网联驾驶环境中基于深度强化学习的交通整形与迟滞缓解

预印本

Rami Ammourah 土木与环境工程系 伊利诺伊大学厄巴纳-香槟分校 厄巴纳，伊利诺伊州 61801 ramiaa2@illinois.edu

Alireza Talebpour 土木与环境工程系 伊利诺伊大学厄巴纳-香槟分校 厄巴纳，伊利诺伊州 61801 ataleb@illinois.edu

2023年2月6日

摘要

本文提出了一种基于多智能体深度强化学习的交通整形框架，即在给定交通密度的情况下维持理想的交通流。与现有拥堵管理策略相比，该框架的一项关键优势是能够缓解滞回现象。换言之，与大多数着眼于预防或延缓交通流崩溃的拥堵管理策略不同，即使在交通流崩溃形成后，该框架仍然极为有效。该框架假设存在部分联网，即网联自动驾驶车辆（CAVs）彼此共享信息。此外，该框架仅要求具备基本的自动化水平（1级自动化），其定义为自动驾驶车辆执行一维纵向控制。该框架主要采用集中式训练、集中式执行的多智能体深度强化学习方法构建，其中纵向控制由加速或减速指令信号定义，再由所有智能体统一执行（即集中式执行）。用于训练和测试该框架的模型基于著名的双重深度Q学习算法（DDQN），将交通流中的平均状态作为模型输入（状态），并输出加速或减速值形式的动作。我们展示了该模型塑造交通状态、缓解滞回的负面影响，甚至将交通流改善至高于原有水平的能力。本文还确定了在假设网联自动驾驶车辆均匀分布于环形系统中的情况下，成功进行交通整形所需的最低网联自动驾驶车辆比例。本文所阐述的框架不仅展示了强化学习在应对这类挑战方面的理论适用性，还提出了一种切实可行的解决方案，该方案只需部分联网，并持续监测系统的平均速度；这一点可通过现成传感器实现，这些传感器能够测量网联自动驾驶车辆附近一定范围内车辆的速度。

关键词：深度强化学习、自动化与网联车辆、交通整形、拥堵管理、滞回现象

1 引言

几十年来，道路拥堵一直是城市环境面临的一项重大挑战，对通勤者和环境都造成了负面影响。因此，过去几十年间，人们广泛研究了交通管理和拥堵缓解策略，并提出了涵盖需求侧和供给侧（通常兼顾两者）的各种解决方案，包括拥堵定价 [1, 2, 3], 速度协调和可变信息标志 [4, 5, 6, 7, 8], 匝道控制 [9, 10, 11, 12, 13, 14], 以及高乘载率车辆（HOV）车道 [15, 16]。

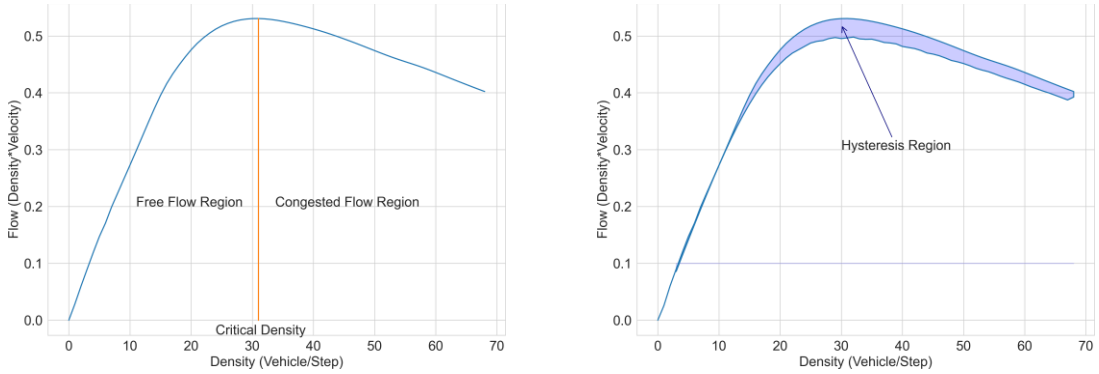
不幸的是，由于道路环境的复杂性以及各类道路使用者之间的相互作用，对于世界各地的研究人员和从业者而言，实现有效的交通管理仍是一项尚待解决的挑战。

尽管传统的交通管理方法取得了一些成效，但它们大多是被动应对的方法，且主要依赖历史数据和/或低分辨率数据（来自基础设施传感器）。网联自动驾驶汽车（CAV）技术和人工智能的最新发展，为应对其中许多挑战提供了契机，也为拥堵和交通管理领域的新研究开辟了道路。Elfar [17] 的一项研究就是这类工作的一个典型例子。他利用逻辑回归、随机森林和神经网络，开发了离线和在线预测模型，借助网联车辆提供的车辆轨迹预测短期交通拥堵。此外，研究人员已开始探索强化学习提升模型性能的能力，并利用这些预测来实现预测性速度协调系统。Mohanty 等人 [18] 使用长短期记忆（LSTM）网络，根据全区域观测数据预测局部区域拥堵。随后，研究人员利用 LSTM 模型估算缓解拥堵所需的最优收费。另一方面，Genser 等人 [19] 通过应用多层神经网络，研究了以交通管理为目的的动态最优拥堵定价。他们利用神经网络模型预测广义成本，并推导最优定价函数。文献中提出的其他方法还包括贝叶斯回归 [20], 、卷积神经网络 [21], 和聚类 [22]。

近年来，强化学习与网联自动驾驶车辆的结合受到了广泛关注。Han 等人提出了一种基于物理信息强化学习（RL）的匝道控制策略。他们结合历史数据和交通流模型生成的合成数据对强化学习模型进行训练，以避免陷入不准确的训练环境。该策略同时应用于局部匝道控制和协同匝道计量控制，结果表明，其能够显著改善交通性能，并且优于经典的基于反馈的匝道计量策略 [23]。早期的实现以经典形式采用强化学习，通过构建Q学习模型并利用动作与奖励对其进行训练 [24, 25]。近年来，深度强化学习的出现也为进一步探索打开了大门，已有多项研究采用深度Q网络（DQN）及其变体来解决交通管理中的多个问题 [26, 27, 28]。Walraven 等人 [29] 将交通流优化问题建模为强化学习问题。他们利用Q学习获得限速策略，并使用神经网络提升其策略学习算法的性能。该方法考虑交通预测，并主动输出控制信号。Wen 等人 [30] 利用深度Q网络（DQN）解决再平衡需求，这对于共享出行按需系统中的有效车队管理至关重要。他们表明，DQN能够通过减少出行者的等待时间并限制车辆行驶距离而发挥良好效果。在另一项研究中，Zhou 等人 [31]采用深度确定性策略梯度算法，构建了用于网联自动驾驶汽车的基于强化学习的跟驰模型，以获得适当的驾驶行为，从而实时改善信号交叉口处的出行效率、燃油消耗和安全性。其他研究则采用多智能体强化学习执行大规模交通信号控制任务。研究人员使用了协作式深度强化学习的不同变体，结果表明，这些框架能够显著缓解交通拥堵 [32, 33, 34]。

