SA-BEV: Generating Semantic-Aware Bird’s-Eye-View Feature for Multi-view 3D Object Detection

Jinqing Zhang¹, Yanan Zhang¹, Qingjie Liu¹,²,³*, Yunhong Wang¹,³
¹State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
²Zhongguancun Laboratory, Beijing, China
³Hangzhou Innovation Institute, Beihang University, Hangzhou, China
{zhangjinqing, zhangyanan, qingjie.liu, yhwang}@buaa.edu.cn

Abstract

Recently, the pure camera-based Bird’s-Eye-View (BEV) perception provides a feasible solution for economical autonomous driving. However, the existing BEV-based multi-view 3D detectors generally transform all image features into BEV features, without considering the problem that the large proportion of background information may submerge the object information. In this paper, we propose Semantic-Aware BEV Pooling (SA-BEVPool), which can filter out background information according to the semantic segmentation of image features and transform image features into semantic-aware BEV features. Accordingly, we propose BEV-Paste, an effective data augmentation strategy that closely matches with semantic-aware BEV feature. In addition, we design a Multi-Scale Cross-Task (MSCT) head, which combines task-specific and cross-task information to predict depth distribution and semantic segmentation more accurately, further improving the quality of semantic-aware BEV feature. Finally, we integrate the above modules into a novel multi-view 3D object detection framework, namely SA-BEV. Experiments on nuScenes show that SA-BEV achieves state-of-the-art performance. Code has been available at https://github.com/mengtan00/SA-BEV.git.

1. Introduction

Camera and LiDAR are the two most commonly used sensors for 3D object detection, which is essential to autonomous driving systems. LiDAR-based methods [4, 12, 33, 39, 34, 35, 42] have attained excellent performance due to the accurate spatial structure information of point clouds, but the expensive LiDAR sensor reduces its universality. In contrast, camera-based methods [30, 29, 31, 18, 19] are relatively low-cost with plentiful semantic information, but are constrained by the lack of geometric depth cues.

*indicates the corresponding author.

Figure 1: Comparison between normal BEV features (upper row) and semantic-aware BEV features (lower row). The brightness reveals the norm of the features and the red / green boxes are the ground truth of the original / pasted frame. The last column shows BEV-Paste, an data augmentation strategy that matches semantic-aware BEV features.

Considering the performance gap between camera and LiDAR, the Bird’s-Eye-View paradigm transforms multi-view image features into the BEV feature to make the following 3D perception easier [22, 10]. This practical and scalable camera-only paradigm is gaining popularity, and numerous advancements have allowed it to reach high perceptual precision [15, 14, 16, 8, 11]. The core step of the BEV paradigm is generating virtual points from image features, which will be projected into the “pillarized” BEV space. The features of the virtual points in the same pillar are then cumulated as the BEV feature. However, this operation does not fully utilize the semantic information of the image features and will inject massive background information that submerges object information.

In order to take full advantage of the valuable semantic
information of image features, we propose Semantic-Aware BEV Pooling (SA-BEVPool) to generate semantic-aware BEV features, which replace the normal BEV feature for 3D detection. Before projecting virtual points into BEV space, the semantic segmentation of image features is first predicted. If a virtual point is generated by the image element that belongs to the background, it will not be projected into BEV space. Similarly, virtual points with low depth scores will also be ignored. The comparison between normal BEV features and semantic-aware BEV features is shown in Fig. 1. SA-BEVPool can obviously filter out most of the background BEV features and alleviate the problem that the large proportion of background information submerges object information, therefore effectively improving the detection performance. Some multi-modal 3D objectors [27, 36] also adopt segmentation on images when combining with LiDAR features, but they generally use powerful instance segmentation networks like CenterNet2 [43] to predict the segmentation of the large-scale image. Instead, SA-BEVPool can be easily applied in current BEV-based detectors like BEVDepth [15] and BEVStereo [14] by using their depth branch to simultaneously predict the semantic segmentation of small-scale image features.

