Talk2Radar: Bridging Natural Language with 4D mmWave Radar for 3D Referring Expression Comprehension

Runwei Guan[1,2,5] *, Ruixiao Zhang[3] *, Ningwei Ouyang[1,2,5] *, Jianan Liu[4] *, Ka Lok Man[5], Xiaohao Cai[3], Ming Xu[2,5], Jeremy Smith[2], Eng Gee Lim[5], *Senior Member, IEEE*, Yutao Yue[6,7,1] †, Hui Xiong[6], *Fellow, IEEE*

***Abstract*—Embodied perception is essential for intelligent vehicles and robots in interactive environmental understanding. However, these advancements primarily focus on vision, with limited attention given to using 3D modeling sensors, restricting a comprehensive understanding of objects in response to prompts containing qualitative and quantitative queries. Recently, as a promising automotive sensor with affordable cost, 4D millimeter-wave radars provide denser point clouds than conventional radars and perceive both semantic and physical characteristics of objects, thereby enhancing the reliability of perception systems. To foster the development of natural language-driven context understanding in radar scenes for 3D visual grounding, we construct the first dataset, Talk2Radar, which bridges these two modalities for 3D Referring Expression Comprehension (REC). Talk2Radar contains 8,682 referring prompt samples with 20,558 referred objects. Moreover, we propose a novel model, T-RadarNet, for 3D REC on point clouds, achieving State-Of-The-Art (SOTA) performance on the Talk2Radar dataset compared to counterparts. Deformable-FPN and Gated Graph Fusion are meticulously designed for efficient point cloud feature modeling and cross-modal fusion between radar and text features, respectively. Comprehensive experiments provide deep insights into radar-based 3D REC. We release our project at https://github.com/GuanRunwei/Talk2Radar.**

I. Introduction

Millimeter-wave (mmWave) radar, as an all-weather and low-cost perception sensor [1, 2] which can capture objects' distance, azimuth, velocity, motion directions, reflected power, etc., has been widely used in perception for autonomous driving [3–8], robotic navigation [9, 10] and Cooperative Intelligent Transportation Systems (C-ITS) [11]. Recently, to enhance detection accuracy, 4D radar emerges and addresses the limitation of conventional radars' inability to measure height and significantly increases Point Cloud (PC) density [12, 13], allowing objects to contain more exploitable features for fruitful scene understanding and better performance in downstream tasks on land [14–21] and maritime [22–26].

The work of Runwei Guan, Ningwei Ouyang, and Ka Lok Man is partially supported by the AI University Research Centre, Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU and Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004). The work of Runwei Guan, Ningwei Ouyang and Yutao Yue is also supported by Jiangsu Industrial Technology Research Institute (JITRI).

* R. Guan, R. Zhang, N. Ouyang and J. Liu contribute equally.
[1] Institute of Deep Perception Technology, JITRI, Wuxi, China
[2] Department of EEE, University of Liverpool, Liverpool, UK
[3] School of ECS, University of Southampton, Southampton, UK
[4] Momoni AI, Gothenburg, Sweden
[5] SAT, Xi'an Jiaotong-Liverpool University, Suzhou, China
[6] Thrust of Artificial Intelligence, HKUST (GZ), Guangzhou, China
[7] Thrust of Intelligent Transportation, HKUST (GZ), Guangzhou, China
† Corresponding author: yutaoyue@hkust-gz.edu.cn

Meanwhile, recent advancements driven by Vision-Language Models (VLMs) [27] in embodied intelligence and human-centric intelligent driving perception systems [28] enable intelligent vehicles and robots to understand human commands, perceive surroundings, and make decisions [29]. However, these advancements are primarily confined to the visual domain [30–32]. While some research has focused on 3D visual grounding, these efforts are predominantly aimed at RGB-D and LiDAR [33–37]. For autonomous driving and navigation, we argue that referring prompts should not only emphasize positional relations, shapes, and categories but also include object motion characteristics, which is precisely the strength of radar. Radar is immune to adverse weather conditions and can provide quantified measurements of distance, velocity, motion, and azimuth, while Radar Cross Section (RCS) and Point Clouds (PCs) offer qualitative semantic features. This dual capability allows for a comprehensive description of objects, incorporating both qualitative and quantitative characteristics. With the emergence of high-resolution 4D radar, this unexplored field is poised to be addressed. Unlike VLM-based environmental understanding, which can leverage web-scale training data, radar data depend on real-world scenarios and physical models [38]. Currently, there are no relevant datasets or research focusing on this area. Hence, we aim to pioneer and unlock the potential of 4D radar for natural language-based multi-modal 3D object localization. Specifically, this enables individuals to locate specific objects using 3D bounding boxes by describing features perceivable by 4D radar.

Based on these foundations, we introduce 3D Referring Expression Comprehension (REC), also known as open-vocabulary 3D object detection or 3D visual grounding [39], which locates specific objects based on textual descriptions. To this end, we establish the first radar-based 3D visual grounding dataset, Talk2Radar, and its corresponding benchmark. Talk2Radar is built upon the well-known 4D radar dataset, View of Delft (VoD) [40]. Each sample in Talk2Radar contains radar PCs, LiDAR PCs, an RGB image, and a textual reference prompt. The dataset possesses the following two key characteristics: **(i)** A prompt can refer to one or more objects, making it more flexible and realistic; **(ii)** The text features contain object attributes perceivable by radar, while excluding any features perceivable by vision. Besides the dataset, we propose a novel radar-text fusion model called T-RadarNet for 3D visual grounding. Within T-RadarNet, **firstly**, we introduce Deformable-FPN to enhance the modeling of irregular radar PCs. **Secondly**, to efficiently fuse the features of radar and text, we devise a module named Gated Graph

