

Radar Enlighten the Dark: Enhancing Low-Visibility Perception for Automated Vehicles with Camera-Radar Fusion

Can Cui, Yunsheng Ma, Juanwu Lu and Ziran Wang

Abstract—Sensor fusion is a crucial augmentation technique for improving the accuracy and reliability of perception systems for automated vehicles under diverse driving conditions. However, adverse weather and low-light conditions remain challenging, where sensor performance degrades significantly, exposing vehicle safety to potential risks. Advanced sensors such as LiDARs can help mitigate the issue but with extremely high marginal costs. In this paper, we propose a novel transformer-based 3D object detection model “REDFormer” to tackle low visibility conditions, exploiting the power of a more practical and cost-effective solution by leveraging bird’s-eye-view camera-radar fusion. Using the nuScenes dataset with multi-radar point clouds, weather information, and time-of-day data, our model outperforms state-of-the-art (SOTA) models on classification and detection accuracy. Finally, we provide extensive ablation studies of each model component on their contributions to address the above-mentioned challenges. Particularly, it is shown in the experiments that our model achieves a significant performance improvement over the baseline model in low-visibility scenarios, specifically exhibiting a 31.31% increase in rainy scenes and a 46.99% enhancement in nighttime scenes. The source code of this study is publicly available¹.

I. INTRODUCTION

Sensor fusion, also known as sensor integration, refers to the process of combining data from various sensors installed in an automated vehicle to obtain a comprehensive and accurate understanding of the surrounding environment [1], it is also a crucial component in digital twin technology [2]. This approach allows the vehicle to obtain a more comprehensive view of the environment, critical for making informed decisions and navigating safely. The sensors commonly used in automated vehicle sensor integration include LiDAR, radar, camera, and global navigation satellite system (GNSS). Integrating information from multiple sensors allows automated vehicles to operate more effectively and safely in various driving conditions.

However, one major challenge in sensor fusion applications is the underperformance of sensors in low visibility environments. Common causes for visibility reduction include adverse weathers, such as rain, snow, and low-light conditions (like driving at night). As a result, automated vehicles that rely on these inaccurate detection results pose substantial risks, which might expedite serious traffic accidents or other devastating consequences. Therefore, there is a need to develop a sensor fusion system that is reliable in

Can Cui, Yunsheng Ma, Juanwu Lu and Ziran Wang are with Lyles School of Civil Engineering, Purdue University, West Lafayette, IN 47906, USA. Email: {cancui, yunsheng, juanwu, ziran@purdue.edu}

¹<https://github.com/PurdueDigitalTwin/REDFormer>

low visibility and can work effectively in diverse practical conditions.

A number of existing research prefers fusing LiDAR and camera sensors due to their complementary nature. However, such a solution comes with the potential limitations [3]:

- **Costliness:** LiDAR can be very expensive compared to other sensors on automated vehicles.
- **Environmental Sensitivity:** Environmental factors such as rain, snow, or fog can reduce LiDAR’s accuracy.
- **High Computations:** LiDAR generates large amounts of data, which can be computationally intensive to process and analyze.

On the contrary, radar sensors can effectively detect objects under a broader range of environmental conditions and are generally less expensive than LiDARs. Therefore, it is more reasonable to investigate sensor fusion of radar and camera data to achieve the 3D object detection task in a practical context. Compared to relying solely on radar sensors, camera-radar fusion offers numerous benefits, including improved object detection by combining high-resolution images from cameras with the ability to be unaffected by environmental conditions of radar to detect objects even in low-light or adverse weather conditions. Particularly, advanced machine learning algorithms such as Transformers empower the perception capability of automated vehicles by its computational efficiency and scalability [4], and allowing us to train multi tasks in one joint model [5]. Such systems are more robust and provide a complete awareness of the environment, contributing to avoiding potential hazards and accidents. Moreover, camera-radar fusion systems are more accessible and cost-effective for all vehicles, promoting the overall penetration rate.

This paper proposes a 3D object detection model applying sensor fusion on multi-camera images and multi-radar point clouds in a bird’s-eye-view (BEV) perspective. Inspired by achievements in natural language processing, we learn the positional representation of multi-radar point clouds using an embedding mechanism and gated linear filters. We achieve spatial-temporal fusion using the attention mechanism with image embedding, radar point cloud embedding, and BEV positional queries. Additionally, we design a multi-task objective to integrate visibility conditions into the end-to-end training procedure, aiming to enable our model to understand different low-visibility conditions and their corresponding weather and time-of-day (TOD) properties comprehensively.

The main contributions of this paper are summarized as follows:

- A novel radar embedding backbone is proposed for camera-radar fusion that significantly improves 3D object detection accuracy compared to using images solely.
- A multi-task learning (MTL) paradigm is developed to incorporate weather and TOD information as additional contextual clues during training.
- Extensive experiments on a 3D object detection benchmark are conducted. The results demonstrate that our model outperforms the state-of-the-art approaches, especially in adverse weather conditions and low-light environments.

