arXiv:2302.01586v1 [cs.RO] 3 Feb 2023

用于高度受限空间冲突消解的强化学习与分布式模型预测控制

徐深和Francesco Borrelli

摘要—本文提出一种分布式算法，用于解决高度受限空间中的协作式多车辆冲突。通过将冲突消解问题建模为多智能体强化学习（RL）问题，我们可以在简化的离散环境中离线训练策略，引导车辆安全、高效地驶向目的地。在线执行期间，每辆车首先利用训练好的策略模拟车辆间的交互，以获得自身策略，并据此指导参考轨迹的计算。随后，本文提出一种分布式模型预测控制器（MPC），用于跟踪参考轨迹并避免碰撞。初步结果表明，强化学习与分布式模型预测控制相结合，有望引导车辆安全、平稳地解决冲突，同时比集中式方法所需的计算量更少。

一、引言

自动驾驶车辆（AV）有望彻底改变我们生活的诸多方面，包括出行方式以及与环境的互动方式。然而，这一潜力也带来了一系列独特挑战：当车辆在高度受限环境中相遇时，如何解决冲突。简单的走走停停逻辑可能导致死锁，因为车辆可能已经彼此阻挡了通行路径。

为应对这些挑战，研究人员开发了多种冲突消解策略。强化学习（RL） [1] 是近年来日益受到关注的一种方法。通过从环境以及与其他智能体的交互中学习，自动驾驶车辆能够更好地理解每种动作所伴随的潜在风险，并做出长期奖励更高的决策。将冲突消解问题表述为马尔可夫决策过程（MDP）后，可以通过以等级-k 博弈的方式优化期望累积奖励 [2]，或利用深度神经网络近似状态-动作价值函数 [3], [4]，从而获得最优动作。现有的强化学习方法虽能探索多个智能体的组合动作，但通常需要对智能体的几何形状或动作空间进行近似。在高度受限空间中，这些低保真度近似会显得不足，因为车辆往往需要充分发挥其运动能力，才能在近距离障碍物之间机动。

基于优化的方法是解决自动驾驶车辆冲突的另一种途径。这些方法将车辆动力学纳入分析，在约束条件下优化轨迹。文献 [5]提出了一种集中式模型预测控制器（MPC），用于优化所有车辆的控制动作

美国加利福尼亚大学伯克利分校，美国加利福尼亚州（{徐深，f bor- relli}@berkeley.edu）。

交叉顺序的排列。为减轻计算负担，文献 [6],设计了一种采用约束优先级排序的分布式模型预测控制方法，并在文献 [7]中设计了一种采用交替方向乘子法（ADMM）的去中心化控制器。在高度受限空间中，可应用强对偶定理，精确表述任意形状两个凸集之间的碰撞规避问题 [8]。随后，文献 [9]将这一表述用于多机器人协同的分布式模型预测控制。

如文献 [10],所述，良好的初始猜测对于车辆在高度受限空间中寻找可行轨迹并完成机动动作至关重要。我们近期提出的方法 [11] 在离散环境中运用强化学习搜索构型策略，以此引导基于模型的集中式优化问题，在非线性、非完整车辆动力学和精确避碰约束下生成无冲突轨迹。然而，集中式优化计算开销很大，且无法适应在线执行过程中的不确定性。

在这项工作中，我们基于文献 [11],提出的方法，给出一种分布式算法，同时不再需要中央轨迹规划器。具体而言，

（i）首先，将冲突消解问题视为离散环境中的多智能体强化学习问题，在该环境中，我们可以收集所有智能体的轨迹采样，并离线训练共享策略。奖励函数旨在鼓励智能体尽快到达目的地，并对任何可能发生的碰撞予以惩罚；（ii）面对具体场景时，每辆车都使用一份训练好的策略副本，在离散环境中模拟所有车辆之间的交互。模拟结果作为高层策略，用于引导参考轨迹的计算；（iii）在线执行期间，每辆车通过实时求解分布式模型预测控制问题来跟踪其参考轨迹。为确保安全，将车辆之间以及车辆与静态障碍物之间的避碰约束纳入考虑。

本文结构安排如下：第 II 节对问题进行形式化定义。第 III 节简要介绍我们在 [11] 中提出的方法，该方法利用深度强化学习生成高层策略以解决冲突。第 IV 节详细阐述分布式模型预测控制器的构建。第 V 节介绍分布式冲突消解的完整算法。第 VI 节展示一个示例场景的初步结果。最后，第 VII 节总结全文。

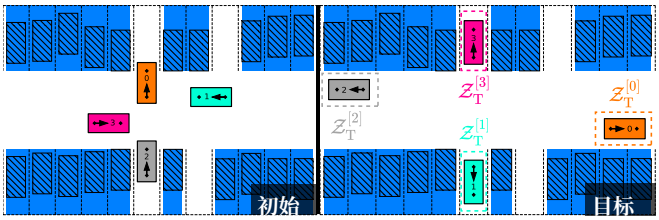


图 1：一个示例场景。静态障碍物以蓝色矩形表示，它们是对带有斜线阴影的停放车辆的过度近似。该场景涉及的四辆车辆 0–3 分别以橙色、青色、灰色和洋红色绘出，短箭头从后轴中心指向前轴中心，表示车辆的航向。车辆从各自的初始状态出发，并以到达其目标集 $\mathcal{Z}^{[i]}{}_{\mathrm{T}}$ $i \in \{0,1,2,3\}$ 为目标。

II. 问题表述

A. 假设

考虑在具有 M 个静态障碍物的高度受限环境中，由 $i \in \mathcal{I} = \{0,\dots,N_{\mathrm{v}}-1\}$ 编号的 N_{v} 辆自动驾驶车辆，$m \in \mathcal{M} = \{0,\dots,M-1\}$。作出以下假设：

