MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation

Xiaoshuai Hao[1†], Ruikai Li[2†], Hui Zhang[1], Dingzhe Li[1], Rong Yin[3], Sangil Jung[4], Seung-In Park[4], ByungIn Yoo[4], Haimei Zhao[5§], and Jing Zhang[5§]

[1] Samsung R&D Institute China–Beijing
[2] State Key Lab of Intelligent Transportation System, Beihang University
[3] Institute of Information Engineering, Chinese Academy of Sciences
[4] Computer Vision TU, SAIT, SEC, Korea [5] The University of Sydney
{xshuai.hao, hui123.zhang, dingzhe.li, byungin.yoo}@samsung.com
ricky_developer@buaa.edu.cn {si14.park, sang-il.jung}@samsung.com
yinrong@iie.ac.cn hzha7798@uni.sydney.edu.au jingzhang.cv@gmail.com

Abstract. Online high-definition (HD) map construction is an important and challenging task in autonomous driving. Recently, there has been a growing interest in cost-effective multi-view camera-based methods without relying on other sensors like LiDAR. However, these methods suffer from a lack of explicit depth information, necessitating the use of large models to achieve satisfactory performance. To address this, we employ the Knowledge Distillation (KD) idea for efficient HD map construction for the first time and introduce a novel KD-based approach called MapDistill to transfer knowledge from a high-performance camera-LiDAR fusion model to a lightweight camera-only model. Specifically, we adopt the teacher-student architecture, *i.e.*, a camera-LiDAR fusion model as the teacher and a lightweight camera model as the student, and devise a dual BEV transform module to facilitate cross-modal knowledge distillation while maintaining cost-effective camera-only deployment. Additionally, we present a comprehensive distillation scheme encompassing cross-modal relation distillation, dual-level feature distillation, and map head distillation. This approach alleviates knowledge transfer challenges between modalities, enabling the student model to learn improved feature representations for HD map construction. Experimental results on the challenging nuScenes dataset demonstrate the effectiveness of MapDistill, surpassing existing competitors by over 7.7 mAP or 4.5× speedup.

Keywords: HD Map Construction · Knowledge Distillation · Lightweight

1 Introduction

Online high-definition (HD) map provides abundant and precise static environmental information about the driving scenes, which is fundamental for planning

† The first two authors contributed equally to this work.
§ Corresponding authors.

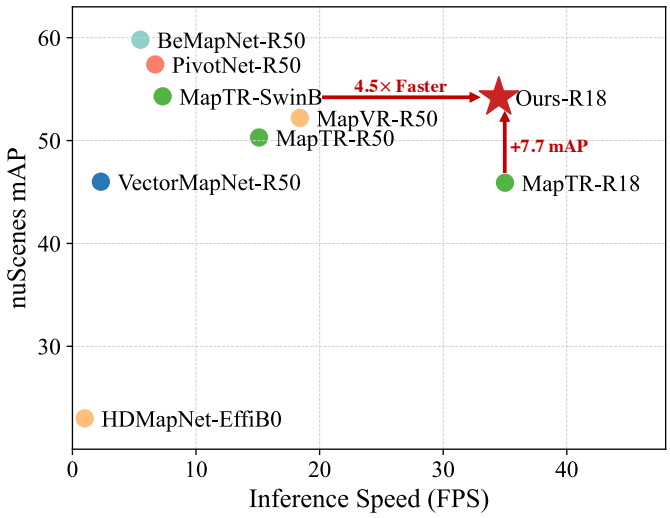

Fig. 1: Comparison of different methods on the nuScenes val dataset. We benchmark the inference speed on a single NVIDIA RTX 3090 GPU. Our method can achieve a better trade-off in both speed (FPS) and accuracy (mAP).

and navigation in autonomous driving systems. Recently, multi-view camera-based [7, 23, 34] HD map construction has gained increasing attention thanks to the significant progress of Bird's-Eye-View (BEV) perception. Compared with LiDAR-based [13, 39] and Fusion-based methods [19, 23, 25], multi-view camera-based methods can be deployed at low cost, while the lack of depth information makes current approaches adopt large models for effective feature extraction and good performance achievement. Therefore, it is crucial to trade off the performance and efficiency of the camera-based model for practical deployment.

To achieve this goal, Knowledge Distillation (KD) [8] has drawn great attention in related fields since it is one of the most practical techniques for training efficient yet accurate models. KD-based methods usually transfer knowledge from a large well-trained model (teacher) to a small model (student) [14], which has made remarkable progress in many fields, such as image classification [31], 2D object detection [3], semantic segmentation [43] and 3D object detection [5, 51, 53]. Previous methods follow the well-known teacher-student paradigm [14], which forces the logits of the student network to match those of the teacher network. Recently, BEV-based KD methods have advanced the field of 3D object detection, which unify the image and LiDAR features in the Bird-Eye-View (BEV) space and adaptively transfer knowledge across non-homogenous representations in a teacher-student paradigm. Existing works use a strong LiDAR teacher model to distill a camera student model, such as BEVDistill [5], UVTR [20], BEVLGKD [18], TiG-BEV [15] and DistillBEV [41]. Furthermore, the latest work UniDistill [53] proposes a universal cross-modality knowledge distillation framework for 3D Object Detection.

Compared to these methods, BEV-based HD map construction KD method differs in two crucial aspects: Firstly, the detection head (DetHead) produces the output of classification and localization for objects, while the output of the map head (MapHead) from a vectorized map construction model, *e.g.* MapTR [23], is the classification and point regression result. Secondly, existing BEV-based KD methods for 3D object detection typically focus on aligning foreground objects' features to mitigate the background environment's adverse impact, which is obviously unsuitable for HD map construction. Therefore, directly applying the BEV-based KD method for 3D object detection to HD map construction fails to achieve satisfying results (see the experiment results in Tab. 1) due to the inherent dissimilarity between the two tasks. To the best of our knowledge, BEV-based KD methods for HD map construction are still under exploration.

To fill this gap, we propose a novel KD-based method named MapDistill to transfer the knowledge from a high-performance teacher model to an efficient student model. First, we adopt the teacher-student architecture, *i.e.*, a camera-LiDAR fusion model as the teacher and a lightweight camera model as the student, and devise a dual BEV transform module to facilitate cross-modal knowledge distillation while maintaining cost-effective camera-only deployment. Building upon this architecture, we propose a comprehensive distillation scheme encompassing cross-modal relation distillation, dual-level feature distillation, and map head distillation, to mitigate the knowledge transfer challenges between modalities and help the student model learn improved feature representations for HD map construction. Specifically, we first introduce the cross-modal relation distillation loss for the student model to learn better cross-modal representations from the fusion teacher model. Second, to achieve better semantic knowledge transfer, we employ the dual-level feature distillation loss on both the low-level and high-level feature representations in the unified BEV space. Last but not least, we specifically introduce a map head distillation loss tailored for the HD map construction task, including classification loss and point-to-point loss, which can make the final predictions of the student closely resemble those of the teacher. Extensive experiments on the challenging nuScenes dataset [2] demonstrate the effectiveness of MapDistill, surpassing existing competitors by over 7.7 mAP or 4.5× speedup as shown in Fig. 1.

