Autonomous Conflict Resolution (AutoCR) Based on Improved Multi-agent Reinforcement Learning

HUANG Xiao, TIAN Yong, LI Jiangchen, ZHANG Naizhong*

College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P. R. China

(Received 20 June 2025; revised 30 August 2025; accepted 10 September 2025)

Abstract: Conflict resolution (CR) is a fundamental component of air traffic management, where recent progress in artificial intelligence has led to the effective application of deep reinforcement learning (DRL) techniques to enhance CR strategies. However, existing DRL models applied to CR are often limited to simple scenarios. This approach frequently leads to the neglect of the high risks associated with multiple intersections in the high-density and multi-airport system terminal area (MAS-TMA), and suffers from poor interpretability. This paper addresses the aforementioned gap by introducing an improved multi-agent DRL model that adopted to autonomous CR (AutoCR) within MAS-TMA. Specifically, dynamic weather conditions are incorporated into the state space to enhance adaptability. In the action space, the flight intent is considered and transformed into optimal maneuvers according to overload, thus improving interpretability. On these bases, the deep Q-network (DQN) algorithm is further improved to address the AutoCR problem in MAS-TMA. Simulation experiments conducted in the "Guangdong-Hong Kong-Macao" greater bay area (GBA) MAS-TMA demonstrate the effectiveness of the proposed method, successfully resolving over eight potential conflicts and performing robustly across various air traffic densities.

Key words: air traffic management; conflict resolution; multi-airport system terminal area(MAS-TMA); multi-agent reinforcement learning

CLC number: V355 **Document code:** A **Article ID:** 1005-1120(2025)S-0091-11

0 Introduction

With the development of aviation, the multi-airport system terminal area (MAS-TMA) has been evolved into an operation environment for high-density air traffic. The complexity of MAS-TMA is exerting safety strain on the existing air traffic management (ATM) system due to the restricted airspace resource. Additionally, the persistent stagnation in the evolution of tactical decision-making over the past half-century has left air traffic controllers (ATCOs) facing operational workload under increasingly complex conditions[1]. These shortcomings imply that the current ATM system is ill-prepared for high-density, complex, and dynamic air traffic. Therefore, enhancing the ATM system is imperative to increase airspace capacity and reduce ATCOs workloads.

As a critical component of air traffic management, conflict resolution (CR) faces challenges in achieving effective and optimal resolution. Fortunately, autonomous air traffic control (ATC) has emerged as a crucial solution to improve the CR. The initial concept of autonomous CR (AutoCR) decision-making was introduced in 2005 within the framework of advanced airspace concept (AAC). It was originally implemented as designed assistant tools, including Autoresolver and TSAFE[2-4]. While these automated tools were not widely employed[5], they had brought substantial benefits by providing tactical advisories for ATCOs. In a scenario exemplified by CR, the advisories are provided to ensure safety separation and assign adaptive

*Corresponding author, E-mail address: tianyong@nuaa.edu.cn.

How to cite this article: HUANG Xiao, TIAN Yong, LI Jiangchen, et al. Autonomous conflict resolution (AutoCR) based on improved multi-agent reinforcement learning[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2025,42(S):91-101.

http://dx.doi.org/10.16356/j.1005-1120.2025.S.008

routes. Therefore, the AutoCR is emphasized again for satisfying the previous requirement in high-density and dynamic MAS-TMA.

To actualize an autonomous system, a robust and reliable autonomous decision-making algorithm should be introduced. Artificial intelligence (AI) is highly suitable for this purpose. Current research has explored various AI approaches in ATM, including automat theory[6], multi-agent formulation[7], and reinforcement learning (RL)[8]. Among these, RL has exhibited auspicious qualities in perceiving air traffic situations and providing effective advisory to agents. Nevertheless, the AutoCR applied with RL still struggles to gain the trust of ATCOs[5,9] due to concerns about system deviations from actual operations. Furthermore, the inherent gap hampers addressing robustness and adaptability issues are evident in the lack of adaptability to highly dynamic environments and the absence of intention learning in agents for effective action execution.

Considering abovementioned shortcomings, this paper presents a framework for scalable AutoCR in the high-density stochastic MAS-TMA environment, leveraging multi-agent DRL techniques. This study presented three primary contributions. First, the proposed AutoCR approach adopts a MARL framework, in which every aircraft is modeled as an independent agent. Second, the dynamic weather conditions are incorporated to ensure the framework's effectiveness to uncertain airspace conditions. Third, the flight intent is emphasized, integrating interpretability into the component of the multi-agent DRL.

1 Related Work

AutoCR has attracted sustained research attention over the past decades. This is attributed to the development of applied methods, including the optimization control method, Markov decision-making processes (MDPs) approach and multi-agent DRL techniques. Consequently, the literature review of AutoCR is categorized into three stages, corresponding to the evolution of the methodology.

In the pre-RL optimal control stage (prior to 2011), the initial research on conflict resolution focused on optimization algorithm. This work was pioneered by Erzberger[2], which automated the control of aircraft. Specifically, the conflict resolution algorithm involved trajectory programming, candidate trajectories evaluating until the conflict-free trajectory was obtained[10]. However, it requires transmitting extensive information to ATCOs within a centralized architecture, which resulted in mishandling the high-density stochastics airspace.

In the proto-RL MDPs-based stage (2011—2017), the MDP gained attention in AutoCR under uncertainties using probabilistic models[11]. It was successfully applied as airborne collision avoidance system X (ACAS-X)[12]. Furthermore, the offline MDP-based methods for autonomous ATC evolved into various forms, including partially observable MDP (POMDP)[12], multi-agent MDP (MMDP)[13], and continuous-time MDP (CTMDP)[14]. While these MDP formulations have shown promise in large-scale simulations, they have drawbacks such as requirements for high computational resources[15] and limited adaptability in dynamic environments, thus limiting their applicability in autonomous ATC.

In the On-RL autonomous control stage (since 2018), RL techniques have been increasingly applied to address the challenges encountered in MDP-based AutoCR. The introduction of RL to AutoCR was initially documented in Ref.[16] for the dynamic enroute sector, characterized by multiple intersections and merging points. Additionally, the interactive conflict resolver using RL was emphasized by considering the preferences of ATCOs[17]. These centralized-based approaches focused on single-agent collisions, which might have limitations in highly coordinated environments. The multi-agent RL formulation represents a collaborative solution that has been explored in Refs.[18-19]. Consequently, the multi-agent frameworks employing techniques such as the actor-critic (A2C) algorithm[20], the long short-term memory (LSTM) model[21], the ORCA algorithm[22], and the DRL model[23] gradually gained application in the field of AutoCR.