GT-Paste [33] is a successful data augmentation strategy that has been frequently adopted by various LiDAR-based 3D detectors. However, due to the modality gap, it cannot directly adapt to camera-based 3D detectors. In our work, thanks to the reliable depth distribution and semantic segmentation predicted on image features, the semantic-aware BEV feature can approximately represent the information of all objects that are located appropriately in BEV space. As a result, adding the semantic-aware BEV features of another frame to the current semantic-aware BEV feature is the same as pasting all objects of another frame into the current frame. This strategy, we called BEV-Paste, enhances data diversity in a similar way to GT-Paste.

Although it is convenient to predict depth distribution and semantic segmentation with the same branch, doing so may result in a sub-optimal semantic-aware BEV feature. Research conclusion in the field of multi-task learning demonstrates that the integration of specific tasks and cross-task information is more conducive to the optimal solution of multiple prediction tasks. Inspired by this, we design a Multi-Scale Cross-Task (MSCT) head to combine the task-specific and cross-task information through multi-task distillation and dual-supervision on multiple scales prediction.

We integrate our proposed modules as a whole and name it SA-BEV. Extensive experiments on nuScenes dataset show that SA-BEV achieves a new state-of-the-art. In summary, the major contributions of this paper are:

- We propose SA-BEVPool, which uses semantic information to filter out unnecessary virtual points and generate the semantic-aware BEV feature, alleviating the problem that the large proportion of background information submerges the object information.
- We propose BEV-Paste, an effective and convenient data augmentation strategy closely matching the semantic-aware BEV feature, which enhances data diversity and further promotes detection performance.
- We propose the MSCT head that combines the task-specific and cross-task information through multi-task learning on multiple scales, facilitating the optimization of the semantic-aware BEV feature.

2. Related Work

2.1. Vision-based 3D Object Detection

Although camera does not provide reliable depth of the surroundings like LiDAR, the plentiful semantic information carried by images still supports vision-based 3D object detectors to achieve considerable precision. Early vision-based 3D detectors predict attributes of 3D objects directly from 2D image features. For instance, CenterNet [44], a 2D detector, can be used to predict 3D objects without many modifications. Lately, FCOS3D [30] detects the 2D centers of the 3D objects, and features around the centers are used to predict the 3D attributes. PGD [29] establishes geometric relation graphs to improve the depth estimation results for better 3D object detection. DETR3D [31] follows DETR [2] and detects 3D objects with Transformer. PETR [18] introduces 3D position-aware representations, ameliorating the detection precision. PETRv2 [19] further brings in temporal information and improves efficiency.

Recently, an approach that transforms the image feature into the BEV feature is proposed by LSS [15], and BEVDet [10] employs the detection head of CenterPoint [35] to predict 3D objects from BEV feature. This paradigm can achieve comparable accuracy without cumbersome operations and is easy to extend, making it gain popularity. BEVDet4D [8] processes multiple key frames to introduce temporal information. BEVFormer [16] utilizes the deformable attention mechanism to generate the BEV feature. BEVDepth [15] applies explicit depth supervision on the predicted latent depth distribution, improving the detection accuracy. BEVStereo [14] further improves the quality of the depth by applying the multi-view stereo on nearby key frames. PolarFormer [11] generates the BEV feature using the polar coordinate for a more accurate location. However, these methods project all image features into BEV features, without considering the problem that the large proportion of background information may submerge the object information. In this paper, we propose SA-BEVPool, which can filter out background information according to the semantic segmentation of image features and generate semantic-aware BEV features.
2.2. Data Augmentation in 3D Object Detection

The diversity of the dataset is crucial to the generalization performance of models. Besides regular data augmentation like random scaling, flipping and rotation, GT-Paste [33] is another effective strategy frequently used by LiDAR-based detectors. It crops the points according to the 3D boxes of ground truth and pastes them to other frames to create new training data. Lately, an improvement in generating a visibility map to correct the wrong occlusion relationship introduced by GT-Paste is proposed in [7]. The augmentation on the individual object is also proposed in [3, 41] which takes object points into parts and applies operations like dropout, swap, and mix on it.

Since GT-Paste shows excellent effectiveness on increase data diversity, there have been some attempts to adopt it in camera-only 3D detectors. Box-Mixup and Box-Cut-Paste proposed by [24] directly cut the objects from images according to their 2D bounding boxes and paste them into other frames. To paste precisely, objects are cropped by their instance masks in [37]. Pointaugmenting [28] utilizes a more complicated way to tackle the occlusion relationship of original objects and pasted objects. However, these attempts to expand the GT-Paste into image space cannot easily overcome the issues caused by the gap between LiDAR and camera. In this paper, we propose BEV-Paste, a convenient way to effectively extend GT-Paste into BEV-based methods with the help of SA-BEVPool.

