

RESEARCH ARTICLE

LGD-BEV: Label-Guided Distillation for BEV 3D Object Detection

JI-YONG LEE^{1,2}, MINHO PARK¹, YOUNGJOO JO¹, AND DONG-OH KANG¹¹Visual Intelligence Laboratory, Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, Republic of Korea²The University of Texas at Austin, Austin, TX 78712, USA

Corresponding author: Dong-Oh Kang (dongoh@etri.re.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government (MSIT), Development of AI Bots Collaboration Platform and Self-Organizing AI, 90%, under Grant RS-2022-II220907; and in part by the Electronics and Telecommunications Research Institute (ETRI) grant funded by Korean Government, Core Technology Research for Self-Improving Integrated Artificial Intelligence System, 10%, under Grant 25ZB1100.

ABSTRACT Knowledge distillation in autonomous driving typically involves transferring information from high-cost sensor modalities such as LiDAR to camera-based models to improve 3D object detection. However, these methods often require access to additional modalities during training or inference, leading to increased engineering complexity and extensive trial-and-error efforts. This limitation hinders their scalability and deployment in real-world applications. In contrast, our approach leverages only 3D ground-truth annotations as an auxiliary intermediate supervision, acting as a lightweight teacher, thereby removing the dependency on external sensors or strong pre-trained teachers. In this paper, we introduce a Label Encoder module to generate dense semantic representations in the Bird's-Eye View (BEV) space to guide the training of camera-based networks through knowledge distillation. Specifically, we train an AutoEncoder on 3D ground-truth annotations and use its encoder to produce informative BEV features, which are then used as an auxiliary loss for the camera-based detection backbone. This yields a lightweight, modality-independent, and scalable solution, as demonstrated by extensive evaluations on the NuScenes dataset.

INDEX TERMS AutoEncoder, autonomous driving, knowledge distillation, multi-view cameras, object detection.

I. INTRODUCTION

In autonomous driving, recent research has explored camera-only 3D perception, which leverages multi-view images captured by standard RGB cameras [1], [2], [3]. Although camera-based systems are more cost-effective and widely deployable, they often underperform compared to their LiDAR-based counterparts due to the absence of explicit depth and geometric cues. To address this performance gap, knowledge distillation (KD) [4] has become a prominent technique. KD enables a compact camera-based student network to learn from a high-capacity LiDAR-based teacher network [5], [6], [7], thereby improving the student's performance without increasing the inference cost.

The associate editor coordinating the review of this manuscript and approving it for publication was Angelo Trotta¹.

Despite its promise, existing knowledge distillation approaches in 3D detection often fall into two categories: relying on knowledge from high-capacity teacher models and leveraging additional sensor modalities such as infrared or LiDAR [8]. The first category often demands complex and computationally intensive teacher networks to effectively guide the student, increasing training cost and model complexity [9], [10], [11]. The second introduces the dependency on additional data sources, which may be expensive, unavailable, or impractical in real-world deployments [12], [13], [14]. Both types of methods create significant engineering overhead, tight coupling between sensors, and limited scalability. Furthermore, they often require extensive fine-tuning and hyperparameter tuning, further complicating the training process.

In this paper, we propose a lightweight and modality-independent knowledge distillation framework tailored for

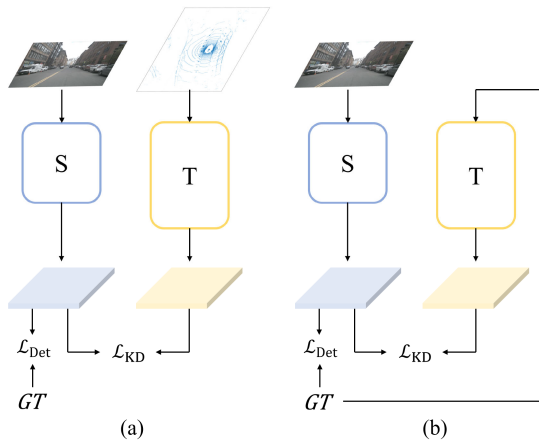


FIGURE 1. Comparison between traditional cross-modal knowledge distillation and our proposed approach. ‘S’ denotes student and ‘T’ denotes teacher network. (a) Traditional methods rely on additional sensor modalities (e.g., LiDAR) during training to distill knowledge. (b) Our method, LGD-BEV, removes the dependency on extra modalities by distilling from a lightweight teacher trained solely with feature representations derived from ground truth annotations.

camera-only 3D perception. Instead of relying on external sensors or pretrained LiDAR models, we leverage the 3D ground-truth annotations as a source of distilled spatial knowledge. Specifically, we design a Label Encoder module that serves as an implicit teacher, trained to capture semantic and spatial distributions of objects in the Bird’s-Eye View (BEV) space using only 3D annotations.

Figure 1 provides a conceptual comparison between traditional cross-modal distillation (a), which requires additional modalities such as LiDAR during training, and our approach (b), which eliminates such dependencies. In our method, the teacher is derived from 3D ground-truth labels alone, avoiding the need for multi-sensor alignment and enabling a sensor-agnostic training pipeline. This design allows the student network to benefit from spatial supervision without access to any external sensor data during training or inference.

The Label Encoder is trained in an autoencoding setup, where it learns to compress 3D label maps into informative 2D BEV representations. These representations are then used to supervise the training of the camera-based detector via an auxiliary distillation loss, following standard knowledge distillation practices. Our method is simple yet efficient, requiring no additional sensors, no access to inference-time teachers, and no heavy pretraining.

To further improve the effectiveness of our distillation framework, we introduce a Gating Fusion Module that dynamically adjusts the weighting between features derived from the AutoEncoder’s label-based reconstruction and those based on geometric box attributes. Rather than simply summing the two feature streams as in previous work [15], the module employs a learnable gating mechanism that produces an element-wise weight map. This map controls the relative contribution of each feature type at every spatial location in the BEV space. As a result, the model can

adaptively emphasize class-specific semantic cues from label or geometric information from box features depending on the region, leading to more precise BEV representations.