II. RELATED WORKS

A. Object Detection

Problem of object detection is a computer vision task that aims to identify and localize objects within an image. During the past decade, the developments of machine learning technologies have led to emerging interests in research on 2D object detection, with a multitude of works in this field [6], [7]. Nevertheless, 2D object detection has limitations in representing the depth of view of the context. As a result, more recent research focuses on designing models for 3D object detection. FCOS3D [8] enhanced a 2D detector FCOS [9] and directly generated 3D bounding boxes. Inspired by DETR [10], DETR3D [11] extracts 2D features from multiple camera images and associates them with 3D positions using sparse 3D object queries and camera transformation matrices. Using multi-camera image inputs, M²M [12] achieves 3D object detection and map segmentation in the BEV space. Other methods [13] directly use multi-layer perception to learn how to project images input from the camera to the BEV plane. BEVformer [14] uses a deformable transformer to compute spatial and temporal features in BEV grid regions of interest across camera views and previous BEV information. Besides, existing methods have also investigated how to incorporate radar information to enhance image-based object detection. For example, CRF-Net [15] converts unstructured radar pins into a pseudo-image format and uses a downstream model to process radar embeddings alongside the camera image.

However, many of these studies need to address the performance degradation under low-visibility conditions, which has been attracting attentions recently. Mirza et al. analyzes the performance degradation of different architectures under adverse weather conditions[16]. Tomy et al. proposes a robust sensor fusion model that combines event-based and frame-based cameras for robust object detection in adverse conditions, utilizing a voxel grid representation for event input and employs a two-parallel feature extractor network for both frames and events [17]. Bijelic et al. proposes a deep fusion architecture that enables robust fusion in foggy and snowy conditions [18]. Sakaridis et al. presents a benchmark dataset comprising authentic images captured in diverse adverse weather conditions [19]. Meanwhile, Kenk provides a benchmark radar dataset for evaluating object-detecting model performance in bad weather [20]. Therefore, our work

takes one step further to investigate how to exploit multi-radar point clouds to enhance object detection algorithms under low-visibility conditions.

B. Camera-Radar Sensor Fusion

A previous study by Waldschmidt et al. shows that radar is not affected by adverse lighting and severe weather conditions, enabling direct measurement of distance, radial velocity, and, with the aid of a suitable antenna system, determination of the angle of distant objects [21]. Hence, the concept of camera-radar sensor fusion aims to exploit the merits of both types of sensors and provide a more comprehensive understanding of the surroundings. Kadow et al. [22] used radar data to limit the search scope for video input and return the distance features and then utilized a simple neural network for vehicle identification. Bertozzi et al. used radar data to refine detecting bounders and apply camera data to detect and classify road obstacles [23]. Other works used similar methods to achieve camera-radar sensor fusion [24], [25]. With the rising deep learning techniques, more efficient sensor fusion models have emerged in recent years. Nobis et al. presented a network model whose layers incorporated both image data and projected sparse radar data and enabled the model to learn features from both sensors [15]. [26] and [27] involved independent object detection by each sensor (camera and radar), followed by the integration of their results to arrive at a final decision. Lekic et al. projected the radar data onto the 2D images generated by the camera and then used Conditional Multi-Generator Generative Adversarial Networks (CMGANs) to implement object detection [28]. Bijelic et al. developed a deep fusion model using camera and radar inputs and validated it on a dataset with adverse weather conditions [18]. A middle-fusion method, CenterFusion, proposed in [29] utilized a center point detection network to detect objects based on their center points in the image. Distinguished from the abovementioned studies, our work draws inspiration from recent achievements in natural language processing and uses embedding along with attention mechanisms [30] to achieve camera-radar sensor fusion.

III. METHODOLOGY

This section presents our camera-radar bird's-eye-view fusion transformer (REDFormer) model for the 3D object detection in detail. The main design idea is to fuse multi-radar point cloud and multi-camera image features in a bird's-eye-view (BEV) plane with addressing the performance drop under low-visibility conditions. We will start with our problem statement and introduce the key components in our model, including learnable BEV queries, the radar backbone (RB) module, and the multi-task learning (MTL) module that address the objectives in the problem statement.

A. Problem Statement and Model Architecture

Our model aims to achieve 3D object detection using camera-radar sensor fusion. Suppose we have N_r radar sensors and N_c cameras. At each timestep t ,

