Constrained Multi-Agent Reinforcement Learning with MAF-Net for Safe Trajectory Planning

Bizhao Pang†
Air Traffic Management Research Institute, Nanyang Technological University
Singapore
bizhao001@e.ntu.edu.sg

Mingcheng Zhang†
Air Traffic Management Research Institute, Nanyang Technological University
Singapore
m200057@e.ntu.edu.sg

Xinting Hu
Air Traffic Management Research Institute, Nanyang Technological University
Singapore
xinting.hu@ntu.edu.sg

Duc-Thinh Pham
Center for AI Research, VinUniversity
Hanoi, Vietnam
Air Traffic Management Research Institute, Nanyang Technological University
Singapore
thinh.pd@vinuni.edu.vn

Sameer Alam*
Air Traffic Management Research Institute, Nanyang Technological University
Singapore
sameeralam@ntu.edu.sg

Guglielmo Lulli
Dept of Informatics, Systems and Communication, University of Milano-Bicocca
Milan, Italy
guglielmo.lulli@unimib.it

ABSTRACT

Multi-agent trajectory planning in safety-critical systems needs to ensure safety while scaling to many agents. Sampling and optimization methods often adapt slowly and scale poorly. Reinforcement learning can improve adaptability, but it often violates safety constraints and suffers sample inefficiency. This work proposes IDDPG-MAF, which integrates Independent Deep Deterministic Policy Gradient (IDDPG) with a pre-trained Multi-head Action Filter Network (MAF-Net). We first cast the problem as a constrained mixed-integer nonlinear program and then reformulate it as a constrained decentralized Markov decision process for real-time adaptability and coordination. IDDPG enables scalable learning, while MAF-Net acts as a differentiable safety filter that masks unsafe actions and penalizes suboptimal behaviors. The IDDPG-MAF method is adapted to a complex multi-aircraft trajectory planning task under dynamic thunderstorm cells. Experimental results show that IDDPG-MAF achieves over 99% safe separation (vs. 82% for the state-of-the-art baseline), 95.5% task success even under moderate uncertainty, and scales safely to 45 aircraft in a compact spatiotemporal window, effectively doubling the maximum capacity of current operations.

KEYWORDS

Multi-agent system; planning under uncertainty; decentralized decision making; deep reinforcement learning; action masking

ACM Reference Format:
Bizhao Pang, Mingcheng Zhang, Xinting Hu, Duc-Thinh Pham, Sameer Alam, and Guglielmo Lulli. 2026. Constrained Multi-Agent Reinforcement Learning with MAF-Net for Safe Trajectory Planning. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/MQDV9851

†Both authors contributed equally to this paper.
*Corresponding author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/MQDV9851

1 INTRODUCTION

Real-time decision-making in safety-critical transportation systems, such as air traffic management (ATM) [4], urban air mobility [37], autonomous driving [42], and traffic light control [23] involves complex challenges arising from operational constraints and environmental uncertainty. In the ATM domain, multi-aircraft trajectory planning during dynamically evolving thunderstorms is a representative case [10, 31], where parts of the airspace are intermittently blocked. In such conditions, aircraft must simultaneously adjust their trajectories within limited maneuvering space while maintaining safe separation from one another. The presence of multiple aircraft in a confined region creates high-density interactions, where the motion of each agent reshapes the feasible action space of the others. This results in a non-stationary and highly coupled decision space, especially under uncertainties.

Reinforcement learning (RL) has shown potential in addressing such dynamic and sequential planning problems [4, 15, 48]. However, applying RL to safety-critical domains remains difficult due to the need to satisfy operational constraints [24]. Most existing RL methods for trajectory planning [32, 35] focus on reward maximization but lack mechanisms to enforce safety constraints. Furthermore, RL in continuous spaces suffers from sample inefficiency and frequently explores irrelevant or unsafe actions [40], impeding learning and safety. Recent approaches, such as constrained RL [45] and action masking [6], have contributed to addressing these issues. However, many rely on non-differentiable or hard masking components that disrupt the gradient flow and hinder stable policy learning.

To address these challenges, this work proposes an IDDPG-MAF method that integrates the Independent Deep Deterministic Policy

Gradient (IDDPG) algorithm with a Multi-head Action Filter Network (MAF-Net). IDDPG supports scalable policy learning, enabling cooperative decision-making across multiple agents. MAF-Net is designed as a pre-trained and gradient-preserving module that filters unsafe and suboptimal actions based on time-dependent observations. By decoupling safety filtering from policy learning, MAF-Net avoids disrupting the gradient flow and enhances learning stability in continuous action spaces. The IDDPG-MAF method is adapted to a complex multi-aircraft trajectory planning task, where multiple high-speed aircraft agents must coordinate in real time, adapt to dynamic thunderstorm movements, and maintain safe separation under uncertainties. Beyond aviation applications, MAF-Net provides a generalizable module for other safety-critical systems by serving as a safety layer that filters unsafe actions without interfering with the learning process. The main contributions are summarized as follows:

- We propose a Multi-Head Action Filter Network (MAF-Net) that introduces a pre-trained and gradient-preserving action filtering mechanism for continuous action spaces. Unlike prior RL methods that rely on non-differentiable or hard filters, MAF-Net enables real-time classification of actions into unsafe, desired, and suboptimal categories, allowing safe exploration without disrupting policy learning.
- We integrate MAF-Net with an IDDPG algorithm under a constrained Decentralized Markov Decision Process (Dec-MDP) framework. We then incorporate Constrained Policy Optimization to enforce safety constraints while preserving scalable policy updates. This formulation provides insights for the modeling of any other multi-robot collision-avoidance and navigation tasks.
- We demonstrate the proposed IDDPG-MAF method in a real-world multi-agent trajectory planning scenario under uncertainty. Experimental results show that our method outperforms baseline algorithms in constraint satisfaction, learning efficiency, and scalability.

2 RELATED WORKS

This section reviews traditional trajectory planning and reinforcement learning methods, as well as recent advances in action masking and constrained reinforcement learning.

Traditional trajectory planning methods primarily use optimization techniques to ensure safe separation [20, 34]. Early work [9] introduced automated conflict detection tools, while later studies incorporated probabilistic factors such as weather and trajectory uncertainties. For example, Ng et al. [29] proposed a dynamic programming for weather-dependent rerouting, Kamgarpour et al. [21] developed constrained optimization models to guarantee safe trajectories, and Maliah et al. [27] extended Conflict-Based Search to minimize makespan in multi-agent path finding, enabling optimal conflict resolution in time-critical tasks. Robust control method [13] has also been employed to manage wind disturbances. However, these approaches often suffer from high computational complexity, making them unsuitable for real-time decision-making problems.

Reinforcement learning (RL) has been applied to trajectory planning as an effective alternative. Early efforts focused on single-agent trajectory planning with deep Q-learning [5], which improved safety in dense traffic but relied on centralized control. Decentralized approaches were later explored: Nguyen et al. [30] proposed A-MCTS, a decentralized Monte Carlo Tree Search method resilient to agent failures; Yan et al. [44] introduced TSCAL, a curriculum-based MARL framework for UAV flocking; Diller et al. [8] modeled drone–UGV cooperation as a convex program optimizing long-horizon coordination and energy sharing; Pham et al. [36] developed multi-agent DDPG for conflict resolution, enhancing efficiency and scalability; and Hua et al. [18] proposed CAMP, a collaborative attention-based MARL model for profiled vehicle routing. Other works validated MARL in simulated environments [35] or emphasized alignment with human preferences [17]. Beyond coordination, Zhao et al. [49] incorporated physics-informed models to improve generalization and interpretability. Further decentralized models leveraged local observations for agent cooperation in structured environments [43, 47]. More generally, Zhu et al. [53] proposed a unified type-based framework for single-agent planning in multi-agent environments that formalizes the exploration–exploitation trade-off. However, most of these approaches lack explicit safety enforcement, which motivates recent interest in action masking and constrained RL to integrate safety-aware filtering with efficient policy learning.

