Edge AI Optimization on Rails—Literally

Matthew Pietrzykowski

Principal Data Scientist

Wabtec Corporation

Wabtec CORPORATION

Who We Are & What We Do

Introduction

ADVANCED TECHNOLOGY LABS

OUR MISSION

*De-risk innovation investments and deliver **breakthrough** solutions that scale for our customers.*

The Advanced Technology Team

PRODUCT FOCUSED INNOVATION

Disrupting today, transforming tomorrow

EXPERTISE:

- MACHINE & DEEP LEARNING
- COMPUTER VISION
- INTELLIGENT SYSTEMS
- ROBOTICS & AUTOMATION

3+

YEAR PRODUCT OUTLOOK

30

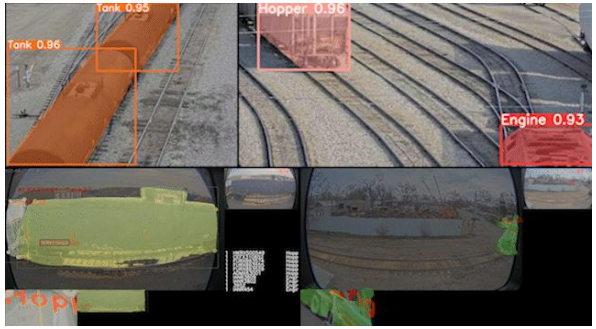
ROBOTICS ENGRS, DATA SCIENTISTS. ML ENGINEERS, SW DEV, UX

