Mono3DBox: Monocular Vision for 3D bounding box detection with direct vertex regression and Convex-Hull based IOU

Motivation:

The advancement of Autonomous Driving (AD) technologies is critically dependent on the ability to accurately perceive and interpret the 3D environment in real-time. Mono3DBox leverages the simplicity and ubiquity of monocular cameras to address this challenge by extending the YOLO architecture for direct regression of 2D projections of 3D bounding box vertices (which can be extended to get 6D pose estimates using these 3D projections as control points) and integrating a convex hull-based IOU for enhanced loss function precision in the regression branch for *box_loss*. This approach is particularly important and relevant for several reasons:

- **Reducing Cost and Complexity**: The reliance on expensive depth-sensing hardware, such as LiDAR and RADAR, significantly increases the cost and complexity of AD systems, making the commercial viability of autonomous vehicles more challenging. By improving the 3D object detection capabilities of monocular images, Mono3DBox proposes a more cost-effective solution that could accelerate the deployment and adoption of AD technologies.
- **Democratizing Research and Development**: The high cost and limited availability of advanced sensing technologies like LiDAR have been barriers to entry for researchers and developers in resource-constrained environments. Mono3DBox's reliance on monocular imagery democratizes access to 3D detection technology, enabling a broader base of innovation and experimentation within the AD field.
- **Enabling Real-time Performance**: The ability to process and interpret environmental data in real-time is paramount for the safety and efficiency of autonomous systems. Mono3DBox's architectural innovations ensure real-time performance without sacrificing accuracy, making it an invaluable tool for AD applications where decision-making speed is critical.
- **Spatial Understanding from 2D Data**: Accurate 3D object localization using monocular images addresses one of the significant challenges in computer vision—extracting depth information from 2D data. This improved spatial understanding is crucial for navigation, obstacle avoidance, and complex interaction tasks in autonomous systems.

Prior Work / Literature Review:

Monocular 3D object detection has been profoundly shaped by a blend of foundational research and cutting-edge advancements, particularly with the goal of enhancing autonomous driving technologies. Early on, Redmon et al. [1] laid the groundwork for real-time object detection frameworks with YOLOv8, spotlighting the importance of detection accuracy. Building on this, Mousavian et al. [2] made pivotal contributions by utilizing geometric constraints for 3D bounding box estimation from monocular images, thus bridging the gap between 2D image features and 3D object properties. The field has seen a further surge of innovation in recent years. For instance, Zhou et al. [3] introduced MonoPair, exploiting spatial relationships to infer depth, while Reading et al. [4] developed CaDDN, a framework that directly estimates depth distribution from a single image to facilitate 3D bounding box predictions. Simonelli et al. [5] tackled the challenge of domain adaptation, enhancing the adaptability of monocular 3D object detection across varying scenes. Furthermore, recent contributions like MonoDETR [6] and MonoPGC [7] have showcased the potential of depth-guided transformers and cross-attention mechanisms to refine the precision and efficiency of detection, marking a significant evolution towards understanding depth from single images and improving the accessibility and effectiveness of 3D object detection for real-world applications.

Submitted by: Yatharth Ahuja
Team Members: Cherry Bhatt, Shahram Najam Syed, Zihan Wan, Yatharth Ahuja

Proposed Setup:

The core idea is to adapt the YOLOv8 architecture to enhance monocular 3D object detection by focusing on regressing the vertices of the 2D projection of the 3D bounding box. This modification aims to accurately capture the spatial dimensions and orientation of objects in 3D space from single-image inputs. To improve the precision of bounding box predictions, we propose utilizing a convex hull approach to compute the Intersection over Union (IoU), which will replace the standard CIoU loss used in YOLOv8. Integrating the regression of 2D projections of 3D bounding box vertices directly into the YOLOv8 framework intuitively addresses the depth and spatial ambiguity challenges inherent in monocular images. By focusing on vertices, the model can more accurately infer the object's 3D structure and orientation, a critical aspect often lost in 2D object detection tasks. The convex hull-based IoU further enhances this by ensuring a more geometrically accurate measure of the overlap between predicted and actual bounding boxes, intuitively accounting for the complex shapes and orientations objects can have in the real world.

