基于 Transformer 的多智能体强化学习：结构化与非结构化空域中的间隔保障

Arsyi Aziz 计算机科学系 乔治华盛顿大学 美国华盛顿特区

Peng Wei 机械与航空航天工程系 乔治华盛顿大学 美国华盛顿特区

arXiv:2601.04401v1 [cs.RO] 7 Jan 2026

摘要—传统的基于优化的流量调度依赖于严格遵循预先计算的时刻表，这限制了先进空中交通（AAM）随机运行所需的灵活性。相比之下，多智能体强化学习（MARL）提供了一种去中心化、自适应的框架，能够更好地应对不确定性，而这对于安全的航空器间隔保障至关重要。尽管具有这一优势，当前的多智能体强化学习方法往往会对特定空域结构过拟合，从而限制其适应新配置的能力。为提高泛化能力，我们将多智能体强化学习问题重新表述为相对极坐标状态空间问题，并在多种交通模式和交叉角条件下训练 Transformer 编码器模型。训练得到的模型可提供速度建议，以解决冲突，同时使航空器保持接近期望巡航速度。在实验中，我们在结构化与非结构化空域中评估了深度为 1、2 和 3 层的编码器，发现单一编码器配置的表现优于更深的变体，其近空中碰撞率接近于零，且间隔丧失违规持续时间短于更深的配置。此外，我们还表明，相同配置的表现优于一个完全基于注意力机制设计的基线模型。总而言之，结果表明，新提出的状态表示、创新的神经网络架构设计以及所提出的训练策略，为结构化与非结构化空域中的航空器间隔保障提供了一种适应性强、可扩展的去中心化解决方案。

关键词—先进空中交通、间隔保障、多智能体强化学习、Transformer

I. 引言

当代空中交通管理（ATM）系统围绕基于优化的时刻表构建。例如，在美国，交通管理顾问系统和基于时间的流量管理（TMA/TBFM）利用时刻表对航空器在流量调节点的排序进行安排。尽管这些时刻表方法适用于当前的空中交通运行，尤其有助于维持航空器之间的安全间隔，但它们可能无法满足新兴的先进空中交通（AAM）需求。点对点航班、更高的交通流量以及快速变化的随机空域状况，可能使计算得出的时刻表失效，从而降低其应对未来空中交通状况的能力。

基于学习的方法，尤其是强化学习（RL），提供了一种颇具前景的替代方案。与基于时刻表的方法相比，强化学习能够学习对动态且不确定的空中交通状况作出实时响应的策略。在单次遭遇场景中，研究已表明其能够学习有效的引导航向策略，用于冲突消解 [1]。在更大范围的控制层面，强化学习可以扩展为多智能体强化学习（MARL），从而支持多个航空器在结构化空域 [2] 和非结构化空域 [3] 中开展协同的航路间隔保障。

尽管近年来取得了进展，泛化能力仍是该领域多智能体强化学习面临的主要障碍。在有限场景集合上训练的模型往往会对特定空域结构参数过拟合，从而在空域结构或交通密度变化时降低其鲁棒性和效能。针对一种空域结构训练的多智能体强化学习策略可能无法适用于另一种结构；针对中等交通流量优化的策略，在高密度情况下可能变得不安全，在交通稀疏时则可能过于保守。这种对场景变化或分布偏移的易感性凸显了开发新方法的必要性，以便更好地适应不同空域配置和交通密度。

为解决这一局限，我们聚焦于航空器间隔保障问题，并提出一种多智能体强化学习框架，以提升其在先进空中交通运行的结构化和非结构化空域中的泛化能力。我们的方法结合Transformer的编码能力与多智能体强化学习的适应性，将本机和侵入航空器的状态编码为Transformer编码器中的词元，再输出加速或减速建议形式的速度建议。我们不让航空器（或智能体）仅在单一航路结构或固定交通模式下训练，而是设计了多样化的训练流程，使学习智能体能够接触不同的空域结构和交通密度。通过接触这些多样情形，智能体能够学习可有效迁移至训练期间未见过的新空域结构的表征。

此外，为进一步提升泛化能力，我们对训练流程进行了调整，改动了强化学习的问题建模和神经网络架构。在问题建模方面，我们采用相对极坐标定义状态空间，以突出自我中心状态，

单层Transformer模型性能演示可在https://youtu.be/DpwM6QDnCuU观看。

并在奖励中加入速度激励项。在神经网络架构方面，我们通过引入一个以本机信息为条件的分类器令牌，从注意力机制转向词元化。

因此，本研究的主要贡献如下：
1）我们使用极坐标状态表示和速度激励奖励函数重新表述强化学习问题；2）我们提出了一种带有条件化分类器令牌的编码器–Transformer架构，可适应数量可变的侵入航空器；3）我们设计了一套训练方案，基于程序化生成的扇区结构，并改变交叉角和交通密度。

这些贡献共同支持间隔保障模型向未来先进空中交通运行的泛化。

本文其余部分组织如下。第*II*节 回顾相关工作。第*III*节 将间隔保障问题形式化为马尔可夫决策过程。第*IV*节 介绍我们改进的Transformer编码器架构。第*V*节介绍模型训练细节。第*VI*节报告实验结果，第*VII*节则进行相关讨论。第*VIII*节总结本文，并讨论本研究更广泛的影响以及未来研究方向。

II. 相关工作

*A. Transformer*网络

Transformer [4] 是一类处理离散词元序列的序列模型。例如，在自然语言处理中，词元可以对应单词或子词；而在航空器间隔保障和冲突消解中，词元则表示对航空器状态向量的编码。Transformer架构的核心是自注意力机制，它使模型能够比较并整合所有词元位置的信息。这与长短期记忆网络（LSTM） [5]和门控循环单元（GRU） [6],等循环架构形成对比，后者按顺序处理词元，并且往往难以保留序列早期位置的信息，或难以对序列进行并行计算 [4]。仅编码器架构（如 BERT） [7]基于Transformer架构，将输入序列映射为包含上下文信息的高层词元嵌入。随后，这些词元级表示可以通过不同方式进行聚合，以用于下游分类任务 [7], [8], ，例如，引入专用的分类词元 [7] ，或对词元应用均值池化或最大池化等池化操作 [9]。

