LiDAR2Map: In Defense of LiDAR-Based Semantic Map Construction Using Online Camera Distillation

Song Wang, Wentong Li, Wenyu Liu, Xiaolu Liu, Jianke Zhu*
Zhejiang University
{songw, liwentong, liuwenyu.lwy, xiaoluliu, jkzhu}@zju.edu.cn

Abstract

Semantic map construction under bird's-eye view (BEV) plays an essential role in autonomous driving. In contrast to camera image, LiDAR provides the accurate 3D observations to project the captured 3D features onto BEV space inherently. However, the vanilla LiDAR-based BEV feature often contains many indefinite noises, where the spatial features have little texture and semantic cues. In this paper, we propose an effective LiDAR-based method to build semantic map. Specifically, we introduce a BEV pyramid feature decoder that learns the robust multi-scale BEV features for semantic map construction, which greatly boosts the accuracy of the LiDAR-based method. To mitigate the defects caused by lacking semantic cues in LiDAR data, we present an online Camera-to-LiDAR distillation scheme to facilitate the semantic learning from image to point cloud. Our distillation scheme consists of feature-level and logit-level distillation to absorb the semantic information from camera in BEV. The experimental results on challenging nuScenes dataset demonstrate the efficacy of our proposed LiDAR2Map on semantic map construction, which significantly outperforms the previous LiDAR-based methods over 27.9% mIoU and even performs better than the state-of-the-art camera-based approaches. Source code is available at: https://github.com/songw-zju/LiDAR2Map.

1. Introduction

High-definition (HD) map contains the enriched semantic understanding of elements on road, which is a fundamental module for navigation and path planning in autonomous driving. Recently, online semantic map construction has attracted increasing attention, which enables to construct HD map at runtime with onboard LiDAR and cameras. It provides a compact way to model the environment around the ego vehicle, which is convenient to obtain the essential information for the downstream tasks.

Most of recent online approaches treat semantic map learning as a segmentation problem in bird's-eye view (BEV), which assign each map pixel with a category label. As shown in Fig. 1, the existing methods can be roughly divided into three groups, including camera-based methods [19, 20, 30, 32, 52], LiDAR-based methods [10, 19] and Camera-LiDAR fusion methods [19, 27, 37]. Among them, camera-based methods are able to make full use of multi-view images with the enriched semantic information, which dominate this task with the promising performance. In contrast to camera image, LiDAR outputs the accurate 3D spatial information that can be used to project the captured features onto the BEV space. By taking advantage of the geometric and spatial information, LiDAR-based methods are widely explored in 3D object detection [18, 38, 47, 57] while it is rarely investigated in semantic map construction.

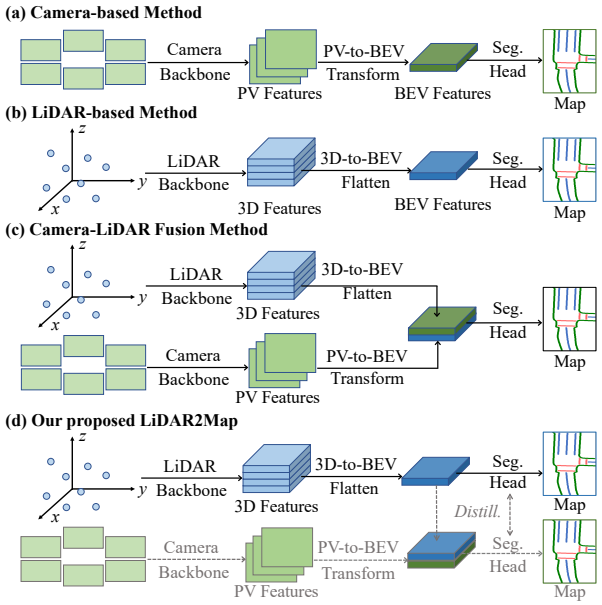

Figure 1. Comparisons on semantic map construction frameworks (camera-based, LiDAR-based, Camera-LiDAR fusion methods) and our proposed LiDAR2Map that presents an effective online Camera-to-LiDAR distillation scheme with a BEV feature pyramid decoder in training.

*Corresponding author is Jianke Zhu.

HDMapNet-LiDAR [19] intends to directly utilize the LiDAR data for map segmentation, however, it performs inferior to the camera-based models due to the vanilla BEV feature with the indefinite noises. Besides, map segmentation is a semantic-oriented task [27] while the semantic cues in LiDAR are not as rich as those in image. In this work, we aim to exploit the LiDAR-based semantic map construction by taking advantage of the global spatial information and auxiliary semantic density from the image features.

In this paper, we introduce an efficient framework for semantic map construction, named LiDAR2Map, which fully exhibits the potentials of LiDAR-based model. Firstly, we present an effective decoder to learn the robust multi-scale BEV feature representations from the accurate spatial point cloud information for semantic map. It provides distinct responses and boosts the accuracy of our baseline model. To make full use of the abundant semantic cues from camera, we then suggest a novel online Camera-to-LiDAR distillation scheme to further promote the LiDAR-based model. It fully utilizes the semantic features from the image-based network with a position-guided feature fusion module (PGF^{2}M). Both the feature-level and logit-level distillation are performed in the unified BEV space to facilitate the LiDAR-based network to absorb the semantic representation during the training. Specially, we suggest to generate the global affinity map with the input low-level and high-level feature guidance for the satisfactory feature-level distillation. The inference process of LiDAR2Map is efficient and direct without the computational cost of distillation scheme and auxiliary camera-based branch. Extensive experiments on the challenging nuScenes benchmark [4] show that our proposed model significantly outperforms the conventional LiDAR-based method (29.5% mIoU *vs.* 57.4% mIoU). It even performs better than the state-of-the-art camera-based methods by a large margin.

Our main contributions are summarized as: 1) an efficient framework LiDAR2Map for semantic map construction, where the presented BEV pyramid feature decoder can learn the robust BEV feature representations to boost the baseline of our LiDAR-based model; 2) an effective online Camera-to-LiDAR distillation scheme that performs both feature-level and logit-level distillation during the training to fully absorb the semantic representations from the images; 3) extensive experiments on nuScenes for semantic map construction including map and vehicle segmentation under different settings, shows the promising performance of our proposed LiDAR2Map.

