Leveraging Vision-Centric Multi-Modal Expertise for 3D Object Detection

Linyan Huang[1] **Zhiqi Li**[2] **Chonghao Sima**[1] **Wenhai Wang**[3]
Jingdong Wang[4] **Yu Qiao**[1] **Hongyang Li**[1]

[1]Shanghai AI Lab [2]Nanjing University [3]CUHK [4]Baidu

Abstract

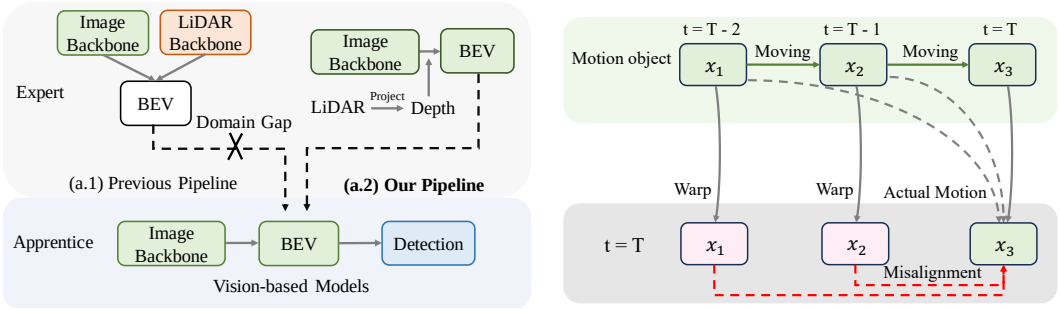
Current research is primarily dedicated to advancing the accuracy of camera-only 3D object detectors (apprentice) through the knowledge transferred from LiDAR- or multi-modal-based counterparts (expert). However, the presence of the domain gap between LiDAR and camera features, coupled with the inherent incompatibility in temporal fusion, significantly hinders the effectiveness of distillation-based enhancements for apprentices. Motivated by the success of uni-modal distillation, an apprentice-friendly expert model would predominantly rely on camera features, while still achieving comparable performance to multi-modal models. To this end, we introduce **VCD**, a framework to improve the camera-only apprentice model, including an apprentice-friendly multi-modal expert and temporal-fusion-friendly distillation supervision. The multi-modal expert **VCD-E** adopts an identical structure as that of the camera-only apprentice in order to alleviate the feature disparity, and leverages LiDAR input as a depth prior to reconstruct the 3D scene, achieving the performance on par with other heterogeneous multi-modal experts. Additionally, a fine-grained trajectory-based distillation module is introduced with the purpose of individually rectifying the motion misalignment for each object in the scene. With those improvements, our camera-only apprentice **VCD-A** sets new state-of-the-art on nuScenes with a score of 63.1% NDS. The code will be released at https://github.com/OpenDriveLab/Birds-eye-view-Perception.

1 Introduction

The camera-only 3D perception has garnered increasing attention in autonomous driving perception tasks [31, 21, 47, 50]. Although camera-only models possess the advantages of low deployment cost and ease of widespread application, they still fall behind state-of-the-art models that leverage LiDAR sensors regarding perception accuracy. Researchers have recently employed distillation methods to transfer knowledge from a powerful expert model into a camera-only apprentice model, with the expectation of leveraging the expertise of these stronger expert models to enhance the capability of the camera-only models. Existing 3D perception distillation methods often adopt expert models with the best performance, such as LiDAR-based models [60] or multi-modal fusion models [2, 59, 33]. However, the presence of the domain gap between LiDAR and camera features hampers knowledge transfer during distillation, resulting in limited improvements in practical applications. An alternative expert model is the large-scale camera-only model [32, 61]. Despite eliminating the domain gap between the camera-only expert model and the apprentice model, the expert model falls short in terms of effectiveness due to the inherent lack of precise geometry information. Likewise, it fails to yield a satisfactory improvement to the apprentice model. Hence, a desirable expert should meet two essential requirements: attaining state-of-the-art performance and minimizing the domain gap.

Furthermore, current distillation methods fall short in compatibility with long-term temporal fusion, which is an essential component in cutting-edge camera-only 3D detectors [18, 43]. Long-term

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a) Comparison of various multimodal fusion methods. (b) The misalignment of moving objects.

Figure 1: (a) The existing pipelines require camera and LiDAR backbones, while our pipeline eliminates the need for LiDAR. Using point cloud depth, we directly transform image features into BEV space to create a vision-centric expert. (b) The warping of an object in the historical frame into the current timestamp results in a false position in the current frame due to assuming the object is stationary. The green rectangle represents true positives, while the pink rectangle indicates false positives. x_i denotes the various positions of the object in the historical timestamp.

temporal modeling has shown considerable potential in enhancing the accuracy of depth estimation and detection performance, but it introduces the issue of motion misalignment. Previous methods for BEV distillation have followed two distinct approaches, either distilling the entire BEV space without sufficient attention to the foreground objects [44] or exclusively distilling the foreground object regions [11, 25], thereby overlooking the motion misalignment issue resulting from the long-term temporal fusion. As shown in Fig. 1 (b), this misalignment occurs when past scenes are transformed into the current scene coordinates based solely on ego-motion, assuming all objects are stationary. While in reality, dynamic objects will cause the misalignment, thus interfering with the temporal fusion features. This is more challenging in the case of long-term temporal fusion. Existing methods, such as StreamPETR [51], introduce LayerNorm [1] for dynamic object modeling, but the effects of incorporating velocity and time variables in the model are relatively minor.

To address the aforementioned challenges, we first propose a vision-centric expert, termed as **VCD-E**, which incorporates LiDAR information to enhance the accuracy of depth input. In this context, the term "vision-centric" refers to the utilization of prominent features derived from camera input, distinguishing it from approaches that heavily rely on LiDAR-based features. This model is distinct from conventional multi-modality fusion techniques by eliminating LiDAR backbone. By solely integrating LiDAR depth and long-term temporal fusion under bird's-eye-view (BEV), our model achieves comparable performance to state-of-the-art multi-modal fusion methods [38] by only encoding image modality. As illustrated in Fig. 1 (a), different from previous fusion methods that adopt two modality-specific backbones, we only leverage a single branch to generate semantic features based on image input, and point clouds only provide depth information. Compared to previous fusion methods, our approach eliminates the need for intricate training strategies or specialized fusion module designs while reaching a comparable performance to current fusion methods [38, 33]. More importantly, the vision-centric multi-modal model has the exact same architecture as the camera-only apprentice model, and the generated BEV spatial representation is solely from image features. The domain gap is significantly alleviated through the distillation of knowledge from the proposed multi-modal expert model to the camera-only apprentice model. Due to the advantageous characteristic of domain consistency, our apprentice model acquires substantial benefits from the expert, surpassing previous distillation methods [11, 46].