B. 用于航空器间隔保障的多智能体强化学习

多智能体强化学习（MARL）是一种机器学习方法，通过让模型在预设奖励函数的引导下与环境交互，使其学会产生预期行为。在间隔保障问题中，环境被建模为多架航空器在共享空域内交互。早期采用多智能体强化学习的方法在此场景中使用全连接网络表示策略，并通过输入填充处理航空器数量可变的情况 [10]。这种设计对模型可处理的航空器数量设定了固定上限。例如，如果将输入层配置为处理最多十架航空器的特征，模型就无法适应航空器数量超过规定的十架的场景。

后续研究引入了注意力机制，以更好地应对不同的交通流量水平 [2]。这些模型能够学习每架侵入航空器的相对重要性，从而适应航空器数量的变化。随后，这种方法扩展为采用 Transformer 架构，并通过将每架侵入航空器的状态编码为相对于本机的状态，进一步加以改进 [11]。在这些进展的基础上，我们提出了进一步的改进，以增强模型对不同空域配置和交通密度的适应能力。

III. 问题建模

我们研究先进空中交通在低空扇区运行时的间隔保障问题，并将其扩展至非结构化空域。在该问题中，我们考虑多架航空器试图保持不同期望巡航速度的情形。为防止碰撞，间隔保障模型必须监测空域中的潜在冲突，并发布离散的速度建议（降低、保持、提高），以维持安全间隔。由于期望巡航速度也纳入考量，在安全条件允许时，模型还应确保每架航空器保持接近期望巡航速度。因此，该问题被设定为一种速度控制的间隔保障问题，包含两个目标：*(i)* 维持航空器之间的安全间隔，以及*(ii)* 尽量减少各航班偏离其期望巡航速度的程度。

A. 安全关键事件

我们定义了空域内可能发生的三类安全关键事件：

a) 冲突：冲突是指在规定的前瞻时域内预计会违反最小间隔标准的情况。换言之，如果航空器保持当前轨迹和速度，预计将在规定的时间窗口内失去安全间隔。在我们的仅速度控制场景中，智能体必须发出加速或减速指令，以主动缓解这些预计的间隔侵犯。

b) 间隔丧失：当实际违反最小间隔标准时，即发生间隔丧失（LoS），这意味着两架或更多航空器同时进入保护区。检测到间隔丧失事件后，航空器必须立即协调进行速度调整（增加或降低速度），以恢复间隔并降低事态升级为近空中碰撞（NMAC）的风险。

c) 近空中碰撞：近空中碰撞（NMAC）是最严重的事件，定义为两架或多架航空器彼此接近至不安全的距离。发生近空中碰撞表明，空中交通管制系统未能维持空域内的安全。如果速度

建议无法在临界阈值内彻底化解冲突，则将启动机载防撞系统——交通警报与防撞系统（TCAS）或机载防撞系统 X（ACAS-X）——发出协调的垂直解脱建议（例如爬升或下降），以避免碰撞。

B. 数学建模

我们将多个航空器（智能体）的扇区级协调建模为一个具有时变群体的完全可观测多智能体马尔可夫决策过程。该过程描述为

$$\mathcal{M} \coloneqq \langle \mathcal{N}_{\max}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}_{\max}}, \mathcal{P}, \mathcal{R}, \gamma \rangle,$$

其中， $\mathcal{N}_{\max}$ 索引可能航空器的超集， $\mathcal{S}$ 是状态空间， $\mathcal{A}_i$ 是智能体 i的动作空间， $\mathcal{P}$ 是转移核（由飞机动力学和大气条件决定）， $\mathcal{R}$ 是奖励函数， $\gamma \in [0,1]$ 是折扣因子。此外，令 $\mathcal{N}_t \subseteq \mathcal{N}_{\max}$ 表示时间 t的活动飞机集合。

在每个离散时间 t，环境具有一个全局状态$s_t \in \mathcal{S}$。基于该全局状态，每架活动航空器 $i \in \mathcal{N}_t$观测到一个自我相对状态

$$s_t^{(i)} \coloneqq \psi_i(s_t) \in \mathcal{S}_i,$$

其中， $\psi_i : \mathcal{S} \to \mathcal{S}_i \subset S$ 是一个确定性的、针对智能体的投影，用于从以 i的相对坐标系表示的全局状态中提取航空器 i的状态。广播式自动相关监视和无误差通信假设这些状态是准确的。

随后，每架航空器根据策略选择一个动作

$$a_t^{(i)} \sim \pi_i(\cdot \mid s_t^{(i)}),$$

并将联合动作配置写作所有动作的集合 $a_t \coloneqq \{a_t^{(i)}\}_{i\in\mathcal{N}_t}$. 下一个状态的演化方式为

$$s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t),$$

这隐含地允许航空器到达/离场，因此航空器数量可能发生变化，通常 $\mathcal{N}_{t+1}$ 不必等于 $\mathcal{N}_t$ 。

每次状态转移后，航空器 i 会获得奖励

$$r_t^{(i)} \coloneqq \mathcal{R}_i\big(s_t^{(i)}, a_t^{(i)}\big).$$

我们将 $r_t \coloneqq \{r_t^{(i)}\}_{i\in\mathcal{N}_t}$ 表示为时间步 t获得的所有奖励的集合。轨迹记为$\tau = (s_0,\ a_0,\ r_0,\ s_1,\ a_1,\ r_0 \ldots,\ s_T)$，定义在有限时域 T上。对于智能体 i，完成一条轨迹后，其期望折扣回报为

$$J_i(\pi) \coloneqq \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T-1} \gamma^t\, r_t^{(i)}\right].$$

由于我们采用一般和多智能体设定，因此最大化所有航空器这些数值的平均值，所以总体目标是找到使其最大化的策略

$$\pi^\star = \arg\max_{\pi} \frac{1}{|\mathcal{N}_{\max}|} \sum_{i\in\mathcal{N}_{\max}} J_i(\pi).$$

C. 状态空间

我们在以自我为中心的坐标系中构建智能体的状态空间，其中包括本机特征和一组入侵机特征。

1) 本机特征

本机特征为智能体提供与速度相关的信息，包括以下元素：

- 校准空速：航空器 i 在时间 t的空速，经仪表误差和位置误差校正后记为 $v_{(i)\ \text{cas}\ t}$。

- 速度偏差：校准空速与航空器 i在时间 t的期望速度之间的绝对差，定义为

$$\Delta v_t^{(i)} \coloneqq \left| v_{\text{cas},t}^{(i)} - v_{\text{des}}^{(i)} \right|.$$

2) 入侵机特征

入侵机特征是相对于本机编码的入侵者信息集合，包括：

- 相对位置：每架侵入航空器相对于本机的位置，以极坐标表示。令 $d_{(i,j)\ t}$ 表示时间 t时航空器i 与 j 之间的欧几里得距离。相对位置包含四个分量：

