Open-Vocabulary BEV Segmentation with 3D-Aware Geometric Constraints

Hojun Choi[1*], Seulbin Hwang[2], Daejung Kim[2], Kisung Kim[2], Hyunjung Shim[1†], and Jinhan Lee[2†]

[1] KAIST AI, South Korea
{hchoi256,kateshim}@kaist.ac.kr
[2] NAVER LABS, South Korea
{h.sb,daejung.kim,ks.kim,jinhan.lee}@naverlabs.com

Abstract. Bird's-eye view (BEV) perception fuses multi-camera images into a unified top-down representation for autonomous driving. Despite recent progress, state-of-the-art methods remain confined to closed-set scenarios, making them vulnerable to unpredictable real-world environments. In this work, we introduce open-vocabulary BEV segmentation (OVBS), which leverages vision-language models (VLMs) to recognize categories beyond the training set while maintaining precise BEV perception and real-time efficiency. A key challenge in OVBS lies in the 3D geometric inconsistency inherent in the ill-posed lifting of 2D VLM semantics into BEV. To address this, we propose OVBEVSeg, a geometry-aware OVBS framework that enhances efficient Gaussian splatting (GS)-based unprojection by leveraging robust 3D geometric constraints across three progressive stages: (1) 2D-to-BEV pseudo-labeling via reliable 3D projection for OV generalization; (2) joint 2D–BEV per-scene optimization with BEV structural constraints for 3D geometric consistency; and (3) 3D geometric distillation for online efficiency. On the nuScenes dataset, OVBEVSeg achieves state-of-the-art performance, outperforming closed-set methods by 15.3 mIoU on unseen categories. Remarkably, even with no novel-class ground-truth labels, it remains competitive with self- and semi-supervised baselines trained with up to 40% of ground-truth annotations. Furthermore, it achieves 2.5× faster inference with only 0.22× the memory consumption of projection-based methods.

Keywords: Autonomous Driving · Bird's-eye view Perception · Open Vocabulary · 3D Gaussian Splatting

1 Introduction

Bird's-eye view (BEV) perception has emerged as a foundational paradigm in autonomous driving (AD), providing a unified ego-centric top-down representation for effective multi-sensor fusion. It now underpins a broad spectrum of 3D

*Work done during an internship at NAVER LABS.
†Co-corresponding authors

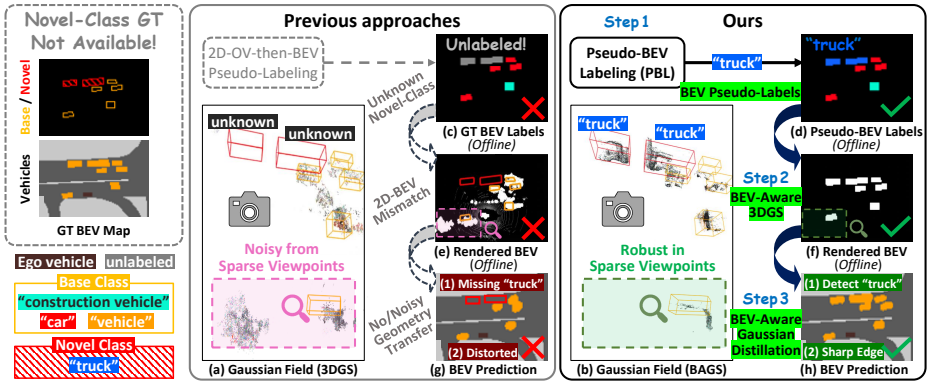

Fig. 1: **(c,d)** Ground-truth or pseudo BEV annotations. **(a,e)** 3D and BEV renderings with plain 3D Gaussian splatting, showing dispersed splats beyond object regions. **(b,f)** 3D and BEV renderings with our BAGS, with splats concentrated along object boundaries. **(g,h)** Online BEV predictions: the baseline misses novel classes (*e.g.*, "truck") and exhibits boundary distortions, which are successfully resolved by our method.

scene understanding tasks, including object detection [25, 29, 49], semantic segmentation [5, 15, 17, 36, 57], and trajectory planning [17–19]. A central challenge in BEV perception is how to establish spatial correspondence between 2D camera features and the BEV representation. This process is fueled by advancements in depth-based [17, 20, 24, 36], projection-based [5, 15], attention-based [25, 57], and Gaussian-based [4, 32] architectures, leading to successful real-time, scalable inference BEV perception critical for modern autonomous platforms.

Despite strict demands for efficient deployment, a crucial yet overlooked assumption persists: existing BEV perception pipelines are fundamentally framed as closed-set systems. That is, both training and inference assume the same, finite set of semantic classes. In real-world AD, this assumption is neither realistic nor safe. In Fig. 1(g) and 3, autonomous vehicles continuously encounter previously unseen objects—*e.g.*, "truck," or atypical movers like "stroller" or "wheelchair." These novel classes, while present in sensor data, lack both human-annotated and ground-truth labels during training and are therefore silently ignored during inference, creating semantically blind regions in the BEV map. This discrepancy is more than a technical shortcoming: it poses a direct threat to downstream tasks such as path planning and collision avoidance [19, 21]. We argue that reframing BEV perception for open-world environments is not optional but essential for ensuring semantic completeness and safety in real-world autonomy.

To this end, we introduce *open-vocabulary BEV segmentation* (OVBS), a new task that extends OV recognition from 2D vision to the BEV domain, leveraging vision-language models (VLMs) to recognize previously unseen categories. The objective of OVBS is to simultaneously achieve three goals: OV generalization to novel classes, precise BEV perception, and inference-time efficiency.

A natural extension to OVBS is to graft OV visual recognition [46, 47] onto the core logic of unprojection-based BEV pipelines. Concretely, one could attach a 2D OV detector [28] to each camera view, lift the predicted 2D regions into 3D space via depth estimation, and project them into BEV—often used as a fair OV baseline. However, this seemingly straightforward "2D-OV-then-BEV" solution remains vulnerable to a longstanding challenge in AD: it relies on direct 2D-to-3D lifting, which is fundamentally brittle in large-scale outdoor scenes with sparse views. Small errors in 2D localization or semantic classification are amplified by the ill-posed lifting process [24,36]—a structural flaw inherent to all unprojection-based methods. In Fig. 1, this direct mapping fails on two fronts: **(L1)** (c) compounded errors hinder the recognition of novel classes in BEV; and **(L2)** (g) inherent geometric instability results in geometrically inconsistent or distorted BEV boxes. In short, naive designs built upon 2D unprojection break down due to the lack of consistent multi-view 3D geometry—a prerequisite for both OV-to-BEV extensions and reliable unprojection-based BEV perception.