Moss et al. [28] introduced adaptive constraints based on failure probabilities, while Zhong et al. [50] applied KL-based discrete masking to eliminate redundant actions. Santana et al. [38] handled chance-constrained POMDPs by propagating risk bounds in belief space for discrete-action planning, while Yu et al. [46] and Ben-Iwhiwhu et al. [2] proposed self-supervised and modulating masks. Cheng et al. [6] (StateMask) and Wu et al. [43] (iterative masking) improved interpretability and scalability in large spaces. Most methods, however, rely on discrete and non-differentiable filters. Stolz et al. [40] extended masking to continuous domains via convex projections, improving efficiency in control tasks. Safety-constrained RL integrates explicit constraints into policy optimization [24]. Yu et al. [45] proposed a reachability-constrained method for feasible set learning. Goodall et al. [14] developed a shielding approach to verify adherence, and Zhou et al. [51] introduced uniformly constrained MDPs to reduce long-tail violations. When constraints are unknown, Wachi et al. [41] designed a safe near-optimal MDP. More recently, Liu et al. [25] leveraged large language models (LLMs) to dynamically prune unsafe state-action pairs in unstructured environments. Complementary to these algorithmic approaches, Shefin et al. [39] proposed xSRL, a framework that integrates explainability into safe RL to enhance trust and robustness. Despite progress, action masking and constrained RL remain limited in safety-critical domains with continuous action spaces, motivating our gradient-preserving MAF-Net for multi-agent trajectory planning under dynamic threats and uncertainties.

3 PROBLEM FORMULATION

Trajectory planning in safety-critical domains requires handling multi-agent interactions, time-sensitive decisions, and dynamic threats. Agents must coordinate in shared and constrained spaces where the actions of one reshape the feasible options of others, creating a highly coupled and non-stationary decision space. We

study real-time rerouting of multiple aircraft to avoid evolving thunderstorm cells while maintaining safe separation. In regions with limited air traffic control services, aircraft must make decentralized decisions under uncertainty, requiring scalable coordination for safety and efficiency.

Assumptions. (1) The environment is modeled in two dimensions at a fixed flight level (RVSM operations) [19], since altitude changes are rarely used for tactical rerouting and thunderstorm heights often exceed the performance limits of civilian aircraft. (2) Aircraft agents share state information (position, heading, speed, route) at one-second intervals via onboard surveillance.

3.1 Mathematical Formulation

Notations: Let $F = \{f_1, f_2, \ldots, f_n\}$ denote the set of aircraft, and $P_{exit} = \{p_{ex_1}, p_{ex_2}, \ldots, p_{ex_l}\}$ denote the set of exit waypoints' position, $p_{ex_l} \in \mathbb{R}^2$. Time is represented by the discrete set $T = \{t_0, t_1, \ldots, t_{\text{final}}\}$, where t_0 and t_{final} are the rerouting start and end time, respectively. The set of thunderstorm cells at time t is denoted by $O(t) = \{o_1(t), o_2(t), \ldots, o_k(t)\}$, with each cell's centroid position represented by $p_{O_k}(t)$. The position of aircraft i at time t is given by $p_i(t) \in \mathbb{R}^2$, corresponding to its latitude and longitude. The minimum separation distance between aircraft is denoted by $d_{\min}^{\text{flight}}$, and between aircraft and thunderstorm cells by $d_{\min}^{\text{cell}}$. Here $R_{\text{major\_axis}}^{\text{cell}}$ denotes the major-axis radius of the ellipse used to capture the irregular shape of the storm cell.

Decision variables: The heading change of aircraft i at time t is represented by the continuous variable $\Delta h_i(t)$. The heading direction $\alpha_i(t)$ is an artificial variable that accumulates heading changes over time. The position $p_i(t)$ is also an artificial variable denoting the location of aircraft i at t.

Constrained optimization problem:

$$\min \sum_{i=1}^{n} \sum_{t=t_0}^{t_{\text{final}}} \|p_i(t+1) - p_i(t)\| \tag{1}$$

$$\min \sum_{i=1}^{n} \sum_{t=t_0}^{t_{\text{final}}} \left| \frac{\Delta h_i(t)}{\Delta h} \right| \tag{2}$$

Subject to:

$$P_r\left(\|p_i(t) - p_j(t)\| < d_{\min}^{\text{flight}}\right) < \epsilon, \forall i, j \in \{1, \ldots, n\}, \quad i \neq j, \quad \forall t \in T \tag{3}$$

$$P_r\left(\|p_i(t) - p_{O_k}(t)\| < (d_{\min}^{\text{cell}} + R_{\text{major\_axis}}^{\text{cell}_k})\right) < \epsilon, \quad \forall i \in \{1, \ldots, n\}, \forall k \in \{1, \ldots, K\} \quad \forall t \in T \tag{4}$$

$$p_i(t_{\text{final}}) \in P_{exit}, \quad \forall i \in \{1, \ldots, n\} \tag{5}$$

$$p_i(t+1) = p_i(t) + v_i(t)\Delta t \begin{bmatrix} \cos(\alpha_i(t+1)) \\ \sin(\alpha_i(t+1)) \end{bmatrix} (1+u) \tag{6}$$

$$\alpha_i(t+1) = \alpha_i(t) + \Delta h_i(t) \tag{7}$$

The problem has two objectives: (i) minimize the total distance from each aircraft's start position $p_i(t_0)$ to its assigned exit waypoint $p_i(t_{\text{final}})$ (Eq. 1); and (ii) minimize the number of heading changes $\Delta h_i(t)$ to reduce space complexity (Eq. 2). Here, Δh is a constant threshold per time step, yielding the normalized term $\left|\frac{\Delta h_i(t)}{\Delta h}\right| \in [0, 1]$. Chance constraints (Eq. 3 and Eq. 4) limit the probability that any aircraft pair f_i, f_j violates the minimum separation with another aircraft or a thunderstorm cell, keeping the violation probability below tolerance ϵ. Constraint (Eq. 5) requires each agent to reach exactly one waypoint in P_{exit}, while permitting multiple agents to share the same destination. The motion dynamics of each agent are defined by position updates based on velocity $v_i(t)$, heading $\alpha_i(t)$, and time step Δt (Eq. 6). The heading direction is updated in Eq. 7 by adding $\Delta h_i(t)$ to the current angle $\alpha_i(t)$.

Real-world position uncertainty is modeled as a dimensionless multiplicative noise $u \sim \mathcal{N}(0, \sigma^2)$ applied to the per-step displacement in Eq. 6. This yields a per-step position standard deviation of $\sigma \|v_i(t)\|\Delta t$ (in nautical miles), providing a direct mapping from σ to spatial error. The uncertainty represents real-world deviations caused by surveillance errors, wind, or sensor noise. Although penalty terms in Eq. 3–4 regulate probabilistic safety violations, observation noise can still distort agents' perception of the environment. We vary σ to control uncertainty levels in experiments.

The problem is a multiobjective, non-linear, and non-convex program with chance constraints [33], rendering it computationally intractable and motivating our use of multi-agent reinforcement learning for safe and cooperative multi-agent trajectory planning.

4 METHODOLOGY AND APPROACH

4.1 Constrained Dec-MDP

We reformulate the optimization problem as a Constrained Decentralized Markov Decision Process (Constrained Dec-MDP), extending the Dec-MDP framework [3] to handle sequential decision-making under chance-constrained safety requirements. Each aircraft is modeled as an independent agent with local observations and state transitions, while all agents collectively pursue a shared objective subject to safety constraints.

