arXiv:2301.01861v1 [cs.RO] 5 Jan 2023

基于强化学习的空中交通冲突消解

Denis Osipychev波音研究与技术 阿拉巴马州亨茨维尔
denis.osipychev@boeing.com

Dragos Margineantu波音研究与技术 华盛顿州西雅图
dragos.margineantu@boeing.com

Girish Chowdhary伊利诺伊大学厄巴纳-香槟分校 girishc@illinois.com

摘要

保持充分间隔，即通过适当的间隔距离使飞机远离危险，是确保无人驾驶航空器在拥挤空域中安全运行的一项关键技术。本文着重研究两架飞机水平间隔的自动化，并将障碍物规避问题表述为一个二维替代优化任务。根据我们的设计，代理任务采用了更为保守的设定，以确保在主域中执行该方案。

我们使用强化学习（RL）优化规避策略，并对动力学、交互过程和决策进行建模。通过递归采样由此得到的策略和代理状态转移，系统将规避策略转化为完整的避让轨迹。随后，求解器通过机器人操作系统（ROS）接口，将该轨迹发布为一组航路点，供飞机跟随。

所提出的系统能够生成快速且可实现的避让轨迹，并满足安全要求。我们在高保真仿真和全尺寸飞机演示中完成了对该系统的评估。此外，本文总结了大量集成工作，这些工作使基于强化学习的系统得以在现实环境中演示。

1 引言

自动化冲突解脱对于无人驾驶航空器（UAV）和空中交通管制（ATC）的安全运行至关重要 [1, 2]。为确保共享空域的安全，美国联邦航空管理局（FAA）在以下概念中规定了无人驾驶航空器的行为要求 [3]:

- 探测与避让（DAA）——在规划和执行飞行的各阶段探测并避让危险；
- 保持充分间隔——使飞机与危险源之间至少保持适当的间隔距离。

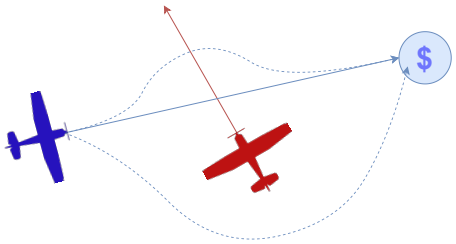
The 监管机构要求所有飞行器保持最小安全距离，并提供垂直 l
or h水平间隔。最常见的方法是垂直间隔——为不同的航班分配不同的飞行 t

本研究由美国国防高级研究计划局（美国国防高级研究计划局）资助。文中表达的观点、意见和/或研究结果均属于作者，不应被解读为代表美国国防部或美国政府的官方观点或政策。

第36届神经信息处理系统大会（NeurIPS 2022）。

高度。这种方法对于交通拥堵严重、能量水平有限的飞机，或反应时间较长的情况，可能效果不佳。水平间隔更为复杂，要求动作之间保持同步，这通常通过空中交通管制的指引，或根据飞机类型或位置规定具体行为的一系列规则来实现。

在这项工作中，我们聚焦于水平间隔，并针对当前方法无法解决的情况，开发了一种快速响应的解决方案：在没有程序间隔的非管制空域中实现水平间隔。



(a) 将两架飞机之间的水平间隔建模为二维类汽车障碍物规避问题。



(b) 单发涡轮螺旋桨飞机塞斯纳208“大篷车”——我们的演示平台。

图1：通过生成一条在飞机之间保持安全距离的避让轨迹，可以化解空中交通冲突。

本研究将冲突消解问题建模为二维动态路径规划问题。该优化是一个数据驱动、基于仿真的序贯决策过程，采用深度强化学习（RL）。这种建模方式能够处理问题中的随机性，而这一点常被基于能量的方法和基于图的轨迹方法所忽视 [4, 5]。尽管自动驾驶汽车领域在冲突消解和碰撞规避方面已有大量研究，本研究仍独树一帜，专门聚焦于航空应用 [6–9]。与汽车应用不同，该方案的解空间仅受飞机动力学限制，因此搜索空间更大。

如图 1a 所示，目标是改变受控飞机（智能体）的冲突航向，与另一架飞机（入侵者）保持安全距离，并在安全后返回原始航线。本研究采用以下强假设，未来研究或可放宽这些假设。首先，入侵者仅限于一架沿被动轨迹飞行的飞机——它不会改变意图，也不会对智能体的任何动作作出反应。其次，智能体和入侵者的状态均完全可观测，并假定能够完全掌握有限的一组参数。这组有限参数足以描述马尔可夫决策过程（MDP）系统的马尔可夫状态。

本研究提出使用针对代理任务求解的预计算避让策略，以缩短响应时间并提供运行时解决方案。代理问题的设计更加保守，以确保该解决方案在实际任务中可行。我们的系统输出一条完整的避让轨迹，以一组供飞机遵循的航路点表示。这种格式在航空航天业中更受青睐，也更容易集成到现有自动驾驶仪功能中。与控制副翼、方向舵和油门的端到端策略相比，我们的方法能够提供一种可视化、易于解释的解决方案，可在应用前进行验证。

