Monocular 3D Object Detection in Roadside Settings: Extending Data and Enhancing Models

Sondos Mohamed

Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale, 72, 09124 ,Cagliari, Italy

Abstract

Understanding three-dimensional objects is crucial in domains like urban autonomous driving, roadside monitoring, augmented and virtual reality. Traditionally, this task required expensive LiDAR sensors and stereo RGB imaging due to the limitations of monocular image-only methods, which could not count on depth information. Recent advances in monocular models based on deep learning have improved this situation, yet real-world challenges persist. For instance, variations in camera properties and object complexity constrain existing monocular 3D object detection. In my PhD research, I focus on monocular 3D object detection from images collected by roadside cameras. Firstly, my objective is to curate diverse datasets that encompass a wide array of scenarios and camera configurations. Secondly, I strive to train and assess detection models, surmounting existing limitations. Thirdly, my goal is to refine these models, fostering adaptability and robustness, thereby empowering them to generalize across diverse scenes and scenarios. This work advances monocular 3D object detection in domains like roadside monitoring.

Keywords

Object Detection, Monocular Models, Roadside Cameras.

1. Introduction

Object detection methods in single images can take two forms: 2D methods [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12] or 3D methods [13, 14, 15, 16, 17, 18, 19]. The application of 3D object detection approaches offers advantages. By providing a better understanding of the scene and enabling the detection of occluded objects, they can enhance the accuracy and reliability of object detection in complex environments. Moreover, they are better suited to describe object pose and shape.

However, the lack of depth information in 2D images makes it challenging to precisely estimate the size and location of objects. 3D object detection has applications in both indoor and outdoor contexts. In outdoor scenarios, recent advancements in autonomous driving have shown promising results [14, 15, 16, 17]. Furthermore, the adoption of an increasing number of datasets [20, 21, 22, 23, 24, 25, 26, 27] has further improved the effectiveness of this technology.

Nevertheless, upon closer examination of outdoor datasets, it becomes evident that most of them are tailored for autonomous driving, with only a limited number focusing on roadside scenarios [28, 29, 30, 31, 32]. A more in-depth analysis of existing monocular models reveals that the majority are trained and tested using a single dataset. Recent efforts have attempted to address these limitations. For instance, a recent work [19] has made significant progress by integrating indoor and outdoor datasets into a large, standardized dataset and training models to

AIxIA 2023 22nd International Conference of the Italian Association for Artificial Intelligence,Rome, Italy

sondoswa.mohamed@unica.it (S. Mohamed)

cover this diversity. However, it is important to note that this work does not encompass roadside datasets, and their pretrained models face challenges when tested on roadside images. Most of the roadside cameras are originally set up as CCTV cameras, and many have been in place on the road for decades. However, some of them lack crucial information. such as focal length, camera coordinate systems, and other configuration details. With this in mind, this PhD thesis aims to generate a diverse dataset that includes different scenes, focal lengths, and resolutions. Subsequently, we aim to develop models that can accommodate these diverse requirements. Our ultimate goal is to determine whether our model can effectively demonstrate its generalization during the testing phase by achieving satisfactory performance on previously unseen scenes and datasets without retraining. This endeavor underscores the need for comprehensive, versatile datasets and models in the field of 3D object detection, especially for real-world scenarios.

2. Research Plan

This research plan outlines my three-year PhD trajectory aimed at advancing the field of monocular 3D object detection for roadside monitoring. It encompasses a systematic journey from building a foundational understanding of the subject to practical application and integration into real-world scenarios. The plan is organized along three distinct years, delineated as follows.

Year 1: Literature Review and Foundation. During the first year of my doctoral program, I established a solid foundation for my research by conducting an extensive literature review. This review encompassed the study of 2D and 3D object detection methodologies, as well as object tracking and person re-identification techniques, with a specific focus on their application in roadside monitoring using CCTV cameras. My primary objectives during this foundational year included gaining a deep understanding of the existing research landscape and familiarizing myself with fundamental concepts and advanced techniques in computer vision for CCTV camera-based monitoring. These efforts provided the groundwork for the subsequent steps.

