arXiv:2604.18476v1 [cs.CV] 20 Apr 2026

# SemLT3D: Semantic-Guided Expert Distillation for Camera-only Long-Tailed 3D Object Detection

Hao Vo[1], Khoa Vo[1], Thinh Phan[1], Ngo Xuan Cuong[1],
Gianfranco Doretto[2], Hien Nguyen[3], Anh Nguyen[4], Ngan Le[1]
[1]AICV Lab, University of Arkansas, USA [2]University of Utah, USA
[3]University of Houston, USA [4]University of Liverpool, UK
{haov,khoavoho,thinhp,cngo,thile}@uark.edu, u6069230@umail.utah.edu
hvnguy35@central.uh.edu, anh.nguyen@liverpool.ac.uk

## Abstract

*Camera-only 3D object detection has emerged as a cost-effective and scalable alternative to LiDAR for autonomous driving, yet existing methods primarily prioritize overall performance while overlooking the severe long-tail imbalance inherent in real-world datasets. In practice, many rare but safety-critical categories such as children, strollers, or emergency vehicles are heavily underrepresented, leading to biased learning and degraded performance. This challenge is further exacerbated by pronounced inter-class ambiguity (e.g., visually similar subclasses) and substantial intra-class diversity (e.g., objects varying widely in appearance, scale, pose, or context), which together hinder reliable long-tail recognition. In this work, we introduce* ***SemLT3D****, a Semantic-Guided Expert Distillation framework designed to enrich the representation space for underrepresented classes through semantic priors. SemLT3D consists of: (1) a* ***language-guided mixture-of-experts*** *module that routes 3D queries to specialized experts according to their semantic affinity, enabling the model to better disentangle confusing classes and specialize on tail distributions; and (2) a* ***semantic projection distillation*** *pipeline that aligns 3D queries with CLIP-informed 2D semantics, producing more coherent and discriminative features across diverse visual manifestations. Although motivated by long-tail imbalance, the semantically structured learning in SemLT3D also improves robustness under broader appearance variations and challenging corner cases, offering a principled step toward more reliable camera-only 3D perception.*

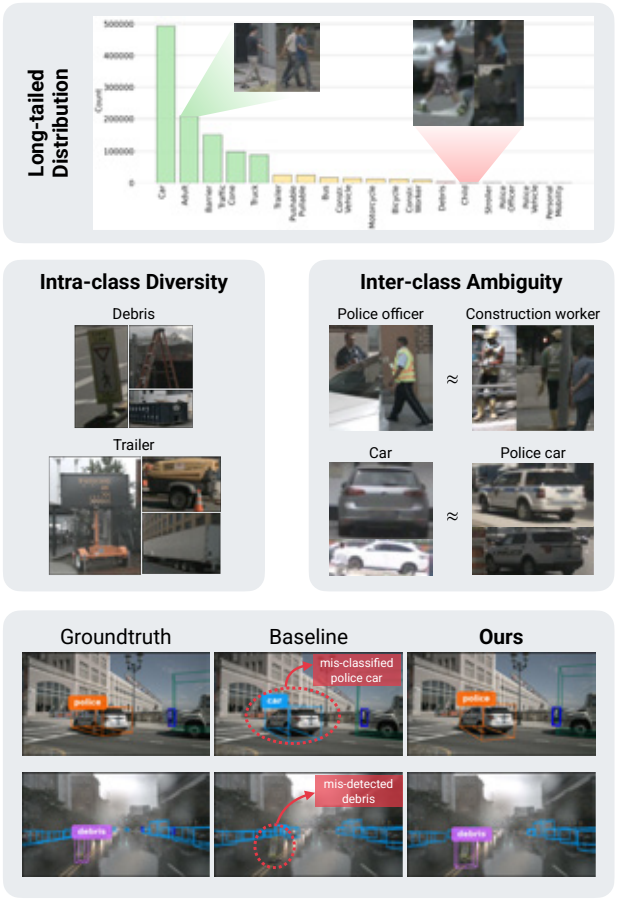


Figure 1. Visualization of long-tailed distribution in nuScenes [3] (top), illustration of inter-intra categories diversity challenges in tail-classes (middle), and representative failure cases of baseline model caused by such inter class ambiguities (bottom-$1^{st}$) and intra-class diversity (bottom - $2^{nd}$).

## 1. Introduction

In recent years, camera-only 3D object detection for autonomous driving (AV) [14–16, 22, 29, 31, 39, 46, 47, 52] has become one of the most active and transformative research topics in perception. Compared with LiDAR-

based counterparts [5, 18, 33], camera-only perception systems offer substantially lower hardware cost and easier scalability, enabling mass deployment at fleet level. A broad spectrum of methods have merged, ranging from depth-estimation–based approaches [15, 21] to depth-free paradigms [23, 31, 47], from single-frame detectors [47] to multi-frame temporal models [29, 46], and from sparse-query [26, 29] to dense-query [22, 23] frameworks.

Despite these advances, existing camera-only 3D object detectors remain biased toward just frequently seen categories. Object occurrences in driving datasets follow a strongly long-tailed (Zipfian) distribution [63], where a few dominant *"head"* classes such as `car` and `adult` appear abundantly, while many *"tail"* classes (e.g., `child`, `emergency vehicle`, `stroller`, `debris`, etc.) are extremely underrepresented [59]. Yet these rare classes are often those with the highest safety stakes: a missed `child` or `emergency vehicle` detection can have catastrophic consequences [44, 50]. In widely adopted benchmarks such as nuScenes [3], evaluation protocols further exacerbate this imbalance by aggregating distinct subcategories–`child`, `police officer`, `construction worker`, `stroller`–into a coarse group of pedestrian, obscuring crucial behavioral and safety nuances. As a result, current models achieve impressive overall performance yet remain unreliable in the rare but safety-critical cases that define real-world robustness.

Beyond data scarcity and imbalance, two inherent challenges further exacerbate the issue in camera-only 3D detection (Fig. 1 middle). First, intra-class diversity: tail categories often cover visually heterogeneous instances. For example, `debris` can refer to trash containers, ladders, or fallen cargo. Second, inter-class ambiguity: visual overlaps between semantically related classes cause frequent confusion. The `police officer` may wear reflective vests that are indistinguishable from the `construction worker`, and `the police car` at certain viewpoints may resemble a typical `car`.

