Cross-Modal Supervision Based Road Segmentation and Trajectory Prediction With Automotive Radar

Zhaoze Wang, Yi Jin, *Graduate Student Member, IEEE*, Anastasios Deligiannis, *Member, IEEE*, Juan-Carlos Fuentes-Michel, and Martin Vossiek, *Fellow, IEEE*

***Abstract*—Automotive radar plays a crucial role in providing reliable environmental perception for autonomous driving, particularly in challenging conditions such as high speeds and bad weather. In this domain the deep learning-based method is one of the most promising approaches, but the presence of noisy signals and the complexity of data annotation limit its development. In this letter, we propose a novel approach to address road area segmentation and driving trajectory prediction tasks by introducing Differential Global Positioning System (DGPS) data to generate labels in a cross-modal supervised manner. Then our method employs a multi-task learning-based CNN trained by radar point clouds or occupancy grid maps without any manual modification. This multi-task network not only boosts processing efficiency but also enhances the performances in both tasks, compared with single-task counterparts. Experimental results on a real-world dataset demonstrate the effect of our implementation qualitatively, achieving decimeter-level predictions within a 100 m forward range. Our approach attains an impressive 91.4% mean Intersection over Union (mIoU) in road area segmentation and exhibits an overall average curve deviation of less than 0.35 m within a range of 100 m forward in trajectory prediction.**

***Index Terms*—Intelligent transportation systems, semantic scene understanding, automotive radar, cross-modal learning.**

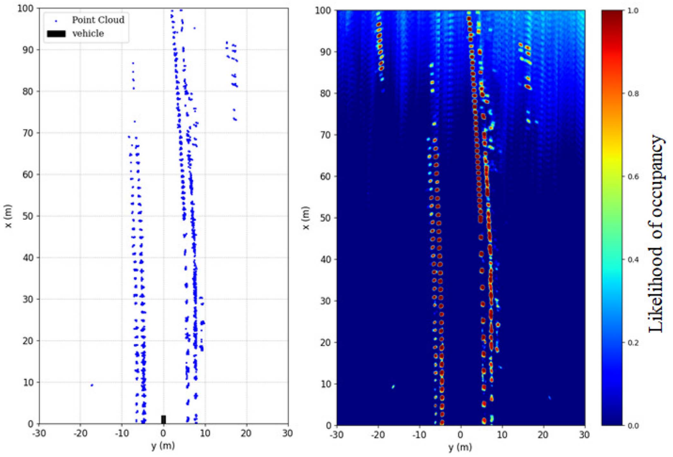


Fig. 1. Visualization of point cloud detection (left) and OGM (right). On the left figure, the blue points represent 2D point cloud of obstacles detected by radar. The black rectangular block in the figure represents the current location of the vehicle. On the right figure, the state of occupancy for the surrounding environment is visualized through the color bar on the right side.

I. Introduction

In OUTDOOR autonomous driving scenarios, cameras struggle with lighting variations and textureless objects, while the performance of lidar is significantly hampered by adverse weather conditions, such as heavy fog, and the widespread integration is limited by its high costs [1]. In contrast, despite a notable drawback of relatively constrained angular resolution, radar sensors exhibit robustness in the face of extreme weather conditions and excel in the instantaneous measurement of Doppler velocity [2]. Radar Occupancy Grid Map (OGM) is one of the relevant technologies, which discretizes the surrounding environment into 2D or 3D cells in a geometric manner and probabilistically represents the occupancy-state estimation of each cell.

While radar OGM is an effective tool for perceiving spatial-physical information, it falls short in providing the underlying semantic details of the environment. In the contemporary landscape of research, prominent scholars are increasingly turning to deep learning-based architectures to address pivotal challenges in radar signal processing [3]. However, the majority of their implementations need manual annotation of ground truth, requiring massive effort and time. Moreover, manual labeling is challenging because the representation of the Range-Doppler-Azimuth tensor is not trivially comprehensible, as shown in [4]. Therefore, the non-intuitive property of radar signals poses a substantial impediment to the advancement of this field.

In recent years, some research efforts have proposed the utilization of unsupervised learning or cross-modal supervised learning methods to alleviate the burdensome and complex tasks in radar data annotation [5], [6], [7], [8] and [9]. These innovative approaches ingeniously transform signals from one modality into supervision signals for another and avoid labeling manually, which showcases a promising avenue for circumventing the labor-intensive process of data annotation in various domains, particularly within the radar field. Therefore, it is meaningful to utilize the advantages of various sensors in a cross-modal

Manuscript received 24 March 2024; accepted 22 July 2024. Date of publication 7 August 2024; date of current version 13 August 2024. This article was recommended for publication by Associate Editor Juan D. Hernandez and Editor Aniket Bera upon evaluation of the reviewers' comments. This work was supported by BMW Group. *(Corresponding author: Yi Jin.)*

Zhaoze Wang, Yi Jin, and Martin Vossiek are with the Institute of Microwaves and Photonics (LHFT), Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany (e-mail: zhaoze.wang@fau.de; yi.y.jin@fau.de; martin.vossiek@fau.de).