2.3. Multi-task Learning

Multi-task learning generally leads to better prediction through interactive learning between multiple tasks. According to [25], both task-specific information and cross-task information are important for getting optimal results on multiple tasks. PAD-Net [32] proposes multi-modal distillation module to automatically supplement cross-task information. PAP-Net [40] extracts cross-task affinity patterns and recursively propagates the pattern by affinity matrices. MTI-Net [26] models the task interactions at different scales and aggregates multi-scale information to make precise predictions. Some methods also introduce multi-task learning into 3D object detection. MMF [17] deeply fuses the features of images and LiDAR through simultaneous supervision made on multiple tasks. Latent support surfaces are estimated in [23] to help improve the precision of 3D detection. A multi-task LiDAR network proposed by [5] makes predictions on 3D detection and road understanding that can complement each other. Some BEV-based 3D detectors [16, 19] also apply BEV segmentation to obtain better BEV representation. In this paper, we propose the MSCT head that combines the task-specific and cross-task information of multiple scales for depth estimation and semantic segmentation, facilitating the optimization of the semantic-aware BEV feature.

3. Method

In this work, we propose SA-BEV, a novel multi-view 3D object detection framework that generates semantic-aware BEV features for better detection performance. It contains the Semantic-Aware BEV Pooling (SA-BEVPool), the BEV-Paste data augmentation strategy and the Multi-Scale Cross-Task (MSCT) head. The overall framework of SA-BEV is shown in Fig. 2.
3.1. Semantic-Aware BEV Pooling

The way of transforming the image features into BEV features for better perception was first proposed by LSS [22]. It predicts the depth distribution $\alpha$ and context feature $c$ of each image feature element. Then each element generates virtual points at different depths. The feature of the point at depth $d$ is represented as $p_d = \alpha_d c$. After that, all virtual points will be projected to the BEV space which is divided into pillars. The features of virtual points in the same pillar will be cumulated as the BEV feature. This process is known as BEV pooling.

Subsequent BEV-based 3D detectors [15, 20, 9] significantly improve the efficiency and accuracy of BEV pooling. But what is unchanged is that these methods insist on projecting all virtual points into BEV space. However, we argue that this is unnecessary for 3D detection tasks. On the contrary, if all virtual points belonging to the background are projected, the foreground virtual points that account for less than 2% of the total virtual points will be submerged. It will confuse the following detection head and reduce the detection accuracy.

To highlight the valuable foreground information in the BEV features, we propose a novel Semantic-Aware BEV Pooling (SA-BEVPool) that is shown in Fig. 3. It applies semantic segmentation on the image features to get the foreground score $\beta$ of each element. The element with low $\beta$ is more possible to carry useless information for detection, and the virtual points generated from it will be ignored during the BEV pooling. Similarly, the virtual points with low $\alpha_d$ provide trivial information and will also be ignored. Denoting the filtering function as:

$$ F(x, y) = \begin{cases} 0, & x < y, \\ 1, & x \geq y \end{cases}, $$

the point features after filtering are changed as:

$$ \hat{p}_d = F(\alpha_d, T_D)F(\beta, T_S)p_d, $$

where $T_D$ and $T_S$ are the threshold for $\alpha_d$ and $\beta$. Only non-empty $p_d$ will construct BEV feature. Since the operation of filtering relatively low-value virtual points attaches semantic information to the generated BEV features, they can be called semantic-aware BEV features.

The difference between the normal BEV feature and the semantic-aware BEV feature is clearly shown in Fig. 1. The normal BEV features in the first row generally have a ring of light in the center, which represents the ground. It accounts for most of the signal strength in the feature without contributing useful information for detection. In contrast, most of the background information is removed in the semantic-aware BEV feature and object information is emphasized. Furthermore, the location of object information in the semantic-aware BEV feature matches the ground truth well, making the detection head easier to predict accurately.