To address these challenges, we introduce SemLT3D, a semantic-guided expert distillation for camera-only long-tailed 3D object detection. Following recent query-based paradigms for 3D detection [16, 46], SemLT3D initializes a set of 3D object queries that aggregate visual information from multiple camera views to form 3D object tokens, from which a detection head decodes into 3D object locations and categories. Building on this foundation, SemLT3D directly targets two major difficulties identified above through two complementary components. **(1) Language-guided mixture-of-experts** addresses intra-class diversity by introducing semantically guided expert specialization. The router leverages language embeddings of object names to guide the assignment of 3D object queries to experts according to their visual-language semantic similarity. This language-guided routing allows semantically related categories to share experts, while distinct categories are handled by different ones, enabling each expert to focus on a coherent and meaningful subset of visual patterns. **(2) Semantic projection distillation** mitigates data scarcity and inter-class ambiguity by transferring semantic knowledge from a powerful pretrained vision-language model (VLM) into our 3D object features. By aligning 3D object tokens with corresponding semantically grounded 2D image embeddings extracted from the pretrained VLM, we inject high-level semantic priors into 3D object tokens. This semantic transfer enriches the visual representations of underrepresented categories and improves discrimination among visually similar classes, leading to more robust long-tailed 3D detection.

To demonstrate the effectiveness and generality of our proposed method, we follow prior work [40] and extend the evaluation from the standard 10 classes to all 18 classes on both the nuScenes [3] and Argoverse 2 (AV2) [49] datasets, providing a more comprehensive assessment of overall performance. Without any bells and whistles, our approach consistently improves upon the corresponding baselines by 2.62%/ 0.96% mAP and 2.75%/ 1.62% NDS on nuScenes with ResNet-50 and ResNet-101 backbones, respectively, and raises 1.5% mAP and 1.3% NDS on the AV2 validation split. In summary, our main contributions are as follows:

- We propose **SemLT3D**, a semantic-guided expert distillation framework, that tackles long-tail imbalance, along with inter-class ambiguity, and intra-class diversity in camera-only 3D detection.
- SemLT3D couples a **language-guided mixture-of-experts** which semantically routes queries for expert specialization with a **semantic projection distillation** that injects CLIP-informed 2D priors into 3D tokens for stronger class discrimination.
- **Extensive evaluation** under the full 18-class long-tailed setting shows consistent gains across baselines, establishing SemLT3D as an effective and scalable solution for robust camera-only 3D perception.

## 2. Related Work

**Camera-based 3D Object Detection** Recently, camera-based 3D object detection has attracted much attention and achieved great progress, due to its advantages of low deployment cost and rich semantic information. The related methods can be roughly categorized into two branches: BEV-based representation [15, 20, 21, 41] and query-based representation [23, 29, 32, 46, 55]. While these methods have significantly advanced overall performance, handling corner cases may be the last mile of perception systems. Recent studies have thus focused on addressing specific challenges within camera-based 3D detection. For instance, RoboBEV [51] adding more challenging attributes to the nuScenes benchmarks. Far3D [16] and LR3D [56] address

the long-range problem by extend the detection range and using support from 2D detection or depth estimation. Motivated by human behavior, CorrBEV [52] resolve the occlusion by propose to use the correlation between text and image. *In contrast, our work turns attention to a largely underexplored yet critical issue, the long-tail distribution in camera-based 3D object detection.*

**3D Long-tail Object Detection** Resolving the long-tail problem remains an inevitable and critical challenge in autonomous driving, particularly for 3D object detection. A large body of prior work has focused on addressing this challenge through LiDAR-based pipelines. For instance, CBGS [62] mitigates data imbalance by up-sampling LiDAR sweeps of rare categories and pasting instances of underrepresented objects from other scenes, while LT3D [40] introduces a hierarchical loss that leverages superclass–subclass relations and camera–LiDAR fusion to improve recall on rare categories. More recently, FOMO-3D [53] integrates OWLv2 [37] for 2D open-vocabulary detection and Metric3Dv2 [12] for depth estimation to enhance long-tail recognition. Although these approaches demonstrate notable progress, their reliance on LiDAR or multi-modal (LiDAR–camera) and multi-stage (e.g., LiDAR–VLM or LiDAR–2D detector) inference pipelines introduces additional computational overhead, synchronization complexity, hardware cost, and latency, making large-scale deployment less feasible. Surprisingly, despite these limitations, the long-tail issue in camera-only 3D detection has received little attention. *Motivated by this observation, we propose **SemLT3D**, a simple yet effective plug-and-play framework that tackles the long-tail problem within a unified camera-only 3D detection paradigm, maintaining a favorable balance between scalability, efficiency, and practical deployment.*

**Connection to 2D Long-tail learning** To better understand imbalance in camera-based 3D detection, it is helpful to draw parallels with the extensive progress made in 2D long-tail learning [59]. In 2D classification, resampling [9, 17] and re-weighting [6, 25, 43] are widely used, yet they remain constrained by the inherently limited samples of tail classes and therefore cannot fully address representation deficiency. For 2D detection, prior work generally falls into two directions. First, multi-stage training pipelines inspired by few-shot detection [8, 13, 36, 45] attempt to transfer head-class knowledge to tail classes; however, such approaches require training multiple models, multiple times and become prohibitively expensive for large-scale 3D benchmarks [3, 49]. Second, other methods leverage additional data or image-level supervision [36, 61], often using CLIP to enrich tail semantics, but these approaches face a significant domain gap when applied directly to 3D detection. *Motivated by these insights, we introduce a language-guided MoE that routes queries by semantic similarity, allowing experts to share strong head-class features while dedicating capacity to tail categories, together with a semantic projection distillation mechanism that enables effective distillation from powerful 2D vision–language models. Taken together, these designs bring the benefits of mature 2D long-tail strategies into a unified and scalable camera-only 3D detection framework.*

# 3. Method

We present SemLT3D (Figure 2) for camera-based 3D object detection. In section 3.1, we propose semantic-guided MoE by using language embeddings. Next, we delineate the process of semantic projection distillation in Section 3.2 and process of visual-language alignment in Section 3.3.

