计算机科学
COMPUTER SCIENCE

基于深度强化学习的多机冲突解决方法的研究

霍丹, 余付平, 沈堤, 韩雪艳

引用本文

霍丹, 余付平, 沈堤, 韩雪艳. 基于深度强化学习的多机冲突解决方法的研究[J]. 计算机科学, 2025, 52(7): 271-278.

HUO Dan, YU Fuping, SHEN Di, HAN Xueyan. Research on Multi-machine Conflict Resolution Based on Deep Reinforcement Learning [J]. Computer Science, 2025, 52(7): 271-278.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240800133

基于深度强化学习的多机冲突解决方法的研究

霍 丹 余付平 沈 堤 韩雪艳
空军工程大学空管领航学院 西安 710000

摘 要 随着军民航及通航飞行活动增多,用空矛盾突出,在同一空域中多架飞机同时飞行成为一种常态,如何通过技术手段提供辅助防撞决策,避免飞行冲突成为亟待解决的问题。针对航空器在飞行过程中的多机飞行冲突解脱问题,提出了一种基于多智能体深度强化学习,结合图卷积神经网络作为扩展框架的图卷积深度强化学习(GDQN)算法。首先构造消息传递功能,建立多智能体的飞行冲突模型,该模型可以在避免冲突和碰撞的同时,引导多架飞机穿越三维的非结构化空域;其次利用基于图卷积神经网络的深度自学习方法为机场调度提供智能化的冲突规避手段,针对多机飞行冲突场景建立多智能体系统(MAS);最后通过在受控的模拟环境中使用广泛的训练集来训练策略函数,对算法的有效性进行了仿真验证。结果表明,优化后的算法是可行的,用于解决冲突时的成功率可达90%以上,且冲突解决决策的计算时间短于3s,发出的空中交通管制(ATC)指令明显减少,效率得到了明显提升。

关键词:深度强化学习;图卷积神经网络;消息传递;多智能体模型;多机飞行;冲突解脱
中图分类号 TP389.1

Research on Multi-machine Conflict Resolution Based on Deep Reinforcement Learning

HUO Dan, YU Fuping, SHEN Di and HAN Xueyan
College of Air Traffic Control and Navigation, Air Force Engineering University, Xi'an 710000, China

Abstract With the increase in military, civilian, and general aviation flight activities, the conflict over airspace use has become prominent, and it has become a normal phenomenon for multiple aircraft to fly simultaneously in the same airspace. Therefore, it is an urgent problem that needs to be solved how to provide assistance in avoiding flight collisions through technical means. To tackle the challenge of resolving conflicts between multiple aircraft in flight, this paper introduces a Graph Convolutional Deep Reinforcement Learning(GDQN) algorithm. This algorithm combins multi-agent deep reinforcement learning with a graph convolutional neural network framework. Initially, it constructs a message-passing function to develop a multi-agent flight conflict model, which can navigate multiple aircraft through three-dimensional, unstructures airspace while avoiding conflicts and collisions. Subsequently, it employes a deep self-learning method based on graph convolutional networks to offer intelligent conflict avoidance solutions for airport scheduling, creats a multi-agent system(MAS) for managing multi-aircraft conflict scenarios. The effectiveness of the algorithm is validated through simulations using extensive training datasets in a controlled environment. The results indicate that the optimized algorithm is effective, achieving a conflict resolution success rate of over 90%, with resolution decision times of less than 3 seconds. Additionally, it significantly reduces the number of air traffic control(ATC) commands issued and improves overall operational efficiency.

Keywords Deep reinforcement learning, Graph convolutional neural network, Message passing, Multi-agent model, Multi-aircraft flight, Conflict resolution

随着近年来民航业的高速发展,空中交通流量日益增加。然而,由于空域资源有限,因流量增加而带来的飞行冲突问题日趋突出。这一问题将会带来空中交通态势紧张、航路拥挤、航班晚点率增加、安全可靠性降低等一系列问题。传统的完全依靠人工经验和技巧进行空域规划配置、流量调控和管制指挥的做法,已不能适应民航业高速发展的需要。随着空管人员的工作负荷和压力的提高,更容易出现"错、忘、漏、低效"现象。因此,发展智能化管制技术,对于维护空中交通的正常运行、提高空域流量、防止航空器之间相撞、保证空中交通安全具有重要的现实意义。

当前对于多机飞行冲突解脱问题的研究成果较多。深度强化学习是人工智能的前沿研究领域,随着深度强化学习在控制决策领域的不断发展,Wang等[1]将深度强化学习应用到飞行冲突检测和调配中,提出了K-Control Actor-Critic算法,在强化学习模拟环境中进行冲突检测与调配,通过调整航向角控制飞行器的方向,改变飞行路径,以最少改变飞行器航

到稿日期:2024-08-23 返修日期:2024-12-02
基金项目:国家社会科学基金(22BGL319)
This work was supported by the National Social Science Foundation of China(22BGL319).
通信作者:霍丹(kannyh2022@163.com)

向角的次数为优化目标，避免飞行器在空域中发生碰撞。Liu等[2]提出了通过混合整数非线性规划来优化速度变化和航向角变化，以解决冲突。Wen等[3]结合多架飞行器进行冲突解脱时具有的连续状态空间和连续动作空间特点，将深度确定性策略梯度算法(Deep Deterministic Policy Gradient，DDPG)应用于飞行冲突解脱任务中，通过仿真实验提高了多机飞行冲突的解脱效率。除此之外，Cai等[4]将K-Means空域聚类算法运用于冲突检测方法，Tong等[5]通过改进混沌蚁群算法解决了自由飞行过程中的多机冲突问题。

上述方法在解决多机冲突问题中取得了较好的效果，但考虑到在实际受限飞行条件下，多数是基于管制操作规程来解决多机之间的飞行冲突，随着空管人员的工作负荷和压力的提高，管制效率会下降。对于传统的图卷积神经网络算法的应用，在航空领域主要集中于对飞机方面的研究，通过控制群体化的飞机，从而模拟出类似动物的群体效应。此外，针对突发事件建立相关的群体决策模型，进行演化仿真。本文利用图卷积神经网络作为扩展框架，构造消息传递功能，推导出解决多智能体飞行冲突模型的GDQN(GNN Deep Q-Network)算法，仿真多机之间的群体行为，并在此基础上结合深度强化学习算法进行改进。改进后，在充分考虑人的主观能动性的基础上，将飞机当作智能体(Agent)来研究，利用图卷积神经网络(Graph Neural Network，GNN)作为扩展框架改进的深度强化学习(Deep Q-Network，DQN)算法进行研究，旨在提高多机冲突解决效率，减少空中交通管制(Air Traffic Control，ATC)指令，提高终端区航空器的安全性和运行效率。