4

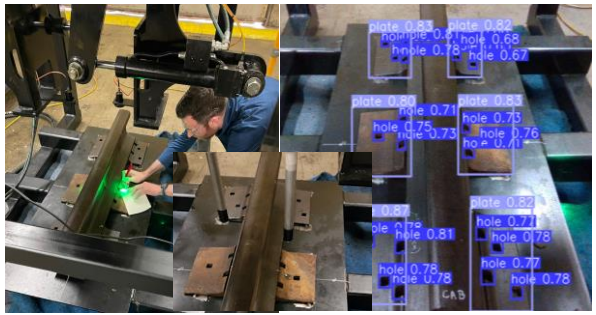
STRATEGIC PARTNERSHIPS

Wabtec CORPORATION

Advanced Technology Labs: Wabtec Innovation

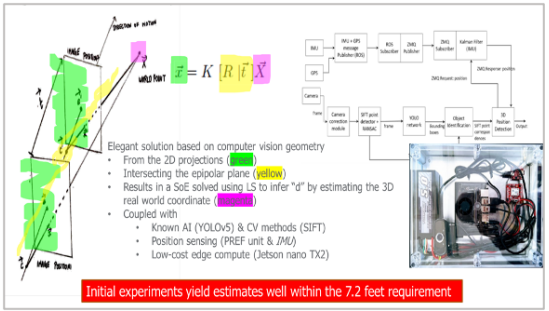


REAL-TIME ASSET & CARLOAD INVENTORY TRACKING SYSTEM



SPIKING VEHICLE ADVANCED SYSTEMS

AI WITH EXISTING CAMERA HW FOR DERAILMENT & ENVIRONMENTAL RISK

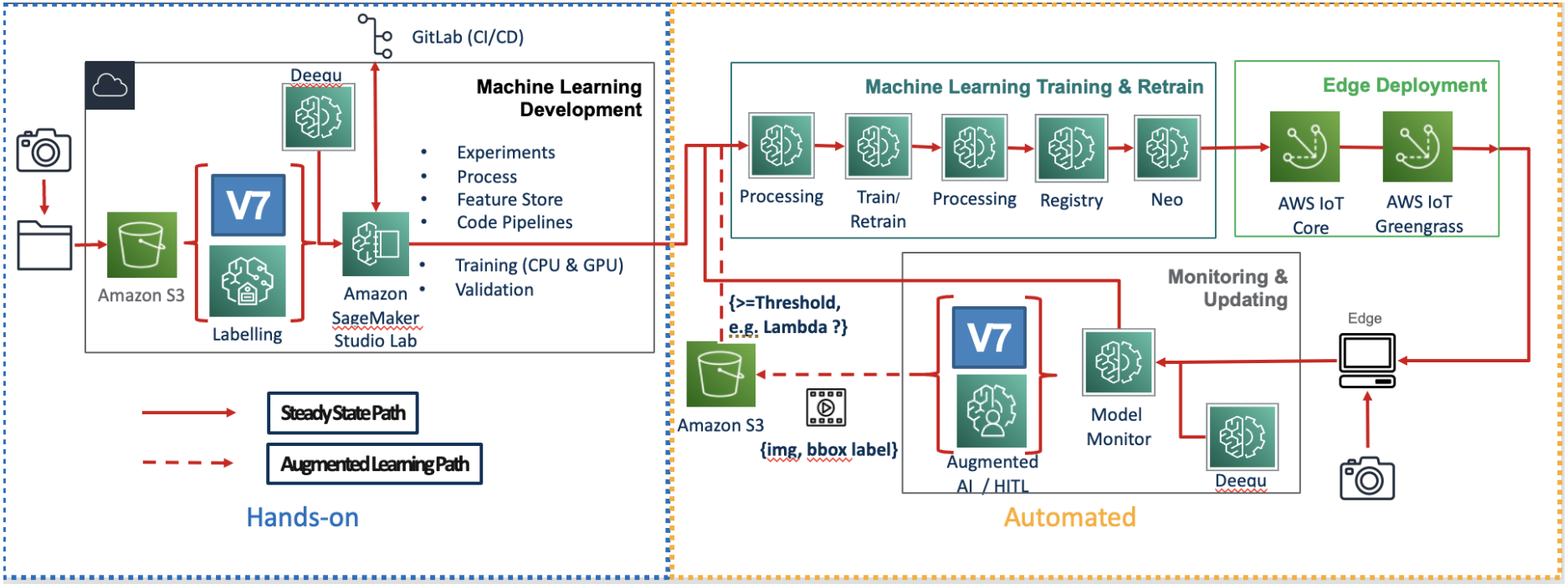


COLLABORATION WITH RESEARCH PARTNERS & ADJACENT INDUSTRIES

PASSENGER DETECTION & DETECTION MAPPING SYSTEMS

ADVANCED PERCEPTION SYSTEMS WITH MULTI-SENSOR FUSION FOR TRAIN AUTOMATION

Hybridized Edge Deployment



Approach to Edge Device Deployment

Metrics to Consider

- Latency
- Throughput
- Memory
- Power consumption
- Model size
- Accuracy
- Parallelism
- Pipelining
- Hardware
- Portability

Approaches for Optimal (or at least Target) Performance

- Model conversion to TensorRT
- Use a smaller model architecture
- Quantization
- GPU-specific libraries
- Model pruning
- Batch size optimization
- Input resolution
- Profiling and benchmarking

Example 1: Transit – Bus Doors

VaporBus – Vision Application

VaporBus is the principal door equipment supplier to the North American bus industry.

Current Product

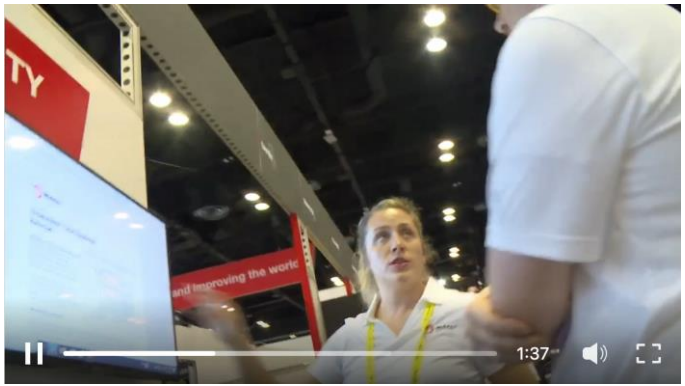
- CLASS – Contactless Acoustic Sensing System.
- Acoustic sensors to detect passengers in doorway.

Challenges

- Coverage of sensors - limited.
- Dead zones in the coverage area to account for environmental changes – E.g., Door handle reflects sound.
- Environment change.

