Structured Knowledge Distillation Towards Efficient Multi-View 3D Object Detection

Linfeng Zhang [1,4]
Ke Wang [2]
Hung-Shuo Tai [2]
Kaisheng Ma [1,3]
Yukang Shi [2,4]
Zhipeng Zhang [2]
Yuan He [2]

[1] Tsinghua University
[2] KargoBot

Abstract

Detecting 3D objects from multi-view images is a fundamental problem in 3D computer vision. Recently, significant breakthrough has been made in multi-view 3D detection tasks. However, the unprecedented detection performance of these vision BEV (bird's-eye-view) detection models is accompanied with enormous parameters and computation, which make them unaffordable on edge devices. To address this problem, in this paper, we propose a structured knowledge distillation framework, aiming to improve the efficiency of modern vision-only BEV detection models. The proposed framework mainly includes: (a) spatial-temporal distillation which distills teacher knowledge of information fusion from different timestamps and views, (b) BEV response distillation which distills teacher response to different pillars, and (c) weight-inheriting which solves the problem of inconsistent inputs between students and teacher in modern transformer architectures. Experimental results show that our method leads to an average improvement of 2.16 mAP and 2.27 NDS on the nuScenes benchmark, outperforming multiple baselines by a large margin.

1 Introduction

Recently, bird's-eye-view (BEV) based multi-camera perception frameworks have greatly narrowed the performance gap with LiDAR based methods for 3D object detection tasks [19, 21, 30, 31]. For example, compared with state-of-the-art LiDAR methods, some recent works have obtained NDS scores within a 10% margin [18, 19].

Such vision-centric BEV frameworks usually involve two stages: single view feature extraction using backbone networks (convnets [25] or transformers [29]), and information fusion across multiple camera views and multiple timestamps using transformers [21, 23, 24] or the lift-splat-shoot paradigm [15, 16, 19]. Once a spatial-temporal coherent feature representation is obtained in the unified BEV space, 3D object detection and semantic segmentation [16, 19, 21] can be done on the BEV feature map with high accuracy.

[3] Corresponding author.

[4] The first two authors have equal contribution.

However, such performance improvements are achieved with a hefty computation overhead. For instance, the 120M parameters in BEVDet [16] require more than 4 TFlops computation, which is almost 20× larger and 10× slower than CenterPoint [32], a state-of-the-art LiDAR-based 3D detector. Practical applications such as self-driving vehicles, usually have limited computation budget but rather strict latency and accuracy requirements. Deployment of such visual BEV models onto edge devices requires a delicate balancing between low computation cost and high detection accuracy. Compared with neural network pruning [10] and quantization [8, 38], knowledge distillation (KD) [3, 13] is more suited for striking such a balance.

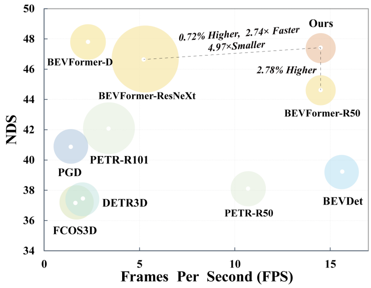


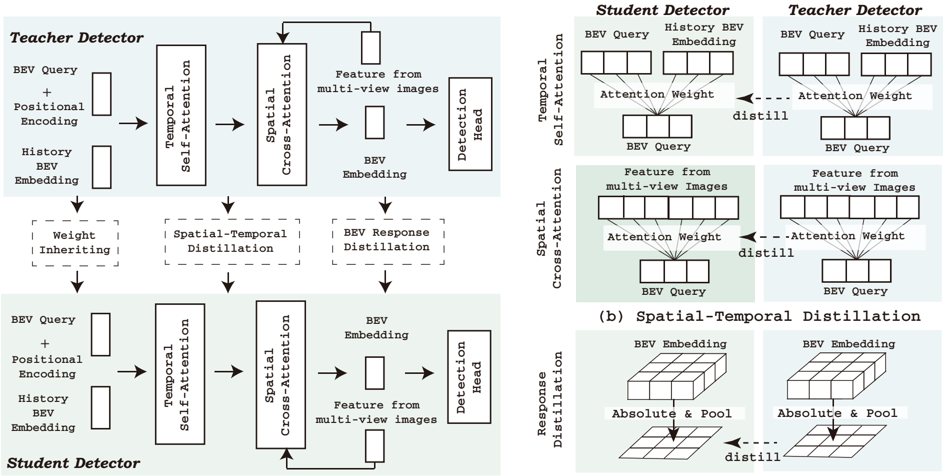
Figure 1: Experimental results on nuScenes. The area of circles indicates the number of parameters. Compared with BEVFormer with a ResNeXt backbone, our method achieves 0.72 NDS improvements, 2.74× acceleration and 4.97× compression.

