Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BEV-Parking: An End-to-End Autonomous Parking Approach Utilizing Bird's Eye View Representations

Xuguang Sun
China North Vehicle Research Institute
Zhengwu Fan
fanzhengwu2008@126.com

Taiyuan University of Technology
Ruiliang Zhang
Taiyuan University of Technology
Jiaxiong Wu
China North Vehicle Research Institute
Xiaoce Sun
China North Vehicle Research Institute

Article

Keywords: Autonomous parking, Bird's Eye View, end-to-end learning, neural networks, CARLA simulator

Posted Date: May 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9105895/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

With the rapid development of autonomous driving, automated parking plays an increasingly significant role for intelligent vehicles. However, when current autonomous parking methods encounter complex scenarios—especially in crowded parking lots—shortcomings such as parking failure or collisions become evident. In such scenarios, BEV-based neural network algorithms appear to outperform traditional rule-based methods. In this paper, we refine BEV sensing to more accurately capture relevant information in complex parking scenarios, while employing an end-to-end model to generate driving trajectories. The network takes as input BEV feature maps generated from six camera views and expert human driving data, and outputs trajectory points with corresponding control signals. We conducted closed-loop experiments in various scenarios of the CARLA simulator to validate the effectiveness of the proposed algorithm, and the results demonstrate its superior performance.

1 Introduction

By 2025, automated parking has become a prominent topic in autonomous driving, playing a significant role in the vehicle transportation industry. As intelligent vehicle technology continues to advance rapidly, enabling vehicles to park autonomously in complex environments and confined spaces has become a core challenge for autonomous driving systems. Drivers who rely solely on their driving experience may struggle to park smoothly or avoid collisions in narrow or crowded parking spaces, which can lead to financial losses and safety risks. Even experienced drivers can face difficulties in such scenarios, resulting in accidents such as scratches or collisions with surrounding obstacles. This issue is particularly acute in environments with limited space and numerous obstacles, where high precision in perception, planning, and control is essential. Recent work by Zhang et al. [27] proposed a trajectory planning method combining Hybrid A* and nonlinear model predictive control for narrow parking scenarios, highlighting the ongoing research efforts in this area.

In this context, the development of autonomous parking systems has become increasingly critical. Traditional parking assist systems typically rely on sensor inputs and predefined geometric rules to guide the vehicle [1]. While these systems can be effective in simpler environments, they often lack flexibility and efficiency when dealing with dynamic or complex parking scenarios. Furthermore, errors from sensors and models tend to accumulate, reducing the robustness and flexibility of these systems [2]. This makes traditional multi-stage approaches fragile and ill-suited for handling complex real-world parking scenarios.

Recently, end-to-end learning has been explicitly applied to autonomous parking tasks. Yang et al. proposed E2E Parking, an end-to-end autonomous parking framework that directly maps multi-camera images and vehicle states to control commands in the CARLA simulator, demonstrating the feasibility of imitation learning for parking scenarios. Subsequently, transformer-based approaches such as Park former further enhanced spatial reasoning and long-range dependency modeling for parking maneuvers.

These works highlight the potential of end-to-end parking systems, while also revealing limitations in spatial consistency and robustness under complex parking layouts.

This paper presents YOPO (You Only Park Once), an end-to-end autonomous parking framework that integrates refined Bird's Eye View (BEV) sensing with auxiliary supervision. The main technical contributions are: (i) a BEV generation module with explicit depth supervision that improves spatial understanding in crowded parking lots; (ii) auxiliary tasks (depth prediction and semantic segmentation) that enhance the interpretability and robustness of the learned representations; and (iii) a trajectory-based control output that produces smooth and precise parking maneuvers. These components are designed to work together within a unified neural network architecture, enabling the vehicle to park autonomously in complex scenarios.

Furthermore, end-to-end learning approaches have shown great potential for autonomous driving tasks, including parking [6]. These systems directly map raw sensor inputs, such as camera images and vehicle motion states, to control outputs like steering angle, acceleration, and gear without relying on explicit handcrafted rules. One particularly promising technique within this framework is Bird's Eye View (BEV) representation, which provides a top-down view of the vehicle’s surrounding environment [18]. BEV has proven to be highly effective in many autonomous driving tasks, including parking, as it captures critical 3D information such as obstacles, lanes, and other traffic participants in a unified format. By using BEV, end-to-end parking systems can better perceive the environment and make more accurate control decisions, especially in tight and dynamic spaces [12].

End-to-end systems, particularly those employing BEV, aim to reduce the reliance on complex multi-stage pipelines by integrating perception, planning, and control into a single unified model. These models can learn to park autonomously by imitating expert human drivers, making them more adaptable to various parking scenarios. However, despite the promise of such methods, challenges remain in scaling up these systems for real-world applications, particularly in dynamic environments where obstacles move and sensor data can be noisy. The combination of BEV representation with real-time control strategies has the potential to significantly improve the accuracy, safety, and efficiency of autonomous parking systems [2]. Motivated by the strengths of end-to-end approaches and BEV representation, we integrate the advantages of both methodologies and propose a novel architecture to enhance the performance of automatic parking systems, which we call YOPO (You Only Park Once), as shown in Fig. 1. In the proposed approach, the only requirement for users is to identify an appropriate parking goal; thereafter, the vehicle, utilizing the YOPO framework, autonomously tracks the designated parking space until the parking task is completed. Our end-to-end automated parking system processes visual input to generate precise vehicle control signals. Through extensive training and parameter optimization, it enables vehicles to autonomously navigate and park in complex parking environments. To assess the effectiveness of the proposed approach, we developed quantitative metrics and conducted closed-loop experiments using the CARLA [3] simulator. The results demonstrate that our approach achieves a 91.6% success parking rate, with an average position error of 0.3 meters and an orientation error of 0.9 degrees.

To address the limitations of traditional rule-based parking methods and improve the performance of autonomous parking systems in complex scenarios, this paper makes the following contributions:

- **Refinement of BEV Sensing**: We enhance the BEV representation by refining the sensing process to more accurately capture critical environmental information in complex parking scenarios. This refinement improves the system's ability to perceive and interpret the environment, especially in crowded parking lots with multiple obstacles.
- **End-to-End Model Integration**: We design an end-to-end neural network model that directly maps BEV feature maps and human driving data to control outputs. Unlike traditional methods that rely on multi-stage processing pipelines, our approach simplifies the overall architecture while maintaining high adaptability and robustness in diverse parking scenarios.
- **Multi-View BEV Feature Input**: The proposed network leverages BEV feature maps generated from six camera views, enabling a comprehensive understanding of the vehicle's surroundings. This multi-view input allows the system to better handle challenging parking tasks by providing richer spatial and contextual information.
- **Trajectory-Based Control**: The output of our network includes trajectory points and control signals (e.g., steering angle, acceleration, and gear), which allow for smooth and precise parking maneuvers. This trajectory-based approach bridges the gap between perception and control in autonomous parking systems.
- **Extensive Validation**: We validate the proposed approach through closed-loop experiments in the CARLA simulator, simulating various parking scenarios with different levels of complexity. The experimental results demonstrate the effectiveness of our algorithm, achieving excellent performance in terms of accuracy, robustness, and success rate.

