Research Statement

Introduction

Over the past few years, my research has focused on advancing autonomous driving through innovative algorithms that integrate perception and planning. My work spans from early reinforcement learning explorations to recent breakthroughs in end-to-end autonomous systems, including the Perception Helps Planning (PHP) framework and the Lane-level camera-LiDAR Fusion Planning (LFP) method. Recently, I have been exploring Vision-Language Models (VLMs) to enhance perception and planning for autonomous systems.

Core Contributions: End-to-End Autonomous Driving

Perception Helps Planning (PHP)

The PHP framework represents a significant step forward in integrating perception and planning for autonomous driving. Traditional end-to-end planning methods often overlook critical traffic elements such as lanes, intersections, and dynamic agents, leading to inefficiencies and non-compliance with traffic regulations. To address this, I proposed PHP, a novel framework that reconciles lane-level planning with perception. PHP focuses on both edges of a lane, considering their positions in Bird's Eye View (BEV) along with attributes related to lane intersections, directions, and occupancy.

The PHP framework begins with a transformer encoding multi-camera images to extract lane-level features. A hierarchical feature early fusion module refines these features for predicting planning attributes. Finally, a late-fusion process integrates lane-level perception and planning information to generate control signals. Experiments on the CARLA demonstrated significant improvements in driving scores, achieving state-of-the-art performance. Additionally, the system was successfully deployed on real vehicles, showcasing its robustness and high frame rate in real-world scenarios.

Lane-level camera-LiDAR Fusion Planning (LFP)

Building on the PHP framework, I developed the LFP method to enhance the efficiency and accuracy of end-to-end autonomous driving through multi-modal fusion. LFP targets driving-relevant lane elements, reducing the volume of LiDAR features while preserving critical information. This approach enhances interaction at the lane level between the image and LiDAR branches, allowing for the extraction and fusion of their respective advantageous features.

LFP introduces three novel modules: an image-guided coarse lane prior generation module, a lane-level LiDAR feature extraction module, and a lane-level cross-modal query integration and feature enhancement module. These modules work together to balance efficiency with performance, using lanes as the unit for sensor fusion. Specifically, the image-guided coarse lane prior generation module extracts lane-level semantic and geometric priors from the image branch, which are then utilized by the lane-level LiDAR feature extraction module to guide sparse sampling of LiDAR pillars, focusing on critical lane regions to capture essential depth information. The lane-level cross-modal query integration and feature enhancement module then facilitate interaction between the image and LiDAR branches at the lane level, enabling the complementary integration of semantic richness from the image branch and depth accuracy from the LiDAR branch. This synergistic interaction enhances the overall perception and planning capabilities of the system. Experiments on the CARLA demonstrated that LFP introduces critical depth information with only a minimal increase in computational latency compared to PHP, while also achieving robust performance in real-world vehicle deployments.

Foundational Work: Reinforcement Learning for Autonomous Driving

Reinforcement Learning for End-to-End Planning

In my early research, I focused on developing end-to-end planning algorithms for autonomous driving using reinforcement learning. I proposed a reinforcement learning framework that addresses challenges such as sparse feedback and high-dimensional state spaces. This framework introduced self-play collaborative training and parameterized path-planning techniques, enabling vehicles to learn from simulated interactions and generate optimized trajectories.

Domain Adaptation in Reinforcement Learning

To bridge the gap between simulated and real-world, I developed a novel prompt-based model transfer algorithm. This approach improved the generalization of reinforcement learning models trained in simulation to dynamic real-world traffic conditions. The algorithm, known as Prompt-based Proximal Policy Optimization (P^3O), leverages pre-trained models to achieve similar performance in target environments with different visual inputs. Experiments on the OpenAI CarRacing video game and CARLA demonstrated that P^3O achieves superior performance, significantly improving training efficiency.

Emerging Directions: Vision-Language Models for Autonomous Systems

VLM-based Robotic Grasping Perception

Recently, I have been involved in exploring the application of Vision-Language Models (VLMs) in robotic grasping tasks (GraspCoT). As part of this research, I contributed to the development of a framework for robotic arm grasping perception, which leverages VLMs combined with RGB-D data. This framework utilizes VLMs to reason about object properties and infer optimal grasping actions.

Extending VLM Reasoning to End-to-End Planning

Building on this work, I am currently exploring the integration of Vision-Language Models (VLMs) into end-to-end planning systems. My approach leverages VLMs as a high-level reasoning module to complement LFP. The VLM branch processes multi-modal inputs (e.g., camera, LiDAR, and textual commands) to generate long-term driving strategies, such as lane-changing decisions and route optimization. These strategies are combined with real-time planning outputs to enhance adaptability and interpretability, paving the way for more intelligent and human-like decision-making in autonomous driving. In this framework, VLM and LFP operate as parallel branches, enabling synchronous reasoning while maintaining computational efficiency. This design ensures that the system benefits from the advanced reasoning capabilities of VLMs without compromising the real-time performance of LFP.

Conclusion

My research has made significant contributions to the field of autonomous driving, from early explorations in reinforcement learning to the development of advanced end-to-end planning frameworks like PHP and LFP. These innovations have improved the efficiency, accuracy, and safety of autonomous systems. Recently, I have extended my work to explore Vision-Language Models (VLMs) for robotic manipulation and autonomous driving, leveraging their reasoning capabilities to enhance adaptability and interpretability. I look forward to continuing my research and contributing to the ongoing development of intelligent and human-centric autonomous technologies.

Related Publications:

1. *Perception Helps Planning: Facilitating Multi-Stage Lane-Level Integration via Double-Edge Structures* (PHP)
2. *LFP: Efficient and Accurate End-to-End Lane-Level Planning via Camera-LiDAR Fusion* (LFP)
3. *P^3O: Transferring Visual Representations for Reinforcement Learning via Prompting* (P^3O)
4. *Collision Avoidance for An Ackermann-Steering Vehicle via Map-Based Deep Reinforcement Learning*
5. *GraspCoT: Integrating Physical Property Reasoning for 6-DoF Grasping under Flexible Language Instructions*
6. *VLMPlanner: Integrating Visual Language Models with Motion Planning*