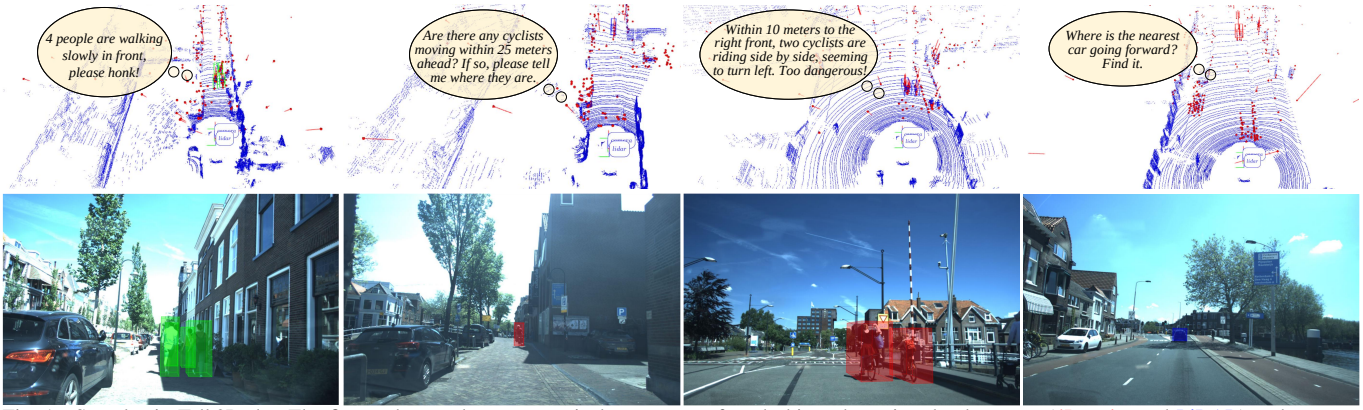

Fig. 1. Samples in Talk2Radar. The first and second row respectively presents referred objects by point cloud sensors (4D radar and LiDAR) and camera.

Fusion (GGF). GGF leverages graph and gating mechanisms for neighborhood feature association and cross-modal feature matching. Notably, GGF can be integrated into most PC detectors for 3D visual grounding, demonstrating impressive generalization capabilities.

The contributions of this paper are summarized as follows:

1) We present Talk2Radar, the first 4D radar-based 3D REC dataset for terrestrial autonomous driving, along with its corresponding benchmark for subsequent studies.
2) For 4D radar-based 3D visual grounding, we propose a novel 3D REC model named T-RadarNet, by incorporating the devised Deformable-FPN and GGF.
3) Comprehensive experiments have been conducted to analyze and enhance 4D radar-based 3D visual grounding, to promote a thorough understanding in this field.

II. Related Works

A. 3D Referring Expression Comprehension in Traffic

As a challenging task based on multi-modal learning and 3D geometry, 3D REC aims to provide a textual prompt and locate one or more objects most closely matching it in the form of 3D bounding boxes. **(i) Datasets.** As Table. I shows, Talk2Car [41] is a 2D image-based REC dataset on driving scenarios upon Nuscenes [42], but providing LiDAR data of the same frame. M3DRef [43] is a monocular 3D grounding dataset upon KITTI [44]. NuPro [45] is a 3D grounding benchmark based on multi-camera and multi-frames for autonomous driving. **(ii) Methods.** [43] proposes a monocular-based 3D REC baseline model called Mono3DVG-TR. [46] raises a model called AFMNet for modeling complex 3D object relation. [35] propose a baseline model for LiDAR-based visual grounding upon Talk2Car dataset. Nevertheless, radar-based 3D REC is still unexplored and lacks full-scale analysis. Moreover, textual prompts of our Talk2Radar dataset contain more attributes of objects and context with qualitative or numerical description, and provide a realistic referring paradigm.

B. 3D Object Detection with Point Cloud

3D object detection is vital for environmental perception for autonomous driving. PC sensors mainly include LiDAR and 4D radar. LiDAR can provide rich geometric and depth features [47], while 4D radar can capture certain semantic features, motion, velocity, and depth features of objects [48–50]. Currently, PC-based 3D detectors can be primarily divided into pillar-based and voxel-based. [51, 52] are three fast pillar-based, which project PCs onto a Bird's-Eye View (BEV) and then extract features like image processing. [53–55] are three voxel-based detectors converting PCs into voxel grids, which are known for simplicity, but they may suffer from low computation efficiency. Additionally, [56, 57] provide an anchor-free paradigm for 3D detection. [58] fuses different representations of PCs dedicated to radar-based 3D detection. Based on the above, we fully consider two aspects in our T-RadarNet: **(i)** Extraction of irregular PC features, **(ii)** Dynamic weighting and separation of objects and clutter for high-quality fusion with text features.

Table I. Comparison of 3D visual grounding datasets in traffic

Datasets	Sensors	Objects	AvgExpr	Context
NuPro [45]	Camera	187,445	-	color, size, relation motion, location
Talk2Car [41]	Camera, LiDAR	10,519	11.01	color, size, location, relation
M3DRef [43]	Camera	8,228	53.24	size, relation, location, color
Talk2Radar	**4D radar, LiDAR**	**20,558**	**14.30**	**size, motion, location, relation, velocity, depth**

III. Talk2Radar Dataset

A. Data Collection and Annotation Process

Thanks to the renowned VoD dataset [40] in the field of radar perception, which is specifically designed for autonomous driving perception and equipped with the ZF FR-Gen21 4D mmWave radar, as well as a LiDAR and a stereo camera, we can engineer textual prompts based on finely annotated 3D bounding boxes of various road objects. Fig. 1 shows some samples, while Fig. 2 presents the entire annotation process.

(1) Select Candidate Object(s): First, utilizing the developed Graphical User Interface for annotation, we present the image with projected 3D bounding boxes for easy observation. Then, experienced annotators select the candidate boxes in

Table II. Statistics of referent object number and point clouds in Talk2Radar

Objects	Pedestrian	Cyclist	Car	motor	truck	bicycle	rider	moped scooter	bicycle rack	human depiction	rider other	vehicle other
Sensor PC	3487 (17.0%)	3234 (15.8%)	9336 (45.6%)	73 (0.4%)	27 (0.1%)	2442 (11.9%)	159 (7.8%)	470 (2.3%)	1165 (5.7%)	48 (2.3%)	11 (-)	3 (-)
$\text{Radar1}_{(1\text{-frame})}$	2.0	3.8	3.6	3.4	17.7	1.4	1.7	1.8	3.0	-	2.5	-
$\text{Radar3}_{(3\text{-frame})}$	5.3	9.6	11.8	19.1	65.2	4.3	3.6	7.1	9.8	-	8.0	-
$\text{Radar5}_{(5\text{-frame})}$	7.4	12.1	18.7	44.4	128.0	7.0	4.2	9.6	14.3	-	9.5	-
LiDAR	258.9	211.5	771.6	1378.3	7338.4	174.2	273.9	232.9	215.5	-	6.0	-

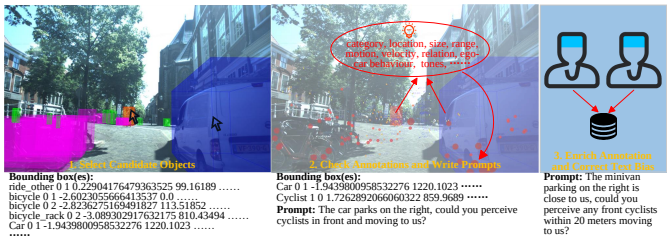

Fig. 2. Annotation process of Talk2Radar dataset.

