

Adaptive Cross-Attention Gated Network for Radar-Camera Fusion in BEV Space

Ji-Yong Lee*, **, Jae-Hyeok Lee*, Dong-oh Kang*

*Visual Intelligent Lab., ETRI, Daejeon, Republic of Korea

**The University of Texas at Austin, Austin, TX 78712, USA

jyiyong.lee@etri.re.kr, heuyklee@etri.re.kr, dongoh@etri.re.kr

Abstract— Fusing multimodal sensors for 3D object detection has been extensively researched in the field of autonomous driving. However, existing multimodal sensor fusion methods still struggle to provide reliable detection across different modalities under diverse environmental conditions. Specifically, straightforward methods like summation or concatenation in radar-camera fusion may lead to spatial misalignment and fail to localize objects in complex scenes. To address this, we propose Adaptive Cross-Attention Gated Network (ACAGN) to enhance radar-camera fusion capabilities in Bird's-Eye View (BEV) space. Our approach integrates a deformable cross-attention and an adaptive gated network mechanism. The deformable cross-attention aligns radar and camera features from BEV with greater spatial precision, handling variations between those features effectively. Meanwhile, the adaptive gated network dynamically filters and prioritizes the most relevant information from each sensor. This dual approach improves stability and robustness of detection, as demonstrated through extensive evaluations on the nuScenes dataset.

Keywords— Autonomous driving, multi-modal sensor fusion, millimeter-wave radar, multi-view camera, 3D object detection

I. INTRODUCTION

In the field of autonomous driving, the most commonly used sensors are camera, radar, and LiDAR for 3D detection in perception systems [25][32]. Among various sensor fusion modalities, LiDAR combined with either camera or radar (or both) has consistently achieved top performance compared to other standalone or fusion combinations, such as camera and radar fusion [1][3][16][27]. This is due to LiDAR's unique capability to capture detailed and sophisticated geometric information, which significantly enhances 3D object detection accuracy and improves precise motion tracking for surrounding objects [2].

However, LiDAR sensors are expensive and require maintenance. While LiDAR provides depth and structural data that contribute to high performance autonomous systems, its high expense poses challenges for large-scale deployment in consumer vehicles [32]. Consequently, radar-camera fusion has emerged as an alternative that offers cost effectiveness and robustness across diverse environmental conditions, including adverse weather such as snow, fog, and rain. From recent works [26][34], the RC fusion leverages the strengths of radar for reliable depth and velocity estimation and those of cameras for capturing color, texture, and semantic detail.

BEVFusion [17] has demonstrated remarkable performance by fusing multimodal features into a unified Bird's Eye View (BEV) representation. Even though recent BEV based radar-camera fusion methods [28] have achieved outstanding results due to their ability to preserve the geometric structure and semantic density, there still remains room for improvement in the fusion process itself. Instead of relying on straightforward fusion operations like summation or concatenation, as used in BEVFusion [17], more advanced fusion techniques are required to resolve spatial correspondences between images and radar inputs, particularly in complex scenes.

To mitigate this issue, we propose the Adaptive Cross Attention Gated Network to enhance radar-camera fusion in BEV space, aiming to improve upon existing fusion methods [28]. Built upon the baseline framework RCBEVDet [16], our proposed method integrates an adaptive gated network [7][22][30] for further refinement. This network selectively focuses on the most relevant features from each sensor modality based on contextual and environmental cues. This integrated mechanism not only improves alignment and coherence of cross-attended features, but also dynamically filters and prioritizes critical information through the gated network. This provides stable and robust fusion in diverse driving scenarios and complex environments for perception tasks. This paper presents an enhanced radar-camera fusion method called Adaptive Cross-Attention Gated Network (ACAGN) for 3D object detection. Our contributions are summarized as follows:

- We introduce ACAGN, integrating cross-attention and an adaptive gated network to enhance stability and robustness for multimodal sensor fusion.
- We design a fusion mechanism where the output features from the cross-attention module are fed into an adaptive gating network that allows to weight and combine features effectively from different sensors.
- We evaluate our proposed enhanced fusion method on nuScenes dataset [2] to verify the effectiveness of ACAGN. This achieves NDS 56.9 and mAP 45.5.