–到近空中碰撞（*NMAC*）边界的距离：

$$d_{\text{nmac},t}^{(i,j)} \coloneqq d_t^{(i,j)} - r_{\text{nmac}},$$

其中 r_{nmac} 表示500 英尺的NMAC半径。– 到间隔丧失（*LOS*）边界的距离：

$$d_{\text{pz, t}}^{(i,j)} \coloneqq d_t^{(i,j)} - r_{\text{rpz}}^{(i)},$$

其中 $r_{\text{pz}}^{(i)}$是航空器 i发生间隔丧失时的保护区半径，默认设为 5 海里。

– 侵入航空器 j 相对于本机 i的相对方位角：

$$\theta_t^{(i,j)} \coloneqq \big(\texttt{arctan2}(\Delta y_t^{(i,j)}, \Delta x_t^{(i,j)}) - \phi_t^{(i)}\big).$$

此处， $\Delta x_t^{(i,j)}$ 和 $\Delta y_t^{(i,j)}$ 是侵入航空器 j 相对于本机i的相对位置分量。双参数反正切函数 arctan2 是四象限反正切函数，返回侵入航空器相对于本机的绝对方位角。减去本机航向 $\phi_t^{(i)}$ 即可得到相对方位角。此特征通过正弦函数和余弦函数进行编码，以消除特征中的角度不连续性。

– 失隔指示变量：

$$b_{\text{los,t}}^{(i)} \coloneqq \mathbb{1}_{\{d_t^{(i,j)} \le r_{\text{pz}}\}},$$

当航空器 i 和j 位于保护区内时取值为 1，否则取值为 0 的二元变量。尽管该条件可以根据 $d_t^{(i,j)}$ 推断得出

相对于 $r_{\text{pz}}^{(i)}$，显式纳入该指示变量有两个目的：（*i*）它直接编码了失隔事件；（*ii*）它允许模型在发生此类失隔事件时有选择地停用或调节其他特征。

- 相对速度。除了位置特征外，我们还纳入了两个速度特征，即径向速度和切向速度。在时间 t，将本机 i与侵入航空器 j 之间的相对位置和相对速度分别记为

$$\mathbf{p}_t^{(i,j)} := \begin{bmatrix} \Delta x_t^{(i,j)} \\ \Delta y_t^{(i,j)} \end{bmatrix}, \qquad \mathbf{v}_t^{(i,j)} := \begin{bmatrix} \Delta v_{x,t}^{(i,j)} \\ \Delta v_{y,t}^{(i,j)} \end{bmatrix},$$

其中，$\Delta x_t^{(i,j)} = x_t^{(j)} - x_t^{(i)}$ 和 $\Delta y_t^{(i,j)} = y_t^{(j)} - y_t^{(i)}$ 分别为东西向和南北向位置差，而 $\Delta v_x^{(i,j)}{}_{、t} = v_x^{(j)}{}_{、t} - v_{x\,t}^{(i)}{}_{、}$，$\Delta v_y^{(i,j)}{}_{、t} = v_y^{(j)}{}_{、t} - v_{y\,t}^{(i)}{}_{、}$为相应的速度分量差。

从 i指向 j的径向单位向量为

$$\hat{\mathbf{e}}_{p,t}^{(i,j)} := \frac{\mathbf{p}_t^{(i,j)}}{\|\mathbf{p}_t^{(i,j)}\|},$$

切向单位向量则由 $\pi/2$逆时针旋转得到：

$$\hat{\mathbf{e}}_{\psi,t}^{(i,j)} := R_{\pi/2}\,\hat{\mathbf{e}}_{p,t}^{(i,j)}.$$

将相对速度投影到这些轴上，可得到两个特征：

- 径向速度：

$$v_{p,t}^{(i,j)} := \mathbf{v}_t^{(i,j)} \cdot \hat{\mathbf{e}}_{p,t}^{(i,j)},$$

- 切向速度：

$$v_{\psi,t}^{(i,j)} := \mathbf{v}_t^{(i,j)} \cdot \hat{\mathbf{e}}_{\psi,t}^{(i,j)}.$$

D. 动作空间

在每个决策步，智能体会从三种速度建议中选择一种：提高、降低或保持当前校准空速。速度调整以固定的 5 节为增量。由于我们将航空器视为具备悬停能力的先进空中交通飞行器，因此校准空速的范围限定为 0 节至航空器允许的最高速度。我们将此范围的端点记为 $v_{\min} = 0$ 和 $v_{\max}$.

E. 奖励函数

为了训练能够在避免安全关键事件的同时，针对航空器的期望速度发出速度建议的策略，我们设计了一种平衡正向分量和负向分量的奖励函数。正向项用于强化朝理想行为迈进的过程，而负向项则用于惩罚不安全或不理想的结果。重要的是，该奖励函数经过精心设计，旨在促进安全且自适应的行为，使其在训练期间未见过的环境中仍然有效。

令$\alpha_{(\cdot)}$ 表示奖励分量的一个超参数。奖励分量随后分为三组：

1) 无冲突：

a) 与期望速度的接近程度：如果在固定的前瞻时域 L内预计不会发生冲突，则智能体会因接近期望速度而获得奖励：

$$\mathcal{R}_v(s_t, a_t) := (+)\ \alpha_v\, \Delta\hat{v}_t,$$

其中 $\Delta\hat{v}_t = 1 - \frac{\Delta v_t}{v_{\max} - v_{min}} \in [0,1]$ 表示与期望速度的归一化接近程度。

2) 发生冲突时：

a) 侵入时间：如果预计在固定的前瞻时域 L 内发生冲突，则智能体会受到与最小归一化侵入时间成正比的惩罚：

$$\mathcal{R}_{\text{conflict}}(s_t, a_t) := (-)\ \alpha_{\text{conflict}}\, \hat{T}_{\text{los},t},$$

其中

$$\hat{T}_{\text{los},t} := \text{clip}\left(\left[\frac{L - \min_{j\in\mathcal{N}\setminus\{i\}} T_{\text{los},t}^{(i,j)}}{L}\right];\ 0, 1\right),$$

且 $T^{(i,j)}{}_{\text{间隔丧失 }t} \in [0, L]$ 表示根据相对速度分量计算得出的航空器 i 与 j之间预计发生间隔丧失的时间。

b) 至 *NMAC* 的距离：如果智能体已经发生间隔丧失，则根据其与最近侵入航空器之间的归一化距离施加额外惩罚：

$$\mathcal{R}_{\text{los}}(s_t, a_t) := (-)\ \alpha_{\text{los}}\, \hat{d}_t^{(i,j)},$$

其中，航空器 i与j 之间的归一化距离定义为

$$\hat{d}_t^{(i,j)} := \text{clip}\left(\left[\frac{r_{\text{pz}} - \min_{j\in\mathcal{N}\setminus\{i\}}\left(d_t^{(i,j)}\right)}{r_{\text{pz}} - r_{\text{nmac}}}\right];\ 0, 1\right),$$

且 $d_t^{(i,j)}$ 表示航空器 i 与 j之间的欧几里得距离。

3) 发生近空中碰撞时：如果智能体处于近空中碰撞状态，我们会施加较大的惩罚，以抑制这种行为：

$$\mathcal{R}_{\text{nmac}} := (-)\ \alpha_{nmac}.$$

在我们的设定中，我们将 α_{nmac} 设为 -100.

