



A Survey on Multi-agent Reinforcement Learning for Adaptive Transportation Solutions

Jintao Liang¹ · Xinqi Du² · Yilin Liu¹ · Ziyue Li³ · Haoyuan Jiang⁴ · Zhishuai Li⁵ · Cheng Long⁶ · Hua Wei⁷ · Wolfgang Ketter³

Received: 16 November 2024 / Accepted: 13 October 2025
© The Author(s) 2025

Abstract

Reinforcement learning (RL) has become an essential tool in optimizing transportation systems, offering promising solutions to dynamic and complex decision-making challenges. This paper explores the application of RL in four representative domains of transportation systems, including road-level traffic signal control, low-altitude unmanned aerial vehicle (UAV) navigation, high-altitude air traffic management (ATM), and maritime autonomous ship (AMS) navigation. These domains span land, low airspace, high airspace, and sea, reflecting key operational environments where multi-agent coordination is crucial. While significant advances have been made in the development of RL-based systems, traditional models often face limitations when applied to diverse, real-world environments due to their reliance on single-scenario training. To address this, we focus on cross-scenario multi-agent reinforcement learning (MARL), which aims to enhance the adaptability and scalability of RL models across multiple, distinct scenarios. Specifically, we examine offline RL and meta RL techniques, which enable agents to transfer policies learned in one environment to new, unseen scenarios. Despite the progress, challenges such as policy transferability, data efficiency, and scalability remain, which need to be addressed for broader application in transportation systems. This paper provides an overview of these challenges, reviews the state-of-the-art methods, and discusses future directions for advancing RL in transportation systems.

Keywords Intelligent transportation · Multi-agent reinforcement learning · Markov decision process · Cross-scenario transferability

J. Liang and X. Du have contributed equally to this work.

✉ Ziyue Li
ziyue.li@tum.de

Jintao Liang
ljt2021@bupt.edu.cn

Xinqi Du
duxq@dlut.edu.cn

Yilin Liu
liuyilin10@bupt.edu.cn

Haoyuan Jiang
jianghaoyuan@zju.edu.cn

Zhishuai Li
zhishuai.li@foxmail.com

Cheng Long
c.long@ntu.edu.sg

Hua Wei
hua.wei@asu.edu

Wolfgang Ketter
ketter@wiso.uni-koeln.de

- ¹ State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, No. 10 Xitucheng Road, Haidian District 100876, Beijing, China
- ² School of Control Science and Engineering, Dalian University of Technology, No.2 Linggong Road, Dalian 116024, Liaoning Province, China
- ³ Department of Operations and Technology, Heilbronn Data Science Center, Technical University of Munich, Munich, Germany
- ⁴ Baidu, Shenzhen, Guangdong Province, China
- ⁵ China University of Petroleum, Qingdao, Shandong Province, China
- ⁶ College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore
- ⁷ School of Computing and Augmented Intelligence, Arizona State University, 699 S Mill Ave, Tempe, AZ 85281, USA

Introduction

Reinforcement learning (RL) is a machine learning framework where an agent learns to make sequential decisions by interacting with an environment and receiving feedback in the form of rewards. Through the balance of exploration and exploitation, RL enables agents to discover strategies that maximize long-term returns. Its capacity to adapt under uncertainty and handle complex decision-making processes makes RL particularly well-suited for transportation systems, which are characterized by dynamic conditions, diverse agent interactions, and real-time operational demands. In this survey, we focus on four representative domains of transportation systems: (1) road-level traffic signal control (TSC), (2) low-altitude unmanned aerial vehicle (UAV) navigation, (3) high-altitude air traffic management (ATM), and (4) autonomous maritime ship (AMS) navigation. These domains span land, low airspace, high airspace, and sea, representing critical operational environments that demand real-time coordination among multiple autonomous agents. Each of these domains is associated with practical performance metrics that can be naturally incorporated into RL reward design. In TSC, metrics such as average delay, queue length, and throughput reflect traffic efficiency. UAV navigation is often evaluated by path optimality, collision avoidance, and energy consumption. ATM performance relies on measures like airspace capacity, flight efficiency, and conflict resolution rates. For AMS navigation, common indicators include safety, fuel efficiency, and regulatory compliance. These domain-specific metrics not only capture real-world objectives but also provide meaningful reward signals for training RL agents. These transportation systems are inherently dynamic and uncertain, characterized by constantly changing environments and complex agent interactions. RL, with its trial-and-error learning paradigm, is particularly well-suited for these types of environments, enabling agents to learn optimal policies over time through interactions with their surroundings.

Building upon the scope outlined above, we now briefly review key RL applications in these four transportation domains. In TSC, RL methods have been widely used to optimize the flow of traffic and minimize congestion by adapting signal timings based on real-time traffic data. Early studies on RL for TSC primarily employed a single centralized Q-learning agent to manage signal phases across multiple intersections [1–3]. Due to the centralized structure, these approaches faced significant scalability limitations. More recent advancements have introduced multi-agent RL frameworks that emphasize cooperative control among intersections [4–10]. For instance, IntelliLight [4] and FRAP [5] integrate deep Q-networks (DQN) with advanced reward design to enhance signal phase scheduling and

accelerate convergence for training a multi-agent RL traffic system. Similarly, RL has been applied in UAV navigation to address challenges such as path planning and obstacle avoidance. Leveraging onboard sensors and cameras, early efforts adopted Q-learning-based control for individual UAVs to learn optimal paths in unknown environments, progressively refining their navigation policies [11–14]. As mission complexity and environmental dynamics increased, more recent studies have advanced toward multi-agent RL frameworks that enable decentralized coordination among multiple UAVs [15–19], supporting complex collaborative behaviors such as formation control, area exploration, and energy-efficient flight under partial observability. Moreover, air traffic management is another area where RL has shown considerable potential, particularly in multi-aircraft conflict resolution (MACR) and traffic flow management (TFM). Recent studies have adopted multi-agent RL frameworks that coordinate multiple aircraft to prevent air conflicts and improve traffic volumes [20–22]. Several RL-based TFM systems further leverage decentralized decision-making to balance airspace demand and support adaptive route planning under dynamic conditions such as fluctuating traffic or weather [22–24]. In these systems, RL agents dynamically adjust aircraft trajectories in real time, enhancing overall traffic flow while reducing delays and bottlenecks. Autonomous maritime ship navigation has also benefited from RL-based methods that address challenges such as dynamic path planning, obstacle avoidance, and compliance with maritime navigation rules. Various centralized RL algorithms have been applied to support both global and local navigation tasks for individual unmanned vessels, enabling them to plan safe and efficient routes in real time under uncertain and dynamic maritime conditions [25–27]. Recent approaches have further integrated domain-specific constraints, such as COLREGS, into the reward design to improve safety and regulatory compliance [28–30]. In addition, multi-agent RL frameworks have shown promise in enabling coordinated behavior among multiple vessels, enhancing collision avoidance and system stability in collaborative navigation scenarios [31–33]. The left part of Fig. 1 presents several efficient RL methods applied to different transportation environments, including TSC, UAV, ATM and AMS.

While these applications of RL in transportation systems have made significant strides, a key challenge remains in the cross-scenario adaptability of trained policies. Here, a *scenario* refers to different road network maps, such as those representing various urban traffic systems or regional infrastructures. Current RL methods often rely on training in a single scenario, which limits the transferability of learned policies to new, unseen environments. This is particularly problematic in transportation systems where conditions,

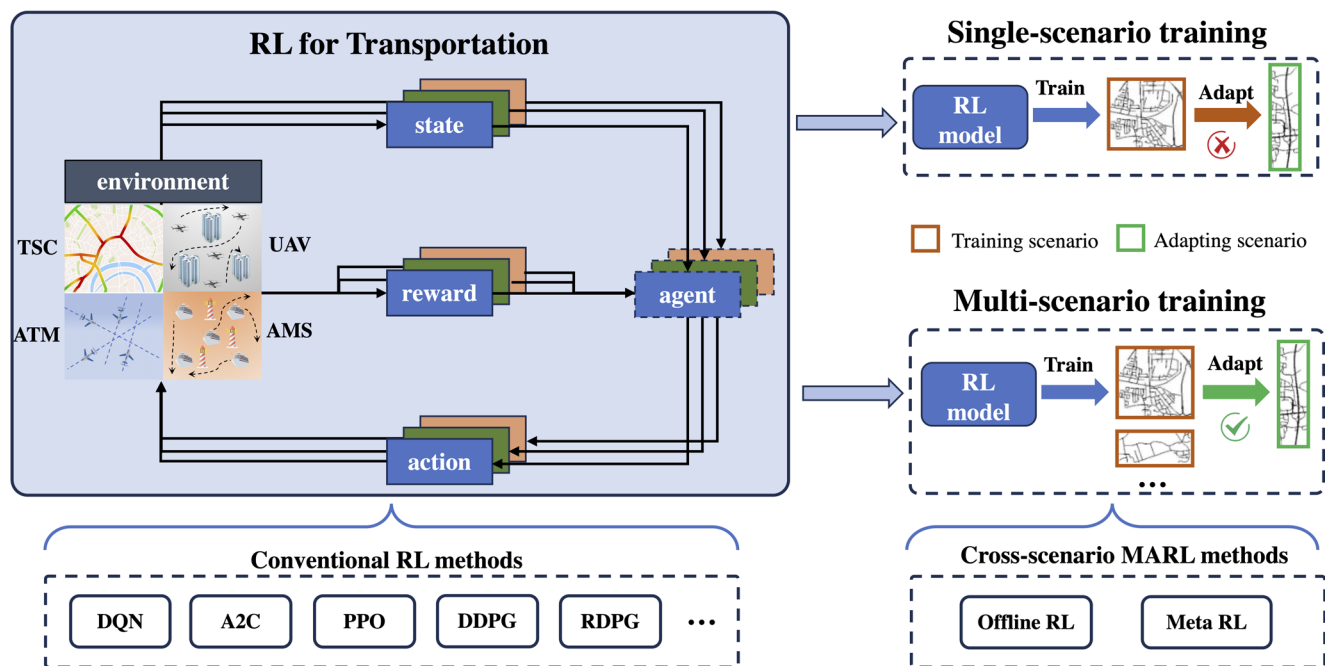


Fig. 1 The overview of our survey. The left part presents various RL methods applied to different transportation environments. The right part compares single-scenario and multi-scenario training, highlighting cross-scenario MARL methods

such as traffic patterns, environmental factors, and agent behaviors, can vary widely across regions or over time. To address this challenge, there has been growing interest in cross-scenario multi-agent reinforcement learning (MARL), which seeks to improve the adaptability and scalability of RL models across diverse environments.