Additionally, we applied a Class-Aware Rendering strategy to further enhance the semantic quality of the BEV label maps generated by the Label Encoder during its autoencoding process. In previous rasterization [15], overlapping objects are often written in arbitrary order, causing rare or small objects to be overwritten by dominant classes like cars. Our method assigns each class a priority level based on class frequency statistics from the nuScenes [16] dataset, following insights from [17], allowing lower-frequency or critical classes such as trailers or bicycles to overwrite higher-frequency ones in the rasterization step. This ensures that underrepresented classes are not suppressed during label encoding. This mechanism is simple to implement yet effective in preserving the representational richness and supervision quality of the label-encoded BEV features.

In summary, our proposed framework enables effective camera-only 3D object detection by distilling label-derived BEV representations in a self-supervised manner, removing the reliance on external sensor teachers. By combining Label Encoder-based supervision, adaptive feature fusion, and class-aware encoding, our method provides a scalable, lightweight, and high-performing solution for monocular 3D detection.

To validate the effectiveness of our proposed framework, we conduct experiments on the nuScenes [16] benchmark for autonomous outdoor 3D detection. Our method significantly improves the performance of camera-only 3D object detectors and achieves competitive results compared to other state-of-the-art approaches. Our key contributions are summarized as follows:

- We propose the Label-Guided Knowledge Distillation framework that enables efficient training of camera-only 3D detectors using only 3D ground-truth annotations, without relying on an external sensor.
- We introduce a Gating Fusion Module that enhances the BEV representations generated by the Label Encoder by dynamically balancing label-derived semantic features and box-derived geometric features through spatially adaptive fusion.
- We apply a Class-Aware Rendering strategy in the Label Encoder to preserve rare object classes during BEV label generation.
- We conduct extensive experiments on the nuScenes benchmark and show that our label-guided distillation achieves significant improvements in camera-only detection without incurring additional costs in the inference.

II. RELATED WORKS

A. CAMERA-ONLY 3D OBJECT DETECTION

Camera-only 3D object detection has garnered increasing interest due to its cost-efficiency and ease of deployment in autonomous driving systems [18], [19], [20], [21]. Unlike

LiDAR-based approaches, which directly capture geometric structure and depth information, camera-based methods rely on inferring spatial cues from RGB images, which are inherently ambiguous and less precise. Early methods such as FCOS3D [22] and MonoDLE [23] attempted to regress 3D boxes from monocular images, but suffered from poor localization accuracy.

To alleviate these limitations, recent works adopt Bird's-Eye View (BEV) representations to align multi-view image features spatially. BEVDet and its extensions [1], [2], [24] project image features into BEV using learned depth priors and calibrated geometry, achieving significant performance gains. BEVDepth [1] further improves detection by leveraging depth supervision, while BEVFormer [2] introduces temporal modeling for feature aggregation. However, these methods still suffer from depth ambiguity and fail to match the precision of LiDAR-based detectors, especially in long-range and occluded scenarios.

B. KNOWLEDGE DISTILLATION FOR 3D OBJECT DETECTION

Knowledge Distillation (KD) is a widely adopted paradigm for transferring knowledge from a high-capacity teacher network to a lightweight student model [4]. In 3D detection, KD is often applied in a cross-modal fashion, where a camera-based detector learns from a LiDAR-based teacher [6], [7], [25], [26]. BEVDistill [26] proposes feature-level and response-level distillation from LiDAR BEV features to camera BEV features, while X3KD [25] incorporates cross-modality guidance using teacher-generated anchors. TiG-BEV [7] enhances this approach with geometric-aware attention. Despite their success, these methods introduce significant engineering complexity and dependency on LiDAR data, which is costly and often unavailable in real-world deployments [8].

Other works like 2DPASS [14] and Drive&Segment [13] explore auxiliary modalities for distillation, such as 2D semantic priors or multi-modal fusion. However, the requirement for additional sensors or cross-modal alignment remains a practical bottleneck. Moreover, fine-tuning large pretrained teachers and balancing multiple losses often demands considerable hyperparameter tuning [9], [10], [11].

Beyond cross-modal KD, several recent works [27], [28], [29] explore voxel-aware or anchor/region-aware KD strategies. Voxel-aware KD typically relies on LiDAR or radar voxel features as privileged teacher signals, while anchor or region-aware KD distills knowledge from detection head outputs at the anchor or proposal level. Although effective, these methods still depend on pretrained or sensor-based teachers, thereby inheriting the complexity and limitations of teacher-dependent frameworks.

In contrast to prior works, our method avoids external sensors and pretrained teachers altogether. Instead, we distill geometric and semantic knowledge directly from noise-free ground-truth 3D annotations provided by the benchmark

nuScenes dataset [16], by learning dense BEV representations through an autoencoder trained purely on label maps.

C. LABEL-GUIDED DISTILLATION AND SUPERVISION

Label-guided distillation offers a compelling alternative by leveraging ground-truth annotations as a source of intermediate supervision. LabelEnc [30] proposes a two-stage training strategy in which a label encoder is learned to approximate the inverse of the detection head, generating semantic embeddings to guide backbone learning. This approach improves generalization and training efficiency, enabling model-agnostic reuse.

LGD [31] extends label-guided distillation to self-distillation, incorporating intra-object relations via label-appearance encoders and relational adapters. Similarly, LG3D [32] uses point-level label clouds and annotation encoders to enrich object-level features in 3D detectors through auxiliary supervision. Although effective, these approaches typically rely on manually designed encoders and are often applied to point cloud-based networks.

Complementary to these, LAD [8] explores label-guided distillation in 2D semantic segmentation by introducing noise to labels as privileged input without pretrained models or extra modalities.

