AGILE3D: Adaptive Contention- and Content-Aware 3D Object Detection for Embedded GPUs

Pengcheng Wang[α], Zhuoming Liu[β], Shayok Bagchi[γ], Ran Xu[δ], Saurabh Bagchi[α], Yin Li[β], Somali Chaterji[α]

α Purdue University, β University of Wisconsin-Madison, γ West Lafayette Jr./Sr. High School, δ NVIDIA

Introduction

3D object detection using LiDAR-generated point clouds is crucial for the perceptual systems in autonomous vehicles, offering detailed environmental insights. However, this process poses significant computational demands.

- **Adaptability:** The system must dynamically adapt to changing latency needs, influenced by diverse environmental conditions.
- **Computational Challenges:** Inference on embedded GPUs is essential to reduce end-to-end latency and maintain data privacy.
- **Contention- and Content-Aware Scheduling:** Choose an execution branch that is optimal based on current resource and input content.

Keywords: Autonomous System, Point Clouds, 3D Object Detection, Embedded GPUs.

Motivational Studies

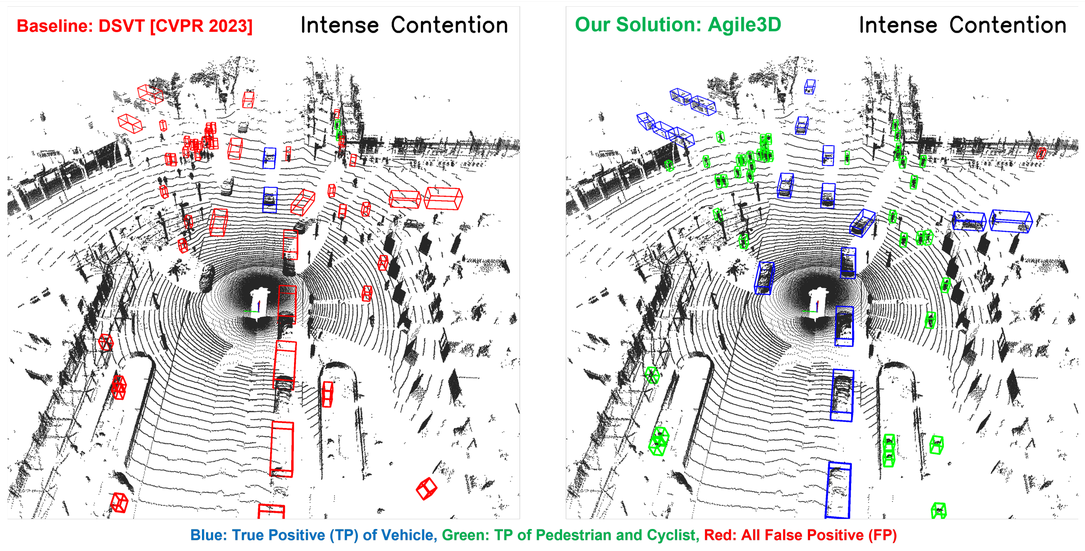

Figure 1. Under resource contention, baseline (DSVT [CVPR2023]) (Left) shows FP results because of accumulated detection lag, while Agile3D (Right) keeps pace with the latency requirement and shows TP results.

3D vs 2D Object Detection: Our study examines the design challenges in 3D object detection using point clouds, comparing 2D and 3D approaches, and highlights the complexities of tuning key parameters. The 3D Encoder—voxelization, voxel encoder, and 3D spatial encoder—dominate inference latency and memory use.