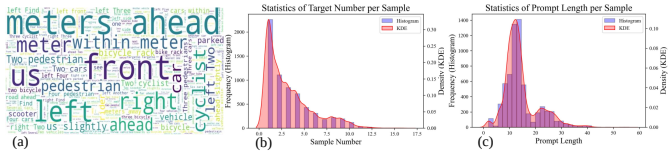

Fig. 3. Statictics of Talk2Radar dataset on referent objects and prompts.

the image that need to be described. The corresponding 3D bounding box annotation is saved. The annotator prioritizes selecting objects within the radar's Field of View (FoV).

(2) Check Annotations and Write Prompts: Based on the chosen 3D bounding box(es), the annotator first checks for any errors in the position or category of bounding boxes. During annotation, we also attach compensated vector velocity and plane depth next to the corresponding PC for reference. Considering detection errors and object size, we generally describe the object depth as a range rather than a fixed value. For individual objects, annotators consider attributes such as category, spatial position relative to the ego vehicle, distance, velocity, motion trend, and size, but exclude color and other information that radar cannot perceive. If multiple objects are selected, their spatial relationships are also considered.

(3) Enrich Annotation and Correct Text Bias: To avoid subjectivity and errors in text prompts, we implement a collaborative review and revision process involving two additional annotators after the initial annotation. This process serves to increase the diversity of descriptions to test the model's robustness and generalization. Additionally, it allows for the discussion of subjective errors made by the initial annotator, followed by corrections upon reaching a consensus.

B. Dataset Statistics

We present quantitative statistics from the perspectives of textual prompts and referent objects.

Textual Prompts: As illustrated in Fig. 3(a), **firstly**, high-frequency words in the word cloud include both qualitative and quantitative descriptions. This rich vocabulary benefits from the 4D radar's multi-dimensional perception capabilities, capturing quantitative features such as object velocity, motion direction, and depth, while also perceiving some degree of semantic features. This enables more detailed descriptions of specific objects and allows for fine-grained object clustering and filtering based on individual or partial object attributes. **Secondly**, as shown in Fig. 3(c), the prompt length distribution in Talk2Radar is broad, with an average value of 14.30 (Table I, **AvgExpr**). This variety allows the model to learn a wide range of sentence patterns, increasing task complexity as the model must handle text prompts of varying patterns, structures, and lengths, correctly understanding their semantic representations to query complex PC scenes.

Referent Objects: Firstly, as presented in Table I, Talk2Radar contains a total of 20,558 referent objects. **Secondly**, Table II shows the distribution of object quantities across 12 categories and the average PC quantity for each category, including radar (1, 3, and 5 frames accumulated) and LiDAR. Cars, pedestrians, and cyclists are the three primary referent categories in Talk2Radar. Despite the improved resolution of 4D radar, there remains a gap in PC density compared to LiDAR. **Thirdly**, as shown in Fig. 3(b), the number of referent objects per sample in Talk2Radar ranges from 1 to 11, presenting a challenge for understanding complex scenes. Specifically, the model must adaptively filter out noise words in prompts, focus on the key characteristics of the referent object(s), and perform multiple Region-Of-Interest (ROI) queries in irregular PC contexts.

C. Metrics and Subset Settings

Metrics: We adopt Average Precision (AP) and Average Orientation Similarity (AOS), including 3D bounding box mAP and mAOS on the entire annotated and driving corridor area, consistent with VoD for comprehensive evaluation.

Subset Settings: We divide Talk2Radar into three subsets, sharing same IDs with VoD. The number of subsets for training, validation and test are 5139, 1296 and 2247, respectively. Since the test set annotation of VoD is not publicly available, models are evaluated on the validation set in this paper.

IV. Methods

A. Overall Pipeline

Fig. 4 illustrates the detailed architecture of T-RadarNet. Given a frame of radar PCs and a textual prompt as input, T-RadarNet provides the 3D grounding prediction of the specified object(s) guided by natural language. Firstly, for primary PC feature extraction, considering computational efficiency and scalability, we adopt a pillar encoder for the basic representation of 3D PC as a 2D pseudo image. Subsequently,

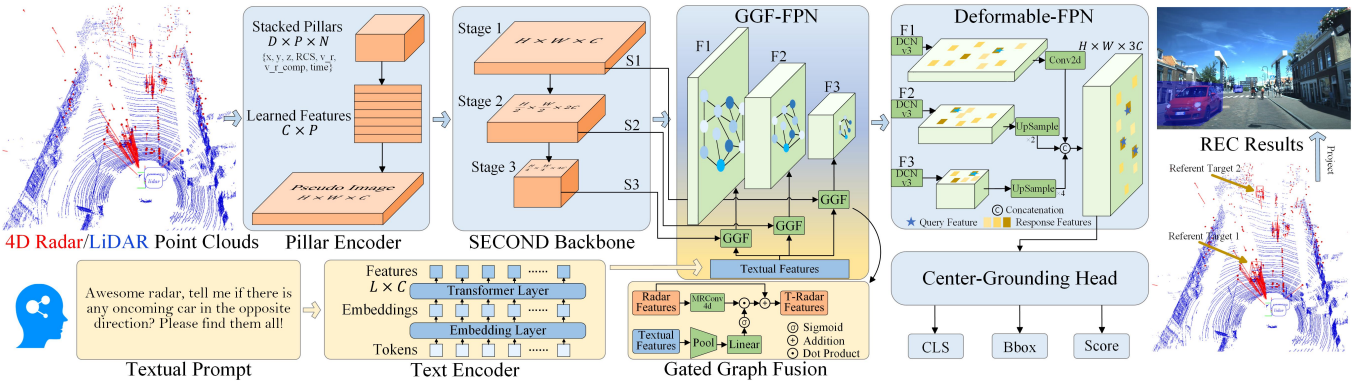

Fig. 4. The architecture of **T-RadarNet**. LiDAR is not the main modality, but it can also be used as the input of T-RadarNet.