LabelDistill [15] introduces a cross-modal distillation framework where a label encoder is trained to approximate the inverse mapping of a pretrained LiDAR detection head. The resulting BEV label embeddings, fused with LiDAR features, serve as distillation targets for the camera-based network. A feature-partitioning strategy divides the BEV representation into LiDAR-guided, label-guided, and self-learned components to balance semantic and modality-specific information.

Inspired from previous works, our work builds upon the label-guided paradigm but eliminates the reliance on external sensor sources such as LiDAR. Instead, we train a Label Encoder directly from 3D bounding box annotations using an autoencoding objective, generating dense 2D BEV maps as a supervision signal without relying on voxel-level features or anchor-level outputs from pretrained or sensor-based teacher networks. We further enhance the label-derived features using a Gating Fusion Module that adaptively balances semantic and geometric cues, and apply a Class-Aware Rendering strategy that ensures rare or occluded object classes are preserved during label rasterization. These contributions result in a scalable and modality-agnostic distillation pipeline that improves camera-only 3D object detection without incurring any inference-time overhead.

III. PROPOSED METHOD

Our proposed framework consists of two stages: (1) as shown in Fig. 2, a label encoding stage that learns to compress structured ground-truth annotations into dense semantic BEV representations; and (2) as shown in Fig. 3, a label-guided distillation stage where the pretrained label encoder provides auxiliary supervision to a camera-only detection network.

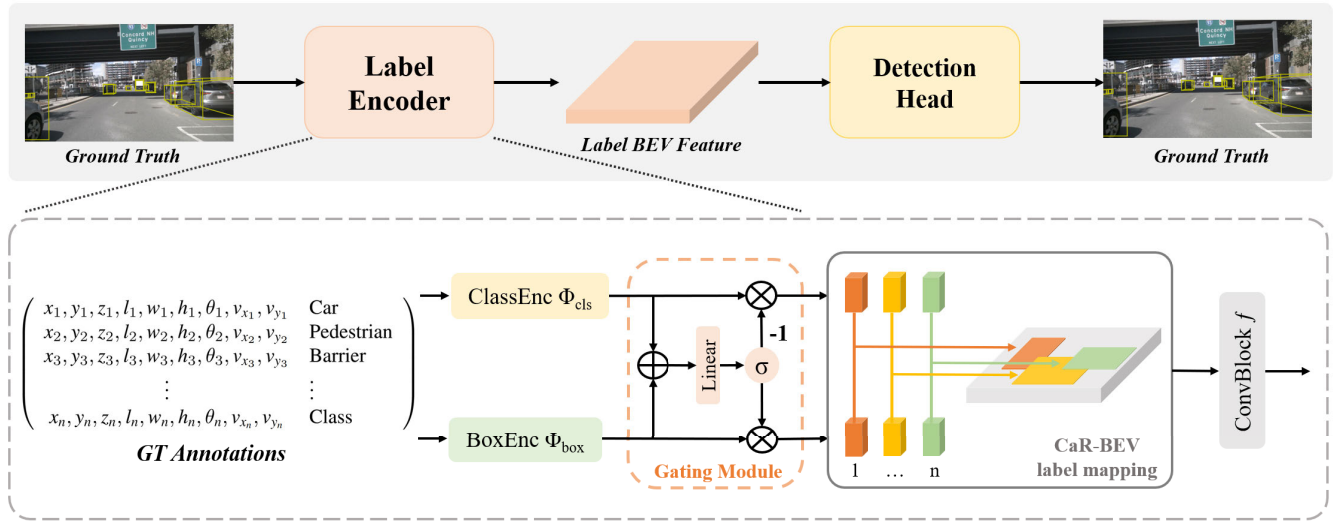


FIGURE 2. Architecture of the label encoder in the autoencoder training setup. The label encoder is jointly trained with a detection head to reconstruct object-level heatmaps and box attributes from 3D ground-truth annotations. Each object's semantic class and bounding box attributes are first embedded using two separate MLPs. A Gating Module adaptively fuses semantic and geometric features via element-wise modulation. The fused embeddings are then spatially projected into the BEV plane using the CaR-BEV label mapping strategy, which applies class-aware priority rules to resolve overlaps. The convolutional block f further refines the rasterized BEV features into the final dense representation F^{label} . Once training is completed, the detection head is discarded and the label encoder is ready to use for distillation.

A. LABEL ENCODER MODULE

As illustrated in Fig. 2, our proposed framework begins by training a label encoder jointly with a detection head in an autoencoding setup. The encoder learns to generate dense BEV features from 3D object annotations. These features are later used to guide the student model via label-guided distillation. Unlike prior works [15] that learn inverse mappings from complex sensor-based detection heads, our method builds a direct forward encoder trained entirely from label-derived supervision, without requiring LiDAR inputs or pretrained teacher dependencies. Once training is completed, the detection head is discarded.

Let the ground-truth annotation set for a scene be denoted as $y = \{(c_n, b_n)\}_{n=1}^N$, where $c_n \in \{0, 1\}^m$ is a one-hot encoded class vector over m semantic categories, and $b_n \in \mathbb{R}^z$ is a 3D bounding box attribute vector containing z attributes such as 3D location, size, orientation, and velocity:

$$b_n = (x_n, y_n, z_n, w_n, l_n, h_n, \theta_n, v_{x_n}, v_{y_n}). \quad (1)$$

Given the structured and noise-free nature of ground-truth annotations, we embed the semantic and geometric information via two separate MLPs:

$$f_n^{\text{cls}} = \Phi_{\text{cls}}(c_n), \quad f_n^{\text{box}} = \Phi_{\text{box}}(b_n). \quad (2)$$

The final fused representation for each object is obtained using a learnable gating mechanism (detailed below), producing $f_n \in \mathbb{R}^C$.