In Section 4, we explore the advancements in cross-scenario MARL, focusing on two key approaches: offline RL and meta RL. Offline RL allows agents to learn from pre-collected datasets, enabling policy transfer across different environments without the need for real-time interactions. Meta RL, on the other hand, enhances policy adaptability by enabling agents to rapidly fine-tune their learned knowledge when faced with new, unseen scenarios. Both approaches have shown promise in overcoming the limitations of traditional MARL methods, but they still face challenges related to policy transferability and generalization. This paper discusses these challenges and provides insights into how they can be addressed to enable more resilient and adaptive transportation solutions. On the right, Fig. 1 compares single-scenario and multi-scenario training approaches, highlighting cross-scenario MARL methods such as offline RL and meta RL.

The paper is organized as follows. Section 2 introduces the fundamental concepts of single-agent and multi-agent reinforcement learning, providing a brief overview of traditional RL foundations, with more detailed discussions in later sections. Section 3 provides an overview of RL applications in practical transportation systems. Section 4 focuses

on cross-scenario MARL, exploring the limitations of traditional RL methods and the potential of offline RL and meta RL to enhance policy transferability and adaptability across diverse environments. Section 5 presents an outlook on the future of cross-scenario MARL in transportation. Finally, Section 6 concludes the paper and suggests directions for future research in MARL for transportation systems.

Reinforcement Learning

Reinforcement learning (RL) enables agents to learn optimal decision-making strategies through interaction with a dynamic environment, guided by a trial-and-error mechanism aimed at maximizing cumulative rewards [34]. This section provides single and multi-agent reinforcement learning methods.

Single-Agent Reinforcement Learning

Single-agent reinforcement learning focuses on how an agent learns to interact with an environment to maximize long-term cumulative rewards. The agent repeatedly observes the environment, selects actions, and receives feedback in the form of rewards, gradually improving its behavior through trial and error. These methods can be roughly divided into three categories: (1) Value-based methods. It makes decisions on actions based on the evaluation results of the action-value function, and the classical algorithms mainly

include the Deep Q-Network (DQN) [35] and its improvement algorithms. For example, Double DQN [36] adopts two independent networks for evaluating the action-value function, which is used to alleviate the problem of estimation error. (2) Policy-based methods. It directly approximates the policy function and updates the parameters of the policy neural network by maximizing the expected return, such as REINFORCE [37] and others. (3) Actor-critic methods. These combine the strengths of value-based and policy-based approaches by maintaining two models: an actor which learns the policy, and a critic which estimates the value function to guide the policy update. Classical actor-critic algorithms include Asynchronous Advantage Actor-Critic (A3C) [38] and Proximal Policy Optimization (PPO) [39], and so on.

Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning (MARL) extends RL to settings involving multiple interacting agents [40]. These agents may cooperate, compete, or both, depending on the task. Each agent learns policies based on its observations and possibly through communication or shared learning. Based on training structures, MARL methods are commonly categorized as: (1) Decentralized: each agent has autonomy, and they independently learn policies based on their observation and reward, such as the (Independent QL) IQL [41]. It is characterized by the simplicity of the deployment, but due to the fact that the observations in a multi-agent system are partially observed, any single agent is unable to make stable and reasonable action decisions with its private observation. (2) Centralized: it uses the joint state-action information of all agents for policy decision-making and model training, which alleviates the potential problems caused by partial observations. However, it suffers from limited scalability. (3) Centralized training and decentralized execution: each agent uses joint information to train the model during the training period; and during the execution, each agent only uses its respective local observation to select actions [42]. Multi-Actor-Attention-Critic (MAAC) [43] applies soft actor-critic to multi-agent systems that can either enable collaboration or competition among multiple agents. Multi-Agent Proximal Policy Optimization (MAPPO) [44] extends the PPO method to multi-agent systems. In addition, some other multi-agent reinforcement learning methods explicitly model the information interaction process between agents, such as CommNet [45].

RL for Transportation System

In recent years, reinforcement learning (RL) has been widely applied in many fields of transportation systems due to its ability to learn optimal policies in dynamic environments. In this section, we will discuss the applications and challenges of RL in traffic signal control, drone navigation, air traffic management, and autonomous ship navigation.

Traffic Signal Control

The issue of traffic congestion has become increasingly severe, and traffic signal control is recognized as an effective approach to alleviate traffic congestion. Substantial research has focused on developing methods for traffic signal control to enhance traffic flow efficiency and maximize system throughput. Existing methods for traffic signal control can be broadly categorized into two types: traditional control methods and advanced data-driven or learning-based approaches.

Traditional traffic signal control methods primarily rely on domain expert knowledge, utilizing pre-defined scheduling rules or signal phase-switching thresholds to determine signal timing plans [46, 47]. They can be broadly categorized into four main types: fixed-time, actuated, traffic-responsive, and adaptive control, as summarized in prior surveys [48–52]. Fixed-time control methods rely on static signal plans based on historical traffic data. Classical examples include Webster's method [53], which optimizes cycle length based on traffic volumes, and other implementations such as FixedTime [54]. Actuated control methods use vehicle detectors to trigger signal changes based on real-time presence detection. Common implementations include [55], where traffic signals adapt locally to detected vehicle arrivals. Traffic-responsive systems such as SCOOT [56] and SCATS [57], use real-time traffic data to adjust cycle times, splits, and offsets dynamically across coordinated intersections. Adaptive control methods aim to optimize network-wide objectives (e.g., delay minimization) using real-time feedback. One notable approach is Max-Pressure control [58], which selects signal phases to maximize discharge rates and balance queue lengths across competing flows. While these traditional methods are widely deployed due to their simplicity and robustness, they often lack flexibility and struggle to adapt to rapidly changing or highly uncertain traffic patterns.

To address the above limitations, researchers began exploring reinforcement learning (RL) for traffic signal control to enhance the flexibility and efficiency. Given that traffic signal control can be formulated as a sequential decision-making problem, RL-based methods have emerged as an effective method. Some centralized RL methods utilized

Q-learning to train a centralized agent to manage the traffic signal phases across multiple intersections [1, 59, 60]. These centralized approaches established foundational techniques for learning-based signal optimization. However, they face significant scalability challenges due to the exponential growth of the joint state-action space as network size increases. These limitations motivated a shift toward decentralized and multi-agent RL solutions. Early efforts in this direction include the work of Wiering [61, 62] and Bazzan [63], who explored distributed RL frameworks in which each intersection acts as an independent agent. These studies laid the groundwork for modern multi-agent TSC systems. We summarize the recent representative studies on RL for TSC, all published in top-tier conferences or journals, in Table 1. In 2018, Wei et al. [4] proposed the IntelliLight framework, which effectively addressed traffic signal control using a DQN. While the method demonstrated strong

adaptability to local traffic state abstraction, it suffers from scalability limitations in large-scale networks. Later, Zheng et al. [5] developed the FRAP model, which improved the modeling process of traffic signal control by designing a phase competition model. Combined with the Ape-X DQN algorithm, it achieved faster convergence, although its performance is highly sensitive to the design of phase and state representations. Wei et al. [9] also incorporated the MaxPressure concept from the transportation field into reinforcement learning reward function design, eliminating the need for heuristic design. However, this method requires accurate real-time pressure data, which may not be available in all deployments. Zhang et al. [64] focused on enhancing the state representation and proposed ExpressionLight, which improves generalization ability through better state encoding. The model, however, places less emphasis on coordination across intersections or scalability. Liang et