综上所述，在国内外研究多机飞行状态预测的热点问题基础上，本文的目标是通过图卷积神经网络建立一种智能化方法，针对飞行冲突情况，综合考虑多架飞机的飞行状态，进行冲突状态的预测并实现冲突解脱，可以为管制员的指挥提供一定依据。

1 研究基础

强化学习(Reinforcement Learning，RL)是一种通过与环境交互来学习决策策略的机器学习方法。它的核心思想是让智能体(Agent)在执行动作(Action)、观察环境(Environment)反馈的状态(State)和奖励(Reward)的过程中，学习到一个最优策略(Optimal Policy)，从而实现长期累积奖励最大化[6]。

1.1 智能体强化学习模型

强化学习解决问题的第一步是建立一个马尔可夫决策过程(Markov Decision Process，MDP)[7]的模型，该模型包括一个状态空间、一个动作空间和一个奖励函数。

1.1.1 状态空间

在模型建立过程中，两个航空器近似为质点，其坐标即为环境状态 s，包括 $2n$ 个状态参数。因此状态空间如式(1)所示：

$$S=\{x_1,y_1,x_2,y_2,\cdots,x_n,y_n\} \tag{1}$$

1.1.2 动作空间

动作为 n 个航空器所选速度的组合，因为每个航空器有 m 个速度选择，所以一共有 m^n 种组合方式，故动作集合如式(2)所示：

$$\boldsymbol{A}=\{a_1,a_2,\cdots,a_{m^n}\} \tag{2}$$

1.1.3 奖励函数

在时间 t，环境处于一个来自可能状态的集合 $S=\{s\}$。在该状态中，智能体 s_t 属于 $a_t\in\{\boldsymbol{A}(S_t)\}$ 状态中可能的动作之一，同时智能体接收一个数值奖励 r_t，环境状态变化为 s_{t+1}。

智能体的目标是将预期收益最大化，与环境进行长时间交互的总奖励的获取方式如式(3)所示：

$$R_t=\sum_{k=0}^{\infty}r^k r_{t+k+1} \tag{3}$$

其中，r 是折扣系数，$0\leqslant r\leqslant 1$，这个系数越接近1，对智能体的未来奖励意义就越大。在学习过程中，根据环境的状态形成的动作选择策略为：$\pi(s,a)$即在状态 s 中选择动作 a 的概率。

本文模型中，智能体除了在与环境的多次交互过程中积累经验之外，没有其他的学习方式。其中价值函数即为智能体的经验积累值。每个智能体的动作值函数所对应策略 π 指当智能体在状态 s 中选择动作 a 并遵循该策略时的预期收益的数学期望，如式(4)所示：

$$\begin{aligned}Q^{\pi}(s,a)&=E_{\pi}\{R_t\mid s_t=s,a_t=a\}\\&=E_{\pi}\{\sum_{k=0}^{\infty}r^k r_{t+k+1}\mid s_t=s,a_t=a\}\end{aligned} \tag{4}$$

明确了 π 策略的价值函数 $V^{\pi}(s)$，就可以通过策略值为每个状态选择最佳动作来改进这个策略。而改进后的策略 π' 相对于价值函数 $V^{\pi}(s)$来说是贪婪的，如式(5)所示：

$$\begin{aligned}\pi'(s)&=\arg\max_a Q^{\pi}(s,a)\\&=\arg\max_a E_{\pi}\{r_{t+1}+\gamma V_k(s_{t+1})\mid s_t=s,a_t=a\}\end{aligned} \tag{5}$$

对于新的策略，π'对价值函数 $V^{\pi}(s)$进行评估，贪婪策略由新的价值函数、策略 π''等组成。在极限情况下，收敛到函数的最优值，对应的是最优的策略。

1.2 基于图的多智能体深度强化学习

图卷积网络(Graph Convolutional Network，GCN)是近年来逐渐流行的一种神经网络结构，它能够处理具有广义拓扑图结构的数据，并深入发掘其特征和规律，可以将有关拓扑结构通过节点之间的依赖关系以及图上的节点来表示[8]。

本文主要结合了基于图的深度强化学习方法。将GCN神经网络层作为训练层，其输入参数的表示如下：

1)特征矩阵 $\boldsymbol{X}\in\hat{R}\{N\times D\}$，其中 N 是节点数，D 是输入特征数。

2)邻接矩阵 $\boldsymbol{A}\in\hat{R}\{N\times N\}$描述了图的连通性。

输出一个矩阵 $\boldsymbol{Z}\in\hat{R}\{N\times F\}$，其中 F 是训练后的输出维度。邻接矩阵 $\boldsymbol{A}[v,w]$处的值可以是节点 v 和节点 w 之间的欧氏距离，如果节点连接，可以是1，如果节点不连接，则为0。考虑两个连接的节点(空域冲突点 v 和 w)，这两个节点之间的连接强度由 α_{vw}给出，而GCN层通过连接节点数加权的平均值来计算 α_{vw}。本文以DQN算法作为基本算法，并利用图卷积神经网络作为扩展框架，结合消息传递神经网络，推导出解决飞行冲突模型的GDQN算法。该模型可以在避免冲突和碰撞的同时，引导任意数量的飞机穿越三维的非结构化空

域，利用基于图卷积的深度自学习方法为机场调度提供智能化的冲突规避手段。

2　多智能体模型建立

多智能体飞行中的冲突解决关键是建立问题模型，对模型进行求解和计算，从而得到理论上最优的冲突解决策略[9]。因此，有必要对多智能体飞行冲突问题进行分析，选择合适的方法对问题进行建模。