The embeddings f_n are rasterized into the BEV plane using a spatial projection operator $q(\cdot)$ based on the bounding box geometry:

$$F_{\text{raw}} = q\left(\{(f_n, b_n)\}_{n=1}^N\right) \in \mathbb{R}^{H \times W \times C} \quad (3)$$

We then refine F_{raw} using a convolutional encoder $f(\cdot)$ composed of stacked Conv-BN-ReLU blocks to produce the final BEV label feature map:

$$F^{\text{label}} = f(F_{\text{raw}}) \in \mathbb{R}^{H \times W \times C} \quad (4)$$

The label encoder is trained to reconstruct object-level heatmaps and regression targets from the label BEV features F^{label} . Following [33], [34], we adopt the same training losses from [1]:

$$\mathcal{L}_{\text{enc}} = \mathcal{L}_{\text{cls}}(\hat{c}, c^{\text{gt}}) + \mathcal{L}_{\text{bbox}}(\hat{b}, b^{\text{gt}}), \quad (5)$$

where $\mathcal{L}_{\text{cls}}$ is a Gaussian focal loss applied to the object heatmap, and $\mathcal{L}_{\text{bbox}}$ is an L1 loss computed on the ground-truth box attributes, including center offset, height, size, orientation, and velocity. $\hat{c}$ and $\hat{b}$ denote the class heatmap and bounding box predictions that are generated by encoding the corresponding ground truth c^{gt} and b^{gt} through the autoencoding pipeline.

1) GATING FUSION MODULE

To better encode semantic and geometric cues, we introduce a Gating Fusion Module that adaptively blends f_n^{cls} and f_n^{box} through a learnable gate. Specifically, the module first concatenates both features:

$$f_n^{\text{cat}} = [f_n^{\text{cls}}, f_n^{\text{box}}] \in \mathbb{R}^{2C}, \quad (6)$$

and generates a gating weight via:

$$\gamma_n = \sigma(W_g f_n^{\text{cat}} + b_g) \in (0, 1)^C, \quad (7)$$

where $\sigma(\cdot)$ is the sigmoid function and W_g, b_g are learnable parameters.

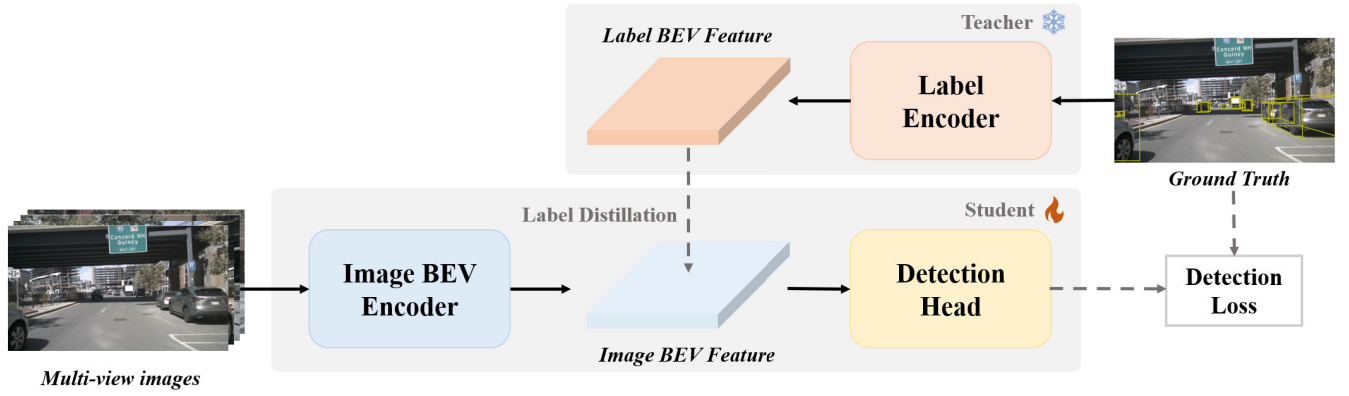


FIGURE 3. Overall training pipeline of the label-guided distillation framework. During training, a camera-only student model [1] receives multi-view images and generates image-based BEV features via an Image BEV Encoder. Through label distillation, the Label Encoder provides noise-free supervision from 3D ground-truth annotations. Finally, the Detection Head predicts and is trained using detection loss computed against ground truth. Note that the label encoder is only used during training and removed at inference time.

The final fused feature is then computed as:

$$f_n = \gamma_n \odot f_n^{\text{box}} + (1 - \gamma_n) \odot f_n^{\text{cls}}, \quad (8)$$

where $\odot$ denotes element-wise multiplication.

This design enables the encoder to dynamically adjust the contribution of geometric and semantic components per object for improving feature expressiveness and spatial reasoning.

2) CLASS-AWARE RENDERING

During the rasterization step $q(\cdot)$, overlapping objects may contend for the same BEV grid locations. To prevent dominant object classes (e.g., cars) from suppressing smaller or rare classes (e.g., bicycles or traffic cones), as observed in [17], we introduce a class-priority rendering scheme, CaR-BEV label mapping.

Each class c_k is assigned a continuous priority score $p(c_k)$ based on its training-set frequency f_{c_k} , following a class-balanced formulation:

$$p(c_k) = \frac{1}{\log(1 + f_{c_k})}, \quad f_{c_k} > 0 \quad (9)$$

When rendering each object n , we update a grid location (i, j) only if the priority of the current content is lower than that of the incoming object. Let $p_{\text{curr}}(i, j)$ denote the priority of the feature currently occupying location (i, j) , and $p(c_n)$ be the priority of the new object. Then:

$$\text{if } p(c_n) > p_{\text{curr}}(i, j), \quad F_{\text{raw}}[i, j, :] \leftarrow f_n \quad (10)$$

This frequency-aware overwrite rule ensures that features corresponding to rare or underrepresented categories are retained, enhancing the semantic richness and diversity of the label-derived BEV representation.

B. LABEL-GUIDED KNOWLEDGE DISTILLATION

After pretraining the label encoder, we freeze its parameters and use its output F^{label} as an auxiliary supervision signal for a student network trained on multi-view images.