This inherent vulnerability persists even in Gaussian splatting (GS)-based unprojection—arguably the most promising candidate due to its superior efficiency and scalability—as it remains susceptible to the same 2D-to-3D lifting errors (L1, L2). Specifically, it amplifies these biases by directly lifting 2D features into 3D Gaussians for BEV splatting. To address these limitations, our key insight is that the direction of geometric reasoning must be reversed: rather than lifting noisy 2D predictions into 3D (unprojection), we should project reliable 3D structure down into 2D and BEV (projection). Following this projective shift, we propose OVBEVSeg, a novel geometry-aware OVBS framework that leverages robust 3D geometric constraints across three progressive stages: (1) generating BEV pseudo-labels by establishing consistent 2D–BEV correspondences and routing 2D VLM semantics via reliable 3D projection of class-agnostic 3D detections [43], thereby bypassing erroneous 2D unprojection and resolving (L1); (2) conducting joint 2D–BEV per-scene optimization constrained by BEV spatial layouts as strong geometric constraints, thereby enabling robust 3D geometric reconstruction from sparse viewpoints; and (3) distilling the recovered geometry into a student model, mitigating (L2) while preserving inference efficiency. Crucially, these stages are not independent modules but successive instantiations of a single principle: leveraging 3D geometric constraints to bypass the ill-posed 2D unprojection—first for semantics, geometry, and efficient knowledge transfer.

Extensive experiments show that OVBEVSeg sets a new state of the art, surpassing prior work [5] by 15.3 mIoU on novel classes while running 2.5× faster with 0.22× memory usage. Without novel-class ground-truth labels, it remains comparable to semi-/self-supervised baselines using up to 40% annotations.

2 Related Work

Bird's-Eye View (BEV) Perception. BEV perception maps multi-sensor data into a unified top-down space for 3D scene understanding. Existing methods fall into four paradigms: (1) depth-based methods [20, 24, 36] explicitly splat

image features with per-pixel depth distributions into a 3D grid, yielding geometrically grounded BEV; (2) projection-based schemes [5,15] query camera features at predefined 3D points; (3) attention-based transformers [25,27,29,44,57] leverage cross-attention to fuse image and BEV tokens; and (4) feed-forward 3D Gaussian splatting methods [4, 32] predict 3D Gaussians to rasterize BEV features. In particular, depth-based approaches are highly sensitive to depth estimation errors, which can propagate into the BEV representation. Despite their architectural diversity, all these paradigms remain strictly constrained to closed-set categories. In this work, we transcend these boundaries by recasting BEV perception as an open-vocabulary task integrated with 3D geometric constraints.

3D Gaussian Splatting (3DGS). 3DGS [22] has emerged as a powerful differentiable renderer for scene representation. In autonomous driving, 3DGS-based methods bifurcate into *offline* per-scene optimization and *online* feed-forward approaches. Within this taxonomy, offline methods [58, 59] incrementally optimize static Gaussian fields while compositing object-level dynamics. While recovering high-fidelity geometry under photometric supervision, they remain computationally prohibitive for real-time perception. Conversely, feed-forward systems [6, 8, 50] enable omnidirectional rendering by computing per-pixel ray directions under equirectangular models. Building on this, recent BEV variants [4, 32] map image features to 3D Gaussians for BEV splatting. Despite their scalability, these methods often compromise 3D geometric precision due to the ill-posed nature of 2D-to-3D unprojection under sparse viewpoints. Drawing on hybrid 3DGS advances [12, 41], we bridge this gap by synergizing the geometric rigor of per-scene optimization with feed-forward efficiency.

Open-Vocabulary Learning (OVL). To overcome the scalability limits of fixed categories, OVL [46] has gained traction, leveraging large-scale vision-language pretraining [38] for unseen category recognition. In autonomous driving, several studies have extended OVL to various 3D tasks: OV-3DETIC [33] addresses novel-class 3D detection via image-level supervision and debiased cross-modal contrastive learning; OpenScene [35] enables OV semantic segmentation (OVSS) by co-embedding 3D features with text and pixels; and OVTrack [23] introduces the first OV multi-object tracking (OV-MOT) framework by distilling vision-language knowledge into a tracker; and LidarCLIP [16] aligns LiDAR features with 2D embeddings for OV recognition in point clouds. In contrast, OV recognition within the BEV space remains underexplored, hindered by the 2D–BEV correspondence ambiguity. To our knowledge, this is the first *open-vocabulary BEV segmentation* (OVBS) framework that leverages robust 3D geometric constraints to map 2D VLM semantics consistently into the BEV space.

3 Methodology

In this work, we introduce *open-vocabulary BEV segmentation* (OVBS), a novel task aimed at enhancing generalization to unseen categories via vision-language

models (VLMs) while preserving precise BEV perception and real-time efficiency. As established in Sec. 1, the core bottleneck for OVBS can be the ill-posed 2D-to-3D lifting shared by all unprojection-based BEV methods. Our 3D projection-first principle addresses this by reversing the direction of geometric reasoning: we project reliable 3D structure down into 2D and BEV.

In Fig. 2, we propose OVBEVSeg, a geometry-aware OVBS framework that enhances Gaussian splatting (GS)-based unprojection by incorporating robust 3D geometric constraints across three synergistic modules: (1) pseudo-BEV labeling (PBL) for OV generalization (Sec. 3.1), (2) BEV-aware 3DGS (BAGS) for geometric reconstruction (Sec. 3.2), and (3) BEV-aware Gaussian distillation (BAGD) for efficiency (Sec. 3.3). In (1), we explicitly establish 2D–BEV correspondence by propagating semantics via the 3D projection of unsupervised 3D clusters [43]. These 3D candidates are projected into the image space, where instance masks are recovered using a segmentation foundation model [39]. For every segmented 2D object, we extract CLIP-based image embeddings and obtain pseudo-class labels via text prompts. These labels are then mapped back to corresponding 3D bounding boxes, which serve as semantic carriers that link the 2D image domain to the 3D scene layout. By projecting these annotated 3D boxes into the BEV space, we generate consistent BEV pseudo-labels for both base (*e.g.*, "car") and novel classes (*e.g.*, "truck"); see Fig. 1(d). Building upon these correspondences, in (2), BAGS leverages the derived BEV structural layouts as strong geometric priors to enforce 3D geometric consistency from sparse viewpoints; see Fig. 1(b,f). In (3), BAGD distills this high-fidelity geometric knowledge into a student network to ensure real-time inference; see Fig. 1(h). These stages successively leverage 3D geometric constraints to bypass 2D unprojection across semantics, geometry, and knowledge transfer, respectively.

3.1 Open-Vocabulary Pseudo-BEV Labeling (PBL)

In this section, we rethink standard BEV perception through the lens of OVBS, targeting robust novel-class recognition under base-class supervision alone. Unlike naive 2D-to-3D lifting, our 3D projection-first principle ensures (1) consistent 2D–BEV semantic propagation by leveraging unsupervised 3D detections as geometric constraints, thereby facilitating (2) BEV pseudo-label generation necessary for OV recognition—a linchpin of safety-critical AD.

