基于深度逆强化学习的无人机实时动态避障航路规划算法

张俊驰，冼彦宁

（江汉大学 人工智能学院，武汉 430056）

摘　要：针对传统强化学习算法对奖励函数敏感性较高、将深度逆强化学习应用于无人机避障航路规划时面临的训练数据不足等挑战，提出一种结合专家演示与深度逆强化学习的无人机实时动态避障航路规划算法。首先，基于最大熵逆强化学习框架训练一个策略网络，从专家演示中学习安全高效的飞行策略，引入熵正则化构建由奖励网络与策略网络组成的对抗学习系统；其次，为了处理连续的路径点，在传统策略网络上结合长短期记忆网络（LSTM）层以增强连续路径点之间的关系建模能力；最后，构建一个模拟动态障碍物运动的环境，结合Informed-RRT*-DWA算法与特征感知的专家行为模拟方法，生成专家演示轨迹。实验表明，该算法在动态避障场景下能实时生成路径长度与专家路径相近的安全航路，相较于基于深度强化学习的D3QN算法而言，在低密度障碍物的环境下路径长度减少13.3%、成功率提升17%；在高密度环境下路径长度减少5.1%、成功率提升20%。

关键词：无人机；航路规划；深度学习；逆强化学习

DOI：10.11907/rjdk.251048　扫描二维码阅读全文：

中图分类号：G434　**文献标识码：**A　**文章编号：**1672-7800（XXXX）0XX-0001-09

Deep Inverse Reinforcement Learning Based Real-time Dynamic Obstacle Avoidance Path Planning Algorithm for UAVs

ZHANG Junchi, XIAN Yanning

（*School of Artificial Intelligence*, *Jianghan University*, *Wuhan* 430056, *China*）

Abstract: Aiming at the challenges of high sensitivity of traditional reinforcement learning algorithms to reward functions and insufficient training data when applying deep inverse reinforcement learning to UAV obstacle avoidance route planning, a real-time dynamic UAV obstacle avoidance route planning algorithm combining expert demonstrations and deep inverse reinforcement learning is proposed. Firstly, a policy network is trained based on the maximum entropy inverse reinforcement learning framework to learn safe and efficient flight strategies from expert demonstrations. Entropy regularization is introduced to construct an adversarial learning system consisting of a reward network and a policy network; Secondly, in order to handle continuous path points, a Long Short Term Memory (LSTM) layer is combined with traditional policy networks to enhance the modeling ability of relationships between continuous path points; Finally, construct an environment that simulates the motion of dynamic obstacles, and combine the Informed-RRT*-DWA algorithm with feature aware expert behavior simulation method to generate expert demonstration trajectories. Experiments have shown that this algorithm can generate safe routes with path lengths similar to expert paths in real-time in dynamic obstacle avoidance scenarios. Compared with the D3QN algorithm based on deep reinforcement learning, the path length is reduced by 13.3% and the success rate is increased by 17% in low-density obstacle environments; Reduce path length by 5.1% and increase success rate by 20% in high-density environments.

Key Words: UAV; path planning; deep learning; inverse reinforcement learning

收稿日期：2025-01-19

基金项目：国家自然科学基金青年基金项目（62106179）

作者简介：张骏驰（1990-），男，博士，江汉大学人工智能学院讲师、硕士生导师，研究方向为机器学习、自然语言处理、路径规划；冼彦宁（1999-），男，江汉大学人工智能学院硕士研究生，研究方向为自然语言处理、数据挖掘、路径规划。本文通讯作者：张俊驰。

0 引言

近年来，无人机已成为各领域不可或缺的一环，通过远程操控来完成侦察监视、物资运输等多种复杂任务，在现代军事冲突中具有实时态势感知能力和精确打击能力，可深入危险区域执行人类难以完成的高风险任务，为己方收集关键的情报，对目标实施精准打击[1-3]。然而，在执行任务过程中，无人机往往会面临来自敌方战机、飞鸟等动态障碍物的威胁，对飞行安全造成严峻威胁[4]。因此，如何确保无人机在复杂多变的战场环境中，安全高效地执行任务便成为当前的一个重要课题。

无人机路径规划算法主要分为基于网格、采样、优化和学习的方法[5]。以 A*算法、Dijkstra 算法为代表的基于网格的方法，将环境离散化为网格的方式去搜索最短路径，计算量较大且实时性较差[6-8]。快速扩展随机树及其变体等基于采样的方法，通过随机采样探索来搜索空间，以快速生成可行路径，但在复杂环境中易陷入局部最优且难以保证路径的平滑性和最优性[9-12]。遗传算法、粒子群优化算法等基于优化的方法，通过迭代优化寻优路径，但对初始解的质量较敏感[13-15]。

近年来，基于学习的方法可从大量数据中学习复杂的决策模式，使无人机在复杂的动态环境中自主学习并作出实时决策。然而，深度学习算法对数据的依赖性极强，需要大量高质量的样本数据进行训练，但在实际飞行场景获取充足、多样化的飞行数据往往成本高昂、耗时费力[16]。此外，强化学习算法对奖励函数的设计极为敏感，若奖励函数设计不合理，可能会导致算法收敛到局部最优解，甚至无法学习到有效的策略[17-21]。为了解决上述问题，本文提出基于深度逆强化学习的无人机航路规划算法，将策略网络与长短期记忆网络（Long Short Term Memory Network，LSTM）相结合，以增强连续路径点之间的关系建模能力，以期提升无人机的避障能力。

1 相关工作

研究者们积极探索各种基于机器学习的方法，包括监督学习、强化学习和深度学习等，开发更具鲁棒性、更高效的无人机路径规划算法，以克服传统方法在复杂动态环境中的局限性。Chen 等[22]结合图像处理与支持向量机，通过构建非线性分离面，先将路径规划问题转化为空间划分问题，再利用支持向量机寻找最优的分离面作为飞行路径，从而高效规划出安全、平滑的飞行路径。Zhao 等[23]采用自适应随机探索方法，结合强化学习思想实现了无人机在未知环境中的自主导航与避障，通过搜索机制引导无人机寻找合理路径，并在不同场景下的模拟实验中取得了良好效果。Yan 等[24]进一步改进 Q-learning 算法，使其适用于对抗性环境。