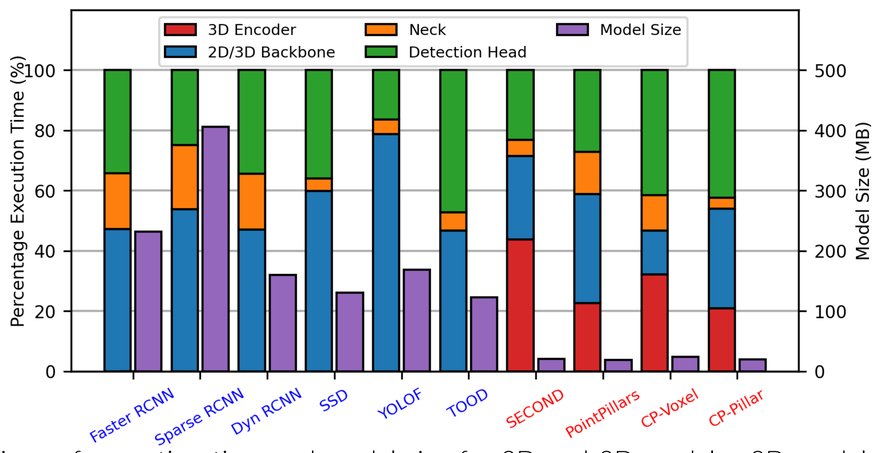

Figure 2. Comparison of execution time and model size for 2D and 3D models. 3D models require higher computation for point clouds but offer better memory efficiency.

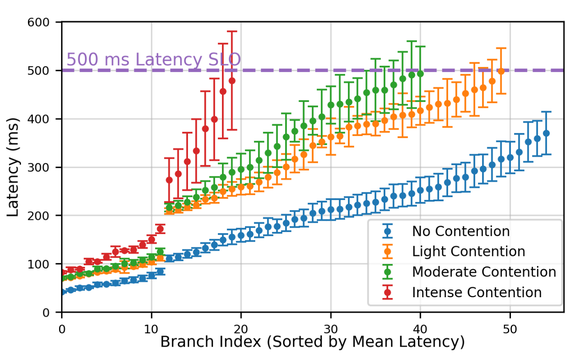

Figure 3. Mean latency with standard deviation across branches. Higher contention increases variability and limits branches within the SLO.

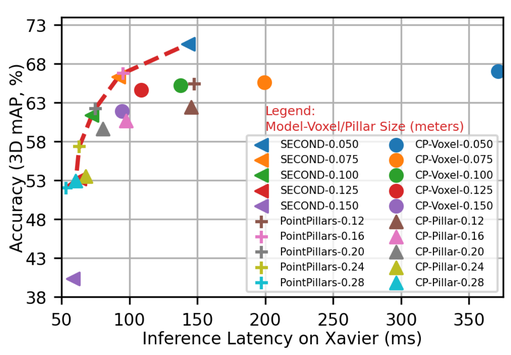

Figure 4. Comparison of 3D models: No single model dominates across all latency ranges, motivating the need for adaptive switching among models.

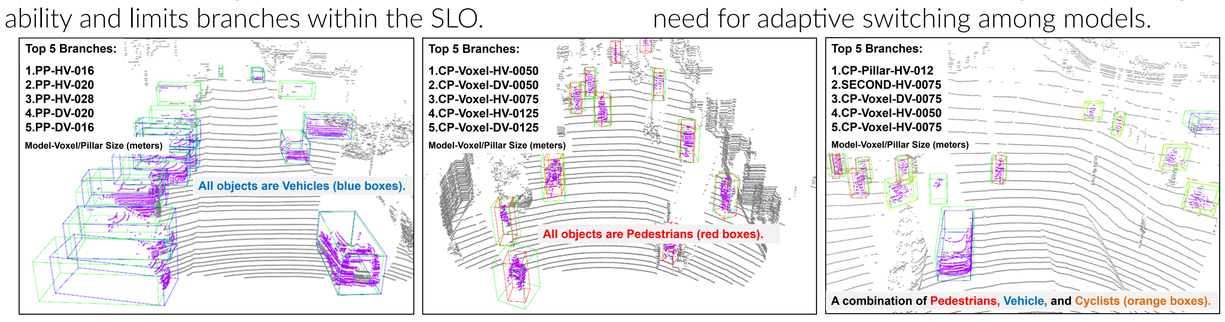

Figure 5. Visualization of diverse point clouds: Vehicles [L], Pedestrians [M], and a mix of Pedestrians, Cyclists, and Vehicles [R]. Ground-truth boxes are green, with top branch predictions for Pedestrians red, Cyclists orange, and Vehicles blue. The top-5 model ranking varies by context.

Challenges

Embedded 3D detection faces three key challenges:

1. Embedded GPUs have tight compute/memory budgets and serve multiple tasks.
2. Voxelization and sparse 3D convolutions in the encoder greatly increase latency and resource usage versus 2D workflows.
3. Scene-dependent sensor inputs and fluctuating co-located applications create rapid content and contention shifts that demand online adaptation.