2. Related Work

Semantic Map Construction. High-definition (HD) maps have the rich information on road layout, which are essential to autonomous vehicles [1, 23, 45]. Traditional offline approaches to HD map construction require lots of manual annotations and regular updates [3, 16, 43, 49, 54], which incur the expensive costs on labeling. Recently, the learning-based methods [19, 24, 56] have been proposed to construct semantic map online with camera image and LiDAR point cloud using an end-to-end network, which can be roughly divided into three groups, including camera-based methods, LiDAR-based approaches and Camera-LiDAR fusion methods. Camera-based methods [30, 32, 52] learn to project the perspective view (PV) features onto BEV space through the geometric prior, which often have the spatial distortions inevitably. Besides, the camera-based methods rely on high-resolution images and large pre-trained models for better accuracy [20, 52], which brings serious challenges to the practical scenarios. LiDAR-based approaches [10, 19] directly capture the accurate spatial information for the unified BEV feature representation. However, they cannot robustly deal with large noises in the vanilla BEV feature. Camera-LiDAR fusion methods [19, 27, 37] make use of both the semantic features from camera and geometric information from LiDAR. They achieve better results than those approaches with single modality under the same setting while having the larger computational burden. In this paper, we intend to construct the semantic map from LiDAR point cloud effectively.

Multi-sensor Fusion. Multi-sensor fusion is always a key issue in autonomous driving, among which camera and LiDAR fusion research is the most in-depth. Previous methods obtain the promising performance on 3D detection and segmentation through a point-to-pixel fusion strategy [42, 48, 58]. However, such pipeline requires the correspondences between points and pixels, which cannot fully utilize the information of whole image and all the point cloud. Recently, multi-modal feature fusion in the unified BEV space has attracted some attention [21, 27]. Converting the semantic features from camera into a BEV representation can be better integrated with spatial features from LiDAR [29, 31]. This provides the enriched information for downstream tasks like planning and decision-making. However, the fusion of multi-sensor may increase the computational burden on the deployment. In this work, we exploit an effective online Camera-to-LiDAR distillation scheme to fully absorb the semantic features for LiDAR-based branch.

Cross-modal Knowledge Distillation. Knowledge distillation is originally proposed for model compression [13], where knowledge can be transferred from a pre-trained model to an untrained small model. In addition to logit-level distillation [5, 7, 53], feature-level distillation has received more attention [11, 12, 34, 46]. Cross-modal knowledge distillation has been validated in many tasks such as LiDAR semantic segmentation [15, 44], monocular 3D object detection [6], 3D hand pose estimation [50] and 3D dense captioning [51]. In this work, we introduce both feature-level and logit-level distillation on BEV representation.

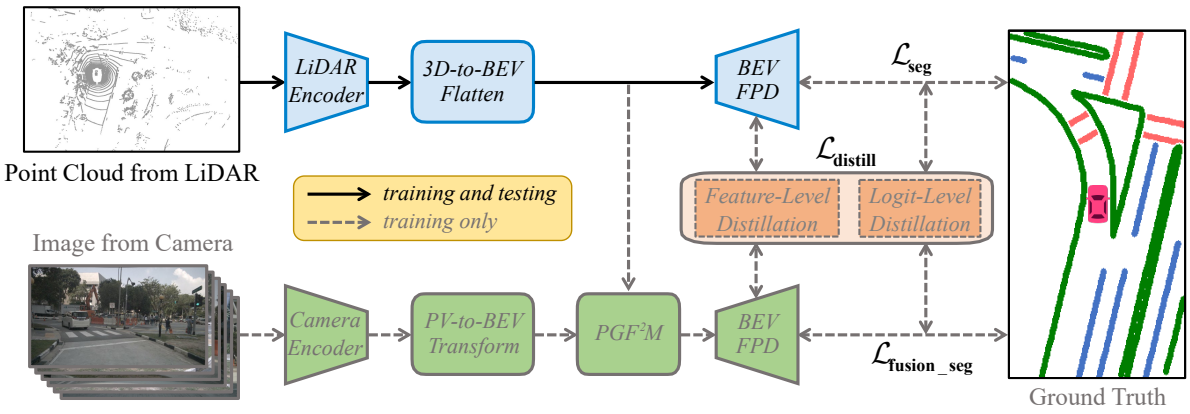

Figure 2. **Overview of LiDAR2Map Framework.** LiDAR2Map employs the LiDAR-based network as the main branch to encode the point cloud feature with a robust BEV feature pyramid decoder (BEV-FPD) for semantic map construction. During the training, the camera-based branch is adopted to extract the semantic image features. Both feature-level and logit-level distillation are performed to allow LiDAR-branch to benefit from the providing image features without the overhead during inference.

3. LiDAR2Map

3.1. Overview

In this work, we aim to explore the potentials of an efficient LiDAR-based model for semantic map construction. Different from the previous LiDAR-based methods [10, 19], we introduce an effective BEV feature pyramid decoder to learn the robust representations from the spatial information of point cloud. To enhance the semantic information of single LiDAR modality, we take into account of the images through an online distillation scheme on the BEV space that employs the multi-level distillation during the training. In the inference stage, we only preserve the LiDAR branch for efficient semantic map prediction. Fig. 2 shows the overview of our proposed LiDAR2Map framework.

3.2. Map-Oriented Perception Framework

Multi-Modal Feature Extractors. LiDAR sensor typically outputs a set of unordered points, which cannot be directly processed by 2D convolution. We investigate the most commonly used backbones in 3D object detection, including PointPillars [18] and VoxelNet [57], which can extract the effective 3D features $\mathbf{F}^{\text{3D}}_{\text{LiDAR}}$ from LiDAR point cloud. Specifically, PointPillars converts the raw point cloud into multiple pillars, and then extracts features from pillar-wise point cloud by 2D convolution. VoxelNet directly voxelizes the point cloud first and uses the sparse convolution to build 3D network to encoder the better 3D feature representation. Then, the unified BEV representation $\mathbf{F}^{\text{BEV}}_{\text{LiDAR}}$ is obtained by pooling the 3D features $\mathbf{F}^{\text{3D}}_{\text{LiDAR}}$.