Table 1 Representative MARL-based methods for traffic signal control

Methods	Key technique	Results	Limitations
Wei et al. [4]	DQN as a RL agent	Adaptive TSC	Poor scalability in large traffic networks
Zheng et al. [5]	Phase competition modeling +Ape-X DQN	Faster convergence	Sensitive to phase/state dynamics
Wei et al. [9]	MaxPressure for reward function design	Eliminates the need for heuristic reward design	Requires real-time pressure estimation
Zhang et al. [64]	Advanced state representation design	Better performance via improved state encoding	No emphasis on coordination or scaling
Liang et al. [65]	3DQN + Cycle control	Optimizes phase duration dynamically	Assumes fixed signal cycling scheme
Zhao et al. [66]	GNN + Pressure-guided DQN	Learns both phase and duration via graph-based reasoning	High computational overhead
Chu et al. [67]	MARL based on A2C	Scalability to large road networks	Coordination among agents not explicitly modeled
Chen et al. [68]	Shared-parameter decentralized DQN	Controlled 1,000+ intersections	Convergence instability
Wu et al.[76]	Spatio-temporal GAT + DQN	Effective dynamic coordination over intersections	Performance depends on temporal graph quality
Xu et al. [78]	Hierarchical control architecture	Multi-level collaboration and policy consistency	Increased model complexity
Ruan et al. [80]	Top-K intersection collaborative selection mechanism	Optimizes the collaborative control of key intersections	High computational overhead
Xiong et al. [81]	Demonstration learning	Faster convergence	Depends on availability of high-quality expert data
Zang et al. [82]	Meta-RL with value-based learning	Improves adaptation and sample efficiency	Limited to similar topology/task spaces
Yang et al. [83]	Influence-aware coordination under missing data	Maintains coordination in incomplete scenarios	Needs inference for missing state info
Chen et al. [84]	Diffusion-based RL with partial rewards	Enables training under incomplete feedback	Struggles with delayed or sparse rewards
Zhang et al. [85]	GAN-based traffic data augmentation	Enhances generalization via diverse data synthesis	Stability of GAN training is a challenge
Oroojlooy et al. [86]	Attention-based RL	Improves generalization to different road structures	High complexity and attention overhead
Jiang et al. [87]	Scenario-agnostic co-training	Learns shared policy across cities	Sensitive to heterogeneous scenarios
Lu et al. [88]	Dual-view co-training	Generalizes across topology	Needs consistent dual observations
Jiang et al. [89]	Transformer-based meta-RL (X-Light)	Achieves zero-shot transfer across cities	High training cost and model complexity

al. [65] proposed a TSC method based on a cyclic traffic signal phase scheduling, and deployed 3DQN for making decisions on the duration of each traffic signal phase to improve transportation throughput. However, this method relies on the assumption of fixed signal phase sequences, limiting its flexibility. Similarly, Zhao et al. [66] introduced a method that combines GNN with pressure-guided DQN to jointly learn signal phase and duration. While achieving high scheduling efficiency, the model comes with increased computational complexity.

With the in-depth study of the multi-intersection traffic signal control problem, researchers have explored from multiple perspectives, deriving the following research directions:

Scalability

In large-scale traffic signal control, scalability is an important research direction. Chu et al. [67] proposed a decentralized multi-agent reinforcement learning algorithm based on the A2C (Advantageous Actor-Critic) method, which achieves the scalability of large-scale traffic signal control. Nevertheless, the method does not explicitly address inter-agent coordination. To further address scalability, Chen et al. [68] designed a decentralized DQN method with parameter sharing, which is suitable for controlling thousands of traffic signals. Despite its efficiency, the model is prone to convergence instability under diverse traffic scenarios.

Collaboration

The collaboration of multiple intersections is crucial to the overall performance of a transportation system. Some earlier efforts in multi-agent TSC explored decentralized collaboration. For example, El-Tantawy et al. [69] proposed the MARLIN-ATSC system, one of the first fully decentralized MARL-based TSC frameworks. Kuyer et al. [70] modeled intersection control as a coordination graph problem, enabling joint optimization across agents. Aslani et al. [71] developed a fuzzy Q-learning-based multi-agent controller that adapts to traffic pattern changes while maintaining distributed control. Abdoos et al. [72] applied distributed Q-learning with function approximation to address the scalability challenge in large-scale urban traffic networks, demonstrating how generalization in Q-values can support more efficient policy learning across similar intersections. Building on these foundations, recent studies have incorporated more advanced collaboration mechanisms using neural networks. Wei et al. [6] constructed the road network as a graph structure and combined the graph attention network with DQN to realize the information exchange and collaborative control among intersections. Since then,

GNN-based methods such as PRGLight [66], IG-RL [73], GraphLight [74], and KeyLight [75] have further improved multi-agent coordination, though most still face computational and scalability bottlenecks. Meanwhile, the collaboration in temporal states has also become a research hotspot [10]. Wu et al. [76] combined spatio-temporal information to support dynamic coordination, but it is heavily reliant on the temporal graph construction quality. In addition, several literature have further explored traffic signal control methods using different collaborative strategies. Ma et al. [77] proposed a Feudal RL-based method to divide the road network into multiple regions, which enhances intra- and inter-region collaboration through hierarchical control. Xu et al. [78] proposed a hierarchical control architecture that improves policy consistency across intersections, at the cost of increased model complexity. The RACS [79] method is based on the RL method and enhances the collaborative effect by dividing the regions in the road network. The CosLight [80] method, introduces the top-K intersection collaborative selection mechanism to further optimize the collaborative control of key intersections. While effective in optimizing critical points in the network, it also incurs additional overhead during coordination.

Data Efficiency

In RL-based traffic signal control methods, model performance primarily relies on a large amount of traffic data. Xiong et al. [81] proposed to use classical expert demonstrations to guide exploration and improve training convergence. However, its effectiveness depends heavily on the availability and quality of expert trajectories. Zang et al. [82] introduced MetaLight, a meta-reinforcement learning (Meta-RL) based method that improves adaptation and sample efficiency across similar tasks. Meta-RL is designed to improve adaptability across different tasks by learning a generalizable meta-policy. Instead of learning each traffic scenario from scratch, the agent is trained to extract transferable knowledge from prior experiences and quickly adapt to new intersections or network conditions, thus improving sample efficiency. However, generalization of the meta-RL-based method may be limited when traffic layouts vary drastically. In addition, other recent methods such as MalLight [83] and DiffLight [84] aim to maintain effective control even in scenarios with incomplete or missing traffic sensor data, further contributing to the data efficiency goal. However, MalLight requires robust inference mechanisms and DiffLight struggles under severe reward delays or noise.

Generalization

In reality, traffic signal flows dynamically vary, which puts higher demands on the generalization ability of traffic signal control methods.

Zhang et al. [85] proposed GeneraLight, which uses generative adversarial networks to produce diverse traffic patterns and improve model generalization. However, GAN training itself may suffer from stability issues. Oroojlooy et al. [86] proposed AttendLight, incorporating attention mechanisms to improve adaptation to various intersection layouts, though the increased model complexity could hinder real-time performance. Besides, some literature focuses on the generalization in different scenarios. For example, the GESA [87] proposed by Jiang et al. and the DuaLight [88] proposed by Lu et al. both adopt Co-Training to facilitate the model to learn generalized dynamic information in different scenarios. However, GESA may be sensitive to the heterogeneity of training and target scenes and DuaLight requires synchronized dual-state information. Jiang et al. [89] introduced X-Light, a transformer-based meta-RL approach achieving strong zero-shot performance across cities, but with a high computational cost due to dual-level transformer design.

Other Issues

In addition to the above research directions, there are some other important research directions in traffic signal control that deserve attention, such as how to provide fair traffic signal phase scheduling [90–92], how to narrow the simulation-to-real performance gap [93], how to ensure traffic safety in the scheduling process [94], and so on. Another emerging topic is explainability in RL-based TSC. For instance, Schreiber et al. [95] studied the trade-offs between model performance and interpretability by comparing explainable and black-box RL approaches in signal control. Gu et al. [96] proposed a programmatic interpretable RL framework, which improves transparency by extracting symbolic decision policies from neural models. These factors cannot be ignored to improve the overall performance of traffic management systems.

Unmanned Aerial Vehicle Navigation

Unmanned aerial vehicles (UAVs) (also known as drones), are increasingly deployed in missions that involve navigating through complex environments, including search-and-rescue operations [97], wildfire monitoring [98], and target tracking [99–102]. UAV navigation requires safe and fast path planning in unknown environments while avoiding obstacles and reducing energy consumption. RL can support

UAV navigation by learning an optimal or suboptimal path under dynamic conditions through continuous interaction with the environment [103, 104]. Acting as an RL agent, the UAV observes environmental states through onboard cameras and sensors, which capture real-time data for state representation. The agent then processes this state information to decide on actions, such as adjusting flight direction or modifying thrust levels of the propellers, maximizing the cumulative rewards. These rewards are tailored to mission objectives, such as minimizing travel time, conserving energy, or avoiding obstacles. Over time, the UAV improves its policy for action control to enhance its ability to autonomously navigate complex, uncertain environments.

In the single-UAV setting, we include representative works that reflect the progression from classical Q-learning to deep reinforcement learning (DRL) in UAV navigation and control. Pham et al. [11] proposed one of the early applications of Q-learning for autonomous UAV navigation, in which the agent adjusts PID controller parameters to navigate within a 2D indoor environment. However, this method is constrained to low-dimensional static tasks and lacks adaptability. Colonnese et al. [13] also applied Q-learning but focused on path planning to improve the quality of experience (QoE) for ground users; however, it does not consider environment dynamics or UAV coordination. Liu et al. [12] introduced a double Q-learning framework for UAV base station control, which enhances stability but assumes relatively stable terrain and demand patterns. Li et al. [14] further advanced control with a DRL framework that uses temporal abstraction and variable decision frequency, though its performance depends on hyperparameter tuning for action abstraction.

