Distilling Future Temporal Knowledge with Masked Feature Reconstruction for 3D Object Detection

Haowen Zheng[1*], Hu Zhu[2], Lu Deng[3], Weihao Gu[4†], Yang Yang[5†], Yanyan Liang[1†]

[1]Macau University of Science and Technology
[2] The Hong Kong Polytechnic University
[3] HAOMO.AI Technology Co., Ltd.
[4] Institute for AI Industry Research, Tsinghua University
[5] MAIS, Institute of Automation, Chinese Academy of Sciences
zhengnayin@gmail.com, guwh22@mails.tsinghua.edu.cn, yang.yang@nlpr.ia.ac.cn, yyliang@must.edu.mo

Abstract

Camera-based temporal 3D object detection has shown impressive results in autonomous driving, with offline models improving accuracy by using future frames. Knowledge distillation (KD) can be an appealing framework for transferring rich information from offline models to online models. However, existing KD methods overlook future frames, as they mainly focus on spatial feature distillation under strict frame alignment or on temporal relational distillation, thereby making it challenging for online models to effectively learn future knowledge. To this end, we propose a sparse query-based approach, Future Temporal Knowledge Distillation (FTKD), which effectively transfers future frame knowledge from an offline teacher model to an online student model. Specifically, we present a future-aware feature reconstruction strategy to encourage the student model to capture future features without strict frame alignment. In addition, we further introduce future-guided logit distillation to leverage the teacher's stable foreground and background context. FTKD is applied to two high-performing 3D object detection baselines, achieving up to 1.3 mAP and 1.3 NDS gains on the nuScenes dataset, as well as the most accurate velocity estimation, without increasing inference cost.

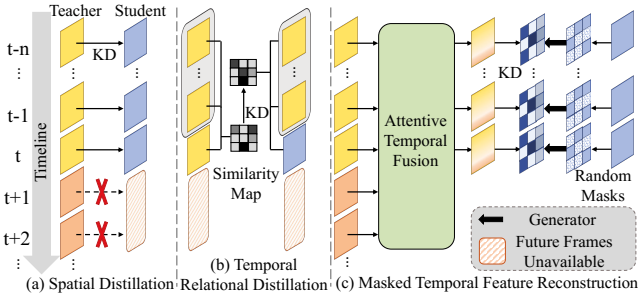


Figure 1: Illustration of using future frames in feature distillation. (a) Spatial feature distillation requires strict alignment of input frames between the teacher and student models, preventing the use of future frame information. (b) Temporal relational distillation focuses on inter-frame relational knowledge but overlooks future frames. (c) In FTKD, information from future frames is aggregated temporally and used as the reconstruction objective for student's masked feature, facilitating effective learning of future knowledge.

Introduction

Camera-based multi-view 3D object detection has attracted much attention in autonomous driving due to its low deployment cost and rich visual information. Recently, bird's-eye-view (BEV) detection achieves promising performance by incorporating temporal information (Li et al. 2022b; Yang et al. 2023; Huang and Huang 2022; Li et al. 2023; Liu et al. 2023c; Park et al. 2022; Han et al. 2023; Lin et al. 2023a, 2022). Furthermore, several offline models (Liu et al. 2023a; Wang et al. 2023a; Lin et al. 2023b) introduce future frames through parallel temporal fusion to further boost accuracy, which aids in detecting small or occluded objects. However, online detection lacks access to future frames, making it challenging to effectively utilize this knowledge.

While knowledge distillation (KD) can be an appealing technique for transferring knowledge from an offline teacher model to an online student model, existing KD methods for 3D object detection still suffer from three key limitations. (i) Most KD methods (Yang et al. 2022b; Zeng et al. 2023; Zhou et al. 2023) primarily concentrate on spatial feature distillation, which fail to effectively exploit the teacher's future knowledge due to strict input frame alignment requirements (see Fig. 1(a)). (ii) Temporal distillation methods (Jang et al. 2023; Wang et al. 2023b) neglect valuable information from future frames, as shown in Fig. 1(b). (iii) They apply to weak baselines or are based on dense BEV representation (Li et al. 2022b; Yang et al. 2023; Li et al. 2023; Huang and Huang 2022; Park et al. 2022). Dense BEV methods suffer from increased latency with more input frames, which hinders real-world applicability and deployment.

Moreover, the selection of an appropriate offline teacher model necessitates careful consideration of several factors,

[*]Work done during the internship at HAOMO.AI Technology.
[†]Corresponding author.

including the domain gap across modalities, alignment with future frame information, and the consistency of temporal feature representations. To this end, the teacher model should meet three key criteria: using a camera-based modality to ensure domain alignment, adopting a parallel temporal fusion strategy to effectively integrate temporal cues, and employing a sparse query representation to maintain consistency with the student model's feature representation.

To address the aforementioned problems, we propose a sparse query-based framework, Future Temporal Knowledge Distillation (FTKD). FTKD overcomes the limitations of strict frame alignment in spatial feature distillation. To achieve this, we introduce a future-aware feature reconstruction mechanism that enables the online student model to distill knowledge from future frames, as illustrated in Fig. 1(c). The online student model reconstructs complete feature representations from its partial features, guided by a teacher model. This enhances the student's representational capacity by integrating future contextual information. Specifically, we first construct the reconstruction objective by aggregating teacher features that contain future information. Then, we introduce random masks to the student features and generate new features using an adaptive generator. Finally, the student features are reconstructed based on the defined reconstruction objective. This design allows the student detector to incorporate long-term temporal (both historical and future) knowledge without increasing inference overhead.

