Research Article

Real-Time 3D Object Detection Using InnovizOne LiDAR and Low-Power Hailo-8 AI Accelerator

Itay Krispin-Avraham[1], Roy Orfaig[2], Ben-Zion Bobrovsky[2]

1. Faculty of Exact Sciences, Tel Aviv University, Israel; 2. School of Electrical Engineering, Tel Aviv University, Israel

Object detection is a significant field in autonomous driving. Popular sensors for this task include cameras and LiDAR sensors. LiDAR sensors offer several advantages, such as insensitivity to light changes, like in a dark setting (Figure 2) and the ability to provide 3D information in the form of point clouds, which include the ranges of objects. However, 3D detection methods, such as PointPillars[1], typically require high-power hardware. Additionally, most common spinning LiDARs are sparse and may not achieve the desired quality of object detection in front of the car. In this paper, we present the feasibility of performing real-time 3D object detection of cars using 3D point clouds from a LiDAR sensor, processed and deployed on a low-power Hailo-8 AI accelerator[2]. The LiDAR sensor used in this study is theInnovizOne sensor[3], which captures objects in higher quality compared to spinning LiDAR techniques, especially for distant objects. We successfully achieved real-time inference at a rate of approximately 5Hz with a high accuracy of 0.91% F1 score, with only -0.2% degradation compared to running the same model on an NVIDIA GeForce RTX 2080 Ti[4]. This work demonstrates that effective real-time 3D object detection can be achieved on low-cost, low-power hardware, representing a significant step towards more accessible autonomous driving technologies. The source code and the pre-trained models are available at https://github.com/AIROTAU/PointPillarsHailoInnoviz/tree/main.

1. Introduction

In the field of autonomous driving, 3D object detection is crucial for understanding the environment and making informed navigation and safety decisions. Traditional approaches often rely on high-power, expensive hardware to achieve real-time performance, posing challenges for scalability and

 doi.org/10.32388/A7AS29

cost-effectiveness. This study investigates the use of the InnovizOne LiDAR sensor in conjunction with the Hailo AI Accelerator, a low-power alternative, to perform real-time 3D object detection.

InnovizOne LiDAR provides high-resolution 3D point cloud data, essential for accurately detecting and classifying objects in various environments. The Hailo-8 AI chip (Figure 3) , designed for edge devices, offers a cost-effective and energy-efficient solution for deploying deep learning models. By leveraging the OpenPCDet framework[5], we adapted state-of-the-art detector, PointPillars, to work with InnovizOne LiDAR data and optimized them for real-time inference on the Hailo chip.

2. Methodology

PointPillars accepts point clouds as input and estimates oriented 3D boxes for objects such as cars, pedestrians and cyclists. It consists of three main stages: (1) A feature encoder network that converts a point cloud to a sparse pseudo-image; (2) a 2D convolutional backbone to process the pseudo-image into a high-level representation; and (3) a detection head that detects and regresses 3D bounding boxes.

2.1. Data Collection

Our autonomous lab vehicle (Figure 1), equipped with an InnovizOne LiDAR, was used by Yasmin Tsiprun in her work[6] to record data in diverse environments, including both static and dynamic scenes across Tel-Aviv University campus and its surrounding roads. This type of LiDAR captures high-resolution 3D point clouds, providing detailed information about the area in front of the car. Unlike cameras, which can struggle in low light, LiDAR offers significant advantages by delivering detailed information about the surroundings in high-resolution 3D (Figure 2).

Figure 1. Left image: The lab vehicle equipped with a multi-sensor kit, including the InnovizOne LiDAR mounted at the front of the roof. Right image: A close-up of the InnovizOne LiDAR mounted on the vehicle's roof

Figure 2. The image shows a traditional camera view at night, where pedestrians are difficult to spot due to poor lighting. The bottom image, generated using Innoviz LiDAR data, reveals a much clearer and detailed scene, detecting pedestrians and vehicles even in complete darkness. This demonstrates how LiDAR technology excels in low-light environments, offering critical advantages for autonomous systems.[7]

doi.org/10.32388/A7AS29

2.2. Data Annotation

The raw LiDAR data was divided into segments, with each frame labeled to identify cars. The labeled data was converted into a format compatible with the OpenPCDet framework, specifically preparing it for the PointPillars detector model. For each object, the annotated data include the 3D center of the objects (cx, cy and cz), the orientation of the objects in relation to the LiDAR (theta) and a 3D bounding box size (width, height and depth). We chose to utilize this dataset in our project.

