CHALMERS
UNIVERSITY OF TECHNOLOGY

SMAB: Simple Multimodal Attention for Effective BEV Fusion

SMAB: Simple Multimodal Attention for Effective BEV Fusion

Amer Mustajbasic[1,3], Shuangshuang Chen[2], Erik Stenborg[3], and Selpi[1]

***Abstract*— Sensor fusion plays a crucial role in accurate and robust environment perception for autonomous driving. Recent works utilize Bird's-Eye-View (BEV) grid as a 3D representation, however, only using a partial set of multimodal signals. This paper introduces Simple-Multimodal-Attention-BEV (SMAB), a novel and simple approach to multimodal sensor fusion in BEV perception. We propose an attention mechanism called BEV feature aggregation that effectively enhances BEV feature representations. It integrates bilinearly interpolated semantic data from cameras with rasterized distance information from radars and/or lidars, and facilitates training with full-modality data or partial-modality data without modification of the method. In addition to the simplicity of the design, we demonstrate that using all sensor modalities improves segmentation accuracy. Meanwhile, SMAB is resilient to sporadic sensor signal loss, which enhances the robustness of the perception system. The proposed method outperforms state-of-the-art methods while simplifying the model.**

I. Introduction

Perception systems are critical for autonomous driving, enabling vehicle for decision-making and control through their interpretation of surroundings. Modern vehicles rely on multimodal sensor suites, integrating lidars, radars and multi-view cameras. Although the comprehensive array of sensors elevates computational demands, it ensures a robust and accurate perception stack for diverse environments. Cameras are vital for providing rich semantic information but they lack depth information. To mitigate this limitation and enhance redundancy, vehicles often incorporate radars and lidars: lidars offer precise depth measurements, while radars excel in challenging weather conditions such as fog or rain. Unlike cameras, these sensors generate sparse point cloud data, necessitating effective sensor fusion to integrate the distance data with the semantic information derived from cameras.

Recent advancements in sensor fusion techniques have focused on unifying data from multiple modalities into a coherent representation, with most methods utilizing a Bird's-Eye-View (BEV) grid to represent the 3D environment around the vehicle in a flattened format [1], [2], [3], [4]. Previous works, e.g., BEVFusion [5], SimpleBEV [3], and CRN [4], have explored multimodal signal fusion in the BEV representations and showed improvements in some vision tasks. However, it is not clear where the improvements come from, due to the use of different and in many cases complex architectures and training methods at the same time. Moreover, these methods utilize only a partial set of multimodal signals, leaving our understanding on the impact and redundancy of using all multimodal signals during inference unclear.

Inspired by SimpleBEV [3], a streamlined fusion architecture with fewer parameters and a straightforward setup offers clear advantages for deployment in real-time autonomous driving, particularly under constraints on computational resources and critical inference times. This simplicity enhances scalability by reducing the need for complex layers and extensive learning parameters, making it easy to incorporate additional sensors with minimal modifications of method.

The main contribution of this paper is the introduction of Simple-Multimodal-Attention-BEV (SMAB), a simple, yet competitive method to handle robust multimodal sensor fusion for images, radar and/or lidar (see Figure 1). The core element of our method is the BEV feature aggregation (BFA), a simple attention mechanism to enhance the BEV feature representations. We demonstrate that:

- SMAB facilitates training with full-modality or partial-modality data resulting in improved segmentation accuracy for each added modality in comparison to methods with similar model complexity,
- BFA extracts an efficient representation especially when the density of radar and/or lidar points is low,
- SMAB handles sporadic multimodal signal loss with only a minor reduction in accuracy.
- SMAB enables scalability without modifying fusion method, therefore, additional sensor modalities can be integrated with minimal increase in model complexity.

II. Related work

BEV perception has become an important approach in computer vision tasks and various autonomous driving functions, as it enables more effective sensor fusion through a unified spatial representation.

Camera-based BEV perception methods, such as [1], [2], [6], handle the lack of external distance signals by focusing on various lifting strategies, to project 2D image features into 3D space, thus allowing for more comprehensive representations in space. The Lift, Splat, Shoot (LSS) [2] parameterizes the depth distribution along projected image rays, scaling image features with these parameters. It is followed by voxel pooling and the cumulative sum trick to aggregate image features into a BEV representation. BEVFormer [6] introduces attention-based lifting that uses deformable cross-attention [7] to learn sampling offsets and attention matrices to align the BEV grid with spatially structured image features. SimpleBEV [3] on the other hand, takes a non-parametric approach by bilinearly pulling features from the image feature space into a voxel grid, compressing the height

[1]Chalmers University of Technology and University of Gothenburg {amermus|selpi}@chalmers.se
[2]Volvo Car Corporation chen.shuangshuang@volvocars.com
[3]Zenseact erik.stenborg@zenseact.com

