Deep Reinforcement Learning Based Tractor-Trailer Tracking Control

Qi Kang, Hartmannsgruber Andreas, Sze-Hui Tan, Xiang Zhang, Chee-Meng Chew

***Abstract*—This paper introduces a method for controlling an autonomous truck-trailer system using policy-based deep reinforcement learning. Approaches for planning and controlling traditional passenger vehicles cannot be easily transferred to this special system because its physical properties like size, kinematics, and dynamics are much more complex. To manage this complexity, we make use of a new and trending artificial intelligence tool: deep reinforcement learning. The proximal policy optimization is adopted for training the agent with a reward function defined to penalize deviation and reward following a given track. We chose a B-spline-based lane representation to improve the (partial) observation of the agent's state in the environment. The policy is trained and validated in the Pybullet simulator and our trained agent achieves good capability in navigating random and complex tracks.**

***Index Terms*—Deep reinforcement learning, POMDP, Tractor-Trailer system.**

I. Introduction

THE use of large dimension vehicles, especially tractor-trailer systems, is ubiquitous in the transportation industry and the efficient and safe operation of these vehicles is crucial for ensuring timely and cost-effective transportation of goods. However, operating these large vehicles in real-world scenarios can be challenging due to their size, weight, and complex dynamics. Due to their size, these vehicles require more space to maneuver, and their heavy weight causes high momentum, which is an added burden to velocity planning and control. Furthermore, large-dimension vehicles have complicated designs. Tractor-trailer systems, for example, consist of multiple articulated segments that are interconnected through a kingpin which allows for greater cargo capacity but makes the vehicle difficult to control. The driver must be able to coordinate the movement of the tractor and trailer segments to ensure that the vehicle is stable and balanced, and the cargo secure. For this reason, algorithms for motion planning and control of large-dimension vehicles need to consider a bigger parameter space (i.e., configuration space).

Deep Reinforcement Learning (DRL) has shown promising results in estimating various complex control tasks, including autonomous vehicles [1] [2] and robotics [3]. In recent years, researchers have started applying DRL algorithms to tractor-trailer systems to enhance their maneuverability and stability [4]. One popular DRL algorithm used in these applications is Proximal Policy Optimization (PPO), which has demonstrated superior performance in continuous control tasks [5], [6].

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

This paper focuses on using the PPO algorithm to control large-dimension vehicles like buses and truck-trailer systems and addresses some of the challenges faced by them. With careful design of reward functions, observation spaces, and formulation of the reference path, we conducted successful training of a unique agent capable of accurately following a reference path For training of the agent, we auto-generate a big variety of reference paths utilizing a B-spline curve representation. Finally, to enhance the generalization capacity of the learned policy, we have adopted an incremental control scheme, complemented by a local coordinate system.

The remainder of this paper is structured as follows: Section II reviews the state-of-the-art related to this publication. We first introduce the background of tractor-trailer systems, including their dynamics and control challenges, followed by a discussion on the principles and fundamentals of DRL, focusing on the PPO algorithm. In Section **??**, we introduce the feature of the training environment and vehicle model we crafted for simulation, as well as the B-spline representation of the reference path and the kinematics of tractor-trailer systems. Section **??** discloses the most vital part of the paper, which are the state and action space we adopted to interact with the environment and the reward functions we designed to train the agent to maneuver along the reference path closely. Section V presents the learning curve with different reward functions and constraints on the policy as well as the performance of our proposed method. Section VI concludes this publication with a discussion of the results and an outlook for future directions.

II. Background

Due to their exceptional load-carrying capabilities, unparalleled freight efficiency, and cost-effective transportation solutions, truck-trailer vehicles have garnered immense demand and have become the backbone of modern commercial material transportation and military equipment loading and transportation operations. The research on truck-trailer control can be traced back to the 1980s [7] and start with the investigation of non-holonomic kinematics, which is the common dynamic structure of the truck-trailer vehicle. In this section, we will discuss the methods applied to truck-trailer including canonical control theory methods and learning-based approaches. Interested audiences are referred to [8], [9]for the comprehensive literature review of Truck Trailer control. Moreover, in the ever-evolving realm of autonomous driving, the challenge of tracking control holds tremendous significance, propelling a surge of interest in Deep Reinforcement Learning (DRL) owing to its extraordinary abilities in

representation and experiential learning [1]. This has sparked an emerging trend wherein DRL is being harnessed for vehicle tracking control, forming the pivotal focus of inquiry in this paper. Within the confines of this study, we embark upon a comprehensive exploration, providing a succinct yet insightful introduction to the development and classification of DRL methods. As the narrative progresses, we aim to culminate this section by shedding light on the exceptional policy-based DRL algorithm known as Proximal Policy Optimization (PPO), an outstanding innovation that has garnered immense attention and acclaim within the field.

