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 widespread adoption of Autonomous Driving (AD) technologies is contingent upon sophisticated capabilities in 3D environmental perception, which critically depend on the precise, real-time interpretation of spatial data. This requirement has traditionally necessitated the integration of advanced depth-sensing technologies, such as LiDAR and RADAR. Although these technologies are effective, they introduce challenges related to cost, complexity, and scalability that significantly hinder the broader deployment and advancement of AD systems, particularly in environments with limited resources. The Mono3DBox approach proposes that monocular vision, due to its ubiquity and cost-efficiency, could serve as a feasible alternative for 3D object detection. It leverages learning-based models and implicit geometric analysis, aiming to democratize AD technologies by reducing dependence on expensive hardware and lowering the barriers to entry for innovation within the sector.

Mono3DBox adapts the YOLO architecture—known for its balance between accuracy and real-time performance in 2D object detection—to 3D applications by directly regressing the vertices of 2D projections of 3D bounding boxes from single-image inputs. This method circumvents the need for direct depth estimation, which is typically a bottleneck in both computational efficiency and accuracy due to the loss of depth information in 2D representations. Moreover, Mono3DBox integrates a convex hull-based Intersection over Union (IoU) metric as a weighted loss function to enhance bounding box accuracy. This novel integration enhances the geometric fidelity of the 3D bounding boxes, offering a more precise and robust measure of detection accuracy, particularly with the complex object shapes frequently encountered in road scenes. This method not only maintains but also enhances the computational efficiency and scalability of AD technologies utilizing monocular imagery for 3D object detection.

The foundational motivation for Mono3DBox is rooted in both a technological necessity and a visionary pursuit of increased accessibility in AD technologies. It represents an integrative approach that combines cutting-edge machine learning techniques with geometric principles to surpass the limitations of existing monocular 3D detection methods. By addressing the critical challenges related to cost, complexity, and performance, Mono3DBox aims to facilitate 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 recent years, advancements have been seen for limited dataset benchmarks where monocular 3D bounding box estimation has outperformed fusion based methods. 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, Mono3DBox, we extend the foundational work seen in references [7] and [8], where the regression of 3D bounding boxes is accomplished using 2D keypoints. Mono3DBox innovates further by directly regressing the vertices of 3D bounding boxes from single-image inputs, and employing a Convex-Hull based Intersection over Union (IoU) as a novel metric for loss calculation during training. Unlike traditional approaches that depend on intermediate depth estimation, our methodology opts for direct geometric inference from 2D projections, effectively addressing the ambiguities inherent in monocular depth perception. Central to Mono3DBox is the use of the geometric properties visible in 2D projections of 3D objects to infer the spatial configuration of their bounding boxes directly. This is facilitated by a specialized version of the YOLOv8 architecture, which has been adapted to decode geometric cues implicit in monocular imagery. This adaptation allows the model to predict the coordinates of the vertices of 3D bounding boxes directly within the image plane, focusing on spatial reasoning and the understanding of complex geometries necessary for accurate localization. The model employs a Convex-Hull based IoU metric to evaluate the accuracy of the bounding box predictions. This metric calculates the IoU using the convex polygon formed by the predicted vertices and compares it to the convex hull of the ground truth vertices. This method offers a more geometrically accurate measure of detection precision, particularly advantageous in complex environments where traditional IoU metrics may struggle to accurately reflect the precision of 3D object detection. Incorporating curriculum learning into Mono3DBox's training regimen represents a significant enhancement. This methodological approach systematically increases the complexity of the training samples as the model's performance improves. Starting with simpler geometric configurations and gradually introducing more complex and realistic scenarios, curriculum learning helps the model to develop a robust understanding of the geometric principles at play. This staged learning strategy is crucial for training the model to discern subtle

geometric cues and relationships that are vital for precise 3D detection from monocular images. Mono3DBox's strategy of bypassing traditional depth estimation steps is grounded in the observation that accurate 3D spatial configurations can often be derived from purely geometric data present in 2D images. By focusing directly on these geometric attributes rather than attempting to infer depth through complex computational processes, the system significantly reduces the computational load and enhances the potential for real-time applications. This approach differs fundamentally from those that rely heavily on semantic segmentation or depth cues derived from stereo vision, offering a more streamlined and potentially more accurate method for 3D object detection in monocular vision systems. Moreover, by not depending on specific scenarios of object occlusion or precise 2D detection, Mono3DBox aims to offer a generalized and robust solution capable of functioning across a variety of operational environments. This methodology not only addresses the critical limitations posed by depth ambiguity and computational inefficiency but also sets a new benchmark for the deployment of monocular 3D detection systems in real-world applications.

Data Curation

We utilize the KITTI 3D Object Detection dataset, which is structured as follows:

- Total Images: 7,481
    - Training Set: 5,000 images
    - Validation Set: 1,000 images
    - Test Set: 1,481 images

The dataset encompasses a diverse range of instances:

- Car Instances: 28,742
- Pedestrian Instances: 4,487
- Cyclist Instances: 1,627

The following details the distribution of object instances across different difficulty levels in each subset:

Training Set (5000 images)			
	Car	**Pedestrian**	**Cyclist**
Easy	8300	1727	570
Moderate	6538	836	309
Hard	4373	500	249
Total	19211	3063	1128

Validation Set (1000 images)			
	Car	**Pedestrian**	**Cyclist**
Easy	1822	281	109
Moderate	1412	220	63
Hard	963	100	57
Total	4197	601	229

Test Set (1481 images)			
	Car	**Pedestrian**	**Cyclist**
Easy	2322	472	140
Moderate	1789	203	72
Hard	1223	148	58
Total	5334	823	270

Results:

In our project, we benchmarked the Mono3DBox framework against the widely recognized KITTI dataset, a cornerstone in the field of autonomous driving research. For evaluating 3D object detection performance, we utilize the Recall 40 point interpolation, similar to the traditional PASCAL Recall 11 criteria used for 2D object detection. This approach enhances our analysis by allowing for finer distinctions in model performance across a spectrum of detection thresholds. Objects are filtered based on their bounding box height within the image plane to ensure only relevant detections are considered. Only objects visible and labeled on the image plane count towards the evaluation, and those in 'don't care' areas are excluded from false positive calculations. It's important to note, however, that our evaluation does not currently exclude detections of non-visible objects on the image plane, which may contribute to false positives.