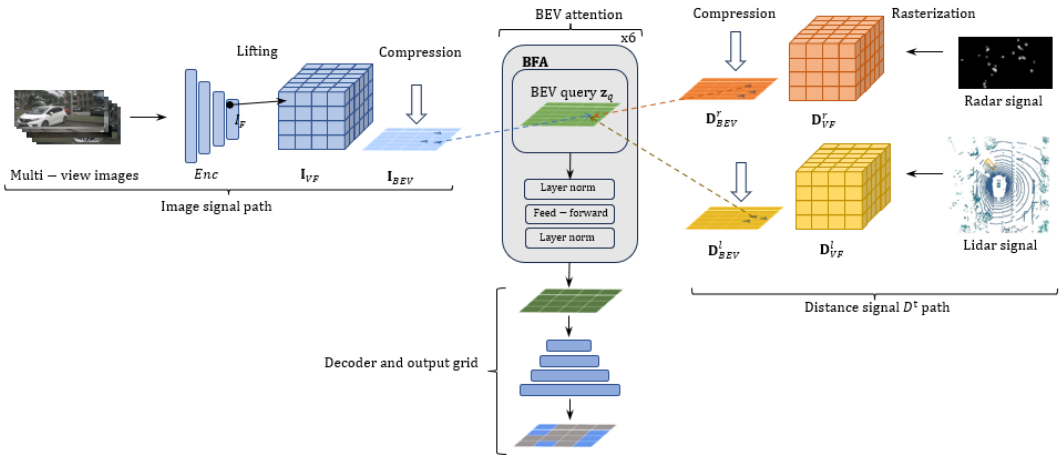

Fig. 1: SMAB architecture. The imaging signal is merged with distance signals from radar and/or lidar to a common voxel grid representation using learnable BEV query and BFA attention mechanism.

dimension into the BEV grid, and implicitly estimating image feature projections into the BEV space.

In contrast, lidar-based perception leverages accurate distance information but it is limited by signal sparsity and a narrow field-of-view. VoxelNet [8] addresses this by applying PointNets [9] and 3D convolutions to voxelized point clouds, achieving strong detection performance at the cost of high computation. SECOND [10] builds on VoxelNet [8], improves inference speed, though it still relies on expensive 3D convolutions. PointPillars [11] further improves efficiency by encoding point cloud features within vertical columns (pillars) using PointNets [9], reducing complexity while maintaining accuracy.

Radar-based perception has also seen significant advancements. NVRadarNet [12] introduces deep convolutional neural networks to process radar signals to estimate free space in the BEV. Inverse Sensor Model [13] is a deep neural network trained in a self-supervised manner to convert raw radar scans into grid maps of occupancy probabilities.

Recent research develops fusion techniques that combine camera data with lidar and/or radar signals to improve the performance and robustness of BEV perception, utilizing the rich semantic information from camera images along with the precise distance measurements from lidar and radar. For instance, Frustum PointNets [14] uses PointNets [14] to integrate both lidar and camera signals, and employs a multi-stage process where the point cloud is segmented and classified based on detection proposals in the image space. BEVFusion [5] employs late fusion by concatenating LSS-lifted image features [2] with lidar features into a shared BEV representation, which is then processed by a convolutional BEV encoder. CRN [4], on the other hand, focuses on radar-camera perception and introduces radar-assisted view transformation (RVT), scaling LSS-lifted image features with radar occupancy maps, followed by cross-attention on image and radar features.

SimpleBEV [3] explores the fusion with image-radar or image-lidar combinations by rasterizing radar or lidar signals in a 3D voxel grid, concatenating them with lifted image features, and performing convolutional BEV compression to reduce the height dimension. Convolutional operations, however, have a limited receptive field, which restricts their ability to capture broader contextual information across modalities. In contrast, DeepInteraction [15] employs a modality-interaction fusion strategy, yet its scalability is limited, as adding an additional modality increases attention complexity quadratically. BEVFusion4D [16] introduces a fusion approach using a Lidar-Guided View Transformer, however, it relies on the availability of all modalities making the fusion process prone to failure if any signal drops.

Leveraging all sensor modalities like camera, radar, and lidar, takes advantage of each sensor's unique strengths and enhances the robustness of a perception system such that the reduced visibility of a camera in adverse conditions like rainy weather can be compensated for by lidar. However, previous works typically rely on only a subset of these modalities during inference, most often combining the camera with either radar or lidar, and using the camera as the primary input. For instance, CRN [4] uses lidar during training to assist depth estimation but excludes it during inference, limiting the reliability of multimodal fusion in real-world scenarios.