The remainder of this paper is organized as follows. Section 2 reviews related work on traditional parking methods and end-to-end learning paradigms. Section 3 describes the methodology of YOPO, including problem definition, overall architecture, and implementation details. Section 4 presents the experimental setup, results, and discussion. Section 5 concludes the paper and outlines future work.

2 Related Work

In recent years, the rapid advancement of artificial intelligence has significantly accelerated the transformation of the autonomous driving industry, particularly in the domain of automated parking. This section examines the limitations of traditional parking methods and explores recent advancements in neural network models and methodologies applied to automated parking.

2.1 Limitations of Traditional Parking Methods

Traditional automated parking methods primarily depend on rule-based systems and multi-stage pipelines, comprising distinct modules for perception, localization, planning, and control. Although

effective in simple and structured environments, these methods face several inherent limitations:

- **Path Planning Constraints**: Traditional parking systems frequently utilize algorithms such as Hybrid A* for path planning. While these methods perform well in structured environments, they often struggle with the unstructured and dynamic nature of real-world parking lots. This can result in suboptimal trajectories that fail to adequately account for moving obstacles or complex layouts [1].
- **Limited Re-planning Capabilities**: Many conventional systems are unable to re-plan paths in real-time when encountering unexpected obstacles. This limitation increases the risk of collisions or failed parking attempts, particularly in dynamic environments with unpredictably moving obstacles [3].
- **Error Accumulation**: The multi-stage architecture of these systems involves distinct modules for tasks such as perception, localization, planning, and control. Errors from one module, such as inaccurate sensor data or localization, can propagate through subsequent stages, reducing system robustness and accuracy, particularly in complex and dynamic real-world environments [2, 4].
- **Sensor Noise**: Traditional methods depend heavily on sensors such as ultrasonic and radar to detect the surrounding environment and guide vehicle movement. However, these sensors are highly susceptible to noise, which can result from factors such as environmental interference, reflective surfaces, or crowded, confined spaces. Such noise often leads to misinterpretation of obstacles or inaccuracies in distance measurements, ultimately affecting the system's reliability and performance in real-world scenarios [5].

2.2 End-to-End Learning in Autonomous Driving

In recent years, end-to-end learning approaches have gained significant popularity among researchers in autonomous driving [6]. These methods simplify the driving process by directly mapping sensor inputs, such as camera images and lidar data, to control signals, eliminating the need for complex, predefined rules and modular pipelines [7]. Unlike traditional approaches, which rely heavily on rule-based systems, end-to-end learning exhibits stronger generalization capabilities, making it more adaptable to diverse and dynamic driving scenarios [8]. These systems have demonstrated notable success across a range of autonomous driving tasks, including lane keeping, obstacle avoidance, and urban navigation, showcasing their potential for more efficient and flexible solutions in real-world applications.

End-to-end models minimize dependence on handcrafted rules and modular designs, enhancing adaptability and generalization across diverse environments. Li et al. provide comprehensive analyses of these advancements, highlighting the integration of perception, prediction, and planning within unified neural network architectures [9]. However, applying end-to-end learning to parking introduces unique challenges, as it demands precise control in confined and often dynamic spaces [10]. Recent research has investigated reinforcement learning-based end-to-end parking approaches, where vehicles learn to generate optimal steering commands through extensive trial-and-error parking attempts. While these methods have shown promise, achieving human-like intelligence in parking tasks remains a complex

endeavor. Key challenges include addressing data scarcity for training robust models, enhancing the ability to make accurate real-time decisions, and improving generalization across diverse parking scenarios [11]. To fully leverage the end-to-end framework, we employ a transformer to directly map visual images to control signals.

2.3 BEV Representation

Bird's Eye View (BEV) has become a crucial data representation in autonomous driving, providing a top-down perspective of the vehicle's surroundings. BEV representations are extensively utilized in tasks such as semantic segmentation [12], object detection [14, 15], and path planning [16]. They facilitate the capture of spatial relationships and contextual information in 3D space, which is crucial for accurate environmental perception. Du et al. have focused on improving BEV generation from multi-camera inputs, enhancing the system's ability to interpret complex driving environments [17]. In parking scenarios, BEV simplifies the representation of critical elements such as obstacles, lanes, and parking slots, making it especially suitable for end-to-end models. For instance, Wang and Huang [28] demonstrated the effectiveness of BEV representations for fast vehicle detection using colored point clouds, achieving high accuracy with low latency. Recent advancements, including BEV Depth [13] and Lift-Splat-Shoot (LSS) [18], have further enhanced the accuracy of BEV generation from multi-camera inputs, enabling more precise environmental perception and control in complex parking tasks. Considering both accuracy and efficiency, our YOPO model ultimately adopts the LSS module as the benchmark for generating BEV representations [18].

2.4 Parking Tasks and Neural Networks

Neural network-based approaches are increasingly being applied to parking tasks in autonomous driving, offering improved adaptability and precision over traditional rule-based methods. These approaches leverage deep learning models, such as CNNs and transformers, to process complex sensor data and directly predict control actions like steering and acceleration [19]. While promising, their application in parking tasks remains in early development, with challenges in handling tight spaces, dynamic obstacles, and ensuring real-time performance [20]. The incorporation of transformer-based models has shown promise in various autonomous driving tasks, including lane detection, segmentation, tracking, planning, and decision-making. These models effectively fuse multi-modal inputs, such as images and LiDAR data, to generate control signals [21, 22, 23]. However, their application specifically optimized for parking tasks remains underexplored. Zhong et al.'s comprehensive survey highlights the potential of transformers in autonomous driving but indicates a gap in their utilization for precise parking scenarios [24].

3 Methodology

This section presents the methodology for the proposed autonomous parking system, which incorporates refined BEV sensing, an end-to-end neural network model, and a trajectory-based control approach. Our system takes advantage of multi-view BEV feature maps generated from six camera

views and uses expert human driving data to train the model, with the aim of achieving high accuracy and robustness in complex parking scenarios.

3.1 Problem Definition

The objective of our autonomous parking system is to enable vehicles to autonomously park into a target parking slot in complex and dynamic environments using BEV perception. With the proposed end-to-end network, the system maps sensor inputs (camera images, vehicle motion states) to control outputs (steering angle, acceleration, and gear). We define the problem as follows:

- **Inputs**: The inputs to the system are BEV feature maps generated from six camera views (left-front, front, right-front, left-rear, rear, and right-rear) and the vehicle's motion states (e.g., speed and acceleration, obtained from wheel speed sensors and IMU). The target parking slot is also provided as part of the input and is tracked throughout the parking task.
- **Outputs**: The outputs include control signals predicted by the network, such as steering angle, acceleration, and gear selection (forward or reverse). Our model predicts vehicle control signals for the next four steps, with a time interval of 0.1 seconds between each step. These control signals are used to guide the vehicle’s trajectory into the target slot.

