various environments. Code and models can be accessed at <https://github.com/getterupper/DiScene>.

Index Terms—3D Occupancy Prediction, Distillation Learning, Scene Understanding

I. INTRODUCTION

OCCUPANCY prediction has gained significant attention in robotics society due to its ability to provide fine-grained geometric and semantic information [1], [2]. Its objective is to estimate the occupancy status of each voxel and their semantic labels within an entire scene from limited observations. Current

Manuscript received: June 26, 2025; Revised: September 8, 2025; Accepted: September 21, 2025.

This paper was recommended for publication by Editor Pascal Vasseur upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by the Tsinghua-TARS Special Program for the Deep Collaboration in Embodied Intelligence. (*Project leader: Yupeng Zheng, †Corresponding author: Wenchao Ding.)

Xiang Li is with the College of AI, Tsinghua University, Beijing 100083, China and TARS, Shanghai 200233, China.

Yupeng Zheng is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China and TARS, Shanghai 200233, China.

Pengfei Li is with the AIR, Tsinghua University, Beijing 100084, China and TARS, Shanghai 200233, China.

Yilun Chen is with the TARS, Shanghai 200233, China.

Ya-Qin Zhang is with the AIR, Tsinghua University, Beijing 100084, China.

Wenchao Ding is with the College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai 200001, China and TARS, Shanghai 200233, China (e-mail: dingwenchao@fudan.edu.cn).

Digital Object Identifier (DOI): see top of this page.

©2026 IEEE

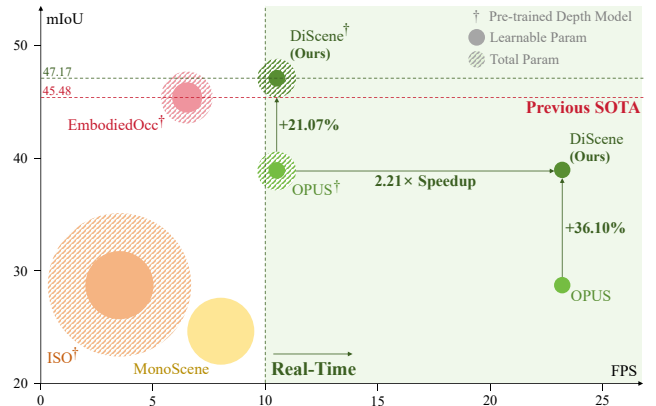


Fig. 1. We compare DiScene with existing indoor occupancy prediction methods in terms of speed and accuracy. All models are evaluated on the Occ-ScanNet [9] validation set and inference speeds are measured on one NVIDIA A800 GPU w/o TensorRT. The size of the circle represents the model's size.

mainstream methods typically employ explicit 3D spatial modeling (e.g., dense voxels [3], [4], Bird's-Eye View [5], [6], Tri-Perspective View [7]), where most computational resources are consumed by empty voxel calculations, resulting in inefficiency. Alternative sparse query-based approaches [8] simultaneously perform spatial occupancy regression and semantic label classification, thereby significantly accelerating inference speeds.

However, these sparse methods underperform in diverse and complex indoor scenes due to insufficient geometric information. Introducing additional vision foundation models [10], [11] to mitigate the ambiguity, in turn, increases latency and compromises real-time performance. Hence, current indoor occupancy prediction methods still fail to achieve a satisfactory balance between model performance and inference speed.

To address this critical limitation, we propose **DiScene**, a novel distillation framework specifically designed for sparse query-based occupancy prediction. We argue that there are several challenges that prevent traditional distillation methods from achieving optimal gains: large feature discrepancy between teacher and student models impedes effective knowledge transfer; there is no natural one-to-one correspondence between teacher and student predictions for vanilla logit- or feature-based distillation; directly distilling both spatial distributions and feature representations from sparse teacher queries introduces excessive learning complexity. To overcome the above challenges, we pioneer a hierarchical distillation strategy that establishes coordinated knowledge transfer between teacher and student models and progressively incorporates guidance

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

information, thus effectively reducing the difficulty of the distillation learning process while maximizing the efficacy of knowledge transfer.

Specifically, we adopt the sparse-centric model OPUS [8] as our baseline and introduce a **Multi-level Consistent Knowledge Distillation** strategy, comprising (1) *Encoder-level Feature Alignment*: We address feature discrepancy caused by heterogeneous encoders through encoder-level alignment loss, enabling effective distillation between teacher and student models. (2) *Query-level Distillation*: We utilize the Hungarian algorithm to establish optimal bipartite matching between student and teacher predictions, allowing coarse feature-based knowledge transfer. (3) *Prior-level Distillation*: We provide the teacher query positions as spatial priors to the student. This addresses unstable bipartite matching resulting from divergent spatial distributions of query embeddings, facilitating more focused feature representation learning. (4) *Anchor-level Distillation*: We sample anchor points from ground-truth occupied voxels and provide them to both models. This ensures selective transfer of high-confidence knowledge, thereby preventing knowledge contamination from low-quality teacher predictions and further enhancing feature representation learning. Moreover, we propose a **Teacher-Guided Initialization** policy, which utilizes well-optimized teacher parameters to accelerate model convergence as free lunch.

DiScene achieves SOTA occupancy prediction performance and real-time inference on the challenging Occ-Scannet benchmark [9]. As demonstrated in Fig. 1, relying solely on distillation and initialization strategy, DiScene maintains a 23.2 FPS inference speed while outperforming the baseline method OPUS by 36.1% and delivers comparable performance to OPUS[†], which leverages pre-trained depth models. When incorporating depth priors, our enhanced DiScene[†] surpasses the previous SOTA method EmbodiedOcc [12] by 3.7% while sustaining inference speeds above 10 FPS.

Moreover, on the Occ3D-nuScenes [13] benchmark, our strategy improves performance by 6.9%, exhibiting robust performance across both indoor and outdoor robotic perception scenarios. We further validate the generality and versatility of our approach on self-collected in-the-wild datasets.

Our main contributions are as follows:

- We propose DiScene, a sparse query-based distillation framework that bridges the accuracy-efficiency gap prevalent in existing indoor occupancy prediction methods.
- We propose a Multi-level Consistent Knowledge Distillation strategy that ensures effective knowledge transfer across multiple complementary levels.
- We introduce a Teacher-Guided Initialization policy that accelerates model convergence at no additional costs.
- We demonstrate the effectiveness and robustness of our method through extensive experiments across indoor and outdoor benchmarks, with additional validation on in-the-wild scenarios.