(i) 所有车辆都拥有当前环境的地图，并且能够准确地定位自身；(ii) 所有车辆均为全自动车辆，能够相互通信并交换信息；(iii) 所有车辆的车身尺寸和动力学特性均相同。

B. 问题陈述

每辆车的状态 z 由其后轴中心的位置（x，y）、航向角 ψ、速度v和前轮转向角 δ_f描述。在冲突消解场景中，每辆车 $i \in \mathcal{I}$ 从初始状态 $z_0^{[i]}$ 静止出发，并以到达目标集 $\mathcal{Z}^{[i]}{}_{\mathrm{T}}$ 为目标，同时避免与其他车辆或障碍物发生碰撞。图 1 展示了一个包含四辆车的示例场景。

每辆车 i 的几何形状是根据其尺寸和实时状态 $z^{[i]}$计算得出的多面体

$$\mathbb{B}(z^{[i]}) := \{p \in \mathbb{R}^2 : G(z^{[i]})p \le g(z^{[i]})\},\ i \in \mathcal{I}, \quad (1)$$

类似地，静态障碍物也用多胞体描述

$$\mathbb{O}^{[m]} := \{p \in \mathbb{R}^2 : A^{[m]}p \le b^{[m]}\},\ m \in \mathcal{M}. \quad (2)$$

由于车辆以低速行驶，其动力学遵循运动学自行车模型：

$$\dot{z} = f(z,u) := \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\psi} \\ \dot{v} \\ \dot{\delta_f} \end{bmatrix} = \begin{bmatrix} v\cos(\psi) \\ v\sin(\psi) \\ \frac{v}{l_{\mathrm{wb}}}\tan(\delta_f) \\ a \\ \omega \end{bmatrix},\ u = \begin{bmatrix} a \\ \omega \end{bmatrix}, \quad (3)$$

其中，输入 u 由加速度 a 和转向角速度 ω组成。参数 l_{wb} 表示轴距。该离散时间动力学

$$z_{k+1} = f_{\mathrm{dt}}(z_k, u_k, \tau) \quad (4)$$

通过使用四阶 Runge-Kutta 方法对式（3）进行离散化，可得到该式，其中 $\tau > 0$ 为采样时间，z_k，u_k分别为离散时间步 k 的状态和输入。

C. 集中式模型预测控制

给定一条参考轨迹 $\mathbf{z}^{[i]}{}_{\mathrm{ref}}$，它从每辆车 $i \in \mathcal{I}$ 的初始状态 $z_0^{[i]}$ 出发并到达目标集 $\mathcal{Z}^{[i]}{}_{\mathrm{T}}$，可以针对控制构建一个集中式模型预测控制问题：

$$\min_{\mathbf{z}^{[i]}_{\cdot|t},\mathbf{u}^{[i]}_{\cdot|t}} \sum_{i\in\mathcal{I}} J^{[i]}\left(\mathbf{z}^{[i]}_{\cdot|t}, \mathbf{u}^{[i]}_{\cdot|t} \mid \mathbf{z}^{[i]}_{\mathrm{ref}}\right)$$

$$\text{s.t. } z^{[i]}_{k+1|t} = f_{\mathrm{dt}}\left(z^{[i]}_{k|t}, u^{[i]}_{k|t}, \tau\right), \quad (5a)$$

$$z^{[i]}_{0|t} = z^{[i]}(t) \quad (5b)$$

$$z^{[i]}_{k|t} \in \mathcal{Z}, u^{[i]}_{k|t} \in \mathcal{U}, \quad (5c)$$

$$\mathrm{dist}\left(\mathbb{B}(z^{[i]}_{k|t}), \mathbb{O}^{(m)}\right) \ge d_{\min}, \quad (5d)$$

$$\mathrm{dist}\left(\mathbb{B}(z^{[i]}_{k|t}), \mathbb{B}(z^{[j]}_{k|t})\right) \ge d_{\min}, \quad (5e)$$

$$\forall i,j \in \mathcal{I}, i \ne j, m \in \mathcal{M}, k \in 0,\dots,N,$$

其中，$\mathbf{z}^{[i]}_{\cdot|t} = \left\{z^{[i]}_{0|t},\dots,z^{[i]}_{N|t}\right\}$ 和 $\mathbf{u}^{[i]}_{\cdot|t} = \left\{u^{[i]}_{0|t},\dots,u^{[i]}_{N|t}\right\}$ 分别表示车辆 i在模型预测控制前瞻时域 N 内的状态序列和控制输入序列。车辆状态和输入受限于可行集 $\mathcal{Z}$ 和 $\mathcal{U}$。约束 (5d) 和 (5e)用于确保车辆与静态障碍物之间的碰撞规避。其中，$\mathrm{dist}(\cdot,\cdot) \in \mathbb{R}$表示两个多面体之间的距离，$d_{\min}$ 为安全阈值。

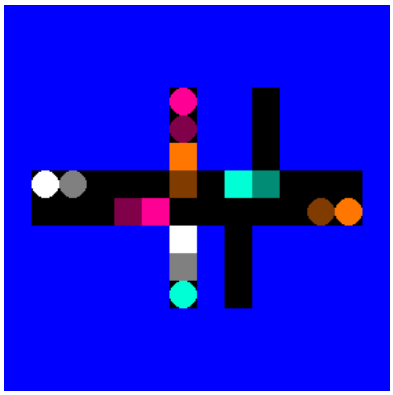
有几个挑战阻碍了我们有效应用这一集中式方法：

(i) 很难找到这样的参考轨迹$\mathbf{z}^{[i]}{}_{\text{参考}}$，使车辆沿着这些轨迹行驶时，能够轻松找到平滑的机动动作以避免碰撞并保持可行；(ii) 约束(5d)和(5e)使问题(5)成为双层优化问题，因此难以使用现成的求解器求解；(iii) 由于需要在所有车辆与所有障碍物之间进行成对避碰，中央指挥者迭代求解问题(5)的计算开销很大，因此不适合在线部署。

