The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W5-2025
13th International Conference on Mobile Mapping Technology (MMT 2025)
"Mobile Mapping for Autonomous Systems and Spatial Intelligence", 20–22 June 2025, Xiamen, China

BEV Space LiDAR-Camera Fusion Methods Based on Attention Driven Feature Fusion Mechanism

Leheng Xu [1], Minglei Li* [1,2], Cong Zhou [1], Jiahui Chai [1], Junnan Zhang [1]

[1] College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, 211106 Nanjing, China - (xuleheng657, minglei_li, iszhouc, jiahui, zhangjunnan)@nuaa.edu.cn
[2] Key Laboratory of Radar Imaging and Microwave Photonics (Nanjing University of Aeronautics and Astronautics), Ministry of Education, 211106 Nanjing, China

Keywords: Multi-scale Attention, Bird's Eye View, LiDAR-Camera Fusion, Object Detection

Abstract

To address perception challenges for autonomous vehicles and drones in complex urban environments, this paper proposes a novel Bird's Eye View (BEV) fusion method MSA-BEVFusion to integrate LiDAR and RGB cameras via multi-scale attention mechanisms. Unlike existing methods that tightly couple image and LiDAR features or BEV-based approaches relying on simplistic convolutional fusion, our method first integrates multi-scale image features through the MFPN module and employs multi-scale attention enhancement to achieve deep fusion between camera and LiDAR features before feeding them into the detection head, ultimately delivering superior detection performance. Experiments on the nuScenes dataset demonstrate excellent performance, achieving 0.2% NDS and 0.4% mAP improvements over BEVFusion-MIT. The method shows robust 3D detection in dark, rainy, and snowy conditions, with enhanced accuracy for small or occluded objects. Attention heatmaps reveal effective cross-modal alignment, synergizing LiDAR's geometric precision with camera texture details. This work bridges modality gaps through bidirectional interaction, advancing robust environmental perception while mitigating spatial discordance in unified BEV representations.

1. Introduction

Driven by the rapid advancement of autonomous driving, 3D object detection methods have undergone significant development. Numerous approaches are based on 2D images, while a considerable number rely on 3D point clouds acquired from LiDAR. 2D images offer abundant texture details but are deficient in depth information which is crucial for 3D detection. Mainstream depth estimation methods, such as the LSS (Philion et al., 2020) which explicitly estimates depth and the BEVDepth (Li et al., 2023) incorporating depth supervision, while demonstrating commendable detection performance, do not attain the same level of accuracy as LiDAR point clouds, which inherently possess precise depth ground truth. While LiDAR point clouds are rich in spatial information, their unordered and sparse nature still poses significant challenges for object detection. Consequently, fusion methods combining these two modalities have gradually emerged as a dominant direction in current research.

Fusion approaches can be broadly classified into three levels based on their integration point: early-fusion, deep-fusion, and late-fusion. Figure 1 illustrates representative examples of these three fusion approaches.