While research in automation operations has yielded promising results, several shortcomings re-

main. First, the effectiveness of multi-agent DRL applied to AutoCR decision-making has not been proven in high-density and complex scenarios such as MAS-TMA. Second, dynamic weather conditions are often not incorporated, reducing the framework's effectiveness in uncertain airspace. Third, flight intent is rarely emphasized, leading to poor interpretability and a lack of transparency in multi-agent DRL.

2 Problem Description

In autonomous ATC systems, multiple aircraft agents operate collaboratively in the MAS-TMA, making decisions according to their respective states. The AutoCR directs aircraft toward the designated destination airport while reducing spatial conflicts. However, the sustained growth in air traffic demand imposes additional complexity on this collaboration process. To comprehensively address this problem, a detailed description is presented in this section, including complex MAS-TMA environment, interactive conflict solver, and autonomous flight scheduling.

2.1 Problem definition

Terminal areas serve as transitional zones between air routes and airports. The intersections of arrival routes of airports create a complex operational environment, making it challenging for aircraft to maneuver while ensuring safety. For improved solutions in the autonomous scheduling of multiple aircraft, it is crucial to model the MAS-TMA properly. The MAS-TMA encompasses structured airspace elements such as arrival points, routes, and airports. Fig.1 depicts a generalized schematic of MAS-TMA.

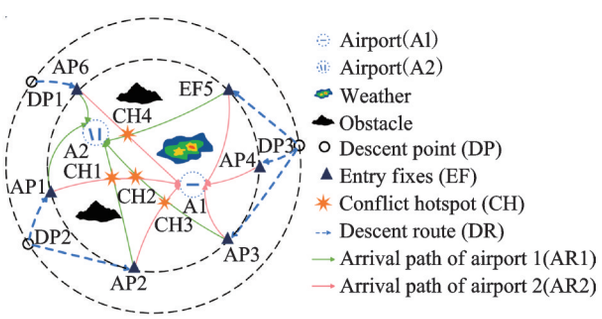

Fig.1 Generalized schematic of MAS-TMA

As illustrated in Fig.1, the MAS-TMA graphically depicted with two concentric circles. Within circular region, there are three descent points, six entry fixes, and two destination airports. Once agents acquire these three positions, the arrival path is nearly finalized. And, multiple flight trajectories $\boldsymbol{p}_{\text{Entry}}\boldsymbol{p}_{\text{Dest}}$ link the entry fixed points $\boldsymbol{p}_{\text{Entry}}$ to their respective destination airports $\boldsymbol{p}_{\text{Dest}}$. Intersections among these routes form potential conflict hotspots, where the risk of collision is elevated. In the arrival phase, aircraft must maintain adequate separation buffers to mitigate potential conflicts. Adverse weather conditions constitute another category of hazardous regions that must be avoided, as they are highly dynamic and subject to real-time variability. Collectively, these factors constrain aircraft maneuverability while ensuring operational safety, and even minor trajectory adjustments during arrival may propagate through the system, resulting in cascading effects. Consequently, the primary challenge is to generate conflict-free trajectories that completely avoid potential hazards.

2.2 Problem formulation

The AutoCR task can be modeled as a MDP defined by a six-component tuple $\langle \boldsymbol{S}, \boldsymbol{A}, \boldsymbol{R}, P, \gamma, \mu \rangle$, where $\boldsymbol{S}$ denotes the state space, $\boldsymbol{A}$ the action space, $\boldsymbol{R}$ the reward function, $P(s'|s,a)$ the state transition function, $\gamma \in (0,1)$ the discount factor, and μ the initial state distribution. In this framework, each aircraft operates as an independent agent with state s_t^i at time step t. Given s_t^i, the agent selects an action a_t^i according to a probabilistic policy $\pi(\cdot|s_t^i)$, aiming to navigate from its current position to the assigned destination while avoiding potential conflicts. The stationary policy $\pi: \boldsymbol{S} \to P(\boldsymbol{A})$ defines the mapping from any given states to probability distributions over possible actions. In MARL, the objective is to determine an optimal policy π^i that maximizes a performance measure $J(\pi^i)$, which is typically defined as the finite-horizon discounted return $J(\pi^i)$, shown as

$$J(\pi^i) = \underset{\tau^i \sim \pi^i}{E}\left[\sum_{t=t_{\text{Entry}}}^{t_{\text{Dest}}} \gamma^t R(s_t^i, a_t^i, s_{t+1}^i)\right] \tag{1}$$

Let τ^i denote the set of training process variables, with $\tau^i=(s^i_{t_{\text{Entry}}}, a^i_{t_{\text{Entry}}}, s^i_{t_{\text{Entry}+1}}, a^i_{t_{\text{Entry}+1}}, \cdots, s^i_{t_{\text{Dest}}}, a^i_{t_{\text{Dest}}})$. The notation $\tau^i \sim \pi^i$ indicates that the trajectory distribution is conditioned on policy π^i: $s^i_{t_{\text{Entry}}} \sim \mu$, $a^i_t \sim \pi(\cdot|s^i_t)$, $s^i_{t+1} \sim P(\cdot|s^i_t, a^i_t)$. The variables t_{Entry} and t_{Dest}, represent the time at which aircraft i reaches its entry fix point and its destination.

For effective conflict resolution, the primary objective is to generate conflict-free arrival trajectories. To this end, each state s^i_t must explicitly incorporate the position of the corresponding aircraft, and τ^i should be represented as a trajectory $\boldsymbol{\mathcal{T}}^i$.

Definition 1 Let $F(\cdot)$ be an injective mapping from τ^i to $(s^i_{t_{\text{Entry}}}, s^i_{t_{\text{Entry}+1}}, \cdots, s^i_{t_{\text{Dest}}})$, denoted concisely as $F(\tau^i)=(s^i_{t_{\text{Entry}}}, s^i_{t_{\text{Entry}+1}}, \cdots, s^i_{t_{\text{Dest}}})$.

