Theory and Practice of Science and Technology
2025, VOL. 6, NO. 6, 69-74
DOI: 10.47297/taposatWSP2633-456914.20250606

Research on Perception and Decision-Making Methods for Autonomous Driving

Qinwen Qiu

School of Management and engineering, Nanjing University, Industrial Intelligence and System Integration, Nanjing, Jiangsu, 210000, China

ABSTRACT
Addressing the issues of insufficient perception robustness and limited decision-making real-time performance in autonomous driving systems under complex conditions (such as extreme weather and dynamic interactive environments), this review focuses on the synergistic optimization of multi-modal perception fusion and deep reinforcement learning (DRL) technologies. First, we review the development of multimodal perception fusion technologies, including perception mechanisms, feature alignment, BEV mapping, and dynamic weighting mechanisms. We then analyze typical models, policy optimization, curriculum learning mechanisms, and practical deployment issues of deep reinforcement learning in autonomous driving decision-making. Finally, we systematically summarize the research progress of perception and decision-making collaborative systems. This review summarizes current mainstream methods, existing issues, and future development directions, providing a reference for research on intelligent autonomous driving systems.

KEYWORDS
Autonomous driving; Multimodal perception; Deep reinforcement learning; Sensor fusion; Feature alignment

1 Introduction

As technology continues to evolve, robotics technology is constantly being updated, with applications in industries such as manufacturing, services, and healthcare becoming increasingly widespread. The variety of scenarios requiring robotic solutions is growing, and the environments in which robots operate are becoming increasingly complex. This places higher demands on a robot's perception and decision-making capabilities. The requirements for perception and decision-making capabilities are particularly stringent in autonomous driving, and the demand for such capabilities in industries such as manufacturing and services is also growing steadily. Perception and decision-making in autonomous driving, as one of the key technologies for achieving autonomous driving, is closely tied to the backdrop of technological advancement, reflecting the modern demand for intelligent and modern solutions. Despite significant progress in artificial intelligence, sensors, and decision-making algorithms, robots still need to improve their perception and decision-making capabilities in more complex environments. This phenomenon is particularly evident in autonomous driving, where vehicles may operate on various road segments with differing road conditions, imposing even stricter demands on sensing and decision-making capabilities. Therefore, this study aims to improve technology to enhance road sensing and decision-making capabilities in autonomous driving, enabling it to handle various complex road conditions. Finally, by enhancing the safety and reliability of autonomous driving through research, the technology can be applied in more fields.

Currently, research on autonomous vehicle perception and decision-making focuses on the continuous improvement of sensor technology and algorithm optimization. However, complex road conditions often reduce the perception and decision-making speed of autonomous vehicles. Existing research still has significant limitations regarding the perception capabilities of sensors, the interactivity of decision-making, and real-time performance issues. Extreme weather conditions such as heavy rainfall and smog can reduce visibility, thereby affecting the perception capabilities of sensors. Zhang and Singh noted in their research on lidar and visual fusion technology that existing sensor systems perform poorly in complex scenarios such as extreme weather and changes in lighting conditions, particularly under adverse weather conditions, where lidar and camera data acquisition can be severely disrupted. In terms of decision-making, the interaction between autonomous vehicles and other traffic participants (such as pedestrians, cyclists, and other vehicles) remains a major challenge. Lefèvre et al. noted that in complex traffic scenarios, there is uncertainty in predicting the behavior of other participants, especially in emergency situations, where existing decision-making models cannot effectively handle such uncertainty. Meanwhile, while deep learning algorithms have played a significant role in perception and decision-making for autonomous driving, their high computational resource requirements conflict with the real-time demands of autonomous driving. Kendalp et al. demonstrated the potential of deep convolutional networks for real-time pose estimation in their PoseNet research, but also noted that computational complexity severely impacts the system's real-time performance.

Autonomous driving perception and decision-making still face the challenge of insufficient sensor perception and

decision-making capabilities. Therefore, this research aims to improve the perception capabilities of sensors as well as the decision-making capabilities and speed of autonomous driving systems. This will address the limitations in perception and decision-making capabilities of autonomous vehicles in complex road environments, enhance the autonomy, adaptability, and decision-making accuracy of autonomous vehicles in such environments; Achieve the simultaneous integration of multi-modal perception mapping and deep reinforcement learning decision-making mechanisms within autonomous driving systems, breaking through the current bottleneck of separated perception and decision-making systems to realize closed-loop collaborative optimization between perception and decision-making; innovatively apply cross-modal Transformer structures to feature fusion and combine Curriculum Learning to guide the stable convergence of deep reinforcement learning models, thereby alleviating training instability in high-dimensional spaces and enhancing the system's ability to adapt to dynamic scenarios.

