CLIPDet3D: Vision-Language Collaborative Distillation for 3D Object Detection

Jiaqi Zhao[1,2], **Huanfeng Hu**[1,2], **Yong Zhou**[1,2*], **Wen-Liang Du**[1,2], **Kunyang Sun**[1,2], **Rui Yao**[1,2], **Qigong Sun**[3]

[1]School of Computer Science and Technology / School of Artificial Intelligence, China University of Mining and Techology, Xuzhou 221116, China

[2]Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou 221116, China

[3]SenseTime Research, Shanghai 200000, China

{jiaqizhao, huanfenghu, yzhou, wldu, kunyang_sun, ruiyao}@cumt.edu.cn, sunqigong@sensetime.com

Abstract

Multi-view 3D object detection plays a vital role in autonomous driving systems due to its ability to perceive complex scenes accurately. However, real-world driving data often exhibits a long-tailed distribution, causing significant drops in detection accuracy for rare categories in existing methods. To mitigate this issue, we propose CLIPDet3D, a novel vision-language collaborative framework for multi-view 3D object detection. First, to tackle the difficulty of capturing the semantic information of rare categories, a Vision-Language Collaborative Learning strategy is proposed to incorporate class-level semantic priors from CLIP. Second, a Depth Feature Contrastive Distillation module is designed to overcome the large depth estimation error for rare categories by aligning depth features between a teacher and a student network. Furthermore, to alleviate the difficulty in focusing on regions of rare categories, a Dual-Stream Prompt Attention mechanism is devised to inject learnable prompts and compute attention along both horizontal and vertical BEV directions. Evaluations on the nuScenes dataset demonstrate that CLIPDet3D achieves state-of-the-art accuracy while maintaining efficient inference.

Code — https://github.com/Rory-Hu/CLIPDet3D

Introduction

3D object detection is crucial for understanding real-world scenes in autonomous driving. Multi-view methods have become popular due to their low cost and strong performance. These approaches project image features from surround view cameras into a shared bird's eye view (BEV) space for spatial reasoning. Recent works have shown promising results on large-scale datasets like nuScenes. However, most existing methods suffer from the long-tailed distribution of object categories. As shown in Fig. 1 (a), existing methods still achieve low mAP for rare categories such as buses and trailers, due to their underrepresentation in the training data, leading to limited detection accuracy and increased safety risks in critical scenarios.

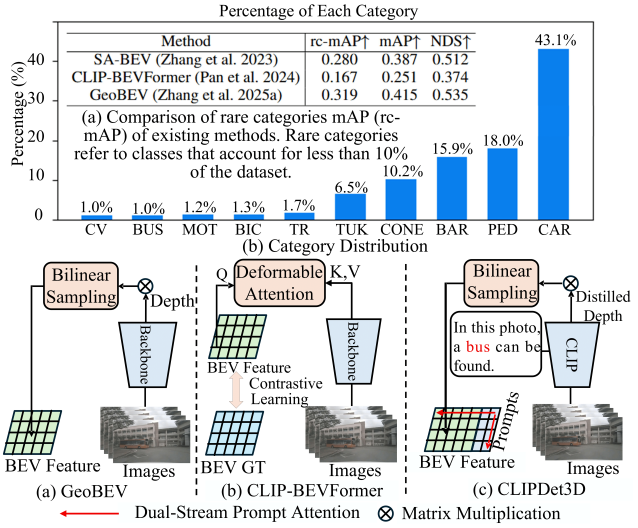

Figure 1: (a) shows rare categories mAP (rc-mAP) of existing methods. (b) shows the long-tailed category distribution in nuScenes. (c–e) Compare three representative approaches. GeoBEV (Zhang et al. 2025a) uses depth prediction with bilinear sampling to build BEV features. CLIP-BEVFormer (Pan et al. 2024) introduces contrastive learning to align BEV features with ground-truth supervision. CLIPDet3D (ours) improves rare-region representation by using CLIP-guided prompts, dual-stream attention, and depth feature contrastive distillation.

Recent multi-view 3D object detection methods are generally divided into dense feature projection-based methods and query-based sparse feature methods, depending on their intermediate representations. Dense feature projection methods project 3D queries onto multi-view 2D image planes to build dense BEV representations. CLIP-BEVFormer (Pan et al. 2024) enhances BEV feature learning by introducing contrastive supervision between image views and BEV elements, improving the alignment with ground-truth signals. WidthFormer (Yang et al. 2024a) uses 3D positional encoding and vertical compression to generate BEV features with a lightweight transformer efficiently. GeoBEV (Zhang

*Corresponding author.

et al. 2025a) designs RC-Sampling to enhance the geometric quality of BEV features with high resolution and fine details. Query-based sparse feature methods interact 3D queries with 2D image features using queries. PETR (Liu et al. 2022) maps image features to 3D space for better query interaction. StreamPETR (Wang et al. 2023a) introduces temporal query propagation to support long-term modeling. OPEN (Hou et al. 2025) enhances object center awareness through position embedding.

Despite great progress in multi-view 3D object detection, current methods still face challenges under long-tailed data distributions common in real-world driving. As shown in Fig. 1 (b), rare categories like buses and trailers appear infrequently, leading to low detection accuracy. Dense projection methods rely on accurate depth estimation, but depth prediction becomes unreliable for rare categories due to limited visual cues and sparse annotations. Although sparse query-based methods are efficient, their lack of semantic priors makes it difficult to detect rare categories accurately. In addition, existing attention mechanisms tend to focus on frequent categories, leaving regions of rare categories overlooked. These issues limit the model's performance on rare categories.

