BEVMOSNet: Multimodal Fusion for BEV Moving Object Segmentation

Hiep Truong Cong[1,3], Ajay Kumar Sigatapu[1], Arindam Das[2,3], Yashwanth Sharma[2], Venkatesh Satagopan[2], Ganesh Sistu[3,4] and Ciarán Eising[3]
[1]*DSW, Valeo Kronach, Germany*
[2]*DSW, Valeo, India*
[3]*University of Limerick, Ireland*
[4]*Valeo Vision Systems, Ireland*
{*firstname.lastname*}*@valeo.com*, {*firstname.lastname*}*@ul.ie*

Keywords: Autonomous Driving, Sensor Fusion, Bird's-Eye-View Perception, Moving Object Segmentation.

Abstract: Accurate motion understanding of the dynamic objects within the scene in bird's-eye-view (BEV) is critical to ensure a reliable obstacle avoidance system and smooth path planning for autonomous vehicles. However, this task has received relatively limited exploration when compared to object detection and segmentation with only a few recent vision-based approaches presenting preliminary findings that significantly deteriorate in low-light, nighttime, and adverse weather conditions such as rain. Conversely, LiDAR and radar sensors remain almost unaffected in these scenarios, and radar provides key velocity information of the objects. Therefore, we introduce BEVMOSNet, to our knowledge, the first end-to-end multimodal fusion leveraging cameras, LiDAR, and radar to precisely predict the moving objects in BEV. In addition, we perform a deeper analysis to find out the optimal strategy for deformable cross-attention-guided sensor fusion for cross-sensor knowledge sharing in BEV. While evaluating BEVMOSNet on the nuScenes dataset, we show an overall improvement in IoU score of 36.59% compared to the vision-based unimodal baseline BEV-MoSeg (Sigatapu et al., 2023), and 2.35% compared to the multimodel SimpleBEV (Harley et al., 2022), extended for the motion segmentation task, establishing this method as the state-of-the-art in BEV motion segmentation.

1 INTRODUCTION

Recent research in accurate modeling of dynamic obstacles (Das et al., 2024) has stimulated rapid progress in achieving autonomous navigation in intricate environments, ensuring effective collision avoidance. Key components for safe and efficient autonomous driving include comprehending the movements of nearby objects and planning the vehicle's trajectory based on their anticipated future states. In recent times, the realm of autonomous driving has experienced notable progress, with leading car manufacturers integrating multiple sensor technologies (Xu et al., 2017; Li et al., 2018; Dasgupta et al., 2022) to enhance the reliability of their autonomous systems.

Rich semantic information in the image pixels has motivated the research community to pursue perception in Bird's-Eye-View (BEV) space (Roddick and Cipolla, 2020; Philion and Fidler, 2020). Despite the low cost and several other advantages, cameras are prone to failure in low illumination (low light, low contrast) and adverse weather conditions (Das et al., 2020). We aim to alleviate the failure in low illumination conditions by using LiDAR technology (another rich semantic 3D information-providing sensor).

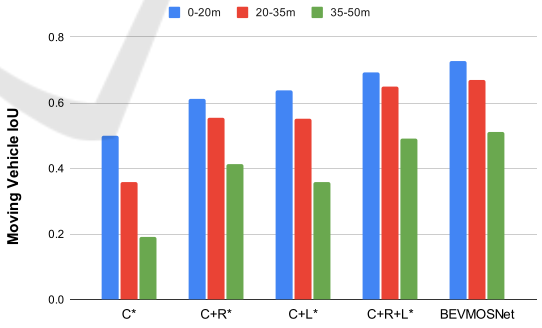

Figure 1: We propose BEVMOSNet for motion understanding within the scene. We demonstrate that our multi-modal fusion encompassing 6-cameras (C), LiDAR (L), and radar (R) yields better IoU across all distance ranges when compared to camera-only and other multimodal models. * denotes the SimpleBEV baseline model extended for the motion segmentation task.

Cong, H. T., Sigatapu, A. K., Das, A., Sharma, Y., Satagopan, V., Sistu, G. and Eising, C.
BEVMOSNet: Multimodal Fusion for BEV Moving Object Segmentation.
DOI: 10.5220/0013383300003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 3: VISAPP*, pages 863-872
ISBN: 978-989-758-728-3; ISSN: 2184-4321

Despite all the advantages of the camera and LiDAR sensor, they are prone to failure in adverse weather conditions (Godfrey et al., 2023). We aim to address this key shortcoming by integrating another cost-effective sensor, such as radar (Dong et al., 2020). It is highly reliable in adverse weather conditions and supports long-range perception. In addition, it holds key velocity information that catalyzes the detection of moving vehicles. However, radar has limitations such as the sparsity of the points in a single frame (Lippke et al., 2023) when compared to LiDAR in the nuScenes dataset (Caesar et al., 2020). We aim to make use of the complimentary features from all three sensors. Figure 1 shows the superiority in terms of distance-based IoU metrics of all three sensors when fused together over camera only, camera + radar, and camera + LiDAR fusion proposals, respectively on the nuScenes dataset.

Recent works such as (Man et al., 2023; Liang et al., 2022) focus on producing an accurate BEV semantic representation of the surrounding 3D space using a fusion of multiple sensors like multi-view cameras, radar, and LiDAR, as the BEV coordinates acts as a common ground for representing the sensor-agnostic information. Methods similar to (Chen et al., 2023) are prone to errors, as the imaging modality requires explicit depth estimation, the reason being the transformation process is quite complex, and any error in this process will have an impact on the subsequent fusion.

