Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.080595

ARTICLE

DA-T3D: Distribution-Aware Cross-Modal Distillation Framework for Temporal 3D Object Detection

Tianzhe Jiao, Yuming Chen, Xiaoyue Feng, Chaopeng Guo and Jie Song*

Software College, Northeastern University, Shenyang, China

*Corresponding Author: Jie Song. Email: songjie@mail.neu.edu.cn

Received: 12 February 2026; Accepted: 23 March 2026; Published: 27 April 2026

ABSTRACT: Knowledge distillation bridges the performance gap between camera-based and LiDAR-based 3D detectors by leveraging the precise geometric information from LiDAR. However, cross-modal knowledge transfer remains challenging due to the inherent modality heterogeneity between LiDAR and camera data, which often leads to instability during training. In this work, we find that these instabilities are closely related to distribution mismatch in the cross-modal feature space and noisy teacher signals. To address this issue, we propose a novel distribution-aware cross-modal distillation framework, named DA-T3D. Specifically, we first explicitly model the LiDAR teacher's Bird's-Eye-View (BEV) feature distribution and use the learned distribution as a statistical prior to guide the student features toward high-density and geometrically stable regions in the teacher's BEV feature space. This ensures feature alignment in BEV space by constraining the student model's feature distribution to match that of the LiDAR teacher model within foreground regions. Next, we further introduce response-level distillation to directly transfer the teacher's prediction behavior to the student detection head, providing direct output-space supervision that complements feature distillation and effectively reduces modality-induced ambiguity, leading to more accurate and stable classification confidence and bounding-box regression. Furthermore, we perform temporal modeling on the distilled cross-modal features to produce fused BEV representations that capture more comprehensive scene context. Finally, we utilize the fused BEV features to generate 3D detection results. Through experiments, we validate the effectiveness and superiority of DA-T3D on the nuScenes dataset, achieving 46.7% mAP and 58.1% NDS.

KEYWORDS: 3D object detection; Bird's-Eye-View perception; cross-modal knowledge distillation; Dirichlet process Gaussian mixture model; temporal modeling

1 Introduction

3D object detection based on multi-view cameras is a fundamental yet challenging task in autonomous driving [1]. In real-world applications such as autonomous driving, accurate 3D object detection directly affects a vehicle's ability to perceive the surrounding environment and make safe driving decisions. However, compared with LiDAR-based methods, camera-only methods often suffer from ambiguous depth estimation and are more sensitive to illumination variations and occlusions, which typically result in degraded 3D localization accuracy and limited robustness. To narrow the performance gap with LiDAR-based methods, researchers have increasingly explored cross-modal knowledge distillation in recent years. Specifically, cross-modal distillation transfers geometric priors from complementary modalities such as LiDAR to a camera-based student, providing reliable 3D structural cues to improve the 3D detection performance [2].

However, the inherent data heterogeneity between LiDAR point clouds and camera images poses challenges for effective cross-modal distillation.

To alleviate the distillation challenges caused by the modality gap between LiDAR and cameras, existing methods typically map data from both modalities into a unified feature space to facilitate feature imitation [3]. Some studies project LiDAR points onto the image plane and perform distillation in 2D space [4]. However, such cross-modal transformations often lead to the loss of intrinsic features of the original data, which limits the student model's ability to learn effective information from the teacher. Consequently, another mainstream method maps both modalities into a unified BEV space [5], enabling the student model to align features with the teacher more directly, as shown in Fig. 1. These works commonly adopt point-wise aligned distillation, which allows fine-grained matching between BEV features from the two modalities. Nevertheless, background regions in BEV space often contain substantial task-irrelevant noise, which can divert the distillation process toward redundant background features and reduce the efficiency of learning key foreground features. To address this issue, Chen et al. [6] proposed a foreground-aware distillation method that has been widely adopted. By focusing knowledge transfer on foreground target regions in the scene, it enhances the model's ability to extract and transfer important features.

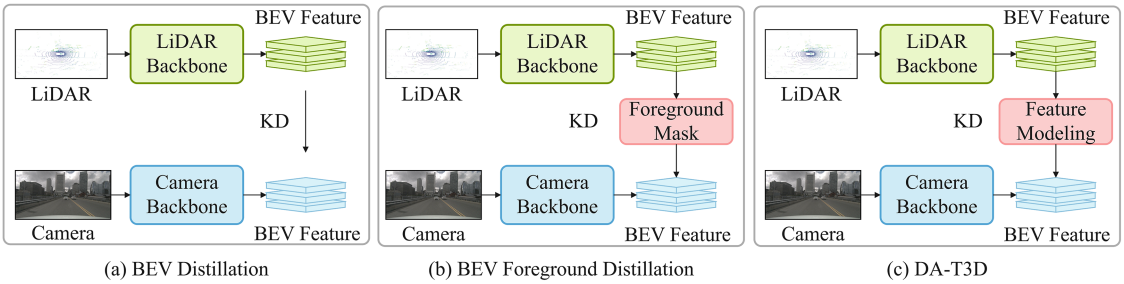

Figure 1: Cross-modal knowledge distillation frameworks.

Despite the promising progress of existing cross-modal distillation methods, the domain gap across modalities remains persists due to differences in imaging mechanisms and spatial resolution. In this context, adopting a point-wise aligned distillation scheme that enforces exact consistency between the BEV features of the two modalities may lead to noise amplification and overly restrictive constraints, thereby affecting the model's detection performance. Moreover, distillation typically depends on high-quality supervisory signals from the teacher model. However, the teacher's features may themselves contain noise and bias, for example, due to false positives, missed detections, or feature jitter. Such noise can be directly transferred to the student during distillation, leading to unstable supervision and reduced distillation effectiveness. Therefore, cross-modal knowledge distillation faces two core challenges: (1) due to inherent modality heterogeneity, using a simple point-to-point distillation method is suboptimal, and (2) the LiDAR teacher's features can be noisy, so naive imitation may introduce erroneous supervision.

In this work, we propose a novel distribution-aware cross-modal distillation framework, which is a carefully designed distribution-level cross-modal distillation strategy that effectively addresses the aforementioned challenges. Specifically, our method first models class-conditional feature distributions of the LiDAR teacher's BEV features. Then, using a distribution-consistency constraint, we encourage the student features to fall into the teacher's high-density and geometrically stable regions, as shown in Fig. 1c. By aligning features at the distribution level, this method effectively narrows the BEV representation gap between the two modalities. Meanwhile, the modeling process naturally suppresses a small number of outlier and noisy teacher features. Distribution-level distillation pulls the student toward aggregated mode centers rather than

individual noisy instances, thereby mitigating the adverse effects of teacher noise. In addition, to reduce interference from factors such as target occlusion and motion blur, we further apply lightweight temporal modeling to the distilled BEV features, improving training stability. The main contributions of this paper are as follows:

1. We propose a novel distribution-aware cross-modal distillation framework (DA-T3D) for 3D object detection, which enables distribution-level knowledge transfer from a LiDAR teacher to a camera-based student. In addition, we introduce response-level distillation to convey task-specific decision knowledge, further improving detection performance.
2. We propose a lightweight temporal fusion module that fuses features from two consecutive frames and introduces a gating mechanism to adaptively balance the contributions of the current and historical frames.
3. Through extensive experiments and ablation studies on the nuScenes benchmark, our framework demonstrates outstanding performance in 3D object detection. Our best model achieves 46.7% mAP and 58.1% NDS on nuScenes.

The remainder of this paper is organized as follows: Section 2 briefly reviews the related work. Section 3 introduces our proposed solutions in detail. Experimental settings and results, along with comparisons to baseline methods, are presented in Section 4 to validate the effectiveness of our approach. Finally, Section 5 presents the conclusion of this paper, summarizing the key contributions and discussing potential future directions.

2 Related Work

2.1 Multi-View 3D Object Detection

Multi-view 3D object detection aims to leverage surround-view camera images to align and fuse multi-view 2D features into a unified 3D space or bird's-eye-view (BEV) representation, thereby enabling 3D object localization and attribute regression. Existing methods mainly follow two paradigms: (1) explicitly constructing a dense BEV representation and then performing detection; and (2) adopting query-based or sparse 3D representations, where 3D queries directly aggregate information from multi-view features to regress 3D bounding boxes [7].

For explicit BEV construction, early studies achieved view transformation and feature fusion by predicting pixel-wise depth distributions (e.g., LSS [8]). Subsequent works have improved this pipeline along several directions, including depth estimation quality and temporal fusion. For example, BEVDepth introduces depth supervision [9], BEVFormer generates BEV features with spatiotemporal attention [10], and GeoBEV enhances geometric details via more efficient BEV sampling and structure-aware depth supervision [11]. In contrast, to avoid the computational overhead of dense BEV, query-based methods use 3D queries to interact with multi-view features. DETR3D samples features by projecting 3D reference points onto 2D views [12]. PETR and its variants strengthen spatial alignment with 3D positional embeddings [13], and Sparse4D aggregates multi-view and temporal information using 4D keypoints [14]. These methods have continually evolved to balance efficiency and accuracy, collectively advancing vision-only 3D detection. However, the performance of multi-view models heavily depends on the quality of depth estimation, lacks robustness to complex conditions such as illumination changes and adverse weather, and typically requires large amounts of accurately annotated data for supervised learning [9].

