

GAFusion: Adaptive Fusing LiDAR and Camera with Multiple Guidance for 3D Object Detection

Xiaotian Li¹ Baojie Fan^{1*} Jiandong Tian² Huijie Fan²

¹Nanjing University of Posts and Telecommunications

²Shenyang Institute of Automation Chinese Academy of Science

f xiaotianli981, jobfbj

g@gmail.com

f tianjd, fanhuijie

g@sia.cn

Abstract

Recent years have witnessed the remarkable progress of 3D multi-modality object detection methods based on the Bird's-Eye-View (BEV) perspective. However, most of them overlook the complementary interaction and guidance between LiDAR and camera. In this work, we propose a novel multi-modality 3D objection detection method, named GAFusion, with LiDAR-guided global interaction and adaptive fusion. Specifically, we introduce sparse depth guidance (SDG) and LiDAR occupancy guidance (LOG) to generate 3D features with sufficient depth information. In the following, LiDAR-guided adaptive fusion transformer (LGAFT) is developed to adaptively enhance the interaction of different modal BEV features from a global perspective. Meanwhile, additional downsampling with sparse height compression and multi-scale dual-path transformer (MSDPT) are designed to enlarge the receptive fields of different modal features. Finally, a temporal fusion module is introduced to aggregate features from previous frames. GAFusion achieves state-of-the-art 3D object detection results with 73.6 AP and 74.9% NDS on the nuScenes test set.

Figure 1. Comparison between BEVFusion and the proposed GAFusion. (a) In BEVFusion, the camera stream and the LiDAR stream separately generate BEV features, which are then concatenated together. (b) In GAFusion, the camera modality BEV features are generated by multiple guidance from the LiDAR stream, and the receptive fields are enhanced by MSDPT. The BEV features are fused by LGAFT. "VT" is view transformer.

1. Introduction

3D object detection is a crucial task in autonomous driving. To cope with the complex road scenarios, multiple sensors (LiDARs or cameras) are usually employed for scene understanding. LiDAR can generate accurate but sparse 3D point clouds, which contains precise spatial information. Images have rich semantic and texture information, but lack depth information. Therefore, a natural operation is to extensively fuse LiDAR and camera to leverage the complementarity of multi-modality information, which can enable the autonomous driving system to achieve higher accuracy and robustness.

Recently, fusing LiDAR and camera has achieved some progress. Early methods [4, 14, 35, 45] achieve LiDAR-camera fusion by projecting 3D LiDAR point clouds (or region proposals) onto 2D images. But these methods overlook the information gap between the two modalities. Recent works [1, 3, 13, 22, 26, 39, 40, 42] adopt different query generation strategies or create a unified Bird's-Eye-View (BEV) [28] intermediate feature to fuse multi-modality features. For instance, TransFusion [1] applies a two-stage pipeline to fuse the camera and LiDAR features, but its performance relies on the query initialization strategy. BEVFusion [22, 26] explores a unified representation for BEV features through view transformation, which not only preserves the spatial information of sparse LiDAR point clouds, but also lifts the 2D images to the 3D features, effectively maintaining the consistency between the two modalities. However, the camera modality still struggles

* Equal Contribution.

† Corresponding Author.

with geometric perception information, which limits the complementarity between LiDAR and camera. As shown in Fig. 1(a), there is no interaction between both modalities.

To tackle the above challenges, we propose an effective 3D multi-modality object detection method, named GAFusion. Within it, a LiDAR guidance module is developed, which consists of sparse depth guidance (SDG) and LiDAR occupancy guidance (LOG). SDG combines the sparse depth generated by LiDAR point clouds and the camera features to produce depth-aware features, enhancing the sensitivity of camera features to depth information. Inspired by the occupancy task [12], LOG guides a 3D feature volume generated by view transformation with occupancy features, and focuses on the targets in the 3D feature volume, thus providing more valuable information for fusion. Then, we construct a multi-scale dual-path transformer (MSDPT) to expand the receptive fields of the 3D feature volume. With the above designs, the camera modality has sufficient semantic features and more accurate depth distribution. In the following, to obtain abundant features in the LiDAR modality, we perform additional downsampling on the LiDAR point clouds and use sparse depth compression to aggregate features from different scales. This operation can provide larger receptive fields with less computation and memory consumption. Moreover, a LiDAR-guided adaptive fusion transformer (LGAFT) module is proposed to effectively fuse BEV features generated by LiDAR point clouds and images. In this module, the LiDAR BEV features adaptively guide the camera BEV features to strengthen the cross-modality interaction from global scope.

All of the above operations are evaluated on single-frame raw data. In order to further explore the target correlation and motion consistency among multiple successive frames, a temporal fusion module is designed. To be specific, we store the BEV features of different frames in memory buffer, which is used to fuse the features of the previous frame and the current frame.