To address these issues, we propose CLIPDet3D, a novel multi-view 3D object detection framework that combines visual and language information. As shown at the bottom of Fig. 1, CLIPDet3D introduces the pretrained CLIP model to provide rich semantic priors. We first design a Vision-Language Collaborative Learning strategy. It uses class-level semantic embeddings from CLIP to help the model distinguish long-tailed categories from the background. Then, we introduce a Depth Feature Contrastive Distillation module. This module guides the student network to learn accurate depth features from a teacher network, which improves depth estimation for rare categories. In addition, we propose a Dual-Stream Prompt Attention mechanism. It injects learnable prompts along both horizontal and vertical BEV directions. This helps the model focus on rare categories and enhance their feature representation. Experiments on the nuScenes dataset show that CLIPDet3D significantly improves the detection of rare categories. It outperforms previous methods by up to 1.2% NDS and 1.8% mAP. It also improves inference speed from 3.9 to 4.5 FPS and reduces memory usage by 0.8 GB, proving its efficiency and robustness in real-world scenarios.

Overall, our significant contributions can be summarized as follows:

- Vision-Language Collaborative Learning is proposed to incorporate class-level semantic priors from CLIP, addressing the difficulty of capturing semantic information for rare categories and enhancing the model's overall semantic understanding.
- Depth Feature Contrastive Distillation is developed to align depth features between a teacher and a student network, overcoming the large depth estimation errors commonly seen in rare categories and improving spatial perception in underrepresented regions.
- Dual-Stream Prompt Attention is designed to inject learnable prompts and compute attention along both horizontal and vertical BEV directions, alleviating the difficulty in focusing on regions belonging to rare categories and strengthening the model's attention to challenging instances.
- Extensive ablation and comparison experiments are conducted on the nuScenes dataset, demonstrating that our model outperforms state-of-the-art methods under the same conditions.

Related Work

Multi-view 3D Object Detection

Bird's-Eye-View (BEV) representation is widely used in multi-view 3D detection. PETR (Liu et al. 2022) and PETRv2 (Liu et al. 2023) introduce 3D-aware queries and temporal modeling for end-to-end learning. BEVFormer (Li et al. 2022) uses spatial-temporal transformers to generate unified BEV features. View Transformation Modules (VTM) link image features to BEV using forward projection (e.g., Lift-Splat-Shoot (Philion and Fidler 2020)) or backward projection (e.g., BEVFormer (Li et al. 2022)), though each has limitations. FB-BEV (Li et al. 2023c) fuses both directions for better feature quality. GeoBEV (Zhang et al. 2025a) improves geometric accuracy through higher resolution and novel sampling.

Vision-Language Models

Vision-Language Models (Zhou et al. 2023; Gondal et al. 2024) have shown strong performance in 2D zero-shot and few-shot tasks. Recent efforts extend CLIP to 3D scene understanding: CLIP2Scene (Chen et al. 2023a) aligns CLIP semantics with point clouds via contrastive learning, while other methods (Parelli et al. 2023; Zhu et al. 2023) project image features into 3D space for distillation into point-based networks. PointCLIP V2 (Zhu et al. 2023) enhances 3D zero-shot classification by prompting CLIP and GPT with shape and text cues. Some works (Zhang et al. 2025b; Chen et al. 2024) explore CLIP for pretraining and open-vocabulary tasks to boost semantic reasoning.

Knowledge Distillation in 3D Vison

Knowledge distillation was initially introduced for model compression by transferring knowledge from a teacher to a student model. In 3D object detection, recent works (Wang et al. 2023b; Chen et al. 2023b; Zhao et al. 2024) leverage LiDAR-based detectors to guide image-based models. For example, DistillBEV (Wang et al. 2023b) transfers multi-scale BEV features with temporal fusion, while BEVDistill (Chen et al. 2023b) aligns BEV features across modalities without depth supervision. In monocular depth estimation, methods (He et al. 2025; Yang et al. 2024b; Shi et al. 2023) improve depth quality by distilling pseudo-labels from strong teacher models. Depth Anything V2 (Yang et al. 2024b) further scales this with synthetic data and large pseudo-labeled datasets.