3.2. BEV-Paste

GT-Paste [33] is a data augmentation strategy commonly used by LiDAR-based 3D detectors. It has been proved that the diversity of the dataset can be effectively increased by sampling the points in 3D boxes and pasting them into other frames. However, several problems prevent the application of GT-Paste in camera-based methods. First, sampling an object by the bounding box on the image can not get its pure data as the point cloud does. Another problem is that pasting objects to another image may wrongly occlude original objects and result in data loss. In addition, the illumination change of different frames also gives the pasted objects unnatural appearances. Some multi-modal 3D detectors [24, 28, 37, 38] make an effort to solve these issues but generally lack convenience and accuracy.

Here, we propose BEV-Paste that successfully applies GT-Paste in camera-only 3D detectors without complicated steps. With SA-BEVPool, the semantic-aware BEV features transformed from image features approximately represent the information of all objects in the frame as shown in Fig. 1. It makes adding arbitrary semantic-aware BEV features of two frames during the training phase equivalent to aggregating the objects contained in two frames into
one frame. While effectively increasing the diversity of the entire training dataset, BEV-Paste does not increasing the computational cost in the inference stage.

In practice, we randomly select the original semantic-aware BEV feature $B_O$ and the pasted semantic-aware BEV feature $B_P$ from the same batch. This is to guarantee $B_O$ and $B_P$ follow the same distribution. Instead of directly pasting $B_P$ to $B_O$, extra BEV data augmentations (BDA) shown in Fig. 4 are first applied to $B_P$ and $B_P$ is obtained. It prevents the data duplication of $B_P$. The same augmentation is also applied to the ground truth of the pasted frame $G_P$ to get $\hat{G}_P$. The detection loss after BEV-Paste can be represented as:

$$L_{det} = L_{det}(Det(B_O + \hat{B}_P), G_O \cup \hat{G}_P),$$  

where $Det$ includes BEV encoder and detection head, $G_O$ is the ground truth of original frame.

### 3.3. Multi-Scale Cross-Task Head

It is a convenient way to obtain semantic-aware BEV features by making the depth branch predict semantic segmentation at the same time, but it generally leads to sub-optimal results. Let us regard the generation of the semantic-aware BEV feature as a multi-task learning application. According to the research conclusion, both task-specific information and cross-task information are important for getting the global optimal solution of multiple tasks. If the depth distribution and semantic segmentation are predicted by the same network branch, the network only extracts cross-task information from the image features and can not perform optimally on each task.

Inspired by the principle of multi-task learning, we design a Multi-Scale Cross-Task (MSCT) head as shown in Fig. 5. In the first stage, the head takes 1/16 scale image feature $F_{16}^{16}$ as input and makes a relatively coarse prediction of depth distribution and semantic segmentation. After that, $F_{16}^{16}$ is transformed into depth feature $F_{16}^{16}$ and semantic feature $F_{16}^{16}$, which carry the task-specific information of their own task. To complement cross-task information, Multi-Task Distillation (MTD) module proposed in [32] is applied between $F_{16}^{16}$ and $F_{16}^{16}$. It is composed of several self-attention blocks, which generate gate map $G$ by

$$G(F) = \sigma(W_G F),$$  

where $W_G$ is the gate convolution and $\sigma$ denotes sigmoid function. The features supplemented by the cross-task information can be formulated as:

$$\hat{F}_{16}^{16} = F_{16}^{16} + G(F_{16}^{16}) \odot (W_I F_{16}^{16})$$  

$$\hat{F}_{16}^{16} = F_{16}^{16} + G(F_{16}^{16}) \odot (W_I F_{16}^{16})$$  

where $W_I$ is the task convolution and $\odot$ denotes element-wise multiplication. It is clear that MTD uses these self-attention blocks to automatically extract cross-task information from one task feature and add it to other task features.

After the task features interaction, $F_{16}^{16}$ and $F_{16}^{16}$ obtain both task-specific information and cross-task information. Before inputting them into the second stage prediction head, they are up-sampled to the 1/8 scale and combined with the 1/8 scale image feature $F_{8}^{8}$ using the same self-attention blocks. The features can be formulated as:

$$\hat{F}_{8}^{16} = Up(F_{16}^{16}) + \hat{G}(F_{8}^{8}) \odot (W_I F_{8}^{8})$$  

$$\hat{F}_{8}^{16} = Up(F_{16}^{16}) + \hat{G}(F_{8}^{8}) \odot (W_I F_{8}^{8})$$

The second stage head then predicts the relatively fine depth distribution and semantic segmentation which will be used to generate the semantic-aware BEV feature.