所提出的系统旨在针对状态转移中的不确定性，提供快速且可行的避让轨迹，并满足安全要求。我们在高保真仿真中的实际任务以及如图 1b 所示、由塞斯纳“大篷车”208 执行的全尺寸飞机演示中，评估了强化学习冲突避让系统（RLCAS）的效率。

2 方法

2.1 技术方法

在这项工作中，避让机动的修正动作由预先计算的解决方案——强化学习（RL）策略模型生成。RL 是一种序贯策略优化方法，需要与环境进行持续且数据丰富的交互 [10]。为了提供所需数量的交互，我们在代理环境而非原始任务上训练策略。由波音开发的代理环境是一个轻量级 Python 环境，并集成了 OpenAI GYM 框架 [11]。

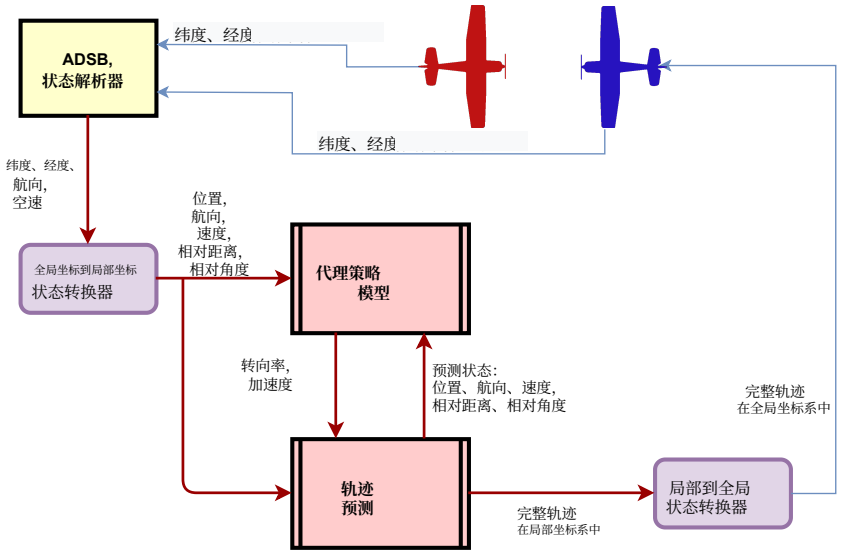


图 2：所提出的基于模型的冲突消解器系统图。

学习到的模型通过在代理环境中提供连续控制指令来降低风险。这些指令利用代理环境转换为几何轨迹。代理策略 $\hat{}\pi$ 是针对代理任务的数据驱动端到端强化学习策略，具有相似的状态转移动力学 T 和交战规则 R。在假设状态转移和条件具有相似性的前提下，将代理策略应用于原始任务。该代理是对实际任务更保守的近似，可确保所得解在实际系统中可行。

$$\pi_{\text{True}} = \pi(R_{\text{True}}, T_{\text{True}}) \tag{1}$$

$$\hat{\pi} = \pi^*(\hat{R}, \hat{T}) \tag{2}$$

$$\pi_{\text{True}} \approx \hat{\pi}, \text{ if } [R_{\text{True}}, T_{\text{True}}] \in [\hat{R} + \epsilon, \hat{T} + \epsilon] \tag{3}$$

2.2 代理环境

参数	代理	C208（真实）	单位
X	-10000 .. 10000	..	m
Y	-10000 .. 10000	..	m
航向	-180 .. 180	-180 .. 180	deg
空速	50 .. 100	31 .. 100	m/s
纵向加速度	-0.5 .. 0.5	-0.8 .. 0.5	m/s^2
偏航角速度	-3 .. 3	-5 .. 5	度/秒

表 1：代理与真实动态参数及其范围。

我们的代理环境是一个简化的二维避障问题，旨在模拟实际任务（空中交通冲突解脱）。飞机控制不同于二维类汽车动力学，且没有传统的制动器、加速度或转向。控制效果取决于飞机的高度、空速等。针对所选飞机的转弯率和纵向加速度所选取的参数较为保守，并确保在其整个飞行包线内均适用。所选的纵向加速度允许飞机在目标飞行高度层内略微偏离高度。

所选转向率可确保飞机能够执行所需的转弯半径。表1概述了仿真飞行器主要动力学参数的差异。研究采用该特定机体的高保真仿真来评估代理策略的性能。指令动作下的详细飞行器响应见附录图8、图9。为便于强化学习训练，所有观测值和控制值均归一化至 [-1..1]。

该环境仿真了两架飞机在20x20千米区域内的运动。受控智能体必须绕过入侵者，保持最小水平间隔，并在仿真结束前重新并入原始航线上的下一个安全航路点。训练仿真的详细参数见附录表2。代理为附录表3中更详细展示的交互提供稀疏奖励反馈。