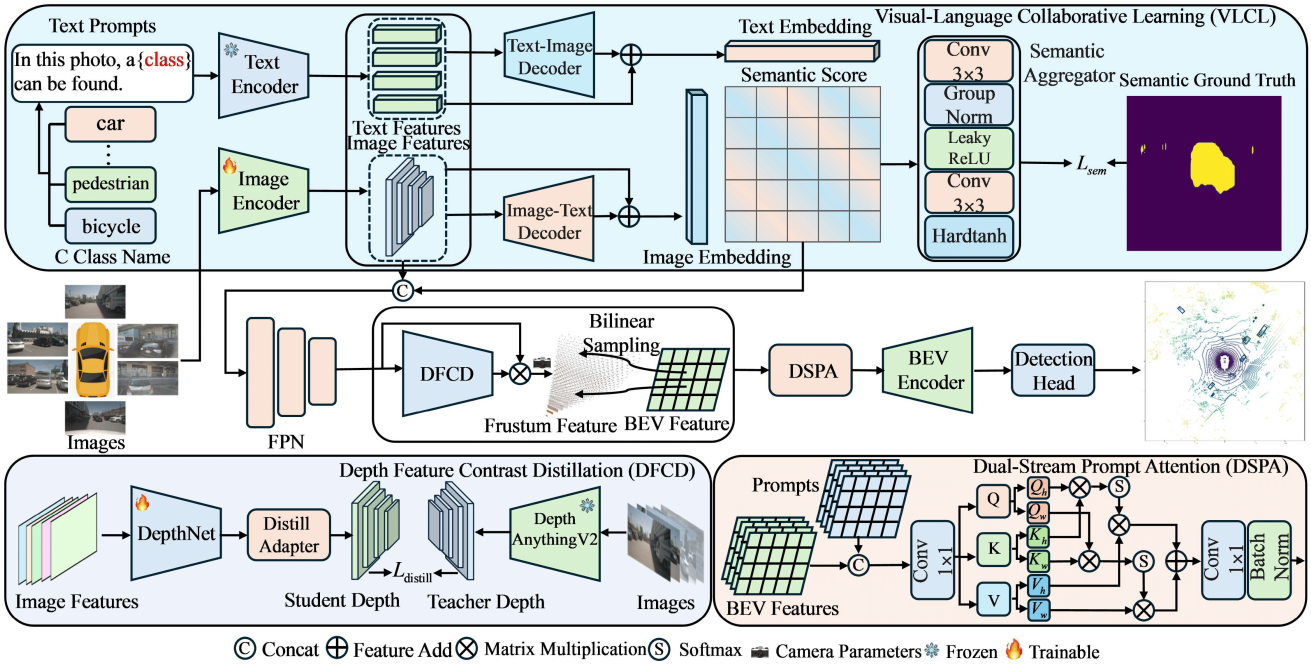

Figure 2: Overall architecture of CLIPDet3D. Multi-view images are first encoded into features and depth maps. Vision-Language Collaborative Learning injects CLIP-based priors to enhance the semantics of rare categories. Depth Feature Contrastive Distillation refines spatial perception. Dual-Stream Prompt Attention further focuses on informative regions for accurate detection.

Method

The overall architecture of our proposed CLIPDet3D is shown in Fig. 2. First, multi-view images are processed by a shared image encoder to extract multi-scale image features. These features are used for both depth estimation and vision-language interaction. Then, the Vision-Language Collaborative Learning module introduces semantic priors from Contrastive Language-Image Pre-training by aligning bird's eye view features with text-guided semantic representations. At the same time, the Depth Feature Contrastive Distillation module aligns the depth features between a student depth network and a teacher model (Depth Anything V2 (Yang et al. 2024b)), enhancing the spatial understanding of underrepresented objects. Finally, the Dual-Stream Prompt Attention module guides the bird's eye view encoder to focus on rare categories regions from both horizontal and vertical directions, improving detection accuracy under long-tailed distributions.

Vision-Language Collaborative Learning

To enhance visual features with class-level semantics and improve recognition of rare categories in long-tailed scenarios, we propose a Vision-Language Collaborative Learning module. It incorporates CLIP-derived textual embeddings as semantic priors to guide spatial visual representations through bidirectional interaction. As illustrated in the top of Fig. 2, the module takes multi-view images and textual prompts as input, extracting textual features via a frozen CLIP Text Encoder and image features via a trainable Image Encoder. The module consists of two stages: Semantic Infusion and Semantic Alignment.

In Semantic Infusion, as shown in Fig. 3, the Image-Text Decoder treats pixel-level visual features $\mathbf{V} \in R^{B\times C\times H\times W}$ as queries and textual features $\mathbf{T} \in \mathbb{R}^{B\times K\times C}$ as keys and values. The visual update is:

$$\Delta\mathbf{V} = Softmax\Big([\mathbf{W}_q\Phi(\mathbf{V})](\mathbf{T}\mathbf{W}_k)^\top\sqrt{C}\Big)(\mathbf{T}\mathbf{W}_v), \quad (1)$$

$$\mathbf{V}_{\text{new}} = \mathbf{V} + \gamma_v \cdot \mathcal{R}(\Delta\mathbf{V}), \quad (2)$$

where $\Phi(\cdot)$ means self-attention, $\mathcal{R}(\cdot)$ reshapes it back to feature maps, and $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{C\times C}$ are learnable projections, γ_v is a learnable scalar. The Text-Image Decoder enhances textual features by treating them as queries and using concatenated global-pixel visual contexts $\mathbf{V}_{\text{ctx}} \in \mathbb{R}^{B\times(1+HW)\times C}$ as keys and values:

$$\Delta\mathbf{T} = Softmax\Big(\frac{[\mathbf{W}_q\Phi(\mathbf{T})](\mathbf{V}_{\text{ctx}}\mathbf{W}_k)^\top}{\sqrt{C}}\Big)(\mathbf{V}_{\text{ctx}}\mathbf{W}_v), \quad (3)$$

$$\mathbf{T}_{\text{new}} = \mathbf{T} + \gamma_t \cdot \Delta\mathbf{T}, \quad (4)$$

where γ_t is a learnable scalar.

In Semantic Alignment, we compute a similarity map between the refined image and text embeddings:

$$S_{\ell,h,w} = \frac{1}{\tau}\sum_{c=1}^{C}\frac{\mathbf{V}_{\text{new}}}{\sqrt{\sum_{c'}\mathbf{V}_{\text{new}}^2}}\cdot\frac{\mathbf{T}_{\text{new}}}{\sqrt{\sum_{c'}\mathbf{T}_{\text{new}}^2}}, \quad (5)$$

where τ is a learnable temperature. The map is passed through a Semantics Aggregator for semantic fidelity. We

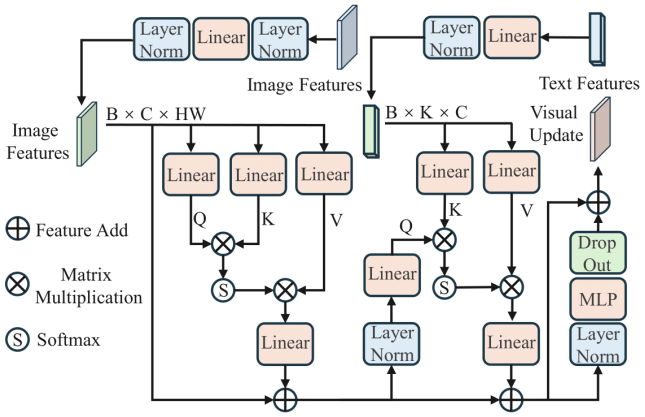

Figure 3: Details of Image-Text Decoder. The computation process of Text-Image Decoder is the same as that of Image-Text Decoder, except that the input features are different.

supervise it using a binary cross-entropy loss:

$$\mathcal{L}_{\text{sem}} = \sum_{\ell,h,w} \left[y_{\ell,h,w} \log S_{\ell,h,w} + (1 - y_{\ell,h,w}) \log(1 - S_{\ell,h,w}) \right], \tag{6}$$

where $y_{\ell,h,w}$ is the sampled ground truth mask. The resulting semantic scores are concatenated with image features and forwarded to downstream modules.