2 Overview of multimodal perception fusion technology

2.1 Sensors

Autonomous driving perception systems use a variety of sensors to collect information about the surrounding environment. Typically, RGB cameras are used to collect information such as texture, color, and semantics; LiDAR (light detection and ranging) is used to provide high-precision distance and spatial structure information; IMUs (inertial measurement units) are used to output acceleration and angular velocity for auxiliary motion estimation; and GNSS is used for positioning and time synchronization assistance.

2.2 Fusion strategies

Based on the characteristics of the collected information and data, the following three different fusion strategies can be used:

Early fusion: Directly concatenate raw data, mostly used for image and point cloud concatenation;

Mid-fusion: First extract modal features, then fuse them in the intermediate network layer. Mainstream methods include cross-modal Transformers, attention fusion, etc.;

Late fusion: Each modality predicts independently, and the results are then fused for decision-making, suitable for redundant mechanisms.

2.3 Fusion models

Depending on the selected fusion strategy, we can choose different fusion models. For example, Zhang et al. proposed VoxelNet, which voxelizes point clouds, encodes features of points within non-empty voxels, learns and aggregates local features, and then performs convolution and output; Project multi-view image features into the BEV space using LSS (Lift, Splat, Shoot) or BEVFormer-style Transformers, extract point cloud BEV features using a VoxelNet variant, and then fuse image BEV features with point cloud BEV features using a bidirectional cross-attention mechanism; introduce a Transformer perception path to achieve deep fusion.

2.4 Challenges in fusion

For multimodal perception fusion technology, the challenges typically include the following aspects: First, the spatio-temporal parameters collected by sensors are mismatched with spatial transformations (SE3), leading to fusion misalignment, which can be addressed through dynamic compensation and online calibration; second, the perception capabilities of sensors degrade under rainy/foggy or nighttime conditions, making accurate perception impossible; finally, implementing perception fusion for autonomous driving requires significant computational resources, posing challenges for real-vehicle deployment.

3 A review of the application of deep reinforcement learning in autonomous driving decision-making

3.1 Types of decision-making models

Deep reinforcement learning (DRL) learns optimal decision strategies by maximizing cumulative rewards through continuous interaction between the agent and the environment. Its core breakthrough lies in the introduction of deep neural networks as function approximators, enabling direct handling of high-dimensional state inputs that traditional reinforcement learning struggles with. Commonly used strategies include: Using Deep Q-Networks (DQN) to estimate

discrete action policies via neural networks; Using Deep Deterministic Policy Gradient (DDPG) and Twin Delayed DDPG (TD3) to address continuous action space control problems; Using the Actor-Critic framework to decouple policy optimization and value assessment, significantly improving the training stability and sample efficiency of reinforcement learning; Using SAC (Soft Actor-Critic), an entropy-regularized maximum entropy reinforcement learning framework, which significantly improves exploration efficiency and policy robustness by introducing an entropy regularization term.

The above strategies can be applied to different aspects of autonomous driving technology: learning to generate optimal trajectories between obstacles and lane lines for path planning; learning to avoid other vehicles/pedestrians in dynamic traffic conditions; being able to handle scenarios requiring strategic decision-making, such as merging and intersections; and dynamically switching strategies according to different objectives and tasks.

3.2 Extension of reinforcement mechanisms

Curriculum Learning: Just as humans learn any skill (such as mathematics, language, or musical instruments), teachers start by teaching basic concepts (simple tasks) and then gradually introduce more complex knowledge and challenges (difficult tasks). Curriculum learning applies this concept to machine learning model training, with the core being the strategic organization of training data. The model is trained starting with the simplest, cleanest, and most typical samples, and then, according to predefined difficulty standards and a progress plan, more complex and challenging samples are gradually introduced.

This gradual approach, which progresses from easy to difficult, simulates human learning and aims to guide the model to optimize more stably and efficiently, learn more robust basic features, and ultimately improve the model's generalization ability and final performance on new data.