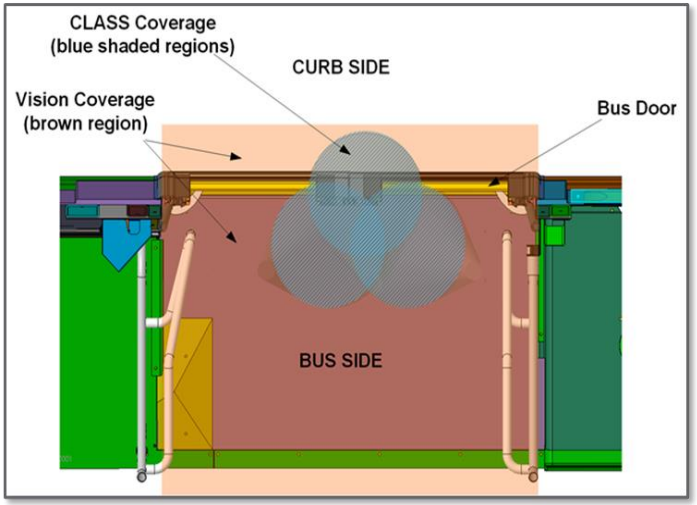
VaporBus Vision Solution Advantages

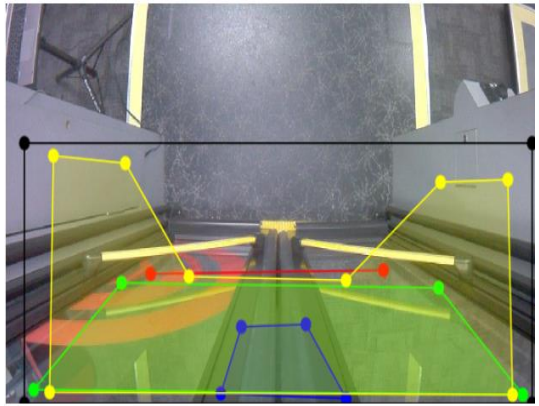
- Better specificity
- Improved coverage area.
- Configurable detection/action zones
- Improved responsiveness.



VaporBus Vision Overview

QUALITY ASSURANCE	PASSENGER DETECTION	OBSTRUCTION DETECTION	PASSENGER COUNTING
• Flag outlier frames before the algorithm	• Detect passengers near door for touchless open.	• Risk mitigation to avoid closing of door on passengers	• Count passengers getting in and out of the bus.