2.1　问题分析

飞机在特定空域内飞行，遇到航线交叉或高度变换时都可能会产生飞行冲突。此时就需要提前感知冲突的存在，并采取恰当的方式进行冲突解脱。飞机在发生冲突时，通常可以通过调整高度、速度以及方向来避免冲突，这就需要在3D环境中完成更多的机动动作，但3D环境的复杂性会使智能体的训练变得困难。虽然高度调整在现行管制过程中适用度较高，但缺点是在既定空域中可供使用的飞行高度层较少，同时若穿越高度层次数较多，则可能会对其他航空器的正常飞行造成影响，反而加大管制员调配压力。此外，节点间建立连边大多只考虑航空器的位置信息，忽略了速度和航向对于飞行冲突的重要性，导致模型对冲突信息的反映不够准确。因此，本文重点考虑航空器在发生冲突时的速度和航向调整的2D控制。

一架飞机具有自主功能，如接收指令、改变飞行状态等，它可以抽象为一个智能体，其内部结构如图1所示。该智能体具有生成飞行航迹、接收控制指令和状态查询信号、传输状态信息等功能。

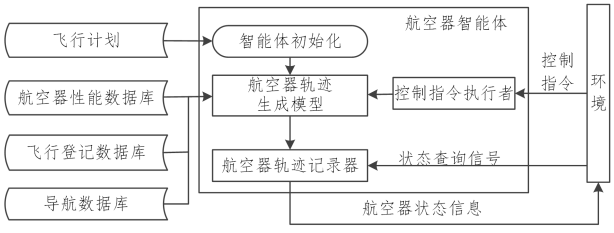

图1　航空器智能体内部结构

Fig. 1　Internal structure of aircraft agent

多机飞行的场景中涉及多个智能体，这些智能体之间相互关联，因此多机冲突场景可以看作一个分布式多智能体系统(Distributed Multi-Agent System，MAS)[10]。MAS的典型建模方法是将问题建模为MDP或随机博弈过程。从实际的控制规则来看，当检测到潜在冲突时，管制员根据飞机当前状态和下一个空域状态发送决议指令，而不考虑飞机执行指令之前的早期状态。因此，在飞机执行冲突解决指令后，下一个环境反馈只与当前状态和动作有关，这符合MDP的前提条件，即马尔可夫性。

假设智能体的初始状态是 s_0，选择一个动作 a_0 执行，智能体按照概率随机转移到下一个状态 s_1。然后执行另一个动作 a_1，转移到下一个状态 s_2，重复这个过程并继续。接着，定义一个图 $G=(V,E)$，V 包含图中所有的顶点，边集 E 表示连接所有顶点的边。边集 E 的元素表示为二元组 (v,w)，基于节点与边就可以构建出一个GCN图的模型。首先，每个节点由一个节点特征向量表示，在编码时设置为隐藏状态 $h_w \in n_h$，边可以是边的特征向量 $\boldsymbol{e}_{vw} \in n_e$，其中包含它们连接节点 v 和 w 的更新信息；然后，节点从相邻节点聚合信息进行学习，随后更新其隐藏状态；最后，经过几次更新迭代后，节点对自己和邻居的特征有了更多的了解，对组建整个图就有了准确的表示。而在整个过程中，从状态 s_t 到状态 $stp1$ 的转换将得到奖励 $strp1$。

2.2　模型建立

通过上述分析，将多机飞行冲突场景建模为一个MAS，每架冲突飞机都被视为一个智能体，具有接收管制员指令、执行指令、按照航迹预测模型飞行的能力。这些飞机彼此独立，将GDQN和MPNN作为智能体的学习方法。其中每个智能体对应于图的节点，边包含了关于它们相对位置的特征。在受控环境中进行了数千次模拟后，智能体学会了在采取联合行动之前进行沟通，以最大限度地提高长期奖励，降低惩罚分离损失，提升飞行效率和减少ATC指令。

2.2.1　初始化

首先设置消息函数，聚合邻居节点特征，为了兼容边的特征，假设消息函数为 M_{vw}，如式(6)所示：

$$M_{vw}=A(e_{vw})h_w^c \tag{6}$$

其中，$A(e_{vw})$ 为边的向量 e_{vw} 映射到 $d\times d$ 矩阵的神经网络。

该消息函数用于聚合目标节点的邻居特征，包括目标节点 $h_v^{(t)}$、邻居节点 $h_w^{(t)}$、边特征 e_{vw}，进而形成一个消息向量并传递给目标节点，如式(7)所示：

$$m_v^{(t+1)}=\sum_{w\in N_v} M_{vw}(h_v^{(t)},h_w^{(t)},e_{vw}) \tag{7}$$

其中，N_v 表示节点 v 的邻居节点集合。

然后，组合当前层节点的特征以及从消息函数中获得的消息，更新自身节点特征，如式(8)所示：

$$\begin{aligned} U_t&=GRU(h_w^t,m_v^{(t+1)})\\ h_v^{(t+1)}&=U_t(h_v^{(t)},m_v^{(t+1)}) \end{aligned} \tag{8}$$

其中，GRU(Gated Recurrent Unit)[11]是门控循环单元。

最后，利用读取函数进行图的读取，利用一个特征向量表示整张图，如式(9)所示：

$$\begin{aligned} \boldsymbol{R}&=\sum_{v\in V}\sigma(f_p(h_v^{(t)},h_v^{(0)}))\odot f_q(h_v^{(t)})\\ \hat{y}&=\boldsymbol{R}(\{h_v^{(t)}\mid v\in N\}) \end{aligned} \tag{9}$$

其中，$\odot$ 表示哈达玛乘积[12]；σ 是sigmoid激活函数；f_p 和 f_q 为前馈神经网络(Feedforward Neural Network，FNN)[13]。

2.2.2　Actor-Critic模型

模型中的飞机可以视为一个单独的智能体，智能体共同组成多智能体系统。在这个系统中，每个智能体之间的相互作用可以通过图的形式体现，即 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$，$\boldsymbol{V}$ 中的每个节点对应一个智能体。图中的边 $\boldsymbol{E}$ 不仅表明了智能体之间能够进行信息交流，还包含了与连接的两个智能体相关的特定属性。