Anastasios Deligiannis and Juan-Carlos Fuentes-Michel are with the BMW Group, 80809 Munich, Germany (e-mail: Anastasios.Deligiannis@bmw.de; Juan-Carlos.Fuentes-Michel@bmw.de).

The source code is available at https://wangzhaoze.github.io/RADxGPS/.

Digital Object Identifier 10.1109/LRA.2024.3440093

fashion and further improve radar capabilities in the perception of the static road environment by incorporating deep learning methodologies.

Contributions: Our contributions are as follows: (i) Development of radar and DGPS signals-based cross-modal supervised learning pipeline for road area segmentation and continuous trajectory prediction tasks, which eliminates the need for manual annotation; (ii) Through comprehensive comparisons, our approach outperforms previous works in both tasks, especially the heatmap regression approach has been demonstrated to improve the performance in the trajectory prediction task, which will be evaluated in Section VI; (iii) Finally, we integrate the aforementioned two tasks using a multi-task learning method, which performs more efficiency and computational-efficiently compared to individually executing each task in comparative results.

In Section II, we'll overview background knowledge and related works involving radar signal processing and cross-modal supervised learning. Subsequently, in Section III the data preprocessing and label generation will be introduced. Then, in Section V we will detail the training procedure, including network architecture and training parameters. In Section VI we will evaluate our results qualitatively and quantitatively, and summarize the approach and conclusion in Section VII.

II. Related Works

A. Radar Occupancy Grid Map

In [10], OGMs were adapted for automotive radar systems, recognizing static and dynamic detections using Doppler velocity, ego-motion, and radar mounting position information utilizing the Inverse Sensor Model (ISM). Typically, static detections are retained to form grid maps and accumulated over time to depict the shape and energy of environment detection [11]. However, this filtering-based approach is time- and resource-intensive, lacking real-time state estimation.

Recent efforts have applied deep learning methods to address these drawbacks in radar OGM. Stojcheski et al. [12] train OGMs with future time-step data to enhance state completion and map perception. Jin et al. [13] proposed a novel radar data-driven ISM for static and dynamic obstacles segmentation on BEV (Bird's Eye View). Moreover, in [5] and [14], OGMs or radar 3D point clouds are utilized to estimate road courses directly, and the latter avoids OGM generation.

Inspired by these works, we map sparse radar point cloud data combining SNR (Signal-Noise-Ratio) onto the BEV plane, feeding transformed data into a CNN-based model. In comparison to using OGM or 3D point cloud as input, our results demonstrate that point cloud clustering can almost entirely replace them, achieving even better outcomes in trajectory prediction tasks.

B. Cross-Modal Supervised Learning

Self-supervised learning typically involves only one class of sensor or signal modality, which means that the signal sampling at one specific time supervises the signals at another time. In contrast, cross-modal supervised learning establishes supervision associations between multi-modal data at the same timestamp for eliminating manual labeling. A camera-based training is extended to road users, using segmented images as supervision for training with radar data [15]. Wang et al. [7] developed a camera-radar fusion framework for training RODNet without human labeling, while Huang et al. [9] introduced contrastive loss to achieve consistency between radar and LiDAR latent representations. Approaches in [5] and [14] resemble ours, estimating discrete positions of road courses or driving trajectories by integrating DGPS signals and generated radar OGM or 3D point clouds. However, we incorporate simpler 2D point clouds with SNR for continuous trajectory predictions instead. Comparative outcomes are discussed in Section VI.

C. Semantic Segmentation

Semantic segmentation in road scenes is a crucial technology in the fields of autonomous driving and deep learning, providing the pixel-wise perception of the categories of all static structures and dynamic objects in the surrounding environment. [11] and [2] have achieved multi-category semantic segmentation of OGM on both simulator and real-world datasets, encompassing categories such as street, curbstone, car, fence, and background. However, their labels are manually annotated by a trained person. In contrast, our labels are derived solely from ECEF coordinates provided by the DGPS. We obtain satellite images from Google Maps, and the pixel-level category outputs are obtained through SAM [16] (Segment Anything Model). None of manual annotations are added in this process, and our approach achieves superior results in evaluation comparisons of road area segmentation.

D. Trajectory Prediction

Autonomous vehicles need to achieve real-time path planning and decision-making based on the surrounding environment. Accurate trajectory prediction can assist vehicles in selecting safer and more efficient driving paths and making appropriate driving decisions, such as lane changes, overtaking, and obstacle avoidance. Typically, few works focus on predicting trajectories using automotive radar. Many rely on algorithms employing analytical models for road course descriptions. For instance, [17] models roads as circles, incorporating dynamic detections through tracking algorithms to determine vehicle trajectories. Alternatively [22] explores grid-based approaches by leveraging the estimated trajectories of other vehicles in the process.

We draw inspiration from [5], [14] and project sequential Latitude, Longitude, and Altitude (LLA) coordinate recordings into the vehicle frame, serving as training labels for the vehicle's future trajectory. A notable distinction lies in the fact that their approach applies an image classification neural network and treats the output vector as estimated positions. In contrast, we consider potential spatial properties, producing a dense heat map for the representation of the probabilistic future trajectory likelihood. This heat map is utilized for regressing continuous driving curves in the second stage, rendering our results more robust for various situations in comparison, and has demonstrated better results qualitatively on both straight and curved roads.