A. Truck Trailer Control

Truck trailer control plays a crucial role in the efficient and safe operation of heavy-duty trucks and trailers. It involves the coordination and management of various maneuvers, such as forward and backward control [10], [11], dynamic obstacle avoidance [12], and parallel parking [13]. To date, researchers have developed many algorithms as Truck trailer control solutions, which significantly contributed to enhancing the performance and reliability of truck trailer control systems. To be more organized, we roughly categorize them into several groups, namely (1)PID control, (2)Linear quadratic regulator control (LQR), (3)Model predictive control (MPC), (4)Sliding mode control (SMC) (5)fuzzy control(6) learning-based methods.

PID is a commonly used control theory-based approach that doesn't need to establish an accurate mathematical model and is easy to implement. [14] shows that the linear reversing problem of tractor–trailer vehicles can be solved by deploying a PD controller based on feedback control theory. As for the non-linear truck trailer control system, Ye [15]managed to combine the traditional PID controller with a neural network to solve the controlling condition with nonlinearity, uncertainty, and disturbance. By considering the kinematic and dynamic model of truck-trailer vehicles, Khanpoor et al. [16] combined the Lyapunov method with PID control and validated the stability and effectiveness of the proposed algorithm with simulation and experimental results. Adjusting PID parameters appropriately can yield satisfactory results in theory. However, in the presence of unknown or time-varying models, parameter adjustment becomes challenging, leading to poor performance and vulnerability to external disturbances.

LQR is generally used in non-linear truck trailer systems. After establishing the vehicle system model and pre-linearizing the truck-trailer system with feedback linearization, the optimal state feedback control can be achieved by solving a linear quadratic optimization objective. Chen et al. [17] applied the optimal control theory to design a linear quadratic controller for the straight-line reversing of truck–trailer vehicles. The linear quadratic optimal control problem has formed a standardized solution process with its superiority proven in different working conditions [18], [19], but it cannot handle the multi-variable constraint problem in the vehicle driving control process, and the solution process is normally computational-complex.

MPC usually solves the optimal control problem of truck trailers by predicting the future output behavior of the system according to the dynamic model and current state. Kayacan et al. [20] combine a nonlinear model predictive controller (NMPC) with a nonlinear motion level estimator (NMHE) to generate a control strategy to improve the tracking accuracy of Truck-Trailer. Moreover, they dig further in [21] by linearizing the input state of the original model, from which they can perform linear model predictive control (LMPC) and propose the ISL-LMPC control scheme for the truck trailer. The experimental results showed that ISL-LMPC possesses higher control efficiency. MPC has good robustness, convenient modeling, and good dynamic performance. However, the computational complexity of the MPC keeps it from a real-world application.

SMC alters the dynamics of the truck trailer systems by applying a discontinuous control signal to force the system to slide along a pre-designed sliding mode surface to enter the system's steady state. [22] proposed an SMC controller with a non-linear disturbance observer for tracking control of tractor-trailer vehicles and compared it with back-stepping control and model predictive control through experiments. [23] combined MPC and SMC on the kinematic and dynamic models and achieved trajectory tracking of tractor-trailer vehicles. In general, SMC can overcome the uncertainty of the control system and has good robustness and a fast response speed. But it needs to overcome the chattering phenomenon that occurs when the state trajectory reaches the sliding mode surface.

Fuzzy control is an intelligent control method that uses multi-valued fuzzy logic and simplified reasoning principles without relying on precise mathematical models during the control process. In terms of tractor-trailer vehicle reversing control, a fuzzy control method was designed using linear matrix inequalities (LMIs) [24]. Two independent fuzzy controllers were designed in [25], which were used for target search and obstacle avoidance respectively to solve the parking problem. Although it has strong adaptability, the fuzzy control algorithm has strong subjectivity, which always produces large errors.