2.3. Real-Time Inference on Hailo AI Accelerator

To achieve real-time performance, we integrated the trained detector with the Hailo AI accelerator. This involved several steps and optimizations to ensure that the model performed efficiently and accurately on the device, given its computational constraints compared to more powerful but costly alternatives like NVIDIA Jetson[8].

The Hailo-8 AI chip is designed for efficient edge computing, offering low power consumption, which makes it suitable for deployment in resource-constrained environments, and high computational efficiency, enabling real-time processing of high-resolution sensor data.

Figure 3. Hailo-8 AI Accelerator.

doi.org/10.32388/A7AS29

3. Pipeline Overview

To integrate the PointPillars model with the Hailo-8 AI accelerator, we utilized Hailo's proof-of-concept (POC)[9], which demonstrated the offloading of computationally intensive 2D-convolutional layers of a 3D object detection network operating on point clouds from the KITTI dataset[10]. We adapted this POC to process data captured by our InnovizOne LiDAR. The pipeline for this POC involves the following steps:

3.1. Data Preparation and Preprocessing

- **Conversion to Compatible Formats:** The raw InnovizOne LiDAR data was converted into a format compatible with the PointPillars model and the Hailo hardware. This involved segmenting the data into frames and labeling objects within each frame for the model training phase.
- **Normalization and Augmentation:** The PCDet framework performs data normalization and augmentation in order to improve model robustness. This included transformations such as random flips and rotations to simulate different environmental conditions.

3.2. Model Adaptation for Hailo

To leverage the Hailo hardware for accelerating the 2D-convolutional parts of the PointPillars network, the model architecture and execution flow was adapted. This involves several critical steps: exporting the PyTorch model to ONNX[11], translating the ONNX model to a Hailo-compatible format, and creating a new PyTorch module that integrates the Hailo-inferred operations. The process included the following kep steps:

3.2.1. Extracting the 2D Backbone and Dense Head

The 2D convolutional layers and the detection head were isolated from the PointPillars network. These components are responsible for most of the computational load and are well-suited for offloading to the Hailo hardware (Figure 4). A new PyTorch module was created to encapsulate these components. This module takes the spatial features as input and produces intermediate features, classification predictions, and bounding box predictions.

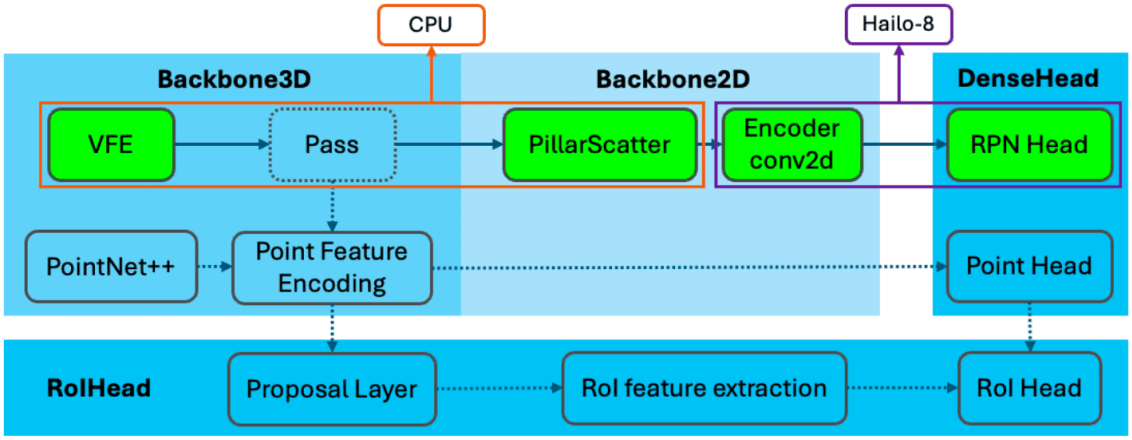

Figure 4. The diagram shows in orange the part of the PointPillars model architecture that is processed by the CPU and in purple the part that is being offloaded to the Hailo-8 accelerator.