Knowledge distillation (KD) is an effective model training technique that improves the performance of a lightweight student model by transferring the knowledge from a pre-trained but over-parameterized teacher model [3, 13]. At deployment time, only the lightweight student model is used. While KD has demonstrated great success in various 2D computer vision tasks, such as classification [35], object detection [6, 17, 34], semantic segmentation [11, 22], and image generation [7, 20, 27, 37], the application of KD distillation on 3D computer vision, especially the camera-based multi-view 3D detection, has not been well-studied. However, it is also brought to our attention that simply applying traditional KD methods to 3D vision tasks usually leads to limited performance gains.

To address the aforementioned problems, this paper proposes a novel KD framework for visual BEV detection models. We start with analyzing the challenges in the multi-view 3D detection task and then propose the corresponding solution as follows:

Information fusion from multiple positions: In multi-view 3D detection, the detector takes input from multiple cameras across different timestamps to identify objects. Hence, the student should be able to learn not only the information from single images but also how to fuse and leverage the information from multiple spatial/temporal positions. To tackle this challenge, we propose spatial-temporal distillation, which improves student performance by allowing it to learn the semantic correspondence between inputs in different spatial (*i.e.,* view) and temporal positions from their teachers. Moreover, we also propose BEV response distillation, which aims to distill teacher response to different positions/pillars in the BEV feature map, which contains high level information on object localization.

Discrepancies between the inputs: The state-of-the-art BEV 3D detectors usually employ a DETR-like architecture, which utilizes self-attention and cross-attention layers for information fusion [5, 31]. Different from traditional convolutional detectors, the input information of DETR-style detectors contains not only images but also trainable queries and positional encodings. Without explicit constraints, student and teacher models could have learned different positional encodings and queries after training. Knowledge distillation will be hindered by such discrepancies [2]. To address this problem, we propose a weight-inheriting scheme which fixes the positional encodings and BEV queries of the student model to the



(a)Overall Framework

(b) Spatial-Temporal Distillation

(c) BEV Response Distillation

Figure 2: The overall framework and details of our method. (a) The proposed knowledge distillation methods mainly include weight-inheriting, spatial-temporal distillation, and BEV response distillation. Weight-inheriting fixes the parameters of BEV queries and positional encoding to their value in the pre-trained teacher detector during the whole training period to guarantee that students and teachers have the same inputs. (b) Spatial-temporal distillation aims to improve student performance on information fusion of images from multiple views and timestamps by transferring teacher knowledge in the attention weights in the temporal self-attention and spatial cross-attention layers. (c) BEV response distillation first computes the response of different positions in BEV map and then distills it to the students.

corresponding values in the teacher detector. In this way, the student detector will benefit from the pre-trained weights of the teacher detector directly. Surprisingly, we find that even without applying any knowledge distillation losses, simply using the weight-inheriting scheme can significantly improve the performance of knowledge distillation for this task.

Without loss of generality, we conduct extensive experiments on the nuScenes dataset [4] using a representative and state-of-the-art BEVFormer model architecture [21]. On average, 2.16 mAP and 2.27 NDS improvements can be observed across three different student-teacher settings, demonstrating the effectiveness of our proposed knowledge distillation framework. Compared with multiple baseline methods [1, 9, 12, 13, 14, 28, 33, 34], our method outperforms them all by a large margin.

In summary, our contributions include: (**1**) We propose a novel spatial-temporal distillation scheme which enables the student detector to learn teacher knowledge on how to fuse information from different camera views and timestamps. (**2**) BEV response distillation is proposed to distill teacher response to different BEV pillars, which transfers teacher knowledge on object localization to the student. (**3**) We identify the problem of inconsistent inputs in knowledge distillation on DETR-style detectors and propose a weight-inheriting scheme to solve it. (**4**) Extensive experiments on nuScenes demonstrate the effectiveness of our method. On average, **2.16** mAP and **2.27** NDS improvements can be obtained compared with the student without KD.

2 Methodology

2.1 Preliminary

Without loss of generality, we conduct our experiments on top of the BEVFormer model, which is a performant and representative multi-view 3D object detection architecture. To recap, BEVFormer consists of four stages, including feature extraction from single images, temporal information fusion, spatial information fusion, and prediction.