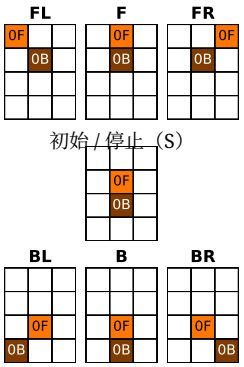
在本文余下部分，我们将详细介绍如何基于强化学习生成的策略计算高质量参考轨迹，并对问题(5)进行分布式重构，使其能够满足实时要求。

III. 使用强化学习生成策略

本节介绍一种为车辆生成策略以计算其参考轨迹的方法。这些策略表示为战术车辆配置序列，并通过将冲突消解问题构建为离散多智能体强化学习问题来计算。本节部分内容摘自 [11]



（a）离散网格环境



（b）网格中的动力学

图 2：离散环境中的车辆。（a）车辆车身以正方形表示，目的地以圆形表示。橙色、青色、白色/灰色和洋红色分别代表不同的车辆，亮度变化用于区分车辆的前侧和后侧。障碍物以蓝色表示，自由空间以黑色表示。（b）自由网格地图中的单车动力学。“0F”和“0B”标记车辆0的前侧和后侧。

此处仅为完整性和可读性起见予以报告。关于策略生成的深入讨论，请读者参阅 [11]。

A. 状态与观测空间

在离散强化学习问题中，每辆车的状态和目的地由其“前侧”和“后侧”在离散网格环境中所占据的方格表示。图 2a 展示了与图 1 场景示例的初始条件和目标集相对应的离散表示。此外，智能体将图 2a 这类 RGB图像作为当前步骤的观测。

B. 动作空间

每辆车在网格地图中可以采取 7 种离散动作：{停止（S）、前进（F）、左前进（FL）、右前进（FR）、后退（B）、左后退（BL）、右后退（BR）}。图 2b 展示了单辆车在自由网格地图中的相应状态转移。当车辆与任何障碍物或其他车辆发生碰撞时，它会被“弹回”，其状态保持不变。

C. 奖励

奖励旨在对碰撞、“停止（S）”动作、车辆与目的地之间的距离以及时间消耗进行惩罚。车辆到达目的地时还会获得巨额正奖励。

D. 策略学习

通过在大量仿真中收集所有车辆的轨迹采样，我们训练了一个参数为 θ 的基于CNN的深度Q网络，以近似观测-动作价值函数$Q(o,a|\theta)$，其中 o 是类似于图2a的RGB图像， a 是第三-B节中介绍的动作之一。策略表示为 $\pi_\theta = \text{argmax}_a Q(o,a|\theta)$。

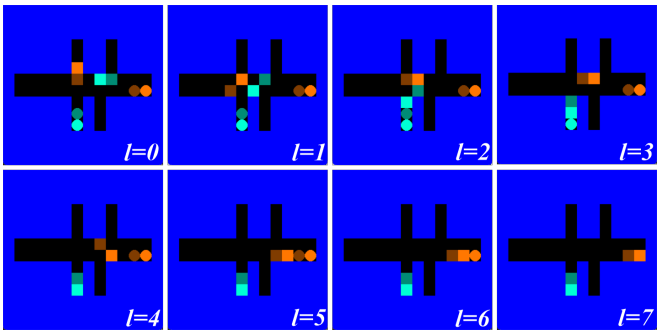


图 3：两辆车解决冲突的策略

E. 策略

给定特定的冲突场景，每辆车都可以使用训练好的策略π_{θ^*}，在离散环境中模拟所有车辆的交互，直至冲突得到解决。图3展示了一个双车辆示例。我们使用“策略”来指代离散网格环境中由方格反映的车辆构型序列。通过应用变换，网格地图中的方格变为地面坐标中的策略引导集合

$$\mathbf{Z}^{[i]} = \left\{ \bar{\mathcal{Z}}^{[i]}_{\mathrm{F},l}, \bar{\mathcal{Z}}^{[i]}_{\mathrm{B},l} \right\}_{l=0}^{L^{[i]}-1}$$

其中， $L^{[i]}$ 是车辆 i到达其目的地所需的步数， $\bar{\mathcal{Z}}^{[i]}{}_{\mathrm{F}\,l} \subset \mathbb{R}^{2\times 2}$ 和 $\bar{\mathcal{Z}}^{[i]}{}_{\mathrm{B}\,l} \subset \mathbb{R}^{2\times 2}$ 分别是车辆 i 在第 l 步时前部和后部的凸集。

将策略步骤 l和 $l+1$之间的时间间隔记为 T_s，连续空间中的车辆构型受到引导，使得在时间 $T_\mathrm{s}l$，车辆后轴中心（$x^{[i]}, y^{[i]}$）位于集合 $\bar{\mathcal{Z}}^{[i]}{}_{\mathrm{B}\,l}$ 内，而车辆前轴中心（$x^{[i]}{}_{\mathrm{F}}$ $y^{[i]}{}_{\mathrm{F}}$）位于集合 $\bar{\mathcal{Z}}^{[i]}{}_{\mathrm{F}\,l}$ 内，形式化表示为

$$\left[x^{[i]}(T_\mathrm{s}l), y^{[i]}(T_\mathrm{s}l)\right]^\top \in \bar{\mathcal{Z}}^{[i]}_{\mathrm{B},l} \tag{6a}$$