Generating 2D–BEV Correspondences via 3D Projection. A straightforward route to consistent 2D–BEV object discovery is to run an off-the-shelf OV 2D detector [28], lift the detections to 3D via depth estimation [17,32], and then project them into BEV. This protocol is widely adopted in OV settings for fairness, since the detector is not trained on the dataset's novel-class ground truth [39,48,54]. However, these pipelines typically rely on ray sampling for depth and inevitably discard information in the 2D-to-3D transformation [24,36]. Despite notable progress in 2D-to-3D understanding, lifting image features into a metrically consistent 3D space remains fundamentally ill-posed and approximate due to occlusions, depth ambiguity, and discretization.

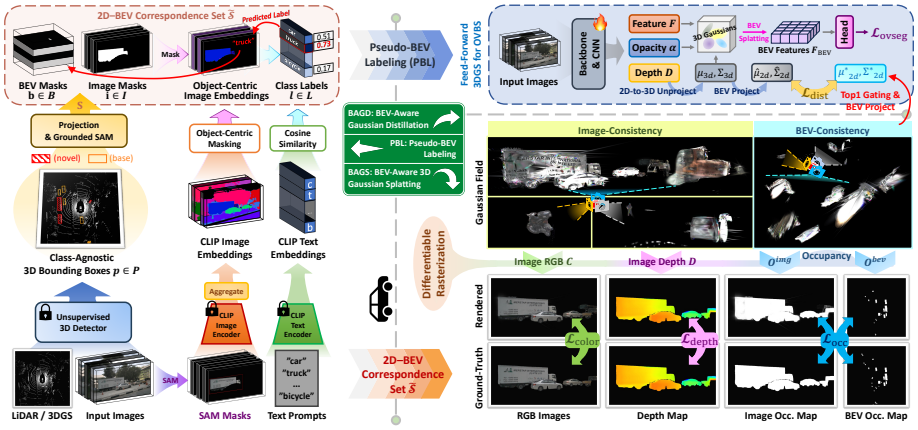

Fig. 2: Our method consists of three modules: **(1) PBL** generates BEV pseudo-labels by leveraging unsupervised 3D object discovery [43], foundational 2D segmentation [39], and VLM grounding [38] to construct a 2D–BEV correspondence set $\tilde{\mathcal{S}}$. **(2) BAGS** performs 3D geometry reconstruction using 2D, depth, and BEV occupancy supervision from $\tilde{\mathcal{S}}$. **(3) BAGD** distills this optimized 3D geometry into a feed-forward 3DGS-based student network in the BEV space, preserving fast inference speeds.

Instead, we avoid this unprojection ambiguity by predicting 3D object boxes with an unsupervised 3D detector and projecting them onto the image and BEV planes to establish reliable 2D–BEV correspondences. Specifically, we adopt the 3D bounding-box pseudo-annotations P from UNION [43], which deliver competitive class-agnostic detection by leveraging multi-sensor and temporal cues. Given these boxes, we invoke a 2D segmentation foundation model [39] to obtain pixel-level masks for dynamic objects within the projected pseudo-boxes. For objects seen by multiple cameras, we retain only the largest 2D instance $i \in I$ to maximize visual representation, enforcing a one-to-one match with a single BEV instance $b \in B$. Finally, we define the 2D–BEV correspondence set $\mathcal{S} = \{(i, b, p) \mid i \in I,\ b \in B,\ p \in P\}$. As shown in Fig. 2, it successfully encodes object-centric correspondences (*e.g.*, "truck") across 2D and BEV domains. This 3D-driven pipeline effectively circumvents the unstable depth estimation inherent in naive 2D unprojection—a consistent cross-modal route for 2D VLM semantics.

To mitigate failure modes arising from noisy 3D detection priors, we discard instances where the 2D mask area significantly deviates from its projected 2D bounding box. This mismatch signals that the 3D box is inadequately sized; specifically, an over-estimated 3D box leads to a substantial area discrepancy, while an under-estimated box typically yields a non-object response from the foundational grounding model. Such proactive pruning ensures that the resulting set $\mathcal{S}$ remains high-fidelity, even when raw 3D predictions are suboptimal.

Generating BEV Pseudo-Labels. Regarding OV pseudo-labeling, we adopt the standard CLIP-based paradigm, leveraging its shared embedding space to

align visual and linguistic concepts. Specifically, for each 2D region proposal, the image encoder generates a visual embedding, which is then compared against a set of text embeddings—encompassing both base and novel classes—to compute cosine similarity scores. These textual prompts can be derived from image captions [14, 53], dataset class names [55], or MLLM-driven zero-shot labels [9].

Following this paradigm, we initialize C categories from the dataset [2] using their text embeddings derived from standard prompt templates [14]. Using $\mathcal{S}$, we first extract 2D instance masks I to generate background-masked, object-centric crops for the CLIP image encoder, thereby mitigating semantic ambiguity near object boundaries. Each visual embedding is then assigned a pseudo-label l based on its maximum cosine similarity to the text embeddings of the C categories. By incorporating these labels into the original set—$\tilde{\mathcal{S}} = \{(i, b, p, l) \mid (i, b, p) \in S,\ l \in L\}$—we associate the pseudo-labels with their respective 3D instances, effectively lifting image-level VLM outputs into the BEV space. In Fig. 2, VLM embeddings for novel classes (*e.g.*, "truck") remain spatially consistent across both domains. As a result, PBL bridges the gap between closed-set training and OV generalization while relying strictly on base-class supervision.

3.2 BEV-Aware 3D Gaussian Splatting (BAGS)

In this section, we directly tackle the ill-posed 2D-to-3D unprojection inherent in GS-based approaches. Our core insight is that high-fidelity geometric cues can serve as strong guidance to recover the degraded geometric structure during training, thereby mitigating the 3D geometric inconsistencies as established in Sec. 1. Inspired by recent hybrid 3DGS designs [12, 41]—combining per-scene optimization for geometric fidelity with feed-forward mechanisms for efficiency and scalability—we leverage per-scene optimization as the source of this precise 3D supervision. To this end, we first revisit the 3DGS optimization process.

3DGS. 3DGS represents scenes as a collection of 3D Gaussian primitives, each defined by a mean μ and covariance Σ governing its spatial extent: $G(x) = \exp(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu))$. To optimize the attributes of these primitives, it employs tile-based rasterization to compute the rendered color $C(v)$ at pixel v:

$$C(v) = \sum_{k \in \mathcal{N}} c_k \alpha_k \prod_{j=1}^{k-1} (1 - \alpha_j), \tag{1}$$

where $\mathcal{N}$ denotes the set of Gaussians within the tile, c_k denotes the color of the k-th Gaussian and $\alpha_k = o_k G_k^{2D}(v)$ represents the contribution of the k-th Gaussian, with o_k being its opacity and $G_k^{2D}(\cdot)$ its 2D projection. The rendered color is subsequently supervised by a photometric loss against the ground truth.