## 3.1. Language-guided Mixture of Experts

As aforementioned, learning from highly long-tailed diversity data with a unified model is inherently challenging, as optimization is dominated by head classes while tail classes receive insufficient supervision. This issue is further exacerbated by substantial intra-class diversity in rare categories - e.g., trailer, debris exhibit wide variations in pose, appearance, and context despite having very few training examples. To alleviate this limitation without increasing overall model complexity, we draw inspiration from recent advances in mixture-of-experts architectures [7, 24] and replace the standard Feed-Forward Network (FFN) layers in transformer blocks with our proposed Language-guided Mixture-of-Experts (LMoE).

Unlike a conventional FFN that applies a single transformation to all object queries, LMoE decomposes the representation space into multiple lightweight experts, each specializing in a subset of semantically related categories. This category-level specialization reduces interference between heterogeneous classes and allows the model to better capture the unique appearance variations within tail categories. Furthermore, instead of treating every category uniformly, LMoE intentionally routes semantically related categories into the same expert based on their shared geometric and contextual characteristics. For example, human-like categories such as `adult`, `police officer` or `debris` tend to share similar height ranges, body proportions, and spatial placement in driving scenes, while vehicle-related categories exhibit their own consistent structural patterns. By assigning these semantically aligned categories to the same expert, LMoE forms meaningful sub-clusters that reduce intra-expert heterogeneity, making it easier to handle large intra-class diversity.

As illustrated in Figure 2, LMoE consists of a router, $M$ specialized experts, and one shared expert, all implemented as lightweight linear layers. For each query token, the router predicts routing weights and selects the top-$k$ experts to process it, after which their outputs are aggregated

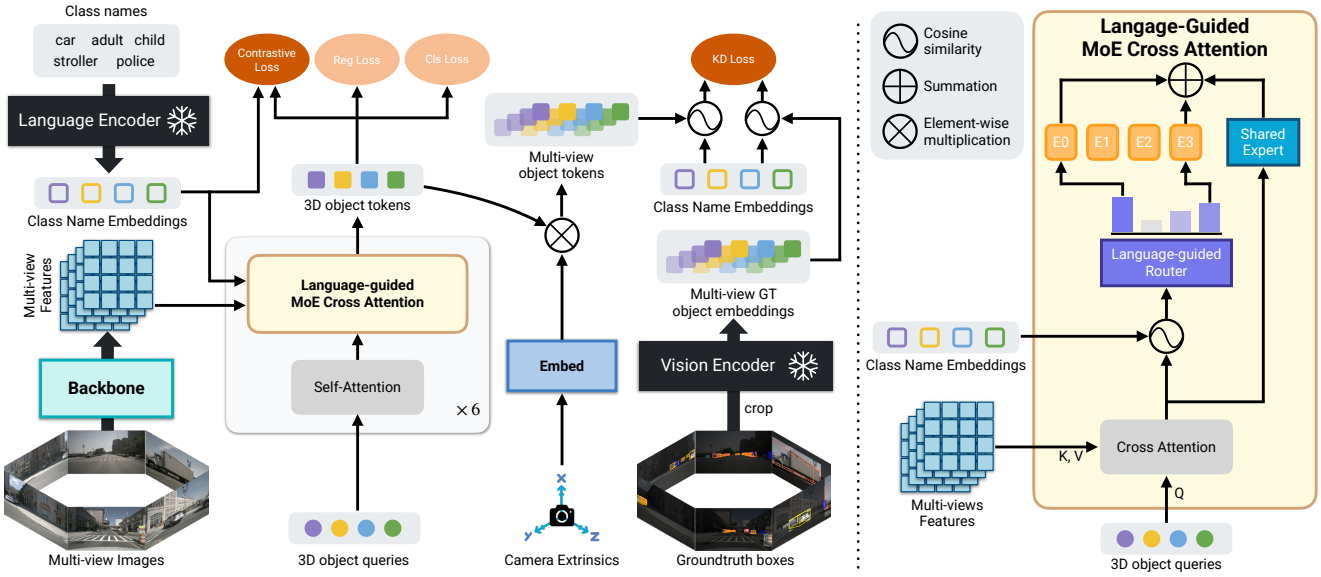


Figure 2. Overall architecture of the proposed SemLT3D for multi-view 3D long-tailed object detection.

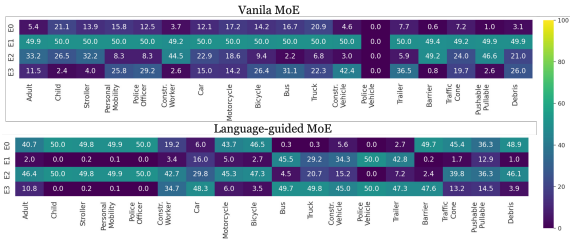


Figure 3. Distribution of queries assigned to activated experts for standard MoE and our LMoE. The x-axis shows category names and the y-axis indicates expert IDs.

according to the routing weights. In parallel, each token is also passed through the shared expert to retain generalizable knowledge across all categories.

Recent DETR-style detectors, which directly adopt MoE layers designed for large-language-model architectures [34, 58], often promote uniform routing [38], limiting their ability to form meaningful semantic specialization. To address this shortcoming, we redesign the routing mechanism by replacing raw query features with language-aware similarity vectors. Instead of feeding high-dimensional query embeddings into the router, we compute their similarity to class name embeddings extracted by CLIP's text encoder, providing an explicit semantic prior that naturally groups queries associated with related meanings. As shown in Figure 3, this design produces clear semantic partitions, yielding expert assignments that align with intuitive structure—such as grouping human-like classes together or routing all vehicle-related categories into another expert. Concretely, after the cross-attention block, the 3D queries $Q \in \mathbb{R}^{k\times D}$ (where $k$ is the number of queries and $D$ is the query dimension) are projected into the language embedding space of dimension $d$, producing semantically aligned queries $\hat{Q} \in \mathbb{R}^{k\times d}$.

$$\hat{Q} = \text{Linear}(Q) \tag{1}$$

then, we extract language-driven semantic embeddings by converting the category names into prompts [60] and feeding into the CLIP language encoder, as result we obtain $P_{\text{language}} \in \mathbb{R}^{n\times d}$, where $n$ denotes the number of classes:

$$P^{\text{language}} = \text{CLIP}_{\text{language}}(\textit{categories}) \tag{2}$$

