



Article

Cooperative MARL-PPO Approach for Automated Highway Platoon Merging

Máté Kolat  and Tamás Bécsi * 

Department of Control for Transportation and Vehicle Systems, Budapest University of Technology and Economics, H-1111 Budapest, Hungary; mate.kolat@edu.bme.hu

* Correspondence: becsi.tamas@kjk.bme.hu

Abstract: This paper presents a cooperative highway platooning strategy that integrates Multi-Agent Reinforcement Learning (MARL) with Proximal Policy Optimization (PPO) to effectively manage the complex task of merging. In modern transportation systems, platooning—where multiple vehicles travel closely together under coordinated control—promises significant improvements in traffic flow and fuel efficiency. However, the challenge of merging, which involves dynamically adjusting the formation to incorporate new vehicles, remains challenging. Our approach leverages the strengths of MARL to enable individual vehicles within a platoon to learn optimal behaviors through interactions. PPO ensures stable and efficient learning by optimizing policies balancing exploration and exploitation. Simulation results show that our method achieves merging with safety and operational efficiency.

Keywords: deep learning; reinforcement learning; platooning; traffic merging; road traffic control; multi-agent systems



Citation: Kolat, M.; Bécsi, T. Cooperative MARL-PPO Approach for Automated Highway Platoon Merging. *Electronics* **2024**, *13*, 3102. <https://doi.org/10.3390/electronics13153102>

Academic Editors: Martin Reisslein and Felipe Jiménez

Received: 6 June 2024

Revised: 11 July 2024

Accepted: 1 August 2024

Published: 5 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the rise of autonomous driving technology and the communication systems between them, such as Vehicle-to-vehicle (V2V) [1], Vehicle-to-infrastructure (V2I) [2], or Vehicle-to-everything (V2X) [3] platooning can play a significant role in the automotive sector, aspiring to provide safety, fuel and travel time optimization on highway environment. Vehicle platooning involves connecting multiple vehicles in a convoy using connectivity technology and automated driving support systems. These vehicles automatically maintain a close, set distance between each other during specific parts of their journey, such as on highways. The lead vehicle acts as the guide, with the following vehicles adjusting to their movements with little to no driver intervention. Initially, drivers will always stay in control, allowing them to leave the platoon and drive independently. Platooning offers numerous advantages. By driving in each other's slipstream, trucks save fuel, reduce costs, and lower CO₂ emissions. This method of travel also improves road efficiency, leading to smoother traffic flow, time savings, and more dependable road transport. Moreover, vehicle platooning allows drivers to perform other tasks while driving, improving labor efficiency. Increased road safety is another benefit, as automated driving in a platoon minimizes the risk of human error, a leading cause of accidents [4–8]. Participating parties also share data on a matching platform to form platoons efficiently. The concept of vehicle platooning dates back to at least the 1970s [9], with research conducted by universities and government organizations over the decades. Early ideas often involved some form of mechanical connection between vehicles, similar to a “road train”, before the development of modern wireless communication, GPS, and radar sensors. Today, with the significant advancements in autonomous vehicle technology, vehicle platooning is becoming feasible using a combination of existing technologies. These include steer-by-wire steering, throttle-by-wire, radar sensors, GPS, cameras, 5G, or different collision avoidance systems.

These technologies would enable vehicles to communicate with each other, allowing a lead vehicle to manage the movements of the trailing vehicles.

1.1. Related Work

Extensive work has been undertaken to carry out field experiments on platooning since the 1970s. The substantial fuel savings associated with vehicle platoons make them a prime candidate for early adoption in roadway automation. Since the 1980s, numerous projects have explored the concepts of platooning. Initially, these projects concentrated on facilitating cooperative driving capabilities and communication-supported cooperative driving behavior. However, over time, the research focus shifted towards platooning coordination and, more recently, multi-brand platooning and live demonstrations. Noteworthy vehicle platoon initiatives encompass projects such as Chauffeur [10], Partners of Advanced Transit and Highways (PATH) [11], KONVOI [12], and Energy ITS [13]. The Path program goal was to develop and implement advanced technologies to improve transportation efficiency, safety, and sustainability. The Chauffeur project was introduced by Daimler-Benz, Iveco, and several automotive companies, resulting in a second project called Chauffeur II [14] concentrating on communication within the platoon. In contrast to these projects, the KONVOI project focuses on the applicability, usability, and legal factors of platooning instead of its establishment. Furthermore, the driver information system installed on the vehicle communicates with a central server via mobile communications to locate platoons, showcasing an early instance of platooning coordination. Energy ITS, initiated in 2008, focuses on utilizing Intelligent Transportation Systems (ITS) technologies to conserve energy and mitigate global warming. Currently, a platoon comprising three fully automated heavy trucks and a fully automated light truck operates at a constant velocity along an expressway specified for testing purposes. The trucks maintain a gap of up to 4 m, even during executing lane changes. Lateral control is achieved through computer vision-based lane marker detection, while longitudinal control relies on radar and lidar for gap measurement, extended by inter-vehicle communications and infrared technology. The radar and lidar also function as obstacle detection systems. Notably, these technologies prioritize high reliability, positioning them for imminent integration into practical applications. In [15], The Cooperative dynamic formation of platoons for safe and energy-optimized goods transportation (COMPANION) project concentrates on fuel consumption. The Safe Road Trains for the Environment (SARTRE) project [16] introduces a platooning system without altering the road infrastructure. The participants of the road communicate through a remote mobile network, which directs drivers to the closest platoon. Platoon tests have been conducted on several types of vehicles, such as passenger cars, trucks, or other types of participants like robots in [17–22]. For the European truck platooning challenge, Ref. [23] proposes a cross-border platoon from their company headquarters to Rotterdam in a real-world highway scenario. At the same time, there was a competition called The Grand Cooperative Driving Challenge (GCDC) [24,25]. Several papers deal with platoon stability, investigating how to maintain the speed and the desired distance between the participants in the platoon, also called string stability [26–28]. As with many other fields of research [29], AI-driven decision-making has also gained importance in this field. Reinforcement learning is also widely used as a framework for vehicle platooning. Ref. [30] introduces a platoon-forming strategy classifying the vehicles based on their speeds. Ref. [31] optimizing gap among the vehicles in the platooning utilizing a Deep Deterministic Policy Gradient (DDPG) algorithm. In contrast, Ref. [32] proposes a multi-agent soft actor-critic (PI-MASAC) control framework with a human-leading automated heavy-duty-truck platoon. Table 1 gives a brief survey on these results.