$$\left[x^{[i]}_\mathrm{F}(T_\mathrm{s}l), y^{[i]}_\mathrm{F}(T_\mathrm{s}l)\right]^\top \in \bar{\mathcal{Z}}^{[i]}_{\mathrm{F},l}, \tag{6b}$$

$$x^{[i]}_\mathrm{F}(T_\mathrm{s}l) = x^{[i]}(T_\mathrm{s}l) + l_\mathrm{wb}\cos\left(\psi^{[i]}(T_\mathrm{s}l)\right) \tag{6c}$$

$$y^{[i]}_\mathrm{F}(T_\mathrm{s}l) = y^{[i]}(T_\mathrm{s}l) + l_\mathrm{wb}\sin\left(\psi^{[i]}(T_\mathrm{s}l)\right) \tag{6d}$$

$$\forall l = 0,\ldots,L^{[i]}-1, i \in \mathcal{I},$$

IV. 策略引导的分布式模型预测控制

本节介绍我们的方法：通过重新表述碰撞规避（CA）约束并利用策略引导约束 (6)，使每辆车辆 i 都能计算其参考轨迹 $\mathbf{z}^{[i]}{}_\mathrm{ref}$。随后，我们为每辆车辆 i 提出一种分布式模型预测控制方法，使其能够实时独立求解控制输入。

A. 针对静态障碍物的碰撞规避（*CA*）

根据 [8], [9],车辆 $i \in \mathcal{I}$ 与静态障碍物 $m \in \mathcal{M}$ 之间的 CA 约束 (5d)，可将其

重构为一个可行性问题：

$$\exists \lambda_m^{[i]} \geq 0, \mu_m^{[i]} \geq 0, s^{[i,m]} \ : \|s^{[i,m]}\| \leq 1, \tag{7a}$$
$$-g(z^{[i]})^\top \lambda_m^{[i]} - b^{[m],\top} \mu_m^{[i]} \geq d_{\min}, \tag{7b}$$
$$G(z^{[i]})\lambda_m^{[i]} + s^{[i,m]} = 0, \tag{7c}$$
$$A^{[m]}\mu_m^{[i]} - s^{[i,m]} = 0, \tag{7d}$$

其中 $G(\cdot)$ $g(\cdot)$ $A^{[m]}$, $b^{[m]}$ 描述了式 (1)、(2) 中引入的多胞体 $\mathbb{B}(z^{[i]})$ $\mathbb{O}^{[m]}$。

B. 基于优化的参考计算

通过结合策略引导约束 (6) 和避碰约束 (7)，我们可以通过求解最优控制问题，计算每辆车辆 i 的参考轨迹 $\mathbf{z}^{[i]}{}_{\text{ref}}$

$$\min_{\substack{\mathbf{z}^{[i]},\mathbf{u}^{[i]} \\ \boldsymbol{\lambda}^{[i]},\boldsymbol{\mu}^{[i]},\mathbf{s}^{[i,\cdot]}}} \quad \int_{t=0}^{T_{\text{s}}^{[i]}L^{[i]}} c\left(z^{[i]}(t), u^{[i]}(t)\right) dt$$
$$\text{s.t.} \quad \dot{z}^{[i]}(t) = f(z^{[i]}(t), u^{[i]}(t)), \tag{8a}$$
$$z^{[i]}(t) \in \mathcal{Z}, u^{[i]}(t) \in \mathcal{U}, \tag{8b}$$
$$z^{[i]}(0) = z_0^{[i]}, \tag{8c}$$
$$z^{[i]}(T_{\text{s}}L^{[i]}) \in \mathcal{Z}_{\text{T}}^{[i]}, \tag{8d}$$
$$\text{策略引导约束 (6)},$$
$$\text{针对障碍物 } m \text{ 的碰撞规避 (7)},$$
$$\forall 0 \leq t \leq T_{\text{s}}L^{[i]}, m \in \mathcal{M}.$$

直观而言，参考轨迹$\mathbf{z}^{[i]}{}_{\text{ref}}$的计算方式满足以下条件：(i) 遵循车辆动力学约束 (8a) 以及状态和输入限制 (8b)，因此运动学可行；(ii) 从初始状态 $z_0^{[i]}$ (8c) 出发，并最终到达目标集 $\mathcal{Z}^{[i]}{}_{\text{T}}$ (8d)；(iii) 按顺序遵循$\mathbf{Z}^{[i]}$ 所定义的策略引导配置，时间间隔为 T_{s}；(iv) 避开环境中的所有静态障碍物。请注意，由于无法获取其他车辆的状态，计算参考轨迹时不会施加车辆间碰撞规避约束。

尽管采用连续时间形式，问题 (8) 仍可通过有限元上的正交配置法进行离散化 [12], ，其中区间数量由策略步骤数量 $L^{[i]}$ 决定，每个区间的长度为 T_{s}。配点法所用的插值多项式也用于在线控制期间对车辆参考状态进行采样。

C. 与其他车辆的避碰

类似于式 (7)，车辆对 i、$j \in \mathcal{I}$、$i \neq j$ 之间的 CA 约束 (5e) 被重新表述为

$$\exists \lambda_j^{[i]} \geq 0, \mu_j^{[i]} \geq 0, s^{[i,j]} \ : \|s^{[i,j]}\| \leq 1, \tag{9a}$$
$$-g(z^{[i]})^\top \lambda_j^{[i]} - g(z^{[j]})^\top \mu_j^{[i]} \geq d_{\min}, \tag{9b}$$
$$G(z^{[i]})\lambda_j^{[i]} + s^{[i,j]} = 0, \tag{9c}$$
$$G(z^{[j]})\mu_j^{[i]} - s^{[i,j]} = 0. \tag{9d}$$