Our Adaptive 3D Object Detection System

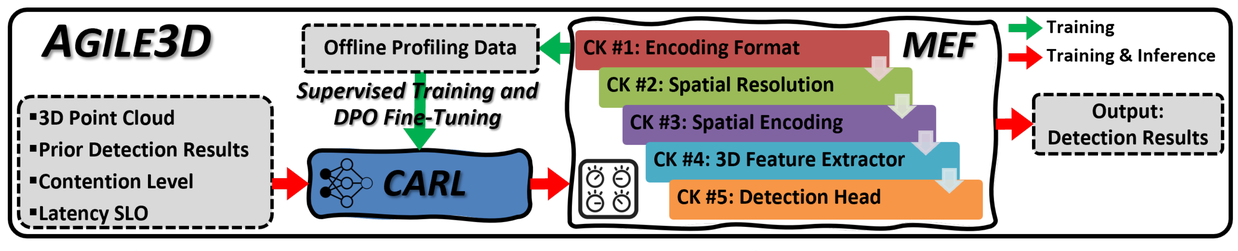

Figure 6. Agile3D integrates MEF and CARL for dynamic branch selection based on input content, contention levels, and latency SLOs. Supervised training with DPO fine-tuning and five control knobs (CK) ensure adaptability across diverse scenarios

Agile3D incorporates a Multi-branch Execution Framework (**MEF**) and a Contention- and Content-Aware RL-based (**CARL**) controller. It tunes key 3D components to balance latency and accuracy by selecting the optimal branch at runtime. CARL aims to select the optimal execution branch that meets the latency SLO and maximizes accuracy, achieved through supervised initial training and Direct Preference Optimization (DPO) fine-tuning.

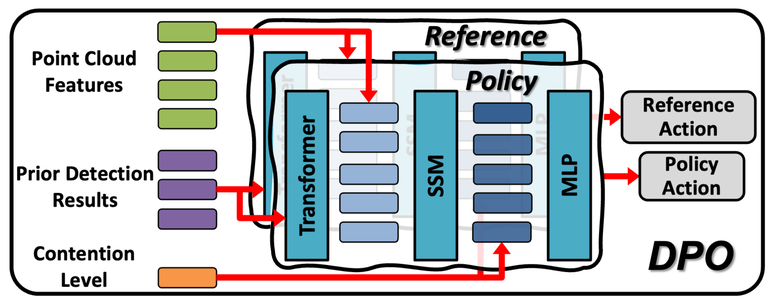

Figure 7. The CARL controller uses a shared architecture for policy and reference models, integrating GD-MAE for 3D features, transformers for prior detection results embedding, SSM for sequence processing, and positional embeddings for latency objectives, enabling adaptive branch selection.

CARL dynamically schedules branches by considering contention levels and frame-specific input content. DPO refines branch selection through preference comparisons instead of absolute scores. We employ an Approximate Oracle controller using Beam Search to provide preference labels, which significantly reduce human labeling efforts.

Evaluation

Two embedded GPUs: NVIDIA AGX Xavier and Orin. Three datasets: Waymo, nuScenes, and KITTI. Baselines: two system controllers: Chanakya [NeurIPS '24], LiteReconfig [EuroSys '22], and six 3D models: CenterPoint [CVPR'21], Part-A2 [TPAMI'20], SSN [ECCV'20], PointRCNN [CVPR'19], PointPillars [CVPR'19], SECOND [Sensors'18].

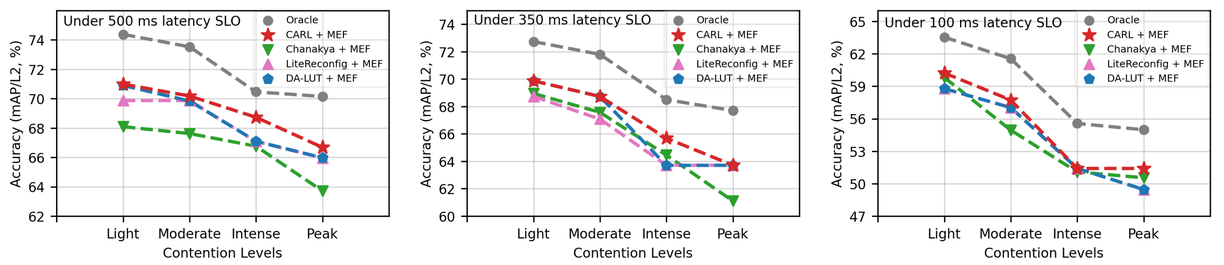

Figure 8. E2E evaluation of Agile3D across varying contention levels and latency SLOs using the Waymo dataset and on Orin GPU. Agile3D consistently achieves superior accuracy, shining on the Pareto frontier across all contention levels and latency SLOs.