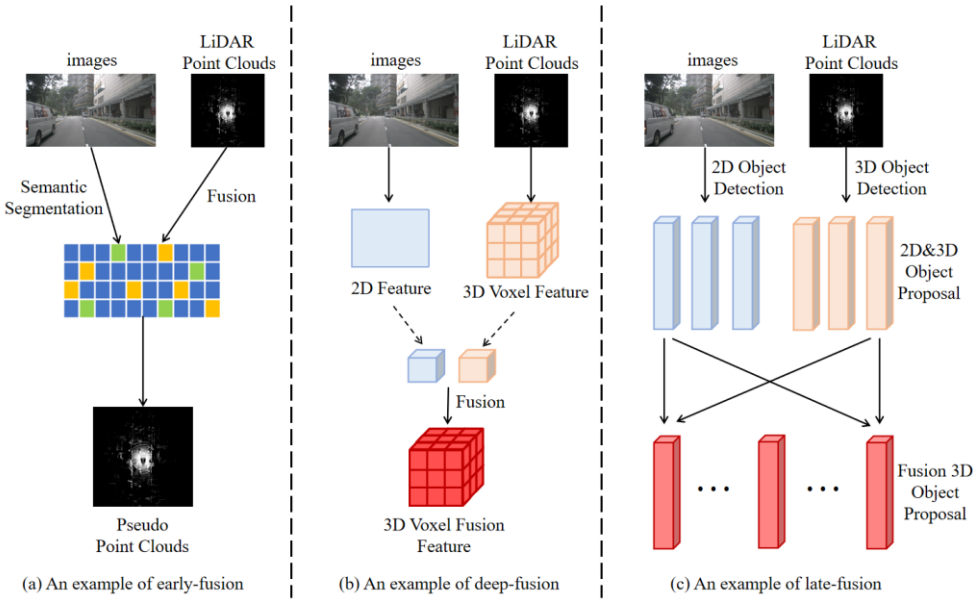

Figure 1. Examples of three different fusion models

Early-fusion primarily occurs at the raw data level. This includes approaches such as projecting rich color information from images onto 3D point clouds or projecting upsampled point clouds back to the image plane. Representative methods include PFF3D (Zhang et al., 2020), Painted PointRCNN (Vora et al., 2020), and others. However, these approaches can lead to the loss of integrity of data from both modalities and imposes strict requirements on sensor calibration. Furthermore, in case of failure in one sensor stream, the entire fusion method becomes ineffective. Deep-fusion, on the other hand, performs integration at the feature level, combining features from both camera and LiDAR point clouds. This results in fused features such as 3D voxels or geometric primitives like pillars. Representative methods include EPNet (Huang et al., 2020), 3D-CVF (Yoo et al., 2020), and others. This approach typically exhibits stronger robustness and has gained popularity in recent years. Late-fusion, also known as object-fusion, refers to approaches that independently predict detection proposals from each modality and then perform post-processing on the combined proposals. Late-fusion can be regarded as a kind of ensemble method that utilizes multi-modal information to optimize the final proposal. A representative method is CLOCS (Pang et al., 2020). However, late fusion's overall prediction performance can be compromised when the detection proposals from one modality are of poor quality.

Bird's Eye View (BEV) appeared as a prominent data representation in autonomous driving algorithms in recent years, and most BEV-based fusion algorithms fall under the deep-fusion paradigm. In the context of object detection, BEV effectively eliminates perspective distortion and enhances 3D spatial perception capabilities by mapping multi-view camera or LiDAR feature to a unified bird's eye view coordinate system. For LiDAR-camera fusion scenarios, the BEV framework provides a naturally aligned intermediate representation for integrating data across modalities, enabling deep integration of visual semantic information with LiDAR's precise space information. This significantly enhances detection robustness in complex dynamic environments. Furthermore, its top-down perspective and capability to consolidate global contextual information help mitigate occlusion problems and optimize the detection performance for distant and occluded targets, rendering the method more practical and reliable in real-world applications like autonomous driving.

We propose MSA-BEVFusion, which optimizes the processes of feature extraction and fusion within the BEV spatial framework. The main contributions we made are summarized as follows:

- We decouple the multi-camera and LiDAR sensors, which ensures a functional detection head with minimal performance degradation upon single sensor failure.
- We introduce a Merged Feature Pyramid Network (MFPN) module for optimizing the image feature extraction pipeline. This module integrates multi-scale image features rather than discarding portions as is common in many existing methods, thus guaranteeing the preservation of features across all scales.
- We propose a Multi-scale Attention (MSA) module, enabling the model to adaptively adjust the weights of camera and LiDAR features via self-attention during training. This lightweight yet effective module leads to a notable improvement in NDS and mAP scores compared to baseline methods.

2. Related Work

2.1 BEV-based object detection

BEV approaches have been widely adopted. Among them, popular BEV-based detection methods are broadly categorized as camera-only methods and LiDAR-visual fusion methods.

2.1.1 Camera-Only Methods: While the detection performance of camera-only methods may not match that of LiDAR-based or LiDAR-visual fusion methods, the high cost of LiDAR sensors drives many autonomous driving manufacturers today to continue extensive research into camera-only algorithms. Consequently, detection algorithms based on multi-camera images still hold significant research value. As previously discussed, a key challenge for 3D object detection from pure images is that two-dimensional images inherently lack depth information. In 3D space, points that are distant due to depth differences can appear very close in the 2D image plane, whereas points with similar depth can be far apart in the image. This renders image context completely uninformative for depth estimation. Lift-Splat-Shoot (Philion et al., 2020) implicitly unprojects multi-view images into 3D feature frustums through the Lift operation, then efficiently aggregates features into BEV grid via Splat for cross-camera fusion. It addresses monocular depth ambiguity through differentiable depth probability modeling, learns BEV semantic representations end-to-end, and enables interpretable motion planning through trajectory Shooting in the BEV space. BEVDet (Huang et al., 2022) innovatively performs multi-camera 3D object detection in BEV space, addresses overfitting through customized BEV-space data augmentation, proposes Scale-NMS to adaptively adjust detection boxes by object categories for small-target precision, and employs modular architecture to decouple image-view encoder from BEV-space learning. BEVDepth (Li et al., 2022) explicitly supervises depth prediction networks using point-cloud-generated depth ground truth, enhances cross-device robustness through camera-parameter-encoded networks, refines BEV-space feature projection via depth-axis convolutions to mitigate semantic drift, and achieves high-precision 3D detection by integrating efficient voxel pooling with multi-frame fusion under unified BEV representation.