In recent years, learning-based approaches have emerged as a powerful tool for truck trailer control. These methods leverage artificial intelligence and machine learning techniques to train algorithms that can adapt and improve their performance over time. Aiming at the formation tracking control problem of tractor-trailer vehicles, Shojaei [26] built a multi-layer neural network controller and adopted a dynamic surface control method, which effectively reduced the complexity of the controller. Similarly, [27] proposes a BP Neural Network-based controller for improving the truck-trailer system's stability and validated its efficiency in handling huge slip angle of the vehicle with Matlab platform [28] proposed a neuro-fuzzy controller based on preview control and deep reinforcement learning to tackle backward parking control of tractor-trailer vehicles. [29] proposes a novel path-planning approach based on semi-supervised learning. By training an encoder-decoder type of deep neural network to plan paths with the objective to minimize the off-track of the tractor-trailer swept area. Learning-based methods have good nonlinear fitting ability, and mapping abilities, which is convenient for computer processing and enables it to be a fashionable trend in the

autonomous truck trailer society.

B. Deep Reinforcement learning

Deep reinforcement learning (DRL) integrates the feature representation ability of deep learning with the decision-making ability of reinforcement learning so that it can achieve powerful end-to-end learning control capabilities. With the advent of deep neural networks and advancements in computing power, attention and success on DRL is rapidly growing these years. One of the breakthrough moments was the introduction of Deep Q-Networks (DQN) by DeepMind in 2013 [30], which achieved human-level performance on several Atari 2600 games. For more applications and details of each reinforcement learning algorithm, readers can refer [31].

Creating a comprehensive taxonomy for modern RL algorithms is challenging due to their modular nature, which defies a tree structure. A crucial division arises based on whether the agent possesses or learns a model of the environment, i.e. the function that predicts state transitions and rewards. Algorithms that use a model are called model-based methods, such as Model-Based Value Expansion (MBVE) [32], and Monte Carlo Tree Search (MCTS) [33], and the rest are called model-free. Although model-free methods sacrifice sample efficiency, they offer easier implementation and tuning. Consequently, model-free methods are more popular, extensively developed, and tested compared to model-based methods. There are two main branches to representing and training agents with model-free RL: Policy-based and value-based methods.

The value-based algorithms aim to estimate the value of different states or state-action pairs. Deep Q-Networks (DQN) [30] is a typical value-based method. It combines deep neural networks with the Q-learning algorithm to estimate the Q-values of state-action pairs. It also uses experience replay and a target network to stabilize training [34]. Several DQN variants like Double DQN [35] and Dueling DQN [36] are proposed to improve the quality of the algorithm.

Policy-Based algorithms directly learn a policy to determine the agent's actions. Methods in this family represent a policy explicitly as $\pi_\theta(a|s)$. They optimize the parameters θ either directly by gradient ascent on the performance objective $J(\pi_\theta)$, or indirectly, by maximizing local approximations of $J(\pi_\theta)$.Within this category, Asynchronous Advantage Actor-Critic (A3C) [37] is an algorithm that combines the advantages of asynchronous training and the Actor-Critic framework. It uses multiple parallel agents to explore the environment independently and updates a global policy and value function based on their experiences. Trust Region Policy Optimization (TRPO) [38] and Proximal Policy Optimization (PPO) [5], [6] also fall into this domain.

The two families mentioned above are not completely incompatible, there still exists a range of algorithms that live between these two extremes: Deep Deterministic Policy Gradient (DDPG) [39], Twin Delayed Deep Deterministic Policy Gradient (TD3) [40], and Soft Actor-Critic (SAC) [41]. Take DDPG as an example, it extends the Q-learning algorithm to continuous action spaces and uses an actor-critic architecture with separate networks for the policy (actor) and Q-value estimation (critic). Furthermore, DDPG employs off-policy learning and can utilize a replay buffer for experience replay to improve training performance.

Having understood the mainstream DRL algorithms, let's delve into PPO, our chosen algorithm for implementation. PPO is a policy optimization algorithm that aims to find an optimal agent policy, striking a balance between exploration and exploitation while ensuring stable and efficient training. It employs a surrogate objective function to update policy parameters, constraining policy updates to maintain stability and prevent drastic changes that could adversely affect performance.PPO excels in handling high-dimensional action spaces and continuous control problems, outperforming other algorithms. It achieves superior sample efficiency by optimizing data utilization and exploration. The inclusion of a clipping mechanism ensures training stability, preventing divergence and catastrophic forgetting. Its simplicity and strong performance have made it widely embraced in research and industry. Additionally, PPO exhibits excellent generalization capabilities across various tasks and environments. The versatility of PPO has led to successful applications in a wide range of reinforcement learning domains, including robotics, game playing, autonomous navigation, and natural language processing. It has proven its efficacy in both simulated environments and real-world scenarios, solidifying its position as a powerful and reliable algorithm.