E. Multi-Task Learning

Multi-task Learning (MTL) is a promising paradigm designed to leverage representational information from multiple related tasks to improve the generalization performance of all the tasks simultaneously [18]. These tasks usually use a shared encoder to extract common latent features, followed by task-specific decoders to generate predictions for each task, which can lead to superior performance for each task compared to single-task models, because the implementation of MTL can reduce overfitting to specific tasks and serve a function similar to regularization [19]. Additionally, MTL models can reduce memory usage, energy consumption, and inference latency compared to single-task models by avoiding the recomputation of features in shared layers [20]. In recent years, several previous works, including [15] and [21], have applied this method to scene understanding tasks in autonomous driving, demonstrating significant potential.

In this work, we hypothesize a connection between road area segmentation and driving trajectory prediction. On one hand, the shape of the road area partially dictates the driving trajectory. On the other hand, when the edges of the road are undetectable by radar, driving trajectory information can help infer the road area or course. We conducted multi-task learning experiments and compared the results with those from single-task experiments. Quantitative results show that simultaneously training these two tasks enhances the network's performance on each specific task and reduces storage and inference time.

III. Data Preparation

In this work, we gathered a dataset from test vehicles operating in Germany and Italy, equipped with a high-precision Differential Global Positioning System (DGPS) and a 76 GHz Frequency Modulated Continuous Wave (FMCW) radars. The maximum range of radar detection is about 300 meters, with a range resolution of 0.05 meters, and azimuth and elevation resolutions of 1° and 2.5°. In comparison to radar, DGPS boasts a sampling frequency 5 times higher and provides precise Earth-centered, Earth-Fixed geographical location information, encompassing longitude, latitude, and altitude. The closest sampling moment in time can be determined based on the radar signal timestamp to synchronize it with the radar signal sampling. With an error at the millisecond level, which can be neglected.

A total of 12 groups of recordings were acquired, with each group spanning approximately 200 seconds. 8 groups were captured on highways, featuring an average speed of approximately $100\,\mathrm{km/h}$. These highway environments were characterized by road boundaries constructed with guardrails. The remaining 4 groups of data were obtained from rural roads, exhibiting an average speed of around $75\,\mathrm{km/h}$. The road boundaries in rural settings encompassed guardrails, concrete barriers, and vegetation such as plants and grass, leading to less distinct radar-detectable road edges.

According to the distribution of data and road environmental conditions, we split the whole dataset into 10 groups for training and validation, and 2 groups for testing. We sample data of each group at every fixed time interval because surrounding environment changes are not significant at high speeds driving states with radar's 50 ms sampling cycle. After removing frames with stationary vehicles and invalid DGPS signals where satellite images are not able to be visited through Google Map, we obtained the training and validation set of 1500 frames and the test set of 200 frames.

IV. Methodology

A. Radar Data Processing

Point Cloud Map: Following signal processing, the sparse 3D point cloud set $\{Pi\} \in \mathbb{R}^{N_{\text{Det}}\times 3}$ with SNR reflected by the surrounding environment is captured in the vehicle coordinate system at a rate of 20 frames of detection per second. The radar point cloud in the vehicle coordinate system is then projected onto BEV. Points are clustered as pillars and the corresponding intensity value is set to the SNR of the point. Subsequent steps include discretization, crop of the generated grid, and creation of a two-dimensional point cloud map in a 1000×1000 format. The sensing range is confined to $[0\,\mathrm{m}, 100\,\mathrm{m}]$ on the x-axis and $[-50\,\mathrm{m}, 50\,\mathrm{m}]$ on the y-axis, with each pixel corresponding to 0.1 meters in the real world. As illustrated in Fig. 1, the point at the bottom center of the image represents the vehicle's position, the vertical vector denotes the instantaneous forward direction, and obstacles are depicted as dots in the image.

Occupancy Grid Map: To generate an occupancy grid map m, the space around the vehicle is discretized into a range of cells. Each cell is assigned a corresponding value $m_{u,v}$ to represent the probability of the point at position (u, v) being occupied. In this work, we configured each cell of the generated radar OGM also to 0.1 meters, the same as PCM, and further cropped it into a resolution of 1000×1000 image. If $m_{u,v} = 1$, this means *occupied*, while $m_{u,v}$ represent $free$ and 0.5 signifies an $unknown$.