首先，在初始化智能体后，进行与相邻智能体之间的多次通信。通信过程中采用注意力机制，该机制的功能类似于每个智能体向其他智能体查询关于它们的观察和操作的信息，并将这些信息合并到其值函数的估计中。消息函数如式(10)所示：

$$m_v^{(t+1)}=\sum_{w\in N_v} e_{vw}^{(t+1)}\alpha_{vw}^{(t+1)} \tag{10}$$

其中，$\alpha_{vw}^{(t+1)}$ 表示智能体 v 给予智能体 $w\in N_v$ 的注意力权重值，计算式如式(11)所示：

$$\alpha_{vw}^{(t+1)}=\frac{\exp(v_a^{(t)}f_a^{(t)}([h_v^{(t)},h_w^{(t)},e_{vw}^{(t)}]))}{\sum_{w\in N_v}\exp(v_a^{(t)},f_a^{(t)}([h_v^{(t)},h_w^{(t)},e_{vw}^{(t)}]))} \tag{11}$$

其中，$f_a^{(t)}$ 为 FNN，$v_a^{(t)}\in n_\alpha$ 是 $\alpha_{vw}^{(t+1)}$ 的参数。

在每个时间步中，将智能体编码为隐藏状态可实现模型边值的动态变化，通过 FNN 编码实现，如式(12)所示：

$$\begin{aligned}h_v^{(0)}&=f_h(o_v)\\ e_{vw}^{(t+1)}&=f_e^{(t)}([h_v^{(t)},h_w^{(t)},e_{vw}^{(t)}])\end{aligned} \tag{12}$$

其中，o_i(节点特征)为每个智能体的可观察状态。

最后，通过式(13)计算出整个图的期望价值 V^π。

$$V^\pi=f_x(\hat{y})=f_x(\sum_{v\in N}f_x([h_v^{(0)},h_v^{(t)}])) \tag{13}$$

2.2.3 训练过程

为了研究 GDQN 模型在空中交通管制任务上的表现，将环境表示为连续的状态。单集训练过程如图 2 所示。

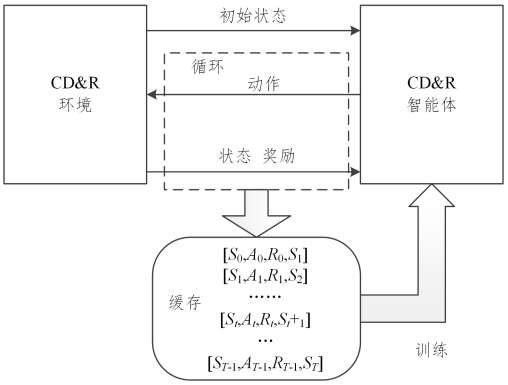

图 2 单集训练过程

Fig. 2 Single-set training process

在每个训练集开始时，将环境中初始化状态 s_0 发送给智能体。在每个时间步 t，智能体接收状态 s_t 并向环境发送动作 A_t。之后，环境根据输入动作 A_t 执行一步，并将下一个状态 S_{t+1} 和奖励 R_t 都发送给智能体。重复这个过程，直到 S_{t+1} 成为最终状态。S_t，A_t，R_t，S_{t+1} 被定义为元组并存储在内存中。智能体从存储器中选择数据，并根据 GDQN 算法进行自我训练。

1)航空器运动方程

定义每一架飞机为单独的智能体，因此每个智能体的状态可以由三维空间坐标值(x_v，y_v，z_v)、飞行速率(V_v)、航迹角 γ、航向角 ψ 组成，每个智能体的所属状态值如式(14)所示：

$$\begin{aligned}x_v(t+1)&=x_v(t)+V_v(t)\sin\psi_v(t)\Delta t\\ y_v(t+1)&=y_v(t)+V_v(t)\cos\psi_v(t)\Delta t\\ Z_v(t+1)&=Z_v(t)+V_v(t)\sin\gamma_v(t)\Delta t\\ V_v(t+1)&=V_v(t)+\Delta V_v\\ \psi_v(t+1)&=\psi_v(t)+\Delta\psi_v\\ \gamma_v(t+1)&=\gamma_v(t)+\Delta\gamma_v\end{aligned} \tag{14}$$

其中，ΔV_v 和 $\Delta\psi_v$ 分别是速度和轨迹变化；Δt 是模拟的步长。

2)状态空间

每个智能体要观察其当前状态，从环境中收集其他状态信息，并预测相对于其邻居 $w\in N_v$ 中的每一个最近接近点(Closest Point of Approach，CPA)，其定义为智能体 v 和 w 更近的点[14]。假设它们保持其当前轨迹和速度。智能体 v 和 w 之间的 CPA 几何用于创建如式(15)所示的边特征。

$$\boldsymbol{e}_{vw}=\begin{bmatrix}t_{\mathrm{CPA}_{vw}}/60\\ d_{\mathrm{CPA}_{vw}}/d_{\min}\\ \cos\theta_{vw}\\ \sin\theta_{vw}\\ \cos(\psi_v-\psi_w)\\ \sin(\psi_v-\psi_w)\end{bmatrix}^{\mathrm{T}} \tag{15}$$

其中，$t_{\mathrm{CPA}_{vw}}$ 和 $d_{\mathrm{CPA}_{vw}}$ 分别为到 CPA 点的时间以及 v 和 w 的智能体到 CPA 点的距离；$d_{\min}$ 为每个时间步长下每对智能体之间保持安全的间隔距离，此间隔是通过对分隔距离低于 $d_{\min}$ 的智能体对给予负奖励(即惩罚)来实现的；θ_{vw} 表示 CPA 从智能体 v 到智能体 w 的方位角，$\Delta\psi=\psi_v-\psi_w$ 为交角。

为了使每个智能体能直接从出发点飞行到目标点，它们之间必须进行合作。在实际飞行过程中飞行器的速度调节范围为正常速度的 20%，其幅度变化有限，并且在每个时间步长下每对智能体之间必须保持安全的间隔距离 $d_{\min}$。因此，两架飞机主要以调节航空器航向进行冲突解脱。在不改变飞行计划的情况下，若检测到潜在冲突，为不影响其他飞机的飞行计划以及避免多方协调出现错误，调整飞机进行航向偏转完成冲突解脱。飞机检测到潜在冲突，在起始点 P 调整航向避免冲突发生，因为调整航向后飞机偏离了原目的地，所以在飞行过程中飞机需时刻检测与目的地之间的偏转角是否符合偏转要求，当检测到该偏转角不会再次引起冲突时便立刻调整航向以最短路径到达各自目的地。图 3 为一个飞机可规避量化角度的示意图。