Year 2: Data Generation and Model Reproduction. In the second year of my doctoral journey, I transitioned from theory to practical implementation. This phase involved two main objectives: data generation and model adaptation. I used simulators to create synthetic datasets, customized for roadside monitoring, allowing for initial model testing. Simultaneously, I worked on adapting and fine-tuning existing 3D object detection models, originally developed for autonomous driving, to suit the specific needs of roadside monitoring. Key achievements included the creation of synthetic datasets with controlled variations in parameters like resolution and focal lengths, as well as the adaptation of state-of-the-art monocular 3D object detection models to the generated dataset, aligning them with requirements of roadside monitoring.

Year 3: Real-World Data Creation, Model Improvement, and Integration. In the third and pivotal year of my research, the focus shifts to working with authentic real-world data. This phase involves several key activities, including the curation of datasets derived from roadside cameras, rigorous evaluations of previously developed models to assess their generalizability, and enhancements to improve model efficiency, accuracy, and robustness. The culmination of this year's efforts will be the seamless integration and thorough evaluation of these advanced models into an existing interface customized for use by local municipalities in Sardinia. The sig-

nificant achievements during this transformative year include the curation of authentic datasets representing real-world scenarios, extensive experimentation to test the models' adaptability, efficiency improvements based on insights from real-world data, and the successful integration and evaluation of the refined models in a practical operational setting. This structured and progressive research plan seeks to bridge the gap between theoretical knowledge and practical application in the domain of 3D object detection for roadside monitoring. It strives to make a comprehensive and impactful contribution to the field of computer vision, particularly within the context of enhancing roadside safety and security.

3. Current Results

The current result is the first version of "MonoRoadCam" [33], a synthetic dataset serving two purposes: facilitating the adaptation of 3D object detection methods for roadside cameras and evaluating existing methods from the autonomous driving domain. MonoRoadCam was created using the CARLA simulation environment [34], ensuring data closely resembling real-world scenarios and adhering to the KITTI format [20]. Our contributions is threefold:

- Synthetic Dataset Generation: MonoRoadCam is designed for monocular 3D object detection, featuring 7,481 development images and 7,518 test images, all annotated with object type, size, location, and orientation, with simulations of three weather conditions.
- Model Reproduction and Evaluation: Our research verifies the reproducibility of state-of-the-art monocular 3D object detection methods (M3DRPN [13], Kinematic [14], SMOKE [15], Monodle [16]) originally designed for autonomous driving in the roadside context.
- Comparative Study: We conducted an extensive comparative study between 3D object detection datasets captured by roadside and frontal cameras, revealing the potential and limitations of applying autonomous driving solutions directly without training to monocular roadside camera images.

Quantitatively, when tested on the original KITTI dataset [20], the reproduced models show a slight decrease in performance compared to the evaluation metric scores reported in their respective papers, with SMOKE exhibiting a notable drop. Comparing results across datasets, the performance measured on our synthetic dataset was higher than the performance measured on the original KITTI dataset. Finally, qualitatively, our evaluation shows the potential of MonoRoadCam as a valuable resource for advancing monocular 3D object detection.

4. Open Challenges and Expected Benefits

Based on the open challenge, I aim to seek answers to the following three key questions:

- **How can I effectively bridge the gap between existing 3D object detection methods developed for autonomous driving and their applicability to the unique challenges of roadside monitoring, especially when working with monocular images from diverse cameras?** This question lies at the heart of my research. The transition from well-established autonomous driving scenarios to roadside monitoring

introduces a set of distinctive challenges. By addressing this question, I aim to develop solutions that not only function effectively but also integrate into real-world indoor and outdoor environments that require object detection from monocular cameras.