Our contributions are summarized as follows:

- 1) We propose GAFusion, a novel 3D object detection method that leverages LiDAR guidance to compensate for depth distribution of the camera features, and provides sufficient spatial information for the camera features.
- 2) We design LiDAR-guided adaptive fusion transformer (LGAFT), which aims to enhance the global features interaction between the two modalities in an adaptive way, facilitating the fusion of semantic and geometric features.
- 3) We conduct extensive experiments on the nuScenes dataset to verify the effectiveness of our GAFusion. The experiments show that without using any augmentation strategies, our model achieves the state-of-the-art performance of 72.1% mAP and 73.5% NDS.

2. Related Work

2.1. Single-modality 3D Object Detection

Single-modality 3D object detection, mainly including LiDAR-based 3D object detection and camera-based 3D object detection, has achieved remarkable progress in recent years. LiDAR-based 3D object detection aims to predict 3D object bounding boxes using the point clouds captured from LiDAR. Existing methods [15, 18, 31, 33, 34, 47] either directly predict on point clouds, or convert point clouds into voxels or pillars. PointNet [31] is the first framework that processes point clouds in an end-to-end manner, by taking unordered point cloud sets as direct inputs and preserving the spatial structure of point clouds. VoxelNet [47] discretizes point clouds into voxels, and uses dense convolution to obtain BEV features. Camera-based 3D object detection, which can be divided into two categories: image-view-based and BEV-based. DETR3D [38] and PETR [23] introduce transformer into the framework, wherein the former aggregates 2D features into 3D Query, and the latter embeds coordinate information into 2D features. They both use transformer to implicitly transform the image features to 3D space. BEVDet [11] and BEVDepth [16] predict the depth distribution to lift the image features to a 3D frustum meshgrid. The semantic or spatial information provided by a single-modality is still limited, despite the impressive performance achieved by the aforementioned detection tasks.

2.2. Multi-modality 3D Object Detection

Multi-sensor fusion has gained great attention in 3D detection due to its superior performance. Previous works [4, 14, 35, 45] fuse 3D point cloud features and 2D image features by projecting the former onto the latter. MV3D [4] associates 3D proposals with 2D RoI features and converts 3D representation into 2D pseudo images, enabling the network to leverage 2D convolutions for geometric refinement. PointPainting [36] enriches point clouds with semantic labels from images. However, the above methods fail to fully exploit the dense semantic information in images. Recently, BEVFusion [22, 26] are proposed to fuse LiDAR features and camera features in BEV space and apply a lift-splat-shoot (LSS) [30] operation to project image features, resulting in semantic-rich features. Before fusing BEV features from two modalities, LiDAR point clouds and images do not interact at all. From another perspective, CMT [40] proposes a novel end-to-end transformer-based 3D object detection framework, which implicitly encodes 3D point clouds into multi-modality tokens. Inspired by the above works, we propose the global interaction and adaptive BEV fusion that achieves significant performance improvement while maintaining the simplicity of the framework.

Figure 2. The overall architecture of GAFusion. The multi-view images and point clouds are fed into the corresponding backbone networks to obtain multi-scale LiDAR features and camera features. For LiDAR guidance, we propose sparse depth guidance (SDG) and LiDAR occupancy guidance (LOG) to guide the 2D camera features by adopting the raw point clouds and LiDAR BEV features, respectively. In addition, we use multi-scale dual-path transformer (MSDPT) to enlarge the receptive fields. Then, LiDAR-guided adaptive fusion transformer (LGAFT) further fuses the two modalities' BEV features. A temporal fusion module is introduced to aggregate the previous frame's BEV features, and finally feeds these BEV features into an encoder and a detection head.

2.3. Occupancy Task

Recently, 3D occupancy prediction (Occ) [12] has been proposed as a novel 3D detection task. Based on FB-BEV [21], FB-OCC [20] emphasizes the importance of model scale and pre-training. OccDepth [29] leverages the implicit depth information from depth images (or RGBD images) to help recover the 3D geometry. VoxFormer [17] adopts a two-stage design and generates a set of sparse visible and occupied voxel queries from depth estimation. OpenOccu-pancy [37] is the first omnidirectional semantic occupancy perception benchmark. We notice that 3D occupancy prediction aims to estimate the occupancy state and semantic label of each voxel in the scene from multi-view images, and it can provide more fine-grained and comprehensive 3D perception capabilities.