This paper presents evidence highlighting the significant impact of automotive sensors beyond cameras for the task at hand. We essentially paid more focus on how to intelligently integrate radar and LiDAR with camera sensors to propose a robust automotive multisensor perception stack. Our first approach is to project the sensor agnostic rich semantic features into a common reference like BEV and combine all of them by simple element-wise fusion as concatenation (Harley et al., 2022). However, the fused features suffer from misalignment due to a significant domain gap between the modalities. For example, the camera has rich semantic features but an inaccurate spatial representation due to an ambiguous transformation process. On the contrary, radar has weak semantic cues but an accurate spatial position. We use the multimodal deformable cross-attention (**MDCA**) to share the cross-modal knowledge in BEV space.

The main contributions of this paper are as follows.

- We present a novel multisensor deep network *BEVMOSNet*, designed specifically for precise motion understanding in a bird's-eye-view. The proposed network combines multi-view cameras, LiDAR, and radar, representing the first known endeavor of its kind.
- Deformable cross attention (DCA) guided design of a sensor fusion module encompassing three modalities to democratize knowledge from individual sensors to cross modalities in BEV.
- Implementation of a single-stage end-to-end trainable network establishing the first state-of-the-art results on the nuScenes dataset, at the same time improving respective state-of-the-art performance for the camera-only proposal.
- We perform thorough ablation studies considering a range of backbones, network components, and diverse feature fusion techniques.

2 RELATED WORK

Moving object segmentation is the task of understanding the dynamic properties in a scene. It includes the detection of moving objects and segmenting them from background or static components. Traditional computer vision techniques such as optical flow have been proposed to estimate the movement at the pixel level in a sequence of images. The limitation of optical flow is that it is only able to estimate relative pixel displacements between two consecutive frames and cannot distinguish dynamic and static components. Many other methods have attempted to overcome this limitation by estimating background motion (Wehrwein and Szeliski, 2017) and using RGB-D data (Menze and Geiger, 2015).

With the emergence of efficient deep learning networks that boost the performance of perception tasks, many researchers have been focusing on leveraging learning-based methods for motion segmentation. In (Patil et al., 2020) an end-to-end, multi-frame multi-scale encoder-decoder adversarial learning network is proposed for moving object segmentation. (Fragkiadaki et al., 2015) uses a CNN with a dual-pathway architecture operating on both RGB images and optical flow to estimate moving objects. InstanceMotSeg (Mohamed et al., 2020) employs the flow field as an extra source of information, guiding a deep learning model to understand object motion at the instance level. In such a multimodal setup, ensuring minimal modality imbalance (Das et al., 2023) is always challenging with automotive sensors.

In the last few years, many large-scale multimodal datasets for autonomous driving have been released, e.g., NuScenes (Caesar et al., 2020), Waymo (Sun et al., 2020). These datasets provide 3D data, such as LiDAR, radar, and surround-view camera

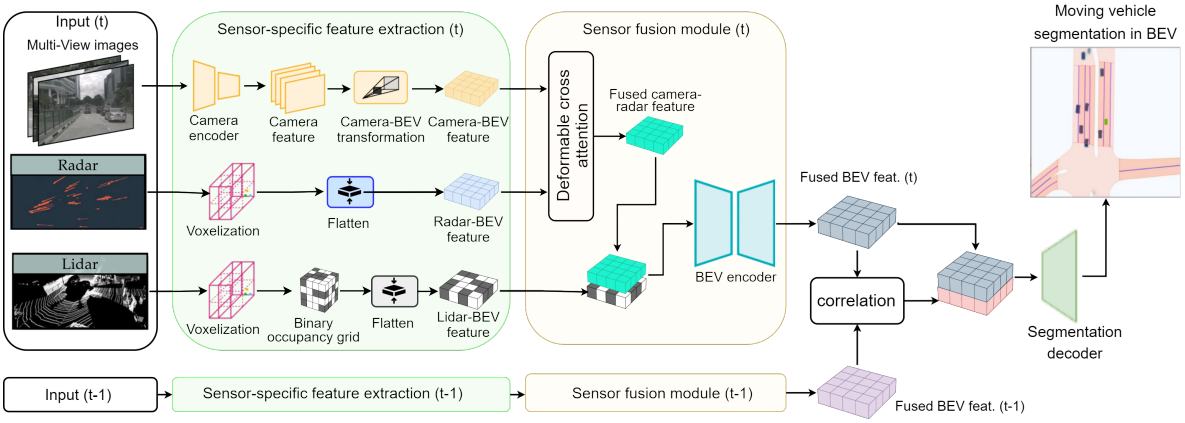

Figure 2: BEVMOSNet extracts features from camera, radar, and LiDAR input and transforms them into BEV, where they are fused together by a sensor fusion module. Consequently, a correlation block is applied to the fused BEV feature maps from current and previous frames to extract motion cues, which are then combined with the current fused BEV feature map as input for the segmentation decoder.

data. This facilitated the study of 3D perception. Many researchers targeted moving object segmentation in 3D LiDAR data. For instance, (Chen et al., 2021) segmented LiDAR points corresponding to moving objects using range images generated from point clouds. Instead of using range images as a secondary input, (Mohapatra et al., 2022) used only point cloud sequences to segment LiDAR points on moving objects and achieved real-time performance. Further studies in LiDAR moving object segmentation are InsMOS (Wang et al., 2023), MotionBEV (Zhou et al., 2023) and MambaMOS (Zeng et al., 2024).