Depth Feature Contrastive Distillation

To address the significant depth estimation errors for rare categories, we design a Depth Feature Contrastive Distillation module. During feature generation, we use Depth Anything V2 to extract multi-channel teacher features from images, and DepthNet to obtain student depth features. A Distill Adapter is introduced to match the feature dimensions. To supervise the alignment, we design a contrastive distillation loss composed of three components: spatial, channel, and feature alignment. The channel alignment loss $\mathcal{L}_{\text{c}}$ is defined as:

$$\mathcal{L}_{\text{c}} = \frac{1}{C} \sum_{c=1}^{C} \left(\frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} s_c - \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} t_c \right)^2, \tag{7}$$

where s_c and t_c are the channel-wise attentions of student and teacher features. The spatial alignment loss $\mathcal{L}_{\text{s}}$ is computed in a similar manner across spatial positions. The feature alignment loss $\mathcal{L}_{\text{f}}$ strengthens local correspondence between features:

$$\mathcal{L}_{\text{f}} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{C} (F_t - A(F_s))^2 \cdot M}, \quad M = t_s \cdot t_c, \tag{8}$$

where $A(F_s)$ is the adapted student features, and F_t is teacher features, and M is a guidance mask derived from teacher's spatial (t_s) and channel (t_c) attentions. The final distillation loss is a weighted sum:

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{f}} + \mathcal{L}_{\text{c}} + \mathcal{L}_{\text{s}}. \tag{9}$$

This contrastive distillation scheme enables effective depth feature transfer from teacher to student, greatly enhancing the model's ability to perceive rare categories and improving overall depth estimation accuracy.

Dual-Stream Prompt Attention

To address the difficulty in focusing on rare category regions, we designed a Dual-Stream Prompt Attention. Dual-Stream Prompt Attention enhances feature representation through spatial prompts and dual orthogonal attention streams. It learns a spatial prompt tensor $\mathbf{P} \in \mathbb{R}^{1 \times C_p \times H \times W}$ initialized with constrained variance to limit early-stage perturbations. During processing, this tensor expands batch-wise to $\mathbf{P}_B$. Input features first undergo channel-wise layer normalization before concatenating with $\mathbf{P}_B$ along the channel dimension:

$$X = Concat(X, \mathbf{P}_B), \tag{10}$$

where X denotes input features. A 1×1 convolution then compresses the combined channels to fuse features and prompts.

The core Dual-Stream attention processes features along two spatial dimensions: the horizontal stream computes row-wise attention $A_H(Q_H, K_H, V_H)$, and the vertical stream computes column-wise attention $A_W(Q_W, K_W, V_W)$. Here, Q_H, K_H, V_H, and Q_W, K_W, V_W are the query, key, and value matrices for horizontal and vertical streams, respectively. Both streams share projection weights to maintain parameter efficiency. Their outputs combine as:

$$\text{Attn} = (A_H(Q_H, K_H, V_H) + A_W(Q_W, K_W, V_W)) \cdot \gamma + x, \tag{11}$$

where A_H and A_W denote horizontal and vertical attention functions, γ is a learnable scaling factor, and x is the original feature input. The result passes through a feature reconstruction block to produce improved representations.

Experiments

Dataset and Metrics. We evaluate our method on the nuScenes (Caesar et al. 2020) benchmark, a large-scale and widely adopted dataset for 3D object detection in autonomous driving. It contains 1,000 urban scenes recorded at 2 Hz, split into 700 training, 150 validation, and 150 testing scenes. Each scene provides multimodal data from 6 cameras, 1 LiDAR, and 5 radars. In this study, we focus on the image-only setting using the 6 cameras, which introduces additional challenges due to the absence of depth measurements. The dataset includes annotations for 10 object categories. We follow the official evaluation protocol, using the nuScenes Detection Score (NDS) as the primary metric. NDS combines mAP with five true positive metrics: mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), and mean Average Attribute Error (mAAE), to jointly assess localization, scale, orientation, motion, and attribute accuracy. Additionally, we report rc-mAP (rare categories mAP) to better evaluate the detection performance of objects at different distances, enabling a more comprehensive assessment under complex conditions.