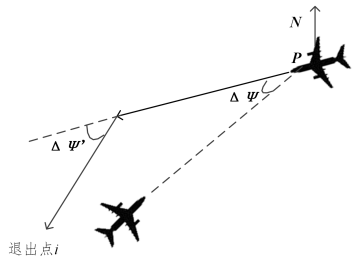

图 3 飞机规避角度示意图

Fig. 3 Schematic diagram of aircraft avoidance angle

3)动作空间

本文中的智能体动作主要包括速度变化和方向变化，动作空间 A_v 动作为 n 个航空器所选速度和方向的组合，因此动作空间模型如式(16)所示：

$$A_v=\{\vec{n}_{-\frac{\pi}{2}}+\vec{n}_{V_{\min}},\cdots,\vec{n}_\psi+\vec{n}_V\} \tag{16}$$

其中，$\vec{n}_\psi$ 和 $\vec{n}_V$ 分别是可能的轨迹和速度变化量，$\vec{n}_\psi\in\left(-\frac{\pi}{2},\frac{\pi}{2}\right)$，$\vec{n}_V\in(V_{\min},V_{\max})$，每个智能体的速度在每个时间步长都被削减到允许的范围内。

4)奖励函数

在采取动作之后，每个智能体生成一个单独的奖励 R_v，该奖励量化的动作是对于当前状态的。奖励函数由速度改变惩罚、航迹改变惩罚、冲突惩罚、预警惩罚、延迟惩罚共同组成。延迟惩罚指惩罚偏离最佳速度的偏差[15]，每当智能体改变航迹或速度时，惩罚函数分别给予负奖励。只要与环境中

存在的其他智能体在任意时刻的距离低于 $d_{\min}$ 就会给予相对较大的惩罚，并且对可能发生的分离损失也会受到惩罚。具体如式(17)所示：

$$\begin{aligned}
r_t' &= -\omega_d \left| \overline{v}_{\in v} \right| \\
r_V' &= -\omega_V L_{\Delta V_v != 0} \\
r_\psi' &= -\omega_\psi L_{\Delta\psi_1 != 0} \\
r_c' &= -\sum_{w \in N \setminus \{v\}} \omega_t L_{d_{vw} < d_{\min}} \\
r_a' &= -\sum_{w \in N \setminus \{v\}} \omega_a L_{d_{\mathrm{CPA}_{vw}} < d_{\min}} L_{t_{\mathrm{CPA}_{vw}} < 2\min}
\end{aligned} \tag{17}$$

其中，正参数 w_t，w_a，w_d，w_V 和 w_ψ 对每个项的相对重要性进行加权，从而确定所学习的策略函数的偏好。奖励由 6 个不同的项组成，其中如果满足条件 x，则指标函数 L_x 为 1，否则为 0。

奖励函数 R_v 计算式如式(18)所示：

$$R_v = \left| r_t' + r_V' + r_\psi' + r_c' + r_a' \right| \tag{18}$$

本模型的总体处理流程如图 4 所示。

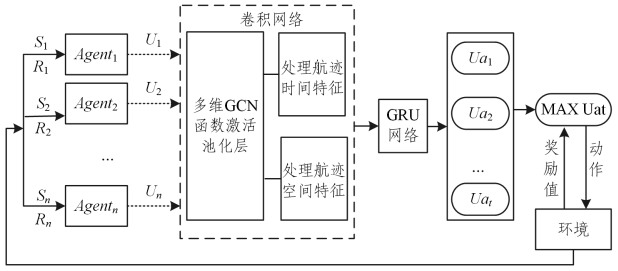

图 4　算法整体处理流程图

Fig. 4　Overall processing flowchart of the proposed algorithm

3　模型仿真实现

3.1　实验设置

首先，可以将管制员的策略表示为一个神经网络。actor 和 critic 共享网络的架构，但不共享参数权重。actor 和 critic 有 4 个输入：飞机状态和 3 个不同的邻接矩阵。每架飞机的状态首先通过两个前馈层进行投影，首先投影到 64 维，然后投影到 128 维，从而嵌入飞机状态。

不同的邻接矩阵根据与其他飞机的接近程度，以不同的方式给模型提供飞机距离的信息。前述嵌入通过使用 3 个不同的邻接矩阵，通过 1 个图卷积或图注意力层传播 3 次，从而产生 3 个 128 维向量。图 5 给出了两架飞机的探测区以及态势示意，不同的邻接矩阵实现 3 个平行的基于图的层，如式(19)所示。可以让飞机对周围环境有多个层次的了解，如果飞机距离更近，就可以提供更多的信息。第一个邻接矩阵将全局信息考虑在内，它为飞机提供了除自身外其他飞机的信息。除了对角线上的元素外，其余元素都为非零元素。第二个邻接矩阵只考虑到探测区域内的飞机，如图中的深色部分。第三个邻接矩阵只考虑了禁区内的飞机，如图中的浅色部分。

$$\begin{aligned}
\boldsymbol{Adjacency}1 &\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\
\boldsymbol{Adjacency}2 &\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\
\boldsymbol{Adjacency}3 &\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}
\end{aligned} \tag{19}$$

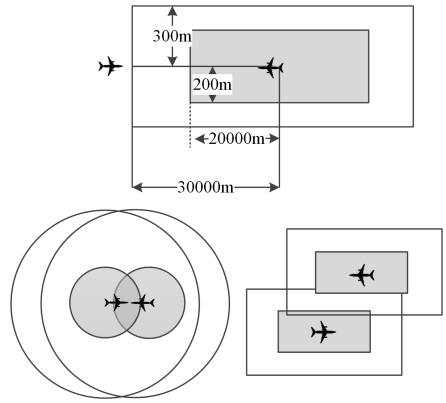

图 5　飞机态势示意图

Fig. 5　Aircraft situation diagram

将跳跃连接应用于 3 个图卷积层和状态的嵌入，各层之间的共享参数配置可能存在差异，每一层的输出节点值也各不相同，因此三层输出值最终需要异或，取最佳的结果。接着对这 4 个向量进行求和，并通过前馈层全链接(Fully Connected，FC)[16]传播到 64 个维度。最后再通过另一个线性层传播，如果考虑了动作头部值，则会产生一个 7 维向量，如果考虑了尾部值，则会产生一个 1 维向量。所有前馈层都有 ReLU 激活函数[17]，除了动作和尾部上的最后一层，其余都可以线性激活。将 4 个 128 维层相加后，应用另一个激活函数。模型总结如图 6 所示。