In this paper, our primary focus lies in the realm of forward tracking control for trucking trailer vehicles within obstacle-free environments. To tackle this challenge, we employ the powerful and sophisticated Proximal Policy Optimization (PPO) learning algorithm.

III. Problem Formulation

A. B-spline Reference path generation

The term "B-spline" was coined by Isaac Jacob Schoenberg and is short for basis spline [TODO: Reference]. A spline function of order n is a piece-wise polynomial function of degree $n-1$ in a variable x. The places where the pieces meet are known as knots. The key property of spline functions is that they and their derivatives may be continuous, depending on the multiplicities of the knots. B-splines of order n are basis functions for spline functions of the same order defined over the same knots, meaning that all possible spline functions can be built from a linear combination of B-splines, and there is only one unique combination for each spline function. In the training environment, we represent the tracks using B-splines to generate curves that are discretized into waypoints. By randomly selecting the control points, we can generate different tracks for each training episode and enhance the model's ability to generalize. We employ 3rd-degree B-spline curves with 4 control points to manipulate its shape.

An example of the sampled curve is shown in Fig.1.

B. Truck Trailer kinematics

To simplify the control problem, we employ a Kinematic Single-Track Model with One On-axle Trailer (KST) to be our vehicle model. The KST depicts a simplified tractor-trailer

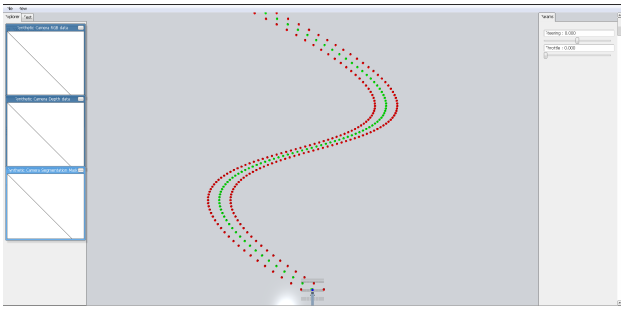

Fig. 1: B-Spline Track for Training.

system with only three wheels, where the left and right wheels are represented by only one wheel in the center of each pair. This simplification is justified since the roll dynamics are not considered. The KST also does not consider any tire slip, so the velocity vector v at the center of the rear axle is always aligned with the link between the front and rear wheel as depicted in Fig. 2. Similar to the point-mass model, the kinematic single-track model is used in many works for motion planning [TODO: References using a KST, preferably for truck-trailer]. To derive the differential equation of the

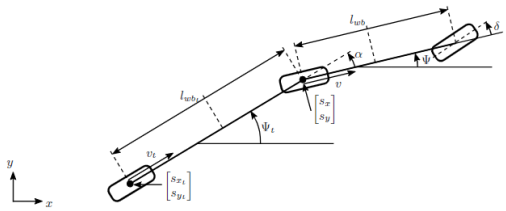

Fig. 2: Kinematics.

KST, we need velocity v, the velocity of the steering angle V_δ, the steering angle δ, the heading ψ, and the parameters l_{wb}, l_{wbt} describing the wheelbase of the truck and the trailer, respectively. The hitch angle α is the difference in orientation between the truck and the trailer.

The differential equations of the KST are:

$$\dot{\delta} = V_\delta \tag{1a}$$

$$\dot{\psi} = \frac{v}{l_{wb}} \tan\delta \tag{1b}$$

$$\dot{\alpha} = -v\left(\frac{\sin\alpha}{l_{wbt}} + \frac{\tan\delta}{l_{wb}}\right) \tag{1c}$$

$$\dot{s_x} = v\cos\psi \tag{1d}$$

$$\dot{s_y} = v\sin\psi \tag{1e}$$

IV. Methodology

A. Observation and Action Space

The mathematical description of our problem includes the vehicle state and lane information. The vehicle is supposed to follow arbitrary lanes, which can turn out to be quite challenging for large, multi-body vehicles. For the vehicle state, we employ the orientation, velocity, and hitch angle between the tractor and the trailer. Our agent uses only two waypoints ahead in the vehicle reference coordination system, which can somehow equip the agent with the ability to anticipate the future condition of the road. The coordinates are all in the vehicle coordinate system, with the origin at the center of the rear axle, the x-axis along the vehicle's longitudinal axis, and the y-axis perpendicular, in the lateral direction. The info we need is these two waypoints' distance to the vehicle coordination origin and their coordination with respect to that [TODO: sentence needs re-write; unclear and grammar issues].