Next, we compute the similarity between $\hat{Q}$ and $P^{\text{language}}$ to obtain $S^l \in \mathbb{R}^{k\times n}$, which serves as the input to the router:

$$S^l = \text{sim}(\hat{Q}, P^{\text{language}}) \tag{3}$$

$$R = \textit{Router}(S^l) \tag{4}$$

where $\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$. Based on the router output $R$, we apply a softmax function to obtain routing weights:

$$W = \text{Softmax}(R) \tag{5}$$

These routing weights determine the contribution of each expert for every query. Specifically, each query is assigned to its top-$k$ experts according to $W$, and their outputs are aggregated to form the routed representation $y^e \in \mathbb{R}^{k\times D}$. In parallel, a shared expert $E^s$ processes all queries to retain generalizable knowledge. The final refined query representation $\bar{Q}$ is then computed as:

$$y^e = \sum_{i\in\mathcal{T}} W_i E_i^R(Q), \qquad \bar{Q} = y^e + E^s(Q), \tag{6}$$

where $\mathcal{T}$ denotes the set of top-$k$ selected expert indices and $E_i^R$ represents the $i$-th routed expert. Following prior

work [1, 7], we further apply an auxiliary balance loss to encourage diverse expert utilization:

$$\mathcal{L}_{\text{balance}} = M \cdot \sum_{i=1}^{M} \mathcal{F}_i \cdot \mathcal{P}_i, \tag{7}$$

where $\mathcal{F}_i$ is the fraction of queries assigned to expert $E_i$, and $\mathcal{P}_i$ is the average routing probability for that expert.

### 3.2. Semantic Projection Distillation

Next, to mitigate inter-class ambiguity, we integrate CLIP [42] to distill large-scale vision–language priors into our 3D detector. Ambiguous categories, such as `police officer` versus `construction worker`, often rely on subtle contextual cues (e.g., standing on a street corner or near a construction site), which CLIP captures effectively. Motivated by this observation, our distillation module is designed with 3 objectives: (1) directly strengthen the feature learning of 3D object tokens from CLIP, (2) inject semantic priors that benefit rare or visually similar categories, and (3) promote class-wise disentanglement through language–visual alignment. To this end, we transform the refined 3D object tokens $\bar{Q}$ into 2D camera-aligned representations by embedding each camera's extrinsic parameters, producing $Q_c \in \mathbb{R}^{C\times k\times d}$, where $C$ is the number of cameras. This alignment enables direct supervision of 3D object tokens using 2D CLIP features at multiple views, consistent with our first objective.

$$Q_c = \text{Linear}(\bar{Q}) \odot \text{Linear}(E) \tag{8}$$

where $\odot$ denotes the Hadamard product, $E \in R^{C\times 16}$ denotes cameras's extrinsics. Next, after performing standard Hungarian matching [31, 47] between ground-truth objects and queries, we project the 3D boxes of the matched ground-truth objects onto the 2D image planes to obtain the ground-truth 3D–2D correspondences for each query. Using these correspondences, we crop the perspective regions of the matched objects in each camera view and feed them through the CLIP image encoder to extract rich visual features that capture not only class semantics but also contextual cues from the surrounding regions (our second objective).

$$P_g^{\text{visual}} = \text{CLIP}_{\text{visual}}(\tilde{I}_g) \tag{9}$$

where $\tilde{I}_g$ denotes the cropped image region g-th in the set of matched 2D ground-truth G. To align semantic across modalities, we compute cosine similarities between the camera-aligned queries $Q_c$ and the language embeddings, as well as between the extracted visual features and the language embeddings. These similarity distributions form the student–teacher pair used in our distillation process:

$$S_g^s = \text{sim}(Q_g^c,\, P^{\text{language}}) \tag{10}$$

$$S_g^t = \text{sim}(P_g^{\text{visual}},\, P^{\text{language}}) \tag{11}$$

We supervise the model using a KL divergence loss, encouraging the student queries to match the teacher's language–visual alignment (our third objective):

$$\mathcal{L}_{\text{KD}} = \frac{1}{G}\sum_{g=1}^{G} \mathcal{L}_{\text{KL}}\big(S_g^s,\, S_g^t\big) \tag{12}$$

### 3.3. Query-Language Alignment

In addition to the two modules described above, to stablize the training, we further encourage alignment between the queries and language embeddings by introducing a contrastive loss. Following prior works [19, 24, 30], we compute the dot-product similarity between each query and the language embeddings to obtain classification logits, and supervise these logits using a focal loss [25]:

$$\mathcal{L}_{\text{contrast}} = \mathcal{L}_{\text{Focal}}\Big(\text{sim}(\hat{Q},\, P_{\text{language}}),\, T\Big) \tag{13}$$

where $T \in \mathbb{R}^{k\times n}$ denotes is the target matching result between queries and classes computed from the Hungarian matching. Consistent with DETR-style architectures, we apply this contrastive loss as an auxiliary loss at each decoder layer to facilitate stable and progressive alignment during optimization.

## 4. Experiment

### 4.1. Datasets and Metrics

We evaluate our approach on two large-scale autonomous driving benchmarks: nuScenes [3] and Argoverse 2 [48].

**nuScenes** consists of 1000 driving scenes (approximately 20 seconds each), with 3D bounding box annotations on sampled keyframes. Following prior work [33, 54], we adopt the extended 18-class taxonomy and additionally report mAP for three frequency groups: Many, Medium, and Few, to better characterize long-tail performance. For standard evaluation, we follow the official nuScenes metrics, reporting mean Average Precision (mAP) and the nuScenes Detection Score (NDS) across 18 categories.

**Argoverse 2** provides 1000 scenes of 15 seconds duration at 10 Hz, split into 700/150/150 for train/val/test. It offers seven high-resolution ring cameras with full 360° coverage and includes 26 object categories within a 150 m perception range. We follow the official protocol and report mAP, the Composite Detection Score (CDS), and true-positive metrics ATE, ASE, and AOE.