仿真更新步遵循OpenAI GYM约定，同时触发以下行为：将指令转换为输入值、向飞行器设置输入、推进飞行器动力学，然后评估冲突。详细算法见附录算法1。

代理中的所有飞行器均采用简化的动力学模型：

- 用于转弯动力学的Dubins车辆模型
- 无质量运动学模型

为便于训练，我们为入侵车辆创建了一个简化的基于PID的航路点控制器。起初，我们不预期入侵车辆会改变方向或速度。该航路点控制器主要用于辅助训练，并在训练场景中生成冲突。不过，改变入侵车辆的意图是未来的工作计划。

2.3 强化学习训练

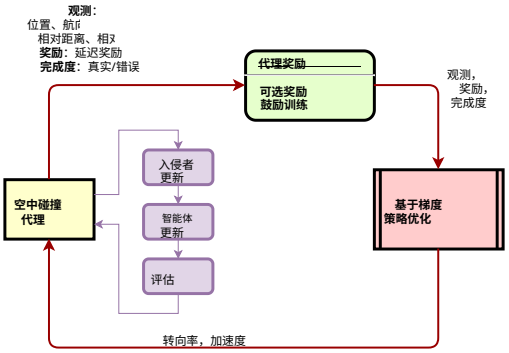
训练封装器允许覆盖某些环境方法和参数，并对环境进行一些重要改动，从而大幅提高策略训练的效率。该封装器包括：

1. 评分，
2. 初始化 - 飞行器出现冲突轨迹的概率很高，
3. 重新打包并归一化观测。

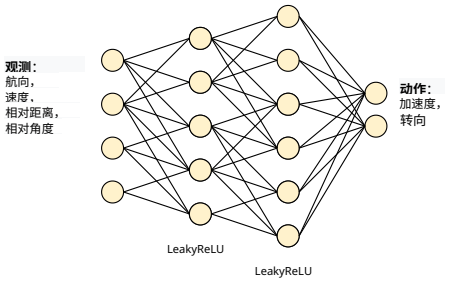
为确保智能体能够泛化该问题，我们重新组织观测，仅关注相对位置，而非绝对位置。此外，我们将数值归一化到 [-1 .. 1]。智能体使用的重新打包后的观测：

航向 空速 到目标的距离 相对于目标的跟踪角	入侵者航向 入侵者空速 与入侵者的距离 相对于入侵者的跟踪角

为更好地覆盖构型空间，我们修改了初始化过程。初始位置从一个虚拟圆中采样，两个飞行器均朝圆心移动。这确保它们的路径相交且飞行器同步。因此，智能体和入侵者出现冲突轨迹的概率更高。



（a）用于
通过迭代采样来优化代理策略
并改进动作。



（b）用于策略模型的神经网络函数逼近由 2 个隐藏层组成，每层有 256 个神经元。

图 3：强化学习策略。

2.4 强化学习策略智能体

强化学习策略智能体通过与仿真交互来学习任务，并使用随机梯度下降（SGD）优化迭代更新策略模型的参数。我们使用图 3b 所示的多层感知机来近似策略。

为了解决作为马尔可夫决策过程（MDP）系统的优化问题，我们将其重构为马尔可夫状态 s、状态转移 $T(s'|s,a)$ 和转移奖励 $R(s'|s,a)$。系统的状态（包括智能体和入侵者）完全可观测，并假设具备完全信息，足以描述 MDP 系统的马尔可夫状态。智能体的状态描述如下

$$s = \{v_a, \psi_a, v_i, \psi_i, \beta_i, d_i, \beta_g, d_g\}$$

其中 v_a - 智能体速度， ψ_a - 智能体航向， v_i - 入侵者速度， ψ_i - 入侵者航向， β_i - 至入侵者的角度， d_i - 与入侵者的距离， β_g - 至目标的角度， d_g - 至目标的距离。

The 优化目标是找到最优策略 $\pi^*(s)$，即一组状态-动作映射， t
max使期望奖励 $V(s)$ [10] 最大化。

$$\pi(s) = P(a|s) \tag{4}$$

$$\pi^*(s) = \arg\max_{\pi} V^{\pi}(s) \tag{5}$$

$$= \arg\max_{a} \left(R(s,a) + \gamma T(s'|s,a)V(s')\right) \tag{6}$$

状态价值是沿轨迹累积的期望未来奖励，由贝尔曼函数定义如下：

$$V(s) = \mathbb{E}[R|s,\pi] \tag{7}$$

$$= \sum_{s'} T(s'|s,a)\left(R(s'|s,a) + \gamma(V(s'))\right) \tag{8}$$

$$= R(s'|s,a) + \gamma \sum_{s'} T(s'|s,a)V^{\pi}(s') \tag{9}$$

$$V^*(s) = \max_a \left(R(s,a) + \gamma \sum_{s'} T(s'|s,a)V^*(s')\right) \tag{10}$$

