Moby: Empowering 2D Models for Efficient Point Cloud Analytics on the Edge

Jingzong Li*, Yik Hong Cai©, Libin Liu•, Yu Mao*, Chun Jason Xue*, Hong Xu©
*City University of Hong Kong; •Zhongguancun Laboratory; ©CUHK

Abstract

3D object detection plays a pivotal role in many applications, most notably autonomous driving and robotics. These applications are commonly deployed on edge devices to promptly interact with the environment, and often require near real-time response. With limited computation power, it is challenging to execute 3D detection on the edge using highly complex neural networks. Common approaches such as offloading to the cloud induce significant latency overheads due to the large amount of point cloud data during trans- mission. To resolve the tension between wimpy edge devices and compute-intensive inference workloads, we explore the possibility of empowering fast 2D detection to extrapolate 3D bounding boxes. To this end, we present **Moby**, a novel system that demonstrates the feasibility and potential of our approach. We design a *transformation* pipeline for Moby that generates 3D bounding boxes efficiently and accurately based on 2D detection results without running 3D detectors. Further, we devise a *frame offloading scheduler* that decides when to launch the 3D detector judiciously in the cloud to avoid the errors from accumulating. Extensive evaluations on NVIDIA Jetson TX2 with real-world autonomous driving datasets demonstrate that Moby offers up to **91.9% latency improvement** with modest accuracy loss over state of the art.