### 4.2. Implementation Details

On nuScenes benchmark [3], we implement our model base on StreamPETR [46], trained with 60 epochs using Resnet50 and Resnet101 [11], for other configs, we follow [46]. On AV2 benchmark [49], we implemented our model base on Far3D [16] with the same settings. We use the

Table 1. Performance comparison on validation split of nuScenes [3] with 18 categories. The best score is in **bold**.

| | Methods | mAP↑ | NDS↑ | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ | FPS↑ |
|---|---|---|---|---|---|---|---|---|---|
| Resnet50 | BEVDet [15] | 13.27 | 19.46 | 0.892 | 0.528 | 0.814 | 0.819 | 0.665 | 13.30 |
| | BEVDet4D [14] | 17.98 | 26.66 | 0.8517 | 0.4557 | 0.8597 | 0.4563 | 0.6095 | 13.18 |
| | BEVFormer [23] | 15.51 | 23.25 | 0.9333 | **0.3200** | 1.00 | 0.6818 | 0.5151 | 18.18 |
| | BEVDepth [21] | 15.43 | 22.01 | 0.7939 | 0.4638 | 0.9624 | 0.7355 | 0.6145 | 10.30 |
| | PETRv2 [32] | 18.13 | 27.63 | 0.8806 | 0.3235 | 0.9013 | 0.5425 | 0.4951 | 14.92 |
| | RayDN [28] | 27.41 | 37.39 | 0.6856 | 0.3516 | 0.7612 | 0.3574 | **0.4762** | 31.25 |
| | SparseBEV [29] | 27.94 | 38.74 | **0.6459** | 0.3496 | 0.7190 | 0.2775 | 0.5312 | 20.8 |
| | StreamPETR [46] | 26.97 | 38.19 | 0.6553 | 0.3597 | 0.6839 | 0.2960 | 0.5351 | **32.36** |
| | Ours | **29.59** | **40.94** | 0.6938 | 0.3335 | **0.6153** | **0.2524** | 0.4909 | 29.40 |
| Resnet101 | BEVFormer [23] | 24.65 | 37.05 | 0.7731 | 0.3228 | 0.6051 | 0.3458 | 0.4807 | 5.01 |
| | SparseBEV [29] | 29.78 | 34.87 | **0.6129** | 0.6082 | 1.6232 | 0.2796 | 0.5016 | 9.21 |
| | StreamPETR [46] | 30.16 | 41.17 | 0.6863 | 0.3209 | 0.6339 | **0.2607** | 0.4892 | **11.76** |
| | Ours | **31.12** | **42.79** | 0.6479 | **0.3019** | **0.5928** | 0.2665 | **0.4677** | 10.80 |

Table 2. Comparisons on the Argoverse 2 [48] validation split. We evaluate 26 object categories with a range of 150 meters. ‡ shows the results take from Far3D [16].

| Methods | Backbone | mAP↑ | CDS↑ | mATE↓ | mASE↓ | mAOE↓ |
|---|---|---|---|---|---|---|
| BEVStereo [20] ‡ | VoV-99 | 14.6 | 10.4 | 0.847 | 0.397 | 0.901 |
| SOLOFusion [39] ‡ | VoV-99 | 14.9 | 10.6 | 0.934 | 0.425 | 0.779 |
| PETR [31]‡ | VoV-99 | 17.6 | 12.2 | 0.911 | 0.339 | 0.819 |
| Sparse4Dv2 [27]‡ | VoV-99 | 18.9 | 13.4 | 0.832 | 0.343 | 0.723 |
| StreamPETR [46] ‡ | VoV-99 | 20.3 | 14.6 | 0.843 | 0.321 | 0.650 |
| Far3D [16]‡ | VoV-99 | 24.4 | 18.1 | 0.796 | **0.304** | 0.538 |
| Ours | VoV-99 | **25.9** | **19.4** | **0.792** | 0.308 | **0.480** |

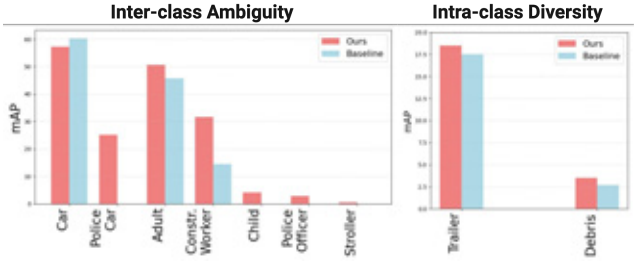


Figure 4. Performance comparison within tail categories has inter-intra diversity challenges.

CLIP-B/16 for knowledge distillation. To maintain complexity, we set the hidden dimensions of shared expert as 1024 and each expert is 512 so with top k = 2 will we able to maintain the origin complexity of the original FFN which is set to 2048. Weights of contrastive loss, knowledge distillation loss and MoE balance loss is set to 1.0, 0.5 and 0.01, respectively. For fair comparison, all the training experiments are performed with 8 × NVIDIA A100(40GB).

### 4.3. State-of-the-art Comparison

**Evaluation of Overall Performance.** To assess the effectiveness of SemLT3D on the nuScenes [3], we re-evaluate eight camera-only baselines under the extended 18-class long-tailed setting, with results summarized in Table 1. Compared with the StreamPETR baseline [46], SemLT3D

Table 3. Long-tail break down mAP evaluation on validation split on nuScenes [3]. L denotes LiDAR. C denotes Camera. † shows the result take from FOMO-3D [54]. * shows method using ViT backbone [10]. Methods shown in grey utilize multi-sensor inputs.

| Method | Modality | All | Many | Medium | Few |
|---|---|---|---|---|---|
| CenterPoint [40, 57] † | L | 39.2 | 76.4 | 43.1 | 3.5 |
| CenterPoint [40, 57]†(w/ hier.) | L | 40.4 | 77.1 | 45.1 | 4.3 |
| BEVFusion-L [33]† | L | 42.5 | 72.5 | 48.0 | 10.6 |
| TransFusion [2]† | L+C | 39.8 | 73.9 | 41.2 | 9.8 |
| BEVFusion [33]† | L+C | 45.5 | 75.5 | 52.0 | 12.8 |
| CMT [33]† | L+C | 44.4 | 79.9 | 53.0 | 4.8 |
| MMF [40]† | L+C | 43.6 | 77.1 | 49.0 | 9.4 |
| MMLF [35]† | L+C | 51.4 | 77.9 | 59.4 | 20.0 |
| FOMO-3D [54]† | L+C | 54.6 | 79.9 | 59.6 | 27.6 |
| BEVDet [15] | C | 13.27 | 27.7 | 14.31 | 0.0 |
| BEVFormer [23] | C | 15.51 | 34.6 | 14.91 | 0.3 |
| BEVDepth [21] | C | 15.43 | 36.46 | 13.51 | 0.13 |
| BEVDet4D [14] | C | 17.98 | 39.34 | 18.07 | 0.0 |
| PETRv2 [32] | C | 18.13 | 38.82 | 14.96 | 0.3 |
| RayDN [28] | C | 27.41 | 54.50 | 31.02 | 0.6 |
| SparseBEV [29] | C | 27.94 | 54.18 | 29.67 | 4.1 |
| StreamPETR [46] | C | 26.97 | 53.32 | 28.53 | 3.22 |
| Ours | C | 29.59 | 50.92 | 34.55 | 6.03 |
| **Ours*** | C | **41.1** | **62.08** | **40.62** | **20.77** |