In addition, since the ground truth contains few foreground bounding boxes, a large portion of the final predictions are assigned to background queries. However, existing KD methods (Jang et al. 2023; Chen et al. 2022) largely overlook the informative cues embedded in these background queries. Benefiting from stable training with access to future frames, the teacher model can provide more accurate guidance for both foreground and background context. Based on this observation, we propose a future-guided logit distillation (FLD) strategy. Since the order of queries from the teacher and student models is not aligned during distillation, we employ the Hungarian algorithm (Kuhn 1955) to establish one-to-one matching between them. This strategy ensures that both foreground and background cues are effectively leveraged, addressing the limitations of prior methods that ignore background supervision. Similarly, the order of student's reconstructed features is also adjusted accordingly before computing the loss. By integrating feature-level and logit-level distillation, the proposed FTKD enables the online student model to effectively learn from future knowledge, while achieving a favorable trade-off between accuracy and efficiency.

In summary, our contributions can be described as

- We propose Future Temporal Knowledge Distillation, a camera-only framework that enables online learning from future frames, balancing accuracy and efficiency.
- We introduce future-aware feature reconstruction and future-guided logit distillation strategies that eliminate the constraint of strict frame alignment and effectively utilize stable background information.
- The proposed KD method is applied to two high-performing baseline models, and experimental results on the nuScenes dataset validate the effectiveness of FTKD. From the qualitative results, we also improve the detection of occluded and distant objects.

Related Work

Camera-based 3D Object Detection

Recently, camera-based 3D object detection has achieved significant success with bird's eye view (BEV) representation. In terms of representation, 3D object detection is generally categorized into dense BEV-based approaches and sparse query-based approaches. One line of dense BEV-based methods (e.g., LSS (Philion and Fidler 2020)) attempts to transform multi-view 2D image features into 3D space based on depth estimation. Building upon LSS, the BEVDet series (Huang et al. 2021; Huang and Huang 2022) further elevate performance by introducing data augmentation and temporal fusion. Different from the 2D-to-3D back-projection methods, BEVFormer (Li et al. 2022b) constructs a dense BEV space to sample multi-view 2D features using a deformable attention mechanism. However, dense BEV methods incur substantial computational overhead when modeling temporal information. As a result, inspired by DETR (Carion et al. 2020), sparse query-based approaches are proposed. For instance, DETR3D (Wang et al. 2022) initializes a set of 3D queries to explore a sparse BEV representation. However, the performance is compromised if projecting 3D query points to 2D space for sampling camera features with a fixed local receptive field. To address this issue, the PETR series (Liu et al. 2022, 2023c) leverage global attention to expand the receptive field. Despite the remarkable progress achieved by the aforementioned approaches, they still entail substantial computational burdens. Consequently, SparseBEV presents a fully sparse detector with a scale-adaptive receptive field for a better trade-off between accuracy and speed.

Temporal Modeling

Integrating long-term temporal knowledge is crucial in autonomous driving. Mainstream temporal fusion mechanisms can be divided into parallel temporal fusion (Yang et al. 2023; Liu et al. 2023c; Huang and Huang 2022; Li et al. 2023; Park et al. 2022; Liu et al. 2023a) and sequential temporal fusion (Li et al. 2022b; Wang et al. 2023a; Han et al. 2023; Lin et al. 2023a). Early works (Liu et al. 2023c; Huang and Huang 2022; Li et al. 2023) fuse short-term memory (2-4 frames), but their performance is not satisfactory. To explore long-term temporal fusion, SOLOFusion (Park et al. 2022) utilizes 17 frames and achieves outstanding performance. Nevertheless, parallel temporal fusion methods commonly grapple with the challenge of balancing accuracy and efficiency. To alleviate this problem, (Wang et al. 2023a; Han et al. 2023; Lin et al. 2023a) carry out sequential temporal fusion instead. They propagate historical features into the current timestamp, largely accelerating the inference speed while maintaining excellent performance. However, sequential temporal fusion is limited to past frames, precluding the use of future frames. Furthermore, existing online temporal

3D object detection methods cannot leverage future information. This paper employs a sparse query-based teacher model with parallel temporal fusion to integrate information from future frames. To ensure consistent sparse representations and a balance between accuracy and efficiency, we select two high-performing sparse query-based student models. The two student models (Liu et al. 2023a; Wang et al. 2023a) adopt parallel and sequential temporal fusion paradigms, respectively, demonstrating the generalizability of the proposed method across different architectures.

Knowledge Distillation for Object Detection

Applying knowledge distillation (Hinton, Vinyals, and Dean 2015) on 2D object detection is a popular topic. Different from distilling global features (Chen et al. 2017), several works emphasize the importance of region selection based on bounding boxes (Wang et al. 2019; Dai et al. 2021; Guo et al. 2021) and the application of attentive masks on features (Zhang and Ma 2020; Huang et al. 2022) to mitigate noise interference. FGD (Yang et al. 2022a) effectively combines both strategies, resulting in further performance enhancements. From another perspective, masked feature reconstruction has proved its effectiveness. Inspired by masked image modeling, MGD (Yang et al. 2022b) generates random masks on student features, and then reconstructs them under the guidance of teacher features. (Huang et al. 2022) generates attentive masks instead of random masks for feature reconstruction. DETRDistill (Chang et al. 2023) is specifically designed for DETR-families, incorporating both feature-level and logit-level distillation.

