Cross-camera Monocular 3D Detection for Autonomous Racing

Anonymous ICCV submission

Paper ID *****

1. Introduction

Autonomous racing represents one of the most challenging and advanced testbeds for autonomous driving technologies. In this context, vehicles must perceive their environment and make decisions at high speeds and under demanding conditions where every millisecond counts.

A robust perception stack is critical, and relying on multiple sensors - such as LiDAR, radar, and cameras - ensures redundancy and accuracy. However, sensor failures can occur due to harsh conditions, hardware faults, or unexpected interferences. Having a reliable monocular 3D detection network acts as a safety net in these scenarios. Even if key sensors like LiDAR fail or degrade, a monocular camera-based system can continue to provide essential 3D environmental understanding while it receives the RGB input. This additional layer of perception enhances system resilience, maintains situational awareness, and ensures safe and continuous operation during autonomous racing tasks. However, 3D autonomous detection for Autonomous Racing (AR) presents some differences with the 3D detection task for city environments:

- range of distance (larger for AR);
- number of classes (only one class in AR);
- occlusions and number of objects in the scene (fewer occlusions and one object per scene in AR);
- cross-camera generalisation abilities required (one model able to generalise for frontal left, frontal right and frontal central cameras in AR).

In this extended abstract, our purpose is to propose a methodology for monocular 3D detection in the Autonomous Racing scenario. Specifically, the contributions involve the following:

- A dataset for 3D detection in the Autonomous Racing scenario;
- A methodology based on MonoDETR [16] with the addition of virtual depth and dimensions to face a cross-camera scenario;
- Quantitative and qualitative experiments.

2. Dataset

Monocular 3D object detection has gained significant traction in recent years, driven by the need for scalable and cost-effective 3D perception systems. Several benchmark datasets have played a crucial role in advancing this field, for example, KITTI [5], which remains one of the most widely used. Other important datasets include nuscenes [3] and waymo [14]. However, the 'Car' category relevant to autonomous racing differs significantly from that of typical city-driving vehicles. While datasets such as BETTY [12] and Racecar [8] exist for autonomous racing, they lack 3D object annotations. Consequently, models trained on datasets like KITTI fail to generalize effectively to the autonomous racing (AR) domain. This highlights a critical gap: the absence of AR-specific datasets for 3D object detection using cameras, underscoring the need to develop a dedicated dataset tailored to this unique setting. Our dataset creation procedure starts from 7 videos acquired during different moments and in different locations. Labels are obtained from a Lidar-based 3D detector named PointPillar [10]. PointPillar is a deep neural network able to predict 3D detection starting from LIDAR input. According to [13], lidar-based 3D detector significantly outperform monocular RGB ones. Therefore, it is reasonable to consider the PointPillar predictions as ground truth. PointPillar was trained on over 10,000 samples from a custom, manually labeled dataset. The point cloud input is formed by merging data from three LiDAR sensors. The resulting 3D bounding boxes are reprojected onto the RGB camera views using intrinsic and extrinsic parameters obtained through an offline calibration process. Each video serves as a dataset for 3D detection in KITTI format, including images, labels, and calibration matrices. The dataset in deep learning is a fundamental block of the entire procedure, since it should represent a balanced and correct distribution of the scenario. For this reason, the first step has been dedicated to the data analysis of the available videos and their labels, in order to create a new significant and representative dataset. The final dataset for training contains:

- 4490 samples:

- inputs taken from three different frontal cameras;
- a depth range between 3.8 and 135.0 meters, depth mean equal to 44.7808 m.

The decision to exclude an entire video from the training set is made to prevent potential bias and overfitting. Using the same video for both training and testing could increase the risk of encountering nearly identical frames, such as a test image captured just milliseconds after one seen during training, which would unfairly boost performance and undermine the validity of the results. The test datasets are separated per camera: one acquired by the frontal central and the other by the frontal left. The range of depth is the same.

3. Method

Monocular 3D object detection methods can be broadly categorised into prior-guided, camera-only, and depth-assisted approaches. Prior-guided methods [1], [17], incorporate shape, geometry, segmentation, or temporal priors to compensate for the ill-posed nature of 3D perception from a single image, often using auxiliary tasks or pretrained modules to enhance spatial understanding and detection robustness. Camera-only methods [16], [15], [11], directly regress 3D bounding boxes from RGB images using neural networks in an end-to-end fashion, drawing inspiration from 2D detectors to learn spatial dimensions and poses without relying on external cues. In contrast, depth-assisted methods [7] utilise pretrained monocular depth estimators [6] to convert images into depth maps or pseudo-LiDAR representations [4], enabling richer geometric reasoning but often facing performance gaps due to depth estimation errors. In this case, we excluded prior-guided methods as their reliance on predefined knowledge could impose constraints and limit adaptability.

We focused our investigation on the most promising methods from both camera-only and depth-assisted categories: respectively, MonoDETR [16] and DEVIANT[9]. While DEVIANT was ultimately excluded from our final approach due to insufficient performance results in racing scenarios (see Section 4), we identified and implemented a crucial adaptation technique derived from [2], called *virtual depth*, that significantly enhances the method's generalization capability across cameras with varying intrinsic parameters. This modification proved essential for our multi-camera racing setup, whereas it was unnecessary for the original MonoDETR implementation since the KITTI dataset on which it was trained utilised a single camera configuration with fixed parameters. The addition of virtual depth alone proved insufficient for our racing application, as MonoDETR's architecture also requires accurate prediction of 3D bounding box dimensions. Consequently, we extended the virtual depth concept to encompass dimension prediction, transforming both depth and size estimations into a unified 'virtual' coordinate system that remains consistent across multiple camera views with different intrinsic parameters. This comprehensive virtualisation approach enables the model to maintain consistent 3D object representations regardless of the camera's position or calibration parameters The effectiveness of this methodology is clearly demonstrated in the experimental results presented in Table 1, which shows improvements in detection accuracy and cross-camera consistency compared to baseline implementations.