Recently, many studies have been investigating the environmental perception in Bird's Eye View (BEV) space. Because BEV is a natural representation of 3D space with the vertical dimension compressed, perception models that operate in BEV are not only more efficient, but also have competitive performance. One of the pioneering works is BEVMODNet (Rashed et al., 2021) which segments moving vehicles in BEV just using a monocular front camera. Similar to (Fragkiadaki et al., 2015) and (Mohamed et al., 2020), this work also leverages optical flow as the second input to predict moving vehicles in BEV. Many other studies utilize the multi-view camera data in large-scale datasets, such as Lift-Splat-Shoot (LSS) (Philion and Fidler, 2020), which proposed a learning-based method to perform semantic segmentation in BEV using multi-view camera data. Following LSS, BEV-MoSeg (Sigatapu et al., 2023) added a correlation layer on top of two BEV feature maps from two consecutive frames to predict feature correspondence in BEV space before utilizing another convolution layer to predict movements of moving vehicles in the current frame. Beyond moving object segmentation in BEV, some other research tackles the dynamic perception problem in the form of motion prediction. Fiery (Hu et al., 2021), PowerBev (Li et al., 2023) and TBP-Former (Fang et al., 2023) utilize only surround views images to target future instance segmentation and motion at the same time.

Recent studies have expanded the perception tasks to radar data, as it contains object velocity, which is valuable information for dynamic perception. RaTrack (Pan et al., 2024) proposes a network for moving object detection and tracking only based on radar. RadarMOSEVE (Pang et al., 2024) proposes a Spatial-Temporal Transformer Network for moving object segmentation and ego-velocity estimation. Radar Velocity Transformer (Zeller et al., 2023) targets moving object segmentation tasks using only a single-scan radar point cloud. This work is also extended for moving instance segmentation tasks as in (Zeller et al., 2024). To our knowledge, there are no previous works that tackle the MOS task in BEV space by leveraging the multisensor data. To address this, we introduce BEVMOSNet, a multimodal deep learning model for precise motion understanding in the BEV space.

3 PROPOSED APPROACH

In this section, we describe our overall architecture and the different fusion methodologies we employed for motion segmentation in BEV space. Our proposed method consists of a multimodal feature extraction module, followed by a sensor fusion module that includes a multi-headed deformable cross-attention strategy. Additionally, a correlation module is used to extract temporal features across multiple frames in BEV, and a segmentation decoder is employed to pre-

cisely segment objects from the correlated features. Our proposed model utilizes camera, radar, and LiDAR sensor data provided in the nuScenes dataset.

3.1 Sensor-Specific Feature Extraction

The module takes the raw camera, LiDAR and radar data as inputs and extracts three sets of feature maps in BEV corresponding to each modality. In this work, we follow the multi-stream setup in SimpleBEV (Harley et al., 2022) to extract multimodal features, which consists of a CNN-based camera feature extractor, a LiDAR, and a radar voxelization module respectively.

Camera Feature Extractor. We reuse the camera feature extractor from (Harley et al., 2022), which consists of an image encoder and a 2D-3D lifting module. The input RGB images, shaped $3 \times H \times W$, are fed into a ResNet-101 (He et al., 2016) backbone. The output from layer 3 is upsampled and concatenated with the layer 2 output before being processed by two additional CNN blocks with instance normalization and ReLU activation. A final convolution layer reduces the number of channels to create image feature maps with shape $C \times H/8 \times W/8$. These 2D feature maps are then transformed into BEV space using the lifting module from (Harley et al., 2022). Where each 3D voxel "pulls" a feature from the 2D map, by projection and subpixel sampling. This results in a 3D feature volume with shape $C \times X \times Y \times Z$. Finally, this volume is rearranged to yield an image BEV feature map with shape $(C \times Y) \times X \times Z$.

LiDAR Feature Extractor. We voxelize the input LiDAR point cloud to create a binary occupancy grid with the shape of $Y \times Z \times X$. In this work, we aim to focus on the sensor fusion and keep the point cloud features as simple as possible. We only leverage the LiDAR points to provide our model with the information about object locations.

Radar Feature Extractor. In our baseline architecture, we rasterize the radar point clouds to create a radar BEV feature map. In the nuScenes dataset, each radar point has 18 attributes; the first 3 positions are point locations, and the remainder consists of velocity, compensated velocity, and other built-in pre-processing information. We use the first three attributes for rasterizing the radar point cloud, and we keep the other 15 attributes as radar features. For the MOS task, we focus on extracting dynamic information about moving objects from velocity attributes rather than relying on the radar point location information.

3.2 Sensor Fusion Module

This module introduces the sensor-specific features from the unimodal encoders. We present several fusion strategies to determine the optimal configuration among the modalities.

Concatenation. We start with the simple concatenation fusion method as proposed in SimpleBEV (Harley et al., 2022). This serves as the baseline for our proposed strategies in the following sections. SimpleBEV (Harley et al., 2022) follows a compression of BEV features to reduce the feature dimension of the unified BEV feature map. Hence, we followed the same techniques for all sensor fusion approaches.

