Transforming Existing NJDOT Traffic Cameras into Autonomous Congestion Detection Systems by Utilizing Video Vision-Language Models

Bhargavi Reddy Kotha, Abolfazl Afshari, Dejan Besenski, Ph.D., Joyoung Lee, Ph.D., Lazar Spasovic, Ph.D.

John A. Reif, Jr. Department of Civil & Environmental Engineering, New Jersey Institute of Technology

bk384@njit.edu, aa3377@njit.edu, besenski@njit.edu, jo.y.lee@njit.edu, spasovic@njit.edu

Introduction

- State DOTs run hundreds to thousands of CCTV cameras, but congestion is still **spotted by people** watching video walls.
- Current automated methods need **manual engineering** per camera: drawn zones, calibration, vehicle boxes.
- **Goal:** turn existing cameras into **automatic congestion sensors**, with no new hardware and no camera calibration.
- **Idea:** a **vision-language model** (Qwen2.5-VL-7B) watches a 30-s clip, answers **Low**/**High** per direction.
- Hand-labeling tens of thousands of clips is too costly, so labels come from **probe travel-time data** instead.

Methodology

- **Record:** 30-s clips, 3 NJDOT cameras, peak hours, 41 days.
- **Auto-label:** travel time above threshold = High (**51,344** clips).
- **Correct:** humans fix labels for ~2 weeks per site.
- **Compare 3 versions** of the model (Fig. 1):
  - **(1)** off-the-shelf Qwen2.5-VL-7B, no training
  - **(2)** trained on automatic labels, one model for all sites
  - **(3)** trained on corrected labels, one model per camera

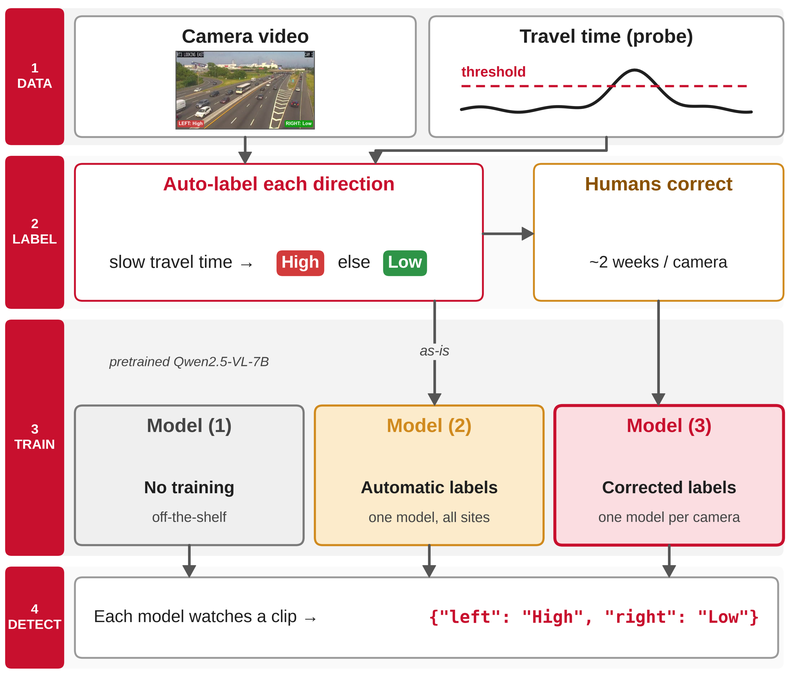


Fig. 1 Pipeline and the 3 compared models

Descriptive Analysis

- **3 sites:** NJ-3 @ NJ Turnpike ramp, I-80 @ Garden State Pkwy, I-80 @ NJ-20.
- Each view shows **2 directions**, each classified separately (Fig. 2).
- **Test:** 4 unseen days per site, **14,450** hand-checked labels.
- **Peak-hour** congestion builds mostly in **one direction**.