仔细审视近期采用强化学习来缓解交通拥堵的研究，可以发现没有一项研究着眼于改善滞回区域内的交通流。例如，Kreidieh 等人 [35] 致力于消散走走停停波：在封闭或开放网络中，利用一定比例采用强化学习控制策略的网联自动驾驶车辆来完成这一任务。（也就是说，该策略能够成功防止走走停停交通不断累积并造成拥堵加剧；然而，它仍是一种主动策略，因为一旦交通拥堵形成，该策略便无用武之地。）Ha 等人同样利用图卷积网络和强化学习来提升瓶颈路段的通行量 [36]。另一方面，Zhu 等人的其他研究 [37] 关注跟驰行为，并利用强化学习在自然驾驶环境中优化驾驶员表现（该环境由 NGSIM数据集模拟），而不以缓解拥堵为重点。

现有大多数文献的一个关键不足在于，它们专注于防止和/或延迟交通流崩溃形成。然而，一旦发生交通拥堵并出现交通流崩溃，几乎没有可用策略来缓解或消除拥堵。事实上，人们普遍认为，一旦发生交通流崩溃，便很难再采取有效措施消除拥堵；如果没有任何干预，交通流恢复只能通过需求降低实现。遗憾的是，在交通流基本图中，当交通流超过崩溃点（即临界流量点）时，如果交通流开始恢复（即研究路段沿线的车辆密度开始降低），流量并不会达到原先的最大值（对于每一个密度值而言），这种现象称为滞回。图 1-a 和图 1-b 阐释了临界密度和滞回的概念。这些图是通过在环形道路环境中模拟人工驾驶车辆生成的，该环境将用于本研究。如图 1-b 所示，在崩溃点之后，通过降低密度，流量无法恢复到先前的水平。本研究的主要贡献是提出一个框架，在不同市场渗透率（MPR）下，利用深度强化学习，并结合网联自动驾驶汽车，专门解决超过临界密度点后的交通流恢复问题。该框架旨在



(a) 基本图　　(b) 滞回现象

图 1：(a) 加载阶段的基本图，以及 (b) 卸载阶段的滞回现象。[38]

防止滞回形成，并且通常可以将流量水平提高到崩溃点时的水平，从而显著提升系统效率。

本文其余部分的结构安排如下：下一节将讨论所提出的深度强化学习（DRL）模型的构建及其细节。随后介绍仿真设置和模型参数。接着呈现详细结果，并讨论本文的研究发现。最后，我们将简要总结全文，并展望未来可能的研究方向。

2 模型构建

2.1 双重深度Q网络

本文所选用的网络架构采用双重深度Q网络（DDQN）[39]的形式。选择这一框架是因为它提供了一种简化的结构，可以直接对我们试图评估的概念进行测试和验证。需要注意的是，该框架已成功应用于许多研究中，包括我们此前的研究工作[40]。双重深度Q网络是深度Q网络的一种变体，它将一种称为Q学习的强化学习方法与深度神经网络相结合，以对智能体执行端到端控制。在DQN和DDQN架构中，都包含两个神经网络：主网络和目标网络。主网络将环境的既定状态作为输入，并在任意给定状态下输出可能的动作之一。与Q学习一样，著名的贝尔曼方程[41] 用于更新网络的权重和Q值。目标网络与主网络相似，但会间歇性地进行更新，而不是在每一步都更新，以提高模型的稳定性并获得更好的收敛速度。在DDQN中，通过将目标网络中的最大值操作分解为动作选择和动作评估，降低了高估现象。这一改进进一步提升了模型的性能[42, 41, 39]。图2展示了深度Q网络的简单示意图：智能体执行动作后，环境会产生并观测到新的状态和奖励；随后，这些新的状态和奖励再次输入神经网络（即智能体），以选择新的动作。

2.2 多智能体结构

在多智能体强化学习中，学习和执行可分为三种主要类别：集中式训练-集中式执行、集中式训练-分散式执行，以及分散式训练-分散式执行。在第一类中，环境中的智能体被视为一个整体，可以共享一个学习模型（即集中式控制器）。集中式控制器决定每个智能体的动作，并根据所有智能体同时提供的输入和奖励进行更新。需要注意的是，模型可以向所有智能体统一输出动作，也可以针对不同智能体输出不同动作。另一方面，集中式训练-分散式执行是指模型利用所有智能体的输入来调整参数，但每个智能体会根据其观测到的局部状态独立采取各自的动作。最后，分散式训练-分散式执行采用这样的结构：每个智能体都有独立的模型，该模型使用特定智能体的输入进行训练；同样，每个智能体按照各自模型的指示执行动作。

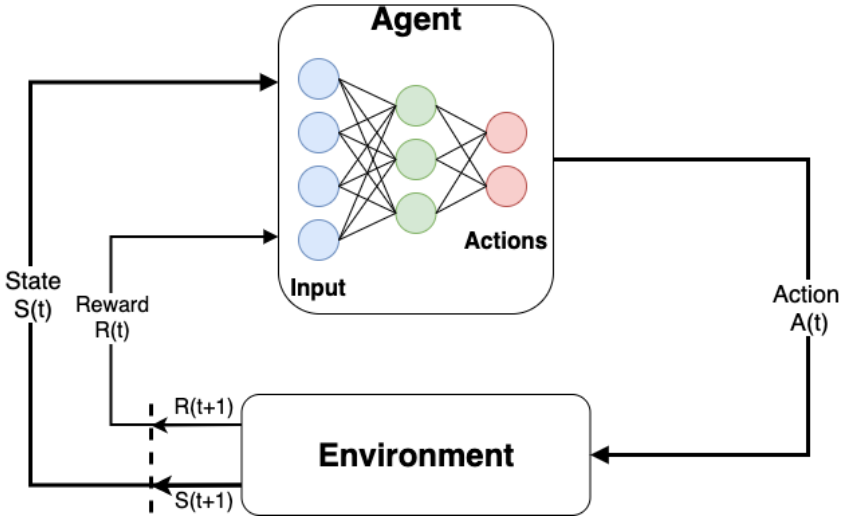


图 2：深度Q网络的简单示意图 [41]