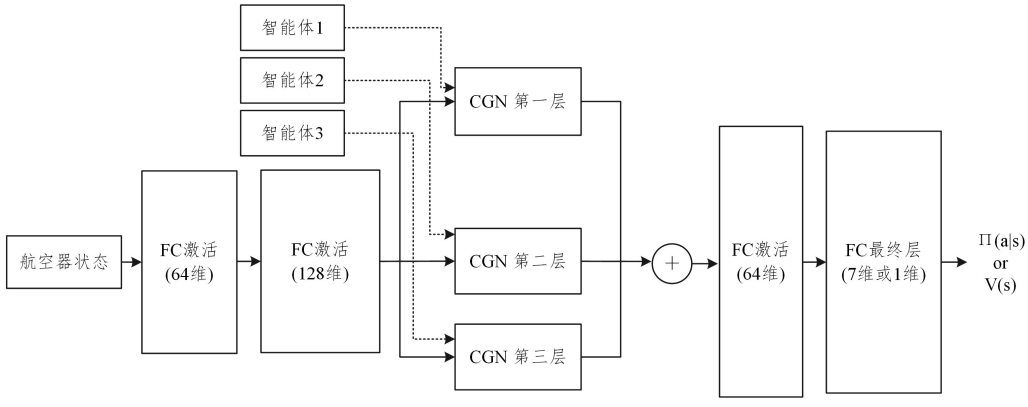

图 6　actor 和 critic 的共享架构

Fig. 6　Shared architecture of actor and critic

表 1 列出了 actor-critic 模型的网络参数。

表 1 actor-critic 模型的网络参数

Table 1 Parameters of actor-critic model

参数	属性
n_h	64 维
n_e	64 维
n_m	64 维
n_a	64 维
通信步长	3
F_a	128-ReLu→64-ReLu→32-ReLu→7-Softmax
F_v	128-ReLu→64-ReLu→32-ReLu→1-Softmax

表 1 中，神经网络的层数顺序用→表示，每层的命名为神经元数量-激活函数。注意，在 A_i 中，F_a 的最后一层有尽可能多的输出动作，其中每个输出对应于相关动作的概率。表 2 列出了仿真参数设置。

表 2 仿真参数设置

Table 2 Environment parameters

参数	属性
最小间隔距离 $d_{\min}$/NM	10
最小速度 $v_{\min_i}$/kt	500
最大速度 $v_{\max_i}$/kt	600
w_d	0.5
w_ψ	1
w_V	1
w_t	10
w_a	5

3.2 仿真与结果

仿真过程是通过连续随机生成一对飞机来模拟一个事件完成的，如果不采取行动，这对飞机肯定会坠毁。以这种方式初始化迫使飞机朝着彼此飞行，并允许学习到的策略对此采取行动。用于训练策略的算法采用第 2.2 节中的模型，根据配置参数，空域内的最大飞机数量被限制为 10 架。当飞机离开空域时，它们就会被移除。如果这样做不会导致最大飞机数量高于 10 架，则增加两架新飞机。在每个动作之后模拟环境 5 s，以允许从一个状态到下一个状态的清晰过渡。在创建 30 架飞机或达到固定数量的步骤后，该训练过程终止。训练达到 5000 集后即终止训练。图 7 给出了训练过程的结果，作为训练集的函数与环境的模拟。

正如预期的那样，每集的总奖励（所有智能体的总和）随着训练集数量的增加而增加，这意味着智能体可以通过从环境中收集经验来改进它们的策略。注意，这里最大可实现奖励大约为 2.4，对应于所有飞机全程按照计划的路线和速度行进，未偏离航线，并保持安全隔离。

冲突和冲突的总数（即预计 2 min 之内）在 3000 集后呈现下降趋势，最终达到 0，表明智能体通过与邻居节点的智能体沟通学习后，已经能预测和解决潜在的分离损失。

在管制员调配飞机冲突的场景下，下发的 ATC 指令对冲突的影响是较大的，每个训练集中管制员 ATC 指令的总数是逐渐减少的。智能体通过训练，每集 ATC 指令的数量大约是 112 条。在 2000 集之后，ATC 指令的数量减少到 20 个。对于每集飞行的额外距离也可以得出类似的结论。图 7 说明了智能体使用注意力机制进行通信并使得共享奖励最大化，能有效避免冲突。

在成功解决的冲突场景中，冲突成功解决率指在测试过程中，飞行冲突成功解决的场景数量占总体测试样本的比例，以百分数的形式表示。这是最直接衡量模型效果的一个指标，指标数值越大，表明模型的解脱效果越好。

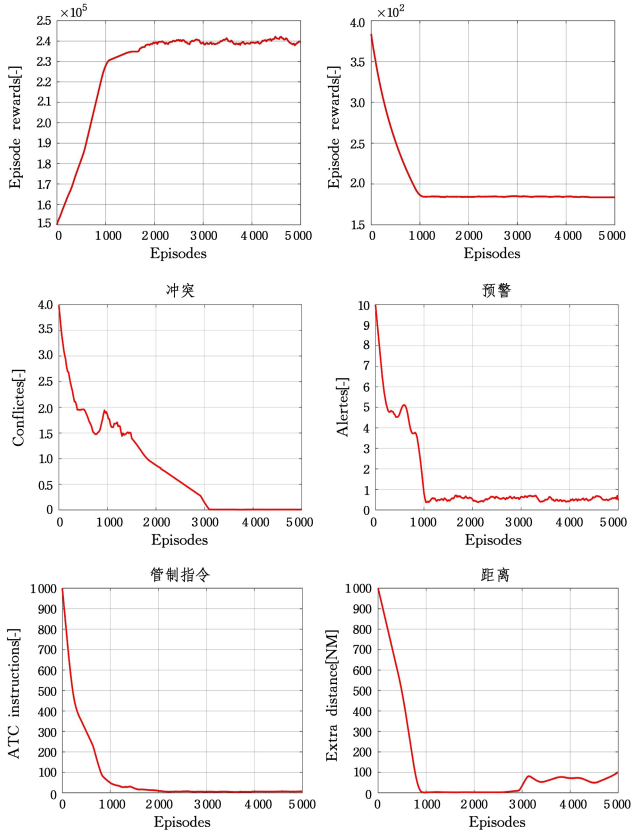

图 7 训练结果

Fig. 7 Training results

图 8 给出了冲突解决成功率与训练步数的正比例关系。同时，使用训练数据和测试数据对模型进行测试，评估模型的泛化能力。当训练步数充足时，模型的冲突解决率更高，可以利用保留的测试数据进行测试。若在某一奖赏值处成功解脱的场景数量越多，表明在冲突场景成功解脱时，落在该奖赏值处的概率最大。该测试结果表明模型的训练获得了一定的稳定性，若想进一步提高模型的稳定性和求解的成功率，则可以通过增加训练样本、增加训练步数等方式提高。