强化学习策略模型基于行动者-评论家架构，有助于提高训练的稳定性 [10]。行动者 θ 和评论家 w 网络的基于随机梯度下降的更新：

$$\delta = R_{t+1} + \gamma \hat{V}(s_{t+1}, w) - \hat{V}(s_t, w) \tag{11}$$

$$w \leftarrow w + \alpha\delta\nabla\hat{V}(s,w) \tag{12}$$

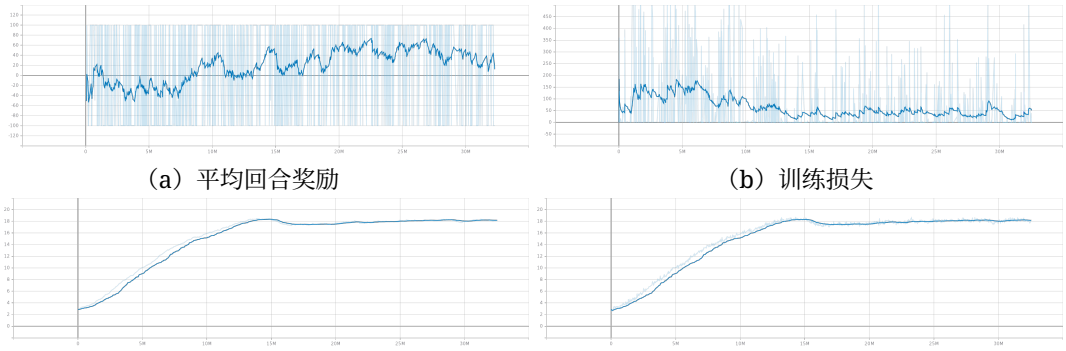
$$\theta \leftarrow \theta + \alpha\delta\nabla \ln \pi(a|s,\theta) \tag{13}$$

强化学习智能体的核心功能采用了 Stable Baselines 库。它是 OpenAI Baselines [12] 的一个广受认可的分支。在探索策略和更新步骤中，本研究使用了近端策略优化（PPO）算法，该算法已成为连续动作智能体领域的最先进方法 [13]。

3 集成

3.1 训练结果

为了让 RLCAS 系统在运行时快速高效，我们预先计算所有解，并将其存储为神经网络模型。为了创建这个解库，神经网络需要经过充分训练，而这是计算开销最大的过程。训练在轻量级 Python 代理环境中进行，该环境将飞行器的动力学简化为二维几何问题。我们最佳的策略是通过运行仿真10^8 步（约 10^6 个回合，或相当于三年的仿真时间）得到的。



（a）平均回合奖励　（b）训练损失

（c）熵损失　(d) 负对数概率

图 4：强化学习智能体训练结果。

该策略在代理任务上的平均得分收敛至 $+75.0$ 。生成的策略存储在一个文件中，其中包含模型的所有参数。我们的 ROS-Agent Runner 将加载并运行此文件。它是一种集成工具，包含 ROS 接口，用于与仿真和硬件通信。

3.2 评估与演示器

将代理策略应用于原始任务需要进行全面调查。由于车辆动力学存在不匹配，直接将策略输出连接到车辆控制器并不可靠，会导致振荡，并加剧系统不稳定性。强化学习策略本身会放大这一影响，因为成功的策略并不能保证动作平滑且一致 [14, 15]。此外，动力学和内部控制器的差异会导致最终轨迹中误差不断累积。

将几何轨迹输入现有车辆控制器，可以弥补这一缺陷。该轨迹表示一组航路点及其估计到达时间。这样，车辆控制器便能补偿状态转移中可能存在的偏差。

我们使用不同的演示平台评估所得策略，以证明该策略可以安全地采用。本研究提出了以下演示系统:

- 采用低保真动力学的 Gazebo 仿真，
- 波音专有的高保真仿真，
- 真实飞机演示系统。

3.3 ROS 集成

CAS 系统集成通过机器人操作系统（ROS）接口 [16] 实现。这使得能够统一连接高保真仿真和实体演示器的接口。ROS-Agent

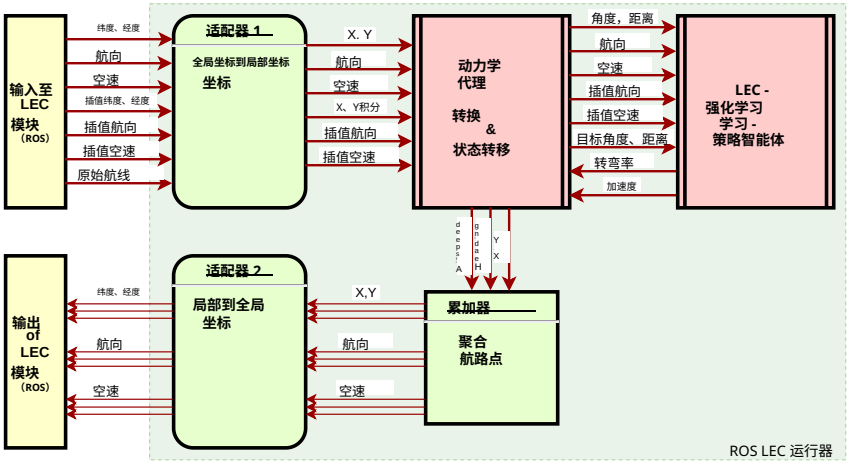


图 5：ROS-LEC（ROS-Agent）运行器的系统图——通过 ROS接口将学习型组件（LEC）集成到物理演示器中。

（ROS-LEC）运行器如图 5 所示，它聚合来自不同领域的数据，并为系统提供重要实用工具。其工作设计如下：

- 接收并累积有关本机状态、入侵机状态、交通警报和 GPS-SRS变换数据的 ROS 消息，
- 将自动相关监视广播和 GPS-NovAtel定位数据转换到局部坐标系，
- 从原始航线中提取目标位置，
- 将观测重新封装为特定于智能体的输入格式，
- 迭代运行智能体以获取纠正动作，
- 迭代运行代理环境以接收状态转移，
- 生成纠正轨迹并检查其是否合理，
- 将轨迹从局部坐标系转换为全球经纬度航路点，
- 将轨迹作为 ROS 消息发布。

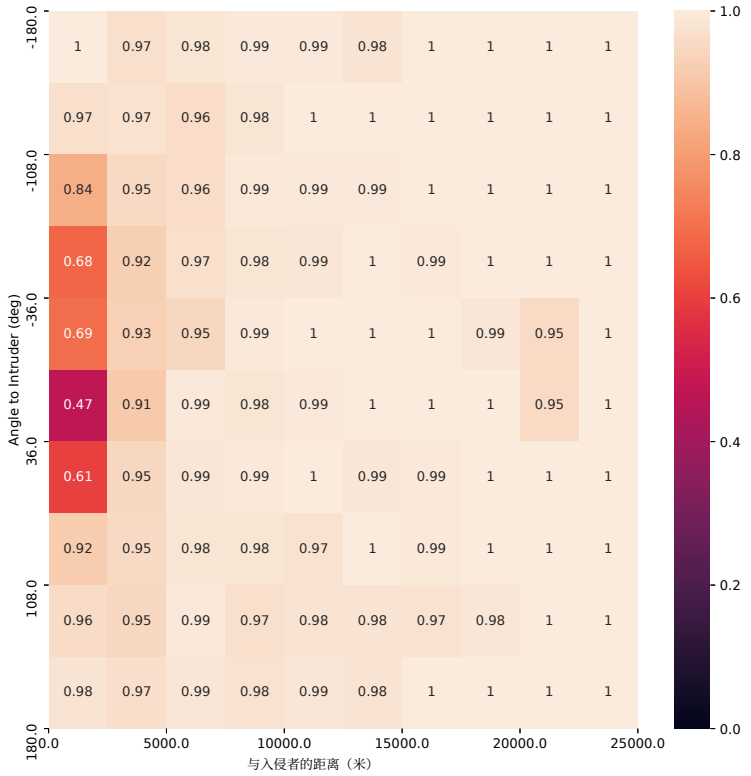
收到外部请求后，运行器会生成一条避让轨迹，并将其作为 ROS 消息发布。该轨迹总共包含 20 个航路点。最后一个航路点取自原始航线，其余 19 个航路点由策略生成。这样可以通过唯一索引关联各航路点，同时保留原始航线的索引。

这些航路点间隔 20 秒，可提供 400 秒的规划时域。由于航路点之间的时间步长较大，策略和代理必须经过 20 次评估才能生成单个航路点。系统生成完整轨迹的总响应时间低于 60 毫秒。

该架构支持通过外部运行时保障监视器进行闭环修正。监视器会跟踪累积的状态转移误差，并请求更新避让计划，或拒绝该操作并切换到备用模式。如需重新规划，可以再次调用运行器。

4 结果

我们将评估结果分为几个部分，以评估整个系统的性能。第一部分是统计评估，我们通过批量仿真计算成功的冲突消解事件总数。第二部分是轨迹性能分析，用于评估控制器遵循代理策略所引入的数值误差。



Fig图 6：统计结果展示了初始条件对系统性能的影响 e.
In 在大多数失败案例中，相对角度以及与入侵者的距离表明冲突不可避免 .