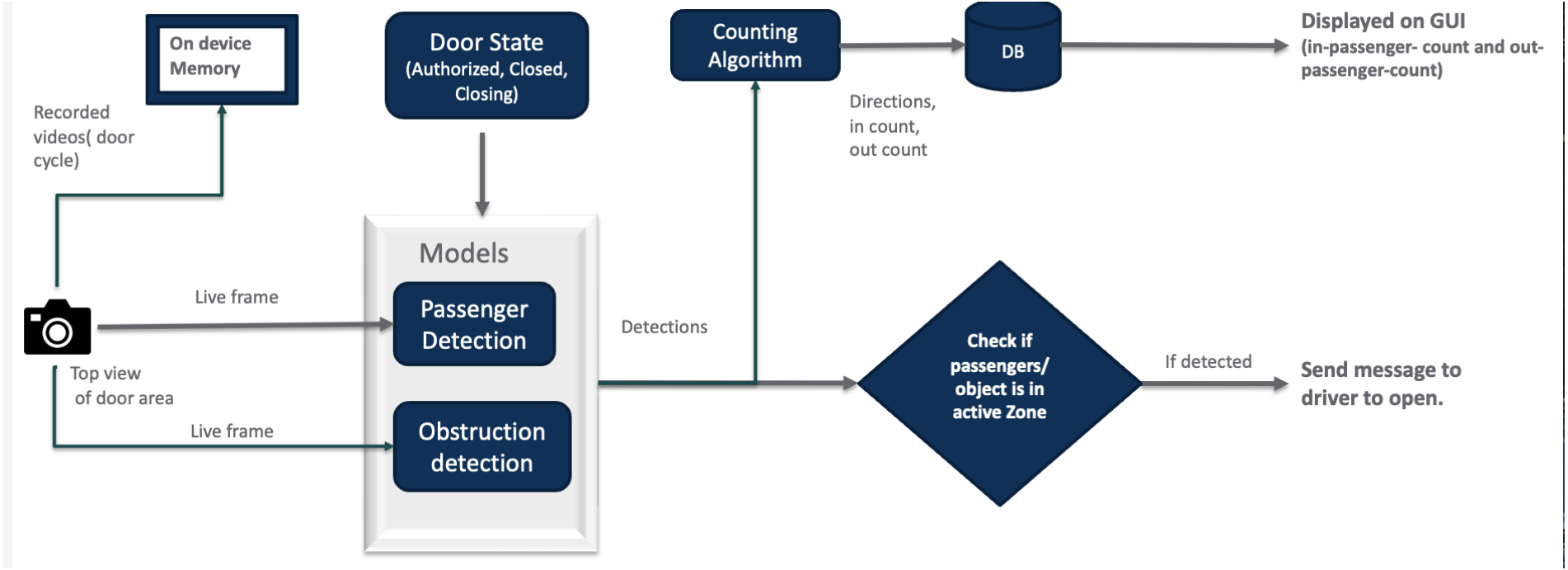




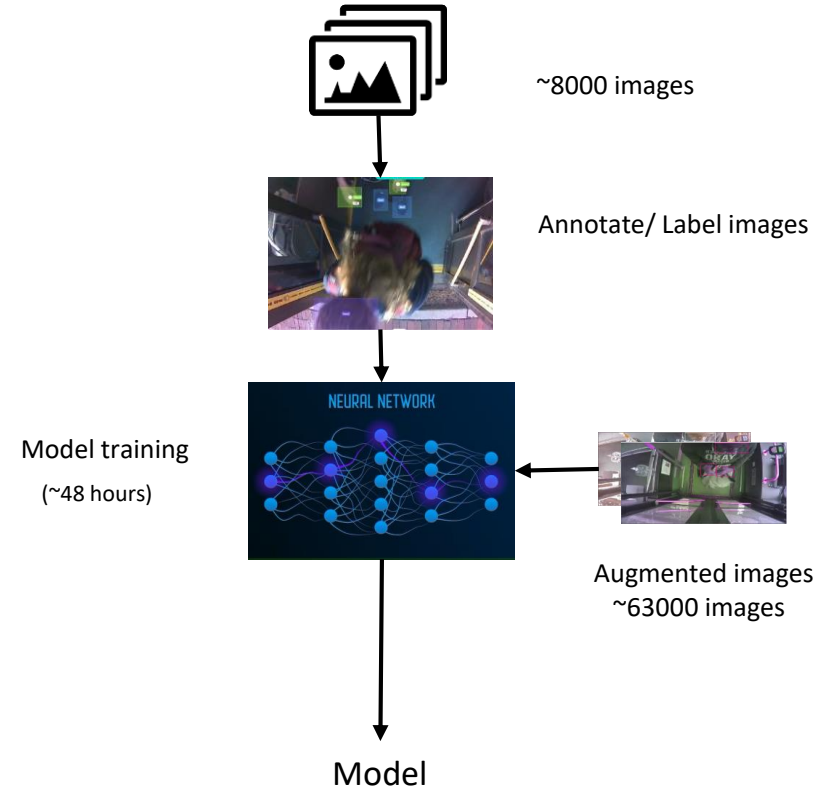
- Yellow line – Collision Zone
- Green line – Drunk Alarm Zone
- Blue line – Touchless Open Zone

Wabtec CORPORATION

VaporBus High Level Solution Diagram



Example: Object Detection



Example: Obstruction Detection Journey

1. Classic CV → 2. Segmented DNN → 3. Flip #2

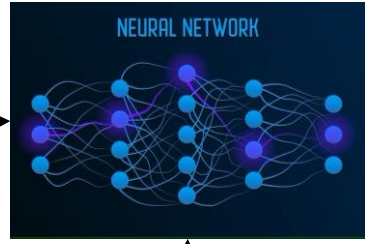
- Model is trained exclusively on background images not on the objects
- Deviation as Anomaly: When inferencing new frames, any departure from this learned background signifies an anomaly or an object of interest.