后期，深度学习为无人机路径规划带来了革命性的变化。Zhang 等[25]针对无人机作为灾后通信网络中的空中基站这一场景，在用户设备能源受限、地理环境复杂的情况下，提出一种基于安全深度 Q 网络的无人机轨迹优化算法，解决了受限马尔可夫决策过程在飞行时间内的最大化上行吞吐量问题，实现了用户设备的能量消耗与上行吞吐量之间的平衡。Guo 等[26]提出的分层深度 Q 网络算法分解任务，利用深度强化学习框架提升算法收敛性和有效性。Yan 等[27]依据全局态势信息，将深度强化学习框架与 D3QN 网络、ε-greedy 策略相结合，在考虑敌方雷达探测和导弹攻击对无人机生存概率的影响下，在动态环境下能高效地规划无人机路径。

然而，由于强化学习对奖励函数的高度敏感性，学习性能很大程度取决于奖励函数的设计，一旦奖励函数设计不合理就会影响学习效果，所以研究者们开始将目光转向逆强化学习算法[28]。You 等[29]结合强化学习与逆强化学习，基于道路几何信息构建了基于马尔可夫决策过程的自动驾驶车辆规划模型，通过学习专家驾驶员的行为实现了兼顾安全性、效率和舒适性的自动驾驶策略。Tung 等[30]基于逆强化学习框架，通过学习人类专家在拉斯维加斯真实路况下的驾驶数据训练了一个轨迹评分模型，通过轻量级安全过滤器，在密集城市的交通中可安全高效地规划、控制无人驾驶车辆的路径。Huang 等[31]通过假设人类驾驶行为由离散的驾驶意图驱动，结合最大熵逆强化学习框架从真实驾驶数据中推断出个性化的奖励函数，从而构建了能模拟人类决策机制的驾驶行为模型，以准确预测、模拟不同驾驶员的驾驶风格。

目前，基于深度逆强化学习的路径规划算法通常采用车辆行驶场景下的专家行为数据，原因为这类数据相较于在动态复杂的作战环境中无人机避障航路规划的专家数据数更容易获取，且后者相关研究尚不成熟。本文针对无人机在动态复杂战场环境下的实时航路规划问题，提出了一套完整的解决方案：

（1）设计了一种基于深度逆强化学习的无人机实时动态避障航路规划算法。具体为，在最大熵逆强化学习（Maximum Entropy Inverse Reinforcement Learning，MaxEnt IRL）框架上，构建了一个由奖励网络与策略网络组成的对抗学习系统。为了处理连续的路径点，在传统策略网络基础上结合了长短期记忆网络层，以增强连续路径点之间的关系建模能力。在训练过程中，算法交替优化奖励网络与策略网络，使奖励网络能准确区分专家轨迹与策略生成轨迹，并引导策略网络生成质量更高的航路。

（2）构建了一个模拟动态障碍物运动的环境。通过模拟敌方无人机、飞鸟等动态障碍物的运动行为，使实验环

境尽可能接近实际应用场景。在此基础上，提出了一种结合 Informed-RRT*-DWA 算法与特征感知的模拟专家行为方法，以生成高质量的专家演示轨迹[32]。

2　无人机动态避障问题建模

2.1　无人机成功避障的定义

路径规划中一个关键的成功评估指标是成功率 S，被定义为成功完成路径的百分比。其中，成功完成路径是指无人机在固定时间限制内到达目标点 (x_g, y_g)，且未与任何障碍物发生碰撞。

2.2　无人机避障算法评价指标

为了全面评估规划的路径性能，考虑了多个反映路径性能的评估标准，包括：①成功率 $S(P)$ 为路径完成的成功率；②路径长度 $L(P)$ 为路径中所有路段长度 d_i 的总和，即 $\sum_{i=1}^{n} d_i$，以衡量路径在最小化飞行距离方面的效率；③累计转角 $A(P)$ 的计算公式为 $\sum_{i=1}^{n} \theta_i$，以量化路径平滑程度，转角越大代表路径越曲折、越不平滑，将增加无人机飞行时的安全风险，加速机械磨损。

2.3　无人机避障算法相关约束

2.3.1　运动学约束

无人机在飞行过程中，转弯角度必须严格控制在规定范围，以确保飞行轨迹的平滑性与安全性。

$$|\theta_i| < \theta_{\max}, \forall i \in [1, n] \tag{1}$$

式中：θ_i 表示路径中每一步的转角；$\theta_{\max}$ 为最大允许转角。

2.3.2　运动学约束

为了确保导航安全、避免碰撞，必须躲避环境中的障碍物。障碍物分为静态障碍物 O_i^{static} 和动态障碍物 $O_i^{dynamic}$，对于前者必须满足式(2)的最小安全距离约束；对于后者，无人机必须始终保持安全距离，并根据障碍物位置的变化进行调整，如式(3)所示。

$$d(P, O_i^{static}) \geqslant d_{safe}, \forall i \tag{2}$$

$$d(P(t), O_i^{dynamic}) \geqslant d_{safe}, \forall i, \forall t \tag{3}$$

式中：$d(P, O_i^{static})$ 表示无人机路径 P 与静态障碍物之间的距离；d_{safe} 为避免碰撞所需的最小安全距离；$d(P(t), O_i^{dynamic})$ 表示 t 时刻无人机路径与动态障碍物的距离。

2.3.3　时间约束

在动态变化的环境中，无人机系统必须实时响应，并在严格的时间约束下完成飞行任务，准确到达预设终点。

$$T_{comp}(P) < T_{\max} \tag{4}$$

式中：$T_{comp}(P)$ 表示计算路径 P 所需的时间；$T_{\max}$ 为最大允许计算时间。

3　本文方法

3.1　逆强化学习原理

逆强化学习(Inverse Reinforcement Learning，IRL)是一种从专家示范中学习的机器学习方法，相较于传统强化学习直接给出奖励函数的不同在于：逆强化学习旨在通过观察专家行为，反向推断出背后的隐含目标，也就是奖励函数。逆强化学习的核心是假设专家行为是为了最大化某个未知的奖励函数，目标是找到一个奖励函数解释智能体的行为。这种方法在许多复杂任务中具有广泛的应用前景，例如自动驾驶、机器人控制等，尤其适用于动态环境下无人机避障算法等难以直接定义奖励函数的场景。