the concise SECOND backbone [54] is used to extract three-stage radar PC features $\{F_R^{S_1}, F_R^{S_2}, F_R^{S_3}\}$ at multiple scales. Secondly, we use transformer-based encoders (e.g., ALBERT) [59] for dynamic context representation of the textual prompt. Thirdly, given the interference in radar PC context from multi-path clutter, aligning and fusing dense textual semantic features with sparse PCs pose significant challenges. To address this, we design a graph-based strategy called Gated Graph Fusion (GGF), which associates ROI in the radar feature space and applies point-wise cross-modal gating between radar and text, yielding the most matched PC regions aligned with the text. The three-stage PC features and textual feature $F_T \in \mathbb{R}^{L\times C}$ are then fed into GGF-FPN, which consists of GGF modules for cross-modal fusion, outputting three-scale text-conditional radar features $\{F_{R|T}^{S_1}, F_{R|T}^{S_2}, F_{R|T}^{S_3}\}$. Fourthly, given the irregularity and sparsity of radar PC, conventional convolutions struggle to interact with them. Hence, we propose Deformable-FPN based on deformable convolutions [60], which provides high-quality PC features queried by the textual prompt to the grounding head for localization. Lastly, considering the efficiency of anchor-free detectors, we adopt the center-based detection head [56] and employ the deformable convolution-based separable localization head for text-referenced 3D detection.

B. Gated Graph Fusion

For one-stage deep radar feature $F_R \in \mathbb{R}^{H\times W\times C_r}$, we first construct a graph map $\mathcal{G} = H(F_R)$ to aggregate neighbour features and associate potential object areas, where $F_R = \{F_R^1, F_R^2, \ldots, F_R^i\}$ denotes a set of unordered nodes initialized upon [61]. We implement the graph convolution $H(\cdot)$ as:

$$\begin{aligned}\mathcal{G} &= H(F_R, \mathcal{W}) \\ &= \text{Update}(\text{Aggregate}(F_R, W_{\text{agg}}), W_{\text{update}}),\end{aligned} \tag{1}$$

where W_{agg} and W_{update} are two learnable weights for feature aggregation and update. Specifically, the aggregation and update operations calculate the representation of the current node i by aggregating features of neighbouring nodes, *i.e.*,

$$\hat{F}_R^i = h(F_R^i, g(F_R^i, \mathcal{N}(F_R^i), W_{\text{agg}}), W_{\text{update}}), \tag{2}$$

where $\mathcal{N}(F_R^i)$ is the set of neighbour nodes of F_R^i. Considering the efficiency of graph feature aggregation, we first adopt Max-Relative Graph Convolution (MRConv4d) [61]. To enrich context, we utilize a residual path with concatenation $\oplus$ to construct radar graph with neighourhood aggregated features:

$$g(\cdot) = \tilde{F}_R^i = \max(\{F_R^i - F_R^j \mid j \in \mathcal{N}(F_R^i)\})W_{\text{agg}}, \tag{3}$$

$$h(\cdot) = \hat{F}_R^i = \tilde{F}_R^i W_{\text{update}} \oplus F_R^i, \tag{4}$$

Based on the above, we construct the graph-based deep radar feature $F_{\mathcal{G}(R)} \in \mathbb{R}^{H\times W\times C_r} = \{\hat{F}_R^1, \hat{F}_R^2, \ldots, \hat{F}_R^i\}$.

For the textual feature $F_T \in \mathbb{R}^{L\times C}$, we first adopt max-pooling along the token dimension to extract abstract high-level semantic information. Then, to align with the radar graph feature $F_{\mathcal{G}(R)}$, a linear transformation is exerted on it and we obtain the abstract textual feature $\hat{F}_T \in \mathbb{R}^{1\times 1\times C}$. Further, to effectively suppress noise and focus on object ROI, we adopt a cross-modal gating strategy. Exactly, we obtain the weights of the linearly transformed textual features through Sigmoid activation and multiply them point-wise with the radar graph feature $F_{\mathcal{G}(R)}$. Finally, the feature is enhanced by addition with $F_{\mathcal{G}(R)}$ by a residual path. The process is shown below:

$$\hat{F}_T = \text{MaxPool}(F_T), \tag{5}$$

$$F_{R|T} = F_{\mathcal{G}(R)} \odot \sigma(\hat{F}_T \cdot W_T) + F_{\mathcal{G}(R)}, \tag{6}$$

where $\odot$ and σ denote the point-wise multiplication and Sigmoid function, respectively.

C. Deformable FPN

To enhance representation of text-conditioned three-stage PC features $\{F_{R|T}^{S_1}, F_{R|T}^{S_2}, F_{R|T}^{S_3}\}$, we first exert the efficient deformable convolutions [60] on three maps to perform adaptive sparse sampling and modeling of key features. For the current element r_0 in the one-stage radar map, the deformable convolution can be formulated as

$$y(r_0) = \sum_{g=1}^{G}\sum_{k=1}^{K} w_g m_g^k x_g(r_0 + r_k + \triangle r_g^k), \tag{7}$$

where G and K denote element aggregation groups and the feature dimension. w_g is the projection weights of the group.

Table III. Overall performances on Talk2Radar (**Best Radar**, **Best LiDAR**, Best Performance. **Car**, **Pedestrian** and **Cyclist** provide specialized **mAPs**.)