In contrast, SMAB introduces a streamlined architecture that incorporates all sensor modalities during inference and can seamlessly scale to include additional sensors without altering the core framework, making it a flexible and effective solution for various perception tasks.

III. Method

We introduce Simple-Multimodal-Attention-BEV (SMAB), a novel fusion method for efficiently combining image, radar, and lidar data in a simple, scalable architecture. SMAB enhances robustness to signal loss and improves inference accuracy. Unlike DeepInteraction [15], which uses 260 times more voxels, SMAB optimizes memory and computation with a fixed-size voxel grid and a constant

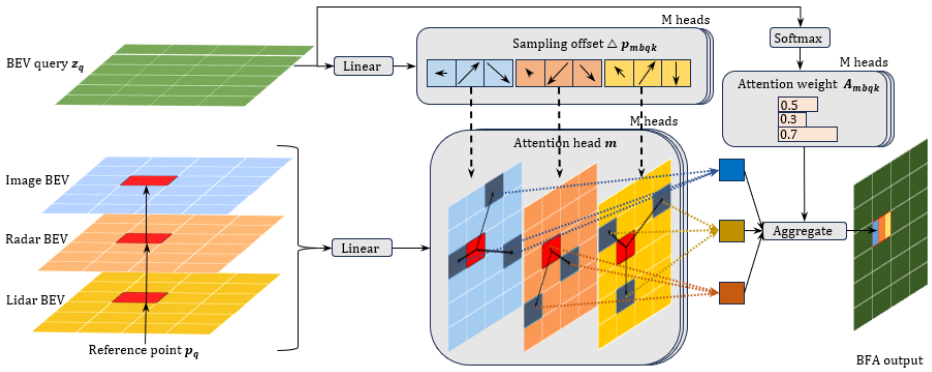

Fig. 2: BFA (BEV feature aggregator) architecture. Compressed multimodal BEV signals are sampled using learned reference point offsets $\Delta\mathbf{p}_{mbqk}$ and aggregated for every attention head with learned attention weights $\mathbf{A}_{mbqk}$. See Section III-C for detailed explanation.

BEV query array, making it more efficient for real-time applications. The fusion method is summarized in Figure 1.

A. Imaging signal path and lifting

Our input signals on the imaging path are multi-view images $\mathbf{I} \in \mathbb{R}^{N\times C\times H\times W}$, where N, C, H, W stand for the number of views, channels, height and width respectively. They are passed through the image encoder $Enc(\cdot)$, which outputs a spatial feature representation of the images by $\mathbf{I}_F = Enc(\mathbf{I}) \in \mathbb{R}^{N\times C_F\times H_F\times W_F}$. The image encoder can be e.g. ResNet [17], where we create spatial feature output similarly to the encoder in [3] where fine-grained features are combined with the coarser ones. In order to have better spatial resolution in feature space, we increase the spatial output size so that the third layer of the encoder, upsamples features and concatenates with second layer features, and then the features are upsampled and concatenated with the first layer features. To achieve transposed convolution, additional convolutional layers are added with instance normalization [18] and ReLU activation [19] after every up-sample and concatenation step. With this design, the resulting spatial feature of the encoder is $1/16$ of the original image resolution where $H_F = H/4$ and $W_F = W/4$. Channel width is $C_F = 128$.

After encoder, image features are sampled to voxel grid. We use the lifting strategy on image features proposed in SimpleBEV [3]. Every voxel along the ray is assigned the same sampled image feature, but only one destination voxel is valid along the ray, so it introduces a noise that is suppressed implicitly during task learning. Using pre-defined homogeneous voxel grid coordinates $\mathbf{V}_R \in \mathbb{R}^{N\times Z_R\times Y_R\times X_R\times 4}$ in the reference camera R coordinate system where Z_R, Y_R, X_R are grid dimensions in the reference frame, image features $\mathbf{I}_F$ are bilinearly sampled from N camera views using extrinsic and intrinsic information in following steps:

1) The homogeneous voxel grid coordinates $\mathbf{V}_N \in \mathbb{R}^{N\times Z_N\times Y_N\times X_N\times 4}$ in camera view N's coordinate system are obtained by applying the transformation matrices $\mathbf{T}^{(n)}_{N\leftarrow R} \in \mathbb{R}^{4\times 4}$ from reference camera R to other camera views N to the corresponding voxels $\mathbf{V}^{(n,z,y,x)}_R \in \mathbb{R}^4$:

$$\mathbf{V}^{(n,z,y,x)}_N = \mathbf{T}^{(n)}_{N\leftarrow R}\mathbf{V}^{(n,z,y,x)}_R; \tag{1}$$