In the multi-UAV setting, several multi-agent deep reinforcement learning (MADRL) solutions have been proposed to enhance coordination among UAVs. Wang et al. [15] modeled UAV navigation as a partially observable Markov decision process (POMDP), enabling learning under partial observability, though real-world deployment may suffer from online instability. Bouhamed et al. [16] applied MADDPG to coordinated tasks like formation control and obstacle avoidance; however, performance degrades in environments with noisy communication or weak cooperation. To support exploration in dynamically evolving environments, Moltajaei et al. [18] adopted MAPPO for on-policy multi-agent learning, though its reliance on high-quality exploration data limits robustness. Xue et al. [17] proposed a recurrent MADRL variant (MARDPG) for more stable learning under temporal dependencies, which adds computational overhead. Chikhaoui et al. [19] incorporated energy-awareness into MAPPO to balance task success with power efficiency, yet struggles when tight energy-delay trade-offs are needed. All these recent studies have been listed in Table 2.

Table 2 Representative MARL-based methods for unmanned aerial vehicle navigation

Methods	Key technique	Results	Limitations
Wang et al. [15]	Online DRL under POMDP	Learns navigation in partially observable environments	Online learning may cause instability in real-world
Bouhamed et al. [16]	MADDPG + formation control	Enables coordinated obstacle avoidance and formation	Performance drops in non-cooperative or noisy settings
Moltajaei et al. [18]	MAPPO + on-policy exploration	Efficiently explores complex dynamic areas	Requires high-quality on-policy sampling
Xue et al. [17]	MARDPG + recurrent coordination	Stabilizes learning in 3D UAV control with memory	Recurrent model increases training cost
Chikhaoui et al. [19]	MAPPO + energy constraints	Balances task execution and energy conservation	May underperform under tight energy-delay trade-offs

Air Traffic Management

Air traffic management (ATM) is fundamental to ensuring flight safety and optimizing airspace utilization. Core tasks in ATM include route planning, conflict detection and resolution (CR), traffic flow management, and real-time decision support. As air traffic density increases, particularly in high-demand airspace, there is a growing need for more intelligent, efficient approaches to manage multi-aircraft conflict resolution (MACR) [20, 22, 105], to prevent mid-air conflicts and allow for higher traffic volumes. RL has shown significant promise in addressing these ATM challenges by enabling autonomous decision-making and real-time adaptation for multi-aircraft in complex, high-density environments, building on early foundational work such as that by Tumer et al. [106, 107] on decentralized RL for airspace flow and coordination.

The application of RL in conflict resolution is one of the most explored areas within ATM. Conflict resolution (CR) aims to ensure safe separation between aircraft by adjusting flight parameters such as heading, altitude and speed. Deep RL (DRL) frameworks, including Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), and Actor-Critic methods, have been widely adopted for this purpose due to their effectiveness in sequential decision-making [108, 109]. For multi-aircraft conflict resolution (MACR), where numerous agents (aircraft) operate simultaneously, multi-agent reinforcement learning (MARL) has proven to be particularly valuable. Chen et al. [20] proposed a MARL framework using adaptive maneuver strategies and Rainbow DQN to dynamically adjust aircraft speed and heading. This reduces conflict risk and fuel consumption but requires careful simulation tuning and may generalize poorly across different airspace patterns. Similarly, Brittain et al. [21] developed a variable attention model combining LSTM with MARL to handle scenarios with varying aircraft counts. While effective in managing dynamic CR, the model's complexity and interpretability remain concerns for real-world integration.

Reinforcement learning has also been applied to traffic flow management (TFM), which involves regulating air traffic to prevent congestion and optimize airspace usage. RL-based TFM models aim to reduce flight delays and manage airspace capacity in response to dynamic conditions, such as weather or varying demand. Dalmau et al. [23] implemented a MARL framework using Message Passing Neural Networks (MPNN) [110] to facilitate decentralized decision-making for large-scale air traffic scenarios. In their model, agents (representing aircraft) communicate with neighboring agents to learn optimal actions, such as rerouting or adjusting speed, thereby optimizing overall traffic flow and reducing bottlenecks. This model improves flow efficiency under dynamic demand but relies on consistent inter-agent communication quality. Route optimization is another area where DRL demonstrates notable efficacy. By training RL agents to dynamically adjust flight paths based on real-time traffic and weather data, these systems achieve reduced travel times and fuel consumption. For instance, Guo et al. [24] applied DRL for autonomous route adjustments, enhancing both safety and operational efficiency. However, its performance can degrade at scale due to increasing complexity. In addition, Zhao's research [22] integrated image-based state representations, allowing agents to interpret traffic density and environmental factors more intuitively, leading to better route optimization decisions without predefined flight paths. Nonetheless, the system's reliability heavily depends on the accuracy of visual inputs. The recent RL methods for ATM mentioned above have been summarized in Table 3.

Autonomous Maritime Ship Navigation

With advancements in automation technology, autonomous maritime ship navigation (AMS) has become a significant research focus in maritime transportation. Path planning for unmanned vessels requires safe and efficient navigation in dynamic and complex marine environments, incorporating functions like obstacle avoidance, path optimization, and conflict prevention with other vessels [25, 26, 111].

Table 3 Representative MARL-based methods for air traffic management

Methods	Key technique	Results	Limitations
Chen et al. [20]	MARL + adaptive maneuver + Rainbow DQN	Minimizes conflict and fuel use in multi-aircraft systems	Requires extensive simulation tuning, sensitive to scenario variance
Brittain et al. [21]	Attention-based MARL + LSTM	Handles varying aircraft numbers and enhances CR stability	High model complexity, limited interpretability
Dalmau et al. [23]	MARL + Message Passing Neural Networks	Decentralized TFM for dynamic traffic scenarios	Requires reliable inter-agent communication
Guo et al. [24]	DRL-based route optimization	Balances safety and efficiency under weather/traffic changes	Limited scalability for large-scale implementation
Zhao et al. [22]	DRL + image-based state representation	Supports path decisions under visual traffic context	Performance sensitive to perception accuracy

Table 4 Representative RL-based methods for autonomous maritime ship navigation

Methods	Key technique	Results	Limitations
Wu et al. [26]	DRL + PPO + Dynamic Window Approach (DWA)	Effective real-time path optimization under dynamic obstacles	Performance relies on precise motion modeling and parameter tuning
Guo et al. [25]	DDPG + Artificial Potential Field	Enhances obstacle avoidance precision and planning speed	Susceptible to local minima due to potential field structure
Wei et al., Zhao et al., Li et al. [28–30]	COLREGS-based reward shaping	Improves legal compliance in collision avoidance	Limited adaptability to ambiguous or multi-agent scenarios
Deraj et al. [27]	DQN under 3-DOF motion model	Accurate heading control in simulation	Lacks validation in real-world marine conditions
Xie et al., Niu et al., Zheng et al. [31–33]	MARL (e.g., A3C) for cooperative navigation	Promotes decentralized collision avoidance in multi-ship settings	Scalability and convergence issues under dense traffic

Numerous RL methods have been applied to enable ships to learn optimal routes across the sea surface. In this section, we include representative studies that address key challenges across global and local path planning, as well as multi-agent coordination, with an emphasis on recent RL-based methods.

Path planning for AMS is typically divided into global and local path planning. In global path planning, an overall route can be generated based on known environmental data, ensuring both optimality and safety [112, 113]. However, due to constantly changing conditions in maritime navigation, global path planning struggles to handle dynamic obstacles effectively. Local path planning, by contrast, focuses on real-time obstacle avoidance to adapt to immediate surroundings [114, 115]. In what follows, we focus on RL-based approaches to local path planning for AMS, which aim to support real-time decision-making under uncertainty. Wu et al. [26] adopted a DRL method based on dynamic windows, integrating the PPO algorithm with the Dynamic Window Approach (DWA). This strategy dynamically optimizes navigation by constructing an information window that includes speed and direction parameters. While effective in complex dynamic environments, it requires accurate motion modeling and careful parameter tuning. For improved obstacle avoidance, Guo et al. [25] proposed a DDPG-based model enhanced by artificial potential fields to guide ships around obstacles. Although the method improves path planning accuracy and convergence, it may suffer from local minima due to the nature of potential field

methods. To ensure legal compliance, other studies have embedded the International Regulations for Preventing Collisions at Sea (COLREGS) into the reward functions [28–30]. These approaches allow agents to learn rule-compliant behavior, but their effectiveness may diminish in ambiguous or densely interactive multi-agent settings. At the control level, Deraj et al. [27] used a DQN algorithm to manage heading under a three-degree-of-freedom (3-DOF) motion model. While demonstrating reliable performance in simulations, the method lacks real-world validation and may face robustness issues in physical environments. In multi-ship navigation scenarios, MARL methods such as A3C have been employed for cooperative control [31–33]. These frameworks enhance decentralized collision avoidance and coordination, though they still face challenges in scalability and stable convergence under dense or uncertain traffic. The recent RL methods for AMS have been summarized in Table 4.

Cross-Scenario Multi-agent RL

In this paper, we focus on multi-agent transportation systems, as interactions among agents significantly impact overall system behavior in real-world settings. Current research on multi-agent reinforcement learning (MARL) has been applied extensively to various transportation systems, as discussed in Section 3. However, existing approaches often face a key limitation: agents are typically trained within a

single scenario, which restricts the transferability and adaptability of their policies to new environments. This section investigates cross-scenario MARL, which aims to address these challenges by enabling agents to transfer knowledge across different scenarios and generalize to unseen environments.