II. RELATED WORKS

A. Multimodal 3D Object Detection

Compared to single-modality detectors, multi-modal 3D object detectors tend to provide better accuracy and robustness,

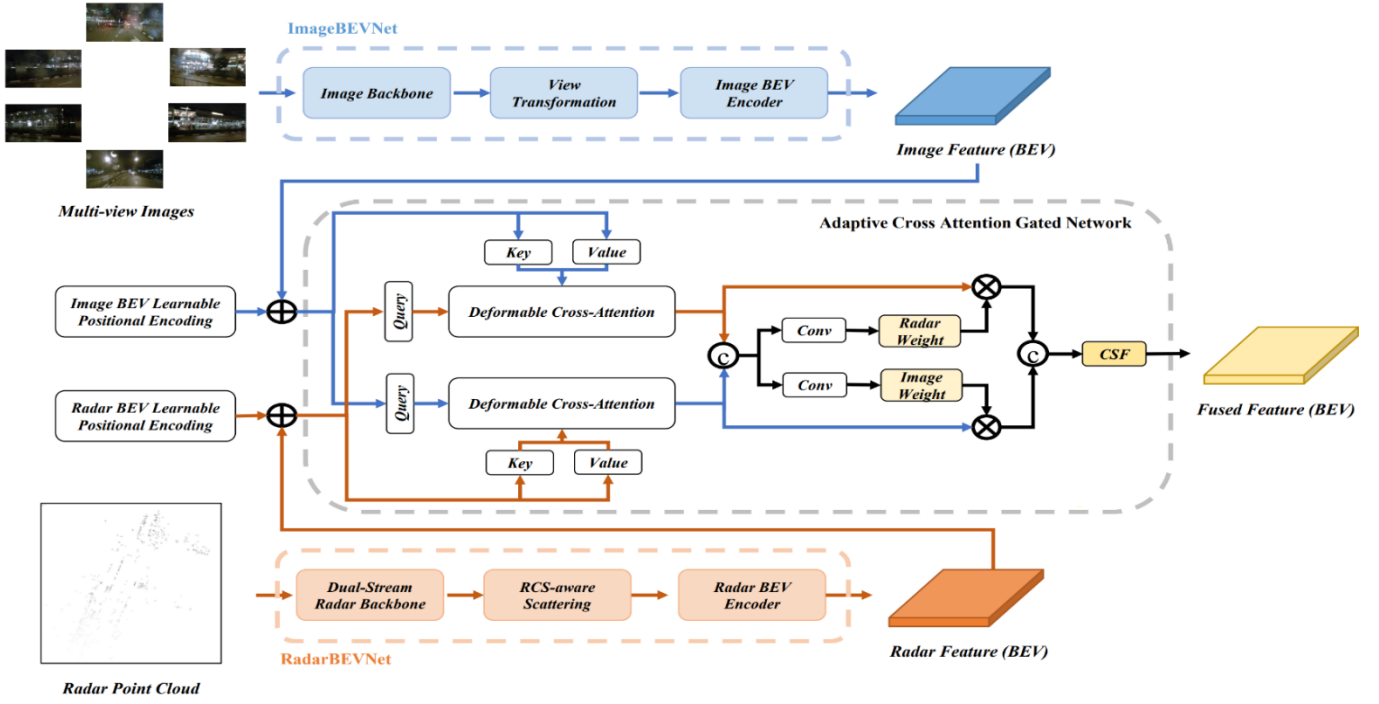


Figure 1. An overview of the proposed **Adaptive Cross Attention Gated Network (ACAGN)** fusion method that enhances sensor fusion in BEV space using radar and camera inputs. This framework builds on the RCBEVDet baseline [16] by incorporating cross-attention and adaptive gating mechanisms to fuse information effectively and dynamically emphasize relevant information from each modality. Finally, the fused BEV output is produced by a CSF module that consists of several CBR (Conv 3×3 , BatchNorm, and ReLU) blocks.

as perceptual sensors like LiDAR, cameras, and radar offer complementary information [25]. In particular, with LiDAR's ability to capture detailed geometric structures more precisely than others, models that incorporate LiDAR consistently achieve leading performance among common sensor setups on most existing datasets [2][3][4][18]. However, deploying LiDAR is challenging due to its substantial cost for consumer cars [32]. Therefore, cost-effective radar-camera (RC) fusion approach stands out as an alternative.