The contributions of this paper are mainly three-fold:

- We present an effective model architecture for distillation-based HD map construction, including a camera-LiDAR fusion teacher model, a lightweight camera-only student model, and a dual BEV transform module, which facilitates knowledge transfer within and between different modalities while enjoying cost-effective camera-only deployment.
- We introduce a comprehensive distillation scheme that supports cross-modal relation distillation, dual-level feature distillation, and map head distillation simultaneously. By mitigating the knowledge transfer challenges between modalities, this distillation scheme helps the student model learn better feature representation for HD map construction.

– MapDistill achieves superior performance than state-of-the-art (SOTA) methods, which could serve as a strong baseline for KD-based HD map construction research.

2 Related Work

Camera-based HD Map Construction. HD map construction is a prominent and extensively researched area within the field of autonomous driving. Recently, camera-based methods [7, 9, 10, 16, 19, 23, 25, 34, 47] have increasingly employed the Bird's-eye view (BEV) representation as an ideal feature space for multi-view perception due to its remarkable ability to mitigate scale-ambiguity and occlusion challenges. Various techniques have been proposed and utilized to project perspective view (PV) features into the BEV space by leveraging geometric priors, such as LSS [33], Deformable Attention [21] and GKT [4]. Furthermore, camera-based methods have come to rely on higher resolution images and larger backbone models to achieve enhanced accuracy [17,21,26,27,40,42,45,50], a practice that introduces substantial challenges for practical deployment. For example, HDMapNet [19] and VectorMapnet [25] employ the Efficient-B0 model [37] and ResNet50 model, respectively, as backbones for feature extraction. Additionally, MapTR [23] explores the impact of various backbones, including the Swin Transformer [27], ResNet50 [12], and Resnet18 [12]. Experimental results demonstrate a direct correlation between the backbone's representation capability and model performance, *i.e.*, larger models generally yield better results. Yet, using larger models leads to slower inference, compromising the cost advantage of camera-based methods. In this paper, we introduce an effective yet efficient camera-based method tailored for practical deployment via knowledge distillation.

Fusion-based HD Map Construction. LiDAR-based methods [11, 13, 19, 25, 39] provide precise spatial data for creating the BEV feature representation. Recently, camera-LiDAR fusion methods [1, 19, 22–24, 28, 35, 38] leverage the semantic richness of camera data and the geometric information from LiDAR in a collaborative manner. This fusion at the BEV level incorporates distinct streams, encoding camera and LiDAR inputs into shared BEV features, surpassing unimodal input approaches in performance. However, this integration may impose significant computational and cost burdens in practical deployment. To address this issue, we leverage KD techniques for efficient HD map construction and introduce a novel approach called MapDistill to transfer knowledge from a high-performance camera-LiDAR fusion model to a lightweight camera-only model, yielding a cost-effective yet accurate solution.

Knowledge Distillation. KD refers to transferring knowledge from a well-trained, larger teacher model to a smaller student [14], which has been widely applied across diverse tasks, such as image classification [31, 48, 49], 2D object detection [3,52], semantic segmentation [36,39,43,46] and 3D object detection [5, 6, 51, 53]. Recently, BEV-based KD methods have gained increasing attention in the field of 3D object detection. Several existing works have adopted cross-modality knowledge distillation frameworks for 3D object detection, including

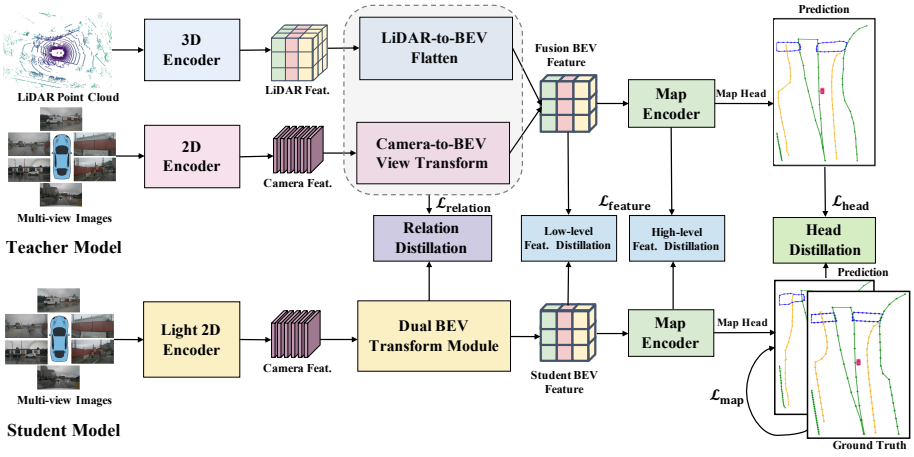

Fig. 2: The overview of our proposed MapDistill. It consists of a fusion-based teacher model (top) and a lightweight camera-based student model (bottom). In addition, three distillation losses are employed to enable the teacher model to transfer knowledge to the student, *i.e.*, by instructing the student model to produce similar features and predictions, which are cross-modal relation distillation ($\mathcal{L}_{relation}$), dual-level feature distillation ($\mathcal{L}_{feature}$), and map head distillation ($\mathcal{L}_{head}$). Note that only the student model is needed for inference.

BEVDistill [5], UVTR [20], BEV-LGKD [18], TiG-BEV [15], DistillBEV [41], and UniDistill [53]. Despite the numerous KD methods for 3D object detection, KD-based HD map construction remains relatively under-explored. In this paper, we fill this gap by proposing a novel KD-based approach called MapDistill to boost efficient camera-based HD map construction via camera-LiDAR fusion model distillation.

3 Methodology

In this section, we describe our proposed MapDistill in detail. We first give an overview of the whole framework in Fig. 2 and clarify the model designs of the teacher and student models in Sec. 3.1. Then, we elaborate details of MapDistill objectives in Sec. 3.2, such as the cross-modal relation distillation, the dual-level feature distillation, and the map head distillation. Finally, we present the overall training procedure in Sec. 3.3.