Multi-Modal Deformable Cross-Attention (MDCA). Cross-attention has proven to yield effective results in multimodal fusion applications. However, the computation cost is quadratic to the length of the input vector $O(N^2)$, where $N = X \times Z$ and X, Z denote the height and width of the Bird's Eye View feature map. Taking into account the computational cost, we use the deformable multimodal cross attention (Kim et al., 2023) in the present task.

We apply deformable multi-modal cross attention shown in Figure 3. This mechanism selectively attends to a small set of keys sampled around a reference point in the spatial dimension for each query. This allows us to effectively identify and track moving vehicles in the neighboring pixels.

Given the sensor-agnostic BEV feature maps, we flatten them to obtain $I \in \mathbb{R}^{C_r \times XZ}$ and $R \in \mathbb{R}^{C_c \times XZ}$, where subscripts r and c denote radar and camera, respectively. $\mathbf{Z_q} \in \mathbb{R}^{XZ \times C}$ is the result of the linear projection of $I^\top \oplus R^\top$. We aim to enrich the BEV feature map using multi-head multi-modal cross-attention (MDCA) as described:

$$\mathbf{MDCA}(\mathbf{Z}_q, \mathbf{P}_q, \mathbf{X}_m) = \sum_{h=1}^{H} \left[\sum_{m=1}^{M} \mathbf{A}_{m,h} \mathbf{X}_m (\mathbf{P} + \Delta \mathbf{P}_{m,h}) \mathbf{W}_m^\top \right] \mathbf{W}_h^\top \quad (1)$$

Where, $\mathbf{A}_{m,h} = softmax(\mathbf{W}_{m,q}\mathbf{Z}_q)$ and $\Delta \mathbf{P}_m = \mathbf{W}'_{m,q}\mathbf{Z}_q$ are obtained by linear projection over the queries.

$\mathbf{P} \in \mathbb{R}^{X \times Y \times 2}$ is the reference point matrix, $\Delta P_{m,h} \in \mathbb{R}^{X \times Y \times 2}$ is the offset matrix, and $A_{m,h} \in \mathbb{R}^{XY \times K}$ is the attention weight matrix of the h-th attention head, where $A_{m,h} X_m (P + \Delta P_{m,h}) \in \mathbb{R}^{XY \times C}$. $\mathbf{W}_h \in \mathbb{R}^{C_v \times C}$ and $\mathbf{W}_m \in \mathbb{R}^{C \times C_v}$, here h and m index the attention head and modality, respectively.

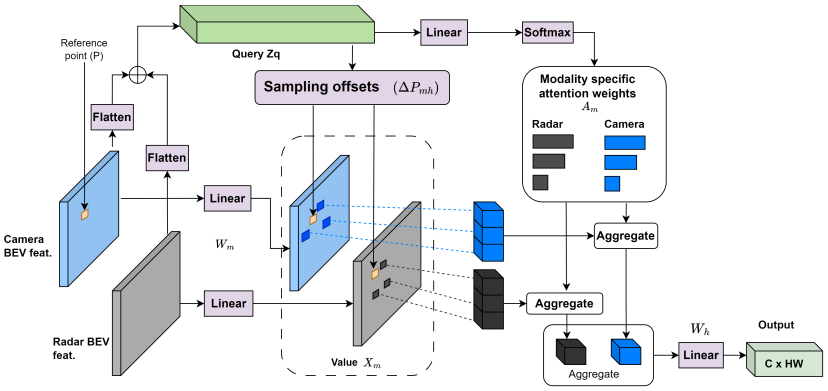

Figure 3: Multimodal deformable cross attention (MDCA) extracts complementary features from camera and radar sensors individually by separately applying attention weights $\mathbf{A}_m$ and learnable sampling offsets $\triangle\mathbf{P}_{m,h}$ in every attention head. $\oplus$ denotes concatenation.

3.3 BEV Encoder

As mentioned in BEVFusion (Liu et al., 2023), despite deformable cross attention being applied for camera and radar and all sensor-specific BEV feature maps being in the same space, there are still local misalignments between them. The camera features in BEV are not accurately located due to errors in the view transformation. Radar and LiDAR BEV feature maps are also not aligned perfectly because they have different sparsity, and radar data is noisy. To this end, we apply a BEV encoder block, which is based on ResNet18 to compensate for the misalignments.

3.4 Correlation for Detecting Motion

We aim to extract the motion cues from the scene by analyzing a pair of consecutive temporal frames similar to BEV-MoSeg (Sigatapu et al., 2023). Recent pixel-based optical flow (Dosovitskiy et al., 2015), initially processes each image independently using convolutional neural networks (CNNs) to extract feature representations. Subsequently, akin to traditional computer vision methods that compare features from image patches, the network correlates these learned representations at a higher level to identify relationships between the two images.

We take pixel-based flow estimation a step further by applying correlation layers to expand it into a higher-dimensional space representing a bird's-eye view (BEV), these correlation layers allow the network to compare sub-regions from f_1 with all other sub-regions in f_2. This enables the network to capture more complex relationships between the two images. The "correlation" between image patches, centered at x_1 in the first feature map and x_2 in the second feature map, was defined as in (Dosovitskiy et al., 2015):

$$\mathbf{c}(x_1,x_2) = \sum_{o\in[-k,k]\times[-k,k]} \langle \mathbf{f_1}(x_1+o), \mathbf{f_2}(x_2+o)\rangle \quad (2)$$

Here, $K := 2k+1$ represented the size of a square kernel. In our experiments, we have used $k = 3$. While the presented equation resembles a single step in a standard neural network convolution, it functions differently. In a typical convolution, the data is processed using trainable filters. Here, however, the correlation layer compares data with other data, eliminating the need for learnable weights.

