Camera-radar 3d object detection

Olga Matykina[1], Dmitry Yudin[1,2]

Abstract—

I. INTRODUCTION

In 3D object detection methods, image features are commonly projected into 3D space with limited depth information, leading to the ill depth problem. Many works that fuse data from cameras and radars [1], [2] do not exploit the valuable spatial information provided by radar points to improve the view transformation from the camera space to the BEV (Bird's Eye View) for image features. Even the few methods that leverage radar for this purpose, such as CRN [3], typically train a shallow neural network for depth prediction.

At the same time, there is a whole subfield in computer vision dedicated to depth estimation, with numerous recent works focused on radar-camera fusion [4], [5], [6]. We believe it is crucial to adopt the latest advances from the depth estimation community to enhance 3D object detection methods.

II. RELATED WORK

A. Depth Estimation

Recent methods for depth estimation using radar and monocular images focus on improving sparse radar data integration. One approach [7] combines radar and RGB images in a two-stage fusion, using coarse depth predictions to filter out noisy radar points and refine depth estimates. Another method [8] extends radar points' height range and filters noisy data, enhancing depth estimation under challenging conditions like rain or nighttime. A more recent approach [9] expands radar depth using a joint bilateral filter, adjusting the window size based on range and leveraging camera images for guidance, which increases radar points and improves depth accuracy.

B. Camera-Radar Fusion

A significant number of recent approaches in radar-camera fusion leverage Bird's Eye View (BEV) representations to enhance 3D object detection. BEV-based methods offer the advantage of transforming multi-modal data into a unified spatial representation, facilitating tasks like detection, segmentation, and tracking. For instance, RCBEVDet [2] encodes radar point clouds into BEV features using RadarBEVNet and combines them with image features via a Cross-attention Multi-layer Fusion module, achieving both robustness and real-time performance. Similarly, CRN [3] transforms perspective image features into BEV with sparse radar points and employs multi-modal deformable attention to address spatial misalignment. Echoes Beyond Points [10] utilizes PolarFormer as a baseline to align raw radar data with images in BEV space, highlighting its adaptability to radar data structure. Finally, RCTrans [1] densifies radar features using a Radar Dense Encoder, connecting multi-scale features in BEV through a U-Net-inspired architecture.

Beyond BEV-based methods, other approaches focus on addressing specific challenges in radar-camera fusion. MVFusion [11] introduces the Semantic-aligned Radar Encoder (SARE) to align radar features with images and employs transformers for robust radar representations. Similarly, SpaRC [12] combines sparse frustum fusion, range-adaptive radar aggregation, and local self-attention to improve alignment and localization precision. Unleashing HyDRa [13] employs a Height Association Transformer for better depth estimation by incorporating radar features directly in perspective view.

III. METHOD

An overview of our proposed approach is shown in Figure 1. We adopt the RCTrans architecture as our baseline and build upon it to better exploit multi-modal information from both cameras and radar. Our method retains the transformer-based detection head of RCTrans but introduces significant modifications to the image feature extraction and view transformation components. In particular, we enhance the image backbone by incorporating DINOv2, a modern vision transformer with strong representation learning capabilities. Furthermore, we propose a dual-branch architecture where image features are processed in both the original camera-view and the BEV (bird's-eye view) space. To enable this, we integrate radar signals to guide a more accurate view transformation from camera to BEV. This radar-assisted transformation aims to improve spatial alignment and depth reasoning in the BEV representation, which is critical for robust 3D object detection. The following subsections describe the components of our approach in detail.

A. Image Backbone

We utilize DINOv2 as the image feature extractor, motivated by its superior performance in visual representation learning compared to traditional ResNet-based models [14]. Features extracted in the camera view are converted into tokens and fed into a transformer. In

[1]Olga Matykina and Dmitry Yudin are with Moscow Institute of Physics and Technology (MIPT) {matykina.ov,yudin.da}@phystech.edu

[2] Dmitry Yudin is also with Artificial Intelligence Research Institute (AIRI) {yudin.da}@airi.net