2.2 Multi-Modal 3D Object Detection

Multi-modal 3D object detection aims to fuse semantic and geometric information from sensors such as cameras, LiDAR, and radar to improve perception performance in complex scenarios. Existing methods can be categorized by fusion stage as early fusion, feature-level fusion, and late fusion. Mainstream directions include BEV-based unified representations, sparse query–based fusion, and unified 3D representations, enabling better cross-modal complementarity [3,15].

Specifically, early fusion injects image semantics directly into point clouds or voxels, as in PointPainting [16] and MVX-Net [17]. However, it is sensitive to calibration errors and point cloud sparsity. Subsequent works such as PPF-Net improve robustness via region-level semantic aggregation [18]. Feature-level fusion maps multi-modal features into a shared BEV space for interaction, with BEVFusion providing a lightweight fusion framework [19,20]. Late fusion performs cross-modal fusion after generating candidate boxes, as in MV3D [21] and CLOCs [22], but the degree of cross-modal interaction is limited. In addition, to improve efficiency and long-range performance, MV2DFusion adopts a sparse query–based fusion scheme, using object queries as carriers for cross-modal interaction [23]. To address sensor disparities, unified 3D representation methods such as FGU3R convert images into pseudo point clouds to enable fine-grained fusion [24]. Although multi-modal fusion methods can effectively mitigate inherent limitations of unimodal methods in depth estimation and robustness under adverse weather conditions [15,16], they face challenges in deployment cost and computational overhead introduced by multiple sensors.

2.3 Cross-Modal Knowledge Distillation for 3D Object Detection

Cross-modal knowledge distillation (CMKD) for 3D object detection aims to use a stronger, information-rich modality (e.g., LiDAR or multimodal fusion) during training to guide a weaker-modality detector (e.g., camera-only or radar-only). In this way, inference can rely solely on low-cost sensors, striking a balance between deployment efficiency and accuracy. Existing studies mainly focus on key issues such as modality representation gaps, spatial alignment, and noise in teacher-generated pseudo labels.

Early works such as MonoDistill [4] distill knowledge by projecting LiDAR features onto the image plane, improving spatial reasoning for monocular 3D detection. BEVDistill [6] and DistillBEV [25] further align image features with LiDAR teacher predictions in BEV space to enhance camera-based BEV detection. UniDistill [26] proposes a generic BEV-oriented CMKD framework that transfers knowledge at multiple levels, including features, predictions, and relations. To alleviate the high cost of 3D annotations, MonoLiG [27] and SCKD [28] combine CMKD with semi-supervised learning, using teacher-generated pseudo labels to train student models and suppressing noisy negative transfer via uncertainty weighting, feature distillation, and related techniques, thus moving CMKD from fully supervised to a semi-supervised training paradigm. In our method, the student is attracted toward dominant modes rather than individual noisy instances. This robustness mechanism is difficult to obtain from moment matching alone, which treats all samples implicitly through aggregated statistics, and it is also less explicit in adversarial alignment, where unstable optimization may itself introduce additional training noise [29]. The effectiveness of cross-modal distillation depends heavily on the teacher model's representational capacity and the accuracy of cross-modal spatial alignment. Calibration errors or large modality discrepancies can easily lead to feature misalignment and negative transfer. To this end, we propose a distribution-level cross-modal distillation method to effectively address the above challenges.

3 Method

In this section, we propose an innovative distribution-aware cross-modal distillation framework that transfers geometric knowledge from a LiDAR-based teacher model to a multi-view camera student model,

improving camera-only 3D object detection. Unlike mainstream point-to-point feature regression for BEV distillation, we model the teacher features with a probabilistic distribution and regularize the student features by enforcing distribution-level consistency. This method alleviates distillation instability caused by cross-modal feature distribution mismatches and noisy teacher signals.

3.1 Overall Architecture

As illustrated in Fig. 2, we first model the teacher's BEV features within each foreground object region as a probabilistic distribution, and encourage the student features to fall into its high-density regions. This strategy couples the supervision strength with the statistical uncertainty of the teacher features, automatically reweighting different feature dimensions. We impose stronger supervision on more stable feature directions, while appropriately relaxing the constraints on directions that are more variable. In this way, the student progressively aligns with the teacher's BEV feature distribution in an overall statistical sense, effectively narrowing the cross-modality feature gap in BEV space. Moreover, distribution-level distillation tends to pull the student toward the aggregated centers of dominant modes rather than individual noisy instances, thus mitigating the adverse impact of teacher noise without modifying the student architecture. Subsequently, we further introduce response distillation to refine output-level supervision and improve distillation quality. Notably, although distillation methods are effective at extracting and transferring knowledge, they cannot eliminate information loss at the physical level. To address this limitation, we incorporate temporal modeling to compensate for missing observations in the current frame by fusing information from historical frames.

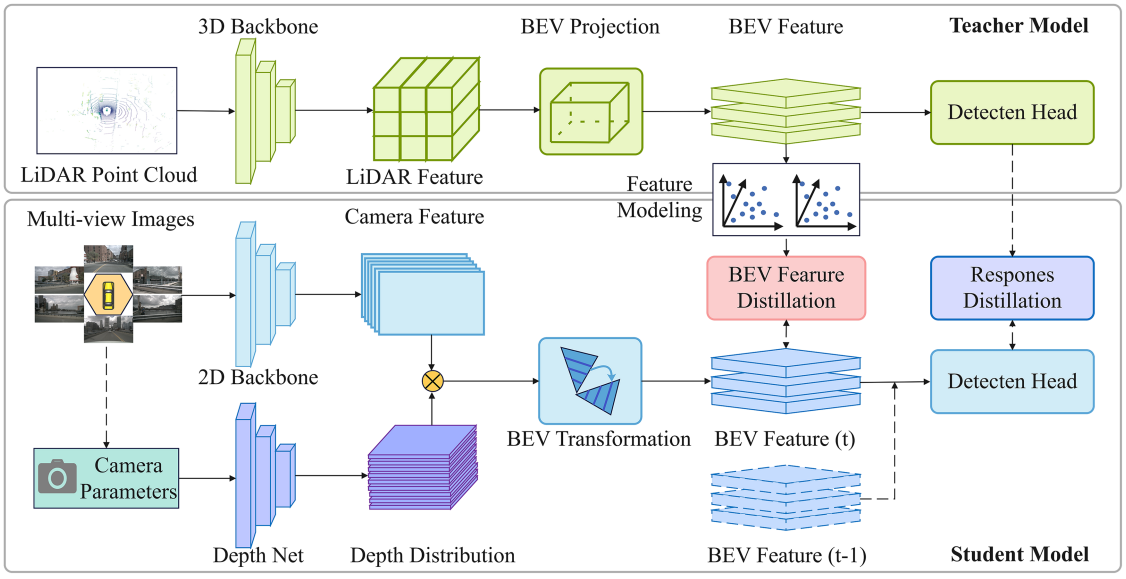

Figure 2: A cross-modal knowledge distillation framework integrating LiDAR and camera modalities for enhanced BEV object detection.

3.2 Distribution-Aware Cross-Modal Distillation Framework

Previous BEV feature distillation methods [6,30] typically use a foreground mask to select target-relevant regions on the BEV plane and perform point-wise alignment between the student and teacher features at these locations. This concentrates the distillation on key spatial positions and reduces interference from background noise. The distillation loss L_{feat} is defined as follows:

$$\mathcal{L}_{\text{feat}} = \frac{1}{N_p} \sum_{i}^{H} \sum_{j}^{W} M_{ij} \| f_{ij}^{t} - \xi(f_{ij}^{s}) \|_2, \tag{1}$$

where H and W are the height and width of the BEV feature map, respectively. $\| \cdot \|$ is the L2 norm. f_{ij}^{t} and f_{ij}^{s} denote the feature at location (i, j) from the teacher and student models, respectively. The foreground mask M is generated from the ground-truth heatmap in the BEV space, and N_p denotes the sum of all non-zero elements in the mask M. The adaptation module ξ uses the convolutional layer to match the dimensionality of the student's features to that of the teacher.

