Simplified Multimodal Imitation Learning for Day–Night Autonomous Driving

Aspen Knox, Mohammadparsa Ghasemi
Autonomous Systems Laboratory, California State Polytechnic University, Pomona

***Abstract*— This study investigates a simplified, conditioned Imitation Learning (IL) model for day-night autonomous driving on a university campus. An initial proof of concept was tested using an end-to-end CNN model, inspired by NVIDIA's approach, which controlled only steering and throttle, performing low-level commands. The model used RGB cameras for daytime driving and a thermal camera for nighttime operation. The test involved a 0.3-mile route from the College of Engineering to the College of Business. Although the trained model for daytime driving achieved 99% accuracy, the model performed worse than its counterpart trained on thermal vision, with a lower accuracy of 94-97%. A total of 3 interventions were made during the nighttime testing. The daytime model failed to perform autonomously as the go-kart kept oversteering toward the bright patches of grass or sidewalks. This failure could be explained by the imbalance in training data, as the go-kart was tested later in the day, diverging from the collected data. Future work will incorporate High-Level Commands, teaching the model to execute specific turns based on input from a path-planning algorithm.**

***Keywords*— Imitation Learning, Autonomous Navigation, Thermal Vision**

I. Introduction

Traditional modular approaches to full autonomy rely on numerous interconnected components requiring extensive computation and complex data processing. These include perception, full SLAM, path planning, and object avoidance subsystems. While effective for high-speed or urban driving, such architectures can be excessive for smaller, slower vehicles operating in structured domains such as university campuses. To investigate a simpler alternative, we explore end-to-end Imitation Learning (IL) to directly map visual inputs to control outputs, bypassing explicit kinematic modeling. End-to-end learning for autonomous driving was pioneered by Bojarski et al. [1], who demonstrated that a convolutional neural network could learn to map raw pixels directly to steering commands. This study focuses on IL for low-level control, with future extensions planned toward conditioned IL [2], enabling both high- and low-level decision-making through the integration of high-level navigational commands. We adopt a CNN-based approach inspired by PilotNet [1][5], demonstrating that pixels can be mapped directly to steering using a compact network.

II. Methodology

Data collection was conducted using a go-kart–based autonomous platform along a fixed 0.3-mile route between the College of Engineering and the College of Business under both daytime and nighttime conditions to enable multimodal training using RGB and thermal imagery. For daytime experiments, two synchronized RGB cameras captured visual data, while vehicle telemetry: including steering angle, throttle, acceleration, and braking was logged through a CAN bus at 20 Hz. The vehicle traversed the route five times in each direction. For nighttime operation, a thermal camera recorded the same route once in each direction. Vehicle commands were sent from a remote keyboard to an onboard laptop (Intel Core i7 CPU, RTX 4090 GPU), which transmitted data to a master CAN node controlling the stepper motor (steering), linear actuator (brakes), and sine-wave motor controller for the three-phase brushless DC motor. The go-kart weighs approximately 250 lb., employs rear-wheel drive with Ackermann steering, and was limited to 5 mph for safety. The RGB cameras operated at 640 × 480 @ 30 Hz, and the thermal at 120 × 96 @ 30 Hz. Cameras were mounted at the front-center of the chassis and timestamp-synchronized with control commands.

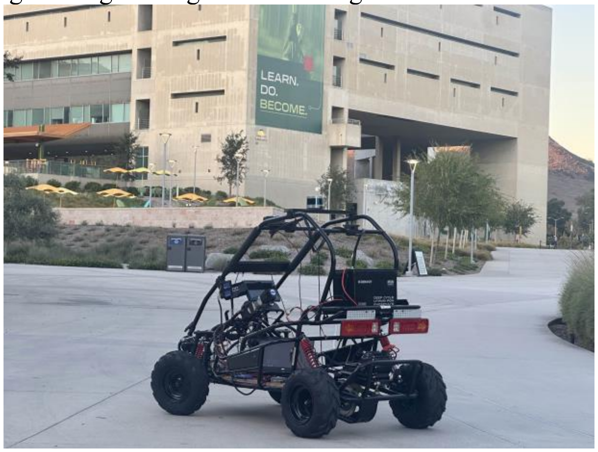

Figure 0. Autonomous go-kart platform with multimodal vision sensors (RGB and thermal cameras) used for day-night imitation learning experiments.

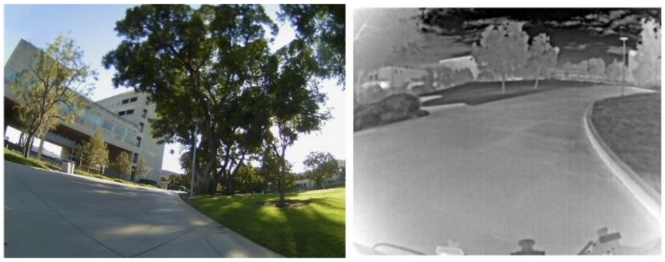

FIGURE 1. *FEATURING RGB AND THERMAL IMAGES.*

III. MODEL ARCHITECTURE

The initial proof-of-concept model processed visual and motion data without high-level commands. The full architecture extends this by incorporating high-level command conditioning. For daytime driving, each image is converted from RGB to YCbCr, normalized (mean=0.5, std=0.5), and augmented with random cropping and rotation before being passed through five convolutional layers for feature extraction [3, 6]. The speed scalar is processed through a small MLP, and high-level commands (0, 1, 2 for right, left, straight) are processed through an embedding layer. All three embeddings are concatenated and passed through dense layers to fuse visual, motion, and intent cues. Steering, brake, and throttle each have a dedicated MLP head, with predictions fine-tuned by linear layers. All models were trained for 50 epochs.

