Multi-Resolution End-to-End Deep Neural Network for Optimizing Latency-Accuracy Tradeoff in Autonomous Driving

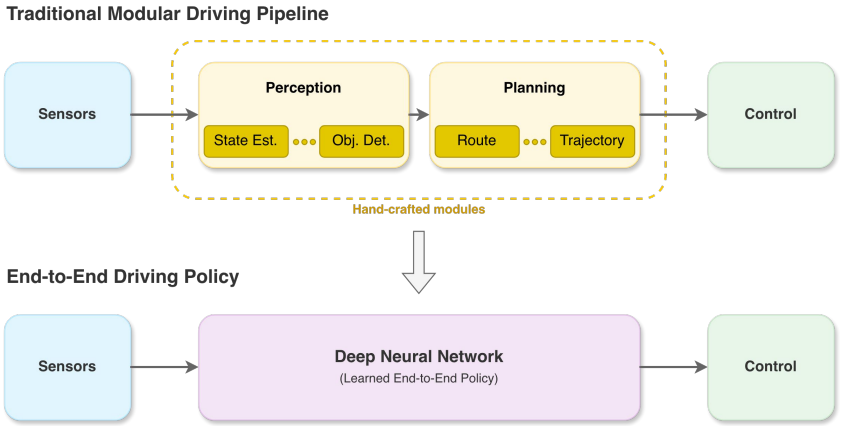
QiTao Weng, Heechul Yun
University of Kansas

KU THE UNIVERSITY OF KANSAS

KU-CSL COMPUTER SYSTEMS LAB

End-to-End DNN based Autonomous Driving

- Autonomous driving is a closed-loop cyber-physical system:
  - sense ⟶ compute ⟶ act ⟶ observe new state
- In end-to-end driving, a learned policy maps raw sensor input directly to vehicle control.
- Key CPS challenge:
  - *accurate enough* to choose the right action
  - *fast enough* to apply it in time

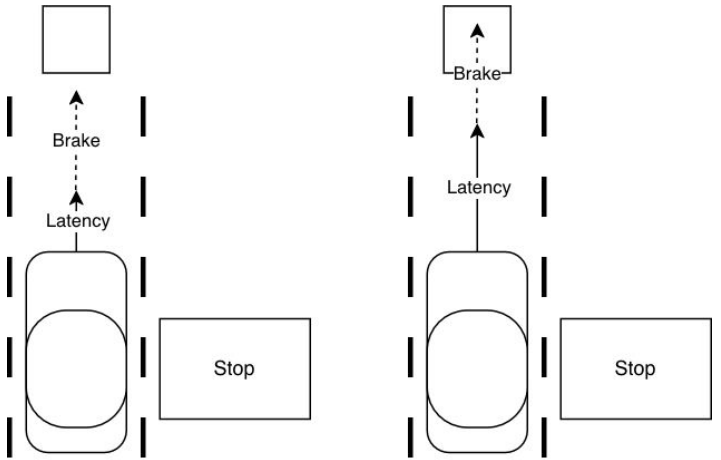


Importance of Time (Latency) in CPS

- Safety depends on both *accuracy* and *latency*
  - Because the vehicle moves during perception–control latency window