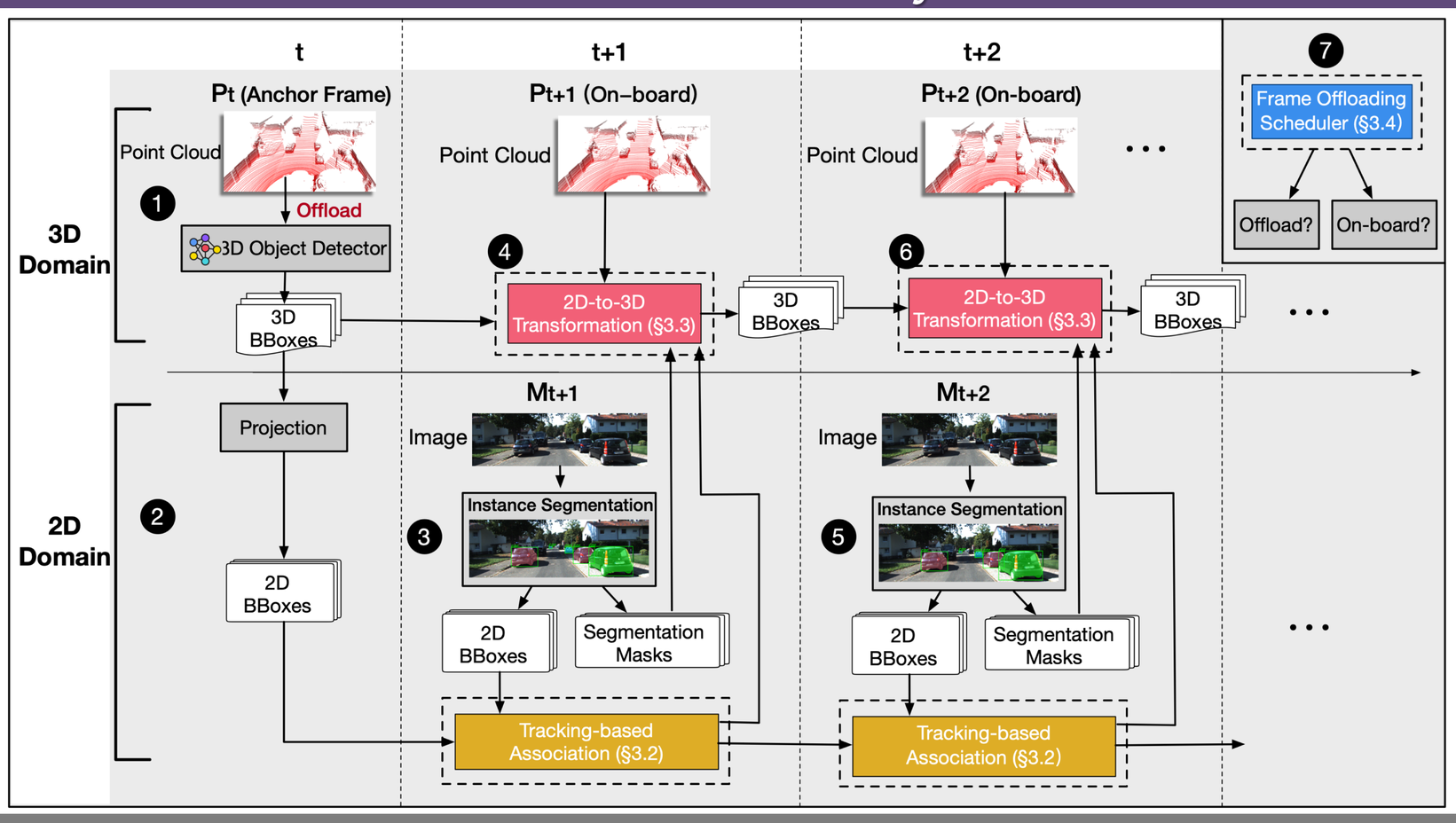
Motivation

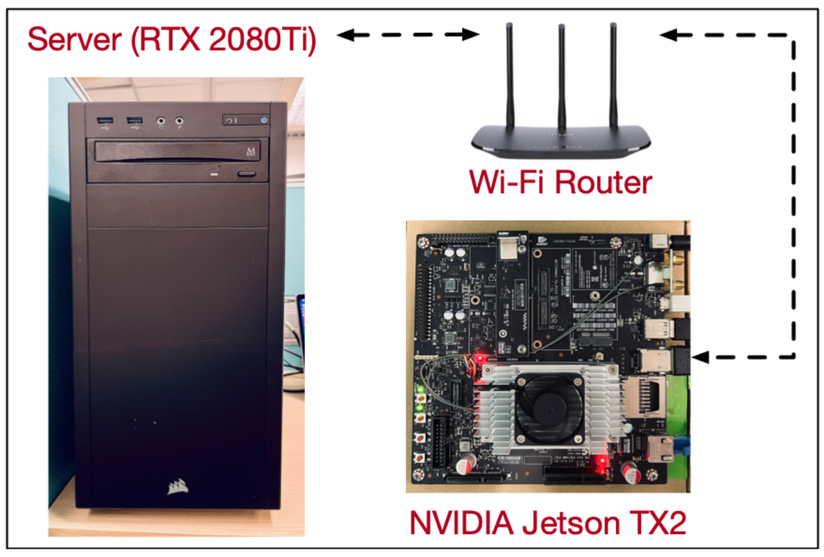
- **Resource on edge devices is constrained.**
- **3D object detection are much more compute-intensive than 2D object detection, and is ill-suited for edge devices.**
- **Transmission of point cloud dominates the end-to-end latency if offloading the 3D detection to server for processing.**

Design

- **Tracking-based Association**
  - Establish the association of 2D bounding boxes in adjacent frames.
- **2D-to-3D Transformation**
  - A light-weight geometric method that takes in 2D results to generate 3D bounding boxes efficiently.
- **Frame Offloading Scheduler**
  - Judiciously decide when to offload a new frame to launch 3D detectors.

System Overview

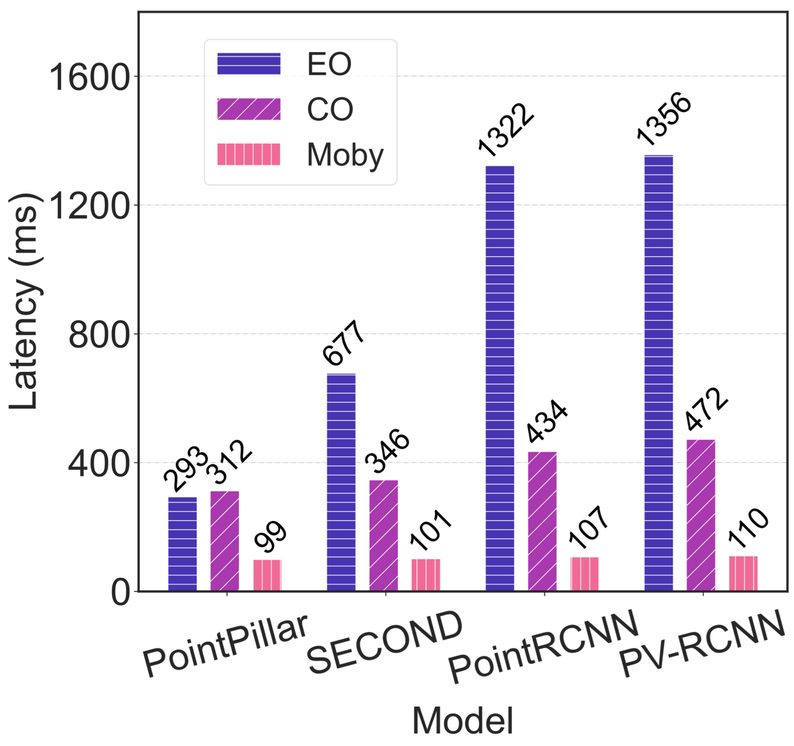




Moby is implemented on an NVIDIA Jetson TX2 and a desktop server with 2080Ti GPU.

Evaluation

- Figure 1: Moby outperforms Edge-Only and Cloud-Only approaches in latency with significant margins ranging from **56.0%** to **91.9%** across models..



- Figure 2: Moby almost maintains the detection accuracy. The accuracy drops slightly between **0.027** to **0.056**, which is negligible.