(I) Feature Extraction of Single Images: In multi-view 3D detection, at timestamp t, the input image set can be denoted as $\mathcal{X}^{(t)} = \{x_1^{(t)}, x_2^{(t)}, ..., x_v^{(t)}\}$, where v denotes the number of views. BEVFormer firstly encodes the feature of each single image with a convolutional 2D backbone f_{2D}, which can be formulated as $F^{(t)} = f_{\text{2D}}(\mathcal{X}^{(t)})$. These features are then fed into spatial cross-attention in Stage III for multi-view feature fusion.

(II) Temporal Information Fusion: Temporal self-attention is utilized to fuse the information between the current input images and the historical images. The input of temporal self-attention layers includes the predefined trainable BEV queries with positional encoding and the previous BEV embedding at timestamp $t-1$, which can be formulated as Q^{BEV} and $E_{\text{BEV}}^{(t-1)}$, respectively. Then, the computation of temporal self-attention can be written as

$$E_{\text{BEV}}^{\prime(t)} = \text{TSA}\left(Q_p^{\text{BEV}}, \{Q^{\text{BEV}}, E_{\text{BEV}}^{(t-1)}\}\right) = \sum\nolimits_{V \in \{Q^{\text{BEV}}, E_{\text{BEV}}^{(t-1)}\}} \text{DeformAttn}(Q_p^{\text{BEV}}, p, V), \quad (1)$$

where DeformAttn indicates the deformable attention layers [39] and Q_p^{BEV} denotes the BEV query located at the position p. TSA and $E_{\text{BEV}}^{\prime(t)}$ indicate temporal self-attention and its outputs, respectively.

(III) Spatial Information Fusion: In the stage of spatial information fusion, BEVFormer samples N_{ref} 3D reference points from each pillar, and then projects them to 2D views. Then, spatial cross-attention is utilized to fuse the BEV embedding output by temporal information fusion with the reference points, which can be formulated as

$$E_{\text{BEV}}^{(t)} = \frac{1}{|v_{\text{hit}}|} \sum_{i \in v_{\text{hit}}} \sum_{j=1}^{N_{\text{ref}}} \text{DeformAttn}\left(E_{\text{BEV}}^{\prime(t)}, \mathcal{P}(p,i,j), F^{(t)}\right), \quad (2)$$

where v_{hit} indicates the number of views that contain the projection of the 3D reference points. $\mathcal{P}(p,i,j)$ is the projection function to get the j-th reference point on the i-th view image. $F^{(t)}$ indicates the feature of single images computed in Stage I.

(IV) Prediction In this stage, BEVFormer predicts the positions, dimensions, headings, and categories of objects based on the two inputs, including the output of spatial cross-attention and a set of object queries, which can be denoted as Q^{Object} and $E_{\text{BEV}}^{(t)}$, respectively. Its computation can be formulated as $\text{B}, \text{P} = \text{Detection Head}(E_{\text{BEV}}^{(t)}, Q^{\text{Object}})$, where "B" and "P" indicate the predicted bounding boxes and the corresponding probability distribution.

2.2 Structured Knowledge Distillation

In this subsection, we introduce the proposed knowledge distillation based on the above four stages in BEVFormer. Note that the Stage I (2D convolutional feature extraction) and Stage IV (prediction) in BEVFormer share quite some similarities with common 2D detectors. Successful attempts have been made to apply knowledge distillation onto these stages [6, 34]. Thus in this paper, we focus on the Stage II and Stage III, which are critical for multi-view 3D detection but rarely explored for knowledge distillation. In particular, our method can be divided into the following three folds.

Spatial-Temporal Knowledge Distillation In Stage II and Stage III, BEVFormer first integrates the BEV queries with the BEV embeddings at the previous timestamp for temporal information fusion, and then fuses the information from different image views for spatial information fusion. Deformable attention layers are utilized during the two processes. Recall that the computation of attention weights in deformable attention layers is obtained by a linear projection over queries followed with a softmax function, which can be formulated as $\mathcal{A}(\mathbf{Q}) = \text{softmax}(\mathbf{WQ})$, where $\mathbf{Q}$ and $\mathbf{W}$ indicate the queries and the trainable parameters in the linear projection layer, respectively. In temporal self-attention, $\mathbf{Q}$ indicates the BEV query Q^{BEV}. And the obtained attention weights are utilized to fuse information from Q^{BEV} and the historical BEV embedding $E_{\text{BEV}}^{(t-1)}$. Hence, the attention weights here show the temporal relation between the information of the current inputs and the previous input. By distilling them, the student is allowed to learn how to fuse temporal information from the teacher detector. In spatial cross-attention, $\mathbf{Q}$ indicates the output of temporal self-attention $E_{\text{BEV}}^{\prime(t)}$. And the obtained attention weights are utilized to fuse the information from the reference points in the multi-view images. Hence, distilling the attention weights here enables the student to learn how to fuse spatial information from the teacher detector. Concretely, we can denote the attention weights in temporal self-attention and temporal self-attention as $\mathbf{A}^{\text{temporal}}$ and $\mathbf{A}^{\text{spatial}}$, respectively, which can be formulated as $\mathbf{A}^{\text{temporal}} = \mathcal{A}(Q^{\text{BEV}})$, and $\mathbf{A}^{\text{spatial}} = \mathcal{A}(E_{\text{BEV}}^{\prime(t)})$, respectively. Then, by distinguishing the student and teacher detector with the scripts $\mathcal{S}$ and $\mathcal{T}$ respectively, spatial-temporal attention can be formulated as