不过，智能体通常会与集中层面的结果保持一定联系，以确保取得理想结果。显然，上述每种方法各有优缺点。例如，在集中式训练-集中式执行中，集中的决策过程可以带来协同且有效的学习[43]。需要注意的是，如果只使用一个模型（所有智能体采取统一动作），就能显著降低训练和执行的复杂度及计算需求。然而，假设所有智能体无论状态和观测有何不同都应采取相同动作，这未免过于简单。由于决策空间复杂，理论上集中式训练会导致策略更新的方差更高[44]。另一方面，通过增加复杂度、定义多个模型并根据各智能体自身的观测分别下达动作，即采用分散式训练-分散式执行，可以降低策略更新的方差。在实践中，分散式训练通常能带来更稳健的性能[44]。不过，分散式算法的一个缺点是，智能体之间的信用分配机制（即谁应因任务完成出色而获得肯定？）很难调整，且往往会导致智能体之间协调失灵[45]。本研究采用相对简单的集中式训练-集中式执行结构：利用道路上所有智能体（即网联自动驾驶汽车，CAV）的状态联合训练一个模型，并由中央控制器在每个时间步下达一个需要所有CAV统一执行的动作。如前所述，这种结构存在不足。最重要的是，显然每辆CAV根据自身局部情况（即与前车和后车的距离）执行自己的动作（即加速或减速），会比遵循一条要求所有车辆统一执行的全局指令更有利。不过，我们证明了该模型能够成功塑造交通状态，并对流量产生积极影响。需要注意的是，当CAV随机分布在交通流中时，这种方法的效果会降低。

2.3 模型参数

为定义智能体，我们创建了一个包含 4 个隐藏层的全连接深度神经网络。前两层各有 512 个神经元，第 3 层和第 4 层则分别有 128 个和 64 个神经元。图 3 展示了网络架构的详细信息。该全连接神经网络能够捕捉任务的复杂性，并且在合理的训练时间内取得了很好的表现。值得注意的是，在选定最终网络结构之前，我们尝试了多种网络结构，并对其他模型参数进行了实验。表 1 列出了所构建的双重深度Q网络模型的超参数设置。在训练和测试过程中，我们尝试了折扣因子（γ）取值在 0.75 到 0.999 之间的情况。最终选定的值为 0.90，这确保强化学习智能体在做出决策时，更加重视当前奖励，而非未来奖励（$\gamma = 1$。0 表示智能体认为当前奖励与未来奖励之间没有差别，也就是说，智能体会变得更有远见 [41]）。这样设置是因为该问题的构建不需要任务完成时的“成功”奖励。如果在另一种情况下需要成功奖励，则可以使用较高的折扣因子，从而充分考虑回合中经过多个时间步后才可能获得的最终奖励。此外，我们将经验回放记忆库 [46] 的大小设为 100,000，以限制训练期间的计算负载。我们还将窗口长度设为 1。窗口大小决定了输入神经网络（即特征）时所考虑的时间步数。在动作取决于系统先前状态的场景中，选择大于 1 的窗口长度更有优势。

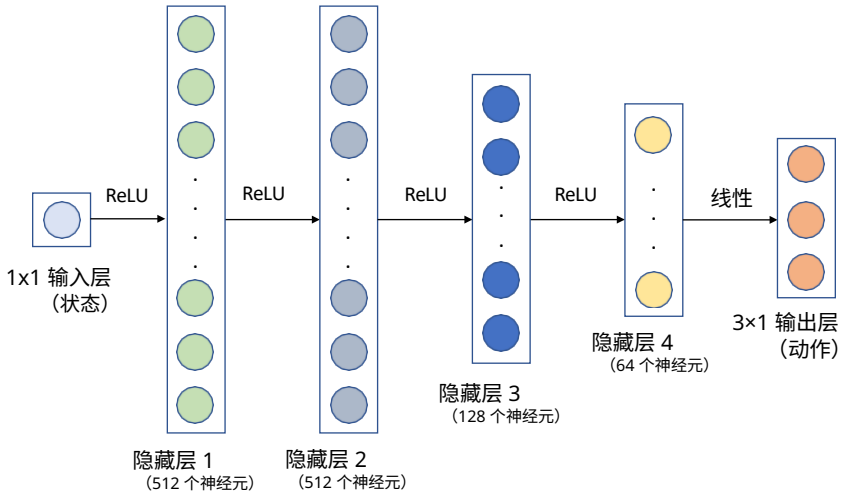


图 3：本研究的深度Q网络架构

在这类场景中，任何时间步采取的动作都取决于系统的先前状态。另一方面，将窗口大小设为 1，意味着只需观测当前状态，就足以准确地选择动作。我们对智能体进行了 100 万步训练，以实现良好的收敛。在训练过程中，我们采用了 ε-贪心策略 [41]，其中 ε 从 1 开始并线性衰减至 0.05。这样可以确保智能体在开始采取贪心动作选择之前，有充足的时间进行探索。这能保证对环境进行充分探索，从而获得良好或近乎最优的表现。请注意，最小值 ϵ 设为 0.05，以确保在整个校准过程中始终保持一定程度的探索。这对于避免系统停留在局部最小值至关重要。最后，我们采用了按时间步衰减的学习率计划，使学习率随着训练推进而逐渐降低。与固定学习率相比，这有助于模型以更稳定的方式收敛。

表 1：深度Q网络超参数设置

超参数	值
层数	4
隐藏单元数	512, 512, 128, 64
学习率	$0.001 \rightarrow 0$
策略	ε-贪心
ϵ	$1.0 \rightarrow 0.05$
折扣因子 γ	0.90
经验回放内存大小	100,000
回合数	5,000
批量大小	32
激活函数	ReLU，线性
优化器	Adam

3 仿真设置

3.1 仿真环境

为了仿真交通流事件并绘制交通流基本图，我们建立了一个单车道环形系统。采用环形系统是因为它能够仿真交通拥堵和滞回现象。尽管环形道路的动力学可能不同于直线道路，但它具备评估本研究主要假设所需的全部特征，而且在这种结构中更容易形成交通拥堵。需要注意的是，单个激波可能在环形系统中无限传播。图4展示了环形系统的简化示意图。环路开始时内部没有车辆，随后车辆逐渐出现，密度随时间递增。人工驾驶车辆采用智能驾驶员模型（IDM）[47]进行建模。第一阶段的所有车辆