For 3D object detection, most methods (Zhou et al. 2023; Klingner et al. 2023; Liu et al. 2023b; Li et al. 2022a; Kim et al. 2024; Huang et al. 2023) focus on cross-modality knowledge distillation, aiming to transfer LiDAR-based features to camera-based features. These methods improve performance by using LiDAR's real-world modeling, but aligning modalities remains challenging, hindering heterogeneous problem processing. To this end, FD3D first proposes camera-only distillation to reconstruct focal knowledge from imperfect teachers. Existing methods largely focus on spatial distillation, leaving temporal distillation underexplored. Although STXD investigates inter-frame relations and DistillBEV distills fused spatiotemporal features, both approaches largely overlook future frame information. In this paper, we propose Future Temporal Knowledge Distillation, effectively transferring future frame information to the online student model, thereby improving performance without introducing any additional inference overhead.

Method

Preliminary: Sparse BEV Models

Recently, two camera-based 3D detectors, SparseBEV (Liu et al. 2023a) and StreamPETR (Wang et al. 2023a), have shown remarkable performance and high efficiency, both employing sparse BEV query representation. A sparse query is defined as a nine-tuple $Q = (x, y, z, w, l, h, \theta, v_x, v_y)$, where (x, y, z) denotes the query's coordinate in the BEV space, while w, l, h represent its width, length, and height, respectively. θ and (v_x, v_y) indicate query's rotation and velocity. The query set consists of N_q queries, each associated with C-dim features. The two models share a similar detection pipeline but differ in temporal fusion strategy. The query features are first passed through self-attention, followed by sampling from the feature pyramid network (FPN) feature maps to extract semantic information, and finally decoded into the final predictions $\hat{y} = \{\hat{c}, \hat{b}\}$ (categories and bounding boxes) through a decoder layer. Following DETR, the predictions are matched to ground truth (GT) using the Hungarian algorithm for optimal bipartite assignment $\hat{\sigma}$. Finally, the classification and regression losses are computed between the optimally matched predictions $\hat{y}_{\hat{\sigma}}$ and the GT.

SparseBEV employs parallel temporal fusion, aligning with the criteria for an effective teacher model as discussed in Sec. 1. Therefore, we adopt SparseBEV as both the teacher and one of the student models in this paper. In contrast, StreamPETR adopts sequential temporal fusion to accelerate inference, making it a suitable student model to further validate the generalizability of our proposed method.

Overall Framework

Incorporating future frames into temporal modeling can provide richer motion information, thereby improving the detection and velocity estimation performance for dynamic objects. However, online models do not have access to future frames. To address this limitation, we propose Future Temporal Knowledge Distillation (FTKD), as illustrated in Fig. 2, which transfers future knowledge from an offline teacher model to an online student model. FTKD consists of two key components: future-aware feature reconstruction and future-guided logit distillation. During the distillation stage, we freeze the teacher detector and retain the original student architecture by removing any extra auxiliary layers post-training. Based on this, there is no additional computational overhead during the inference.

Future-aware Feature Reconstruction

To ensure perceptual consistency, both perspective view (PV) and sparse query features require future-aware feature reconstruction. A key challenge is bridging the gap between the teacher and student models' input frame requirements. This necessitates effectively fusing the teacher's extensive temporal knowledge, encompassing both past and future information. We address this using distinct strategies for PV and sparse query features. Formally, let $\mathbf{F}^T$ and $\mathbf{F}^S$ denote the PV or sparse query features from the teacher and student model, conforming to $\mathbf{F}^T = \{\{\mathbf{F}_i^{T\_his}\}_{i=1}^{M^{his}}, \{\mathbf{F}^{T\_cur}\}, \{\mathbf{F}_i^{T\_fut}\}_{i=1}^{M^{fut}}\}$ and $\mathbf{F}^S = \{\{\mathbf{F}_i^{S\_his}\}_{i=1}^{N^{his}}, \{\mathbf{F}^{S\_cur}\}\}$, where M^{his} and M^{fut} indicate the number of historical and future frames of teacher model. $M = M^{his} + 1 + M^{fut}$ and $N = N^{his} + 1$ are the total numbers of input frames for the teacher and student, respectively. In general, M^{his} is equal to N^{his}.

Temporal self-attention on PV features. To capture high-level semantic information from future frames while reducing computational overhead, we apply temporal self-attention (TSA) to the final-layer features of the FPN.