As illustrated in Fig. 3, the student model extracts a BEV representation F^{img} through its image backbone and transforms it from the perspective view to the BEV using the view transformation module adopted from [2], [35]. The student is optimized using both standard detection loss and feature-level distillation loss from the label encoder.

We adopt the feature-level label distillation loss, formulated as:

$$\mathcal{L}_{\text{feat}} = \frac{1}{N_p} \sum_{i,j} \mathcal{M}_{ij} \left\| \alpha \left(F_{ij}^{\text{img}} \right) - F_{ij}^{\text{label}} \right\|_2^2, \quad (11)$$

where $F_{ij}^{\text{img}} \in \mathbb{R}^C$ and $F_{ij}^{\text{label}} \in \mathbb{R}^C$ denote the BEV features at spatial location (i, j) from the student and the frozen label encoder, respectively. The mask $\mathcal{M} \in \{0, 1\}^{H \times W}$ focuses the distillation only on foreground regions, which are derived from the Gaussian heatmap of ground-truth object centers.

The function $\alpha(\cdot)$ is an adaptation module consisting of convolutional layers that align the channel dimensions of F^{img} to match F^{label} . $N_p = \sum_{i,j} \mathcal{M}_{ij}$ is the number of foreground pixels used for normalization.

Additionally, we adopt the feature partitioning strategy introduced in [15] for consistency and fair comparison, in order to preserve the inherent semantic capacity of image features while allowing supervision from the label encoder. This approach divides F^{img} along the channel dimension into disjoint groups, allowing certain partitions to be exclusively trained via detection loss while others receive supervision from the label encoder.

The detection loss follows the same formulation used during the autoencoder training:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{bbox}}, \quad (12)$$

The total training loss for camera network becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda_1 \mathcal{L}_{\text{feat}} + \lambda_2 \mathcal{L}_{\text{depth}}, \quad (13)$$

where λ_1 and λ_2 are balancing coefficients for the feature distillation and depth supervision [1] terms, respectively.

This composite loss encourages the student to align with the spatially grounded, label-derived BEV features while benefiting from explicit depth learning to enhance geometric reasoning from multi-view inputs.

IV. EXPERIMENTS

A. DATASET AND METRICS

We evaluated our approach, LGD-BEV, on the nuScenes dataset [16], a widely used large-scale outdoor benchmark for 3D perception in autonomous driving. The dataset consists of 1,000 driving scenes, each approximately 20 seconds long with annotations provided at 2 Hz. Each scene includes sensor data from six cameras, one LiDAR, and five radars. This dataset is officially divided into 700 training, 150 validation, and 150 test scenes across 10 object categories.

For evaluation, we followed two official primary metrics: mean Average Precision (mAP) and the nuScenes Detection Score (NDS). For detection quality and accuracy, we report five True Positive (TP) metrics as well: Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE). The formula for NDS is as follows,

$$NDS = \frac{1}{10} \left[5 \text{mAP} + \sum_{\text{mTP} \in \mathcal{TP}} (1 - \min(1, \text{mTP})) \right] \quad (14)$$

, where $\mathcal{TP}$ is the set of the five mean True Positive metrics.

B. IMPLEMENTATION DETAILS

Our framework is built upon the BEVDepth [1] codebase. All experiments were conducted on 4 NVIDIA A40 GPUs, each equipped with 46 GB of memory. We used a total batch size of 16 without the CBGS [17] strategy.

Training is conducted in two stages: (1) we first train the label encoder together with a detection head in an autoencoding setup for 12 epochs using the AdamW [47] optimizer with a learning rate of $1e-3$, and (2) we then freeze the label encoder and train the student network for 24 epochs using a learning rate of $4e-4$.

We adopt the image and BEV data augmentation strategies introduced in [1] and follow the label configuration defined in [15]. Unless stated otherwise, we use a ResNet-50 backbone for the student network. The input image resolution is resized to 256×704 , and 5 temporal frames (1 key frame and 4 preceding frames) are used during training. Note that the computational cost in terms of GFLOPs and FPS is determined solely by the input resolution and number of frames, and LGD-BEV introduces no additional inference-time modules beyond BEVDepth [1].

C. MAIN RESULTS

We compare our proposed method, LGD-BEV, with various state-of-the-art BEV-based 3D object detection approaches on the nuScenes [16] validation set. As shown in Table 1,

our model demonstrates strong performance across multiple evaluation metrics, achieving 40.2 mAP and 51.5 NDS.

Despite not using CBGS [17], an augmentation technique that expands the training data per epoch by approximately 4.5 times, or larger input resolutions, our method matches or surpasses more complex models such as StreamPETR [42] (40.2 mAP, 50.5 NDS) and SOLOFusion [41] (40.6 mAP, 49.7 NDS), while maintaining a simpler architecture and training pipeline, under a ResNet-50 backbone and input resolution of 256×704 .

We also benchmark LGD-BEV against other LiDAR-guided cross-modal knowledge distillation methods in Table 2. Performance improvements over the baselines (Δ) are reported for both mAP and NDS. Without relying on external LiDAR data during training, LGD-BEV surpasses most existing distillation methods and achieves comparable performance to top-performing CBGS-trained models such as DistillBEV [6] (40.3 mAP, 51.0 NDS), further validating the effectiveness of label-guided distillation.

D. ABLATION STUDIES

1) TRAINING LABEL ENCODER

We train the Label Encoder in an autoencoding setup to directly learn BEV feature representations from structured 3D ground-truth annotations. Unlike prior label guidance methods that rely on student features or LiDAR-based detection heads for supervision, our approach uses only label-derived inputs and supervision.

To evaluate the effectiveness of our training, we compare with prior label-encoder-based baselines, as shown in Table 3. LabelEnc [30] employs an autoencoding setup but incorporates an additional encoding mechanism that depends on the student network's features during training. LabelDistill [15] constructs label embeddings by approximating the inverse mapping of a pretrained LiDAR detection head, introducing a dependency on LiDAR during training. In contrast, our approach eliminates the dependency on pretrained LiDAR detectors and instead directly reconstructs object heatmaps and box attributes purely from label semantics and geometry.