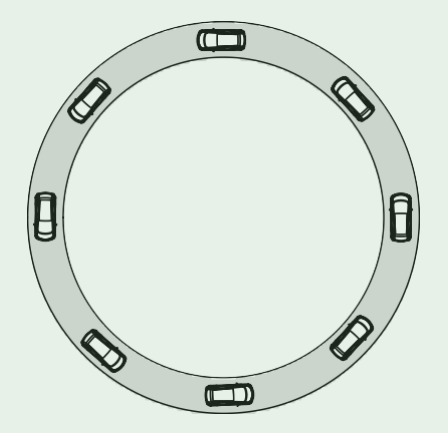


图4：环形道路环境

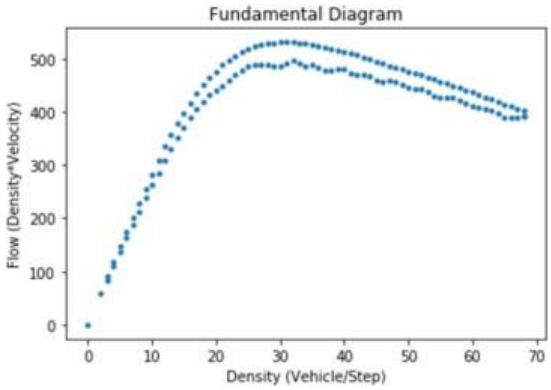


图5：人工驾驶车辆形成的滞回

仿真均为人工驾驶车辆。系统持续运行，直至达到临界密度，并在此后继续运行，超过交通崩溃状态。环路总长度设定为1000米，达到所需拥堵程度约需68辆车（请注意，该数量可通过改变环路长度进行调整）。达到这一密度点后，通过移除部分车辆对系统施加冲击，以观察其响应。图5展示了逐步移除车辆时系统的响应。在此情况下，车辆以随机方式移除。多次随机移除仿真的结果表明，系统的响应保持不变。可以看出，直到环路中几乎没有车辆剩余时，流量才接近其原始最大值（即滞回现象）。

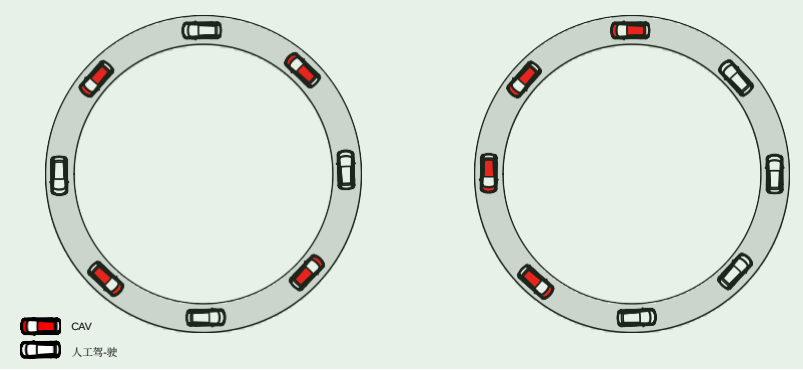


图6：网联自动驾驶汽车的一般编队策略

或者，可以将部分人工驾驶车辆替换为网联自动驾驶汽车，由先前定义的强化学习模型通过集中式动作选择对其进行控制，并假设这些车辆与基础设施实现完全通信。当系统受到冲击时，强化学习智能体的任务是将交通流恢复到高于原始观测最大值的水平（针对特定密度）。需要注意的是，在实践中，立即将人工驾驶车辆替换为网联自动驾驶汽车可能会带来一些问题；不过，预计所提出的系统能够应对切换过程中的轻微延迟（类似于车辆切换到自适应巡航控制时的延迟）。无论如何，本研究的重点并非评估这种改变的可行性，也不在于驾驶员遵从性（尽管两者都是关键的决定因素）。

假设强化学习智能体均匀分布在环形道路上的车辆之间。研究发现，让网联自动驾驶汽车均匀分布在车辆之间能够取得较好的效果，因此采用了这一方案。需要注意的是，尽管这可以被视为一项限制性假设，但对于任何所需的强化学习智能体渗透率，只要网联自动驾驶汽车的数量达到最低要求且其在交通流中随机分布，就可以仅使用其中的一个子集，形成足够接近均匀分布的队形。不过，所需的网联自动驾驶汽车最低数量可能会随目标强化学习智能体渗透率而变化。然而，有一种特殊队形无法通过这种方法实现（除非网联自动驾驶汽车的市场渗透率非常高）：强化学习智能体组成车队。因此，我们对这两种队形都进行了测试。图 6 展示了上述队形策略。需要注意的是，当多辆强化学习车辆前后相随时，它们对交通流的影响会减弱。因此，强化学习智能体仅略微偏离均匀分布时，系统的动态特性可能会与本研究所呈现的情况有所不同。由于在现实环境中，我们对强化学习车辆的可用情况（受其位置、驾驶员遵从性等多种因素影响）几乎无法控制，因此必须设计一种能够处理强化学习车辆任意队形的控制系统。基于此，本研究被视为对所提控制策略的概念验证，而更一般的情况留待未来研究。我们测试了不同的强化学习智能体市场渗透率（介于 15% 和 66% 之间），以检验网联自动驾驶汽车影响系统的能力。此外，我们还将这一框架与其他常见的现有拥堵管理策略进行了比较。

3.2 强化学习环境 **t**

网联自动驾驶车辆的任务是在发生交通流崩溃后恢复流量。车辆离开环形道路的一部分后对恢复情况进行测量，强化学习车辆的任务是将流量水平提高到理想值（如果可能），即防止基本图中的滞回现象。假设网联自动驾驶车辆与基础设施保持完全通信，并配备了一些传感器，例如惯性测量单元（IMU）、全球定位系统（GPS）和感知传感器（如激光雷达或雷达），通过这些传感器可以测量前后方车辆的大致速度（系统将利用这些数据估计系统速度）。随后，系统的平均速度被用作深度强化学习（DRL）模型的状态（输入）。另一方面，网联自动驾驶车辆选择动作的动作空间定义为加速度/减速度值，取值范围为 -1、0 和 1 米/秒²。需要注意的是，这是双重深度Q网络（DDQN）算法的一个缺点：其动作空间是离散的，因此我们将动作空间限制为三个值；这也是深度Q网络（DQN）架构中的输出层神经元数量设为三个的原因（分别对应三种可能的动作）。此外，还可以选择更激进的加速度/减速度值，这可能会更快达到理想流量，但代价是牺牲乘客舒适度。这样的状态-动作空间定义使深度强化学习能够根据给定的平均速度加快或减慢网联自动驾驶车辆的速度，并最终改变整个系统的流量。随后，我们定义了一个奖励系统，使流量增加时获得最大化奖励，并对系统中发生碰撞的事件进行惩罚。流量增加所获得的奖励根据以下方程计算：

$$r(v) = v_a \tag{1}$$

其中，

- r(v) 是奖励
- v_a 是系统的平均速度

方程 1 中的奖励在一个回合内的每个时间步都会计算。因此，如果在整个回合期间系统的平均速度达到最大值，则该回合的总奖励也会最大化。相反，如果车辆的平均速度低于最大可达速度，则获得的奖励会低于最优值。奖励的另一个组成部分是与碰撞相关的负奖励。这是一种每回合只施加一次的惩罚：如果环形道路上的任何车辆与其他车辆发生碰撞，就会触发该惩罚。发生碰撞时，整个回合将终止。最后定义了一项成功奖励，旨在驱使强化学习模型收敛到期望的解决方案。此处的“成功”定义为