逆强化学习的实现过程通常是一个迭代优化过程，包括最大熵逆强化学习、学徒学习这两种经典方法。首先，算法根据某种初始化方式(随机初始化或基于先验信息)构建一个初始奖励函数；其次，在该奖励函数下利用强化学习或其他近似方法生成一个策略；再次，计算生成策略的行为(状态分布或特征分布等)与专家示范行为之间的差异，并据此更新奖励函数；最后，不断重复直到生成的策略与专家行为在一定准则下足够接近。

3.2　方法概述

图 1 为基于深度逆强化学习的无人机航路规划算法的实现过程。首先，构建了包含各种接近实际动态障碍物与静态障碍物的障碍区；其次，结合 Informed-RRT*-DWA 算法与传感器特征感知数据生成专家轨迹。

算法的核心部分以最大熵逆强化学习理论为基础，构建奖励网络与策略网络的对抗性训练框架。其中，奖励网络用于建模环境中状态与动作的价值函数，通过学习专家轨迹中的隐含意图逐步接近专家行为；策略网络通过生成当前策略下的最优轨迹，尽量接近奖励网络定义的高价值区域，从而完成任务。

训练过程中，本文采用动态学习率调度机制、熵正则化策略提升训练的稳定性与泛化能力。同时，为了应对复杂环境中的不确定性，在数据生成阶段引入随机采样与动态步长调整策略，使训练样本覆盖更多可能性，增强网络对环境变化的适应性。此外，为了缓解了模型过拟合和梯度爆炸问题，还引入梯度裁剪和正则化约束，在训练结束后加载训练好的策略网络进行路径规划。

3.3　模拟环境构建与专家数据生成方法

在深度逆强化学习框架中，为了生成高质量的专家数据及解决实际专家训练数据不足的问题，设计了一个尽可能接近战场上动态障碍物的模拟环境。同时，采用结合 Informed-RRT*-DWA 与特征感知的专家行为模拟方法，通过整合静态、动态障碍物的特征信息，构建了包含实时状态观测与动作决策的数据，为后续的深度逆强化学习算法提供数据支撑。

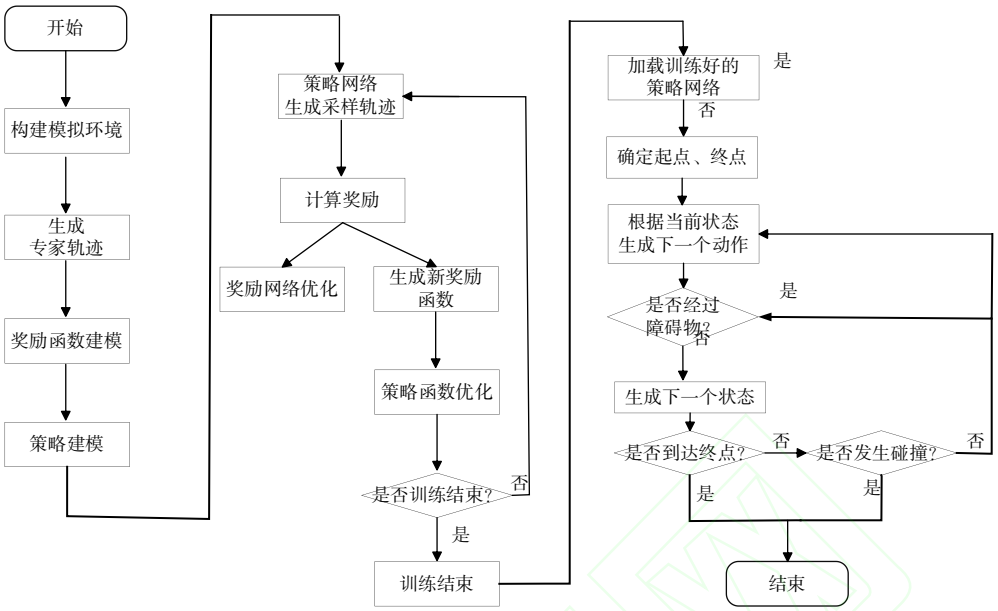

Fig. 1 Overall process of the algorithm

图1 算法总流程

3.3.1 *动态障碍物*

本文针对现代战场环境中多类型、高动态的威胁，设计了一种动态障碍物生成方法，旨在生成能模拟真实环境中各类不同速度、加速度、运动轨迹的动态障碍物，例如高速突袭的敌方战机、不规则移动方向的飞鸟等。

(1)敌方战机。敌方战机在空战中的机动性极强，运动轨迹呈现出高度的不确定性，通常会非匀速地从不同方向发起攻击。为了准确预测敌机的飞行意图，假设敌机的初始速度、加速度等关键参数可通过机载传感器实时获取，并基于实时数据建立敌机运动模型。考虑到敌机在飞行过程中受到各种因素影响(风力、气流等)，其运动方向可视为传感器当前检测到的方向。因此，本文基于经典运动学公式，结合随机扰动因素构建了一个既包含确定性运动特征，又能模拟环境不确定性的综合运动模型。

根据经典的运动学理论，物体位置可通过初始位置、速度和加速度进行表示。具体为，假设敌机初始速度为 $p_{enemy}(0)$，其方向由初始角度 θ_0 给定，则初始速度在 x、y 方向的分量为：

$$\begin{cases} v_{x_{enemy}}(0) = v_{enemy}(0)\cos\theta_0 \\ v_{y_{enemy}}(0) = v_{enemy}(0)\sin\theta_0 \end{cases} \tag{5}$$

敌机的加速度 $a_{enemy}(t)$ 也可分解为 x、y 两个方向。

$$\begin{cases} a_{x_{enemy}}(t) = a_{enemy}(t)\cos\theta(t) \\ a_{y_{enemy}}(t) = a_{enemy}(t)\sin\theta(t) \end{cases} \tag{6}$$

式中：$\theta(t)$ 为敌机在时间 t 的飞行方向。

为了模拟真实环境中的随机性，还引入了随机扰动 $\delta a(t)$，此时敌机在 x、y 方向的位置分量为：

$$\begin{cases} x_{enemy}(t) = x_{enemy}(0) + v_{enemy}(0)\cos\theta_0 \cdot t + \dfrac{1}{2}a_{enemy}(t)\cos\theta(t)\cdot t^2 + \displaystyle\int_0^t \delta a(t)\cos\varphi(t)dt' \\ y_{enemy}(t) = y_{enemy}(0) + v_{enemy}(0)\sin\theta_0 \cdot t + \dfrac{1}{2}a_{enemy}(t)\sin\theta(t)\cdot t^2 + \displaystyle\int_0^t \delta a(t)\sin\varphi(t)dt' \end{cases} \tag{7}$$