To mitigate the incompatibility arising from the motion misalignment of dynamic objects, we further propose a trajectory-based distillation module. In this paper, our primary focus revolves around foreground objects, while simultaneously incorporating a meticulous consideration of their historical trajectories. Specifically, by warping dynamic targets from history to the current frame, we derive the motion trajectory associated with each individual object. Then we use the trajectory of each object to query BEV features of the apprentice and expert models respectively. By leveraging the trajectory features of the expert model to optimize the corresponding features of the apprentice model, the latter can acquire the ability to mitigate the interference arising from motion misalignment. In addition, to

enhance the depth perception ability, we diffuse the depth of the foreground part into the 3D space, modeling occupancy to obtain grid-based supervision to assist in depth prediction for objects.

In summary, we propose the multi-modality expert and camera-only apprentice models, termed as VCD-E and VCD-A, respectively. Our contributions are summarized as follows:

- We construct a vision-centric multi-modal expert that solely encodes the image modality, eliminating the need for a LiDAR backbone. For the first time, we demonstrate that the expert can deliver performance on par with other state-of-the-art multi-modal methods while being significantly simpler.
- Due to its homogeneous characteristics and superior performance, the vision-centric expert has been proven effective in distilling knowledge to vision-based models. The effects are significant across a range of model sizes, from compact to more extensive architectures.
- We propose trajectory-based distillation and occupancy reconstruction modules, which supervise both static and dynamic objects to alleviate misalignment during long-term temporal fusion. Combined with the constructed expert model, we enhance the performance of the vision-based models and achieve state-of-the-art on the nuScenes val and test leaderboard.

2 Related Work

In this section, we review previous studies in the areas of 3D object detection, multi-modality fusion, knowledge distillation, focusing on the techniques and methods most relevant to our research.

3D Object Detection. 3D object detection has recently gained significant popularity in the context of autonomous driving and robotics. Detection methods generally fall into two categories: vision-based 3D object detection [64, 65, 5, 37, 52, 36, 53, 58, 22, 56] and LiDAR-based 3D detection [30, 60, 45]. Vision-based approaches [34, 54, 15, 35] exploit image information and frequently involve deep learning techniques to estimate depth. In contrast, LiDAR-based methods capitalize on precise geometric information from LiDAR sensors to achieve superior object detection accuracy. Our proposed vision-centric expert shares the same modality as vision-based detectors, while exhibiting superior performance compared to LiDAR-based detectors.

Long-term modeling has been employed to improve the performance of 3D object detection models [18, 43]. SOLOFusion [43] utilizes long-term temporal modeling to achieve excellent performance. VideoBEV [18] maintains comparable performance with SOLOFusion while being more efficient, using long-term recurrent temporal modeling. However, previous research [21, 51] has highlighted that long-term temporal fusion can lead to inadequate detection of dynamic objects. Although StreamPETR [51] proposes the propagation transformer, [48, 7, 9, 8] to conduct object-centric temporal modeling, the improvement of dynamic object modeling remains relatively modest. Our proposed trajectory-based distillation module alleviates this limitation, enabling accurate detection of both static and dynamic objects by camera-based 3D object detection models that utilize long-term modeling.

Multi-modality Fusion. Multi-modality fusion techniques [49, 2, 33, 16] have been extensively investigated to enhance 3D object detection performance by integrating complementary information from different sensor modalities, such as cameras and LiDAR sensors. These methods [38, 63] typically require complex training strategies and the development of specialized fusion modules to effectively merge the distinct sources of information. In contrast, we propose a streamlined architecture that utilizes an image backbone for feature extraction, obviating the need for a LiDAR backbone. This efficient approach augments vision-based models by incorporating LiDAR information, maintaining homogeneity with vision-based models while preserving exceptional performance.

Knowledge Distillation. Knowledge distillation [44, 46] is a technique facilitating the transfer of knowledge from a larger, more complex model (expert) to a smaller, more efficient model (apprentice). This approach has been successfully applied in various domains, including image classification [14], natural language processing [39], and policy learning [55, 27, 26, 6]. Therefore, recent distillation methods [17, 25, 32, 24, 29] build upon 3D object detection aims to transfer the accurate geometry knowledge from LiDAR to camera. MonoDistill [12] projects the LiDAR points into the image plane to serve as the expert to transfer knowledge. BEVSimDet [52] simulates fusion-based methods to alleviate the domain gap between the two different modalities. BEVDistill [11] projects the LiDAR points and images into the BEV space to align the LiDAR feature and image feature. Due to the

non-homogenous nature between the LiDAR and camera, transferring knowledge from LiDAR to images is challenging. Instead, our work constructs a vision-centric expert, which possesses a homogeneous modality with vision-based models. The vision-centric expert can leverage knowledge distillation to transfer the geometric perception capabilities to various vision-based models, hence enhancing performance accordingly.

3 Method

In this section, we present our approach in detail. The overall architecture is presented in Sec. 3.1. Our method involves two main components: (1) the vision-centric expert in Sec. 3.2, and (2) the trajectory-based distillation and occupancy reconstruction modules as elaborated in Sec. 3.3. The pipeline of our method is depicted in Fig. 2.

3.1 Overall Architecture

In this paper, we construct a pair of harmonious expert and apprentice models. The expert and apprentice models adopt the consistent model architecture. The only difference is that the expert additionally leverages the accurate depth map generated from the point cloud, while the apprentice model predicts the depth map from the image. Although our expert model only uses an image backbone to encode high-level scene information, it is on par with state-of-the-art multi-model fusion methods that use several modality-speci c backbones and complex interaction strategies. More importantly, we eliminate the domain gap between the multi-model expert and the camera-only apprentice model, which is deemed as one of the most challenging topics in the cross-modality distillation literature.

As illustrated in Fig. 2, we construct a distillation framework between the expert network and the apprentice network. The vision-centric expert fuses feature F^E_{img} extracted from the image backbone and the temporal depth map D projected from LiDAR points to create a uni ed BEV representation F^E used for 3D object detection. Therefore, although we adopt a cross-modality approach for 3D object detection, the resulting representation remains homogeneous with image modality features.

After obtaining the pretrained vision-centric expert and corresponding apprentice network, we freeze the expert network and leverage its intermediate features as auxiliary supervision for the apprentice network. Since current advanced vision-based detectors employ long-term temporal modeling to attain state-of-the-art performance, we utilize a standard long-term temporal vision-based detector based on BEVDepth as the apprentice model in our context. The expert model also utilizes long-term temporal modeling to ensure consistency and achieve higher performance.

3.2 The Generation of Expert Model

We construct a vision-centric expert by integrating LiDAR information as accurate depth input into a vision-based model. Nevertheless, given the sparsity of LiDAR depth data, we rely on the predicted depth values obtained from images for pixels lacking LiDAR depth information. We also utilize future frames to further improve the performance of the expert model in the of ine 3D detection setting. The vision-based detector serves as the primary model, and LiDAR information complements it by providing precise depth information. This approach eliminates the need for complex training strategies or custom-designed fusion modules, streamlining the fusion process.