Although existing methods project both features maps onto the BEV plane to alleviate cross-view discrepancies, a domain gap still remains due to differences in imaging mechanisms and spatial resolution. Moreover, teacher features often contain noise and bias. Directly forcing the student to mimic the teacher's feature maps can weaken the distillation effectiveness. To address this, we employ a Dirichlet Process Gaussian Mixture Model (DPGMM) to model the distribution of the teacher's BEV features, approximating it as a mixture of Gaussian components. DPGMM can adaptively infer the effective number of active components for each class from the data, thereby avoiding per-class manual tuning and providing a more flexible prior for distribution-level distillation. Each component is parameterized by a mean and a covariance matrix, which describe the feature center and its variation across directions. This shifts teacher supervision from point-wise distillation to distribution-level distillation. We then introduce a distribution-consistency constraint to encourage the student features to match the teacher's mixture distribution in a probabilistic manner. Compared with purely point-wise regression, our method provides stronger and more structure-aware supervision. It avoids noise amplification and overly restrictive constraints caused by point-to-point alignment, leading to more robust BEV feature transfer.

Teacher model. The teacher model adopts CenterPoint, a LiDAR-based 3D object detector that performs detection in the BEV space. Given an input LiDAR point cloud, it first quantizes the 3D space into regular bins (voxels or pillars) and encodes points within each bin into learned features. A standard LiDAR-based backbone network (e.g., VoxelNet [31] or PointPillars [32]) then produces a BEV feature map $\mathbf{F}_{\text{bev}}^{t} \in \mathbb{R}^{C_t \times H \times W}$. The CenterPoint detection head predicts object centers and regresses 3D box attributes from this BEV feature map. Notably, the teacher model is used only during training to provide supervision for BEV feature distillation.

Student model. The student model is based on BEVDepth [9], a camera-only BEV detector that explicitly lifts multi-view image features into the BEV space using depth-aware projection. It first extracts image features with a image backbone and predicts per-pixel depth distributions using a depth network. The features are then lifted to 3D space and projected onto a predefined BEV grid through a lift-splat-shoot operation, followed by a 2D BEV backbone for further encoding, producing the student BEV feature map $\mathbf{F}_{\text{bev}}^{s} \in \mathbb{R}^{C_s \times H \times W}$. It has the same feature size as $\mathbf{F}_{\text{bev}}^{t}$. For distillation, we align the channel dimension by applying a 1×1 convolution.

Distribution-aware feature distillation (DAFD). For each ground-truth 3D bounding box b_k with a class label c, we project b_k onto the BEV plane and extract a foreground region Ω_k from the teacher model. Then, we apply average pooling to aggregate the features and obtain D-dimensional feature vectors:

$$\mathbf{F}_t = \{ f_k^t \in \mathbb{R}^D | k = 1, 2, \ldots, K \}, \tag{2}$$

where K is the number of foreground objects. f_k^t is the feature vector of the k-th object from the teacher model.

Because the feature distribution is highly class-dependent, mixing different semantic categories would lead to ambiguous high-density regions that provide misleading supervision for distillation. Therefore, we model each class c separately. For a given class c, we collect the corresponding teacher features:

$$\mathbf{D}_c^t = \{f_{k,c}^t \in \mathbb{R}^D\}_{k=1}^{N_c}, \tag{3}$$

where N_c denotes the number of teacher features belonging to class c. Due to variations in viewpoint, distance, and occlusion, features from the same category follow a multimodal distribution. To explicitly capture these intra-class modes, we model the teacher features of each class c with a DPGMM. In this way, different appearance and geometry patterns are separated into distinct sub-modes:

$$p(D_c^t) = \prod_{k=1}^{N_c} \sum_{m=1}^{\infty} \pi_{c,m} \mathcal{N}\left(f_{k,c}^t \mid \mu_{c,m}, \Sigma_{c,m}\right), \tag{4}$$

where $\{\pi_{c,m}\}_{m=1}^{\infty}$ denote the mixture weights of Gaussian components for class c, generated from a Dirichlet Process, satisfying $\sum_{m=1}^{\infty} \pi_{c,m} = 1$. Each component is parameterized by a mean $\mu_{c,m}$ and a covariance matrix $\Sigma_{c,m}$. $\mathcal{N}$ is the Gaussian distribution.

For each class c, we independently perform Collapsed Variational Inference (CVI) to infer the posterior over latent assignments $z_{k,c}$, where $z_{k,c} = m$ indicates that feature $f_{k,c}^t$ is generated from the m-th Gaussian component. We approximate the posterior as:

$$q(Z_c) = \prod_{k=1}^{N_c} q(z_{k,c}), \tag{5}$$

with categorical factors:

$$q(z_{k,c} = m) = r_{k,c,m}, \tag{6}$$

where $r_{k,c,m}$ is the responsibility of the m-th Gaussian component for feature $f_{k,c}^t$. The collapsed variational updates for assignment responsibilities are:

$$\tilde{r}_{k,c,m} \propto n_{-k,c,m}\, p(f_{k,c}^t \mid \mathcal{S}_{-k,c,m}^t), \tag{7}$$

$$\tilde{r}_{k,c,\text{new}} \propto \gamma\, p(f_{k,c}^t \mid \mathcal{S}_0^t), \tag{8}$$

and the normalized responsibilities are

$$r_{k,c,m} = \frac{\tilde{r}_{k,c,m}}{\sum_{m' \in \mathcal{M}_c} \tilde{r}_{k,c,m'} + \tilde{r}_{k,c,\text{new}}}, \tag{9}$$

$$r_{k,c,\text{new}} = \frac{\tilde{r}_{k,c,\text{new}}}{\sum_{m' \in \mathcal{M}_c} \tilde{r}_{k,c,m'} + \tilde{r}_{k,c,\text{new}}}, \tag{10}$$

where $n_{-k,c,m} = \sum_{i \neq k} r_{i,c,m}$ denotes the expected number of features assigned to the m-th component excluding the current sample k. γ is the Dirichlet process concentration parameter, controlling the trade-off between creating a new component and reusing existing ones. $r_{k,c,\text{new}}$ denotes the assignment responsibility that the feature $f_{k,c}^t$ in class c is assigned to a potential new component. $\mathcal{M}_c$ is the set of currently instantiated components in class c. $p(f_{k,c}^t \mid \mathcal{S}_{-k,c,m}^t)$ is the posterior predictive of the feature under component m, given the posterior hyperparameters $\mathcal{S}_{-k,c,m}^t$ computed from all other samples (again excluding k). $p(f_{k,c}^t \mid \mathcal{S}_0^t)$ is the prior predictive under the prior hyperparameters $\mathcal{S}_0^t$ for a new component.

Using the collapsed sufficient statistics aggregated over all samples, we obtain the posterior hyperparameters $\{\tilde{\pi}_{c,m}, \tilde{\mu}_{c,m}, \tilde{\Sigma}_{c,m}\}$, which characterize the class-wise multi-modal teacher feature distribution. For each teacher feature $f^t_{k,c}$, we define its dominant mode:

$$m^*_k = \arg\max_{m \in \mathcal{M}_c} r_{k,c,m}. \tag{11}$$

In rare cases, some mixture components are supported by only a few teacher features, which leads to unreliable density estimates. Enforcing distribution-aware distillation on such poorly-supported components may introduce noisy supervision. Therefore, we apply a tiny-component filter. For each class c and component m, we compute the expected number of assigned samples $N_{c,m}$. If $N_{c,m} < n_{\min}$, we mark component m as tiny and exclude it from distribution-aware distillation. For samples whose dominant mode $m^\star_k$ is a tiny component, we fall back to the basic feature regression loss in Eq. (14).

Next, we extract student BEV features f^s_k from the corresponding foreground regions Ω'_k of the student model, forming the set $\mathbf{F}_s = \{f^s_k \in \mathbb{R}^D | k = 1, 2, \ldots, K\}$. For each teacher object $f^t_{k,c}$ with class c and dominant mode $m^\star_k$, we regard the Gaussian distribution $\mathcal{N}\left(\cdot \mid \tilde{\mu}_{c,m^\star_k}, \tilde{\Sigma}_{c,m^\star_k}\right)$ as the target distribution for the feature f^s_k. The mode-aware loss is defined as follows:

$$\mathcal{L}^{\text{mode}}_k = \frac{1}{2}\left(f^s_k - \tilde{\mu}_{c,m^\star_k}\right)^\top \tilde{\Sigma}^{-1}_{c,m^\star_k}\left(f^s_k - \tilde{\mu}_{c,m^\star_k}\right) + \frac{1}{2}\log\left|\tilde{\Sigma}_{c,m^\star_k}\right|. \tag{12}$$

$\mathcal{L}^{\text{mode}}_k$ provides the primary mode-level distillation signal, imposing stronger supervision along critical directions with small teacher variance while adaptively relaxing the constraints along high-variance, noise-dominated directions.

In addition, we introduce a mixture-level regularization term to further align class-wise feature distributions across different modes:

$$\mathcal{L}^{\text{mix}}_k = -\log p_c(f^s_k) = -\log\left(\sum_{m \in \mathcal{M}_c} \tilde{\pi}_{c,m}\mathcal{N}(f^s_k \mid \tilde{\mu}_{c,m}, \tilde{\Sigma}_{c,m})\right). \tag{13}$$

To stabilize early training and ensure robustness, we include a standard pair-wise feature loss:

$$\mathcal{L}^{\text{pair}}_k = \|f^s_k - f^t_k\|^2_2. \tag{14}$$

$\mathcal{L}^{\text{pair}}_k$ preserves instance-level details, ensuring that the student does not ignore the specific teacher representation for the current sample. The weight of this loss is gradually decayed to avoid interfering with the probabilistic distillation loss.