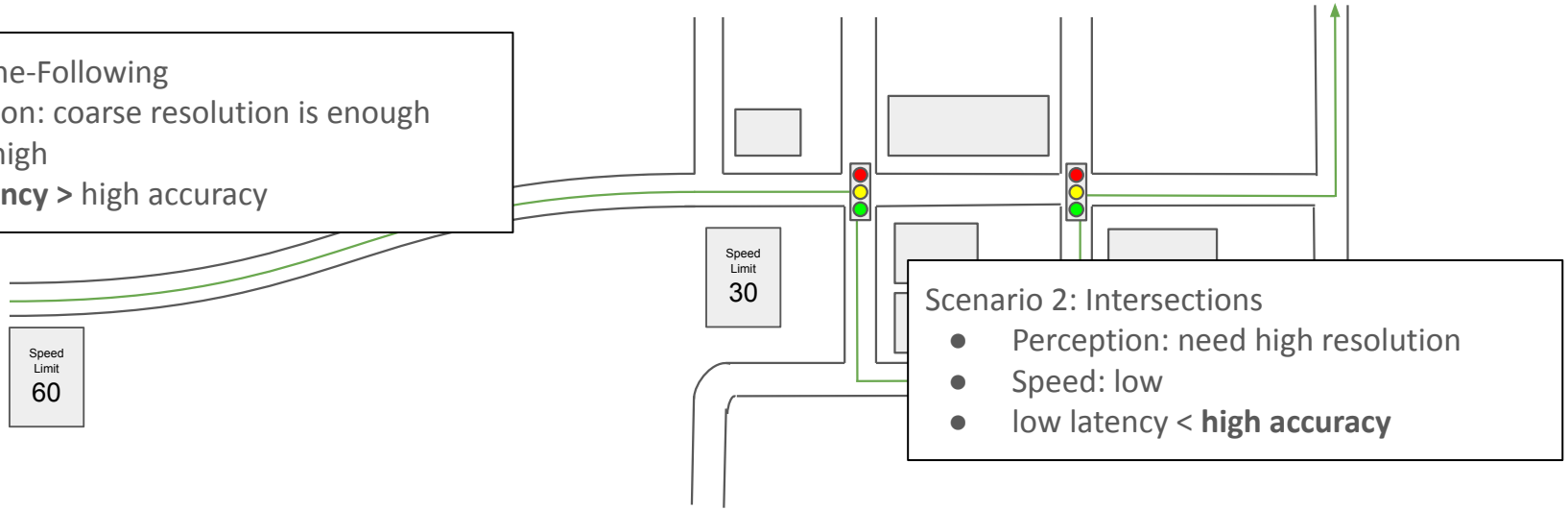
- E.g., Stopping Distance
  - Latency adds reaction time needed to stop when breaking

$$d_{\text{stop}} = \frac{v^2}{2a_{\max}} + vL$$



Environment Dependent *Accuracy-Latency* Tradeoffs

Scenario 1: Lane-Following
- Perception: coarse resolution is enough
- Speed: high
- **low latency >** high accuracy



Scenario 2: Intersections
- Perception: need high resolution
- Speed: low
- low latency < **high accuracy**

The optimal balance between latency and accuracy is not constant; it depends on the driving context. This motivates a system that can *adapt the environment at runtime*

KU THE UNIVERSITY OF KANSAS

Outline

- Motivation
- **Our Approach**
- Evaluation
- Conclusion

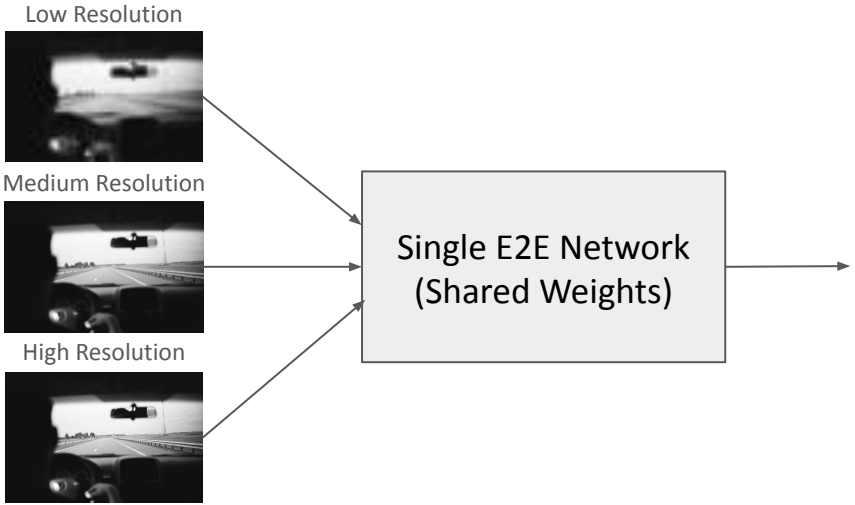
Goal

A **single model** that can **adapt** to different environmental constraints **at runtime**