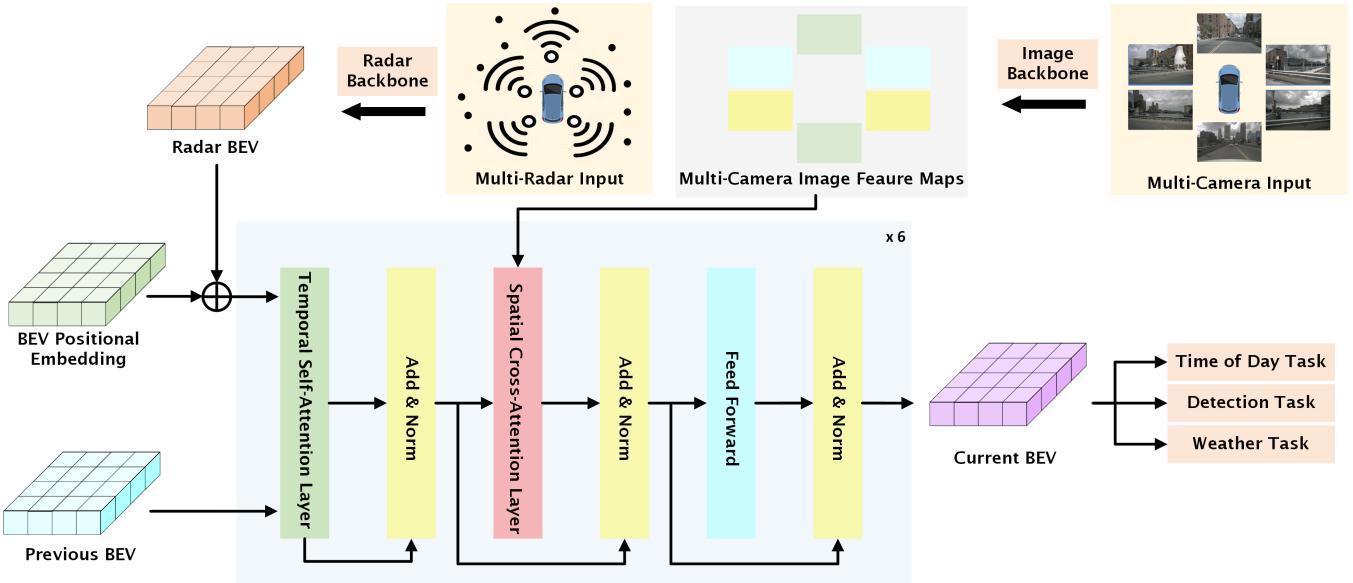


Fig. 1. Illustration of the proposed REDFormer. First, the radar backbone generates radar BEV embedding from the multi-radar point cloud. The radar BEV combines with BEV positional embeddings, and the combination and the learnable BEV features from the previous layer is the input for a temporal self-attention layer. Then, the image backbone extracts multi-view image features. BEV queries from the upstream selectively search within the regions of interest associated with image feature maps in the consecutive spatial cross-attention layer. Finally, a downstream model uses the BEV features from the attention layer to derive predictions that simultaneously minimize multi-task prediction loss.

we observe of a collection of multi-camera images $\mathcal{F}_t = \left\{ f_i^{(t)} \in \mathbb{R}^{3 \times H \times W} \mid i = 1, \dots, N_c \right\}$, where H and W are the height and width of each image. Besides, we also observe multi-radar point clouds $\mathcal{P}_t = \left\{ p_{ij}^{(t)} \in \mathbb{R}^{D_r} \mid i = 1, \dots, N_r \right\}$, where j is the number of radar points for radar sensor i , and D_r the attribute dimension for each radar point. Suppose we have N_k context objects. The bounding box of a context object k is defined by a vector of D_b parameters $b_k^{(t)} \in \mathbb{R}^{D_b}$, and the type of the object is $c_k^{(t)} \in \mathbb{R}$. The objective of our problem is to search for the optimal model inside the model space $\mathcal{M} = \left\{ m \mid m : \mathbb{R}^{D_r} \times \mathbb{R}^{3 \times H \times W} \rightarrow \mathbb{R}^{D_b} \times \mathbb{R} \right\}$ that minimizes the prediction and classification errors

$$m^* = \operatorname{argmin}_{m \in \mathcal{M}} \frac{1}{N_k} \sum_{k=1}^{N_k} \operatorname{err}(b_k^{(t)}, \hat{b}_k^{(t)}) + \operatorname{err}(c_k^{(t)}, \hat{c}_k^{(t)}),$$

where $(\hat{b}_k^{(t)}, \hat{c}_k^{(t)}) = m(\mathcal{F}_t, \mathcal{P}_t \mid \theta)$. (1)

Given the problem statement above, there are three essential problems :

- **Sensor Fusion** How to resolve the differences in feature spaces between multi-radar point clouds and multi-camera images?
- **Temporal Dependencies** How to learn the temporal dependencies between consecutive observations?
- **Vulnerability to Low-visibility Conditions** How to incorporate low-visibility information into the end-to-end learning pipeline?

As illustrated in Fig. 1, we approach sensor fusion and temporal dependencies by proposing a novel embedding-

based RB and a temporal self-attention layer to unify multi-radar point clouds and multi-camera images in a BEV perspective. The RB handles extracting the positional saliency signal of each region of interest, while the attention mechanism learns temporal correlations between current radar signals and previously observed BEV features.

Nevertheless, the original objective function [14] failed to consider vulnerability to low-visibility conditions. To compensate, we train our model by optimizing a multi-task objective. Specifically, instead of only minimizing the prediction and classification error, our model minimizes the binary classification error of current weather conditions $\hat{w}^{(t)}$ and the time-of-day label (i.e., night or daytime) $T^{(t)}$ simultaneously. The overall architecture of the proposed model is inspired by the existing state-of-the-art model BEVFormer [14]. However, we distinguish ourselves with the three innovative components explicitly designed to address the three critical problems.

B. Radar Backbone

It is critical to resolving the difference in feature spaces between multi-radar point clouds and multi-camera image inputs for sensor fusion. To that end, we design the RB module, which projects and helps unify the features under the BEV scope.