II. RELATED WORK

A. Occupancy Prediction

Occupancy prediction has achieved notable progress in recent years. Conventional 3D [3], [14], [13], [15], [16], [17] or

4D [18], [19], [20] methods predominantly employ dense voxels as feature representation ; however, such an approach incurs heavy and redundant computational costs. Consequently, recent research in outdoor driving scenarios has seen the emergence of numerous acceleration techniques utilizing alternative representations, such as Bird’s-Eye View [21], [22], Tri-Perspective View [7], 3D Gaussians [23], [24] and sparse 3D queries [4], [25], [8].

On the contrary, similar efforts have not yet been observed in indoor scenarios. Methods like ISO [9] and EmbodiedOcc [12] leverage pre-trained depth models [26], [10] to estimate depth information, which is then fused with scene features to enhance model performance by mitigating depth ambiguity. Nevertheless, the incorporation of such depth models substantially increases inference overhead, hindering their practical deployment in the real world.

In this letter, we attempt to strike a balance between performance and real-time inference for indoor occupancy prediction. Our solution adopts a sparse query-based architecture as the primary framework while integrating knowledge distillation to boost performance without introducing additional costs.

B. Knowledge Distillation

As a classical method for model compression and accuracy enhancing, the concept of knowledge distillation was first introduced by [27], where students are trained to mimic the soft label predictions of teachers. According to the objective of mimicking, subsequent works can be broadly categorized into two types, distilling from output logits [28], [29], [30], [31], [32] and intermediate features [33], [34], [35], [36], [37]. Researchers have applied knowledge distillation to various vision tasks and modality and lead to consistent effectiveness, including image generation [38], [39], 2D semantic segmentation [40], 2D object detection [41], [42], [43], LiDAR semantic segmentation [44] and 3D object detection [45], [46].

Prior works such as SCPNet [47] and MonoOcc [48] have adopted this strategy for occupancy prediction, transferring geometric and semantic knowledge from multi-frame teachers to single-frame students. However, these methods employ dense 3D feature representation, which makes it easier for student models to imitate teachers due to the explicit correspondence between voxels. The application of knowledge distillation to sparse queries for occupancy prediction remains an unexplored and challenging task.

III. METHODOLOGY

A. Preliminaries

Problem Formulation. Following OPUS [8], we reformulate occupancy prediction as a set-to-set matching task to better leverage the sparsity inherent in indoor scenes. Given M occupied voxels of the current scene, we denote them as a ground-truth set $\mathcal{S}^{\text{GT}} = \{\mathcal{P}^{\text{GT}}, \mathcal{C}^{\text{GT}}\} = \{p_i^{\text{GT}}, c_i^{\text{GT}}\}_{i=1}^M$, where p_i^{GT} denotes the 3D coordinates of a voxel center, and c_i^{GT} represents its corresponding semantic class. Our model predicts M' point positions and semantic classes, denoted as the prediction set $\mathcal{S}^{\text{Pred}} = \{\mathcal{P}^{\text{Pred}}, \mathcal{C}^{\text{Pred}}\}$, where M and M' are not necessarily equal. Consequently, the goal of our method is to

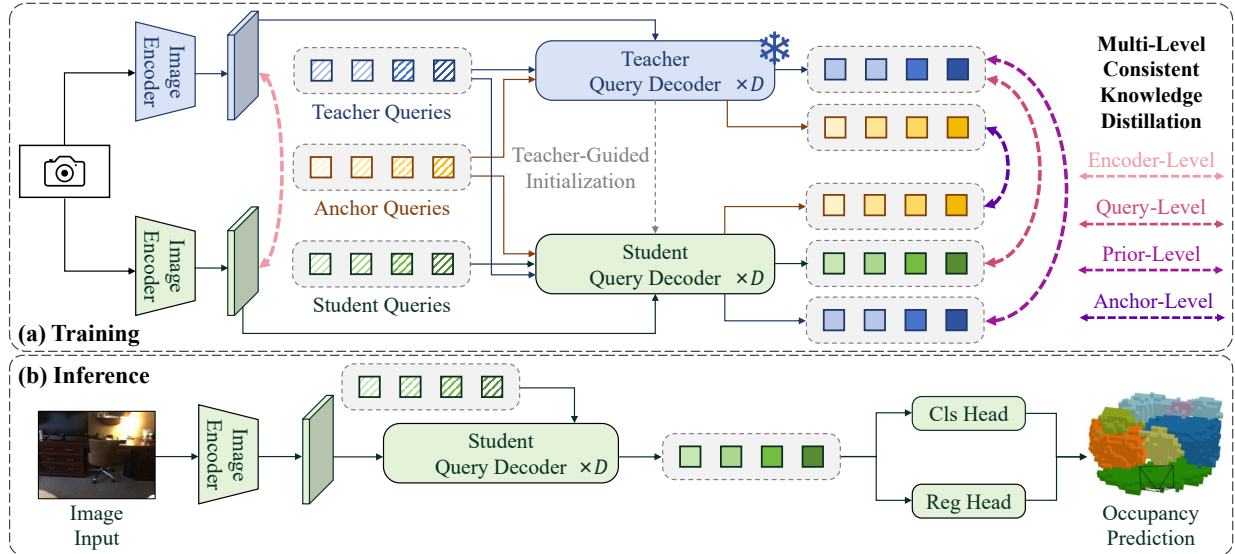


Fig. 2. (a) The illustration of our proposed knowledge distillation strategies. (b) The architecture of our primary framework. Best viewed in color.

minimize the discrepancy between the distributions of the two sets, \mathcal{S}^{GT} and $\mathcal{S}^{\text{Pred}}$.

Vanilla Pipeline. As illustrated in Fig. 2 (b), our baseline method leverages a transformer encoder-decoder architecture. Taking a set of N learnable queries $\{q_i\}_{i=1}^N$ as input, the model utilizes an image encoder to extract 2D features from the image input, and subsequently employs a multi-layer decoder to iteratively refine the queries using image features. At the end of each decoder layer, a regression and a classification head are used to output updated position and semantic predictions. To address the computational costs incurred by an excessive number of queries, each query q_i predicts the distribution of a point set $\{r_{i,j}\}_{j=1}^R$, where R denotes the number of points generated from one query, which is progressively increased across successive decoder layers in a coarse-to-fine manner.