在加载过程中，流量超过基本图所定义的原始最大流量值。这样可以避免模型陷入局部极小值：模型虽然不断累积奖励，却未达到系统实际期望的目标。正如后续章节将展示的，这种架构可用于交通整形，并实现期望的流量和速度。换言之，我们不必将期望流量设定为超过原始最大流量，而可以将其设定为某个特定流量值，并训练模型将系统调节至该流量水平。这种方法有多种用途，包括交通整形、速度协调、碰撞缓解等。表 2 详细概述了强化学习和道路环境的参数。

表 2：强化学习与道路环境设置

参数	数值
状态数量	1
动作数量	3
碰撞奖励	-3000
平均速度奖励	系统平均速度/步
成功奖励	+1000
加速度/减速度值 (m/s^2)	-1, 0, 1
期望速度（米/秒）	30
时间步长（秒）	0.1

值得注意的是，我们测试了不同的奖励元素，并通过反复试验选定了本文所展示的结构。对于我们所研究的特定问题，这种奖励值组合的表现名列前茅。不过，这种结构并非唯一，其他奖励结构也能成功完成任务。另一方面，需要牢记，场景或环境发生某些变化时，可能需要调整奖励结构或学习过程。因此，在实践中，构建能够适应不同场景并可扩展至仿真环境之外的稳健模型至关重要。这通常是基于强化学习的方法的一项弱点，因为它们可能对奖励结构非常敏感。

4 结果与讨论

在测试所提出的结构之前，第一步是推断能够影响系统所需的最小 MRP。例如，一辆网联自动驾驶汽车无法影响由 60 多辆车辆组成的车队的状态。下面将讨论这种方法。

假设某一路段在流量 q^k 定常的车辆系统，且已知的情况下密度降低（即密度突然下降），流量将直接降低（表示为 $q = ku$）。鉴于流量（q）与车头时距（h）成反比，在这种情况下，部分车辆驶离后缩短时间车头时距（即保持驶离前的车头时距），即可维持系统流量。因此，可以使用以下公式计算维持系统流量所需的网联自动驾驶汽车数量。

$$\bar{h}_{t-1} = \frac{(NT_t - NN_t) \cdot \bar{h}_t + h_{CAV} \cdot NN_t}{NT_t} \tag{2}$$

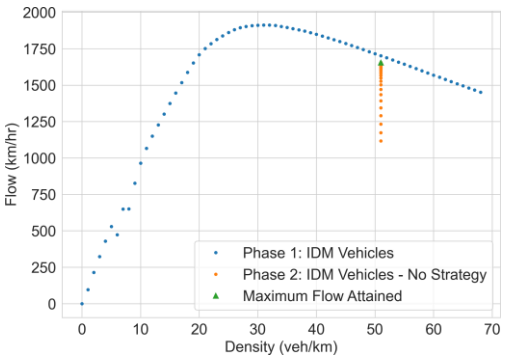
其中，

- $\bar{h}_{t-1}$ 表示前一时间步的平均时间车头时距
- $\bar{h}_t$ 表示当前时间步的平均时间车头时距
- NT_t 表示时间 t 的车辆总数
- NN_t 是维持 h_t 所需的网联自动驾驶车辆数量 -1
- $h_{网联自动驾驶汽车}$ 的期望时间车头时距

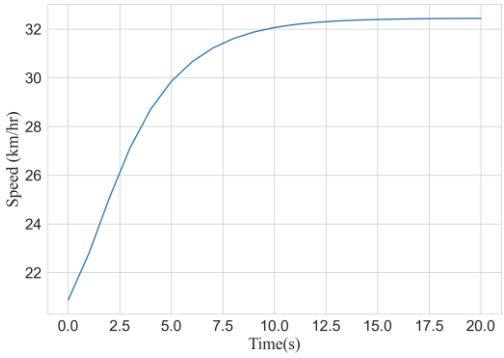
根据上面的公式 2，求出维持先前时间车头时距所需的网联自动驾驶车辆总数（NN_T），即可得到实现有效框架所需的最低市场渗透率。因此，以一个总共有 68 辆车、当前时间车头时距为 2.549 秒、期望的网联自动驾驶车辆车头时距为 2 秒的案例为例，如果有 1 辆车驶离系统，即 NT_t 变为 67，当前时间车头时距变为 2.5779，则我们得出结论：需要 5 辆网联自动驾驶车辆才能

$\bar{q}$ 维持先前的时间车头时距，进而维持流量。如前所述，可以利用反比关系（$\bar{h} = 1$），根据系统的平均流量（$\bar{q}$）推算平均时间车头时距（h）。

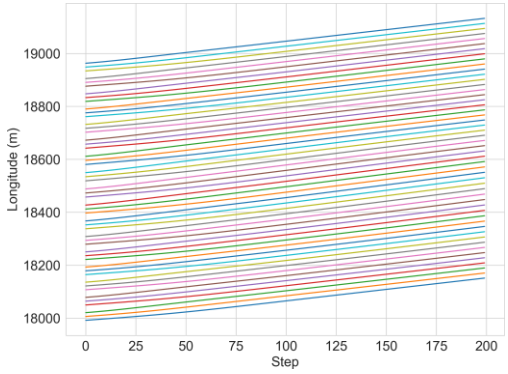
我们使用三种主要的市场渗透率测试框架的能力：15%、33% 和 66%。此外，我们通过两个交通密度变化回合，测试模型执行交通流塑形任务的能力。首先，我们展示智能驾驶员模型作为基准模型的表现。在这项试验中，环形道路上的所有车辆均由人工驾驶（使用智能驾驶员模型进行建模）。下方的图 7 展示了模拟人工驾驶的智能驾驶员模型在部分车辆驶离环形道路后恢复交通流的表现。从图 7-a 可以看出，流量始终未能恢复到第一阶段时的状态（即基本图上的状态）。还可以看出，达到的最高速度稳定在约 9 米/秒。