Table 1. Related works for the proposed method.

Authors	Title	Ref.
van Nunen et al. (2016)	Sensor safety for the european truck platooning challenge	[23]
Englund et al. (2016)	The Grand Cooperative Driving Challenge 2016: boosting the introduction of cooperative automated vehicles	[25]
Feng et al. (2019)	String stability for vehicular platoon control: Definitions and analysis methods	[27]
Guo et al. (2020)	Adaptive fault-tolerant control of platoons with guaranteed traffic flow stability	[28]
Boubakri et al. (2021)	Platoons formation management strategies based on reinforcement learning	[30]
Farag et al. (2020)	Reinforcement learning based approach for multi-vehicle platooning problem with nonlinear dynamic behavior	[31]
Lian et al. (2024)	Predictive Information Multiagent Deep Reinforcement Learning for Automated Truck Platooning Control	[32]

1.2. Contribution

Several papers address Platooning problems utilizing Multi-Agent Reinforcement Learning (MARL); however, they mainly consider behavior in a formed platoon. This study shows a new highway platooning approach considering merging vehicles into the existing platooning. This paper introduces a novel reward mechanism, including the following distance that dynamically adjusts to the platoon's velocity and the goal lane. The following time is used as a following distance rather than relying only on a specified distance. This adjustment of the following distance is defined by using the following time rather than relying exclusively on a specified distance. The research paper uses Proximal Policy Optimization to illustrate how this reward approach is well-suited for a multi-agent environment. It structures platooning scenarios with minimal velocity differences among participating vehicles, demonstrating its capability for effective operation. The paper is structured as follows: Section 2 provides a review of the literature background regarding the employed methods. Section 3 introduces the implemented environment and the relevant RL properties. In Section 4, the findings are presented. Lastly, Section 5 offers a summary of the research and proposes avenues for future development.

2. Methodology

This research introduces a highway platooning technique utilizing the Multi-Agent Reinforcement Learning framework coupled with the Proximal Policy Optimization algorithm. In Section 2.1, the background literature on Reinforcement Learning is discussed. Section 2.2 provides an overview of MARL fundamentals. Section 2.3 introduces the Stable Baseline3 library developed by OpenAI, which provides a stable and reliable set of implementations for various RL algorithms, while Section 2.4 delves into the rationale behind employing PPO.

2.1. Reinforcement Learning

Reinforcement Learning (RL) has garnered considerable attention from researchers due to its notable achievements in various demanding control tasks, including video games [33], robotics [34], autonomous driving [35] or even controlling wastewater treatment plants [36]. Reinforcement learning constitutes a branch of Machine Learning wherein an agent, acting as the decision-maker, makes decisions based on environmental observations. The agent's actions are either rewarded or penalized depending on their desirability. This learning paradigm is structured around the Markov Decision Process (MDP), characterized by $\langle S, A, R, P \rangle$:

- S: Set of observable states.

- A : Set of possible actions the agent can take.
- R : Set of rewards, what the agent receives as a consequence of its action.
- P : Policy is to select an action at a given state.

Reinforcement Learning harnesses the reward signal to refine the agent's Neural Network, aiming to achieve the desired behavior. The objective is adjusting a network capable of generalizing its responses to act in situations not encountered during training appropriately.

2.2. Multi-Agent Reinforcement Learning

Multi-Agent Reinforcement Learning represents a subset of RL wherein multiple agents can influence the environment's internal state through actions. The mathematical framework of this approach is formalized as a Markov Game $\langle N, X, U^i, P, R^i, \gamma \rangle$. Figure 1 illustrates the interactions between the agents and the environment.

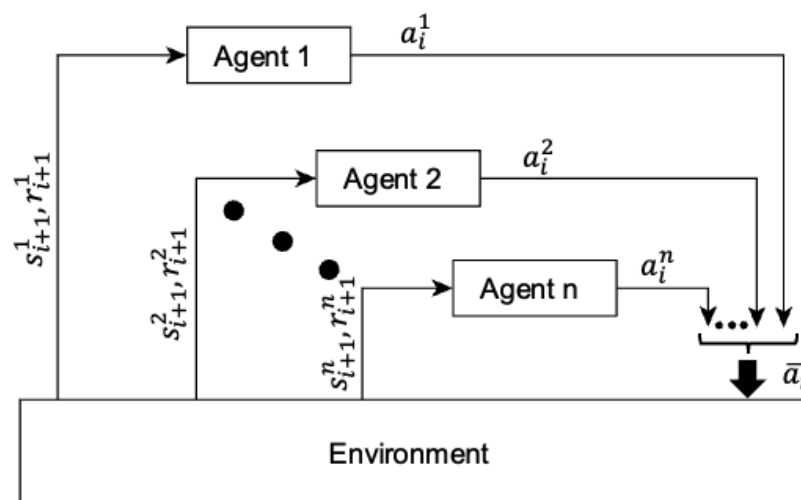


Figure 1. Multi-Agent Reinforcement Learning training loop.

