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 present 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.
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 (PGF²M). 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.
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 $F_{\text{LiDAR}}^{3D}$ 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 voxelize the point cloud first and uses the sparse convolution to build 3D network to encode the better 3D feature representation. Then, the unified BEV representation $F_{\text{BEV}}^{\text{LiDAR}}$ is obtained by pooling the 3D features $F_{\text{LiDAR}}^{3D}$.

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 $F_{\text{Camera}}^{\text{PV}}$ from each input image $I \in \mathbb{R}^{3 \times H \times W}$ by 2D convolution and predict the depth distribution of $D$ equally spaced discrete points associated with each pixel. Second, we assign the perspective features $F_{\text{Camera}}^{\text{PV}}$ to $D$ points along the camera ray direction to obtain a $D \times H \times W$ pseudo point cloud features $F_{\text{Camera}}^{\text{PV}}$. Finally, the pseudo point cloud features are flatten to the BEV space $F_{\text{BEV}}^{\text{PV}}$ 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 $F_{\text{BEV}}$ from the LiDAR or camera branch, we firstly perform $7 \times 7$ convolution on the BEV features to generate the global features with the large receptive field. The multi-scale BEV features $\{F_{\text{BEV}}^{N} \}_{i=1}^{N}$ 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 $\{F_{\text{BEV}}^{N} \}_{i=1}^{N}$ 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-
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 (PGF²M), Feature-level Distillation (FD) and Logit-level Distillation (LD).

Position-Guided Feature Fusion Module. PGF²M 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 \( F_{\text{BEV}}^{\text{LiDAR}} \) and camera image feature \( F_{\text{BEV}}^{\text{Camera}} \). Then, we perform the preliminary fusion through a \( 3 \times 3 \) convolutional layer to obtain \( F_{\text{BEV}}^{\text{Fusion}_1} \) as below,

\[
F_{\text{BEV}}^{\text{Fusion_1}} = \text{Conv} \times 3 \times 3 \left( [F_{\text{BEV}}^{\text{Camera}}, F_{\text{BEV}}^{\text{LiDAR}}] \right). \tag{1}
\]

Secondly, we calculate the relative coordinates of \( x \)-axis and \( y \)-axis \( F_{\text{Pos}}^{\text{BEV}} \) with the same size. Then, we concatenate it with the fusion result \( F_{\text{BEV}}^{\text{Fusion}_1} \) at the previous stage along the channel dimension to encode the spatial information, and perform \( 3 \times 3 \) convolution:

\[
F_{\text{BEV}}^{\text{Fusion}_2} = \text{Conv} \times 3 \times 3 \left( [F_{\text{BEV}}^{\text{Fusion}_1}, F_{\text{Pos}}^{\text{BEV}}] \right). \tag{2}
\]

\( F_{\text{BEV}}^{\text{Fusion}_2} \) 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 \( F_{\text{BEV}}^{\text{Fusion}_3} \) is obtained by

\[
F_{\text{BEV}}^{\text{Fusion}_3} = \sigma \left( \text{MLP} \left( \text{Avg}(F_{\text{BEV}}^{\text{Fusion}_2}) \right) \right) \odot F_{\text{BEV}}^{\text{Fusion}_2}. \tag{3}
\]

Finally, we add \( F_{\text{BEV}}^{\text{Fusion_3}} \) with the original BEV feature from camera \( F_{\text{BEV}}^{\text{Camera}} \) to obtain the fusion features \( F_{\text{BEV}}^{\text{Fusion}} \).

\[
F_{\text{BEV}}^{\text{Fusion}} = F_{\text{BEV}}^{\text{Camera}} + F_{\text{BEV}}^{\text{Fusion_3}}. \tag{4}
\]

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 \( \{F_{\text{BEV}}^{\text{m}}\}_{m=1}^{N} \) 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 \( F_{\text{BEV}}^{\text{m}} \) from LiDAR backbone and multi-scale BEV features \( \{F_{\text{BEV}}^{\text{m}}\}_{m=1}^{N} \) 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 \( M_{i}^{\text{BEV}} \) for the corresponding \( i \)-th scale BEV features \( \tilde{F}_{i}^{\text{BEV}} \) as following,

\[
M_{i}^{\text{BEV}} = \mathcal{F} \left( \mathcal{F} \left( F_{i}^{\text{BEV}}, F_{\text{BEV}}^{\text{low}} \right), \tilde{F}_{i}^{\text{BEV}} \right). \tag{5}
\]

We compute the affinity similarity between each \( M_{i}^{\text{BEV}} \) from the LiDAR branch and \( M_{i}^{\text{BEV}} \) 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,

\[
\mathcal{L}_{\text{feature}} = \sum_{i=1}^{N} \left\| M_{i}^{\text{BEV}} - M_{i}^{\text{BEV}_{\text{LiDAR}}} \right\|_{1}. \tag{6}
\]