2.4. Temporal Fusion

Temporal fusion adopts multiple frames of images or point clouds to improve the performance of 3D object detection, as it can enhance the perception system's understanding and prediction of dynamic scenes. 3D-VID [46] employs a bidirectional recurrent neural network (Bi-RNN) [32] to model the temporal sequences of multiple point clouds, capturing the motion information and state changes of the targets. BEVDet4D [10] fuses BEV features from different time sequences by coordinate alignment. BEVformer [19] is a transformer-based 3D object detection model that uses images for spatial-temporal information fusion. After all, temporal fusion is an effective technique that enhances the continuity and relevance among different frames, and it can

utilize the information from multiple frames to enrich the feature representation of each frame.

3. Method

The overall architecture of GAFusion is illustrated in Fig. 2. We feed the LiDAR point clouds and the corresponding multi-view images into the backbone to extract dual-stream features. The LiDAR stream first uses additional down-sampling and sparse depth compression to obtain BEV features (Sec. 3.1). The design of LiDAR guidance, which includes sparse depth guidance (SDG) and LiDAR occupancy guidance (LOG), is detailed in Sec. 3.2. After LiDAR guidance, we adopt multi-scale dual-path transformer (MSDPT) (Sec. 3.3) to enlarge the receptive fields of camera features. Then, the proposed LiDAR-guided adaptive fusion transformer (LGAFT) module is utilized to fuse different modalities of the BEV features (Sec. 3.4). We also introduce the temporal fusion module to appropriately fuse the information from the previous frame (Sec. 3.5).

3.1. LiDAR and Camera Features Extraction

In the high-level feature extraction stage, we adopt a dual-stream approach to process the LiDAR point clouds and the multi-view images separately.

For the LiDAR stream, a common method [1, 22, 26] is to use 3D sparse convolution [41] to extract single-scale features from the voxelized point clouds, which has a weak feature representation with limited receptive fields. Therefore, we use additional downsampling layers to compensate for this deficiency. The common sparse convolution features have strides of 1, 2, 4, 8, and the output sparse features

Figure 3. Additional downsampling and sparse height compression. This operation enlarges the receptive fields of the features and reduces the computational cost.

are named F_1, F_2, F_3, F_4 respectively. We adopt two additional downsampling layers with strides of 16, 32 to obtain the F_5, F_6 features. Finally, to effectively combine different scales of F_4, F_5, F_6 features and maintain geometric and positional information, we use sparse depth compression to process different scales of the features. Specifically, we first align the spatial resolutions of F_5, F_6 with F_4 . For stage i , F_i is a set of individual features, P_i is a position in 3D space, with the coordinate (x_p, y_p, z_p) . In addition, we design a BEV grid of size (x_p, y_p) that only contains P_c , which aggregates the sparse features of different scales at the same height, and forms a rich BEV feature. The whole process is shown in Fig. 3. Sparse features F_s and their positions P_c are obtained as follows:

$$\begin{aligned} F_c &= F_4 \cup (F_5 \cup F_6); \\ P_6^0 &= f(x_p - 2^2; y_p - 2^2; z_p - 2^2) \mid p \in P_6 g \\ P_5^0 &= f(x_p - 2^1; y_p - 2^1; z_p - 2^1) \mid p \in P_5 g \\ P_c &= P_4 \cup (P_5^0 \cup P_6^0); \end{aligned} \quad (1)$$

For the camera stream, following the previous works [22, 26], we input multi-view images into backbone to obtain 2D image features $F_c \in \mathbb{R}^{N_c \times C \times H \times W}$ with sufficient semantic information, where N_c, C, H, W denote the number of cameras, feature size, image height and image width respectively.

3.2. LiDAR Guidance

To integrate the camera high-level features into a unified BEV space, a view transformation is required, which first needs to project the 2D image features into 3D space. During this process, it is often difficult to estimate the depth distribution accurately, resulting in the loss of a lot of useful information in the BEV feature generated by the camera transformer (MSDPT). Dual-path transformer (DPT) con-stream. To obtain a reliable depth distribution, our proposed LiDAR guidance consists of two parts: sparse depth guidance (SDG) and LiDAR occupancy guidance (LOG). They

Figure 4. The architecture of sparse depth guidance (SDG) and LiDAR occupancy guidance (LOG). These two modules guide the 2D camera features to generate 3D features that contain sufficient semantic information and accurate depth information.

can help the image features better capture accurate geometric and depth information.

Sparse Depth Guidance As shown in Fig. 4, SDG first projects each point of the input LiDAR point clouds into multi-view images, and obtains sparse multi-view depth maps. Then, they are fed into a shared encoder to extract depth features, which are concatenated with image features to form the depth-aware camera features. They are used as the input of view transformation, and finally voxel pooling [9] is employed to generate the image 3D feature volume, which is denoted as $F_c^0 \in \mathbb{R}^{C \times Z \times H \times W}$. SDG can effectively incorporate LiDAR depth information and generate more accurate and reliable depth.