3.5 Moving Object Segmentation Decoder

We concatenate the correlation map and BEV feature map of the current frame, thereby providing the BEV features of the current as a context to the motion cues from the correlation map, and the final stage of the decoder is the linear projection to reduce the number of filters, which consists of a 3×3 convolutional layer followed by a 1×1 convolutional layer to achieve the final output BEV segmentation map of moving vehicles.

4 EXPERIMENTATION DETAILS

To evaluate our model, we conduct experiments with different sensor combinations and different fusion methods on the publicly available nuScenes dataset (Caesar et al., 2020).

4.1 Dataset

The nuScenes dataset contains a rich collection of point cloud data and image data from 1,000 scenes,

each spanning a duration of 20 seconds collected over a wide range of weather and time-of-day conditions. The data acquisition vehicle is equipped with 6 cameras, 5 radar sensors, and a 360-degree, 32-beam LiDAR scanner. We use the official nuScenes training/validation split, which contains 28,130 samples in the training set and 6,019 samples in the validation set.

4.2 Setup

In our baseline model, we use ResNet-101 (He et al., 2016) for the image backbone. We downsample all input images to a resolution of 224×400. For the 2D-3D transformation, we use the same lifting strategy as described in section 3.1. In the LiDAR path, we voxelize point clouds and create binary occupancy 3D grids. In the radar path, we also apply the rasterized radar BEV feature map as described in section 3.1.

We use the setup described in SimpleBEV (Harley et al., 2022) as a baseline for our experiments. We use a $100m \times 100m$ region around the ego-vehicle (+/- 50m in front of and behind, +/- 50m left and right of the ego-vehicle) with a grid cell size of 50cm. This results in a 2D BEV grid map with a shape of 200×200. Along the vertical axis, we set the range to 10 m and discretize at a resolution of 8. The 3D grid volume then is shaped as $200 \times 8 \times 200$ ($X \times Y \times Z$). We orient this 3D grid according to the reference camera. To evaluate our predicted segmentation output, we use the Intersection-over-Unition (IoU) and pixel precision metrics. IoU is the score between the prediction and the ground truth (GT) of the moving vehicle in the current frame. Precision score is the number of true positive pixels divided by the number of all positive pixels. Since these GTs are not available in the nuScenes dataset, we follow BEV-MoSeg (Sigatapu et al., 2023) to generate them. First, we filter 3D bounding boxes for vehicles within our defined grid area. Next, we utilize the 'vehicle.moving' attribute on the filtered bounding boxes to identify vehicles that are in motion. At the end, we project the filtered 3D bounding boxes into the BEV space to generate binary masks. We train our baseline model with the Adam optimizer, a learning rate of 3e-4, and a weight decay of 1e-7 using 4 A100 GPUs. For all experiments, we use a batch size of 40 and train for 75,000 iterations. We use the standard binary cross-entropy loss to supervise the moving vehicle segmentation:

$$\mathcal{L}_{BCE} = \frac{-1}{N} \sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \quad (3)$$

where p_i denotes the prediction at pixel $i \in [1, N]$, and $y_i \in \{0, 1\}$ denotes the binary ground truth label at pixel i, which specifies whether the pixel i belongs to the vehicle class.

4.3 Baseline Experiments

Due to the limited number of state-of-the-art baseline models for moving object segmentation in BEV, we extended SimpleBEV (Harley et al., 2022) for the moving vehicle segmentation task (SimpleBEV_Motion) as the second baseline model besides BEV-MoSeg (Sigatapu et al., 2023) to investigate the impact of each sensor modality on the MOS task. In the fusion module, we use the simple concatenation fusion method. We start with experiments for the camera-only model. Next, we train the model using two sensor modalities, e.g., cameras with radar and cameras with LiDAR. Finally, we train our model with all sensor data (camera, radar, LiDAR). Table 1 shows experiment results with the SimpleBEV_Motion model. Compared with BEV-MoSeg (Sigatapu et al., 2023), the SimpleBEV_Motion model outperforms BEV-MoSeg (Sigatapu et al., 2023) with all sensor configurations. Even in the camera-only scenario, the model uses a simple lifting strategy without any learnable parameters and achieves an 8.04% improvement. The SimpleBEV_Motion model achieves state-of-the-art results in camera + radar + LiDAR fusion scenarios for the moving vehicle segmentation task.

4.4 Fusion Experiments

Based on the baseline experiments in 4.3, we selected the best candidate model as SimpleBEV-Motion (C+L+R) for further experiments with the MDCA fusion method. To evaluate the MDCA for the moving object segmentation task, we apply MDCA with 3 different fusion strategies: 1. Camera-radar fusion using DCA, then the fused feature map is concatenated with the LiDAR BEV feature map; 2. Camera-LiDAR fusion using DCA, then the fused feature map is concatenated with the radar BEV feature map; 3. Radar and LiDAR feature maps are concatenated, then fused with the camera feature map using DCA. Table 1 shows our experiment results with these configurations. We confirm that the MDCA strategy can boost the performance of moving object segmentation tasks in all fusion configurations compared to the baseline. We also observe that by applying MDCA for camera and radar features, we can utilize the dynamic information inherited in radar data to help the model focus on more useful camera features, which helps to improve the model performance.