(a) 平均流量



(b) 平均速度



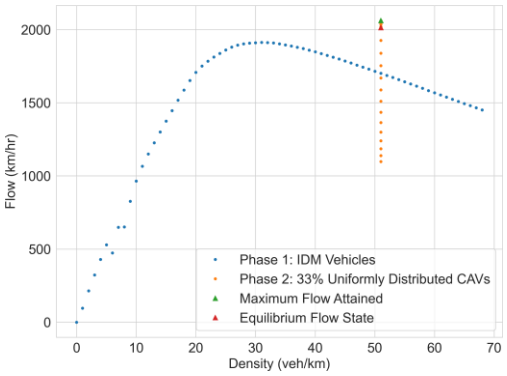
(c) 车辆轨迹

图 7：系统中 100% 车辆采用智能驾驶员模型时的表现。

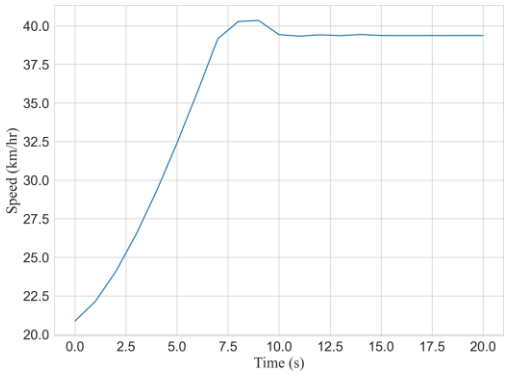
相反，我们进行了相同的实验，但在第二阶段开始时（即部分车辆驶离环形道路后），将 33.3% 的车辆切换为网联自动驾驶车辆。驶离发生前，环形道路上的车辆总数为 68 辆。17 辆车驶离环形道路，留下 51 辆车。其中，17 辆为网联自动驾驶车辆，并均匀分布在环形道路上。图 8 展示了环形道路上有 33.3% 网联自动驾驶车辆时系统的表现。可以看出，经深度强化学习训练的网联自动驾驶车辆能够显著影响系统状态。图 8-a 显示，流量不仅与基本图上该点所代表的初始流量相当，而且远远超过了它。达到的最大平均速度提高至 11 米/秒，与完全由人工驾驶车辆组成的系统的 9 米/秒相比，增幅超过 22%。图 8-c 中的车辆轨迹也表明，网联自动驾驶车辆能够利用驶离车辆所留下的间隙，并创造和维持改善后的交通状态。图 8-d 展示了双重深度Q网络模型训练过程中奖励的变化趋势。训练期间，奖励虽有一些下降——这是智能体在学习过程中发生碰撞所致——但奖励总体稳步提升，并在训练接近结束时趋于平稳。

值得注意的是，在本研究的结果中，不同场景所需的最高奖励值和回合数各不相同。这是因为每项任务的收敛（即停止准则）都根据各场景的变化进行了不同设置。例如，在密度首次变化后改善流量的任务，

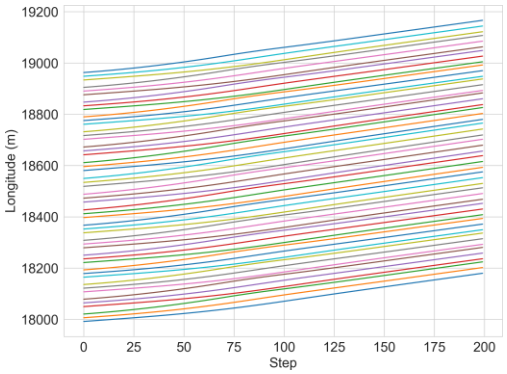
不同于第二种场景中的任务：在该场景中，密度从不同的初始值再次降低。因此，为确保比较有效，成功与否是根据是否实现实际目标来衡量，而不是根据是否达到某个奖励值或完成特定步数来判断。