Initially, since each radar sensor observes points from its perspective, we need to unify the overall point clouds in a shared local coordinate system centered at the current position of the ego vehicle. For each radar i , we project its observed point cloud using an affine transformation matrix $p'_{ij}^{(t)} = T_i \cdot p_{ij}^{(t)}$. Then, we aggregate these projected points in a BEV plane by region of interest (RoI) and create a saliency

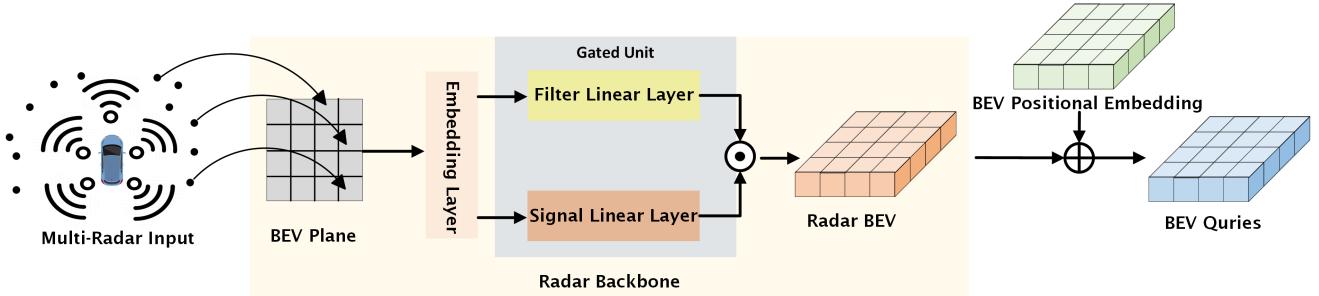


Fig. 2. Illustration of the radar backbone in the REDFormer. The backbone projects multi-radar point clouds onto the BEV plane and then aggregates local point clouds regarding each region of interest (RoI). An embedding layer learns the representation of the saliency signal within each RoI. A gated unit filters the signal and generates the Radar BEV embeddings.

Algorithm 1 Train REDFormer

```

Input: Multi-radar Point  $\mathcal{P}_t$ , Multi-camera Images  $\mathcal{F}_t$   

    Learning Rate  $\eta$ , Data  $\{\{b_k^{(t)}\}, \{c_k^{(t)}\}, w^{(t)}, T^{(t)}\}$ .  

1: Randomly initialize the model  $m$  with parameters  $\theta$ .  

2: while not converged do  

3:    $(\{\hat{b}_k^{(t)}\}, \{\hat{c}_k^{(t)}\}, \hat{w}^{(t)}, \hat{T}^{(t)}) \leftarrow m(\mathcal{P}_t, \mathcal{F}_t; \theta)$ .  

4:   Compute the detection loss  $\mathcal{L}_{\text{det}}(\hat{b}_k^{(t)}, b_k^{(t)}, \hat{c}_k^{(t)}, c_k^{(t)})$ .  

5:   Compute the weather predict loss  $\mathcal{L}_{\text{rain}}(\hat{w}^{(t)}, w^{(t)})$ .  

6:   Compute the time-of-day predict loss  $\mathcal{L}_{\text{tod}}(\hat{T}^{(t)}, T^{(t)})$ .  

7:    $\mathcal{L}_{\text{joint}} \leftarrow \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{rain}} + \mathcal{L}_{\text{tod}}$ .  

8:    $\theta \leftarrow \theta + \eta \nabla_{\theta} \mathcal{L}_{\text{joint}}$ .  

9: end while  

10: return Model  $m(\mathcal{P}_t, \mathcal{F}_t | \theta)$ 

```

matrix. Suppose we denote the BEV regional saliency matrix by $S \in \mathbb{R}^{X \times Y}$, where X and Y are the numbers of ROI on each row and column, respectively. The value of each cell $s_{ij} \in \mathbb{R}$ is the number of projected multi-radar points presented within the corresponding region of interest.

Drawing our concepts from natural language processing, we propose a novel approach incorporating an embedding layer in conjunction with a gated unit layer for processing the structured radar point data. The embedding layer $E(s_{ij}) : \mathbb{R} \rightarrow \mathbb{R}^C$ treats each cell salience s_{ij} as a token and creates a look-up table mapping from the cell salience to a C -dimensional learnable vector. We use a gated unit layer alongside the embedding layer to reduce the noise in raw radar saliency and squash the value space. The gated unit consists of a sigmoid-linear function and a tanh-linear function. If to denote the feature in a output Radar BEV cell by e_{ij} , we can express the overall operation as follows:

$$e_{ij} = \sigma(W_1 \cdot E(s_{ij}) + b_1) \odot \tanh(W_2 \cdot E(s_{ij}) + b_2), \quad (2)$$

where (W_1, W_2) and (b_1, b_2) are the weights and biases for sigmoid-linear and tanh-linear functions, and \odot denotes the element-wise product of two matrices.

C. Temporal Self-Attention and BEV Queries

Object detection models can take advantage of sequential input consisting of previous observations in practical cases.

To address the temporal correlations, we implement the temporal self-attention layer with the help of BEV queries. As shown in Fig. 2, the BEV queries matrix $Q \in \mathbb{R}^{X \times Y \times C}$ derives by adding the radar BEV from RB with a learnable BEV positional embedding. We then fed the BEV queries into the temporal self-attention layer to learn correlations with the BEV representations from preceding timesteps.