3.2.2. Exporting to ONNX

The newly created module was exported to the ONNX format. This format serves as an intermediary representation that can be parsed and optimized by Hailo tools. The ONNX model was simplified using onnxsim to ensure compatibility and efficiency in subsequent steps.

3.2.3. Translating to Hailo Internal Representation

Using the Hailo SDK, the simplified ONNX model was translated into Hailo's internal format (HAR). This process involves parsing the ONNX model and mapping its operations to Hailo's hardware capabilities. The resulting HAR file encapsulates the 2D convolutional layers and the detection head, ready for execution on Hailo hardware.

3.2.4. Creating a PyTorch Module for Hailo Execution

A new PyTorch module was defined to replace the original 2D backbone and dense head in the PointPillars network. This module integrates with Hailo's hardware to perform the 2D convolution and detection head operations. It handles the conversion of data formats and interfaces with the Hailo SDK. This ensures that the spatial features are processed by the Hailo hardware and the results are seamlessly integrated back into the PyTorch model flow.

The system architecture was designed to maximize the efficiency of real-time 3D object detection while maintaining low power consumption. Key components included:

- **Data Acquisition Module:** Captures LiDAR data and converts it into a format suitable for the processing pipeline.
- **Preprocessing Module:** Applies necessary transformations and augmentations to the raw data, ensuring it is ready for inference.
- **Inference Engine:** Runs the optimized PointPillars model on the Hailo-8 AI chip, performing real-time 3D object detection.
- **Post-Processing Module:** Refines the model output, filtering and merging detections to provide accurate and reliable results.

4. Integration and Testing

Overview: After adapting the model for Hailo, we integrated it into the overall inference pipeline and performed testing. This ensured that the adapted model produced consistent and accurate results and leveraged the Hailo hardware effectively. The process included the following kep steps:

4.1. Quantization and Optimization

- To further optimize the model for Hailo hardware, a quantization process was performed. This process involved creating a calibration dataset from the spatial features input to the 2D backbone.
- Using the Hailo SDK, the model based on the calibration dataset was optimized. This step ensured that the model was numerically efficient and ready for execution on Hailo hardware.
- The optimized model was saved as a quantized HAR (q-HAR) file.

4.2. Compiling for Hailo Hardware

- The quantized model was then compiled for execution on Hailo hardware. This involved generating a hardware executable file (HEF) that encapsulates the optimized model.
- The compilation process included creating an allocation plan and optimizing the resource utilization on the Hailo hardware. The resulting HEF file was ready for deployment.

4.3. End-to-End Integration

- The Hailo hardware execution was integrated into the overall inference pipeline using HailoRT's asynchronous send/receive functionality. This allowed for efficient and pipelined processing.
- Two new PyTorch modules were defined to handle the operations before and after the Hailo-mapped parts. These modules encapsulated the preprocessing and postprocessing steps, ensuring a seamless flow of data.
- A multiprocessing setup was implemented to manage the data transfer between PyTorch and Hailo hardware. Separate processes handled the sending and receiving of data, enabling efficient and parallel execution.

4.4. Final Testing

- Extensive testing was performed to verify the accuracy and performance of the integrated model. We compared the results with the original model to ensure consistency.
- The end-to-end inference pipeline was validated to ensure it met the required performance metrics and utilized the Hailo hardware effectively.

Model	F1 Score	Recall	Precision	AP
PVRCNN	**0.96**	**0.97**	0.96	**0.97**
PointPillars	0.92	0.87	**0.97**	0.87
PointPillars+Hailo (ours)	0.91	0.87	0.96	0.85

Table 1. The result metrics[12] were evaluated using an IoU threshold and a confidence threshold of 0.3. We evaluated the results for two models (PVRCNN and PointPillars) and for the new pipeline created with the offloaded computation to the Hailo chip based on the PointPillars model.