IV. DISCUSSION OF RESULTS

The daytime model achieved high training accuracy, approximately 99% across steering, throttle, acceleration, and braking, (fig.3) but failed to generalize during autonomous testing. When evaluated later in the day under altered lighting, the model overfitted strongly, oversteering toward bright grass patches and following illumination rather than the road surface. This behavior exemplifies a fundamental limitation of behavior cloning: models may exploit spurious correlations in the training data rather than learning driving policies [3]. The sensitivity to lighting variations highlights the challenge of dataset shift in vision-based imitation learning systems [4]. The thermal model achieved slightly lower training accuracy (97% throttle, 95% braking, 94% acceleration), (fig.2) yet exhibited greater control. The vehicle completed the 0.3-mile route with only three manual interventions, one of which occurred at a fork where the system hesitated between two pathways. Thermal imagery proved inherently more stable, as temperature gradients remain constant across time and weather. The results align with findings in urban autonomous driving where environmental invariance is critical for deployment [4]. All data were recorded on a clear, dry day under sunny conditions with no pedestrian activity along the route. Thermal imaging has shown promise in autonomous navigation due to its invariance to lighting conditions and ability to detect heat signatures in low-visibility scenarios [7].

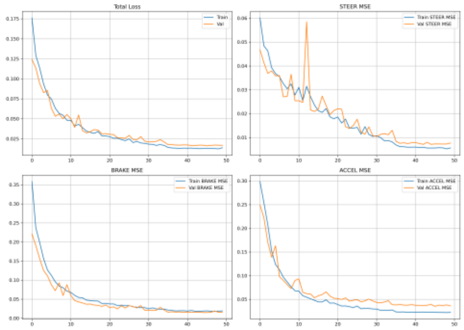

FIGURE 2. *TRAINING AND VALIDATION LOSS CURVES FOR THE THERMAL CAMERA IMITATION LEARNING MODEL.*

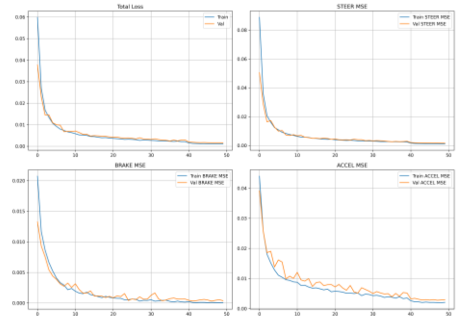

FIGURE 3. *TRAINING AND VALIDATION LOSS CURVES FOR THE DUAL RGB CAMERA IMITATION LEARNING MODEL.*

V. CONCLUSION

This study demonstrates that end-to-end imitation learning can effectively drive a small autonomous platform using thermal perception, despite limited data. The comparison between RGB and thermal modalities highlights the critical sensitivity of visible-light models to illumination changes, while thermal imagery offers a stable perceptual foundation for control. This work validates thermal-based IL as a lightweight solution for autonomous navigation in structured environments and provides a foundation for future architectures incorporating conditioned IL for high-level planning. Following conditional imitation learning frameworks [2], future work will integrate high-level navigation commands (e.g., turn left, turn right, go straight) to enable goal-directed behavior. Improved data diversity and domain randomization will be essential for scaling to more complex environments [3].

VI. ACKNOWLEDGMENT

This work was supported in part by the Ganpat and Manju Center for International Collaboration and the Department of Education under the Industry 4.0: Career Advancement through Research and Education in Modeling and Simulation (iCARE-M&S) program (MSP Absolute Priority Two). We extend our sincere gratitude to our research advisor, Dr. Behnam Bahr, for his invaluable guidance, support, and mentorship throughout this project. We are also deeply thankful to Adam Sorenson for his essential assistance with the CAN bus integration and Johann W. Fernandez for his insights on the IL architecture.

REFERENCES

[1] M. Bojarski et al., "End to End Learning for Self-Driving Cars," arXiv:1604.07316, 2016. doi:10.48550/arXiv.1604.07316.

[2] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end Driving via Conditional Imitation Learning," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), Brisbane, Australia, 2018, pp. 4693-4700. doi:10.1109/ICRA.2018.8460487.

[3] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the Limitations of Behavior Cloning for Autonomous Driving," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, 2019, pp. 9329-9338. doi:10.1109/ICCV.2019.00942.

[4] J. Hawke, R. Shen, C. Gurau, S. Sharma, D. Reda, N. Nikolov, P. Mazur, S. Micklethwaite, N. Griffiths, A. Shah, and A. Kendall, "Urban Driving with Conditional Imitation Learning," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), Paris, France, 2020, pp. 251-257. doi:10.1109/ICRA40945.2020.9196844.

[5] M. Bojarski et al., "The NVIDIA PilotNet Experiments," arXiv:2010.08776, 2020. doi:10.48550/arXiv.2010.08776.

[6] C. Chen et al., "Learning by Cheating," in Proc. Conf. Robot Learn. (CoRL), 2020, pp. 66-75.

[7] A. Rathinam et al., "Autonomous Navigation using Thermal Imaging for Mobile Robots in Low-Light Conditions," in Proc. IEEE Int. Conf. Intell. Robots Syst. (IROS), 2019, pp. 3456-3462.