In complex scenarios, a single agent may prove insufficient to tackle specific challenges. Multi-Agent Reinforcement Learning (MARL) systems offer a solution by employing multiple interacting agents within a shared environment. However, as in robotics [37,38], the effectiveness of MARL systems depends significantly on agent coordination.

2.3. Stable Baselines3

Stable Baselines3 (SB3) is a PyTorch-based library designed to help research in reinforcement learning. It offers a collection of well-validated algorithms implemented as pre-built components. This focus on pre-built functionality minimizes development time and guarantees the reliability of the included algorithms, such as Deep Deterministic Policy Gradient (DDPG) and Proximal Policy Optimization (PPO). While DDPG is suitable for problems with continuous action spaces, PPO offers greater flexibility by handling both continuous and discrete action domains [39], which is essential for this paper's method, as it contains both types of actions. Furthermore, SB3 offers a comprehensive suite of tools for researchers. These tools encompass functionalities for training new agents, both with DDPG and PPO, as well as performance evaluation and hyperparameter optimization. Consequently, Stable Baselines3 is a powerful and versatile toolkit that streamlines the entire reinforcement learning research workflow, particularly for those working within the PyTorch framework. By providing pre-built, well-validated algorithms, comprehensive training and evaluation tools, and functionalities for hyperparameter optimization, SB3 significantly reduces development time, promotes research reproducibility, and fosters a deeper understanding of the learning process within RL agents.

2.4. PPO

In the past few years, significant advancement has been achieved in the field of reinforcement learning, notably driven by the Proximal Policy Optimization (PPO) algorithm. PPO has earned widespread credit for its robust empirical performance and simple implementation. Operating in the domain of model-free RL, PPO primarily targets policy gradient methods. These algorithms rely on an approximation of the policy gradient for policy updates within a stochastic gradient ascent framework. The commonly used gradient estimator can be formulated as:

$$\hat{g} = \hat{\mathbb{E}}_t \left[\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \hat{A}_t^i \right] \quad (1)$$

where π_{θ} illustrates a stochastic policy, while \hat{A}_t and $\hat{\mathbb{E}}_t$ indicate the estimator for the advantage function at time step t and the expectation. The second part involves calculating the average outcome from a limited set of samples in an algorithm that changes between data collection and development. Using automatic differentiation tools creates a function to optimize, and its gradient represents the policy gradient estimator. The estimator \hat{g} is calculated as follows:

$$L_{PG}(\theta) = \hat{\mathbb{E}}_t \left[\theta \log \pi_{\theta}(a_t | s_t) \hat{A}_t^i \right] \quad (2)$$

Even though the idea of improving the loss L_{PG} multiple times using the trajectory sounds good, it does not have solid reasoning behind it. In real-world tests, doing this often leads to making policy changes that are too big. PPO stands out because it focuses on improving policies while ensuring learning stays stable and efficient. It does this by putting “proximal” limits on how much policies can change so learning does not get thrown off course. The main idea of PPO is to keep refining the policy over and over by fine-tuning a substitute objective function that has a built-in limit on changes:

$$L_{PPO}^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)} \hat{A}_t, \text{clip} \left(1 - \epsilon, 1 + \epsilon, \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)} \right) \hat{A}_t \right) \right] \quad (3)$$

where:

- $\pi_{\theta}(a_t | s_t)$ is the probability of taking action a_t given state s_t under policy π_{θ} ,
- $\pi_{\theta_{old}}(a_t | s_t)$ is the old policy’s probability,
- \hat{A}_t is the advantage function’s estimator at t ,
- $\text{clip}(a, b, x)$ clips x to the interval $[a, b]$,
- ϵ is a hyperparameter, the magnitude of the policy updates.

This method limits how much policies can change, ensuring they stay close to the previous policy. This stops significant changes that might worsen performance and helps smoother convergence. PPO has many good points and is a good choice for different RL tasks. It is good at using samples efficiently and learning the right policies utilizing a few environmental interactions. Also, PPO works with both kinds of actions—discrete and continuous—so it is helpful for many different problems. PPO includes entropy regularization, which encourages exploration, preventing early convergence for suboptimal policies. Recently, PPO has been utilized in a variety of real-world scenarios, such as robotics, autonomous agents, and gaming agents, demonstrating its adaptability and strength across diverse fields.

3. Environment

Reinforcement Learning (RL) has gained significant attention in recent years as a promising technique for training agents to make decisions in complex environments. The effectiveness of RL heavily relies on the environment in which it is trained, as it impacts the quality and diversity of the training data. This paper presents a custom environment for RL training, where ten vehicles travel in two neighboring highway lanes. The goal for the vehicles in the inner lane is to merge into the outer lane, creating a platooning of

10 vehicles. The proposed method leverages PettingZoo, a multi-agent framework based on the Gymnasium library. The positions of the vehicles are randomized at the beginning of each training episode, which helps create a diverse training process. This approach introduces a range of traffic scenarios throughout the training process, leading to a resilient and adaptable outcome.