2.1.2 LiDAR-camera fusion methods: LiDAR-camera fusion approaches achieve high detection precision by capitalizing on the respective strengths of camera and LiDAR modalities. While practical implementations are generally costly, their inherent reliability has nonetheless established them as a dominant research direction in recent years. However, traditional methods are often relied on accurate and invariant sensor calibration. When a sensor fails or calibration drifts due to vibrations or other physical disturbances, the model often fails catastrophically. BEVFusion-MIT (Liu et al., 2023) and BEVFusion-ADLAB (Liang et al., 2022) introduced a pioneering approach by proposing an innovative unified BEV framework that integrates multi-modal features into a shared BEV space while preserving both geometric structure and semantic richness. By addressing computational bottlenecks in view transformation through optimized BEV pooling operations, the method achieves efficient cross-modal feature projection. A fully-convolutional encoder further aligns heterogeneous sensor data, enabling seamless multi-task learning for diverse perception objectives. Its core advancement lies in preserving dense semantic information from cameras through ray-based projection, overcoming the information loss caused by sensor density disparities in conventional fusion approaches. This design offers a versatile and computationally effective paradigm

for multimodal perception systems. BEVFormer (Li et al., 2025) even pushes it further by designing grid-shaped BEV queries to extract features from LiDAR point clouds and multi-view images via cross-modal attention in spatial domain, while recurrently fusing historical BEV features via self-attention for temporal coherence. This approach transcends conventional feature concatenation by enabling progressive interaction and mutual enhancement between modalities through query-level iterative optimization. The unified BEV representation flexibly supports multiple collaborative tasks including 3D detection, object tracking, and map construction, providing an efficient environmental perception framework for autonomous driving systems.

2.2 Channel and spatial attention mechanisms

One of the key research hotspots in computer vision in recent years is attention mechanisms applying weighting operations across channel and spatial dimensions of tensors. Squeeze-and-excitation (SENet) (Hu et al., 2018) represents a seminal work in this area. It operates on the channel dimension of features by compressing and expanding through linear layers, and then learns per-channel weights via a sigmoid function. These weights are subsequently multiplied with the original feature maps, enabling the model during training to adaptively suppress less important channels and amplify more important ones. However, subsequent work has gradually revealed that modeling cross-channel relationships by reducing channel dimensionality can introduce detrimental side effects on deep visual feature extraction. CBAM (Woo et al., 2018) adaptively refines features through two sequential sub-modules: channel attention and spatial attention. The channel attention captures cross-channel information using both global average-pooling and max-pooling to generate channel weights, while the spatial attention aggregates features along channel dimensions and employs convolutional layers to produce spatial weights. CA (Hou et al., 2018) decomposes channel attention into two independent one-dimensional feature encoding processes aggregating features along horizontal and vertical axes respectively. By preserving directional coordinate information during spatial encoding, this method generates position-sensitive attention maps that enable precise localization of target regions. EMA (Ouyang et al., 2023) groups channel dimensions into sub-features while preserving per-channel information and reducing computational costs. It introduces a novel cross-spatial learning mechanism that fuses spatial attention maps from different branches using matrix dot-product operations, enabling pixel-level cross-dimensional interaction.

3. Methodology

We propose MSA-BEVFusion, a multi-scale attention with cross-spatial driven fusion model which provides high-performance detection in BEV space.

3.1 Overview

The workflow of our approach consists of three core components: multi-modal BEV feature extraction, attention-based BEV feature fusion, and detection head. First, we extract features from both LiDAR point cloud and RGB images and project them both into the shared BEV space, aligning the 3D LiDAR data with the 2D camera features. Then, we apply a multi-scale attention based fusion model to combine the multi-modal BEV features. Finally, a detection head is applied to obtain the final result of 3D object detection. An illustration of this approach is shown in Figure 2.