B. Radar-Camera Fusion

In recent years, incorporating data of different modalities into a shared Bird's Eye View (BEV) representation has gained traction for multi-sensor fusion in 3D perception tasks. Despite advancements in projecting multiple sensor features into BEV space, further exploration of fusion methods is needed to combine these BEV features effectively. Generic fusion approaches often use simple elementwise concatenation or summation of sensor features before feeding into a BEV encoder [17]. CRKD [32], on the other hand, uses an adaptive gated network to fuse BEV sensor features. CRN [9] and BEV-Radar [33] utilize deformable cross-attention [14] to achieve spatial alignment between radar and camera BEV features. The Unified Feature Fusion (UFF) module from UniBEVFusion [31] combines a channel unifier, shared feature encoder, and fused feature encoder to retain high-quality information from each modality. Additionally, RCBEVDet [15][16] integrates deformable cross-attention with CBR blocks to create fused BEV from image and radar features, whereas RCFusion [34] applies CBAM [24] with CBR blocks for fusion.

Inspired by these approaches, our proposed method combines deformable cross-attention [14] with an adaptive gated network [7][22][30], marking the first attempt at an enhanced fusion approach that effectively integrates features from radar and camera BEV representations.

III. PROPOSED METHOD

A. Pre-processing

The overview of ACAGN is shown in Fig. 1. Following the RCBEVDet pipelines [16], multi-view image features are extracted from an image encoder. These features are then passed through a view transform module to integrate information from multiple camera views and project onto a top-down map of the surrounding area. The transformed features are subsequently processed by a BEV encoder to obtain the BEV features.

Concurrently, radar features are extracted through a dual-stream radar backbone which captures local and global radar information in early stage. This backbone consists of two streams: a point-based encoder for local feature and a transformer-based encoder for global context. These features are then passed through the RCS-aware scattering module to adjust radar's object-specific reflection strengths. Final radar BEV features are obtained using a radar BEV encoder.

The processed image and radar BEV features are fused through an adaptive cross-attention gated network module for the 3D object detection tasks.

B. Adaptive Cross Attention Gated Network

TABLE I. COMPARISON ON THE NUSCENES VALIDATION SPLIT. * DENOTES REPRODUCED AND THE BEST IS **BOLDED**. **C** AND **R** REFER TO CAMERA AND RADAR, RESPECTIVELY.

Method	Modality	Backbone	Resolution	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVDet [6]	C	R50	256 × 704	39.2	31.2	0.691	0.272	0.523	0.909	0.247
BEVDepth [13]	C	R50	256 × 704	47.5	35.1	0.639	0.267	0.479	0.428	0.198
SOLOFusion [20]	C	R50	256 × 704	53.4	42.7	0.567	0.274	0.411	0.252	0.188
StreamPETR [23]	C	R50	256 × 704	54.0	43.2	0.581	0.272	0.413	0.295	0.195
CenterFusion [19]	C+R	DLA34	448 × 800	45.3	33.2	0.649	0.263	0.535	0.540	0.142
CRAFT [8]	C+R	DLA34	448 × 800	51.7	41.1	0.494	0.276	0.454	0.486	0.176
RCBEV [35]	C+R	SwinT	256 × 704	48.2	37.7	0.534	0.271	0.558	0.493	0.209
BEVFusion [17]	C+R	SwinT	256 × 704	54.1	43.2	0.489	0.269	0.512	0.313	0.171
CRKD student [32]	C+R	SwinT	256 × 704	55.9	44.9	0.464	0.267	0.458	0.304	0.165
CRN [10]	C+R	R50	256 × 704	56.0	49.0	0.487	0.277	0.542	0.344	0.197
RCBEVDet* [16]	C+R	R50	256 × 704	56.7	45.2	0.496	0.276	0.407	0.222	0.190
ACAGN (Ours)	C+R	R50	256 × 704	56.9	45.5	0.494	0.278	0.409	0.218	0.190

Traditional sensor fusion involves concatenating the BEV features of different modalities and then passing them into attention or convolutional modules to align their spatial relationships. However, along with being sparse and noisier compared to LiDAR, radar typically has low angular resolution, particularly in azimuth direction. Hence, this can potentially lead to spatial misalignment of BEV features between images and radar after processing. Furthermore, it may inaccurately distinguish between spaced objects and fail to localize objects in cluttered environments. To address this issue, we introduce a module comprising cross-attention, gated network blocks, and CBR blocks for final fusion to dynamically align multimodal BEV features.