3.1. State Representation

State representation is one of the vital parts of Reinforcement Learning for agents to understand their environment better, which has an enormous impact on the decision-making process. This choice becomes particularly critical, especially within the scope of platooning.

The research presents a state space representation defined by the movements and positions of the surrounding vehicles of the ego, including information about the vehicles' relative distances, lane occupancy, velocities, and accelerations, which are close to the ego vehicles. More precisely, the two closest vehicles in front and behind per each lane. Figure 2 shows state space representations from the ego point of view, which is indicated by a blue dot.

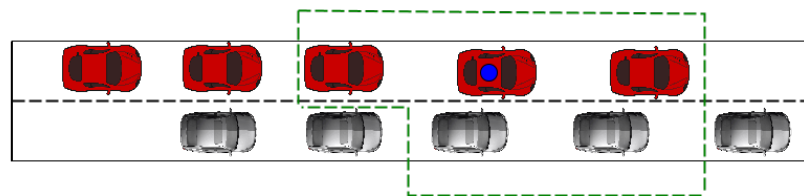


Figure 2. State space representation from the ego (blue dot) point of view.

In scenarios where the ego vehicle lacks companions (front or back) in any lane, the state information is initialized by setting all values to zero in the place of the absent vehicle. This comprehensive state representation provides the RL agent with a detailed understanding of the dynamics within the platoon and its surrounding environment. Considering the movements and positions of nearby vehicles in both lanes, it enables agents—inside or outside the platoon based on their actual positions—to cooperate effectively. This representation allows vehicles outside the platoon to predict the optimal moment for merging, assists vehicles within the platoon in facilitating the entry of new vehicles, or helps maintain the stability of the platoon. Additionally, the state representation includes the ego vehicle's data, which strengthens the platoon's cohesion through self-awareness. This allows the agent to consider its own actions during the decision-making process, whether moving within the platoon or attempting to merge. The features mentioned above meet with the fundamentals of Cooperative Multi-Agent Reinforcement Learning. To achieve safe and efficient driving with optimal platoon stability, vehicles share information through the state representation. This approach aligns perfectly with Cooperative MARL, where agents collaborate towards a common objective. Since each vehicle's actions affect others, this information sharing is crucial for cooperation within the platoon. Utilizing ego data in state representation supports self-awareness, which can significantly contribute to collective decision-making processes in Cooperative MARL.

3.2. Action Space

In this dynamic setting, every vehicle, except the leader, is endowed with two distinct actions. The first action operates inside a continuous action space from -1 to 1 , signifying acceleration and deceleration. Notably, the value of acceleration or deceleration correlates with the selected action value. The second action pertains to lane change, indicated by a boolean value: 0 signifies remaining in the current lane, while 1 denotes that the conditions are met for initiating a lane change maneuver. It is important to emphasize that the lane change has been simplified, resulting in the absence of a lane-changing trajectory. This

implies that when a lane change action is initiated, the vehicle undergoes lateral movement without following a specific trajectory, as shown in Figure 3. To ensure smooth operation, vehicles are prevented from coming to a complete stop with a minimum speed limit of 10 m/s. They also have a ceiling on their speed, capped at 36.1 m/s. These actions are updated every 0.1 s, allowing for precise and coordinated movement among multiple cars within the system.

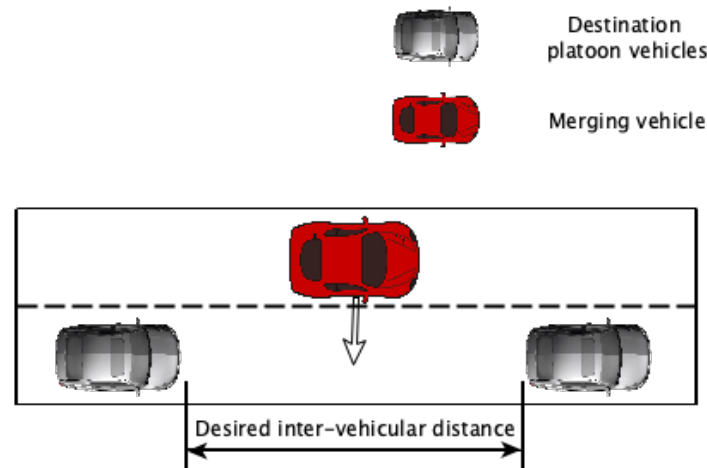


Figure 3. Changing lane.

3.3. Reward

An adequately chosen reward has a vital role in Reinforcement Learning, which measures and evaluates the agent's actions, acting as feedback. The value of the reward represents the quality of the actions taken within the environment, encouraging the optimization of the decision-making process by punishing the undesired actions and rewarding the desired ones. Earning rewards helps agents learn and adapt, aiming to maximize cumulative rewards, representing the elemental goal of RL. The rewarding strategy presented in this paper focuses on optimizing the following time among vehicles. Additionally, it encourages agents to either join the platoon or, if they are already part of it, assist other vehicles in joining the platoon in the correct lane if they have not done so yet. The primary goal is to maintain a target following time among the vehicles. A 0.5-s threshold is introduced in the rewarding strategy, meaning that any positive or negative deviation with this value from the target following time would be penalized. If the actual following time of the agent(s) falls within the threshold, which assumes that the vehicles maintain a desired following time, a reward with a value of 1 is awarded. This positive value motivates the participants of the platoon to maintain the optimal following distance, which keeps collision risk low and promotes traffic efficiency. Nonetheless, if the following time differs from the target following time, exceeding the threshold, a penalty is given in the reward. The penalty equation for the following time is described as:

$$penalty = max_penalty(-1) \times min(0.7, deviation/target_time) \quad (4)$$

We define a maximum penalty value, $max_penalty$, set to -0.7 . This represents the largest penalty an agent can receive. The penalty itself is calculated using the min function and is directly related to how much the agent deviates from the target time. If the difference equals or goes beyond the target time, the penalty reaches its highest value of -0.7 , which strongly discourages that behavior. Including the actual lane in the reward system helps encourage the agents to merge into the proper lane when possible. A -0.3 penalty is introduced in the case when a vehicle is not in the desired lane yet. It encourages the participants to merge as soon as possible to the target lane, designing a rewarded strategy within the limit of -1 and 1 .

3.4. Training

A diverse dataset, as previously mentioned, is crucial for ensuring the network's stable performance and generalization. This study achieves this by utilizing randomly generated traffic scenarios within each training episode. The traffic generation process involved varying initial traffic situations, such as density, speed variability, or the initial number of platoon vehicles. Although the total number of vehicles was fixed, traffic density was varied by adjusting the spatial distribution of these vehicles. This variation in vehicle spacing created different traffic flow dynamics within the fixed number of cars. Moreover, each vehicle was assigned a random speed within a specified range to introduce speed variability. Last but not least, the number of vehicles in the platoon was randomized at the start of each training episode, with a minimum value of 3. Combining these variations creates diverse traffic scenarios that help the RL model during training. A competitive Multi-Agent Reinforcement Learning (MARL) technique is explored in this paper, with a reliance on immediate rewards. Each agent independently seeks to maximize its own reward at each timestep. Episodes are terminated upon reaching a predefined time limit or when a collision is detected. The core of the system is a neural network that receives each agent's state representation and is tasked with suggesting an action in response.

4. Results

Safety is the major goal of road transportation, especially in high-speed highway situations. This paper presents a highway merging and platooning method. The vehicles travel at the maximum speed of 36.1 m/s, emphasizing the importance of the chosen actions and their position in the platoon. The evaluation of the platoon control strategy's performance involved measuring the deviation between the target following time and the actual following time experienced by the vehicle platoon. Hence, the target following distance could vary depending on the actual velocities of the vehicles. Following this evaluation, the proposed approach was subjected to 1000 unique testing episodes. Each episode featured randomly generated traffic patterns, facilitating an analysis of the control strategy's effectiveness and generalizability established during the training phase. Besides, the longitudinal performance of a selected merging vehicle and its surroundings is also presented. The optimal hyperparameters were chosen by grid search and introduced in Table 2.

Table 2. The training parameters.

Parameter	Value
Learning rate (α)	0.00005
Discount factor (γ)	0.95
Num. of ep. after params are upd. (ξ)	20
Num. of hidden layers	4
Num. of neurons for each layer	128, 256, 256, 128
Hidden layers activation function	RELU
Layers	Dense
Optimizer	Adam
kernel initializer	Xavier normal

Performance Evaluation

Figure 4 presents how the average rewards per episode improve during training as time passes. The tendency suggests the training convergence towards the targeted reward level, indicating that the learning algorithm efficiently tunes the policy, ensuring decisions leading to positive rewards on average. Achieving a value above 0.9 in this context indicates that the participants are well maneuvering through the environment and taking appropriate actions.

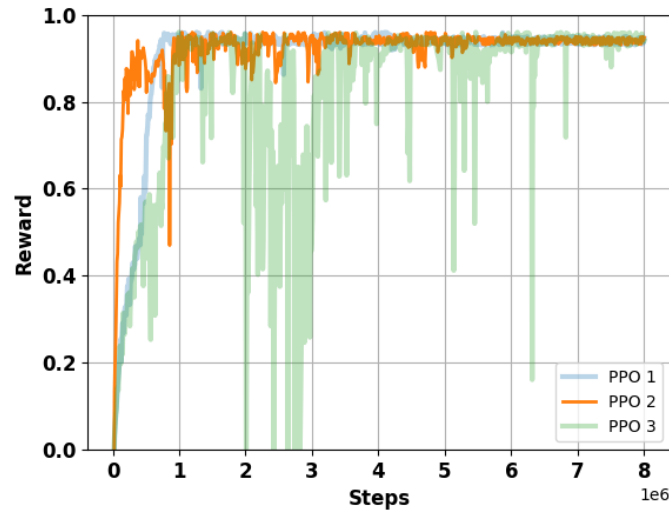


Figure 4. The average reward during training.

As it was mentioned before, the deviation from the target following time is used as an evaluation. Figure 5 presents these values in seconds (s) through the 1000 test episodes.