A Constrained Dec-MDP for n agents is defined by the tuple (S, A, R, P, C), where $S = S_1 \times \cdots \times S_n$ and $A = A_1 \times \cdots \times A_n$ are the joint state and action spaces, R is the joint reward, $P = (P_1, \ldots, P_n)$ the transition functions, and $C = (C_1, \ldots, C_n)$ the agent-specific safety costs that capture violations such as the loss of separation from other agents. The joint transition probability is given by:

$$P(s_{t+1}^1, \ldots, s_{t+1}^n \mid s_t^1, \ldots, s_t^n, a_t^1, \ldots, a_t^n) = \prod_{i=1}^{n} P_i(s_{t+1}^i \mid s_t^i, a_t^i) \tag{8}$$

We adopt the standard transition factorization $\prod_i P_i(s_{t+1}^i \mid s_t^i, a_t^i)$ under CTDE [26]. Inter-agent coupling is captured via joint reward R, shared weather in observations, and safety costs C, which we find empirically sufficient for coordination in this task (see Sec. 5.1 for scalability and Sec. 5.3 for robustness under uncertainty).

4.1.1 State Space. For each aircraft i at time t, the state includes position $p_i(t) = [x_i(t), y_i(t)]$, velocity $v_i(t) = [v_{x_i}(t), v_{y_i}(t)]$, and thunderstorm cell information $O(t)$. The state is represented as:

$$s_t^i = (p_i(t), v_i(t), O(t)) \tag{9}$$

4.1.2 Action Space. The action for each aircraft is the heading change $\Delta h_i(t)$, constrained to $[-30^\circ, 30^\circ]$ per time step:

$$a_t^i = \Delta h_i(t) \tag{10}$$

4.1.3 Reward Function. The reward is derived from the task objectives and safety constraints, combining multiple components into a unified formulation:

$$r_i(t) = \begin{cases} r_i^{\text{dist}}(t), & \text{distance penalty (Eq. 1)} \\ r_i^{\text{heading}}(t), & \text{heading change penalty (Eq. 2)} \\ r_i^{\text{sepa}}(t), & \text{if separation violation occurs (Eq. 3)} \\ r_i^{\text{cell}}(t), & \text{if storm-cell violation occurs (Eq. 4)} \\ r_i^{\text{exit}}(t), & \text{if } p_i(t) \in P_{\text{exit}} \text{ (Eq. 5)} \end{cases} \tag{11}$$

The distance term penalizes longer reroutes relative to the shortest path and the heading term penalizes larger course changes, both normalized to $[0, 1]$. Separation, storm-cell violation, and exit rewards are binary, taking value 1 when a violation or goal reach occurs and 0 otherwise.

The total reward for agent i is calculated:

$$\begin{aligned} r_{i(t)}^{\text{total}} = \omega_{\text{dist}} r_{i(t)}^{\text{dist}} + \omega_{\text{heading}} r_{i(t)}^{\text{heading}} + \omega_{\text{sepa}} r_{i(t)}^{\text{sepa}} \\ + \omega_{\text{cell}} r_{i(t)}^{\text{cell}} + \omega_{\text{exit}} r_{i(t)}^{\text{exit}} \end{aligned} \tag{12}$$

The weight vector is:

$$(\omega_{\text{dist}}, \omega_{\text{heading}}, \omega_{\text{sepa}}, \omega_{\text{cell}}, \omega_{\text{exit}}) = (-0.5, -1, -10, -8, 10) \tag{13}$$

Negative weights penalize long reroutes, frequent heading changes, loss of separation, and storm incursions, while the positive waypoint reward encourages task completion. The weights were empirically tuned to prioritize safety first, with efficiency as secondary.

4.2 IDDPG with Constrained Policy Optimization

We employ an Independent Deep Deterministic Policy Gradient (IDDPG) algorithm combined with Constrained Policy Optimization (CPO) to solve the Constrained Dec-MDP.

4.2.1 IDDPG Algorithm Structure. The IDDPG method enables each aircraft to learn an optimal policy using a shared actor-critic architecture within a Centralized Training and Decentralized Execution (CTDE) setup [26]. During training, a centralized critic network $Q(s_t, a_t \mid \theta^Q)$ evaluates the quality of actions of all agents by leveraging global state information s_t, which includes states of all aircraft. This allows the critic to learn comprehensive value functions, facilitating effective multi-agent coordination. In contrast, each agent i maintains a decentralized actor network $\mu_i(s_t^i \mid \theta^{\mu_i})$, which maps its local state s_t^i to a continuous action a_t^i. During execution, the actor networks operate independently, relying solely on local observations to make decisions. The objective of IDDPG for agent i is to optimize the actor and critic networks to maximize the expected cumulative reward, denoted as:

$$\max \mathbb{E}\left[\sum_{t=0}^{T} \gamma r_{i(t)}\right] \tag{14}$$

where γ is the discount factor. The policy gradient and the critic loss function are calculated as follows:

$$\nabla_{\theta^{\mu_i}} J(\mu_i) = \mathbb{E}_D\left[\nabla_{a_t^i} Q(s_t, a_t \mid \theta^Q) \nabla_{\theta^{\mu_i}} \mu_i(s_t^i \mid \theta^{\mu_i})\right] \tag{15}$$

$$\begin{aligned} L(\theta^Q) = \mathbb{E}_D\Big[\Big(r_{i(t)} + \gamma Q'(s_{t+1}, \mu_i'(s_{t+1}^i \mid \theta^{\mu_i'}) \mid \theta^{Q'}) \\ - Q(s_t, a_t \mid \theta^Q)\Big)^2 \Big] \end{aligned} \tag{16}$$

where $\mathbb{E}_D$ is the expectation over the state action pairs sampled from the experience replay buffer D. Here, μ_i' and Q' represent the target actor and the critic networks, respectively.

4.2.2 Constrained Policy Optimization (CPO). To ensure the learned policies respect chance constraints, the CPO introduces a constraint term into the policy optimization process [1]. The constrained problem is formulated as follows.

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \gamma r_{i(t)}\right] \quad \text{subject to} \quad \mathbb{E}\left[C_{i(t)}\right] < \epsilon \tag{17}$$

where $C_{i(t)}$ represents the expected cumulative cost of safety violations defined in Eq. 3 and Eq. 4. To incorporate the CPO problem into policy learning, we adopt a Lagrangian-based formulation:

$$L(\pi, \lambda) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma r_{i(t)}\right] - \lambda\left(\mathbb{E}\left[C_{i(t)}\right] - \epsilon\right) \tag{18}$$

Here, $\lambda \geq 0$ is a Kühn–Tucker multiplier, which governs the trade-off between reward maximization and constraint satisfaction, corresponding to the inequality constraints on safety cost defined in Eq. 3 and Eq. 4. The non-negativity constraint on λ ensures that the constraint penalty is only applied when the expected safety cost exceeds the predefined threshold ϵ, in line with the Karush–Kühn–Tucker (KKT) conditions. The policy gradient update is therefore augmented with the constraint term to ensure feasible and safe policy improvement, defined as follows:

$$\begin{aligned} \nabla_{\theta^{\mu_i}} J(\mu_i) = \mathbb{E}_D\Big[\nabla_{a_t^i} Q(s_t, a_t \mid \theta^Q) \nabla_{\theta^{\mu_i}} \mu_i(s_t^i \mid \theta^{\mu_i}) \\ - \lambda \nabla_{\theta^{\mu_i}} C_{i(t)}\Big] \end{aligned} \tag{19}$$

4.3 MAF-Net with Action Mask

To improve IDDPG in large continuous action spaces, a Multi-head Action Filter Network (MAF-Net) is proposed with a time-dependent action mask (Fig. 1). The mask partitions the action space A_t^{mask} into unsafe (A_t^{unsafe}), desired (A_t^{desired}), and undesired ($A_t^{\text{undesired}}$) actions based on real-time observations. MAF-Net processes these via three heads: Head-1 projects unsafe actions to safe alternatives (a_t^{safe}), Head-2 preserves desired actions (a_t^{desired}), and Head-3 penalizes undesired actions. The filtered action a_t^{MAF} is stored in the replay buffer $(s_t, a_t^{\text{MAF}}, r_t, s_{t+1})$, so only safe and efficient actions guide learning. As a differentiable module, MAF-Net preserves gradient flow, improving training efficiency and stability.