式中：$\delta a(t)$ 为由于敌机机动、外部扰动和传感器不确定性引起的加速度扰动；$\varphi(t)$ 为随机扰动的方向角。

(2)飞鸟障碍物。飞鸟的运动轨迹呈现出高度的不规则性和随机性，受风力、气流、障碍物、捕食者及自身生理状态等因素影响。为了更逼真地模拟这一复杂行为，引入了随机过程模型，运动轨迹可通过随机微分方程(Stochastic Differential Equation，SDE)表示。该运动方程通常包含多个随机变量，用于描述位置、速度、加速度等随时间的变化。

$$p_{bird}(t) = v_{bird}(0) + \int_0^t v_{bird}(t')dt' + \int_0^t f_{env}(t')dt' \tag{8}$$

式中：$p_{bird}(t)$ 表示飞鸟在 t 时刻的位置；$v_{bird}(t')$ 表示飞鸟在 t' 时刻的位置；$f_{env}(t')$ 表示外部环境的扰动加速度，作为随机变量引入。

值得注意的是，飞鸟的速度不仅受自身动力驱动，还受到环境干扰与随机波动影响，$v_{bird}(t')$ 可表示为：

$$v_{bird}(t) = v_{bird}(0) + \int_0^t a_{bird}(t')dt' + \int_0^t \sigma_{env}(t')dW(t') \tag{9}$$

式中：$a_{bird}(t')$ 为飞鸟自身的加速度；$\sigma_{env}(t')$ 为环境扰动的强度因子；$W(t)$ 为标准布朗运动，表示随机扰动。

由于飞鸟加速度 $a_{bird}(t')$ 通常表现为目标驱动(例如趋向某处或避开某物)与随机波动的结合，如式(10)所示。

为此，飞鸟障碍物的实时位置如式(11)所示。结合式(5)、式(6)即可得到飞鸟障碍物在x、y方向上的位置分量。

$$a_{bird}(t)=a_{goal}(t)+a_{avoid}(t)+\sigma_{acc}(t)dW(t) \tag{10}$$

$$p_{bird}(t)=p_{bird}(0)+v_{bird}(0)t+\frac{1}{2}a_{bird}(0)t^2+\int_0^t f_{env}(t')dt'+\int_0^t \sigma_{env}(t')dW(t') \tag{11}$$

式中：$a_{goal}(t)$为飞鸟趋向静态目标点(如陆地、灯塔)的加速度；$a_{avoid}(t)$为飞鸟避开障碍物或捕食者的加速度；$\sigma_{acc}(t)$为随机加速度的强度因子。

3.3.2　专家轨迹生成

本文基于Informed-RRT*-DWA算法的路径规划与数据生成方法，结合动态环境下的传感器特点进行改进，将静态、动态障碍物的特征信息融入算法，以实时响应随环境变化的专家轨迹。同时，为了实现无人机对环境的全面感知，还综合利用了多种传感器数据。

假设，当前点坐标为$P(p_x,p_y)$，无人机从起点$S(s_x,s_y)$行驶至终点$E(e_x,e_y)$。首先，摄像头数据c_i以布尔数组的形式表示，感知周围是否存在障碍物；其次，通过地理信息系统处理高精度地图数据，为无人机提供当前距离静态环境(高山、建筑物、地形等)的距离信息s_i；最后，通过全球定位系统提供无人机在全局坐标系中的精确位置，并进行定位和导航。此外，通过激光雷达实时感知动态障碍物距离d_i、速度v_i、加速度a_i，为无人机提供即时的环境变化信息。

$$X_{in1}=\begin{bmatrix} c_1,c_2,..,c_{120} \\ s_1,s_2,..,s_{120} \\ d_1,d_2,..,d_{120} \\ v_1,v_2,..,v_{120} \\ a_1,a_2,..,a_{120} \\ long,lat,R_x,R_y,0,...,0 \end{bmatrix} \tag{12}$$

$$R_x=\frac{p_x-s_x}{e_x-s_x} \tag{13}$$

$$R_y=\frac{p_y-s_y}{e_y-s_y} \tag{14}$$

式中：R_x、R_y分别表示当前点在x、y方向相较于起点、终点的归一化位置。

随后，调用Informed-RRT*-DWA算法规划从起点s_{start}到目标点s_{goal}的最优路径，若规划失败(因障碍物阻挡而无法前进或与障碍物发生碰撞)，将重新生成障碍物分布并再次规划，直至生成足够多的可用数据为止。

3.4　深度逆强化学习模型

图2为基于深度逆强化学习的无人机航路规划模型，基于对抗训练的机制，通过协同优化奖励网络(Reward-Net)与策略网络(PolicyNet)实现逆强化学习目标。奖励网络在模型中扮演着判别器的角色，主要评估输入的行为轨迹，区分来源是专家演示还是策略网络所生成。具体而言，奖励网络通过深度神经网络对专家行为背后的隐含奖励函数进行参数化建模，将状态和动作对作为输入，输出一个标量值表示状态—动作对的奖励，优化目标是最大化专家演示的奖励值，最小化策略网络生成轨迹的奖励值。如此，奖励网络能学习到专家行为中蕴含的奖励模式，为优化策略网络提供明确的指导信号。综上，奖励网络通过不断调整自身参数，以准确评估不同行为轨迹的质量，推动策略网络生成与专家行为更相似的轨迹。

策略网络通过不断优化生成轨迹，使其在奖励网络的评估下获得尽可能高的奖励分数，以逼近专家行为。具体地，策略网络以当前状态为输入，通过一系列非线性变换生成与之对应的动作来构建完整的轨迹。训练过程中，策略网络与奖励网络构成一个对抗学习框架，目的是生成能欺骗奖励网络的轨迹，使其误以为这些轨迹由专家生成。

奖励网络通过不断提升自身区分专家演示与策略网络生成轨迹的能力，促使策略网络不断学习专家行为的特征分布，并生成与专家行为高度相似的轨迹。为了实现该目标，策略网络的优化目标包含两个方面：①最小化一个