We focus on two complementary directions that have shown significant promise in cross-scenario MARL: (1) Offline RL-based methods, which leverage historical datasets from multiple environments to pretrain transferable policies without requiring risky or expensive online interactions; and (2) Meta RL-based methods, which aim to train adaptable agents that can quickly generalize and adapt to new tasks or environments with limited retraining. Together, these methods represent core strategies in advancing the robustness and scalability of MARL in transportation systems. Sections 4.1 and 4.2 detail these two categories, explaining their mechanisms, applications, and limitations.

Offline RL-based Method

Offline RL has emerged as a promising approach to tackle the limitations of cross-scenario MARL in transportation systems. Offline RL enables multiple agents to learn policies from previously collected datasets without requiring real-time interactions with the environment. This capability is particularly advantageous in transportation systems such as traffic signal control, where direct online learning poses risks including traffic congestion and operational delays. By leveraging offline data from multiple scenarios, offline RL-based methods can improve the transferability of learned policies across varied environments.

The key mechanism of offline RL is its reliance on a static dataset containing state-action-reward trajectories, collected from expert policies or simulation environments. Offline RL uses these trajectories to approximate optimal policies without active exploration, often framing the decision-making process as a supervised learning task. In methods such as Decision Transformers and trajectory-based sequence modeling [116–118], offline RL converts the problem into a sequential modeling problem, treating actions as predicted tokens based on past states and rewards.

In cross-scenario applications, offline RL supports policy transferability by pre-training on diverse historical datasets collected from multiple scenarios. For instance, in the TSC domain, TransformerLight [119] addresses the out-of-distribution (OOD) problem by dynamically adjusting a gating mechanism, allowing the model to focus on relevant past experiences during decision making. This improves both learning stability and generalization across intersections. Similarly, the CrossLight framework [120] is explicitly designed for cross-city TSC. It uses offline trajectory data

from source cities to pre-train agents, extracting transferable policy patterns across different urban layouts. During training, CrossLight incorporates auto-regressive masking and sequential prediction techniques to build robust representations, which are later fine-tuned with limited online data in the target city. These applications demonstrate how offline RL can be used effectively in high-risk or resource-limited transportation scenarios such as TSC. Beyond TSC, similar offline RL approaches have been explored in UAV navigation and multi-craft maritime coordination, where simulation-generated datasets are used to bootstrap agents for improved stability and safer deployment in real-world environments.

Limitations

While offline RL significantly reduces reliance on costly or risky online interactions, it faces several challenges in transportation settings. First, it assumes the availability of high-quality, diverse datasets that capture a wide range of state-action-reward dynamics across scenarios. When the offline dataset lacks sufficient coverage, the learned policies may fail to generalize. Second, offline RL is sensitive to distributional shifts—differences between training data and actual deployment conditions can result in policy degradation. Finally, selecting proper architectures (e.g., trajectory models vs. value-based estimators) and hyperparameters remains non-trivial, especially under high-dimensional multi-agent settings. These limitations highlight the need for improved dataset design and adaptive model architectures in future work.

Meta RL-based Method

Meta reinforcement learning (Meta RL) is designed to improve the adaptability and transferability of MARL models across diverse scenarios by training agents to generalize over multiple environments. Unlike offline RL, which relies on pre-collected data, meta RL enhances policy adaptability by allowing agents to quickly fine-tune their learned knowledge in new scenarios, optimizing policies through experience across a wide range of training environments. This is particularly valuable in transportation systems where dynamic interactions among agents vary widely based on context, such as in different cities, road networks, or airspace configurations.

Several meta RL approaches have been proposed to address the generalization challenges in transportation systems. For instance, GeneraLight [85] and AttendLight [86] aim to generalize TSC policies by using generative adversarial networks (GANs) and attention mechanisms respectively, to adapt to diverse traffic flow patterns without

extensive retraining. MetaLight uses gradient-based meta-learning to create adaptable policies across multiple scenarios, enabling TSC agents to transfer their learned policies to new intersections within a scenario by updating local gradients based on nearby traffic states [82]. GESA [87] further extends MetaLight by using a general plug-in module (GPI) that reads unlabeled maps from various cities, co-training an actor-critic agent across these scenarios to enhance policy generalization. MetaVIM [121] and MetaGAT [122] introduce additional methods for inter-agent coordination, using context-based meta-learning and graph attention networks (GATs), respectively, to incorporate information from neighboring intersections. MetaGAT extends MetaLight with GAT and Gated Recurrent Units (GRU) to aggregate observation data from neighboring intersections, though it faces limitations with training stability and performance consistency across highly distinct scenarios [122].

Practical Constraints in Meta RL-based Transfer

While meta RL methods have shown promising generalization capabilities across diverse scenarios, their practical deployment in real-world TSC systems is subject to several structural and operational constraints. A key challenge lies in the structural heterogeneity of intersections. Successful policy transfer typically assumes consistency in the action space (e.g., number and configuration of signal phases, timing options) and observation space (e.g., sensor placement, traffic state representation) between the source and target intersections. For example, a policy trained on a four-phase intersection may not transfer effectively to a three-phase one unless the model architecture is explicitly designed to handle such discrepancies. In addition, variations in traffic demand profiles, road layouts, and inter-agent coordination requirements can significantly hinder the reusability of learned policies or meta-learned embeddings. These challenges are particularly pronounced in value-based meta RL approaches, where even minor mismatches in state-action distributions can lead to unstable training and degraded performance. Consequently, meta RL models often struggle when applied to structurally dissimilar urban settings without additional adaptation mechanisms.

Transformer as Meta Learner

Transformers, originally introduced for natural language processing tasks, have recently gained popularity in reinforcement learning due to their ability to model long-range dependencies and complex sequential patterns. In the context of meta RL, they serve as powerful sequence encoders that can generalize agent behaviors across diverse environments by attending to structured historical interactions. Recent

advancements also leverage Transformers as meta-learning agents, which have shown promising results in multi-scenario MARL applications. Transformer-based architectures, such as X-Light [89] and DTPPO [123], employ hierarchical Transformer structures to enhance policy transferability. For example, X-Light for TSC utilizes a Transformer-on-Transformer (TonT) framework, with a Lower Transformer aggregating states, actions, and rewards for enhanced local coordination, while the Upper Transformer captures generalized decision trajectories across different cities. This dual-transformer structure not only improves inter-agent collaboration but also enhances the adaptability of TSC policies to new city topologies. Notably, X-Light demonstrates exceptional transferability by achieving robust zero-shot performance, allowing the model to effectively operate in unseen urban scenarios without additional retraining. In UAV navigation and coordination, DTPPO employs a Spatial Transformer to model inter-agent interactions among UAVs and a Temporal Transformer to capture dynamic changes over time. This architecture significantly improves cross-scenario policy transfer, enabling UAVs to adapt seamlessly to complex, previously unseen environments. Like X-Light, DTPPO excels in zero-shot transfer settings, where it maintains high navigation efficiency and obstacle avoidance performance even when deployed in entirely new scenarios, highlighting the capability of Transformer-based meta RL methods to generalize across diverse and unpredictable environments.

Multi-scenario co-training has also become a valuable training scheme for meta RL, as it enables agents to learn robust policies by training across multiple environments simultaneously [87–89, 123]. This co-training approach diversifies the agents' experience during training, allowing for more effective generalization and stability when deployed in unseen scenarios. More importantly, by exposing agents to diverse intersection configurations, including variations in phase structures, action spaces, and state encodings, co-training helps the model develop shared policy representations that are less sensitive to structural differences. This is crucial for enabling policy transfer across heterogeneous traffic environments. For example, X-Light co-trains on multiple intersections from different cities in each batch, encouraging the model to learn generalized patterns that can be applied to a wide range of urban traffic scenarios. In DTPPO, UAVs are trained across various complex environments, which improves their ability to adapt to new terrain and obstacle configurations, thus enhancing the overall robustness and transferability of the model.

Limitations

Despite the adaptability that meta RL brings to cross-scenario MARL, there are notable limitations. Meta RL approaches, including Transformer-based methods like X-Light and DTPPO, require significant computational resources for multi-scenario co-training, which can be costly and time-consuming. Additionally, while meta RL models improve generalization across similar scenarios, they still face challenges when transferred to environments with drastically different dynamics or state-action distributions. Moreover, methods that rely heavily on inter-agent coordination, such as MetaGAT, may experience training instability and performance fluctuations, especially in heterogeneous environments where inter-agent dependencies are highly variable. These limitations highlight the need for continued research to improve training stability and reduce computational overhead in meta RL applications for transportation systems.

Outlook

As MARL continues to evolve, several promising directions emerge for advancing cross-scenario adaptability in transportation systems. One potential area is adaptive meta-learning, which could enable models to not only transfer policies but also dynamically adjust to varying environmental characteristics or changes in real time. This might involve developing methods that blend online and offline data to refine policies continuously, thereby enhancing performance in unpredictable scenarios. Another important avenue is to improve data efficiency in MARL. High-quality data collection in transportation systems can be costly and time-intensive, particularly for offline RL approaches. Future research may explore data-efficient training algorithms, such as few-shot learning or semi-supervised learning techniques, to maximize performance with limited data while maintaining high levels of cross-scenario generalization.