Mono3DBox represents an innovative synergy of advancements in monocular 3D object detection, refining the core principles established by the YOLOv3 architecture [1] and extending them to 3D spatial analysis. Unlike traditional methods that may rely on explicit depth estimation, Mono3DBox emphasizes the regression of the vertices of 2D projections of 3D bounding boxes. This approach, which draws inspiration from the geometric constraint-based 3D estimations of Mousavian et al. [2], allows for an intuitive understanding of object depth and orientation without direct depth measurement. Mono3DBox further embraces depth-informed strategies, as seen in Zhou et al.'s MonoPair [3] and Reading et al.'s CaDDN [4], to enhance object localization through the implicit interpretation of depth cues. The adoption of a convex hull-based Intersection over Union (IoU) calculation positions Mono3DBox alongside recent depth-guided models like MonoDETR [6] and MonoPGC [7], aiming for precise object boundary representation. Through this synthesis of geometric reasoning and indirect depth inference via custom loss functions, Mono3DBox aspires to redefine the standards for detecting 3D objects from single images, marking a significant evolution in the field.

Experimental Evaluations:

- **Benchmark Datasets**: Utilizing KITTI and nuScenes for diverse urban scenarios.
- **Evaluation Metrics**: Average Precision (AP) for 3D detections, convex hull-based IoU for bounding box accuracy, and average 3D distance of vertices (referred to as ADD metric) as in [17, 18, 19].
- **Ablation Studies**: Analyzing impacts of vertex regression, convex hull-based IoU, and loss function components.
- **Robustness Tests**: Evaluating performance under different environmental conditions and datasets.

Baselines:

We plan to evaluate Mono3DBox against baselines including MonoPair [3], CaDDN [4], MonoDETR [6], MonoPGC [7], RTM3D [8], and SMOKE [10] for its performance in monocular 3D object detection.

Submitted by: Yatharth Ahuja
Team Members: Cherry Bhatt, Shahram Najam Syed, Zihan Wan, Yatharth Ahuja

References:

[1] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." *arXiv preprint arXiv:1804.02767* (2018).
[2] Mousavian, Arsalan, et al. "3d bounding box estimation using deep learning and geometry." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2017.
[3] Chen, Yongjian, et al. "Monopair: Monocular 3d object detection using pairwise spatial relationships." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
[4] Reading, Cody, et al. "Categorical depth distribution network for monocular 3d object detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
[5] Simonelli, Andrea, et al. "Disentangling monocular 3d object detection: From single to multi-class recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*44.3 (2020): 1219-1231.
[6] Zhang, Renrui, et al. "MonoDETR: Depth-guided transformer for monocular 3D object detection." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
[7] Wu, Zizhang, et al. "Monopgc: Monocular 3d object detection with pixel geometry contexts." *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.
[8] Li, Peixuan, et al. "Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving." *European Conference on Computer Vision*. Cham: Springer International Publishing, 2020.
[9] Cai, Yingjie, et al. "Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 07. 2020.
[10] Liu, Zechen, Zizhang Wu, and Roland Tóth. "Smoke: Single-stage monocular 3d object detection via keypoint estimation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020.
[11] Zhang, Chenyangguang, et al. "Sst: Real-time end-to-end monocular 3d reconstruction via sparse spatial-temporal guidance." *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023.
[12] Ding, Mingyu, et al. "Learning depth-guided convolutions for monocular 3d object detection." *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops*. 2020.
[13] Park, Dennis, et al. "Is pseudo-lidar needed for monocular 3d object detection?." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
[14] Xu, Zhenbo, et al. "Zoomnet: Part-aware adaptive zooming neural network for 3d object detection." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 07. 2020.
[15] Lee, Hyo-Jun, et al. "BAAM: Monocular 3D pose and shape reconstruction with bi-contextual attention module and attention-guided modeling." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
[16] Wu, Di, et al. "6d-vnet: End-to-end 6-dof vehicle pose estimation from monocular rgb images." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019.
[17] E. Brachmann, F. Michel, A. Krull, M. Ying Yang, S. Gumhold, et al. Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image. In *CVPR*, 2016.
[18] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again. In *ICCV*, 2017.
[19] M.RadandV.Lepetit.BB8:AScalable,Accurate,Robust Partial Occlusion Method for Predicting the 3D Poses of Challenging Ob- jects without Using Depth. In *ICCV*, 2017.

Submitted by: Yatharth Ahuja
Team Members: Cherry Bhatt, Shahram Najam Syed, Zihan Wan, Yatharth Ahuja