The problem can be formally defined as:

$$\min_{\pi} \mathrm{E}_{(X,C)\sim D}\left[L(\,C, \pi\,(X)\,)\right]$$

where X is the input feature (BEV feature map and motion states), C is the corresponding control signal (steering angle, acceleration, gear), and L is the loss function. The network learns to minimize the discrepancy between predicted control signals and the ground truth.

3.2 Network Architecture

Our proposed approach leverages an end-to-end neural network model to directly map BEV feature maps and vehicle motion states to control signals, eliminating the need for manual feature extraction or separate planning modules. As shown in Fig. 2, the architecture consists of the following key components: BEV Generation, Feature Fusion, and Control Prediction.

3.2.1 BEV Generation

Drawing inspiration from the nosecones dataset [25], a widely used benchmark for automated driving research, we utilize six camera views (left-front, front, right-front, left-rear, rear, and right-rear) to generate the BEV feature map, as shown in Fig. 3. Our BEV perception module is broadly divided into three key components: Feature Extraction, Depth Estimation, and BEV Mapping. The overall process is structured as follows:

- **Feature Extraction**: Each camera image is processed by a CNN backbone (EfficientNet [26] is used in this paper) to extract hierarchical features. Initial layers capture basic textures and edges, while deeper layers encode complex patterns like objects and spatial relationships. Advanced techniques like multi-scale feature extraction and attention mechanisms can further enhance the representation, ensuring accurate and efficient perception for downstream tasks. This also forms the foundation for subsequent sub-tasks, including object detection and semantic segmentation.
- **Depth Estimation**: Leveraging the Lift-Splat-Shoot (LSS) method [18], we estimate the depth distribution for each pixel in the camera view, facilitating the generation of a more accurate BEV representation. This depth estimation significantly enhances the system's ability to comprehend the surrounding environment.
- **BEV Mapping**: The extracted features and depth maps are transformed into a bird's-eye view using the camera intrinsic parameters and extrinsic transformations. The resulting BEV feature map provides a top-down view of the environment, where the vehicle’s location, obstacles, and target parking slot are all represented in a unified grid format [12, 18]. The BEV representation for the parking lot is shown in Fig. 4.

3.2.2 Feature Fusion

For a vehicle to autonomously complete a parking task, it requires information about its current motion state and the target parking space, alongside visual data about the surrounding environment. The feature fusion module combines the BEV feature map with additional vehicle motion information (e.g., speed and acceleration). This module performs the following:

- **Concatenation of Features**: The BEV feature map is concatenated with the target parking slot information and vehicle motion data, creating a unified feature map. The target slot information is represented as a one-hot vector, where the target slot position is marked as '1', and all other positions are marked as '0'.
- **Attention Mechanism**: The fused feature map is passed through a transformer encoder that uses self-attention mechanisms to capture both spatial and temporal dependencies in the feature map. This enables the model to focus on key features in the environment, such as the target slot and surrounding obstacles.

3.2.3 Control Prediction

The control prediction module employs a transformer decoder to produce control signals, including steering angle, acceleration, and gear selection, based on the fused features from previous stages. By utilizing an attention mechanism, the transformer decoder focuses on relevant aspects of the input features, such as the target parking slot and nearby obstacles, while accounting for the vehicle's current motion state. The process involves:

- **Auto-Regressive Prediction**: The decoder generates a sequence of control signals, including steering, acceleration, and gear state, in a step-by-step manner using an auto-regressive approach. Each prediction is conditioned on both the input features and the previously generated outputs, ensuring smooth transitions between steps. Control signals are predicted for a short horizon of 4 steps (approximately 0.4 seconds), striking a balance between computational efficiency and responsiveness to dynamic environmental changes.
- **Control Signals Mapping**: The output control signals are normalized to predefined ranges to ensure consistent and safe operation. Steering values are scaled between − 1 and 1, where − 1 corresponds to the maximum left turn and 1 to the maximum right turn. Acceleration values are similarly normalized, with − 1 representing full braking and 1 representing full throttle. The gear state is encoded as a binary signal, where one state indicates forward motion and the other represents reverse. This normalization and discretization process ensures compatibility with the vehicle's physical constraints and facilitates precise control during parking maneuvers.
- **Loss Function**: Cross-entropy loss is utilized to evaluate the accuracy of the predicted control sequence by comparing it with the ground truth, ensuring precise predictions for steering, acceleration, and gear states. Additionally, a semantic segmentation loss is applied to improve the detection of the target parking slot and surrounding obstacles. This dual-loss approach ensures the model effectively understands the environment and prioritizes critical features for safe and efficient parking.

3.2.4 Network Configuration Details

To ensure reproducibility of the proposed YOPO framework, the detailed network configuration and training parameters are summarized as follows. The BEV feature map is constructed with a spatial resolution of 0.2 m per cell, resulting in a grid size of 200 × 200. EfficientNet-B3 is adopted as the image backbone for feature extraction from each camera view. The transformer encoder consists of 4 layers with a hidden dimension of 256 and 8 attention heads, while the transformer decoder includes 4 layers with identical hidden dimensions and attention heads. The control prediction horizon is set to 4 steps with a temporal resolution of 0.1 s per step.

The network is trained using the Adam optimizer with an initial learning rate of 1e-4 and a batch size of 16. Training is conducted for 50 epochs on the collected expert parking dataset. During training, gradient clipping with a maximum norm of 5.0 is applied to stabilize optimization. These configurations allow the proposed method to be fully reproducible and fairly compared with existing end-to-end parking approaches.

3.3 Auxiliary Tasks

To enhance the model's interpretability and accuracy, auxiliary tasks are incorporated to provide additional supervisory signals during training. These tasks are designed to complement the primary control prediction objective, enabling the model to learn richer and more meaningful feature

representations. By encouraging the network to capture multiple aspects of the environment, these auxiliary tasks enhance generalization to diverse scenarios and edge cases.

- **Depth Prediction**: We integrate a depth prediction task, where the network is trained to estimate scene depth under the supervision of ground truth depth maps. These maps are generated using the LSS method during data collection. Depth prediction enhances the BEV representation by enabling the model to better understand relative distances between the vehicle and surrounding obstacles, thereby improving spatial awareness and navigation accuracy.
- **Semantic Segmentation**: The network performs segmentation to classify the scene into distinct categories, including drivable areas, obstacles, and parking slots. This task allows the model to focus on critical regions in the environment while disregarding irrelevant details. Accurate segmentation is essential for identifying the target parking slot and avoiding collisions with surrounding objects, thereby enhancing the system's safety and effectiveness.

These auxiliary tasks not only enhance the model's interpretability by offering insights into intermediate predictions but also bolster its capacity to make accurate and robust decisions in complex environments. By incorporating these supplementary learning objectives, the system attains a more comprehensive understanding of the parking task, resulting in safer and more reliable performance.