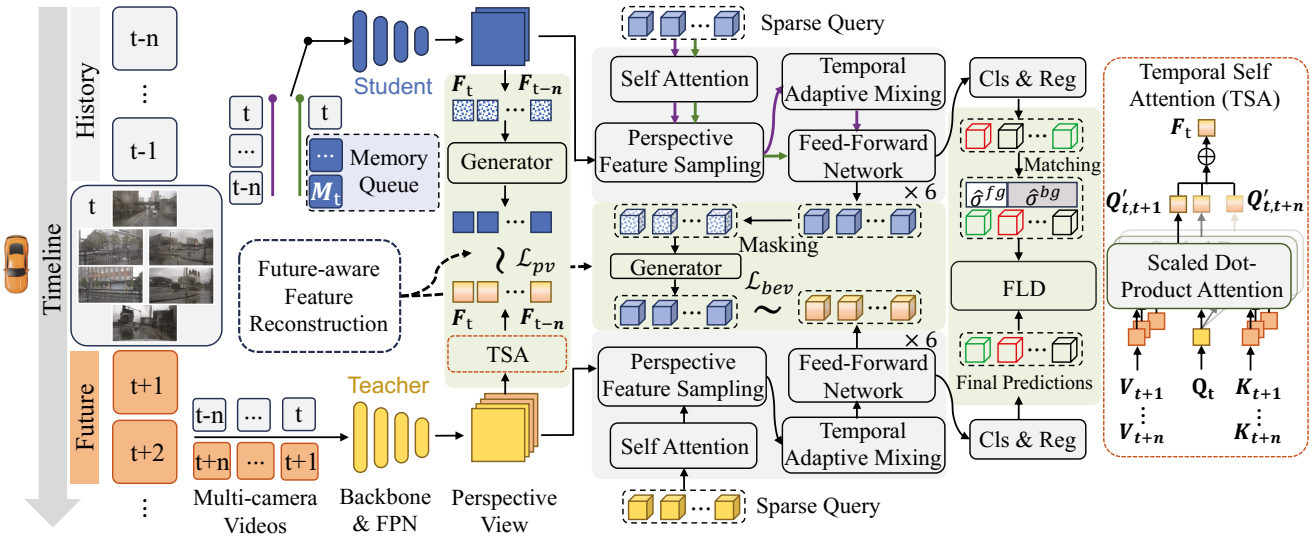


Figure 2: Overall framework of Future Temporal Knowledge Distillation (FTKD). FTKD consists of two core distillation components: future-aware feature reconstruction (FFR) and future-guided logit distillation (FLD), which facilitate the transfer of future knowledge from the offline teacher to the online student model. Specifically, FFR conducts masked feature reconstruction on perspective features and sparse BEV query features, while FLD guides the student in capturing both foreground and background cues embedded in the sparse queries.

$\mathbf{F}_{pv}^{T_0} = \{\{\mathbf{F}_{pv,i}^{T\text{-}his}\}_{i=1}^{M^{his}}, \{\mathbf{F}_{pv}^{T\text{-}cur}\}\}$ is used as the query, while $\{\mathbf{F}_{pv,i}^{T\text{-}fut}\}_{i=1}^{M^{fut}}$ serves as the key and value to extract informative cues from future frames. Therefore, the aggregation of teacher temporal knowledge can be formulated as

$$\mathbf{F}_{pv,i}^{T\text{-}agg} = \sum_{j=1}^{M^{fut}} \text{TSA}(\mathbf{F}_{pv,i}^{T_0}, \mathbf{F}_{pv,j}^{T\text{-}fut}, \mathbf{F}_{pv,j}^{T\text{-}fut}), \tag{1}$$

where $i = 1, 2, ..., M^{his} + 1$ and TSA is based on the scaled dot-product attention operation (Vaswani et al. 2017).

Temporal adaptive mixing on sparse query features. Since AdaMixer (Gao et al. 2022) and SparseBEV propose an efficient and adaptive sparse query decoding mechanism, we reuse it to fuse temporal (both historical and future) query features and finally obtain $\mathbf{F}_{bev}^{T\text{-}agg}$.

Once the reconstruction objective is defined, we perform masked reconstruction on the student features. Random mask is generated on $\mathbf{F}^S$ with mask ratio λ, which can be formulated as

$$M_{k,i} = \begin{cases} 0, \text{if } R_{k,i} < \lambda \\ 1, \text{otherwise} \end{cases}, \tag{2}$$

where $R_{k,i}$ is a random number in $(0, 1)$ and k indicates the index of query or pixel. i denotes the i-th frame. Subsequently, we recover masked student features using a generation layer $\mathcal{G}$:

$$\hat{\mathbf{F}}^S = \mathcal{G}(\mathbf{F}^S \cdot M). \tag{3}$$

Since the PV features and sparse query features of the student model have different dimensions, we design separate generation layers for them accordingly. For PV features, $\mathcal{G}$ is composed of two 2D 3×3 convolutional layers and one $ReLU$ layer, whereas for sparse query features, $\mathcal{G}$ consists of a feed-forward network (FFN) followed by a layer normalization. Then the generated student features $\hat{\mathbf{F}}^S$ are reconstructed under the supervision of the teacher's temporally aggregated features $\mathbf{F}^{T\text{-}agg}$ using Mean Squared Error (MSE). Thus, the temporal reconstruction loss for PV features can be represented as

$$\mathcal{L}_{pv} = \frac{1}{n} \sum_{i=1}^{N} \sum_{l=1}^{L} \sum_{c=1}^{C} \|\hat{\mathbf{F}}_{pv,i,l,c}^{S} - \mathbf{F}_{pv,i,l,c}^{T\text{-}agg}\|_2^2, \tag{4}$$

where L denotes the product of the height and width of the feature map and $n = N \times L \times C$.