Models	Sensors	Text Encoder	Fusion	Entire Annotated Area (EAA)					Driving Corridor Area (DCA)				
				Car	Pedestrian	Cyclist	mAP	mAOS	Car	Pedestrian	Cyclist	mAP	mAOS
PointPillars (SFPN)	Radar5	ALBERT [62]	HDP	18.92	**9.79**	12.47	13.73	12.91	39.20	**10.25**	14.93	21.46	20.19
CenterFormer	Radar5	ALBERT	HDP	17.26	6.79	9.27	11.11	10.79	19.56	9.13	12.03	13.57	13.02
CenterPoint (Voxel-SFPN)	Radar5	ALBERT	HDP	18.98	5.30	14.96	13.08	12.20	40.53	8.57	15.66	21.59	20.25
PointPillars (SFPN)	Radar5	ALBERT	MHCA	5.18	5.76	6.63	5.86	3.58	13.34	4.36	8.79	8.83	7.66
CenterFormer	Radar5	ALBERT	MHCA	4.53	3.48	4.00	4.00	2.03	8.77	3.52	6.69	6.33	5.92
CenterPoint (Voxel-SFPN)	Radar5	ALBERT	MHCA	5.21	4.57	5.13	4.97	3.12	12.70	4.07	7.70	8.16	7.51
MSSG [35]	Radar5	GRU [63]	-	12.53	5.08	8.47	8.69	7.03	18.93	7.88	9.40	12.07	11.67
AFMNet [46]	Radar5	GRU	-	11.98	6.87	9.16	9.34	7.72	18.62	8.21	10.06	12.30	11.79
MSSG [35]	Radar5	ALBERT	-	16.03	5.86	10.57	10.82	8.96	25.79	8.69	12.55	15.68	14.12
AFMNet [46]	Radar5	ALBERT	-	16.31	6.80	10.35	11.15	9.46	26.82	8.71	12.45	15.99	14.18
EDA [64]	Radar5	RoBERTa [65]	-	13.23	6.60	8.63	9.49	8.93	23.55	8.80	11.95	14.77	13.07
T-RadarNet (Ours)	Radar5	ALBERT	**GGF**	**24.68**	9.71	**15.74**	**16.71**	**14.88**	**42.58**	10.13	**17.82**	**23.51**	**22.37**
CenterPoint (Pillar-SFPN)	LiDAR	ALBERT	HDP	**28.16**	6.21	17.46	17.28	16.03	43.43	6.87	27.18	25.83	24.93
CenterPoint (Pillar-SFPN)	LiDAR	ALBERT	MHCA	6.56	5.04	5.33	5.64	4.86	13.60	4.52	7.32	8.48	7.89
MSSG [35]	LiDAR	GRU	-	15.38	7.52	11.67	11.52	9.76	23.27	8.68	13.51	15.15	14.75
AFMNet [46]	LiDAR	GRU	-	16.13	7.68	12.51	12.11	9.92	24.50	9.07	13.87	15.81	15.11
MSSG [35]	LiDAR	ALBERT	-	18.19	7.66	11.63	12.49	10.91	29.61	10.98	14.66	18.42	16.23
AFMNet [46]	LiDAR	ALBERT	-	19.50	7.92	13.56	13.66	12.18	31.68	9.23	18.90	19.94	17.59
EDA [64]	LiDAR	RoBERTa	-	16.10	6.91	12.88	11.96	10.10	25.10	9.28	15.73	16.70	14.91
T-RadarNet (Ours)	LiDAR	ALBERT	**GGF**	24.91	**12.74**	**18.67**	**18.77**	**17.20**	**48.98**	**14.69**	**27.24**	**30.30**	**29.89**

m_g^k is the modulation scalar of the k-th sampling point while $\triangle r_g^k$ is the offset of sampling position r_k in the g-th group.

Then $F_{R|T}^{S_2} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 2C}$ and $F_{R|T}^{S_3} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 4C}$ are up-sampled to the same size as $F_{R|T}^{S_1}$ by transposed convolution. Finally, three-scale feature maps are concatenated to obtain high-resolution aggregated PC features $F_{\text{Agg}} \in \mathbb{R}^{H\times W\times 3C}$ with multi-receptive fields.

D. Training Objectives

The center-based detection head first produces class-wise heatmaps to predict the center location of the detected objects for each class. Afterward, the following properties of each object are supervised following CenterPoint [56]: the sub-voxel location refinement, the height-above-ground, the 3D size, and the orientation angle. The proposed T-RadarNet is trained with the following loss

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{hm}} + \beta \sum_{r\in\mathbf{\Lambda}} \mathcal{L}_{\text{smooth}-\ell_1}(\widehat{\Delta r^a}, \Delta r^a), \tag{8}$$

where $\mathcal{L}_{\text{hm}}$ is the classification loss supervising the heatmap quality of the center-based detection head using a focal loss [66]; $\mathbf{\Lambda} = \{x, y, z, l, h, w, \theta\}$ indicates the smooth-ℓ_1 loss supervising the regression of the box center (for slight modification based on heatmap peak guidance), dimensions, and orientation; and β is the weight to balance the two components of the loss, which is set to 0.25 by default.

V. Experiments

A. Implementation Settings

Model Settings: For T-RadarNet, we set the token length as 30 upon ALBERT [62] and other text encoders for comparison. N in the pillar encoder is set as 10 and 32 for radar and LiDAR, respectively, while C is 64. For comparison, we **firstly** select PC-based detectors with various paradigms including PointPillars (pillar-based) [51], CenterPoint (voxel and anchor-free) [56], and CenterFormer (transformer-based) [57]. **Secondly**, fusion methods including inductive bias-based HDP [67], attention-based MHCA [32] are implemented to compare with our GGF. **Thirdly**, different necks, *i.e.*, SecondFPN [54], ASPP [52] and CSP-FPN [68] are included for comparison with proposed Deformable FPN. **Lastly**, we also compare proposed T-RadarNet with dedicated SOTA 3D PC grounding model MSSG [35], AFMNet [46] and EDA [64].

Fig. 5. Prediction by T-RadarNet. The first row presents correct cases while the second shows problematic cases (FN: False Negative, FP: False Positive).

Dataset Settings: We train and test models on Talk2Radar dataset with three type of objects, *i.e.*, Car, Cyclist, and Pedestrian, using both 4D radar and LiDAR. Moreover, to validate the generalization of T-RadarNet, we also train and test it on Talk2Car dataset [41], which was built upon nuScenes [42] by providing textual prompt for LiDAR PC.