Mathematically, the OGM algorithm aims to estimate the posterior likelihood

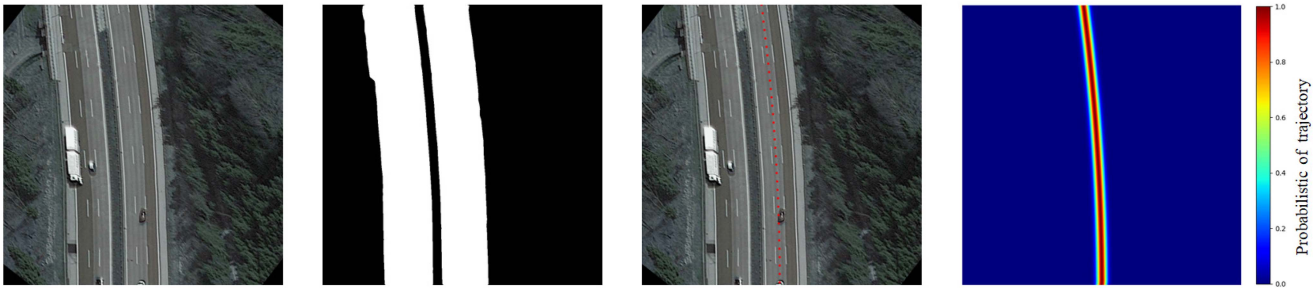
$$p(\mathbf{m}|z_{1:t}, x_{1:t}) \tag{1}$$

where $z_{1:t}$ and $x_{1:t}$ represent the sensor measurements and vehicle positions respectively, from the moment 1 to t.

The generation of OGM relies on two fundamental components: the ISM and Bayesian filtering. The ISM estimates the occupancy likelihood of a cell based on a single measurement, while Bayesian filtering recursively updates this prediction. Directly estimating the probability 1 is computationally infeasible, therefore it is conventionally assumed that the states of individual cells are independent to simplify the calculations. Then the probability 1 can be computed as following:

$$p(\mathbf{m}|z_{1:t}, x_{1:t}) = \prod_{u,v} p(\mathbf{m}_{u,v}|z_{1:t}, x_{1:t}) \tag{2}$$

To avoid numerical instability caused by extremely small or large likelihood values during calculations, a log-odds representation



(a) Satellite image (b) Segmentation label (c) Vehicle ego-motion (d) Heat-map trajectory label

Fig. 2. Visualization of an example of the dataset. Each figure from left to right represents a directional aligned satellite image, a segmented road area label (white means road and black for background), the vehicle ego-motion trajectory (marked as red points on the satellite image), and a generated probabilistic dense heat map of trajectory, where red area means a higher probability that vehicle will pass by.

is preferred:

$$s_t(\mathbf{m}|z_{1:t}, x_{1:t}) = \log \frac{p(\mathbf{m}_{u,v}|z_{1:t}, x_{1:t})}{1 - p(\mathbf{m}_{u,v}|z_{1:t}, x_{1:t})} \tag{3}$$

Finally, when using Bayesian filtering to recursively update the occupancy likelihood at cell (u, v) recursively, the following equation could be derived:

$$s_t(\mathbf{m}|z_{1:t}, x_{1:t}) = s_t(\mathbf{m}|z_t, x_t) + s_{t-1}(\mathbf{m}|z_{1:t-1}, x_{1:t-1}) - s_0 \tag{4}$$

where s_0 as the occupancy log-odd prior is usually zero, and the term $s_t(\mathbf{m}|z_t, x_t)$ denotes the ISM.

B. DGPS Data Processing

Satellite Image Segmentation: For the training procedure in a cross-modal manner, the DGPS signal must be converted into road-segmented satellite images at every synchronized timestamp to match the corresponding 2D point cloud map (PCM) or OGM input in the network. Therefore, we input vehicle locations into the Google Map API, which provides users with an interface to download high-resolution satellite images based on the provided parameters. We set the scale factor to 2 and zoom to 20 respectively and downloaded satellite images with a scale of approximately 0.0962 m in the real world for each pixel, which is almost consistent with 2D PCM and OGM images used as network input. Then we can rotate and correct the satellite image around the image center according to the approximate current driving direction vector of the vehicle, crop it to a suitable size, as shown in Fig. 2(a), and input it into the pre-trained Segment Anything Model (SAM) [16] for getting semantic segmentation labels at the same time-stamp of the trained radar data, like Fig. 2(b).

Heat-map Trajectory Generation: To represent trajectory in heat-map, we could firstly sample N_s ECEF coordinates of future positions and transform them into the current vehicle coordinate frame until out of the range of radar detection. This process yields N_s discrete 2D point-wise ground truth trajectory described as $T_{gt} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{N_s}, y_{N_s})\} \in \mathbb{R}^{N_s \times 2}$, where x-coordinate represents the forward direction and y-coordinate the horizontal direction of driving. The red points annotated in Fig. 2(c) are examples of these positions with the sampling interval of 0.1 s. Then we convert the trajectory T_{gt} into a set of points $T'_{gt} = \{(1, y_1), (2, y_2), \ldots, (1000, y_{1000})\} \in \mathbb{R}^{1000 \times 2}$ using quadratic interpolation, which means for each $x_i \in \{1, 2, \ldots, 1000\}$, there is a corresponding ground truth horizontal position $y_{\mathrm{gt},\mathrm{x}_i}$ in the vehicle coordinate. Finally, the trajectory label is no longer expressed as 1000 positions, but a dense heat map where the probabilities at each row index x in the image follow a one-dimensional Gaussian distribution

$$\mathcal{H}(x, y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y - y_{\mathrm{gt},\mathrm{x}})^2}{2\sigma^2}} \tag{5}$$

with the mean as $y_{\mathrm{gt},\mathrm{x}}$ and sigma as 10 cells of the model. An example of obtained dense trajectory heat-map labels is visualized in Fig. 2(d) with the magnitude normalized as 1.