achieves substantial improvements of 2.62% and 0.96% mAP, as well as 2.75% and 1.62% NDS, on the ResNet50 and ResNet101 backbones, respectively. Importantly, our method introduces only a modest increase in inference time of 9.1% for ResNet50 and 8.16% for ResNet101, highlighting its efficiency. Similar trends are observed on the AV2 benchmark [48], where SemLT3D improves Far3D [16] by 1.5% mAP and 1.30% CDS, as shown in Table 2. These results demonstrate the strong effectiveness of our framework, along with its generalization ability and plug-and-play compatibility across datasets and model architectures, underscoring its practical deployment potential.

**Evaluation of Long-tailed Performance.** Beyond improving overall detection accuracy, our framework substantially enhances performance on classes with limited training samples. As shown in Table 3, the proposed method yields no-

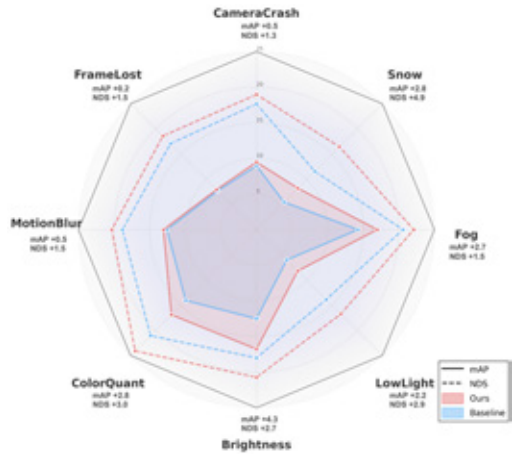


Figure 5. Performance comparison under different corner cases.

table gains of +6.02 mAP on Medium categories and +2.81 mAP on Few categories, with only a marginal decrease on Many. This aligns with our objective: by incorporating semantic-guided MoE and semantic projection distillation, the model leverages richer representations and strong semantic priors to better generalize to tail classes despite their scarcity. To further compare with multimodal approaches, we evaluate our framework using a ViT backbone [10]. The results demonstrate competitive performance relative to CenterPoint (+0.7 mAP) and BEVFusion (-1.3 mAP). Remarkably, even without relying on high-cost LiDAR sensors, our camera-only method achieves a +10.17 mAP improvement over BEVFusion on Few categories, illustrating the scalability and practical value of our approach for long-tailed detection.

**Evaluation on Inter–Intra Diverse Categories.** Figure 4 compares our SemLT3D with the baseline StreamPETR [46]. We consistently see better performance in both inter-class ambiguity cases and intra-class diversity cases, showing that our method is aligned with its objective, help the model distinguish confusing classes and handle the wide variations within each category. There is a small drop in the `car` class, likely due to overall optimization trade-offs, but it does not affect the general trend. Overall, the results highlight that SemLT3D brings clear and meaningful gains, especially for tail classes where diversity and ambiguity pose the biggest challenges.

**Evaluation on Other Corner Cases.** As discussed above, our motivation arises from the inherent diversity and ambiguity among categories. To further assess the robustness of our approach, we evaluate performance under challenging scenarios [51, 52] where object appearances exhibit higher variability, such as `Snow` or `ColorQuant`. As illustrated in Figure 5, integrating our model into the baseline [46] consistently improves detection performance across all scenarios, particularly where visual degradation or domain shifts weaken object representations. These results demonstrate that our model enhances the baseline's generalization ability, enabling more reliable recognition under diverse and adverse conditions.

Table 4. Effectiveness of each module on nuScenes val split.

| Mixture of Expert | Semantic-Guided Router | Semantic Projection Distillation | Query-Language Alignment | mAP | NDS | Many | Medium | Few |
|---|---|---|---|---|---|---|---|---|
| – | – | – | – | 26.97 | 38.19 | 53.32 | 28.53 | 3.22 |
| – | – | ✓ | – | 28.30 | 39.13 | 49.6 | 32.1 | 5.82 |
| ✓ | – | – | – | 26.37 | 37.47 | 51.48 | 30.73 | 0.32 |
| ✓ | ✓ | – | – | 27.50 | 37.59 | 51.7 | 32.16 | 2.25 |
| ✓ | ✓ | ✓ | – | 28.56 | 40.26 | 49.48 | 32.2 | 6.9 |
| ✓ | ✓ | ✓ | ✓ | **29.59** | **40.94** | **50.92** | **34.56** | **6.03** |

Table 5. Configurations of LMoE ablation on nuScenes val split.

| Experts | Top-*k* | mAP | NDS | Many | Medium | Few |
|---|---|---|---|---|---|---|
| 4 | 1 | 28.56 | 40.26 | 49.50 | 33.20 | 5.90 |
| 4 | 2 | **29.59** | **40.94** | **50.92** | **34.55** | 6.03 |
| 8 | 1 | 28.24 | 40.4 | 47.84 | 31.76 | 8.6 |
| 8 | 2 | 28.54 | 39.24 | 47.96 | 32.06 | **8.21** |

Table 6. Ablation of prior semantic provided by different CLIP models on nuScenes validation split.

| CLIP models | mAP | NDS | Many | Medium | Few |
|---|---|---|---|---|---|
| ViT-B/32 | 28.29 | 39.13 | 50.5 | 32.30 | 5.82 |
| ViT-B/16 | 29.59 | 40.94 | 50.92 | 34.55 | 6.03 |
| ViT-L/14 | 29.63 | 40.67 | 49.50 | 33.90 | 8.06 |

## 4.4. Ablation Study

**Component-wise Analysis** Table 4 summarizes the effectiveness of each module and their impact on long-tail performance. Replacing the FFN with a vanilla MoE slightly decreases overall mAP (–0.6) and severely harms tail categories due to unguided expert routing. Introducing our semantic-guided router reverses this effect, improving both overall accuracy and rare-class performance by stabilizing expert assignment under long-tail distributions. CLIP-based semantic projection distillation further boosts performance from 26.97 to 28.30 mAP, particularly benefiting medium and tail classes that depend on strong contextual semantics. When combined with semantic-guided MoE, the synergy improves mAP to 28.56. Adding the query–language contrastive loss yields the best results, 29.59 mAP and 40.94 NDS, demonstrating that semantic routing, CLIP priors, and language-aligned supervision together produce more discriminative and balanced representations across Many/Medium/Few splits.