Equations (1) and (2) show the formula to calculate the local coordinates with the help of the waypoints' global coordinates (X_w, Y_w), vehicle yaw angle ψ, and vehicle world coordinates (X_v, Y_v). In summary, the observation space for our proposal is listed in Table 1:

$$X_{local} = (X_w - X_v) * cos\psi + (Y_w - Y_v) * sin\psi \tag{2}$$

$$Y_{local} = -(X_w - X_v) * sin\psi + (Y_w - Y_v) * cos\psi \tag{3}$$

In order to mimic the driving behavior of human drivers and simplify the control problem, we use two control commands to control the vehicle, i.e., throttle and steering angle, for the action space.

TABLE I: Observation Sapce

$(\cos\psi, \sin\psi)$	Orientation of the tractor
V	Velocity of tractor
ω	angular Velocity of tractor
D_1	Distance between vehicle and the 1st waypoint
D_2	Distance between vehicle and the 2nd waypoint
(X_1, Y_1)	Local Coordination of the 1st waypoint
(X_2, Y_2)	Local Coordination of the 2nd waypoint
α	Hitch angle between tractor and trailer

B. reward function

Since the target of our proposed method is to let the vehicle proceed without off-tracking as far as possible and keep to the reference line, a binary reward structure cannot fulfill our needs. In that case, We formulate a hybrid reward system as follows:

$$R = r_g + r_p + r_t + r_d + r_o \tag{4}$$

The whole reward structure is composed of 5 parts which are the reward for reaching waypointsr_g, the reward for proceedingr_p, the punishment for exceeding the max training timestepr_e, the punishment for deviation from the reference liner_d, and the punishment for the off-track error r_o. The full equation and definition of each part are shown below, and the specific notation used in the reward function is listed in TableII :

$$r_g = \begin{cases} 10, & \text{if } D_1 < D_r, \\ 0, & \text{otherwise.} \end{cases}$$

$$r_e = \begin{cases} -100, & \text{if } t > \text{Max Timesteps}, \\ 0, & \text{otherwise.} \end{cases}$$

$$r_o = \begin{cases} -100, & \text{if } Max(d_1, d_2) > D_o, \\ 0, & \text{otherwise.} \end{cases}$$

$$r_d = (d_1 + w * d_2)^2 * V$$

$$r_p = (DG_t - DG_{t-1})^2 * K$$

For large dimension vehicles such as tractor-trailer systems, it is hard to get an optimal policy that can follow the reference path without off-tracking by only taking the deviation of a single center point to the reference path. In that case, we employ the center points of both the tractor's and trailer's rear axles to calculate the lateral deviation, which are d_1 and d_2. Since the calculation is the same, we take d_1 as an example and treat the distance between the center point of the tractor's rear axle and the nearest segment of the reference path as the front deviation. Since it is difficult to directly calculate the distance between a point and the segment on the B-spline curve, we use the vector pointing from the first waypoint to the second waypoint ahead of the center point on the tractor to represent it [TODO: minor re-write for clarity]. Equation 5 calculates the lateral deviation with (x_1, y_1) and (x_2, y_2) as the coordinates of the first and second waypoint ahead of the truck's center point denoted by (x_0, y_0). An illustration of our way of calculating the lateral deviation is in Fig. 3

TABLE II: Notations in Reward Function

D_r	Determine criteria for waypoint reaching
t	Current timestep
d_1	Lateral deviation of the tractor
d_2	Lateral deviation of trailer
D_o	Safety Range for Lateral Deviation
V	Coefficient of deviation reward
K	Coefficient of proceeding reward
DG	Distance between tractor and the first ahead point

$$d = \frac{|(x_2 - x_1) * (y_1 - y_0) - (x_1 - x_0) * (y_2 - y_1)|}{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}} \quad (5)$$

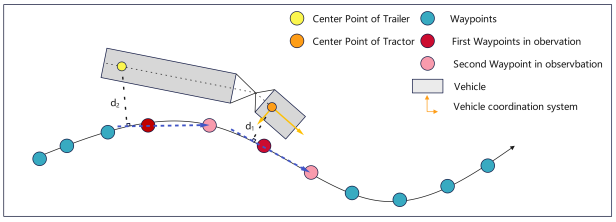

Fig. 3: Illustration of the Deviation Calculation.