During training, both predictions on the 1/16 and 1/8 scale are supervised. It ensures the first stage head can extract task-specific information and the second stage head can combine the task-specific information with cross-task information. The supervision signals are obtained by projecting point clouds on images following BEVDepth [15]. The depth values of the projected points are the depth labels and the points in the 3D boxes are regarded as the foreground. The total loss can be formulated as:

$$L = L_{det} + \lambda_1 \frac{2}{2} (L_{S}^{16} + L_{S}^{8}) + \lambda_2 \frac{2}{2} (L_{D}^{16} + L_{D}^{8})$$  

### 4. Experiments

In this section, we first introduce our experimental settings. Then, comparisons with previous state-of-the-art multi-view 3D detectors are shown. Finally, comprehensive experiments with detailed ablation studies are conducted on SA-BEV to show the effectiveness of each component, i.e. SA-BEVPool, BEV-Paste and MSCT head.
4.1 Experimental Settings

4.1.1 Dataset and Metrics

nuScenes [1] dataset is a large-scale autonomous driving benchmark. It contains 750 scenarios for training, 150 scenarios for validation and 150 scenarios for testing. Each scenario lasts for around 20 seconds and the key samples are annotated at 2Hz. The data collected from six cameras, one LiDAR and five radars are provided to every sample. For 3D object detection, nuScenes Detection Score (NDS) is proposed to capture all aspects of the nuScenes detection tasks. Except mean average precision (mAP), NDS is also related to five types of true positive metrics (TP metrics), including mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), mean Average Attribute Error (mAAE).

Table 1: Comparison with previous state-of-the-art multi-view 3D detectors on the nuScenes test set.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Resolution</th>
<th>mAP↑</th>
<th>NDS↑</th>
<th>mATE↓</th>
<th>mASE↓</th>
<th>mAOE↓</th>
<th>mAVE↓</th>
<th>mAAE↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCOS3D [30]</td>
<td>ResNet-101</td>
<td>900×1600</td>
<td>0.358</td>
<td>0.428</td>
<td>0.690</td>
<td>0.249</td>
<td>0.452</td>
<td>1.434</td>
<td>0.124</td>
</tr>
<tr>
<td>PGD [29]</td>
<td>ResNet-101</td>
<td>900×1600</td>
<td>0.386</td>
<td>0.448</td>
<td>0.626</td>
<td>0.245</td>
<td>0.451</td>
<td>1.509</td>
<td>0.127</td>
</tr>
<tr>
<td>DETER3D [31]</td>
<td>V2-99</td>
<td>900×1600</td>
<td>0.412</td>
<td>0.479</td>
<td>0.641</td>
<td>0.255</td>
<td>0.394</td>
<td>0.845</td>
<td>0.133</td>
</tr>
<tr>
<td>BEVDet [10]</td>
<td>Swin-B</td>
<td>900×1600</td>
<td>0.424</td>
<td>0.488</td>
<td>0.524</td>
<td>0.242</td>
<td>0.373</td>
<td>0.950</td>
<td>0.148</td>
</tr>
<tr>
<td>PETR [18]</td>
<td>V2-99</td>
<td>900×1600</td>
<td>0.441</td>
<td>0.504</td>
<td>0.593</td>
<td>0.249</td>
<td>0.383</td>
<td>0.808</td>
<td>0.132</td>
</tr>
<tr>
<td>BEVFormer [16]</td>
<td>V2-99</td>
<td>900×1600</td>
<td>0.481</td>
<td>0.569</td>
<td>0.582</td>
<td>0.256</td>
<td>0.375</td>
<td>0.378</td>
<td>0.126</td>
</tr>
<tr>
<td>BEVDet4D [8]</td>
<td>Swin-B</td>
<td>640×1600</td>
<td>0.451</td>
<td>0.569</td>
<td>0.511</td>
<td>0.241</td>
<td>0.386</td>
<td>0.301</td>
<td>0.121</td>
</tr>
<tr>
<td>PolarFormer [11]</td>
<td>V2-99</td>
<td>900×1600</td>
<td>0.493</td>
<td>0.572</td>
<td>0.556</td>
<td>0.256</td>
<td>0.364</td>
<td>0.440</td>
<td>0.127</td>
</tr>
<tr>
<td>PETRv2 [19]</td>
<td>V2-99</td>
<td>640×1600</td>
<td>0.490</td>
<td>0.582</td>
<td>0.561</td>
<td>0.243</td>
<td>0.361</td>
<td>0.343</td>
<td>0.120</td>
</tr>
<tr>
<td>BEVDepth [15]</td>
<td>V2-99</td>
<td>640×1600</td>
<td>0.503</td>
<td>0.600</td>
<td>0.445</td>
<td>0.245</td>
<td>0.378</td>
<td>0.320</td>
<td>0.126</td>
</tr>
<tr>
<td>BEVStereo [14]</td>
<td>V2-99</td>
<td>640×1600</td>
<td>0.525</td>
<td>0.610</td>
<td>0.431</td>
<td>0.246</td>
<td>0.358</td>
<td>0.357</td>
<td>0.138</td>
</tr>
<tr>
<td>SA-BEV</td>
<td>V2-99</td>
<td>640×1600</td>
<td>0.533</td>
<td>0.624</td>
<td>0.430</td>
<td>0.241</td>
<td>0.338</td>
<td>0.282</td>
<td>0.139</td>
</tr>
</tbody>
</table>