3.3 Challenges

Artificially designed reward functions often contain prior biases, leading to agents "learning shortcuts." The design of strategy reward functions is highly subjective and requires Inverse Reinforcement Learning(IRL) to address this issue, i.e., inferring reward functions from expert demonstrations to reduce subjective design biases; High-dimensional inputs, such as image processing, impose heavy computational burdens. Methods like Discrete Bottleneck must be employed to map continuous features to discrete latent spaces, reducing information redundancy and enabling perceptual feature compression; Real-world deployment requires domain adaptation and security mechanisms. This can be achieved through Domain Randomization and the implementation of security constraint layers.

4 Perception and decision-making collaboration system

4.1 System structure

The perception-decision integrated system primarily includes:

1) Perception Frontend (Modal Alignment + BEV Mapping + Feature Fusion): First, data is processed using cross-modal contrast learning and feature distillation techniques. Then, 2D-to-3D perspective conversion technology based on Transformers is employed for mapping. Finally, gated adaptive weighting (Gated Fusion) is used for feature fusion to obtain the final processed data.

2) Strategy Network (Actor-Critic Model + Multi-Objective Reward Function): The processed data is fed into a deep reinforcement learning model for training. By combining the Actor-Critic model with a multi-objective reward function, the system achieves discrete action decision-making and continuous trajectory prediction.

3) Training Platform (CARLA Simulation): The processed data and established model are tested on the CARLA platform for dynamic weather and traffic flow simulation to simulate real-world weather conditions and road conditions encountered in actual traffic. The simulation results are then used to estimate real-world road conditions.

4.2 Fusion key mechanisms

1) BEV Generation as a State Input Bridge: Using BEV generation as a state input bridge unifies multi-modal distributions, maps camera/radar/lidar data to a bird's-eye view coordinate system, eliminates perspective distortion, and improves target localization accuracy.

2) Cross-Attention: Cross-Attention is used to align camera features and radar features before feeding them into the decision network, reducing feature redundancy and lowering computational load.

3) Dynamic Reward Function Weights: Adjust weights in real-time based on different road conditions and weather scenarios, considering parameters such as safety, smoothness, and comfort.

4) Data Synchronization: Multiple sensors sample simultaneously, and the sampled data is calibrated in real-time using SE(3) to achieve spatio-temporal consistency across data modalities.

5 Experimental platform and data support

CARLA Simulation Environment :

1) Supports RGB, LiDAR, and IMU sensor mounting and synchronization: The CARLA simulation environment automatically adjusts dynamic exposure for RGB cameras; establishes noise models for LiDAR to simulate rain and fog interference; and establishes temperature drift models for IMU. Additionally, the CARLA simulation environment synchronizes information from different sensors, resulting in more rigorous simulation results.

2) Various complex terrains can be set:The CARLA simulation platform can be configured to meet different research needs, dynamically configuring urban roads (two-way lanes/T-junctions/commercial areas), highways (6 lanes + ramps + tunnels), and multi-directional intersections (turbo-type/roundabouts). The CARLA platform includes 12 pre-set scenarios and allows XML customization; terrain complexity is adjustable (levels 1-5), and it integrates dynamic weather, traffic flow density, pedestrian systems, and adaptive traffic lights. Combined with the OpenStreetMap road network import function, it achieves millimeter-level precision in the SE(3) coordinate system construction in the Town series maps, providing progressive training scenarios for autonomous driving algorithms ranging from simple roads to rainy nighttime eight-way roundabouts.

3) Supports Python API + Gym interface for end-to-end control training: CARLA deeply integrates with the Python API and Gym standard interface to enable end-to-end reinforcement learning training from multi-modal sensor inputs (RGB/LiDAR/IMU) to vehicle-level controls (steering/throttle/brakes). Supports 100Hz high-frequency command issuance in continuous control space, providing real-time feedback on dynamic states; compatible with frameworks such as Stable-Baselines3/RLlib, achieving parallel training throughput exceeding 8, 000 frames per minute with single-step latency under 15ms.

6 Research trends and challenges outlook

6.1 Potential breakthroughs

1) Large-scale model perception fusion: Combine cross-modal large-scale model technology for semantic extraction and planning understanding. For example, GPT-Vision can be used to encode visual and lidar point clouds into token sequences, generate explainable driving instructions, thereby improving semantic understanding accuracy, and achieving real-time inference optimization and decision optimization.