V. Training Details

A. Pipeline

Fig. 3 illustrates the complete pipeline of implementing the DGPS cross-modal supervised Radar procedure, expressed as ‘Teacher’ and ’Student’ respectively. Initially, the radar signal is processed to generate OGM or 2D Point Cloud Map as input. Simultaneously, the DGPS signal undergoes processing using methods explained in Section III to generate satellite image segmentation and trajectory heat maps under BEV at the corresponding time, serving as pseudo labels for supervised training.

B. Data Module

To expand the dataset, we subsequently added random rotations, or mirroring operations along the x-axis of the vehicle coordinate system as data augmentations. Due to the relatively small dataset size, a 3-fold cross-validation is performed, resulting in three different splits for training and test data. Since the radar detection range is limited to the fences on both sides of the current road, to allow the network to focus on the lane in which the vehicle is currently traveling, we crop the middle of the image to a width of 400 pixels and height 1000 pixels, which means 100 meters in front of the vehicle and 20 meters on the left and right sides.

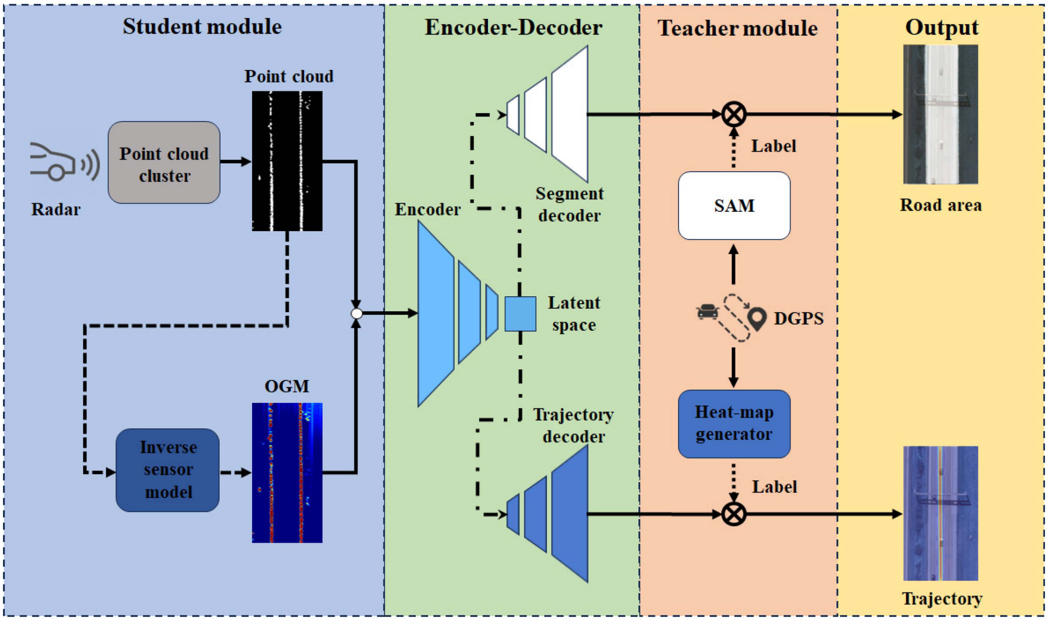


Fig. 3. Schematic pipeline of cross-modal supervised multi-task learning structure based on radar as ‘Student’ and DGPS as ’Teacher’. First, the 2D point cloud cluster from the processed radar signal, or the previous generated OGM, is input into an encoder with shared weights. The associated feature will be fed into different decoders to get the corresponding task-specific output. Labels involve road area and continuous trajectory heat-map, and both are derived from the signals of DGPS, resulting in different losses.

C. Network Structure

In single-task, only one encoder-decoder network structure is configured. Here we chose the U-Net structure [23]. In multi-task learning, to extract multi-task features of radar data and support the simultaneous output of multiple results, two networks share a common weight of encoder for extracting multi-task features and assigning a specific decoder to each output to support multiple task-specific outputs.

D. Hyperparameters

The entire training process uses the Adam optimizer, with a learning rate of 1e-5, weight decay of 1e-8, and training for 50 epochs. The experiment was conducted on a Linux system with an Nvidia GTX3060 graphics card.

E. Loss Funtions

In the road segmentation task, since this is a pixel-level binary classification task, we use Binary Cross-Entropy Loss:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \log(p_i) + (1-y_i)\cdot \log(1-p_i)) \quad (6)$$

In the trajectory prediction task, where the vehicle trajectory does not occupy a significant proportion of the image, there is an imbalance in the sample proportion between the curve part and the background part. Consequently, Focal Loss [24] is selected to address the category imbalance issue effectively.