1) Deformable Cross-Attention We conducted deformable cross-attention [14] to selectively focus on spatial regions for sensor features. Radar and camera BEV features denoted by $F_r \in \mathbb{R}^{C_r \times H \times W}$ and $F_c \in \mathbb{R}^{C_c \times H \times W}$, where H and W represent the height and width of the BEV feature map, and C_r and C_c denote the number of channels for the radar and camera features, respectively. To incorporate spatial positional information, we first add learnable positional encodings to both radar and camera BEV features:

$$\begin{aligned}\tilde{F}_r &= F_r + P_r, \\ \tilde{F}_c &= F_c + P_c,\end{aligned}$$

where P_r and P_c represent the learnable positional encodings for the radar and camera features, respectively. Next, $\tilde{F}_r$ is transformed into queries z_{q_r} and reference points p_{q_r} , while $\tilde{F}_c$ is used as keys and values in the cross-attention mechanism. The multi-head deformable cross-attention [36] is defined as follows:

$$\text{DeformAttn}(z_{q_r}, p_{q_r}, \tilde{F}_c) = \sum_{m=1}^M W_m \left[\sum_{k=1}^K A_{mqk} \cdot W'_m \tilde{F}_c(p_{q_r} + \Delta p_{mqk}) \right]$$

, where M is the number of attention heads, W_m is the weight matrix for each head, A_{mqk} denotes the attention weight calculated by z_{q_r} , W'_m is a linear transformation on $\tilde{F}_c$, and $\tilde{F}_c(p_{q_r} + \Delta p_{mqk})$ refers to the feature at a spatial offset Δp_{mqk} from the p_{q_r} within the camera BEV map. Similarly, $\tilde{F}_c$ is

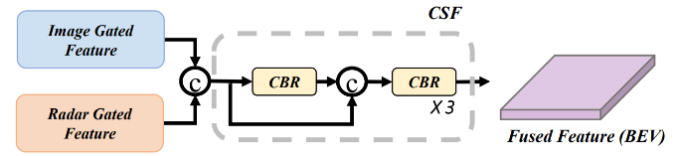


Figure 2. Channel and Spatial Fusion (CSF) module. The concatenated features of gated image and camera inputs are processed through the CSF module, consisting of several CBR (Conv (3 × 3), BatchNorm, ReLU) blocks, to produce the fused BEV feature.

used as the query to extract features from radar BEV by exchanging roles.

The cross-attention mechanism can therefore be described as:

$$\begin{cases} \tilde{F}_c \leftarrow \text{DeformAttn}(z_{q_r}, p_{q_r}, \tilde{F}_c) \\ \tilde{F}_r \leftarrow \text{DeformAttn}(z_{q_c}, p_{q_c}, \tilde{F}_r) \end{cases}$$

2) Gated Network Next, we added a gated network [7][22][30][32] to perform sensor fusion on the radar and image outputs. Specifically, the gated fusion network produces attention maps by applying 3×3 convolution followed by a sigmoid activation to the concatenated inputs. These maps are multiplied element-wise with the radar and camera features. The following formulas use the attention maps to gate the radar and camera features:

$$F_{r_g} = \tilde{F}_r \times \underbrace{\sigma(\text{Conv}_r(\tilde{F}_r \oplus \tilde{F}_c))}_{\text{RadarAttentionMap}},$$

$$F_{c_g} = \tilde{F}_c \times \underbrace{\sigma(\text{Conv}_c(\tilde{F}_c \oplus \tilde{F}_r))}_{\text{CameraAttentionMap}},$$

where F_{r_g} and F_{c_g} are the gated features for radar and camera respectively. Generated from the attention maps, the sigmoid function produces values between 0 and 1. They are used to emphasize the relative importance of each modality features while diminishing the less relevant ones. By applying these values as weights into the original input feature maps, the network can adjust the strength of each feature channel dynamically.