Besides, we build another network branch to encode the pixel-level semantic features in perspective view from the images, which is used in our presented online distillation scheme (see Sec. 3.3). As in [31], we adopt a similar 2D-3D transformation manner. Firstly, we extract the perspective features $\mathbf{F}^{\text{PV}}_{\text{Camera}}$ from each input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ by 2D convolution and predict the depth distribution of D equally spaced discrete points associated with each pixel. Secondly, we assign the perspective features $\mathbf{F}^{\text{PV}}_{\text{Camera}}$ to D points along the camera ray direction to obtain a $D\times H\times W$ pseudo point cloud features $\mathbf{F}^{\text{3D}}_{\text{Camera}}$. Finally, the pseudo point cloud features are flatten to the BEV space $\mathbf{F}^{\text{BEV}}_{\text{Camera}}$ through the pooling as the LiDAR branch.

BEV Feature Pyramid Decoder. BEV features are regarded as the unified representation in our framework, which can absorb both the geometric structure from LiDAR and semantic features from the images. Based on BEV features, the current LiDAR-based method in [19] employs a fully connected layer as the segmentation head to obtain segmentation results directly. Since the vanilla BEV feature from LiDAR backbone contains the ambiguous noise response, it obtains the inferior performance compared with camera-based models [30, 52].

In this work, we develop a BEV feature pyramid decoder (BEV-FPD) to capture the multi-scale BEV features with less noises from LiDAR data for better semantic map construction. Fig. 3 shows the architecture of the BEV-FPD. Based on the BEV features $\mathbf{F}^{\text{BEV}}$ from the LiDAR or camera branch, we firstly perform 7×7 convolution on the BEV features to generate the global features with the large receptive field. The multi-scale BEV features $\{\tilde{\mathbf{F}}^{\text{BEV}}_i\}^N_{i=1}$ are obtained by the six successive layers, and each layer consists of two standard residual block [9] to better transmit the feature representation. The N-scale features $\{\tilde{\mathbf{F}}^{\text{BEV}}_i\}^N_{i=1}$ represent the different level of semantic features in the BEV space. As the feature size decreases, the number of channels increases. The bilinear interpolation is used to up-sample the each low-resolution semantic maps and obtain the feature representations with the same resolution. We then concatenate the feature maps at all scales with the same reso-

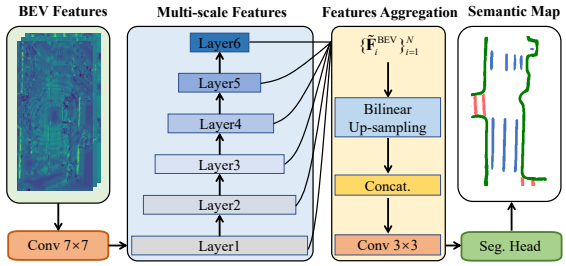

Figure 3. Illustration of **BEV Feature Pyramid Decoder (BEV-FPD)**. BEV-FPD collects the multi-scale BEV features from six layers to perform the feature aggregation for semantic map with a segmentation head.

lution to perform the multi-scale feature aggregation. The final semantic map is obtained by a segmentation head with the *softmax* function to account for the probability distribution of each category. As the layer number increases, the corresponding BEV features $\tilde{\mathbf{F}}_i^{\text{BEV}}$ can better capture the robust spatial features with accurate responses. It plays an essential role in improving our proposed LiDAR-based model (see Sec. 4.3).

3.3. Online Camera-to-LiDAR Distillation

To enhance the semantic representation for our LiDAR-based model, we introduce an effective online Camera-to-LiDAR distillation scheme in BEV space, which enables the LiDAR-based branch to learn the semantic cues from the images. It consists of three components, including Position-Guided Feature Fusion Module (PGF^{2}M), Feature-level Distillation (FD) and Logit-level Distillation (LD).

Position-Guided Feature Fusion Module. PGF^{2}M is introduced to better fuse the features from camera and LiDAR in BEV space, as shown in Fig. 4. Firstly, we concatenate the BEV features along the channel dimension between two modalities, *i.e.* LiDAR point cloud feature $\mathbf{F}_{\text{LiDAR}}^{\text{BEV}}$ and camera image feature $\mathbf{F}_{\text{Camera}}^{\text{BEV}}$. Then, we perform the preliminary fusion through a 3×3 convolutional layer to obtain $\mathbf{F}_{\text{Fusion\_s1}}^{\text{BEV}}$ as below,

$$\mathbf{F}_{\text{Fusion\_s1}}^{\text{BEV}} = \text{Conv } 3 \times 3 \left(\left[\mathbf{F}_{\text{Camera}}^{\text{BEV}}, \mathbf{F}_{\text{LiDAR}}^{\text{BEV}} \right] \right). \quad (1)$$

Secondly, we calculate the relative coordinates of x-axis and y-axis $\mathbf{F}_{\text{Pos}}^{\text{BEV}}$ with the same size. Then, we concatenate it with the fusion result $\mathbf{F}_{\text{Fusion\_s1}}^{\text{BEV}}$ at the previous stage along the channel dimension to encode the spatial information, and perform 3×3 convolution:

$$\mathbf{F}_{\text{Fusion\_s2}}^{\text{BEV}} = \text{Conv } 3 \times 3 \left(\left[\mathbf{F}_{\text{Fusion\_s1}}^{\text{BEV}}, \mathbf{F}_{\text{Pos}}^{\text{BEV}} \right] \right). \quad (2)$$