3.4 Training Procedure

The model is trained using a supervised learning approach, leveraging a dataset of parking trajectories collected from expert human drivers. This dataset includes synchronized inputs such as multi-camera images, vehicle motion states (e.g., speed, acceleration, orientation), and corresponding control signals (steering, acceleration, and gear) at each time step. The training process employs backpropagation to iteratively update the network weights by minimizing a composite loss function. This loss function integrates the control signal prediction loss, ensuring accurate trajectory generation, and the auxiliary task losses from depth prediction and semantic segmentation, which improve environmental understanding and spatial awareness. This combined training strategy enhances both accuracy and robustness, enabling the model to generalize effectively to diverse parking scenarios.

During training, the model is exposed to a diverse set of parking scenarios, ranging from straightforward open spaces to challenging environments, such as narrow parking lots with multiple obstacles. This variety helps the model learn to generalize and adapt to different conditions, ensuring robustness and reliability. After training, the model is validated through closed-loop experiments in the CARLA simulator, where its performance is evaluated in real-time parking tasks. These experiments test the system's ability to navigate, avoid obstacles, and accurately park in various simulated scenarios, providing a comprehensive assessment of its effectiveness and adaptability.

3.5 Closed-Loop Experimentation

We evaluate the performance of our proposed system in a variety of parking scenarios simulated within the CARLA simulator. These scenarios feature diverse parking lot layouts, dynamic obstacles, and

varying levels of environmental complexity to comprehensively assess the system's capabilities. The evaluation focuses on the system’s ability to accurately and efficiently park the vehicle into the target slot while minimizing errors.

4 Experiments

This section describes the experimental setup, data collection process, evaluation metrics, and results obtained from testing the proposed autonomous parking system. The experiments were conducted using the CARLA simulator, which provides a realistic environment for simulating various parking scenarios. The goal is to validate the effectiveness of the proposed method and demonstrate its superiority over traditional rule-based approaches.

4.1 Experimental Setup

All experiments in this study are conducted in the CARLA simulator. While the simulator provides realistic physics and sensor modeling, the results presented are based on simulated environments and may differ from real-world deployment. The system was tested for its ability to park the vehicle autonomously into a target parking slot in these simulated environments.

Key aspects of the experimental setup include:

- **Simulated Environment**: The CARLA simulator (version 0.9.11) was utilized to create a diverse set of virtual parking environments for testing. These environments included various parking lot layouts, ranging from simple open spaces to highly complex configurations with tightly spaced parking slots. To simulate real-world conditions, dynamic obstacles such as moving vehicles and pedestrians were introduced. This variability in complexity and dynamic elements allowed for a thorough evaluation of the system's ability to adapt and perform effectively across different parking scenarios.
- **Vehicle**: The simulated vehicle was equipped with six strategically positioned cameras to capture a comprehensive view of its surrounding environment. These cameras covered the front, rear, sides, and additional perspectives, ensuring robust perception for the BEV generation module. The captured images were used to construct feature maps that provided spatial and contextual information critical for accurate parking maneuvers. Additionally, the vehicle’s motion states, including speed, acceleration, and orientation, were continuously logged during the simulation to provide dynamic context. This combination of visual and motion data enabled the system to make precise and informed control decisions.
- **Training Data**: The network was trained using a dataset collected from expert human drivers, consisting of parking trajectories that included synchronized images from six cameras, the vehicle's motion states (such as speed, acceleration, and orientation), and the corresponding control signals (steering, acceleration, and gear). The dataset comprised a total of 22,000 frames captured across

various parking scenarios, encompassing diverse layouts and complexities. This extensive dataset allowed the network to learn robustly and generalize effectively to a wide range of parking tasks.

- **Control Signals**: The network outputs control signals that include steering angle, acceleration, and gear state (forward or backward). These signals are predicted for a sequence of 4 steps, with each step representing 0.1 seconds of driving time. This sequential output allows the system to plan and execute smooth, precise maneuvers in real-time, adapting dynamically to the parking environment.

The driving dataset used in this study was generated entirely in the CARLA simulator. No human participants were involved, and no real-world human data were collected. As such, informed consent was not applicable.

Since this research did not involve any human subjects, animal subjects, or human tissue, ethical approval was not required under the regulations of the authors' institutions or applicable national guidelines.

All experiments and methods were conducted in accordance with the relevant guidelines and regulations of the Taiyuan University of Technology and the China North Vehicle Research Institute.

4.2 Evaluation Metrics

In this study, we adopt the evaluation metrics proposed in E2E Parking [20] to quantitatively evaluate the performance of the automated parking task. The specific parameters include:

- **Target Success Rate (TSR)**: The percentage of successful parking attempts where the vehicle was parked within a certain distance of the target parking slot. A successful parking attempt is defined as when the vehicle's final position is within 0.6 meters of the target slot and the orientation error is less than 10 degrees.
- **Average Position Error (APE)**: The average distance between the final position of the vehicle and the center of the target parking slot among successful parking attempts. A smaller APE indicates more precise parking.
- **Average Orientation Error (AOE)**: The average angular deviation between the vehicle's final orientation and the desired orientation for parking. The lower the AOE, the more accurately the vehicle aligns with the target parking slot.
- **Average Parking Time (APT)**: The average time taken to complete a successful parking attempt. This metric helps assess the efficiency of the system.
- **Non-Target Rate (NTR)**: The percentage of times the vehicle parks in a non-target parking slot. A lower NTR indicates better precision in selecting the correct parking slot.
- **Collision Rate (CR)**: The percentage of parking attempts where the vehicle collides with an obstacle (e.g., other vehicles, walls, or dynamic objects). A lower CR is indicative of the system's ability to avoid collisions.

- **Timeout Rate (TR)**: The percentage of parking attempts where the vehicle fails to park within a specified time limit (e.g., 30 seconds). A lower TR indicates faster and more reliable parking performance.

These metrics provide a comprehensive framework for evaluating the system's accuracy, safety, and efficiency in performing automated parking tasks.

4.3 Closed-Loop Experiment Results

As shown in Fig. 5, we conducted closed-loop experiments to assess the performance of the proposed system in realistic parking scenarios. These experiments were carried out across 16 distinct parking lot layouts, designed to represent a wide range of real-world conditions. Each layout was tested 24 times to account for variations in initial conditions, ensuring a robust evaluation of the system's adaptability, consistency, and reliability in diverse environments.

Table I. Performance Across Different Initial Conditions

	TSR (%)	TFR (%)	NTR (%)	CR (%)	TR (%)	APE (m)	AOE (deg)	APT (s)
Average	90.72	2.42	3.01	2.59	1.26	0.32	0.95	16.07

Note: Target Success Rate (TSR), Collision Rate (CR), Non-Target Rate (NTR), and Timeout Rate (TR) are computed over all trials, while Average Position Error (APE) and Average Orientation Error (AOE) are calculated only for successful parking episodes. Timeout Rate (TR) denotes the percentage of trials where the vehicle fails to complete parking within the predefined time limit of 30 seconds.