Definition 2 Let $\boldsymbol{G}(\cdot)$ be an injective mapping from s^i_t to (x^i_t, y^i_t), expressed as $\boldsymbol{G}(\cdot): s^i_t \rightarrow \boldsymbol{p}^i_t$, where $\boldsymbol{p}^i_t$ corresponds to the spatial coordinates (x^i_t, y^i_t) of the aircraft i at time t.

Then, the $\boldsymbol{\mathcal{T}}^i$ associated with policy π^i is formulated as

$$\boldsymbol{\mathcal{T}}^i=\left\{\boldsymbol{p}^i_t\boldsymbol{p}^i_{t+1}\middle|\begin{array}{l}a^i_t\sim\pi(\cdot|s^i_t), s^i_{t+1}\sim P(\cdot|s^i_t,a^i_t)\\ t=t_{\text{Entry}}, t_{\text{Entry}+1}, \cdots, t_{\text{Dest}}\\ F(\tau^i)=(s^i_{t_{\text{Entry}}}, s^i_{t_{\text{Entry}+1}}, \cdots, s^i_{t_{\text{Dest}}})\\ \boldsymbol{G}(s^i_t)=\boldsymbol{p}^i_t, \boldsymbol{p}^i_t=(x^i_t, y^i_t)\end{array}\right\} \quad (2)$$

where $\boldsymbol{p}^i_t\boldsymbol{p}^i_{t+1}$ represents the trajectory segment from point $\boldsymbol{p}^i_t$ to point $\boldsymbol{p}^i_{t+1}$. The spatial relation between adjacent points is defined as

$$\boldsymbol{p}^i_{t+1}=\boldsymbol{p}^i_t+\boldsymbol{v}^i_t\Delta t \quad (3)$$

And the any trajectory points $\boldsymbol{p}^i_{t+\lambda\Delta t}$ along $\boldsymbol{p}^i_t\boldsymbol{p}^i_{t+1}$ can be expressed as

$$\begin{gathered}\boldsymbol{p}^i_{t+\lambda\Delta t}=\boldsymbol{p}^i_t+\boldsymbol{v}^i_t\times\lambda\Delta t\\ \lambda\in(0,1), \boldsymbol{v}^i_t\in[\boldsymbol{v}^i_{\min}, \boldsymbol{v}^i_{\max}]\end{gathered} \quad (4)$$

where $\boldsymbol{p}^i_{t+\lambda\Delta t}$ denotes the spatial coordinates of trajectory point of aircraft i at $(t+\lambda)\Delta t$ and $\boldsymbol{v}^i_t$ the airspeed executed under action a^i_t.

Eq.(2) provides the trajectory of the aircraft. Nevertheless, this representation does not account for potential hazards created by surrounding aircraft. As a result, an action a^i_t sampled from the policy $a^i_t\sim\pi(\cdot|s^i_t)$ often directs the aircraft directly toward its destination, which may maximize the reward function. However, this outcome is unrealistic in actual operations. After an action is executed, the subsequent aircraft state s^i_{t+1} must also comply with safety-separation requirements. These requirements include maintaining adequate separation between aircraft and avoiding proximity to hazardous weather regions, which can be formally described as

Constraint 1 $\|\boldsymbol{p}^m_{t+\lambda\Delta t}\boldsymbol{p}^n_{t+\lambda\Delta t}\|\geqslant D_{\text{Sep1}}, m, n\in\mathcal{I}, \lambda\in(0,1)$.

Constraint 2 $\|\boldsymbol{p}^i_{t+\lambda\Delta t}\boldsymbol{p}^{\text{w}}_{t+\lambda\Delta t}\|-D^{\text{w}}\cdot\rho^{\text{w}}_t\geqslant D_{\text{Sep2}}, i\in\mathcal{I}, \lambda\in(0,1)$.

where $\|\boldsymbol{p}^m_t\boldsymbol{p}^n_t\|$ denotes the Euclidean distance between aircraft m and n; the term $\|\boldsymbol{p}^i_t\boldsymbol{p}^{\text{w}}_t\|$ represents the distance between aircraft i and the center of a hazardous weather; the two parameters, D_{Sep1} and D_{Sep2}, specify the minimum separation requirements, with one defining aircraft to aircraft separation and the other defining aircraft to weather separation; D^{w} is the radius and ρ^{w}_t is the scaling factor, which together define the weather-influenced region of airspace; the set $\mathcal{I}$ denotes all arriving aircraft operating within the MAS-TMA.

Conflict resolution is not executed at every time step. Instead, it is activated only when a conflict is detected, determined by comparing the calculated separation with the prescribed safety separation thresholds.

In summary, the conflict-free trajectory set for all arriving aircraft is formulated as

$$T=\left\{\boldsymbol{\mathcal{T}}^i\middle|\begin{array}{l}i\in I, t=t_{\text{Entry}}, t_{\text{Entry}+1}, \cdots, t_{\text{Dest}}\\ \lambda\in(0,1), \boldsymbol{v}^i_t\in[\boldsymbol{v}^i_{\min}, \boldsymbol{v}^i_{\max}]\\ \text{Constraint 1}\wedge\text{Constraint 2}\end{array}\right\} \quad (5)$$

Hence, the objective of AutoCR task is achieved once conflict-free trajectories are obtained. To derive a shared optimal policy for all aircraft, we adopt the objective of minimizing the flight distance within the MAS-TMA, which is formally expressed as

$$\arg\min_{\pi}\left[\sum_{i\in I}\|\boldsymbol{\mathcal{T}}^i\|\right] \quad (6)$$

3 Improved Multi-agent Reinforcement Learning

Given the aforementioned problem, this section focuses on constructing a multi-agent conflict resolver for aircraft flying to destination airports with-

in the MAS-TMA. The framework is commenced with significant model elements, including state observation, action space, and reward function.