Recall that the upstream RB module ensures that the center of the BEV features corresponds to the position of the ego vehicle. Using the intrinsic matrices of cameras, we can determine the corresponding points on the multi-view images for each query $Q_p \in \mathbb{R}^C$. As a result, we build the connection between points on the multi-view images and points on the BEV queries. Such a connection is vital for fusing radar features with camera features in the consecutive spatial cross-attention layer [14].

D. Multi-Task Learning

Multi-task learning (MTL) is an inductive transfer mechanism that aims to enhance generalization performance by leveraging shared information across multiple related tasks [31]. In contrast to traditional machine learning methods, which often focus on learning a single task, MTL capitalizes on the potential richness of information embedded within training signals of related tasks originating from the same domain.

In our proposed REDFormer model, we employ an MTL strategy to bolster its adaptability to diverse environmental circumstances, such as different weather and night scenarios. To achieve this, we apply two additional recognition heads on the unified BEV representation, allowing the model to simultaneously learn and account for weather and TOD conditions alongside the primary 3D object detection task. These recognition heads are designed as separate linear output branches, each responsible for predicting a specific task-related output. The inclusion of these additional learning goals allows the model to capture nuanced relationships and shared features across the tasks. This fosters better generalization and enhanced performance under diverse real-world conditions, resulting in a more robust and versatile 3D object detection system.

Following [10], [11], we employ the set-to-set loss function to quantify the disparity between the predicted and

ground truth values for the 3D object detection task. For environmental variables, we utilize the standard cross-entropy loss function to perform binary classification, identifying whether it's raining or not, and whether it's nighttime or not. The joint loss function is defined as $\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{rain}} + \mathcal{L}_{\text{tod}}$.

IV. EXPERIMENTS

A. Dataset

Our model is implemented and evaluated on the nuScenes dataset, a large-scale public dataset designed by Motional for autonomous driving research [35]. The nuScenes dataset consists of approximately 1,000 scenes, each with a duration of approximately 20 seconds. For every sample, six high-definition cameras are positioned to capture images at a resolution of 1600×900 pixels covering 360 degree field of view. And five radar sensors provide 360-degree coverage around the vehicle, operating at a frequency of 77GHz with a range up to 250 meters.

The nuScenes dataset uses a comprehensive set of evaluation metrics to assess the performance of models and algorithms for object detection task. NuScenes dataset uses mean average precision (mAP) as its primary evaluation metric, which compute the 2D center distance on the ground plane for object matching rather than 3D intersection over union affinities. The nuScenes dataset also includes a set of true positive metrics, such as average translation error (ATE), average scale error (ASE), average orientation error (AOE), average velocity error (AVE), and average attribute error (AAE), which are used to evaluate the accuracy of object detection in terms of translation, scale, orientation, velocity, and attribute errors. In addition, the nuScenes dataset introduced a nuScenes detection score (NDS) to consolidate all the above evaluation metrics. The NDS is calculated as follow:

$$\text{NDS} = 0.5 \times \text{mAP} + \sum 0.1 \times \max((1 - \text{mTP}), 0) \quad (3)$$

B. Baseline

We select several state-of-the-art 3D object detection models as baselines to compare the performance of our approach. The baseline models include DETR3D [11], FCOS3D [32], BEVFusion [33], BEVFormer (small variant) [14] and BEVDet (tiny variant) [34], which rely solely on cameras. Additionally, we compare our approach to CenterFusion[29] and CRF-Net [15], which incorporate both radar and camera data.

C. Main Results

We conduct extensive evaluations of our proposed REDFormer model on the nuScenes dataset object detection task, and compare its performance with several state-of-the-art object detection approaches that use either camera, radar, or camera-radar fusion models. Our results, summarized in Tab. I, demonstrate that our model achieves a significant improvement over the baselines. Specifically, we obtained a 6.7% improvement in NDS and a 16.1% improvement in mAP compared to CenterFusion (camera-radar Fusion) [29],

an 1.5% improvement in NDS and a 4.1% improvement in mAP compared to BEVFormer (Camera only) [14]. These findings indicate the effectiveness of our novel RB and MTL modules, and their abilities to capture the nuances of multi-sensor data and enhance object detection performance in challenging scenarios.

To assess the robustness of our model under adverse weather conditions, we conduct experiments in both low-light nighttime and rainy scenarios. Specifically, we evaluate our model's performance in these adverse conditions to verify its ability to maintain high accuracy and reliability even in low visibility environments. To do this, we create two sub-datasets from the nuScenes dataset, one containing only nighttime scenarios and the other containing only rainy scenes. We then perform experiments on these sub-datasets to evaluate our model's performance in these challenging conditions. We use BEVFormer [14] and BEVDet [34] as our baseline model. The object detection performance in low-visibility scenarios is generally not as good as in normal conditions, as indicated in Tab. II. However, our REDFormer demonstrates significant improvements in both rainy and nighttime conditions when compared to BEVFormer [14]. Specifically, we observe a 31.31% improvement in NDS on rainy scenes and a 46.99% improvement in NDS on night scenes highlighting the effectiveness of our approach in adverse weather conditions. Our model also showcases significant improvements than BEVFormer [14] in mAP across challenging conditions. Particularly, we achieve a notable 14.5% improvement in mAP for rainy scenes and an impressive 11.5% improvement for night scenes. Furthermore, compared to the baseline model BEVDet [34], our model exhibits a substantial 19.8% improvement in mAP for rainy scenes and a remarkable 50% improvement for night scenes. Hence, the notable enhancement in low-visibility scenarios demonstrates the radar's ability to provide precise object localization. This signifies the effectiveness of our RB in guiding the model towards object localization especially in low-visibility conditions, while the MTL heads effectively incorporate light conditions into the predictions. In rainy or nighttime conditions, the model appropriately assigns higher importance to radar inputs compared to sunny or other high-visibility conditions.