While standard 3DGS yields robust 3D reconstructions that often act as a LiDAR surrogate [3, 7, 59], the resulting explicit 3D Gaussians fundamentally misalign with the BEV representation in real-world AD. As shown in Fig. 1(a,e), vanilla 3DGS faithfully reconstructs near-field camera views (*image-consistency*)

driven by the photometric loss; however, its BEV renderings often scatter splats across irrelevant, non-object regions (*BEV-inconsistency*)—a critical flaw for reliable BEV perception. This failure is particularly pronounced in driving scenarios characterized by sparse camera coverage and minimal multi-view overlap.

BAGS. Beyond image-level cues, we contend that BEV structural layouts—driven by our 3D projection-first principle—can impose stringent xy-plane spatial priors to mitigate the depth ambiguity inherent in sparse 2D supervision. By providing such robust geometric constraints, this novel paradigm offers reliable guidance for precise 3D localization where visual data alone is insufficient. To this end, we propose BAGS, which enforces cross-modal consistency by synchronizing 2D and BEV representations from the correspondence set $\tilde{\mathcal{S}}$. As illustrated in Fig. 2, BAGS prioritizes the learning of object-centric 3D geometry through occupancy maps as opposed to photometric attributes, effectively leveraging the color-agnostic nature of the BEV domain. Specifically, we optimize the accumulated occupancy $O(v)$ via a differentiable rasterization pipeline:

$$O(v) = 1 - \prod_{k \in \mathcal{N}} (1 - \alpha_k). \tag{2}$$

The resulting 2D and BEV occupancy maps, O^{img} and O^{bev}, are aligned with the mask priors in $\tilde{\mathcal{S}}$ via a smooth L1 loss. This joint optimization over instance masks $i \in I$ and BEV instances $b \in B$ ensures cross-modal consistency:

$$\mathcal{L}_{\text{occ}} = \left\| O^{\text{img}}(v) - i(v) \right\|_{\text{SL}_1} + \left\| O^{\text{bev}}(v) - b(v) \right\|_{\text{SL}_1}. \tag{3}$$

Moreover, motivated by depth regularization techniques for outdoor scenes [10, 59], we optimize the per-pixel depth $D(v)$ via the rasterization pipeline:

$$D(v) = \sum_{k \in \mathcal{N}} d_k \alpha_k \prod_{j=1}^{k-1} (1 - \alpha_j), \tag{4}$$

where d_k is the camera-space depth of the k-th Gaussian center. We supervise $D(v)$ using a dense depth prior D^*_{dense}, obtained by regularizing the predictions of a depth estimator F_θ [1] with metrically consistent and sparse LiDAR data:

$$\mathcal{L}_{\text{depth}} = \left\| D(v) - D^*_{\text{dense}}(v) \right\|_1, \quad \text{where } D^*_{\text{dense}} = s^* \cdot F_\theta(\mathbf{I}) + t^*, \tag{5}$$

where s^* and t^* are optimized scale and translation parameters, and $\mathbf{I}$ is the input image. Finally, the total objective for BAGS is defined as:

$$\mathcal{L}_{\text{BAGS}} = (1 - \lambda_{\text{ssim}})\mathcal{L}_{\text{color}} + \lambda_{\text{ssim}}\mathcal{L}_{\text{D-SSIM}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{occ}}\mathcal{L}_{\text{occ}}, \tag{6}$$

where $\mathcal{L}_{\text{color}}$ and $\mathcal{L}_{\text{D-SSIM}}$ follow the original 3DGS formulation [22]. This joint supervision effectively constrains the 3D primitives within a well-defined spatial volume, preventing the geometric collapse often observed in sparse-view settings. To suppress background noise, we mask the image with the instance map I, thereby preventing the proliferation of redundant floater Gaussians in irrelevant regions. As shown in Fig. 1(b,f) and 2, BAGS yields compact and geometrically faithful 3D Gaussians, establishing a robust source of geometric supervision.

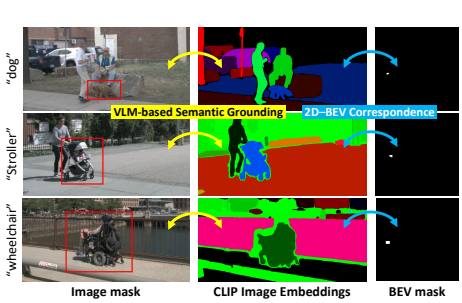

Fig. 3: Open-world 2D–BEV auto-labeling.

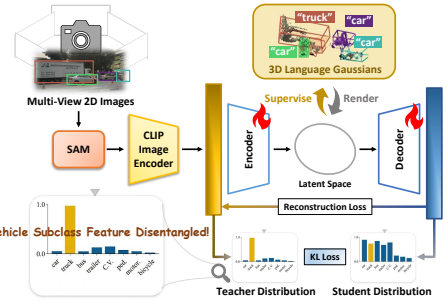

Fig. 4: Language-embedded BAGS.

Language-Embedded BAGS. The rising demand for OV 3D scene understanding in AD underscores the extensibility of BAGS, where 3D Gaussians inherit VLM attributes. A key challenge in this task is reducing the memory overhead incurred by processing high-dimensional VLM embeddings during 3DGS optimization. As shown in Fig. 4, recent work [37] mitigates this by establishing a lower-dimensional latent space via an encoder-decoder pretraining; it optimizes compact latent features as ground-truth targets. However, this innovative design consistently fails in AD scenarios, where features of vehicle subclasses (*e.g.*, "car" and "truck") become entangled during the projection into the latent space. Consequently, the decoded features reveal previously unobserved disentanglement problems among semantically similar subclasses.

To validate the extensibility of language-embedded BAGS—Lang–BAGS—we propose a disentanglement distillation loss that aligns the categorical similarity distributions of reconstructed features with those of original VLM embeddings over subclass embeddings $\mathbf{W}$, inducing discriminative latent projections:

$$\mathcal{L}_{\text{Lang–BAGS}} = \tau^2 \cdot \text{KL}\left[\sigma\left(\frac{s \cdot \mathbf{x}_{\text{orig}}\mathbf{W}^\top}{\tau}\right) \parallel \sigma\left(\frac{s \cdot \mathbf{x}_{\text{rec}}\mathbf{W}^\top}{\tau}\right)\right], \tag{7}$$

where $\sigma(\cdot)$ denotes the softmax function; s and τ are the logit scale and temperature, respectively; $\mathbf{x}_{\text{orig}}$ is the original VLM embedding; and $\mathbf{x}_{\text{rec}}$ is the reconstructed feature. As shown in Figs. 3 and 6, Lang–BAGS facilitates the extension of BAGS to OV 3D scene grounding, where its sparse-view robustness maintains language consistency across image and BEV domains.