Table 1: MOS with different sensor setups and fusion strategies. 'C', 'R', and 'L' represent camera, radar, and LiDAR, ↑ indicates that a higher value is better. ⊗ denotes multimodal deformable cross attention, ⊕ denotes concatenation. * denotes our baseline model for experiments with the MDCA fusion method.

Method	Modality	Image backbone	Fusion method	Precision (%)↑	mIoU (%)↑
MoSeg(Sigatapu et al., 2023)	C	EfficientNet-b0	-	-	26.0
SimpleBEV_Motion(Harley et al., 2022)	C	ResNet-101	-	49.43	34.04
SimpleBEV_Motion(Harley et al., 2022)	C+R	ResNet-101	C ⊕ R	66.65	51.52
SimpleBEV_Motion(Harley et al., 2022)	C+L	ResNet-101	C ⊕ L	67.44	50.27
SimpleBEV_Motion(Harley et al., 2022)	C+R+L	ResNet-50	C ⊕ R ⊕ L	73.03	59.86
SimpleBEV_Motion(Harley et al., 2022)	C+R+L	EfficientNet-b4	C ⊕ R ⊕ L	73.25	59.90
SimpleBEV_Motion(Harley et al., 2022)*	C+R+L	ResNet-101	C ⊕ R ⊕ L	73.79	60.24
BEVMOSNet	C+R+L	ResNet-101	C ⊗ (L ⊕ R)	73.22	61.82
BEVMOSNet	C+R+L	ResNet-101	(C ⊗ L) ⊕ R	74.93	60.91
BEVMOSNet	C+R+L	EfficientNet-b4	(C ⊗ R) ⊕ L	75.05	62.22
BEVMOSNet (ours)	C+R+L	ResNet-101	(C ⊗ R) ⊕ L	**75.35**	**62.59**

4.5 Ablation Study

For each ablation experiment, we only train models for 50,000 iterations for faster convergence. Table 2 shows the IoU of segmented moving objects over different distances. The performance of the camera-only model drops significantly when objects are far from the ego-vehicle. With LiDAR data, the camera + LiDAR model performs better on far objects. Due to the sparsity and noise of radar data, the camera-radar model performs slightly worse at closer distances compared to camera-LiDAR, but it improves significantly at long distances.

In Table 3, we show the result of experiments with different numbers of aggregated LiDAR and radar sweeps. We observe that aggregation of sweeps helps to improve the model performance, and the baseline model achieves the best performance with an aggregation of five sweeps; therefore, we conduct all experiments in Table 1 with the five-sweep aggregation. It has been observed that when we consider more than five sweeps, then we see performance degradation. Although the corresponding results are not reported, this is a critical finding, as we observed accumulating more sweeps shows almost zero overlap of the moving object across frames. This phenomenon can be attributed to the predominance of urban scenarios in nuScenes dataset, where dynamic objects such as vehicles traverse considerable distances over time. In Table 4 we show the model performance in daytime, nighttime and rain driving conditions where the performance in rain scenes is significantly improved by using multimodal sensor fusion. The performance gap of the camera-only model between rain and daytime conditions is 6.05%. By adding radar data, this gap is shrunk to 2.81%. The camera-LiDAR-radar model closes this gap and increases the performance in rain conditions by 0.84% compared to the performance in daytime conditions. This confirms that radar and LiDAR are useful for perception in adverse weather conditions. Generally, adding radar and LiDAR helps improve the performance of the camera-only model in all driving conditions.

Table 2: Ablation study on the usage of unimodal vs. multimodal sensor using mIoU metric with respect to distances for moving object segmentation task.

	0-20m	20-35m	35-50m
C	50.14	35.91	19.11
C+R	61.10	55.57	41.31
C+L	63.73	55.11	35.96
C+R+L (BEVMOSNet)	**72.68**	**67.04**	**51.21**

Table 3: Exploring the influence of varying numbers of aggregated LiDAR and radar sweeps with camera data. The best sweep aggregation was chosen for our experiments in Table 1.

	C+L	C+R	C+R+L (baseline)	BEVMOSNet
1 frame	46.73	50.13	58.69	-
3 frames	48.20	51.01	59.68	-
5 frames	50.27	51.52	60.24	**62.59**

Table 4: Performance analysis of BEVMOSNet and comparison with other sensor proposals in adverse weather scenarios and low illumination conditions using mIoU.

	Camera-only	C+L	C+R	C+R+L (BEVMOSNet)
Rain	28.75	50.88	49.04	**63.41**
Day	34.80	49.67	51.85	**62.57**
Night	35.10	52.76	51.44	**59.64**