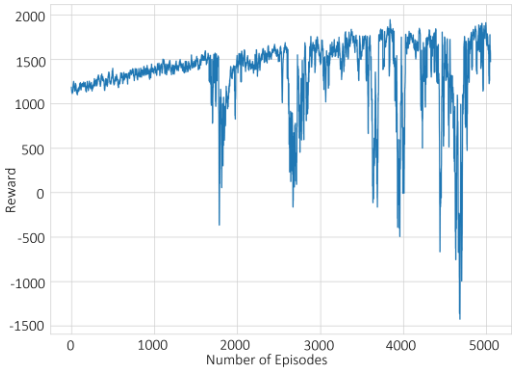
(a) 平均流量



(b) 平均速度



(c) 车辆轨迹



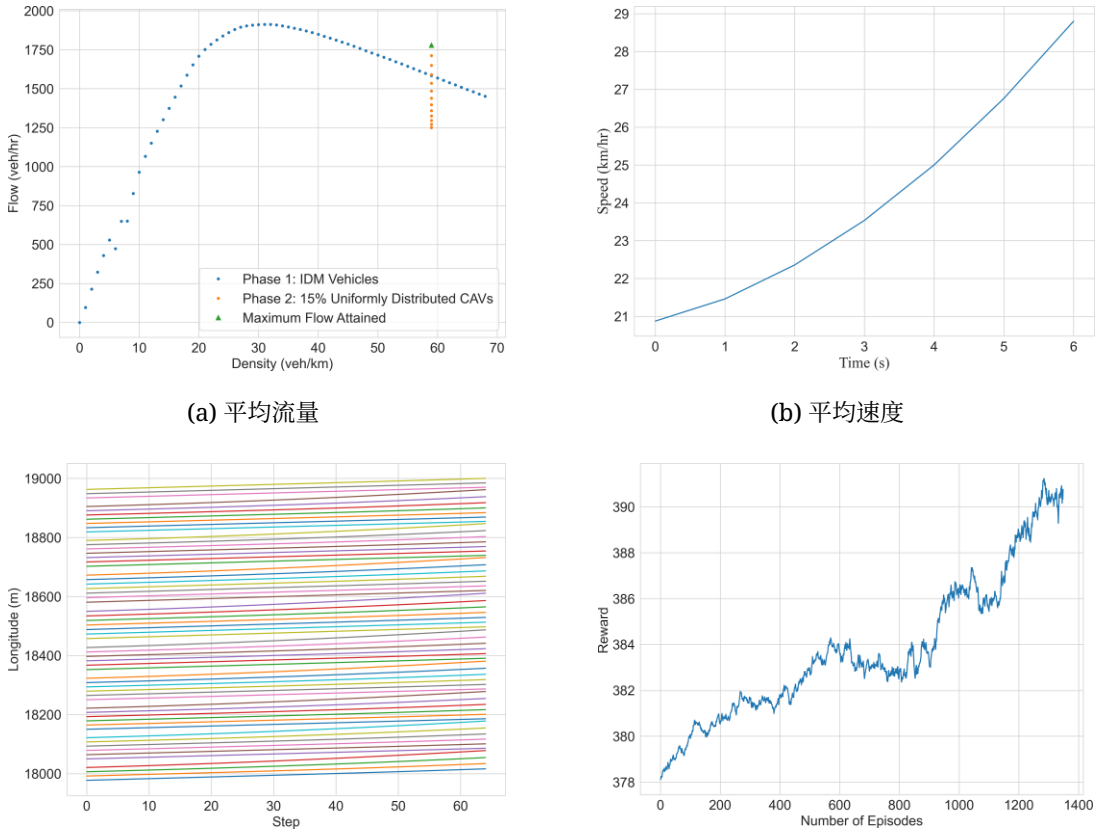
(d) 奖励

图 8：33.3% 网联自动驾驶车辆由强化学习控制时系统的性能

随后，我们测试所提出框架在较低和较高 MPR 下影响交通流的能力，即 15% 和 66%。因此，在 15% 的情形下，从环形道路上的 68 辆车辆中移除 9 辆，使环形道路上剩下 59 辆，其中 9 辆为网联自动驾驶车辆。准确地说，这相当于 15.25% 的 MPR。图 9 展示了框架在环形道路上有 15% 网联自动驾驶车辆时的表现。图 9-a 展示了这 9 辆网联自动驾驶汽车对系统产生显著影响的能力。系统能够以较大的幅度超过原有流量。鉴于此情形下环形道路上剩余车辆数为 59，最大可达速度低于前一种情形（即 33%）。图 9-b 显示，达到的最高平均速度约为 8 米/秒。图 9-c 和 d 中的车辆轨迹和奖励图与前一种情形一致：网联自动驾驶车辆利用间隙提升系统流量，而学习过程在训练回合接近结束时逐渐收敛。

另外，图 10 展示了系统在 66% 市场渗透率下的表现。与市场渗透率为 33% 的第一种情况类似，17 辆车驶离环形道路，但随后有 34 辆车转变为网联自动驾驶车辆。结果表明，随着市场渗透率的提高，系统能够更高效地实现目标。

此外，我们通过在市场渗透率为 33.3% 的情景中增加第二个步骤，进一步探索该框架的潜力。第二个步骤定义为车辆再次驶离环形道路的一个阶段，随后网联自动驾驶车辆尝试再次改善交通状态。在这一情景中，首批 17 辆车驶离后，又有 12 辆车驶离。环形道路上剩下 39 辆车，其中 13 辆（33.3%）是网联自动驾驶汽车。随后，我们重新训练双重深度 Q网络模型以优化输出。该情景是一个概念验证，表明无论初始设置如何，所提出的框架都能够改善整体交通状态，并能在交通密度变化时持续发挥作用。从图 11-a 至 d 可以看出，模型能够收敛至最优解，此时流量达到最大，且



(a) 平均流量

(b) 平均速度

(c) 车辆轨迹

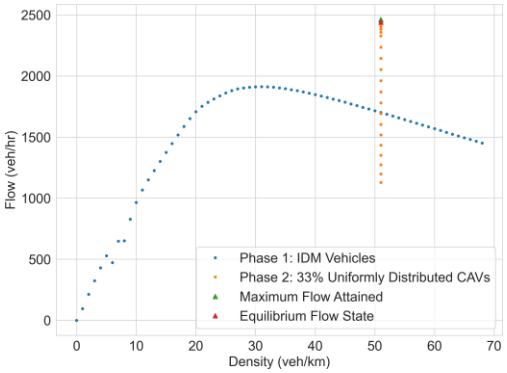
(d) 奖励

图 9：15% 网联自动驾驶车辆由强化学习控制时的系统性能

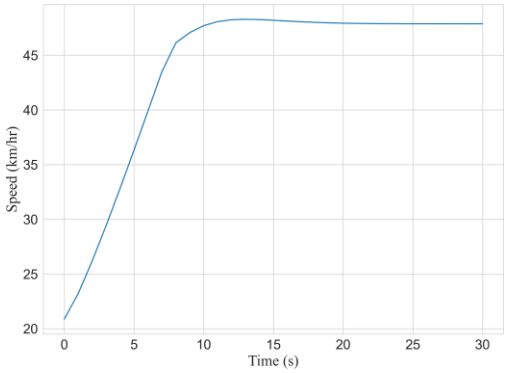
事实上，还超过了初始状态。平均速度超过 12 米/秒。有趣的是，轨迹显示，尽管速度和流量仍在上升，环形道路上车辆之间的车头时距却正趋于平衡。

另一方面，我们探究网联自动驾驶车辆维持其所达到的流量的能力，并考察其积极效果是否并非转瞬即逝。为此，我们对第一个场景稍作调整：网联自动驾驶车辆达到最大流量后，不再让它们保持原状，而是再次将其切换为人工驾驶车辆，并从该时刻起额外运行 200 个时间步。由于每个时间步为 0.1 秒，因此总共增加了 20 秒。图 12 对比了如上所述的两种模式各额外运行 20 秒的结果。显然，网联自动驾驶车辆不仅能够达到前文所示的理想高流量水平，还能够在相当长的一段时间内维持非常接近的流量水平。另一方面，从图 12-b 可以看出，将车辆切换回智能驾驶员模型（即切换为人工驾驶）后，网联自动驾驶车辆所达到的流量迅速下降，交通状态也回到了比 20 秒前显著更低的水平。

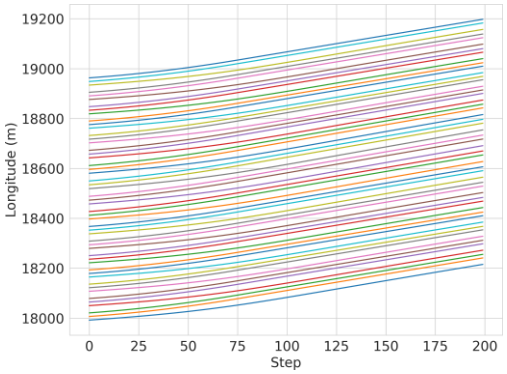
最后，为了说明该框架相较于其他拥堵管理策略的优势，我们测试了一种由 Allaby等人提出、并被许多其他研究采用的热门速度协调策略。[48] 该策略应用于环形道路上的所有车辆，并与一个 100% 人工驾驶（即采用智能驾驶员模型）的系统进行对比。图 13 展示了实验结果。可以看出，速度协调策略（正如预期的那样）削弱了系统在交通崩溃后恢复的能力，而非提升这一能力。这是意料之中的，因为速度协调往往会降低限速，从而降低可实现的最大流量。这种拥堵管理系统只有在交通流崩溃形成之前才有效，能够防止流量达到临界水平。匝道计量控制也可作类似解释，只是匝道控制限制的是密度，而非速度。请注意，由于环形系统的结构特点，本文未对匝道计量控制进行测试。