4.3.1 Time-Dependent Action Mask. The time-dependent action mask operates on the candidate *absolute* heading, computed from the policy's heading change as $\alpha_i^{\text{cand}}(t) = (\alpha_i(t) + \Delta h_i(t)) \bmod 360^\circ$, rather than on $\Delta h_i(t)$ itself. We normalize $\alpha_i^{\text{cand}}(t)$ to $[-1, 1]$ and partition this range into 12 equal bins of width $1/6$ (30° each), aligned with 12 sensor probes with a 30 NM detection radius. At

each time step t, actions are classified as: (1) Unsafe (A_t^{unsafe}): leading to loss of separation with aircraft or storms, value 0. (2) Desired (A_t^{desired}): guiding toward the exit within a 30° buffer, value 1. (3) Undesired ($A_t^{\text{undesired}}$): safe but outside the buffer, value 0.5. The mask A_t^{mask} encodes these values as labels to pre-train MAF-Net, guiding the network to classify actions and learn appropriate responses.

$$A_t^{\text{mask}}(a_t) = \begin{cases} 0, & a_t \in A_t^{\text{unsafe}} \\ 1, & a_t \in A_t^{\text{desired}} \\ 0.5, & a_t \in A_t^{\text{undesired}} \end{cases} \tag{20}$$

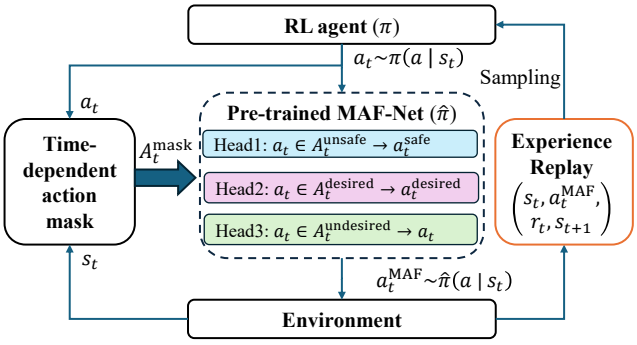

Figure 1: Framework of pre-trained MAF-Net. Head-1 projects unsafe actions to safe ones, Head-2 preserves desired actions, and Head-3 penalizes undesired actions, filtering actions to improve safety and efficiency.

4.3.2 MAF-Net Framework. MAF-Net is a pre-trained neural network that filters unsafe and suboptimal actions using the time-dependent mask A_t^{mask}. Pre-training decouples action filtering from policy optimization, a strategy often adopted in complex models [52] to improve stability and reduce computational overhead. During pre-training, random actions are fed into MAF-Net and mapped to filtered outputs that satisfy safety and efficiency constraints [50].

(1) Head-1: Projects unsafe actions $a_t \in A_t^{\text{unsafe}}$ to the nearest safe alternative a_t^{safe} via a differentiable projection:

$$a_t^{\text{safe}} = \text{Proj}_{A_t^{\text{unsafe}}}(a_t) = \arg\min_{a \in A_t^{\text{safe}}} \left(1 - \frac{1}{e^{\|a - a_t\|}}\right), \tag{21}$$

with loss

$$L_{\text{Head-1}} = \mathbb{E}\left[1 - \frac{1}{e^{\|a_t^{\text{safe}} - a_t\|}}\right]. \tag{22}$$

(2) Head-2: Preserves desired actions through identity mapping, minimizing Mean Squared Error (MSE) with target a_t^{desired}:

$$L_{\text{Head-2}} = \mathbb{E}\left[\|a_t^{\text{desired}} - a_t\|^2\right] \tag{23}$$

(3) Head-3: Penalizes undesired actions $a_t \in A_t^{\text{undesired}}$ to discourage selection, with penalty

$$R_{\text{penalty}}(a_t) = \zeta \|a - a_t\|^2, \quad a \in A_t^{\text{undesired}} \tag{24}$$

and loss

$$L_{\text{Head-3}} = \mathbb{E}\left[\zeta \|a_t^{\text{desired}} - a_t\|^2\right]. \tag{25}$$

where ζ is a penalty coefficient, set to 1 in this work to balance safety compliance and learning exploration.

4.4 Combined Learning Policy

We integrate MAF-Net with IDDPG into a combined policy framework IDDPG-MAF to enable safe and efficient learning. Although Constrained Policy Optimization enforces safety constraints during training, it does not guarantee safe actions during execution. MAF-Net addresses this gap by filtering unsafe actions from the IDDPG policy outputs based on time-dependent observations.

The IDDPG policy $a_i(t) = \mu_i(s_i(t))$ produces a raw action, which is passed through MAF-Net along with a time-dependent action mask A_t^{mask} to produce a filtered action:

$$a_i^{\text{MAF}}(t) = \text{MAF}_\varrho(a_i(t), s_i(t), A_t^{\text{mask}}) \tag{26}$$

This filtered action $a_i^{\text{MAF}}(t)$, rather than the raw action, is executed in the environment and stored in the replay buffer as part of the experience tuple $(s_i(t), a_i^{\text{MAF}}(t), r_i(t), s_i(t+1))$. During training, the actor parameters θ^{μ_i} are updated using the constrained policy gradient defined in Eq. 19. During learning, MAF-Net filters raw actions while preserving differentiability, allowing gradients from the critic loss to propagate through the filtered action back to the actor network. The final actor update is:

$$\theta_{\text{new}}^{\mu_i} = \theta^{\mu_i} + \alpha \left(\nabla_{\theta^{\mu_i}} J(\mu_i) - \lambda \nabla_{\theta^{\mu_i}} C_i(t)\right) \tag{27}$$

where θ^{μ_i} and $\theta_{\text{new}}^{\mu_i}$ represent the current and updated parameters of the actor network for agent i, respectively; α is the learning rate; $\nabla_{\theta^{\mu_i}} J(\mu_i)$ is the gradient of the expected cumulative reward with respect to the actor parameters; λ is a non-negative multiplier that scales the penalty term; and $\nabla_{\theta^{\mu_i}} C_i(t)$ is the gradient of the expected constraint cost.

Algorithm 1 summarizes the integrated IDDPG-MAF training. At each timestep, the IDDPG policy μ_{θ_i} outputs a raw action $a_i(t)$. A time-dependent mask A_t^{mask}, computed from real-time observations, together with $a_i(t)$, is passed through MAF-Net to obtain the filtered action $a_i^{\text{MAF}}(t)$. Gaussian noise η is then added to $a_i^{\text{MAF}}(t)$ to encourage policy exploration, and the resulting action is executed, producing transition $(s_i(t+1), r_i(t))$. Each tuple $(s_i(t), a_i^{\text{MAF}}(t), r_i(t), s_i(t+1))$ is stored in the replay buffer D. Policy parameters θ_i are updated from mini-batches sampled from D using Eq. 27, which balances reward maximization and constraint satisfaction. MAF-Net remains fixed during training but continuously shapes executed actions, enabling safe exploration and convergence to effective, risk-aware policies in dynamic multi-agent settings.

5 EXPERIMENTAL RESULTS

We evaluate the proposed IDDPG-MAF methods on the safety-critical task of multi-aircraft trajectory planning under dynamic thunderstorms and position uncertainty, with comparisons against the baseline methods FMT and DDPG for safe trajectory planning.