$\mathbf{F}_{\text{Fusion\_s2}}^{\text{BEV}}$ is further fed into an attention layer that is composed of a 2D adaptive average pooling, two-layer MLP and a *sigmoid* function to build the global pixel affinity. Thus, its result $\mathbf{F}_{\text{Fusion\_s3}}^{\text{BEV}}$ is obtained by

$$\mathbf{F}_{\text{Fusion\_s3}}^{\text{BEV}} = \sigma \left(\text{MLP} \left(\text{Avg}(\mathbf{F}_{\text{Fusion\_s2}}^{\text{BEV}}) \right) \right) \odot \mathbf{F}_{\text{Fusion\_s2}}^{\text{BEV}}. \quad (3)$$

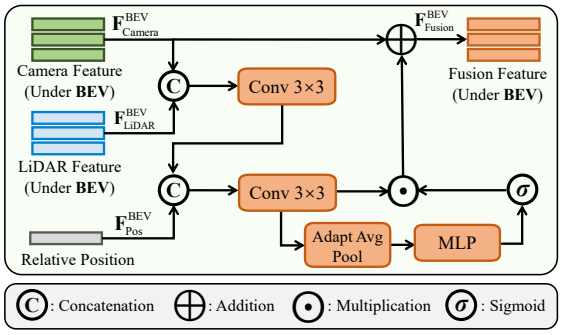

Figure 4. Illustration of **Position-Guided Feature Fusion Module (PGF^{2}M)**. In PGF^{2}M, the camera image features and LiDAR features in BEV space are in integration with the relative position information.

Finally, we add $\mathbf{F}_{\text{Fusion\_s3}}^{\text{BEV}}$ with the original BEV feature from camera $\mathbf{F}_{\text{Camera}}^{\text{BEV}}$ to obtain the fusion features $\mathbf{F}_{\text{Fusion}}^{\text{BEV}}$:

$$\mathbf{F}_{\text{Fusion}}^{\text{BEV}} = \mathbf{F}_{\text{Camera}}^{\text{BEV}} + \mathbf{F}_{\text{Fusion\_s3}}^{\text{BEV}}. \quad (4)$$

Feature-level Distillation. To facilitate the LiDAR branch to absorb the rich semantic features from the images, we take advantage of the multi-scale BEV features $\{\tilde{\mathbf{F}}_i^{\text{BEV}}\}_{i=1}^{N}$ from BEV-FPD for the feature-level distillation. Generally, it is challenging to directly distill high-dimensional features between camera and LiDAR modalities, which lack the global affinity of BEV representation. The straightforward feature distillation on these dense feature often fails to achieve the desired results. To address this issue, we employ the tree filter [22, 39] as the transform function $\mathcal{F}$ to model the long-range dependencies of dense BEV features in each modality by minimal spanning tree. Specifically, the shallow pillar/voxel features $\mathbf{F}_{\text{low}}^{\text{BEV}}$ from LiDAR backbone and multi-scale BEV features $\{\tilde{\mathbf{F}}_i^{\text{BEV}}\}_{i=1}^{N}$ are treated as the low-level and high-level input guidance of tree filter. With these low-level and high-level guidance, the feature transform is performed by tree filter in the cascade manner to obtain the global affinity map $\mathbf{M}_i^{\text{BEV}}$ for the corresponding i-th scale BEV features $\tilde{\mathbf{F}}_i^{\text{BEV}}$ as following,

$$\mathbf{M}_i^{\text{BEV}} = \mathcal{F} \left(\mathcal{F} \left(\tilde{\mathbf{F}}_i^{\text{BEV}}, \mathbf{F}_{\text{low}}^{\text{BEV}} \right), \tilde{\mathbf{F}}_i^{\text{BEV}} \right). \quad (5)$$

We compute the affinity similarity between each $\mathbf{M}_{\text{LiDAR},i}^{\text{BEV}}$ from the LiDAR branch and $\mathbf{M}_{\text{Fusion},i}^{\text{BEV}}$ of the Camera-LiDAR fusion branch to achieve the feature-level distillation. More specifically, a simple L_1 distance is used to accumulate them at all the scales as below,

$$\mathcal{L}_{\textbf{feature}} = \sum_{i=1}^{N} \left\| \mathbf{M}_{\text{Fusion},i}^{\text{BEV}} - \mathbf{M}_{\text{LiDAR},i}^{\text{BEV}} \right\|_1. \quad (6)$$

We employ $\mathcal{L}_{\textbf{feature}}$ as one of the loss terms to enable the LiDAR-based branch to benefit from the image feature implicitly through the network optimization.

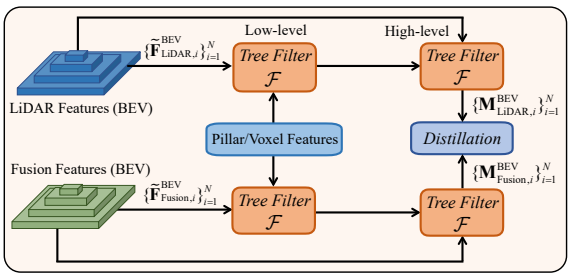

Figure 5. Illustration of **Feature-level Distillation**. Multi-scale BEV features are fed into two successive tree filters to generates the affinity map with the low-level and high-level guidance. The feature-level distillation is performed on generated affinity maps between LiDAR and fusion branch.

Logit-level Distillation. The semantic map predictions of segmentation head represent the probability distribution of each modality. We further suggest the logit-level distillation to make the LiDAR-based "Student" prediction learn from the soft labels generated by Camera-LiDAR fusion model as a "Teacher".