For the expert model, we project the last sweep LiDAR frame with the current LiDAR frame onto images to obtain corresponding depth maps D. Since the depth maps D generated from point clouds can not cover every pixel of the images, we also predict depth distribution for each pixel based on the image features. Then we will project the image features onto the BEV space to obtain BEV features F^E based on their depth. Furthermore, we transform the BEV features F^E_{t-N} from previous timestamps to the current BEV features F^E_t, modeling the long-term relationship. N denotes the time interval between the current frame and the history frame. The uni ed BEV features F^E are then combined to produce 3D object detection predictions P. The model is trained using a multi-task loss function that considers both 3D detection loss L_{Det} and depth estimation loss L_{Depth} .

Figure 2: Algorithm Overview. Expert utilizes LiDAR data to enhance depth estimation accuracy before view transformation in BEV pipeline. Apprentice represents a standard long-term vision-based detection model. The occupancy is built using the depth information from the expert model, which serves as the supervision for the apprentice model. Motion trajectory is constructed by warping the time series GT query into the current timestamp. Projecting the motion trajectory of each object into BEV space can rectify the misalignment of object motion. With the knowledge transferred from the expert, the apprentice can deliver higher performance than before.

3.3 The Procedure of Distillation

In this section, we elucidate the methodology adopted to overcome the constraints inherent in long-term modeling for multi-camera 3D object detection. This is achieved through the incorporation of two innovative modules within the distillation process: the trajectory-based distillation module and the occupancy reconstruction module.

Trajectory-based Distillation. The trajectory-based distillation module aims to improve the detection of dynamic objects by focusing on the inconsistent portion of the objects' motion. For i-th historical frame at timestamp t_i that contains K objects, extract the j-th ground truth object position $P_i^j = (x_i^j, y_i^j, z_i^j, 1)^T$ in the ego coordinate system. Determine the actual ego motion matrix M_i between the current frame t_0 and each historical frame t_i. Apply the ego-motion transformation matrix M_i to the ground truth object positions P_i^j to obtain the transformed positions $P_i^{j\,0}$ in the current frame's coordinate system:

$$P_i^{j\,0} = M_i \quad P_i^j : \tag{1}$$

We amalgamate the transformed ground truth object positions, $P_i^{j\,0}$, from all historical frames to construct the motion trajectories. The trajectory of j-th object can thus be represented as a sequence of object positions within the current frame as $(P_1^{j\,0}, P_2^{j\,0}, : : : , P_N^{j\,0})$, where N is the number of points on each motion trajectory. Let F_{ij}^E, F_{ij}^A represent the sampled features on identical point $P_i^{j\,0}$ from the expert BEV feature F^E and apprentice BEV feature F^A, respectively. They are sampled via bilinear interpretation and then are normalized as :

$$F_{ij}^E = \mathrm{norm}(F^E(P_i^{j\,0})); \quad F_{ij}^A = \mathrm{norm}(F^A(P_i^{j\,0})): \tag{2}$$

Finally, the trajectory-based distillation loss L_{TD} is computed between the normalized key sampled features:

$$L_{TD} = \frac{1}{K \quad N} \sum_{j=1}^{K} \sum_{i=1}^{N} L_2(F_{ij}^A, F_{ij}^E): \tag{3}$$

Ultimately, by using the motion trajectory as queries, we conduct trajectory-based distillation on these representative positions. This approach enables the expert to rectify the motion misalignment in the apprentice.

Table 1: Comparison among the camera-only methods on the nuScenes val set. denotes the long-term baseline implemented by us based on BEVDet4D-Depth [22]. y depicts that the size of BEV feature is 256 256. VCD-A surpasses previous SOTA by 2 points in NDS and achieves SOTA under the same setting.

Methods	Backbone	Image Size		Frames	mAP"	NDS"	mATE#	mASE#	mAOE#	mAVE#	mAAE#
BEVDet [23]	ResNet-50	256	704	1	0.298	0.379	0.725	0.279	0.589	0.860	0.245
PETR [39]	ResNet-50	384	1056	1	0.313	0.381	0.768	0.278	0.564	0.923	0.225
BEVDet4D [22]	ResNet-50	256	704	2	0.322	0.457	0.703	0.278	0.495	0.354	0.206
BEVDepth [35]	ResNet-50	256	704	2	0.351	0.475	0.639	0.267	0.479	0.428	0.198
BEVStereo [34]	ResNet-50	256	704	2	0.372	0.500	0.598	0.270	0.438	0.367	0.190
STS [54]	ResNet-50	256	704	2	0.377	0.489	0.601	0.275	0.450	0.446	0.212
VideoBEV [19]	ResNet-50	256	704	8	0.422	0.535	0.564	0.276	0.440	0.286	0.198
SOLOFusion [43]	ResNet-50	256	704	16+1	0.427	0.534	0.567	0.274	0.411	0.252	0.188
StreamPETR [51]	ResNet-50	256	704	8	0.432	0.540	0.581	0.272	0.413	0.295	0.195
Baseline	ResNet-50	256	704	8+1	0.401	0.515	0.595	0.279	0.489	0.291	0.198
VCD-A	ResNet-50	256	704	8+1	0.426	0.540	0.547	0.271	0.433	0.268	0.207
Baseline^{y}	ResNet-50	256	704	8+1	0.418	0.542	0.522	0.267	0.428	0.262	0.188
VCD-A^{y}	ResNet-50	256	704	8+1	0.446	0.566	0.497	0.260	0.350	0.257	0.203

Occupancy Reconstruction. The expert model demonstrates outstanding performance in 3D object detection, resulting in a more precise 3D geometric representation of the objects. We utilize depth estimation modules to predict the depth, denoted $D(u, v)$, for each image pixel (u, v). Subsequently, the depth map D is back-projected into a 3D point cloud, each image pixel (u, v) is transformed into a 3D coordinate (x, y, z). Then the scores from different pixels in the multi-camera within the same occupancy region, are accumulated to make decisions about the presence of an object in that region. In this way, we can generate the occupancy O_e and O_a for expert and apprentice model, respectively.