Depth Branch. Building upon this, we introduce a simple yet effective depth branch inspired by [12], leveraging depth predictions from a pre-trained depth model as prior information. For a given query q and its corresponding point set center \bar{r} , We project \bar{r} into the camera space and subsequently onto the image plane, yielding its projected depth value d_q and corresponding 2D image coordinates (u, v) . Letting \tilde{I}_d denote the 2D depth prediction from the pre-trained depth model, we thus obtain the prior depth value at this location as $d_p = \tilde{I}_d(u, v)$. We then encode the projected and prior depth values into a depth-prior feature f_d using a simple MLP, which is eventually used to enhance the query feature f via channel-wise addition:

$$f_d = \text{MLP}(d_q, d_p), \hat{f} = f + f_d. \quad (1)$$

Loss Function. Both the vanilla and depth-prior settings utilize identical loss functions. We employ the Chamfer distance loss [49] to supervise the position predictions $\mathcal{P}^{\text{Pred}}$. For semantic supervision, a matched set of ground-truth semantic labels \mathcal{C}^{GT} is assigned to the predicted semantic labels $\mathcal{C}^{\text{Pred}}$ via nearest-neighbor matching. The semantic predictions are then optimized using focal loss [50]. Therefore, the overall

loss function for the task can be formulated as:

$$\mathcal{L}_{\text{task}} = \sum_{d=1}^D \mathcal{L}_{\text{CD}}(\mathcal{P}_d^{\text{Pred}}, \mathcal{P}_d^{\text{GT}}) + \mathcal{L}_{\text{focal}}(\mathcal{C}_d^{\text{Pred}}, \hat{\mathcal{C}}_d^{\text{GT}}), \quad (2)$$

where D is the number of decoder layers.

B. Multi-Level Consistent Knowledge Distillation

1) *Encoder-Level Feature Alignment:* In practice, our student and teacher models employ heterogeneous image encoders with different scales, leading significant divergence in their image feature representations. Since query features are substantially influenced by encoder outputs, we empirically find that this discrepancy largely hinders effective knowledge transfer and even compromises student performance. To cope with this issue, we adopt a simple yet effective feature alignment loss:

$$\mathcal{L}_{\text{EFA}} = \mathcal{L}_{\text{MSE}}\left(\frac{F^S}{\|F^S\|_2}, \frac{F^T}{\|F^T\|_2}\right), \quad (3)$$

where F^S and F^T denotes student and teacher image features.

2) *Query-Level Distillation:* A straightforward approach for knowledge distillation is to directly align the predictions of corresponding queries between the two models. However, our student and teacher queries lack ordered one-to-one correspondence, presenting a fundamental challenge for direct application. To resolve this misalignment, we establish an optimal bipartite matching $\hat{\sigma}$ between the N student queries $\{q_i^S\}_{i=1}^N$ and teacher queries $\{q_j^T\}_{i=1}^N$ using the Hungarian algorithm [51]. In practice, we employ the L2 distance between the point set centers from the student and teacher queries as the pair-wise matching cost of the cost matrix:

$$c_{ij} = c(q_i^S, q_j^T) = \|\bar{r}_i^S - \bar{r}_j^T\|_2. \quad (4)$$

This matching ensures consistent pairing of teacher-student predictions. Thus, the query-level distillation loss can be formulated as:

$$\mathcal{L}_{\text{QL}} = \frac{1}{N} \sum_{d=1}^D \sum_{i=1}^N \mathcal{L}_{\text{match}}(q_i^S, q_{\hat{\sigma}_i}^T). \quad (5)$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Since each query governs multiple point positions and semantics, we investigate two granularities of \mathcal{L}_{match} , respectively termed fine-grained logit-based distillation and coarse feature-based distillation, denoted as \mathcal{L}_{FLD} and \mathcal{L}_{CFD} . For \mathcal{L}_{FLD} , we supervise the 3D coordinates $r_{i,j}$ and output semantic logits $c_{i,j}$ for all R points within the point set associated with a matched query pair $(q_i^S, q_{\hat{\sigma}_i}^T)$. While for \mathcal{L}_{CFD} , we only supervise the point set center position \bar{r}_i and the query feature f_i . We use L1 loss, Kullback-Leibler Divergence loss and MSE loss for position, semantic and feature distillation, respectively:

$$\begin{aligned} \mathcal{L}_{FLD} &= \frac{1}{R} \sum_{j=1}^R \mathcal{L}_{L1}(r_{i,j}^S, r_{\hat{\sigma}_i,j}^T) + \mathcal{L}_{KL}(c_{i,j}^S, c_{\hat{\sigma}_i,j}^T), \\ \mathcal{L}_{CFD} &= \mathcal{L}_{L1}(\bar{r}_i^S, \bar{r}_{\hat{\sigma}_i}^T) + \mathcal{L}_{MSE}\left(\frac{f_i^S}{\|f_i^S\|_2}, \frac{f_{\hat{\sigma}_i}^T}{\|f_{\hat{\sigma}_i}^T\|_2}\right). \end{aligned} \quad (6)$$

Empirically, we observe that coarse feature-based distillation facilitates more effective knowledge transfer to the student model. Therefore, our final query-level distillation loss can be represented as follows:

$$\begin{aligned} \mathcal{L}_{QL} &= \frac{1}{N} \sum_{d=1}^D \sum_{i=1}^N \mathcal{L}_{CFD}(q_i^S, q_{\hat{\sigma}_i}^T) \\ &= \frac{1}{N} \sum_{d=1}^D \sum_{i=1}^N \mathcal{L}_{L1}(\bar{r}_i^S, \bar{r}_{\hat{\sigma}_i}^T) + \mathcal{L}_{MSE}\left(\frac{f_i^S}{\|f_i^S\|_2}, \frac{f_{\hat{\sigma}_i}^T}{\|f_{\hat{\sigma}_i}^T\|_2}\right). \end{aligned} \quad (7)$$

3) *Prior-Level Distillation*: Given that the student query embeddings are randomly initialized, their spatial distribution inherently diverges from the well-optimized teacher query embeddings. This discrepancy can cause unstable and sub-optimal bipartite matching during early training phases, which diminishes the effectiveness of query-level distillation and impedes model convergence. Since our coarse feature-based distillation aims to transfer the spatial distributions and feature representations of teacher queries, we decouple this process by first aligning spatial distributions between the two models using spatial priors as guidance, thus enabling the student to focus on feature learning and alleviating the mismatch problem. Based on this insight, we propose prior-level distillation.