Through BEV-FPD with the segmentation head, the corresponding semantic map predictions $\mathbf{P}^{\text{BEV}}_{\text{LiDAR}}$ and $\mathbf{P}^{\text{BEV}}_{\text{Fusion}}$ can be obtained. As in [15], we adopt KL divergence to measure the the similarity on the probability distribution, which makes the $\mathbf{P}^{\text{BEV}}_{\text{LiDAR}}$ of LiDAR closer to $\mathbf{P}^{\text{BEV}}_{\text{Fusion}}$ of fusion "Teacher" as below,

$$\mathcal{L}_{\mathbf{logit}} = \boldsymbol{D}_{\text{KL}}\left(\mathbf{P}^{\text{BEV}}_{\text{Fusion}} \| \mathbf{P}^{\text{BEV}}_{\text{LiDAR}}\right). \tag{7}$$

3.4. Training and Inference

Overall Loss Function for Training. In this work, we treat the semantic map construction task as a pixel-level classification problem with segmentation loss in network optimization. Overall, the total training loss of our proposed framework consists of three terms:

$$\mathcal{L} = \mathcal{L}_{\mathbf{seg}} + \mathcal{L}_{\mathbf{fusion\_seg}} + \mathcal{L}_{\mathbf{distill}}, \tag{8}$$

where $\mathcal{L}_{\mathbf{seg}}$ and $\mathcal{L}_{\mathbf{fusion\_seg}}$ are the segmentation losses of LiDAR-branch and Camera-LiDAR fusion branch, respectively. $\mathcal{L}_{\mathbf{distill}}$ consists of $\mathcal{L}_{\mathbf{feature}}$ and $\mathcal{L}_{\mathbf{logit}}$ for online Camera-to-LiDAR distillation.

The segmentation loss for semantic map construction is composed of two items including $\mathcal{L}_{\mathbf{ce}}$ and $\mathcal{L}_{\mathbf{ls}}$ as following,

$$\mathcal{L}_{\mathbf{seg}} = \mathcal{L}_{\mathbf{ce}} + \mathcal{L}_{\mathbf{ls}}, \tag{9}$$

where $\mathcal{L}_{\mathbf{ce}}$ is the cross-entropy loss. $\mathcal{L}_{\mathbf{ls}}$ is employed to maximize the Intersection-over-Union (IoU) score as below,

$$\mathcal{L}_{\mathbf{ls}} = \frac{1}{|C|}\sum_{c\in C}\overline{\Delta_{J_c}}(\mathbf{m}(c)), \tag{10}$$

where $|C|$ is the total number of classes. $\mathbf{m}(c)$ denotes the vector of pixel errors on class $c \in C$. $\overline{\Delta_{J_c}}$ is the Lovász extension [2] for $\mathbf{m}(c)$ as the surrogate loss. The calculation of $\mathcal{L}_{\mathbf{fusion\_seg}}$ is the same as $\mathcal{L}_{\mathbf{seg}}$.

Inference. The LiDAR-based branch is fully optimized during training, which not only captures the spatial geometric features but also absorbs the enriched semantic information from the camera images. It is worthy of noting that we only preserve the LiDAR-branch for the predictions. The inference process is direct and efficient without incurring the computational cost on distillation and the camera-based branch.

4. Experiments

4.1. Implementation Details

Dataset. To evaluate the efficacy on semantic map construction, we conduct comprehensive experiments on nuScenes benchmark [4] that is a general and authoritative dataset. It contains 1,000 driving scenes collected in Boston and Singapore. The vehicle used for data collection is equipped with a 32-beam LiDAR, five long range RADARs and six cameras. There are 700 and 150 complete scenes for training and validation, respectively.

Evaluation. In this paper, we evaluate the performance on map and vehicle segmentation under different evaluation settings. For map segmentation, we adopt the same setting as HDMapNet [19], which uses a 60m $\times$ 30m area around the ego vehicle and samples a map at a 15cm resolution with three classes, including Divider (Div.), Ped Crossing (P. C.) and Boundary (Bound.). For vehicle segmentation, we utilize two commonly used settings proposed in PON [33] and Lift-Splat [31]. Setting 1 for vehicle segmentation employs a 100m $\times$ 50m map around the ego vehicle and samples at a 25cm resolution. Setting 2 adopts a 100m $\times$ 100m map at 25cm resolution. The mean Intersection-over-Union (mIoU) is used for the performance evaluation.

Training. For camera-branch, we choose Swin-Tiny [26] pre-trained on ImageNet [35] as the image backbone. For LiDAR-branch, PointPillars [18] and VoxelNet [57] are used to extract the point cloud feature. We train the whole network with 30 epochs using Adam optimizer [17] having a weight decay of $1e^{-7}$ on 4 NVIDIA Tesla V100 GPUs. The learning rate is $2e^{-3}$ for PointPillars and $1e^{-4}$ for VoxelNet, which decreases with a factor of 10 at the 20th epoch. The image size is set to 352 $\times$ 128 for PointPillars and 704 $\times$ 256 for VoxelNet during training. More training details under different settings are given in our supplementary material.

4.2. Main Results

Map Segmentation. For quantitative evaluation, we compare our method with the state-of-the-art camera-based

Table 1. Performance comparison on the validation set of nuScenes with the 60m × 30m setting for map segmentation. "∗" means the results reported from HDMapNet [19]. "†" denotes the results reported from UniFusion [32].

Method	Image Size	Modality	Backbone	Divider	Ped Crossing	Boundary	mIoU
VPN* [29]	352×128	Camera	EfficientNet-B0 [40]	36.5	15.8	35.6	29.3
Lift-Splat* [31]	352×128	Camera	EfficientNet-B0	38.3	14.9	39.3	30.8
HDMapNet-Camera [19]	352×128	Camera	EfficientNet-B0	40.6	18.7	39.5	32.9
BEVSegFormer [30]	800×448	Camera	ResNet-101	51.1	32.6	50.0	44.6
BEVFormer† [20]	1600×900	Camera	ResNet-50	53.0	36.6	54.1	47.9
BEVerse [52]	1408×512	Camera	Swin-Tiny	56.1	44.9	58.7	53.2
UniFusion [32]	1600×900	Camera	Swin-Tiny	58.6	43.3	59.0	53.6
HDMapNet-Fusion [19]	352×128	Camera & LiDAR	EfficientNet-B0 & PointPillars	46.1	31.4	56.0	44.5
HDMapNet-LiDAR [19]	-	LiDAR	PointPillars	26.7	17.3	44.6	29.5
LiDAR2Map	-	LiDAR	PointPillars	60.4	45.5	66.4	57.4
LiDAR2Map	-	LiDAR	VoxelNet	**61.5**	**46.3**	**68.1**	**58.6**

Table 2. Performance comparison on the validation set of nuScenes with two commonly used settings for vehicle segmentation without masking invisible vehicles. Setting 1 is with the 100m × 50m at 25cm resolution. Setting 2 is with the 100m × 100m at 50cm resolution.