The occupancy reconstruction module improves the model's capability to discern the 3D geometric properties of objects. The grid-based supervisory signals effectively direct the model to enhance its prediction accuracy for object depths. Different voxels within the occupancy structure aggregate the fused depth distribution from various perspective views, thereby enhancing robustness against depth errors. Drawing inspiration from CenterPoint [61], we simply extend the Gaussian distribution applied to each target into the 3D space for more focused 3D object modeling:

$$G_{xyz} = \exp\left(\frac{(x - p_x)^2 + (y - p_y)^2 + (z - p_z)^2}{2\sigma_p^2}\right); \quad (4)$$

where (p_x, p_y, p_z) represents the center of the 3D object, while σ_p denotes the standard deviation of each object's size. The model utilizes the expert model's occupancy O_e as supplementary supervision by adopting a straightforward l_1 regularization loss for the occupancy status O_a of the apprentice model, which optimizes depth prediction capabilities for both static and dynamic objects. The occupancy reconstruction loss can be formulated as

$$L_{OR} = L_1(G_{xyz} \cdot O_e, G_{xyz} \cdot O_a): \quad (5)$$

Training Loss. During the distillation phase, the joint training loss L_{total} is formulated as

$$L_{Total} = L_A + \alpha_1 L_{TD} + \alpha_2 L_{OR}; \quad (6)$$

where L_A is the perceptual loss of the apprentice model. Besides, α_1 and α_2 represent hyperparameters employed to effectively balance the scales of the respective loss functions. The utilization of trajectory-based distillation loss L_{TD} and occupancy reconstruction loss L_{OR} collectively facilitates the transfer of semantic and geometry knowledge from the expert model to the apprentice model.

4 Experiments

In this section, we outline the experimental setup and assess our proposed VCD-E model, as well as the newly introduced trajectory-based distillation and occupancy reconstruction modules, using the nuScenes dataset [3]. This includes a presentation of the evaluation metrics, baseline models, ablation studies, and a comparative analysis of our approach with the current state-of-the-art methods.

Table 2: Comparison among the camera-only methods on the nuScenes test set. Methods marked with † denote long-term baseline implemented by us, based on BEVDet4D-Depth [22]. ‡ depicts test time augmentation adopted during the inference phase. VCD-A achieves SOTA under critical metrics and surpasses its baseline by 2 points in NDS.

Methods	Backbone	Image Size		mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FCOS3D† [52]	R101-DCN	900	1600	0.358	0.428	0.690	0.249	0.452	1.434	0.124
DETR3D† [53]	V2-99	900	1600	0.412	0.479	0.641	0.255	0.394	0.845	0.133
UVTR [33]	V2-99	900	1600	0.472	0.551	0.577	0.253	0.391	0.508	0.123
BEVDet4D† [22]	Swin-B [41]	900	1600	0.451	0.569	0.511	0.241	0.386	0.301	0.121
BEVFormer [36]	V2-99	900	1600	0.481	0.569	0.582	0.256	0.375	0.378	0.126
PolarFormer [28]	V2-99	900	1600	0.493	0.572	0.556	0.256	0.364	0.439	0.127
BEVDistill [11]	ConvNeXt-B	900	1600	0.496	0.594	0.475	0.249	0.378	0.313	0.125
PETRv2 [40]	RevCol [4]	640	1600	0.512	0.592	0.547	0.242	0.360	0.367	0.126
BEVDepth [35]	ConvNeXt-B	640	1600	0.520	0.609	0.445	0.243	0.352	0.347	0.127
AeDet† [15]	ConvNeXt-B	640	1600	0.531	0.620	0.439	0.247	0.344	0.292	0.130
SOLOFusion [43]	ConvNeXt-B	640	1600	0.540	0.619	0.453	0.257	0.376	0.276	0.148
StreamPETR [51]	ConvNeXt-B	640	1600	0.550	0.631	0.493	0.241	0.343	0.243	0.123
Baseline	ConvNeXt-B	640	1600	0.522	0.610	0.457	0.253	0.391	0.271	0.142
VCD-A	ConvNeXt-B	640	1600	0.548	0.631	0.436	0.244	0.343	0.290	0.120

Table 3: Comparison among the multi-modality methods on the nuScenes val set. Our proposed VCD-E only adopts image backbone while achieving comparable performance with state-of-the-art multi-modal methods who predominately rely on the LiDAR backbone.

Methods	Venue	Backbone	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVFusion [38]	NeurIPS 2022	LiDAR & Image	0.642	0.680	-	-	-	-	-
FUTR3D [10]	Arxiv 2022	LiDAR & Image	0.645	0.683	-	-	-	-	-
UVTR [33]	NeurIPS 2022	LiDAR & Image	0.654	0.702	0.332	0.258	0.268	0.212	0.177
CMT [57]	Arxiv 2023	LiDAR & Image	0.679	0.708	-	-	-	-	-
VCD-E	-	Image	0.677	0.711	0.308	0.254	0.317	0.189	0.201

4.1 Main Results

Camera-only 3D detection. In order to evaluate the effectiveness of our proposed Revenge model and the trajectory-based distillation modules, we have conducted rigorous experiments on the nuScenes validation and test sets. As presented in Tab. 1, the performance of VCD-A surpasses other cutting-edge methods, achieving a record of 44.6% and 56.6% on the nuScenes benchmark. This provides robust evidence of the effectiveness of our approach. Utilizing an image resolution of 256 704 and a ResNet-50 backbone, VCD outperforms the state-of-the-art method, DORT [46], by improving mAP by 1.9% and NDS by 3.2%. Additionally, VCD-A exhibits an improvement over our baseline by 2.8% mAP and 2.4% NDS, thereby indicating that our methods can considerably advance state-of-the-art results. To further ascertain the generalizability of our methods, we conducted experiments on the nuScenes test set, as shown in Tab. 2. With the adoption of the ConvNext-B backbone [42], VCD achieved 54.8% mAP and 63.1% NDS, outperforming the state-of-the-art detector, SOLOFusion [43], by 0.8% mAP and 1.2% NDS. Furthermore, our proposed approach led to a 2.6% increase in mAP and a 2.1% improvement in NDS compared to our baseline, thereby demonstrating the efficacy of our method for large-scale vision-based models.

Multi-modality Fusion. We first compare our proposed vision-centric expert with other state-of-the-art multi-modality fusion methods. VCD-E does not require any complex fusion strategies or three-stage training schedule, yet it achieves 67.7% mAP and a 71.1% NDS as shown in Tab. 3. Despite employing only a single image backbone, the model's performance is comparable to state-of-the-art methods such as BEVFusion [38], and it even surpasses UVTR [33] by 2.3% mAP and 0.9% NDS, which utilizes two backbones, by a considerable margin. Owing to its homogeneous nature and compatibility with camera-only models, the VCD-E can serve as a powerful expert for knowledge transfer, facilitating performance improvements for small to large vision detectors.

4.2 Ablation Study

To verify the effectiveness and necessity of each component, we conduct various ablation experiments on the nuScenes validation set.

Table 4:Ablation study of the proposed distillation framework on different temporal lengths. The design works with different length of time windows and the gain grows with the temporal length.

Temporal Length	Distill	mAP (%)"	NDS (%)"	mATE#	mAOE#	mAVE#
1	7	26.6	37.9	0.815	0.645	0.556
	3	30.1(+3.5)	41.5(+3.6)	0.732	0.629	0.476
2	7	26.9	38.4	0.804	0.706	0.461
	3	31.3(+4.4)	43.2(+4.8)	0.717	0.615	0.403
4	7	28.4	39.8	0.748	0.739	0.432
	3	33.0(+4.6)	44.1(+4.3)	0.707	0.632	0.389
8	7	29.7	40.9	0.762	0.714	0.415
	3	35.4(+5.7)	45.9(+5.0)	0.690	0.625	0.370

Table 5:The performance gains of the apprentice which bene t from different experts. CM denotes cross-modal and UM represents Uni-modal. It indicates the success of uni-modal distillation remains.