As shown in Table I, the proposed approach achieves a high Target Success Rate (TSR) of 90.72%, with an average position error (APE) of just 0.32 meters and a low orientation error (AOE) of 0.95 degrees. The average parking time (APT) was 16.07 seconds, which is competitive compared to existing systems. Moreover, the non-target rate (NTR) was relatively low at 3.01%, indicating that the system reliably parks the vehicle in the correct slot. The collision rate (CR) was 2.59%, suggesting that the system generally avoids collisions, although there is still room for improvement in scenarios with complex obstacles.

4.4 Ablation Studies

We conducted several ablation studies to evaluate the contributions of different system components, such as depth supervision, speed limitations, and the number of control steps, to the overall performance. These studies provide insight into how each element affects the system’s accuracy, efficiency, and robustness. The key findings are as follows:

Table II. Impact of Different Components of the System

Evaluation Metric	YOPO	Extrinsic reference matrix distortion	Non-fused multimodal	No speed limit	Step size control with step size 1
TSR	**90.72%**	76.31%	77.08%	82.17%	83.07%
CR	**2.59%**	10.33%	6.25%	5.33%	2.86%
APT	16.07 s	18.5 s	16.37 s	**13.23 s**	17.68 s

- **Without Depth Supervision**: Removing depth supervision led to a significant reduction in the Target Success Rate (TSR), emphasizing the critical role of accurate depth information in producing reliable BEV representations. Depth supervision enables the system to better understand spatial relationships and distances, which are essential for precise obstacle detection and parking accuracy.
- **Without Speed Limitation**: When the speed limit was removed, the Target Success Rate (TSR) dropped, and the collision rate increased significantly. This outcome highlights the crucial role of speed control in maintaining stability and ensuring precise parking maneuvers, especially in complex and confined environments.
- **Control Steps**: Testing various control step configurations showed that a 4-step setup achieved the optimal balance between accuracy and efficiency. Fewer steps resulted in reduced precision, as the model struggled to capture finer details of the parking trajectory. Conversely, increasing the number of steps offered negligible performance gains while introducing higher computational overhead.

The ablation results also provide insight into the learning process. The significant performance drop when removing depth supervision, as discussed above, indicates that the model relies heavily on depth information to understand spatial relationships. This suggests that the auxiliary depth prediction task effectively guides the BEV generation module to encode distance cues, which are essential for accurate parking in crowded scenarios. Similarly, the degradation without speed limitation (TSR from 90.72% to 82.17%, as shown in Table II) shows that the model learns to balance precision and stability through appropriate speed control. These observations demonstrate that the proposed components contribute meaningfully to the learned representations and decision-making.

4.5 Comparison with Expert and Rookie Drivers

To further validate the performance of our system, we conducted a comparative analysis against the parking performance of expert and rookie human drivers. The evaluation was based on key metrics, including Target Success Rate (TSR), Average Position Error (APE), Average Orientation Error (AOE), and Average Parking Time (APT), as detailed in Table III.

Table III. Comparison with Expert and Rookie Drivers

Operator	TSR (%) ↑	TFR (%) ↓	CR (%) ↓	APE (m) ↓	AOE (deg) ↓	APT (s) ↓
YOPO	90.72	2.42	2.59	0.32	0.95	16.07
Expert Drivers	100.0	0.0	0.0	0.25	0.51	15.02
Rookie Drivers	75.0	18.75	6.25	0.35	4.00	20.13

This comparison highlights the system's ability to match or exceed human-level performance in terms of precision, efficiency, and consistency across various parking scenarios. The proposed system achieved performance comparable to expert drivers in terms of TSR, APE, and APT. Compared to rookie drivers, the system demonstrated significantly higher accuracy and faster execution, highlighting its potential to assist novice drivers by providing precise and efficient parking capabilities.

4.6 Comparison with Existing End-to-End Parking Methods

To further evaluate the effectiveness of the proposed YOPO framework, we compare our method with a representative end-to-end parking baseline, namely E2E Parking [20]. Both methods are evaluated under the same CARLA simulation environment and employ similar evaluation metrics. According to Yang et al. [20], E2E Parking achieves a target success rate of approximately 86% with larger average position and orientation errors. In contrast, YOPO achieves a higher TSR of 90.72%, along with lower APE and AOE, indicating improved spatial consistency and control accuracy. These results demonstrate that refined BEV sensing and auxiliary supervision contribute to more reliable parking performance in complex scenarios.

5 Conclusion

5.1 Limitations and Discussion

Despite the promising performance demonstrated by the proposed YOPO framework, several limitations should be acknowledged. First, the training strategy relies primarily on imitation learning from expert demonstrations. Although imitation learning enables efficient training, it may suffer from covariate shift when the vehicle encounters states that deviate from the expert distribution, potentially leading to suboptimal behavior in rare or unseen scenarios. Future work will investigate data aggregation strategies such as Dagger and hybrid imitation and reinforcement learning to mitigate this issue.

Second, all evaluations in this study are conducted within the CARLA simulator. While CARLA provides realistic physics and diverse parking layouts, a gap inevitably exists between simulation and real-world deployment. Factors such as sensor noise, illumination changes, weather conditions, and hardware latency are not fully captured in simulation. Extensive real-world experiments will be required to validate the robustness and generalization of the proposed approach.

Third, although refined BEV sensing improves spatial consistency, the current system assumes accurate camera calibration. Performance degradation may occur under extrinsic parameter disturbances, as

reflected in the ablation studies. Incorporating online calibration or self-supervised adaptation mechanisms represents an important direction for future research.

Finally, the proposed method focuses on static or mildly dynamic parking scenarios. Highly dynamic environments with dense pedestrian interactions remain challenging and warrant further investigation. Addressing these limitations will be essential for deploying end-to-end autonomous parking systems in real-world applications.

Regarding real-time applicability, the YOPO framework is designed with computational efficiency in mind. The BEV representation reduces 3D data to 2D grids, enabling standard 2D convolutions, and the transformer modules are kept shallow (4 layers). These design choices suggest potential for real-time deployment with appropriate optimization. However, actual inference speed depends on hardware implementation and has not been measured in this study; this will be investigated in future work.

5.2 Conclusion

This paper proposed an end-to-end autonomous parking system utilizing refined BEV sensing and a neural network-based architecture. By directly mapping multi-view BEV features and vehicle motion states to control signals, the system eliminates the need for traditional rule-based methods and demonstrates strong adaptability to complex parking scenarios. The main contributions are as follows:

1. Improved environmental perception by refining BEV representations, enabling precise parking in tight and dynamic spaces.
2. Developed an end-to-end neural network that simplifies system design while improving robustness and accuracy.
3. Demonstrated high performance through closed-loop experiments in the CARLA simulator, achieving a 90.72% Target Success Rate (TSR) with low position error (0.32 m) and orientation error (0.95°).
4. Directly mapped sensor inputs to control signals, reducing error accumulation and enhancing system reliability.