Method	Image Size	Modality	Backbone	Setting 1	Setting 2	#Params(M)	FPS
VED [28]	800×600	Camera	ResNet-50	8.8	-	-	-
PON [33]	800×600	Camera	ResNet-50	24.7	-	38	30
VPN [29]	800×600	Camera	ResNet-50	25.5	-	18	-
STA [36]	1280×720	Camera	ResNet-50	36.0	-	-	-
Lift-Splat [31]	352×128	Camera	EfficientNet-B0	-	32.1	14	25
FIERY Static [14]	448×224	Camera	EfficientNet-B4	37.7	35.8	7.4	8
PolarBEV [25]	960×448	Camera	EfficientNet-B4	45.4	41.2	7.4	10
SimpleBEV [8]	800×448	Camera	ResNet-101	-	47.4	37	7.3
TransFuseGrid [37]	352×128	Camera & LiDAR	EfficientNet-B0 & PointPillars	-	35.9	-	18.4
Pillar feature Net [37]	-	LiDAR	PointPillars	-	23.4	-	-
LiDAR2Map	-	LiDAR	PointPillars	**58.9**	**52.1**	8.8	**35**

models, including BEVSegFormer [30], BEVFormer [20], BEVerse [52] and UniFusion [32], as shown in Tab. 1. LiDAR2Map outperforms all the existing methods significantly and boosts the performance of the LiDAR-based models from 29.5% mIoU to 57.4% mIoU. Our model with PointPillars [18] outperforms the state-of-the-art camera-based methods by 3.8% mIoU. With the stronger backbones like VoxelNet [57], LiDAR2Map even achieves a segmentation accuracy of 58.6% mIoU. It is worthy of noting that LiDAR2Map achieves the promising results in the case of Boundary class. It indicates that the accurate height information from LiDAR is important for map segmentation. Furthermore, we visualize the results of LiDAR2Map in some typical driving scenarios including cloudy and rainy conditions as shown in Fig. 6. More visualization results are included into the supplementary material.

Vehicle Segmentation. Vehicle segmentation is one of the most important task among the moving elements in autonomous driving. In order to examine the scalability of our method, we evaluate LiDAR2Map under two different settings for vehicle segmentation. We only adopt PointPillars as the LiDAR backbone and report the inference speed of LiDAR2Map on single NVIDIA RTX 2080Ti GPU for a fair comparison. As shown in Tab. 2, our method not only outperforms the state-of-the-art camera-based models by a large margin in accuracy, but also has the small model parameters with 35 FPS speed in inference. These promising results indicate the efficacy of our proposed LiDAR2Map approach and defend the strength of LiDAR on semantic map construction. We provide visual results on vehicle segmentation in the supplementary material.

4.3. Ablation Studies

BEV Feature Pyramid Decoder. In our experiments, we find that the layer number to obtain multi-scale features in the BEV-FPD has the substantial impact on the performance of LiDAR2Map for map segmentation. As shown in Tab. 3, the results of Camera-LiDAR fusion model and LiDAR2Map using PointPillars have been greatly improved with the increasing number of layers. With the 2-layer model in BEV-FPD, our LiDAR2Map achieves 43.8% mIoU. For the 4-layer model in BEV-FPD, a large performance improvement with +10.5% mIoU is obtained, where LiDAR2Map achieves the comparable results against the recent camera-based methods like BEVerse [52] and UniFusion [32]. As the number of layers is increased to 6, the ac-

Figure 6. Visualization of LiDAR2Map on the validation set of nuScenes with cloudy and rainy condition scenes. The left shows the surrounding views from cameras with the six views, which are just used for visualization. The middle is the input LiDAR data for inference. The right is the predicted semantic map and the corresponding ground truth.

curacy is boosted to 57.4% mIoU and achieves the best performance. We further visualize the feature maps to analyze our LiDAR2Map with different layer number in BEV-FPD. As shown in Fig. 7, the model with 6-layer BEV-FPD holds the distinct response map in the region, where the target element appears with little noise for semantic map construction. Furthermore, Tab. 3 reports the performance of fusion model as the "Teacher" in our LiDAR2Map. Notably, LiDAR2Map with 6-layer BEV-FPD as a "Student" network has achieved the 98.8% performance of fusion model with 2× faster inference speed.

Layer Num.	Div.	P. C.	Bound.	mIoU	FPS
2	49.3	34.1	58.4	47.3	8.2
	45.4	30.5	55.6	43.8	23.3
4	56.9	45.1	64.0	55.3	7.2
	55.7	43.9	63.2	54.3	16.3
6	60.8	47.2	66.3	58.1	6.3
	60.4	45.5	66.4	57.4	12.6

Table 3. Accuracy and speed performance with different layer number of BEV-FPD. At each row, the upper one is the results of the Camera-LiDAR fusion model ("Teacher"), and the lower one corresponds to the result of LiDAR2Map ("Student") in gray.