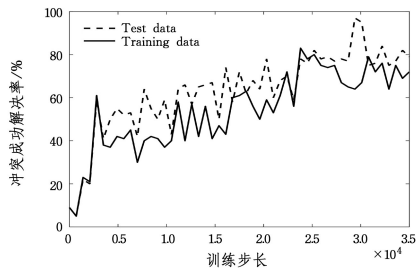

图 8 成功率随训练步数的变化曲线

Fig. 8 Curve of resolution change with the number of training steps

3.3 对比分析

深度强化学习的一大优势是离线训练，在线求解的方式能极大地提高求解问题的效率[18]。在众多的飞行冲突解决方法中，基于飞行冲突网络和遗传算法（Network and Genetic

Algorithm，Net_GA）的飞行冲突解决方法适用范围较广[19]，解决方式采用的是航向、速度、航向-速度混合 3 种冲突解脱方式，可对空域内多架飞机等复杂冲突场景中的飞行冲突进行快速消解。针对动作维度不能满足现实管制场景的需求，存在多维度状态、动作的单智能体短期冲突问题，多维深度确定型策略梯度算法（Deep Deterministic Policy Gradient，DDPG）模型利用强化学习中的深度确定型策略梯度算法建立了一个自动解决飞行冲突的模型[20]，实验表明该方法在不确定性条件下也能有效解决飞机间的冲突，证明了其在解决飞行冲突问题上的应用前景，但其并未研究多机飞行冲突场景。对于多机飞行冲突的场景，每个飞机都看作是合作型智能体，航空器智能体需要通过合作共同完成解脱任务，不存在牺牲某一架航空器智能体的利益而使得其他航空器智能体利益最大化的可能性。独立深度 Q 网络（Independent Deep Q-Network，IDQN）算法[21]也是用来解决多机飞行冲突的问题，它扩展了 DQN 算法解决多智能体问题，每架飞机是一个智能体被控制。对于有 n 架飞机的场景，需要 n 个与之对应的智能体。独立性意味着智能体之间没有耦合关系，没有通信行为。所有智能体的神经网络共享相同的参数，从环境中采样数据，独立更新信息网络。

本文设定场景大小为 200 km * 200 km * 6 km 的空域。每个场景包含 3 架、4 架或 5 架飞机的冲突，以及其他飞机，飞机数量随机分布在 20～100 架之间。设定训练集数为 5000 集的训练过程，并记录上述场景范围内的多机飞行冲突情况进行对比分析。利用冲突检测算法共检测出 6000 个多机飞行冲突场景，其中 3 机冲突场景占比 70%，4 机冲突场景占比 20%，5 机冲突场景占比 10%。将冲突场景分为训练场景和测试场景（分别为 5300 个和 700 个），每类冲突场景所占比例与上文相同。模型一次训练 500 步，训练后的模型表现出良好的稳定性和收敛性。成功解决情景所获得的奖励分布如图 9 所示。

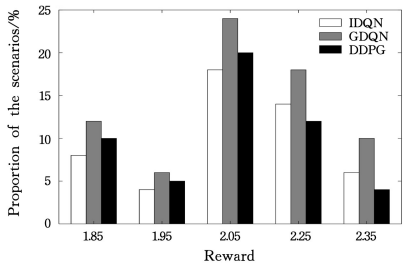

图 9　成功解决情景所获得的奖励分布

Fig. 9　Distribution of rewards for successful scenario resolution

在 700 个冲突场景中，本文的 GDQN 算法模型冲突解决成功率约为 97.14%，成功解决了 680 个冲突场景，其中 3 机冲突 500 次，4 机冲突 120 次，5 机冲突 60 次，共计 2280 架次。IDQN 冲突解决成功率约为 85.71%，成功解决了 600 个冲突场景。其中 3 机冲突 460 次，4 机冲突 110 次，5 机冲突 30 次，共计 1970 架次。Net_GA 冲突解决成功率为80.93%，成功解决 567 个场景，而多维度 DDPG 冲突解决成功率只有 62.84%，成功解决约 440 个场景，远低于其他 3 种算法，说明随着飞机架次 n 的数量增加，冲突解脱的成功率相差较大，相同机动次数的条件下，DDPG 在解决冲突的过程中更需要 ATC 的引导。统计 700 个冲突场景下模型收敛后冲突解决成功的数量，并计算对应的冲突解决平均计算时间、管制员发出的 ATC 指令数量比率以及平均冲突解决成功率，如表 3 所列。

表 3　解决冲突场景的结果

Table 3　Results of the conflict resolution scenarios

方法	平均计算时间/s	ATC/%	冲突/%
Net_GA	18.01	90	80.93
DDPG	15.80	98	62.84
IDQN	7.02	87	85.71
GDQN	**2.24**	**62**	**97.14**

通过在 700 个冲突场景中训练的结果表明，GDQN 算法和 IDQN 算法的求解计算时间明显短于前两种算法的求解时间，同时计算时间的分布也相对集中，IDQN 算法的冲突解决策略的平均计算时间为 7 s，而所提出的 GDQN 算法在求解中，冲突解决策略的计算时间分布在 2～3s，说明 GDQN 算法在求解时间上有很大的优势。这有助于提升冲突调配策略的求解效率，更好地适应实际管制运行情况，适合空域环境不断变化的动态冲突解决策略求解。

针对管制员发出的 ATC 指令，Net_GA 算法和 DDPG 在多机冲突解决场景下，更加依赖于管制员 ATC 指令，但是管制员所能关注和调配的飞机数量是有限的，而空域中少量飞机对构成飞行冲突有着重要的影响，如果能对这些飞机进行自适应互相学习，则能够快速、有效地消解飞行冲突。因此，在接收管制员 ATC 指令方面，IDQN 接收指令的占比为 87%，而 GDQN 接收发出指令的比率仅为 62%，说明飞机在遇到冲突预警后，通过基于图的自学习消息传递机制能自适应地避开冲突风险，降低与管制员的通话频次，从而可减轻管制员的工作负担。