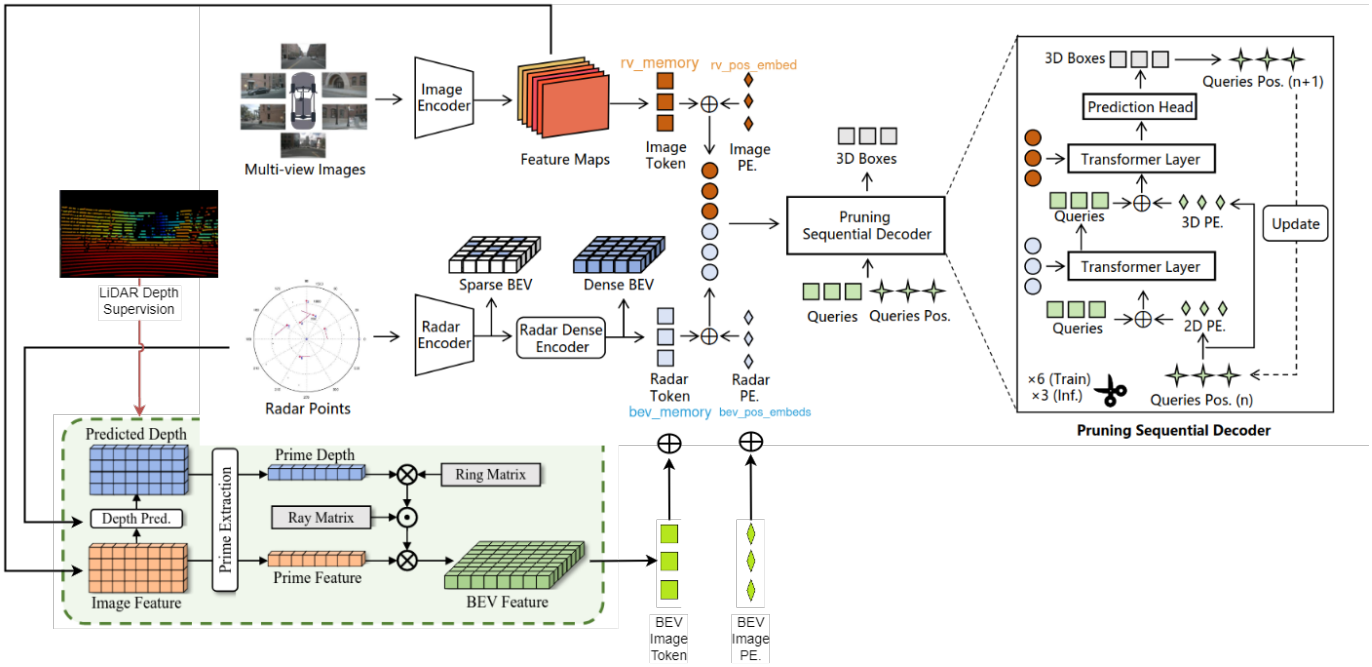

Figure 1. An overview of our proposed architecture based on RCTrans. The model consists of two main modules: (1) a standard RCTrans pipeline with DINOv2 as the image backbone and transformer-based token processing, and (2) a novel Radar Depth View Transformation module that uses radar-guided depth estimation for accurate image feature projection into the BEV space. The BEV and RV features from both branches are fused and passed to the transformer head for final 3D object detection.

parallel, we introduce a second branch where image features are transformed into the BEV space using radar information. We refer to this module as Radar Depth View Transformation.

B. Radar Depth View Transformation

Accurate feature transformation from the camera view to BEV space is critical for precise 3D detection. To achieve this, we utilize radar data, which provides accurate spatial coordinates and is robust to adverse weather and lighting conditions.

In view transformation approaches such as LSS [15], it is essential to estimate depth for every pixel in the feature map. We adopt the Sparse-beats-Dense [6] method, a state-of-the-art approach in depth estimation, to generate the depth map. This depth map is then used in the MatrixVT [16] module, which enables computationally efficient and accurate view transformation from the camera to the BEV perspective. The resulting BEV features are fused with radar features, transformed into tokens, and fed into the transformer head.

IV. EXPERIMENTS

A. Dataset and Metrics

Following previous approaches [3], [2], [1], we conduct experiments on the large-scale autonomous driving dataset nuScenes [17], designed for 3D object detection using radar and camera sensors. The dataset includes 1000 driving scenes, split into 700/150/150 for training, validation, and testing. Each frame contains 6 camera images and 5 radar sweeps covering a 360° field of view. There are approximately 1.4 million annotated 3D bounding boxes across ten object categories.

We use the official evaluation metrics: mean Average Precision (mAP) and nuScenes Detection Score (NDS). The NDS is a weighted sum of mAP and other metrics including Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE).

Our initial experiments are conducted on the nuScenes-mini dataset. This subset contains 10 scenes, with 323 frames in the training split and 81 frames for validation.

B. Implementation Details

We implement our model based on the RCTrans framework. Training is conducted for 90 epochs with a batch size of 16 on NVIDIA RTX 4060 and RTX 2080 Ti GPUs. We use the AdamW optimizer with a weight decay of 10^{-2} and a cyclic learning rate policy with an initial value of 4×10^{-4}.