Training and Evaluation Settings: For Talk2Radar, all models are trained on four RTX A4000 with a batch size of 4 for 80 epochs. The initial learning rate is set 1e-3 with a cosine scheduler. We choose AdamW for optimization with a weight

Table IV. Comparison (**mAP**) of radar and LiDAR upon various prompts

Prompt	Motion			Depth			Velocity		
Sensors	Car	Pedes	Cyclist	Car	Pedes	Cyclist	Car	Pedes	Cyclist
Radar1	26.10	5.33	11.07	32.68	12.69	24.72	26.92	13.99	18.45
Radar3	35.92	11.05	15.43	40.68	18.66	31.69	33.78	20.51	**25.78**
Radar5	**36.72**	**11.54**	**16.27**	42.50	19.63	32.03	**35.50**	**20.79**	25.73
LiDAR	33.56	8.80	13.58	**45.68**	**20.54**	**38.60**	12.63	7.63	8.24

Table V. Statistics of predicted objects **mAP** by depth upon 5-frame radar

Objects	0-10 (m)	10-20	20-30	30-40	40-50	50+
Car	42.57	44.57	23.26	18.36	10.04	1.32
Pedestrian	15.54	7.17	3.14	2.06	4.54	0.0
Cyclist	29.15	11.39	14.14	3.94	1.78	0.0

decay of 5e-4. For Talk2Car, we set the initial learning rate as 1e-2 while other settings are kept the same with Talk2Radar.

B. Quantitative Results

Overall Performances: As Table III shows, on the whole, T-RadarNet outperforms other models in most aspects whatever on 4D radar or LiDAR. For other models, we find that the models based on pillar encoding perform better than those based on voxels, and the models based on self-attention do not significantly outperform those based on convolution. Regarding fusion, models using MHCA perform worse than those using gating and dot-product operations. This phenomenon is particularly evident with five-frame radar data (Radar5). **Performances on Various Prompts:** As Table IV shows, when we compare the performance of T-RadarNet on prompts with queries of types on motion, depth and velocity, the radar-based model outperforms LiDAR-based on prompts regarding motion and velocity, which is more obvious on velocity-based prompts. Besides, as LiDAR can reason the object motion upon PC object appearances, the gap between radar and LiDAR on motion query is not large. For the depth-related prompts, LiDAR obtains the advantage to some extent. **Precision by Piece-wise Depths:** Table V presents the mAP of predicted objects by segmented depths upon 5-frame radar. Small-object grounding is doubtless a challenge. **Generalization Performances:** Table VI shows the generalization evaluation results of T-RadarNet on Talk2Car for 3D REC, our T-RadarNet outperforms the baseline and the other three models, which proves the effectiveness and generalization of T-RadarNet for PC-based 3D REC.

C. Ablation Experiments

Table VII presents the ablation results of T-RadarNet. **GGF:** When we replace the MRConv4D (GConv) with normal convolution, mAP drops remarkably, which is the same circumstance with pooling for textual semantics abstraction. **Fusion Methods:** Our proposed GGF presents a better result than other fusion methods, implying a more efficient approach to aligning and embedding textual semantics to PCs. Although MHCA can model global cross-modal similarity, its performance is sub-optimal when dealing with irregular PCs containing both objects and ghosting in radar data. **Necks:** Deformable FPN outperforms all other necks, indicating its preeminent capability of representing PCs. **Text Encoders:** Pretrained transformer-based text encoders outperform RNN-based text encoders remarkably, indicating the significance of attention weighting on text encoding for the following fusion.

Table VI. Performances on Talk2Car benchmark for LiDAR-based 3D REC, where $\mathbf{AP_A}$ and $\mathbf{AP_B}$ follow MSSG [35] that define different IoU thresholds.

Models	BEV AP		3D AP	
	$\mathbf{AP_A}$	$\mathbf{AP_B}$	$\mathbf{AP_A}$	$\mathbf{AP_B}$
Talk2Car [41]	30.6	24.4	27.9	19.1
MSSG [35]	27.8	26.1	31.9	20.3
EDA [64]	37.0	29.8	37.2	20.4
AFMNet [46]	45.3	33.1	41.9	20.7
T-RadarNet	**52.8**	**39.9**	**47.2**	**30.5**

Table VII. Ablation comparison of T-RadarNet upon 5-Frame radar data

Methods	EAA			DCA		
	Car	Ped	Cyc	Car	Ped	Cyc
Gated Graph Fusion						
GConv $\rightarrow$ Conv	20.03	7.56	13.70	36.56	7.99	13.58
MaxPool $\rightarrow$ AvgPool	21.40	8.04	13.88	38.67	7.87	13.78
GConv + MaxPool	**24.68**	**9.71**	**15.74**	**42.58**	**10.13**	**17.82**
Fusion Methods						
HDP [67]	20.07	6.52	14.86	39.53	8.62	15.85
MHCA [32]	7.79	5.02	5.36	13.69	5.81	9.35
GGF	**24.68**	**9.71**	**15.74**	**42.58**	**10.13**	**17.82**
Necks (Feature Pyramid Networks)						
ASPP [52]	23.01	9.56	14.59	38.06	9.27	16.60
SecondFPN [54]	23.57	9.26	14.03	41.56	9.02	16.42
CSP-FPN [68]	23.53	9.70	14.38	41.58	9.31	17.03
DeformableFPN	**24.68**	**9.71**	**15.74**	**42.58**	**10.13**	**17.82**
Text Encoders						
Bi-GRU [63]	18.74	7.57	10.89	24.60	7.98	11.50
RoBERTa [65]	24.22	9.53	**16.12**	**42.62**	10.03	16.66
ALBERT	**24.68**	**9.71**	15.74	42.58	**10.13**	**17.82**

D. Visualization and Discussion

Fig. 5 presents the prediction results by T-RadarNet based on the 5-frame accumulated radar PC, including visualization on the image plane and BEV view in the PC context. For the correct prediction results in the first row, we can see our T-RadarNet can understand textual prompts and localize the correct objects among radar PC contexts, including single and multiple referent objects. However, there are still some problems with this task, in the second row, we can see the serious false positive cases in the predicted bounding boxes, implying that efficient representation of radar PC features and adaptive filtering of non-object clutter are still a challenge.