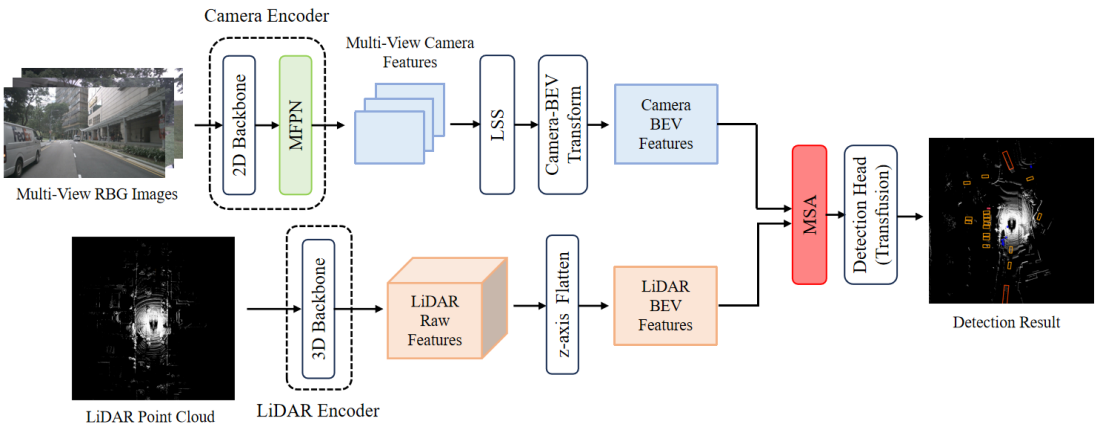

Figure 2. Structure of MSA-BEVFusion

3.2 multi-modal BEV feature extraction branch

For the feature extraction module, we employ several classic and high-performing backbones to extract 2D and 3D features respectively. We also propose a neck module for integrating multi-scale features and utilize an efficient BEV pooling method when projecting image features into the BEV space.

3.2.1 Camera image encoder: we select Swin-Transformer (Liu et al., 2021) as our 2D backbone. Unlike other transformer-based encoders, Swint-T centers around hierarchical window attention, unifying computational efficiency and multi-scale feature representation while maintaining the global modeling capability of Transformers. The processing pipeline starts by partitioning the input image into non-overlapping 4×4 pixel patches, which are mapped through a learnable linear projection to a high-dimensional embedding space of dimension C, resulting in a feature map of size $H/4\times W/4$. Following this, hierarchical representations are constructed through four stages of progressive downsampling. In each stage, a patch merging operation is applied to concatenate and then reduce the dimension of features from adjacent 2×2 patches, successively reducing the resolution to $H/8\times W/8$, $H/16\times W/16$, and ultimately $H/32\times W/32$, forming multi-scale feature outputs. During this process, the Swin-Transformer module employs alternating standard window self-attention (W-MSA), which partitions the feature map into non-overlapping local windows of size $M\times M$, and performs self-attention computation within

each window. To overcome the limitations imposed by window boundaries on modeling long-range dependencies, a shifted window partitioning (SW-MSA) is introduced between adjacent layers. This is achieved by cyclically shifting the windows $\lfloor M/2 \rfloor$ pixels towards the bottom-right before re-partitioning, enabling attention computation to dynamically capture contextual information across window boundaries. All in all, the outputs of the Swin-T module are three tensors f_1, f_2, f_3 with shape of $N \times 2C \times H/8 \times W/8$, $N \times 4C \times H/16 \times W/16$ and $N \times 8C \times H/32 \times W/32$. To preserve and fuse features from three scales, the Merged Feature Pyramid Network (MFPN) module we proposed innovatively fuses features for subsequent spatial alignment purposes.

Figure 3 illustrates the process of the MFPN module. Specifically, features from all scales are upsampled to a common base resolution $H/8 \times W/8$, and their channels are adjusted to C_{camera} using 1×1 convolutions. Finally, these features are concatenated along the channel dimension and then passed through a 3×3 convolution to extract deeper features, resulting in camera features with shape of $N \times C_{camera} \times H/8 \times W/8$. Then following BEVFusion-MIT (Liu et al., 2023), we utilize LSS (Philion et al., 2020) to create 3D point cloud frustum indices by explicitly estimating the depth value for each feature pixel using the model. Then, using the pose relationships between the camera, LiDAR and ego-car coordinates, we employ an optimized BEV pooling method to accelerate the process of projecting camera features into the BEV space.