Specifically, we input the teacher query embeddings into the student model to obtain an additional group of prior queries $\{q_i^P\}_{i=1}^N$ and their predictions. Since $\{q_i^P\}_{i=1}^N$ and $\{q_i^T\}_{i=1}^N$ share identical initialization distributions, we approximate their consistency and establish pairwise correspondences between the two sets of queries. This approach omits the bipartite matching process and further reduces the training time. Therefore, our prior-level distillation loss can be represented as:

$$\mathcal{L}_{PL} = \frac{1}{N} \sum_{d=1}^D \sum_{i=1}^N \mathcal{L}_{CFD}(q_i^P, q_i^T). \quad (8)$$

4) *Anchor-Level Distillation*: While our previous strategy enhances feature representation learning, directly distilling low-confidence predictions of the teacher may be harmful to the student model. Rather than manually filtering these suboptimal outputs, we introduce anchor-level distillation to ensure high-quality knowledge transfer. To achieve this goal, we sample N anchor points from the ground-truth set $\{\mathcal{P}^{GT}, \mathcal{C}^{GT}\}$ with

rebalanced weight according to the frequency distribution across different semantic classes. These anchors initialize the spatial distribution of a set of anchor queries, which are then fed into both models, obtaining updated student anchor queries $\{a_i^S\}_{i=1}^N$ and teacher anchor queries $\{a_i^T\}_{i=1}^N$.

This approach simultaneously guarantees spatial distribution consistency and restricts distillation exclusively to the high-confidence predictions of the teacher at anchor locations, thus establishing robust knowledge transfer by distilling only the most reliable knowledge representations. Analogous to our prior-level distillation strategy, knowledge transfer between corresponding anchor queries bypasses the need for bipartite matching due to their shared initialization, written as:

$$\mathcal{L}_{AL} = \frac{1}{N} \sum_{d=1}^D \sum_{i=1}^N \mathcal{L}_{CFD}(a_i^S, a_i^T). \quad (9)$$

5) *Distillation Loss*: To sum up, the overall loss function for knowledge distillation can be formulated as:

$$\mathcal{L}_{distill} = \lambda_1 \mathcal{L}_{EFA} + \lambda_2 \mathcal{L}_{QL} + \lambda_3 \mathcal{L}_{PL} + \lambda_4 \mathcal{L}_{AL}. \quad (10)$$

C. Teacher-Guided Initialization

Inspired by [52], we empirically find that the parameters of the decoder layers in the teacher model also serve as a source of knowledge. Despite employing heterogeneous encoders, the spatial and feature representations within the decoders exhibit inherent cross-model consistency. By initializing the student decoder with pre-trained weights from the teacher decoder, we significantly accelerate convergence while obtaining performance gains at no additional computational cost.

IV. EXPERIMENT

A. Experimental Setup

1) *Benchmark*: We adopt Occ-ScanNet [9] as the indoor occupancy prediction benchmark, which provides voxelized scenes in $60 \times 60 \times 36$ grids with 0.08m resolution, representing $4.8\text{m} \times 4.8\text{m} \times 2.88\text{m}$. Each voxel is annotated with 12 classes (11 semantic classes and 1 empty). Following common practices, we use mIoU and IoU as evaluation metrics.

2) *Implement Details*: For teacher model, we employ InternImage-XL [53] as the image backbone. The input image is resized to a resolution of 480×640 . For student model, we adopt a lightweight encoder ResNet-50 [54]. We train the model for 10 epochs on 8 A800 GPUs with a total batch size of 8 using the AdamW [55] optimizer. We set the learning rate to 2×10^{-4} and the hyperparameters as follows: $\lambda_1 = 1, \lambda_2 = 0.2, \lambda_3 = 0.2, \lambda_4 = 0.5$.

B. Quantitative and Qualitative Results

Comparison with SOTA methods. We first compare our method with competitive baselines on the validation set of Occ-ScanNet benchmark. As shown in Table I, our vanilla DiScene already surpasses most existing methods. Solely through distillation and initialization strategies, it elevates the mIoU of our baseline model OPUS by 36.1%, from 28.70 to 39.06. This performance marginally exceeds that of OPUS†,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
QUANTITATIVE COMPARISON ON THE OCC-SCANNET VALIDATION SET

Method	PDM	IoU	mIoU	ceiling	floor	wall	window	chair	bed	sofa	table	tv	furniture	objects	FPS
MonoScene [3]	×	41.60	24.62	15.17	44.71	22.41	12.55	26.11	27.03	35.91	28.32	6.57	32.16	19.84	8.0
ISO† [9]	DAv1	42.16	28.71	19.88	41.88	22.37	16.98	29.09	42.43	42.00	29.60	10.62	36.36	24.61	3.5
EmbodiedOcc† [12]	DAv2	53.95	45.48	40.90	50.80	41.90	33.00	41.20	55.20	61.90	43.80	35.40	53.50	42.90	6.5
OPUS [8]	×	35.58	28.70	15.37	37.75	20.60	18.64	26.43	44.55	45.63	30.79	14.63	35.80	25.49	23.2
OPUS† [8]	DAv2	45.62	38.96	39.06	45.04	34.97	28.63	35.92	49.27	54.39	37.93	23.93	45.04	34.42	<u>10.5</u>
DiScene (Ours)	×	43.68	39.06	29.66	45.28	28.70	28.73	35.90	53.13	56.89	39.90	30.07	44.97	36.38	23.2
DiScene† (Ours)	DAv2	<u>51.99</u>	47.17	45.21	<u>50.63</u>	<u>40.38</u>	36.73	42.28	59.68	62.04	45.60	41.17	<u>52.42</u>	<u>42.72</u>	<u>10.5</u>

† represents the result with pre-trained depth model, denoted as PDM. DAv1 and DAv2 are short for Depth Anything v1 [26] and v2 [10] respectively.