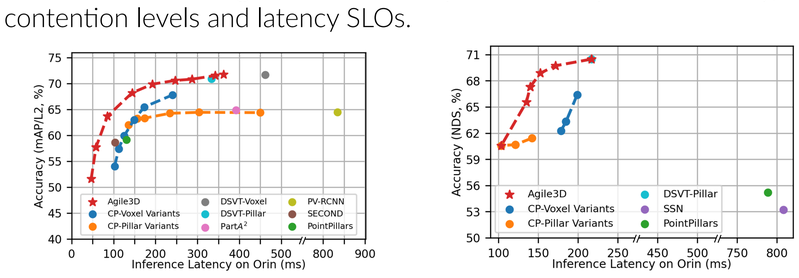

Figure 9. Agile3D vs. baselines on Waymo (Orin). Ours achieves 1-2.5% higher accuracy while adapting to latency SLOs of 50-350 ms, outperforming baselines.

Figure 10. Agile3D vs. baselines on nuScenes (Orin). Ours has 7-16% higher accuracy than CP-Pillar, PP, and SSN, and meets SLOs of 100-250 ms.

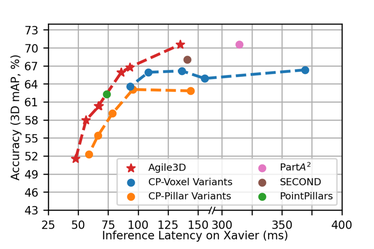

Figure 11. Agile3D vs. baselines on KITTI (Xavier). Agile3D adapts to latency SLOs of 50-150 ms, achieving 5-7% higher accuracy than PP and CP under the 100 ms SLO.

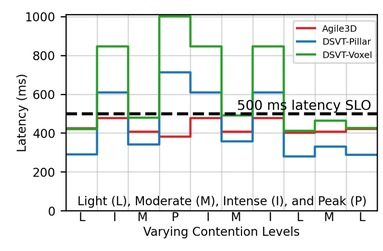

Figure 12. Agile3D adapts to changing contention levels under 500 ms latency SLO. Baselines fail to adapt.

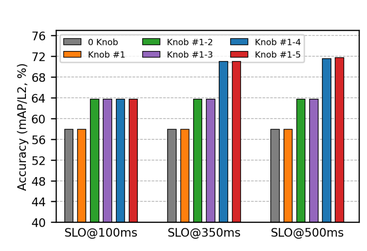

Figure 13. Activating more control knobs improves accuracy and satisfies lower latency SLOs.

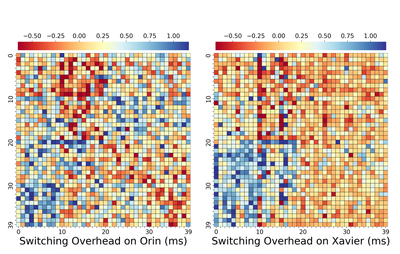

Figure 14. Switching overhead between branches. Mean overhead <1 ms with pre-buffered models.

Conclusion

- We design the first adaptive 3D object detection system for embedded GPUs, which excels in achieving SOTA accuracy while consistently meeting stringent runtime latency SLOs across diverse resource contention levels.
- The system features two complementary and innovative controllers: CARL controller for high contention scenarios and DA-LUT controller for contention-free scenarios.
- By leveraging the MEF and CARL controller, Agile3D efficiently buffers all 3D models in GPU memory, enabling rapid model switching within 1 ms.
- Across multiple datasets and hardware platforms, Agile3D demonstrates superior adaptability and better accuracy.

The 23rd ACM International Conference on Mobile Systems, Applications, and Services.