Table 2: Comparison with previous state-of-the-art multi-view 3D detectors on the nuScenes val set.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Resolution</th>
<th>mAP↑</th>
<th>NDS↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCOS3D [30]</td>
<td>ResNet-101</td>
<td>900×1600</td>
<td>0.343</td>
<td>0.415</td>
</tr>
<tr>
<td>DETER3D [31]</td>
<td>ResNet-101</td>
<td>900×1600</td>
<td>0.303</td>
<td>0.374</td>
</tr>
<tr>
<td>PGD [29]</td>
<td>ResNet-101</td>
<td>900×1600</td>
<td>0.369</td>
<td>0.428</td>
</tr>
<tr>
<td>PETR [18]</td>
<td>ResNet-101</td>
<td>512×1408</td>
<td>0.357</td>
<td>0.421</td>
</tr>
<tr>
<td>BEVDepth [15]</td>
<td>ResNet-101</td>
<td>900×1600</td>
<td>0.416</td>
<td>0.517</td>
</tr>
<tr>
<td>BEVDepth [16]</td>
<td>ResNet-101</td>
<td>900×1600</td>
<td>0.421</td>
<td>0.524</td>
</tr>
<tr>
<td>BEVFormer [11]</td>
<td>Swin-B</td>
<td>640×1600</td>
<td>0.421</td>
<td>0.545</td>
</tr>
<tr>
<td>PETRv2 [19]</td>
<td>ResNet-101</td>
<td>900×1600</td>
<td>0.432</td>
<td>0.528</td>
</tr>
<tr>
<td>BEVStereo [14]</td>
<td>ConvNeXt-B</td>
<td>512×1408</td>
<td>0.462</td>
<td>0.558</td>
</tr>
<tr>
<td>SA-BEV</td>
<td>ConvNeXt-B</td>
<td>512×1408</td>
<td>0.479</td>
<td>0.579</td>
</tr>
</tbody>
</table>

4.1.2 Implementation Detail

We accomplish our proposed improvements on the network structure of BEVDepth [15]. Our experiments are implemented based on MMDetection3D with 8 NVIDIA GeForce RTX 3090 GPUs. Models are trained with AdamW [21] optimizer and gradient clip is utilized. The universal data augmentation we adopt on the image and BEV feature follows the configuration in [10]. For the ablation study, we use ResNet-50 [6] as the image backbone and the image size is downsampled to 256×704. The models are trained for 24 epochs without CBGS strategy [45] for the ablation study. When compared to other methods, the models are trained for 20 epochs with CBGS strategy.