Future work will focus on real-world testing, improving dynamic obstacle handling, integrating additional sensors, and optimizing system performance for real-time applications. This work provides a robust foundation for advancing autonomous parking technologies and adapting them to real-world challenges.

Declarations

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Funding

This work was supported by the Shanxi Provincial Major Science and Technology Special Project (Grant No. 202501060302002).

Author Contribution

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Xuguang Sun, Zhengwu Fan, Ruiliang Zhang, Jiaxiong Wu and Xiaoce Sun. The first draft of the manuscript was written by Xuguang Sun and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgement

This research was supported by the Shanxi Provincial Major Science and Technology Special Project (Grant No. 202501060302002).

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

References

1. Xiong, L., Gao, J., Fu, Z. & Xiao, K. Path planning for automatic parking based on improved hybrid A* algorithm, in Proc. 5th CAA Int. Conf. Vehicular Control and Intelligence (CVCI), pp. 1–5. (2021).
2. Qin, T., Chen, T., Chen, Y. & Su, Q. AVP-SLAM: Semantic visual mapping and localization for autonomous vehicles in the parking lot, in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), pp. 5939–5945. (2020).
3. Jang, C. et al. Re-plannable automated parking system with a standalone around view monitor for narrow parking lots. *IEEE Trans. Intell. Transp. Syst.* **21** (2), 777–790 (2019).
4. Zhou, B. & Krähenbühl, P. Cross-view transformers for real-time map-view semantic segmentation, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 13760–13769. (2022).
5. Shen, X. et al. ParkPredict: Motion and intent prediction of vehicles in parking lots, in Proc. IEEE Intell. Veh. Symp. (IV), pp. 1170–1175. (2020).
6. Bojarski, M. et al. End to end learning for self-driving cars. *arXiv preprint arXiv* :1604.07316, 2016.
7. Fan, H. et al. Baidu Apollo EM motion planner. *arXiv preprint arXiv* :180708048, (2018).

8. Prakash, A., Chitta, K. & Geiger, A. Multi-modal fusion transformer for end-to-end autonomous driving, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 7077–7087. (2021).
9. Chen, L. et al. End-to-end autonomous driving: Challenges and frontiers. *IEEE Trans. Pattern Anal. Mach. Intell.* **46** (2), 1234–1250 (2024).
10. Hu, Y. et al. Planning-oriented autonomous driving, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 17853–17862. (2023).
11. Zhang, P. et al. Reinforcement learning-based end-to-end parking for automatic parking system. *Sensors* **19** (18), 3996 (2019).
12. Xie, E. et al. M²BEV: Multi-camera joint 3D detection and segmentation with unified birds-eye view representation. *arXiv preprint* (2022). arXiv:2204.05088.
13. Li, Y. et al. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection, in Proc. AAAI Conf. Artif. Intell., vol. 37, no. 2, pp. 1477–1485. (2023).
14. Huang, J. et al. BEVDet: High-performance multi-camera 3D object detection in bird-eye-view, arXiv preprint arXiv:2112.11790, (2021).
15. Li, Z. et al. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *Eur. Conf. Comput. Vis. (ECCV)*, pp. 1–18. (2022).
16. Codevilla, F. et al. Exploring the limitations of behavior cloning for autonomous driving, in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 9329–9338. (2019).
17. Du, J., Su, S., Fan, R. & Chen, Q. Bird's eye view perception for autonomous driving, in Autonomous Driving Perception: Fundamentals and Applications, Springer, pp. 323–356. (2023).
18. Philion, J. & Fidler, S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. *Eur. Conf. Comput. Vis. (ECCV)*, pp. 194–210. (2020).
19. Wu, P. et al. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. *Adv. Neural Inf. Process. Syst.* **35**, 6119–6132 (2022).
20. Yang, Y. et al. E2E parking: Autonomous parking by the end-to-end neural network on the CARLA simulator, in Proc. IEEE Intell. Veh. Symp. (IV), pp. 2375–2382. (2024).
21. Li, R. et al. An end-to-end fully automatic bay parking approach for autonomous vehicles, in Proc. Dyn. Syst. Control Conf., vol. 51906, pp. V002T15A004. (2018).
22. Shen, X. et al. ParkPredict+: Multimodal intent and motion prediction for vehicles in parking lots with CNN and transformer, in Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC), pp. 3999–4004. (2022).
23. Shan, Y. et al. DeepParking: Deep learning-based planning method for autonomous parking, in Int. Conf. Spatial Data Intell., Springer, pp. 21–33. (2023).
24. Zhong, J., Liu, Z. & Chen, X. Transformer-based models and hardware acceleration analysis in autonomous driving: A survey. *arXiv preprint* (2023). arXiv:2304.10891.
25. Caesar, H. et al. nuScenes: A multimodal dataset for autonomous driving, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), (2020).

26. Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks, in Proc. Int. Conf. Mach. Learn. (ICML), PMLR, (2019).
27. Zhang, J. et al. Automatic parking trajectory planning in narrow spaces based on Hybrid A* and NMPC. *Sci. Rep.* **15**, 1384. https://doi.org/10.1038/s41598-025-85541-x (2025).
28. Wang, L. & Huang, Y. Fast vehicle detection based on colored point cloud with bird 's eye view representation. *Sci. Rep.* **13** https://doi.org/10.1038/s41598-023-34479-z (2023).