3.1 Model Overview

Fusion-based Model (Teacher). To enable the knowledge transfer from the camera-LiDAR fusion teacher model to the student model, we first establish a baseline of fusion-based HD map construction based on the state-of-the-art MapTR [23] model. The fused MapTR model has two branches, as depicted in

the top part of Fig. 2. For the camera branch, it firstly utilizes **Resnet50** [12] as the backbone to extract multi-view features. Next, it uses GKT [4] as the 2D-to-BEV transformation module to convert the multi-view features into the BEV space. The generated camera BEV features can be denoted as $\mathbf{F}^T_{C_{bev}} \in \mathbb{R}^{H\times W\times C}$, where H, W, C represents the height, width and the number of channels of BEV features respectively, and the superscript T is short for "teacher". For the LiDAR branch, it adopts **SECOND** [44] for point cloud voxelization and LiDAR feature encoding. The LiDAR features are projected to BEV space using a flattening operation as in [28], to obtain the LiDAR BEV representation $\mathbf{F}^T_{L_{bev}} \in \mathbb{R}^{H\times W\times C}$. Then, MapTR concatenates $\mathbf{F}^T_{C_{bev}}$ and $\mathbf{F}^T_{L_{bev}}$ and processes the features with the fully convolutional network to produce the fused BEV features $\mathbf{F}^T_{fused} \in \mathbb{R}^{H\times W\times C}$.

The following step is to use a Map Encoder (MapEnc), which takes the fused BEV features $\mathbf{F}^T_{fused}$ as input, to further generate the high-level feature $\mathbf{F}^T_{high}$:

$$\mathbf{F}^T_{high} = \text{MapEnc}(\mathbf{F}^T_{fused}), \tag{1}$$

Then, the teacher Map head (MapHead) employs the classification and point branches to produce the final predictions of map elements categories $\mathbf{F}^T_{cls}$ and point positions $\mathbf{F}^T_{point}$:

$$\mathbf{F}^T_{cls}, \mathbf{F}^T_{point} = \text{MapHead}(\mathbf{F}^T_{high}). \tag{2}$$

During the overall training procedure, the teacher model will continuously produce diverse features $\mathbf{F}^T_{C_{bev}}$, $\mathbf{F}^T_{L_{bev}}$, $\mathbf{F}^T_{fused}$, $\mathbf{F}^T_{high}$, $\mathbf{F}^T_{cls}$ and $\mathbf{F}^T_{point}$.

Camera-based Model (Student). To realize real-time inference speed for practical deployment, we adopt MapTR's camera branch as the base for the student model. Note that we employ **Resnet18** [12] as the backbone to extract the multi-view features, which can make the network lightweight and easy to deploy. On the base from MapTR, to mimic the multimodal fusion pipeline of the teacher model, we propose a Dual BEV Transform module to convert the multi-view features into two distinct BEV subspaces, whose effect will be verified in the ablation experiments. Specifically, we firstly use GKT [4] to generate BEV features in the first subspace $\mathbf{F}^S_{C_{sub1}} \in \mathbb{R}^{H\times W\times C}$, where the superscript S is short for "student". Then, we utilize LSS [33] to generate BEV features in the second subspace $\mathbf{F}^S_{C_{sub2}} \in \mathbb{R}^{H\times W\times C}$. Then, we concatenate $\mathbf{F}^S_{C_{sub1}}$ and $\mathbf{F}^S_{C_{sub2}}$ and process the features with the fully convolutional network to produce the fused BEV features $\mathbf{F}^S_{fused} \in \mathbb{R}^{H\times W\times C}$.

Then, employing the same process as the teacher model, we can generate $\mathbf{F}^S_{high}$, $\mathbf{F}^S_{cls}$ and $\mathbf{F}^S_{point}$ from $\mathbf{F}^S_{fused}$ with Eq. 1 and Eq. 2. Therefore, the student model will consistently produce $\mathbf{F}^S_{C_{sub1}}$, $\mathbf{F}^S_{C_{sub2}}$, $\mathbf{F}^S_{fused}$, $\mathbf{F}^S_{high}$, $\mathbf{F}^S_{cls}$ and $\mathbf{F}^S_{point}$ during the procedure of map construction.

3.2 MapDistill Objectives

Cross-modal Relation Distillation. The teacher model, a camera-LiDAR fusion model, combines semantic-rich information from camera data with explicit

geometric data from LiDAR. In contrast, the student model, a camera-based model, focuses mainly on capturing semantic information from the camera. The essential factor contributing to the teacher model's superior performance is cross-modal interaction, which the student model lacks. Therefore, we encourage the student model to develop this cross-modal interaction capability through imitation.

To this end, we introduce a cross-modal attention distillation objective. The core idea is to let the student model imitate the cross-modal attention of the teacher model during training. More specifically, for the teacher model, we begin by reshaping the camera BEV features $\mathbf{F}^T_{C_{bev}} \in \mathbb{R}^{H\times W\times C}$ and the LiDAR BEV features $\mathbf{F}^T_{L_{bev}} \in \mathbb{R}^{H\times W\times C}$ into sequences of 2D patches represented as $\mathbf{Fp}^T_{C_{bev}} \in \mathbb{R}^{N\times(P^2C)}$ and $\mathbf{Fp}^T_{L_{bev}} \in \mathbb{R}^{N\times(P^2C)}$, respectively. Here, the patch size is denoted as $P \times P$, and the number of patches is given by $N = HW/P^2$.

Then, we calculate the cross-modal attention from the teacher, including camera-to-lidar attention $\mathbf{A}^T_{c2l} \in \mathbb{R}^{N\times N}$ and lidar-to-camera attention $\mathbf{A}^T_{l2c} \in \mathbb{R}^{N\times N}$ as follows:

$$A^T_{c2l} = \text{softmax}\left(\frac{\mathbf{Fp}^T_{C_{bev}}\ \text{Transpose}\big(\mathbf{Fp}^T_{L_{bev}}\big)}{\sqrt{D_k}}\right), \tag{3}$$

$$A^T_{l2c} = \text{softmax}\left(\frac{\mathbf{Fp}^T_{L_{bev}}\ \text{Transpose}\big(\mathbf{Fp}^T_{C_{bev}}\big)}{\sqrt{D_k}}\right), \tag{4}$$

where $\frac{1}{\sqrt{D_k}}$ is a scaling factor for preventing the softmax function from falling into a region with extremely small gradients when the magnitude of dot products grow large.