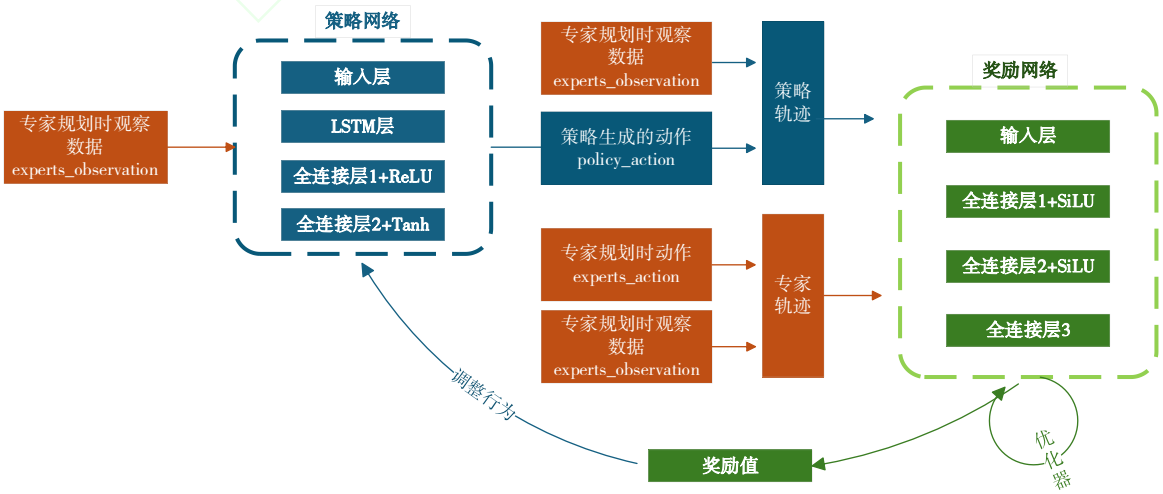

Fig. 2　UAV route planning generative adversarial network model

图2　无人机航路规划生成对抗网络模型

与奖励值相关的损失函数去最大化生成轨迹的累积奖励；②引入熵正则化项增加生成轨迹的多样性。熵正则化项有助于防止策略网络过早陷入局部最优解，可提升模型在面对复杂环境时产生多种应对策略的能力。通过不断迭代优化，策略网络可逐渐学习到与专家行为相似轨迹的策略，从而自动控制复杂任务。此外，本文模型还采用梯度裁剪、L2正则化和动态学习率调整等技术，增强模型的训练效率与鲁棒性。

3.4.1 奖励函数模型

奖励函数模型采用深度神经网络对专家行为的隐式奖励函数进行参数化建模，核心是构建一个奖励网络（RewardNet），实现从状态—动作对(s,a)到标量奖励值$r(s,a)$的映射。奖励网络的输入是状态和动作的拼接向量$x=[s,a]$，通过全连接层序列实现特征提取与非线性变换。具体地，模型结构由3层全连接层构成：第一层、第二层有256个隐藏单元，采用SiLU激活函数$\mathrm{SiLU}(x)=x\cdot\sigma(x)$（$\sigma(x)$为Sigmoid函数以增强非线性表达能力）；最后一层输出一个标量值$r(s,a)$，表示当前状态与动作的奖励。

在训练过程中，奖励网络的目标是通过对抗优化最大化专家行为的奖励，最小化由策略网络生成行为的奖励。首先，给定专家轨迹$\tau_E=\{(s_i,a_i)\}_{i=1}^{N}$和生成轨迹$\tau_G=\{(s'_i,a_i')\}_{i=1}^{N}$；其次，分别计算两种轨迹的累积奖励。

$$\hat{r}(\tau)=\frac{\exp(R(\tau)}{\sum_{\tau'}\exp(R(\tau'))} \tag{15}$$

式（15）是为了构造一个近似的最大熵分布，使奖励网络从概率的角度区分不同轨迹的优劣。同时，奖励网络的优化目标是通过最大化专家轨迹的归一化概率，并最小化生成轨迹的归一化概率。具体的损失函数为：

$$L_r=-\log\hat{r}(\tau_E)+\lambda\|\theta_r\|^2 \tag{16}$$

式中：$-\log\hat{r}(\tau_E)$用于提升专家轨迹得分的对数损失项；$\lambda\|\theta_r\|^2$为L2的正则化项，防止过拟合并增强模型泛化能力；在优化过程中通过反向传播计算梯度，并应用梯度裁剪确保参数更新的稳定性。

3.4.2 策略网络

策略模型基于深度神经网络，引入长短期记忆网络增强连续路径点之间的关系建模能力，使模型适应动态环境下复杂的策略学习任务。模型输入是一段连续的状态序列$\{s_t\}_{t=1}^{T}$，输出为当前路径点的动作a_t。首先，将观察数据$observations\in R^{N\times T\times state_dim}$（$N$为批次大小，$T$为时间步数，state_dim为单个观察状态的维度）作为时间序列的状态$\{s_t\}_{t=1}^{T}$，输入长短期记忆网络模块；其次，通过长短期记忆网络处理连续路径点，以提取依赖关系特征h_T；最后，通过全连接层生成动作值a_t。

策略网络的目标是最大化由奖励网络评估的轨迹累计奖励，引入熵正则化保证动作生成的多样性。具体的损失函数为：

$$L_{policy}=-E_{\tau\sim\pi}[R(\tau)+\alpha H(\pi)] \tag{17}$$

式中：$R(\tau)$为轨迹的累计奖励，$H(\pi)$为策略的熵，α为熵系数。

4 实验结果与分析

4.1 实验设置与训练数据

为了全面评估算法在复杂多变的军事作战环境中的导航性能，构建了一个100×100的模拟环境，随机生成并动态配置了多种类型的障碍物，以模拟真实战场的复杂多变性。首先，随机确定了动态障碍物的初始位置，根据预设规则随机选取动态障碍物的类型；其次，利用上述方法设定多样的运动轨迹，使其在模拟环境中呈现出不规则的运动状态；最后，基于随机分布、形状各异的静态障碍物模拟复杂的地形地貌，以全面考验算法在各种突发情况、复杂地形时的鲁棒性，为算法实际应用提供保障。相关参数如表1所示。

Table 1 Experimental parameters

表1 实验参数

参数	参数值
深度学习框架	Pytorch
GPU	NVIDIA RTX 3090
Python版本	3.8
初始学习率	1e-3
正则化参数	1e-3
熵系数	0.015
熵系数衰减率	0.995
成功率	无人机在10 s内到达目标点，且未与任何障碍物发生碰撞