Thus, the final BEV feature distillation loss is defined as follows:

$$\mathcal{L}_{\text{feat}} = \frac{1}{K}\sum_{k=1}^{K}(\lambda_{\text{mode}}\mathcal{L}^{\text{mode}}_k + \lambda_{\text{mix}}\mathcal{L}^{\text{mix}}_k + \lambda_{\text{pair}}\mathcal{L}^{\text{pair}}_k), \tag{15}$$

where λ_{mode}, λ_{mix} and λ_{pair} are scalar weights. With the above distribution-level constraints, the student progressively aligns with the LiDAR BEV feature distribution in a global statistical sense, effectively narrowing the feature gap between the two modalities. The pseudocode of the proposed distribution-aware feature distillation is presented in Algorithm 1.

Algorithm 1: Distribution-aware cross-modal knowledge distillation algorithm

Require: 3D coordinates of LiDAR point clouds: (x_{3D}, y_{3D}, z_{3D}), BEV features of the teacher model: F_t, BEV features of the student model: F_s, Hyperparameters: $\gamma, \mu_0, \Sigma_0, n_{min}$

Ensure: Feature distillation loss: $\mathcal{L}_{\text{feat}}$

1. 3D to BEV Coordinate Transformation
$(w, h, 1)^T \leftarrow M_{3D\to BEV} \cdot (x_{3D}, y_{3D}, z_{3D}, 1)^T$
Define ROI region Ω on the BEV feature map
2. DPMM Modeling for Teacher Features (CVI)
Extract teacher features from ROI Ω: $F_t = \{f_t^k \in \mathbb{R}^D | k = 1, 2, \ldots, K\}$
Initialize priors: γ, μ_0, Σ_0
for iteration in 1 to MaxIterations **do**
    Update Assignments (CVI Step)
    for each feature point f_t^k in F_t **do**
        Update responsibilities $r_{k,m}$ (Eqs. (7)–(10))
        Normalize $r_{k,m}$ such that $\sum_m r_{k,m} = 1$
    end for
end for
Compute Posterior Hyperparameters & Filter
for each Gaussian component $m \in \{1, \ldots, M\}$ **do**
    Update Component Parameters:
    $N_m \leftarrow \sum_{k=1}^{K} r_{k,m}$
    Update $\tilde{\pi}_m, \tilde{\mu}_m, \tilde{\Sigma}_m$
    Tiny-Component Filter:
    if $N_m \geq n_{min}$ **then**
        Add component m to active set $\mathcal{M}_{active}$
    else
        Discard component m
    end if
end for
3. Distribution-Aware Feature Distillation (DAFD)
Extract student features from ROI Ω': F_s
Initialize $\mathcal{L}_{\text{feat}} \leftarrow 0$
for each sample k in 1 to K **do**
    Identify dominant mode m_k^* for teacher feature f_t^k (Eq. (11))
    $\mathcal{L}_k^{\text{pair}} \leftarrow \|f_s^k - f_t^k\|_2^2$
    if $m_k^* \in \mathcal{M}_{active}$ **then**
        Compute $\mathcal{L}_k^{\text{mode}}$ using $\tilde{\mu}_{m_k^*}, \tilde{\Sigma}_{m_k^*}$ (Eq. (12))
        Compute $\mathcal{L}_k^{\text{mix}}$ using $\{\tilde{\pi}_m, \tilde{\mu}_m, \tilde{\Sigma}_m\}_{m\in\mathcal{M}_{active}}$ (Eq. (13))
        $\mathcal{L}_k \leftarrow \lambda_{\text{mode}}\mathcal{L}_k^{\text{mode}} + \lambda_{\text{mix}}\mathcal{L}_k^{\text{mix}} + \lambda_{\text{pair}}\mathcal{L}_k^{\text{pair}}$
    else
        $\mathcal{L}_k \leftarrow \lambda_{\text{pair}}\mathcal{L}_k^{\text{pair}}$
    end if
    $\mathcal{L}_{\text{feat}} \leftarrow \mathcal{L}_{\text{feat}} + \mathcal{L}_k$
end for
$\mathcal{L}_{\text{feat}} \leftarrow \frac{1}{K}\mathcal{L}_{\text{feat}}$

Response-Level Distillation. To transfer knowledge from the teacher's detection head to the student's detection head with the same architecture, we introduce response-level loss, which directly encourages the student head's outputs to match the teacher's responses. We also apply ground truth guided head distillation to prevent background dominated, uninformative locations from propagating noise.

For the classification branch, we distill the teacher's soft responses in foreground regions and define the classification distillation term using a Gaussian focal loss, following [30]:

$$\mathcal{L}_{\text{cls}} = L_{\text{GFocal}}\left(h^{s}_{\text{cls}},\, h^{t}_{\text{cls}} \odot M\right), \tag{16}$$

where h^{s}_{cls} and h^{t}_{cls} denote the classification heatmap outputs of the student and teacher models, respectively, and $\odot$ denotes element-wise multiplication. The foreground mask M is generated from the ground-truth Gaussian heatmap.

For the regression branch, following the training scheme of CenterPoint, we compute the regression distillation term using a L_1 loss only at the center locations of positive samples:

$$\mathcal{L}_{\text{reg}} = L_1\left(r^{s}_{\text{reg}}, r^{t}_{\text{reg}}\right), \tag{17}$$

where r^{s}_{reg} and r^{t}_{reg} denote the regression vectors predicted by the student and teacher models at the corresponding center locations, and $\mathcal{W}$ denotes the weight matrix. Finally, we obtain the response-level distillation loss:

$$\mathcal{L}_{\text{resp}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}}. \tag{18}$$

3.3 Overall Loss

To summarize, we improve the camera-based student detector by distilling knowledge from a LiDAR teacher at two complementary levels. First, we perform distribution-aware feature distillation, which aligns the student's BEV representations with the teacher via distribution-consistency constraints. Second, we apply response-level distillation on the detection head to further transfer the teacher's prediction behavior, providing direct output-level guidance. These distillation objectives are jointly optimized with the student's original training losses [9], including the standard 3D detection loss $\mathcal{L}_{\text{det}}$ and the depth supervision loss $\mathcal{L}_{\text{depth}}$. The overall loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{depth}} + \lambda_{\text{feat}}\,\mathcal{L}_{\text{feat}} + \lambda_{\text{resp}}\,\mathcal{L}_{\text{resp}}, \tag{19}$$

where λ_{feat} and λ_{resp} balance the contributions of feature-level and response-level distillation, respectively. During training, we assign a larger λ_{feat} in the early stage to facilitate feature-level knowledge transfer, and then gradually decrease λ_{feat} while increasing λ_{resp}, thereby shifting the optimization focus to $\mathcal{L}_{\text{resp}}$ to refine the final predictions.

3.4 Temporal Multi-View 3D Object Detection

While several existing methods achieve competitive 3D perception using a single image frame, relying solely on single-frame cues inevitably leads to performance bottlenecks. First, a single frame provides only static geometric and appearance information, which can result in unstable motion estimation. Second, objects that are occluded or only partially observed in one frame are more likely to be missed or localized inaccurately, hindering reliable detection. Incorporating temporal context improves the completeness and robustness of the representation. To this end, we introduce a lightweight, plug-and-play two-frame temporal fusion module that leverages distilled BEV features from the previous frame as historical compensation

and injects cross-frame information into the current-frame representation through explicit alignment and adaptive fusion, thereby improving detection stability.

We take the current-frame BEV feature $B_t \in \mathbb{R}^{C\times H\times W}$ as the primary representation and use the previous-frame feature B_{t-1} to provide cross-frame information compensation. As illustrated in Fig. 3, we first map the historical feature into the current coordinate system to eliminate the effect of ego-motion. We then selectively incorporate historical information through motion-aware gating and suppress dynamic and inconsistent regions. Finally, we achieve stable fusion in a residual manner.

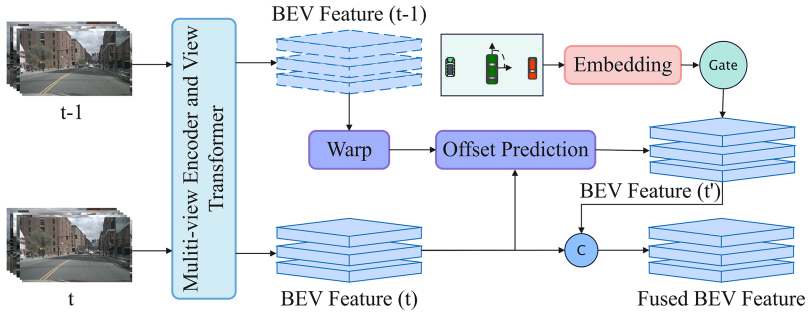

Figure 3: Architecture of the two-frame temporal fusion model.