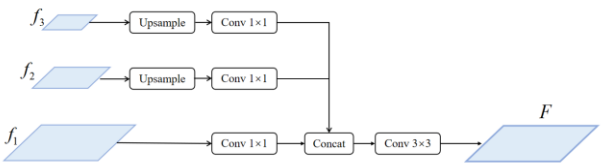

Figure 3. The structure of MFPN module

3.2.2 LiDAR point cloud encoder: For the LiDAR point cloud encoder, we employ a classic voxelization encoding approach. This involves first establishing a voxel grid for the point cloud, followed by performing feature extraction through sparse convolution (Graham et al., 2018). This method employs a collaborative mechanism of hash tables and feature matrices to efficiently process 3D sparse data. The workflow encodes input data into spatial coordinate hash tables and active-site feature matrices, dynamically establishing convolutional kernel position mapping rules between input and output sites through traversal operations. The core submanifold sparse convolution restricts feature computation exclusively to cases where the kernel center aligns with active input sites, preserving identical sparsity patterns between input and output layers. Usually, matrix multiply-add operations are rigorously confined to active regions, while integrated batch normalization and pooling operations enable full-spectrum sparse computation from feature extraction to multi-scale information fusion. This architecture fundamentally resolves computational redundancy in 3D space, achieving substantial reductions in computational overhead and memory consumption while maintaining high accuracy. The extracted voxel features are then compressed and flattened along the z-axis to align with the camera features in the BEV space. Finally, features from both modalities are now aligned in the BEV space.

3.3 Multi-scale attention-based feature fuser

Once features from both modalities are represented in the same BEV space, how to effectively fuse them becomes a topic of considerable interest. The easiest approach is direct concatenation. However, this method typically exhibits poor robustness. Convolutional is also a plausible choice, however, prior methods have shown that they struggle to effectively address the issue of spatial misalignment arising from depth estimation errors. To mitigate this issue, following EMA (Ouyang et al., 2023), we introduce the Multi-scale Attention (MSA) module, Extending this approach from 2D image processing to 3D tasks. The module is illustrated in Figure 4. The MSA module is designed to enhance feature representation through a cross-spatial learning approach while maintaining computational efficiency. MSA takes an input feature map $X \in \mathbb{R}^{N \times C \times H \times W}$ and divides it into G sub-feature groups $[X_0, X_1, \ldots, X_{G-1}]$ along the channel dimension. To circumvent the influence of batch dimensions on the number of convolution kernels in traditional convolutions, MSA cleverly reshapes and permutes these groups into the batch dimension, forming input tensors with a shape of $N \times C/G \times H \times W$, where G is much smaller than C . The module incorporates three parallel processing paths internally to extract attention weight descriptors.

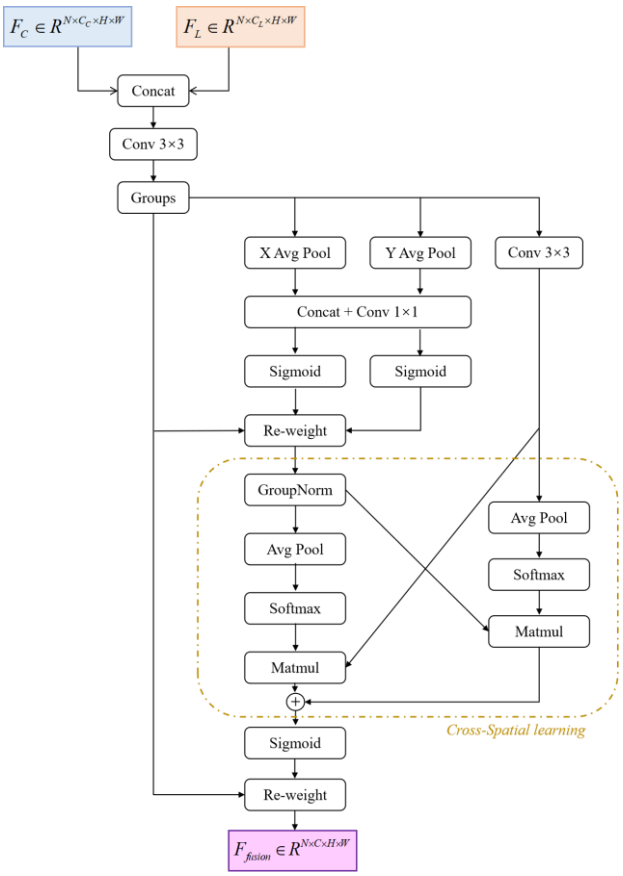

Figure 4. The structure of MSA module.