Similarly, the reconstruction loss for sparse query features can be formulated as

$$\mathcal{L}_{bev} = \frac{1}{n} \sum_{i=1}^{N} \sum_{q=1}^{N_q} \sum_{c=1}^{C} \|\hat{\mathbf{F}}_{bev,i,\hat{\sigma}_q,c}^{S} - \mathbf{F}_{bev,i,q,c}^{T\text{-}agg}\|_2^2, \tag{5}$$

where $n = N \times N_q \times C$. $\hat{\sigma}_q$ represents the optimal permutation of N_q elements obtained by applying the Hungarian algorithm to match the teacher's final predictions with those of the student.

Future-guided Logit Distillation

Logit distillation has been widely adopted in KD. In cross-modal distillation (Chen et al. 2022; Jang et al. 2023), these methods typically assume that the teacher's final predictions vary in importance. Therefore, they assign quality scores as weights to the teacher predictions. However, such approaches often neglect background information. To address

Method	Frames	NDS ↑	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓	FPS ↑
Results without distillation schemes									
BEVDet4D † (Huang and Huang 2022)	2	45.7	32.2	0.703	0.278	0.495	0.354	0.206	30.7
PETRv2 (Liu et al. 2023c)	2	45.6	34.9	0.700	0.275	0.580	0.437	0.187	-
SOLOFusion † (Park et al. 2022)	16+1	53.4	42.7	0.567	0.274	0.511	0.252	0.181	15.7*
VideoBEV † (Han et al. 2023)	8	53.5	42.2	0.564	0.276	0.440	0.286	0.198	-
Sparse4Dv2 (Lin et al. 2023a)	-	53.9	43.9	0.598	0.270	0.475	0.282	0.179	17.3
Results with distillation schemes									
T:SparseBEV-R101 (Liu et al. 2023a)	15	63.8	55.1	0.493	0.265	0.275	0.163	0.181	3.1
S:SparseBEV-R50	8	55.5	44.7	0.585	0.271	0.391	0.251	0.188	20.2
+MGD (Yang et al. 2022b)	8	55.1 (↓0.4)	44.8 (↑0.1)	0.591	0.270	0.425	0.248	0.192	20.2
+CWD (Shu et al. 2021)	8	55.1 (↓0.4)	44.6 (↓0.1)	0.591	0.270	0.408	0.251	0.190	20.2
+FD3D (Zeng et al. 2023)	8	55.0 (↓0.5)	44.6 (↓0.1)	0.598	0.270	0.423	0.251	0.187	20.2
+STXD (Jang et al. 2023)	8	55.6 (↑0.1)	45.0 (↑0.3)	0.588	0.271	0.398	0.247	0.187	20.2
+FTKD (ours)	8	**56.5** (↑1.0)	**46.0** (↑1.3)	**0.579**	**0.270**	**0.372**	**0.234**	**0.179**	20.2
S:StreamPETR-R50 (Wang et al. 2023a)	8	55.0	45.0	0.613	0.267	0.413	0.265	0.196	**33.9**
+MGD (Yang et al. 2022b)	8	55.1 (↑0.1)	45.0 (↑0.0)	0.618	0.269	0.402	0.260	0.194	**33.9**
+CWD (Shu et al. 2021)	8	54.8 (↓0.2)	44.8 (↓0.2)	0.610	0.270	0.428	0.263	0.189	**33.9**
+FD3D (Zeng et al. 2023)	8	55.4 (↑0.4)	45.3 (↑0.3)	0.600	0.268	0.405	0.259	0.190	**33.9**
+STXD (Jang et al. 2023)	8	55.6 (↑0.6)	45.5 (↑0.5)	0.597	0.269	0.411	0.254	0.184	**33.9**
+FTKD (ours)	8	**56.3** (↑1.3)	**46.3** (↑1.3)	**0.589**	**0.268**	**0.398**	**0.252**	**0.182**	**33.9**

Table 1: Comparison on the nuScenes validation set. † denotes methods with CBGS (Zhu et al. 2019). Only the teacher model has access to future frames. The student baselines benefit from perspective-view pretraining. FPS is measured on RTX4090 with fp32 without cuda acceleration. * represents inference with fp16. The input size for ResNet101 (R101) and ResNet50 (R50) are 512 × 1408 and 256 × 704, respectively.

this limitation, inspired by DETRDistill (Chang et al. 2023), we introduce FLD.

With the guidance of future knowledge, the teacher model is well optimized and can consistently provide a large number of true negatives. Therefore, we apply the Hungarian algorithm to perform bipartite matching between the teacher's predictions $\hat{y}^T$ and the student's predictions $\hat{y}^S$. This yields the optimal permutation of both foreground and background samples, denoted as $\hat{\sigma}^{fg}$ and $\hat{\sigma}^{bg}$, respectively. Then the logit distillation can be formulated as

$$\mathcal{L}_{logits} = \sum_{q=1}^{N_q} \alpha \mathcal{L}_{cls}(\hat{c}^S_{\hat{\sigma}_q}, \hat{c}^T_q) + \beta \mathcal{L}_{bbx}(\hat{b}^S_{\hat{\sigma}_q}, \hat{b}^T_q), \quad (6)$$

where $\hat{\sigma} = \{\hat{\sigma}^{fg}, \hat{\sigma}^{bg}\}$. $\mathcal{L}_{cls}$ and $\mathcal{L}_{bbx}$ is FocalLoss (Lin et al. 2017) and L1 loss, respectively. α and β are weights to balance the logit KD loss (set to 2.0 and 0.25 by default).