2) The pixel coordinates (u_N, v_N) in the image features plane of N camera views are obtained by projecting the 3D coordinates $\mathbf{V}^{(n,z,y,x)}_N$, obtained in Equation 1 using the intrinsic matrix $\mathbf{K}^{(n)}_N \in \mathbb{R}^{4\times 4}$ for the n-th camera view:

$$\begin{pmatrix} u^{(n,z,y,x)}_N \\ v^{(n,z,y,x)}_N \\ 1 \end{pmatrix} = \mathscr{P}\left(\mathbf{K}^{(n)}_N \mathbf{V}^{(n,z,y,x)}_N\right), \tag{2}$$

where $\mathscr{P}(\cdot)$ is an operator that scales on z coordinate and removes the last dimension.

3) Image features voxel grid $\mathbf{I}_{VF} \in \mathbb{R}^{N\times C_F\times Z_R\times Y_R\times X_R}$ is obtained by bilinear sampling and invalid projection filtering of the image features, indexed by pixel coordinates from the Equation 2:

$$\mathbf{I}_{VF} = \mathbf{I}_F\langle(u_N, v_N, 0)\rangle \odot \mathbf{M}, \tag{3}$$

where $\langle\rangle$ is bilinear sampling operator, $\odot$ is element-wise multiplication and $\mathbf{M} \in \mathbb{R}^{N\times 1\times Z_R\times Y_R\times X_R}$ is a mask to filter out features from outside the camera frustum, constructed as:

$$\mathbf{M}(u_N, v_N, d_N) = \mathbb{I}(0 \le u_N < W) \odot \mathbb{I}(0 \le v_N < H) \odot \mathbb{I}(d_N > 0), \tag{4}$$

where $\mathbb{I}$ is an indicator function and d_N is the depth coordinate of $\mathbf{V}_N$. The generated voxel grid $\mathbf{I}_{VF}$, obtained in the Equation 3, is then averaged over dimension N and summed over dimension Y_R, resulting in the BEV grid $\mathbf{I}_{BEV} \in \mathbb{R}^{C_F\times Z_R\times X_R}$.

B. Distance signal path

On the distance signal path, we use general notation for both radar and lidar as $\mathbf{D}^o \in \mathbb{R}^{S\times E\times C_D}$, where o is signal type, S is sequence length, E is the number of distance signal signatures and C_D is the number of distance signal features.

The distance signal $\mathbf{D}^o$ is first transformed into the reference camera R coordinate system using the extrinsics transformation $\mathbf{T}^{(n)}_{R\leftarrow D^o} \in \mathbb{R}^{4\times 4}$ for each of N camera views:

$$\mathbf{D}^{o(n,e)}_R = \mathbf{T}^{(n)}_{R\leftarrow D^o}\mathbf{D}^{o(s,e)}. \tag{5}$$

Using positional information in the signal, every distance signal data point, transformed in Equation 5, is assigned with a voxel index i_R and placed to the distance feature voxel grid $\mathbf{D}^o_{VF} \in \mathbb{R}^{C_F \times Z_R \times Y_R \times X_R}$ using corresponding voxel indices.

The distance feature voxel grid $\mathbf{D}^o_{VF}$ is compressed on the Y_R dimension, using a convolutional layer with instance normalization and GELU activation [20], turning to BEV grid $\mathbf{D}^o_{BEV} \in \mathbb{R}^{C_F \times Z_R \times X_R}$.

C. Fusion mechanism

We employ attention mechanism to fuse contextual information from different BEV representations, capitalizing on their complementary strengths. For instance, image signals offer rich semantic context but lack depth information, making spatial positioning challenging, while radar and lidar signals provide precise spatial data but lack the detailed scene context. These issues require us to introduce an attention mechanism to better integrate contextual and structural information, resulting in a more comprehensive and cohesive representation.

We adopt a similar architecture to the transformer in [21], where blocks consisting of multi-head attention, feature normalization, and feed-forward operations are repeatedly stacked. We base our attention design on multi-scale deformable attention in Deformable DETR [7], but instead of using multiple scales, we leverage multimodal signal features represented in multimodal BEV grids. Our method uses a learnable BEV array as the initial query to generate sampling offsets and attention weights for cross-attending to imaging and distance signals. A new BEV array is produced at each attention block and used as the query for the next. The final BEV array is then fed into the decoder. This allows us to effectively integrate information from multiple sources while maintaining the spatial consistency of the BEV representation. One significant advantage of the BEV array is that it retains the same number of parameters, regardless of the number of additional sensor signals, ensuring efficient scalability without increasing computational complexity. It is also possible to optimize resource sharing across computational units and distributed sensor processing by allowing a single BEV array to be shared across these units. The operation can be explained through BEV feature aggregator **BFA** (see Figure 2). Given BEV features from imaging and distance signal paths $\mathbf{X}_b = \{\mathbf{I}_{BEV}, \mathbf{D}^o_{BEV}\}$, learnable BEV query content feature $\mathbf{z}_q \in \mathbb{R}^{Z \times X \times C}$ (C is BEV feature size and same for imaging and distance BEVs), reference points in BEV grid $\mathbf{p}_q \in \mathbb{R}^{Z \times X}$ and reference point offsets $\Delta p_{mbqk} \in \mathbb{R}^{Z \times X \times M \times B \times K \times 2}$, the BEV feature aggregator **BFA** is defined as:

$$\mathbf{BFA}(\mathbf{z}_q, \mathbf{p}_q, \mathbf{X}_b) = \sum_{m=1}^{M} \sum_{b=1}^{B} \sum_{k=1}^{K} \mathbf{A}_{mbqk} \mathbf{W}_{mb} \mathbf{X}_b (\mathbf{p}_q + \Delta \mathbf{p}_{mbqk}), \quad (6)$$

where M is number of attention heads, B is number of BEV feature grids, K is number of sampling points. The total number of sampling points, $K \ll ZX$, improves the efficiency of the attention mechanism as noted in [7]. $\mathbf{A}_{mbqk} \in \mathbb{R}^{Z \times X \times M \times B \times K}$ is an attention weight that attends to both signals and sampling points. Both $\mathbf{A}_{mbqk}$ and $\Delta \mathbf{p}_{mbqk}$ are generated by linear projection of the learnable query feature $\mathbf{z}_q$. $\mathbf{W}_{mb} \in \mathbb{R}^{Z \times X \times M \times B}$ is a parameter matrix applied to sampled signals per attention head, and is generated through linear projection of the multimodal signals. Sampling coordinates $(\mathbf{p}_q + \Delta \mathbf{p}_{mbqk}) \in [0,1]^2$ are normalized for bilinear sampling operator. Finally, the aggregated **BFA** representation is processed through two layer normalization steps and a feed-forward layer before being passed to segmentation task decoder $Dec(\cdot)$.

IV. Experiment setup

To compare with the most relevant state-of-the-art approaches, we choose vehicle segmentation as downstream task and IOU as performance metric like in FIERY [1] and SimpleBEV [3]. We also add LSS [2], and CRN [4] to the comparison. Vehicle segmentation is particularly important for short-range perception in low-speed environments in which BEV segmentation is more appropriate as it captures the object layout in the immediate surroundings, without the need for high vertical resolution.

A. Data

Dataset. We train and evaluate our approach on the multimodal dataset nuScenes [22]. For training, we use input image size 256×704 and pre-trained ResNet-50 for image encoder as the baseline of our method. For the radar data, we aggregate it using multiple signal sweeps from the timestamps $(t, t-1, t-2)$ as in SimpleBEV [3]. We utilize all 16 meta-data channels from the radar and disable nuScenes built-in outlier filtering. For the lidar data, we use only the sweep at current time t and distance information.

Data representation Following [3], we use a 3D voxel representation of size $200 \times 8 \times 200$ (Z, Y, X), with a feature channel size of 128. The coordinate for the BEV grid is right-handed where the Z axis points forward, the X axis points left, and the Y axis points up. Each voxel represents $0.5\,\mathrm{m} \times 1.25\,\mathrm{m} \times 0.5\,\mathrm{m}$ in practice, thus the total voxel grid corresponds to $100\,\mathrm{m} \times 10\,\mathrm{m} \times 100\,\mathrm{m}$. The final output of segmentation is 200×200 (Z, X).

Augmentations We follow similar signal augmentations as in [3]. For camera signals, we randomly resize and crop the input along the intrinsics in a scale range of [0.8, 1.2], and change the reference camera that randomizes the 3D volume orientation together with the orientation of the rasterized annotations. We randomly drop one of the six cameras or completely drop any individual signal path (10% of the time) to improve the robustness of the network for missing signals.

B. Downstream task

Following LSS [2], we perform vehicle segmentation by projecting the 3D bounding boxes from the vehicle meta-category in nuScenes onto the BEV plane, assigning labels

TABLE I: Comparison of BEV vehicle segmentation on the nuScenes validation set with previous works and baselines. Using SMAB signal fusion is better than baseline and previous work. Using a full multimodal setup with camera, radar and lidar is better than using a single or two signal combinations. All FPS values are measured on the same GPU.