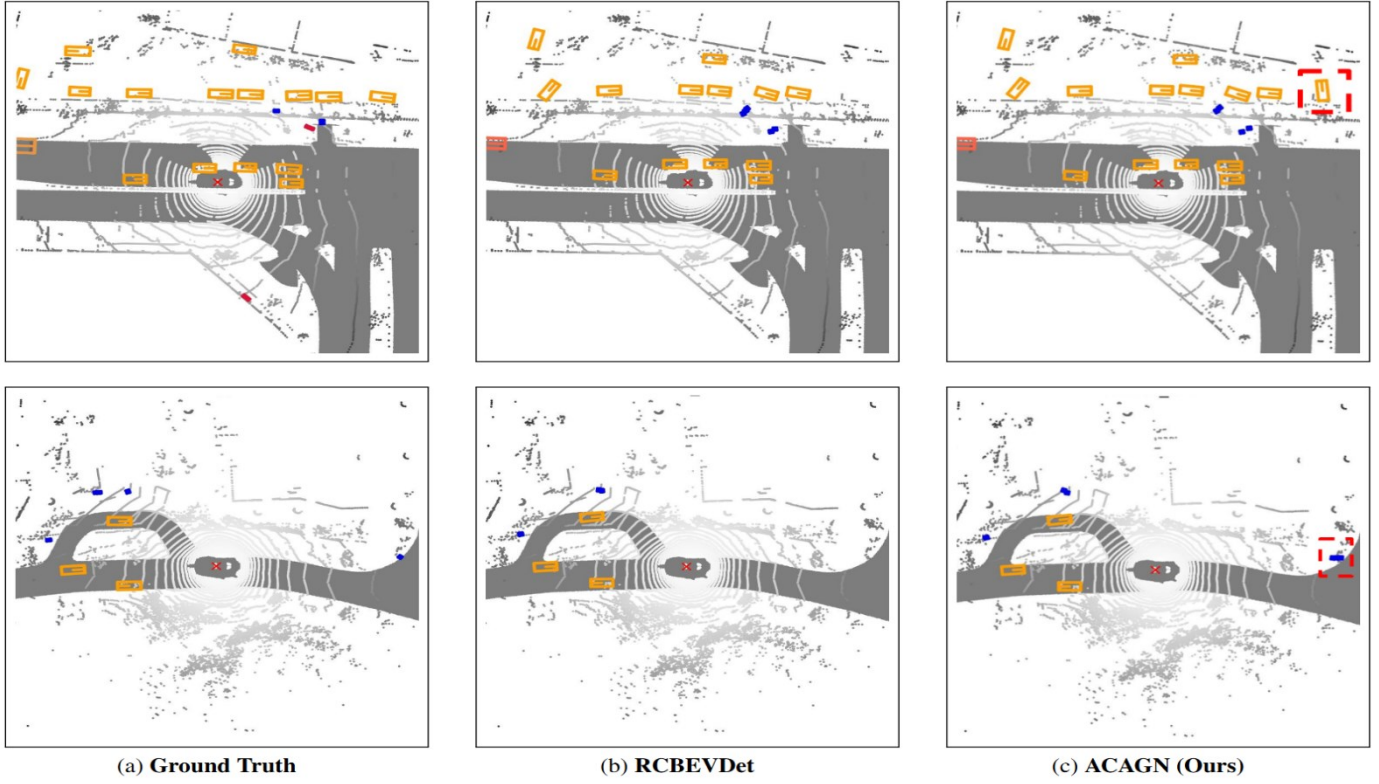


Figure 3. Qualitative analysis of detection results on nuScenes. From left to right: Ground Truth, RCBEVDet, and ACAGN. **Orange** boxes indicate cars, **red** boxes denote construction vehicle, and **blue** boxes represent pedestrians. **Red dotted** boxes indicate difference detection results between RCBEVDet and ACAGN.