将各个奖励分量汇总后，整体奖励函数可简洁地表示为：

$$\mathcal{R}(s_t, a_t) = \begin{cases} \mathcal{R}_v(s_t, a_t), & \text{if } \textbf{not in} \text{ conflict}, \\ \mathcal{R}_{\text{conflict}}(s_t, a_t) + \mathcal{R}_{\text{los}}(s_t, a_t), & \text{if } \textbf{in} \text{ conflict}, \\ \mathcal{R}_{\text{nmac}}, & \text{if } \textbf{in} \text{ NMAC}. \end{cases}$$

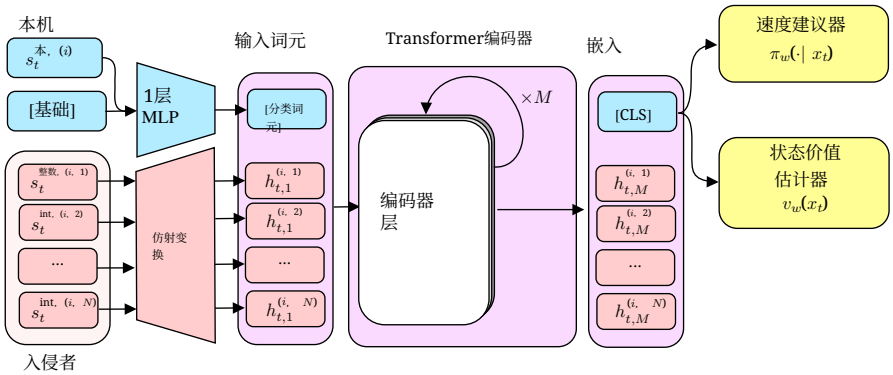


图1. 所实现的网络采用编码器–Transformer架构，其中入侵者信息表示为入侵者令牌。由本机信息生成的分类器（[CLS]）令牌会附加到该集合中。完整的标记序列由包含 M 层的Transformer编码器处理。最后一层的［CLS］令牌用于生成策略和状态价值估计。

IV. 神经网络架构

图1展示了我们的神经网络架构。该架构基于自然语言处理中的编码器-Transformer设计 [7]，，并经过调整，使其处理入侵者令牌而非词令牌。在这一设计中，本机信息由一个条件化分类器令牌编码，该令牌汇总来自入侵者令牌的信息。下面详细介绍各个组成部分：

本机特征适配：本机特征被编码为条件化的分类器令牌。为构建该令牌，我们首先将基础令牌定义为一个从标准高斯分布初始化的可学习参数。随后，通过拼接将该可学习先验与本机信息融合，再依次进行仿射变换、GELU激活[12], 和层归一化 [13]，以显式地根据自车状态对令牌进行条件化。所得表示即为分类器令牌，记为［CLS］，后续将作为令牌空间中信息的聚合器。

入侵者特征适配：入侵机特征被编码为入侵者令牌。为获得这些令牌，我们分别通过带有层归一化的仿射变换，将每架侵入航空器的特征映射到嵌入空间。此过程会生成一组入侵者令牌，每架侵入航空器对应一个。

令牌处理与聚合：将本机和侵入航空器的特征适配到嵌入空间后，我们将［CLS］词元与入侵者令牌拼接，形成一个词元集合。随后，该集合由一个 M层Transformer编码器进行处理，编码器通过自注意力机制、 [4]，层归一化和GELU激活生成经过注意力机制精炼的词元集合。我们从编码器的输出中提取[CLS］词元，并将其用作策略分布（速度建议器）和状态价值函数估计器头的输入。

这里，我们明确使用 ［CLS］词元作为全局聚合器，而不是对入侵者词元进行池化，原因有三。首先，这样可以让编码器将所有入侵者信息汇集到一个表征中。这样便无需对入侵者词元采用诸如最小池化或最大池化等临时的池化方案。其次，由于 ［CLS］词元以本机特征为条件，因此它会在所有编码器层中保留这一上下文。其他设计则需要通过拼接，在 Transformer 头处稍后注入本机信息，这可能会将本机上下文限制在处理的最后阶段，并削弱其对较早的低层表征的影响。第三，这种设计天然支持数量可变的入侵者，包括不存在入侵者的情况，无需对入侵者词元进行显式填充。

V. 模型训练

本节概述我们的训练方法。首先简要介绍用于训练模型的近端策略优化算法，然后介绍训练环境，最后说明训练细节。

A. 近端策略优化

我们使用近端策略优化（PPO）的裁剪变体来训练网络 [14]。PPO 扩展了策略梯度框架，其中通过沿期望回报的梯度上升方向优化随机策略 $\pi_w(a \mid s)$ 的参数。在标准策略梯度方法中，更新方向为

$$\nabla_w J(w) = \mathbb{E}_t\Big[\nabla_w \log \pi_w(a_t \mid s_t)\, G_t\Big],$$

其中， G_t 为后续累积回报， $\mathbb{E}_t$ 表示在当前策略下收集的多个时间步上的期望。此更新会提高获得较高回报的动作的概率，并降低效果较差的动作的概率。

然而，这种朴素更新中的大步长可能会使学习过程不稳定。PPO 通过最大化代理目标函数来解决这一问题，该目标函数会限制新旧策略之间的偏差。代理目标函数定义如下：

$$\max_w \ \hat{\mathbb{E}}\Big[\min\big(r_t(w)\,\hat{A}_t,\ \text{clip}\big(r_t(w), 1-\epsilon, 1+\epsilon\big)\,\hat{A}_t\big)\Big],$$