3.3 BEV-Aware Gaussian Distillation (BAGD)

In this section, we introduce OVBEVSeg, a hybrid 3DGS-based framework addressing three key objectives: open-world recognition through PBL, 3D geometric reconstruction via BAGS, and efficient knowledge transfer using BAGD. Following recent hybrid paradigms [12, 41], we first optimize teacher Gaussians offline with BAGS on temporally downsampled sequences to capture discriminative scene semantics. These high-fidelity primitives provide a rigorous geometric foundation, serving as a dense supervisory signal to guide the student model.

A fundamental challenge in this geometric distillation lies in the inherent lack of one-to-one correspondence between student and teacher Gaussians. While the student predicts a grid-aligned feature tensor $\hat{\mathbf{G}} \in \mathbb{R}^{N \times H_F \times W_F \times d}$, the teacher conversely optimizes an unstructured, dense set of primitives. To bridge this structural gap, we leverage per-pixel opacity α to identify the most influential teacher Gaussian for each ray—a natural choice since objects in driving scenes are rarely stacked vertically in BEV. This allows a single dominant Gaussian to effectively represent the local 3D geometry. Specifically, for each pixel v, we select a representative Gaussian $G^*(v) := G_{k^*(v)}$ by maximizing its rendering contribution: $k^*(v) = \arg\max_{k \in \mathcal{N}(v)} \alpha_k$. By mapping the attributes of these selected primitives into the supervision tensor $\mathbf{G}$, we ensure pixel-perfect spatial alignment, effectively anchoring discrete grids to continuous 3D Gaussians.

On the student side, the predicted 3D Gaussians are projected onto the same BEV plane to obtain $\hat{\mathbf{G}}$. We then train the student via a symmetrized distillation objective [26] that enforces teacher–student distribution alignment:

$$\mathcal{L}_{\text{BAGD}} = \frac{1}{2|\mathcal{V}|} \sum_{v \in \mathcal{V}} \Big[\text{KL}(\hat{G}(v) \| G^*(v)) + \text{KL}(G^*(v) \| \hat{G}(v)) \Big], \tag{8}$$

where $\mathcal{V}$ denotes the set of pixel coordinates. This distillation strategy effectively harmonizes the two 3DGS variants, synergizing their respective strengths to satisfy stringent real-time requirements within a unified OVBS framework.

4 Experiments

Datasets and evaluation metrics. OVBEVSeg is evaluated on the official nuScenes dataset [2], a large-scale AD benchmark with synchronized multi-sensor data across diverse weather and times of day. It contains 1,000 scenes split into 700 training, 150 validation, and 150 test scenes. Each 20-second scene includes data from six cameras providing a full 360-degree view around the ego vehicle. For OVBS, we follow the category split in prior work [11], dividing vehicle categories into five base and three novel classes. For fair comparison with prior work [32], we apply a visibility filter that selects objects with at least 40% coverage. This focuses the evaluation on adequately visible objects. The BEV grid has size (200×200), covering ([-50 m, 50 m]) along both the X (forward) and Y (sideways) axes relative to the ego vehicle. Each cell corresponds to (0.5 m×0.5 m).

Implementation details. The online framework operates exclusively on camera inputs with LiDAR priors confined to offline, and is optimized using a weighted sum of a segmentation loss on cosine-similarity logits [38] and a KL divergence loss [26], with weights set to 1.0 and 0.01, respectively. For category prompting, we utilize the hand-crafted templates from ViLD [14]. Using AdamW [30] with a learning rate of 3×10^{-4} and weight decay of 1×10^{-7}, the model is trained for 50 epochs with a batch size of 8 on two A100 GPUs. Input images are resized to 224×480 without augmentation, processing multi-scale BEV features at resolutions of 50×50, 100×100, and 200×200. By default, we use an EfficientNet-B4 backbone [42] and disable visibility filtering during inference.

Table 1: Comparison of OVBS in the multi-class setting on the nuScenes *validation* set [2]. Note that † denotes visibility filtering. Δ_B and Δ_S denote the relative improvements over the baseline [32] and the SOTA model [56], respectively.

Method	Memory (MiB) ↓	FPS ↑	Novel (IoU) ↑			Base (IoU) ↑					Mean IoU ↑
			truck	bus	motorcycle	car	trailer	construction_vehicle	pedestrian	bicycle	
Fully Supervised† [32]	33.0	80.2	29.7	38.4	10.4	-	-	-	-	-	-
OVBS											
VED† [31]	-	-	0.0	0.0	0.0	7.4	0.0	0.0	0.0	0.0	0.9
VPN† [34]	-	-	0.0	0.0	0.0	16.6	4.9	7.1	0.0	4.4	4.1
PON† [40]	38.6	43.8	0.0	0.0	0.0	24.7	16.6	12.3	8.2	9.4	8.9
DiffBEV† [61]	-	-	0.0	0.0	0.0	38.9	21.1	8.4	9.6	13.2	11.4
GaussianLSS† [32]	33.0	80.2	0.0	0.0	0.0	40.0	25.1	11.6	14.4	10.4	12.7
TaDe† [56]	41.5	51.4	0.0	0.0	0.0	**42.8**	26.3	11.4	14.0	**14.2**	13.6
OVBEVSeg (Ours)†	32.4	79.6	**19.0**	**20.3**	**6.6**	41.8	**26.9**	**13.1**	**15.0**	12.0	**19.3**
			Δ_B (+19.0)	(+20.3)	(+6.6)	(+1.8)	(+1.8)	(+1.5)	(+0.6)	(+1.6)	(+6.6)
			Δ_S (+19.0)	(+20.3)	(+6.6)	(-1.0)	(+0.6)	(+1.7)	(+1.0)	(-2.2)	(+5.7)

Table 2: Single-class OVBS comparison. ∗: novel classes withheld during training.

Method	Memory (MiB) ↓	FPS ↑	IoU ↑		
			vehicle∗	pedestrian†	mIoU
Fully Supervised [32]	33.0	80.2	38.3	18.0	28.1
OVBS					
CVT [57]	35.0	107.6	24.6	14.2	19.4
FIERY static [17]	40.0	27.3	28.0	17.2	22.6
SimpleBEV [15]	331.0	37.1	28.9	17.1	23.3
GaussianLSS [32]	33.0	80.2	29.7	18.0	23.9
PointBeV [5]	126.0	32.0	30.3	**18.5**	24.4
OVBEVSeg (Ours)	28.7	80.2	**34.7**	18.4	**26.6**
			Δ_B (+5.0)	(+0.4)	(+2.7)
			Δ_S (+4.4)	(-0.1)	(+2.2)

Table 3: OVBS results with self- and semi-supervised baselines on nuScenes.