3) Channel and Spatial Fusion Finally, the gated features F_{rg} and F_{cg} are concatenated and further fused by the Channel and Spatial Fusion (CSF) module [16], as shown in Fig. 2. Specifically, the concatenated features are passed through a CBR block (Convolution 3×3 , BatchNorm, and ReLU) with a residual connection followed by three additional CBR blocks to produce the final fused feature. This process can be formulated as:

$$F_{rc} = \text{CSF}(F_{rg} \oplus F_{cg}),$$

where F_{rc} represents the fused output feature map integrated from the gated radar and camera features.

IV. EXPERIMENTS

A. Experimental Setup

We evaluated our proposed Adaptive Cross Attention Gated Network (ACAGN) on the nuScenes dataset [2]. This dataset is a large-scale benchmark for autonomous driving 3D perception tasks that provide official split for 700 scenes for training and 150 scenes for validation.

For evaluation, we used two primary metrics: mean Average Precision (mAP) and the nuScenes Detection Score (NDS). To comprehensively assess detection quality and accuracy, we report five True Positive (TP) metrics as well: Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and

Average Attribute Error (AAE). The formula for NDS is as follows,

$$NDS = \frac{1}{10} \left[5 \text{mAP} + \sum_{\text{mTP} \in \mathcal{TP}} (1 - \min(1, \text{mTP})) \right]$$

, where $\mathcal{TP}$ is the set of the five mean True Positive metrics.

Our implementation is based on RCBEVDet codebase [16]. We utilized standard loss functions for 3D detection, specifically Gaussian Focal Loss [29] for classification and L1 Loss [12] for bounding box regression.

We used 4 NVIDIA A40 GPUs with 46GB of memory per GPU for all experiments. We set the point cloud range $[-51.2\text{m}, -51.2\text{m}, -5.0\text{m}] \times [51.2\text{m}, 51.2\text{m}, 5.0\text{m}]$ to ensure comprehensive coverage of 3D space. Following baseline's training pipeline, models need to be trained in two-stage process. In the first stage, we train the camera stream independently as outlined in [13] to learn image-based features. Secondly, we train the radar-camera fusion model with the parameters of the camera stream frozen.

We augmented the data using a combination of random sizing $[-0.06, 0.11]$, random rotation $[-5.4^\circ, 5.4^\circ]$, horizontal flipping with a probability of 0.5, and global rotation and scaling for radar features. In addition, we used multi-sweep radar fusion with up to 8 sweeps per frame. We set camera backbone network as ResNet R50 [5], image resolution as 256×704 , and the BEV size as 128×128 . We used AdamW [11] optimizer with a learning rate of $1e-4$, weight decay of $1e-2$

TABLE II. EVALUATION OF THE ACAGN'S ABLATION STUDY ON THE nuScenes VALIDATION SPLIT.

Method	CSF	Deform Cross Attention	Gated Network	NDS $\uparrow$	mAP $\uparrow$	mATE $\downarrow$	mASE $\downarrow$	MAOE $\downarrow$	mAVE $\downarrow$	mAAE $\downarrow$
ACAGN	$\checkmark$	-	-	56.5	45.3	0.500	0.279	0.408	0.238	0.193
	$\checkmark$	$\checkmark$	-	56.7	45.2	0.496	0.276	0.407	0.222	0.190
	$\checkmark$	$\checkmark$	$\checkmark$	56.9	45.5	0.494	0.278	0.409	0.218	0.190

TABLE III. PERFORMANCE ANALYSIS BY WEATHER AND LIGHTING CONDITIONS ON THE nuScenes VAL DATASET.

Method	Modality	NDS $\uparrow$			mAP $\uparrow$		
							
RCBEVDet [16]	C+R	56.49	34.40	59.50	45.29	25.33	47.14
ACAGN	C+R	56.77 (+0.28)	34.46 (+0.06)	59.88 (+0.38)	45.62 (+0.33)	25.45 (+0.12)	47.78 (+0.64)

, and trained for 15 epochs using a cyclic learning rate policy [21].

B. Quantitative Results

As shown in Tab. 1, we present an overall comparison of ACAGN with existing camera-only (C) and camera-radar (C+R) modality detectors on nuScenes [2] validation dataset. Our proposed ACAGN fusion method enhances the baseline, RCBEVDet [6], and achieves top performance across multiple evaluation metrics.