其中， $r_t(w) = \frac{\pi_w(a_t|s_t)}{\pi_{w_{\text{old}}}(a_t|s_t)}$ 表示更新后的策略与旧策略之间的概率比， A_t 是优势函数的估计值（通常通过获得的奖励 R_t 与学习得到的状态价值函数 $v(s_t)$之差计算），而 ϵ是裁剪参数。裁剪参数 ϵ可防止更新过于激进，从而使训练更加稳定。

B. 训练环境

我们在基于 BlueSky 仿真器 [15]构建的模拟空中交通管制环境中训练策略。该环境模拟了一个圆形飞行扇区，其中两条航线在中心点交叉。每条航线在固定距离内延伸，扇区半径约为 30 海里。智能体每隔一秒采取一次行动。为提高泛化能力，我们随机调整航线之间的交叉角，并确保航线端点之间的最小间隔为 5 海里。训练环境的可视化结果见图2。

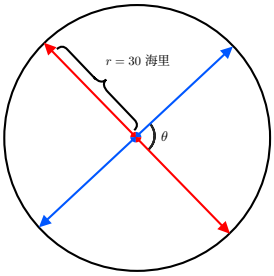


图2. 仿真训练环境由两条航线组成，其端点在半径为30海里的圆内均匀采样。航线可能以 $\theta \in [0°, 180°]$ 角相交，并且可以沿任一方向飞行。

与传统的固定翼场景不同，我们考虑具有垂直起降能力的航空器，它们既能前向运动，也能悬停。这一额外自由度不仅会改变状态动力学，还会带来悬停行为这一挑战，而这可能导致死锁（即环境可能无法终止）。

为确保环境能够终止，我们考虑航空器可能从环境中移除的两种情况。第一种情况是航空器到达目的地所用时间过长。为此，我们根据每架航空器的期望速度，预先计算其预计到达时间，并额外增加20分钟缓冲时间；考虑到电池限制，我们将缓冲时间设为20分钟。如果航空器超过这一预先计算的时间，就会从环境中移除。第二种情况是发生近空中碰撞。此时，违规航空器将被移除，并受到高额惩罚（见第III-E节）。

C. 训练细节

我们使用 PPO，在八个并行环境中训练第 IV 节所述的网络，这些环境共同模拟了约 75 天的交通流量。固定长度的滚动缓冲区汇总了所有智能体的轨迹，每个环境生成 4096 秒的序列。状态转移会被随机打乱并采样，以优化 PPO 目标函数。为稳定训练，我们采用价值裁剪和优势归一化。

总共对嵌入维度为128、编码器前馈维度为512的Transformer编码器网络进行了200次更新。我们进一步训练了三个模型变体，其编码器深度分别为1、2和3层。为

表 I. 实验中采用的基于 Transformer 的神经网络配置。

组件	设置
编码器层	[1, 2, 3]
词元/嵌入维度	128
前馈网络（FFN）维度	512
注意力头	16

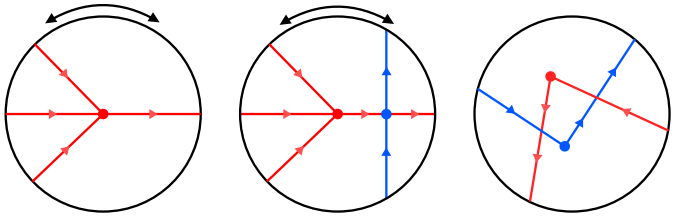


图3. 在模拟结构化空域和非结构化空域的三种场景下对模型进行评估。情况 (a) 和情况 (b) 表示结构化空域，但存在旋转变化。情况（c）作为非结构化空域的代理，其中包含一个中间航路点。为便于可视化，我们仅展示情况（c）中的两条航线。然而，可能的航线数量由表III中定义的生成方案决定。

确保评估公平并降低对初始化的敏感性，每个模型都使用不同的随机种子训练三次。相关神经网络配置和超参数的摘要分别列于表I和表II。

表II. Transformer 网络 PPO 训练所采用的超参数。

超参数	值
更新次数	200
并行环境	8
时间范围（T）	4096
批量大小	128
更新轮数	4
折扣因子 (γ)	0.99
GAE-λ	0.95
裁剪系数 (ϵ)	0.2
熵系数	0.01
价值函数系数	0.5
最大梯度范数	0.5
优势归一化	✓
价值函数裁剪	✓

VI. 仿真实验

本节介绍仿真实验的结果。首先，我们将考察训练曲线，比较三种神经网络配置，它们分别包含 1、2 和 3 个编码器层。随后，我们将评估各个模型在结构化与非结构化空域中适应未见场景的能力。最后，我们将比较基于Transformer的架构与一种完全由注意力机制构成的神经网络架构。

A. 训练曲线

我们分别使用 1 层、2 层和 3 层编码器层的三种 Transformer 配置进行了训练，每种配置训练 200 次更新，并运行了三次独立实验

。在整个训练过程中，我们使用两个指标评估模型性能：

1) λ-回报：通过广义优势估计（GAE） [16]计算得到的自助式回报；以及 2）策略熵：用于衡量学习到的策略的随机性（即可预测性）。

我们先来看图 4a，该图绘制了 200 次训练更新期间的自助式回报。在这张图中，我们观察到所有编码器配置的 λ-回报都随时间稳步上升。各条曲线从接近 0 开始，在不同配置下上升至约 75 到 83。其中，1层编码器达到最高的最终 λ-回报 83.1，其次是 2层编码器的 78.8，以及 3层编码器的 75.1。

策略的熵如图 4b 所示。在所有 Transformer网络配置中，熵都随时间降低，反映出策略逐渐收敛到熵更低的状态。1层编码器在大约第25^{th} 次和第 160^{th} 次更新之间表现出明显偏离，此时熵上升并稳定在接近 0.75 的水平，随后又回落到较低值。2层编码器在各次运行中呈现出最宽的范围，其上界约为 0.8，并在训练后期出现。相比之下，3层编码器表现出更加平稳的下降趋势，且波动范围更窄。在最后几次更新时，所有模型都收敛到相近的熵值，约在 0.35 到 0.5 之间。

B. 未见过的扇区结构配置

在训练环境中对三种不同的 Transformer 配置完成训练后（见图2），我们随后在三种未见过的空域结构上评估训练后模型的性能，如图3所示：

- 情况 (a)：三个航线汇聚于一个汇合点；
- 情况 (b)：三个航线的单一汇合点下游设有一个四岔交叉口；以及
- 情况（c）：一个包含程序生成航线的空域，这些航线的端点从扇区边界上采样，并在扇区内随机放置一个中间航路点。