**Semantic-guided MoE Configuration Analysis.** Table 5 evaluates different configurations of our LMoE module. Increasing the number of experts consistently improves performance on tail categories, suggesting that a larger expert pool captures richer and more fine-grained semantics crucial for rare classes. However, excessively expanding the expert set or reducing the top-*k* selection leads to imbalanced routing and under-utilized experts, which negatively affects overall performance, especially on Many and Medium categories. Balancing specialization and stable expert usage, the configuration with 4 experts and top-*k*=2

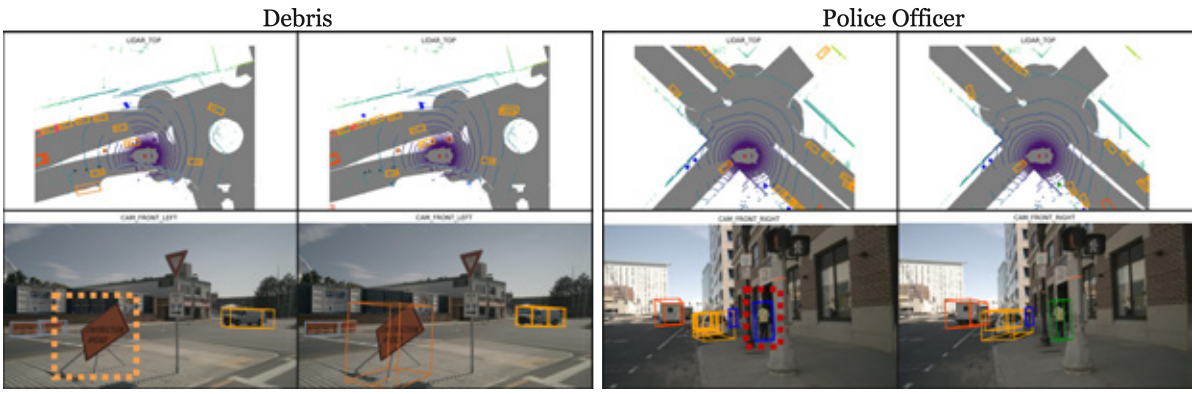


Figure 6. Quanlitative result on inter-intra diversity cases (Debris and Police Officer), compare between our method and baseline [46] on tail-class. Left: our basline, Right: our proposed SemLT3D. The orange dashed box highlight failure cases caused by intra-class diversity, while the red dashed box indicate failure cases arising from inter-class ambiguity.

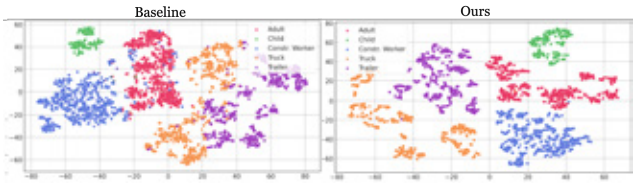


Figure 7. Inter-Intra class diversity visualization in embedding space via t-SNE. Left: our baseline [46], Right: our SemLT3D.

achieves the best trade-off, yielding the strongest results across both head and tail splits.

**Different CLIP Model Analysis.** Table 6 presents an ablation study on different variants of CLIP models. Although using a larger backbone such as ViT-L/14 yields the highest performance, it also significantly increases both training and inference time. To balance efficiency and accuracy, we adopt the ViT-B/16 variant as the teacher model in our final framework, which provides strong performance with moderate computational cost.

### 4.5. Further Analysis

**Qualitative Results** Figure 6 provides a qualitative comparison between the baseline StreamPETR and our proposed SemLT3D. Our focus is on addressing the persistent failures on tail classes and mitigating the challenges posed by both intra-class diversity and inter-class similarity. The figure highlights two representative failure cases where SemLT3D delivers clear improvements. First, SemLT3D successfully localizes the "debris," whereas the baseline either misses it entirely (orange box), illustrating how scarce samples and strong intra-class diversity can lead to confusion. Second, for the inter-class ambiguity distinction between "police officer" and "adult" SemLT3D correctly identifies the police officer while the baseline misclassifies it as an adult (red box), reflecting improved robustness against inter-class ambiguity. Overall, these qualitative results demonstrate the effectiveness of SemLT3D in handling tail classes and improving reliability in safety-critical scenarios for autonomous driving.

**Inter–Intra Class Distribution.** Figure 7 compares the feature distributions of StreamPETR [46] and our SemLT3D, illustrating how our method reduces both inter-class ambiguity and intra-class variability—two challenges that are especially pronounced in tail categories. Using t-SNE [4], we visualize the features of three commonly confused classes (`adult`, `child`, `construction worker`) and observe that SemLT3D forms more compact and clearly separated clusters, indicating that our 2D–3D semantic distillation and query–language contrastive alignment effectively mitigate inter-class ambiguity. In parallel, the LMoE module enhances modeling of intra-class diversity by routing queries to specialized experts, enabling SemLT3D to produce distinct sub-clusters for different "trailer" variants, whereas the baseline remains scattered. These results demonstrate that our semantic alignment and expert specialization strategies substantially strengthen feature discrimination and reduce tail-class confusion.