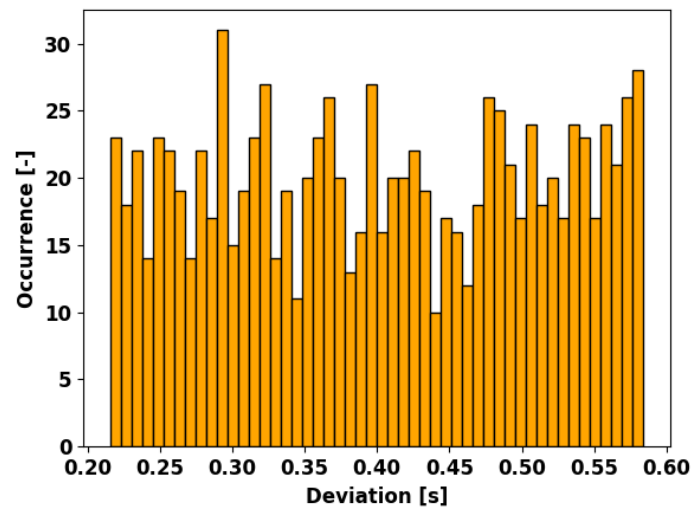


Figure 5. The difference between the vehicle’s average and the target following time.

The figure shows that the deviation of the agents from the target following time stays below the predefined threshold, defined in 0.5 s. It can be seen that these values change between 0.28–0.42. These values confirm that the agents are able to stick to the following time averagely, maintaining a solid platoon structure, given the relatively small deviations from the recommended 3 s following time.

Figure 6 presents the velocity profile of the two vehicles in the recipient lane and the merging vehicle. The maneuver starts at the beginning of the plot when the vehicles in the platoon start to open a gap for the incoming merging vehicle when it arrives next to them, which can be seen as a deceleration maneuver in Figure 6. Upon detecting this maneuver, the merging vehicle also begins to reduce its velocity and smoothly integrates into the gap, gradually accelerating to match the speed of the other two vehicles.

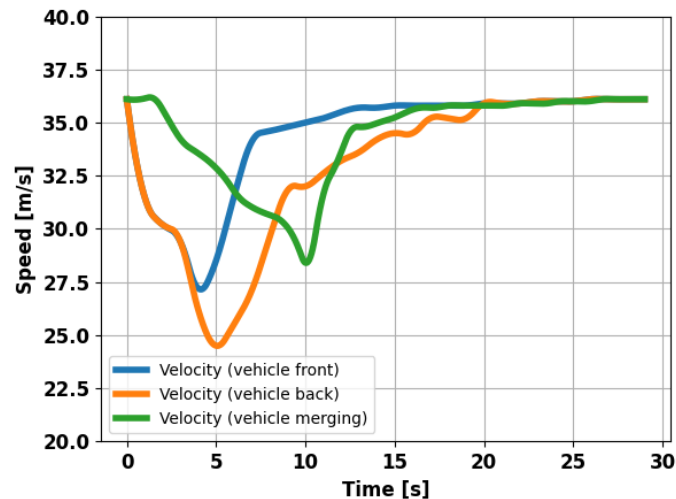


Figure 6. Vehicles' speed profile during merging.

Figure 7 shows the distance profiles of the three aforementioned vehicles. As illustrated, in conjunction with the speed profile figure, the two vehicles in the platoon cover the same distance within the first three seconds. The merging vehicle, moving at a higher speed, travels a greater distance at the same time to achieve proper positioning. Subsequently, the two platoon vehicles create a gap for the merging vehicle, with the merge occurring around the 15th second. After merging, the new vehicle slowly matches the speed of the other two.

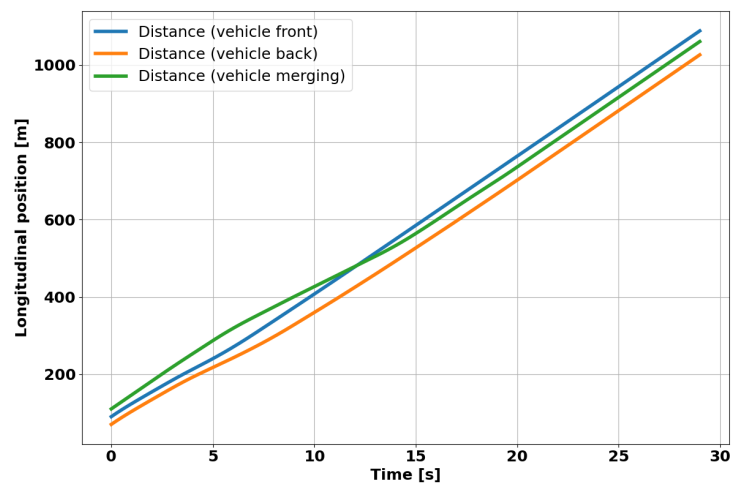


Figure 7. Distance profile during merging.

5. Conclusions

The rise of self-driving cars offers a big chance to make transportation more efficient and safer. Platooning, where vehicles move together in sync over short distances, could help reduce traffic jams, save fuel, and make roads safer overall. The paper introduces a promising idea: using Multi-Agent Reinforcement Learning with Proximal Policy Optimization for platooning, showing its potential benefits. The study findings reveal that the proposed method effectively maintains strong stability within the platooning system. This study presents a reward strategy utilizing the following time, where the following distance changes based on the speed rather than sticking to a fixed distance. Moreover, the strategy introduces a penalty value until the vehicle does not travel in the desired lane in the suggested platoon. As mentioned before, lane change has been simplified, meaning that when a lane change action is initiated, the vehicles operate a lateral movement without following a specific trajectory during lane change. For future endeavors, it is important to improve the dynamic of the lane change, implementing a valid and safe trajectory. The evaluation of

the algorithm presented in this paper is limited to platoon formation, as this article aimed to provide the feasibility of cooperating agents based on the independent learner paradigm. The results show that the approach is viable, though it naturally needs further investigation. For future work, it is recommended to integrate different sustainability metrics into the algorithm's implementation process. This would enhance the comprehensiveness and applicability of the evaluation, making it a crucial area for future development.