D. 分布式模型预测控制

在时间 t，每辆车辆 i 独立计算自身在时域 N 内的状态和输入轨迹 $\mathbf{z}_{\cdot|t}^{[i]}$、$\mathbf{u}_{\cdot|t}^{[i]}$，方法是求解以下优化问题

$$\min_{\substack{\mathbf{z}_{\cdot|t}^{[i]},\mathbf{u}_{\cdot|t}^{[i]} \\ \boldsymbol{\lambda}_{\cdot|t}^{[i]},\boldsymbol{\mu}_{\cdot|t}^{[i]},\mathbf{s}_{\cdot|t}^{[i,\cdot]}}} \quad J^{[i]}\left(\mathbf{z}_{\cdot|t}^{[i]}, \mathbf{u}_{\cdot|t}^{[i]} \mid \mathbf{z}_{\text{ref}}^{[i]}\right)$$
$$\text{s.t.} \quad z_{k+1|t}^{[i]} = f_{\text{dt}}\left(z_{k|t}^{[i]}, u_{k|t}^{[i]}, \tau\right), \tag{10a}$$
$$z_{0|t}^{[i]} = z^{[i]}(t) \tag{10b}$$
$$z_{k|t}^{[i]} \in \mathcal{Z}, u_{k|t}^{[i]} \in \mathcal{U}, \tag{10c}$$
$$\text{针对障碍物 } m \text{ 的碰撞规避 (7)}$$
$$\text{针对车辆 } j \text{ 的碰撞规避 (9)，给定 } \bar{\mathbf{z}}_{\cdot|t}^{[j]}$$
$$\forall j \in \mathcal{I}\backslash\{i\}, m \in \mathcal{M}, k = 0, \ldots, N,$$

其中，$\bar{\mathbf{z}}_{\cdot|t}^{[j]}$ 表示来自另一辆车辆的已知前瞻预测 j，该预测通过车辆间通信获得。上述分布式形式解决了集中式形式 (5) 面临的挑战，具体方式如下：

(i) 遵循由策略引导的“高质量”参考轨迹$\mathbf{z}^{[i]}{}_{\text{ref}}$，该轨迹已通过策略引导编码了解决冲突所需的机动动作；(ii) 使用平滑的 CA 约束，使问题能够通过基于梯度或海森矩阵的求解器求解；(iii) 仅对车辆 i 求解决策变量，从而具备实时处理能力。此外，问题的复杂度仅随障碍物和其他车辆的数量线性增长。

V. 分布式冲突消解算法

本节介绍用于高度受限空间中分布式冲突消解的算法1。

第2行至第5行对应于第III节所述的过程。在获取所有车辆的初始状态 $z_0^{[i]}$ 和目标集 $\mathcal{Z}^{[i]}{}_{\text{T}}$ 后，每辆车辆 i 都会如图2a所示建立离散网格环境，以模拟所有车辆之间的交互，直至其到达目的地。需要注意的是，对于每个场景，所有车辆都会创建相同的离散环境，并为所有车辆实施相同的策略 π_{θ^*}。因此，一辆车辆获得的结果将与其他任何车辆获得的结果完全一致。车辆 i 自身执行的离散步序列会被记录为一种策略，用于生成集合$\mathbf{Z}^{[i]}$，以引导车辆构型。

根据 (8) 计算出参考轨迹 $\mathbf{z}^{[i]}{}_{\text{ref}}$ 后，每辆车辆 i 将首先沿其模型预测控制时域，从 $\mathbf{z}^{[i]}{}_{\text{ref}}$ 插值出状态 N，作为其初始前瞻预测 $\mathbf{z}_{\cdot|-1}^{[i]}$，并将其广播给所有其他车辆，使所有车辆都能从 $t = 0$ 开始求解问题 (10)。

在时间 t，从时间 $t-1$接收到的关于另一辆车辆 j 的最新信息是 $\mathbf{z}_{\cdot|t-1}^{[j]} = \left\{z_{0|t-1}^{[j]}, \ldots, z_{N|t-1}^{[j]}\right\}$。因此，在第 13 行，我们将 $\mathbf{z}_{\cdot|t-1}^{[j]}$ 向前移动一步，使得 $\bar{\mathbf{z}}_{\cdot|t}^{[j]} = \left\{z_{1|t-1}^{[j]}, \ldots, z_{N|t-1}^{[j]}, z_{N|t-1}^{[j]}\right\}$。此外，由于

表 I：初始位姿和最终位姿的约束

i	初始位姿 $(x,\ y,\ \psi)$	最终 x	最终 y	最终 ψ
0	$(16.25,\ 18.75,\ -\frac{1}{2}\pi)$	$[27.5, 30]$	$[15, 17.5]$	0
1	$(23_{\circ}75,\ 18_{\circ}75,\ \pi)$	$[15, 17.5]$	$[10, 12.5]$	$-\frac{1}{2}\pi$
2	$(16.25, 11.25, -\frac{1}{2}\pi)$	$[5, 7.5]$	$[17.5, 20]$	π
3	$(11.25, 16.25, 0)$	$[15, 17.5]$	$[22.5, 25]$	$\frac{1}{2}\pi$

非凸非线性规划问题 (10) 的解在很大程度上取决于初始猜测，因此我们将从时间 $t-1$ 起的最优解 $\mathbf{z}^{[i]}_{\cdot|t-1}, \mathbf{u}^{[i]}_{\cdot|t-1}, \boldsymbol{\lambda}^{[i]}_{\cdot|t-1}, \boldsymbol{\mu}^{[i]}_{\cdot|t-1}, \mathbf{s}^{[i,\cdot]}_{\cdot|t-1}$ 向前移动，以便在时间 t 热启动求解器。