为了便于评估本文算法的优势，本文构建了一个多样化、具有代表性的数据集。在100×100像素的地图上利用Informed-RRT*-DWA与特征感知的专家行为模拟方法生成了65 536万个路径样本，每个样本包含随机生成的3~8个动态障碍物和2~15个静态障碍物，为模型提供了丰富的训练数据。

4.2 比较算法

专家路径：Informed-RRT*-DWA算法是一种基于改进的Informed-RRT*与动态窗口方法相结合的无人机路径规划算法，可在未知或动态变化的环境中进行路径规划，能在机器人运动过程中实时感知环境变化并快速调整路径，以适应新的环境信息[33]。D3QN算法是一种结合双Q学习与对偶架构的深度强化学习算法，通过分离动作选择、评价状态值与优势值，以提升Q值估计的准确性和训练稳定性，适用于复杂动态环境中的路径规划问题[27]。

本文在低密度、高密度的障碍物环境中分别进行100次实验，通过成功率、路径长度和平均转弯角度等指标，比较本文方法与Informed-RRT*-DWA、D3QN在具有动态障碍物的环境中的表现，实验结果如表2所示。

图3为本文实验的可视化案例（彩图扫OSID可见，下同）。实线为本文方法路径，虚线为Informed-RRT*-DWA算法路径（专家路径），点线为D3QN算法的路径，黑色不

Table 2　Algorithm comparison results

表 2　算法比较结果

环境	低密度障碍物环境（20%–45%）			高密度障碍物环境（50%–70%）		
评价指标	成功率	路径长度	平均转角	成功率	路径长度	平均转角
Informed-RRT*–DWA（专家行为来源）	87	143.04	18.91	81	168.71	30.6
D3QN	76	167.62	30.07	67	179.65	34.4
本文方法	93	145.25	20.95	87	170.5	29.57

规则物体为静态障碍物，大半径的深黄色圆圈为敌方战机，浅黄色圆环为可能与敌机发生碰撞的范围，红色小圆圈为飞鸟，浅红色圆环为可能与飞鸟发生碰撞的范围。由此可见，所提方法所生成的路径质量接近专家行为，明显优于深度强化学习方法 D3QN。

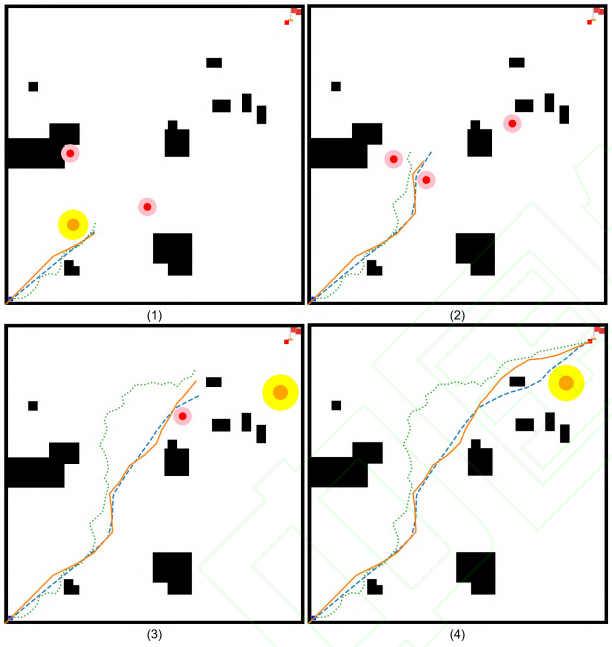

Fig. 3　Comparison of visualization results of different algorithms

图 3　不同算法的可视化结果比较

由表 2 数据可知，所提方法在路径长度和转弯角度上，相较于专家行为十分接近；在低、高密度障碍物环境下，路径长度相较于 D3QN 分别短 13.3%、5.1%。此外，本文方法在具有多种不同运动规律、不同大小的动态障碍物的环境下，随着障碍物密度增大仍能保持较高的成功率。

4.3　算法拓展性分析

图 4 为大小为 256×256 环境下算法的可视化结果。由此可见，基于 100×100 地图训练数据构建的策略模型在更大规模的 256×256 地图场景中泛化能力良好，规划路径与原始训练场景相近，原因为实验的训练数据处理的是无人机与周围障碍物的相对位置关系，与终点的距离也是相对位置，因此实验效果对绝对地图尺寸的依赖性较低。

4.4　算法训练性能分析

本文详细分析了基于逆强化学习的无人机轨迹规划

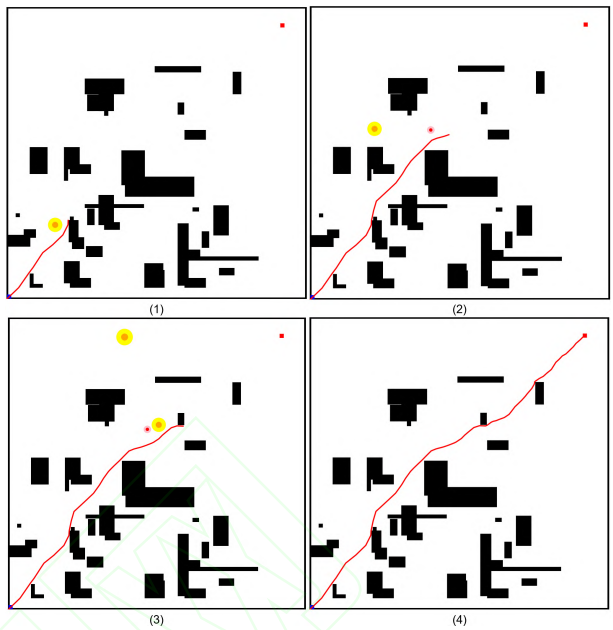

Fig. 4　Visualization results of algorithms in an environment with a size of 256x256

图 4　大小为 256x256 环境下算法的可视化结果

算法的训练过程，展示了不同训练轮次时专家演示奖励值（Expert Reward）与采样奖励值（Sampled Reward）的变化趋势，如图 5 所示。由此可见，训练初期两种奖励值差别较大，随着训练进行两条曲线逐渐重合，在第 40 轮左右稳定在一个数值；专家演示奖励值始终略高于采样奖励值，表明算法能较好地学习专家演示的特征，但仍存在优化空间。