Two of these paths belong to the 1×1 branch, which utilizes 1D global average pooling along the horizontal and vertical directions, respectively, to encode features. This 1D pooling operation effectively captures long-range dependencies in the corresponding direction and preserves precise positional

information in the other. The pooling output for the horizontal direction is represented as

$$z_c^H(H) = \frac{1}{W} \sum_{0 \le i \le W} x_c(H, i) \quad (1)$$

where x_c indicates the input features at c -th channel. With such encoding processes, MSA can capture the long-range dependencies at the horizontal direction and preserve precise positional information at the vertical direction. Similarly, the other one parallel route is directly from 1D global average-pooling along the horizontal dimension direction and hence can be viewed as a collection of positional information along the vertical dimension direction. Then, the route utilizes the 1D global average-pooling along the vertical dimension direction to capture long-range interactions spatially and preserve the precise positional information along the horizontal dimension direction. The pooling output in C at W can be formulated as

$$z_c^W(W) = \frac{1}{H} \sum_{0 \le j \le H} x_c(j, W) \quad (2)$$

where x_c denotes the input feature at the c -th channel. These two sets of encoded features are concatenated along the image height dimension and processed by a shared 1×1 convolution. Subsequently, they undergo a nonlinear transformation via the Sigmoid function, and the intra-group channel attention maps are aggregated through simple multiplication, enabling adaptive calibration and interaction between channels. Since the positional information preserved along different spatial directions is complementary, this method helps MSA learn fine-grained low-level features.

The third path is the 3×3 branch, which employs a 3×3 convolution to expand the feature space and capture multi-scale features. To further merge information from different spatial scales, MSA introduces cross-spatial information aggregation between the outputs of the 1×1 and 3×3 branches. In the 1×1 branch, 2D global average pooling, which is formulated as

$$z_c = \frac{1}{H \times W} \sum_{j}^{H} \sum_{i}^{W} x_c(i, j) \quad (3)$$

is used to encode global spatial information. The pooled output is nonlinearly processed by a *Softmax* function and then multiplied with the output of the 3×3 branch using matrix dot-product, yielding the first spatial attention map. Similarly, by performing 2D global average pooling in the 3×3 branch and multiplying with the output of the 1×1 branch, a second spatial attention map is obtained, which retains complete spatial positional information. Finally, these two spatial attention maps are aggregated and processed through a Sigmoid function to generate attention weights A_{final} used to weight the original feature map. The final output of MSA is represented as

$$X_{out} = X \otimes A_{final} \quad (4)$$

where $\otimes$ is element-wise multiplication. The output has the same size as the input feature map, and its cross-spatial learning approach effectively combines long-range dependencies and precise positional information, thereby enhancing feature representation capabilities.

3.4 Decoder and detection head

For the decoding part, we employ the classic SECOND (Yan et al., 2018) to perform further processing and preparation on the output features, with TransFusion (Bai et al., 2022) serving as the final detection head, which gives the final 2D and 3D bounding box results. We utilize FocalLoss as our classification and heatmap loss functions, and employ L1Loss as the bounding box regression loss function.

4. Experiments

4.1 Experimental settings

4.1.1 Dataset: We use Nuscenes (Caesar et al., 2020) for our 3D detection. The dataset covers diverse urban road scenarios and weather conditions. Each frame has six surround-view images at 1600×900 resolution and one 32-beam rotating LiDAR point cloud, synchronously capturing multi-modal data at 20Hz. It contains 1,000 driving sequences with approximately 20 seconds duration, including 400 k keyframes and 1.4 million precisely annotated 3D bounding boxes spanning 23 subcategories such as vehicles, pedestrians, and traffic cones. All annotations include 8 motion state attributes and 6 visibility levels. The dataset is divided into 700 training scenes,150 validation scenes and 150 testing scenes, ensuring coverage of environmental diversities. We use nuScenes detection score (NDS) and mean average precision (mAP) as evaluation metrics.