- **Fast Marching Tree (FMT)**: A state-of-the-art cooperative trajectory optimization algorithm that generates collision-free paths by incrementally expanding a search tree in space-time for aircraft rerouting in thunderstorms [16].
- **Deep Deterministic Policy Gradient (DDPG)**: A widely adopted model-free RL algorithm for continuous control. It has been applied to aircraft conflict resolution tasks in deterministic settings without weather dynamics [36]. We

Algorithm 1 Training Setup of IDDPG-MAF

Input: Agent batch size B, replay buffer size N, update frequency C, max. timesteps per episode T, number of agents n, pre-trained MAF_ϱ, policy params $\{\theta_i\}_{i=1}^n$, target policy params $\{\theta_i'\}_{i=1}^n$, critic params ϕ, target critic params ϕ', soft-update factor τ, learning rate α, number of training episodes M.
Output: Trained actor parameters $\{\theta_i\}_{i=1}^n$
Initialise replay buffer as empty, $D \leftarrow \emptyset$ (capacity N)
Initialise critic ϕ, target critic ϕ'
Initialise policies θ_i and target policies θ_i', $i = 1, \ldots, n$
for episode = 1 to M **do**
  Obtain initial states $s_i(0)$ for all agents i
  for $t = 1$ to T **do**
    for each agent $i = 1$ to n **do**
      Select action: $a_i(t) = \mu_{\theta_i}(s_i(t))$
      Compute time-dependent mask: $A_t^{\text{mask}} = \text{Mask}(s_i(t))$
      Filter action: $a_i^{\text{MAF}}(t) = \text{MAF}_\varrho(a_i(t), s_i(t), A_t^{\text{mask}})$
      Add exploration noise: $a_i^{\text{MAF}}(t) = a_i^{\text{MAF}}(t) + \eta, \quad \eta \sim \mathcal{N}(0, \epsilon^2)$
      Execute actions $a_i^{\text{MAF}}(t)$ to obtain $s_i(t+1)$ and $r_i(t)$
      Store transition $(s_i(t), a_i^{\text{MAF}}(t), r_i(t), s_i(t+1))$ into reply buffer D
      $s_i(t) \leftarrow s_i(t+1)$
      Sample minibatch of size B from D
      Update each policy θ_i via loss (Eq. (15) and Eq. (27))
      Update critic ϕ via loss (Eq. (16))
      if $t \bmod C = 0$ **then**
        Soft-update targets:
        $\theta_i' \leftarrow \tau\,\theta_i + (1-\tau)\,\theta_i'$
        $\phi' \leftarrow \tau\,\phi + (1-\tau)\,\phi'$
      end if
    end for
  end for
end for

adapted its structure to address problems with dynamic thunderstorms for a fair comparison.

- **Independent DDPG (IDDPG)**: A decentralized multi-agent version of DDPG that we propose in this research to improve scalability and trajectory coordination.
- **IDDPG with MAF-Net (IDDPG-MAF)**: An enhanced version of IDDPG integrated with the MAF-Net for real-time action filtering to improve safety and training efficiency.

All methods were trained and evaluated using a custom-built simulator environment, where a 200 × 200 square nautical mile (NM^2) en route airspace was created (illustrated in Figure 2). Multiple aircraft flew at a speed of 400 knots with a minimum separation of 5 NM. Aircraft entered at randomized time intervals and aimed to reach exit waypoints while avoiding dynamic thunderstorm cells and other aircraft. The thunderstorm cells moved at speeds of 50–90 knots, with radii of 10–25 NM, updated every five time steps. Each time step lasted 12 seconds, and simulations were capped at 150 steps (i.e., 30 minutes). Performance metrics included (1) Aircraft loss of separation (LOS) rate, (2) Thunderstorm LOS rate, (3) Goal reach rate, and (4) distance ratio (actual path / nominal straight-line). The hyperparameters used in IDDPG-MAF are provided in Table 1.

5.1 Effectiveness and Scalability of IDDPG-MAF

The experimental results, summarized in Table 2, demonstrate the comparative performance of the four methods, and an illustration of multi-aircraft rerouting thunderstorms and other aircraft is provided in Figure 2. The results show that the proposed IDDPG method consistently outperforms the state-of-the-art FMT and DDPG models in terms of safety and goal reach rate. For aircraft densities of 6 and 8, IDDPG achieves a 0% aircraft LOS rate, compared to 12% and 17% for FMT, and 2% and 1% for DDPG. Similarly, IDDPG reduces thunderstorm LOS rates to 2% and 5%, significantly outperforming DDPG's 8% and 15%, respectively. These results emphasize the scalability and robustness of IDDPG in multi-agent cooperative decision-making, particularly under increased traffic density and adverse weather conditions.

Table 1: Hyperparameters used in IDDPG-MAF.

Parameter	Value
Batch size (B)	512
Soft update rate (τ)	0.01
Target update frequency (C)	1
Learning rate (α)	0.0001
Discount factor (γ)	0.95
Replay buffer size (N)	100000
Number of training episodes (M)	20000
Max time steps per episode (T)	150
Exploration noise (ϵ)	$1 \rightarrow 0.03$

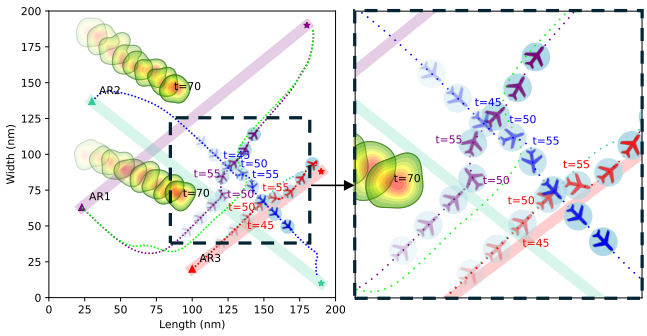

Figure 2: Illustration of multiple aircraft rerouting around dynamic thunderstorm cells. Entry and exit waypoints are marked by solid triangles and stars, respectively, which connects by air route (AR). Thunderstorm cells are depicted as evolving contours. The dashed lines represent the rerouted trajectories, color-coded for individual aircraft agents. The inset zooms into a critical area of the airspace, demonstrating how conflicts were resolved by the proposed IDDPG-MAF.

The integration of MAF-Net into IDDPG results in significant safety improvements, achieving a 0% aircraft LOS rate and reducing thunderstorm LOS rates to 1% or less across all tested aircraft

density scenarios. These results highlight MAF-Net's effectiveness in filtering unsafe actions, ensuring adherence to safety constraints. In addition to safety, IDDPG-MAF also achieves the highest goal reach rates across all scenarios, with 100% at a density of 4 and maintaining 99% at densities of 6 and 8. These findings show that the proposed IDDPG outperforms FMT and DDPG in scalability, robustness, and trajectory optimization. The integration of MAF-Net further reduces separation violations, demonstrating its potential for safety-critical multi-robot conflict resolution tasks.

Our approach scales to 8 aircraft under dynamic thunderstorms, significantly outperforming the baseline [16] at 4 aircraft. In current operations, typical density per sector per flight level is below 3, whereas multi-sector operations [12] require handling a higher number of aircraft. To assess scalability, we evaluated larger airspace structures (Table 3): across 100 tested scenarios, the method achieved 100% goal reach in smaller sectors (200×200–300×300 NM^2, up to 12 aircraft) and 99% in larger airspace (600×600 NM^2, 45 aircraft), effectively doubling the maximum capacity of current practice and supporting future higher-demand concepts of operation.

Table 2: The proposed IDDPG-MAF outperforms baseline algorithms in 100 runs under varying number of agents. The rates of loss of separation (LOS) are represented as a combined percentage (Aircraft/Thunderstorm).