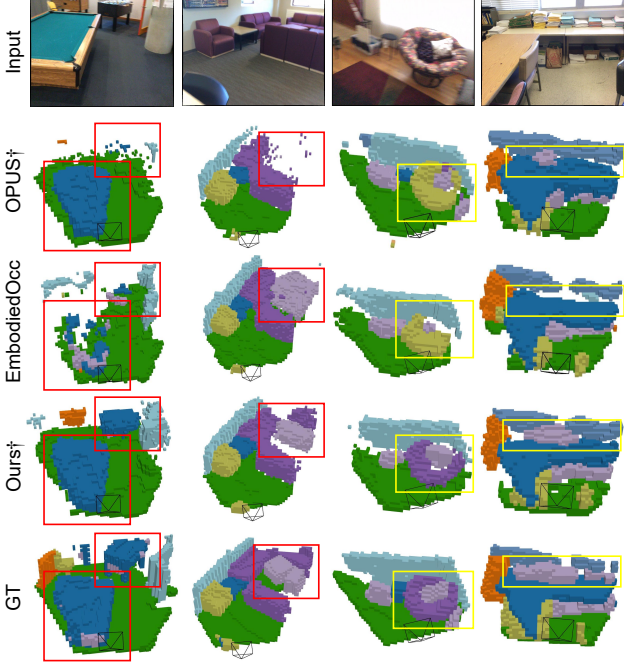


Fig. 3. Qualitative results of occupancy prediction on the Occ-ScanNet validation set. Compared with existing methods, DiScene demonstrates superior geometric awareness and semantic comprehension, visually highlighted by red and yellow boxes respectively.

which requires a pre-trained depth model, while maintaining the highest inference speed at 23.2 FPS. This demonstrates the effectiveness of our approach in balancing accuracy and real-time efficiency. Furthermore, when integrating the depth branch, our model advances the mIoU of the baseline OPUS† from 38.96 to 47.17, establishing a new SOTA and outperforming the previous best method EmbodiedOcc by 3.7%. Crucially, it retains real-time capability at 10.5 FPS, further validating the superiority of our method.

To illustrate the performance of our method more intuitively, we also provide qualitative visualizations in Fig. 3. Compared to the previous SOTA method EmbodiedOcc, our model demonstrates superior comprehension of geometry and semantics in complex and diverse indoor scenes. For instance, it accurately recognizes and reconstructs objects like the *table* in the first column and the *sofa* in the third column. Similarly, our approach outperforms the baseline method OPUS† in these scenarios, demonstrating enhanced capability in comprehending global scene structures and capturing finer local details. For example, our model successfully identifies the *sofa* at a distance in the second column and the *books* on the table in the last column, while existing methods fail in both cases. These findings underscore the efficacy of knowledge distillation in

TABLE II
EFFECTIVENESS OF KNOWLEDGE DISTILLATION

Method	PDM	Param (M)	IoU	mIoU
Teacher	×	379.1	52.79	48.42
OPUS	×	73.7	35.58	28.70
DiScene	×	73.7	43.68	39.06
Teacher†	M3Dv2-G	379.7 (1757.4)	59.84	56.58
OPUS†	DAv2	74.3 (172.0)	45.62	38.96
DiScene†	DAv2	74.3 (172.0)	51.99	47.17

We report learnable (w/o bracket) and total (w/ bracket) param. M3Dv2 denotes Metric3D v2 [11].

TABLE III
ABLATION STUDY OF EACH COMPONENT IN DISCENE

Query-Level	MCKD			TGI	IoU	mIoU
	Prior-Level	Anchor-Level				
✓					45.62	38.96
	✓				48.32	42.83
		✓			48.06	42.87
			✓		48.19	42.82
				✓	49.61	44.44
✓				✓	50.16	45.27
✓	✓			✓	50.35	45.61
✓	✓	✓		✓	51.99	47.17

strengthening scene understanding capabilities.

Effectiveness of knowledge distillation. The results are illustrated in Table II. Through knowledge distillation, we achieve substantial mIoU improvements of 36.10% and 21.07% for the student model under both settings. Concurrently, our approach reduces learnable parameters by over 80% compared to the teacher model, with nearly 90% total parameter reduction when incorporating the pre-trained depth model. These results validate the effectiveness of our method in balancing accuracy and computational costs, demonstrating strong suitability for practical deployment.

We further showcase the effectiveness of knowledge distillation in Fig. 4, which compares predictions from the non-distilled student, distilled student, and teacher models. For each row, we visualize occupancy predictions across models as well as the spatial distributions of activated queries for a specific semantic category. As demonstrated, the activated queries of the distilled student model exhibit significantly closer alignment with the teacher’s spatial distribution, which is particularly evident in the first row. After distillation, we observe a substantial increase in the quantity of activated queries. These queries concentrate closer to ground-truth regions, accompanied by remarkably improved prediction accuracy compared to the non-distilled baseline. These results confirm that our distillation strategy enables the student model to effectively learn the teacher’s spatial distributions and feature representations, thereby achieving performance gains. This validates both the correctness and efficacy of our Multi-level Consistent Knowledge Distillation framework.