VI. Conclusion

This paper proposes a novel task: 4D mmWave radar-based 3D referring expression comprehension (3D visual grounding), aimed at enhancing embodied intelligence and interactive perception with an all-weather, low-cost sensor. We introduce the first dataset for this task, Talk2Radar, which fully leverages the object detection capabilities of 4D mmWave radar. Talk2Radar includes a rich collection of textual prompts and object distributions, alongside LiDAR data. Through extensive experiments, we establish a corresponding benchmark

by proposing an efficient radar-based 3D REC model, T-RadarNet. Within T-RadarNet, we design an effective module, Gated Graph Fusion, for the alignment and fusion of textual and 4D radar point cloud features. Additionally, we propose Deformable FPN to adequately model irregular and sparse point cloud features. With these contributions, we aim to advance the interactive perception capabilities of 4D radar for environmental understanding in autonomous driving.

REFERENCES

[1] Shanliang Yao, Runwei Guan, Xiaoyu Huang, Zhuoxiao Li, Xiangyu Sha, Yong Yue, Eng Gee Lim, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu, et al., "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 2094–2128, 2024.

[2] Shanliang Yao, Runwei Guan, Zitian Peng, Chenhang Xu, Yilu Shi, Yong Yue, Eng Gee Lim, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu, et al., "Radar perception in autonomous driving: Exploring different data representations," *arXiv:2312.04861*, 2023.

[3] Jianan Liu, Weiyi Xiong, Liping Bai, Yuxuan Xia, Tao Huang, Wanli Ouyang, and Bing Zhu, "Deep instance segmentation with automotive radar detection points," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 84–94, 2023.

[4] Matthias Zeller, Vardeep S Sandhu, Benedikt Mersch, Jens Behley, Michael Heidingsfeld, and Cyrill Stachniss, "Radar velocity transformer: Single-scan moving object segmentation in noisy radar point clouds," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 7054–7061.

[5] Alexander Popov, Patrik Gebhardt, Ke Chen, and Ryan Oldja, "Nvradarnet: Real-time radar obstacle and free space detection for autonomous driving," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 6958–6964.

[6] Yanlong Yang, Jianan Liu, Tao Huang, Qing-Long Han, Gang Ma, and Bing Zhu, "RaLiBEV: Radar and LiDAR BEV fusion learning for anchor box free object detection systems," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–15, 2024, 10.1109/TCSVT.2024.3521375.

[7] Jisong Kim, Minjae Seong, Geonho Bang, Dongsuk Kum, and Jun Won Choi, "Rcm-fusion: Radar-camera multi-level fusion for 3d object detection," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[8] 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)*, 2023, pp. 17615–17626.

[9] Ziyang Hong, Yvan Petillot, and Sen Wang, "Radarslam: Radar based large-scale slam in all weathers," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5164–5170.

[10] Yeong Sang Park, Young-Sik Shin, Joowan Kim, and Ayoung Kim, "3d ego-motion estimation using low-cost mmwave radars via radar velocity factor for pose-graph slam," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7691–7698, 2021.

[11] Tao Huang, Jianan Liu, Xi Zhou, Dinh C Nguyen, Mostafa Rahimi Azghadi, Yuxuan Xia, Qing-Long Han, and Sumei Sun, "V2X cooperative perception for autonomous driving: Recent advances and challenges," 2023, *arXiv:2310.03525*.

[12] Zeyu Han, Jiahao Wang, Zikun Xu, Shuocheng Yang, Lei He, Shaobing Xu, and Jianqiang Wang, "4d millimeter-wave radar in autonomous driving: A survey," *arXiv preprint arXiv:2306.04242*, 2023.

[13] Lili Fan, Junhao Wang, Yuanmeng Chang, Yuke Li, Yutong Wang, and Dongpu Cao, "4d mmwave radar for autonomous driving perception: A comprehensive survey," *IEEE Transactions on Intelligent Vehicles*, pp. 1–15, 2024, doi: 10.1109/TIV.2024.3380244.

[14] Lianqing Zheng, Sen Li, Bin Tan, Long Yang, Sihan 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*, vol. 72, pp. 1–14, 2023.

[15] Weiyi 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*, vol. 9, no. 1, pp. 79–92, 2024.

[16] Jun Zhang, Huayang Zhuge, Zhenyu Wu, Guohao Peng, Mingxing Wen, Yiyao Liu, and Danwei Wang, "4dradarslam: A 4d imaging radar slam system for large-scale environments based on pose graph optimization," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8333–8340.

[17] Shouyi Lu, Guirong Zhuo, Lu Xiong, Xichan Zhu, Lianqing Zheng, Zihang He, Mingyu Zhou, Xinfei Lu, and Jie Bai, "Efficient deep-learning 4d automotive radar odometry method," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 879–892, 2024.

[18] Jianan Liu, Guanhua Ding, Jinping Xia, Yuxuan Sun, Tao Huang, Lihua Xie, and Bing Zhu, "Which framework is suitable for online 3D multi-object tracking for autonomous driving with automotive 4D imaging radar?," in *Proceedings of the IEEE 35th Intelligent Vehicles Symposium (IV)*, 2024, *arXiv:2309.06036*.

[19] Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu, "Rcbevdet: Radar-camera fusion in bird's eye view for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[20] Felix Fent, Andras Palffy, and Holger Caesar, "Dpft: Dual perspective fusion transformer for camera-radar-based object detection," *arXiv:2404.03015*, 2024.

[21] Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya, "K-radar: 4d radar object detection for autonomous driving in various weather conditions," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3819–3829, 2022.

[22] Runwei Guan, Shanliang Yao, Xiaohui Zhu, Ka Lok Man, Eng Gee Lim, Jeremy Smith, Yong Yue, and Yutao Yue, "Achelous: A fast unified water-surface panoptic perception framework based on fusion of monocular camera and 4d mmwave radar," in *Proceedings of the IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 182–188.

[23] Runwei Guan, Shanliang Yao, Lulu Liu, Xiaohui Zhu, Ka Lok Man, Yong Yue, Jeremy Smith, Eng Gee Lim, and Yutao Yue, "Mask-vrdet: A robust riverway panoptic perception model based on dual graph fusion of vision and 4d mmwave radar," *Robotics and Autonomous Systems*, vol. 171, pp. 104572, 2024.