D. Ablation Study

a) Module analysis: We conduct ablation experiments on the nuScenes validation set to analyze the impact of different modules, specifically the MTL and the RB. Each component is removed individually while the other is kept fixed. Our ablation study includes both the full nuScenes dataset and low-visibility subsets to comprehensively evaluate the effectiveness of our approach. Tab. III demonstrates that both the RB and MTL modules significantly improve model performance. Additionally, when compared to the baseline model, Tab. IV highlights the improvements made by each module in night and rainy scenes. Notably, the RB module improves NDS by 31.21% and 43.33% in rainy and night subsets, respectively, while the MTL module

TABLE I

COMPARISON OF THE PROPOSED REDFORMER WITH VARIOUS STATE-OF-THE-ART METHODS. THE TOP-PERFORMING METHOD FOR EACH SETTING IS HIGHLIGHTED IN **BOLD**. \downarrow : LOWER VALUES ARE BETTER. \uparrow : HIGHER VALUES ARE BETTER. *: RESULTS OBTAINED FROM THE ORIGINAL PAPERS.

Modality	Method	#param.	NDS (\uparrow)	mAP (\uparrow)	mATE (\downarrow)	mASE (\downarrow)	mAOE (\downarrow)	mAVE (\downarrow)	mAAE (\downarrow)
Camera only	FCOS3D*[32]	$\geq 52.5M$	0.415	0.343	0.725	0.263	0.422	1.292	0.153
	DETR3D*[11]	51.3M	0.425	0.346	0.773	0.268	0.383	0.842	0.216
	BEVFusion*[33]	-	0.412	0.356	-	-	-	-	-
	BEVDet*[34]	53.7M	0.392	0.312	0.691	0.272	0.523	0.909	0.247
	BEVFormer*[14]	56.8M	0.479	0.370	0.725	0.272	0.391	0.802	0.200
Camera and Radar	CenterFusion*[29]	-	0.453	0.332	0.649	0.263	0.535	0.540	0.142
	REDFormer (ours)	56.8M	0.486	0.385	0.726	0.282	0.407	0.427	0.218

TABLE II

COMPARISON OF THE PROPOSED REDFORMER WITH THE BEST STATE-OF-THE-ART METHOD IN LOW-VISIBILITY SUBSETS. *: RESULTS OBTAINED FROM THE ORIGINAL PAPERS

Subset	Method	NDS (\uparrow)	mAP(\uparrow)
Rainy Scenes	BEVDet* [34]	-	0.3370
	BEVFormer [14]	0.3877	0.3524
	REDFormer (ours)	0.5091	0.4036
Night Scenes	BEVDet* [34]	-	0.1350
	BEVFormer[14]	0.1913	0.1819
	REDFormer (ours)	0.2812	0.2028

improves NDS by 30.51% and 43.44% in rainy and night subsets, respectively. Individually, both of these components contribute to enhancing the object detection performance of the model, and their combined usage yields the highest performance, indicating the synergistic benefits of integrating both components in the model. These findings highlight the effectiveness of our proposed approach and underscore the value of the consequential modules in enhancing the performance of object detection systems in challenging scenarios.

b) *Influence of the capacity limit of the embedding dictionary:* We discuss the outcomes of selecting the capacity limit K for the embedding dictionary based on the experiment presented in Tab. V. We restrict our focus to $K \geq 10$, as this value represents the maximum number of radar points within a single grid throughout the entire nuScenes dataset. The optimal performance in terms of mAP is achieved when $K = 10$, while the NDS values show near identical best performances for $K = 10$ and $K = 20$. These results suggest that K should be selected as close as possible to the radar point capacity per grid, contingent on the specific radar setup of the vehicle.

E. Visualization

In order to provide a comprehensive insight into the performance of the REDFormer model in 3D object detection tasks under low-visibility conditions, we present detailed visualizations showcasing the outstanding performance and the significant improvement of our model compared to the baseline (BEVFormer) model in such scenarios. The comparison of prediction results between RedFormer, the baseline (BEVFormer), and ground truth in the rainy subset and night subset are visualized in Fig.3 and Fig.4, respectively.

TABLE III

THE PERFORMANCE OF REDFORMER IS COMPARED WITH AND WITHOUT THE INCLUSION OF RADAR BACKBONE (RB) AND MULTI-TASK LEARNING (MTL) IN THE WHOLE NUSCENES DATASET.

With RB	With MTL	Module	NDS (\uparrow)	mAP (\uparrow)
\times	\times		0.4787	0.3700
\times	\checkmark		0.4851	0.3838
\checkmark	\times		0.4833	0.3816
\checkmark	\checkmark		0.4863	0.3853

TABLE IV

THE PERFORMANCE OF REDFORMER IS COMPARED WITH AND WITHOUT THE INCLUSION OF RADAR BACKBONE (RB) AND MULTI-TASK LEARNING (MTL) IN RAINY AND NIGHT-TIME SUBSET.