$$\mathcal{L}_{\text{spatial-temp}} = \|\mathbf{A}_{\mathcal{S}}^{\text{temporal}} - \mathbf{A}_{\mathcal{T}}^{\text{temporal}}\|^2 + \|\mathbf{A}_{\mathcal{S}}^{\text{spatial}} - \mathbf{A}_{\mathcal{T}}^{\text{spatial}}\|^2. \tag{3}$$

BEV Response Distillation Besides distilling teacher knowledge on the fusion of the information from different timestamps and views, we also propose BEV response distillation to distill teacher responses to different object queries, which correspond to different pillars in 3D space. In this paper, we define the BEV response as the average score across the channel dimension on the absolute value of BEV embedding, which can be written as

$$\mathcal{R}(E_{\text{BEV}(i,j)}) = \sum\nolimits_{j=1}^{C} \frac{1}{C} |E_{\text{BEV}(i,j)}|, \tag{4}$$

where C denotes the number of channels. The scripts (i, j) denotes the value on the i_{th} BEV query (*i.e.,* pillar) of the j_{th} channel. As pointed out by abundant research [33, 34, 36], the response of features demonstrates the importance of their corresponding spatial positions. Hence, by distilling the BEV response from the teacher, the student model can better correlate between the learned semantic features and the potential object spatial occupancies. An example of BEV response is visualized in Fig. 3. An L2 loss is adopted for BEV response distillation: $\mathcal{L}_{\text{response}} = \|\mathcal{R}(E_{\text{BEV}}^{\mathcal{S}}) - \mathcal{R}(E_{\text{BEV}}^{\mathcal{T}})\|^2$, where $\mathcal{S}$ and $\mathcal{T}$ denote the student detector and the teacher detector, respectively. Based on the above notations, the overall training loss of the detector $\mathcal{L}$ becomes:

$$\mathcal{L} = \mathcal{L}_{\text{original}} + \lambda \cdot (\mathcal{L}_{\text{spatial-temp}} + \mathcal{L}_{\text{response}}), \tag{5}$$

where L_{original} indicates the original training loss of BEVFormer. λ is a hyper-parameter to balance the magnitudes of knowledge distillation loss, which is set to 1×10^{-2} in all the experiments. Please refer to the supplementary material for its sensitivity study.

Table 1: Comparison with other KD methods on the nuScenes [4] dataset with BEVFormer. Note that a higher mAP and NDS, as well as a lower ATE, ASE, AOE, and AAE indicate better performance. Params: the number of parameters (M). FPS: Frame per second. FPS is measured with one A100 GPU. Please refer to [4] for detailed metrics definitions.