LiDAR Occupancy Guidance Due to the sparsity and measurement noises of LiDAR point clouds, the depth information of some pixels is inaccurate. Therefore, LOG is proposed to address the aforementioned drawbacks, as shown in Fig. 4. Specifically, we first map LiDAR BEV features to 3D space to obtain the 3D features, and then attach an occupancy prediction head that estimates occupancy states to obtain the LiDAR 3D occupancy voxel, denoted as $O_L \in \mathbb{R}^{1 \times Z \times H \times W}$. It is worth noting that the resolution of O_L is the same as that of F_c^0 . The LiDAR 3D occupancy voxel is then multiplied by F_c^0 to obtain the LiDAR occupancy-guided image 3D feature volume $F_c^{00} \in \mathbb{R}^{C \times Z \times H \times W}$ using the following equation:

$$F_c^{00} = \text{Mul}(F_c^0, O_L) \quad (2)$$

Where Mul denotes element-wise multiplication with broadcasting operation. With the above designs, the 2D camera features contain sufficient semantic information and accurate depth information, which provide an excellent reference for subsequent module interactions.

3.3. Multi-Scale Dual-Path Transformer

To effectively aggregate semantic information and enlarge the receptive fields, we introduce multi-scale dual-path transformer (DPT). Dual-path transformer (DPT) consists of a local path and a global path, which uses 3D convolution to perform downsampling to obtain features of different scales. The detailed structures of DPT are shown

Figure 5. The schema of dual-path transformer (DPT), which effectively aggregates semantic information and expands the receptive fields of the camera features.

in Fig. 5. The local path is mainly used to extract fine-grained semantic structures. Since the height direction has less variation in 3D object detection, the local path only slices and processes the 3D feature volume extracted from the multi-view images in parallel along the horizontal direction. The global path attempts to acquire the semantic layout of the scene accurately. It first obtains BEV features by average pooling along the height dimension, and then interacts with the basic information of the BEV features. To improve computational efficiency, they both use windowed self-attention [25], and share weights. Finally, the 3D feature volume from the local path merges the sufficient semantic features from the global path. The dual-path outputs are $F_{\text{local}} \in \mathbb{R}^{C \times Y \times Z}$ and $F_{\text{global}} \in \mathbb{R}^{C \times Y}$, the combined output F_{out} is computed as:

$$F_{\text{out}} = F_{\text{local}} + (W_H F_{\text{local}}) \text{unsqueeze}(F_{\text{global}}; 1) \quad (3)$$

where W_H refers to the aggregation weights along the height dimension generated by the FFN, $\sigma(\cdot)$ is the sigmoid function, and “unsqueeze” expands the global 2D features along the height.

3.4. LiDAR-guided Adaptive Fusion Transformer

Recent works [22, 26] simply concatenate different modalities of BEV features to obtain a shared BEV representation, which does not consider the information interaction and global spatial relevance among different modalities. To this end, LGAFT is developed to adaptively enhance the interaction of LiDAR BEV features F_{LB} and camera BEV features F_{CB} from a global perspective. The detailed architecture is illustrated in Fig. 6. We use 1 convolution to expand F_{LB} and F_{CB} to appropriate channels, and concatenate the expanded BEV features F_{LB}^0 and F_{CB}^0 to obtain feature weights W_F from a sigmoid function. Then, we adopt W_F to fuse the LiDAR and camera BEV features adaptively, and the fused features are denoted as F_{BEV} . The weights W_F can be expressed as:

Figure 6. The overview of LiDAR-guided adaptive fusion transformer (LGAFT). LGAFT adaptively enhances the interaction between LiDAR and camera BEV features from a global perspective.

$$\begin{aligned} F_{\text{LB}}^0 &= \text{conv}_1(F_{\text{LB}}) \\ F_{\text{CB}}^0 &= \text{conv}_1(F_{\text{CB}}) \\ W_F &= \text{Conca}(F_{\text{LB}}^0; F_{\text{CB}}^0) \end{aligned} \quad (4)$$

Where “Conca” denotes the concatenate operation. To reduce the computation cost, we do not use the multi-head attention module in transformer structure. Specifically, we adopt F_a as the query of the cross-attention module. The adaptive camera features are regarded as the keys and values to avoid the gradient explosion convergence problem. Therefore, the final fused features F_{BEV} can be presented as:

$$\begin{aligned} Q &= \text{Conca}((1 - W_F)F_{\text{LB}}^0; W_F(F_{\text{CB}}^0 + P))W_Q \\ K &= W_F(F_{\text{CB}}^0 + P)W_K \\ V &= (F_{\text{CB}}^0 + P)W_V \\ F_{\text{BEV}} &= \text{MLP}(\text{LN}(\text{Softmax}(\frac{QK^T}{C})V)) \end{aligned} \quad (5)$$

Where Q , K , and V denote the query, key, and value. W_Q , W_K and W_V are learnable parameters, P stands for the learnable position embedding, LN means layer normalization and MLP is the multi-layer perception block.