No. of Aircraft	Methods	LOS Rate	Goal Reach Rate
4	FMT	6% / 1%	93%
	DDPG	0% / 3%	97%
	IDDPG	0% / 1%	99%
	IDDPG-MAF	**0% / 0%**	**100%**
6	FMT	12% / 1%	87%
	DDPG	2% / 8%	90%
	IDDPG	0% / 2%	98%
	IDDPG-MAF	**0% / 1%**	**99%**
8	FMT	17% / 1%	82%
	DDPG	1% / 15%	84%
	IDDPG	0% / 5%	95%
	IDDPG-MAF	**0% / 1%**	**99%**

5.2 Enhanced Performance of IDDPG-MAF

Figure 3 (a) and (b) present the benefits of integrating MAF-Net with IDDPG for improved learning efficiency and safety. In Figure 3(a), IDDPG-MAF achieves faster convergence and an early reward leap compared to IDDPG and DDPG, driven by MAF-Net's ability to filter unsafe actions and promote goal-oriented decisions early in the training process. The reduced variance in IDDPG-MAF's reward curve reflects increased stability by eliminating irrelevant actions. While IDDPG outperforms DDPG, IDDPG-MAF slightly surpasses IDDPG, demonstrating its effectiveness in policy optimization. In Figure 3(b), IDDPG-MAF maintains a total loss of separation (LOS) rate below the acceptable risk threshold ($\epsilon = 0.01$ [28]) after 3,000 episodes and maintains this level throughout training. In contrast, IDDPG reaches similar safety levels at 4,000 episodes but intermittently exceeds the threshold after 14,000 episodes. DDPG lags significantly in achieving safety compliance, reflecting its limitations in satisfying constraints. Overall, IDDPG-MAF enhances learning efficiency and stability, while ensuring consistent safety compliance, making it a effective method for dynamic and safety-critical multi-agent systems.

Table 3: Scalability over 100 test runs in larger pre-defined airspace structure with increasing number of agents.

Airspace (NM^2)	No. of Aircraft	Goal Reach Rate
200×200	5	99%
300×300	12	100%
400×400	20	99%
500×500	30	99%
600×600	45	99%

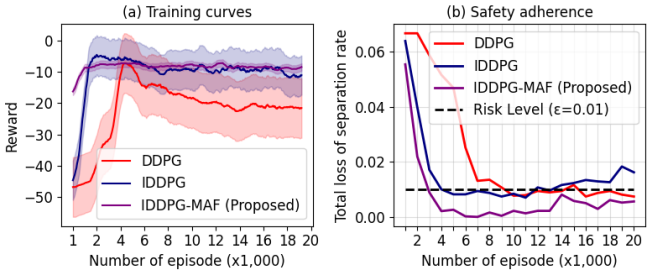

Figure 3: Training performance and safety adherence. (a) Training curves illustrate the episodic reward over time for DDPG, IDDPG, and the proposed IDDPG-MAF. IDDPG-MAF achieves faster convergence and reduced variance compared to baselines, indicating improved learning efficiency and stability. (b) Safety adherence is evaluated by the total loss of separation rate, where IDDPG-MAF consistently stays below the risk threshold ($\epsilon = 0.01$).

5.3 Robustness under Uncertainty

We assess robustness by varying the noise standard deviation σ and analyzing its effect on safety, efficiency, and trajectory stability. Table 4 reports model performance under different σ. As σ increases from 0.0 to 1.0, aircraft and storm loss-of-separation (LOS) rates rise, while goal reach drops from 99% to 75%. This trend reflects growing positional drift that limits agents' ability to maintain safe separation and reach destinations. The model remains robust under moderate noise: at $\sigma = 0.5$, goal reach remains 95.5% with only 1.5% aircraft LOS, confirming that probabilistic penalties in Eq. 3–4 effectively regulate safety. Beyond $\sigma > 0.6$, storm LOS increases sharply, reaching 17.5% at $\sigma = 1.0$, indicating that excessive uncertainty overwhelms avoidance capability. The distance ratio also increases with σ, showing that aircraft adopt longer reroutes under high uncertainty scenarios.

5.4 Ablation Analysis of MAF-Net

This ablation evaluates the contribution of each MAF-Net head to the safety, efficiency, and learning performance of IDDPG. Five models are trained under identical CTDE settings to isolate the effects of different action-filtering mechanisms.

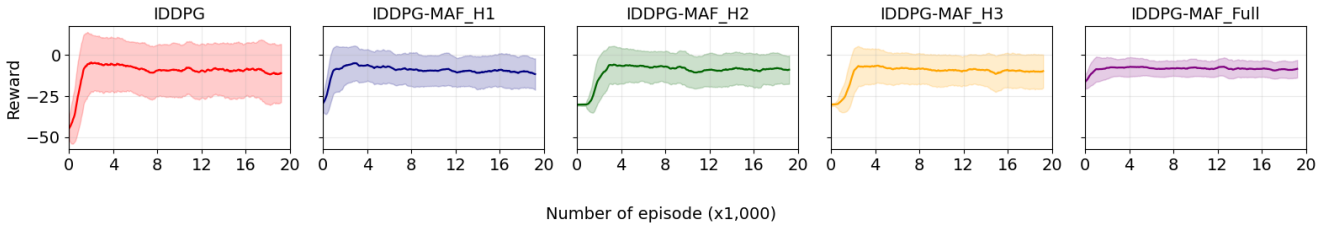

Figure 4: Reward learning curves for IDDPG and IDDPG-MAF variants. Masked variants (MAF_H1/H2/H3) converge faster and show lower variance than IDDPG, while MAF_Full is the most stable and later achieves the best safety–efficiency trade-off.

Table 4: Performance under different position uncertainty levels σ.

Uncertainty level σ	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Aircraft LOS rate (%)	1.0	1.5	2.5	2.0	3.0	1.5	1.0	3.0	3.0	5.5	7.5
Storm LOS rate (%)	0.0	0.5	0.5	1.5	2.0	3.0	5.0	6.0	4.5	10.5	17.5
Goal reach rate (%)	99.0	98.0	97.0	96.5	95.0	95.5	94.0	91.0	92.5	84.0	75.0
Distance ratio (mean±SD)	1.33±0.32	1.34±0.33	1.33±0.33	1.35±0.33	1.36±0.35	1.40±0.33	1.47±0.35	1.47±0.39	1.59±0.40	1.55±0.46	1.64±0.47

Fig. 4 presents learning curves for five variants under identical CTDE settings. Relative to IDDPG, IDDPG–MAF_H1 converges faster (≈1k episodes) with lower variance, as Head 1 accelerates learning by projecting unsafe actions to feasible a_t^{safe} and thereby avoids unproductive exploration. By contrast, IDDPG–MAF_H2 and IDDPG–MAF_H3 converge more slowly (≈4k episodes) with higher dispersion: without projection, identity preservation alone (H2) provides limited corrective signal early in training, while penalization without projection or preservation (H3) can induce conservative detours. The full model (IDDPG–MAF_Full) is the most stable and achieves the best return. Consistent with the curves, evaluation indicates that MAF_Full provides the best safety–efficiency trade-off (aircraft LOS 0, storm LOS ≈ 1%, goal reach 99%, distance ratio 1.08 ± 0.13), followed by MAF_H1 and MAF_H2; MAF_H3 underperforms on safety (storm LOS 8%). These results indicate that single-head variants are not intended as standalone solutions; rather, they demonstrate how safety projection (H1), preservation (H2), and penalization (H3) contribute to safe exploration and convergence when combined in MAF_Full. The ablation further reveals sensitivity to mask specification and head composition. Future work could explore adaptive head weighting and learned mask thresholds for improved policy robustness.