Subset	Module		NDS (\uparrow)	mAP (\uparrow)
	With RB	With MTL		
Rainy Scenes	\times	\times	0.3877	0.3524
	\times	\checkmark	0.5060	0.3959
	\checkmark	\times	0.5087	0.3966
	\checkmark	\checkmark	0.5091	0.4036
Night Scenes	\times	\times	0.1913	0.1819
	\times	\checkmark	0.2742	0.2079
	\checkmark	\times	0.2744	0.2067
	\checkmark	\checkmark	0.2812	0.2028

TABLE V

ABLATION STUDY INVESTIGATING THE INFLUENCE OF VARIOUS CAPACITY LIMITS K OF THE EMBEDDING DICTIONARY.

Capacity Limit K	NDS (\uparrow)	mAP (\uparrow)
10	0.4834	0.3854
20	0.4836	0.3839
30	0.4802	0.3814

The visualization results clearly demonstrate that our model, REDFormer, outperforms the baseline model under challenging visibility conditions, as evidenced by the more precise 3D bounding boxes it generates. Notably, our model avoids entirely non-existent predictions, whereas the baseline model occasionally produces erroneous predictions. These results validate the effectiveness of our MTL approach, which enables our model to account for visibility conditions



Fig. 3. Visualization results of REDFormer and BEVFormer on nuScenes rainy validation subset, exclusively including figures when objects are present. Vehicles are marked by orange bounding boxes, motorcycles by red, pedestrians by blue, and barriers by gray.



Fig. 4. Visualization results of REDFormer and BEVFormer on nuScenes night validation subset, exclusively including figures when objects are present. Vehicles are marked by orange bounding boxes, motorcycles by red, pedestrians by blue, and barriers by gray.

and achieve superior performance in such scenes.

Furthermore, the incorporation of multi-radar input proves to be advantageous, as it remains unaffected by environmental conditions and provides reliable cues in adverse environments. Additionally, the inclusion of radar points imparts additional depth information, resulting in more accurate and realistic 3D bounding boxes.

V. CONCLUSIONS

In this paper, our proposed approach to 3D object detection represents a significant improvement over existing state-of-the-art (SOTA) methods, leveraging a middle fusion technique that combines a transformer-based bird's-eye-view (BEV) encoder. We introduced an innovative radar backbone (RB) to extract features from multi-radar points and employ multi-task learning (MTL) to enable the model

to consider the impact of weather and time-of-day (TOD) on object detection. The transformer-based BEV approach enables us to effectively utilize comprehensive environmental information, leading to high performance in object detection. Our approach enhances the accuracy and robustness of the system in diverse environments, including those with reduced visibility due to adverse weather conditions or low light. By combining the benefits of MTL, the novel RB, and the transformer-based middle fusion approach, our method demonstrates significant improvements in performance. Our experiments reveal that our model outperforms the SOTA baseline model (BEVFormer), achieving a 31.31% higher (NDS) in rainy scenes and a 46.99% higher (NDS) in low-visibility night scenes. Overall, our approach represents a valuable contribution to the field of 3D object detection, with potential applications in a wide range of industries and

use cases. One limitation of our work is that the current frames per second (FPS) of our model are relatively high. As a result, our future research will focus on optimizing the FPS and enhancing the model's deployability for real-time predictions in real automated vehicles.