[24] Shanliang Yao, Runwei Guan, Zhaodong Wu, Yi Ni, Wen Ryan Liu, Zile Huang, Xiaohui Zhu, Yong Yue, Eng Gee Lim, Hyungjoon Seo, et al., "Waterscenes: A multi-task 4d radar-camera fusion dataset and benchmark for autonomous driving on water surfaces," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[25] Runwei Guan, Liye Jia, Shanliang Yao, Fengyufan Yang, Sheng Xu, Erick Purwanto, Xiaohui Zhu, Ka Lok Man, Eng Gee Lim, Jeremy Smith, et al., "Watervg: Waterway visual grounding based on text-guided vision and mmwave radar," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[26] Runwei Guan, Shanliang Yao, Ka Lok Man, Xiaohui Zhu, Yong Yue, Jeremy Smith, Eng Gee Lim, and Yutao Yue, "Asy-vrnet: Waterway panoptic driving perception model based on asymmetric fair fusion of vision and 4d mmwave radar," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 12479–12486.

[27] Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus, "Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models," *arXiv:2310.17642*, 2023.

[28] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al., "A survey on multimodal large language models for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 958–979.

[29] Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li, "Large multimodal agents: A survey," *arXiv:2402.15116*, 2024.

[30] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li, "Drivelm: Driving with graph visual question answering," *arXiv:2312.14150*, 2023.

[31] Chang Liu, Henghui Ding, and Xudong Jiang, "Gres: Generalized referring expression segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 23592–23601.

[32] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu

Zhang, and Jianbing Shen, "Referring multi-object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14633–14642.

[33] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes," in *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 2020, pp. 422–440.

[34] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu, "3dvg-transformer: Relation modeling for visual grounding on point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2928–2937.

[35] Wenhao Cheng, Junbo Yin, Wei Li, Ruigang Yang, and Jianbing Shen, "Language-guided 3d object detection in point cloud for autonomous driving," *arXiv:2305.15765*, 2023.

[36] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang, "Lidar-llm: Exploring the potential of large language models for 3d lidar understanding," *arXiv:2312.14074*, 2023.

[37] Georg Hess, Adam Tonderski, Christoffer Petersson, Kalle Åström, and Lennart Svensson, "Lidarclip or: How i learned to talk to point clouds," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 7438–7447.

[38] Mariia Pushkareva, Yuri Feldman, Csaba Domokos, Kilian Rambach, and Dotan Di Castro, "Radar spectra-language model for automotive scene parsing," 2023.

[39] Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll, "Vision language models in autonomous driving: A survey and outlook," *IEEE Transactions on Intelligent Vehicles*, 2024, doi:10.1109/TIV.2024.3402136.

[40] Andras Palffy, Ewoud Pool, Srimannarayana Baratam, Julian FP Kooij, and Dariu M Gavrila, "Multi-class road user detection with 3+ 1d radar in the view-of-delft dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, 2022.

[41] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie Francine Moens, "Talk2car: Taking control of your self-driving car," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 2088–2098.

[42] 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," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11621–11631.

[43] Yang Zhan, Yuan Yuan, and Zhitong Xiong, "Mono3dvg: 3d visual grounding in monocular images," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024, vol. 38, pp. 6988–6996.

[44] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.

[45] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen, "Language prompt for autonomous driving," *arXiv preprint arXiv:2309.04379*, 2023.

[46] Ali Solgi and Mehdi Ezoji, "A transformer-based framework for visual grounding on 3d point clouds," in *the 20th IEEE CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)*, 2024, pp. 1–5.

[47] Jiuming Liu, Guangming Wang, Weicai Ye, Chaokang Jiang, Jinru Han, Zhe Liu, Guofeng Zhang, Dalong Du, and Hesheng Wang, "Difflow3d: Toward robust uncertainty-aware scene flow estimation with iterative diffusion-based refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15109–15119.

[48] Liye Jia, Runwei Guan, Haocheng Zhao, Qiuchi Zhao, Ka Lok Man, Jeremy Smith, Limin Yu, and Yutao Yue, "Radarnext: Real-time and reliable 3d object detector based on 4d mmwave imaging radar," *arXiv preprint arXiv:2501.02314*, 2025.

[49] Lianqing Zheng, Jianan Liu, Runwei Guan, Long Yang, Shouyi Lu, Yuanzhe Li, Xiaokai Bai, Jie Bai, Zhixiong Ma, Hui-Liang Shen, et al., "Doracamom: Joint 3d detection and occupancy prediction with multi-view 4d radars and cameras for omnidirectional perception," *arXiv preprint arXiv:2501.15394*, 2025.

[50] Yuanyuan Zhang, Runwei Guan, Lingxiao Li, Rui Yang, Yutao Yue, and Eng Gee Lim, "radarode: An ode-embedded deep learning model for contactless ecg reconstruction from millimeter-wave radar," *arXiv preprint arXiv:2408.01672*, 2024.

[51] 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)*, 2019, pp. 12697–12705.

[52] Jinyu Li, Chenxu Luo, and Xiaodong Yang, "Pillarnext: Rethinking network designs for 3d object detection in lidar point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 17567–17576.

[53] Yin Zhou and Oncel Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4490–4499.

[54] Yan Yan, Yuxing Mao, and Bo Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, pp. 3337, 2018.

[55] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia, "Voxelnext: Fully sparse voxelnet for 3d object detection and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 21674–21683.

[56] 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)*, 2021, pp. 11784–11793.

[57] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh, "Centerformer: Center-based transformer for 3d object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 496–513.

[58] Jianan Liu, Qiuchi Zhao, Weiyi Xiong, Tao Huang, Qing-Long Han, and Bing Zhu, "SMURF: Spatial multi-representation fusion for 3d object detection with 4d imaging radar," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 799–812, 2024.

[59] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[60] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al., "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14408–14419.

[61] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem, "Deepgcns: Can gcns go as deep as cnns?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9267–9276.

[62] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv:1909.11942*, 2019.

[63] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[64] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang, "Eda: Explicit text-decoupling and dense alignment for 3d visual grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19231–19242.

[65] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv:1907.11692*, 2019.

[66] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[67] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji, "Seqtr: A simple yet universal network for visual grounding," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 598–615.

[68] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.