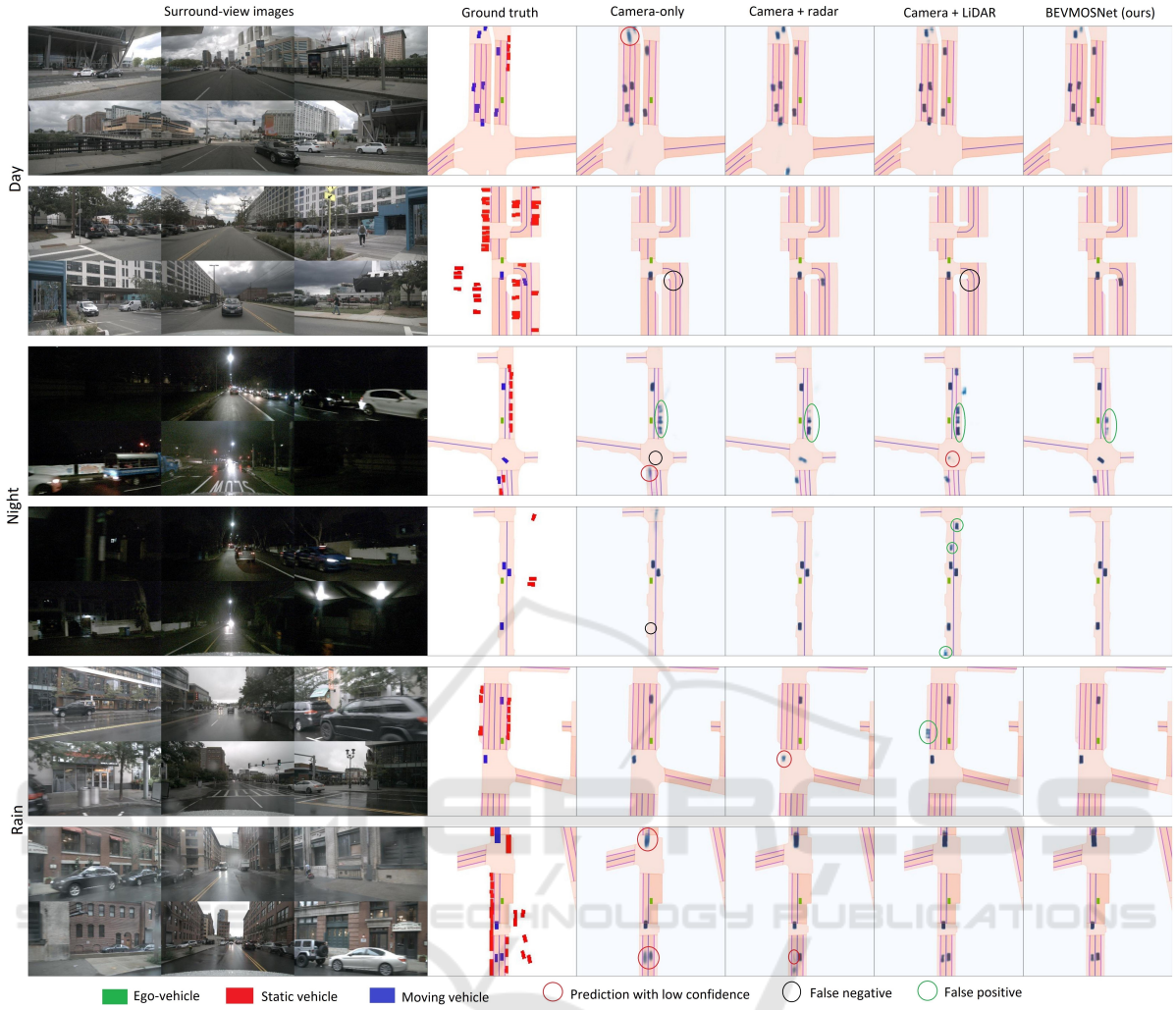

Figure 4: Qualitative results on MOS in various weather conditions. The camera-only model predicts distant moving objects with lower confidence (blurred region, marked with red circles). It also fails to segment occluded moving objects, or when operating in low light conditions, such as at night (regions marked with black circles). Generally, LiDAR helps to locate object positions and estimate object orientation accurately; radar improves the segmentation of distant objects. By combining camera, LiDAR, and radar we can leverage the advantages of each modality to build a robust model, which reduces false positive predictions (marked with green circles).

4.6 Qualitative Results

Figure 4 shows qualitative results of moving vehicle segmentation. We observe that the prediction confidence of the camera-only model is lower, particularly for objects far from ego vehicles. The camera-LiDAR model predicts more precise object locations and also increases prediction confidence. The camera-radar model helps greatly predict objects at far distances. Besides that, radar sensors also provide information about dynamic objects through velocity attributes, which greatly improves the prediction of occluded moving objects. On the other hand, due to the sparse and noisy nature of radar data, the camera-radar model produces more noisy predictions. By combining all three sensor modalities, we achieve a more robust model, which compensates for the weaknesses of each sensor type and increases the overall segmentation performance.

5 CONCLUSION

In this work, we introduce a novel multi-sensor, multi-camera architecture for motion understanding in BEV, achieving a 62.59% IoU score on the nuScenes moving object detection dataset. Our investigation includes extensive experiments aimed at assessing the impact of each sensor modality on overall

performance during the feature fusion stage and optimal configuration for sensor fusion. Additionally, we integrate deformable cross-attention to improve the extraction of robust camera features, leveraging the complementary information from LiDAR and radar modalities. Due to the limited availability of moving object labels within nuScenes, which are currently restricted to the vehicle class, our experimental validation solely focuses on this category. However, it is possible to boost the performance further and extend the motion detection task to more classes, such as bicyclists and pedestrians with the label availability. We leave this work to future research.

REFERENCES

Caesar, H. et al. (2020). nuscenes: A multimodal dataset for autonomous driving. In *CVPR*.

Chen, X. et al. (2021). Moving Object Segmentation in 3D LiDAR Data: A Learning-based Approach Exploiting Sequential Data. *IEEE Robotics and Automation Letters (RA-L)*, 6:6529–6536.