Our Approach: *Anytime* E2E Driving

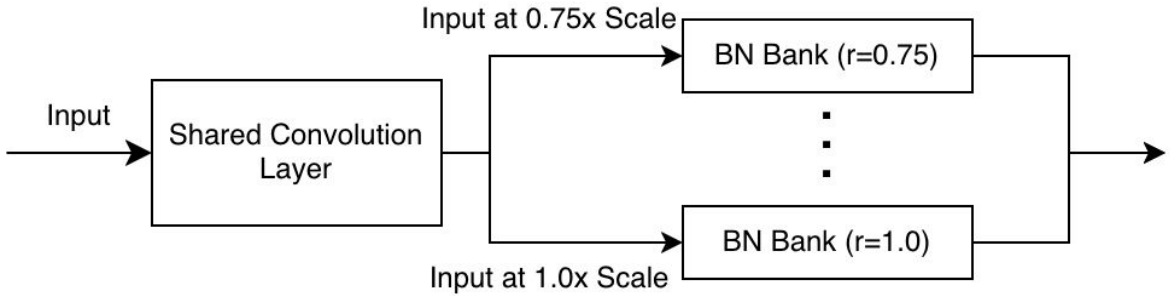
Resolution as the *knob* – more detail costs latency; less detail risks missing cues like traffic lights

- **Multiple Resolution:**
  runtime selection based on driving context
- **Single Model:**
  efficient, no multiple models in memory
- **How do we do this?**



Resolution-Aware Batch Normalization

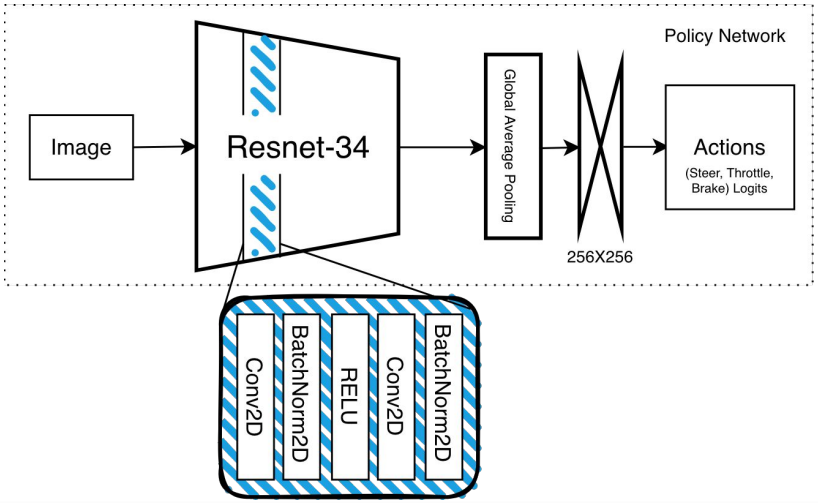
- Shared Convolution Weights: the feature extraction layers are shared across resolutions. This maintains efficiency and model size.
- Replicated Batch Normalization (BN) Banks: for each resolution (r), we create a separate set of parameters (γ, β) and statistics (μ, σ)



Mingjian Zhu, Kai Han, Enhua Wu, Qiulin Zhang, Ying Nie, Zhenzhong Lan, and Yunhe Wang. 2021. Dynamic resolution network. Advances in Neural Information Processing Systems 34 (2021).
Yikai Wang, Fuchun Sun, Duo Li, and Anbang Yao. 2020. Resolution switchable networks for runtime efficient image recognition. In ECCV. Springer.