Method	Modality	Backbone	Image Size	FPS	IOU
Lift-Splat-Shoot (LSS) [2]	Camera	EffNetB0	128 x 352	44	32.1
FIERY [1]	Camera	EffNetB4	224 x 480	1.7	35.8
BEVFormer [6]	Camera	ResNet-101	900 x 1600	4.3	46.7
SimpleBEV [3]	Camera	ResNet-101	448 x 800	16.8	47.4
SimpleBEV†	Camera + Radar	ResNet-50	256 x 704	18.5	53.0
SMAB	Camera + Radar	ResNet-50	256 x 704	**22**	**55.1 (+2.1)**‡
SimpleBEV [3]	Camera + Radar	ResNet-101	448 x 800	17	55.7
Camera Radar Net (CRN) [4]	Camera + Radar (Lidar*)	ResNet-50	256 x 704	♢	58.8
SimpleBEV [3]	Camera + Lidar	ResNet-101	448 x 800	16.6	60.8
SimpleBEV†	Camera + Lidar	ResNet-50	256 x 704	17.3	62.1
SMAB	Camera + Lidar	ResNet-50	256 x 704	22	**63.4 (+1.3)**‡
SimpleBEV†	Camera + Radar + Lidar	ResNet-50	256 x 704	17.5	62.9
SimpleBEV†	Camera + Radar + Lidar	ResNet-101	448 x 800	16	64.9
SMAB	Camera + Radar + Lidar	ResNet-50	256 x 704	**21**	**64.9 (+2.0)**‡

∗ only trained with lidar supervision
† done in this work
‡ compared to SimpleBEV [3] with same backbone & image size
♢ segmentation code/model not available

accordingly and transforming them to the ego vehicle using the provided extrinsics in nuScenes.

For the task decoder, we follow SimpleBEV [3] to process the attended BEV feature using three blocks of ResNet-18 [17] to generate three feature maps. To bring coarser features to the input resolution of 200×200, skip connections are added with bilinear upsampling. Two convolutional layers are applied to generate the final segmentation.

We also implement additional task heads to predict centerness and offset as in FIERY [1] to better regularize model. Centerness indicates the likelihood of a voxel center that is the center of the object mask while the offset is a vector field where each vector in the object mask points to the center of the object.

C. Training and evaluation setup

We train the image encoder, distance signal compression module, BFA components, and task decoder. With ResNet-50 as encoder, our model has 31M parameters. Following [3], we use the cross-entropy loss for segmentation, $L1$ loss for centerness and offset prediction, and learn uncertainty weights [23], [3] to balance three losses. All training uses Adam-W [24] with a 1-cycle schedule [25] with initial rate $5e-4$, batch size 16, and gradient aggregation over 5 steps. The complete model is trained end-to-end for $15k$ steps, with early stopping occurring around $12k$ steps. Training takes $24h$ on 4 A100 GPUs. Unlike [6], [4], we do not use temporally aggregated BEV representations. We use IOU to evaluate the performance on the full BEV grid, with the front camera as a reference and without cropping. For redundancy analysis, signals are selectively dropped according to a predefined rate. To evaluate computational cost of various methods, we measure inference speed using a single GPU (A100) without any additional accelerators, and evaluate only the model's performance, excluding any pre- or post-processing steps.

V. Results and discussion

In this section, we demonstrate that SMAB effectively handles a complete multimodal signal setup and remains resilient to sporadic signal loss while maintaining a simple architecture.

A. Multimodality

We demonstrate the advantage of using the full set of multimodal signals. Table I compares different BEV vehicle segmentation methods on nuScenes, showing that SMAB with camera, radar, and lidar outperforms variants with partial signals, even with a smaller image size and encoder. Despite radar's sparsity, integrating radar with multi-view images and lidar improves IOU from 63.4 to 64.9 when using SMAB, which highlights the benefit of leveraging all available signals. To our knowledge, it is the first time the performance is reported using all three modalities at inference time, achieving significant gains in accuracy.

B. BEV fusion

We further investigate our proposed BFA attention fusion to the closest baseline - SimpleBEV [3] under different combinations of modalities. Table I also shows that SMAB with BFA improves IOU against the baseline by 2.1 when using only camera and radar, 1.3 only camera and lidar, and 2.0 using all modalities. BFA attention fusion shows the ability to capture contextual information from multimodal BEV feature grids across all available signals by integrating multimodal features through learned BEV query. We show a more detailed analysis in the next section.

We also examine the effect of image size on the segmentation performance. Table II shows the segmentation IOU

TABLE II: Effect of image size on BEV vehicle segmentation. All experiments use ResNet-50 as image encoder. “C”, “R”, and “L” stand for camera, radar, and lidar, respectively.