Inter-agent coordination is also an area where further advancements could yield significant improvements, especially in complex transportation environments. Methods that allow agents to communicate selectively and adaptively, perhaps through improved attention mechanisms or adaptive communication protocols, may enhance collaboration without overwhelming the system's computational resources. Additionally, enhancing robustness through simulation-to-real transfer methods will be critical in bridging the gap between simulated training environments and real-world applications, ensuring that policies developed in controlled conditions can perform reliably in diverse, real-world settings.

Finally, scalability and computational efficiency remain critical challenges. Transformer-based architectures and meta-learning techniques offer powerful solutions but require substantial computational resources. Optimizing these models to reduce resource consumption—potentially through more efficient neural architectures or distributed computing frameworks—will be essential for their practical deployment in large-scale transportation systems.

Conclusion

This paper reviewed the development of reinforcement learning in transportation systems, focusing on its growing application across complex, dynamic environments and the unique cross-scenario challenges it faces. We first discussed the applications and challenges of RL in traffic signal control, drone navigation, air traffic management, and autonomous ship navigation. Then we highlighted key advancements in multi-agent RL for transportation, including offline RL and meta RL methods, both of which aim to enhance the adaptability and scalability of MARL models across diverse scenarios. Despite these advancements, cross-scenario MARL still encounters significant obstacles, particularly in maintaining policy transferability and robustness when applied to new or unseen environments. Addressing these challenges will require continued innovation in data efficiency, generalization methods, and scalable architectures, paving the way for more resilient and adaptive transportation solutions.

Author Contributions Conceptualization: Jintao Liang, Xinqi Du, Ziyue Li; Formal analysis and investigation: Jintao Liang, Xinqi Du; Writing—original draft preparation: Jintao Liang, Xinqi Du, Yilin Liu; Writing—review and editing: Ziyue Li, Haoyuan Jiang, Zhishuai Li; Supervision: Ziyue Li, Cheng Long, Hua Wei, Wolfgang Ketter; All authors have read and agreed to the published version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Not Applicable.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Approval Not applicable.

Research Involving Human and/or Animals Not applicable.

Informed Consent Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Prashanth L, Bhatnagar S. Reinforcement learning with function approximation for traffic signal control. *IEEE Trans Intell Transp Syst.* 2010;12(2):412–21.
- Genders W, Razavi S. Using a deep reinforcement learning agent for traffic signal control. 2016. arXiv preprint [arXiv:1611.01142](https://arxiv.org/abs/1611.01142).
- Mousavi SS, Schukat M, Howley E. Traffic light control using deep policy-gradient and value-function-based reinforcement learning. *IET Intel Transp Syst.* 2017;11(7):417–23.
- Wei H, Zheng G, Yao H, Li Z. IntelliLight: a reinforcement learning approach for intelligent traffic light control. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18*. New York, NY, USA: Association for Computing Machinery. 2018;2496–2505. <https://doi.org/10.1145/3219819.3220096>.
- Zheng G, Xiong Y, Zang X, Feng J, Wei H, Zhang H, Li Y, Xu K, Li Z. Learning phase competition for traffic signal control. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19*. New York, NY, USA: Association for Computing Machinery; 2019;1963–1972. <https://doi.org/10.1145/3357384.3357900>.
- Wei H, Xu N, Zhang H, Zheng G, Zang X, Chen C, Zhang W, Zhu Y, Xu K, Li Z. CoLight: learning network-level cooperation for traffic signal control. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19*. New York, NY, USA: Association for Computing Machinery; 2019; 1913–1922. <https://doi.org/10.1145/3357384.3357902>.
- Liu Y, Luo G, Yuan Q, Li J, Jin L, Chen B, Pan R. Gplight: grouped multi-agent reinforcement learning for large-scale traffic signal control. In: *IJCAI*. 2023;199–207.
- Liu Y, Liang J, Zhang Y, Gong P, Luo G, Yuan Q, et al. Global-light: exploring global influence in multi-agent deep reinforcement learning for large-scale traffic signal control. *Neurocomput.* 2025;637:130065.
- Wei H, Chen C, Zheng G, Wu K, Gayah V, Xu K, Li Z. Press-Light: learning max pressure control to coordinate traffic signals in arterial network. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19*. New York, NY, USA: Association for Computing Machinery; 2019. pp. 1290–1298. <https://doi.org/10.1145/3292500.3330949>.
- Wang Y, Xu T, Niu X, Tan C, Chen E, Xiong H. STMARL: a Spatio-temporal multi-agent reinforcement learning approach for cooperative traffic light control. *IEEE Trans Mob Comput.* 2022;21(06):2228–42. <https://doi.org/10.1109/TMC.2020.3033782>.
- Pham HX, La HM, Feil-Seifer D, Van Nguyen L. Reinforcement learning for autonomous uav navigation using function approximation. In: *2018 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. 2018;1–6. IEEE.
- Liu X, Chen M, Yin C. Optimized trajectory design in uav based cellular networks for 3d users: a double q-learning approach. *J Commun Inf Netw.* 2019;4(1):24–32.
- Colonnese S, Cuomo F, Pagliari G, Chiaraviglio L. Q-square: a q-learning approach to provide a qoe aware uav flight path in cellular networks. *Ad Hoc Netw.* 2019;91:101872.
- Li C-C, Shuai H-H, Wang L-C. Efficiency-reinforced learning with auxiliary depth reconstruction for autonomous navigation of mobile devices. In: *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*. 2022;458–463. IEEE.
- Wang C, Wang J, Shen Y, Zhang X. Autonomous navigation of uavs in large-scale complex environments: a deep reinforcement learning approach. *IEEE Trans Veh Technol.* 2019;68(3):2124–36.
- Bouhamed O, Ghazzai H, Besbes H, Massoud Y. Autonomous uav navigation: a ddpq-based deep reinforcement learning approach. In: *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2020;1–5. IEEE.
- Xue Y, Chen W. Multi-agent deep reinforcement learning for uavs navigation in unknown complex environment. *IEEE Trans Intell Vehicles.* 2023;9(1):2290–303.
- Farid AM, Roshanian J, Mouhoub M. On-policy actor-critic reinforcement learning for multi-uav exploration. 2024. [arXiv:2409.11058](https://arxiv.org/abs/2409.11058).
- Chikhaoui K, Ghazzai H, Massoud Y. Ppo-based reinforcement learning for uav navigation in urban environments. In: *2022 IEEE 65th International Midwest Symposium on Circuits and Systems (MWSCAS)*. 2022;1–4. IEEE.
- Chen Y, Hu M, Yang L, Xu Y, Xie H. General multi-agent reinforcement learning integrating adaptive manoeuvre strategy for real-time multi-aircraft conflict resolution. *Transp Res Part C Emerg Technol.* 2023;151:104125.
- Brittain MW, Wei P. One to any: distributed conflict resolution with deep multi-agent reinforcement learning and long short-term memory. In: *AIAA Scitech 2021 Forum*. 2021;1952.
- Zhao P, Erzberger H, Liu Y. Multiple-aircraft-conflict resolution under uncertainties. *J Guid Control Dyn.* 2021;44(11):2031–49.
- Dalmau R, Allard E. Air traffic control using message passing neural networks and multi-agent reinforcement learning. In: *Proc of the 10th SESAR Innov Days, Virtual Event*. 2020;7–10.
- Guo W, Brittain M, Wei P. Safety enhancement for deep reinforcement learning in autonomous separation assurance. In: *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. 2021;348–354. IEEE.
- Guo S, Zhang X, Zheng Y, Du Y. An autonomous path planning model for unmanned ships based on deep reinforcement learning. *Sensor.* 2020;20(2):426.
- Wu C, Yu W, Li G, Liao W. Deep reinforcement learning with dynamic window approach based collision avoidance path planning for maritime autonomous surface ships. *Ocean Eng.* 2023;284:115208.
- Deraj R, Kumar RS, Alam MS, Somayajula A. Deep reinforcement learning based controller for ship navigation. *Ocean Eng.* 2023;273:113937.
- Wei G, Kuo W. Colregs-compliant multi-ship collision avoidance based on multi-agent reinforcement learning technique. *J Marine Sci Eng.* 2022;10(10):1431.
- Zhao L, Roh M-I. Colregs-compliant multiship collision avoidance based on deep reinforcement learning. *Ocean Eng.* 2019;191:106436.
- Li L, Wu D, Huang Y, Yuan Z-M. A path planning strategy unified with a colregs collision avoidance function based on deep reinforcement learning and artificial potential field. *Appl Ocean Res.* 2021;113:102759.