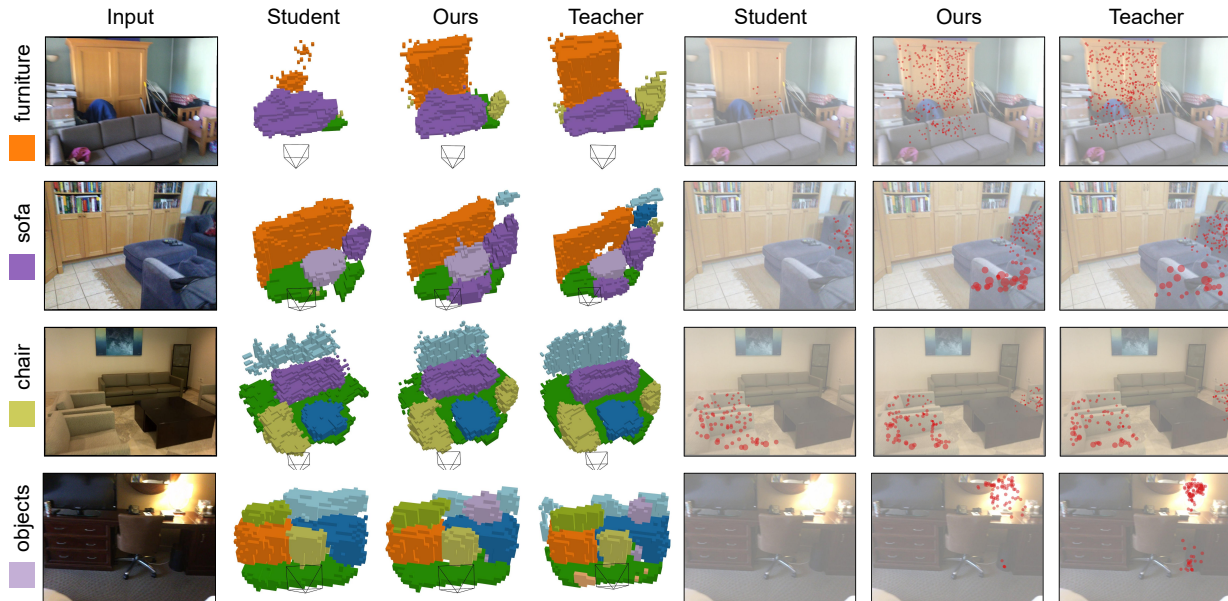


Fig. 4. Visualization of occupancy predictions and activated query distributions across non-distilled student, distilled student, and teacher models. Activated queries are highlighted in red, with higher density near ground-truth regions indicating superior performance. The size of the circle represents the distance of the query center: larger circles are closer to the camera. We adjust the opacity of certain figures for better illustration. Best viewed in color.

TABLE IV
ABLATION STUDY OF DISTILLATION STRATEGY

Distillation	IoU	mIoU
×	45.62	38.96
Query-Level	48.32	42.83
w/ FLD	47.07	41.70
w/o EFA	46.01	40.44
Prior-Level	48.06	42.87
w/ FLD	47.34	41.39
w/o EFA	46.45	40.40
Anchor-Level	48.19	42.82
w/ FLD	47.36	41.82
w/o EFA	45.73	38.45

TABLE V
ABLATION STUDY OF DIFFERENT PRE-TRAINED DEPTH MODEL

Model	FPS	IoU	mIoU
×	23.2	35.58	28.70
Depth Anything v1 [26]	7.1	41.13	34.46
Depth Anything v2 [10]	<u>10.5</u>	<u>45.62</u>	<u>38.96</u>
Metric3D v2-S [11]	8.4	43.77	37.82
Metric3D v2-G [11]	1.2	47.45	41.17

C. Ablation Study

Effects of each component. To investigate the impact of each component in DiScene, we report the performance of each module in Table III. When individually applying query-level, prior-level, and anchor-level distillation, we observe mIoU gains of 3.87, 3.91, and 3.86, respectively. This demonstrates the effectiveness of knowledge transfer at each distinct level. Furthermore, solely applying Teacher-Guided Initialization policy significantly boosts model performance by 5.48 mIoU, confirming its simplicity and efficacy. As all four components are progressively integrated into the framework, the model achieves steady performance improvements, culminating in a total gain of 8.21 mIoU. These findings underscore the importance and contribution of each component in our approach.

Selection of distillation strategy. In Table IV, we compare the impact of different distillation strategies on model performance. When replacing coarse feature-based distillation with fine-grained logit-based distillation (FLD), performance degradation is observed across all three levels. This decline is likely attributable to non-strict correspondence between point sets and voxels, where overly rigid fine-grained constraints may impede student learning. In contrast, coarse distillation imposes minimal restrictions on internal point distributions within the point set, thereby facilitating more effective model optimization. Furthermore, removing the encoder-level feature

alignment loss causes significant performance drops at all levels, with the mIoU performance of anchor-level distillation even falling below that of the non-distilled baseline. These results validate that direct knowledge distillation between models with heterogeneous encoders suffers from substantial feature discrepancy, while our feature alignment loss effectively mitigates this issue.

Selection of pre-trained depth model. Table V presents model performance using different pre-trained depth models. We evaluated two models producing relative depth estimations, Depth Anything v1 [26] and v2 [10] (both *fine-tuned* on indoor scenes to get metric outputs), alongside Metric3D v2 [11], a *zero-shot* model producing metric estimations. These results reveal that models integrating Depth Anything v2 achieve the fastest inference speed among all depth-enhanced variants while delivering the second-best mIoU performance. Conversely, models utilizing Metric3D v2-G attain peak accuracy but suffer from severely constrained inference speeds. Based on these observations, our DiScene[†] strategically employs Depth Anything v2 in the student model to strike an optimal accuracy-speed balance, while adopting Metric3D v2-G in the teacher model to ensure demonstrably more robust performance.

D. Robustness Analysis

In this section, we investigate the robustness of our distillation strategy in outdoor driving scenarios.

Experimental setup. Our experiments are conducted on the Occ3D-nuScenes benchmark [13], which provides dense

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

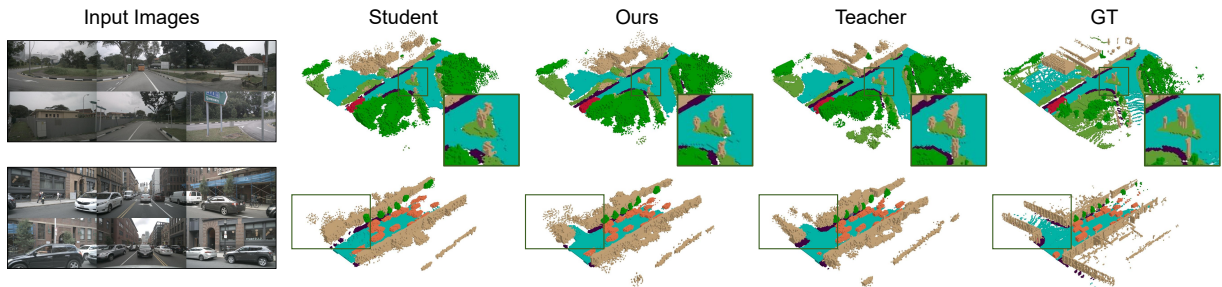


Fig. 5. Qualitative results of occupancy prediction on the Occ3D-nuScenes validation set. The boxes highlight finer local detail capture (Row 1) and enhanced scene structure reconstruction (Row 2) achieved by our distilled model.