Author Contributions: Conceptualization, T.B.; Methodology, M.K. and T.B.; Software, M.K.; Writing—original draft, M.K.; Writing—review & editing, T.B.; Visualization, M.K.; Supervision, T.B.; Funding acquisition, T.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the European Union within the framework of the National Laboratory for Autonomous Systems (RRF-2.3.1-21-2022-00002). Project no. TKP2021-NVA-02 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme. T.B. was supported by BO/00233/21/6, János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

V2V	Vehicle-to-Vehicle
V2I	Vehicle-to-Infrastructure
V2X	Vehicle-to-Everything
GPS	Global Positioning System
ITS	Intelligent Transportation Systems
LIDAR	Light Detection and Ranging
COMPANION	Cooperative Dynamic Formation of Platoons for Safe and Energy-Optimized Goods Transportation
SARTRE	Safe Road Trains for the Environment
GCDC	Grand Cooperative Driving Challenge
MARL	Multi-Agent Reinforcement Learning
RL	Reinforcement Learning
MDP	Markov Decision Process
SB3	Stable Baselines3
PPO	Proximal Policy Optimization
DDPG	Deep Deterministic Policy Gradient

References

- Demba, A.; Möller, D.P.F. Vehicle-to-Vehicle Communication Technology. In Proceedings of the 2018 IEEE International Conference on Electro/Information Technology (EIT), Rochester, MI, USA, 3–5 May 2018; pp. 0459–0464. [\[CrossRef\]](#)
- Van Phu, C.N.; Farhi, N.; Haj-Salem, H.; Lebacque, J.P. A vehicle-to-infrastructure communication based algorithm for urban traffic control. In Proceedings of the 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), Naples, Italy, 26–28 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 651–656.
- Hasan, M.; Mohan, S.; Shimizu, T.; Lu, H. Securing vehicle-to-everything (V2X) communication platforms. *IEEE Trans. Intell. Veh.* **2020**, *5*, 693–713. [\[CrossRef\]](#)
- Krizsik, N.; Sipos, T. Social Perception of Autonomous Vehicles. *Period. Polytech. Transp. Eng.* **2023**, *51*, 133–139. [\[CrossRef\]](#)
- Maiti, S.; Winter, S.; Kulik, L. A conceptualization of vehicle platoons and platoon operations. *Transp. Res. Part Emerg. Technol.* **2017**, *80*, 1–19. [\[CrossRef\]](#)
- Hu, M.; Zhao, X.; Hui, F.; Tian, B.; Xu, Z.; Zhang, X. Modeling and analysis on minimum safe distance for platooning vehicles based on field test of communication delay. *J. Adv. Transp.* **2021**, *2021*, 1–15. [\[CrossRef\]](#)
- Wu, J.; Ahn, S.; Zhou, Y.; Liu, P.; Qu, X. The cooperative sorting strategy for connected and automated vehicle platoons. *Transp. Res. Part Emerg. Technol.* **2021**, *123*, 102986. [\[CrossRef\]](#)
- Cao, D.; Wu, J.; Wu, J.; Kulcsár, B.; Qu, X. A platoon regulation algorithm to improve the traffic performance of highway work zones. *Comput.-Aided Civ. Infrastruct. Eng.* **2021**, *36*, 941–956. [\[CrossRef\]](#)