Baseline E2E Driving Model

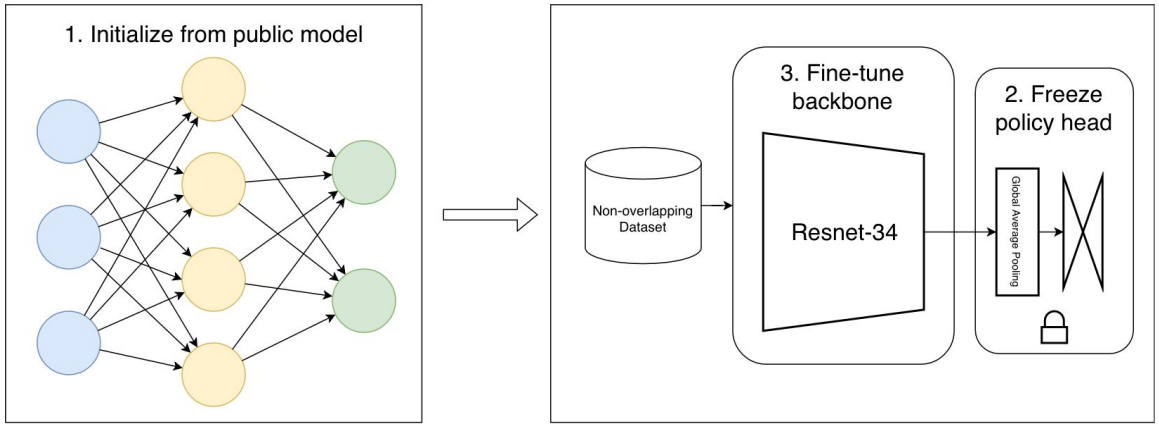
- We adopt a monocular camera–based E2E controller from *Learning to Drive from a World on Rails*[1]
  - Modify static model to support multiple resolutions



[1] Dian Chen, Vladlen Koltun, and Philipp Krähenbühl. 2021. Learning to drive from a world on rails. In ICCV. IEEE.

Resolution Retargeting

Challenge: can this be done without full retraining? without original dataset?



Segmentation (from dataset)

Constrain post GAP to frozen model

$$L_{\text{retarget}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left(L_{\text{act}}^{(r)} + \lambda_{\text{seg}} L_{\text{seg}}^{(r)} + \lambda_{\text{KD}} L_{\text{KD}}^{(r)} + \lambda_{\text{feat}} L_{\text{feat}}^{(r)} \right),$$

Ground truth (from dataset)

Constrain output to frozen model

Outline

- Motivation
- Our Approach
- **Evaluation**
- Conclusion

Evaluation

- Simulator: CARLA, synchronous mode at 40Hz
- Latency conditions: controller period and injected delay varied across {50, 100, 150, 200} ms via zero-order hold
- Scenarios:
  - 2 towns: Town01 (seen during training), Town02 (unseen)
  - 3 traffic densities: (0, 0), (5, 10), (20, 40) vehicles/pedestrians
  - 12 runs per timing configuration

Effects of Resolution Retargeting

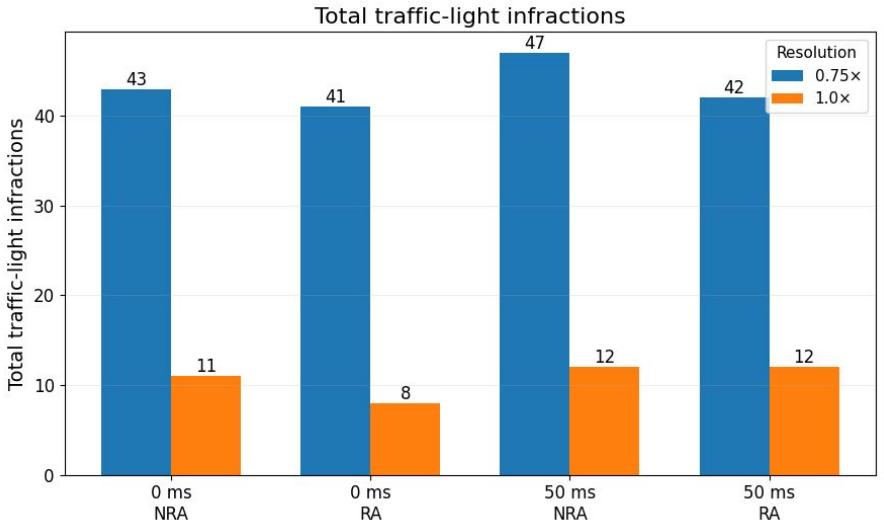
- Performance of NRA (baseline) and RA (multi-resolution) are comparable on both towns.
- Only a small increase in minor, recoverable collisions on the unseen test town.