3.1 State

In MARL, the agent makes decisions based on the state received from the environment. This reliance on environmental state information underscores the need for the state space to encompass all pertinent data required by the agent for decision-making. Hence, the state space $\mathbf{S}$ should contain at least two critical components: The state information of aircraft s_t^{a} and the dynamic weather s_t^{w}.

$$s_t^{\mathrm{a}}=\left\{(\boldsymbol{p}_t,\boldsymbol{p}_{\mathrm{Entry}},\boldsymbol{p}_{\mathrm{Dest}},\mathbf{hd}_t,\boldsymbol{v}_t)\mid \boldsymbol{p}_t\in\mathcal{K}\wedge\boldsymbol{p}_t\notin\mathcal{K}_t\right\} \tag{7}$$

$$s_t^{\mathrm{w}}=\{(\boldsymbol{p}_t^{\mathrm{w}},\boldsymbol{v}_t^{\mathrm{w}},D^{\mathrm{w}},\rho_t^{\mathrm{w}})\mid \boldsymbol{p}_t^{\mathrm{w}}\in\mathcal{K}_t,\rho_t^{\mathrm{w}}\in[0,2]\} \tag{8}$$

The state information of aircraft includes its current position $\boldsymbol{p}_t$, designated entry fix $\boldsymbol{p}_{\mathrm{Entry}}$, destination airport $\boldsymbol{p}_{\mathrm{Dest}}$, heading $\mathbf{hd}_t$, and velocity $\boldsymbol{v}_t$. The state information of weather consists of the center location $\boldsymbol{p}_t^{\mathrm{w}}$, effective influence radius D^{w}, and velocity $\boldsymbol{v}_t^{\mathrm{weather}}$. The $\mathcal{K}$ denotes the MAS-TMA airspace. $\mathcal{K}_t$ is the hazardous weather region.

3.2 Action

Drawing inspiration from recent developments in advanced action space design within MARL, this study refines the representation of flight intention and embeds it adaptively into the action space. Here, the flight intention is defined in two dimensions, represented by two sub-actions: Airspeed intention a_t^{Speed} and heading intention a_t^{Heading}. These sub-actions can be formulated as

$$a_t^{\mathrm{Speed}}=\begin{cases}1 & \text{Lower speed}\\ 2 & \text{Normal speed}\\ 3 & \text{Higher speed}\end{cases} \tag{9}$$

$$a_t^{\mathrm{Heading}}=\begin{cases}1 & \text{Yaw left}\\ 2 & \text{Aiming}\\ 3 & \text{Yaw right}\end{cases} \tag{10}$$

Once the action space is defined, agent maneuvers are modeled to resolve potential conflicts. The aircraft are capable of performing different maneuvers within their performance envelopes. Thus, maneuverability is characterized through predefined bounds, expressed in terms of aircraft overload to quantify the feasible maneuver set for each agent.

Overload is defined as the ratio of the resultant aerodynamic and thrust forces to the gravity acting on the aircraft. The upper and lower maneuvering limits are determined by the maximum permissible overload. According to CCAR-25 regulation, the overload experienced during maneuvering must remain within $[-1, 2.5]$. The maneuvering space can therefore be expressed as

$$\mathcal{M}=(m_t^{\mathrm{Speed}},m_t^{\mathrm{Heading}}) \tag{11}$$

where $\mathcal{M}$ denotes a two-dimensional space for the agent's actions. The limits on speed maneuvers are defined by constraining the allowable range of acceleration and deceleration within $[-6\%, 3\%]$. Accordingly, the speed adjustment available to each agent can be formulated as

$$m_t^{\mathrm{Speed}}=\begin{cases}0.94v_t^i & a_t^{\mathrm{Speed}}=1\\ v_t^i & a_t^{\mathrm{Speed}}=2\\ 1.03v_t^i & a_t^{\mathrm{Speed}}=3\end{cases} \tag{12}$$

As indicated in Eq.(12), the adjustment of airspeed is formulated in terms of a rate of change rather than maintaining a constant speed. For example, when agent i repeatedly applies a deceleration command over three steps (from $t-1$ to $t+1$), its speed v_{t+1}^i will evolve to $0.94^2\cdot v_{t-1}^i$ which is not constant. Thus, the speed is continuously updated over time instead of remaining constant.

The heading maneuver limits are specified by constraining the allowable range of directional changes within $[-\pi/9, \pi/9]$. Accordingly, the heading adjustment for each agent can be expressed as

$$m_t^{\mathrm{Heading}}=\begin{cases}\mathbf{hd}_t^i-\dfrac{\pi}{9} & a_t^{\mathrm{Heading}}=1\\ \mathbf{hd}_t^i & a_t^{\mathrm{Heading}}=2\\ \mathbf{hd}_t^i+\dfrac{\pi}{9} & a_t^{\mathrm{Heading}}=3\end{cases} \tag{13}$$

As described in Eq.(13), the adjustment of heading is implemented through interpolation between discrete action points with a fixed step of $\pi/9$. For instance, if the current heading is $\mathbf{hd}_t^i$, the heading at the next step may become $\mathbf{hd}_{t+1}^i=\mathbf{hd}_t^i-\pi/9$, $\mathbf{hd}_{t+1}^i=\mathbf{hd}_t^i$ (maintain), or $\mathbf{hd}_{t+1}^i=\mathbf{hd}_t^i+\pi/9$ (right turn), and consecutive iterations continue with the same $\pi/9$ incremental change, thereby ensuring a smooth transition rather than a fixed heading.

3.3 Reward

The reward functions incorporated in this study are consisted of destination-reached reward, destination-approaching reward, weather-entered reward, and collision reward.

In the CR task within the MAS-TMA, aircraft should receive a reward upon successfully reaching their destinations. More specifically, the reward r^{Reach} is assigned according to the proportion of aircraft that arrive on time. The punctual ratio of airport j can be calculated as

$$\mu^j = \frac{\sum_{i \in \mathcal{I}_j} \delta^{i,j}}{f^j}, \delta^{i,j} = \begin{cases} 1 & \| \boldsymbol{p}_t^{i,j} - \boldsymbol{p}_{\text{Dest}}^{i,j} \| = 0 \\ 0 & \text{Otherwise} \end{cases} \tag{14}$$

where $\delta^{i,j}$ is an indicator variable that equals "1" if aircraft successfully arrives at destination airport, and "0" otherwise; f^j denotes the total number of aircraft. Accordingly, the on-time arrival rate μ^j can be defined as the ratio of successful arrivals. Then, the reward r^{Reach} is represented as

$$r^{\text{Reach}} = \begin{cases} R^{\text{Reach}} & \forall j \in \mathcal{J}, \mu^j = 1 \\ \dfrac{\sum_{j \in \mathcal{J}} \mu^j}{g} \times R^{\text{Reach}} & \exists j \in \mathcal{J}, \mu^j \neq 0 \\ 0 & \text{Otherwise} \end{cases} \tag{15}$$