As a baseline, we implement a simple autoencoder in which both the encoder and decoder are trained from scratch using label annotations as both inputs and targets. Building upon this, we further introduce two architectural enhancements to the encoder: (1) Gating Fusion Module (GFM) that dynamically blends semantic class and geometric box features and (2) Class-Aware Rendering (CaR) that prevents dominant classes from overwriting rare ones during rasterization. As illustrated in Table 3, each of these components contributes measurable performance gains. The full model with both CaR and GFM achieves the best results, validating the benefit of structured supervision and adaptive feature fusion.

This label-driven training strategy produces high-quality BEV features that are later used to guide the student network through feature-level distillation. Because no external

TABLE 1. Comparison of BEV-based 3D object detection methods on the nuScenes validation set. †: trained with CBGS. Best is in bold.

Set	Method	Backbone	Size	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
Validation	BEVDet4D [36]	ResNet50	256×704	32.3	45.3	0.674	0.272	0.503	0.429	0.208
	BEVDepth [1]	ResNet50	256×704	33.3	44.1	0.683	0.276	0.545	0.526	0.226
	BEVStereo [3]	ResNet50	256×704	34.1	44.9	0.659	0.276	0.579	0.503	0.216
	PETR v2 [37]	ResNet50	256×704	34.9	45.6	0.700	0.275	0.580	0.437	0.187
	FB-BEV [38]	ResNet50	256×704	35.0	47.9	0.642	0.275	0.459	0.391	0.193
	P2D [39]	ResNet50	256×704	37.4	48.8	0.631	0.272	0.508	0.384	0.212
	BEVFormer v2 [†] [40]	ResNet50	640×1600	38.9	49.6	0.679	0.276	0.417	0.403	0.189
	SOLOFusion [41]	ResNet50	256×704	40.6	49.7	0.609	0.284	0.650	0.315	0.204
	StreamPETR [42]	ResNet50	256×704	40.2	50.5	0.660	0.273	0.470	0.316	0.210
	LGD-BEV (Ours)	ResNet50	256×704	40.2	51.5	0.598	0.271	0.408	0.373	0.214
Validation	DETR3D [†] [43]	ResNet101	900×1600	34.9	43.4	0.716	0.268	0.379	0.842	0.200
	BEVDepth [1]	ResNet101	512×1408	40.6	49.0	0.626	0.278	0.513	0.489	0.226
	BEVFormer [2]	ResNet101	900×1600	41.6	51.7	0.673	0.274	0.372	0.394	0.198
	PolarFormer [44]	ResNet101	900×1600	43.2	52.8	0.648	0.270	0.348	0.409	0.201
	P2D [39]	ResNet101	512×1408	43.3	52.8	0.619	0.265	0.432	0.364	0.211
	LGD-BEV (Ours)	ResNet101	512×1408	43.2	53.9	0.595	0.265	0.327	0.385	0.201

TABLE 2. Performance comparison of LiDAR-guided cross-modal knowledge distillation methods. †: trained with CBGS.

Model	Baseline	Image Size	mAP (Δ)	NDS (Δ)
UniDistill [45]	BEVDet	704×256	29.6 (3.2)	39.3 (3.2)
BEVSimDet [27]	BEVFusion	704×256	37.3 (1.7)	43.8 (2.6)
BEVDistill [5]	BEVDepth	704×256	33.0 (1.3)	45.2 (1.2)
TiG-BEV [7]	BEVDepth	704×256	36.6 (3.7)	46.1 (3.0)
X ³ KD [†] [46]	BEVDepth	704×256	39.0 (3.1)	50.5 (3.3)
DistillBEV [†] [6]	BEVDepth	704×256	40.3 (3.9)	51.0 (2.6)
LGD-BEV	BEVDepth	704×256	40.2 (3.4)	51.5 (3.2)

TABLE 3. Evaluation of different label encoder training methods. CaR: Class-Aware Rendering, GFM: Gating Fusion Module.

Method	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓
LabelEnc [30]	90.22	86.33	0.209	0.265	0.105
LabelDistill [15]	93.43	89.01	0.202	0.142	0.105
Label Encoder	93.47	89.04	0.192	0.161	0.069
+ GFM	93.59	89.39	0.192	0.169	0.066
+ GFM + CaR	93.70	90.00	0.190	0.131	0.061

TABLE 4. Ablation study on student performance under different Label Encoder configurations.

Label Encoder		Student Model				
GFM	CaR	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓
		39.7	50.4	0.623	0.270	0.461
✓		39.8	51.0	0.604	0.268	0.425
✓	✓	40.2	51.5	0.598	0.271	0.408

modalities or pretrained networks are involved, the resulting label encoder serves as a lightweight, interpretable, and modality-agnostic teacher model.

2) COMPARISON WITH STUDENT MODEL

As shown in Table 4, we evaluate the student trained with BEV features generated from the label encoder. Without any enhancements, the student model achieves 39.7 mAP and

TABLE 5. Performance comparison across different channel splits between label and image BEV features.

Channel Ratio	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓
F_{label} F_{image}					
3 1	38.9	50.4	0.621	0.273	0.423
2 2	40.0	51.2	0.610	0.269	0.412
1 3	40.2	51.5	0.598	0.271	0.408

TABLE 6. Detection performance at near ($\leq 30\text{m}$) and far ($\geq 30\text{m}$) ranges with different Label Encoder configurations.

Distance	GFM	CaR	mATE ↓	mASE ↓	mAOE ↓
$\leq 30\text{m}$	✓		0.548	0.285	0.421
	✓	✓	0.535	0.269	0.392
$\geq 30\text{m}$	✓		0.917	0.296	0.582
	✓	✓	0.908	0.287	0.558
			0.912	0.283	0.520

50.4 NDS. Incorporating GFM alone boosts performance to 39.8 (0.1+) mAP and 51.0 (0.6+) NDS.

When both GFM and CaR are employed, the student model achieves its best results with 40.2 mAP and 51.5 NDS, surpassing the versions with only CaR or no enhancements. This indicates that the student network not only successfully learns from the label encoder's supervision but can even generalize beyond it.