31. Xie S, Chu X, Zheng M, Liu C. A composite learning method for multi-ship collision avoidance based on reinforcement learning and inverse control. *Neurocomput.* 2020;411:375–92.
32. Niu Y, Zhu F, Wei M, Du Y, Zhai P. A multi-ship collision avoidance algorithm using data-driven multi-agent deep reinforcement learning. *J Marine Sci Eng.* 2023;11(11):2101.
33. Zheng K, Zhang X, Wang C, Zhang M, Cui H. A partially observable multi-ship collision avoidance decision-making model based on deep reinforcement learning. *Ocean Coastal Manag.* 2023;242:106689.
34. Li Y. Deep reinforcement learning: an overview. 2017. arXiv preprint [arXiv:1701.07274](https://arxiv.org/abs/1701.07274).
35. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *Nature.* 2015;518(7540):529–33.
36. Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* 2016;30.
37. Sutton RS. *Reinforcement learning: an introduction.* Bradford: A Bradford Book; 2018.
38. Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K. Asynchronous methods for deep reinforcement learning. In: *International Conference on Machine Learning.* 2016; 1928–1937. PmLR.
39. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017. arXiv preprint [arXiv:1707.06347](https://arxiv.org/abs/1707.06347).
40. Du X, Chen H, Xing Y, Philip SY, He L. A contrastive-enhanced ensemble framework for efficient multi-agent reinforcement learning. *Expert Syst Appl.* 2024;245:123158.
41. Tan M. Multi-agent reinforcement learning: Independent vs. cooperative agents. In: *Proceedings of the Tenth International Conference on Machine Learning.* 1993;330–337.
42. Du X, Chen H, Wang C, Xing Y, Yang J, Philip SY, et al. Robust multi-agent reinforcement learning via bayesian distributional value estimation. *Pattern Recogn.* 2024;145:109917.
43. Iqbal S, Sha F. Actor-attention-critic for multi-agent reinforcement learning. In: *International Conference on Machine Learning.* 2019;2961–2970. PMLR.
44. Yu C, Velu A, Vinitzky E, Gao J, Wang Y, Bayen A, et al. The surprising effectiveness of ppo in cooperative multi-agent games. *Adv Neural Inf Proc Syst.* 2022;35:24611–24.
45. Sukhbaatar S, Fergus R, et al. Learning multiagent communication with backpropagation. *Adv in Neural Inf Proc Syst.* 2016;29.
46. Wei H, Zheng G, Gayah V, Li Z. A survey on traffic signal control methods. 2019. arXiv preprint [arXiv:1904.08117](https://arxiv.org/abs/1904.08117).
47. Haydari A, Yilmaz Y. Deep reinforcement learning for intelligent transportation systems: a survey. *IEEE Trans Intell Transp Syst.* 2020;23(1):11–32.
48. Bazzan AL. Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Auton Agent Multi-Agent Syst.* 2009;18:342–75.
49. Bazzan AL, Oliveira D, Silva BC. Learning in groups of traffic signals. *Eng Appl Artif Intell.* 2010;23(4):560–8. <https://doi.org/10.1016/j.engappai.2009.11.009>.
50. Mannion P, Duggan J, Howley E. An experimental review of reinforcement learning algorithms for adaptive traffic signal control. *Auto Road Transp Support Syst.* 2016;47–66.
51. Yau K-LA, Qadir J, Khoo HL, Ling MH, Komisarczuk P. A survey on reinforcement learning models and algorithms for traffic signal control. *ACM Comput Surv (CSUR).* 2017;50(3):1–38.
52. Noaen M, Naik A, Goodman L, Crebo J, Abrar T, Abad ZSH, et al. Reinforcement learning in urban network traffic signal control: a systematic literature review. *Expert Syst Appl.* 2022;199:116830.
53. Webster FV. Traffic signal settings. In: *Technical report.* 1958.
54. Singh L, Tripathi S, Arora H. Time optimization for traffic signal control using genetic algorithm. *Int J Recent Trends Eng.* 2009;2(2):4.
55. Yarger BW. Fully actuated vs. semi-actuated traffic signal systems. Yarger Eng. Indianapolis. 1993.
56. Hunt P, Robertson D, Bretherton R, Royle MC. The scoot on-line traffic signal optimisation technique. *Traffic Eng Cont.* 1982;23(4).
57. Lowrie P. SCATS. Sydney co-ordinated adaptive traffic system: a traffic responsive method of controlling urban traffic. 1990.
58. Varaiya P. Max pressure control of a network of signalized intersections. *Transp Res Part C Emerg Technol.* 2013;36:177–95.
59. Abdulhai B, Pringle R, Karakoulas GJ. Reinforcement learning for true adaptive traffic signal control. *J Transp Eng.* 2003;129(3):278–85.
60. Arel I, Liu C, Urbanik T, Kohls AG. Reinforcement learning-based multi-agent system for network traffic signal control. *IET Intel Transp Syst.* 2010;4(2):128–35.
61. Wiering MA. et al. Multi-agent reinforcement learning for traffic light control. In: *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000).* 2000;1151–1158.
62. Wiering MA. Convergence and divergence in standard and averaging reinforcement learning. In: *European Conference on Machine Learning.* Springer. 2004;477–488.
63. Bazzan AL. A distributed approach for coordination of traffic signal agents. *Auton Agent Multi-Agent Syst.* 2005;10:131–64.
64. Zhang L, Wu Q, Shen J, Lü L, Du B, Wu J. Expression might be enough: representing pressure and demand for reinforcement learning based traffic signal control. In: *Proceedings of the 39th International Conference on Machine Learning.* 2022;26645–26654. PMLR.
65. Liang X, Du X, Wang G, Han Z. A deep reinforcement learning network for traffic light cycle control. *IEEE Trans Veh Technol.* 2019;68(2):1243–53. <https://doi.org/10.1109/TVT.2018.2890726>.
66. Zhao C, Hu X, Wang G. PRGLight: a novel traffic light control framework with Pressure-based-Reinforcement Learning and Graph Neural Network: IJCAI2021.
67. Chu T, Wang J, Codecà L, Li Z. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Trans Intell Transp Syst.* 2020;21(3):1086–95. <https://doi.org/10.1109/TITS.2019.2901791>.
68. Chen C, Wei H, Xu N, Zheng G, Yang M, Xiong Y, et al. Toward a thousand lights: decentralized deep reinforcement learning for large-scale traffic signal control. *Proc AAAI Conf Artif Intell.* 2020;34(04):3414–21. <https://doi.org/10.1609/aaai.v34i04.5744>.
69. El-Tantawy S, Abdulhai B, Abdelgawad H. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (marlin-atsc): methodology and large-scale application on downtown toronto. *IEEE Trans Intell Transp Syst.* 2013;14(3):1140–50.
70. Kuyer L, Whiteson S, Bakker B, Vlassis N. Multiagent reinforcement learning for urban traffic control using coordination graphs. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15–19, 2008, Proceedings, Part I 19.* Springer. 2008;656–671.
71. Aslani M, Mesgari MS, Wiering M. Adaptive traffic signal control with actor-critic methods in a real-world traffic network with different traffic disruption events. *Transp Res Part C Emerg Technol.* 2017;85:732–52.
72. Abdoos M, Mozayani N, Bazzan AL. Traffic light control in non-stationary environments based on multi agent q-learning. In: *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC).* 2011;1580–1585. IEEE.