For the student model, we adopt the same operation as the teacher model to generate $\mathbf{Fp}^S_{C_{sub1}} \in \mathbb{R}^{N\times(P^2C)}$ and $\mathbf{Fp}^S_{C_{sub2}} \in \mathbb{R}^{N\times(P^2C)}$ from $\mathbf{F}^S_{C_{sub1}}$ and $\mathbf{F}^S_{C_{sub2}}$, respectively, and then compute the cross-modal attention of the student $\mathbf{A}^S_{c2l}$, $\mathbf{A}^S_{l2c}$ as follows:

$$A^S_{c2l} = \text{softmax}\left(\frac{\mathbf{Fp}^S_{C_{sub1}}\ \text{Transpose}\big(\mathbf{Fp}^S_{C_{sub2}}\big)}{\sqrt{D_k}}\right), \tag{5}$$

$$A^S_{l2c} = \text{softmax}\left(\frac{\mathbf{Fp}^S_{C_{sub2}}\ \text{Transpose}\big(\mathbf{Fp}^S_{C_{sub1}}\big)}{\sqrt{D_k}}\right). \tag{6}$$

To this end, we propose the cross-modal relation distillation and employ a KL-divergence loss to align the cross-modal attention $\mathbf{A}^S_{c2l}$ and $\mathbf{A}^S_{l2c}$ of the student with $\mathbf{A}^T_{c2l}$ and $\mathbf{A}^T_{l2c}$ of the teacher model:

$$\mathcal{L}_{relation} = D_{KL}(A^T_{c2l}||A^S_{c2l}) + D_{KL}(A^T_{l2c}||A^S_{l2c}). \tag{7}$$

Dual-level Feature Distillation. To facilitate the student model to absorb the rich semantic/geometric knowledge from the teacher model, we take advantage of the fused BEV features for the feature-level distillation. Specifically, we leverage

the low-level fused BEV feature of the teacher $\mathbf{F}^T_{fused}$ as the supervisory signal for learning the counterpart of the student $\mathbf{F}^S_{fused}$ via an MSE loss, *i.e.*,

$$\mathcal{L}_{low} = \text{MSE}(\mathbf{F}^T_{fused}, \mathbf{F}^S_{fused}). \tag{8}$$

In addition, we further propose the high-level feature distillation $\mathcal{L}_{high}$ to align $\mathbf{F}^T_{high}$ and $\mathbf{F}^S_{high}$, which are generated by the Map Encoder. $\mathcal{L}_{high}$ is defined as:

$$\mathcal{L}_{high} = \text{MSE}(\mathbf{F}^T_{high}, \mathbf{F}^S_{high}). \tag{9}$$

Formally, the dual-level feature distillation loss $\mathcal{L}_{features}$ is the sum of low-level distillation loss $\mathcal{L}_{low}$ and high-level distillation loss $\mathcal{L}_{high}$, *i.e.*,

$$\mathcal{L}_{feature} = \mathcal{L}_{low} + \mathcal{L}_{high}. \tag{10}$$

We use $\mathcal{L}_{feature}$ as one of the distillation objectives to enable the student model to benefit from the teacher model implicitly during training.

Map Head Distillation. After the Map Encoder, the high-level BEV feature in the student model is fed into the HD Map Head to produce the prediction in the same way as the teacher model. To make the final prediction of the student close to that of the teacher, we further propose the map head distillation. Specifically, we use the predictions generated by the teacher model as pseudo labels to supervise the student model via the $\mathcal{L}_{head}$ loss. To achieve the goal, we need to construct the correspondence between the predictions of the student and the teacher. Suppose the classification and point predictions from the teacher model are $\mathbf{F}^T_{cls}$ and $\mathbf{F}^T_{point}$ respectively, and those from the student can be represented as $\mathbf{F}^S_{cls}$ and $\mathbf{F}^S_{point}$ respectively. The $\mathcal{L}_{head}$ loss consists of two parts, *i.e.*, the classification loss $\mathcal{L}_{cls}$ for map elements classification and the point2point loss $\mathcal{L}_{point}$ for point position regression:

$$\begin{aligned} \mathcal{L}_{head} &= \mathcal{L}_{cls} + \mathcal{L}_{point} \\ &= \mathcal{L}_{Focal}(\mathbf{F}^T_{cls}, \mathbf{F}^S_{cls}) + \mathcal{L}_{p2p}(\mathbf{F}^T_{point}, \mathbf{F}^S_{point}), \end{aligned} \tag{11}$$

where $\mathcal{L}_{Focal}$ denotes the Focal loss [32] and $\mathcal{L}_{p2p}$ denotes the Manhattan distance [30] between $\mathbf{F}^T_{point}$ and $\mathbf{F}^S_{point}$.

3.3 Overall Training

To facilitate knowledge transfer from the multi-modal fusion-based teacher model to the camera-based student model, we integrate the map loss $\mathcal{L}_{map}$ with the above distillation losses, including the cross-modal relation distillation loss ($\mathcal{L}_{relation}$), the dual-level feature distillation loss ($\mathcal{L}_{feature}$), and the map head distillation loss ($\mathcal{L}_{head}$). The overall training objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_{map} + \lambda_1 \mathcal{L}_{relation} + \lambda_2 \mathcal{L}_{feature} + \lambda_3 \mathcal{L}_{head}, \tag{12}$$

where λ_1, λ_2 and λ_3 are hyper-parameters for balancing these terms. The map loss $\mathcal{L}_{map}$ is calculated following [23], which is composed of three parts, *i.e.*, classification loss, point2point loss, and edge direction loss.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our method on the widely-used challenging nuScenes [2] dataset following the standard setting of previous methods [19, 23, 25]. The nuScenes dataset contains 1,000 sequences of recordings collected by autonomous driving cars. Each sample is annotated at 2Hz and contains 6 camera images covering 360° horizontal FOV of the ego-vehicle. Following [19,23,25], three kinds of map elements are chosen for fair evaluation – pedestrian crossing, lane divider, and road boundary.

Evaluation Metrics. We adopt the evaluation metrics used in previous works [19,23,25]. Specifically, average precision (AP) is used to evaluate the map construction quality. Chamfer distance $D_{Chamfer}$ is used to determine whether the prediction and GT are matched or not. We calculate the AP_τ under several $D_{Chamfer}$ thresholds ($\tau \in T = \{0.5, 1.0, 1.5\}$), and then average across all thresholds as the final mean AP (mAP) metric:

$$mAP = \frac{1}{|T|} \sum_{\tau \in T} AP_\tau. \tag{13}$$

The perception ranges are [-15.0m, 15.0m]/[-30.0m, 30.0m] for X/Y-axes.

Model and Training Details. MapDistill is trained with 8 NVIDIA RTX A6000 GPUs. For the teacher model, we first establish a baseline method of fusion-based HD map construction based on MapTR [23]. The fused MapTR model uses ResNet50 [12] and SECOND [44] as the backbone and employ GKT [4] as the default 2D-to-BEV module. For the student model, we adopt MapTR's camera branch as the base, and introduce the dual BEV transform module to facilitate cross-modal knowledge distillation. Note that, the student model adopts ResNet18 [12] as the backbone. Moreover, we adopt the AdamW optimizer [29] for all our experiments. The setting of hyper-parameters λ_1, λ_2, and λ_3 is discussed extensively in the ablation studies. We set the mini-batch size to 64, and use a step-decayed learning rate with an initial value of $4e^{-3}$.