5. Results

The optimized model on the Hailo chip achieved a processing rate of approximately 5 Hz, with detection accuracy comparable to running on more powerful hardware. This demonstrated the

feasibility of deploying advanced 3D object detection models on low-power edge devices. In addition to the PointPillars model, we also trained the PV-RCNN model[13], which is a more complex and powerful 3D object detection architecture. PV-RCNN is known for its superior accuracy due to its multi-scale feature aggregation and region-based refinement. However, the model is computationally heavier, and we recognized early on that it would not be feasible to run PV-RCNN on the Hailo chip, given its resource constraints. Despite this, we conducted the comparison to evaluate the potential trade-offs between model complexity and performance. While PV-RCNN achieved higher accuracy on the same dataset, its processing speed was significantly slower, further justifying our choice to optimize the lighter PointPillars model for real-time edge inference on the Hailo platform.

The experimental results, summarized in Table 1, highlight the performance metrics for the PV-RCNN and PointPillars models. Evaluations were conducted using an IoU threshold and a confidence threshold of 0.3, with comparisons made between the models' performance with and without the Hailo-optimized pipeline. Notably, the PointPillars model, optimized for the Hailo chip, demonstrated competitive accuracy while achieving faster processing speeds. Detection metrics for PointPillars at these thresholds showed minimal variation between the standard and Hailo-optimized pipelines.

6. Discussion

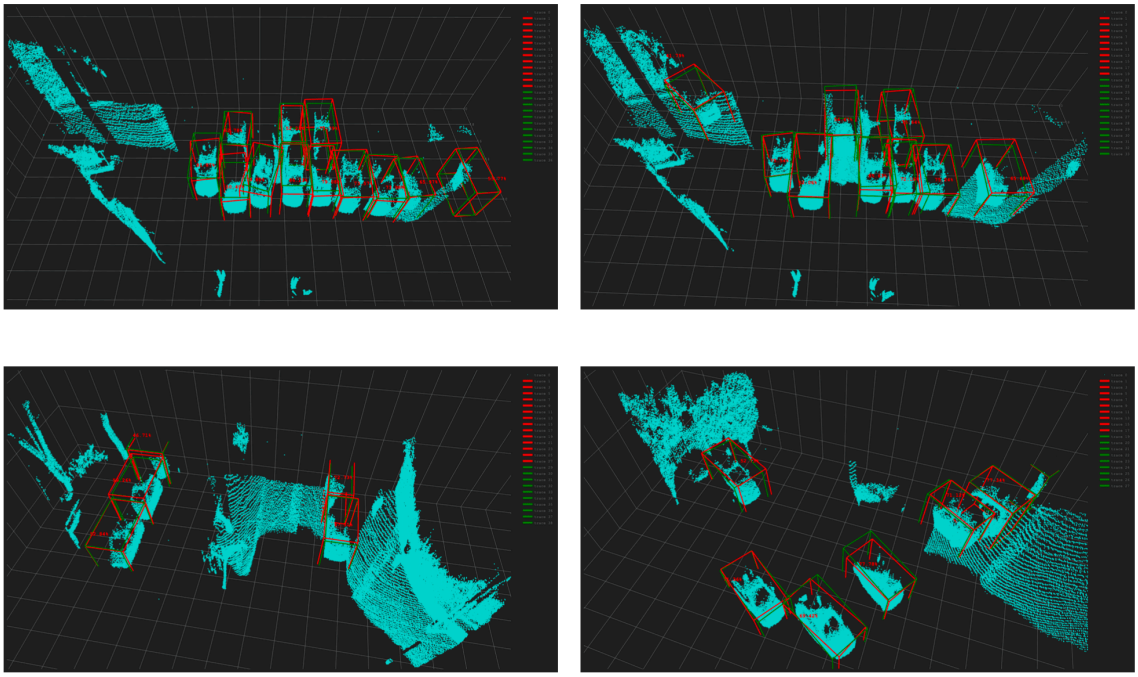

Figure 5. Above are several detection of cars generated by the PointPillars model using Hailo on point cloud data from Innoviz LiDAR. Red bounding boxes represent the detector results, while green bounding boxes indicate the ground truth.