V. Simulation Results

VI. Conclusion and Future Work

In this paper, we present a novel approach to tackle the challenges of large-dimension vehicle tracking control policies for truck trailer systems using deep reinforcement learning methods. Our work encompasses the design of a dedicated DRL training environment for the truck trailer vehicle model. Through meticulous crafting of reward functions, observation spaces, and reference path formulation, we achieve successful training of the truck trailer to flawlessly track randomly generated B-spline reference lanes and execute sharp turning maneuvers while avoiding off-tracking. The effectiveness of the trained policy's generalization is verified across various scenarios. Notably, we extend the applicability of our control policy and reward structure to develop a tracking control system for bus vehicles. Looking ahead, we aim to enhance our control scheme by integrating the speed profile and incorporating velocity-based observation spaces. Furthermore, we intend to explore the realm of obstacle avoidance in truck trailer control, which presents intriguing avenues for future research.

REFERENCES

[1] B. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. Sallab, S. Yogamani, and P. Perez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, 2022.

[2] S. Aradi, "Survey of deep reinforcement learning for motion planning of autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 740–759, 2022.

[3] J. Kober, J. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[4] C.-J. Hoel, K. Wolff, and L. Laine, "Automated speed and lane change decision making using deep reinforcement learning," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2148–2155.

[5] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv:1707.06347*, 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[6] N. Heess, D. TB, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. M. A. Eslami, M. A. Riedmiller, and D. Silver, "Emergence of locomotion behaviours in rich environments," *arXiv:1707.02286*, 2017. [Online]. Available: http://arxiv.org/abs/1707.02286

[7] J.-P. Laumond, "Feasible trajectories for mobile robots with kinematic and environment constraints," in *Intelligent Autonomous Systems, An International Conference*. NLD: North-Holland Publishing Co., 1986, p. 346–354.

[8] J. David and P. Manivannan, "Control of truck-trailer mobile robots: A survey," *Intelligent Service Robotics*, vol. 7, no. 4, pp. 245–258, 2014.

[9] Q. Liu, X. Li, Y. Zhu, and S. Li, "A review of tracking control method for tractor-trailer vehicles," in *Proceedings of the 2022 2nd International Conference on Control and Intelligent Robotics*, ser. ICCIR '22. Association for Computing Machinery, 2022, p. 863–867.

[10] A. C. Manav, I. Lazoglu, and E. Aydemir, "Adaptive path-following control for autonomous semi-trailer docking," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 1, pp. 69–85, 2022.

[11] J. Kolb, G. Nitzsche, S. Wagner, and K. Robenack, "Path tracking of articulated vehicles in backward motion," 2020, pp. 489–494.

[12] M. Aali and J. Liu, "Multiple control barrier functions: An application to reactive obstacle avoidance for a multi-steering tractor-trailer system," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 6993–6998.

[13] Z. Li, H. Cheng, J. Ma, and H. Zhou, "Research on parking control of semi-trailer truck," in *2020 4th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, 2020, pp. 424–429.

[14] Z. Zhu, J. Chen, and R. Torisu, "Automatic backward driving system of tractor-trailer combination along rectilinear path using pd control," *Nongye Jixie Xuebao/Transactions of the Chinese Society of Agricultural Machinery*, vol. 37, no. 7, pp. 98–100+73, 2006, cited By 6.

[15] J. Ye, "Adaptive control of nonlinear pid-based analog neural networks for a nonholonomic mobile robot," *Neurocomputing*, vol. 71, no. 7-9, pp. 1561–1565, 2008.

[16] A. Khanpoor, A. Khalaji, and A. Moosavian, "Modeling and control of an underactuated tractor-trailer wheeled mobile robot," *Robotica*, vol. 35, no. 12, pp. 2297–2318, 2017.