Specifically, ACAGN achieves an NDS of 56.9 (0.2 $\uparrow$) and a mAP of 45.5 (0.3 $\uparrow$) that surpass the reproduced baseline. Furthermore, it also achieves a mAVE of 0.218 which outperforms other state-of-the-art methods. This improvement demonstrates the effectiveness of our proposed method in enhancing multimodal fusion.

C. Ablation Studies

To analyze the contribution of each component, we conducted ablation studies on the nuScenes [2] validation split, as shown in Tab. 2. The results report the impact of key fusion components which include CSF, Deformable cross-attention, and gated network on detection performance. We set a baseline with CSF only, as it simply concatenates two BEV features from camera and radar modalities before feeding into CBR blocks. Consequently, adding deformable cross-attention and gate network sequentially improves model performance across key metric. Specifically, adopting the gated network further refines the fusion process by adaptively weighting the contributions of camera and radar features. It results in improvements in most of all metric except for mASE and mAOE.

D. Qualitative Results

In Fig. 3, we present visualization of 3D detection results to demonstrate effectiveness of the proposed multimodal fusion method, Adaptive Cross-Attention Gated Network (ACAGN). As shown, the visual detection results of ACAGN and RCBEVDet are mostly identical but ACAGN has a higher alignment with ground truth.

E. Weather Robustness Analysis

We evaluated the performance of ACAGN in various weather conditions on the nuScenes validation dataset, as

shown in Tab. 3. This table compares ACAGN with the baseline RCBEVDet [16] using radar and camera modalities (C+R) under day, night, and rain weather scenarios. With consistent improvements across diverse weather scenes, it shows significant improvements in rainy conditions compared to daytime and nighttime. Specifically, ACAGN achieves NDS of 59.88 and mAP of 47.78 in rainy conditions, with huge improvements of 0.38 and 0.64, respectively. These results demonstrate notable performance of ACAGN in clear and low-light conditions while preserving competitive results in adverse weather.

V. CONCLUSION

In this paper, we have introduced the Adaptive Cross-Attention Gated Network (ACAGN) to improve radar-camera fusion for 3D object detection in BEV space. Building upon the RCBEVDet baseline, our method integrates deformable cross-attention and an adaptive gated network to dynamically align and prioritize multimodal features. The deformable cross-attention mechanism facilitates dynamic alignment of radar and camera features that allows the model to focus on relevant regions spatially and handle the inherent discrepancies between these modalities. Meanwhile, the adaptive gated network selectively filters and prioritizes the most informative features from each sensor dynamically. This integrated mechanism addresses the challenges of spatial misalignment and noise in radar-camera fusion, and improves the accuracy of multimodal feature fusion. Furthermore, ACAGN achieved state-of-the-art performance on the nuScenes dataset, with NDS of 56.9 and mAP of 45.5, demonstrating its effectiveness in autonomous driving scenarios.

ACKNOWLEDGMENT

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-II220907, Development of AI Bots Collaboration Platform and Self-organizing AI).

REFERENCES

- [1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers, 2022.