Expert	Paradigm	mAP	NDS
-	-	0.297	0.409
CenterPoint [60]	CM	0.281	0.420
Transfusion [2]	CM	0.292	0.435
BEVDepth [35]	UM	0.341	0.442
VCD-E	UM	0.354	0.459

Table 6:Effect of different distillation methods. All models are trained with VCD-E as expert. Our proposal signi cantly surpasses previous SOTA methods.

Methods	mAP	NDS
Baseline [35]	0.297	0.409
FitNet [44]	0.318	0.421
CWD [46]	0.311	0.412
BEVDistill [11]	0.316	0.439
VCD-A	0.354	0.459

Effectiveness of the General Framework. We rst validate the effectiveness of the general framework for various temporal lengths. In Tab. 4, we examine different timestamps for temporal modeling and nd that our method signi cantly outperforms the baseline. By employing the general framework, we achieve 5.0% NDS and 5.7% mAP performance gains with 8 temporal modeling instances. For other temporal modeling scenarios, our method continues to exhibit substantial performance improvements, ranging from 3.6% to 5.0% NDS. As the duration of temporal fusion extends, the advantages of the model become increasingly pronounced, suggesting that this framework exhibits higher compatibility with extended sequences. This indicates the effectiveness of our proposed trajectory-based distillation module, which alleviates the issue of motion misalignment.

Effectiveness of VCD-E. In this study, it is crucial to verify the effectiveness of our proposed vision-centric experts. We select various expert models for a fair comparison, including LiDAR-based expert Centerpoint[60], vision-based expert BEVDepth[35], and fusion-based expert Transfusion, based on our proposed distillation strategy. In Tab. 5, we observe that homogeneous expert models signi cantly in uence the success of knowledge transfer. The camera-based expert BEVDepth demonstrates superior performance gains compared to the other two heterogeneous expert models. Our advanced model, VCD-E, operates within the same modality as vision-based models, and it is equipped with precise geometric information. Consequently, the distillation effect outperforms the Transfusion[2] expert by a signi cant margin, achieving a 6.2% increase in mAP. Moreover, it surpasses the BEVDepth expert by an additional 1.7% in NDS. This demonstrates the superior performance and effectiveness of our model.

Comparison to SOTA. To further demonstrate the effectiveness of our proposed distillation strategy, we compare our method with other state-of-the-art methods. To ensure a fair comparison, we conduct all experiments based on our vision-centric expert VCD-E. We nd that our method consistently outperforms other distillation methods by a signi cant margin. In 2D detection, FitNet[44] and CWD [46] are two classic distillation methods, we adapt them for 3D object detection to facilitate a fair comparison. BEVDistill[11] represents a state-of-the-art distillation method for multi-view 3D object detection, and we compare our results with this approach as well. As shown in Tab. 6, our method achieves the best results when compared to these state-of-the-art distillation strategies, demonstrating the effectiveness of our approach. Our method surpasses BEVDistill by 2% NDS and 3.8% mAP.

Table 7: Gains of different image backbone on multi-modal models [18]. The stronger backbone still demonstrates better performance in this case.

Methods	Backbone	mAP	NDS
BEVFusion	ResNet-50	0.598	0.662
BEVFusion	ConvNext-B	0.597	0.665
VCD-E	ResNet-50	0.611	0.656
VCD-E	ConvNext-B	0.664	0.693

Table 8: Gains of different depth fusion strategy. The proposed fusion depth is optimal among different methods.

Methods	mAP	NDS
Predicted depth	0.495	0.585
LiDAR depth	0.638	0.687
Weighted depth	0.644	0.690
Fusion depth	0.646	0.690

Table 9: The performance gains of different trajectory length for trajectory-based distillation. As the trajectory length increases, the bene ts derived from the distillation process become more pronounced.

Trajectory Length	Distill	mAP (%)	NDS (%)
-	7	29.7	40.9
0	3	31.8	42.1
1	3	33.1	44.5
3	3	34.6	45.6
5	3	35.4	45.9
9	3	33.9	44.7

Gains of the Image Backbone. To verify that VCD-E bene ts from contemporary image backbones, we conducted experiments outlined in Tab. 7. We selected ResNet-50 and ConvNext-B as two distinct modern image backbones for BEVFusion and VCD-E. Our ndings indicate that VCD-E achieves substantial improvements of 5.3% mAP and 3.7% NDS when using ConvNext-B compared to ResNet-50, demonstrating that VCD-E can indeed bene t from modern image backbones. However, existing fusion-based methods, such as BEVFusion, which primarily rely on the capabilities of LiDAR backbones, show limited gains from employing modern image backbones.

Fusion Strategy of Expert. To determine the optimal fusion strategy for LiDAR points and images, we propose four distinct depth fusion approaches, including (1) Predicted depth that directly uses predicted depth based on image features. (2) LiDAR depth that utilizes the projected depth from LiDAR points. (3) Fusion depth that employs the projected depth when corresponding LiDAR points exist for each pixel, otherwise using the predicted scores based on image features, and (4) Weighted depth. Applying the weighted average depth of the predicted depth and LiDAR depth. As demonstrated in Tab. 8, fusion depth yields the best results which indicates that using corresponding predicted scores where LiDAR depth is unavailable is advantageous.

The Effectiveness of Trajectory-based Distillation The results presented in Table 9 indicate that as the trajectory length increases, the bene ts derived from the distillation process become more pronounced. The temporal fusion length for this experiment is set at eight. The rst row denotes the baseline model which does not use the distillation method. The second row depicts that the VCD-A model directly conducts distillation under the full BEV feature without the Trajectory-based distillation module. When the trajectory length is set to 1, we only distill ground truth locations in the current frame. However, when the trajectory length exceeds ve, there is a noticeable decrease in accuracy. We hypothesize that this decrease may be attributed to the model's distracted attention towards distant motions. The density of traf c can lead to distant motion locations being occupied by other objects, which may not necessarily require additional trajectory supervision. This suggests that the application of excessive trajectory supervision in such scenarios could be unnecessary and inef cient.

Ablation Study of Each Component. In Tab. 10, we conduct an ablation study on the components employed in VCD to verify their contributions to the nal result. These components include (1) Long-term temporal fusion, which allows our model to bene t from historical information, thereby enhancing its performance. (2) Trajectory-based distillation, which transfers knowledge from VCD-E to VCD-A in areas with motion misalignment to mitigate this issue. (3) Occupancy reconstruction,

Table 10:Ablation study of important components in VCD-A. It shows that the motion module contributes most to the performance improvement, indicating the potential of future research here.