Backbone	FPS	Params	KD Method	mAP(↑)	NDS(↑)	mATE(↓)	mASE(↓)	mAOE(↓)	mAVE(↓)	mAAE(↓)
ResNet101	10.2	56.57	Teacher w/o KD	36.31	47.49	69.21	28.16	46.08	43.87	19.32
ResNet50	14.5	40.45	Student w/o KD	33.56	44.61	71.41	28.65	54.17	46.44	21.03
			+ Hinton *et al.* [13]	33.57	45.23	71.17	28.50	49.04	46.52	20.33
			+ Zagoruyko *et al.* [33]	33.68	45.69	70.13	**27.74**	47.87	45.45	20.26
			+ Heo *et al.* [12]	33.87	45.82	69.92	27.79	47.78	45.55	20.09
			+ Park *et al.* [26]	33.77	45.87	70.88	27.78	48.18	43.47	19.83
			+ Pung *et al.* [28]	34.01	45.36	71.21	28.06	50.49	45.77	20.88
			+ Ahn *et al.* [1]	34.11	46.36	70.69	28.02	46.16	42.09	20.04
			+ Zhang *et al.* [34]	34.25	46.34	70.84	28.44	47.06	**41.68**	19.82
			+ Guo *et al.* [9]	34.10	46.22	70.39	28.39	46.75	42.52	20.22
			+ Ours	**34.91**	**46.87**	**69.77**	28.07	**46.31**	42.23	**19.43**
ResNeXt-Large	5.3	201.2	Teacher w/o KD	37.69	46.67	70.44	28.52	56.89	45.81	20.12
ResNet50	14.5	40.45	Student w/o KD	33.56	44.61	71.41	28.65	54.17	46.44	21.03
			+ Hinton *et al.* [13]	33.84	45.68	72.72	28.16	46.54	44.50	20.48
			+ Zagoruyko *et al.* [33]	34.10	46.26	70.99	28.24	46.12	42.45	20.05
			+ Heo *et al.* [12]	34.30	46.50	70.36	27.94	**44.78**	43.06	20.39
			+ Park *et al.* [26]	33.98	46.40	71.82	28.07	45.84	39.86	20.24
			+ Pung *et al.* [28]	34.23	46.23	70.05	28.32	47.33	43.13	20.04
			+ Ahn *et al.* [1]	34.16	46.25	70.37	28.08	46.43	42.73	20.66
			+ Zhang *et al.* [34]	34.56	46.61	70.11	28.01	46.00	42.39	20.14
			+ Guo *et al.* [9]	34.35	46.06	69.92	**27.79**	47.78	45.55	20.09
			+ Ours	**35.58**	**47.39**	**68.97**	28.25	48.06	**39.79**	**18.93**
ResNet101	3.5	65.93	Teacher w/o KD	41.01	51.88	67.45	27.36	34.92	37.57	18.97
ResNet50	5.2	47.56	Student w/o KD	35.77	46.74	73.61	28.26	45.85	43.79	19.94
			+ Hinton *et al.* [13]	35.89	46.93	73.45	**28.02**	45.46	43.66	19.58
			+ Zagoruyko *et al.* [33]	35.98	46.98	73.30	28.22	45.32	43.68	19.60
			+ Heo *et al.* [12]	36.23	47.16	73.09	28.18	45.28	43.34	19.69
			+ Park *et al.* [26]	36.30	47.18	72.94	28.17	45.48	43.43	19.64
			+ Pung *et al.* [28]	36.42	47.26	72.96	28.23	45.48	43.37	19.51
			+ Ahn *et al.* [1]	36.38	47.20	73.02	28.25	45.51	43.50	19.60
			+ Zhang *et al.* [34]	36.64	47.38	73.12	28.15	**45.28**	43.11	19.53
			+ Guo *et al.* [9]	36.77	47.40	73.14	28.25	45.34	43.43	19.74
			+ Ours	**38.88**	**48.52**	**71.53**	28.24	47.34	**42.91**	**19.17**

Weight-Inheriting Convnets-based detectors usually only require images as input. But modern transformer-based detection models require additional learned queries and positional encodings as input. The teacher and the student model tend to have different query and positional encoding values after training converges. Intuitively, knowledge distillation works by aligning the output of the student with the teacher given the same input. Such paradigm is likely to fail for transformer-based detectors, as the teacher and student can have different learned queries and position encodings. The discrepancies between the transformer inputs must be resolved to make the underlying assumptions of knowledge distillation hold true.

Hence, in this paper, we propose a weight-inheriting scheme that fixes the value of the BEV queries and positional encoding in the student with their values from the teacher detector *during the whole training period*. Hence, the student detector can have consistent inputs with its teacher detector. Surprisingly, we find that simply performing this weight-inheriting scheme can make a significant difference in the effectiveness of knowledge distillation, which has been discussed in the ablation study.

3 Experiment

Dataset and Model: The nuScenes dataset is a large-scale autonomous driving dataset, which has 3D bounding boxes for 1000 scenes collected from six cameras [4]. The scenes are

Table 2: Student-teacher settings in our experiments. Please refer to the supplementary material for more details.

Model	FPS	Params	2D Backbone	BEV Query	Decoder Depth
Student-1	14.5	40.45	ResNet50	(150, 150)	3
Teacher-1	10.2	56.57	ResNet101	(150, 150)	3
Student-2	14.5	40.45	ResNet50	(150, 150)	3
Teacher-2	5.3	201.20	ResNeXt-Large	(150, 150)	3
Student-3	5.2	47.56	ResNet50	(200, 200)	6
Teacher-3	3.5	65.93	ResNet101	(200, 200)	6

Table 3: Average precision in different classes on nuScenes. "KD" indicates whether our method is applied. Experiments of the three groups are conducted with student-teacher settings in Table 2.