Overall Distillation Loss

Finally, the overall KD loss for the online student model is formulated by integrating Eq. 4, Eq. 5, and Eq. 6:

$$\mathcal{L}_{KD} = \lambda_1 \mathcal{L}_{pv} + \lambda_2 \mathcal{L}_{bev} + \lambda_3 \mathcal{L}_{logits}, \quad (7)$$

where λ_1, λ_2 and λ_3 are loss weights to balance the KD losses. In summary, we train the student model with original classification and regression loss as well as KD loss $\mathcal{L}_{KD}$.

Model	Future Frame	NDS ↑	mAP ↑
SparseBEV-R50	✗	55.5	45.0
	✓	**56.5**	**46.0**
StreamPETR-R50	✗	55.2	45.1
	✓	**56.3**	**46.3**

Table 2: Ablation on the utilization of future frames in KD. We use SparseBEV-R101 as the teacher model, with 8 input frames (7 historical frames and 1 current frame).

Experiments

Datasets and Metrics

A large-scale surround-view autonomous driving benchmark, nuScenes (Caesar et al. 2020), is utilized to evaluate our approach. It comprises 700/150/150 scenes for training/-validation/testing. Each scene spans approximately 20 seconds, with annotations available for key frames at 0.5s intervals. For 3D object detection, it includes 1.4M 3D bounding boxes across 10 categories. Following the official evaluation metrics, we report nuScenes detection score (NDS), mean Average Precision (mAP), and five true positive (TP) metrics, including ATE, ASE, AOE, AVE, and AAE for measuring translation, scale, orientation, velocity, and attributes, respectively. The NDS combines mAP and five TP metrics to provide a comprehensive evaluation score.

Implementation Details

The experimental results are reported based on 8 A100 GPUs, and FPS measurements are conducted on RTX4090

FFR PV	FFR BEV	FLD	NDS ↑	mAP ↑	mAVE ↓
			55.5	44.7	0.251
✓			55.7	44.9	0.250
	✓		55.8	45.2	0.243
		✓	55.9	45.3	0.247
	✓	✓	56.3	45.6	0.235
✓	✓		55.9	45.4	0.243
✓	✓	✓	**56.5**	**46.0**	**0.234**

Table 3: Effectiveness of loss components. FFR indicates future-aware feature reconstruction, which is applied on perspective view (PV) and sparse bird's-eye-view (BEV) query, respectively. FLD denotes future-guided logit distillation.

with fp32. We train all the models with AdamW (Loshchilov and Hutter 2017) optimizer for 24 epochs on SparseBEV and 60 epochs on StreamPETR, using perspective pretraining on nuImage (Caesar et al. 2020). The initial learning rate is set to 2×10^{-4} and is decayed with a cosine annealing policy. The global batch size is fixed to 8. For supervised training, the Hungarian algorithm is used for label assignment. FocalLoss and L1 loss are employed for classification and 3D bounding boxes regression, respectively. We initialize $N_q = 900$ queries and set the channel of query features $C = 256$. The mask ratio λ is fixed to 0.5. Loss weights λ_1, λ_2, and λ_3 are set to $1e^{-3}$, 16 and 1, respectively. We set the number of frames $M^{his} = M^{fut} = N^{his} = 7$ by default. All data preprocessing follows the corresponding baselines.

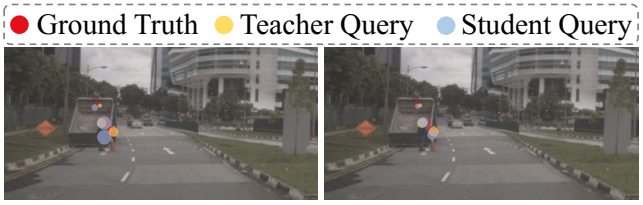
Main Results

Comparison with spatial distillation methods. Conventional 2D spatial distillation methods (Yang et al. 2022b; Shu et al. 2021) struggle to improve temporal 3D object detection, possibly due to their inability to capture temporal dynamics and 3D spatial context. FD3D is the first camera-only KD method that is designed for the spatial domain. We reimplement it for a fairer comparison. Our method outperforms FD3D with improvements of 1.4 mAP and 1.5 NDS on SparseBEV. Note that FD3D requires strictly aligned input frames between the student and teacher models, which prevents the transfer of information from future frames. Hence, only spatial knowledge from past and current frames can be distilled. This limitation, combined with its neglect of background information, likely contributes to its suboptimal performance in temporal 3D object detection.

Comparison with temporal distillation method. Although STXD is a cross-modal KD method, its temporal distillation component is highly relevant to our temporal student detectors. Therefore, we reproduce STXD on both SparseBEV and StreamPETR. For a fair comparison, the reproduced STXD is also adapted to explore relational knowledge from future frames. Our FTKD outperforms STXD by 0.9 NDS and 1.0 mAP on SparseBEV, and by 0.7 NDS and 0.8 mAP on StreamPETR. The insufficient utilization of future knowledge and background information may be the main factors limiting the performance of STXD.