The position reward $r^{\text{Approaching}}$ is designed to encourage agents to reach their destination airports efficiently. It can be formulated as

$$r^{\text{Approaching}} = \frac{\| \boldsymbol{p}_{t-1} - \boldsymbol{p}_{\text{Dest}} \| - \| \boldsymbol{p}_t - \boldsymbol{p}_{\text{Dest}} \|}{\| \boldsymbol{p}_{\text{Entry}} - \boldsymbol{p}_{\text{Dest}} \|} \tag{16}$$

When an aircraft violates the separation constraint (Constraint 1) with another aircraft, it receives a substantial penalty specified in r^{Collide}, shown as

$$r^{\text{Collide}} = \begin{cases} -R^{\text{Collide}} & \| \boldsymbol{p}_{t+\lambda\Delta t}^m \boldsymbol{p}_{t+\lambda\Delta t}^n \| \leqslant D_{\text{Sep1}} \\ 0 & \text{Otherwise} \end{cases} \tag{17}$$

When an aircraft violates the separation constraint (Constraint 2) with hazardous weather, it receives a substantial penalty specified in r^{Enter}, shown as

$$r^{\text{Enter}} = \begin{cases} -R^{\text{Enter}} & \| \boldsymbol{p}_{t+\lambda\Delta t}^i \boldsymbol{p}_{t+\lambda\Delta t}^{\text{w}} \| - D^{\text{w}} \cdot \rho_t^{\text{w}} \leqslant D_{\text{Sep2}} \\ 0 & \text{Otherwise} \end{cases} \tag{18}$$

In Eqs. (15—18), R^{Reach}, R^{Collide}, and R^{Enter} are constant values.

4 Simulation Results

This section presents simulations performed on the deep Q-network (DQN) algorithm. The aim is to validate the performance of the scalable autonomous conflict resolution within MAS-TMA. Specifically, this section introduces the relevant content of simulation, including environmental parameters and results discussion.

4.1 Environment setting

The "Guangdong-Hong Kong-Macao" greater bay area (GBA) MAS-TMA ranks among the busiest airspace in China, and provides a representative case for modeling MAS-TMA operations. The simulation environment is constructed primarily on the basis of the airspace configuration, as depicted in Fig.2.

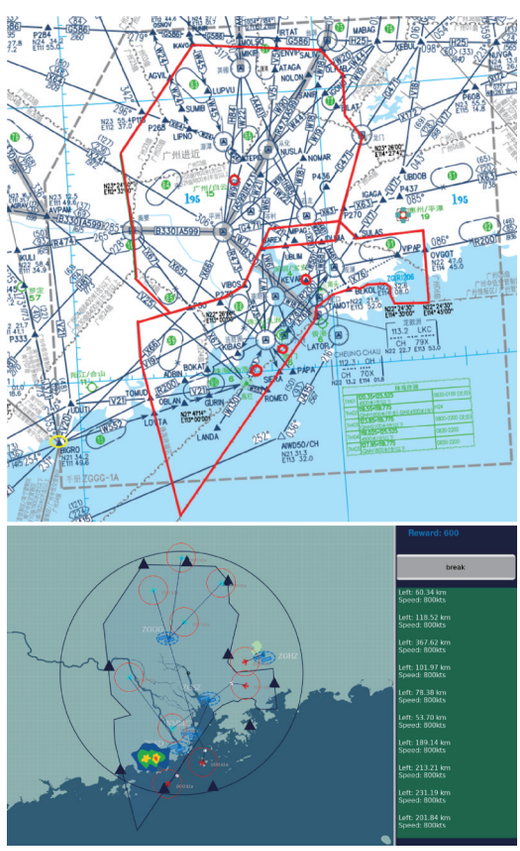

Fig.2 Simulation scenario

Within the realistic simulation environment, the hyperparameters were adjusted to satisfy algorithmic requirements while ensuring suitability for the AutoCR task. The principal hyperparameter settings are summarized in Table 1.

Table 1 Key hyperparameters of overall DQN

Parameter	Notation	Value
Learning rate	α	0.5
Discount factor	γ	0.9
Exploration rate	$1-\varepsilon$	0.1
Greedy rate	ε	0.9
Replay buffer size	N	20 000
Batch size	N_{BS}	100
Replay sampling times	N_{RST}	10
Target network update frequency	N_{TN}	20

Since training independent Q-networks are more suited to heterogeneous tasks, the employed DQN algorithm adopts a shared neural network with two hidden layers. This design enables parameter updates to be applied across all agents, thereby lowering computational overhead, accelerating convergence by exploiting multi-agent experiences, and promoting cooperative decision-making under a unified policy. The shared DQN processes local observations collected by agents through their interactions with the environment and outputs Q-values for the respective available actions. During training, gradients from all agents are aggregated to update the shared parameters using mini-batch stochastic gradient descent with experience replay and a target network. The Adam optimizer is applied for training, and the loss is evaluated using the mean squared error (MSE) criterion.

4.2 Results and discussion

(1) Training curves

Fig.3 presents the training results of conflict resolution. Two key evaluation metrics, namely the conflict rate and the resolution rate, are employed to assess system performance. Both indicators converge after roughly 2 000 episodes. At the outset of training, the resolution rate is about 20%, but it rises steadily and stabilizes near 96% after convergence. The resolution curve remains relatively smooth, reflecting the robustness of the algorithm. By contrast, the conflict rate decreases from approximately 16% to below 2%, although the curve exhibits noticeable fluctuations, which may be attributed to the structural complexity of the airspace. Overall, the training curves demonstrate that the algorithm achieves reliable conflict resolution performance.

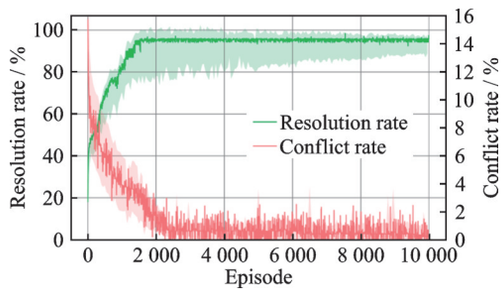

Fig.3 Training curves

(2) CR sample results

To evaluate the AutoCR capability of the proposed method, representative samples were drawn from the 1 000th, 5 000th, and 10 000th training episodes, corresponding to the early, middle, and final stages of training, respectively. Fig.4 illustrates the comparison between the initial planning trajectories and the conflict-free trajectories within the MAS-TMA.