TABLE VI
PERFORMANCE ON THE OCC3D-NUSCENES DATASET

Method	Param (M)	mIoU	RayIoU	RayIoU _{1m, 2m, 4m}		
Teacher	382.9	33.82	37.9	30.9	38.9	43.9
OPUS	77.5	28.31	32.8	26.2	33.7	39.0
DiScene	77.5	30.27	34.7	27.9	35.5	40.7

semantic occupancy annotations for the widely used nuScenes dataset [56]. Each voxel is annotated with 18 classes (17 semantic classes and 1 free). Following [8], we use mIoU and RayIoU as evaluation metrics. Implementation details remain consistent with Section IV-A2.

Quantitative results. As evidenced in Table VI, the integration of distillation and initialization strategies yields a 6.92% mIoU improvement and 5.79% RayIoU gain over the baseline, while reducing learnable parameters by nearly 80% compared to the teacher model. These results demonstrate the efficacy of our method in outdoor scenarios, achieving performance gains with reduced computational overhead, thus confirming its robustness in both indoor and outdoor environments.

Qualitative results. We further visualize the prediction results in Fig. 5. Our distilled model demonstrates markedly superior capabilities over the baseline in capturing local details and obtaining holistic structures. In the first row, the baseline erroneously predicts the distribution of *poles* at the intersection center, while the second row reveals its inaccurate *road structure* reconstruction. These limitations are effectively resolved through distillation, yielding predictions that closely align with the teacher model and exhibit enhanced scene comprehension capabilities. Collectively, these results validate the effectiveness and robustness of our method across diverse perception scenarios. Furthermore, the demonstration of in-the-wild scenes in Fig. 6 indicates the versatility of our approach.

E. Failure Cases

Fig. 7 illustrates several failure cases of our approach, in which the student model still struggles to effectively learn from the teacher through distillation. These cases typically occur when objects of a certain category exhibit both high density and large spatial distribution in the image, often accompanied by partial occlusion, such as the *books* on the bookshelf in the first row and the *chairs* in the second row. Such scenarios provide an abundance of intricate visual cues, which significantly increases the difficulty of learning both spatial distributions and feature representations, thereby limiting the efficacy of knowledge distillation. We believe that this issue could be addressed by incorporating instance-level priors, meriting deeper investigation in future work.

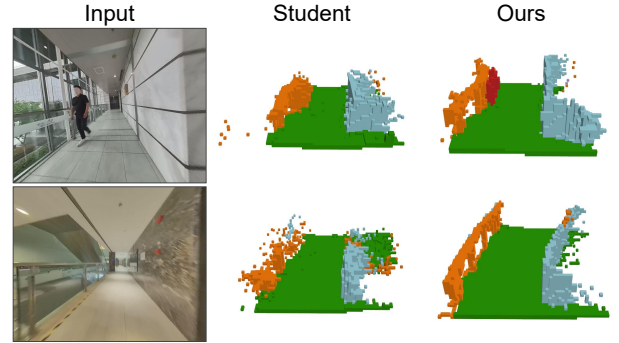


Fig. 6. Qualitative results of occupancy prediction on self-collected in-the-wild datasets. Our method demonstrates enhanced capabilities in geometric and semantic understanding.

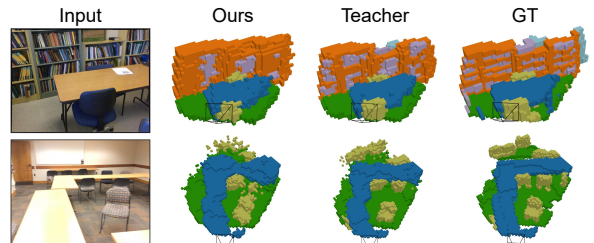


Fig. 7. Failure cases of our approach.

V. CONCLUSION

In this paper, we present DiScene, a novel framework for sparse query-based occupancy prediction. We propose Multi-level Consistent Knowledge Distillation, a hierarchical distillation strategy incorporating coordinated distillation across multiple complementary levels. This approach ensures consistent feature alignment and robust knowledge transfer, significantly boosting the performance of student model. Moreover, we introduce a Teacher-Guided Initialization policy that significantly accelerates convergence and enhances model performance without incurring additional computational costs. Our method optimally balances real-time efficiency with prediction accuracy, establishing new SOTA performance on the Occ-ScanNet benchmark while demonstrating robustness across diverse environments. We hope that DiScene can establish a practical paradigm for enhancing 3D perception in resource-constrained and complex indoor scenarios.

REFERENCES

- [1] M. Popović, F. Thomas, S. Papatheodorou, N. Funk, T. Vidal-Calleja, and S. Leutenegger, "Volumetric occupancy mapping with probabilistic depth completion for robotic navigation," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5072–5079, 2021.
- [2] L. Wang, H. Ye, Q. Wang, Y. Gao, C. Xu, and F. Gao, "Learning-based 3d occupancy prediction for autonomous navigation in occluded environments," in *IROS*, 2021.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