Online Camera-to-LiDAR Distillation Scheme. To examine the effect of each module in the online Camera-to-LiDAR distillation, we conduct the ablation experiments on nuScenes, including map and vehicle segmentation. For vehicle segmentation, we adopt Setting 2 for performance evaluation. As shown in Tab. 4, our baseline model achieves 52.2% mIoU on map segmentation by the design on 4-layer BEV-FPD. The proposed Position-Guided Feature Fusion

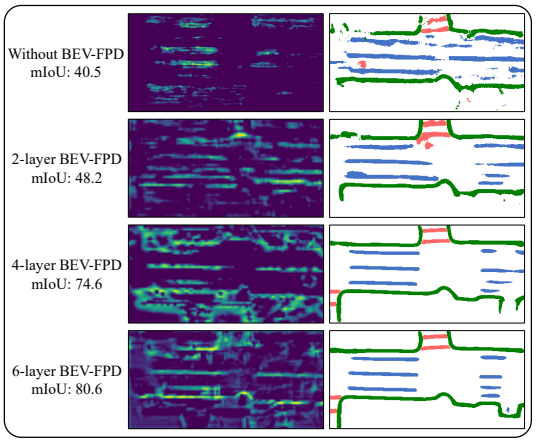

Figure 7. Visualization comparisons of LiDAR2Map with different BEV-FPDs and the corresponding semantic map predictions. The mIoU value means the evaluation score of the *single frame*. Besides the baseline model without BEV-FPD, we provide the results of the second layer's output with 2-, 4- and 6-layer BEV-FPD based models, respectively. LiDAR2Map with 6-layer BEV-FPD obtains the best segmentation performance and its feature map has more accurate responses with less noises.

Module (PGF^{2}M) improves the baseline around 0.4% mIoU and 1.5% mIoU on map and vehicle, respectively. This demonstrates that multi-modality fusion is effective with both spatial features from LiDAR and semantic features from camera. Moreover, Feature-level Distillation (FD) and Logit-level Distillation (LD) achieve over 0.9/1.2% mIoU and 1.1/0.7% mIoU performance gains on map/vehicle segmentation, respectively. These encouraging results demonstrate that our proposed online distillation scheme can effectively improve the model accuracy.

Baseline	PGF^{2}M	FD	LD	Map	Vehicle
✓				52.2	49.1
✓	✓			52.6	50.6
✓	✓	✓		53.5	51.8
✓	✓		✓	53.7	51.3
✓	✓	✓	✓	54.3	52.1

Table 4. The effectiveness of our online Camera-to-LiDAR distillation scheme with different settings on the nuScenes dataset.

Method	Div.	P. C.	Bound.	mIoU
Baseline	53.9	41.2	61.6	52.2
MonoDistill [6]	47.2	31.4	55.1	44.6
MGD [46]	52.0	38.7	59.6	50.1
xMUDA [15]	54.7	42.6	62.5	53.3
2DPASS [44]	55.3	43.0	62.4	53.6
LiDAR2Map (Ours)	55.7	43.9	63.2	54.3

Table 5. Performance comparison with different knowledge distillation strategies on the nuScenes dataset.

Cam. Num.	Div.	P. C.	Bound.	mIoU
0	53.9	41.2	61.6	52.2
1	55.4	43.1	63.3	53.9
2	56.3	43.7	63.4	54.5
4	56.0	43.1	63.0	54.0
6	55.7	43.9	63.2	54.3

Table 6. Performance comparison with different camera number during training on the nuScenes dataset.

Comparison with Other Distillation Schemes. To further investigate the effectiveness of our online distillation scheme, we compare it with current knowledge distillation strategies. We have re-implemented these methods in the BEV feature space under the same setting to facilitate a fair comparison. Tab. 5 shows the comparison results. Among these methods, MonoDistill [6] and MGD [46] are feature-based distillation methods. Their results are even worse than the baseline model, which indicates the difficulty of the cross-modal knowledge distillation on high-dimensional BEV features. xMUDA [15] and 2DPASS [44] are the logit-level distillation methods, which obtain better results over the baseline. Our Camera-to-LiDAR distillation scheme provides a more effective way compared against other distillation schemes and achieves the best performance.

Different Number of Cameras. Tab. 6 reports the results to compare the performance with the camera branch using the different number of cameras. The performance is not linearly related to the number of camera like those camera-based methods [55]. The LiDAR2Map model with two cameras of front and rear performs the best with 54.5% mIoU while the models with all six cameras achieves 54.3% mIoU. These results show that it is unnecessary to use so many cameras when the LiDAR is adopted.

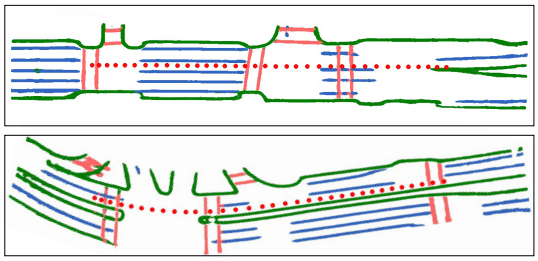

Figure 8. Scene-level semantic map obtained by accumulating 20s single frame maps with Bayesian filtering. The red dots indicate the trajectory of the ego vehicle.

4.4. Scene-Level Semantic Map Construction

Semantic map construction in a single frame is limited for self-driving. It is necessary to fuse the keyframes in a whole scene for scene-level map construction. We construct the scene-level semantic map on nuScenes [4], which is a typical dataset collected in driving scenes. Each scene lasts for 20s, and around 40 keyframes are sampled at 2Hz. We introduce a temporal accumulation method to build the scene-level semantic map. More precisely, the local semantic maps are warped to the global coordinate system with the extrinsic matrix. Then, the coincident regions are optimized by Bayesian filtering [33, 41] to obtain a smooth global map. The visual examples shown in Fig. 8 demonstrate that our LiDAR2Map approach is able to generate the consistent maps and provide more information for downstream tasks such as navigation and planning.