算法1 分布式冲突消解

要求： 初始状态 $z^{[i]}_0$ 和目标集 $\mathcal{Z}^{[i]}_{\mathrm{T}}$ 所有车辆的 $i \in \mathcal{I}$、静态障碍物、训练好的强化学习策略 π_{θ^*}

对所有车辆 $i \in \mathcal{I}$ 并行**执行**
  构建包含 $z^{[i]}_0, \forall i \in \mathcal{I}$ 和障碍物
  当车辆 i 尚未到达其目的地时**执行**
    使用 π_{θ^*} 模拟车辆交互
    将每一步记录为车辆 i 的策略
  根据记录的策略计算集合 $\mathbf{Z}^{[i]}$
  计算参考轨迹 $\mathbf{z}^{[i]}_{\mathrm{ref}}$ 通过求解 (8)
  计算初始 $\mathbf{z}^{[i]}_{\cdot|-1}$ 沿 $\mathbf{z}^{[i]}_{\mathrm{ref}}$ 采样
  向其他车辆广播 $\mathbf{z}^{[i]}_{\cdot|-1}\ \ j \in \mathcal{I}\backslash\{i\}$
  初始化 $\boldsymbol{\lambda}^{[i]}_{\cdot|-1}, \boldsymbol{\mu}^{[i]}_{\cdot|-1}, \mathbf{s}^{[i,\cdot]}_{\cdot|-1}$ 用于 CA

对t = 0, 1, . . ., ∞ **do**
  对所有车辆 $i \in \mathcal{I}$ 并行**执行**
    移动最新的 $\mathbf{z}^{[j]}_{\cdot|t-1}$ 以获得 $\mathbf{z}^{[j]}_{\cdot|t}$，对于所有 $j \in \mathcal{I}\backslash\{i\}$
    计算 $\mathbf{z}^{[i]}_{\cdot|t}, \mathbf{u}^{[i]}_{\cdot|t}, \boldsymbol{\lambda}^{[i]}_{\cdot|t}, \boldsymbol{\mu}^{[i]}_{\cdot|t}, \mathbf{s}^{[i,\cdot]}_{\cdot|t}$ 通过求解 (10)
    向其他车辆广播 $\mathbf{z}^{[i]}_{\cdot|t}\ \ j \in \mathcal{I}\backslash\{i\}$
    应用 $u^{[i]}_{0|t}$ 向前移动

VI. 结果

下一节将详细介绍图 1 所示的冲突场景，以及所提出的分布式算法 1 的仿真结果。为使说明尽可能清晰，我们选取了图 1 中的场景，因为它是所有测试场景中最复杂的一个。其他场景的源代码和结果请参见：https://bit.ly/rl-cr。

整个停车场区域的大小为 30m × 20m，其中大多数车位已被占用，因此形成了高度受限的环境。车辆在其初始状态 $z^{[i]}_0$ 和目标集 $\mathcal{Z}^{[i]}_{\mathrm{T}}$ 中的位姿如表 I 所示。所有其他状态和输入分量在初始时间步均为 0。车辆车身为 3.9m × 1.8m 的矩形，轴距为 2.5m。车辆的状态和输入均受运行限制约束，即 $v^{[i]} \in [-2.5, 2.5]$m/s、$\delta^{[i]}_f \in [-0.85, 0.85]$ rad、$a^{[i]} \in [-1.5, 1.5]$m/s^2、$w^{[i]} \in [-1, 1]$rad/s。在本文中，我们将安全阈值设为 $d_{\min} = 0.05$m。优化问题采用 CasADi [13] 编写，并通过 IPOPT [14] 和线性求解器 HSL MA97 [15] 求解。

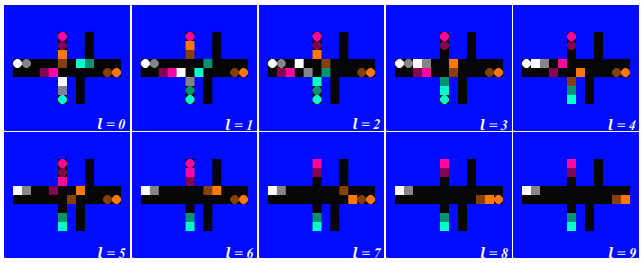


图4：每辆车在离散环境中解决冲突的步骤。

A. 策略引导的参考轨迹

通过使用强化学习策略 π_{θ^*} 在离散环境中模拟车辆交互，每辆车都会获得一系列离散步序列，作为其解决冲突的策略，如图4所示。从图中可以看出，车辆0（橙色）首先向前驶入上方空位，为其他车辆腾出空间，然后倒车，并利用下方空位改变航向角；车辆1（青色）和车辆2（灰色）立即驶向各自的目的地；车辆3（洋红色）则先倒车以避免碰撞，然后驶向上方空位。每辆车所需的策略步数 $L^{[i]}$ 分别为 $L^{[0]} = 10,\ L^{[1]} = 5,\ L^{[2]} = 6,\ L^{[3]} = 8$。

对应图4中的策略，计算出不同车辆的策略引导集合 $\mathbf{Z}^{[i]}$，并在图5中以橙色、青色、灰色、洋红色方块表示。我们使用问题（8）中的阶段代价 $c(z,\ u) = \psi^2 + v^2w^2 + a^2 + 1$ 来描述乘坐舒适性、执行量大小和时间消耗。两个策略步骤之间的时间间隔为 $T_{\mathrm{s}} = 3$ 秒。我们采用 5 阶拉格朗日插值多项式和 Gauss-Radau根进行配点。通过求解问题（8），得到所有车辆的参考轨迹 $\mathbf{z}^{[i]}_{\mathrm{ref}}$，如图5所示。我们还绘制了 $t = lT_{\mathrm{s}}$、$l = 0$、.....、9时刻的车辆构型，以展示策略引导约束（6）的效果。请注意，由于我们仅在问题（8）中施加避碰约束（7），因此无法保证这些参考轨迹彼此之间不会发生碰撞；图5中时间 $t = 3T_{\mathrm{s}}$ 时的车辆构型也反映了这一点。