The Focal Loss balances the weight term $-\alpha_t(1-p_t)^\gamma$ of easily and hard discriminable samples with a tunable focusing parameter γ, allowing the model to concentrate more on learning challenging categories. This approach proves particularly beneficial when dealing with the uneven distribution between the road curve and the background parts of the samples.

$$L_{\text{Focal}}(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \quad (7)$$

In multi-task learning networks, we weighted the two losses in the same way as [15]. For finding the optimal solution, we adjusted the weights from $\alpha_1 = 0.5$ and $\alpha_2 = 0.5$ as the start, and observed that assigning a slightly higher weight to α_2 yields better results. Finally, we got the best performance with the weights of α_2 as 0.7 and α_1 as 0.3.

$$L_{\text{Multi}} = \alpha_1 \cdot L_{\text{BCE}} + \alpha_2 \cdot L_{\text{Focal}} \quad (8)$$

VI. Experimental Results

A. Road Area Segmentation

For the pixel-wise classification task, we choose mean Intersection over Union (mIoU) and recall as evaluation metrics and compare our results with [2], which contained multiple datasets and various input forms. Thus, we selected the most similar setup to our process (input form E1 with 3 input channels and dataset 3 consisting of 1536 × 512 high-resolution images). The results are presented in the Table I, with the best values highlighted in bold.

The overall performance of our approach is superior to the baseline. Although the results using 2D point cloud as input are marginally inferior to those using OGM, the difference is insignificant. This indicates that the grid map generation step can be bypassed to obtain task-specific output directly, even though the grid map contains more probabilistic space occupation information.

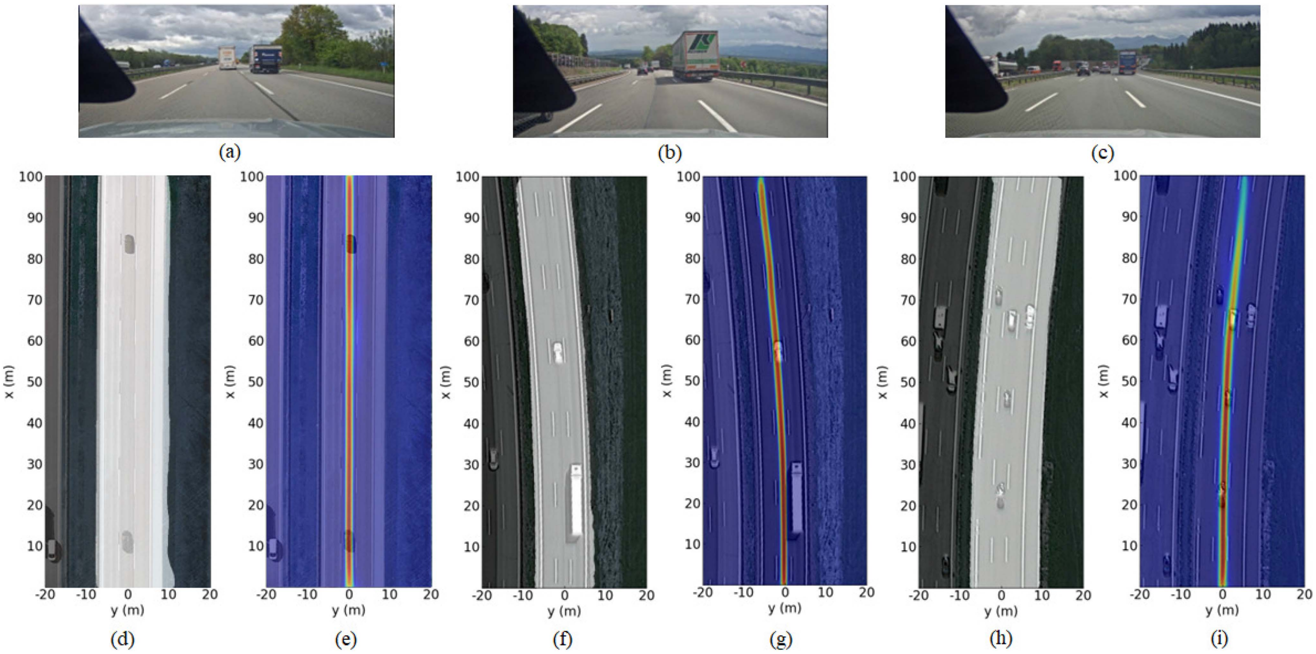


Fig. 4. Qualitative visualization of results. (a)–(c) Represents the capture of the front camera, and (d)–(i) represents the masked satellite image by road area segmentation and trajectory prediction from the output at the same time-stamp to the above three captures of the front camera. The segmented road area in satellite images (d), (f), and (h) is masked with white. In (e), (g), and (i) the estimated trajectories heat maps are visualized as masks on corresponding satellite images which red represents higher probabilities of the vehicle's position in the future.