Method	Modality	Image Size	IOU
SMAB	C+R	256 x 704	55.1
SMAB	C+R	448 x 800	**55.5 (+0.4)**
SMAB	C+R+L	256 x 704	64.9
SMAB	C+R+L	448 x 800	**65.7 (+0.8)**

TABLE III: Effect of image encoder size. All experiments are done using all modalities i.e. camera, radar and lidar.

Method	Backbone	Image Size	IOU
SMAB	R-101	256 x 704	64.5(-0.4)
SMAB	R-50	256 x 704	**64.9**

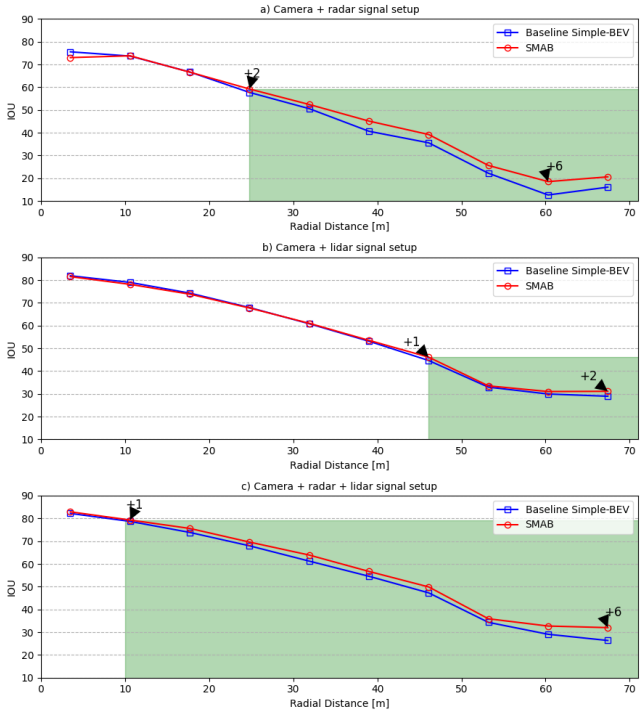

Fig. 3: IOU across radial distances for different sensor setups. The blue square-marked line represents the baseline, while the red circle-marked line represents SMAB. Arrows indicate IOU gains over the baseline, with the green area highlighting improvements. BFA enhances representation, especially when radar/lidar density is low.

under different image size and modalities. It can be seen that larger image size, e.g., increased from 256x704 to 448x800, slightly improves IOU: +0.4 using camera and radar only; +0.8 using all modalities but at the cost of almost doubled pixels and thus longer inference and training time.

Like SimpleBEV [3], we evaluate the choice of the image encoder on the performance. Table III shows that a larger image encoder, e.g. ResNet-101, does not improve performance when all modalities are used.

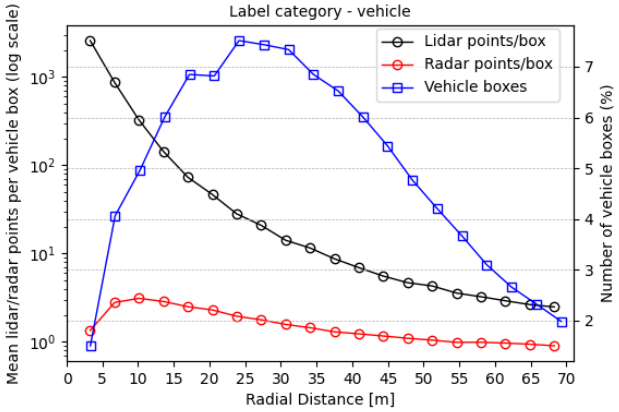

Fig. 4: Comparison of radar and lidar point density in vehicle box versus radial distance w.r.t. the ego-vehicle in the nuScenes [22] dataset. The black and red, circle marked curves represent the mean number of lidar and radar points respectively, for a given distance while the blue, square marked curve indicates the number of boxes of vehicle.

C. Radial distance analysis

In addition to overall IOU across various models (Table I), we assess SMAB’s fusion mechanism by examining IOU variations across radial distances with different sensor combinations. Figure 3 shows IOU gains (green area), when comparing SMAB to SimpleBEV [3]. SMAB better captures radar features even when its density decreases (Figure 4) as shown in Figure 3a, where SMAB improves 6 points over the SimpleBEV [3] baseline at larger radial distances. SMAB also improves camera-lidar combination (Figure 3b) at far distances.

Figure 3 also illustrates that IOU decreases with distance. This can be attributed to image resolution limitations, as objects at farther distances occupy fewer pixels. Interestingly, density of the distance signals, i.e. lidar in Figure 4, plays a less significant role at short distances (Figure 3); for example, camera-radar fusion achieves 75 IOU at 5m even with sparse radar points while camera-lidar fusion achieves 82 IOU at 5m with significantly higher density of the lidar signal. Overall, IOU performance is primarily affected by image resolution and distance signal quality, while the model is relatively insensitive to vehicle box distribution.