GT Ratio	Method	truck	bus	motorcycle	mIoU ↑
0%	2D-OV-then-BEV	3.7	2.6	0.0	2.1
	OVBEVSeg (Ours)	19.0	20.3	6.6	15.3
(1, 5%]	LetsMap [13]	13.6	-	5.8	9.7
	UniMatch [51]	10.4	5.7	1.6	5.9
	Semi-BEVseg [60]	14.7	9.8	1.6	8.7
20%	UniMatch [51]	16.8	10.5	2.6	10.0
	Semi-BEVseg [60]	20.4	16.8	6.7	14.6
40%	UniMatch [51]	18.0	17.0	5.6	13.5
	Semi-BEVseg [60]	**22.7**	**21.6**	**6.8**	**17.0**

4.1 Quantitative Research

Comparison with closed-set methods. We evaluate OVBEVSeg against state-of-the-art (SOTA) closed-set BEV segmentation models on nuScenes [2] under OVBS settings, focusing on novel classes. In Tabs. 1 and 2, SC and MC denote training on a single class (*e.g.*, "vehicle") and multiple classes jointly, respectively; notably, the "vehicle" category in SC uses only base-class annotations. Leveraging our BEV pseudo-labels, OVBEVSeg sets a new SOTA on novel classes, surpassing previous projection-based methods [5, 56] by 4.4 and 15.3 mIoU in SC and MC, while being 2.7× faster and consuming only 0.22× the memory. It also outperforms the feed-forward 3DGS-based baseline [32] by 2.7 and 6.6 mIoU, while maintaining competitive online efficiency. These results strongly validate that our OVBS framework paves the way for practical deployment.

Comparison with semi- and self-supervised methods. Despite the clear necessity of OVBS, the current scarcity of direct competitors in this emerging domain leads us to broaden our evaluation. We include two additional baselines with objectives relevant to OVBS. First, we compare against self- and semi-supervised methods [13, 51, 60] trained with varying ground-truth (GT) ratios. In Tab. 3, our method—operating with 0% GT—significantly outperforms these baselines with $< 20\%$ GT, and remains competitive with those using up to 40%. Second,

Table 4: OVBS comparison with different 3D priors and non-vehicle classes (4 novel classes).

Method	Novel (IoU) ↑			
	truck	bus	motorcycle	cone
UNION [43]	**19.0**	**20.3**	6.6	-
F&B [11]	14.2	15.5	**7.9**	**11.5**

Table 5: Per-sample submodule runtime on an A100 GPU. Inference-only modules are shared with the baseline. ♮ uses 10% of per-scene samples.

Online Method	Backbone	Neck	Encoder	Head	BAGD♮ (ms)
Ours (BAGD♮)	29.3	5.6	5.7	1.38	0.37

Offline Method	3D DET	SAM	CLIP	BAGS♮	Total (min)
Ours (PBL + BAGS♮)	< 0.01	0.23	0.32	1.54	2.1

a "2D-OV-then-BEV" pipeline—comprising a LiDAR-regularized depth estimator [10] and a strong OV detector [28]—is evaluated as a naive OVBS baseline. Our approach markedly outperforms this alternative, particularly in sparse-view settings where 2D unprojection errors are amplified. These fairness comparisons underscore the need for our method in the critical yet underserved OVBS task.

4.2 Ablation Study

Ablation of 3D priors and non-vehicle classes. Table 4 shows that PBL consistently improves OVBS across unsupervised [43] and VLM-based OV priors [11], while reflecting their complementary strengths, such as better recognition of small static objects with the OV prior. This highlights the generality of PBL beyond vehicle-centric categories and its compatibility with future 3D priors.

Submodule runtime. Table 5 details the per-scene submodule runtimes on an A100 GPU. Our model extends the baseline with a lightweight BAGD module completely removed during inference. The remaining components (*e.g.*, PBL, BAGS, and cached 3D detections) are precomputed and cached offline prior to online training. Each submodule incurs negligible computational overhead via asynchronous batch parallelism—SAM producing masks for batch-wise CLIP aggregation—with BAGS leveraging our online efficient distillation scheme. The offline preprocessing of the entire training set—roughly 30,000 samples—takes approximately 4 days using 8 A100 GPUs, while online efficiency remains on par with the baseline. These results highlight that OVBEVSeg significantly boosts OVBS performance while preserving competitive computational online efficiency.

Ablation of the individual modules. Table 6 evaluates the contribution of each component within OVBEVSeg. While Group II establishes a new SOTA for novel classes, thereby facilitating OV generalization, the substantial performance degradation in Group III underscores a critical limitation: integrating vanilla 3DGS supervision with BAGD propagates inaccurate 3D geometry, leaving the 2D unprojection dilemma unresolved. In contrast, Group IV demonstrates that combining BAGS with BAGD ensures geometric consistency across camera and BEV views, providing constructive supervision that boosts overall OVBS performance by refining boundary details across all classes. These results confirm that the synergistic integration of all proposed modules yields a robust OVBS framework for real-world driving scenarios.

Table 6: Component ablation.

Group	Components			IoU ↑		
	PBL	BAGS	BAGD	Novel	Base	Mean
I				0.0	20.3	10.2
II	✓			14.0	20.5	17.3
III	✓		✓	6.1	9.8	7.9
IV	✓	✓	✓	**15.3**	**21.8**	**19.3**

Table 7: BAGD ratio.

Ratio	IoU ↑		
	Novel	Base	Mean
0	14.0	20.5	17.8
0.05	14.8	21.5	18.9
0.1	**15.3**	**21.8**	**19.3**

Table 8: Map segmentation.

Method	drivable	crossing	walkway
TaDe	65.9	40.9	42.3
CVT	74.3	36.8	39.9
LSS	75.4	38.8	46.3
GaussianLSS	76.3	46.3	50.2
Ours	**78.1**	**49.4**	**51.9**

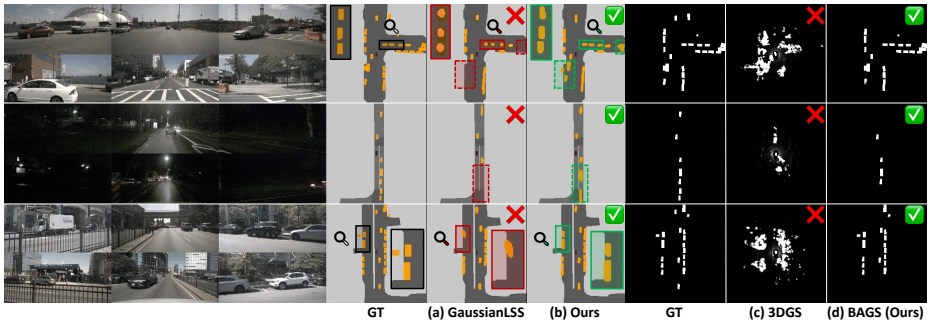

Fig. 5: (a–b) Our BEV predictions vs. the baseline: orange regions denote both base and novel vehicle classes. **(c–d)** Our BEV renderings (BAGS) vs. vanilla 3DGS.