2) Lightweight deployment of end-to-end decision-making systems: The lightweight deployment of end-to-end decision-making systems is a key bottleneck in the implementation of autonomous driving. Its essence lies in resolving the fundamental contradiction between the computational demands of large models and the resource constraints of in-vehicle hardware. Current mainstream solutions employ a three-tier compression strategy: first, task-specific distillation is used to compress a hundred-billion-parameter visual-language model (such as GPT-4V) into a million-parameter 3D convolutional network, achieving over 95% knowledge transfer through KL divergence and feature imitation loss; second, dynamic structured sparsification is applied to dynamically activate network channels based on the importance of driving scenarios; Finally, it employs INT4 quantization-hardware co-design, leveraging the sparse tensor core and instruction set optimizations of the Tesla D1 chip. However, this field still faces three major challenges: first, the accumulation of quantization errors in dynamic scenarios amplifies control deviations; second, cross-modal alignment losses in multi-sensor fusion models, with a 3-8% accuracy gap between lidar and camera feature distillation; Third, safety certification barriers, with ISO 26262 requiring a failure rate of <10 FITs (1 FIT = 1 failure per 1 billion hours), and current compressed models not yet meeting robustness standards under extreme conditions. The industry is seeking breakthroughs through heterogeneous computing architectures (such as the FP8 sparse units in NVIDIA's Thor chip) and lifecycle learning frameworks (incremental fine-tuning after deployment).

3) Improved generalization capability: Traditional static rule-driven approaches, which have limited scene coverage and struggle to adapt to dynamic environments, are gradually being replaced by a new paradigm combining task-driven approaches and few-shot generalization. Its technical core involves three breakthroughs: First, task-driven architectures decompose driving objectives into composable atomic skills (e.g., the 23-skill library defined by Mobileye ChauffeurNet RL) through hierarchical reinforcement learning, enabling cross-scenario strategy reuse—for example, the "construction zone detour" task automatically invokes the "lane change" + "obstacle avoidance" skill combination; Second, the meta-learning framework (such as Berkeley DriveAdapter) relies on model-agnostic meta-learning (MAML) mechanisms to achieve domain adaptation with ⩽5 new scene samples. The core lies in decoupling scene-invariant features (road

topology) from domain-variant features (lighting/texture); Finally, generative few-shot learning, represented by Wayve GAIA-1, uses spatio-temporal diffusion models to synthesize physically validated extreme scenarios (e.g., vehicles driving against traffic/road collapses) at the million-scale level, ultimately yielding synthetic data with probability coverage > 99.9%. However, cognitive uncertainty in open-world environments remains a key challenge, as existing methods inadequately model implicit environmental logic (e. g., "maintaining lane position based on reference shoulder geometry"). Cutting-edge research is integrating neuro-symbolic architectures, combining the causal reasoning of large language models (e.g., scene semantic analysis in DriveGPT4) with decision optimization from reinforcement learning.

6.2 Actual deployment challenges

1) Asynchronous interference and drift between sensors: Multi-modal sampling clock misalignment (such as a 30Hz camera and a 20Hz lidar timing mismatch) makes it difficult to align cross-sensor data in the spatiotemporal dimension; IMU zero bias temperature drift (>0.1°/s) causes attitude angle estimation to deviate by over 3.6° per hour, while LiDAR external parameter offset (caused by vibration/thermal expansion and contraction, resulting in >2cm/24h rigid transformation error) leads to failed point cloud and image registration. More critically, the coupled amplification effect of errors traps the perception-decision chain in a vicious cycle: asynchronous interference reduces feature fusion confidence, forcing the system to increase computational load to compensate for uncertainty, while sustained high-load operation exacerbates hardware temperature rise, further amplifying IMU zero bias and lens distortion.

2) Deep models have high hardware resource dependencies and require heterogeneous computing platforms for collaboration: The extreme hardware resource dependencies of autonomous driving deep models have become a core bottleneck for industry implementation. A BEV perception model with hundreds of billions of parameters requires extremely high computing power to support real-time inference, placing triple compound pressure on in-vehicle computing platforms. In terms of computing power, the sparse convolutional network for processing lidar point clouds requires 2.7 trillion operations per second, far exceeding the processing limits of traditional CPUs. In terms of energy efficiency, multi-modal fusion models on the NVIDIA Orin platform reach a peak power consumption of 220W, causing a 60° C temperature rise in the in-vehicle cooling system and significantly reducing electric vehicle range; in terms of hardware costs, dual-chip solutions meeting ASIL-D safety redundancy requirements (such as Tesla's HW4.0) have a unit price exceeding $2,500.