3.5. Temporal Fusion Module

Temporal information is crucial for the visual system to understand the surrounding environment. Temporal information can better help detect the motion states of the objects and occluded objects. We follow the fusion scheme of BEVDet4D [10] and store the BEV features of historical frames in a memory buffer, and fuse the BEV features of the previous frame at each time. The detailed operation can be found in [10]. Finally, we feed the fused BEV features into the BEV encoder and detection head to obtain the final detection results.

4. Experiments

4.1. Dataset and Metrics

Similar to previous works [1, 22, 26], we conduct extensive synthetic experiments on the nuScenes dataset. The

| Method | Modality | mAP | NDS | Car | Truck | C.V. | Trailer | Bus | Barrier | Motor. | Bike | Ped. | T.C. |
|----------------------|----------|------|------|------|-------|------|---------|------|---------|--------|------|------|------|
| PointPillars [15] | L | 30.5 | 45.3 | 68.4 | 23.0 | 4.1 | 23.4 | 28.2 | 38.9 | 27.4 | 1.1 | 59.7 | 30.8 |
| CenterPoint [44] | L | 60.3 | 67.3 | 85.2 | 53.5 | 20.0 | 56.0 | 63.6 | 71.1 | 59.5 | 30.7 | 84.6 | 78.4 |
| TransFusion-L [1] | L | 65.5 | 70.2 | 86.2 | 56.7 | 28.2 | 58.8 | 66.3 | 78.2 | 68.3 | 44.2 | 86.1 | 82.0 |
| LargeKernel3D [5] | L | 65.3 | 70.5 | 85.9 | 55.3 | 26.8 | 60.2 | 66.2 | 74.3 | 72.5 | 46.6 | 85.6 | 80.0 |
| FocalFormer3D [6] | L | 68.7 | 72.6 | 87.8 | 59.4 | 37.8 | 65.7 | 73.0 | 77.8 | 77.4 | 52.4 | 90.0 | 83.4 |
| PointPainting [36] | LC | 46.4 | 58.1 | 77.9 | 35.8 | 15.8 | 37.3 | 36.2 | 60.2 | 41.5 | 24.1 | 73.3 | 62.4 |
| 3D-CVF [45] | LC | 52.7 | 62.3 | 83.0 | 45.0 | 15.9 | 49.6 | 48.8 | 65.9 | 51.2 | 30.4 | 74.2 | 62.9 |
| MVP [43] | LC | 66.4 | 70.5 | 86.8 | 58.5 | 26.1 | 57.3 | 67.4 | 74.8 | 70.0 | 49.3 | 89.1 | 85.0 |
| TransFusion [1] | LC | 68.9 | 71.7 | 87.1 | 60.0 | 33.1 | 60.8 | 68.3 | 78.1 | 73.6 | 52.9 | 88.4 | 86.7 |
| AutoAlignV2 [7] | LC | 68.4 | 72.4 | 87.0 | 59.0 | 33.1 | 59.3 | 69.3 | - | 72.9 | 52.1 | 87.6 | - |
| BEVFusion [26] | LC | 70.2 | 72.9 | 88.6 | 60.1 | 39.3 | 63.8 | 69.8 | 80.0 | 74.1 | 51.0 | 89.2 | 85.2 |
| BEVFusion [22] | LC | 71.3 | 73.3 | 88.1 | 60.9 | 34.4 | 62.1 | 69.3 | 78.2 | 72.2 | 52.2 | 89.2 | 86.7 |
| CMT [40] | LC | 70.4 | 73.0 | 87.2 | 61.5 | 37.5 | 62.8 | 72.4 | 86.9 | 79.4 | 58.3 | 86.9 | 83.1 |
| DeepInteraction [42] | LC | 70.8 | 73.4 | 87.9 | 60.2 | 37.5 | 63.8 | 70.8 | 80.4 | 75.4 | 54.5 | 91.7 | 87.2 |
| FocalFormer3D [6] | LC | 71.6 | 73.9 | 88.5 | 61.4 | 35.9 | 66.4 | 71.7 | 79.3 | 80.3 | 57.1 | 89.7 | 85.3 |
| MSMDFusion [13] | LC | 71.5 | 74.0 | 88.4 | 61.0 | 35.2 | 66.2 | 71.4 | 80.7 | 76.9 | 58.3 | 90.6 | 86.6 |
| GAFusion(ours) | LC | 73.6 | 74.9 | 89.4 | 65.3 | 42.4 | 65.8 | 73.7 | 79.2 | 80.8 | 60.2 | 92.3 | 87.0 |