Figures

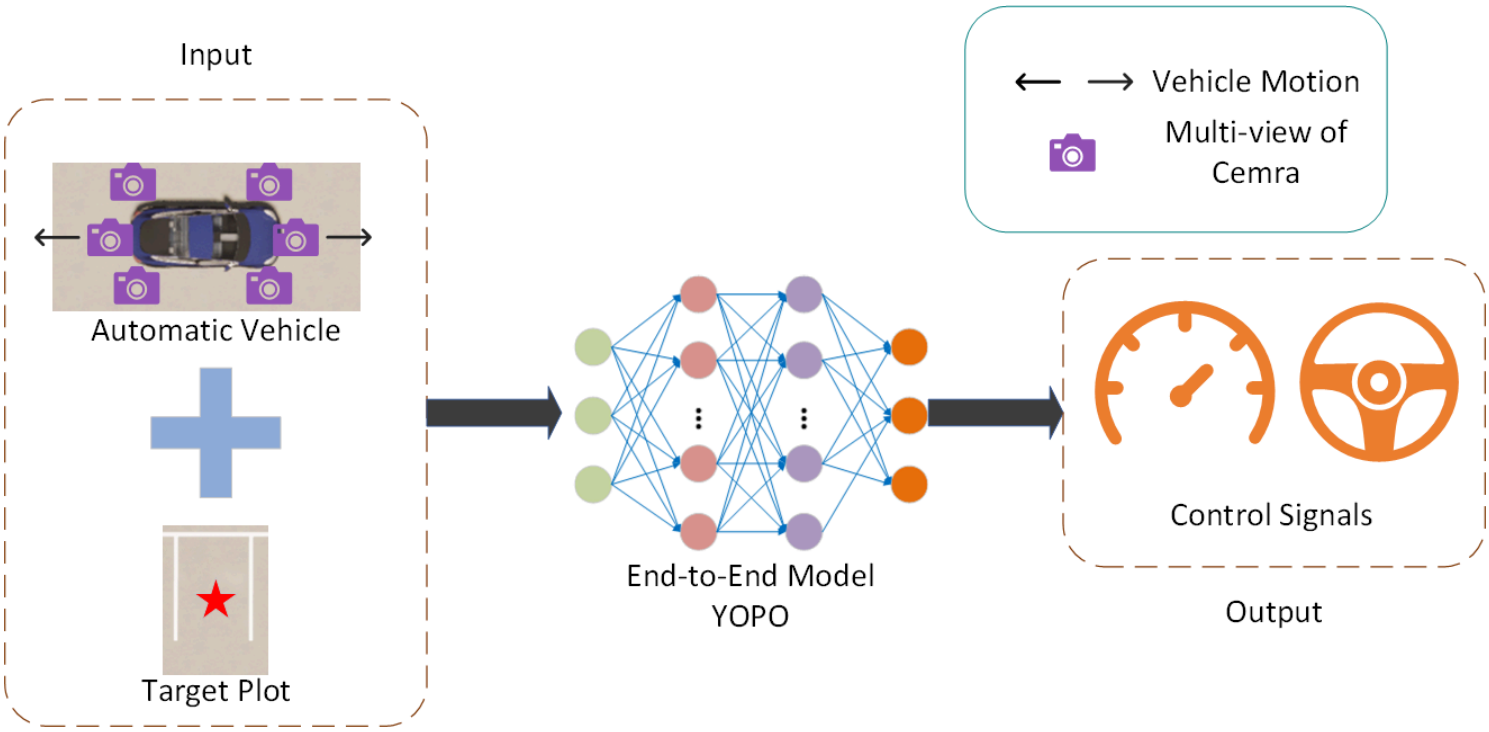


Figure 1

Overall structure of the YOPO automatic parking system.

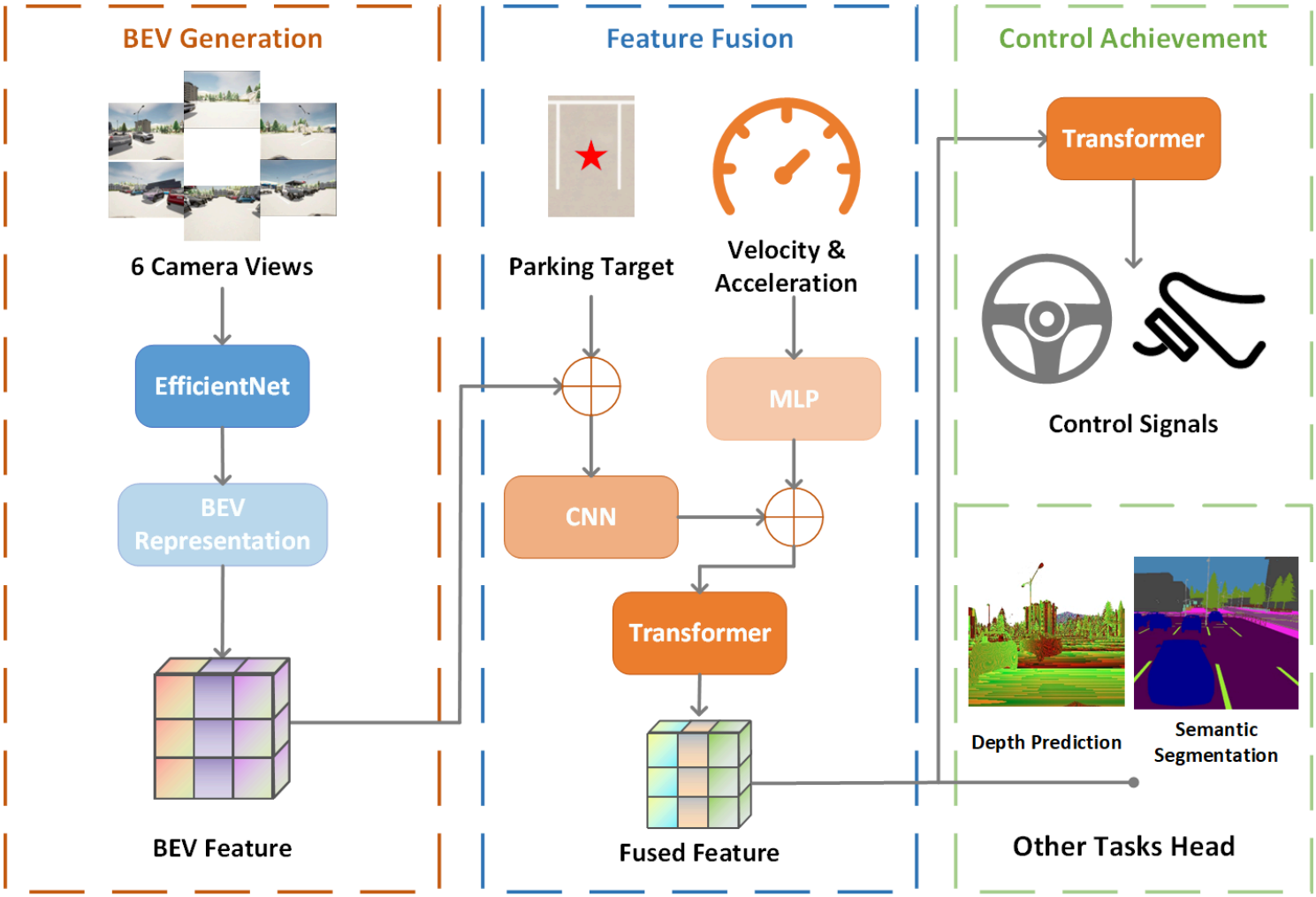


Figure 2

YOPO system architecture including BEV generation, feature fusion, and control.