TABLE I
ROAD AREA SEGMENTATION

	Input	Network	mIoU	Recall
[2]	Occupancy Grid Map	SegNet-1	0.850	0.959
	Occupancy Grid Map	SegNet-2	0.835	0.953
	Occupancy Grid Map	DeepLabv3+	0.876	0.946
Ours	Occupancy Grid Map	U-Net	0.908	0.966
	Occupancy Grid Map	Multi-Task	**0.914**	**0.969**
	2D Point Cloud Map	U-Net	0.893	0.944
	2D Point Cloud Map	Multi-Task	0.862	0.955

TABLE II
TRAJECTORY PREDICTION

	Input	Network	(9)	(10)
[5]	Occupancy Grid Map	AlexNet	0.3763m	0.6270m
	Occupancy Grid Map	DenseNet-201	0.3638m	0.5837m
	Occupancy Grid Map	NasNet	0.3779m	0.6032m
	Occupancy Grid Map	ShuffleNet	0.8621m	1.0964m
[14]	3D Point Cloud [1]	PointNet++	0.3458m	0.5630m
	3D Point Cloud [2]	PointNet++	0.3573m	0.5680m
	3D Point Cloud [3]	PointNet++	0.4552m	0.7822m
Ours	Occupancy Grid Map	U-Net	0.3669m	0.5141m
	Occupancy Grid Map	Multi-Task	0.3437m	0.5049m
	2D Point Cloud Map	U-Net	0.3576m	0.4932m
	2D Point Cloud Map	Multi-Task	**0.3141m**	**0.4853m**

B. Trajectory Prediction

In the trajectory prediction task, to ensure a fair comparison between our implementation and the baseline, we also adopt Formulas (9) and (10) from [5] as the evaluation criteria. Similarly, we set the reference point for evaluation to $y_{est,i,j}$ for $i \in [1, 6]$ when $i \in [1, 200]$ and $x_{est,i,j} \in \{\pm 75, \pm 50, \pm 25\}$. Given that our approach only estimates the path in front of the vehicle, we assume symmetry with the vehicle as the origin in the negative x-direction.

$$\frac{1}{6}\cdot\sum_{j=1}^{6}\sqrt{\frac{1}{N_T}\cdot\sum_{i=1}^{N_T}\left(y^{(vc)}_{\text{pred},i,j}-t_{i,j}\right)^2} \tag{9}$$

$$\frac{1}{2}\cdot\sum_{j=1,6}\sqrt{\frac{1}{N_T}\cdot\sum_{i=1}^{N_T}\left(y^{(vc)}_{\text{pred},i,j}-t_{i,j}\right)^2} \tag{10}$$

Fig. 4 shows that the predicted trajectory is presented as dense heat map regression masked on satellite images. To derive the final path from the output image, each row in the image (x-axis of the vehicle coordinate system) is normalized as the weight for each y-index using Softmax. Multiplying these weights by the corresponding y-coordinates yields predicted trajectory $\{T_{\text{pred},i}\} \in \mathbb{R}^{N_s \times 2}$. The quantitative results in Table II compared with [5] and [14], which output discrete points as trajectory directly, demonstrates that our method results in a more stable and robust estimation.

Additionally, we reached another conclusion that OGM is not always the best option, and point clouds with SNR information can achieve more appealing results. From another perspective, the results in [14] outperform those in [5], further supporting this observation.

C. Multi-Task Learning

Compared to setting up a dedicated network model for each task, we quantitatively demonstrate in the Table III that by constructing a multi-task learning network, we only need one

TABLE III
MULTI-TASK ANALYSIS

Task	Time for 10 frames	Total parameters
Multi-Tasks	1.549 s	85.45 M
Road Segmentation	1.232 s	82.66 M
Trajectory Prediction	1.275 s	82.66 M
Sum of Above	2.507 s	165.32 M

encoder for extracting features and a task-specific decoder for each output, thus reducing the total trainable parameters and inference time of the model. Additionally, in the Tables I and II, we observe that when the tasks are selected appropriately, or when there is a correlation between tasks, they can be mutually reinforcing during the training of the multi-task learning network, resulting in a certain performance improvement compared to the single-task approach.

At the same time, we observed from the results that our method remains effective even when the road edges are partially or completely undetectable. An example is shown in Fig. 4. In the view of the vehicle's front camera in Fig. 4(a), there is a lack of radar-reflective obstacles such as fences on the right side of the road. Despite the absence of radar detection returns, the network still performs well and derives robust predictions for road area segmentation and trajectory prediction, as qualitatively shown in Fig. 4(d) and (e).

VII. CONCLUSION

In this letter, we employ a cross-modal supervised approach using DGPS signals to indirectly label radar data in autonomous driving. This overcomes labeling challenges and achieves competitive results in road segmentation and trajectory prediction tasks compared to previous works. In road segmentation, we download and correct satellite images with Google Map API and ECEF coordinates for obtaining pixel-level ground truth by a pre-trained model. In trajectory prediction, we map global coordinates to the vehicle frame and use a probabilistic heat map regression for robust path prediction. Finally, we exploit the potential correlation between these two perceptual tasks and use multi-task learning, coupled with task-specific loss functions to achieve simultaneous multi-task outputs more efficiently, which provides a feasible solution for autonomous driving applications. While our focus is on static environmental perception at high speeds, addressing dynamic road participants remains a future challenge.