KD	Car	Truck	Bus	Trailer	Con.Veh.	Pedest.	Motor.	Bicycle	Barrier	Tra.Cone
×	54.3	26.0	32.3	8.9	7.4	41.8	31.8	28.2	53.8	51.0
✓	55.1	27.4	34.2	10.1	6.8	43.4	34.2	31.1	53.9	52.9
×	54.3	26.0	32.3	8.9	7.4	41.8	31.8	28.2	53.8	51.0
✓	56.5	29.3	37.5	13.3	10.3	45.6	34.4	34.4	43.8	50.8
×	55.8	28.7	35.0	9.7	6.5	46.6	37.5	37.3	54.6	46.0
✓	58.7	33.1	36.2	12.8	10.1	47.4	40.4	40.8	57.6	51.9

officially split into 700, 150, and 150 scenes for training, validation, and testing, respectively, including 1.4 million annotated 3D bounding boxes belonging to 10 classes. BEVFormer models of different sizes are utilized as the student and teacher detectors in our experiments. As shown in Table 2, We mainly reduce the model size by using fewer BEV queries and smaller 2D backbones. Please refer to the supplementary material for more details on the models ,training settings and the implementation of comparison methods.

Experimental results of our method and eight previous knowledge distillation methods in three different student-teacher settings are shown in Table 1. It is observed that: (i) On average, 2.16 mAP and 2.27 NDS improvements can be observed with our method in the three student-teacher settings, which are 1.26 mAP and 0.80 NDS higher than the second-best knowledge distillation methods. (ii) In all three student-teacher settings, our method leads to performance improvements in terms of most of the performance metrics, including mAP, NDS, mATE, mATE, mASE, mAOE, mAVE, and mAAE, indicating that our method benefits students in estimating the translation, scale, orientation, velocity and attributes of the objects. (iii) The performance of our method in different categories is shown in Table 3. It is observed that our method leads to consistent improvements in most of the categories. (iv) The first student achieves 0.67 higher mAP than the second student, indicating that our method benefits from a strong teacher.

Table 4: Ablation study of different modules in our method. “Spatial-Temporal”, “BEV Response”, “Weight-Inherit” indicates spatial-temporal distillation, BEV response distillation, and the weight-inheriting scheme, respectively.

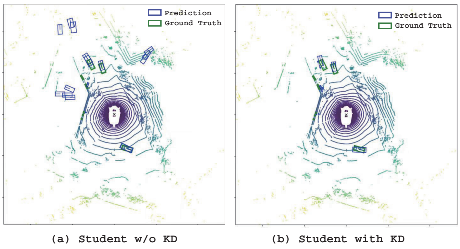
Modules in Our Method			mAP(↑)	NDS(↑)	mATE(↓)	mASE(↓)	mAOE(↓)	mAVE(↓)	mAAE(↓)
Spatial-Temporal	BEV Response	Weight-Inherit							
×	×	×	33.56	44.61	71.41	28.65	54.17	46.44	21.03
×	×	✓	34.52	46.60	70.97	28.05	46.00	41.74	19.86
×	✓	✓	34.99	47.17	70.22	27.75	46.58	39.47	19.24
✓	×	✓	34.91	47.02	70.62	27.90	47.26	39.60	18.93
✓	✓	×	35.00	46.68	71.07	28.43	46.09	42.46	20.19
✓	✓	✓	35.58	47.39	68.97	28.25	48.06	39.79	18.93

4 Discussion

4.1 Ablation Study

The proposed knowledge distillation methods mainly have three modules, including spatial-temporal distillation, the BEV embedding distillation, and the weight-inheriting scheme. Table 1 gives the ablation study of the three modules. It is observed that: (i) By simply using the weight-inheriting scheme without applying any knowledge distillation loss, 0.96 mAP and 1.99 NDS improvements can be obtained, indicating that the student detector can benefit from using the pre-trained weights from teachers on the BEV queries and positional encoding. (ii) By applying BEV response distillation and weight-inheriting, 1.43 mAP and 2.56 NDS improvements can be observed, which are 0.47 and 0.57 higher than only using weight-inheriting, indicating BEV response distillation can successfully transfer teacher knowledge to the student. (iii) 1.35 mAP and 2.41 NDS improvements can be obtained by using spatial-temporal distillation and weight-inheriting, which are 0.39 and 0.42 higher than only using weight-inheriting, indicating spatial-temporal distillation allows the student to learn how to fuse information from different timestamps and views from its teacher. (iv) By combining the three modules together, 1.67 mAP and 2.58 NDS improvements can be obtained, which demonstrates that the benefits of spatial-temporal distillation and BEV response distillation are orthogonal. (v) By only using the two knowledge distillation while disabling the weight-inheriting scheme, 1.44 mAP and 2.07 NDS improvements can be observed, which are 0.58 and 0.71 lower than performing knowledge distillation with weight-inheriting, indicating weight-inheriting is also indispensable even if knowledge distillation losses are applied. In summary, these experimental results demonstrate that the three modules in our method have their own effectiveness and their merits are orthogonal.