73. Devailly F-X, Larocque D, Charlin L. IG-RL: inductive graph reinforcement learning for massive-scale traffic signal control. *IEEE Trans Intell Transp Syst.* 2022;23(7):7496–507. <https://doi.org/10.1109/TITS.2021.3070835>.
74. Zeng Z. GraphLight: graph-based reinforcement learning for traffic signal control. In: 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS). 2021;645–650. <https://doi.org/10.1109/ICCCS52626.2021.9449147>.
75. Sun Y, Lin K, Bashir AK. KeyLight: intelligent traffic signal control method based on improved graph neural network. *IEEE Trans Consum Electron.* 2024;70(1):2861–71. <https://doi.org/10.1109/TCE.2023.3272524>.
76. Wu L, Wang M, Wu D, Wu J. DynSTGAT: dynamic spatial-temporal graph attention network for traffic signal control. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. CIKM '21. New York, NY, USA: Association for Computing Machinery. 2021;2150–2159. <https://doi.org/10.1145/3459637.3482254>.
77. Ma J, Wu F. Feudal multi-agent deep reinforcement learning for traffic signal control. In: Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems. 2020;816–824.
78. Xu B, Wang Y, Wang Z, Jia H, Lu Z. Hierarchically and cooperatively learning traffic signal control. *Proc AAAI Conf Artif Intell.* 2021;35(1):669–77. <https://doi.org/10.1609/aaai.v35i1.16147>.
79. Wang M, Wu L, Li J, He L. Traffic signal control with reinforcement learning based on region-aware cooperative strategy. *IEEE Trans Intell Transp Syst.* 2022;23(7):6774–85. <https://doi.org/10.1109/TITS.2021.3062072>.
80. Ruan J, Li Z, Wei H, Jiang H, Lu J, Xiong X, Mao H, Zhao R: CoSLight: co-optimizing collaborator selection and decision-making to enhance traffic signal control. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '24. New York, NY, USA: Association for Computing Machinery. 2024;2500–2511. <https://doi.org/10.1145/3637528.3671998>.
81. Xiong Y, Zheng G, Xu K, Li Z. Learning traffic signal control from demonstrations. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19. New York, NY, USA: Association for Computing Machinery; 2019;2289–2292. <https://doi.org/10.1145/3357384.3358079>.
82. Zang X, Yao H, Zheng G, Xu N, Xu K, Li Z. MetaLight: value-based meta-reinforcement learning for traffic signal control. *Proc AAAI Conf Artif Intell.* 2020;34(01):1153–60. <https://doi.org/10.1609/aaai.v34i01.5467>.
83. Yang Q, Xie Z, Wei H, Zhang D, Yang Y. MalLight: influence-aware coordinated traffic signal control for traffic signal malfunctions. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. CIKM '24. New York, NY, USA: Association for Computing Machinery. 2024;2879–2889. <https://doi.org/10.1145/3627673.3679605>.
84. Chen H, Jiang Y, Guo S, Mao X, Lin Y, Wan H. DiffLight: a partial rewards conditioned diffusion model for traffic signal control with missing data. *Adv Neural Inf Process Syst.* 2024;37:123353–78.
85. Zhang H, Liu C, Zhang W, Zheng G, Yu Y. GeneralLight: improving environment generalization of traffic signal control via meta reinforcement learning. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. CIKM '20. New York, NY, USA: Association for Computing Machinery; 2020;1783–1792. <https://doi.org/10.1145/3340531.3411859>.
86. Oroojlooy A, Nazari M, Hajinezhad D, Silva J. AttendLight: universal attention-based reinforcement learning model for traffic signal control. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20. 2020;4079–4090.
87. Jiang H, Li Z, Li Z, Bai L, Mao H, Ketter W, et al. A general scenario-agnostic reinforcement learning for traffic signal control. *IEEE Trans Intell Transp Syst.* 2024;25(9):11330–44. <https://doi.org/10.1109/TITS.2024.3377106>.
88. Lu J, Ruan J, Jiang H, Li Z, Mao H, Zhao R. DuaLight: enhancing traffic signal control by leveraging scenario-specific and scenario-shared knowledge. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. AAMAS '24. 2024;1283–1291.
89. Jiang H, Li Z, Wei H, Xiong X, Ruan J, Lu J, et al. X-Light: cross-city traffic signal control using transformer on transformer as meta multi-agent reinforcement learner. *Thirty-Third Int Joint Conf on Artif Intell.* 2024;1:94–102. <https://doi.org/10.24963/ijcai.2024/11>.
90. Chen J, Zhang Z, Feng J, Zhu K. FIT: fairness-aware intelligent traffic signal control with deep reinforcement learning. In: 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys). 2021; 846–852. <https://doi.org/10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00136>.
91. Ye Y, Ding J, Wang T, Zhou J, Wei X, Chen M. FairLight: fairness-aware autonomous traffic signal control with hierarchical action space. *IEEE Trans Comput Aided Des Integr Circ Syst.* 2023;42(8):2434–46. <https://doi.org/10.1109/TCAD.2022.3226673>.
92. Du X, Li Z, Long C, Xing Y, Philip SY, Chen H. Felight: fairness-aware traffic signal control via sample-efficient reinforcement learning. *IEEE Trans on Knowl and Data Eng.* 2024;36(9):4678–90.
93. Da L, Gao M, Mei H, Wei H. Prompt to transfer: sim-to-real transfer for traffic signal control with prompt learning. *Proc AAAI Conf Artif Intell.* 2024;38:82–90.
94. Du W, Ye J, Gu J, Li J, Wei H, Wang G. Safelight: a reinforcement learning method toward collision-free traffic signal control. *Proc AAAI Conf Artif Intell.* 2023;37:14801–10.
95. Schreiber LV, Alegre LN, Bazzan AL, Ramos GO. On the explainability and expressiveness of function approximation methods in rl-based traffic signal control. In: 2022 International Joint Conference on Neural Networks (IJCNN). 2022;1–8. <https://doi.org/10.1109/IJCNN55064.2022.9892422>. IEEE.
96. Gu Y, Zhang K, Liu Q, Gao W, Li L, Zhou J. π -light: programmatic interpretable reinforcement learning for resource-limited traffic signal control. *Proc AAAI Conf Artif Intell.* 2024;38:21107–15. <https://doi.org/10.1609/aaai.v38i19.30103>.
97. Tomic T, Schmid K, Lutz P, Domel A, Kassecker M, Mair E, et al. Toward a fully autonomous uav: research platform for indoor and outdoor urban search and rescue. *IEEE Robot Auto Mag.* 2012;19(3):46–56.
98. Pham HX, La HM, Feil-Seifer D, Deans M. A distributed control framework for a team of unmanned aerial vehicles for dynamic wildfire tracking. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2017. pp. 6648–6653. IEEE.
99. Woods AC, La HM, Ha QP. A novel extended potential field controller for use on aerial robots. In: 2016 IEEE International Conference on Automation Science and Engineering (CASE). 2016;286–291. IEEE.
100. Woods AC, La HM. Dynamic target tracking and obstacle avoidance using a drone. In: Advances in Visual Computing: 11th International Symposium, ISVC 2015, December 14–16, 2015, Proceedings, Part I 11. Las Vegas, NV, USA: Springer; 2015;857–866.

101. Muñoz F, Espinoza Quesada ES, La HM, Salazar S, Commuri S, Garcia Carrillo LR. Adaptive consensus algorithms for real-time operation of multi-agent systems affected by switching network events. *Int J Robust Nonlinear Control*. 2017;27(9):1566–88.
102. Bondi E, Fang F, Hamilton M, Kar D, Dmello D, Choi J, et al. Spot poachers in action: augmenting conservation drones with automatic detection in near real time. *Proc of the AAAI Conf on Artif Intell*. 2018;32. <https://doi.org/10.1609/aaai.v32i1.11414>.
103. Kurunathan H, Huang H, Li K, Ni W, Hossain E. Machine learning-aided operations and communications of unmanned aerial vehicles: a contemporary survey. *IEEE Commun Surv Tutor*. 2023;26(1):496–533.
104. Azar AT, Koubaa A, Ali Mohamed N, Ibrahim HA, Ibrahim ZF, Kazim M, et al. Drone deep reinforcement learning: a review. *Electronics*. 2021;10(9):999.
105. Razzaghi P, Tabrizian A, Guo W, Chen S, Taye A, Thompson E, et al. A survey on reinforcement learning in aviation applications. *Eng Appl Artif Intell*. 2024;136:108911.
106. Tumer K, Agogino A. Distributed agent-based air traffic flow management. In: *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*. 2007. pp. 1–8.
107. Tumer K, Agogino A. Improving air traffic management with a learning multiagent system. *IEEE Intell Syst*. 2009;24(1):18–21.
108. Wang Z, Pan W, Li H, Wang X, Zuo Q. Review of deep reinforcement learning approaches for conflict resolution in air traffic control. *Aerospace*. 2022;9(6):294.
109. Agogino AK, Tumer K. A multiagent approach to managing air traffic flow. *Auton Agent Multi-Agent Syst*. 2012;24:1–25. <https://doi.org/10.1007/s10458-010-9142-5>.
110. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: *International Conference on Machine Learning*. 2017;1263–1272. PMLR.
111. Öztürk Ü, Akdağ M, Ayabakan T. A review of path planning algorithms in maritime autonomous surface ships: navigation safety perspective. *Ocean Eng*. 2022;251:111010.
112. Yun S-W, Kim D-H, Kim S-W, Kim D-J, Kim H-J. Global path planning for autonomous ship navigation considering the practical characteristics of the port of ulsan. *J Marine Sci Eng*. 2024;12(1):160.
113. Wang G, Wang J, Wang X, Wang Q, Han J, Chen L, et al. A method for coastal global route planning of unmanned ships based on human-like thinking. *J Marine Sci Eng*. 2024;12(3):476.
114. Wang G, Wang J, Wang X, Wang Q, Chen L, Han J, et al. Local path planning method for unmanned ship based on encounter situation inference and colregs constraints. *J Marine Sci Eng*. 2024;12(5):720.
115. Ohn S-W, Namgung H. Requirements for optimal local route planning of autonomous ships. *J Marine Sci Eng*. 2022;11(1):17.
116. Chen L, Lu K, Rajeswaran A, Lee K, Grover A, Laskin M, et al. Decision transformer reinforcement learning via sequence modeling. *Adv Neural Inf Proc Syst*. 2021;34:15084–97.
117. Mao H, Zhao R, Li Z, Xu Z, Chen H, Chen Y, Zhang B, Xiao Z, Zhang J, Yin J. Pdit: interleaving perception and decision-making transformers for deep reinforcement learning. 2023. arXiv preprint [arXiv:2312.15863](https://arxiv.org/abs/2312.15863).
118. Janner M, Li Q, Levine S. Offline reinforcement learning as one big sequence modeling problem. *Adv in Neural Inf Proc Syst*. 2021;34:1273–86.
119. Wu Q, Li M, Shen J, Lü L, Du B, Zhang K. Transformerlight: a novel sequence modeling based traffic signaling mechanism via gated transformer. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023;2639–2647.
120. Sun Q, Zha R, Zhang L, Zhou J, Mei Y, Li Z, Xiong H. Cross-light: offline-to-online reinforcement learning for cross-city traffic signal control. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024;2765–2774.
121. Zhu L, Peng P, Lu Z, Tian Y. Metavim: meta variationally intrinsic motivated reinforcement learning for decentralized traffic signal control. *IEEE Trans Knowl Data Eng*. 2023;35(11):11570–84.
122. Lou Y, Wu J, Ran Y. Meta-reinforcement learning for multiple traffic signals control. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022; 4264–4268.
123. Wei A, Liang J, Lin K, Li Z, Zhao R. Dtpo: dual-transformer encoder-based proximal policy optimization for multi-uav navigation in unseen complex environments. 2024. arXiv preprint [arXiv:2410.15205](https://arxiv.org/abs/2410.15205).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.