Method	Backbone	Image Size	Frames	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
PETRv2 (Liu et al. 2023)	ResNet50	256×704	2	0.349	0.456	0.700	0.275	0.580	0.437	0.187
BEVDepth (Li et al. 2023b)	ResNet50	256×704	2	0.351	0.475	0.639	0.267	0.479	0.428	0.198
BEVStereo (Li et al. 2023a)	ResNet50	256×704	2	0.372	0.500	0.598	0.270	0.438	0.367	0.190
SA-BEV (Zhang et al. 2023)	ResNet50	256×704	2	0.387	0.512	0.613	0.266	0.352	0.382	0.199
BEVFormerv2 (Yang et al. 2023)	ResNet50	256×704	-	0.423	0.529	0.618	0.273	0.413	0.333	0.188
SOLOFusion* (Park et al. 2023)	ResNet50	256×704	17	0.427	0.534	0.567	0.274	0.511	**0.252**	**0.181**
StreamPETR* (Wang et al. 2023a)	ResNet50	256×704	8	0.450	0.550	0.613	0.267	0.413	0.265	0.196
BEVNeXt (Li et al. 2024)	ResNet50	256×704	8	0.437	0.548	0.550	0.265	0.427	0.260	0.208
CLIP-BEVFormer (Pan et al. 2024)	ResNet50	256×704	-	0.421	0.534	0.612	0.273	**0.348**	0.341	0.192
RecurrentBEV* (Chang et al. 2024)	ResNet50	256×704	8	0.445	0.549	0.555	0.272	0.451	0.256	0.204
GeoBEV (Zhang et al. 2025a)	ResNet50	256×704	2	0.415	0.535	0.533	0.265	0.419	0.298	0.214
BEVMamba* (Liu et al. 2025)	ResNet50	256×704	10	0.427	0.542	0.562	0.273	0.432	0.264	0.188
CLIPDet3D	ResNet50	256×704	2	0.433	0.547	0.532	**0.259**	0.385	0.302	0.213
CLIPDet3D*	ResNet50	256×704	8	**0.462**	**0.566**	**0.507**	0.266	0.405	**0.252**	0.206
PETRv2 (Liu et al. 2023)	ResNet101	900× 1600	2	0.421	0.524	0.681	0.267	0.357	0.377	**0.186**
BEVDepth (Li et al. 2023b)	ResNet101	512×1408	2	0.412	0.535	0.565	0.266	0.358	0.331	0.190
HoP* (Zong et al. 2023)	ResNet101	512×1408	8	0.454	0.558	0.565	0.265	0.329	0.337	0.194
CLIP-BEVFormer (Pan et al. 2024)	ResNet101	512×1408	-	**0.467**	0.562	0.605	0.253	0.331	0.336	0.187
VectorFormer (Pan et al. 2024)	ResNet101	512×1408	-	0.437	0.540	0.643	0.270	0.363	0.324	**0.186**
BEVHeight++ (Yang et al. 2025)	ResNet101	512×1408	2	0.423	0.554	0.541	0.262	**0.307**	**0.277**	0.187
BEVMamba* (Liu et al. 2025)	ResNet101	512×1408	7	0.453	0.567	**0.505**	0.261	0.345	0.292	0.191
CLIPDet3D	ResNet101	512×1408	2	**0.467**	**0.573**	0.513	**0.252**	0.328	0.300	0.208

Table 1: Comparison with previous state-of-the-art multi-view 3D detectors on the nuScenes val set. The best is in bold. The second-best model is underlined. * indicates the use of more than 2 frames.

Method	Backbone	Image Size	Frames	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVDet (Huang et al. 2021)	Swin-B	900×1600	1	0.424	0.488	0.524	**0.242**	0.373	0.950	0.148
SA-BEV (Zhang et al. 2023)	ResNet50	256×704	2	0.393	0.515	0.567	0.257	0.457	0.402	0.130
GeoBEV (Zhang et al. 2025a)	ResNet50	256×704	2	0.434	0.552	0.497	0.255	0.403	0.358	0.133
BEVFormer* (Li et al. 2022)	ResNet101	512×1408	4	0.445	0.535	0.631	0.257	0.405	0.435	0.143
PETRv2 (Liu et al. 2023)	ResNet101	640×1600	2	0.456	0.553	0.601	0.249	0.391	0.382	**0.123**
CLIP-BEVFormer (Pan et al. 2024)	ResNet101	512×1408	-	0.447	0.547	0.591	0.257	0.417	0.371	0.128
BEVMamba* (Liu et al. 2025)	ResNet101	900×1600	5	0.412	0.518	0.681	0.269	0.345	0.391	0.192
CLIPDet3D	ResNet50	256×704	2	0.439	0.560	**0.480**	0.255	0.386	**0.339**	0.131
CLIPDet3D	ResNet101	512×1408	2	**0.462**	**0.574**	0.495	0.247	**0.357**	0.340	0.126

Table 2: Comparison with previous state-of-the-art multi-view 3D detectors on the nuScenes test set.

Implementation Details

We conducted experiments on the nuScenes dataset using ResNet101 (He et al. 2016) and ResNet50 (He et al. 2016) as backbone networks. To ensure a fair comparison, we did not apply any test-time augmentation or incorporate future frame information during evaluation. In training CLIPDet3D, we avoided using pre-training from 2D object detection or perspective-view tasks. Instead, we adopted the same data augmentation strategies as the original method and employed the AdamW optimizer (Loshchilov and Hutter 2019). Except for regular data augmentation, we also adopted BEV-Paste (Zhang et al. 2023) to alleviate overfitting during the long training process. Training was performed on four NVIDIA RTX A6000 GPUs with a batch size of 8, following the CBGS strategy (Cui et al. 2019). For comparison with other methods on the nuScenes validation set, CLIPDet3D was trained for 20 epochs with an initial learning rate of 1×10^{-4}. In the ablation study, we chose GeoBEV (Zhang et al. 2025a) and BEVDet (Huang et al. 2021) as the baseline, used ResNet50 as the backbone, and trained for 20 epochs. All experiments were conducted under consistent settings to ensure reproducibility and fair evaluation.

Experimental Results

Main Results. We compare CLIPDet3D with state-of-the-art multi-view 3D detectors on the nuScenes validation and test sets. On the validation set (Tab. 1), CLIPDet3D achieves 56.6% NDS and 46.2% mAP with 8-frame input, surpassing StreamPETR by 1.6% NDS and 1.2% mAP. In the 2-frame setting, it outperforms BEVHeight++ (Yang et al. 2025) by up to 1.9% NDS and 4.4% mAP. In the test set (Tab. 2), CLIPDet3D reaches 57.4% NDS and 46.2% mAP, outperforming CLIP-BEVFormer (Pan et al. 2024) by 2.7% NDS and 1.5% mAP. It also achieves the best mATE, mAOE, and mAVE, validating the benefit of our depth and semantic designs, with clear improvements on rare categories from prompt-guided attention.

Model Efficiency. As shown in Tab. 6, CLIPDet3D achieves a strong balance between accuracy and efficiency.