~1000 Images

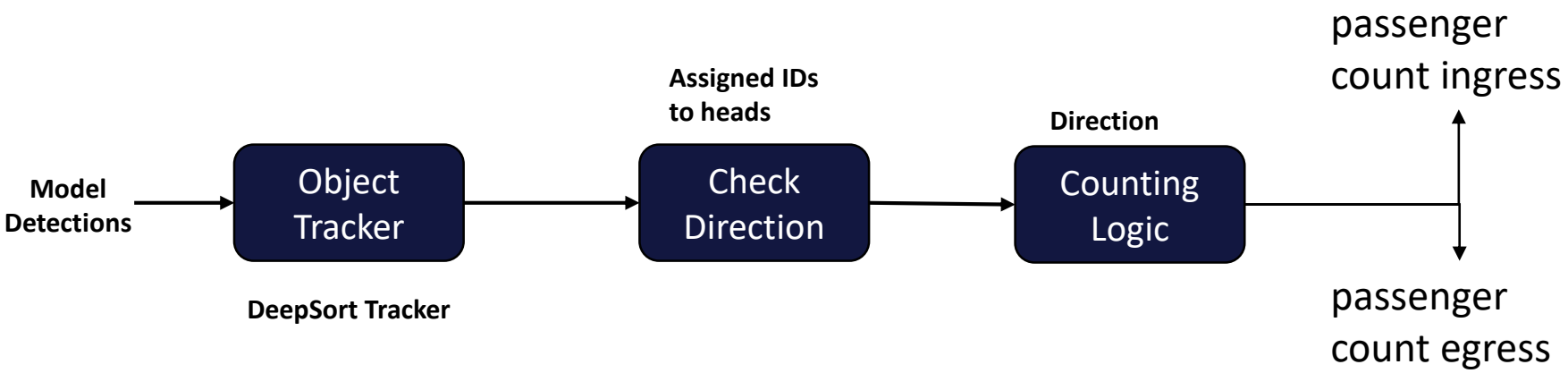
Annotate/ Label Objects in images

Binary Masking to get background mask

Segmentation Model



Example: Passenger Counting



Edge Compute

[Jetson Orin for Next-Gen Robotics | NVIDIA]

NVIDIA® Tegra® X1 series SoC
- NVIDIA Maxwell GPU
- ARM® quad-core Cortex®- A57 CPU Complex

4GB LPDDR4 memory
16GB eMMC 5.1 storage
Gigabit Ethernet (10/100/1000 Mbps)
PMIC, regulators, power and voltage monitors
260-pin keyed connector (exposes both high-speed and low-speed industry standard I/O)
On-chip temperature sensors

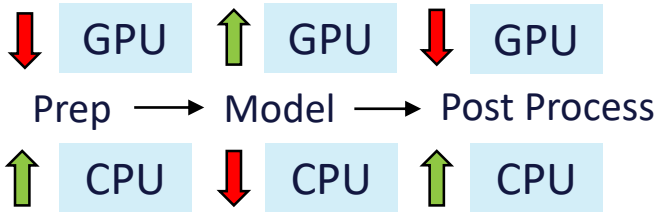
Constraining Issues

- Target cost
- Model size, Image size, core programs
- Competition for CPU
  - Vision tasks
    - Camera driver
    - Video feed
    - Pre & post processing
  - Core tasks
    - Web GUI
    - State Machine
      - Hardware
      - I/O
      - CAN msgs
      - Etc.

Holistic Approach to Compute Constraints

How to ingest & supply the incoming video stream as quickly as the system can handle it…

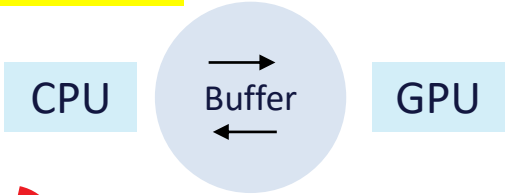
Serially:



- ✓ Train -> ONNX -> TensorRT -> Docker
- ✓ Architecture Sizing
- ✓ Quantization
- ✓ GPU-Specific Libraries