Table 1. Comparison on the nuScenes test set. The models in the table are without ensemble or test-time augmentation. “L” is LiDAR, “C” is camera.

nuScenes [2] dataset is a large-scale autonomous driving benchmark, which includes 1000 scenes with images from 6 cameras with surrounding views, points from 5 Radars and 1 LiDAR. The scenes are officially split into 700/150/150 scenes for training/validation/testing. Each scene lasts for about 20 seconds, where key frames are annotated at 2Hz. Each frame of point cloud data corresponds to 6 RGB images with 360° horizontal FOV.

For the 3D detection task, we adopt the nuScenes Detection Score (NDS) and mean Average Precision (mAP) to evaluate the performance of the proposed model. In addition, the evaluation metrics of nuScenes also include True Positive (TP) metrics, namely mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE) and mean Average Attribute Error (mAAE), which assess the performance of the model from different perspectives. NDS is the weighted average of mAP and TP metrics.

4.2. Implementation Details

The developed model is implemented based on the MMDetection3D [8] framework. For the LiDAR stream, we utilize the additional downsampling operation on top of VoxelNet [47] as our backbone. For the camera stream, we adopt Swin-T [25] and FPN as the image backbone, and use the pre-trained model of Swin-T. Similar to most models, the image resolution is 448x800, and the voxel size is (0.075m, 0.075m, 0.2m). The whole training process follows the previous work [22, 26]. Firstly, a LiDAR detector is trained as 3D backbone for 20 epochs. Then, we freeze the pre-trained

| Methods | Modality | mAP | NDS |
|----------------------|----------|------|------|
| BEVFormer [19] | C | 41.6 | 51.7 |
| PETrv2 [24] | C | 45.6 | 35.0 |
| CenterPoint [44] | L | 60.3 | 67.3 |
| TransFusion-L [1] | L | 65.5 | 70.2 |
| TransFusion [1] | LC | 67.5 | 71.3 |
| BEVFusion [26] | LC | 68.5 | 71.4 |
| CMT [40] | LC | 67.9 | 70.8 |
| MSMDFusion [13] | LC | 69.3 | 72.1 |
| DeepInteraction [42] | LC | 69.9 | 72.6 |
| SparseFusion [39] | LC | 70.5 | 72.8 |
| GAFusion(ours) | LC | 72.1 | 73.5 |

Table 2. Comparison on the nuScenes val set. The models in the table are without ensemble or test-time augmentation.

LiDAR components and jointly train for another 6 epochs according to the proposed framework. During the training stage, we use AdamW [27] optimizer with an initial learning rate of 5×10^{-5} and a weight decay of 10^{-4} . GAFusion is trained on two 3090 GPUs with batch size of 4. In the inference stage, we do not use test-time augmentation (TTA) or multi-model ensemble.

4.3. Results and Comparison

As shown in Table 2 and Table 1, we report the results of GAFusion on the nuScenes validation and test sets, and compare them with other state-of-the-art models. The results show that, on the test set, GAFusion surpasses all the existing methods with 73.6% mAP and 74.9% NDS, such as MSMDFusion [13] and CMT [40]. It also achieves excel-

Figure 7. Visualization results of BEVFusion and GAFusion on the nuScenes validation set. The red circles and boxes show the detection ability of GAFusion for small and occluded objects.

lent performance on the validation set. In addition, we also provide the visualization results of GAFusion and BEVFusion to demonstrate the superiority of the proposed method, and they can be seen in Fig. 7. This is attributed to better guidance mechanisms, larger receptive fields and a more suitable fusion method.

4.4. Ablation Studies

To demonstrate the effectiveness and rationality of GAFusion, we conduct comprehensive ablation studies for each of the proposed components.

Additional downsampling and sparse height compression. To prove the validity and generalization of this module, we separately insert the developed module into TransFusion [1] and BEVFusion [26], as shown in Table 3. We do not use any augmentation strategies or multi-model ensemble during testing. The results illustrate that it can significantly improve the performance of different models. It enhances 1.7% mAP and 0.6% NDS in TransFusion, and 0.8% mAP and 0.5% NDS in BEVFusion, which indicates that it can aggregate multi-scale information.