Distillation ratio. We evaluate various BAGD ratios to assess their efficacy in 3D geometric knowledge transfer. A typical nuScenes [2] scene consists of approximately 40 frames captured at 2 Hz. Given the high temporal redundancy between consecutive frames [45,52], we sample frames at uniform intervals to mitigate 3DGS optimization overhead. This selective distillation strategy enhances the online efficiency of BAGS by focusing on key scene-discriminative frames. As shown in Tab. 7, OVBS performance scales near-linearly with the distillation ratio. Notably, a ratio of 0.1 (*i.e.*, 4 frames per scene) achieves the optimal efficiency–accuracy balance, underscoring our framework's robust scalability.

Map segmentation. Table 8 evaluates map classes requiring precise road geometry. Comparing our BAGD-enhanced model against existing methods, we predict categories such as drivable areas, pedestrian crossings, and walkways following the same setup of prior work [32]. Our approach achieves SOTA performance across all classes, effectively reinforcing the underlying scene structure.

4.3 Qualitative Research

Qualitative results. As illustrated in Fig. 5, the baseline [32] **(a)** fails to detect novel-class objects (*e.g.*, "truck") and collapses geometries into unstructured, amorphous volumes, compromising reliability in safety-critical AD. In contrast, OVBEVSeg **(b)** successfully captures novel classes with sharp boundaries, underscoring its efficacy for OVBS. Regarding geometric fidelity under sparse ego-car viewpoints, vanilla 3DGS **(c)** suffers from inaccurate geometry and misaligned

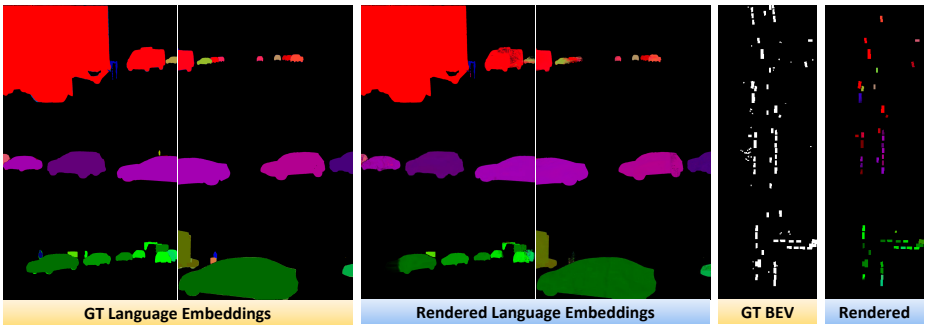

Fig. 6: Language-embedded OV 3D scene understanding via the Lang–BAGS protocol.

BEV renderings, whereas BAGS **(d)** effectively resolves these inconsistencies, ensuring robustly aligned 3D geometries as reliable OVBS supervision.

Open-world 2D–BEV auto-labeling. Given the increasing scale of datasets, automatic multimodal annotation is essential for real-world AI applications like AD. In contrast to closed-set approaches, our PBL system facilitates OV auto-labeling (*e.g.*, "stroller" in Fig. 3. By projecting cross-modal similarity signals into BEV space via the precomputed correspondence set, our system successfully yields synchronized 3D, 2D, and BEV annotations. This framework establishes a scalable pipeline for high-fidelity, multimodal OV auto-labeling in 3D.

Open-world 3D scene understanding via 3DGS. Figure 5 demonstrates that Lang–BAGS effectively embeds CLIP features into 3D Gaussians, ensuring their projections in both image and BEV domains consistently align with GT embeddings. Beyond standard 3DGS, BAGS enables language-aware 3D scene understanding in AD scenarios, maintaining robustness even with sparse viewpoints.

5 Conclusion

In this work, we introduced open-vocabulary BEV segmentation (OVBS) to address the inherent safety-critical limitations of closed-set driving systems. To mitigate the 3D geometric inconsistencies common in 2D-to-3D semantic lifting—identified as a primary obstacle in OVBS—we proposed OVBEVSeg, the first OVBS framework that enhances Gaussian splatting (GS)-based unprojection by leveraging robust 3D geometric constraints through three synergistic modules—(1) PBL: pseudo-BEV labeling via unsupervised 3D projection for OV generalization; (2) BAGS: 2D–BEV-consistent 3D Gaussian splatting for 3D geometry recovery; and (3) BAGD: efficiency-preserving geometric distillation. Extensive evaluations validate that OVBEVSeg achieves state-of-the-art novel-class recognition with a competitive memory footprint and real-time speed. By enforcing geometric consistency across the 2D-3D-BEV continuum, this work establishes a scalable foundation for robust open-world multimodal 3D perception.

Acknowledgements

This research was supported by a grant (code RS-2023-00233952) from the R&D Program funded by the Ministry of Land, Infrastructure and Transport of the Korean government.

References

1. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. CoRR **abs/2302.12288** (2023)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 11618–11628 (2020)
3. Cao, Y., Jv, Y., Xu, D.: 3dgs-det: Empower 3d gaussian splatting with boundary guidance and box-focused sampling for 3d object detection (2024)
4. Chabot, F., Granger, N., Lapouge, G.: Gaussianbev: 3d gaussian representation meets perception models for bev segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025. pp. 2250–2259 (2025)
5. Chambon, L., Zablocki, É., Chen, M., Bartoccioni, F., Pérez, P., Cord, M.: Pointbev: A sparse approach to bev predictions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 15195–15204 (2024)
6. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: Pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 19457–19467 (2024)
7. Chen, Q., Yang, S., Du, S., Tang, T., Chen, P., Huo, Y.: Lidar-gs:real-time lidar re-simulation using gaussian splatting. CoRR **abs/2410.05111** (2024)
8. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXI. vol. 15079, pp. 370–386 (2024)
9. Choi, H., Lim, Y., Shin, J., Shim, H.: Mspl: Multi-step pseudo-labeling for open-vocabulary object detection (2026)
10. Chung, J., Oh, J., Lee, K.M.: Depth-regularized optimization for 3d gaussian splatting in few-shot images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024. pp. 811–820 (2024)
11. Etchegaray, D., Huang, Z., Harada, T., Luo, Y.: Find n' propagate: Open-vocabulary 3d object detection in urban environments. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XL. vol. 15098, pp. 133–151 (2024)
12. Fan, Z., Cong, W., Wen, K., Wang, K., Zhang, J., Ding, X., Xu, D., Ivanovic, B., Pavone, M., Pavlakos, G., Wang, Z., Wang, Y.: Instantsplat: Sparse-view gaussian splatting in seconds (2025)
13. Gosala, N., Petek, K., Kiran, B.R., Yogamani, S.K., Drews-Jr, P.L.J., Burgard, W., Valada, A.: Letsmap: Unsupervised representation learning for label-efficient semantic BEV mapping. In: ECCV (2024)