Methods	Long	Occ	Motion	mAP"	NDS"	mATE#	mAVE#
Baseline				0.271	0.319	0.767	1.065
Longer Temporal Fusion	3			0.297	0.409	0.762	0.415
Trajectory-based Distillation			3	0.283	0.356	0.772	0.870
All but Longer Temporal		3	3	0.290	0.367	0.724	0.834
Occupancy Reconstruction	3	3		0.308	0.416	0.734	0.379
VCD	3	3	3	0.341	0.449	0.715	0.389

which serves as dense depth supervision in the perspective view, improving the performance of VCD-A. Long-term temporal modeling enables our model to reference prior information from history, assisting in object velocity estimation and resulting in a signi cantly lower mATE. We also evaluate the impact of the trajectory-based distillation module when all other components are incorporated. As shown in Tab. 10, trajectory-based distillation increases NDS by 3.3% and mAP by 3.3%, contributing to the majority of the improvement.

5 Conclusion

In this paper, we presented a novel vision-centric model as the expert, which leverages the strengths of both modalities, our proposed model achieves exceptional performance, rivaling that of state-of-the-art multimodal fusion models, while also mitigating domain gap issues, making it suitable for distilling vision-based models. Furthermore, we introduced two innovative modules, the trajectory-based distillation and occupancy reconstruction modules. These modules enhance the geometric perception capabilities of multi-camera 3D object detection models and improve the detection of both static and dynamic objects in the scene. Our experiments demonstrate the effectiveness of our proposed methods on the widely-used nuScenes 3D object detection benchmark.

Limitation. In this work, we have not delved into more details of the vision-centric expert, which may hold signi cant potential for improvement. Additionally, we have not explored further applications, such as automatic dimension estimation based on VCD-E.

Acknowledgements

This work was supported by National Key R&D Program of China (2022ZD0160104) and NSFC (62206172). We would like to thank anonymous reviewers for active discussions.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016. 2

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust lidar-camera fusion for 3d object detection with transformers. In CVPR, 2022. 1, 3, 8

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In CVPR, 2020. 6, 14

[4] Yuxuan Cai, Yizhuang Zhou, Qi Han, Jianjian Sun, Xiangwen Kong, Jun Li, and Xiangyu Zhang. Reversible column networks. In ICLR, 2023. 7

[5] Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In ECCV, 2022. 3

[6] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. arXiv preprint arXiv:2306.16927, 2023. 3

[7] Mengzhao Chen, Mingbao Lin, Ke Li, Yunhang Shen, Yongjian Wu, Fei Chao, and Rongrong Ji. Cf-vit: A general coarse-to- ne method for vision transformer. AAAI, 2023. 3

[8] Mengzhao Chen, Mingbao Lin, Zhihang Lin, Yuxin Zhang, Fei Chao, and Rongrong Ji. Smmix: Self-motivated image mixing for vision transformers. ICCV, 2023. 3

[9] Mengzhao Chen, Wenqi Shao, Peng Xu, Mingbao Lin, Kaipeng Zhang, Fei Chao, Rongrong Ji, Yu Qiao, and Ping Luo. Diffrate: Differentiable compression rate for ef cient vision transformers. arXiv preprint arXiv:2305.17997, 2023. 3

[10] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A uni ed sensor fusion framework for 3d detection. arXiv preprint arXiv:2203.10642, 2022. 7

[11] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. arXiv preprint arXiv:2211.09386, 2022. 2, 3, 7, 8

[12] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. MonoDistill: Learning spatial features for monocular 3d object detection. arXiv preprint arXiv:2201.10830, 2022. 3

[13] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d , 2020. 14

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In CVPR, 2009. 3

[15] Chengjian Feng, Zequn Jie, Yujie Zhong, Xiangxiang Chu, and Lin Ma. Aedet: Azimuth-invariant multi-view 3d object detection. arXiv preprint arXiv:2211.12501, 2022. 3, 7

[16] Yulu Gao, Chonghao Sima, Shaoshuai Shi, Shangzhe Di, Si Liu, and Hongyang Li. Sparse dense fusion for 3d object detection. In IROS, 2023. 3

[17] Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. LIGA-Stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In ICCV, 2021. 3

[18] Chunrui Han, Jianjian Sun, Zheng Ge, Jinrong Yang, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. Exploring recurrent long-term temporal fusion for multi-view 3d perception. arXiv preprint arXiv:2303.05970, 2023. 1, 3

[19] Chunrui Han, Jianjian Sun, Zheng Ge, Jinrong Yang, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. Exploring recurrent long-term temporal fusion for multi-view 3d perception. arXiv preprint arXiv:2303.05970, 2023. 6

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016. 14

[21] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In CVPR, 2023. 1

[22] Junjie Huang and Guan Huang. BEVDet4D: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054, 2022. 3, 6, 7

[23] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. BEVDet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790, 2021. 6

[24] Linyan Huang, Huijie Wang, Jia Zeng, Shengchuan Zhang, Liujuan Cao, Rongrong Ji, Junchi Yan, and Hongyang Li. Geometric-aware pretraining for vision-centric 3d object detection. arXiv preprint arXiv:2304.03105, 2023. 3

[25] Peixiang Huang, Li Liu, Renrui Zhang, Song Zhang, Xinli Xu, Baichao Wang, and Guoyi Liu. Tig-bev: Multi-view bev 3d object detection via target inner-geometry learning. arXiv preprint arXiv:2212.13979, 2022. 2, 3

[26] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In ICCV, 2023. 3

[27] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In CVPR, 2023. 3

[28] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. arXiv preprint arXiv:2206.15398, 2022. 7

[29] Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, and Fatih Porikli. X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. arXiv preprint arXiv:2303.02203, 2023. 3

[30] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. CVPR, 2019. 3

[31] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Enze Xie, Zhiqi Li, Hanming Deng, Hao Tian, et al. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. arXiv preprint arXiv:2209.05324, 2022. 1

[32] Jianing Li, Ming Lu, Jiaming Liu, Yandong Guo, Li Du, and Shanghang Zhang. Bev-lgkd: A uni ed lidar-guided knowledge distillation framework for bev 3d object detection. arXiv preprint arXiv:2212.00623, 2022. 1, 3

[33] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. arXiv preprint arXiv:2206.00630, 2022. 1, 2, 3, 7

[34] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. arXiv preprint arXiv:2209.10248, 2022. 3, 6

[35] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3d object detection. arXiv preprint arXiv:2206.10092, 2022. 3, 6, 7, 8, 14

[36] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. arXiv preprint arXiv:2203.17270, 2022. 3, 7

[37] Zhuoling Li, Chuanrui Zhang, Wei-Chiu Ma, Yipin Zhou, Linyan Huang, Haoqian Wang, SerNam Lim, and Hengshuang Zhao. Voxelformer: Bird's-eye-view feature generation based on dual-view attention for multi-view 3d object detection. arXiv preprint arXiv:2304.01054, 2023. 3