6 CONCLUSIONS

This work proposes IDDPG-MAF, a constrained policy learning framework that integrates Independent Deep Deterministic Policy Gradient (IDDPG) with a pre-trained Multi-Head Action Filter Network (MAF-Net) for decentralized multi-agent trajectory planning under uncertainty. The framework is adapted to a complex multi-aircraft trajectory planning task, where multiple high-speed aircraft agents must coordinate in real time, adapt to dynamic thunderstorm movements, and maintain safe separation under uncertainties. Compared with baseline methods, IDDPG-MAF achieves faster convergence, lower reward variance, and consistently satisfies safety constraints across various multi-aircraft planning test cases. More broadly, this work contributes to constrained multi-agent reinforcement learning by demonstrating how pre-trained safety-aware modules can potentially be embedded into gradient-based policy learning for safety-critical and real-time decision-making tasks.

The proposed IDDPG-MAF is an early simulation-based step toward MARL with an integrated safety filter for multi-aircraft rerouting under fast-evolving convective weather. Operationally, it could provide an added safety layer by resolving conflicts beyond 5 nautical miles, reducing the likelihood of triggering the ACAS X system that handles near mir-air collisions in 500 ft ranges [22]. It could also serve as AI decision support for controllers in remote regions and for pilots during coordinated storm avoidance.

Empirically, IDDPG-MAF converges reliably across the tested scenarios, but shared-network training may become harder to stabilize in more complex environments [7]. We also assume storm tracks are observed without uncertainty; incorporating probabilistic storm forecasts in decision making is an important next step. In addition, MAF-Net currently relies on a rule-based action mask; replacing it with more adaptive masking, for example using human feedback [11] or large language models (LLMs) [52], may improve generalization. Beyond aviation, extending IDDPG-MAF to cooperative multi-robot collision avoidance and navigation, such as autonomous driving [42] and uncrewed aerial vehicles, can further test its generality.

ACKNOWLEDGMENTS

This research is supported by the Italian Ministry of Foreign Affairs and International Cooperation (MAECI) and the Agency for Science, Technology and Research (A*STAR), Singapore, under (R23I0IR034) the First Executive Programme of Scientific and Technological Cooperation between Italy and Singapore for the years 2023–2025. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Italian Ministry of Foreign Affairs and International Cooperation or the Agency for Science, Technology and Research (A*STAR), Singapore.

REFERENCES

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained Policy Optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 22–31.

[2] Eseoghene Ben-Iwhiwhu, Saptarshi Nath, Praveen K. Pilly, Soheil Kolouri, and Andrea Soltoggio. 2022. Lifelong Reinforcement Learning with Modulating Masks. *arXiv preprint* arXiv:2212.11110 (2022).

[3] Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The Complexity of Decentralized Control of Markov Decision Processes. *Mathematics of Operations Research* 27, 4 (2002), 819–840.

[4] Marc Brittain and Peng Wei. 2019. Autonomous Air Traffic Controller: A Deep Multi-Agent Reinforcement Learning Approach. In *Proceedings of the International Conference on Machine Learning (ICML), Reinforcement Learning for Real Life Workshop*.

[5] Marc W. Brittain and Peng Wei. 2018. Towards Autonomous Air Traffic Control for Sequencing and Separation: A Deep Reinforcement Learning Approach. In *Proceedings of the 2018 Aviation Technology, Integration, and Operations Conference*. 3664.

[6] Zelei Cheng, Xian Wu, Jiahao Yu, Wenhai Sun, Wenbo Guo, and Xinyu Xing. 2023. Statemask: Explaining Deep Reinforcement Learning through State Mask. *Advances in Neural Information Processing Systems (NeurIPS)* 36 (2023), 62457–62487.

[7] Filippos Christianos, Lukas Schäfer, and Stefano Albrecht. 2020. Shared Experience Actor-Critic for Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), 10707–10717.

[8] Jonathan Diller, Qi Han, Robert Byers, James Dotterweich, and James Humann. 2025. Hitchhiker's Guide to Patrolling: Path-Finding for Energy-Sharing Drone-UGV Teams. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 611–619.

[9] Heinz Erzberger. 2005. Automated Conflict Resolution for Air Traffic Control. In *25th International Congress of the Aeronautical Sciences (ICAS 2006)*. 1–27.

[10] Heinz Erzberger, Tasos Nikoleris, Russell A. Paielli, and Yung-Cheng Chu. 2016. Algorithms for control of arrival and departure traffic in terminal airspace. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering* 230, 9 (July 2016), 1762–1779.

[11] Flint Xiaofeng Fan, Cheston Tan, Yew-Soon Ong, Roger Wattenhofer, and Wei-Tsang Ooi. 2025. FedRLHF: A Convergence-Guaranteed Federated Framework for Privacy-Preserving and Personalized RLHF. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (Detroit, MI, USA) *(AAMAS '25)*. International Foundation for Autonomous Agents and Multiagent Systems, Detroit, MI, USA, 713–721.

[12] Pierre Flener, Justin Pearson, Magnus Ågren, Carlos Garcia-Avello, Mete Celiktin, and Søren Dissing. 2007. Air-Traffic Complexity Resolution in Multi-Sector Planning. *Journal of Air Transport Management* 13, 6 (2007), 323–328.

[13] Daniel González-Arribas, Manuel Soler, and Manuel Sanjurjo-Rivo. 2018. Robust Aircraft Trajectory Planning Under Wind Uncertainty Using Optimal Control. *Journal of Guidance, Control, and Dynamics* 41, 3 (2018), 673–688.

[14] Alexander W. Goodall and Francesco Belardinelli. 2023. Approximate Model-Based Shielding for Safe Reinforcement Learning. In *Proceedings of the European Conference on Artificial Intelligence (ECAI 2023)*. IOS Press, 883–890.

[15] D. J. Groot, Joost Ellerbroek, and Jacco M. Hoekstra. 2024. Analysis of the Impact of Traffic Density on Training of Reinforcement Learning Based Conflict Resolution Methods for Drones. *Engineering Applications of Artificial Intelligence* 133 (2024), 108066.

[16] Andréas Guitart, Daniel Delahaye, Félix Mora Camino, and Eric Feron. 2023. Collaborative Generation of Local Conflict-Free Trajectories With Weather Hazards Avoidance. *IEEE Transactions on Intelligent Transportation Systems* 24, 11 (November 2023), 12831–12842.

[17] Yash Guleria, Duc-Thinh Pham, Sameer Alam, Phu N. Tran, and Nicolas Durand. 2024. Towards Conformal Automation in Air Traffic Control: Learning Conflict Resolution Strategies through Behavior Cloning. *Advanced Engineering Informatics* 59 (2024), 102273.

[18] Chuanbo Hua, Federico Berto, Jiwoo Son, Seunghyun Kang, Changhyun Kwon, and Jinkyoo Park. 2025. CAMP: Collaborative Attention Model with Profiles for Vehicle Routing Problems. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 1015–1024.

[19] ICAO. 2003. *Appendix G to Part 91—Operations in Reduced Vertical Separation Minimum (RVSM) Airspace.* Regulation Report. The International Civil Aviation Organization.

[20] Paveen Juntama, Supatcha Chaimatanan, Sameer Alam, and Daniel Delahaye. 2020. A Distributed Metaheuristic Approach for Complexity Reduction in Air Traffic for Strategic 4D Trajectory Optimization. In *Proceedings of the 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT)*. IEEE, 1–9.

[21] Maryam Kamgarpour, Vera Dadok, and Claire Tomlin. 2010. Trajectory Generation for Aircraft Subject to Dynamic Weather Uncertainty. In *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 2063–2068.

[22] Sydney M Katz, Robert J Moss, Dylan M Asmar, Wesley A Olson, James K Kuchar, and Mykel J Kochenderfer. 2025. Aircraft Collision Avoidance Systems: Technological Challenges and Solutions on the Path to Regulatory Acceptance. *arXiv preprint arXiv:2510.20916* (2025).