B. 分布式在线控制

计算出参考轨迹 $\mathbf{z}^{[i]}_{\mathrm{ref}}$ 后，车辆通过求解分布式模型预测控制问题（10）来跟踪参考值，具体如算法1第11行至第16行所述。离散时间车辆动力学（4）的采样时间为 $\tau = 0.1$秒，模型预测控制前瞻时域为 $N = 30$。代价函数 $J^{[i]}$ 定义为

$$J^{[i]}\left(\mathbf{z}^{[i]}_{\cdot|t}, \mathbf{u}^{[i]}_{\cdot|t} \mid \mathbf{z}^{[i]}_{\mathrm{ref}}\right) = \sum_{k=0}^{N} \|z^{[i]}_{k|t} - z^{[i]}_{\mathrm{ref},k|t}\|^2_Q \left(\psi^{[i]}_{k|t}\right)^2 + \left(v^{[i]}_{k|t} w^{[i]}_{k|t}\right)^2 + \left(a^{[i]}_{k|t}\right)^2,$$

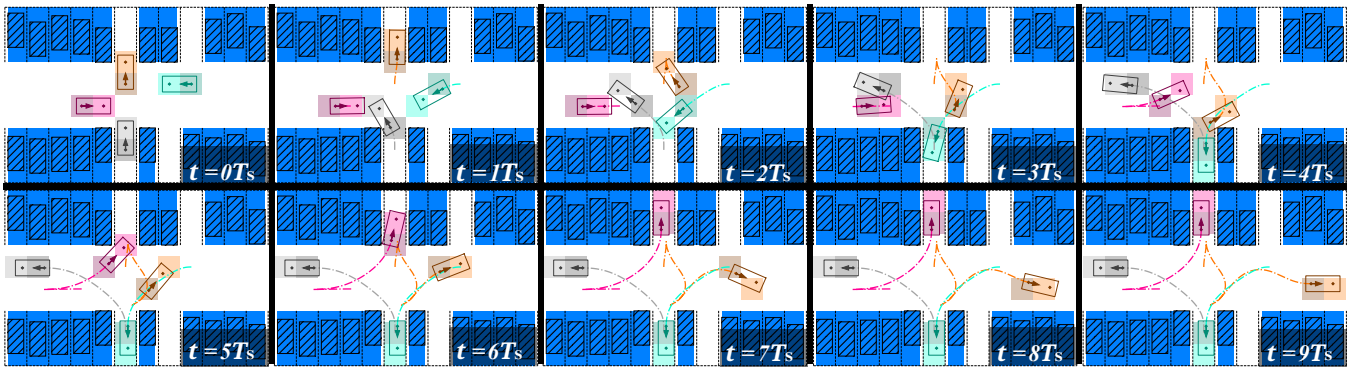


图5：策略引导集合$\mathbf{Z}^{[i]}$时间 $t = lT_s$，$l = 0, \ldots, 9$ 以及由此得到的车辆参考轨迹。集合、车辆车身和轨迹分别以橙色、青色、灰色和洋红色表示车辆0、车辆1、车辆2和车辆3。

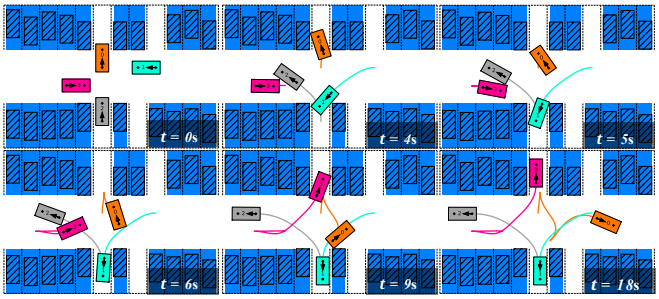


图6：在线控制期间的快照

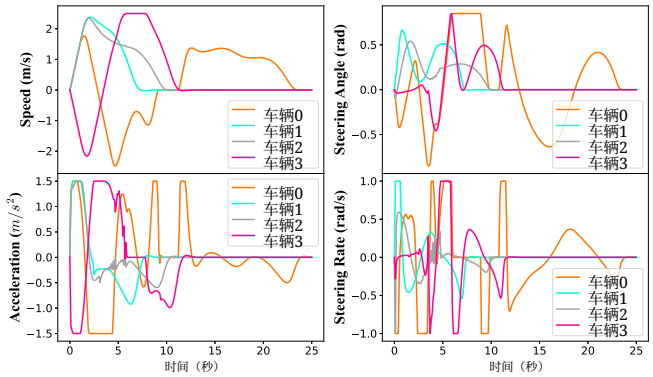
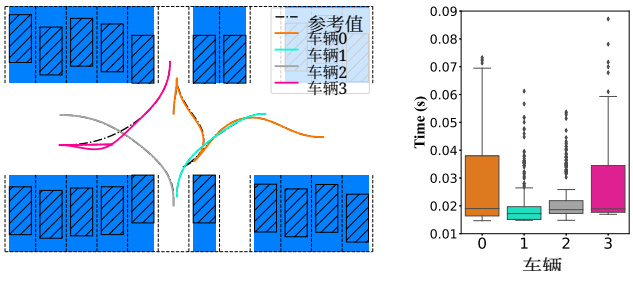