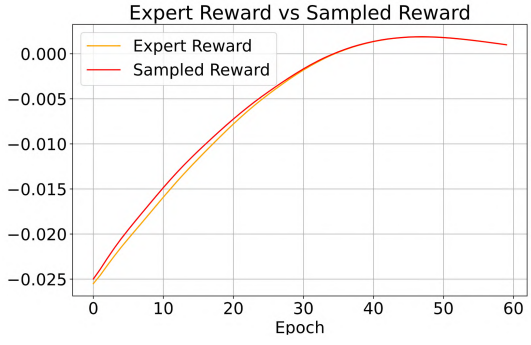

Fig. 5　Comparison between expert demonstration reward value and sampling reward value

图 5　专家演示奖励值与采样奖励值比较

图 6 为策略损失（Policy Loss）的变化过程，训练初期策略损失较高，呈现出明显的下降趋势，反映了算法在初始阶段会对策略进行大幅调整的特点；在 60 轮后趋于稳定，维持在较低水平，说明算法的策略网络已逐渐收敛到了一个相对稳定的状态。

图 7 为奖励损失（Reward Loss）的训练过程，奖励损失从较高数值开始持续下降，在整个训练过程中表现出平滑的衰减特性，表明奖励函数的学习过程有效，算法能逐步提取、理解专家演示中潜在的奖励机制；训练后期（60 轮以

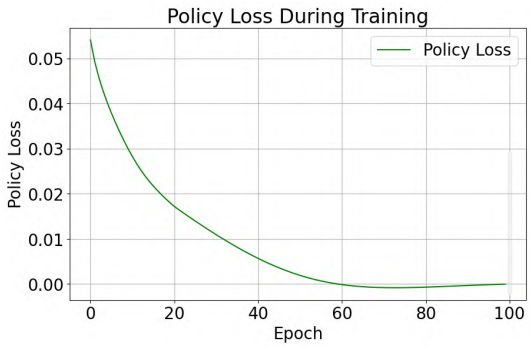

Fig. 6 Strategy loss training curve

图6 策略损失训练曲线

后),奖励损失的下降速度明显放缓,最终稳定在0.693左右,表明奖励函数的学习已基本完成,证实了基于逆强化学习的无人机轨迹规划算法具有良好的收敛性和学习效果,在训练过程中能有效提取专家演示中的特征,可通过策略网络和奖励函数优化路径,以逐步模仿专家行为。

Fig. 7 Reward loss training curve

图7 奖励损失训练曲线

4.4 消融实验

本文设计了一系列消融实验,以深入探讨熵正则化对策略多样性的影响。实验环境包含5-15个随机位置、随机类型的障碍物,障碍物密度控制在30%-50%。每个实验重复20次,通过区间表示收敛轮次(见表3),如果在100轮内成功收敛则计算成功率,如果成功率高于60%则对平均长度求平均值。

首先,尝试移除熵正则化项,即观察不包含熵项的目标函数的学习效果;其次,调整熵系数大小,以探究不同程度的熵正则化对策略多样性的影响。由表3可知,当熵系数为0、0.005时会限制策略随机性,生成的轨迹集中于局部最优路径,导致奖励网络更容易区分专家轨迹与生成轨迹,使奖励损失快速收敛。然而,由于探索行为不足,策略网络生成的轨迹分布缺乏多样性,奖励网络提供的梯度信号不足以有效引导策略优化,导致策略损失长时间未能收敛。这样的训练动态不平衡会使策略陷入局部最优,降低生成路径的多样性和鲁棒性,难以应对复杂环境中的变化和扰动,成功率极低。当熵系数较大时,模型也将难以收敛。为此,熵系数的取值为0.01-0.015时最佳。

为了进一步验证最大熵强化学习对算法性能的贡献程度,对奖励正则化机制进行消融实验。由表3可知,当移除奖励正则化时,未正则化的奖励可能使策略对极端奖励值过于敏感或忽略较小的奖励信号,导致奖励网络不稳定,使策略损失的波动性更大,最终使策略质量下降。当奖励正则化强度为1e-4时,生成的路径较短但缺乏稳定性、成功率较低;当奖励正则化强度为1e-2时,会出现欠拟合问题。因此,奖励正则化强度设定为1e-3时最合理。

Table 3 Ablation experiment results

表3 消融实验结果

环境	熵系数(奖励正则化强度为1e-3)					奖励正则化强度(熵系数为0.015)			
参数	0	0.005	0.01	0.015	0.025	无(0)	1e-4	1e-3	1e-2
Reward Loss收敛轮次	40-50	45-65	60-80	60-80	100+	震荡	50-65	60-80	欠拟合
Policy Loss收敛轮次	100+	100+	60-80	60-70	100+	震荡不收敛	100+	60-70	-
航路规划成功率/%	20	15	75	85	80	-	60	85	-
平均长度	-	-	154.02	150.72	156.4	-	148.98	150.72	-

5 结语

本文提出了适用于无人机航路规划的对抗性网络,在最大熵逆强化学习框架下实现了实时动态避障航路规划算法。针对无人机航路规划领域缺乏大量专家数据的问题,通过模拟现实敌机、飞鸟等动态障碍物创建模拟环境,并结合InformedRRT*-DWA与实时传感数据方法生成专家轨迹。实验表明,所提方法能生成与专家路径相似的航路,相较于深度强化学习算法D3QN而言缩短了路径长度,减少了平均转弯角度,提升了算法在动态复杂场景中规划路径能力。

然而,当前算法并未考虑更高维度、更复杂的环境,生成的模拟专家数据与实际专家数据仍存在一定差异。未来,将重点探索更先进的技术以生成高质量的无人机导航路径,以进一步提升效率、安全性和适应性;将研究范围扩展至更复杂、更动态、更高维度的环境,以应对实际应用中的挑战。

参考文献:

[1] ZHANG T, YU X J, SHUAI H, et al. UAV track planning algorithm[J]. Journal of Naval Aviation University, 2022, 37(2): 172-178, 208.
张婷,鱼小军,帅欢,等. 无人机航迹规划算法[J]. 海军航空大学学报, 2022, 37(2): 172-178, 208.

[2] ZHU X. Analysis of military application of UAV swarm technology[C]//

3rd International Conference on Unmanned Systems, 2020: 1200-1204.

[3] FAHLSTROM P G, GLEASON T J, SADRAEY M H. Introduction to UAV systems[M]. Hoboken : John Wiley & Sons, 2022.