Specifically, we compute the relative transformation matrix $T_{t-1\to t}$ from the ego poses of two consecutive frames. Based on this transformation, we construct a sampling function $G(x)$ that maps a BEV grid location x in the current frame to the corresponding feature coordinates in the previous frame. We then align the previous-frame BEV feature to the current coordinate system via spatial sampling:

$$\tilde{B}_{t-1} = \mathrm{Warp}(B_{t-1}, G(x)), \tag{20}$$

where $\mathrm{Warp}(\cdot)$ denotes bilinear interpolation on the BEV grid [33]. To compensate for local misalignment caused by discretized interpolation and dynamic objects, we introduce a learnable refinement on top of rigid alignment and predict a small offset increment $\Delta u(x)$ for each BEV location:

$$\Delta u(x) = f_{\max} \cdot \tanh(\mathrm{Conv}([B_t, \tilde{B}_{t-1}])), \tag{21}$$

where $[\cdot]$ denotes concatenation along the channel dimension. $f_{\max}$ is used to constrain the maximum displacement magnitude, ensuring that the refinement only compensates for local errors without introducing unstable deformations.

Unlike a cascaded two-stage warp (first obtaining $\tilde{B}_{t-1}$ and then resampling on $\tilde{B}_{t-1}$), we directly add the rigid grid and the residual increment to obtain a joint sampling grid, and perform only a single interpolated sampling on the original B_{t-1} to obtain the final aligned historical features:

$$\hat{B}_{t-1} = \mathrm{Warp}(B_{t-1}, G(x) + \Delta u(x)). \tag{22}$$

This design ensures that the entire alignment process performs only one interpolation, numerically avoiding the extra smoothing and amplification of systematic bias introduced by a second resampling.

Next, we introduce a pixel-wise gating weight $\mathcal{G}_{t-1} \in \mathbb{R}^{1\times H\times W}$ to estimate the contribution of history based on the current features, the aligned historical features, and motion priors (relative pose increment and time interval):

$$\mathcal{G}_{t-1} = \sigma(\mathrm{Conv}([B_t, \hat{B}_{t-1}, \Delta p_{t-1}, \Delta t_{t-1}])), \tag{23}$$

where Δp_{t-1} is a low-dimensional encoding of the relative translation and yaw, and Δt_{t-1} is a time-interval encoding. $\sigma(\cdot)$ is the sigmoid function. Then, we construct an inconsistency map from the aligned cross-frame differences and explicitly suppress the gating:

$$\mathcal{G}'_{t-1} = \mathcal{G}_{t-1} \cdot (1 - \sigma(\mathrm{Conv}(|B_t - \hat{B}_{t-1}|))). \tag{24}$$

Finally, we aggregate information from the previous frame in a residual manner:

$$B_t^{\mathrm{fused}} = B_t + \mathcal{G}'_{t-1} \odot \phi(\hat{B}_{t-1}),$$

where $\phi(\cdot)$ is a 1×1 convolution for channel alignment, and $\odot$ denotes element-wise multiplication.

4 Experiments

In this section, we present the evaluation setup, including the datasets used, evaluation metrics, and implementation details. We conduct a series of ablation studies and related analyses to thoroughly investigate the role and contribution of each component in our method. Finally, we perform comprehensive comparisons between our method and current state-of-the-art methods on widely used benchmark datasets.

4.1 Dataset and Metrics

We evaluate our method on the nuScenes datasets, covering diverse scenarios and sensor configurations.

nuScenes Dataset contains 1000 scenes (700 train, 150 val, 150 test) captured with 6 cameras and a 32-beam LiDAR at 20 Hz/10 Hz. Annotations include 1.4M 3D bounding boxes for 10 classes: *car, truck, bus, trailer, construction vehicle, pedestrian, motorcycle, bicycle, barrier, traffic cone*. We use the official metrics: nuScenes Detection Score (NDS), mean Average Precision (mAP), and 5 True Positive (TP) metrics: Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE). The NDS is calculated as follows:

$$\mathrm{NDS} = \frac{1}{10}\left(5 \times \mathrm{mAP} + \sum_{TP \in TP}(1 - \min(1, TP))\right) \tag{25}$$

4.2 Implementation Details

Our framework is implemented using the MMDetection3D toolkit and trained on 4 NVIDIA GeForce RTX 4090 GPUs. We employ the AdamW optimizer with a cosine-scheduled learning rate of 2×10^{-4} and a batch size of 8. Models are trained for 20 epochs on nuScenes using the Class-Balanced Group Sampling (CBGS) strategy. Data augmentations follow [9,34], including random flipping, scaling, rotation, and noise injection. We use ResNet-50/101 pre-trained on ImageNet-1K as image backbones. For LiDAR data during training, we adopt a pre-trained CenterPoint model. CenterPoint was selected because it is a representative and widely used 3D detector with strong performance, providing a reliable baseline for evaluating DA-T3D and facilitating fair comparison with other models. For ablation studies, we train models for 24 epochs. When comparing with state-of-the-art methods, we extend training to 60 epochs for convergence. During inference, we process 2 frames and apply motion compensation using ego-vehicle pose information. For temporal data augmentation, we randomly skip 1 frame in the training sequence.

In our DPGMM-based feature modeling, we set the tiny-component filter threshold to $n_{\min} = 5$ and the Dirichlet Process concentration parameter to $\gamma = 1.0$. For the DPGMM base prior, (μ_0, Σ_0) is specified as a simple data-adaptive weak prior. For each class, we compute the mean and diagonal covariance of the teacher ROI-aggregated features over the training set, use them as μ_0 and Σ_0, and add a small diagonal

jitter for numerical stability. In our implementation, the DPGMM is fitted offline, and the resulting mixture parameters remain fixed throughout distillation training. Moreover, the DPGMM is used only during training. At inference time, neither the teacher model nor the DPGMM fitting process is involved, and thus our method introduces no additional computational cost compared with the student detector. We set the initial loss weights to $\lambda_{\text{feat}} = 1.0$ and $\lambda_{\text{resp}} = 0.1$, and linearly adjust them during training by gradually decreasing λ_{feat} while increasing λ_{resp}, so that training emphasizes feature distillation in the early stage and shifts the focus to response distillation in the later stage. We fix $\lambda_{\text{mode}} = 1.0$ and $\lambda_{\text{mix}} = 0.1$, while λ_{pair} is decayed with a cosine schedule from 1.0 to 0.

4.3 Comparison with Other Models

We first report the main comparison results on the nuScenes validation set under the standard evaluation protocol. For a fair comparison, we group methods by backbone and input setting (image resolution and the number of frames), and summarize the overall 3D detection performance using the official metrics mAP and NDS. Table 1 compares our method with representative camera-based 3D detectors, and Table 2 further benchmarks different cross-modal distillation strategies under comparable student/teacher settings.

Table 1: Comparison of methods on the nuScenes val set for 3D object detection.

Method	Backbone	Image Size	Frames	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAEE↓
BEVDet [34]	ResNet50	256 × 704	1	0.298	0.379	0.725	0.279	0.589	0.860	0.245
BEVDet4D [33]	ResNet50	256 × 704	2	0.322	0.457	0.703	0.278	0.495	0.354	0.206
PETRv2 [13]	ResNet50	256 × 704	2	0.349	0.456	0.700	0.275	0.580	0.437	0.187
BEVDepth [9]	ResNet50	256 × 704	2	0.351	0.475	0.639	0.267	0.479	0.428	0.198
BEVStereo [35]	ResNet50	256 × 704	2	0.372	0.500	0.598	0.270	0.438	0.367	0.190
BEVFormerv2 [36]	ResNet50	–	–	0.423	0.529	0.618	0.273	0.413	0.333	0.188
SOLOFusion [37]	ResNet50	256 × 704	16 + 1	0.427	0.534	0.567	0.274	0.511	0.252	0.181
BEVPoolv2 [38]	ResNet50	256 × 704	8 + 1	0.406	0.526	0.572	0.275	0.463	0.275	0.188
DA-T3D	ResNet50	256 × 704	2	0.421	0.543	0.532	0.223	0.398	0.347	0.175
DETR3D [12]	ResNet101-DCN	900 × 1600	1	0.349	0.434	0.716	0.268	0.379	0.842	0.200
Focal-PETR [39]	ResNet101-DCN	512 × 1408	1	0.390	0.461	0.678	0.263	0.395	0.804	0.202
PETR [40]	ResNet101-DCN	512 × 1408	1	0.366	0.441	0.717	0.267	0.412	0.834	0.190
BEVFormer [41]	ResNet101-DCN	900 × 1600	4	0.416	0.517	0.673	0.274	0.372	0.394	0.198
PolarDETR [42]	ResNet101-DCN	900 × 1600	2	0.383	0.488	0.707	0.269	0.344	0.518	0.196
Sparse4D [14]	ResNet101-DCN	900 × 1600	4	0.436	0.541	0.633	0.279	0.363	0.317	0.177
CenterNet [43]	DLA	–	–	0.306	0.328	0.716	0.264	0.609	1.426	0.658
FCOS3D [44]	ResNet101	900 × 1600	–	0.343	0.415	0.725	0.263	0.422	1.292	0.153
PGD [45]	ResNet101	900 × 1600	–	0.369	0.428	0.683	0.260	0.439	1.268	0.185
BEVDepth	ResNet101	512 × 1408	2	0.412	0.535	0.565	0.266	0.358	0.331	0.190
SOLOFusion	ResNet101	512 × 1408	16 + 1	0.483	0.582	0.503	0.264	0.381	0.246	0.207
BEVFormer-En [46]	–	–	–	0.418	0.529	0.631	0.268	0.328	0.373	0.194
BEVDiffuser [47]	–	–	–	0.430	0.537	0.638	0.274	0.333	0.355	0.179
DA-T3D	ResNet101	512 × 1408	2	0.467	0.581	0.501	0.239	0.315	0.295	0.178