9. Hoberock, L.; Rouse, R., Jr. Emergency control of vehicle platoons: System operation and platoon leader control. *J. Dyn. Syst. Meas. Control* **1976**, *98*, 245–251. [[CrossRef](#)]
10. Gehring, O.; Fritz, H. Practical results of a longitudinal control concept for truck platooning with vehicle to vehicle communication. In Proceedings of the Conference on Intelligent Transportation Systems, Boston, MA, USA, 12 November 1997; IEEE: Piscataway, NJ, USA, 1997; pp. 117–122.
11. Shladover, S.E. PATH at 20—History and major milestones. *IEEE Trans. Intell. Transp. Syst.* **2007**, *8*, 584–592. [[CrossRef](#)]
12. Kunze, R.; Ramakers, R.; Henning, K.; Jeschke, S. Organization and operation of electronically coupled truck platoons on German motorways. In *Automation, Communication and Cybernetics in Science and Engineering 2009/2010*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 427–439.
13. Tsugawa, S. Results and issues of an automated truck platoon within the energy ITS project. In Proceedings of the 2014 IEEE Intelligent Vehicles Symposium Proceedings, Dearborn, MI, USA, 8–11 June 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 642–647.
14. Franke, U.; Bottiger, F.; Zomotor, Z.; Seeberger, D. Truck platooning in mixed traffic. In Proceedings of the Intelligent Vehicles' 95. Symposium, Detroit, MI, USA, 25–26 September 1995; IEEE: Piscataway, NJ, USA, 1995; pp. 1–6.
15. Eilers, S.; Mårtensson, J.; Pettersson, H.; Pillado, M.; Gallegos, D.; Tobar, M.; Johansson, K.H.; Ma, X.; Friedrichs, T.; Borojeni, S.S.; et al. COMPANION—Towards Co-operative Platoon Management of Heavy-Duty Vehicles. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Gran Canaria, Spain, 15–18 September 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1267–1273.
16. Jootel, P.S. SARTRE project final report. In *Eur. Commission under Framework 7 Programme Project 233683*; Publication Office of the European Union: Luxembourg, 2012.
17. Li, Y.; Tang, C.; Peeta, S.; Wang, Y. Integral-sliding-mode braking control for a connected vehicle platoon: Theory and application. *IEEE Trans. Ind. Electron.* **2018**, *66*, 4618–4628. [[CrossRef](#)]
18. Zhang, Y.; Hu, J.; Wu, Z. Cooperative adaptive cruise control: A field experiment. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–2.
19. Kita, E.; Sakamoto, H.; Takaeue, H.; Yamada, M. Robot vehicle platoon experiment based on multi-leader vehicle following model. In Proceedings of the 2014 Second International Symposium on Computing and Networking, Shizuoka, Japan, 10–12 December 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 491–494.
20. Guo, G.; Yue, W. Autonomous platoon control allowing range-limited sensors. *IEEE Trans. Veh. Technol.* **2012**, *61*, 2901–2912. [[CrossRef](#)]
21. Knoop, V.L.; Wang, M.; Wilmlink, I.; Hoedemaeker, D.M.; Maaskant, M.; Van der Meer, E.J. Platoon of SAE level-2 automated vehicles on public roads: Setup, traffic interactions, and stability. *Transp. Res. Rec.* **2019**, *2673*, 311–322. [[CrossRef](#)]
22. Ding, J.; Pei, H.; Hu, J.; Zhang, Y. Cooperative adaptive cruise control in vehicle platoon under environment of i-VICS. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1246–1251.
23. van Nunen, E.; Koch, R.; Elshof, L.; Krosse, B. Sensor safety for the european truck platooning challenge. In Proceedings of the Intelligent Transportation Systems World (ITS), 2016 23rd World Congress, Melbourne, Australia, 10–14 October 2016; pp. 306–311.
24. Ploeg, J.; Shladover, S.; Nijmeijer, H.; van de Wouw, N. Introduction to the special issue on the 2011 grand cooperative driving challenge. *IEEE Trans. Intell. Transp. Syst.* **2012**, *13*, 989–993. [[CrossRef](#)]
25. Englund, C.; Chen, L.; Ploeg, J.; Semsar-Kazerooni, E.; Voronov, A.; Bengtsson, H.H.; Didoff, J. The grand cooperative driving challenge 2016: Boosting the introduction of cooperative automated vehicles. *IEEE Wirel. Commun.* **2016**, *23*, 146–152. [[CrossRef](#)]
26. Li, D.; Guo, G. Prescribed performance concurrent control of connected vehicles with nonlinear third-order dynamics. *IEEE Trans. Veh. Technol.* **2020**, *69*, 14793–14802. [[CrossRef](#)]
27. Feng, S.; Zhang, Y.; Li, S.E.; Cao, Z.; Liu, H.X.; Li, L. String stability for vehicular platoon control: Definitions and analysis methods. *Annu. Rev. Control* **2019**, *47*, 81–97. [[CrossRef](#)]
28. Guo, G.; Li, P.; Hao, L.Y. Adaptive fault-tolerant control of platoons with guaranteed traffic flow stability. *IEEE Trans. Veh. Technol.* **2020**, *69*, 6916–6927. [[CrossRef](#)]
29. Nguyen, V.T.T.; Vo, T.M.N. *Using Traditional Design Methods to Enhance AI-Driven Decision Making*; IGI Global: Hershey, PA, USA, 2024. [[CrossRef](#)]
30. Boubakri, A.; Matali Gmmar, S. Platoons formation management strategies based on reinforcement learning. In Proceedings of the International Conference on Systems Engineering, Wroclaw, Poland, 14–16 December 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 57–66.
31. Farag, A.; AbdelAziz, O.M.; Hussein, A.; Shehata, O.M. Reinforcement learning based approach for multi-vehicle platooning problem with nonlinear dynamic behavior. In Proceedings of the Machine Learning for Autonomous Driving Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020.
32. Lian, R.; Li, Z.; Wen, B.; Wei, J.; Zhang, J.; Li, L. Predictive Information Multiagent Deep Reinforcement Learning for Automated Truck Platooning Control. *IEEE Intell. Transp. Syst. Mag.* **2024**, *16*, 116–131. [[CrossRef](#)]
33. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [[CrossRef](#)]

34. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
35. Fehér, Á.; Aradi, S.; Bécsi, T. Hierarchical Evasive Path Planning Using Reinforcement Learning and Model Predictive Control. *IEEE Access* **2020**, *8*, 187470–187482. [[CrossRef](#)]
36. Hernández-del Olmo, F.; Gaudio, E.; Duro, N.; Dormido, R. Machine Learning Weather Soft-Sensor for Advanced Control of Wastewater Treatment Plants. *Sensors* **2019**, *19*, 3139. [[CrossRef](#)] [[PubMed](#)]
37. Guinaldo, M.; Dimarogonas, D.V. A hybrid systems framework for multi agent task planning and control. In Proceedings of the 2017 American Control Conference (ACC), Seattle, WA, USA, 24–26 May 2017; pp. 1181–1186. [[CrossRef](#)]
38. Guinaldo, M.; Farias, G.; Fabregas, E.; Sánchez, J.; Dormido-Canto, S.; Dormido, S. An interactive simulator for networked mobile robots. *IEEE Netw.* **2012**, *26*, 14–20. [[CrossRef](#)]
39. Zhu, J.; Wu, F.; Zhao, J. An overview of the action space for deep reinforcement learning. In Proceedings of the 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence, Sanya, China, 22–24 December 2021; pp. 1–10.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.