REFERENCES

- [1] S. Campbell, N. O'Mahony, L. Krpalcova, D. Riordan, J. Walsh, A. Murphy, and C. Ryan, "Sensor Technology in Autonomous Vehicles : A review," in *2018 29th Irish Signals and Systems Conference (ISSC)*, June 2018, pp. 1–4.
- [2] Z. Wang, R. Gupta, K. Han, H. Wang, A. Ganlath, N. Ammar, and P. Tiwari, "Mobility digital twin: Concept, architecture, case study, and future challenges," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17452–17467, 2022.
- [3] Z. Liu, Y. Cai, H. Wang, L. Chen, H. Gao, Y. Jia, and Y. Li, "Robust Target Recognition and Tracking of Self-Driving Cars With Radar and Camera Information Fusion Under Severe Weather Conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6640–6653, July 2022. [Online]. Available: <https://ieeexplore.ieee.org/document/9363012/>
- [4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," June 2021, arXiv:2010.11929 [cs]. [Online]. Available: <http://arxiv.org/abs/2010.11929>
- [5] Y. B. Can, A. Liniger, D. P. Paudel, and L. Van Gool, "Structured Bird's-Eye-View Traffic Scene Understanding from Onboard Images," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 15 641–15 650. [Online]. Available: <https://ieeexplore.ieee.org/document/9711322/>
- [6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Jan. 2016, arXiv:1506.01497 [cs]. [Online]. Available: <http://arxiv.org/abs/1506.01497>
- [7] R. Girshick, "Fast R-CNN," Sept. 2015, arXiv:1504.08083 [cs]. [Online]. Available: <http://arxiv.org/abs/1504.08083>
- [8] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection," Sept. 2021, arXiv:2104.10956 [cs]. [Online]. Available: <http://arxiv.org/abs/2104.10956>
- [9] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," Aug. 2019, arXiv:1904.01355 [cs]. [Online]. Available: <http://arxiv.org/abs/1904.01355>
- [10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," May 2020, arXiv:2005.12872 [cs]. [Online]. Available: <http://arxiv.org/abs/2005.12872>
- [11] Y. Wang, V. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries," Oct. 2021, arXiv:2110.06922 [cs]. [Online]. Available: <http://arxiv.org/abs/2110.06922>
- [12] E. Xie, Z. Yu, D. Zhou, J. Philion, A. Anandkumar, S. Fidler, P. Luo, and J. M. Alvarez, "M²bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation," 2022.
- [13] K. Chitta, A. Prakash, and A. Geiger, "NEAT: Neural Attention Fields for End-to-End Autonomous Driving," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 15 773–15 783. [Online]. Available: <https://ieeexplore.ieee.org/document/9710855/>
- [14] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers," July 2022, arXiv:2203.17270 [cs]. [Online]. Available: <http://arxiv.org/abs/2203.17270>
- [15] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A Deep Learning-based Radar and Camera Sensor Fusion Architecture for Object Detection," May 2020, arXiv:2005.07431 [cs]. [Online]. Available: <http://arxiv.org/abs/2005.07431>
- [16] M. J. Mirza, C. Buerkle, J. Jarquin, M. Opitz, F. Oboril, K.-U. Scholl, and H. Bischof, "Robustness of object detectors in degrading weather conditions," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 2719–2724.
- [17] A. Tomy, A. Paigwar, K. S. Mann, A. Renzaglia, and C. Laugier, "Fusing event-based and rgb camera for robust object detection in adverse conditions," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 933–939.
- [18] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather," June 2020, arXiv:1902.08913 [cs]. [Online]. Available: <http://arxiv.org/abs/1902.08913>
- [19] C. Sakaridis, D. Dai, and L. V. Gool, "Semantic foggy scene understanding with synthetic data," *CoRR*, vol. abs/1708.07819, 2017. [Online]. Available: <http://arxiv.org/abs/1708.07819>
- [20] M. KENK, "Dawn: Vehicle detection in adverse weather nature dataset," 2020. [Online]. Available: <https://data.mendeley.com/datasets/766ygrbt8y/3>
- [21] C. Waldschmidt, J. Hasch, and W. Menzel, "Automotive radar — from first efforts to future systems," *IEEE Journal of Microwaves*, vol. 1, no. 1, pp. 135–148, 2021.
- [22] U. Kadov, G. Schneider, and A. Vukotich, "Radar-Vision Based Vehicle Recognition with Evolutionary Optimized and Boosted Features," in *2007 IEEE Intelligent Vehicles Symposium*. Istanbul, Turkey: IEEE, June 2007, pp. 749–754, iSSN: 1931-0587. [Online]. Available: <http://ieeexplore.ieee.org/document/4290206/>
- [23] M. Bertozzi, L. Bombini, P. Cerri, P. Medici, P. C. Antonello, and M. Miglietta, "Obstacle detection and classification fusing radar and vision," in *2008 IEEE Intelligent Vehicles Symposium*, June 2008, pp. 608–613, iSSN: 1931-0587.
- [24] J. Kocic, N. Jovicic, and V. Drndarevic, "Sensors and Sensor Fusion in Autonomous Vehicles," in *2018 26th Telecommunications Forum (TELFOR)*. Belgrade: IEEE, Nov. 2018, pp. 420–425. [Online]. Available: <https://ieeexplore.ieee.org/document/8612054/>
- [25] S. Han, X. Wang, L. Xu, H. Sun, and N. Zheng, "Frontal object perception for Intelligent Vehicles based on radar and camera fusion," in *2016 35th Chinese Control Conference (CCC)*, July 2016, pp. 4003–4008, iSSN: 1934-1768.
- [26] H. Jha, V. Lodhi, and D. Chakravarty, "Object Detection and Identification Using Vision and Radar Data Fusion System for Ground-Based Navigation," in *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, Mar. 2019, pp. 590–593.
- [27] K.-E. Kim, C.-J. Lee, D.-S. Pae, and M.-T. Lim, "Sensor fusion for vehicle tracking with camera and radar sensor," in *2017 17th International Conference on Control, Automation and Systems (ICCAS)*, Oct. 2017, pp. 1075–1077.
- [28] V. Lekic and Z. Babic, "Automotive radar and camera fusion using generative adversarial networks," *Computer Vision and Image Understanding*, vol. 184, 04 2019.
- [29] R. Nabati and H. Qi, "CenterFusion: Center-based Radar and Camera Fusion for 3D Object Detection," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 1526–1535. [Online]. Available: <https://ieeexplore.ieee.org/document/9423268/>
- [30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
- [31] R. Caruana, "Multitask Learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, July 1997.
- [32] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection," Sept. 2021, arXiv:2104.10956 [cs]. [Online]. Available: <http://arxiv.org/abs/2104.10956>
- [33] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation," June 2022, arXiv:2205.13542 [cs].
- [34] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View," June 2022, arXiv:2112.11790 [cs]. [Online]. Available: <http://arxiv.org/abs/2112.11790>
- [35] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," May 2020, arXiv:1903.11027 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/1903.11027>