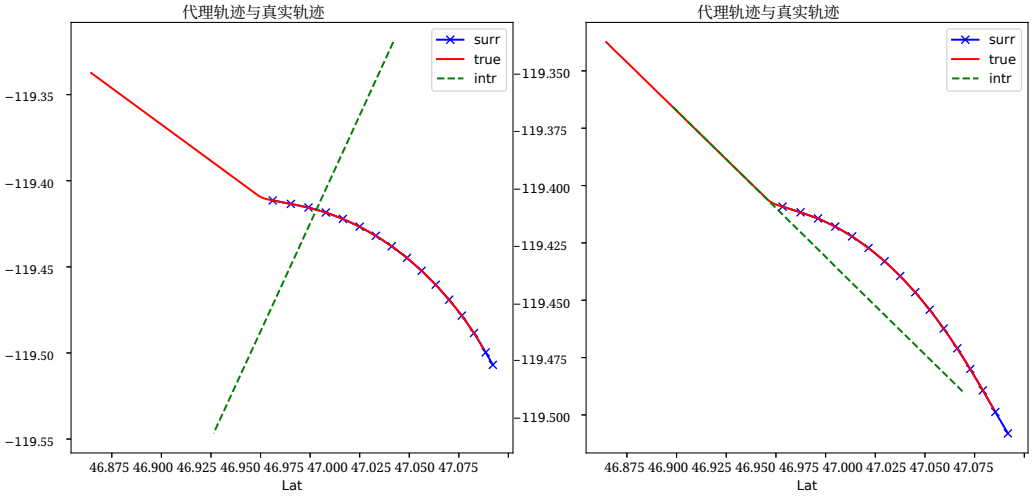
4.1 统计结果

为了评估规避策略本身的性能，我们针对超过 10^4 种随机条件进行了批量仿真。我们的 RLCAS 成功规避与入侵者之间的交通冲突并返回原始航线，表现达到 98%。在失败的案例中，即仿真总数的 2%，智能体未能在保持安全距离的同时返回航线。图 6 显示，在未能解决冲突的案例中，智能体与入侵者的初始距离过近。由于失败率相对较高，该系统并非设计为独立运行的系统，因此需要在其基础上构建安全层。

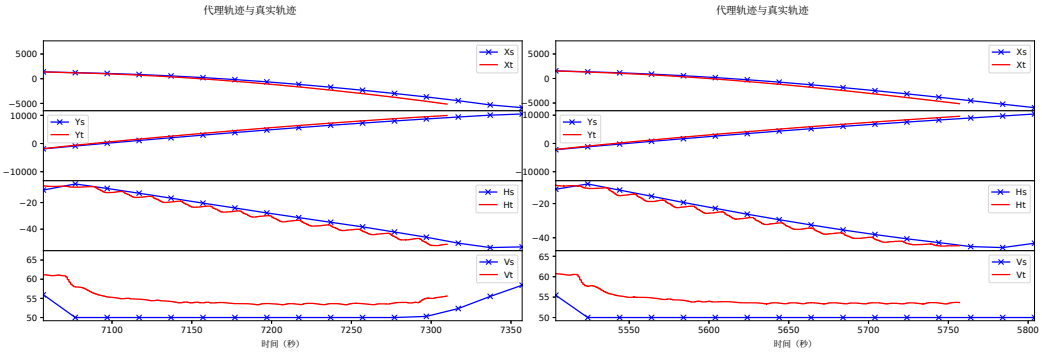
4.2 轨迹评估

本节评估演示器能否紧密跟随指令轨迹。我们设置了四项实验，旨在促使智能体展现不同的行为。这些实验包括以下场景：入侵者从左向右、从右向左穿越规划航线，入侵者沿迎头航向飞行，以及入侵者沿相同航向飞行。

在所有场景中，无人机的控制器都弥补了物理能力与替代模型假设之间的差异。图7a和图7b展示了无人机如何在跟随航路点时补偿距离误差和到达时间误差。



(a) 入侵者沿冲突的交叉航向飞行。 (b) 入侵者沿冲突的迎头航向飞行。



(c) 轨迹误差的累积。 (d) 轨迹误差的累积。

Fig图 7：代理与真实状态转移不匹配导致的状态转移误差 .

5 结论

本研究提出了一种基于强化学习框架训练的自动化冲突解脱系统。规避策略在代理任务上离线训练，并以神经网络模型的形式存储。运行时的RLCAS系统携带预先计算的解库，可快速、高效地提供规避路线，以化解潜在的空中交通冲突。所有优化都在线下完成，不在飞行器上进行，因此可以在低功耗节点上运行。该系统非常适合无法搭载运行时路径规划求解器的小型高速无人机。

该系统与 ROS 集成，可从仿真过渡到实体演示器。2021 年在格兰特县国际机场进行的现场演示生成了飞行数据，该机场位于华盛顿州格兰特县摩西湖西北六英里处。相关数据将另行分析并发表。由于故障率相对较高，RLCAS 需要配备一个在其上运行的保障监控器。我们概述了两种适用于 RLCAS 系统的潜在保障方法——内部代理验证和外部运行时监控。有关保障的更多信息，请参阅另一篇论文 [17]。