[23] Sunbowen Lee, Hongqin Lyu, Yicheng Gong, Yingying Sun, and Chao Deng. 2025. MacLight: Multi-scene Aggregation Convolutional Learning for Traffic Signal Control. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 1263–1271.

[24] Yongshuai Liu, Avishai Halev, and Xin Liu. 2021. Policy Learning with Constraints in Model-Free Reinforcement Learning: A Survey. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*.

[25] Zhihao Liu, Xianliang Yang, Zichuan Liu, Yifan Xia, Wei Jiang, Yuanyu Zhang, Lijuan Li, Guoliang Fan, Lei Song, and Bian Jiang. 2024. Knowing What Not to Do: Leverage Language Model Insights for Action Space Pruning in Multi-Agent Reinforcement Learning. *arXiv preprint* arXiv:2405.16854 (2024).

[26] Ryan Lowe, Yi I. Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

[27] Amir Maliah, Dor Atzmon, and Ariel Felner. 2025. Minimizing Makespan with Conflict-Based Search for Optimal Multi-Agent Path Finding. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 1418–1426.

[28] Robert J. Moss, Arec Jamgochian, Johannes Fischer, Anthony Corso, and Mykel J. Kochenderfer. 2024. ConstrainedZero: Chance-Constrained POMDP Planning using Learned Probabilistic Failure Surrogates and Adaptive Safety Constraints. *arXiv preprint* arXiv:2405.00644 (2024).

[29] Hok Kwan Ng, Shon Grabbe, and Avijit Mukherjee. 2009. Design and evaluation of a dynamic programming flight routing algorithm using the convective weather avoidance model. In *AIAA Guidance, Navigation, and Control Conference and Exhibit*. 5862.

[30] Nhat Nguyen, Duong Nguyen, Gianluca Rizzo, and Hung Nguyen. 2024. United We Stand: Decentralized Multi-Agent Planning With Attrition. In *Proceedings of the European Conference on Artificial Intelligence (ECAI 2024)*. IOS Press, 3421–3428.

[31] Bizhao Pang, Xinting Hu, Mingcheng Zhang, Sameer Alam, and Guglielmo Lulli. 2025. Decentralized Deep Reinforcement Learning for Cooperative Multi-Agent Flight Trajectory Planning in Adverse Weather. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 2705–2707.

[32] Bizhao Pang, Xinting Hu, Mingcheng Zhang, Sameer Alam, and Guglielmo Lulli. 2025. A multi-aircraft co-operative trajectory planning model under dynamic thunderstorm cells using decentralized deep reinforcement learning. *Advanced Engineering Informatics* 65 (2025), 103157.

[33] Bizhao Pang, Kin Huat Low, and Vu N. Duong. 2024. Chance-Constrained UAM Traffic Flow Optimization with Fast Disruption Recovery Under Uncertain Waypoint Occupancy Time. *Transportation Research Part C: Emerging Technologies* 161 (April 2024), 104547.

[34] Bizhao Pang, Kin Huat Low, and Chen Lv. 2022. Adaptive Conflict Resolution for Multi-UAV 4D Routes Optimization Using Stochastic Fractal Search Algorithm. *Transportation Research Part C: Emerging Technologies* 139 (2022), 103666.

[35] George Papadopoulos, Alevizos Bastas, George A. Vouros, Ian Crook, Natalia Andrienko, Gennady Andrienko, and Jose Manuel Cordero. 2024. Deep Reinforcement Learning in Service of Air Traffic Controllers to Resolve Tactical Conflicts. *Expert Systems with Applications* 236 (February 2024).

[36] Duc-Thinh Pham, Phu N. Tran, Sameer Alam, Vu Duong, and Daniel Delahaye. 2022. Deep Reinforcement Learning Based Path Stretch Vector Resolution in Dense Traffic with Uncertainties. *Transportation Research Part C: Emerging Technologies* 135 (February 2022).

[37] Gauthier Picard. 2022. Trajectory Coordination Based on Distributed Constraint Optimization Techniques in Unmanned Air Traffic Management. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 1065–1073.

[38] Pedro Santana, Sylvie Thiébaux, and Brian C. Williams. 2016. RAO*: An Algorithm for Chance-Constrained POMDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 3308–3314.

[39] Risal Shahriar Shefin, Md Asifur Rahman, Thai Le, and Sarra Alqahtani. 2025. xSRL: Safety-Aware Explainable Reinforcement Learning - Safety as a Product of Explainability. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (Detroit, MI, USA) *(AAMAS '25)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1932–1940.

[40] Roland Stolz, Hanna Krasowski, Jakob Thumm, Michael Eichelbeck, Philipp Gassert, and Matthias Althoff. 2024. Excluding the Irrelevant: Focusing Reinforcement Learning through Continuous Action Masking. *arXiv preprint* arXiv:2406.03704 (2024).

[41] Akifumi Wachi and Yanan Sui. 2020. Safe Reinforcement Learning in Constrained Markov Decision Processes. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 9797–9806.

[42] Jingda Wu, Zhiyu Huang, Wenhui Huang, and Chen Lv. 2022. Prioritized Experience-Based Reinforcement Learning with Human Guidance for Autonomous Driving. *IEEE Transactions on Neural Networks and Learning Systems* 35, 1 (2022), 855–869.

[43] Zheng Wu, Yichuan Li, Wei Zhan, Changliu Liu, Yun-Hui Liu, and Masayoshi Tomizuka. 2024. Efficient Reinforcement Learning of Task Planners for Robotic Palletization through Iterative Action Masking Learning. *arXiv preprint* arXiv:2404.04772 (2024).

[44] Chao Yan, Chang Wang, Xiaojia Xiang, Kin Huat Low, Xiangke Wang, Xin Xu, and Lincheng Shen. 2023. Collision-Avoiding Flocking with Multiple Fixed-Wing UAVs in Obstacle-Cluttered Environments: A Task-Specific Curriculum-Based MADRL Approach. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[45] Dongjie Yu, Haitong Ma, Shengbo Li, and Jianyu Chen. 2022. Reachability Constrained Reinforcement Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 25636–25655.

[46] Tao Yu, Zhizheng Zhang, Cuiling Lan, Yan Lu, and Zhibo Chen. 2022. Mask-based Latent Reconstruction for Reinforcement Learning. *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), 25117–25131.

[47] Han Zhang, Jingkai Chen, Jiaoyang Li, Brian Williams, and Sven Koenig. 2022. Multi-Agent Path Finding for Precedence-Constrained Goal Sequences. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 1464–1472.

[48] Mingcheng Zhang, Bizhao Pang, Chao Yan, Mir Feroskhan, and Chen Lv. 2025. Real-Time Avoidance of Obstacles and Emergent Geo-Fences for Urban Air Mobility Using Deep Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems* (2025).

[49] Peng Zhao and Yongming Liu. 2022. Physics Informed Deep Reinforcement Learning for Aircraft Conflict Resolution. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (July 2022), 8288–8301.

[50] Dianyu Zhong, Yiqin Yang, and Qianchuan Zhao. 2024. No Prior Mask: Eliminate Redundant Action for Deep Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 38. 17078–17086.

[51] Zihan Zhou, Jonathan Booher, Khashayar Rohanimanesh, Wei Liu, Aleksandr Petiushko, and Animesh Garg. 2024. Uniformly Safe RL with Objective Suppression for Multi-Constraint Safety-Critical Applications. *arXiv preprint* arXiv:2402.15650 (2024).

[52] Zihao Zhou, Bin Hu, Chenyang Zhao, Pu Zhang, and Bin Liu. 2024. Large Language Model as a Policy Teacher for Training Reinforcement Learning Agents. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 5671–5679.

[53] Fengming Zhu and Fangzhen Lin. 2025. Single-Agent Planning in a Multi-Agent System: A Unified Framework for Type-Based Planners. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 2382–2391.