The successful deployment of real-time 3D object detection on low-power hardware is a major step forward for the future of autonomous driving technologies. This advancement not only demonstrates the feasibility of using efficient, power-constrained hardware in complex perception tasks but also highlights the potential for broader adoption across various autonomous platforms. By addressing challenges related to scalability and cost, our approach enables more affordable access to advanced detection capabilities, which could benefit a wide range of applications, from commercial autonomous vehicles to agricultural and industrial automation.

In addition, our study identifies key areas for further development, including optimization techniques that could reduce processing latency and enhance detection accuracy. Expanding the applicability of our approach to integrate with various sensor types, such as other LiDAR models and radar, could improve system robustness across diverse driving conditions and environments. These ongoing

doi.org/10.32388/A7AS29

improvements will be essential for advancing the reliability and versatility of real-time 3D object detection in autonomous systems

7. Conclusion

This study illustrates that high-performance real-time 3D object detection is achievable using cost-effective, low-power hardware. By leveraging the capabilities of the InnovizOne LiDAR sensor in combination with the Hailo AI chip, and through careful optimization of the PointPillars model architecture, we achieved a processing rate of 5Hz with substantial accuracy. This setup strikes a balance between performance and energy efficiency, showcasing the feasibility of deploying advanced perception systems without the need for expensive, high-power GPUs traditionally associated with autonomous vehicle technologies.

The results underscore the potential for more accessible and scalable autonomous driving solutions. This approach can be particularly impactful in applications where both cost and power consumption are limiting factors, such as compact or lightweight robotic systems, autonomous delivery vehicles, and agricultural or industrial automation. Moreover, it opens opportunities to integrate high-resolution 3D perception in environments with constrained resources, making autonomous technology more adaptable and affordable across a wider range of sectors. Future work could extend these optimizations to support additional sensor types and processing frameworks, further broadening the scope and accessibility of autonomous systems.

Acknowledgments

We thank both Innoviz and Hailo for their generous donations of the LiDAR and the AI accelerators. We also thank Yasmin Tsiprun for her work on collecting and creating the point-cloud dataset and allowing us to use it in this project. We would also like to thanks Roi Raich for all the work and support he provided us along the way, including in the creation of the dataset.

References

1. ^ *Lang AH, Vora S, Caesar H, Zhou L, Yang J, Beijbom O (2019). "PointPillars: Fast Encoders for Object Detection from Point Clouds". arXiv. arXiv:1812.05784 [cs.LG].*
2. ^ *Hailo-8 AI Accelerator overview.*

3. ^ *InnovizOne website*.
4. ^ *Nvidia2080*.
5. ^ *OpenPCDet - LiDAR-based 3D object detection*.
6. ^ *Innoviz pointcloud dataset -link. Available from: https://docs.google.com/document/d/1-_x3RI1r_vA-NzFSG2OaRcuC8rQ49QxU/edit?usp=drive_link&ouid=116133536233416768299&rtpof=true&sd=true*.
7. ^ *Innoviz - what is lidar blog*.
8. ^ *NVIDIA Jetson description*.
9. ^ *Hailo Application Code Examples*.
10. ^ *Liao Y, Xie J, Geiger A (2022). "KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D". arXiv. arXiv:2109.13410 [cs.CV]*.
11. ^ *ONNX*.
12. ^ *Powers DMW (2020). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". arXiv. arXiv:2010.16061 [cs.LG]*.
13. ^ *Shi S, Guo C, Jiang L, Wang Z, Shi J, Wang X, Li H (2021). "PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection". arXiv. arXiv:1912.13192 [cs.CV]*.

Declarations

Funding: We thank both Innoviz and Hailo for their generous donations of the LiDAR and the AI accelerators.

Potential competing interests: No potential competing interests to declare.