[38] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. BEVFusion: A simple and robust lidar-camera fusion framework. arXiv preprint arXiv:2205.13790, 2022. 2, 3, 7, 9

[39] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3d object detection. arXiv preprint arXiv:2203.05625, 2022. 6

[40] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETRv2: A uni ed framework for 3d perception from multi-camera images. arXiv preprint arXiv:2206.01256, 2022. 7

[41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. ICCV, 2021. 7

[42] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. IJCVPR, 2022. 7, 14

[43] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. arXiv preprint arXiv:2210.02443, 2022. 1, 3, 6, 7, 15

[44] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550, 2014. 2, 3, 8

[45] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3d object detection. arXiv preprint arXiv:2102.00463, 2021. 3

[46] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In ICCV, 2021. 2, 3, 8

[47] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, and Hongyang Li. Scene as occupancy. In ICCV, 2023. 1

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017. 3

[49] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. PointPainting: Sequential fusion for 3d object detection. In CVPR, 2020. 3

[50] Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Yuting Wang, Shengyin Jiang, Peijin Jia, Bangjun Wang, Feng Wen, Hang Xu, Ping Luo, Junchi Yan, Wei Zhang, and Hongyang Li. Openlane-v2: A topology reasoning benchmark for scene understanding in autonomous driving, 2023. 1

[51] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for ef cient multi-view 3d object detection. arXiv preprint arXiv:2303.11926, 2023. 2, 3, 6, 7

[52] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3d object detection. In ICCV, 2021. 3, 7

[53] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3d object detection from multi-view images via 3d-to-2d queries. In CoRL, 2022. 3, 7

[54] Zengran Wang, Chen Min, Zheng Ge, Yinhao Li, Zeming Li, Hongyu Yang, and Di Huang. STS: Surround-view temporal stereo for multi-view 3d detection. arXiv preprint arXiv:2208.10145, 2022. 3, 6

[55] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In Advances in Neural Information Processing Systems (NeurIPS), 2022. 3

[56] Kaixin Xiong, Shi Gong, Xiaoqing Ye, Xiao Tan, Ji Wan, Errui Ding, Jingdong Wang, and Xiang Bai. Cape: Camera view position embedding for multi-view 3d object detection. In CVPR, pages 21570–21579, June 2023. 3

[57] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer via coordinates encoding for 3d object dected ion. arXiv preprint arXiv:2301.01283, 2023. 7

[58] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. arXiv preprint arXiv:2211.10439, 2022. 3

[59] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. DeepInteraction: 3d object detection via modality interaction. arXiv preprint arXiv:2208.11112, 2022. 1, 3

[60] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In CVPR, 2021. 1, 3, 8

[61] Jia Zeng, Li Chen, Hanming Deng, Lewei Lu, Junchi Yan, Yu Qiao, and Hongyang Li. Distilling focal knowledge from imperfect expert for 3d object detection. In CVPR, 2023. 1

[62] Haimei Zhao, Qiming Zhang, Shanshan Zhao, Jing Zhang, and Dacheng Tao. Bevsimdet: Simulated multi-modal distillation in bird's-eye view for multi-view 3d object detection. arXiv preprint arXiv:2303.16818, 2023. 3

[63] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. arXiv preprint arXiv:1904.07850, 2019. 6

[64] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In CVPR, 2021. 3

[65] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monoef: Extrinsic parameter free monocular 3d object detection. In IEEE TPAMI, 2022. 3

A Experiment Details

A.1 Dataset and Evaluation Metrics

We conduct our experiments on the nuScenes dataset [2], a widely used benchmark for autonomous driving tasks. The dataset encompasses diverse driving scenarios captured using cameras and LiDAR sensors, offering rich information for both visual and LiDAR-based 3D object detection. The dataset comprises 700 training scenes, 150 validation scenes, and 150 testing scenes. Each scene spans approximately 20 seconds, with key frames annotated at a 2 Hz frequency.

The two dominant metrics for the nuScenes detection task are the nuScenes Detection Score (NDS) and mean Average Precision (mAP). The mAP for nuScenes is computed based on the center distance between predictions and ground truth annotations on the ground plane. Moreover, the nuScenes dataset de nes ve true positive metrics (mATE, mASE, mAOE, mAVE, mAAE) for measuring translation, scale, orientation, velocity, and attribute, respectively. The NDS for nuScenes is a weighted sum of mAP and the ve true positive metrics, de ned as: $\text{NDS} = \frac{1}{10}[5\text{mAP} + \sum_{\text{mTP}} (1 - \min(1, \text{mTP}))]$.

A.2 Implementation Details

We conduct experiments on BEVDepth [35]. The codebase is developed upon MMDetection3D [3]. Main experiments are trained on 8 NVIDIA A100 GPUs, while ablation experiments are conducted on 8 NVIDIA V100 GPUS. For BEVDepth, the model is trained for 20 epochs with an initial learning rate of 2e-4. In the distillation process, the per-GPU batch size is set to 4, whereas during the training of the baseline model, it is set to 8. Normal data augmentations are introduced in the training process such as ip and rotate. In our apprentice models, future frames are not incorporated into the long-term temporal fusion throughout the training phase to ensure a fair comparison. As for ablation study, we conduct experiments with an online training strategy, and we have employed the ResNet-50 con guration in the absence of CBGS.

In our research, we implement distinct temporal modeling strategies for both apprentice and expert models. For the apprentice models, we incorporate a sequence of eight frames into the temporal modeling process. In contrast, the expert models integrate four future frames into the temporal modeling as demonstrated in our primary results. However, in our ablation study, we deviate from this approach and instead employ eight historical frames for temporal modeling. Besides, We use the last sweep LiDAR frame with the current LiDAR frame to create the depth map.

Table 11: Experiment settings.denotes that the training schedule for VCD-E is approximately one-fourth of the original schedule. This reduction was implemented to expedite the training process during the ablation study. The rst group is engaged in training on the main results, whereas the second group is utilized in the ablation study.

Method	Backbone	Image Size		Frames	mAP (%)	NDS (%)
VCD-E	ConvNext-B [42]	512	1408	8+1	67.7	71.1
VCD-A	Res-50 [20]	256	704	8+1	41.8	54.2
VCD-E	ConvNext-B [42]	256	704	8+1	54.2	58.8
VCD-A	Res-50 [20]	256	704	8+1	29.7	40.9

A.3 Experiments Settings

The setting of adopted expert-apprentice pairs is depicted in Tab. 11. We categorize the distillation setting into two distinct groups. The primary group is engaged in training on the main results, whereas the second group is utilized for the ablation study. Both our baseline and VCD-A adopt 8 history frames for temporal fusion.