4.2 Main Results

4.2.1 Comparison with State-of-the-Arts

We compare SA-BEV with state-of-the-art multi-view 3D detectors on nuScenes test set and show the results in Table 1. We take 640×1600 resolution image as input and VoVNet-99 [13] as the image backbone. SA-BEV achieves the best mAP and NDS, 3.0% and 2.4% higher than its baseline (i.e. BEVDepth [15]). It also exceeds BEVStereo [14] by 0.8% mAP and 1.4% NDS, which adopts the complicated multi-view stereo structure for more accurate depth estimation. The comparison on nuScenes val set is shown in 2. It can be found SA-BEV also achieves the best detection precision. The decent results highlight the advantage of the proposed SA-BEV.

4.2.2 Visualization

We visualize the detection results on images and BEV features in Fig. 6. Compared to BEVDepth, SA-BEV can make more precise predictions with the help of semantic-aware BEV features. For instance, the orange dashed rect-
Figure 6: Visualization of detection results on images and BEV features. The red boxes and green boxes on BEV features represent the ground truth and the predicted boxes, respectively. The dashed rectangles illustrate that the prediction of SA-BEV is more precise than BEVDepth.

Table 3: Ablation study of component in SA-BEV on the nuScenes val set. pool, paste and head denotes SA-BEVPool, BEV-Paste and MSCT head, respectively.

<table>
<thead>
<tr>
<th>Baseline</th>
<th>pool</th>
<th>paste</th>
<th>head</th>
<th>mAP↑</th>
<th>NDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEVDepth [15]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.330</td>
<td>0.436</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.340</td>
<td>0.449</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>0.354</td>
<td>0.464</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>0.365</td>
<td>0.483</td>
</tr>
<tr>
<td>BEVDet [10]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.278</td>
<td>0.322</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>0.304</td>
<td>0.348</td>
</tr>
<tr>
<td>BEVStereo [14]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.349</td>
<td>0.454</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>0.364</td>
<td>0.467</td>
</tr>
</tbody>
</table>

angles show that the filtration of the background prevents SA-BEV from making false detection. The yellow dashed rectangles indicate that the semantic-aware BEV feature correctly emphasizes the location of the truck, which results in precise detection. In addition, the green / blue dashed rectangles display that SA-BEV can successfully recall the missed object and remove the redundant detection box.

4.3. Ablation Study

4.3.1 Component Analysis

We individually evaluate the contributions of SA-BEVPool, BEV-Paste and MSCT head with BEVDepth [15] as the baseline. The results are shown in Table 3. After applying SA-BEVPool, the performance is boosted by 1.0% and 1.3% on mAP and NDS. It is further improved by 1.4% / 1.5% and 1.1% / 1.9% through incorporating BEV-Paste and MSCT head respectively. Finally, we obtain the full model of SA-BEV, which gains 3.5% and 4.7% in total, validating its effectiveness. SA-BEVPool and BEV-Paste are also applied to BEVDet [10] and BEVStereo [15], increasing 2.6% / 2.6% and 1.5% / 1.3% respectively on mAP and NDS. It demonstrates that these components can be easily embedded into the existing BEV-based detectors and bring about noticeable precision improvement.

4.3.2 Semantic-Aware BEV Pooling

The semantic threshold \( T_S \) and semantic threshold \( T_D \) in SA-BEVPool control the scale of the valid virtual points. We vary thresholds and show the results in Table 4. The results indicate that even a low \( T_D \) can sharply reduce the
Table 4: Ablation study of the semantic threshold used in SA-BEVPool. “Percentage” denotes the average proportion of valid virtual points.

<table>
<thead>
<tr>
<th>$T_D$</th>
<th>$T_S$</th>
<th>mAP↑</th>
<th>NDS↑</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>0.330</td>
<td>0.436</td>
<td>100%</td>
</tr>
<tr>
<td>0.0085</td>
<td>-</td>
<td>0.338</td>
<td>0.438</td>
<td>7.92%</td>
</tr>
<tr>
<td>0.0085</td>
<td>0.10</td>
<td>0.339</td>
<td>0.444</td>
<td>3.26%</td>
</tr>
<tr>
<td>0.0085</td>
<td>0.25</td>
<td>0.430</td>
<td>0.449</td>
<td>1.80%</td>
</tr>
<tr>
<td>0.0085</td>
<td>0.50</td>
<td>0.329</td>
<td>0.432</td>
<td>0.89%</td>
</tr>
</tbody>
</table>

Table 5: Ablation study of BEV-Paste strategy. $N_F$ denotes the average number of frames that are pasted to the original frame.