- [2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2020.
- [3] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. <https://github.com/open-mmlab/mmdetection3d>, 2020.
- [4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 3354–3361, 2012.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [6] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view, 2022.
- [7] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.
- [8] Youngseok Kim, Sanmin Kim, Junwon Choi, and Dongsuk Kum. Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer. In AAAI Conference on Artificial Intelligence, 2022.
- [9] Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. Crn: Camera radar net for accurate, robust, efficient 3d perception. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 17615–17626, 2023.
- [10] Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. Crn: Camera radar net for accurate, robust, efficient 3d perception, 2023.
- [11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- [12] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12697–12705, 2019.
- [13] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection, 2022.
- [14] Zhiqi Li, Wenhao Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers, 2022.
- [15] Zhiwei Lin, Zhe Liu, Yongtao Wang, Le Zhang, and Ce Zhu. Rcbvdet++: Toward high-accuracy radar-camera fusion 3d perception network, 2024.
- [16] Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. Rcbvdet: Radar-camera fusion in bird's eye view for 3d object detection, 2024.
- [17] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation, 2024.
- [18] Jieru Mei, Alex Zihao Zhu, Xinchun Yan, Hang Yan, Siyuan Qiao, Yukun Zhu, Liang-Chieh Chen, Henrik Kretschmar, and Dragomir Anguelov. Waymo open dataset: Panoramic video panoptic segmentation, 2022.
- [19] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), page 1526–1535. IEEE, 2021.
- [20] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection, 2022.
- [21] Leslie N. Smith. Cyclical learning rates for training neural networks. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 464–472, 2017.
- [22] Jingyu Song, Lingjun Zhao, and Katherine A. Skinner. Lirafusion: Deep adaptive lidar-radar fusion for 3d object detection, 2024.
- [23] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection, 2023.
- [24] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018.
- [25] L. xilinx Wang, Xinyu Zhang, Ziyang Song, Jiangfeng Bi, Guoxin Zhang, Haiyue Wei, Liyao Tang, Lei Yang, Jun Li, Caiyan Jia, and Lijun Zhao. Multi-modal 3d object detection in autonomous driving: A survey taxonomy. *IEEE Transactions on Intelligent Vehicles*, 8:3781–3798, 2023.
- [26] Weiyei Xiong, Jianan Liu, Tao Huang, Qing-Long Han, Yuxuan Xia, and Bing Zhu. Lxl: Lidar excluded lean 3d object detection with 4d imaging radar and camera fusion. *IEEE Transactions on Intelligent Vehicles*, 9(1):79–92, 2024.
- [27] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction, 2022.
- [28] Shanliang Yao, Runwei Guan, Xiaoyu Huang, Zhuoxiao Li, Xiangyu Sha, Yong Yue, Eng Gee Lim, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu, and Yutao Yue. Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *IEEE Transactions on Intelligent Vehicles*, 9(1):2094–2128, 2024.
- [29] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11784–11793, 2021.
- [30] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3D-CVF: Generating Joint Camera and LiDAR Features Using Cross-view Spatial Feature Fusion for 3D Object Detection, page 720–736. Springer International Publishing, 2020.
- [31] Haocheng Zhao, Runwei Guan, Taoyu Wu, Ka Lok Man, Limin Yu, and Yutao Yue. Unibevfusion: Unified radar-vision bevfusion for 3d object detection, 2024.
- [32] Lingjun Zhao, Jingyu Song, and Katherine A. Skinner. Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15470–15480, 2024.
- [33] Yuan Zhao, Luting Zhang, Jiajun Deng, and Yanyong Zhang. Bev-radar: Bidirectional radar-camera fusion for 3d object detection. *JUSTC*, 2023.
- [34] Lianqing Zheng, Sen Li, Bin Tan, Long Yang, Sihao Chen, Libo Huang, Jie Bai, Xichan Zhu, and Zhixiong Ma. Rcfusion: Fusing 4-d radar and camera with bird's-eye view features for 3-d object detection. *IEEE Transactions on Instrumentation and Measurement*, 72:1–14, 2023.
- [35] Taohua Zhou, Yining Shi, Junjie Chen, Kun Jiang, Mengmeng Yang, and Diange Yang. Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection, 2022.
- [36] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021.



Ji-Yong Lee received his B.S. degree in Electrical and Computer Engineering at the University of Texas at Austin in 2021 and is currently pursuing a master's degree in Artificial Intelligence. He is currently a research associate at the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. His research interests include machine learning, knowledge distillation, and multi-modal sensor fusion.



Jae-Hyeok Lee received the Ph.D. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2024. He is currently a researcher at the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. His current research interests include machine learning, computer vision, 3D vision, and interpretable deep learning.



Dong-oh Kang received his B.S. degree in electronic engineering from Yonsei University, Korea, in 1994, and his M.S. degree and Ph.D. in electronic engineering from Korea Advanced Institute of Science and Technology (KAIST), Korea, in 1996 and 2001, respectively. Since 2001, he has been working at Electronics and Telecommunications Research Institute (ETRI) and is currently a principal researcher of the Visual Intelligence Research Section. His research interests include multimodal knowledge representation, knowledge graph completion, and self-growing intelligent agents.