Method	CV	BUS	MOT	BIC	TR	TUK	CONE	BAR	PED	CAR	rc-mAP
Total Num	650	657	748	857	1114	4215	6591	10263	11564	27727	-
Percentage	1.0%	1.0%	1.2%	1.3%	1.7%	6.5%	10.2%	15.9%	18.0%	43.1%	-
BEVFormer (Li et al. 2022)	5.8	23.3	21.4	20.3	6.6	19.2	38.4	37.9	33.2	45.7	16.1
SA-BEV (Zhang et al. 2023)	8.7	41.3	34.3	32.7	19.8	31.6	57.3	60.5	43.0	57.0	28.0
CLIP-BEVFormer (Pan et al. 2024)	7.1	28.0	26.1	21.6	8.1	20.9	41.1	40.0	33.9	46.8	18.6
GeoBEV (Zhang et al. 2025a)	10.2	43.9	40.7	39.3	**21.9**	35.9	64.2	59.2	51.0	63.2	31.9
CLIPDet3D	**15.1**	**46.5**	**48.0**	**47.1**	17.5	**37.4**	**66.9**	**61.7**	**56.6**	**65.0**	**35.2**

Table 3: Per-class performance comparison on nuScenes validation set. rc-mAP denotes the mean Average Precision (mAP) for rare categories (those accounting for less than 10%). CV represents Construction Vehicle, BUS represents Bus, MOT represents Motorcycle, BIC represents Bicycle, TR represents Trailer, TUK represents Truck, CONE represents Traffic Cone, BAR represents Barrier, PED represents Pedestrian, CAR represents Car.

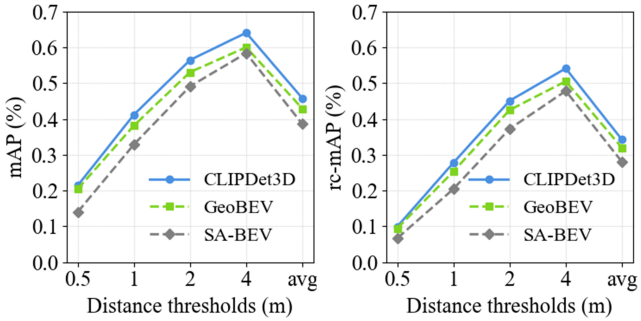

Figure 4: Comparison of mAP and rc-mAP at different distance thresholds. rc-mAP denotes the mean Average Precision (mAP) for rare categories (those accounting for less than 10%). The values 0.5, 1.0, 2.0, 4.0 represent the matching distance thresholds for mAP evaluation.

VLCL	DFCD	DSPA	mAP↑	NDS↑	mASE↓	mAVE↓
			0.415	0.535	0.265	**0.298**
✓			0.426	0.538	0.260	0.301
✓	✓		0.430	0.543	0.261	0.304
✓	✓	✓	**0.433**	**0.547**	**0.259**	0.302

Table 4: Performance metrics for different configurations of Vision-Language Collaborative Learning (VLCL), Depth Feature Contrastive Distillation (DFCD), and Dual-Stream Prompt Attention (DSPA).

With higher BEV resolution (256×256), it improves over BEVFormer (200×200) by 1.7% mAP, 3.0% NDS, increases FPS from 3.9 to 4.5, and reduces memory usage by 0.8 GB. Compared to VectorFormer, it achieves +0.8 mAP, +1.5 NDS, faster inference (4.5 vs. 3.4 FPS), and lower memory (3.9 vs. 4.85 GB), demonstrating better computational efficiency with finer spatial granularity.

Visualization. Fig. 5 shows that CLIPDet3D delivers more accurate and complete detections, especially for distant or small objects, while GeoBEV misses many targets. These improvements result from the enhanced capacity of CLIPDet3D to perceive and understand complex scenes. Notably, it can detect rare category instances that are often ignored by conventional methods, thanks to the injected semantic priors and guided attention.

Setting	rc-mAP↑	mAP↑	NDS↑	mASE↓
BEVDet (Huang et al. 2021)	0.173	0.277	0.336	0.304
+ VLCL (Without $\mathcal{L}_{\text{sem}}$)	0.151	0.270	0.346	**0.291**
+ VLCL (With $\mathcal{L}_{\text{sem}}$)	**0.197**	**0.311**	**0.359**	0.292

Table 5: Performance metrics for different configurations of Vision-Language Collaborative Learning (VLCL).

Method	Pub.	BEV Dim.	mAP↑	NDS↑	FPS	Mem.(GB)
BEVFormer	ECCV22	200×200	0.416	0.517	3.9	4.70
VectorFormer	ECCV24	200×200	0.425	0.532	3.4	4.85
CLIPDet3D	-	256×256	0.433	0.547	4.5	3.90

Table 6: Effectiveness and efficiency comparisons between BEVFormer (Li et al. 2022), VectorFormer (Pan et al. 2024) and our proposed CLIPDet3D.

Long-tail Detection Results

To evaluate long-tail performance, we compare CLIPDet3D with CLIP-BEVFormer (Pan et al. 2024) and GeoBEV (Zhang et al. 2025a) on the nuScenes validation set. As shown in Tab. 3, rare categories each account for less than 10% of the data. CLIPDet3D achieves substantial gains over CLIP-BEVFormer by 8.0%, 18.5%, 21.9%, 25.5%, 9.4%, and 16.5%, respectively. Overall, it obtains the highest rc-mAP of 35.2%, exceeding CLIP-BEVFormer and GeoBEV by 16.6% and 3.3%. As shown in Fig. 4, CLIPDet3D also consistently outperforms GeoBEV and SA-BEV across different matching thresholds. The improvement is particularly evident under stricter conditions, with +2.45% rc-mAP at 1.0 m and +2.57% at 2.0 m than GeoBEV, indicating better localization for rare categories. These results benefit from the combined effects of CLIP-guided prompts, depth-aware contrastive distillation, and dual-stream attention, which together improve recognition, geometry, and focus on rare categories regions.