14. Gu, X., Lin, T., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022 (2022)
15. Harley, A.W., Fang, Z., Li, J., Ambrus, R., Fragkiadaki, K.: Simple-bev: What really matters for multi-sensor BEV perception? In: IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023. pp. 2759–2765 (2023)
16. Hess, G., Tonderski, A., Petersson, C., Åström, K., Svensson, L.: Lidarclip or: How I learned to talk to point clouds. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024. pp. 7423–7432 (2024)
17. Hu, A., Murez, Z., Mohan, N., Dudas, S., Hawke, J., Badrinarayanan, V., Cipolla, R., Kendall, A.: FIERY: future instance prediction in bird's-eye view from surround monocular cameras. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 15253–15262 (2021)
18. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: ST-P3: end-to-end vision-based autonomous driving via spatial-temporal feature learning. In: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVIII. vol. 13698, pp. 533–549 (2022)
19. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., Jia, X., Liu, Q., Dai, J., Qiao, Y., Li, H.: Planning-oriented autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 17853–17862 (2023)
20. Huang, J., Huang, G., Zhu, Z., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. CoRR **abs/2112.11790** (2021)
21. Jia, X., Gao, Y., Chen, L., Yan, J., Liu, P.L., Li, H.: Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 7919–7929 (2023)
22. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**, 139:1–139:14 (2023)
23. Li, S., Fischer, T., Ke, L., Ding, H., Danelljan, M., Yu, F.: Ovtrack: Open-vocabulary multiple object tracking. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 5567–5577 (2023)
24. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023. pp. 1477–1485 (2023)
25. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX. pp. 1–18 (2022)
26. Lin, Y., Wang, C., Chang, C., Sun, H.: An efficient framework for counting pedestrians crossing a line using low-cost devices: the benefits of distilling the knowledge in a neural network. Multim. Tools Appl. **80**(3), 4037–4051 (2021)

27. Liu, H., Teng, Y., Lu, T., Wang, H., Wang, L.: Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 18534–18544 (2023)
28. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII. vol. 15105, pp. 38–55 (2024)
29. Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X.: Petrv2: A unified framework for 3d perception from multi-camera images. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 3239–3249 (2023)
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019 (2019)
31. Lu, C., van de Molengraft, M.J.G., Dubbelman, G.: Monocular semantic occupancy grid mapping with convolutional variational encoder-decoder networks. IEEE Robotics Autom. Lett. **4**(2), 445–452 (2019)
32. Lu, S., Tsai, Y., Chen, Y.: Toward real-world BEV perception: Depth uncertainty estimation via gaussian splatting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 17124–17133 (2025)
33. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-vocabulary 3d detection via image-level class and debiased cross-modal contrastive learning. CoRR **abs/2207.01987** (2022)
34. Pan, B., Sun, J., Leung, H.Y.T., Andonian, A., Zhou, B.: Cross-view semantic segmentation for sensing surroundings. IEEE Robotics Autom. Lett. **5**(3), 4867–4873 (2020)
35. Peng, S., Genova, K., Jiang, C.M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.A.: Openscene: 3d scene understanding with open vocabularies. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 815–824 (2023)
36. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV. vol. 12359, pp. 194–210 (2020)
37. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 20051–20060 (2024)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. vol. 139, pp. 8748–8763 (2021)
39. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded SAM: assembling open-world models for diverse visual tasks. CoRR **abs/2401.14159** (2024)

40. Roddick, T., Cipolla, R.: Predicting semantic map representations from images using pyramid occupancy networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 11135–11144 (2020)
41. Tan, K., Shen, Y., Tu, M., Zhu, H., Wang, B., Chen, G., Ye, H., Sun, H.: Ufo: Unifying feed-forward and optimization-based methods for large driving scene modeling (2026)
42. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. vol. 97, pp. 6105–6114 (2019)
43. de Vries Lentsch, T., Caesar, H., Gavrila, D.: UNION: unsupervised 3d object detection using object appearance-based pseudo-classes. In: Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024)
44. Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 3598–3608 (2023)
45. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 20310–20320 (2024)
46. Wu, J., Li, X., Xu, S., Yuan, H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., Jiang, X., Ghanem, B., Tao, D.: Towards open vocabulary learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **46**(7), 5092–5113 (2024)
47. Wu, J., Li, X., Xu, S., Yuan, H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., Jiang, X., Ghanem, B., Tao, D.: Towards open vocabulary learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **46**, 5092–5113 (2024)
48. Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., Cheng, X., Zhao, C., Feng, H., Ding, E., Wang, J., Zhang, J.: Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. In: Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024)
49. Xie, E., Yu, Z., Zhou, D., Philion, J., Anandkumar, A., Fidler, S., Luo, P., Álvarez, J.M.: M^2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. CoRR **abs/2204.05088** (2022)
50. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depthsplat: Connecting gaussian splatting and depth. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 16453–16463 (2025)
51. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: CVPR (2023)
52. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 2633–2642 (2020)
53. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.: Open-vocabulary object detection using captions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 14393–14402 (2021)

54. Zeng, Z., Boehm, J.: Exploration of an open vocabulary model on semantic segmentation for street scene imagery. ISPRS Int. J. Geo Inf. **13**(5), 153 (2024)
55. Zhao, S., Zhang, Z., Schulter, S., Zhao, L., Kumar, B.G.V., Stathopoulos, A., Chandraker, M., Metaxas, D.N.: Exploiting unlabeled data with vision and language models for object detection. In: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX. vol. 13669, pp. 159–175 (2022)
56. Zhao, T., Chen, Y., Wu, Y., Liu, T., Du, B., Xiao, P., Qiu, S., Yang, H., Li, G., Yang, Y., Lin, Y.: Improving bird's eye view semantic segmentation by task decomposition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 15512–15521 (2024)
57. Zhou, B., Krähenbühl, P.: Cross-view transformers for real-time map-view semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 13750–13759 (2022)
58. Zhou, H., Shao, J., Xu, L., Bai, D., Qiu, W., Liu, B., Wang, Y., Geiger, A., Liao, Y.: HUGS: holistic urban 3d scene understanding via gaussian splatting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 21336–21345 (2024)
59. Zhou, X., Lin, Z., Shan, X., Wang, Y., Sun, D., Yang, M.: Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 21634–21643 (2024)
60. Zhu, J., Liu, L., Tang, Y., Wen, F., Li, W., Liu, Y.: Semi-supervised learning for visual bird's eye view semantic segmentation (2024)
61. Zou, J., Tian, K., Zhu, Z., Ye, Y., Wang, X.: Diffbev: Conditional diffusion model for bird's eye view perception. In: Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada. pp. 7846–7854 (2024)