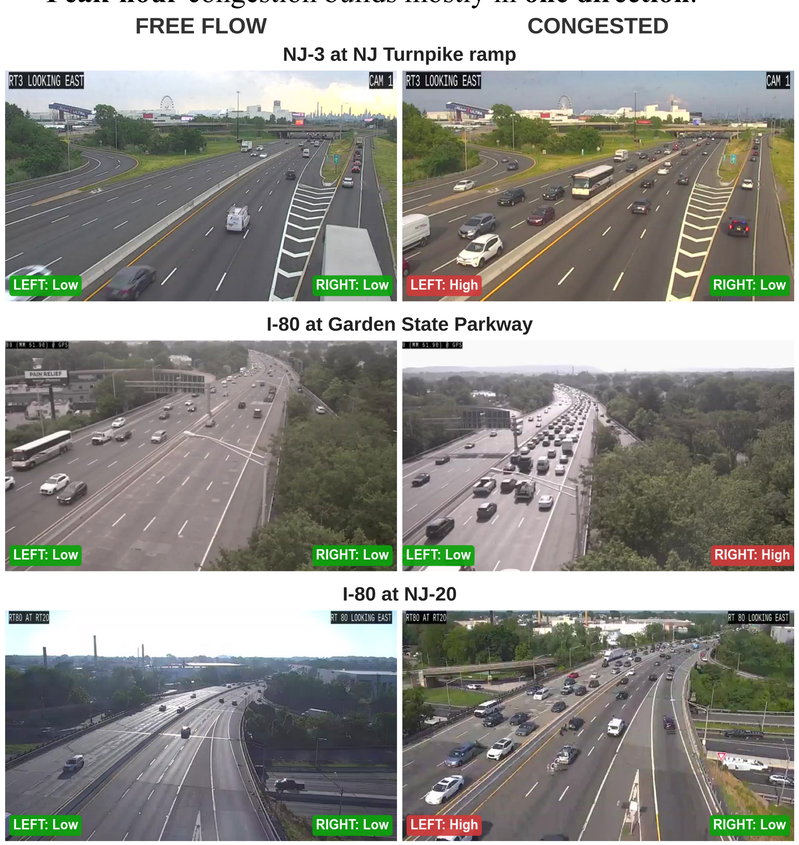


Fig. 2 Free flow vs. congested views

Model Result

Model	Accuracy	Precision	Recall
(1) No training	62%	50%	69%
(2) Automatic labels	78%	98%	54%
(3) Corrected labels	**90%**	**96%**	**83%**

Accuracy: both directions right · Precision: "High" calls correct · Recall: congestion caught

- Model (2) learned from noisy labels and became **too cautious**: few false alarms, but it **misses about half of real congestion**.

Model Result (cont'd)

90% accuracy

using cameras NJDOT already has, with no new hardware

- Model (1) works at GSP (**86%**) but **fails** at NJ-20 (**38%**).
- Model (3) has the **highest accuracy at every site**; NJ-20 rises to **81%** (Fig. 3a).
- Congested direction: model (3) **catches 91%** (NJ-3) and **79%** (NJ-20) vs 41% and 50% for model (2) (Fig. 3b).

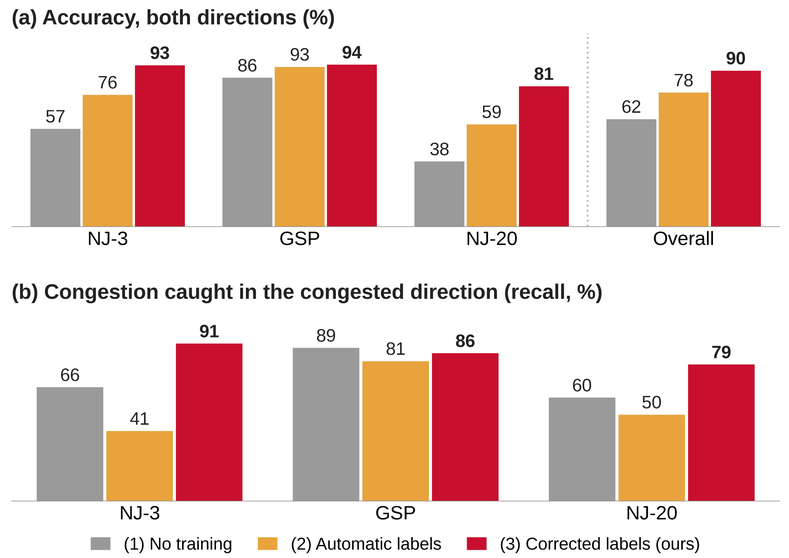


Fig. 3 Results on held-out test days

Conclusion & Recommendations

- **No new sensors:** existing cameras can spot congestion on their own.
- **Small, corrected data won:** ~2 weeks of corrected video per camera outperformed a 5–8× larger noisy set.
- **Lightweight per camera:** a small add-on to the shared model, trained in under a day on one GPU.
- **How to use it:** label automatically from travel-time data → fix the labels briefly → train one model per camera.
- Already running live on 3 NJDOT cameras. Next: more cameras, night and bad weather, more congestion levels.

- Funded by NJDOT ITS Resource Center, CY2025 (TO-119).