Learning Robust Multimodal Control for Resource-Constrained Platforms

Many robotic platforms already have limited computational resources and strict latency constraints. Therefore, the typical approach of scaling up model size is often impractical for these systems. Instead, we ask: *can we deploy ML-based controllers that require minimal training and computation?* Prior end-to-end work mapping RGB to actions ([1, 7, 5, 6]) reports strong results on datasets and static deployment environments, but reliance on RGB alone limits robustness. We show that multimodality consistently guides model attention and remedies failures of unimodal controllers. In our experiments, augmenting the same controllers used in [5] with depth information consistently corrects failure modes of RGB-only agents, as depth data guides model attention toward distant horizons and away from obstacles. Moreover, this suggests that some attention cues and performance differences previously attributed to model architecture can be substantially mitigated by enhanced perception.

Approach. We implement a modular pipeline with a multimodal fusion module and a small recurrent controller. We experiment with:

- Fusion techniques: *early fusion* (depth as 4th channel); *late fusion* (separate RGB and depth encoders); and *depth-adaptive* fusion using deformable convolutions guided by depth.
- Recurrent controllers: one-layer RNN for sequential decision making using either LSTM or biologically inspired architectures LTC[4], CfC[3], LRC[2].
- Minimal training for deployment: learning from a small set of expert demonstrations and controlling a small-scale electric vehicle.

Results. We use pipe-lined circuits to assess navigation autonomy and robustness under realistic sensor perturbations (Gaussian noise with variance up to $\sigma^2 = 0.3$, intermittent frame-rate loss). Our main findings highlight:

- **Perception matters.** Unimodal agents fail to react in sudden turns; adding depth improves obstacle anticipation, achieving 100% autonomy. Depth-adaptive fusion improves obstacle sensitivity and often produces steering trajectories closer to the human expert, but incurs higher latency.
- **Early fusion is best.** Under high noise, early data fusion retains full autonomy and stable behavior.
- **Bio-inspired controllers achieve faster inference.** The biologically-inspired controllers LRC and CfC frequently match LSTM robustness while reducing inference time in several configurations.
- **Lighter models are better for deployment.** Some models (e.g., late fusion CfC) that are not top-ranked by dataset results nevertheless succeed on hardware, highlighting that offline losses do not fully predict deployment performance.

Contributions. This work makes the following contributions:

1. An empirical demonstration that multimodality improves navigation, correcting where unimodality fails.
2. A systematic comparison of sensor fusion strategies under strict computational and latency constraints.
3. Results showing that perception dominates performance over architecture choice, while the choice of controller remains important for inference latency and agent behavior.

Conclusion. For resource-constrained platforms, robust autonomy follows from smart model design that fuses the sensors already available on robots. In our navigation tests, RGB–D fusion matters more than the choice of recurrent controller: numerical metrics, attention maps, and video analysis of vehicle trajectory all show that depth consistently guides controllers toward the track horizon and away from obstacles.

[1] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016.

[2] Mónika Farsang, Sophie A Neubauer, and Radu Grosu. Liquid resistance liquid capacitance networks. *arXiv preprint arXiv:2403.08791*, 2024.

[3] Ramin Hasani, Mathias Lechner, Alexander Amini, Lucas Liebenwein, Aaron Ray, Max Tschaikowski, Gerald Teschl, and Daniela Rus. Closed-form continuous-time neural networks. *Nature Machine Intelligence*, 4(11):992–1003, November 2022.

[4] Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. Liquid time-constant networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7657–7666, 2021.

[5] Mathias Lechner, Ramin Hasani, Alexander Amini, Thomas A. Henzinger, Daniela Rus, and Radu Grosu. Neural circuit policies enabling auditable autonomy. *Nature Machine Intelligence*, 2(10):642–652, October 2020.

[6] Fouad Makiyeh, Mark Bastourous, Anass Bairouk, Wei Xiao, Mirjana Maras, Tsun-Hsuan Wangb, Marc Blanchon, Ramin Hasani, Patrick Chareyre, and Daniela Rus. Optical flow matters: an empirical comparative study on fusing monocular extracted modalities for better steering, 2024.

[7] Yanqiu Zhang, Ruiquan Ge, Lei Lyu, Jinling Zhang, Chen Lyu, and Xiaojuan Yang. A virtual end-to-end learning system for robot navigation based on temporal dependencies. *IEEE Access*, 8:134111–134123, 2020.