Location	Mask ratio	NDS ↑	mAP ↑	mAVE ↓
BEV	0.4	55.4	44.8	0.251
	0.5	**55.8**	**45.2**	**0.243**
	0.6	55.5	45.1	0.247
	0.75	54.9	45.1	0.253
	0.9	54.8	44.3	0.268
BEV & PV	0.5 & 0.5	**55.9**	**45.4**	**0.243**
	0.5 & 0.65	55.0	45.2	0.261
	0.5 & 0.75	55.2	45.1	0.255

Table 4: Ablation of the mask ratio.



(a) Without FFR (b) With FFR

Figure 3: Visualization of sparse queries (a) with and (b) without future-aware feature reconstruction (FFR). Larger points denote shallower depth. It is evident that, with FFR, the sparse queries are more aligned with the ground truth.

Ablation Study

Ablation on future frames. We first validate the effectiveness of incorporating future frames in knowledge distillation. While high-performance 3D object detectors have demonstrated strong capabilities in capturing information from historical frames, our results show that future cues can further benefit online student models, demonstrating the strength of our approach.

Effects of loss components. In Table 3, we evaluate the effects of each loss component on NDS, mAP, and mAVE. Notably, employing FFR solely on BEV yields higher accuracy (NDS: ↑ 0.3, mAP: ↑ 0.5) than the other two loss terms. Moreover, the other two loss terms can also improve 0.2-0.4 NDS. When applying FLD and FFR (on sparse BEV query), we observe significant enhancements in NDS and mAP, reaching 56.3 and 45.6, respectively. This emphasizes the efficacy of FFR and FLD, providing the model with future knowledge and background information. As a result, the combination of these two types of distillation can lead to certain performance gains. We also provide a visualization of the sparse queries in Fig. 3, revealing that the teacher model effectively guides their feature reconstruction.

Impacts of mask ratio. We investigate various mask ratios for FFR in both sparse BEV query and PV. As shown in Table 4, we can identify the optimal mask ratio to be 0.5 for both BEV and PV. The core idea of masked feature reconstruction is to use the residual features to reconstruct complete feature maps. A high mask ratio results in a poor representation of residual features, while a low mask ratio simplifies the generator's learning process, enabling shortcuts that

Figure 4: Qualitative results over three consecutive frames (front camera) in two scenes. The first and third row show the prediction made by the baseline model, while the second and fourth row demonstrate the predictive results of FTKD. In the last column, the LiDAR point cloud in BEV is display for frame $t+1$, except the last row (for $t+2$) due to the limited BEV distance. FTKD successfully predicts an occluded car merging into the main road and a pedestrian crossing the street in the distance, highlighted by red dotted circles.

selections of FLD	NDS ↑	mAP ↑	mAVE ↓
foreground	55.4	44.9	0.254
background	55.5	45.1	0.251
foreground+background	**55.9**	**45.3**	**0.247**

Table 5: Ablation on the selections of FLD.

lead to local optima. Moreover, a low mAVE demonstrates that temporal feature reconstruction is conducive to focusing on dynamic objects and estimating their velocity.

Selections of future-guided logit distillation. Our ablation study on logit distillation reveals that using only the positive foreground queries yields limited performance improvement, potentially because GT already provides foreground information. However, performance improvements from using only background queries suggest that exploiting these queries is beneficial. The results suggest that future-guided teacher models can enhance student performance by offering both stable foreground and background cues.

Qualitative Results

We provide qualitative results to visualize the prediction of the model with and without FTKD, which highlights the importance of future knowledge and the superior performance of the proposed FTKD. Specifically, FTKD exhibits enhanced capabilities in detecting occluded objects and distant targets. As shown in the second row in Fig. 4, an occluded vehicle about to merge onto the main road can be detected earlier. Furthermore, FTKD successfully identifies a distant pedestrian in advance (see the last row in Fig. 4). These examples demonstrate that FTKD can effectively capture knowledge from future frames, contributing to improved detection of occluded and distant objects.

Conclusion

In this work, we introduce Future Temporal Knowledge Distillation (FTKD), a sparse query-based framework that effectively transfers knowledge from future frames encoded by an offline teacher model to an online student model. FTKD leverages future-aware feature reconstruction to overcome frame alignment constraints in spatial distillation, facilitating effective future temporal knowledge transfer. Additionally, future-guided logit distillation enriches the student model's understanding of background information. Experiments conducted on two strong detectors validate the effectiveness of FTKD and show that it maintains a desirable trade-off between accuracy and efficiency.

Limitations. FTKD is only validated on the 3D object detection task under the camera modality. For autonomous driving, exploring how to effectively learn future knowledge in multi-modal settings and other 3D perception tasks (e.g., 3D occupancy prediction) remains an important research topic.

Acknowledgments

This work was supported by the Science and Technology Development Fund of Macau Project 0096/2023/RIA2, and in part by Chinese National Natural Science Foundation Projects 62206276.

References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chang, J.; Wang, S.; Xu, H.-M.; Chen, Z.; Yang, C.; and Zhao, F. 2023. Detrdistill: A universal knowledge distillation framework for detr-families. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6898–6908.

Chen, G.; Choi, W.; Yu, X.; Han, T.; and Chandraker, M. 2017. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; and Zhao, F. 2022. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*.