<table>
<thead>
<tr>
<th>Method</th>
<th>$N_F$</th>
<th>mAP↑</th>
<th>NDS↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o extra BDA</td>
<td>0</td>
<td>0.340</td>
<td>0.449</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.348</td>
<td>0.453</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.349</td>
<td>0.453</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.349</td>
<td>0.452</td>
</tr>
<tr>
<td>w/ extra BDA</td>
<td>1</td>
<td>0.354</td>
<td>0.464</td>
</tr>
</tbody>
</table>

Table 6: Ablation study of MSCT head. “MTD”, “DS” and “MS” denote the multi-task distillation module, the dual supervision and the utilization of multi-scale image features.

<table>
<thead>
<tr>
<th>MTD</th>
<th>DS</th>
<th>MS</th>
<th>mAP↑</th>
<th>NDS↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td></td>
<td></td>
<td>0.354</td>
<td>0.464</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.358</td>
<td>0.468</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.361</td>
<td>0.473</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.361</td>
<td>0.478</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.365</td>
<td>0.483</td>
</tr>
</tbody>
</table>

scale of valid virtual points and an appropriate $T_S$ can effectively improve the detection precision. However, a too-high semantic threshold may lead to the loss of foreground information and damage the precision. We set the semantic threshold to 0.25 as a good trade-off. It only needs 1.8% valid virtual points, resulting in 1% mAP and 1.3% NDS performance improvement.

4.3.3 BEV-Paste

We conduct experiments with different settings when applying BEV-Paste, including the number of frames pasted to each original frame and whether to utilize extra BDA. The results are shown in Table 5. Setting $N_F$ to 0.5 means half of the original frames are augmented by one pasted frame while the others are not augmented. As shown in Table 5, the BEV-Paste is not sensitive to the number of pasted frames. Considering that pasting too many frames will cost more time on training detection head because the ground truth objects are increased, setting $N_F$ as 1 is enough. Besides, extra BDA can effectively alleviate data duplication and further improve detection performance. The cooperation of these two points contributes to the performance improvement of 1.4% mAP and 1.5% NDS, confirming the effectiveness of BEV-Paste.

4.3.4 Multi-Scale Cross-Task Head

The MSCT head contains the Multi-Task Distillation (MTD) module and the Dual Supervision (DS) on prediction from Multi-Scale (MS) image features. A number of experiments are carried out to further verify the effectiveness of each module and the results are shown in Table 6. Our MTD, DS and MS modules improve NDS performance by 0.4%, 0.5% and 1.0% respectively. We attribute this improvement to the fact that the multi-task distillation module supplements cross-task information, and the dual supervision further promotes the extraction and fusion of task-specific information and cross-task information, as well as the participation of multi-scale image features.

5. Conclusion and Discussion

In this paper, we propose SA-BEV to fully utilize the semantic information of images. SA-BEVPool filters out background virtual points and generates semantic-aware BEV features. BEV-Paste then pastes the semantic-aware BEV features of two frames to enhance data diversity. MSCT head introduces multi-task learning and facilitates the optimization of semantic-aware BEV features.

Our proposed components show strong universality. SA-BEVPool and BEV-Paste can be easily embedded into most BEV-based detectors and bring stable improvements. Besides, we believe that introducing multi-task learning into the generation of semantic-aware BEV features adds a valuable perspective and will inspire future works.

Still, there are limitations in SA-BEV. The thresholds used in SA-BEVPool are manually set, making it hard to achieve optimal performance. BEV-Paste may cause incorrect object overlaps and occlusions when pasting the semantic-aware BEV feature of one frame to another. Those are what we will tackle next. We also would like to extend SA-BEV into a multi-modal detector to activate the complementarity between image and LiDAR.

Acknowledgment

This paper was supported by National Natural Science Foundation of China (Grant No. 62176017 and U20B2069) and the Fundamental Research Funds for the Central Universities.
References