4.2 Comparison with the State-of-the-Arts

We compare our method with several state-of-the-art baselines across two categories, *i.e.*, camera-based HD map construction methods, and customized KD methods which were originally designed for BEV-based 3D object detection. For camera-based HD map construction methods, we directly report the results from the corresponding papers. For KD-based methods, we implement three methods for BEV-based 3D object detection and modify them for the HD map construction task, including BEV-LGKD [18], BEVDistill [5], and UnDistill [53]. For fairness, we use the same teacher and student models as our method.

Tab. 1 shows that: (1) KD methods originally designed for BEV-based 3D object detection fail to achieve satisfying results due to task discrepancies between 3D object detection and HD map construction. (2) Intra-modal distillation between camera-only teacher and student models cannot learn accurate

Table 1: Performance analysis of MapDistill on nuScenes val set. "L" and "C" represent the LiDAR and camera, respectively. "Effi-B0", "R18", "R50", and "Sec" are short for EfficientNet-B0 [37], ResNet18 [12], ResNet50 [12], and SECOND [44], respectively. We adopt the MapTR method to build the teacher model and the student model. Note that the directly-trained MapTR models in the red region are selected as teachers. Our proposed MapDistill outperforms all existing approaches in both single-class APs and the overall mAP by a significant margin. † denotes our re-implementation following the setting in the paper. Best viewed in color.

Method	Student Modality	Teacher Modality	Backbone	Epochs	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
HDMapNet [19]	C	–	Effi−B0	30	14.4	21.7	33.0	23.0
VectorMapNet [25]	C	–	R50	110	36.1	47.3	39.3	40.9
MapVR [47]	C	–	R50	24	47.7	54.4	51.4	51.2
PivotNet [7]	C	–	R50	30	58.5	53.8	59.6	57.4
BeMapNet [34]	C	–	R50	30	62.3	57.7	59.4	59.8
MapTR [23]	C	–	R50	24	45.3	51.5	53.1	50.3
MapTR [23]	L	–	Sec	24	48.5	53.7	64.7	55.6
MapTR [23]	C & L	–	R50 & Sec	24	55.9	62.3	69.3	62.5
MapTR [23]	C	–	R18	110	39.6	49.9	48.2	45.9
BEV-LGKD† [18]	C	C	R18	110	42.2	47.6	49.7	$46.5_{+0.6}$
BEVDistill† [5]	C	L	R18	110	42.4	48.5	50.2	$47.1_{+1.2}$
UniDistill† [53]	C	C&L	R18	110	43.9	48.6	52.1	$48.2_{+2.3}$
MapDistill	C	C	R18	110	43.3	48.8	51.9	$48.0_{+2.1}$
MapDistill	C	L	R18	110	45.9	50.7	53.6	$50.1_{+4.2}$
MapDistill	**C**	**C & L**	**R18**	**110**	**49.2**	**54.5**	**57.1**	$\mathbf{53.6}_{+7.7}$

3D information due to the limited capacity of the teacher model for inferring 3D geometry, and the gain is only 0.6 mAP by BEV-LGKD and 2.1 mAP by our MapDistill. (3) Cross-modal distillation between the LiDAR teacher and the camera student enables learning useful 3D information from the teacher but suffers from the large cross-modal gap, achieving the improved gain of 1.2 mAP by BEVDistill and 4.2 mAP by our MapDistill. (4) Our MapDistill with the fusion-based teacher enables effective knowledge distillation within/between modalities while enjoying cost-effective camera-only deployment, achieving the most significant gain of 7.7 mAP and surpassing UniDistill by 5.4 mAP.

4.3 Ablation Study

Effect of $\mathcal{L}_{relation}$, $\mathcal{L}_{feature}$, and $\mathcal{L}_{head}$. We conduct an ablation study on the components in MapDistill and summarize our results in Tab. 2. We evaluate model variants using different combinations of the proposed distillation losses, including $\mathcal{L}_{relation}$, $\mathcal{L}_{feature}$, and $\mathcal{L}_{head}$.

We first investigate the effect of each distillation loss function. In model variants (a), (b), and (c), we use different distillation losses to distill the student model separately. The experimental results show that all model variants get improved performance compared to the baseline model, verifying the effectiveness of the proposed distillation losses. Moreover, the results of model variants (d), (e), and (f) prove that different distillation losses are complementary to each other. Finally, using all the proposed distillation losses together, we arrive at the full MapDistill method, which achieves the overall best performance of 53.6

Table 2: Ablation study on the components in MapDistill.

Setting	$\mathcal{L}_{relation}$	$\mathcal{L}_{feature}$	$\mathcal{L}_{head}$	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
Baseline	✗	✗	✗	39.6	49.9	48.2	45.9
a	✔	✗	✗	44.1	49.7	52.4	48.8
b	✗	✔	✗	44.3	49.4	51.5	48.4
c	✗	✗	✔	44.2	50.1	52.7	49.0
d	✔	✔	✗	45.4	51.4	54.1	50.3
e	✗	✔	✔	46.3	51.8	54.3	50.8
f	✔	✗	✔	46.5	52.3	54.5	51.1
g	✔	✔	✔	**49.2**	**54.5**	**57.1**	**53.6**

mAP, significantly surpassing the baseline's performance of 45.9 mAP. The ablation study results show that each of the distillation losses in MapDistll provides a meaningful contribution to improving the student model performance. Notably, these losses are only calculated during training, which brings no computational overhead during inference.

Ablations on the cross-modal relation distillation. We investigate the choice of relation distillation loss in our method. The ablation variants include training without relation distillation loss (MapDistill (w/o $\mathcal{L}_{relation}$)), uni-modal relation distillation (Uni-modal Rel.), cross-modal relation distillation (Cross-modal Rel.), and the hybrid relation distillation (hybrid Cross-modal and Uni-modal). Note that uni-modal relation distillation means replacing the cross-modal attention matrices $\mathbf{A}_{c2l}^{S/T}$ and $\mathbf{A}_{l2c}^{S/T}$ in Eq. 7 with the uni-modal ones $\mathbf{A}_{c2c}^{S/T}$ and $\mathbf{A}_{l2l}^{S/T}$. We explore which relation (cross-modal or uni-modal) is more critical. As shown in Tab. 3a, employing cross-modal relation distillation achieves more improvements. Furthermore, we find that using only cross-modal relation for distillation performs better than using both cross-modal and uni-modal relations. These observations validate that cross-modal interactions encode useful knowledge and can be transferred to the student model for improving HD Map construction.