Map	Original	NRA 0.75×	NRA 1.0×	RA 0.75×	RA 1.0×
Success (%)					
Town01	100	100	100	100	100
Town02	100	100	100	100	100
Collision incidence					
Town01	0	0	0	0	0
Town02	0	1	1	1	2

One model

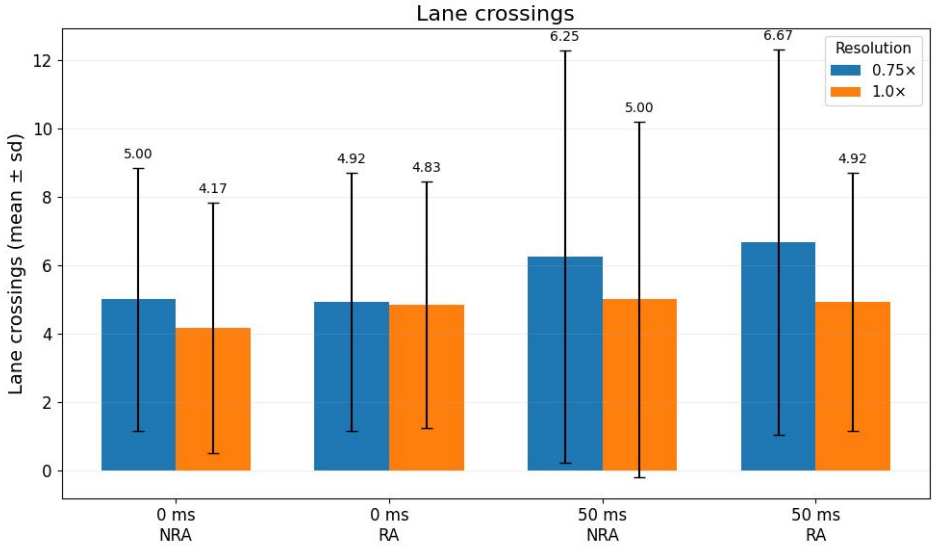
Effects of Input Resolution

- At the same latency, lower-resolution inputs cause a large increase in traffic-light infractions

