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

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

Motivation:

The ubiquitous adoption of Autonomous Driving (AD) technologies necessitates advanced 3D environment perception capabilities, critically dependent on the precise, real-time interpretation of spatial data. Historically, this necessity has advocated for the integration of sophisticated depth-sensing technologies such as LiDAR and RADAR, which, despite their efficacy, introduce significant cost, complexity, and scalability challenges. These challenges represent a substantial barrier to the widespread adoption and further development of AD systems, particularly in resource-constrained research environments. The Mono3DBox approach is predicated on the hypothesis that monocular vision—owing to its ubiquity and cost-effectiveness—can serve as a viable alternative for 3D object detection by exploiting learning-based models and implicit geometric analysis. This work is motivated by the need to democratize AD technologies by minimizing reliance on expensive hardware, thus lowering the entry barrier for innovation in this field. Mono3DBox extends the YOLO architecture, renowned for its balance between accuracy and real-time performance in 2D object detection, to the domain of 3D by directly regressing the vertices of 2D projections of 3D bounding boxes from single-image inputs. This methodological innovation aims to resolve the depth and spatial loss of information inherent in 2D representations without the direct estimation of depth, a common bottleneck in computational efficiency and accuracy.

Further, the incorporation of a convex hull-based Intersection over Union (IoU) metric as a weighted loss function for bounding box accuracy represents a novel utility in the field. This integration addresses the geometric fidelity of 3D bounding boxes, providing a more accurate and robust measure of detection precision, especially in the context of varied and complex object shapes encountered in the road scene. Such an approach not only aligns with but also improves the computational efficiency and scalability of AD technologies through monocular imagery leveraging 3D object detection. The foundational motivation for Mono3DBox thus lies in the confluence of technological necessity and the visionary pursuit of accessibility in AD technologies. It embodies an integrative approach that synthesizes state-of-the-art machine learning techniques with geometric principles to transcend the limitations of current monocular 3D detection methodologies. By addressing the critical challenges of cost, complexity, and performance, Mono3DBox aims to catalyze a transformative shift in the AD landscape, making autonomous technologies more accessible, scalable, and efficient.

Related Work

3D object detection from monocular images has been an open-research problem partially because of the accuracy of spatial sensing LiDARs accuracy and majorly attributed to the complexity in inferring 3D geometry and spatial understanding through learning based approaches without the requirement of LiDAR even as a pre-training or training step. But in the recent years advancements have been seen for limited dataset benchmarks where monocular 3D bounding box estimation has outperformed fusion based method. These methodologies, each with its unique perspective on tackling the inherent challenges of monocular vision, aim to bridge the gap between the rich contextual information available in 2D imagery and the spatial depth required for accurate object localization and pose estimation. Early contributions by Mousavian et al. [2] broke new ground by regressing stable 3D object properties through deep convolutional neural networks and combining these estimates with 2D object bounding boxes. Their approach, leveraging a novel hybrid discrete-continuous loss, marked a significant departure from methods reliant on semantic segmentation, simplifying 3D pose recovery with high stability across various object types. However, while effective, this methodology's reliance on geometric constraints limits its adaptability in highly dynamic or cluttered environments. Chen et al.'s MonoPair [3] introduced an innovative technique for enhancing detection accuracy in occluded and cluttered scenarios by leveraging spatial relationships between objects. Similarly, Jiang et al.'s SKD-WM3D [20] utilized self-knowledge distillation to enhance 3D object localization with minimal computational overhead. These approaches underscore the potential of spatial cues in improving detection but often face challenges in densely occluded scenes where pairwise relationships or self-distilled features may not fully capture complex spatial dynamics. Significant advancements have been made in direct depth estimation and object dimension disentanglement, with works such as Reading et al.'s CaDDN [4] and Simonelli et al.'s [5] technique focusing on multi-class recognition. These methods demonstrate the efficacy of directly inferring depth distribution from imagery and separately optimizing object dimensions for domain adaptation. However, the accuracy of depth estimation from monocular images remains a challenge, limiting the overall effectiveness of these approaches in scenarios with variable illumination or at greater distances. Wu et al.'s MonoPGC [7] and Li et al.'s RTM3D [8] address the need for real-time detection capabilities by incorporating pixel geometry contexts and predicting 3D bounding boxes from perspective keypoints. These models offer compact, efficient solutions for real-time applications but struggle with the accuracy of keypoint estimation and the integration of pixel-level geometry information under rapid scene changes. Xu et al.'s ZoomNet [14] presents the use of stereo imagery to improve disparity estimation through adaptive zooming, offering enhanced detection of distant and occluded objects. While these stereo-based methods set new benchmarks, they inherently require more complex hardware setups and may not be directly applicable to monocular vision systems. Approaches by Ma et al. and Kehl et al. [18], focusing on multi-level feature fusion and the detection of 3D model instances, respectively, highlight the advantages of integrating features across network depths and utilizing synthetic model data for training. These strategies, aiming to improve depth and dimension estimation, face limitations in scenarios lacking precise depth cues or where synthetic and real-world data discrepancies impact detection accuracy. Wang et al.'s PVFusion [21] and Hai et al.'s VirConvNet [22] (top of the leaderboard for KITTI benchmark) pushes the boundaries of fusing LiDAR and image data to detect small or distant objects, addressing the sparsity of point cloud data. While innovative, the computational complexity and reliance on LiDAR data for fusion with image features may not be feasible for all applications, particularly those constrained by hardware capabilities or operational environments.