Ablations on the dual-level feature distillation. To explore the impact of BEV feature distillation at different levels, we train the model by using low-level or high-level feature distillation solely and present the results in Tab. 3b. We design the following model variants: (1) MapDistill (w/o $\mathcal{L}_{feature}$): we remove the feature distillation loss from MapDistill. (2) Low-level (only) and High-level (only) mean that the MapDistill model is trained only using low-level BEV feature distillation or high-level BEV feature distillation, respectively. (3) Dual-level (ours): we use dual-level feature distillation (the default setting in our MapDistill) to train the model. The results of Low-level (only) and High-level (only) are inferior to the Dual-level (ours), verifying the effectiveness of distilling both low-level and high-level BEV features simultaneously.

Ablations on the map head distillation. In this ablation, we conduct detailed experiments on the loss selection for both map elements classification and point position regression. We design the following model variants: (1) MapDistill (w/o $\mathcal{L}_{head}$): we train the model without the map head distillation loss; (2)

Table 3: Ablation experiments to validate our distillation losses.

(a) Cross-modal relation distillation loss

Method	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
MapDistill (w/o $\mathcal{L}_{relation}$)	46.3	51.8	54.3	50.8
+Uni-modal Relation	48.0	52.9	55.1	52.0
+Hybrid Relation	48.3	53.4	55.5	52.4
+Cross-modal Relation	**49.2**	**54.5**	**57.1**	**53.6**

(b) Dual-level feature distillation loss

Method	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
MapDistill (w/o $\mathcal{L}_{feature}$)	46.5	52.3	54.5	51.1
+Low-level (only)	48.4	53.7	56.0	52.7
+High-level (only)	48.7	53.9	56.1	52.9
+Dual-level (ours)	**49.2**	**54.5**	**57.1**	**53.6**

(c) Map head distillation loss

$\mathcal{L}_{cls}$	$\mathcal{L}_{point}$	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
✗	✗	45.4	51.4	54.1	50.3
✔	✗	47.3	52.8	55.3	51.8
✗	✔	47.1	53.0	55.6	51.9
✔	✔	**49.2**	**54.5**	**57.1**	**53.6**

$\mathcal{L}_{head}$ (w/o $\mathcal{L}_{point}$): we remove the point2point loss from the map head distillation loss; (3) $\mathcal{L}_{head}$ (w/o $\mathcal{L}_{cls}$): we remove the classification loss from the map head distillation loss; (4) Using both $\mathcal{L}_{cls}$ and $\mathcal{L}_{point}$ (the default setting in our MapDistill). As shown in Tab. 3c, the results of $\mathcal{L}_{head}$ (w/o $\mathcal{L}_{cls}$) and $\mathcal{L}_{head}$ (w/o $\mathcal{L}_{point}$) are inferior to the default setting, verifying the effectiveness of transferring knowledge of both map elements categories and point positions from the teacher to the student.

Table 4: Ablation study of Dual BEV Transform Module.

	subspace1	subspace2	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
(a)	GKT	✗	44.9	49.6	52.8	49.1
(b)	LSS	LSS	45.9	51.2	54.4	50.5
	GKT	GKT	46.7	51.6	54.5	50.9
	Deform.	Deform.	46.8	51.6	54.6	51.0
(c)	GKT	Deform.	47.1	53.2	56.2	52.1
	Deform.	GKT	47.3	53.4	56.1	52.3
	LSS	Deform.	48.9	53.9	56.2	53.0
	Deform.	LSS	48.7	53.8	55.9	52.8
	LSS	GKT	49.1	54.2	56.7	53.3
	GKT	**LSS**	**49.2**	**54.5**	**57.1**	**53.6**

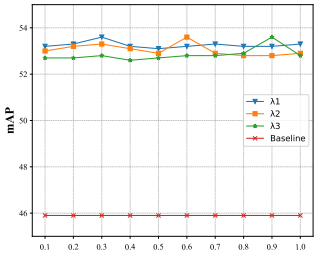

Fig. 3: Sensitivity of hyper-parameters.

Ablation study of the Dual BEV Transform Module. We further conduct ablation studies on the design choice of the Dual BEV Transform Module, *i.e.*, using different 2D-to-BEV methods to obtain subspace BEV features, to verify which combination performs most effectively. We choose three state-of-the-art 2D-to-BEV methods in this study, including LSS [33], Deformable Attention [21] and GKT [4]. The experiment consists of two groups, (1) both subspaces use the same 2D-to-BEV method, and (2) the two subspaces use different 2D-to-BEV methods. As shown in Tab. 4, the experimental results reveal some interesting findings: (1) Using the same 2D-to-BEV method only slightly outperforms the single-branch baseline in (a), implying that using the homogeneous BEV feature space makes it difficult to imitate the cross-modal interactions in the teacher model. (2) Using different 2D-to-BEV methods consistently outperforms the baseline and all model variants in (b). It is reasonable since the cross-modal relation of the teacher is calculated based on BEV features from different modal-

Table 5: Ablation experiments to verify the generalization ability of MapDistill.

(a) Different HD map construction methods

Method	Backbone	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
Teacher model-1	SwinT&Sec	57.5	63.3	70.9	63.9
Student model-1	R18	39.6	49.9	48.2	45.9
MapDistill	R18	**50.2**	**55.5**	**57.8**	**54.5**$_{+8.6}$
Teacher model-2	R50&Sec	65.6	66.5	74.8	69.0
Student model-2	R18	46.9	55.1	54.9	52.3
MapDistill	R18	**53.9**	**62.2**	**61.5**	**59.2**$_{+6.9}$

(b) Different student models

Method	Backbone	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
MapTR-Teacher	R50&Sec	55.9	62.3	69.3	62.5
Student model-I	R50	46.3	51.5	53.1	50.3
MapDistill	R50	**51.3**	**56.4**	**57.6**	**55.1**$_{+4.8}$
MapTR-Teacher	R50&Sec	55.9	62.3	69.3	62.5
Student model-II	Swin-T	45.2	52.7	52.3	50.1
MapDistill	Swin-T	**50.1**	**56.8**	**58.9**	**55.2**$_{+5.1}$

ities, using heterogeneous BEV feature spaces makes it possible to learn distinct BEV features and thus could imitate the cross-modal interactions. Specifically, the combination of LSS and GKT achieves the best results. These observations validate the motivation for devising dual BEV spaces using different 2D-to-BEV methods.

Ablation study of various HD map construction methods. To explore the compatibility of MapDistill with different HD map construction methods, we comprehensively investigate two popular methods and show the results in Tab. 5a. Specifically, Teacher model-1 and Teacher model-2 mean the MapTR variant model whose camera branch uses Swin-Tiny backbone to extract image features and the most advanced MapTRv2 (improving MapTR with both network design and training strategy techniques), respectively. Note that, both student models employ Resnet18 as the backbone to extract the multi-view features. The experimental results demonstrate that "Great teachers spawn exceptional students". As the proficient teacher model has acquired valuable knowledge for HD map construction, the student model can effectively leverage this knowledge through KD techniques (*e.g.*, the proposed MapDistill), enhancing its ability to perform the same task. Moreover, the results of consistent performance improvements show that our method is effective with different teacher models.