最重要的是，这项工作为解决动态轨迹生成任务提供了一种替代方法。关键组件的多样化是整个系统风险缓解的重要理念。

致谢

作者感谢波音的 Alex Chen 和 Michael McGivern 为数据收集工作提供帮助。

参考文献

[1] Steve Calandrillo、Jason Oh 和 Ari Webb。致命无人机：为何美国联邦航空管理局的法规未能保障无人机安全。斯坦福科技法律评论，23：182，2020年。[2] Ewen MacPherson。世界准备好迎接无人机了吗？《航空与航天法》，43(2)，2018年。[3] 美国联邦航空管理局（FAA）AJV-115 新兴技术团队。国家空域系统中的无人驾驶航空器运行（NAS）。美国联邦航空管理局（*FAA*），JO 7210.891，2015年。[4] Yoshiaki Kuwata、Gaston A Fiore、Justin Teo、Emilio Frazzoli 和 Jonathan P How。利用快速探索随机树进行城市驾驶运动规划。载于*2008*年 *IEEE/RSJ*智能机器人与系统国际会议，第1681–1686页。IEEE，2008年。[5] Charles W Warren。利用人工势场进行全局路径规划。载于*1989*年 *IEEE*机器人与自动化国际会议，第316–317页。IEEE计算机学会，1989年。[6] Joohyun Woo 和 Nakwan Kim。利用深度强化学习实现无人水面艇碰撞规避。《海洋工程》，199：107001，2020年。[7] Meha Kaushik、Vignesh Prasad、K Madhava Krishna 和 Balaraman Ravindran。利用深度强化学习进行模拟高速公路驾驶中的超车机动。载于*2018*年 *IEEE*智能车辆研讨会（*IV*），第1885–1890页。IEEE，2018年。[8] Luman Zhao、Myung-Il Roh 和 Sung-Jun Lee。基于深度强化学习的自主船舶路径跟踪与碰撞规避控制方法。《海洋科学与技术杂志》，27(4)：1，2019年。[9] Denis Osipychev、Duy Tran、Weihua Sheng 和 Girish Chowdhary。面向自动驾驶汽车的基于人类意图的碰撞规避。载于*2017*年美国控制会议（*ACC*），第2974–2979页。IEEE，2017年。[10] Richard S Sutton 和 Andrew G Barto。强化学习：导论。麻省理工学院出版社，2018年。[11] Greg Brockman、Vicki Cheung、Ludwig Pettersson、Jonas Schneider、John Schulman、Jie Tang 和 Wojciech Zaremba。OpenAI Gym。*arXiv*预印本 *arXiv:1606.01540*，2016年。[12] Ashley Hill、Antonin Raffin、Maximilian Ernestus、Adam Gleave、Anssi Kanervisto、Rene Traore、Prafulla Dhariwal、Christopher Hesse、Oleg Klimov、Alex Nichol 等。Stable Baselines，2018年。[13] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法。*arXiv*预印本 *arXiv:1707.06347*，2017年。[14] Siddharth Mysore、Bassel Mabsout、Renato Mancuso 和 Kate Saenko。利用强化学习对动作策略进行正则化，以实现平滑控制。*arXiv*预印本 *arXiv:2012.06644*，2020年。[15] Richard Cheng、Abhinav Verma、Gabor Orosz、Swarat Chaudhuri、Yisong Yue 和 Joel Burdick。用于降低方差的强化学习控制正则化。载于国际机器学习会议，第1141–1150页。机器学习研究论文集，2019年。[16] Morgan Quigley、Ken Conley、Brian Gerkey、Josh Faust、Tully Foote、Jeremy Leibs、Rob Wheeler 和 Andrew Y Ng。ROS：一个开源机器人操作系统。载于*ICRA*开源软件研讨会，第3卷，第5页。日本神户，2009年。[17] Darren Cofer、Ramachandra Sattigeri、Isaac Amundson、J unaid Babar、Saqib Hasan、Eric W Smith、Karthik Nukala、Denis Osipychev、Matthew A Moser、James L Paunicka 等。配备运行时保障的碰撞规避神经网络飞行测试。载于*2022*年 *IEEE/AIAA* 第*41*届数字航空电子系统会议（*DASC*），第1–10页。IEEE，2022年。