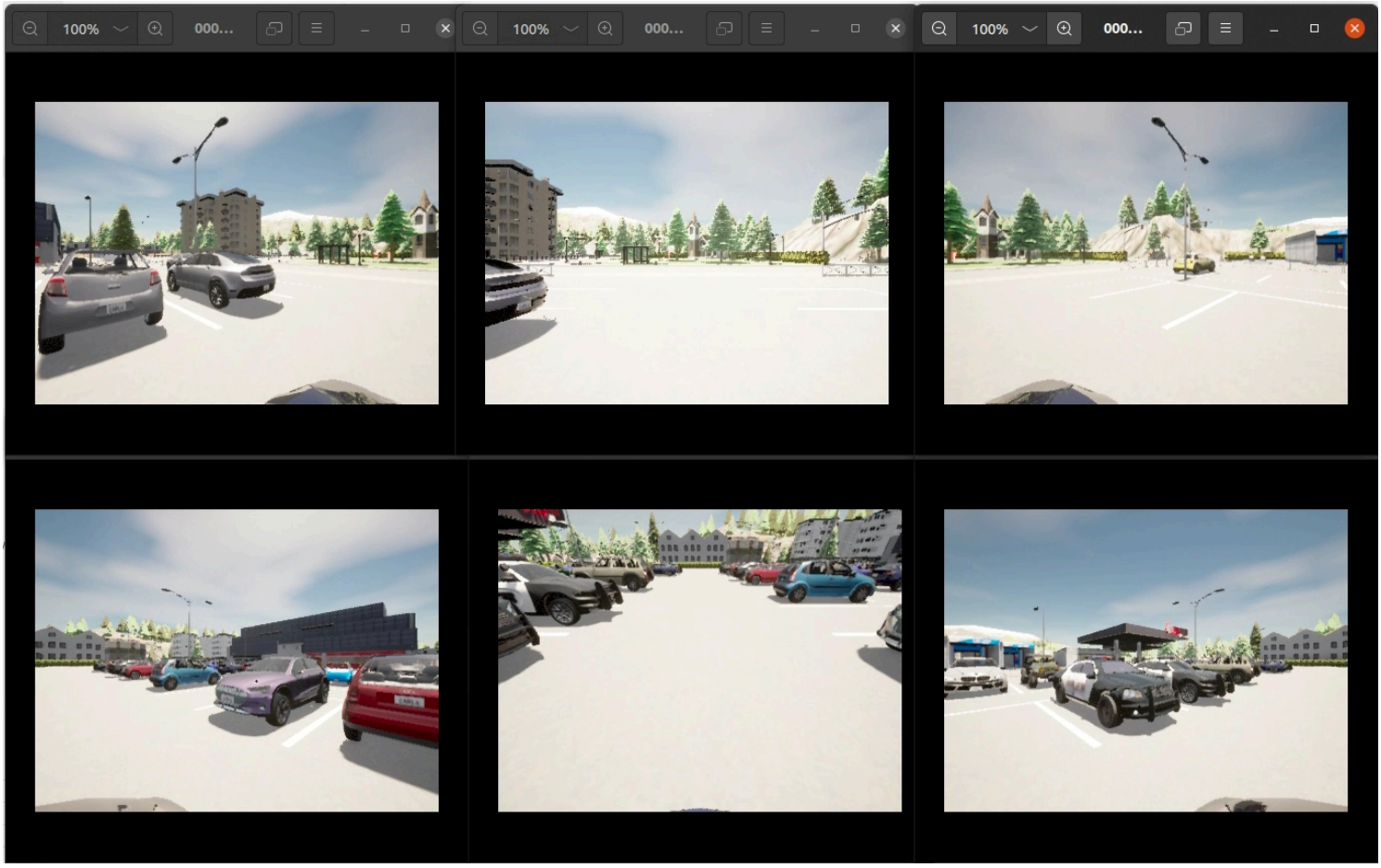


Figure 3

BEV perception using six camera views.

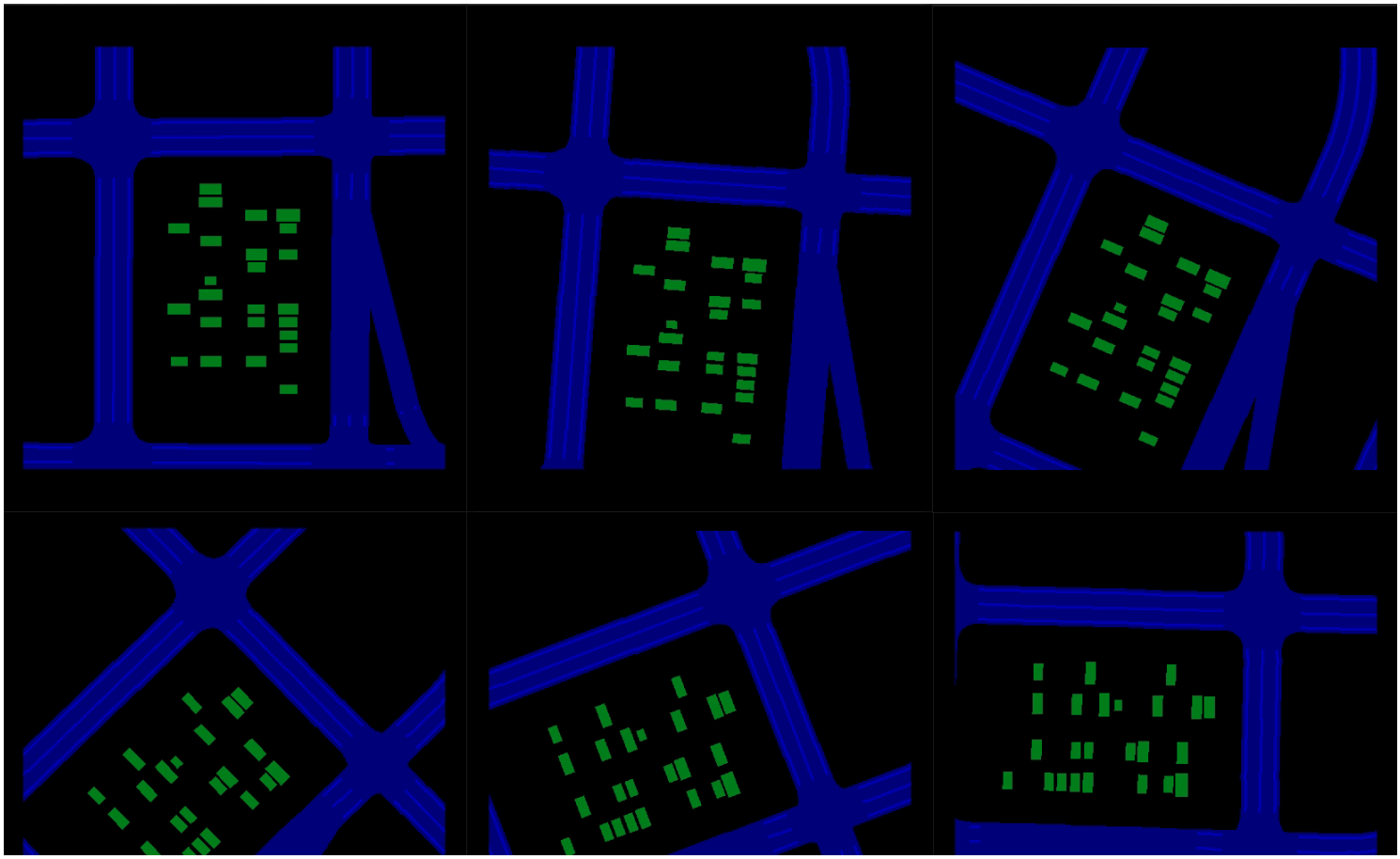


Figure 4

BEV representation in parking scenarios.



Figure 5

Full automatic parking process with six stages.