Ablation on Weight-Inheriting To facilitate the training of the student model, some previous knowledge distillation methods propose initializing the parameters of the student with the parameters of the teacher (*i.e.,* initialization scheme), which sometimes leads to slight performance improvements. In contrast, the proposed weight-inheriting scheme in this paper not only initializes the parameters of BEV queries and positional encoding with their value from the teacher but also freezes them during the whole training period (*i.e.,* weight-inheriting scheme). To study their difference, we have conducted several experiments and found that (i) By using the initialization scheme, after the training of the student, the parameters of BEV queries and positional encoding in the student are totally different from them in the teacher, indicating the inconsistency problem between the students and the teachers in knowledge distillation still exist. (ii) Experimental results show that by only using the



(a) Student w/o KD

(b) Student with KD

Figure 3: Visualization of detection results. Boxes in blue and green indicate prediction and the grounding truth, respectively.

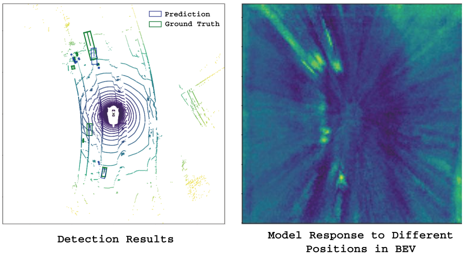
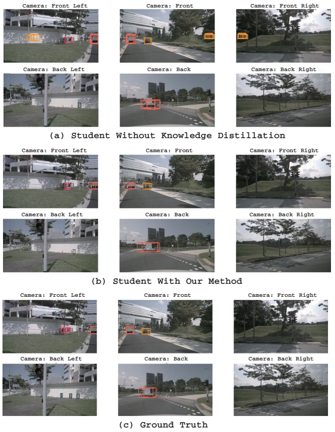


Figure 4: Visualization of model response to different BEV positions and the corresponding detection results.

traditional initialization scheme, the student detector (student-1 in Table 2) achieves 33.63 mAP and 44.70 NDS, which are 0.7 and 0.9 higher than the baseline, but still 0.89 and 1.90 lower than the weight-inheriting scheme. These observations indicate that using such a weight-inheriting scheme which exactly guarantees the consistency between the inputs of the students and teachers is indispensable.

4.2 Visualization



(a) Student Without Knowledge Distillation

(b) Student With Our Method

(c) Ground Truth

Figure 5: Visualization of detection results in different views.

Detection Results Figure 5 and Figure 3 visualize the detection results of the student detector trained without and with our method from the perspective of different camera views and bird-eye-view, respectively. Note that the used student detector has 6.4 FPS and 40.45M parameters. It is observed that the student trained by our method produces impressive results which are similar to the ground truth. In contrast, the student trained without knowledge distillation generates incorrect predictions in the cameras of the front-left view, the front-view, and the front-right view. As shown in their BEV visualization, the mistakes made by the student trained without knowledge distillation have a relatively long distance from the car, indicating the student trained without knowledge distillation is unable to detect the faraway objects while our method can address this problem.

BEV Response Figure 4 shows the BEV response and the corresponding detection results from the student detector. Note that a lighter pixel in BEV response map indicates the detector has a higher response. It is observed that the detector tends to show a higher response in the position where objects exist, indicating that BEV response contains valuable semantic information about the localization of objects. Hence, the proposed BEV response distillation can improve the ability of localization of the student detector by training it to imitate the BEV response from its teacher.

Attention Weights The proposed spatial-temporal distillation enables the student to learn teacher knowledge on information fusion by training it to mimic the attention weights in temporal self-attention and spatial cross-attention. Figure 6 gives the visualization results of attention weights from the teacher, the student trained with knowledge distillation and the

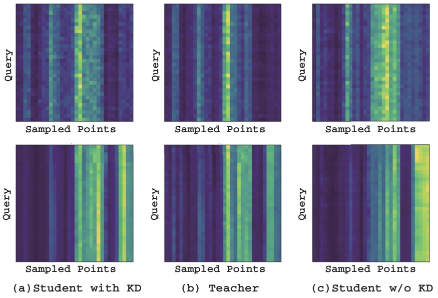


Figure 6: Visualization of attention weights in temporal cross-attention from the teacher, the student trained with and without KD.

student trained without knowledge distillation. It is observed that: (i) Compared with the student detectors, the attention weights from the teacher tend to concentrate more on several sampled points, indicating the teacher detector is able to leverage the information from certain images. (ii) Compared with the student trained without knowledge distillation, the attention weights from the student trained with knowledge distillation are more similar to the attention weights of the teacher, indicating that the spatial-temporal distillation successfully enables the student to mimic the teacher detector.