3) Safety and explainability: Industry reports show that 31% of AEB false triggers in Euro NCAP tests in 2023 were attributed to the uninterpretable decisions of deep learning models (e. g., Transformer attention weights cannot be mapped to specific traffic rules), making it impossible for automakers to identify the root cause of the fault; Additionally, in extreme scenarios, counterintuitive behaviors (such as sudden braking in heavy rain) occur, requiring engineers to reverse-engineer activation paths to barely correlate a small portion of fault samples. More critically, the chaotic temporal sequence of fault occurrence makes post-event tracing nearly impossible, as the causal chain from sensor misdetection to vehicle execution of incorrect actions involves multiple heterogeneous computational modules, rendering traditional diagnostic tools incapable of identifying the root cause.

7 Conclusion

This paper reviews the current state of research on multimodal perception fusion and deep reinforcement learning decision-making mechanisms in the field of autonomous driving. By systematically organizing the structures of perception fusion methods and deep reinforcement learning strategies in autonomous driving, and focusing on the integration optimization and experimental feasibility of autonomous driving collaborative systems, the paper clarifies the strengths and weaknesses of current technologies.

As artificial intelligence algorithms continue to evolve and sensing hardware upgrades, autonomous driving systems are advancing toward higher levels of intelligence and autonomy. Future efforts should prioritize system collaboration, model generalization, experimental standardization, and safety assurance to drive the widespread application and deployment of perception-decision integration in real-world road environments.

Funding

Supports RGB, LiDAR, and IMU sensor mounting and synchronization: The CARLA simulation environment automatically adjusts dynamic exposure for RGB cameras; establishes noise models for LiDAR to simulate rain and fog interference; and establishes temperature drift models for IMU. Additionally, the CARLA simulation environment synchronizes information from different sensors, resulting in more rigorous simulation results.

About the Author

Qinwen Qiu, 221870217@smail.nju.edu.cn

References

[1] J. Zhang and S. Singh, "LOAM: Lidar Odometry and Mapping in Real-time," in Robotics: Science and Systems (RSS), 2014.

[2] S. Lefèvre, D. Vasquez, and C. Laugier, "A survey on motion prediction and risk assessment for intelligent vehicles," in Robotics and Autonomous Systems, vol. 58, no. 9, pp. 1372-1385, 2010.

[3] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[4] Y. Zhou and O. Tuzel, "VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[5] Z. Liu et al., "BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[6] K. Chitta et al., "Multi-Modal Fusion Transformer for End-to-End Autonomous Driving," in Conference on Robot Learning (CoRL), 2021.

[7] Mnih, V. et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529 – 533.

[8] Lillicrap, T. P., et al. "Continuous control with deep reinforcement learning." ICML 2016.

[9] Mnih, V., et al. "Asynchronous methods for deep reinforcement learning."ICML 2016.

[10] Haarnoja, T., et al. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor." ICML 2018.

[11] Florensa, C., et al."Automatic Goal Generation for Reinforcement Learning Agents."ICML 2017.

[12] Finn, C., et al. "Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization." ICML 2016.

[13] Van den Oord, A., et al. "Neural Discrete Representation Learning." NeurIPS 2017.

[14] Tobin, J., et al. "Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World." IROS 2017.

[15] Dalal, G., et al. "Safe Exploration in Continuous Action Spaces." AISTATS 2018.

[16] Li, Y., et al."DeepFusion: Lidar-Camera Fusion via Depth-aware Gating for 3D Object Detection."CVPR 2022.

[17] Philion, J., & Fidler, S."Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D."ECCV 2020.

[18] Dosovitskiy, A., et al."CARLA: An Open Urban Driving Simulator."CoRL 2017.

[19] Huang, Z., et al."BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via SpatiotemporalTransformers."ECCV 2022.

[20] Chen, Z., et al."SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot."International Conference on Machine Learning (ICML), 2023.

[21] Shalev-Shwartz, S., Shammah, S., Shashua, A."On a Formal Model of Safe and Scalable Self-driving Cars"Robotics: Science and Systems (RSS), 2018.

[22] Zhou, B., et al."MetaDrive: Sim-to-Real Generalization via Task-Agnostic Meta-Learning"Robotics: Science and Systems (RSS), 2023.

[23] Hawke, J., et al."GAIA-1: Generative World Model for Autonomous Driving" Wayve Technical Report, 2023.

[24] Euro NCAP Test Protocol v3.1 Annex F (Table F.2, 2023).