The overall airspace environment in the figure is the same as the simulation scenario. The dotted lines in the solution diagram represent the deconflict trajectories of the aircraft. Each point on the dotted lines indicates the decision-making of the agents. The green dotted lines indicate high speed and the blue ones indicate low speed. The pink barred area indicates the conflict hot zone (higher risk of conflict), while the red bar indicates that flying the initially planned path at the current speed and heading conditions will generate a conflict. The pink circle indicates hazardous weather.

Overall, the model is capable of selecting appropriate evasive actions during the training process to resolve potential conflicts in the initial planned path. There are multiple conflict pairs and conflict groups in the initial planning trajectories in the early, middle, and late phases. The conflicts can be effectively detected in the middle and late phases and effectively avoided in accordance with the priority strategy. However, the action selected is not the optimal extrication maneuver under the conflict in the scenarios of a small training episode. Consequently, there remains a risk of conflict at the point where trajectories intersect even after executing a

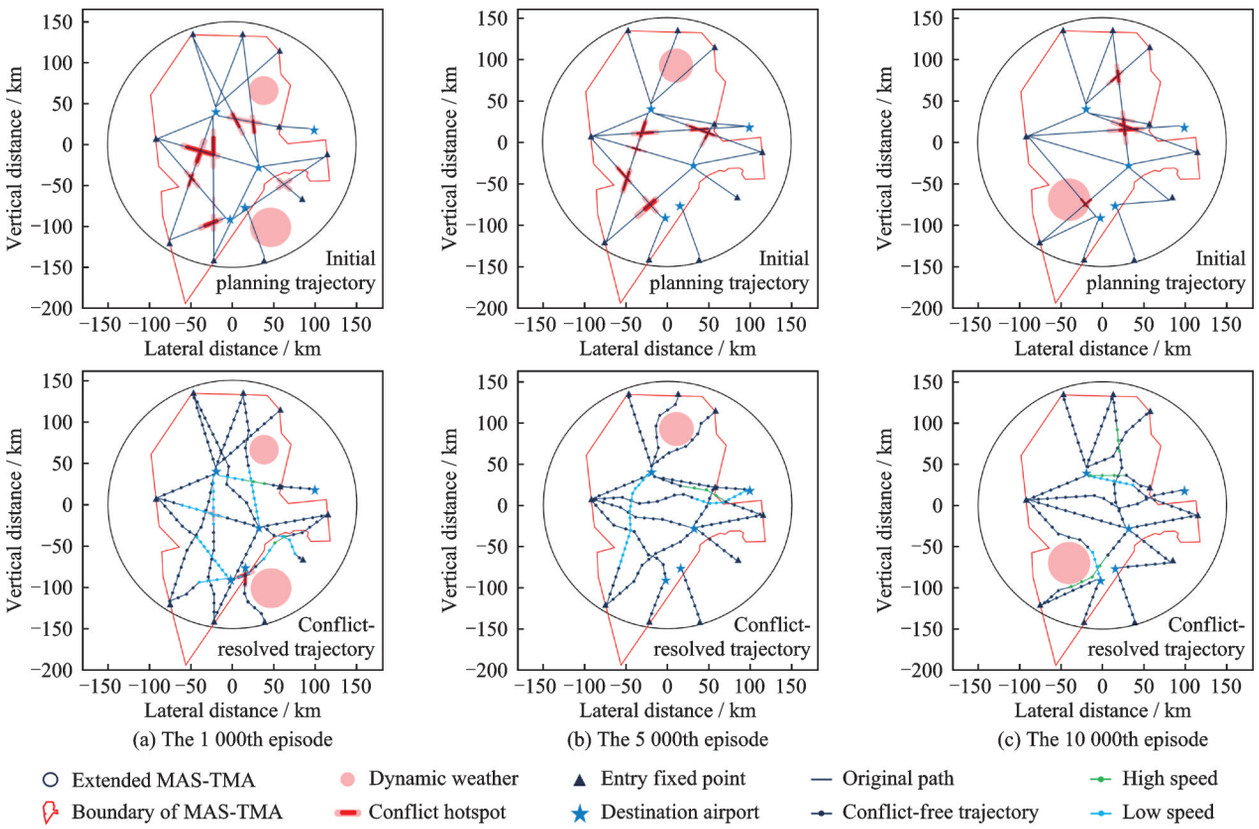

Fig.4 Sampling results of training comparison between original planned trajectories and conflict-resolved trajectories

maneuver. Moreover, the impact of dynamic weather conditions and preemptive conflict resolution can lead to secondary flight conflicts, even when flight conflicts in the initial planned trajectory are effectively resolved. As depicted in Fig.4(a), there is no flight conflict between Aircraft 1, departing from the entry point (116, −11) bound for the airport (−1, −93), and Aircraft 2, departing from the entry point (42, −144) heading towards the airport (14, −78). However, due to Aircraft 1 altering its path and speed to resolve a conflict with Aircraft 3 near point (57, −49), and Aircraft 2 changing path to evade adverse weather conditions, a secondary conflict arises near point (14, −89). This is consistent with the conflict resolution rate shown in Fig.3, which indicates that the rate approaches 100% only after nearly 2 000 training episodes.

From a detailed perspective, the transition of conflict resolution actions from the early to the late stages of training reveals that there is an increase in heading-changing actions and a decrease in the use of deceleration actions. At the 1 000th, 5 000th, and 10 000th training episodes, the aircraft are able to successfully avoid the weather. If the weather cannot be avoided, the training episode is terminated.