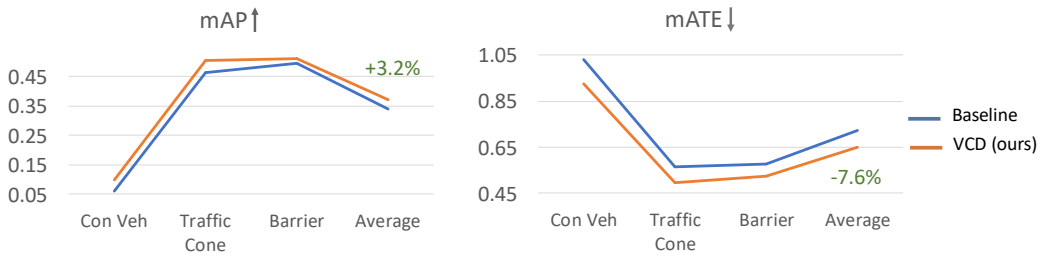


Figure 3: Effects of VCD on static objects. Our distillation framework VCD can still consistently improves static objects, demonstrating 3.2% and 7.6% improvements in precision-recall (mAP) and localization quality (mATE), respectively.

B The Analysis of Temporal Fusion

B.1 The Misalignment of Motion Objects

As highlighted in preceding studies [43], long-term temporal fusion may face misalignment issues in motion estimation, which can be discerned through a reduction in performance on metrics like mATE. Let's consider a moving object and analyze the impact of inaccurate motion estimation on its position in the fused frame. We will assume that the environment is static, except for the moving object. Let the position of the moving object in the world coordinate system be represented by $P_i^w = (x_i^w; y_i^w; z_i^w; 1)^T$ in each of the N frames captured at times $t_1; t_2; : : : ; t_N$. The actual motion of the moving object between frames is represented by M_i^{obj}, and the estimated motion is represented by $\hat{M}_i^{obj}$. The difference between the estimated and actual motion of the object can be denoted as:

$$M_i^{obj} = M_i^{obj} \quad \hat{M}_i^{obj}: \tag{7}$$

As we have already computed the transformation matrix T_i based on the estimated ego motion, we can calculate the transformed object position in the current frame, considering its actual motion, as:

$$P_i^{w^\theta} = T_i \quad M_i^{obj} \quad P_i^w: \tag{8}$$

The error in the transformed object position can be computed as:

$$e_i^{obj} = P_i^{w^\theta} \quad \hat{P}_i^{w^\theta}: \tag{9}$$

In the long-term fusion process, we integrate the information from all N frames. Assuming we use a fusion function F, the fused position in the current frame can be represented as:

$$P_{fusion}^{obj} = F(P_1^{w^\theta}; P_2^{w^\theta} : : : : ; P_N^{w^\theta}): \tag{10}$$

The inaccuracies in the motion estimation of the moving object for each frame can propagate through the fusion function and result in a misaligned object in the fused frame. The overall error in the fused position can be represented as a function of the errors in each frame:

$$e_{fusion}^{obj} = G(e_1^{obj}; e_2^{obj} : : : : ; e_N^{obj}); \tag{11}$$

where G represents a function that combines the errors from each frame. The fused position of the moving object will be less accurate due to these motion estimation errors, leading to a decline in object detection performance in the long-term setting. To address the issue mentioned earlier, we introduce the trajectory-based distillation module, which compensates for the misalignment of moving objects. We will provide further details in the subsequent discussion.

B.2 The Improvements of Static and Dynamic Objects

In this section, we present visualizations to demonstrate the improvements achieved in dynamic objects. Particularly noteworthy is the significant enhancement in the representation of dynamic objects through trajectory-based distillation, thereby highlighting the effectiveness of the trajectory-based module. As depicted in Fig. 3 and Fig. 4, our distillation framework consistently enhances static and dynamic objects across various metrics.

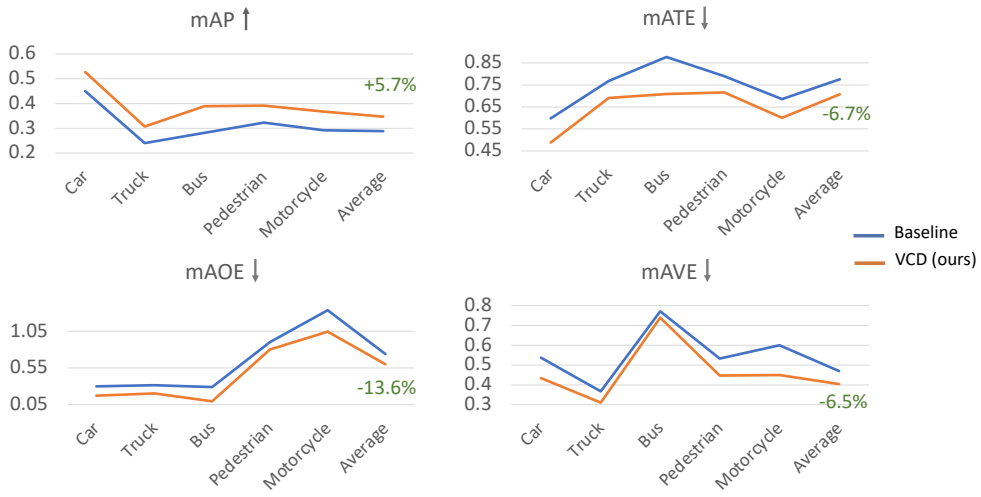


Figure 4: Effects of VCD on movable objects. Our distillation framework VCD consistently improves dynamic objects across a range of metrics.

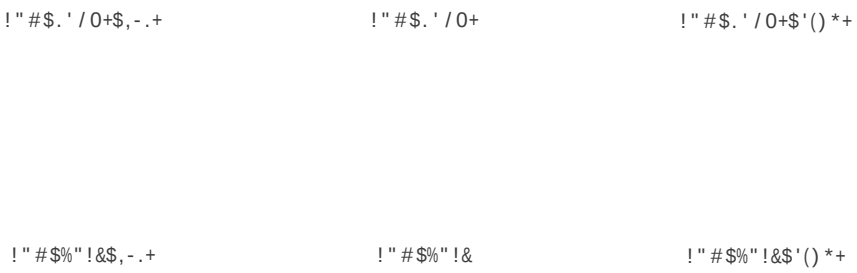


Figure 5: Visualization of the predictions for 3D object detection generated by the VCD-A.

C Visualization

We have performed several visualizations in Fig. 5 to showcase the advancements achieved by our distillation framework. Our findings indicate that our models excel in accurately predicting 3D bounding boxes for the target objects.

D Broader Impact

Our research introduces a novel perspective for multi-modal methodologies and a fresh distillation paradigm for camera-only techniques. We believe that it can establish a robust baseline for the broader scientific community. However, while our methods contribute to the enhancement of autonomous driving, they are not yet capable of addressing more complex corner cases. Consequently, these limitations could potentially introduce risks in real-world autonomous systems.