结束语　本文提出了一种基于图卷积的多智能体深度强化学习，结合消息传递神经网络来最大限度地化解多机在某空域存在的冲突风险。结果表明，本文提出的图卷积深度强化学习模型通过多飞机相邻之间的互相通信自学习，利用强化学习的共享奖励功能，能最大程度地提高飞行效率，减少 ATC 指令，提前预报预警，有效规避冲突。此外，消息传递神经网络可以使飞机共享必要的信息，并选择性地相互自沟通调配，经过通信协议允许在采取最佳行动之前就达成共识，更加有利于多机共享飞行空域中的冲突化解。在未来的工作中，可将所提出的方法与基线方法进行比较，对行动者-批评者（actor-critic）模型和算法的参数进行微调，以进一步提高策略的性能，还可以实现与其他算法的融合，如多演员注意力评论家算法。最后，在本文研究中，航空器节点间建立连边考虑航空器从当前节点向下一节点转移时，只会选择当前节点四周的节点。下一步可将该本文模型扩展到 3D 领域，结合实际的管制规则与管制经验，从多方面考虑航空器冲突化解的优化目标，建立在任意时刻两架航空器的距离约束条件，在三维空间中设置航路威胁，增加冲突解决策略的多样性，再通过实践性研究，使算法应用更加合理。

参 考 文 献

[1] WANG Z, LI H, WANG J, et al. Deep reinforcement learning basedconflict detection and resolution in air traffic control[J]. IET Intelligent Transport Systems, 2019, 13(6): 133-142.

[2] LIU X, XIAO G. Flight Conflict Resolution and Trajectory Recovery Through Mixed Integer Nonlinear Programming Based on Speed and Heading Angle Change[J]. Transportation Research Record, 2024, 2678(4): 751-775.

[3] WEN H. Research on Flight Conflict Resolution based on Deep reinforcement Learning [D]. Chengdu: Sichuan University, 2021: 12-23.

[4] CAI M, WAN L J, GAO Z Z, et al. Conflict detection method based on K-Means spatial clustering in grid system [C]// China Association for Science and Technology, Ministry of Transport, Chinese Academy of Engineering, Hubei Provincial People's Government. Proceedings of World Transport Congress 2022 (WTC2022). Air traffic Control and Navigation College, Air Force Engineering University, 2022: 666-673.

[5] TONG L, YANG J, GAN X S, et al. More chaotic ant colony algorithm based on improved machine conflict free simulation [J]. Journal of system simulation, 2025, 5(1): 155-166.

[6] BRITTAIN M, WEI P. Autonomous separation assurance in an high density en-route sector: A deep multi-agent reinforcement learning approach[C]// Proceedings of Institute of Electrical and Electronics Engineers (IEEE) Intelligent Transportation Systems Conference. Piscataway, NJ: IEEE Press, 2019: 3256-3262.

[7] MOHAMMAD E, FARIBORZ H HAGHIIGHAT, et al. Transfer learning for occupancy-based HVAC control: A data-driven approach using unsupervised learning of occupancy profiles and deep reinforcement learning [J]. Energy & Buildings, 2023, 300(9): 356-362.

[8] VSWANI A, SHAZEER N. Attention is all you need[C]// The 31st Annual Conference on Neural Information Processing Systems (NeurIPS). Long Beach, CA, 2017: 5999-6009.

[9] LI S, EGOROV M, KOCHENDERFER J. Optimizing Collision Avoidance in Dense Airspace using Deep ReinforcementLearning [C]// The 13th USA/Europe Air Traffic Management Research and Development Seminar (ATM2019), Vienna, Austria, 2019: 45-49.

[10] XIE X F, SMITH S, BARLOW G. Coordinated look-ahead scheduling for real-time traffic signal control[C]// International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS). Valencia, Spain, 2012: 1271-1272.

[11] LIU Y P, SUI D, LIN Y D. Research on the learning behavior of Controller Agent based on Q Learning [J]. Journal of Harbin University of Commerce (Natural Science Edition), 2016, 32(6): 763-768.

[12] SILVER D, HUANG A. mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[13] BROWN N, SANDHOLM T. Superhuman AI for multiplayer poker[J]. Science, 2019, 365(6456): 885-890.

[14] SUI D, XU W, ZHANG K. Study on the resolution of multi-aircraft flight conflicts based on an IDQN[J]. Chinese Journal of Aeronautics, 2022, 35(2): 195-213.

[15] BI K X, WU M G, WEN X X, et al. Conflict resolution strategy based on flight conflict network and genetic algorithm [J]. Systems Engineering and Electronics, 2023, 45(5): 1429-1440.

[16] RASHID T, SAMVELYAN M. Monotonic value function factorisation for deep multi-agent reinforcement learning[C]// The 35th International Conference on Machine Learning (ICML). Stockholm, Sweden, 2018: 1335-1345.

[17] SUKHBAATAR S, SZLAM A, FERGUS R. Learning Multiagent Communication with Backpropagation[C]// The 30th Annual Conference on Neural Information Processing Systems (NeurIPS). Barcelona, Spain, 2016: 2244-2252.

[18] JIANG J, DUN C. Graph convolutional reinforcement learning [C]// The 8th International Conference on Learning Representations(ICLR). Addis Ababa Ethiopia, 2020: 1265-1273.

[19] JUNPENG Y, YIYU C. A practical reinforcement learning framework for automatic radar detection[J]. ZTE Communications, 2023, 21(3): 22-28.

[20] ZHANG Z, CUI P, ZHU W. Deep learning on graphs: A survey [C]// Proceedings of Institute of Electrical and Electronics Engineers (IEEE) Transactions on Knowledge and Data Engineering. Piscataway, NJ: IEEE Press, 2020.

[21] LI S, EGOROV M, KOCHENDERFER J. Optimizing collision avoidance in dense airspace using deep reinforcement learning [C]// The 13th USA/Europe Air Traffic Management Research and Development Seminar (ATM2019). Vienna, Austria, 2019: 45-49.

HUO Dan, born in 1990, master, lecturer. Her main research interests include air traffic control and collision prevention safety.

(责任编辑:喻藜)