For detection accuracy, we enforce category-specific 3D bounding box overlap thresholds:

- Cars: A minimum of 70% overlap is required.
- Pedestrians and Cyclists: A minimum of 50% overlap is required.

The detection difficulties are defined as follows, with adjustments based on visibility, occlusion, and truncation:

- Easy: Minimum bounding box height of 40 pixels, fully visible, maximum truncation of 15%.
- Moderate: Minimum bounding box height of 25 pixels, partly occluded, maximum truncation of 30%.
- Hard: Minimum bounding box height of 25 pixels, difficult visibility, maximum truncation of 50%.

Following are the results for the benchmark against our data distribution used to train and test Mono3DBox:

Method	**Class: Car (IOU > 0.7)**			
	mAP Easy (@R40)	**mAP Moderate (@R40)**	**mAP Hard (@R40)**	**mAP Overall (@R40)**
VirConv-S	0.9479	0.9114	0.8786	0.9126
VirConv-T	**0.9491**	0.9061	0.8720	0.9090
TED	0.8667	0.9240	0.8250	0.8719
LogoNet	0.8939	0.9417	0.8631	0.8995
Mono3DBox (our solution)	0.9168	**0.9543**	0.8927	0.9212
Mono3DBox + Curriculum Learning	0.9182	0.9504	**0.9152**	**0.9279**

Method	Class: Pedestrian (IOU > 0.5)			
	mAP Easy (@R40)	**mAP Moderate (@R40)**	**mAP Hard (@R40)**	**mAP Overall (@R40)**
CasA++	0.6519	0.5958	0.5752	0.6076
TED	0.6685	0.6187	0.5985	0.6285
PiFeNet	0.6494	0.5716	0.5395	0.5868
LoGoNet	0.6219	0.5764	0.5586	0.5856
Mono3DBox (our solution)	**0.6785**	**0.7700**	0.6000	0.6828
Mono3DBox + Curriculum Learning	0.6761	0.7695	**0.6097**	**0.6850**

Method	Class: Cyclist (IOU >0.5)			
	mAP Easy (@R40)	**mAP Moderate (@R40)**	**mAP Hard (@R40)**	**mAP Overall (@R40)**
TED	**0.9029**	**0.7754**	0.7122	**0.7968**
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	0.8136	0.5686	0.4473	0.6098
Mono3DBox + Curriculum Learning	0.8988	0.7264	**0.7397**	0.7883

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 project introduces a novel approach to 3D object detection using monocular vision by adapting the YOLO architecture to directly regress the vertices of 3D bounding boxes from single-image inputs. This method circumvents traditional depth estimation, relying instead on geometric properties visible in 2D projections to infer spatial configurations of objects. The integration of a Convex-Hull based Intersection over Union (IoU) metric for loss

calculation during training enhances the geometric fidelity of detection assessments, proving particularly advantageous in complex traffic scenarios. Through rigorous testing on the KITTI 3D Object Detection dataset, Mono3DBox demonstrated promising results in detecting cars and pedestrians with varying degrees of occlusion and truncation. The system employs a Recall 40 point interpolation metric, allowing for a refined analysis of detection performance across a spectrum of thresholds. This is in contrast to traditional methods which use a Recall 11 point interpolation, providing less granularity in performance assessment. Through these experiments we demonstrate exceptional performance of our model against the best performing models in Cars and Pedestrians classes, without a lot of modifications to the YoloV8 model. We attribute this to inherent capabilities of a generic model such as YoloV8 along with careful fine-tuning.

However, challenges remain, particularly in the detection of cyclists, where the model underperformed compared to other classes. This suggests potential areas for improvement in the model's capability to handle objects with complex spatial dynamics. The adoption of curriculum learning in the training process has shown potential in addressing some of these challenges by progressively introducing more complex detection scenarios, thereby enhancing the model's understanding of intricate geometric relationships.

In summary, Mono3DBox represents a significant technical advancement in the field of autonomous driving by reducing reliance on high-cost sensory equipment and improving scalability and efficiency. Future work will focus on refining the detection capabilities for complex object types and integrating advanced training techniques to further enhance the robustness and accuracy of the model in diverse operational environments.

Future Direction

The promising results demonstrated by the Mono3DBox project thus far, following are the key directions for future development:

- Reformat the nuScenes or BDD100k datasets to match the KITTI format, thereby enlarging the training set for pedestrians and cyclists and overcoming the limitations of the KITTI dataset.
- Experiment with pre-trained 2D bounding box RESNET backbone for YoloV8 that better captures the complex interactions and joint movements typical of cyclists and then fine-tuning it for 3D.
- Develop algorithms that explicitly use symmetry detection and other geometric cues to better infer the orientation and depth of objects from monocular images.
- Implement adaptive weighting schemes in the IoU calculation that change based on object types or scene density, improving the relevance and discrimination of the overlap measure for various detection scenarios.

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.