Dai, X.; Jiang, Z.; Wu, Z.; Bao, Y.; Wang, Z.; Liu, S.; and Zhou, E. 2021. General instance distillation for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7842–7851.

Gao, Z.; Wang, L.; Han, B.; and Guo, S. 2022. Adamixer: A fast-converging query-based object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5364–5373.

Guo, J.; Han, K.; Wang, Y.; Wu, H.; Chen, X.; Xu, C.; and Xu, C. 2021. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2154–2164.

Han, C.; Sun, J.; Ge, Z.; Yang, J.; Dong, R.; Zhou, H.; Mao, W.; Peng, Y.; and Zhang, X. 2023. Exploring recurrent long-term temporal fusion for multi-view 3d perception. *arXiv preprint arXiv:2303.05970*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, J.; and Huang, G. 2022. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*.

Huang, J.; Huang, G.; Zhu, Z.; Ye, Y.; and Du, D. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.

Huang, L.; Li, Z.; Sima, C.; Wang, W.; Wang, J.; Qiao, Y.; and Li, H. 2023. Leveraging vision-centric multi-modal expertise for 3d object detection. *Advances in Neural Information Processing Systems*, 36: 38504–38519.

Huang, T.; Zhang, Y.; You, S.; Wang, F.; Qian, C.; Cao, J.; and Xu, C. 2022. Masked distillation with receptive tokens. *arXiv preprint arXiv:2205.14589*.

Jang, S.; Jo, D. U.; Hwang, S. J.; Lee, D.; and Ji, D. 2023. Stxd: Structural and temporal cross-modal distillation for multi-view 3d object detection. *Advances in Neural Information Processing Systems*, 36: 29323–29342.

Kim, S.; Kim, Y.; Hwang, S.; Jeong, H.; and Kum, D. 2024. Labeldistill: Label-guided cross-modal knowledge distillation for camera-based 3d object detection. In *European Conference on Computer Vision*, 19–37. Springer.

Klingner, M.; Borse, S.; Kumar, V. R.; Rezaei, B.; Narayanan, V.; Yogamani, S.; and Porikli, F. 2023. 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*, 13343–13353.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Li, Y.; Chen, Y.; Qi, X.; Li, Z.; Sun, J.; and Jia, J. 2022a. Unifying voxel-based representation with transformer for 3d object detection. *Advances in Neural Information Processing Systems*, 35: 18442–18455.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2023. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1477–1485.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022b. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, X.; Lin, T.; Pei, Z.; Huang, L.; and Su, Z. 2022. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*.

Lin, X.; Lin, T.; Pei, Z.; Huang, L.; and Su, Z. 2023a. Sparse4D v2: Recurrent Temporal Fusion with Sparse Model. *arXiv preprint arXiv:2305.14018*.

Lin, X.; Pei, Z.; Lin, T.; Huang, L.; and Su, Z. 2023b. Sparse4d v3: Advancing end-to-end 3d detection and tracking. *arXiv preprint arXiv:2311.11722*.

Liu, H.; Teng, Y.; Lu, T.; Wang, H.; and Wang, L. 2023a. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18580–18590.

Liu, J.; Wang, T.; Liu, B.; Zhang, Q.; Liu, Y.; and Li, H. 2023b. GeoMIM: Towards Better 3D Knowledge Transfer via Masked Image Modeling for Multi-view 3D Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17839–17849.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, 531–548. Springer.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, A.; Wang, T.; and Zhang, X. 2023c. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3262–3272.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Park, J.; Xu, C.; Yang, S.; Keutzer, K.; Kitani, K.; Tomizuka, M.; and Zhan, W. 2022. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 194–210. Springer.

Shu, C.; Liu, Y.; Gao, J.; Yan, Z.; and Shen, C. 2021. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5311–5320.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, S.; Liu, Y.; Wang, T.; Li, Y.; and Zhang, X. 2023a. Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. *arXiv preprint arXiv:2303.11926*.

Wang, T.; Yuan, L.; Zhang, X.; and Feng, J. 2019. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4933–4942.

Wang, Y.; Guizilini, V. C.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2022. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, 180–191. PMLR.

Wang, Z.; Li, D.; Luo, C.; Xie, C.; and Yang, X. 2023b. DistillBEV: Boosting Multi-Camera 3D Object Detection with Cross-Modal Knowledge Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8637–8646.

Yang, C.; Chen, Y.; Tian, H.; Tao, C.; Zhu, X.; Zhang, Z.; Huang, G.; Li, H.; Qiao, Y.; Lu, L.; et al. 2023. 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*, 17830–17839.

Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; and Yuan, C. 2022a. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4643–4652.

Yang, Z.; Li, Z.; Shao, M.; Shi, D.; Yuan, Z.; and Yuan, C. 2022b. Masked generative distillation. In *European Conference on Computer Vision*, 53–69. Springer.

Zeng, J.; Chen, L.; Deng, H.; Lu, L.; Yan, J.; Qiao, Y.; and Li, H. 2023. Distilling Focal Knowledge from Imperfect Expert for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 992–1001.

Zhang, L.; and Ma, K. 2020. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*.

Zhou, S.; Liu, W.; Hu, C.; Zhou, S.; and Ma, C. 2023. 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*, 5116–5125.

Zhu, B.; Jiang, Z.; Zhou, X.; Li, Z.; and Yu, G. 2019. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*.