前两种情况代表结构化空域，而第三种情况则作为非结构化空域的代理。为评估模型的泛化能力并确保旋转不变性，情况 (a) 和情况 (b) 会随机旋转一个角度，该角度从 $0°$ 到 $360°$ 的均匀分布中采样。由于情况（c）包含随机采样的中间航路点，某些实例可能无法实现，除非要求所有航空器悬停。为解决这一问题，如果航空器处于间隔丧失状态且能够减速至停止，我们允许提前终止其轨迹。有关航空器采样分布的更多细节见表III。

C. 未见过的扇区结构上的性能

为了分析模型在未见过的扇区配置中的性能，我们关注三个指标：*(i)*近空中碰撞（NMACs）的数量，*(ii)* 所有航空器对在所有时间步中间隔丧失（LOS）的累计时间，以及

表III. 各环境中的飞机分布。

环境	N 航空器	间隔（s）	期望速度（节）
训练	$U(1,\ 20)$	$U(60,\ 1200)$	$U(60,\ 120)$
情况 (a)	$U(1,\ 20)$	$U(60,\ 1200)$	110
情况 (b)	$U(1,\ 20)$	$U(60,\ 1200)$	110
情况（c）	$U(1,\ 10)$	$U(60, 1200)$	$U(60,\ 120)$

† 此处， N aircraft 表示在飞行扇区中生成的航空器数量。$U(a, b)$ 表示从 $[a, b]$ 上的均匀分布中抽取的样本。

(iii) 对期望速度的遵循程度，以航空器速度与期望巡航速度相差不超过 10 节的时间步比例来衡量。基准测试结果汇总于表 IV，数据为 1,000个评估回合的平均值；每种神经网络配置均采用三组独立训练的权重进行评估。

我们首先分析模型防止近空中碰撞的能力。如表 IV 所示，不同的 Transformer 配置均表现出相对可靠的安全间隔能力。具体而言，我们的 1 层和 2 层编码器层网络几乎没有发生近空中碰撞，其平均侵入值分别为 0.002 和 0.001。相比之下，三层网络发生近空中碰撞的次数略高，平均近空中碰撞次数为 0.037。

接着考察间隔丧失（LoS），我们发现，所有 Transformer 配置在情况 (a) 和情况 (b) 的结构化空域中通常发生较少的间隔丧失事件，而在情况（c）的非结构化空域中，侵入次数则明显较高。对于 1层编码器，总体间隔丧失平均值为 678.154，情况 (a) 和情况 (b) 的平均值则显著较低，分别为 30.327 和 96.785。2层编码器的总体间隔丧失平均值达到 1152.971，这主要是受情况（c）中高达 2002.420 的平均值推高。3层编码器的总体平均值为 966.157，且在情况 (a) 和情况 (b) 的结构化场景中表现最差。需要指出的是，在所有场景中，航空器只能进行速度调整，这可能限制其缓解安全侵入的能力，尤其是在情况（c）中。

最后，在遵循期望巡航速度方面，三层编码器的遵循程度最高，达到 72%。与此同时，一层和两层编码器的数值略低，分别为 69.8% 和 65.2%。这一比例降低可能归因于更高的安全因素（近空中碰撞和间隔丧失更少），这会促使航空器更频繁地减速或加速，以防止发生侵入。

D. 与纯注意力的比较

我们将所提出的方法与纯注意力模型 [2]进行了比较。如表 IV 所示，我们的基于 Transformer 的神经网络，尤其是单层编码器配置，在近空中碰撞更少、间隔丧失持续时间通常更短的情况下，实现了更高的安全水平。此外，所有采用 1、2 和 3 个编码器层的 Transformer 网络配置，都比仅注意力基线更贴近期望巡航速度。

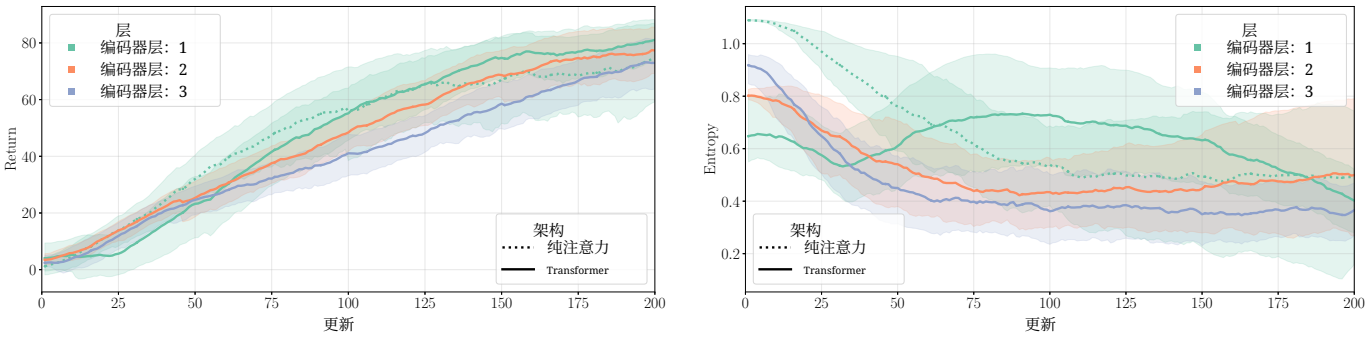


图 4. 训练期间 Transformer 网络的平均 λ-回报 (a) 和熵 (b)，展示了具有 1、2 和 3 个编码器层的配置。每种配置均使用三个随机种子进行训练。实线表示平均值，阴影区域表示最小值与最大值之间的范围。曲线使用 $\alpha = 0.05$ 的指数移动平均进行平滑。