4.1.2 Implementation details: We utilize an NVIDIA RTX 4090 GPU with 24GB of VRAM and implement the model using PyTorch 1.10 on Ubuntu 20.04 under the open-sourced MMDetection3D. The input images from the cameras were downsampled to a resolution of 256×704 pixels. The LiDAR point clouds were voxelized at a resolution of 0.075m×0.075m ×0.2m to convert the raw data into a structured format suitable for processing and fusion with the camera data. This setup allowed for efficient training and evaluation of the proposed fusion-based perception model. We train the camera encoder branch and the LiDAR encoder branch separately, training each for 20 epochs using the resolutions mentioned above. Subsequently, we train the fusion module, freezing the weights of the camera encoder. This stage involves a training duration of 6 epochs, a batch size of 4, an initial learning rate of $2.5e^{-5}$, and utilizes a cosine annealing learning rate schedule. We also employ BEV space data augmentation following BEVFusion-MIT (Liu et al., 2023) for better result. No test-time augmentation is used during testing.

4.2 Detection result

Table 1 presents a comparison of the detection results of MSA-BEVFusion in NuScenes `val` set with mainstream methods. 'L' means LiDAR-only and 'L+C' stands for LiDAR-camera fusion.

Methods	Modality	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
PointPillars(Lang et al., 2019)	L	30.5	45.3	0.517	0.290	0.500	0.316	0.368
CenterPoint(Yin et al., 2021)	L	60.3	67.3	0.262	0.239	0.361	0.288	0.136
MVP(Yin et al., 2021)	L+C	66.4	70.5	0.263	0.238	0.321	0.313	0.134
FusionPainting(Xu et al., 2021)	L+C	67.3	70.7	**0.256**	0.236	0.346	0.274	**0.132**
TransFusion(Bai et al., 2022)	L+C	68.1	70.9	0.259	0.243	0.359	0.288	0.127
BEVFusion-MIT(Liu et al., 2023)	L+C	68.5	71.4	0.261	0.239	0.329	0.260	0.134
MSA-BEVFusion(Ours)	L+C	**68.9**	**71.6**	0.263	**0.235**	**0.318**	**0.258**	0.133

Table 1. MSA-BEVFusion in NuScenes `val` set compared with mainstream methods.

Benefiting from the multi-scale feature fusion facilitated by our designed MFPN module, as well as the multi-scale attention and cross-spatial properties of the MSA module, MSA-BEVFusion achieves relatively superior results on the NuScenes dataset. Our method improves mAP and NDS by 0.4% and 0.2% respectively, compared to BEVFusion-MIT, the state-of-the-art method without incorporating temporal context. For mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), and mean Average Velocity Error (mAVE), our method also ranks first among the mainstream methods compared. Figure 5 presents a set of visualization results under rainy night conditions. It is evident that even when the raw image and point cloud data are of very poor quality due to extremely challenging lighting and reflection conditions, our method still accurately detects most objects, fully demonstrating the robustness of our proposed method.

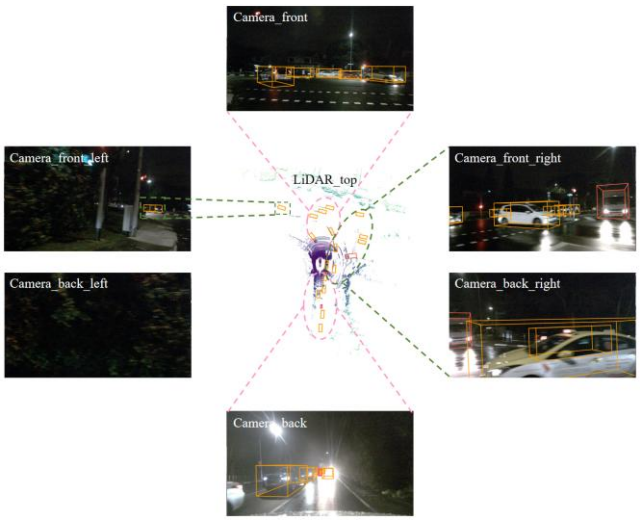

Figure 5. A frame of visualization result under night rainy condition

4.3 Ablation study

As previously stated, the MFPN and MSA modules we designed play significant roles in the superior performance of the model. In this section, we specifically demonstrate the individual contributions of these two modules through ablation studies. Table 2 summarizes the impact of different modules on the model's performance. Specifically, the baseline employs a standard FPN module as the camera feature neck and uses 3×3 convolution for feature fusion. The validation of the ablation studies was conducted on the NuScenes `val` set.

MFPN	MSA	mAP ↑	NDS ↑
		68.1	70.9
√		68.3	71.1
	√	68.4	71.3
√	√	68.9	71.6

Table 2. The impact of different modules on the model's performance in NuScenes `val` set.