D. Redundancy

To further investigate the robustness of our trained model, i.e., its capability in handling transient signal losses caused by environmental disturbance, we deliberately drop signals and present incomplete data to model. Table IV shows the performance of the trained models evaluated under various signal drop scenarios. When 5% signal drop occurs, IOU decreases minimally for radar (from 64.9 to 64.5), slightly more for lidar (to 63.9), and most significantly for camera (to 61.8). The results suggest that the radar drop has the least effect on performance regression while camera signal plays the most important role for accuracy of segmentation task. Notably, when drop rate increases from 5% to 10%, the

TABLE IV: Redundancy analysis: Impact of randomly dropping different sensor signals at various rates on the IOU values. The evaluation sequence is run for each signal drop scenario and corresponding drop rate.

Method	Drop rate (%)	Only L drop	Only R drop	Only C drop
SMAB	5	63.9	64.5	61.8
SMAB	10	63.4	64.4	59.7

performance of model deteriorates, however, still remaining relatively minor.

E. Performance

We evaluate the performance of SMAB and SimpleBEV [3] by comparing Frames per Second (FPS) against IOU. Figure I demonstrates that SMAB outperforms SimpleBEV [3] when using image and radar as input modalities, as well as when incorporating image, radar, and lidar. This improvement holds for both the larger SimpleBEV [3] model, which utilizes a larger image resolution and image backbone, and the smaller SimpleBEV [3] model, which uses the same image resolution and image backbone as SMAB. Additionally, we observe that incorporating lidar only marginally reduces FPS, indicating that SMAB can efficiently scale across modalities without significant performance loss.

VI. Conclusion

SMAB provides an efficient and robust solution for multimodal sensor fusion in autonomous vehicles, integrating diverse sensor inputs with a customized attention mechanism. Unlike prior work, it utilizes all sensor modalities during inference, improving segmentation accuracy while maintaining model simplicity by using learnable parameters only where necessary. Despite its efficiency, SMAB has room for improvement. Exploring multimodal signal grouping into sub-modules and reusing shared learnable BEV queries could enhance performance. Future work will investigate these optimizations and extend SMAB to different datasets.

ACKNOWLEDGMENT

This work was conducted as part of the project "Deep Multimodal Learning for Automotive Applications", funded by Sweden's Innovation Agency Vinnova, grant no. 2023-00763. The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre in Sweden.

References

[1] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "Fiery: Future instance prediction in bird's-eye view from surround monocular cameras," in *Proc. of the IEEE/CVF International Conf. on Computer Vision*, 2021, pp. 15 273–15 282.

[2] J. Philion and S. 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*. Springer, 2020, pp. 194–210.

[3] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, "Simple-bev: What really matters for multi-sensor bev perception?" in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2023, pp. 2759–2765.

[4] Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi, and D. Kum, "Crn: Camera radar net for accurate, robust, efficient 3d perception," in *Proc. IEEE/CVF Int. Conf. on Computer Vision*, 2023, pp. 17 615–17 626.

[5] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *IEEE Int. Conf. on Robotics and Automation (ICRA*, 2023, pp. 2774–2781.

[6] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022, pp. 1–18.

[7] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[8] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[9] C. R. Qi, H. Su, K. Mo, and L. 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*, 2017, pp. 652–660.

[10] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[11] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proc. IEEE CVPR*, 2019, pp. 12 697–12 705.

[12] A. Popov, P. Gebhardt, K. Chen, and R. Oldja, "Nvradarnet: Real-time radar obstacle and free space detection for autonomous driving," in *IEEE ICRA*, 2023, pp. 6958–6964.

[13] R. Weston, S. Cen, P. Newman, and I. Posner, "Probably unknown: Deep inverse sensor modelling radar," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 5446–5452.

[14] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proc. IEEE CVPR*, 2018, pp. 918–927.

[15] Z. Yang, J. Chen, Z. Miao, W. Li, X. Zhu, and L. Zhang, "Deepinteraction: 3d object detection via modality interaction," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 1992–2005.

[16] H. Cai, Z. Zhang, Z. Zhou, Z. Li, W. Ding, and J. Zhao, "Bevfusion4d: Learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation," *arXiv preprint arXiv:2303.17099*, 2023.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[18] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE CVPR*, 2017, pp. 6924–6932.

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. ICML 2010*, pp. 807–814.

[20] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[21] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[22] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proc. IEEE CVPR*, 2020, pp. 11 621–11 631.

[23] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE CVPR*, 2018, pp. 7482–7491.

[24] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[25] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006. SPIE, 2019, pp. 369–386.