Impacts of LiDAR guidance. To demonstrate that the contributions of LiDAR guidance indeed improve the model performance, we introduce SDG and LOG into BEVFusion [26]. Table 4 presents the impacts of different combinations of the guidance modules in BEVFusion. We observe that the model performance brings about 1.4% mAP and 0.8% NDS gain with both SDG and LOG. When nei-

| Backbone + Sparse2Dense | TransFusion | | BEVFusion | |
|-------------------------|------------------|------------------|------------------|------------------|
| | mAP ["] | NDS ["] | mAP ["] | NDS ["] |
| Voxelnet + HC | 68.9 | 71.6 | 70.2 | 72.9 |
| VoxelNet, AD + SHC | 69.9 | 72.4 | 71.0 | 73.4 |

Table 3. Performance and generalization of additional downsampling (AD) and sparse height compression (SHC) on other common models. NDS/mAP comparison on nuScenes test set. "HC" is height compression.

ther SDG nor LOG module is used, the model scores drop significantly. It can be attributed to the lack of guidance in the camera stream, and results in unreliable depth information. Moreover, it realizes 0.3% mAP and 0.4% NDS with SDG alone and 1.3% mAP and 0.6% NDS with LOG alone. The interaction effects of LOG are more remarkable, so we conjecture that directly interacting among 3D features can provide sufficient located information. SDG and LOG play their respective roles: the former integrates sparse depth information into 2D image features, and the latter guides depth information in 3D feature volume, which enables the camera stream to obtain rich geometric information.

BEV features fusion strategy. We explore the impacts of different fusion methods, including addition, concatenation, LiDAR-guide fusion transformer (LGFT) and LGFT. As shown in Table 5, LGFT achieves a noticeable improvement over addition and concatenation, with about 0.7

Figure 8. Receptive elds of the preliminary fused BEV features from different modalities. The colored dots indicate effective receptive elds.

| | SDG | LOG | mAP | NDS |
|-----|-----|-----|-------|-------|
| (1) | | | 68.52 | 71.38 |
| (2) | X | | 69.33 | 71.76 |
| (3) | | X | 69.49 | 72.04 |
| (4) | X | X | 69.93 | 72.24 |

Table 4. Ablation study of LiDAR guidance on nuScenes val set. (1) represents the performance of the original BEVFusion model.

mAP and 0.4% NDS. LGAFT further enhances 0.1% mAP and 0.1% NDS against LGFT due to the addition of adaptive mechanism. It presents that enhancing the interaction between LiDAR and camera BEV features from a global scope and the adaptive mechanism can sufficiently improve global spatial relevance.

Effects of MSDPT. We illustrate the related results to prove that MSDPT can effectively enlarge the receptive elds of camera features and aggregate semantic information. In Table 6, (1)-(5) are using output features to fuse the features of different scales. Without MSDPT, the model performance drops by about 0.5% mAP and 0.4% NDS. Different scales of features also affect the model accuracy, which is due to the fact that the multi-scale operation can enlarge the receptive elds of camera features. However, redundant scales also cause too much computation and the performance enhancement is not obvious. Therefore, we select 3 scales to combine the different scale features for the balance of performance and computation.

Larger receptive elds. As shown in Fig. 8, (a) and (b) illustrate the effective receptive elds of the fused features from the camera and LiDAR BEV features by BEVFusion [26] and GAFusion, respectively. We observe that GAFusion achieves larger effective receptive elds than BEVFusion. This is attributed to the additional downsampling and MSDPT modules, which indicate that multi-scale features can provide more contextual information. Besides, the global and local interaction of LGAFT contributes to enlarging the feature receptive elds to some extent.

Temporal fusion. In Table 5, the temporal fusion improves about 0.2% mAP and 0.1% NDS. We integrate two

| | BEV Fusion | Tem | mAP | NDS |
|-----|------------|-----|-------|-------|
| (1) | ADD. | X | 70.85 | 72.82 |
| (2) | Cat. | X | 70.92 | 72.88 |
| (3) | LGFT | X | 71.63 | 73.32 |
| (4) | LGAFT | X | 71.79 | 73.43 |
| | | | 72.08 | 73.53 |

Table 5. Ablation study of BEV fusion strategy and temporal fusion on nuScenes val set. "Tem" is temporal fusion.

| | C1 | C2 | C3 | C4 | mAP | NDS |
|-----|----|----|----|----|-------|-------|
| (1) | | | | | 71.60 | 73.11 |
| (2) | X | | | | 71.92 | 73.39 |
| (3) | | X | | | 72.01 | 73.45 |
| (4) | | | X | | 72.08 | 73.53 |
| (5) | | | | X | 72.07 | 73.54 |

Table 6. Ablation study of MSDPT on nuScenes val set. C1-C4 denote the number of 3D convolution layers (1-4) applied to the 3D feature volume, respectively.

frames of the BEV features, and the approach enables partial features alignment between adjacent frames, which leads to a marginal performance improvement. For multiple frames, it can attain higher enhancement, which is our future work.