Ablation study of various student models. To explore the generalization capability of MapDistill with different student models, we comprehensively investigate two popular backbone networks as the backbone of the student model and show the results in Tab. 5b. Specifically, Student model-I and Student model-II mean that the student model employs Resnet50 and Swin-Tiny as the backbone to extract the multi-view features, respectively. And here we use MapTR-Teacher, which is the R50&Sec fusion model in Tab. 1, as the teacher model. Experimental results show that our method consistently achieves excellent results, proving the effectiveness and generalization ability of our method.

Sensitivity of hyper-parameters. We conduct experiments to investigate the impact of different hyper-parameter settings and report results on nuScenes val set, as shown in Fig. 3. When one hyper-parameter is varied within a feasible range, the remaining hyper-parameters retain their default values: $\lambda_1 = 0.3$, $\lambda_2 = 0.6$, and $\lambda_3 = 0.9$. The results indicate that the performance remains relatively stable across a wide range (0.1 to 0.9) of values for λ_1, λ_2, and λ_3, suggesting its robustness to different hyper-parameters. Note that, "Baseline" indicates the directly trained camera-based student model.

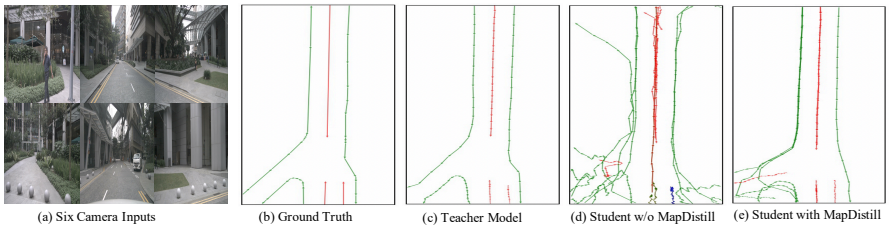

Fig. 4: Qualitative results on nuScenes val set. (a) Six camera inputs. (b) Ground-truth vectorized HD map. (c) Result of the camera-LiDAR-based teacher model. (d) Result of the camera-based student model without MapDistill (Baseline). (e) Result of the camera-based student model with MapDistill. MapDistill helps correct substantial errors in the Baseline's predictions and improves its accuracy.

4.4 Qualitative Results

We present visualizations of vectorized HD map predictions to demonstrate the efficacy of MapDistill. As depicted in Fig. 4, we compare predictions from various models, namely, the camera-LiDAR-based teacher model, the camera-based student model without MapDistill (referred to as "Baseline"), and the camera-based student model with MapDistill. The mAP values of these models are 62.5, 45.9, and 53.6 respectively, as shown in Tab. 1. Note that a common threshold, which is set to 0.4, is employed to filter low-confidence map elements for visualizing the prediction results of all models. We observe significant inaccuracies in the predictions made by the Baseline model. However, employing the MapDistill method substantially corrects these errors and enhances prediction accuracy.

5 Conclusion

In this paper, we present a novel method called MapDistill for boosting efficient camera-based HD map construction via camera-LiDAR fusion model distillation, yielding a cost-effective yet accurate solution. MapDistill is built upon a camera-LiDAR fusion teacher model, a lightweight camera-only student model, and a specifically designed Dual BEV Transform module. In addition, we present a comprehensive distillation scheme encompassing cross-modal relation distillation, dual-level feature distillation, and map head distillation, which facilitates knowledge transfer within and between different modalities and helps the student model achieve better performance. Extensive experiments and analysis validate the design choice and the effectiveness of our MapDistill.

Limitations and Societal Impact. With the KD methodology, the student model may inherit the weakness of the teacher model. More specifically, if the teacher model is biased, or not robust to adverse weather conditions and/or long-tail scenarios, the student model is likely to behave similarly. MapDistill enjoys cost-effective camera-only deployment, showing great potential in practical applications, such as autonomous driving.

Acknowledgement. This work was supported by the National Natural Science Foundation of China No.62106259 and the Beijing Natural Science Foundation under Grant L243008.

References

1. Borse, S., Klingner, M., Kumar, V.R., Cai, H., Almuzairee, A., Yogamani, S., Porikli, F.: X-align: Cross-modal cross-view alignment for bird's-eye-view segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3287–3297 (2023) 4
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11618–11628 (2020) 3, 9
3. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. Advances in neural information processing systems (2017) 2, 4
4. Chen, S., Cheng, T., Wang, X., Meng, W., Zhang, Q., Liu, W.: Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer. arXiv preprint arXiv:2206.04584 (2022) 4, 6, 9, 12
5. Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., Zhao, F.: Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. arXiv preprint arXiv:2211.09386 (2022) 2, 4, 5, 9, 10
6. Cho, H., Choi, J., Baek, G., Hwang, W.: itkd: Interchange transfer-based knowledge distillation for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13540–13549 (2023) 4
7. Ding, W., Qiao, L., Qiu, X., Zhang, C.: Pivotnet: Vectorized pivot learning for end-to-end hd map construction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3672–3682 (2023) 2, 4, 10
8. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision pp. 1789–1819 (2021) 2
9. Hao, X., Wei, M., Yang, Y., Zhao, H., Zhang, H., Zhou, Y., Wang, Q., Li, W., Kong, L., Zhang, J.: Is your hd map constructor reliable under sensor corruptions? arXiv preprint arXiv:2406.12214 (2024) 4
10. Hao, X., Yang, Y., Zhang, H., Wei, M., Zhou, Y., Zhao, H., Zhang, J.: Team samsung-ral: Technical report for 2024 robodrive challenge-robust map segmentation track. arXiv preprint arXiv:2405.10567 (2024) 4
11. Hao, X., Zhang, H., Yang, Y., Zhou, Y., Jung, S., Park, S.I., Yoo, B.: Mbfusion: A new multi-modal bev feature fusion method for hd map construction. In: IEEE International Conference on Robotics and Automation (2024) 4
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016) 4, 6, 9, 10
13. Hendy, N., Sloan, C., Tian, F., Duan, P., Charchut, N., Xie, Y., Wang, C., Philbin, J.: FISHING net: Future inference of semantic heatmaps in grids. arXiv preprint arXiv:2006.09917 (2020) 2, 4
14. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 2, 4
15. Huang, P., Liu, L., Zhang, R., Zhang, S., Xu, X., Wang, B., Liu, G.: Tig-bev: Multi-view bev 3d object detection via target inner-geometry learning. arXiv preprint arXiv:2212.13979 (2022) 2, 5
16. Kong, L., Xie, S., Hu, H., Niu, Y., Ooi, W.T., Cottereau, B.R., Ng, L.X., Ma, Y., Zhang, W., Pan, L., et al.: The robodrive challenge: Drive anytime anywhere in any condition. arXiv preprint arXiv:2405.08816 (2024) 4