[4] UDEANU G, DOBRESCU A, OLTEAN M. Unmanned aerial vehicle in military operations [J]. Scientific Research and Education in the Air Force, 2016, 18(1): 199-206.

[5] WANG Q, LIU M W, REN J W, et al. Overview of common algorithms for UAV path planning[J]. Journal of Jilin University(Information Science Edition),2019,37(1):58-67.
王琼,刘美万,任伟建,等. 无人机航迹规划常用算法综述[J]. 吉林大学学报(信息科学版),2019,37(1):58-67.

[6] HART P E, NILSSON N J, RAPHAEL B. A formal basis for the heuristic determination of minimum cost paths[J]. IEEE Transactions on Systems Science and Cybernetics, 1968, 4(2): 100-107.

[7] DIJKSTRA E W. A note on two problems in connexion with graphs[J]. Numerische Mathematik, 1959,1:269-271.

[8] MEDEIROS F L L, DA S J D S. A Dijkstra algorithm for fixed-wing UAV motion planning based on terrain elevation[C]// 20th Brazilian Symposium on Artificial Intelligence, 2010: 213-222.

[9] LAVALLE S. Rapidly-exploring random trees: a new tool for path planning [EB/OL]. https://www. semanticscholar. org/paper/Rapidly-exploring-random-trees-% 3A-a-new-tool-for-LaValle/d967d9550f831a8b3f5cb00f8835a4c866da60ad.

[10] NOREEN I, KHAN A, HABIB Z. A comparison of RRT, RRT* and RRT*-smart path planning algorithms[J]. International Journal of Computer Science and Network Security, 2016, 16(10): 20-27.

[11] GAMMELL J D, SRINIVASA S S, BARFOOT T D. Informed RRT*: optimal sampling-based path planning focused via direct sampling of an admissible ellipsoidal heuristic[C]// International Conference on Intelligent Robots and Systems, 2014: 2997-3004.

[12] YAO K W, ZHOU F, LI N, et al. Improved informed-RRT* based path planning algorithm[J]. Software Guide,2024,23(7):80-86.
姚凯文,周锋,李楠,等. 基于改进 Informed-RRT*的路径规划算法[J]. 软件导刊,2024,23(7):80-86.

[13] SONMEZ A, KOCYIGIT E, KUGU E. Optimal path planning for UAVs using genetic algorithm[C]// International Conference on Unmanned Aircraft Systems, 2015: 50-55.

[14] ROBERGE V, TARBOUCHI M, LABONTE G. Comparison of parallel genetic algorithm and particle swarm optimization for real-time UAV path planning[J]. IEEE Transactions on Industrial Informatics, 2012, 9(1): 132-141.

[15] KANG H I, LEE B, KIM K. Path planning algorithm using the particle swarm optimization and the improved Dijkstra algorithm[C]// IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2008: 1002-1004.

[16] GAUTAM S A, VERMA N. Path planning for unmanned aerial vehicle based on genetic algorithm & artificial neural network in 3D[C]// International Conference on Data Mining and Intelligent Computing, 2014: 1-5.

[17] CUI Z, WWANG Y. UAV path planning based on multi-layer reinforcement learning technique[J]. IEEE Access, 2021, 9: 59486-59497.

[18] JING Z M, LIU H J, ZHOU Y L. Research on path planning of mobile robot with improved SARSA algorithm [J]. Software Guide, 2024, 23(12):119-124.
井征淼,刘宏杰,周永录. 改进SARSA算法的移动机器人路径规划研究[J]. 软件导刊,2024,23(12):119-124.

[19] HE L, AOUF N, SONG B. Explainable deep reinforcement learning for UAV autonomous path planning[J]. Aerospace Science and Technology, 2021, 118: 107052.

[20] WU J, FENG J, HE J Z, et al. Review of UAV track planning algorithms considering mission scenarios[J]. Aeronautical Computing Technique,2024,54(5):130-134.
吴建,冯君,何佶哲,等. 考虑任务场景的无人机航迹规划算法综述[J]. 航空计算技术,2024,54(5):130-134.

[21] MA C Q, XIE W, SUN W J. Research on reinforcement learning technology: a review[J]. Command Control & Simulation,2018,40(6):68-72.
马骋乾,谢伟,孙伟杰. 强化学习研究综述[J]. 指挥控制与仿真,2018,40(6):68-72.

[22] CHEN Y, ZU W, FAN G, et al. Unmanned aircraft vehicle path planning based on SVM algorithm[C]// Proceedings of the First International Conference on Cognitive Systems and Information Processing , 2014: 705-714.

[23] ZHAO Y, ZHENG Z, ZHANG X, et al. Q learning algorithm based UAV path learning and obstacle avoidance approach[C]// 36th Chinese Control Conference, 2017: 3397-3402.

[24] YAN C, XIANG X. A path planning algorithm for UAV based on improved Q-learning[C]// 2nd International Conference on Robotics and Automation Sciences, 2018: 1-5.

[25] ZHANG T, LEI J, LIU Y, et al. Trajectory optimization for UAV emergency communication with limited user equipment energy: a safe-DQN approach[J]. IEEE Transactions on Green Communications and Networking, 2021, 5(3): 1236-1247.

[26] GUO T, JIANG N, LI B, et al. UAV navigation in high dynamic environments: a deep reinforcement learning approach[J]. Chinese Journal of Aeronautics, 2021, 34(2): 479-489.

[27] YAN C, XIANG X, WANG C. Towards real-time path planning through deep reinforcement learning for a UAV in dynamic environments [J]. Journal of Intelligent & Robotic Systems, 2020, 98: 297-309.

[28] ADAMS S, CODY T, BELING P A. A survey of inverse reinforcement learning[J]. Artificial Intelligence Review, 2022, 55(6): 4307-4346.

[29] YOU C, LU J, FILEV D, et al. Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning [J]. Robotics and Autonomous Systems, 2019, 114: 1-18.

[30] TUNG P M, HOWINGTON F, CHU T S, et al. Driving in real life with inverse reinforcement learning [DB/OL]. https://arxiv. org/abs/2206. 03004.

[31] HUANG Z, WU J, LYU C. Driving behavior modeling using naturalistic human driving data with inverse reinforcement learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(8): 10239-10251.

[32] WU T, ZHANG Z, JING F, et al. A dynamic path planning method for UAVs based on improved informed-RRT* fused dynamic windows[J]. Drones, 2024, 8(10): 539.

(责任编辑:刘嘉文)