图 7：车辆速度、前轮转向角、加速度和转向速率的变化曲线



（a）参考值和最终轨迹　　（b）迭代时间

图 8：跟踪性能和迭代时间统计

该函数反映了与参考轨迹的偏差、乘坐舒适性以及执行量的大小。$\left\{z^{[i]}{}_{\text{ref}}\,0|t, \ldots, z^{[i]}{}_{\text{ref}}\,N|t\right\}$ 沿着$\mathbf{z}^{[i]}{}_{\text{ref}}$进行采样，权重矩阵 $Q = \text{diag}([100, 100, 100, 0, 0])$。图6展示了在线控制过程中车辆构型的一些快照。从图7可以看出，控制器为车辆生成了平滑的速度和转向曲线，同时将输入保持在运行限制范围内。图8a比较了车辆的最终轨迹与其参考轨迹，可以发现，车辆大多数时候都能准确跟踪参考值，只有车辆3为避免与车辆2发生碰撞，不得不暂时偏离轨迹。

我们在图 8b 中记录了每辆车辆求解问题（10）所需的计算时间。数据通过在一台配备四核 Intel® CoreTM i9-9900K 中央处理器（@ 3.60GHz）的笔记本电脑上测试所提出的算法获得。所有车辆的迭代时间中位数约为 0.02 秒，其中 98.2% 低于 0.1 秒。异常值中最长的一次迭代持续了 0.5 秒。

VII. 结论

本文提出一种分布式算法，结合深度多智能体强化学习（RL）与分布式模型预测控制（MPC），以解决高度受限环境中的冲突。

离线时，我们可以使用深度强化学习训练策略，在离散环境中探索车辆的组合动作以消解冲突。训练好的策略可以针对特定场景提供离散指导，从而生成高质量的参考轨迹。在线时，我们构建分布式模型预测控制，以跟踪参考轨迹并规避车辆间的碰撞。在每个时间步，车辆都会计算前瞻时域内的状态和输入，并与其他车辆共享预测结果。仿真结果表明，所提出的算法能够实时控制车辆，以平稳的运动曲线安全、高效地消解冲突。

参考值

[1] Z. Wang、W. Pan、H. Li、X. Wang 和 Q. Zuo，《空中交通管制中冲突消解的深度强化学习方法综述》，《航空航天》，第 9 卷，第 6 期，第 294 页，2022 年 6 月，期号：6，出版方：多学科数字出版研究所。 [在线]。网址：https://www.mdpi.com/ 2226 4310 9 6 294 - / / / N. Li、I. Kolmanovsky、A. Girard 和 Y. Yildiz，《无信号交叉路口车辆交互的博弈论建模及其在自动驾驶车辆控制中的应用》，载于*2018* 年美国控制会议（*ACC*），2018 年 6 月，第 3215–3220 页，ISSN：2378-5861。[3] S. Li、M. Egorov 和 M. Kochenderfer，《使用深度强化学习优化繁忙空域中的碰撞规避》，2019 年 12 月。[4] M. Yuan、J. Shan 和 K. Mi，《基于深度强化学习的自动驾驶车辆博弈论决策》，*IEEE*机器人与自动化快报，第 7 卷，第 2 期，第 818–825 页，2022 年 4 月。[5] L. Riegger、M. Carlander、N. Lidander、N. Murgovski 和 J. Sj¨oberg，《用于自动交叉路口通行的集中式模型预测控制》，载于*2016* 年 *IEEE* 第 *19* 届智能交通系统国际会议（*ITSC*），2016 年 11 月，第 1372–1377 页，ISSN：2153-0017。[6] A. Katriniok、P. Kleibaum 和 M. Jo˘sevski，《使用并行化优化方法实现交叉路口自动化的分布式模型预测控制》，*IFAC-PapersOnLine*，第 50 卷，第 1 期，第 5940–5946 页，2017 年 7 月。[7] F. Rey、Z. Pan、A. Hauswirth 和 J. Lygeros，《用于协调与碰撞规避的完全分散式交替方向乘子法》，载于*2018* 年欧洲控制会议（*ECC*），2018 年 6 月，第 825–830 页。[8] X. Zhang、A. Liniger 和 F. Borrelli，《基于优化的碰撞规避》，*IEEE*控制系统技术汇刊，第 1–12 页，2020 年，arXiv：1711.03449。[9] R. Firoozi、L. Ferranti、X. Zhang、S. Nejadnik 和 F. Borrelli，《用于狭窄环境导航的分布式多机器人协调算法》，2020 年 6 月，arXiv:2006.11492 [cs]。 [在线]。网址：http://arxiv.org/abs/ 2006 11492 。X. Zhang、A. Liniger、A. Sakai 和 F. Borrelli，《使用基于优化的碰撞规避实现自主泊车》，载于*IEEE*决策与控制会议论文集，第 2018-Decem 卷。IEEE，2019 年 12 月，第 4327–4332 页，ISSN：07431546。[11] X. Shen 和 F. Borrelli，《融合最优控制与强化学习，在高度受限空间中实现多车辆冲突消解》，2022 年 11 月，arXiv:2211.01487 [cs, eess]。 [10] [ ] 2211 01487 在线。网址：http://arxiv.org/abs/。L. T. Biegler，《非线性规划：概念、算法及其在化学过程中的应用》。美国：工业与应用数学学会，2010 年。[13] J. A. E. Andersson、J. Gillis、G. Horn、J. B. Rawlings 和 M. Diehl，《CasADi：用于非线性优化和最优控制的软件框架》，《数学规划计算》，第 11 卷，第 1 期，第 1–36 页，2019 年 3 月。[14] A. W¨achter 和 L. T. Biegler，《大型非线性规划问题的内点滤波线搜索算法实现》，《数学规划》，第 106 卷，第 1 期，第 25–57 页，2006 年 3 月。[15] T. Rees，《HSL：用于大规模科学计算的 Fortran代码集》。2022 年 7 月。 [在线]。网址：http://www.hsl.rl.ac.uk/ [12]