Ablation Study

Effectiveness of Vision-Language Collaborative Learning. This module introduces CLIP-derived semantic priors to enhance recognition of rare categories. As shown in Tab. 4, adding it to GeoBEV improves NDS by 0.3% and mAP by 1.1%. Tab. 5 further shows that removing the semantic constraint reduces rc-mAP and mAP by 2.2% and

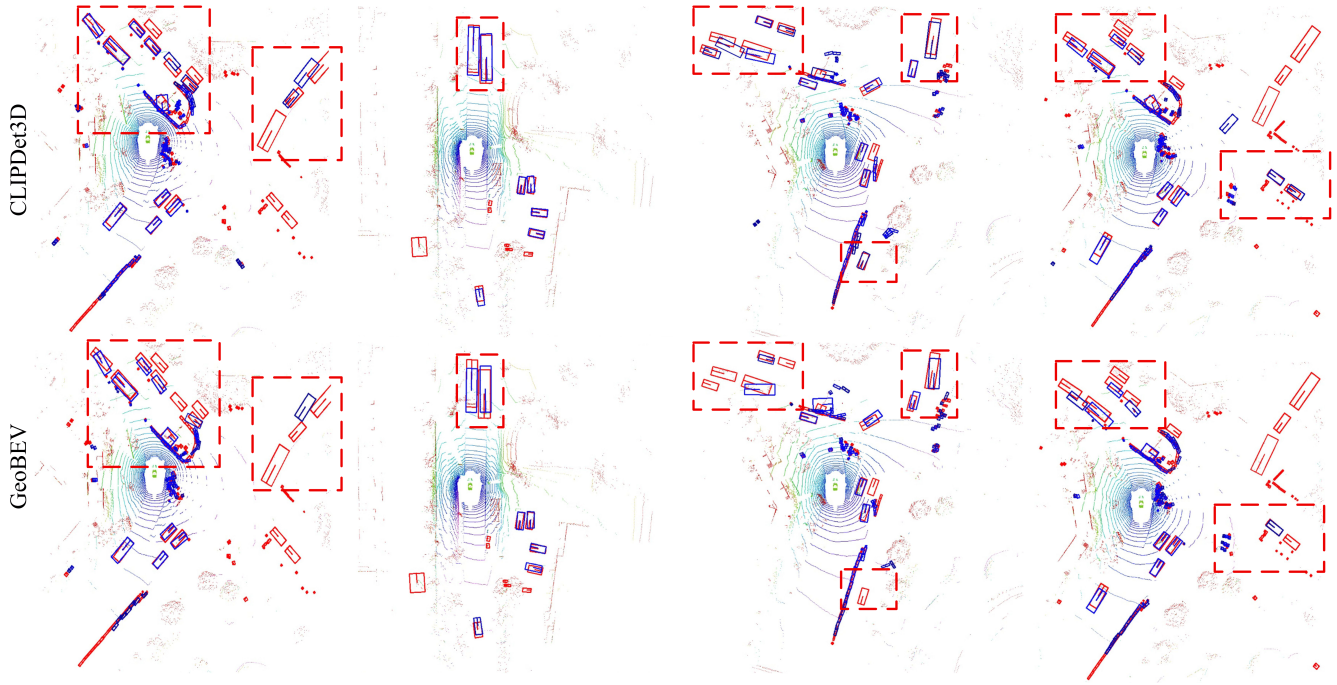

Figure 5: Visualization results on BEV representation of GeoBEV and CLIPDet3D(ours). The red boxes and blue boxes represent the ground truth and the predicted boxes, respectively. The dashed red rectangles illustrate that CLIPDet3D results in higher detection accuracy.

$\mathcal{L}_c$	$\mathcal{L}_s$	$\mathcal{L}_f$	mAP↑	NDS↑	mASE↓	mAVE↓	mAAE↓
✓			0.429	0.545	0.261	**0.291**	0.212
✓	✓		0.432	0.546	**0.258**	0.302	**0.210**
✓	✓	✓	**0.433**	**0.547**	0.259	0.302	0.213

Table 7: Performance metrics for different loss of Depth Feature Contrastive Distillation (DFCD).

Setting	rc-mAP↑	mAP↑	NDS↑	mASE↓
BEVDet (Huang et al. 2021)	0.173	0.277	0.336	0.304
+ DSPA (Without prompts)	0.164	0.268	0.341	**0.302**
+ DSPA (With prompt)	**0.178**	**0.280**	**0.349**	0.293

Table 8: Performance metrics for different configurations of Dual-Stream Prompt Attention (DSPA).

0.7%, while including it boosts rc-mAP by 2.4% and mAP by 3.4%. These results confirm the benefit of semantic guidance.

Effectiveness of Depth Feature Contrastive Distillation. Depth Feature Contrastive Distillation enhances spatial perception by aligning depth features between teacher and student networks. Adding this module yields gains of 0.5% NDS and 0.4% mAP (Tab. 4). As shown in Tab. 7, removing either the channel-wise loss $\mathcal{L}_c$ or the spatial component reduces performance, validating the necessity of both and confirming the effectiveness of the distillation module. In particular, the channel-wise contrast emphasizes semantic consistency, while the spatial component preserves geometric structures.

Effectiveness of Dual-Stream Prompt Attention. This module guides attention along both BEV axes using learnable prompts. Tab. 4 shows that it improves NDS by 0.4% and mAP by 0.3% and achieves the best mASE. As reported in Tab. 8, removing the prompts leads to a 0.9% drop in rc-mAP and a slight degradation of mAP. Including prompts improves rc-mAP, mAP, and NDS by 0.5%, 0.3%, and 1.3%, respectively, demonstrating their contribution to rare categories detection.

Conclusion

In this paper, we propose CLIPDet3D, a novel vision-language collaborative framework for multi-view 3D object detection. It mitigates the long-tailed distribution problem in autonomous driving by improving the recognition of rare categories and depth estimation. CLIPDet3D includes three key modules. Vision-Language Collaborative Learning uses CLIP-based semantic priors to help identify rare categories. Depth Feature Contrastive Distillation aligns depth features between teacher and student networks to enhance spatial reasoning. Dual-Stream Prompt Attention uses learnable prompts to guide the model toward rare category regions in the BEV space. Experiments on the nuScenes benchmark show that CLIPDet3D achieves higher accuracy, better efficiency, and stronger performance on rare categories compared to previous methods. In future work, we will integrate LiDAR input and extend the method to multi-task settings such as BEV segmentation and occupancy prediction.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62272461, Grant 62172417, and Grant 62277046; in part by the Double First-Class Project of China University of Mining and Technology for Independent Innovation and Social Service under Grant 2022ZZCX06; in part by the Six Talent Peaks Project in Jiangsu Province under Grant 2015-DZXX-010 and Grant 2018-XYDXX-044.