3) FEATURE PARTITIONS

To study the effect of label-derived features, we adopt the feature partitioning strategy from [15] with different channel ratios. In this setup, only a portion of the student BEV feature map receives distillation loss from the label encoder. This allows us to examine how partial label supervision affects performance and whether keeping some channels for image-specific features helps maintain semantic diversity.



FIGURE 4. Visualization of detection results on the nuScenes dataset, comparing the BEVDepth [1] baseline with our proposed method. Predicted 3D bounding boxes are shown in the camera views using class-specific colors: blue for car, gray for bicycle, yellow for pedestrian, and pink for motorcycle. On the BEV maps, predicted boxes are shown in green and ground-truth boxes in red. Numbered circles indicate key differences.

Table 5 shows the performance under different channel ratios between label and image features. As the number of image channels increases and label channels decrease, mAP and NDS consistently improve. This indicates that image features contribute more to the final performance, while label features used in smaller amounts for guidance.

4) OBJECT DISTANCES

We evaluate detection performance at different distances by dividing objects into near ($\leq 30\text{m}$) and far ($\geq 30\text{m}$) groups based on their distance to the ego vehicle. The distance is calculated from the 3D center of each object to the ego pose at the time of annotation.

As shown in Table 6, using both the GFM and CaR improves results for both near and far objects. For near objects, the combined setup gives better localization and orientation, especially in crowded areas. For far objects, GFM improves overall spatial features, while CaR reduces orientation errors by keeping class-specific features more consistent.

5) QUALITATIVE RESULTS

We present qualitative comparisons against the BEVDepth baseline on the nuScenes dataset in Fig. 4. Red boxes

denote ground-truth and green boxes denote predictions from BEVDepth and our method on right, respectively.

As illustrated, our approach produces more accurate predictions in urban scenes compared to baseline. Notably, in the numbered region (1), baseline produces a false positive. In region (2), baseline detects overlapping or cluttered boxes, including a motorcycle (pink) and a pedestrian (yellow).

V. CONCLUSION

In this paper, we introduced LGD-BEV, a label-guided distillation framework for camera-only 3D object detection. Unlike conventional cross-modal distillation methods that rely on external sensor modalities or heavy teacher networks, LGD-BEV eliminates such dependencies by using a lightweight Label Encoder trained solely on 3D ground-truth annotations. This encoder provides dense and semantically rich BEV representations that serve as supervision signals for training the student network.

To enhance the supervision quality and spatial expressiveness of the label features, we proposed two key components: (1) Gating Fusion Module that adaptively balances semantic class features and geometric box features during label encoding, and (2) Class-Aware Rendering strategy that preserves

rare or small object classes during BEV rasterization through frequency-aware overwriting.

Extensive experiments on the nuScenes [16] benchmark demonstrate that LGD-BEV outperforms or matches the performance of more complex distillation methods, including those using CBGS [17] and additional sensor data, all while maintaining a simple, scalable, and sensor-agnostic training pipeline. Our ablation studies further validate the contribution of each module and the effectiveness of our approach.

However, its effectiveness inherently depends on the reliability of ground-truth annotations. When the dataset contains noisy or low-quality labels, performance may degrade accordingly. Finally, this study does not examine the adaptability of LGD-BEV to semi-supervised or synthetic label scenarios, which remains an important direction for future work.