Chen, X. et al. (2023). Futr3d: A unified sensor fusion framework for 3d detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 172–181.

Das, A., Das, S., Sistu, G., Horgan, J., Bhattacharya, U., Jones, E., Glavin, M., and Eising, C. (2023). Revisiting modality imbalance in multimodal pedestrian detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1755–1759. IEEE.

Das, A., Křížek, P., Sistu, G., Bürger, F., Madasamy, S., Uřičář, M., Kumar, V. R., and Yogamani, S. (2020). Tiledsoilingnet: Tile-level soiling detection on automotive surround-view cameras using coverage metric. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE.

Das, A., Paul, S., Scholz, N., Malviya, A. K., Sistu, G., Bhattacharya, U., and Eising, C. (2024). Fisheye camera and ultrasonic sensor fusion for near-field obstacle perception in bird's-eye-view. *arXiv preprint arXiv:2402.00637*.

Dasgupta, K., Das, A., Das, S., Bhattacharya, U., and Yogamani, S. (2022). Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*.

Dong, X., Wang, P., Zhang, P., and Liu, L. (2020). Probabilistic oriented object detection in automotive radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 102–103.

Dosovitskiy, A. et al. (2015). Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766.

Fang, S. et al. (2023). Tbp-former: Learning temporal bird's-eye-view pyramid for joint perception and prediction in vision-centric autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1368–1378.

Fragkiadaki, K. et al. (2015). Learning to segment moving objects in videos. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4083–4090.

Godfrey, J., Kumar, V., and Subramanian, S. C. (2023). Evaluation of flash lidar in adverse weather conditions towards active road vehicle safety. *IEEE Sensors Journal*.

Harley, A. W., Fang, Z., Li, J., Ambrus, R., and Fragkiadaki, K. (2022). A simple baseline for bev perception without lidar. In *arXiv:2206.07959*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hu, A. et al. (2021). FIERY: Future instance segmentation in bird's-eye view from surround monocular cameras. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Kim, Y. et al. (2023). Crn: Camera radar net for accurate, robust, efficient 3d perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17615–17626.

Li, C., Song, D., Tong, R., and Tang, M. (2018). Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference 2018, BMVC 2018*. BMVA Press.

Li, P. et al. (2023). Powerbev: A powerful yet lightweight framework for instance prediction in bird's-eye view. In Elkind, E., editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 1080–1088. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Liang, T. et al. (2022). Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434.

Lippke, M. et al. (2023). Exploiting sparsity in automotive radar object detection networks. *arXiv preprint arXiv:2308.07748*.

Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., and Han, S. (2023). Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Man, Y., Gui, L.-Y., and Wang, Y.-X. (2023). Bev-guided multi-modality fusion for driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Menze, M. and Geiger, A. (2015). Object scene flow for autonomous vehicles. In *In Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 245, pages 3061–3070.

Mohamed, E., Ewaisha, M., Siam, M., Rashed, H., Yogamani, S., and El-Sallab, A. (2020). Instancemotseg: Real-time instance motion segmentation for autonomous driving. *arXiv preprint arXiv:2008.07008*.

Mohapatra, S. et al. (2022). Limoseg: Real-time bird's eye view based lidar motion segmentation.

Pan, Z., Ding, F., Zhong, H., and Lu, C. X. (2024). Moving object detection and tracking with 4d radar point cloud. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

Pang, C. et al. (2024). Radarmoseve: A spatial-temporal transformer network for radar-only moving object segmentation and ego-velocity estimation.

Patil, P. W. et al. (2020). An end-to-end edge aggregation network for moving object segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8146–8155.

Philion, J. and Fidler, S. (2020). Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of the European Conference on Computer Vision*.

Rashed, H., Essam, M., Mohamed, M., Ei Sallab, A., and Yogamani, S. (2021). Bev-modnet: Monocular camera based bird's eye view moving object detection for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1503–1508.

Roddick, T. and Cipolla, R. (2020). Predicting semantic map representations from images using pyramid occupancy networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11138–11147.

Sigatapu, A. K. et al. (2023). Bev-moseg: Segmenting moving objects in bird's eye view. In *2023 21st International Conference on Advanced Robotics (ICAR)*, pages 381–386. IEEE.

Sun, P. et al. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, N. et al. (2023). InsMOS: Instance-Aware Moving Object Segmentation in LiDAR Data. *arXiv preprint*, 2303.03909.

Wehrwein, S. and Szeliski, R. (2017). Video segmentation with background motion models. *In BMVC*, 245:246.

Xu, D. et al. (2017). Learning cross-modal deep representations for robust pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5363–5371.

Zeller, M. et al. (2023). Radar velocity transformer: Single-scan moving object segmentation in noisy radar point clouds. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*.

Zeller, M. et al. (2024). Radar instance transformer: Reliable moving instance segmentation in sparse radar point clouds. *IEEE Transactions on Robotics*, 40:2357–2372.

Zeng, K., Shi, H., Lin, J., Li, S., Cheng, J., Wang, K., Li, Z., and Yang, K. (2024). Mambamos: Lidar-based 3d moving object segmentation with motion-aware state space model. In *ACM International Conference on Multimedia (MM)*.

Zhou, B., Xie, J., Pan, Y., Wu, J., and Lu, C. (2023). Motionbev: Attention-aware online lidar moving object segmentation with bird's eye view based appearance and motion features. *IEEE Robotics and Automation Letters*, 8(12):8074–8081.