- **Which strategies can be employed to systematically diversify roadside datasets, by incorporating variations in parameters such as resolution, focal lengths, and full orientation for the object in outdoor scenarios (instead of yaw only)? How will these controlled variations impact the adaptability and performance of the current 3D object detection models?** The quality of data is paramount in training robust computer vision models. However, the dynamic nature of roadside scenarios demands datasets that are both diverse and reflective of real-world conditions. By exploring controlled variations in critical parameters such as resolution and focal lengths, I aim to not only enrich my data but also understand how these variations influence the adaptability and overall performance of my 3D object detection models.
- **In the context of my research, how can I assess the generalizability of models effectively, ensuring that they perform satisfactorily on previously unseen scenes and datasets without requiring extensive retraining, and what are the best practices for achieving this?** The ability of computer vision models to generalize across different scenarios is pivotal for their practical utility. Assessing and ensuring the generalizability of my models is a central aspect of my research objectives. I aim to develop models that not only excel in controlled environments but also demonstrate robustness and reliability when applied to unpredictable roadside situations.

Addressing these questions in my research aims to provide a foundational framework for other researchers striving to develop a zero-shot model that overcomes current challenges and limitations in 3D object detection for roadside monitoring in the next-generation smart cities.

Acknowledgements. The author extends gratitude to Prof. Salvatore Carta and Dr. Mirko Marras for their supervision. Special thanks to Marco Sau for his support in part of this project.

References

[1] S. Ren, K. He, R. B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: NIPS 2015, 2015, pp. 91–99.

[2] J. Cao, H. Cholakkal, R. M. Anwer, F. S. Khan, Y. Pang, L. Shao, D2det: Towards high quality object detection and instance segmentation, in: CVPR 2020, Computer Vision Foundation / IEEE, 2020, pp. 11482–11491.

[3] K. He, G. Gkioxari, P. Dollár, R. B. Girshick, Mask R-CNN, CoRR abs/1703.06870 (2017). arXiv:1703.06870.

[4] Y. Li, Y. Chen, N. Wang, Z. Zhang, Scale-aware trident networks for object detection, in: ICCV 2019, IEEE, 2019, pp. 6053–6062.

[5] J. Redmon, S. K. Divvala, R. B. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: CVPR 2016, IEEE, 2016, pp. 779–788.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, A. C. Berg, SSD: single shot multibox detector, in: ECCV 2016, volume 9905, Springer, 2016, pp. 21–37.