4. Experiments

We conducted several experiments to prove the choice of MonoVDETR as the most promising method. Table 1 shows that MonoVDETR obtains better metrics than the original MonoDETR and DEVIANT. The table, for a better interpretation, shows not only 2D-AP, 3D-AP and BEV-AP, typically used in 3D detection. It also includes metrics like the rotation error, the mean depth error and the median depth error (both expressed in meters and percentage with respect to the ground truth distance). Specifically, the errors are computed on correctly predicted 2D bounding boxes only (with a IoU threshold of 0.70). These metrics help us interpret the final results. We observed that MonoVDETR demonstrates greater robustness, as it produces fewer outlier predictions compared to baseline methods. We attribute this improvement to the use of virtual depth and virtual dimensions, which help the model generalize more effectively across varying camera intrinsics. In our evaluation, FC refers to performance on a test set captured from a frontal central camera, while FR indicates performance on data from a frontal right camera. These settings allow us to assess the model's ability to generalize across different viewpoints and camera configurations.

Additionally, qualitative results in Figure 1 confirmed our numerical observations.

5. Conclusions

In this work, we explored the adaptation of monocular 3D object detection to the autonomous racing domain, which presents significant differences from conventional urban driving scenarios. We developed a model capable of generalising across different camera perspectives, demonstrating robustness in cross-camera settings. Additionally, we introduced a well-distributed and representative dataset tailored to monocular 3D detection in autonomous racing. Our results highlight the feasibility and potential of applying monocular 3D detection to autonomous racing, showing that with domain-aware design and data preparation, models can achieve reliable spatial understanding even in this high-speed context. Looking forward, we aim to further enhance our system by optimising the model for deployment

Table 1. Trained on Autonomous Racing Dataset (Threshold 0.5)

Method	2D-AP	BEV	3D-AP	Deg ROT	D. Mean (m)	D. Mean (%)	D. Med (m)	D. Med (%)
MonoDETR	FC 90.43 FR 90.54	FC 49.46 FR 53.85	FC 40.88 FR 43.40	FC 3.34 FR 3.06	FC 4.175 FR 1.32	FC 10.58 FR 2.42	FC 1.035 FR 0.86	FC 2.67 FR 1.97
MonoVDETR	FC **99.84** FR **90.75**	FC **56.67** FR **55.78**	FC **46.93** FR **44.31**	FC 3.36 FR 2.79	FC 1.65 FR 1.31	FC FR 3.10 2.34	FC 0.82 FR 0.78	FC 2.08 FR 1.71
DEVIANT	FC 75.67	FC 25.51	FC 11.68	FC 3.22	FC 1.32	FC 2.7	FC 0.76	FC 1.93

on efficient embedded hardware and reducing inference latency to meet the real-time requirements of onboard racing applications. These improvements will bring monocular 3D detection closer to practical use in competitive autonomous driving environments.

References

[1] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9287–9296, 2019. 2

[2] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13154–13164, 2023. 2

[3] 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*, pages 11621–11631, 2020. 1

[4] Xiaomeng Chu, Jiajun Deng, Yao Li, Zhenxun Yuan, Yanyong Zhang, Jianmin Ji, and Yu Zhang. Neighbor-vote: Improving monocular 3d object detection through neighbor distance voting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5239–5247, 2021. 2

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 1

[6] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 2

[7] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022. 2

[8] Amar Kulkarni, John Chrosniak, Emory Ducote, Florian Sauerbeck, Andrew Saba, Utkarsh Chirimar, John Link, Madhur Behl, and Marcello Cellina. Racecar-the dataset for high-speed autonomous racing. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11458–11463. IEEE, 2023. 1

[9] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. Deviant: Depth equivariant network for monocular 3d object detection. In *European Conference on Computer Vision*, pages 664–683. Springer, 2022. 2

[10] 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*, pages 12697–12705, 2019. 1

[11] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 996–997, 2020. 2

[12] Micah Nye, Ayoub Raji, Andrew Saba, Eidan Erlich, Robert Exley, Aragya Goyal, Alexander Matros, Ritesh Misra, Matthew Sivaprakasam, Marko Bertogna, Deva Ramanan, and Sebastian Scherer. Betty dataset: A multi-modal dataset for full-stack autonomy, 2025. 1

[13] Ziying Song, Lin Liu, Feiyang Jia, Yadan Luo, Caiyan Jia, Guoxin Zhang, Lei Yang, and Li Wang. Robustness-aware 3d object detection in autonomous driving: A review and outlook. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 1

[14] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1

[15] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 2

[16] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In

(a) MonoDETR (b) MonoVDETR

(c) MonoDETR (d) MonoVDETR

(e) MonoDETR (f) MonoVDETR

Figure 1. Examples of predictions (red) and ground truth (green). Both BEV visualization, 2D and 3D bounding box projection are visible.

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9155–9166, 2023. 1, 2

[17] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021. 2