C. Performance Comparison

We first compare our modified models against the RCTrans baseline on the nuScenes-mini validation set. As shown in Table I, the best results among models with randomly initialized weights are achieved by our model that uses image BEV features enhanced with SBD-based radar-camera depth fusion. This configuration outperforms others in most metrics, including mAP and NDS.

Table I
Performance comparison on the nuScenes-mini validation set. The best values in each group are underlined.

Model	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
ResNet (init weights)	0.0011	0.0236	1.0296	0.9784	1.0777	0.9103	0.8810
DINO (init weights)	0.0000	0.0055	1.0312	0.9586	1.0478	1.0041	0.9864
DINOv2 + Adapter (init weights)	0.0016	0.0322	1.0455	0.9443	1.0948	0.8897	0.8521
Image's BEV feats + SBD (init weights)	0.0038	0.0435	1.0395	0.9069	1.1144	0.8449	0.8318
DINO (from res18.pth)	0.1546	0.2867	0.8591	0.4945	0.7550	0.4966	0.3011
DINOv2 + Adapter (from res18.pth)	0.0205	0.1554	1.0800	0.5976	1.1369	0.5363	0.4146

When initialized from res18.pth (ResNet-18 pretrained on the full nuScenes dataset), the best overall performance in terms of mAP and NDS is obtained by the model using DINO as the image backbone.

V. CONCLUSION AND FUTURE WORK

References

[1] Y. Li, Y. Yang, and Z. Lei, "Rctrans: Radar-camera transformer via radar densifier and sequential decoder for 3d object detection," arXiv preprint arXiv:2412.12799, 2024.

[2] Z. Lin, Z. Liu, Z. Xia, X. Wang, Y. Wang, S. Qi, Y. Dong, N. Dong, L. Zhang, and C. 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, 2024, pp. 14 928–14 937.

[3] Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi, and D. Kum, "Crn: Camera radar net for accurate, robust, efficient 3d perception," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 17 615–17 626.

[4] A. D. Singh, Y. Ba, A. Sarker, H. Zhang, A. Kadambi, S. Soatto, M. Srivastava, and A. Wong, "Depth estimation from camera image and mmwave radar point cloud," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 9275–9285.

[5] H. Li, Y. Ma, Y. Gu, K. Hu, Y. Liu, and X. Zuo, "Radarcam-depth: Radar-camera fusion for depth estimation with learned metric scale," arXiv preprint arXiv:2401.04325, 2024.

[6] H. Li, M. Jing, W. Jin, S. Dong, J. Liang, H. Fan, and R. Ji, "Sparse beats dense: Rethinking supervision in radar-camera depth completion," in European Conference on Computer Vision. Springer, 2024, pp. 127–143.

[7] J.-T. Lin, D. Dai, and L. Van Gool, "Depth estimation from monocular images and sparse radar data," in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 10 233–10 240.

[8] C.-C. Lo and P. Vandewalle, "Depth estimation from monocular images and sparse radar using deep ordinal regression network," in 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021, pp. 3343–3347.

[9] C.-C. Lo and P. Vandewalle, "Expanding sparse radar depth based on joint bilateral filter for radar-guided monocular depth estimation," Sensors, vol. 24, no. 6, p. 1864, 2024.

[10] Y. Liu, F. Wang, N. Wang, and Z. Zhang, "Echoes Beyond Points: Unleashing the Power of Raw Radar Data in Multi-modality Fusion," in NeurIPS, 2023.

[11] Z. Wu, G. Chen, Y. Gan, L. Wang, and J. Pu, "Mvfusion: Multi-view 3d object detection with semantic-aligned radar and camera fusion," in 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 2766–2773.

[12] P. Wolters, J. Gilg, T. Teepe, F. Herzog, F. Fent, and G. Rigoll, "Sparc: Sparse radar-camera fusion for 3d object detection," arXiv preprint arXiv:2411.19860, 2024.

[13] P. Wolters, J. Gilg, T. Teepe, F. Herzog, A. Laouichi, M. Hofmann, and G. Rigoll, "Unleashing hydra: Hybrid fusion, depth consistency and radar for unified 3d perception," arXiv preprint arXiv:2403.07746, 2024.

[14] J. Schramm, N. Vödisch, K. Petek, B. R. Kiran, S. Yogamani, W. Burgard, and A. Valada, "Bevcar: Camera-radar fusion for bev map and object segmentation," in 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024, pp. 1435–1442.

[15] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. Springer, 2020, pp. 194–210.

[16] H. Zhou, Z. Ge, Z. Li, and X. Zhang, "Matrixvt: Efficient multi-camera to bev transformation for 3d perception," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8548–8557.

[17] 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," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11 621–11 631.