5. Conclusion

In this work, an efficient semantic map construction framework named LiDAR2Map, is presented with an effective BEV feature pyramid decoder and an online Camera-to-LiDAR distillation scheme. Unlike previous camera-based methods that have achieved excellent performance on this task, we mainly use LiDAR data and only extract image features as auxiliary network during training. The designed distillation strategy can make the LiDAR-based network well benefit from the semantic features of the camera image. Eventually, our method achieves the state-of-the-art performance on semantic map construction including map and vehicle segmentation under several competitive settings. The distillation scheme in LiDAR2Map is a general and flexible cross-modal distillation method. In the future, we will explore its application in more BEV perception tasks such as 3D object detection and motion prediction.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants (61831015). It is also supported by Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

[1] Sven Bauer, Yasamin Alkhorshid, and Gerd Wanielik. Using high-definition maps for precise urban vehicle localization. In *ITSC*, pages 492–497, 2016. 2

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, pages 4413–4421, 2018. 5

[3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606, 1992. 2

[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 *CVPR*, pages 11621–11631, 2020. 2, 5, 8

[5] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *ICCV*, pages 4794–4802, 2019. 2

[6] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. In *ICLR*, 2022. 2, 8

[7] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, pages 1607–1616, 2018. 2

[8] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? *arXiv preprint arXiv:2206.07959*, 2022. 6

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[10] Noureldin Hendy, Cooper Sloan, Feng Tian, Pengfei Duan, Nick Charchut, Yuesong Xie, Chuang Wang, and James Philbin. Fishing net: Future inference of semantic heatmaps in grids. *arXiv preprint arXiv:2006.09917*, 2020. 1, 2, 3

[11] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, pages 1921–1930, 2019. 2

[12] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *AAAI*, pages 3779–3787, 2019. 2

[13] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2

[14] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *ICCV*, pages 15273–15282, 2021. 6

[15] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *CVPR*, pages 12605–12614, 2020. 2, 5, 8

[16] Kitae Kim, Soohyun Cho, and Woojin Chung. Hd map update for autonomous driving with crowdsourced data. *IEEE Robotics and Automation Letters*, 6(2):1895–1901, 2021. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[18] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019. 1, 3, 5, 6

[19] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *ICRA*, pages 4628–4634, 2022. 1, 2, 3, 5, 6

[20] 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. In *ECCV*, pages 1–18, 2022. 1, 2, 6

[21] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. In *NeurIPS*, 2022. 2

[22] Zhiyuan Liang, Tiancai Wang, Xiangyu Zhang, Jian Sun, and Jianbing Shen. Tree energy loss: Towards sparsely annotated semantic segmentation. In *CVPR*, pages 16907–16916, 2022. 4

[23] Rong Liu, Jinling Wang, and Bingqi Zhang. High definition map for automated driving: Overview and analysis. *The Journal of Navigation*, 73(2):324–341, 2020. 2

[24] Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. *arXiv preprint arXiv:2206.08920*, 2022. 2

[25] Zhi Liu, Shaoyu Chen, Xiaojie Guo, Xinggang Wang, Tianheng Cheng, Hongmei Zhu, Qian Zhang, Wenyu Liu, and Yi Zhang. Vision-based uneven bev representation learning with polar rasterization and surface estimation. In *CoRL*, 2022. 6

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 5

[27] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, 2023. 1, 2

[28] Chenyang Lu, Marinus Jacobus Gerardus van de Molengraft, and Gijs Dubbelman. Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks. *IEEE Robotics and Automation Letters*, 4(2):445–452, 2019. 6

[29] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 5(3):4867–4873, 2020. 2, 6

[30] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. In *WACV*, pages 5935–5943, 2023. 1, 2, 3, 6

[31] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210, 2020. 2, 3, 5, 6

[32] Zequn Qin, Jingyu Chen, Chao Chen, Xiaozhi Chen, and Xi Li. Unifusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view. *arXiv preprint arXiv:2207.08536*, 2022. 1, 2, 6

[33] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *CVPR*, pages 11138–11147, 2020. 5, 6, 8

[34] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 2

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 5

[36] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Enabling spatio-temporal aggregation in birds-eye-view vehicle estimation. In *ICRA*, pages 5133–5139, 2021. 6

[37] Gustavo Salazar-Gomez, David Sierra González, Manuel Alejandro Diaz-Zapata, Anshul Paigwar, Wenqian Liu, Özgür Erkent, and Christian Laugier. Transfusegrid: Transformer-based lidar-rgb fusion for semantic grid prediction. In *ICARCV*, 2022. 1, 2, 6

[38] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, pages 10529–10538, 2020. 1

[39] Lin Song, Yanwei Li, Zeming Li, Gang Yu, Hongbin Sun, Jian Sun, and Nanning Zheng. Learnable tree filter for structure-preserving feature transform. In *NeurIPS*, volume 32, 2019. 4

[40] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019. 6

[41] Sebastian Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002. 8

[42] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *CVPR*, pages 4604–4612, 2020. 2

[43] Song Wang, Jianke Zhu, and Ruixiang Zhang. Meta-rangeseg: Lidar sequence semantic segmentation using multiple feature aggregation. *IEEE Robotics and Automation Letters*, 7(4):9739–9746, 2022. 2

[44] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shenghui Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *ECCV*, pages 677–695, 2022. 2, 8

[45] Bin Yang, Ming Liang, and Raquel Urtasun. Hdnet: Exploiting hd maps for 3d object detection. In *CoRL*, pages 146–155, 2018. 2

[46] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *ECCV*, pages 53–69, 2022. 2, 8

[47] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 1

[48] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multimodal virtual point 3d detection. In *NeurIPS*, volume 34, pages 16494–16507, 2021. 2

[49] Fisher Yu, Jianxiong Xiao, and Thomas Funkhouser. Semantic alignment of lidar data at city scale. In *CVPR*, pages 1722–1731, 2015. 2

[50] Shanxin Yuan, Bjorn Stenger, and Tae-Kyun Kim. 3d hand pose estimation from rgb using privileged learning with depth data. In *ICCVW*, pages 2866–2873, 2019. 2

[51] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *CVPR*, pages 8563–8573, 2022. 2

[52] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 1, 2, 3, 6

[53] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, pages 11953–11962, 2022. 2

[54] Yiming Zhao, Lin Bai, and Xinming Huang. Fidnet: Lidar point cloud semantic segmentation with fully interpolation decoding. In *IROS*, pages 4453–4458, 2021. 2

[55] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *CVPR*, pages 13760–13769, 2022. 8

[56] Yiyang Zhou, Yuichi Takeda, Masayoshi Tomizuka, and Wei Zhan. Automatic construction of lane-level hd maps for urban scenes. In *IROS*, pages 6649–6656, 2021. 2

[57] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, pages 4490–4499, 2018. 1, 3, 5, 6

[58] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *ICCV*, pages 16280–16290, 2021. 2