[7] H. Law, J. Deng, Cornernet: Detecting objects as paired keypoints, in: ECCV 2018, volume 11218, Springer, 2018, pp. 765–781.
[8] X. Zhou, D. Wang, P. Krähenbühl, Objects as points, CoRR abs/1904.07850 (2019). arXiv:1904.07850.
[9] T. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: ICCV 2017, IEEE, 2017, pp. 2999–3007.
[10] A. Atzori, G. Fenu, M. Marras, Explaining bias in deep face recognition via image characteristics, in: IJCB 2022, IEEE, 2022, pp. 1–10.
[11] A. Atzori, G. Fenu, M. Marras, Demographic bias in low-resolution deep face recognition in the wild, IEEE J. Sel. Top. Signal Process. 17 (2023) 599–611.
[12] G. Fenu, H. Lafhouli, M. Marras, Exploring algorithmic fairness in deep speaker verification, in: ICCSA 2020, volume 12252, 2020, pp. 77–93.
[13] G. Brazil, X. Liu, M3D-RPN: monocular 3d region proposal network for object detection, in: ICCV 2019, IEEE, 2019, pp. 9286–9295.
[14] G. Brazil, G. Pons-Moll, X. Liu, B. Schiele, Kinematic 3d object detection in monocular video, in: ECCV 2020, volume 12368 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 135–152.
[15] Z. Liu, Z. Wu, R. Tóth, SMOKE: single-stage monocular 3d object detection via keypoint estimation, in: CVPR 2020, IEEE, 2020, pp. 4289–4298.
[16] X. Ma, Y. Zhang, D. Xu, D. Zhou, S. Yi, H. Li, W. Ouyang, Delving into localization errors for monocular 3d object detection, in: CVPR 2021, IEEE, 2021, pp. 4721–4730.
[17] T. Wang, X. Zhu, J. Pang, D. Lin, FCOS3D: fully convolutional one-stage monocular 3d object detection, in: ICCV 2021, IEEE, 2021, pp. 913–922.
[18] T. Wang, X. Zhu, J. Pang, D. Lin, Probabilistic and geometric depth: Detecting objects in perspective, in: Conference on Robot Learning, volume 164, PMLR, 2021, pp. 1475–1485.
[19] G. Brazil, A. Kumar, J. Straub, N. Ravi, J. Johnson, G. Gkioxari, Omni3d: A large benchmark and model for 3d object detection in the wild, CoRR abs/2207.10660 (2022). arXiv:2207.10660.
[20] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the KITTI vision benchmark suite, in: ICPR 2012, IEEE, 2012, pp. 3354–3361.
[21] A. Patil, S. Malla, H. Gang, Y. Chen, The H3D dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes, in: ICRA 2019, IEEE, 2019, pp. 9552–9557.
[22] R. S. P. H. Z. C. H. P. Y. C. A. M. V. C. J. L. Quang-Hieu Pham, Pierre Sevestre, A*3d dataset: Towards autonomous driving in challenging environments, in: ICRA 2020, 2020.
[23] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, nuscenes: A multimodal dataset for autonomous driving, CoRR abs/1903.11027 (2019). arXiv:1903.11027.
[24] M. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, J. Hays, Argoverse: 3d tracking and forecasting with rich maps, in: CVPR 2019, Computer Vision Foundation / IEEE, 2019, pp. 8748–8757.
[25] J. Mao, M. Niu, C. Jiang, H. Liang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu, C. Xu, H. Xu, One million scenes for autonomous driving: ONCE dataset, in: NIPS 2021, 2021.
[26] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, R. Yang, The apolloscape

dataset for autonomous driving, in: CVPR 2018, IEEE, 2018, pp. 954–960.
[27] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, D. Anguelov, Scalability in perception for autonomous driving: Waymo open dataset, CoRR abs/1912.04838 (2019). arXiv:1912.04838.
[28] E. Strigel, D. A. Meissner, F. Seeliger, B. Wilking, K. Dietmayer, The ko-per intersection laserscanner and video dataset, in: ITSC 2014, 2014, pp. 1900–1901.
[29] X. Ye, M. Shu, H. Li, Y. Shi, Y. Li, G. Wang, X. Tan, E. Ding, Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task, in: CVPR 2022, IEEE, 2022, pp. 21309–21318.
[30] Y. Deng, D. Wang, G. Cao, B. Ma, X. Guan, Y. Wang, J. Liu, Y. Fang, J. Li, BAAI-VANJEE roadside dataset: Towards the connected automated vehicle highway technologies in challenging environments of china, CoRR abs/2105.14370 (2021). arXiv:2105.14370.
[31] J. Sochor, J. Špaňhel, A. Herout, Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance, IEEE Transactions on Intelligent Transportation Systems PP (2018) 1–12.
[32] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, Z. Nie, DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection, CoRR abs/2204.05575 (2022). arXiv:2204.05575.
[33] S. Barra, M. Marras, S. Mohamed, A. S. Podda, R. Saia, Can existing 3d monocular object detection methods work in roadside contexts? A reproducibility study, in: AIxIA 2023, volume 14318, Springer, 2023, pp. 321–335.
[34] A. Dosovitskiy, G. Ros, F. Codevilla, A. M. López, V. Koltun, CARLA: an open urban driving simulator, in: CoRL 2017, volume 78, PMLR, 2017, pp. 1–16. URL: http://proceedings.mlr.press/v78/dosovitskiy17a.html.