Table 1 compares our method with representative camera-based 3D object detection approaches on the nuScenes validation set, evaluated by mAP, NDS, and five TP error metrics. Overall, our approach achieves strong performance under two common settings: with ResNet50 at 256 × 704 using 2 frames, we obtain 0.421 mAP and 0.543 NDS; with ResNet101 at 512 × 1408 using 2 frames, we reach 0.467 mAP and 0.581 NDS. Beyond the headline scores, we consistently reduce key geometric errors (especially mATE and mAOE), indicating improvements in localization and orientation stability rather than a metric-specific trade-off.

Table 2: Comparison to other cross-modal knowledge distillation methods on the nuScenes val set.

Method	Backbone	Image Size	Student Model	Teacher Modality	mAP↑	NDS↑
MemDistill [48]	ResNet50	256 × 704	BEVDepth	Lidar	0.425	0.531
LabelDistill [30]	ResNet50	256 × 704	BEVDepth	Lidar	0.419	0.528
X3KD [49]	ResNet50	256 × 704	BEVDepth	Lidar	0.390	0.505
Set2Set [50]	ResNet50	256 × 704	–	Lidar	0.375	0.479
MonoDistill [4]	ResNet50	256 × 704	MonoDLE	Lidar	0.390	0.495
BEVDistill [51]	ResNet50	900 × 1600	BEVFormer	Lidar	0.407	0.515
DA-T3D	ResNet50	256 × 704	BEVDepth	Lidar	0.421	0.543
UVTR [5]	ResNet101	900 × 1600	–	Lidar	0.392	0.488
DistillBEV [25]	ResNet101	512 × 1408	BEVDepth	Lidar	0.450	0.547
PromptDet [52]	ResNet101	256 × 704	BEVDepth	Lidar	0.433	0.569
TiG-BEV [53]	ResNet101	512 × 1408	BEVDepth	Lidar	0.440	0.544
SimDistill [54]	SwinT	256 × 704	BEVFusion-C	Lidar&Camera	0.404	0.453
DA-T3D	ResNet101	512 × 1408	BEVDepth	Lidar	0.467	0.581

Under the ResNet50 configuration, many competing methods operate with similar input resolution and typically 2 frames. Our method attains a higher NDS while simultaneously lowering geometry-related errors (mATE: 0.532, mASE: 0.223, mAOE: 0.398). These gains align with our design motivation: instead of enforcing strict point-wise matching, we perform distribution-aware cross-modal distillation that guides the student toward high-density and geometrically stable regions of the teacher feature space, which helps mitigate modality mismatch and suppress noisy teacher outliers. In addition, response-level distillation further transfers reliable decision behavior to the student, contributing to the overall quality improvements across TP metrics.

Some methods achieve strong performance by aggregating many historical frames (e.g., 16 + 1). In contrast, our model uses only 2 frames yet achieves competitive or better NDS and notably improved localization, demonstrating that lightweight temporal fusion with alignment and selective information injection can effectively compensate for occlusions and missing observations without relying on long sequences.

The ResNet101 results further confirm the scalability of our framework. With 2 frames, we achieve 0.467 mAP and 0.581 NDS, together with lower errors compared to the 2-frame baseline. These consistent gains support our conclusion that distribution-aware distillation provides robust geometric supervision under modality gaps, while the lightweight temporal design improves robustness in dynamic and occluded scenarios at low temporal overhead.

Table 2 compares our approach with representative cross-modal knowledge distillation and multi-modal baselines on the nuScenes validation set. Existing distillation methods typically reduce the camera–LiDAR gap via foreground feature imitation, label and response distillation, or multi-stage alignment. However, their gains can be affected by modality-induced distribution mismatch and noisy teacher signals (e.g., missed or false detections and feature jitter), which may limit robustness. In contrast, our method performs distribution-aware cross-modal distillation and combines it with a lightweight temporal design, aiming to transfer more stable geometric knowledge while keeping the student model efficient.

Under the ResNet50, 256 × 704 setting (with BEVDepth as the student), our method achieves 0.421 mAP and 0.543 NDS, yielding the best NDS among the listed ResNet50 baselines. Notably, while some methods may obtain slightly higher mAP in this group, the improved NDS suggests better overall detection quality

when both accuracy and error-related components are jointly considered. This supports that distribution-aware alignment can mitigate strict point-wise matching issues under cross-modal heterogeneity and suppress outlier teacher supervision, resulting in more reliable knowledge transfer.

With a stronger backbone and higher resolution (ResNet101, 512 × 1408), our method achieves 0.467 mAP and 0.581 NDS, which are the best results in the ResNet101 group. Compared with prior distillation approaches under similar student/teacher modality settings, this indicates that explicitly addressing distribution mismatch and teacher noise is crucial. Simply introducing additional modalities or branches does not necessarily guarantee consistent improvements. Fig. 4 shows the visualization results of our method and other methods. Overall, the results demonstrate that our framework scales favorably across backbones and resolutions, providing stable gains for camera-based 3D detection.

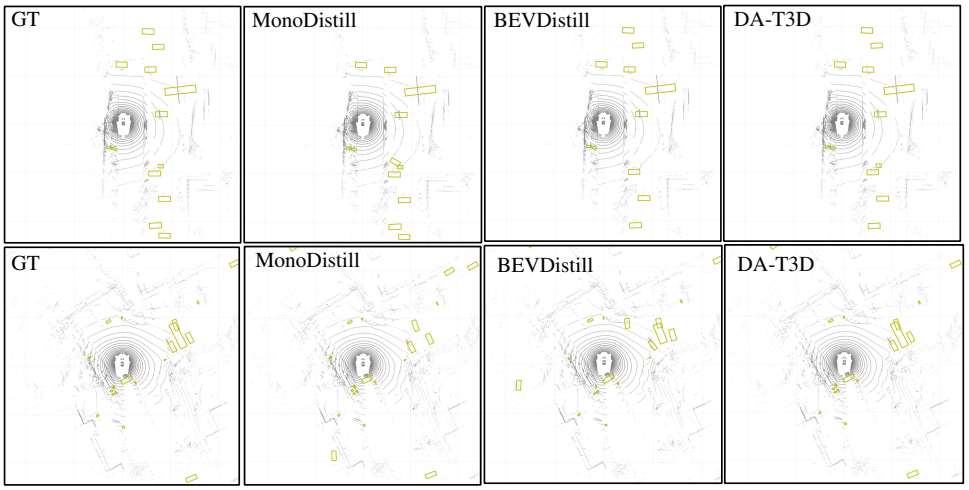

Figure 4: Visualization results of different cross modal knowledge distillation methods.

4.4 Ablation Study

To better understand where the improvements come from, we conduct controlled ablations on the nuScenes validation set by enabling one component at a time. We use BEVDepth as the camera-only baseline student, adopt CenterPoint as the teacher for distillation, and employ a lightweight 2-frame temporal modeling strategy. The results in Table 3 progressively quantify the contribution of feature distillation, response distillation, and temporal modeling.

Table 3: Ablation study of each component.

Setting	Baseline	Feature Distill.	Response Distill.	Temporal (2-Frame)	mAP↑	NDS↑
1	✓				0.412	0.535
2	✓	✓			0.443	0.565
3	✓	✓	✓		0.451	0.572
4	✓	✓	✓	✓	0.467	0.581

Table 3 reports an ablation study on the nuScenes validation set to quantify the contribution of each component in our framework. In Setting 1, the baseline achieves 0.412 mAP and 0.535 NDS. In Setting 2, after introducing feature distillation, the performance improves to 0.443 mAP and 0.565 NDS, indicating

that intermediate BEV representation guidance from the LiDAR teacher helps narrow the modality gap and provides more reliable geometric cues for the camera model, thereby improving spatial feature quality and overall 3D detection performance. Further enabling response distillation on top of feature distillation yields 0.451 mAP and 0.572 NDS, bringing consistent gains. This suggests that output-level supervision complements feature-level alignment by refining the student's prediction distribution, which improves the final detection heads. Finally, incorporating 2-frame temporal modeling achieves the best performance of 0.467 mAP and 0.581 NDS. Temporal fusion aggregates complementary observations across consecutive frames, mitigating single-frame noise and partial occlusions and producing more stable BEV features and more consistent localization, which is reflected in both mAP and NDS. In order to more intuitively demonstrate the difference between the baseline model and the optimal setting, we visualized their inference results, as shown in Fig. 5.