Proposed Setup:

In our proposed framework, inspired by the results in [7] and [8] for regressing 3D bounding boxes using 2D keypoints, we further push this by proposing a methodology for direct regression of 3D bounding box vertices from single images and the application of a Convex-Hull based Intersection over Union (IoU) as a novel loss metric. This methodology diverges from conventional approaches by circumventing intermediate depth estimation in favor of a direct geometric inference, aiming to mitigate the inherent ambiguities associated with monocular depth cues. The foundational principle of Mono3DBox is to leverage the geometric properties inherent in the 2D projection of 3D objects to directly infer the vertices of their bounding boxes. This direct regression approach can be facilitated by any learning-based model (in our case YOLOV8) that has been specifically tailored to implicitly interpret these geometric cues from monocular imagery. The model is trained to predict the coordinates of the 3D bounding box vertices in the image space, utilizing a network architecture with a focus on spatial reasoning and geometric understanding. To accurately evaluate the precision of the detected 3D bounding boxes, Mono3DBox employs a Convex-Hull based IoU metric. This metric offers a more geometrically faithful evaluation of detection accuracy by considering the convex shape formed by the predicted vertices against the ground truth. This approach provides a novel assessment of detection performance, particularly in complex environments where traditional IoU metrics may not fully capture the intricacies of 3D object detection accuracy. Intuitively, Mono3DBox addresses the critical challenge of depth ambiguity in monocular vision by directly mapping the apparent geometric properties of objects in 2D images to their 3D spatial configurations. This methodological pivot from depth estimation to direct geometric inference is grounded in the observation that the spatial arrangement and orientation of objects can be determined from their 2D projections through geometric modeling learned during the training process. This direct vertex regression approach sets apart Mono3DBox approach from the prevalent depth-estimation strategies, such as those employed by Reading et al. in CaDDN, circumventing the inherent ambiguities and computational complexities associated with inferring depth from monocular cues. Unlike methods reliant on geometric constraints or spatial relationships for bounding box estimation, like the works of Mousavian et al. and Chen et al.'s MonoPair, Mono3DBox's strategy does not predicate on the accuracy of 2D detection or specific scenarios of object occlusion, aiming instead for a more generalized and robust framework for 3D object localization. By circumventing intermediary depth inference in favor of a geometry-focused detection mechanism, Mono3DBox through direct regression of 2D projections of 3D bounding boxes along with integration of a Convex-Hull based IoU metric enhances both the efficiency and accuracy of monocular 3D detection, addressing the critical limitations of depth ambiguity and computational overhead that have constrained previous approaches.

Results:

In our project, we benchmarked the Mono3DBox framework against the widely recognized KITTI dataset, a cornerstone in the field of autonomous driving research. Given that the KITTI dataset's test set does not come with publicly available labels, we partitioned the 7481 images from the original training set into subsets: 5000 images for training, 1000 for validation, and the remaining 1481 for testing and benchmarking our models. This distribution ensured a proportional representation of instances across categories—Cars, Pedestrians, and Cyclists—mirroring their occurrence in the original training set. Specifically, the instances of Cars, Pedestrians, and Cyclists in the training set were carefully sampled to maintain a consistent distribution across the training, validation, and test sets derived from the original 7481 images. For the following results, the metrics for VirConv-S, UDeerPEP, VirConv-T, TSSTDet, etc. are derived directly from the KITTI dataset's test benchmark, contrasting with our evaluation conducted on a subset partitioned from the original 7481 training images. Given the consistent and uniform distribution of instances across our training, validation, and test sets—a methodology ensuring proportional representation of instances—our following results posit an inferred comparability in performance metrics.