To further validate the effectiveness of the proposed framework, four additional DRL algorithms (DDQN, DDPG, PPO, and SAC) were compared with the proposed AutoCR method under varying traffic scales. Fig.5 presents the comparative results for conflict resolution. The results in Fig.5

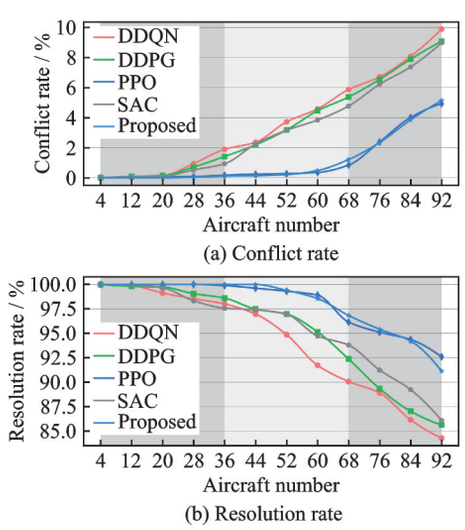

Fig.5 Comparison of conflict resolution performance between the DRL and proposed method

confirm that the proposed method achieves superior conflict resolution performance across most traffic conditions. The difference between the proposed method and SAC is relatively small, as both exhibit closely aligned conflict rates and resolution rates, and both outperform the other three DRL methods. In medium-density cases (up to 60 aircraft), both approaches maintain a conflict rate below 0.35% and a resolution rate above 98.9%. The other methods demonstrate acceptable performance only in low-density scenarios (no more than 36 aircraft).

(3) Discussion

We further analyze how variations in air traffic density influence the number of conflicts. Fig.6 presents the conflict counts observed under different density levels. Combining results in Fig.3, it can be concluded that the number of conflicts per episode stabilizes, once a sufficient number of training episodes has been completed. Therefore, only data from the first 2 000 episodes are considered here to more clearly examine the relationship between conflict occurrence and airspace density. In addition, published results from other studies are incorporated to further demonstrate the performance advantages of the joint approach proposed in this work.

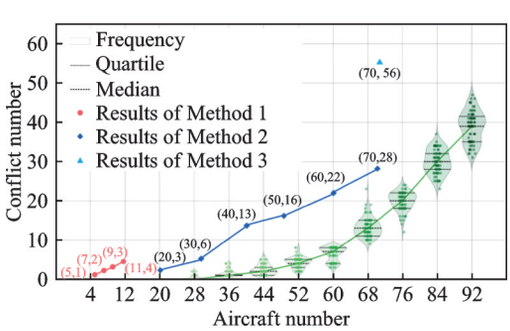

Fig.6 Variation in conflict trends under different air traffic densities

The results reveal that conflict occurrence grows exponentially with increasing airspace density. This sharp escalation can be attributed to congestion effects at high traffic levels, which increase the probability of separation loss between aircraft. For instance, when the MAS-TMA density is set to 52, the total number of potential conflicts remains below 10, with minimal variation between the maximum and minimum cases. In contrast, at a density of 92, the maximum number of conflicts observed during training reaches 50, with each aircraft experiencing an average of 0.32—0.54 conflicts. Since such traffic density exceeds that encountered in current operational environments, the pronounced increase in conflicts is consistent with expectations. These findings underscore the importance of implementing pre-tactical conflict-free trajectory planning in future high-density scenarios, rather than relying solely on tactical-level conflict resolution.

In comparison with other approaches reported in previous studies, the proposed method provides a more robust solution for conflict resolution, as shown in Fig.6. In low-density scenarios, results obtained with Method 1 exhibit an increase in conflicts, whereas the proposed method keeps conflict levels close to zero, demonstrating its effectiveness in globally mitigating conflicts in sparse traffic environments. For medium-density scenarios, the proposed approach consistently achieves approximately eight fewer conflicts on average than that of Method 2. At a traffic level of 70 aircraft, the number of potential conflicts is reduced by about 40 relative to the results obtained using Method 3. Methods 1, 2, and 3 are adopted from Refs.[24-26], respectively. Importantly, the proposed method maintains its effectiveness even in high-density scenarios, yielding on average eight fewer conflicts across low-, medium-, and high-density scenarios compared with the benchmark methods.

5 Conclusions

This paper introduces an autonomous conflict resolution framework designed for complex and high-density MAS-TMA environments. The AutoCR problem is formulated as a multi-agent reinforcement learning, with improvements made to the DQN algorithm by incorporating dynamic weather conditions and flight intent. Simulation experiments were conducted in the GBA MAS-TMA to demonstrate the effectiveness of the proposed method. The algorithm's performance was validated using conflict rate and resolution rate metrics. Early-, middle-, and late-stage training results were sampled to better verify the algorithm's conflict resolution capability in high-density MAS-TMA. Comparisons with

other methods also indicate that the proposed method is effective in resolving conflicts, successfully eliminating over eight potential conflicts.

References

[1] COUNCIL N R. Autonomy research for civil aviation: Toward a new era of flight[M]. Washington, DC: National Academies Press, 2014.

[2] ERZBERGER H. Automated conflict resolution for air traffic control[C]//Proceedings of International Congress of Aeronautical Sciences (ICAS 2006). Hamburg, Germany: [s.n.], 2005.

[3] ERZBERGER H, HEERE K. Algorithm and operational concept for resolving short-range conflicts[J]. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering, 2010, 224(2): 225-243.

[4] FARLEY T, ERZBERGER H. Fast-time simulation evaluation of a conflict resolution algorithm under high air traffic demand[C]//Proceedings of the 7th USA/Europe ATM 2007 R & D Seminar. [S.l.]: ATM, 2007.

[5] WESTIN C, BORST C, HILBURN B. Automation transparency and personalized decision support: Air traffic controller interaction with a resolution advisory system[J]. IFAC-PapersOnLine, 2016, 49(19): 201-206.

[6] BAYEN A, GRIEDER P, MEYER G, et al. Lagrangian delay predictive model for sector-based air traffic flow[J]. Journal of Guidance, Control, and Dynamics, 2005, 28(5): 1015-1026.

[7] WOLFE S R. Supporting air traffic flow management with agents[C]//Proceedings of the AAAI Spring Symposium: Interaction Challenges for Intelligent Assistants. [S.l.]: AAAI, 2007: 137-138.

[8] ALVES D, LI W, SOUZA B B. Reinforcement learning to support meta-level control in air traffic management[M]//Reinforcement Learning. Vienna, Austria: Tech Education and Publishing, 2008.

[9] WESTIN C, HILBURN B, BORST C. The effect of strategic conformance on acceptance of automated advice: Concluding the MUFASA project[J]. Proceedings of SESAR Innovation Days, 2013. DOI: 10.5772/5293.

[10] PONTANI M, CONWAY B A. Particle swarm optimization applied to space trajectories[J]. Journal of Guidance, Control, and Dynamics, 2010, 33(5): 1429-1441.

[11] CHRYSSANTHACOPOULOS J P, KOCHENDERFER M J. Accounting for state uncertainty in collision avoidance[J]. Journal of Guidance, Control, and Dynamics, 2011, 34(4): 951-960.