We employ \( \mathcal{L}_{\text{feature}} \) as one of the loss terms to enable the LiDAR-based branch to benefit from the image feature implicitly through the network optimization.
Fusion Features (BEV)

LiDAR Features (BEV)

Tree Filter

\( \text{Distillation} \)

\[ \text{Tree Filter} \]

\[ \text{Pillar/Voxel Features} \]

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 \( P^\text{BEV}_{\text{LiDAR}} \) and \( P^\text{BEV}_{\text{Fusion}} \) can be obtained. As in [15], we adopt KL divergence to measure the similarity on the probability distribution, which makes the \( P^\text{BEV}_{\text{LiDAR}} \) of LiDAR closer to \( P^\text{BEV}_{\text{Fusion}} \) of fusion “Teacher” as below,

\[
L_{\text{logit}} = D_{\text{KL}}(P^\text{BEV}_{\text{Fusion}} \parallel P^\text{BEV}_{\text{LiDAR}}). \tag{7}
\]

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:

\[
L = L_{\text{seg}} + L_{\text{fusion seg}} + L_{\text{distill}}, \tag{8}
\]

where \( L_{\text{seg}} \) and \( L_{\text{fusion seg}} \) are the segmentation losses of LiDAR-branch and Camera-LiDAR fusion branch, respectively. \( L_{\text{distill}} \) consists of \( L_{\text{feature}} \) and \( L_{\text{logit}} \) for online Camera-to-LiDAR distillation.

The segmentation loss for semantic map construction is composed of two items including \( L_{\text{ce}} \) and \( L_{\text{ls}} \) as following,

\[
L_{\text{seg}} = L_{\text{ce}} + L_{\text{ls}}, \tag{9}
\]

where \( L_{\text{ce}} \) is the cross-entropy loss. \( L_{\text{ls}} \) is employed to maximize the Intersection-over-Union (IoU) score as below,

\[
L_{\text{ls}} = \frac{1}{|C|} \sum_{c \in C} \Delta_{\text{c}} \left( m(c) \right), \tag{10}
\]
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].

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

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Image Size</th>
<th>Modality</th>
<th>Backbone</th>
<th>Setting 1</th>
<th>Setting 2</th>
<th>#Params(M)</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>VED [28]</td>
<td>800×600</td>
<td>Camera</td>
<td>ResNet-50</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PON [33]</td>
<td>800×600</td>
<td>Camera</td>
<td>ResNet-50</td>
<td>24.7</td>
<td>-</td>
<td>38</td>
<td>30</td>
</tr>
<tr>
<td>VPN [29]</td>
<td>800×600</td>
<td>Camera</td>
<td>ResNet-50</td>
<td>25.5</td>
<td>-</td>
<td>18</td>
<td>-</td>
</tr>
<tr>
<td>STA [36]</td>
<td>1280×720</td>
<td>Camera</td>
<td>ResNet-50</td>
<td>36.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lift-Splat [31]</td>
<td>352×128</td>
<td>Camera</td>
<td>EfficientNet-B0</td>
<td>-</td>
<td>32.1</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>FIERY Static [14]</td>
<td>448×224</td>
<td>Camera</td>
<td>EfficientNet-B4</td>
<td>37.7</td>
<td>35.8</td>
<td>7.4</td>
<td>8</td>
</tr>
<tr>
<td>PolarBEV [23]</td>
<td>960×448</td>
<td>Camera</td>
<td>EfficientNet-B4</td>
<td>45.4</td>
<td>41.2</td>
<td>7.4</td>
<td>10</td>
</tr>
<tr>
<td>SimpleBEV [8]</td>
<td>800×448</td>
<td>Camera</td>
<td>ResNet-101</td>
<td>-</td>
<td>47.4</td>
<td>37</td>
<td>7.3</td>
</tr>
<tr>
<td>TransFuseGrid [37]</td>
<td>352×128</td>
<td>Camera &amp; LiDAR</td>
<td>EfficientNet-B0 &amp; PointPillars</td>
<td>-</td>
<td>35.9</td>
<td>-</td>
<td>18.4</td>
</tr>
<tr>
<td>Pillar feature Net [37]</td>
<td>-</td>
<td>LiDAR</td>
<td>PointPillars</td>
<td>-</td>
<td>23.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LiDAR2Map</td>
<td>-</td>
<td>LiDAR</td>
<td>PointPillars</td>
<td><strong>58.9</strong></td>
<td><strong>52.1</strong></td>
<td>8.8</td>
<td><strong>35</strong></td>
</tr>
</tbody>
</table>

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.

Accuracy 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.

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

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 Module (PGF2M) 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.
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.

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


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


[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


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


[58] 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


[66] 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


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


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