5. Conclusion

We propose GAFusion, a more effective 3D object detection method in BEV representation, which is equipped with excellent guidance and fusion mechanisms. Additional downsampling and MSDPT are developed to enlarge the receptive elds of different modal features. Then SDG and LOG are employed to transform the 2D camera features into 3D features with sufficient located and spatial information. Afterward, we propose LGAFT to facilitate the fusion of LiDAR and camera BEV features. Finally, a temporal fusion module is adopted to aggregate features from different frames. Extensive experiments demonstrate the effectiveness and generality of our developed modules and GAFusion achieves state-of-the-art performances on the nuScenes dataset. We hope that the proposed components of GAFusion could provide more insights for subsequent research in this field.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China under Grant U2013210 and in part by the Guangxi Key Laboratory of Machine Vision and Intelligent Control under Grant 2022B09.

References

- [1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 1090–1099, 2022. 1, 3, 5, 6, 7
- [2] 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 multi-modal dataset for autonomous driving. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pages 11621–11631, 2020. 6
- [3] Hongxiang Cai, Zeyuan Zhang, Zhenyu Zhou, Ziyin Li, Wenbo Ding, and Jiuha Zhao. Bevfusion4d: Learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation. *arXiv preprint arXiv:2303.17099* 2023. 1
- [4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* pages 1907–1915, 2017. 1, 2
- [5] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Largekernel3d: Scaling up kernels in 3d sparse cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 13488–13498, 2023. 6
- [6] Yilun Chen, Zhiding Yu, Yukang Chen, Shiyi Lan, Anima Anandkumar, Jiaya Jia, and Jose M Alvarez. Focalformer3d: Focusing on hard instance for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* pages 8394–8405, 2023. 6
- [7] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection. *arXiv preprint arXiv:2207.10316* 2022. 6
- [8] MMDetection3D Contributors. MMDetection3D: Open-MMLab next-generation platform for general 3D object detection. <https://github.com/open-mmlab/mmdetection3d>, 2020. 6
- [9] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence* pages 1201–1209, 2021. 4
- [10] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054* 2022. 3, 5
- [11] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790* 2021. 2
- [12] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 9223–9232, 2023. 2, 3
- [13] Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 21643–21652, 2023. 1, 6
- [14] Jason Ku, Melissa Mozi an, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* pages 1–8. IEEE, 2018. 1, 2
- [15] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pages 12697–12705, 2019. 2, 6
- [16] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* pages 1477–1485, 2023. 2
- [17] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 9087–9098, 2023. 3
- [18] Zhichao Li, Feng Wang, and Naiyan Wang. Lidar r-cnn: An efficient and universal 3d object detector. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 7546–7555, 2021. 2
- [19] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 1–18. Springer, 2022. 3, 6
- [20] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492* 2023. 3
- [21] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. *Proceedings of the IEEE/CVF International Conference on Computer Vision* pages 6919–6928, 2023. 3
- [22] 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. *Advances in Neural Information Processing Systems* 35:10421–10434, 2022. 1, 2, 3, 4, 5, 6
- [23] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *Computer Vision—ECCV 2022: 17th Eu-*

- ropean Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV, pages 531–548. Springer, 2022. 2
- [24] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A uni ed framework for 3d perception from multi-camera images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 6
- [25] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5, 6
- [26] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with uni ed bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 1, 2, 3, 4, 5, 6, 7, 8
- [27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [28] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, Dinesh Manocha, and Xinge Zhu. Vision-centric bev perception: A survey. *arXiv preprint arXiv:2208.02797*, 2022. 1
- [29] Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*, 2023. 3
- [30] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2
- [31] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2
- [32] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 3
- [33] 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. 2
- [34] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *IJCV*, pages 1–21, 2022. 2
- [35] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvxnet: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. IEEE, 2019. 1, 2
- [36] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020. 2, 6
- [37] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17850–17859, 2023. 3
- [38] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2
- [39] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17591–17602, 2023. 1, 6
- [40] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18268–18278, 2023. 1, 2, 6
- [41] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 3
- [42] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. In *NeurIPS*, 2022. 1, 6
- [43] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbühl. Multi-modal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021. 6
- [44] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbühl. Center-based 3d object detection and tracking. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 6
- [45] Jin Hyeok Yoo, Yeocheol Kim, Ji Song Kim, and J. Choi. 3D-CVF: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *ECCV*, 2020. 1, 2, 6
- [46] Zhenyu Zhai, Qiantong Wang, Zongxu Pan, Zhentong Gao, and Wenlong Hu. Multi-frame point cloud feature fusion based on attention mechanisms for 3d object detection. *Sensors*, 22(19):7473, 2022. 3
- [47] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 2, 6