Method	Class: Car			
	mAP Easy	mAP Moderate	mAP Hard	mAP Overall
VirConv-S	0.9248	**0.8720**	0.8245	0.8738
UDeerPEP	0.9177	0.8672	**0.8257**	0.8702
VirConv-T	**0.9254**	0.8625	0.8124	0.8668
TSSTDet	0.9184	0.8547	0.8065	0.8599
Mono3DBox (our solution)	-	-	-	**0.9381**

Method	Class: Pedestrian			
	mAP Easy	mAP Moderate	mAP Hard	mAP Overall
CasA++	0.5633	**0.4929**	**0.4670**	0.5077
TED	0.5585	0.4921	0.4652	0.8702
PiFeNet	**0.5639**	0.4671	0.4271	0.4860
LoGoNet	0.5307	0.4743	0.4522	0.4857
Mono3DBox (our solution)	-	-	-	**0.6829**

Method	Class: Cyclist			
	mAP Easy	mAP Moderate	mAP Hard	mAP Overall
TED	**0.8882**	**0.7412**	**0.6684**	**0.7659**
CasA++	0.8776	0.7379	0.6684	0.7613
CasA	0.8791	0.7347	0.6617	0.7585
LoGoNet	0.8331	0.7071	0.6467	0.7323
Mono3DBox (our solution)	-	-	-	0.60.99

This underperformance raises pertinent considerations about the model's capacity to discern between closely related classes, such as Pedestrians and Cyclists. The observed challenge in accurately detecting Cyclists suggests a potential confusion within the model between these two categories. One argument for the underperformance in detecting Cyclists could relate to the model's spatial reasoning capabilities. Cyclists, unlike Cars or Pedestrians, present a unique spatial configuration due to the bicycle's structure, which may not be adequately captured by the model's implicit geometric inference. To address this, we propose the integration of augmentation techniques aimed at enhancing the model's training dataset. Specifically, the employment of augmentation techniques proposed in [23] [24] [25], which should introduce a broader spectrum of visual variability and context, thereby refining the model's ability to distinguish between Pedestrians and Cyclists more effectively.

Summary

The Mono3DBox framework presents a methodology for directly regressing the 2D projections of vertices of 3D bounding boxes from single images, coupled with a Convex-Hull based IoU for implicit geometry learning. Distinct from traditional depth-based methods, Mono3DBox side steps depth ambiguity challenges, emphasizing geometric fidelity in detection. Our benchmarks on the KITTI dataset demonstrate superior performance for Cars, while revealing areas for improvement in Pedestrian and Cyclist detection—attributed to the model's current limitations in differentiating similar object classes.

Future Work and Timeline

- April 6th - 8th: Categorize the Mono3DBox test set into easy, moderate, and hard categories, based on criteria similar to the KITTI benchmark using KITTI MATLAB dev toolkit.
- April 9th - 11th: Benchmark State-of-the-Art (SOTA) algorithms (VirConv-S, UDeerPEP, VirConv-T, TSSTDet, CasA++, TED, PiFeNet, LoGoNet) on the Mono3DBox dataset split for direct comparison.
- April 12th - 14th: Experiment with adjusting the weights between L1 loss and Convex Hull IoU in the regression loss formula to refine accuracy.
- April 15th - 20th: Expand the training set by incorporating the NuScenes dataset, aiming to enhance Mono3DBox's adaptability and performance across diverse driving conditions.

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 Intelligence44.3 (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.
[20] Jiang, Xueying, et al. "Weakly Supervised Monocular 3D Detection with a Single-View Image." arXiv preprint arXiv:2402.19144 (2024).
[21] Wang, Ke, and Zhichuang Zhang. "Point-voxel fusion for multimodal 3D detection." 2022 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2022.
[22] Wu, Hai, et al. "Virtual sparse convolution for multimodal 3d object detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[23] Zhong, Zhun, et al. "Random erasing data augmentation." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 07. 2020.
[24] Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." arXiv preprint arXiv:2303.05499 (2023).
[25] Dwibedi, Debidatta, Ishan Misra, and Martial Hebert. "Cut, paste and learn: Surprisingly easy synthesis for instance detection." Proceedings of the IEEE international conference on computer vision. 2017.