- [3] A.-Q. Cao and R. De Charette, "Monoscene: Monocular 3d semantic scene completion," in *CVPR*, 2022.
- [4] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion," in *CVPR*, 2023.
- [5] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.
- [6] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *AAAI*, 2023.
- [7] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *CVPR*, 2023.
- [8] J. Wang, Z. Liu, Q. Meng, L. Yan, K. Wang, J. Yang, W. Liu, Q. Hou, and M. Cheng, "Opus: occupancy prediction using a sparse set," *NeurIPS*, 2024.
- [9] H. Yu, Y. Wang, Y. Chen, and Z. Zhang, "Monocular occupancy prediction for scalable indoor scenes," in *ECCV*, 2024.
- [10] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *NeurIPS*, 2024.
- [11] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 579–10 596, 2024.
- [12] Y. Wu, W. Zheng, S. Zuo, Y. Huang, J. Zhou, and J. Lu, "Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding," in *ICCV*, 2025.
- [13] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *NeurIPS*, 2023.
- [14] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *ICCV*, 2023.
- [15] W. Tong, C. Sima, T. Wang, L. Chen, S. Wu, H. Deng, Y. Gu, L. Lu, P. Luo, D. Lin, *et al.*, "Scene as occupancy," in *ICCV*, 2023.
- [16] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez, "Fb-occ: 3d occupancy prediction based on forward-backward view transformation," *arXiv preprint arXiv:2307.01492*, 2023.
- [17] M. Pan, J. Liu, R. Zhang, P. Huang, X. Li, H. Xie, B. Wang, L. Liu, and S. Zhang, "Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision," in *ICRA*, 2024.
- [18] X. Li, P. Li, Y. Zheng, W. Sun, Y. Wang, and Y. Chen, "Semi-supervised vision-centric 3d occupancy world model for autonomous driving," in *ICLR*, 2025.
- [19] S. Gu, W. Yin, B. Jin, X. Guo, J. Wang, H. Li, Q. Zhang, and X. Long, "Dome: Taming diffusion model into high-fidelity controllable occupancy world model," *arXiv preprint arXiv:2410.10429*, 2024.
- [20] B. Jin, S. Gu, X. Hu, Y. Zheng, X. Guo, Q. Zhang, X. Long, and W. Yin, "Occtens: 3d occupancy world model via temporal next-scale prediction," *arXiv preprint arXiv:2509.03887*, 2025.
- [21] Z. Yu, C. Shu, J. Deng, K. Lu, Z. Liu, J. Yu, D. Yang, H. Li, and Y. Chen, "Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin," *arXiv preprint arXiv:2311.12058*, 2023.
- [22] J. Hou, X. Li, W. Guan, G. Zhang, D. Feng, Y. Du, X. Xue, and J. Pu, "Fastocc: Accelerating 3d occupancy prediction by fusing the 2d bird's-eye view and perspective view," in *ICRA*, 2024.
- [23] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction," in *ECCV*, 2024.
- [24] Y. Huang, A. Thammatadatrakoon, W. Zheng, Y. Zhang, D. Du, and J. Lu, "Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction," in *CVPR*, 2025.
- [25] H. Liu, Y. Chen, H. Wang, Z. Yang, T. Li, J. Zeng, L. Chen, H. Li, and L. Wang, "Fully sparse 3d occupancy prediction," in *ECCV*, 2024.
- [26] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.
- [27] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [28] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *CVPR*, 2018.
- [29] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *AAAI*, 2020.
- [30] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *CVPR*, 2022.
- [31] S. Yang, J. Yang, M. Zhou, Z. Huang, W.-S. Zheng, X. Yang, and J. Ren, "Learning from human educational wisdom: A student-centered knowledge distillation method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4188–4205, 2024.
- [32] S. Yang, X. Yang, J. Ren, L. Xu, J. Yang, Z. Huang, Z. Gong, and W. Wang, "Adaptive temperature distillation method for mining hard samples' knowledge," *Neurocomputing*, vol. 636, p. 129745, 2025.
- [33] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *ICCV*, 2019.
- [34] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *ICLR*, 2020.
- [35] L. Zhang, Y. Shi, Z. Shi, K. Ma, and C. Bao, "Task-oriented feature distillation," *NeurIPS*, 2020.
- [36] C. Pham, V.-A. Nguyen, T. Le, D. Phung, G. Carneiro, and T.-T. Do, "Frequency attention for knowledge distillation," in *WACV*, 2024.
- [37] J. Fan, C. Li, X. Liu, and A. Yao, "Scalekd: Strong vision transformers could be excellent teachers," in *NeurIPS*, 2024.
- [38] Q. Jin, J. Ren, O. J. Woodford, J. Wang, G. Yuan, Y. Wang, and S. Tulyakov, "Teachers do more than teach: Compressing image-to-image models," in *CVPR*, 2021.
- [39] L. Zhang, X. Chen, X. Tu, P. Wan, N. Xu, and K. Ma, "Wavelet knowledge distillation: Towards efficient image-to-image translation," in *CVPR*, 2022.
- [40] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *CVPR*, 2019.
- [41] J. Guo, K. Han, Y. Wang, H. Wu, X. Chen, C. Xu, and C. Xu, "Distilling object detectors via decoupled features," in *CVPR*, 2021.
- [42] G. Li, X. Li, Y. Wang, S. Zhang, Y. Wu, and D. Liang, "Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation," in *AAAI*, 2022.
- [43] Y. Wang, X. Li, S. Weng, G. Zhang, H. Yue, H. Feng, J. Han, and E. Ding, "Kd-detr: Knowledge distillation for detection transformer with consistent distillation points sampling," in *CVPR*, 2024.
- [44] Y. Hou, X. Zhu, Y. Ma, C. C. Loy, and Y. Li, "Point-to-voxel knowledge distillation for lidar semantic segmentation," in *CVPR*, 2022.
- [45] Z. Chong, X. Ma, H. Zhang, Y. Yue, H. Li, Z. Wang, and W. Ouyang, "Monodistill: Learning spatial features for monocular 3d object detection," in *ICLR*, 2022.
- [46] L. Zhang, R. Dong, H.-S. Tai, and K. Ma, "Pointdistiller: Structured knowledge distillation towards efficient and compact 3d detection," in *CVPR*, 2023.
- [47] Z. Xia, Y. Liu, X. Li, X. Zhu, Y. Ma, Y. Li, Y. Hou, and Y. Qiao, "Scpnet: Semantic scene completion on point cloud," in *CVPR*, 2023.
- [48] Y. Zheng, X. Li, P. Li, Y. Zheng, B. Jin, C. Zhong, X. Long, H. Zhao, and Q. Zhang, "Monoocc: Digging into monocular semantic occupancy prediction," in *ICRA*, 2024.
- [49] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *CVPR*, 2017.
- [50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017.
- [51] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
- [52] G. Li, W. Wang, X. Li, Z. Li, J. Yang, J. Dai, Y. Qiao, and S. Zhang, "Distilling knowledge from large-scale image models for object detection," in *ECCV*, 2024.
- [53] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *CVPR*, 2023.
- [54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.
- [56] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.