[17] J. Chen, B. Han, Z. Zhu, and J.-I. Takeda, "Optimal control for automatic backward driving system of tractor-trailer combination along rectilinear path," *Nongye Jixie Xuebao/Transactions of the Chinese Society of Agricultural Machinery*, vol. 39, no. 1, pp. 102–105, 2008.

[18] M. Sever, E. Kaya, M. Arslan, and H. Yazici, "Active trailer braking system design with linear matrix inequalities based multi-objective robust lqr controller for vehicle-trailer systems," vol. 2016-August, 2016, pp. 796–801.

[19] T. Sun, Y. He, E. Esmailzadeh, and J. Ren, "Lateral stability improvement of car-trailer systems using active trailer braking control," *Journal of Mechanics Engineering and Automation*, vol. 2, no. 9, pp. 555–562, 2012.

[20] E. Kayacan, H. Ramon, and W. Saeys, "Robust trajectory tracking error model-based predictive control for unmanned ground vehicles," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 2, pp. 806–814, 2016.

[21] E. Kayacan, W. Saeys, H. Ramon, C. Belta, and J. Peschel, "Experimental validation of linear and nonlinear mpc on an articulated unmanned ground vehicle," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 5, pp. 2023–2033, 2018.

[22] J. Taghia, X. Wang, S. Lam, and J. Katupitiya, "A sliding mode controller with a nonlinear disturbance observer for a farm vehicle operating in the presence of wheel slip," *Autonomous Robots*, vol. 41, no. 1, pp. 71–88, 2017.

[23] M. Yue, X. Hou, R. Gao, and J. Chen, "Trajectory tracking control for tractor-trailer vehicles: a coordinated control approach," *Nonlinear Dynamics*, vol. 91, no. 2, pp. 1061–1074, 2018.

[24] K. Tanaka, S. Hori, and H. Wang, "Multiobjective control of a vehicle with triple trailers," *IEEE/ASME Transactions on Mechatronics*, vol. 7, no. 3, pp. 357–368, 2002.

[25] M. Sharafi, A. Zare, and S. Nikpoor, "Intelligent parking method for truck in presence of fixed and moving obstacles and trailer in presence of fixed obstacles: Advanced fuzzy logic technologies in industrial applications," vol. 2, 2010, pp. V2268–V2272.

[26] K. Shojaei, "Neural network formation control of a team of tractor-trailer systems," *Robotica*, vol. 36, no. 1, pp. 39–56, 2018.

[27] Z. Yu, D. Xing, Q. Cao, and S. Li, "Research on the stability of the semi-trailer based on neural network control." IEEE Press, 2016, p. 1472–1476. [Online]. Available: https://doi-org.libproxy1.nus.edu.sg/10.1109/ICMA.2016.7558781

[28] E. Bejar and A. Moran, "A preview neuro-fuzzy controller based on deep reinforcement learning for backing up a truck-trailer vehicle," 2019.

[29] X. Zhang, J. Eck, and F. Lotz, "A path planning approach for tractor-trailer system based on semi-supervised learning," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 2022, pp. 3549–3555.

[30] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv:1312.5602*, 2013. [Online]. Available: http://arxiv.org/abs/1312.5602

[31] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction. 2nd edition," Cambridge, MA, USA: MIT Press, 2018.

[32] V. Feinberg, A. Wan, I. Stoica, M. I. Jordan, J. E. Gonzalez, and S. Levine, "Model-based value estimation for efficient model-free reinforcement learning," *arXiv:1803.00101*, 2018. [Online]. Available: http://arxiv.org/abs/1803.00101

[33] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 7559–7566.

[34] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba, "Hindsight experience replay," *arXiv:1707.01495*, 2017. [Online]. Available: http://arxiv.org/abs/1707.01495

[35] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[36] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *International conference on machine learning*, 2016, pp. 1995–2003.

[37] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," *arXiv:1602.01783*, 2016. [Online]. Available: http://arxiv.org/abs/1602.01783

[38] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," *arXiv:1502.05477*, 2015. [Online]. Available: http://arxiv.org/abs/1502.05477

[39] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv:1509.02971*, 2019. [Online]. Available: https://arxiv.org/abs/1509.02971

[40] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," *arXiv:1802.09477*, 2018. [Online]. Available: http://arxiv.org/abs/1802.09477

[41] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *arXiv:1801.01290*, 2018. [Online]. Available: http://arxiv.org/abs/1801.01290