Figure 5: Visualization results of baseline and our method.

To further investigate the contribution of the proposed distribution-aware feature distillation objective, we additionally perform a loss-level ablation study, as reported in Table 4.

Table 4: Ablation study of each loss.

Setting	$\mathcal{L}_{\text{pair}}$	$\mathcal{L}_{\text{mode}}$	$\mathcal{L}_{\text{mix}}$	mAP↑	NDS↑
1	✓			0.448	0.563
2	✓	✓		0.464	0.579
3	✓	✓	✓	0.467	0.581

Table 4 further analyzes the effect of each loss term in the proposed distribution-aware feature distillation. Starting from the pair-wise loss $\mathcal{L}_{\text{pair}}$, the model achieves 0.448 mAP and 0.563 NDS, indicating that instance-level feature regression provides a stable optimization basis for cross-modal alignment. After introducing the mode-aware loss $\mathcal{L}_{\text{mode}}$, the performance increases to 0.464 mAP and 0.579 NDS. This notable improvement shows that constraining the student features toward the dominant high-density mode of the teacher distribution is more effective than relying only on point-wise matching, as it better captures the

intrinsic structure of teacher features while reducing the influence of noisy samples. When the mixture-level regularization $\mathcal{L}_{\text{mix}}$ is further added, the performance is further improved to 0.467 mAP and 0.581 NDS. Although the gain is relatively smaller, it demonstrates that mixture-level distribution alignment provides complementary supervision by encouraging global consistency across different intra-class modes. Overall, the best performance is achieved by jointly using $\mathcal{L}_{\text{pair}}$, $\mathcal{L}_{\text{mode}}$, and $\mathcal{L}_{\text{mix}}$, validating the effectiveness and complementarity of the three loss terms.

We further analyze the design of the temporal modeling module by ablating its key components, including the learnable refinement and the motion-aware gating mechanism, as shown in Table 5.

Table 5: Ablation study of each component in temporal modeling.

Setting	Baseline	Learnable Refinement	Motion-Aware Gating	mAP↑	NDS↑
1	✓			0.460	0.576
2	✓	✓		0.464	0.579
3	✓	✓	✓	0.467	0.581

Table 5 presents an ablation study of each component in the temporal modeling module. Here, the baseline denotes the basic two-frame fusion design with ego-motion compensation, while removing the learnable refinement offset and the motion-aware gating mechanism. Under this setting, the model achieves 0.460 mAP and 0.576 NDS, showing that simple temporal aggregation already provides useful historical context. After introducing the learnable refinement, the performance improves to 0.464 mAP and 0.579 NDS. This gain indicates that compensating for local misalignment beyond rigid ego-motion warping is beneficial, since discretization errors and dynamic scene variations cannot be fully handled by geometric transformation alone. When the motion-aware gating mechanism is further incorporated, the performance reaches 0.467 mAP and 0.581 NDS. This result shows that adaptively controlling the contribution of historical features is important for suppressing inconsistent or noisy temporal information, especially in regions affected by object motion or partial occlusion.

5 Conclusion

This paper presents a novel distribution-aware cross-modal distillation framework that transfers geometric priors from a LiDAR-based teacher to a camera-only student for temporal 3D object detection in the BEV space. To address distillation instability caused by modality heterogeneity and noisy teacher features, we propose distribution-aware BEV feature distillation that explicitly models class-conditional BEV feature distributions of the teacher using a DPGMM and constrains student features to match the teacher's distribution in a probabilistic manner. Next, we introduce response-level distillation to transfer task-specific decision behavior at the detection head, improving output calibration and localization refinement. Furthermore, we design a lightweight two-frame temporal fusion module with ego-motion compensation, residual alignment refinement, and motion-aware gating to robustly aggregate complementary observations from consecutive frames. Although our study achieves promising results, the proposed framework may be less effective in adverse environments (e.g., low light, rain, or fog) where both camera and LiDAR signals degrade, making the teacher's predictions unreliable and the student's inputs severely corrupted. In such cases, distillation may propagate erroneous supervision and reduce overall performance. In future work, we will systematically investigate robustness under severe sensor degradation. In addition, we will also focus on uncertainty issues caused by long-tail categories and complex motion patterns, and explore more adaptive mixture distribution modeling and uncertainty characterization methods.

Acknowledgement: None.

Funding Statement: This paper is supported by the National Natural Science Foundation of China (Grant No. 62302086).

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Tianzhe Jiao and Jie Song; methodology, Tianzhe Jiao and Yuming Chen; software, Xiaoyue Feng ; validation, Yuming Chen, Tianzhe Jiao and Chaopeng Guo; formal analysis, Tianzhe Jiao; investigation, Yuming Chen; resources, Tianzhe Jiao; data curation, Xiaoyue Feng; writing—original draft preparation, Tianzhe Jiao; writing—review and editing, Jie Song; visualization, Yuming Chen; supervision, Jie Song; project administration, Chaopeng Guo; funding acquisition, Chaopeng Guo. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, upon reasonable request. The original data presented in the study are openly available in publicly accessible repositories: nuScenes at https://www.nuscenes.org/ and KITTI at http://www.cvlibs.net/datasets/kitti/eval_object.php.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang Y, Jiang H, Chen G, Zhang T, Zhou J, Qing Z, et al. Efficient and robust multi-camera 3D object detection in Bird-Eye-View. Image Vis Comput. 2025;154:105428. doi:10.1016/j.imavis.2025.105428.
2. Zhang C, Liang L, Zhou J, Xu Y. Multi-view depth estimation based on multi-feature aggregation for 3D reconstruction. Comput Graph. 2024;122(4):103954. doi:10.1016/j.cag.2024.103954.
3. Li J, Xu J, Zhi M, Zhang J, Zhuo L. BEV-CMHF: a cross-modality hybrid fusion framework for BEV 3D object detection with feature interaction and temporal fusion, Early access. IEEE Trans Intell Transp Syst. 2026. doi:10.1109/TITS.2026.3651793.
4. Chong Z, Ma X, Zhang H, Yue Y, Li H, Wang Z, et al. MonoDistill: learning spatial features for monocular 3D object detection. In: The Tenth International Conference on Learning Representations, ICLR 2022; 2022 Apr 25–29; Virtual Event.
5. Li Y, Chen Y, Qi X, Li Z, Sun J, Jia J. Unifying voxel-based representation with transformer for 3D object detection. Adv Neural Inf Process Syst. 2022;35:18442–55.
6. Chen Z, Li Z, Zhang S, Fang L, Jiang Q, Zhao F. BEVDistill: cross-modal BEV distillation for multi-view 3D object detection. In: The Eleventh International Conference on Learning Representations; 2023 [cited 2026 Jan 1]. Available from: https://openreview.net/forum?id=-2zfgNS917.
7. Liu J, Wang P, Liu W, Du S, Li Y, Hu J, et al. DMPG-BEV: diffusion model-based point clouds features generation for efficient camera-based BEV perception. IEEE Sens J. 2025;25(15):28905–18.
8. Philion J, Fidler S. Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In: Proceedings of the Computer Vision—ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK. p. 194–210.
9. Li Y, Ge Z, Yu G, Yang J, Wang Z, Shi Y, et al. Bevdepth: acquisition of reliable depth for multi-view 3D object detection. Proc AAAI Conf Artif Intell. 2023;37(2):1477–85.
10. Li Z, Wang W, Li H, Xie E, Sima C, Lu T, et al. Bevformer: learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European conference on computer vision. Berlin/Heidelberg, Germany: Springer; 2022. p. 1–18.
11. Zhang J, Zhang Y, Qi Y, Fu Z, Liu Q, Wang Y. GeoBEV: learning geometric BEV representation for multi-view 3D object detection. Proc AAAI Conf Artif Intell. 2025;39(9):9960–8.
12. Wang Y, Guizilini VC, Zhang T, Wang Y, Zhao H, Solomon J. Detr3d: 3D object detection from multi-view images via 3D-to-2D queries. Proc Mach Learn Res. 2022;164:180–91.