References

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.

Chang, M.; Zhang, X.; Zhang, R.; Zhao, Z.; He, G.; and Liu, S. 2024. Recurrentbev: A long-term temporal fusion framework for multi-view 3d detection. In *European Conference on Computer Vision*, 131–147.

Chen, L.; Wang, X.; Lu, J.; Lin, S.; Wang, C.; and He, G. 2024. Clip-driven open-vocabulary 3d scene graph generation via cross-modality contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27863–27873.

Chen, R.; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; and Wang, W. 2023a. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7020–7030.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; and Zhao, F. 2023b. BEVDistill: Cross-Modal BEV Distillation for Multi-View 3D Object Detection. In *The Eleventh International Conference on Learning Representations*.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Gondal, M. W.; Gast, J.; Ruiz, I. A.; Droste, R.; Macri, T.; Kumar, S.; and Staudigl, L. 2024. Domain aligned clip for few-shot classification. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, 5721–5730.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

He, X.; Guo, D.; Li, H.; Li, R.; Cui, Y.; and Zhang, C. 2025. Distill any depth: Distillation creates a stronger monocular depth estimator. *arXiv preprint arXiv:2502.19204*.

Hou, J.; Wang, T.; Ye, X.; Liu, Z.; Gong, S.; Tan, X.; Ding, E.; Wang, J.; and Bai, X. 2025. Open: Object-wise position embedding for multi-view 3d object detection. In *European conference on computer vision*, 146–162.

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*.

Li, Y.; Bao, H.; Ge, Z.; Yang, J.; Sun, J.; and Li, Z. 2023a. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1486–1494.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2023b. 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.; Lan, S.; Alvarez, J. M.; and Wu, Z. 2024. Bevnext: Reviving dense bev frameworks for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20113–20123.

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

Li, Z.; Yu, Z.; Wang, W.; Anandkumar, A.; Lu, T.; and Alvarez, J. M. 2023c. Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6919–6928.

Liu, X.; Zhong, J.; Sun, C.; and Zhang, X. 2025. BEV-Mamba: Time Sequence Dense Bird's-Eye-View Perception Modeling With State Space Model. *IEEE Transactions on Intelligent Transportation Systems*, 1–11.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022. Petr: Position embedding transformation for multi-view 3d object detection. In *European conference on computer vision*, 531–548.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, A.; Wang, T.; and Zhang, X. 2023. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3262–3272.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Pan, C.; Yaman, B.; Velipasalar, S.; and Ren, L. 2024. Clip-bevformer: Enhancing multi-view image-based bev detector with ground truth flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15216–15225.

Parelli, M.; Delitzas, A.; Hars, N.; Vlassis, G.; Anagnostidis, S.; Bachmann, G.; and Hofmann, T. 2023. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5607–5612.

Park, J.; Xu, C.; Yang, S.; Keutzer, K.; Kitani, K.; Tomizuka, M.; and Zhan, W. 2023. Time Will Tell: New Outlooks and A Baseline for Temporal Multi-View 3D Object Detection. In *International Conference on Learning Representations*.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, 194–210.

Shi, X.; Dikov, G.; Reitmayr, G.; Kim, T.-K.; and Ghafoorian, M. 2023. 3d distillation: Improving self-supervised monocular depth estimation on reflective surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9133–9143.

Wang, S.; Liu, Y.; Wang, T.; Li, Y.; and Zhang, X. 2023a. Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3621–3631.

Wang, Z.; Li, D.; Luo, C.; Xie, C.; and Yang, X. 2023b. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8637–8646.

Yang, C.; Chen, Y.; Tian, H.; Tao, C.; Zhu, X.; Zhang, Z.; Huang, G.; Li, H.; Qiao, Y.; Lu, L.; et al. 2023. BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17830–17839.

Yang, C.; Lin, T.; Huang, L.; and Crowley, E. J. 2024a. Widthformer: Toward efficient transformer-based bev view transformation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8457–8464.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024b. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911.

Yang, L.; Tang, T.; Li, J.; Yuan, K.; Wu, K.; Chen, P.; Wang, L.; Huang, Y.; Li, L.; Zhang, X.; et al. 2025. Bevheight++: Toward robust visual centric 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J.; Zhang, Y.; Liu, Q.; and Wang, Y. 2023. SA-BEV: Generating Semantic-Aware Bird's-Eye-View Feature for Multi-view 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3348–3357.

Zhang, J.; Zhang, Y.; Qi, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2025a. Geobev: Learning geometric bev representation for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9960–9968.

Zhang, Z.; Gao, B.; Ye, J.; Jin, H.; Jiang, L.; and Yang, W. 2025b. CLIP prior-guided 3D open-vocabulary occupancy prediction. *Pattern Recognition*, 162: 111347.

Zhao, H.; Zhang, Q.; Zhao, S.; Chen, Z.; Zhang, J.; and Tao, D. 2024. Simdistill: Simulated multi-modal distillation for bev 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 7460–7468.

Zhou, Z.; Lei, Y.; Zhang, B.; Liu, L.; and Liu, Y. 2023. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11175–11185.

Zhu, X.; Zhang, R.; He, B.; Guo, Z.; Zeng, Z.; Qin, Z.; Zhang, S.; and Gao, P. 2023. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2639–2650.

Zong, Z.; Jiang, D.; Song, G.; Xue, Z.; Su, J.; Li, H.; and Liu, Y. 2023. Temporal enhanced training of multi-view 3d object detector via historical object prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3781–3790.