5 Conclusion

Most advanced multi-view 3D detectors suffer from low inference efficiency, which has limited their applications in edge devices. To address this problem, we propose a series of knowledge distillation methods to achieve model compression, which includes (1) spatial-temporal distillation which allows the student to learn how to fuse information from different timestamps and views (2) BEV response distillation which enables the student to learn the localization-aware knowledge, and (3) a weight-inheriting scheme which fixes the BEV queries and positional encoding to guarantee that students and teachers have the same inputs. Comparison experiments with 8 previous methods and sufficient ablation studies demonstrate the significant performance of our method in three different student-teacher settings. On average, 2.16 mAP and 2.27 NDS improvements can be observed on the nuScenes. We hope that this paper may promote more research on efficient multi-view 3D detection.

6 Acknowledgement

This research was partially supported by National Key R&D Program of China (2022YFB 2804103), Key Research and Development Program of Shaanxi (2021 ZDLGY01-05), Tsinghua University Dushi Program, National Natural Science Foundation of China (2021 1710187), and Tsinghua University Talent Program.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 9163–9171, 2019.

[2] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10925–10934, 2022.

[3] Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, pages 535–541. ACM, 2006.

[4] 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 *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 11618–11628. Computer Vision Foundation / IEEE, 2020.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Eur. Conf. Comput. Vis. (ECCV)*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020.

[6] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pages 742–751, 2017.

[7] Hanting Chen, Yunhe Wang, Han Shu, Changyuan Wen, Chunjing Xu, Boxin Shi, Chao Xu, and Chang Xu. Distilling portable generative adversarial networks for image translation. In *AAAI Conf. Artif. Intell. (AAAI)*, volume 34, pages 3585–3592, 2020.

[8] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. PACT: parameterized clipping activation for quantized neural networks. *CoRR*, abs/1805.06085, 2018.

[9] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2154–2164. Computer Vision Foundation / IEEE, 2021.

[10] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *Int. Conf. Learn. Represent. (ICLR)*, 2016.

[11] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 578–587, 2019.

[12] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Int. Conf. Comput. Vis. (ICCV)*, pages 1921–1930, 2019.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2014.

[14] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3588–3597, 2018.

[15] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.

[16] 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.

[17] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 6356–6364, 2017.

[18] 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.

[19] 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.

[20] Zeqi Li, Ruowei Jiang, and Parham Aarabi. Semantic relation preserving knowledge distillation for image-to-image translation. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 648–663. Springer, 2020.

[21] 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.

[22] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2604–2613, 2019.

[23] 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.

[24] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022.

[25] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[26] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3967–3976, 2019.

[27] Yuxi Ren, Jie Wu, Xuefeng Xiao, and Jianchao Yang. Online multi-granularity distillation for GAN compression. In *Int. Conf. Comput. Vis. (ICCV)*, pages 6773–6783. IEEE, 2021.

[28] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Int. Conf. Comput. Vis. (ICCV)*, pages 1365–1374, 2019.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pages 5998–6008, 2017.

[30] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021.

[31] 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 *Conference on Robot Learning*, pages 180–191. PMLR, 2022.

[32] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[33] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Int. Conf. Learn. Represent. (ICLR)*, 2017.

[34] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *Int. Conf. Learn. Represent. (ICLR)*, 2021.

[35] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *arXiv preprint:1905.08094*, 2019.

[36] Linfeng Zhang, Xin Chen, Runpei Dong, and Kaisheng Ma. Region-aware knowledge distillation for efficient image-to-image translation. *arXiv preprint arXiv:2205.12451*, 2022.

[37] Linfeng Zhang, Xin Chen, Xiaobing Tu, Pengfei Wan, Ning Xu, and Kaisheng Ma. Wavelet knowledge distillation: Towards efficient image-to-image translation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022.

[38] Aojun Zhou, Anbang Yao, Yiwen Guo, Lin Xu, and Yurong Chen. Incremental network quantization: Towards lossless cnns with low-precision weights. *arXiv preprint arXiv:1702.03044*, 2017.

[39] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.