表 IV. 不同网络配置在 1000 个回合中的性能和安全指标。

环境密度		纯注意力 近空中碰撞	纯注意力 速度遵从性	纯注意力 视线时间	**1 层 Transformer** 近空中碰撞	**1 层 Transformer** 速度遵从性	**1 层 Transformer** 视线时间	2 层 Transformer 近空中碰撞	2 层 Transformer 速度遵从性	2 层 Transformer 视线时间	3 层 Transformer 近空中碰撞	3 层 Transformer 速度遵从性	3 层 Transformer 视线时间
训练	≤5	0.008 ± 0.125	0.653 ± 0.175	108.338 ± 527.907	0.000 ± 0.000	0.843 ± 0.182	119.300 ± 570.666	0.000 ± 0.000	0.819 ± 0.170	190.182 ± 813.784	0.011 ± 0.145	0.868 ± 0.132	92.371 ± 461.763
	6-10	0.080 ± 0.437	0.578 ± 0.205	1771.115 ± 4060.318	0.008 ± 0.124	0.709 ± 0.212	1119.789 ± 2463.339	0.002 ± 0.067	0.646 ± 0.235	2215.736 ± 4111.380	0.049 ± 0.329	0.726 ± 0.202	1017.803 ± 2379.765
	11+	0.337 ± 0.927	0.371 ± 0.218	11727.380 ± 14699.918	0.000 ± 0.000	0.490 ± 0.206	4059.592 ± 5150.277	0.000 ± 0.000	0.418 ± 0.217	6455.796 ± 6595.136	0.091 ± 0.417	0.539 ± 0.218	3161.449 ± 4144.970
	平均值	0.081 ± 0.453	0.583 ± 0.211	2191.441 ± 6422.559	0.005 ± 0.097	0.736 ± 0.223	1016.157 ± 2585.519	0.001 ± 0.052	0.685 ± 0.241	1874.085 ± 4020.136	0.040 ± 0.294	0.756 ± 0.206	898.693 ± 2344.195
情况 (a)	≤5	0.000 ± 0.000	0.671 ± 0.110	3.524 ± 43.188	0.000 ± 0.000	0.828 ± 0.081	0.515 ± 5.325	0.000 ± 0.000	0.789 ± 0.128	3.973 ± 19.792	0.000 ± 0.000	0.858 ± 0.082	131.431 ± 804.446
	6-10	0.001 ± 0.045	0.645 ± 0.138	55.615 ± 326.128	0.000 ± 0.000	0.759 ± 0.075	13.360 ± 145.403	0.001 ± 0.045	0.696 ± 0.152	97.871 ± 650.170	0.010 ± 0.143	0.781 ± 0.071	387.216 ± 1531.111
	11+	0.014 ± 0.166	0.560 ± 0.169	650.686 ± 1550.735	0.002 ± 0.070	0.572 ± 0.111	78.825 ± 192.645	0.000 ± 0.000	0.542 ± 0.175	744.886 ± 2412.845	0.063 ± 0.363	0.660 ± 0.075	1544.828 ± 3230.180
	平均值	0.005 ± 0.097	0.622 ± 0.152	225.269 ± 916.817	0.001 ± 0.037	0.712 ± 0.123	30.327 ± 158.063	0.001 ± 0.037	0.659 ± 0.175	273.182 ± 1412.005	0.025 ± 0.230	0.750 ± 0.095	713.279 ± 2217.769
情况 (b)	≤5	0.000 ± 0.000	0.677 ± 0.114	0.000 ± 0.000	0.000 ± 0.000	0.839 ± 0.066	0.000 ± 0.000	0.000 ± 0.000	0.814 ± 0.094	3.889 ± 18.900	0.000 ± 0.000	0.868 ± 0.047	4.791 ± 25.383
	6-10	0.002 ± 0.066	0.593 ± 0.148	156.961 ± 509.802	0.000 ± 0.000	0.700 ± 0.080	37.113 ± 113.309	0.000 ± 0.000	0.641 ± 0.129	206.468 ± 863.310	0.019 ± 0.194	0.731 ± 0.082	249.403 ± 895.354
	11+	0.023 ± 0.216	0.496 ± 0.159	1245.955 ± 1747.369	0.000 ± 0.000	0.551 ± 0.079	215.158 ± 408.550	0.000 ± 0.000	0.482 ± 0.125	962.233 ± 2023.816	0.177 ± 0.594	0.599 ± 0.081	1271.186 ± 2338.114
	平均值	0.010 ± 0.141	0.558 ± 0.159	555.713 ± 1250.620	0.000 ± 0.000	0.652 ± 0.110	96.785 ± 268.506	0.000 ± 0.000	0.590 ± 0.151	462.197 ± 1418.890	0.073 ± 0.390	0.688 ± 0.105	598.839 ± 1626.276
情况（c）	≤3	0.000 ± 0.000	0.535 ± 0.210	149.869 ± 527.145	0.000 ± 0.000	0.771 ± 0.179	223.657 ± 679.895	0.000 ± 0.000	0.759 ± 0.150	306.918 ± 802.779	0.000 ± 0.000	0.774 ± 0.153	217.779 ± 659.005
	4-6	0.014 ± 0.166	0.537 ± 0.179	1415.154 ± 1793.011	0.003 ± 0.078	0.688 ± 0.109	1480.423 ± 1883.388	0.000 ± 0.000	0.667 ± 0.104	1996.577 ± 2150.875	0.009 ± 0.135	0.678 ± 0.108	1602.484 ± 2009.602
	7+	0.018 ± 0.191	0.522 ± 0.179	3525.313 ± 3123.919	0.009 ± 0.132	0.614 ± 0.093	3626.186 ± 2983.981	0.000 ± 0.000	0.600 ± 0.085	4420.622 ± 3195.953	0.018 ± 0.186	0.617 ± 0.083	3681.700 ± 3049.316
	平均值	0.012 ± 0.154	0.534 ± 0.185	1485.924 ± 2133.126	0.003 ± 0.082	0.692 ± 0.132	1569.346 ± 2183.789	0.000 ± 0.000	0.676 ± 0.122	2002.420 ± 2447.275	0.009 ± 0.131	0.687 ± 0.125	1653.816 ± 2273.653
总体	平均值	0.027 ± 0.256	0.574 ± 0.181	1114.587 ± 3556.437	0.002 ± 0.066	0.698 ± 0.156	678.154 ± 1817.575	0.001 ± 0.032	0.652 ± 0.181	1152.971 ± 2676.009	0.037 ± 0.279	0.720 ± 0.143	966.157 ± 2173.696

† 此处，密度是指在一个回合的任意时刻出现在扇区中的航空器最大数量，而速度遵循度是指航空器速度处于其期望速度±10节范围内的时间比例。绿色高亮的单元格表示总体性能最佳，橙色高亮的单元格表示总体性能最差。粗体配置表示性能最佳的模型。

VII. 讨论

我们以两个目标研究安全间隔保障：(*i*) 保持安全间隔，以及 (*ii*) 尽量减少与期望巡航速度的偏差。我们将该问题设定为多智能体强化学习问题，并提出了重新构建的状态空间和神经网络架构。为提升泛化能力，我们在具有不同交叉角和交通密度的多种空域结构上训练模型。我们的神经网络架构是一个 Transformer 网络，其中包含一个以本机特征为条件的 [CLS] 标记。此外，我们对采用 1 、 2 和 3 个编码器层配置的模型进行了基准测试，并在三个未见过的空域结构上将其与仅注意力基线进行了比较。