[12] CHRYSSANTHACOPOULOS J P, KOCHENDERFER M J. Decomposition methods for optimized collision avoidance with multiple threats[C]//Proceedings of 2011 IEEE/AIAA 30th Digital Avionics Systems Conference. Seattle, WA, USA: IEEE, 2011: 1D2-1-1D2-11.

[13] ONG H Y, KOCHENDERFER M J. Markov decision process-based distributed conflict resolution for drone air traffic management[J]. Journal of Guidance, Control, and Dynamics, 2017, 40(1): 69-80.

[14] MAHBOUBI Z, KOCHENDERFER M J. Continous time autonomous air traffic control for non-towered airports[C]//Proceedings of the 54th IEEE Conference on Decision and Control (CDC). Osaka, Japan: IEEE, 2015: 3433-3438.

[15] MUELLER E R, KOCHENDERFER M. Multi-rotor aircraft collision avoidance using partially observable Markov decision processes[C]//Proceedings of AIAA Modeling and Simulation Technologies Conference. Washington, DC: AIAA, 2016.

[16] BRITTAIN M, WEI P. Autonomous air traffic controller: A deep multi-agent reinforcement learning approach[EB/OL]. (2019-05-02). https://arxiv. org/abs/1905.01303.

[17] TRAN P N, PHAM D T, GOH S K, et al. An interactive conflict solver for learning air traffic conflict resolutions[J]. Journal of Aerospace Information Systems, 2020, 17(6): 271-277.

[18] YANG X, WEI P. Scalable multi-agent computational guidance with separation assurance for autonomous urban air mobility[J]. Journal of Guidance, Control, and Dynamics, 2020, 43(8): 1473-1486.

[19] BERTRAM J, WEI P. Distributed computational guidance for high-density urban air mobility with cooperative and non-cooperative collision avoidance[C]//Proceedings of AIAA Scitech 2020 Forum. Orlando, FL: AIAA, 2020: AIAA2020-1371.

[20] BRITTAIN M, WEI P. Autonomous separation assurance in an high-density en route sector: A deep multi-agent reinforcement learning approach[C]//Proceedings of 2019 IEEE Intelligent Transportation Systems Conference (ITSC). Auckland, New Zealand: IEEE, 2019: 3256-3262.

[21] BRITTAIN M W, WEI P. One to any: Distributed conflict resolution with deep multi-agent reinforcement learning and long short-term memory[C]//Proceedings of AIAA Scitech 2021 Forum. [S.l.]: AIAA, 2021: AIAA2021-1952.

[22] NIU H, MA C, HAN P, et al. An airborne approach for conflict detection and resolution applied to civil aviation aircraft based on ORCA[C]//Proceedings of 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). Chongqing, China: IEEE, 2019: 686-690.

[23] BRITTAIN M, YANG X, WEI P. A deep multi-agent reinforcement learning approach to autonomous separation assurance[EB/OL].(2020-03-17).https://arxiv.org/abs/2003.08353.

[24] LAI J, CAI K, LIU Z, et al. A multi-agent reinforcement learning approach for conflict resolution in dense traffic scenarios[C]//Proceedings of 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC). San Antonio, TX, USA: IEEE, 2021: 1-9.

[25] SHI K, CAI K, LIU Z, et al. A distributed conflict detection and resolution method for unmanned aircraft systems operation in integrated airspace[C]//Proceedings of 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC). San Antonio, TX, USA: IEEE, 2020: 1-9.

[26] GHOSH S, LAGUNA S, LIM S H, et al. A deep ensemble method for multi-agent reinforcement learning: A case study on air traffic control[C]//Proceedings of the International Conference on Automated Planning and Scheduling. Palo Alto, California, USA: AAAI, 2021: 468-476.

Acknowledgements This work was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No.KYCX25_0621), and the Foundation of Interdisciplinary Innovation Fund for Doctoral Students of Nanjing University of Aeronautics and Astronautics (No. KXKCXJJ202507).

Authors

The first author Mr. **HUANG Xiao** received the B.S. degree in traffic transportation from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2018, and is currently pursuing the Ph.D. degree in transportation planning and management at NUAA. His research interests include autonomous operation decision-making in smart civil aviation and autonomous decision-making for urban low-altitude UAVs.

The corresponding author Prof. **TIAN Yong** received the Ph.D. degree in transportation planning and management from NUAA, China, in 2009. He is currently a professor with College of Civil Aviation, NUAA. His research interests include air traffic management and airspace planning.

Author contributions Mr. **HUANG Xiao** conceived and designed the study, developed the models, performed the analyses, interpreted the results, and drafted the manuscript. Mr. **LI Jiangchen** contributed to the study background and scholarly discussion. Prof. **TIAN Yong** provided the data used for the Guangdong-Hong Kong-Macao Greater Bay Area (GBA) MAS-TMA analysis. Dr. **ZHANG Naizhong** contributed to formal analysis, investigation, and manuscript writing. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Executive Editor: ZHANG Huangqun)

基于改进多智能体强化学习的自主冲突解脱方法研究

黄 潇，田 勇，李江晨，张乃中

（南京航空航天大学民航学院，南京 211106，中国）

摘要：冲突解脱(Conflict resolution，CR)是空中交通管理的重要组成部分。近年来，人工智能的快速发展推动了深度强化学习(Deep reinforcement learning，DRL)技术在CR策略优化中的有效应用。然而，现有应用于CR的DRL模型大多局限于简单场景，且往往忽视高密度、多机场终端区(Multi-airport system terminal area，MAS-TMA)中众多航迹交叉点所带来的高风险。针对上述问题，本文提出了一种改进的多智能体DRL模型，用于MAS-TMA内的自主飞行冲突解脱(Autonomous conflict resolution，AutoCR)。该模型在状态空间中引入动态气象条件，完成MAS-TMA动态环境的精准建模，以增强适应性；在动作空间中引入飞行意图，并依据过载约束将其转化为最优机动方案，从而提升可解释性。基于粤港澳大湾区(Guangdong-Hong Kong-Macao greater bay area，GBA)MAS-TMA的仿真结果表明，AutoCR方法至少可有效消除8个潜在冲突，并在不同空中交通密度下表现出良好的鲁棒性。

关键词：空中交通管理；冲突解脱；多机场终端区；多智能体强化学习