Evidently, both the MFPN and MSA modules contribute significantly to the improvement of the mode's detection performance. When the MFPN module is used alone to optimize the encoding pipeline for camera features, the model's performance on the NuScenes `val` set shows an improvement of 0.2% in both mAP and NDS compared to the baseline. Using the MSA module alone to optimize the feature fusion module with multi-scale attention, mAP and NDS are improved by 0.3% and 0.4%, respectively. The full MSA-BEVFusion model shows even greater improvements of 0.8% in mAP and 0.7% in NDS compared to the baseline. This clearly highlights the positive impact of these two modules on enhancing detection performance and the rationale for combining them together for better performance.

5. Conclusion

In this paper, we present MSA-BEVFusion, a BEV object detection framework utilizing multi-scale attention for feature fusion. Our method decouples the interdependency between camera and LiDAR sensors present in traditional methods and innovatively utilizes the proposed MFPN and MSA modules for the decoding of camera features and the fusion of multimodal features. Our results demonstrate that the multi-scale feature

fusion and multi-scale attention approaches have a positive impact on the effectiveness and robustness of the model. We hope this work will advance further exploration into robust multimodal fusion techniques in the field of autonomous driving and other fields.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 42271343) and the Open Project Funds for the Joint Laboratory of Spatial Intelligent Perception and Large Model Application (Grant No. SIPLMA-2024-YB-06).

References

Philion, J., & Fidler, S., 2020: 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* (pp. 194-210). Springer International Publishing.

Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., ... & Li, Z., 2023: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 37, No. 2, pp. 1477-1485).

Zhang, X., Bai, L., Zhang, Z., & Li, Y., 2022: Multi-scale keypoints feature fusion network for 3d object detection from point clouds. *Hum.-Cent. Comput. Inf. Sci*, *12*, 12-29.

Vora, S., Lang, A. H., Helou, B., & Beijbom, O., 2020: Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4604-4612).

Huang, T., Liu, Z., Chen, X., & Bai, X., 2020: Epnet: Enhancing point features with image semantics for 3d object detection. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XV 16* (pp. 35-52). Springer International Publishing.

Yoo, J. H., Kim, Y., Kim, J., & Choi, J. W., 2020: 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16* (pp. 720-736). Springer International Publishing.

Pang, S., Morris, D., & Radha, H., 2020: CLOCs: Camera-LiDAR object candidates fusion for 3D object detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 10386-10393). IEEE.

Huang, J., Huang, G., Zhu, Z., Ye, Y., & Du, D.,2022: *BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View* (No. arXiv:2112.11790).

Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., & Li, Z., 2022: *BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection* (No. arXiv:2206.10092).

Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D. L., & Han, S., 2023: BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. *2023 IEEE International Conference on Robotics and Automation*, 2774–2781.

Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., & Tang, Z., 2022: *BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework* (No. arXiv:2205.13790).

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., & Dai, J., 2025: BEVFormer: Learning Bird's-Eye-View Representation From LiDAR-Camera via Spatiotemporal Transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(3), 2020–2036.

Hu, J., Shen, L., & Sun, G., 2018: Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S., 2018: CBAM: Convolutional Block Attention Module. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11211, pp. 3–19). Springer International Publishing.

Hou, Q., Zhou, D., & Feng, J., 2021: Coordinate Attention for Efficient Mobile Network Design. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13708–13717.

Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., & Huang, Z., 2023: Efficient Multi-Scale Attention Module with Cross-Spatial Learning. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B., 2021: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision*, 9992–10002.

Graham, B., Engelcke, M., & Maaten, L. V. D., 2018: 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9224–9232.

Yan, Y., Mao, Y., & Li, B., 2018: SECOND: Sparsely embedded convolutional detection. *Sensors*, *18*(10), 3337.

Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., & Tai, C.-L., 2022: TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1080–1089.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuScenes: A Multimodal Dataset for Autonomous Driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11618–11628.

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O, 2019: PointPillars: Fast encoders for object detection from point clouds. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12689–12697.

Yin, T., Zhou, X., & Krahenbuhl, P, 2021: Center-based 3D object detection and tracking. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11779–11788.

Yin, T., Zhou, X., & Krähenbühl, P, 2021: Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, *34*, 16494-16507.

Xu, S., Zhou, D., Fang, J., Yin, J., Bin, Z., & Zhang, L, 2021: Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *2021 IEEE International Intelligent Transportation Systems Conference* (pp. 3047-3054). IEEE.