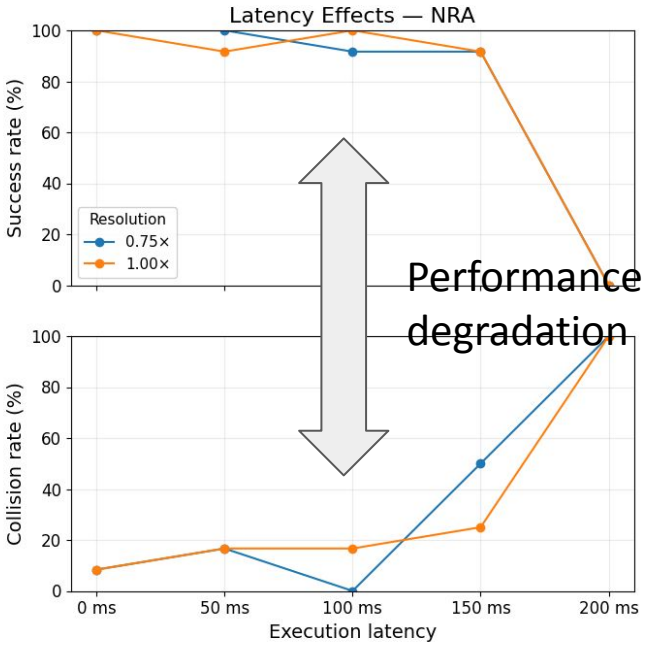


Effects of Input Resolution

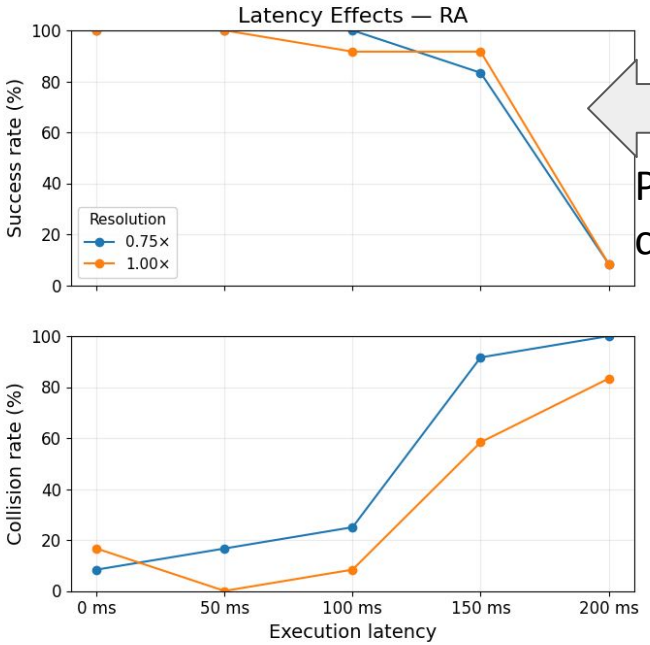
- Similar lane invasions across all models.



Effects of Latency



Performance degradation



Performance cliff

Effects of Dynamic Resolution Switching

- Proximity trigger
  - Use high resolution near traffic lights
  - Low resolution elsewhere
- Assumption
  - Proximity to lights is known (e.g., from map information)

Latency envelope $L_{high} \rightarrow L_{low}$	Configuration	Runs	Success (%)	Collision (%)	Mean lane crossings	Red lights (total)
	Fixed 0.75× @ 50 ms	12	100.00	16.65	6.67 ± 5.63	42
100 → 50 ms	**RA switcher (50/100 ms)**	**12**	**100.00**	**0.00**	5.50 ± 5.21	15
	Fixed 1.0× @ 100 ms	12	91.65	8.35	9.84 ± 9.33	**14**
	Fixed 0.75× @ 100 ms	12	100.00	25.00	11.25 ± 9.14	43
150 → 100 ms	**RA switcher (100/150 ms)**	**12**	**100.00**	**0.00**	**10.84 ± 9.10**	**17**
	Fixed 1.0× @ 150 ms	12	91.65	58.35	66.50 ± 76.01	18

KU THE UNIVERSITY OF KANSAS

KU THE UNIVERSITY OF KANSAS

Outline

- Motivation
- Our Approach
- Evaluation
- **Conclusion**

Conclusion

- A static, fixed resolution E2E model is suboptimal in dynamic environments
  - Low-resolution: fast, but miss small object cues and details.
  - High-resolution: accurate, but too slow to react at high speeds.
- Our approach: Anytime E2E model for runtime adaptation
  - Per-resolution batch normalization to enable multi-resolution support with shared weights
  - Resolution retargeting to enable multi-resolution support with limited training data
  - A traffic-light proximity based resolution selection policy
- CARLA based evaluation in urban-driving scenarios
  - Demonstrated that our anytime E2E DNN approach improves the *latency-safety frontier* compared to fixed resolution baselines across driving performance metrics
- Future work
  - Better resolution selection policy (e.g., learning based algorithms)
  - Real-world evaluation in actual systems

KU THE UNIVERSITY OF KANSAS

Questions?