## 5. Conclusion

In this paper, we introduce SemLT3D, a plug-and-play framework designed to address the long-tail challenge in camera-based multi-view 3D object detection. The central idea is to enhance representation learning for rare and visually ambiguous categories by using semantic priors through language-guided expert routing and 3D–2D knowledge distillation. With these components, SemLT3D delivers consistent improvements across multiple baseline detectors. We also provide a comprehensive evaluation under long-tail distributions and various corner-case scenarios, offering practical insights into the behavior of camera-only 3D perception systems. We believe that our semantic-guided design aligns with real deployment needs in autonomous driving and encourages further attention toward improving long-tail robustness in camera-based 3D detection.

## Acknowledgement

This material is based upon work supported by the National Science Foundation under Award No. #2443877 (NSF CAREER), #1946391 (NSF RII Track-1 DART), #2445877 (NSF E-RISE). We also acknowledge the Arkansas High-Performance Computing Center for providing GPUs.

## References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 5

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022. 6

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2, 3, 5, 6

[4] T. Tony Cai and Rong Ma. Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *Journal of Machine Learning Research*, 23(301):1–54, 2022. 8

[5] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Largekernel3d: Scaling up kernels in 3d sparse cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13488–13498, 2023. 2

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 3

[7] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024. 3, 5

[8] Na Dong, Yongqiang Zhang, Mingli Ding, and Gim Hee Lee. Boosting long-tailed object detection via step-wise learning on smooth-tail data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6940–6949, 2023. 3

[9] Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1):18–36, 2004. 3

[10] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. 6, 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[12] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie 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*, 2024. 3

[13] Xinting Hu, Yi Jiang, Kaihua Tang, Jingyuan Chen, Chunyan Miao, and Hanwang Zhang. Learning to segment the tail. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14045–14054, 2020. 3

[14] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 1, 6

[15] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2, 6

[16] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2561–2569, 2024. 1, 2, 5, 6

[17] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 3

[18] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 2

[19] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 5

[20] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1486–1494, 2023. 2, 6

[21] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1477–1485, 2023. 2, 6

[22] Zhenxin Li, Shiyi Lan, Jose M Alvarez, and Zuxuan Wu. Bevnext: Reviving dense bev frameworks for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20113–20123, 2024. 1, 2

[23] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer:

learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 6

[24] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 3, 5

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 5

[26] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 2

[27] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d v2: Recurrent temporal fusion with sparse model. *arXiv preprint arXiv:2305.14018*, 2023. 6

[28] Feng Liu, Tengteng Huang, Qianjing Zhang, Haotian Yao, Chi Zhang, Fang Wan, Qixiang Ye, and Yanzhao Zhou. Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 200–217. Springer, 2024. 6

[29] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18580–18590, 2023. 1, 2, 6

[30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 5

[31] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European conference on computer vision*, pages 531–548. Springer, 2022. 1, 2, 5, 6

[32] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3262–3272, 2023. 2, 6

[33] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 2, 5, 6

[34] Yehao Lu, Minghe Weng, Zekang Xiao, Rui Jiang, Wei Su, Guangcong Zheng, Ping Lu, and Xi Li. Dynamic-dino: Fine-grained mixture of experts tuning for real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20847–20856, 2025. 4

[35] Yechi Ma, Neehar Peri, Shuoquan Wei, Achal Dave, Wei Hua, Yanan Li, Deva Ramanan, and Shu Kong. Long-tailed 3d detection via multi-modal fusion. *arXiv preprint arXiv:2312.10986*, 2023. 6

[36] Lingchen Meng, Xiyang Dai, Jianwei Yang, Dongdong Chen, Yinpeng Chen, Mengchen Liu, Yi-Ling Chen, Zuxuan Wu, Lu Yuan, and Yu-Gang Jiang. Learning from rich semantics and coarse locations for long-tailed object detection. *Advances in Neural Information Processing Systems*, 36:78082–78094, 2023. 3

[37] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023. 3

[38] Nabil Omi, Siddhartha Sen, and Ali Farhadi. Load balancing mixture of experts with similarity preserving routers. *arXiv preprint arXiv:2506.14038*, 2025. 4

[39] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022. 1, 6

[40] Neehar Peri, Achal Dave, Deva Ramanan, and Shu Kong. Towards long-tailed 3d detection. In *Conference on Robot Learning*, pages 1904–1915. PMLR, 2023. 2, 3, 6

[41] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pages 194–210. Springer, 2020. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5

[43] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 3

[44] Araz Taeihagh and Hazel Si Min Lim. Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport reviews*, 39(1):103–128, 2019. 2

[45] Phi Vu Tran. Simltd: Simple supervised and semi-supervised long-tailed object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4672–4681, 2025. 3

[46] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3621–3631, 2023. 1, 2, 5, 6, 7, 8

[47] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on robot learning*, pages 180–191. PMLR, 2022. 1, 2, 5

[48] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2:

Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021. 5, 6

[49] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 2, 3, 5

[50] Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying unknown instances for autonomous driving. In *Conference on Robot Learning*, pages 384–393. PMLR, 2020. 2

[51] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robobev: Towards robust bird's eye view perception under corruptions. *arXiv preprint arXiv:2304.06719*, 2023. 2, 7

[52] Ziteng Xue, Mingzhe Guo, Heng Fan, Shihui Zhang, and Zhipeng Zhang. Corrbev: Multi-view 3d object detection by correlation learning with multi-modal prototypes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27413–27423, 2025. 1, 3, 7

[53] Anqi Joyce Yang, James Tu, Nikita Dvornik, Enxu Li, and Raquel Urtasun. Fomo-3d: Using vision foundation models for long-tailed 3d object detection. In *9th Annual Conference on Robot Learning*. 3

[54] Anqi Joyce Yang, James Tu, Nikita Dvornik, Enxu Li, and Raquel Urtasun. Fomo-3d: Using vision foundation models for long-tailed 3d object detection. In *9th Annual Conference on Robot Learning*, 2025. 5, 6

[55] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. 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*, pages 17830–17839, 2023. 2

[56] Zetong Yang, Zhiding Yu, Chris Choy, Renhao Wang, Anima Anandkumar, and Jose M Alvarez. Improving distant 3d object detection using 2d box supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14853–14863, 2024. 2

[57] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 6

[58] Chang-Bin Zhang, Yujie Zhong, and Kai Han. Mr. detr: Instructive multi-route training for detection transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9933–9943, 2025. 4

[59] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023. 2, 3

[60] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022. 4

[61] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision*, pages 350–368. Springer, 2022. 3

[62] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 3

[63] George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, 2013. 2