= 24 - 28 fps

Parallel:



max(CPU time, GPU time) rather than sum(CPU time, GPU time)

Example 2: Freight Rail Audit

Freight Rail Audit: Need

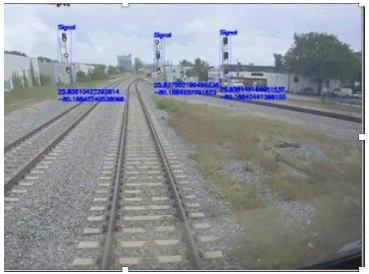
Key Infrastructure Detection & Dead Reckoning (DR)
Utilize the existing onboard camera to identify key Positive Train Control (PTC) and wayside assets... as well as, using those assets for DR.

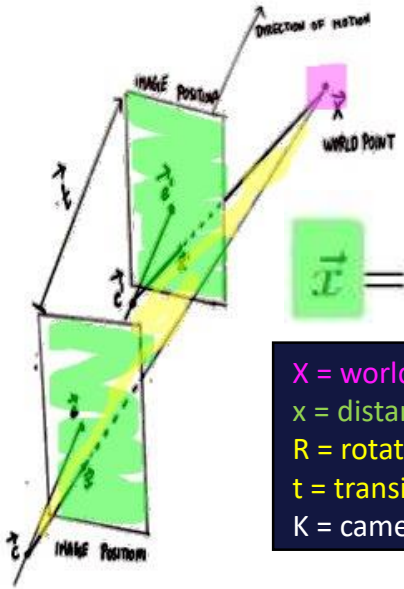
Track Inspection
Identify track and equipment irregularities that can cause a derailment

Environmental
Identify environmental factors around the track that pose a safety risk

Freight Rail Audit: Approach

Defect & Asset Detection	Localization	Classification & Risk	Action: Onboard & Offboard
• ML Algorithm • Tag & Time Stamp	• Onboard PREF • Epipolar Geometry	• Track Health • Vegetation • Asset	• Onboard display • Urgent Alert • Onboard record • Upload for remote health assessment





$$\vec{x} = K \begin{bmatrix} R & |\vec{t} \end{bmatrix} \vec{X}$$

X = world point
x = distance to the object
R = rotational matrix
t = transition displacement
K = camera matrix

- Most solutions found in the literature rely on expensive vision **sensor ensembles making the value proposition questionable**
- Initial attempts focused on calibrating the FOV to infer the AOV and finally to calculate the distance to the object. This led **to large error propagation**.
- The solution developed focuses on elegant hybridization of epipolar geometry, artificial intelligence, and classic CV methods.
- The algorithm uses a hybridized pipeline of of pre-trained, size appropriate algorithms & onboard locomotive hardware to keep the solution as simple as possible

Wabtec CORPORATION

Edge Compute – Onboard Optimization Approach

Defect & Asset Detection → Localization → Action: Onboard & Offboard → Classification & Risk Metrics

- Object Detection Model sizing
  - ❿Yolo V5 Nano
  - MobileNet V2

- Correspondence Point Detection
  - SIFT
  - BRISK
  - ❿FAST - BRIEF

- Onboard Device
  - ❿GoLink
    - Memory
      - 16GB of DDR4 RAM
      - ❿32GB of DDR4 RAM
    - CPU
      - i7-6600U
      - ❿i7-1185FRE

- Detection Accuracy
  - Track Health
  - Vegetation
  - Asset
- Risk Probability
- Frame Rate

Recap

- Performance is dependent on system understanding
- Keep the target in mind when performance tuning
- Effective team collaboration with appropriate skill complementarity
- All contributors to the algorithm should be considered
- Performance targets are a function of:
  - Functional specifications
  - Solution Architecture
  - Hardware
  - Software
  - Cost

Resources

- A Comparative Study of SIFT and its Variants
- Brief: Binary robust independent elementary features
- BRISK: Binary Robust Invariant Scalable Keypoints
- Quick Start Guide :: NVIDIA Deep Learning TensorRT Documentation
- Rail Insider-Locomotive technology 2020: What's next is now
- Rail Insider-On-board locomotive monitoring technology (parts 1 & 2)
- Designing Machine Learning Systems