13. Liu Y, Yan J, Jia F, Li S, Gao Q, Wang T, et al. Petrv2: a unified framework for 3D perception from multi-camera images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 2–3; Paris, France. p. 3262–72.
14. Lin X, Lin T, Pei Z, Huang L, Su Z. Sparse4D: multi-view 3D object detection with sparse spatial-temporal fusion. arXiv:2211.10581. 2022.
15. Chen J, Chen H, Zhu Z, Shen Z, Ling X, Zhu Y. DDCFusion: dynamic depth compensation fusion for camera–radar 3-D object detection. IEEE Sens J. 2026;26(3):4561–74.
16. Vora S, Lang AH, Helou B, Beijbom O. Pointpainting: sequential fusion for 3D object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 4604–12.
17. Sindagi VA, Zhou Y, Tuzel O. MVX-Net: multimodal voxelNet for 3D object detection. In: Proceedings of the 2019 International Conference on Robotics and Automation (ICRA); 2019 May 20–24; Montreal, QC, Canada. p. 7276–82.
18. Zhang L, Li C. PPF-Net: efficient multimodal 3D object detection with pillar-point fusion. Electronics. 2025;14(4):685.
19. Liang T, Xie H, Yu K, Xia Z, Lin Z, Wang Y, et al. BEVFusion: a simple and robust LiDAR-camera fusion framework. Adv Neural Inf Process Syst. 2022;35:10421–34.
20. Liu Z, Tang H, Amini A, Yang X, Mao H, Rus DL, et al. Bevfusion: multi-task multi-sensor fusion with unified bird's-eye view representation. In: Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA); 2023 May 29–Jun 2; London, UK. p. 2774–81.
21. Chen X, Ma H, Wan J, Li B, Xia T. Multi-view 3D object detection network for autonomous driving. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017; 2017 Jul 21–26; Honolulu, HI, USA. p. 6526–34.
22. Pang S, Morris D, Radha H. CLOCs: camera-LiDAR object candidates fusion for 3D object detection. In: Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2020 Oct 24–2021 Jan 24; Las Vegas, NV, USA. p. 10386–93.
23. Wang Z, Huang Z, Gao Y, Wang N, Liu S. MV2DFusion: leveraging modality-specific object semantics for multi-modal 3D detection. IEEE Trans Pattern Anal Mach Intell. 2026;48(1):609–23.
24. Zhang G, Song Z, Liu L, Ou Z. FGU3R: fine-grained fusion via unified 3D representation for multimodal 3D object Detection. In: Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025; 2025 Apr 6–11; Hyderabad, India. p. 1–5.
25. Wang Z, Li D, Luo C, Xie C, Yang X. Distillbev: boosting multi-camera 3D object detection with cross-modal knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 1–6; Paris, France. p. 8637–46.
26. Zhou S, Liu W, Hu C, Zhou S, Ma C. UniDistill: a universal cross-modality knowledge distillation framework for 3D object detection in bird's-eye view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. p. 5116–25.
27. Hekimoglu A, Schmidt M, Marcos-Ramiro A. Monocular 3D object detection with LiDAR guided semi supervised active learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024; 2024 Jan 3–8; Waikoloa, HI, USA. p. 2335–44.
28. Xu R, Xiang Z, Zhang C, Zhong H, Zhao X, Dang R, et al. SCKD: semi-supervised cross-modality knowledge distillation for 4D radar object detection. Proc AAAI Conf Artif Intell. 2025;39(9):8933–41.
29. Dong N, Zhang YQ, Ding ML, Xu SB, Bai YC. One-stage object detection knowledge distillation via adversarial learning. Appl Intell. 2022;52(4):4582–98. doi:10.1007/s10489-021-02634-6.
30. Kim S, Kim Y, Hwang S, Jeong H, Kum D. LabelDistill: label-guided cross-modal knowledge distillation for camera-based 3D object detection. In: European Conference on Computer Vision. Berlin/Heidelberg, Germany: Springer; 2024. p. 19–37.
31. Zhou Y, Tuzel O. Voxelnet: end-to-end learning for point cloud based 3D object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 4490–9.

32. Lang AH, Vora S, Caesar H, Zhou L, Yang J, Beijbom O. PointPillars: fast encoders for object detection from point clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019; 2017 Jun 16–20; Long Beach, CA, USA. p. 12697–705.
33. Huang J, Huang G. BEVDet4D: exploit temporal cues in multi-camera 3D object detection. arXiv:2203.17054. 2022.
34. Huang J, Huang G, Zhu Z, Du D. BEVDet: high-performance multi-camera 3D object detection in bird-eye-view. arXiv:2112.11790. 2021.
35. Li Y, Bao H, Ge Z, Yang J, Sun J, Li Z. BEVStereo: enhancing depth estimation in multi-view 3D object detection with temporal stereo. Proc AAAI Conf Artif Intell. 2023;37(2):1486–94.
36. Yang C, Chen Y, Tian H, Tao C, Zhu X, Zhang Z, et al. BEVFormer v2: adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023; 2023 Jun 17–24; Vancouver, BC, Canada. p. 17830–9.
37. Park J, Xu C, Yang S, Keutzer K, Kitani KM, Tomizuka M, et al. Time will tell: new outlooks and a baseline for temporal multi-view 3D object detection. In: Proceedings of the The Eleventh International Conference on Learning Representations, ICLR 2023; 2023 May 1–5; Kigali, Rwanda.
38. Huang J, Huang G. BEVPoolv2: a cutting-edge implementation of BEVDet toward deployment. arXiv:2211.17111. 2022.
39. Wang S, Jiang X, Li Y. Focal-PETR: embracing foreground for efficient multi-camera 3D object detection. IEEE Trans Intell Veh. 2024;9(1):1481–9.
40. Liu Y, Wang T, Zhang X, Sun J. PETR: position embedding transformation for multi-view 3D object detection. In: Avidan S, Brostow GJ, Cissé M, Farinella GM, Hassner T, editors. Computer vision—ECCV 2022. Berlin/Heidelberg, Germany: Springer; 2022. p. 531–48. doi:10.1007/978-3-031-19812-0_31.
41. Li Z, Wang W, Li H, Xie E, Sima C, Lu T, et al. BEVFormer: learning bird's-eye-view representation from LiDAR-camera via spatiotemporal transformers. IEEE Trans Pattern Anal Mach Intell. 2025;47(3):2020–36.
42. Chen S, Wang X, Cheng T, Zhang Q, Huang C, Liu W. Polar parametrization for vision-based surround-view 3D detection. arXiv:2206.10965. 2022.
43. Yin T, Zhou X, Krähenbühl P. Center-based 3D object detection and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021; 2021 Jun 19–25; Virtual. p. 11784–93. doi:10.1109/cvpr46437.2021.01161.
44. Wang T, Zhu X, Pang J, Lin D. FCOS3D: fully convolutional one-stage monocular 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021; 2021 Oct 11–17; Montreal, QC, Canada. p. 913–22.
45. Wang T, Zhu X, Pang J, Lin D. Probabilistic and geometric depth: detecting objects in perspective. Proc Mach Learn Res. 2022;164:1475–85.
46. Nachkov A, Paudel DP, Danelljan M, Gool LV. Diffusion-based particle-DETR for BEV perception. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025; 2025 Feb 26–Mar 6; Tucson, AZ, USA. p. 2725–35.
47. Ye X, Yaman B, Cheng S, Tao F, Mallik A, Ren L. BEVDiffuser: plug-and-play diffusion model for BEV denoising with ground-truth guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025; 2025 Jun 11–15; Nashville, TN, USA. p. 1495–504.
48. Kwon D, Yoon Y, Son H, Kwak S. MemDistill: distilling LiDAR knowledge into memory for camera-only 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2025 Oct 19–23; Honolulu, HI, USA. p. 6828–38.
49. Klingner M, Borse S, Kumar VR, Rezaei B, Narayanan V, Yogamani SK, et al. X3KD: knowledge distillation across modalities, tasks and stages for multi-camera 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023; 2023 Jun 17–24; Vancouver, BC, Canada. p. 13343–53.
50. Wang Y, Solomon JM. Object DGCNN: 3D object detection using dynamic graphs. In: Proceedings of the Neural Information Processing Systems 2021, NeurIPS 2021; 2021 Dec 6–14; Virtual. p. 20745–58.

51. Wang S, Liu Y, Wang T, Li Y, Zhang X. Exploring object-centric temporal modeling for efficient multi-view 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 3621–31.
52. Guo K, Ling Q. PromptDet: a lightweight 3D object detection framework with LiDAR prompts. Proc AAAI Conf Artif Intell. 2025;39(3):3266–74.
53. Huang P, Liu L, Zhang R, Zhang S, Xu X, Wang B, et al. TiG-BEV: multi-view BEV 3D object detection via target inner-geometry learning. arXiv:2212.13979. 2022.
54. Zhao H, Zhang Q, Zhao S, Chen Z, Zhang J, Tao D. SimDistill: simulated multi-modal distillation for BEV 3D object detection. Proc AAAI Conf Artif Intell. 2024;38(7):7460–8. doi:10.1609/aaai.v38i7.28577.