我们的训练曲线表明，增加编码器层的数量并不能可靠地提升模型性能。单层编码器取得了最高回报，其次是双层编码器配置，而三层配置获得的回报最低。这一模式表明，单个编码器层可能足以通过仅调整速度来学习安全间隔策略。

两个因素可能解释了较深层模型表现较弱的原因。首先，训练更新次数较少可能使较深层模型处于不利地位，因为它们拥有更多的参数。这些较深层模型可能需要更多更新，才能达到较浅层网络的表现。其次，仅速度动作可能导致较深层模型过度优化密集的速度奖励。相比之下，较小的模型似乎不太容易利用这些奖励，而是学会避免后果更严重的冲突惩罚。综合来看，这些因素与我们观察到的单层网络回报更高这一结果相符。

在考察安全指标时也观察到了类似趋势。在三个飞行扇区的 1,000 个评估回合中，单层 Transformer 网络取得了最佳的综合安全性能：近空中碰撞（NMACs）次数接近于零，且在测试配置中间隔丧失（LoS）的平均时间最低。在情况 (a) 和情况 (b) 的结构化空域中，其可靠性最高，同时还能保持与期望速度的高度一致。然而，在情况（c）的非结构化空域中，性能略有下降，这与仅速度动作空间在间隔保障方面的局限性一致。两层编码器配置取得了相近的安全结果，但 LoS 时间略长。三层配置的表现最差，出现次数最高

的近空中碰撞（NMACs）。

最后，我们展示了我们的Transformer架构优于仅注意力基线。具体而言，1层编码器配置的NMAC次数更少、LoS时间更短，且比基线更贴近期望速度。这一结果可能源于额外采用了多头注意力，以及使用了一个以本机信息为条件的 [CLS] 标记。不过，这一假设可能还需要进一步研究。

VIII. 结论

本研究表明，在精心设计的相对状态表示和针对性奖励函数的支持下，多智能体强化学习框架能够解决航空器冲突，同时使航空器保持接近期望巡航速度。对编码器深度的评估显示，采用1层编码器的Transformer架构优于更深的配置，实现了近乎为零的近空中碰撞率，并缩短了间隔丧失时间。此外，我们还证明，该配置优于仅采用纯注意力构建的基线模型。未来可通过改进训练策略进一步提升我们方法的有效性，例如增加交叉航线的数量、允许改变航向或高度，或调整奖励函数。总体而言，我们的方法为未来高密度先进空中交通运行中的航空器间隔保障，推动更完善的多智能体强化学习方案和神经网络架构迈出了有希望的一步。

致谢

本研究得到美国国家科学基金会资助，资助编号 2312092。本文中表达的任何观点、研究结果、结论或建议均为作者所有，并不一定反映美国国家科学基金会的观点。

参考文献

[1] D.-T. Pham、N. P. Tran、S. Alam、V. Duong 和 D. Delahaye，《一种用于存在不确定性的密集交通场景冲突消解的机器学习方法》，载于第十三届美欧空中交通管理研究与发展研讨会，2019 年。[2] M. Brittain 和 P. Wei，《利用异构多智能体强化学习实现可扩展的自主间隔保障》，*IEEE*自动化科学与工程汇刊，第 19 卷，第 4 期，第 2837–2848 页，2022 年。[3] D. Groot、J. Ellerbroek 和 J. Hoekstra，《比较基于注意力的方法与长短期记忆在强化学习间隔管理中的状态编码效果》，《人工智能工程应用》，第 159 卷，文章编号 111592，2025 年。[4] A. Vaswani 等，《Attention is all you need》，载于神经信息处理系统进展，第 30 卷，2017 年。

[5] S. Hochreiter 和 J. Schmidhuber，《长短期记忆》，《神经计算》，第 9 卷，第 8 期，第 1735–1780 页，1997 年。[6] K. Cho 等，《使用 RNN编码器-解码器学习统计机器翻译的短语表征》，载于*2014* 年自然语言处理实证方法会议论文集，A. Moschitti、B. Pang 和 W. Daelemans 编，多哈，卡塔尔：计算语言学协会，2014 年 10 月，第 1724–1734 页。[7] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言理解的深度双向Transformer预训练》，载于*2019* 年北美计算语言学协会分会年会论文集：人类语言技术，第 *1* 卷（长篇与短篇论文），明尼阿波利斯，明尼苏达州：计算语言学协会，2019 年 6 月，第 4171–4186 页。[8] A. Dosovitskiy，《一张图像胜过 16×16 个单词：用于大规模图像识别的 Transformer》，*arXiv*预印本 *arXiv:2010.11929*，2020 年。[9] N. Reimers 和 I. Gurevych，《Sentence-BERT：使用孪生BERT网络的句子嵌入》，载于*2019* 年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议论文集（*EMNLP-IJCNLP*），2019 年。[10] M. Brittain 和 P. Wei，《高密度航路扇区中的自主间隔保障：一种深度多智能体强化学习方法》，载于*2019* 年 *IEEE*智能交通系统会议（*ITSC*），2019 年，第 3256–3262 页。[11] D. J. Groot、J. Ellerbroek 和 J. M. Hoekstra，《在空中交通管制中使用相对状态Transformer模型进行多智能体强化学习》，载于*SESAR*创新日，西班牙塞维利亚，2022 年 11 月。[12] D. Hendrycks 和 K. Gimpel，《高斯误差线性单元（GELU）》，*arXiv*预印本 *arXiv:1606.08415*，2016 年。[13] J. L. Ba、J. R. Kiros 和 G. E. Hinton，《层归一化》，*arXiv*预印本 *arXiv:1607.06450*，2016 年。[14] J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov，《近端策略优化算法》，*arXiv*预印本 *arXiv:1707.06347*，2017 年。[15] J. M. Hoekstra 和 J. Ellerbroek，《BlueSky 空管模拟器项目：一种开放数据与开源方案》，载于第七届国际航空运输研究会议论文集，美国联邦航空管理局/欧洲航空安全组织，美国华盛顿特区，第 131 卷，2016 年，第 132 页。[16] J. Schulman、P. Moritz、S. Levine、M. I. Jordan 和 P. Abbeel，《使用广义优势估计进行高维连续控制》，载于国际学习表征会议（*ICLR*）论文集，2016 年。