17. Li, D., Jin, Y., Yu, H., Shi, J., Hao, X., Hao, P., Liu, H., Sun, F., Fang, B., et al.: What foundation models can bring for robot learning in manipulation: A survey. arXiv preprint arXiv:2404.18201 (2024) 4
18. Li, J., Lu, M., Liu, J., Guo, Y., Du, L., Zhang, S.: Bev-lgkd: A unified lidar-guided knowledge distillation framework for bev 3d object detection. arXiv preprint arXiv:2212.00623 (2022) 2, 5, 9, 10
19. Li, Q., Wang, Y., Wang, Y., Zhao, H.: Hdmapnet: An online hd map construction and evaluation framework. In: IEEE International Conference on Robotics and Automation. pp. 4628–4634 (2022) 2, 4, 9, 10
20. Li, Y., Chen, Y., Qi, X., Li, Z., Sun, J., Jia, J.: Unifying voxel-based representation with transformer for 3d object detection. Advances in Neural Information Processing Systems pp. 18442–18455 (2022) 2, 5
21. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European Conference on Computer Vision. pp. 1–18 (2022) 4, 12
22. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: Bevfusion: A simple and robust lidar-camera fusion framework. Advances in Neural Information Processing Systems pp. 10421–10434 (2022) 4
23. Liao, B., Chen, S., Wang, X., Cheng, T., Zhang, Q., Liu, W., Huang, C.: Maptr: Structured modeling and learning for online vectorized hd map construction. In: International Conference on Learning Representations (2023) 2, 3, 4, 5, 8, 9, 10
24. Liao, B., Chen, S., Zhang, Y., Jiang, B., Zhang, Q., Liu, W., Huang, C., Wang, X.: Maptrv2: An end-to-end framework for online vectorized HD map construction. arXiv preprint arXiv:2308.05736 (2023) 4
25. Liu, Y., Yuan, T., Wang, Y., Wang, Y., Zhao, H.: Vectormapnet: End-to-end vectorized hd map learning. In: International Conference on Machine Learning. pp. 22352–22369 (2023) 2, 4, 9, 10
26. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin transformer v2: Scaling up capacity and resolution. In: International Conference on Computer Vision and Pattern Recognition. pp. 11999–12009 (2022) 4
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9992–10002 (2021) 4
28. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: IEEE International Conference on Robotics and Automation. pp. 2774–2781 (2023) 4, 6
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019) 9
30. Malkauthekar, M.: Analysis of euclidean distance and manhattan distance measure in face recognition. In: Third International Conference on Computational Intelligence and Information Technology (CIIT 2013). pp. 503–507. IET (2013) 8
31. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: Proceedings of the AAAI conference on artificial intelligence. pp. 5191–5198 (2020) 2, 4
32. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. Advances in Neural Information Processing Systems pp. 15288–15299 (2020) 8

33. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: European Conference on Computer Vision. pp. 194–210 (2020) 4, 6, 12
34. Qiao, L., Ding, W., Qiu, X., Zhang, C.: End-to-end vectorized hd-map construction with piecewise bezier curve. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13218–13228 (2023) 2, 4, 10
35. Salazar-Gomez, G., González, D.S., Diaz-Zapata, M., Paigwar, A., Liu, W., Erkent, Ö., Laugier, C.: Transfusegrid: Transformer-based lidar-rgb fusion for semantic grid prediction. In: International Conference on Control, Automation, Robotics and Vision. pp. 268–273 (2022) 4
36. Shang, C., Li, H., Meng, F., Wu, Q., Qiu, H., Wang, L.: Incrementer: Transformer for class-incremental semantic segmentation with knowledge distillation focusing on old class. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7214–7224 (2023) 4
37. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114 (2019) 4, 10
38. Tang, K., Cao, X., Cao, Z., Zhou, T., Li, E., Liu, A., Zou, S., Liu, C., Mei, S., Sizikova, E., et al.: Thma: Tencent hd map ai system for creating hd map annotations. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 15585–15593 (2023) 4
39. Wang, S., Li, W., Liu, W., Liu, X., Zhu, J.: Lidar2map: In defense of lidar-based semantic map construction using online camera distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5186–5195 (2023) 2, 4
40. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. arXiv preprint arXiv:2211.05778 (2022) 4
41. Wang, Z., Li, D., Luo, C., Xie, C., Yang, X.: Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8637–8646 (2023) 2, 5
42. Xiong, X., Liu, Y., Yuan, T., Wang, Y., Wang, Y., Zhao, H.: Neural map prior for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17535–17544 (2023) 4
43. Yan, X., Gao, J., Zheng, C., Zheng, C., Zhang, R., Cui, S., Li, Z.: 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In: European Conference on Computer Vision. pp. 677–695. Springer (2022) 2, 4
44. Yan, Y., Mao, Y., Li, B.: SECOND: sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018) 6, 9, 10
45. Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao, Y., Lu, L., et al.: Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. arXiv preprint arXiv:2211.10439 (2022) 4
46. Yang, Z., Li, R., Ling, E., Zhang, C., Wang, Y., Huang, D., Ma, K.T., Hur, M., Lin, G.: Label-guided knowledge distillation for continual semantic segmentation on 2d images and 3d point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18601–18612 (2023) 4
47. Zhang, G., Lin, J., Wu, S., Song, Y., Luo, Z., Xue, Y., Lu, S., Wang, Z.: Online map vectorization for autonomous driving: A rasterization perspective. arXiv preprint arXiv:2306.10502 (2023) 4, 10

48. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18802–18812 (2022) 4
49. Zhang, Q., Cheng, X., Chen, Y., Rao, Z.: Quantifying the knowledge in a dnn to explain knowledge distillation for classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 5099–5113 (2022) 4
50. Zhang, Y., Zhu, Z., Zheng, W., Huang, J., Huang, G., Zhou, J., Lu, J.: Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. arXiv preprint arXiv:2205.09743 (2022) 4
51. Zhao, H., Zhang, Q., Zhao, S., Chen, Z., Zhang, J., Tao, D.: Simdistill: Simulated multi-modal distillation for bev 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7460–7468 (2024) 2, 4
52. Zheng, Z., Ye, R., Hou, Q., Ren, D., Wang, P., Zuo, W., Cheng, M.M.: Localization distillation for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 4
53. Zhou, S., Liu, W., Hu, C., Zhou, S., Ma, C.: Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5116–5125 (2023) 2, 4, 5, 9, 10