REFERENCES

- [1] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," 2022, *arXiv:2206.10092*.
- [2] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," 2022, *arXiv:2203.17270*.
- [3] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li, "BEVStereo: Enhancing depth estimation in multi-view 3D object detection with dynamic temporal stereo," 2022, *arXiv:2209.10248*.
- [4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
- [5] T. Jiao, Y. Chen, Z. Zhang, C. Guo, and J. Song, "MMDistill: Multi-modal BEV distillation framework for multi-view 3D object detection," *Comput., Mater. Continua*, vol. 81, no. 3, pp. 4307–4325, 2024.
- [6] Z. Wang, D. Li, C. Luo, C. Xie, and X. Yang, "DistillBEV: Boosting multi-camera 3D object detection with cross-modal knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 8603–8612.
- [7] P. Huang, L. Liu, R. Zhang, S. Zhang, X. Xu, B. Wang, and G. Liu, "TiG-BEV: Multi-view bev 3D object detection via target inner-geometry learning," 2022, *arXiv:2212.13979*.
- [8] S. Qiu, J. Chen, X. Li, R. Wan, X. Xue, and J. Pu, "Make a strong teacher with label assistance: A novel knowledge distillation approach for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 371–388.
- [9] D. Chen, J.-P. Mei, Y. Zhang, C. Wang, Y. Feng, and C. Chen, "Cross-layer distillation with semantic calibration," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 7028–7036.
- [10] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 578–587.
- [11] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5291–5300.
- [12] Z. Liu, X. Qi, and C.-W. Fu, "3D-to-2D distillation for indoor scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4462–4472.
- [13] A. Vobecky, D. Hurrych, O. Siméoni, S. Gidaris, A. Bursuc, P. Pérez, and J. Sivic, "Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 478–495.
- [14] X. Yan, J. Gao, C. Zheng, C. Zheng, R. Zhang, S. Cui, and Z. Li, "2DPASS: 2D priors assisted semantic segmentation on LiDAR point clouds," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 677–695.
- [15] S. Kim, Y.-S. Kim, S. Hwang, H. Jeong, and D. Kum, "LabelDistill: Label-guided cross-modal knowledge distillation for camera-based 3D object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 19–37.
- [16] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. Erin Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," 2019, *arXiv:1903.11027*.
- [17] B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu, "Class-balanced grouping and sampling for point cloud 3D object detection," 2019, *arXiv:1908.09492*.
- [18] G. Brazil and X. Liu, "M3D-RPN: Monocular 3D region proposal network for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9286–9295.
- [19] Z. Liu, Z. Wu, and R. Tóth, "SMOKE: Single-stage monocular 3D object detection via keypoint estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 4289–4298.
- [20] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecká, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5632–5640.
- [21] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, "Is pseudo-LiDAR needed for monocular 3D object detection?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3122–3132.
- [22] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 913–922.
- [23] X. Ma, Y. Zhang, D. Xu, D. Zhou, S. Yi, H. Li, and W. Ouyang, "Delving into localization errors for monocular 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4719–4728.
- [24] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance multi-camera 3D object detection in bird-eye-view," 2022, *arXiv:2112.11790*.
- [25] J. Yang, H. Xu, X. Liang, L. Zhang, and L. Lin, "X3-KD: Improving multi-sensor 3D object detection via cross-modal cross-level cross-task knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9966–9976.
- [26] Y. Huang, Q. Li, Z. Cao, X. Ma, H. Xu, and D. Lin, "Bevdistill: Cross-modal bev distillation for multi-view 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16345–16354.
- [27] H. Zhao, Q. Zhang, S. Zhao, Z. Chen, J. Zhang, and D. Tao, "Simdistill: Simulated multi-modal distillation for bev 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 7460–7468.
- [28] Y. Li, Y. Wang, L. Zhang, H. Wang, X. Sun, X. Huang, and C. R. Qi, "Distillbev: Boosting 3D object detection with cross-modal knowledge distillation for bev representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9997–10006.
- [29] J. Li, M. Lu, J. Liu, Y. Guo, Y. Du, L. Du, and S. Zhang, "BEV-LGKD: A unified LiDAR-guided knowledge distillation framework for multi-view BEV 3D object detection," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 2489–2498, Jan. 2024.
- [30] M. Hao, Y. Liu, X. Zhang, and J. Sun, "LabelEnc: A new intermediate supervision method for object detection," 2020, *arXiv:2007.03282*.
- [31] P. Zhang, Z. Kang, T. Yang, X. Zhang, N. Zheng, and J. Sun, "LGD: Label-guided self-distillation for object detection," 2021, *arXiv:2109.11496*.
- [32] Y. Huang, X. Liu, Y. Zhu, Z. Xu, C. Shen, Z. Che, G. Zhang, Y. Peng, F. Feng, and J. Tang, "Label-guided auxiliary training improves 3D object detector," in *Proc. Comput. Vision—ECCV: 17th Eur. Conf.*, Oct. 2022, pp. 684–700.
- [33] T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11779–11788.
- [34] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
- [35] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 194–210.
- [36] J. Huang and G. Huang, "BEVDet4D: Exploit temporal cues in multi-camera 3D object detection," 2022, *arXiv:2203.17054*.
- [37] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "PETRv2: A unified framework for 3D perception from multi-camera images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3239–3249.
- [38] Z. Li, Z. Yu, W. Wang, A. Anandkumar, T. Lu, and J. M. Alvarez, "FB-BEV: BEV representation from forward-backward view transformations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6919–6928.
- [39] S. Kim, Y. Kim, I.-J. Lee, and D. Kum, "Predict to detect: Prediction-guided 3D object detection using sequential images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 18011–18020.

- [40] C. Yang, Y. Chen, H. Tian, C. Tao, X. Zhu, Z. Zhang, G. Huang, H. Li, Y. Qiao, L. Lu, J. Zhou, and J. Dai, "BEVFormer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17830–17839.
- [41] J. Park, C. Xu, S. Yang, K. Keutzer, K. Kitani, M. Tomizuka, and W. Zhan, "Time will tell: New outlooks and a baseline for temporal multi-view 3D object detection," 2022, *arXiv:2210.02443*.
- [42] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3D object detection," 2023, *arXiv:2303.11926*.
- [43] Y. Wang, V. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, 2021, pp. 180–191.
- [44] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, and Y.-G. Jiang, "Polarformer: Multi-camera 3D object detection with polar transformers," in *Proc. AAAI*, 2023, pp. 1–17.
- [45] S. Zhou, W. Liu, C. Hu, S. Zhou, and C. Ma, "UniDistill: A universal cross-modality knowledge distillation framework for 3D object detection in bird's-eye view," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5116–5125.
- [46] M. Klingner, S. Borse, V. R. Kumar, B. Rezaei, V. Narayanan, S. Yogamani, and F. Porikli, "X3KD: Knowledge distillation across modalities, tasks and stages for multi-camera 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13343–13353.
- [47] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019, *arXiv:1711.05101*.



JI-YONG LEE received the B.S. degree in electrical and computer engineering from The University of Texas at Austin, in 2021. He is currently pursuing the master's degree in artificial intelligence. He is a Research Associate with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. His research interests include machine learning, knowledge distillation, and multi-modal sensor fusion.



MINHO PARK received the B.S. and M.S. degrees from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2018 and 2020, respectively. He is currently a Researcher with the Visual Intelligence Research Section, Electronics and Telecommunications Research Institute (ETRI). His research interests include generative models, multi-modal representation, and image/video analysis.



YOUNGJOO JO received the M.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2016. He is currently a Researcher with the Electronics and Telecommunications Research Institute (ETRI), Daejeon. His current research interests include machine learning, computer vision, and generative models.



DONG-OH KANG received the B.S. degree in electronic engineering from Yonsei University, Korea, in 1994, and the M.S. and Ph.D. degrees in electronic engineering from Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 1996 and 2001, respectively. Since 2001, he has been with the Electronics and Telecommunications Research Institute (ETRI). He is currently a Principal Researcher with the Visual Intelligence Research Section. His research interests include multimodal knowledge representation, knowledge graph completion, and self-growing intelligent agents.

...