A 附录

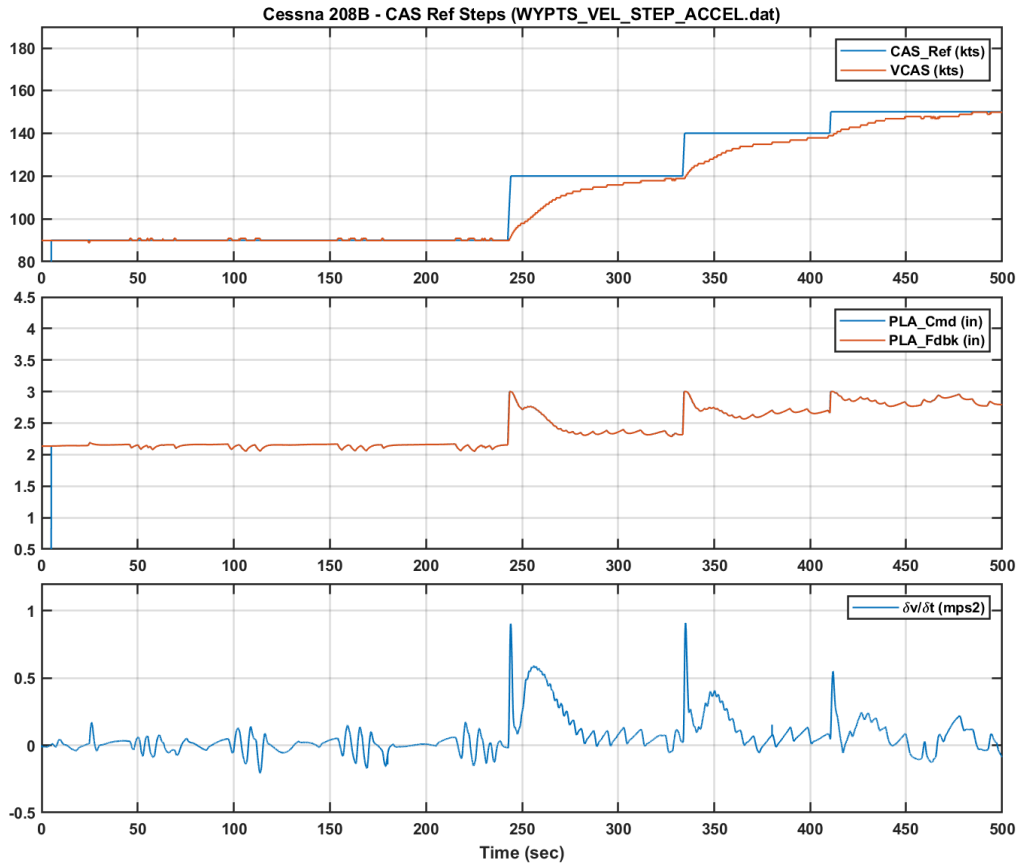
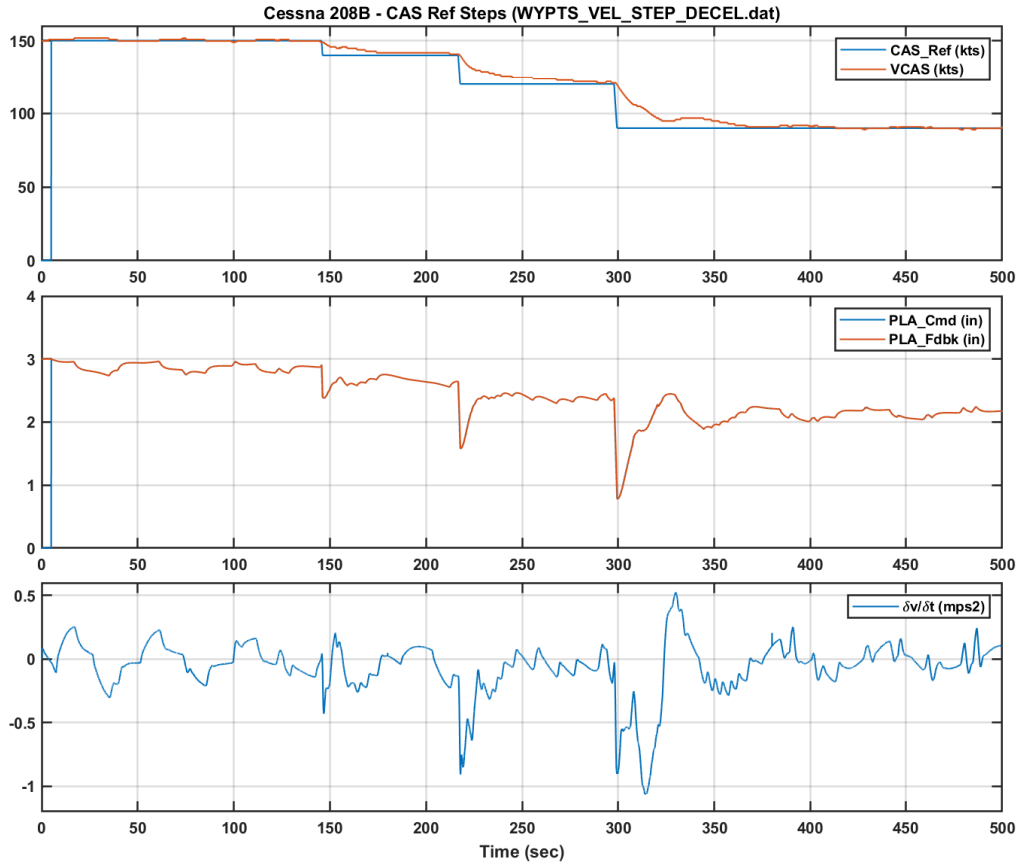


图8：演示器对指令阶跃输入的空速和纵向加速度响应。

参数	范围	单位
时间步长 dT	1.0	sec
最大时间 t_{MAX}	300.0	sec
距离范围	-10000 .. 10000	m
水平间隔	1000	m

T代理环境仿真参数

米。



Figu图 9：演示器对指令的空速和纵向减速响应
输出阶跃。

条件	奖励
距离违规	-100
无违规，缺少原始路径	-10
无违规，返回原始路径	100

表 3：代理环境奖励函数。

算法 1： 代理仿真更新

重置为随机初始状态;

当 $t < t_{max}$ **do**

更新入侵车辆:

- 调用车辆控制器以计算控制量
- 向车辆发送控制信号
- 调用车辆更新以移动该车辆

更新智能体车辆:

- 向车辆传递控制信号
- 调用车辆更新以移动该车辆

评估状态:

- 冲突
- 终止
- 得分

返回 观测 ，奖励 ，终止

end