REFERENCES

[1] Y. Li and J. Ibanez-Guzman, "LiDAR for autonomous driving: The principles, challenges, and trends for automotive LiDAR and perception systems," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 50–61, Jul. 2020.

[2] R. Prophet, A. Deligiannis, J.-C. Fuentes-Michel, I. Weber, and M. Vossiek, "Semantic segmentation on 3D occupancy grids for automotive radar," *IEEE Access*, vol. 8, pp. 197917–197930, 2020.

[3] Z. Geng, H. Yan, J. Zhang, and D. Zhu, "Deep-learning for radar: A survey," *IEEE Access*, vol. 9, pp. 141800–141818, 2021.

[4] O. Schumann, J. Lombacher, M. Hahn, C. Wöhler, and J. Dickmann, "Scene understanding with automotive radar," *IEEE Trans. Intell. Veh.*, vol. 5, no. 2, pp. 188–203, Jun. 2020.

[5] R. Prophet, Y. Jin, J.-C. Fuentes-Michel, A. Deligiannis, I. Weber, and M. Vossiek, "CNN based road course estimation on automotive radar data with various gridmaps," in *Proc. IEEE Int. Conf. Microw. Intell. Mobility*, 2020, pp. 1–4.

[6] L. Sless, B. El Shlomo, G. Cohen, and S. Oron, "Road scene understanding by occupancy grid learning from sparse radar clusters using semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019.

[7] Y. Wang, Z. Jiang, X. Gao, J.-N. Hwang, G. Xing, and H. Liu, "RODNet: Radar object detection using cross-modal supervision," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 504–513.

[8] Y. Jin et al., "Radar and LiDAR deep fusion: Providing doppler contexts to time-of-flight LiDAR," *IEEE Sensors J.*, vol. 23, no. 20, pp. 25587–25600, Oct. 2023.

[9] J.-T. Huang et al., "Cross-modal contrastive learning of representations for navigation using lightweight, low-cost millimeter wave radar for adverse environmental conditions," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 3333–3340, Apr. 2021.

[10] R. Prophet, H. Stark, M. Hoffmann, C. Sturm, and M. Vossiek, "Adaptions for automotive radar based occupancy gridmaps," in *Proc. IEEE MTT-S Int. Conf. Microw. Intell. Mobility*, 2018, pp. 1–4.

[11] R. Prophet, G. Li, C. Sturm, and M. Vossiek, "Semantic segmentation on automotive radar maps," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 756–763.

[12] J. Stojcheski, T. Nürnberg, M. Ulrich, T. Michalke, C. Gläser, and A. Geiger, "Self-supervised occupancy grid map completion for automated driving," in *Proc. IEEE Intell. Veh. Symp.*, 2023, pp. 1–7.

[13] Y. Jin, M. Hoffmann, A. Deligiannis, J.-C. Fuentes-Michel, and M. Vossiek, "Semantic segmentation-based occupancy grid map learning with automotive radar raw data," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 216–230, Jan. 2024, doi: 10.1109/TIV.2023.3322353.

[14] Y. Jin, R. Prophet, A. Deligiannis, J.-C. Fuentes-Michel, and M. Vossiek, "Point-cloud-based road course estimation on automotive radar data," in *Proc. IEEE Int. Conf. Microwaves, Antennas, Commun. Electron. Syst.*, 2021, pp. 29–34.

[15] Y. Jin, A. Deligiannis, J.-C. Fuentes-Michel, and M. Vossiek, "Cross-modal supervision-based multitask learning with automotive radar raw data," *IEEE Trans. Intell. Veh.*, vol. 8, no. 4, pp. 3012–3025, Apr. 2023.

[16] A. Kirillov et al., "Segment Anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[17] C. Adam, R. Schubert, N. Mattern, and G. Wanielik, "Probabilistic road estimation and lane association using radar detections," in *Proc. IEEE 14th Int. Conf. Inf. Fusion*, 2011, pp. 1–8.

[18] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5586–5609, Dec. 2022.

[19] D. Wang, Y. Yu, S. Li, W. Dong, J. Wang, and L. Qing, "MulCode: A multi-task learning approach for source code understanding," in *Proc. IEEE Int. Conf. Softw. Anal., Evol. Reeng.*, 2021, pp. 48–59.

[20] M. Neseem, A. Agiza, and S. Reda, "AdaMTL: Adaptive input-dependent inference for efficient multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 4729–4738.

[21] R. Kuga, A. Kanezaki, M. Samejima, Y. Sugano, and Y. Matsushita, "Multi-task learning using multi-modal encoder-decoder networks with shared skip connections," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 403–411.

[22] F. Sarholz, J. Mehnert, J. Klappstein, J. Dickmann, and B. Radig, "Evaluation of different approaches for road course estimation using imaging radar," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2011, pp. 4587–4592.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *2015 18th Int. Conf., Med. Image Comput. Comput.-Assist. Interv.–MICCAI Munich, Germany, 2015, Proc., Part III 18*, 2015, pp. 234–241.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.