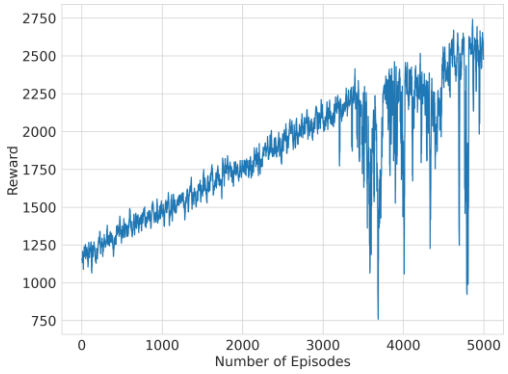
(a) 平均流量



(b) 平均速度

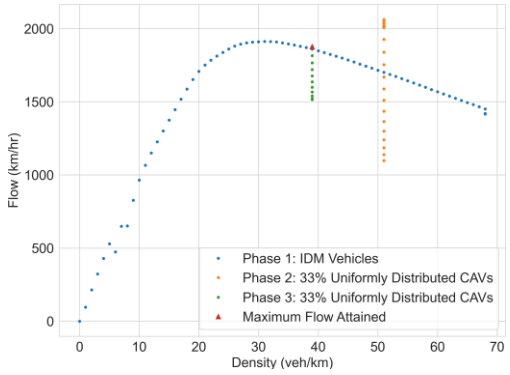


(c) 车辆轨迹

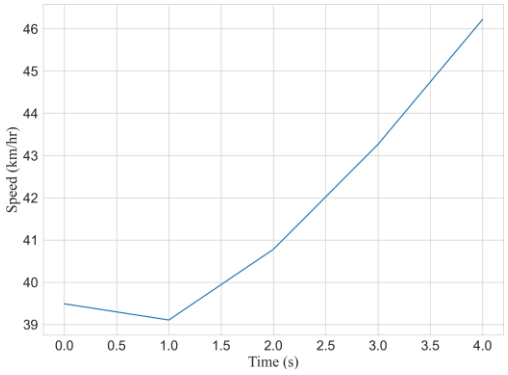


(d) 奖励

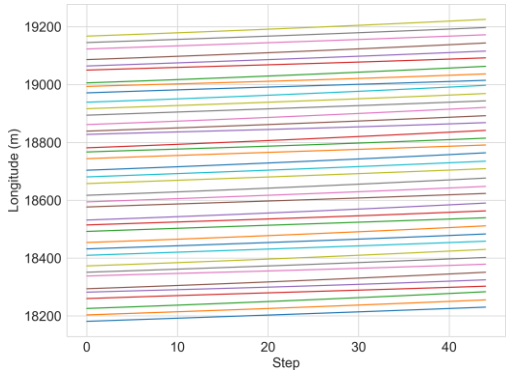
图 10：66% 强化学习控制的网联自动驾驶车辆下系统的性能



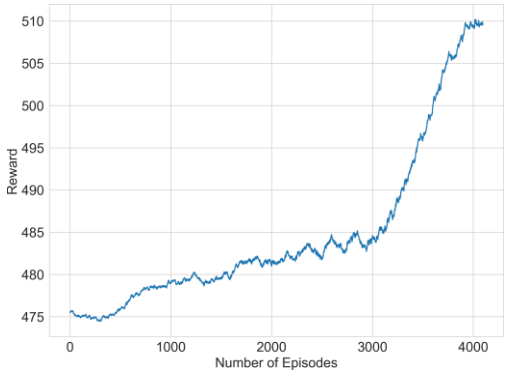
(a) 平均流量



(b) 平均速度

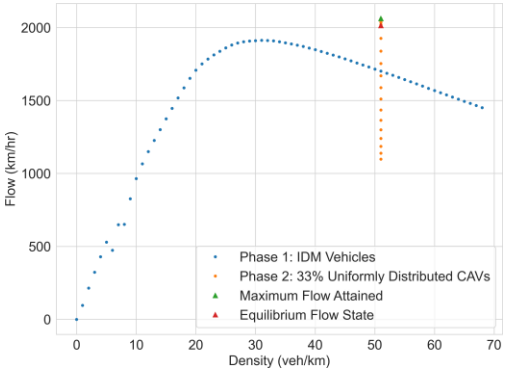


(c) 车辆轨迹

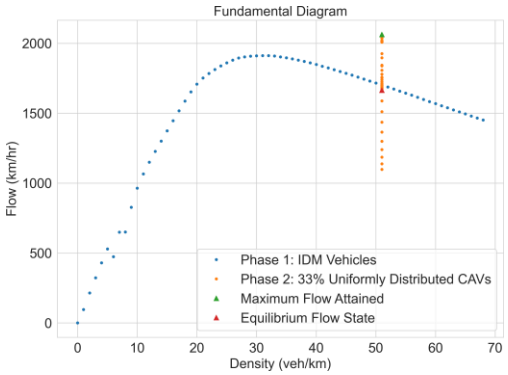


(d) 奖励

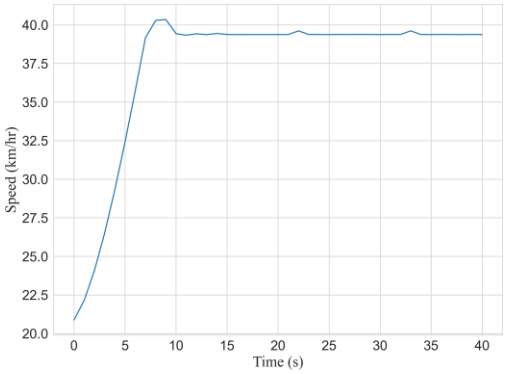
图 11：33.3% 网联自动驾驶车辆的性能——第二种场景



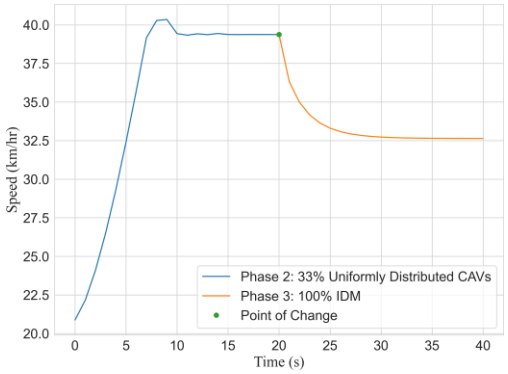
(a) 平均流量 - 网联自动驾驶车辆



(b) 平均流量 - 智能驾驶员模型

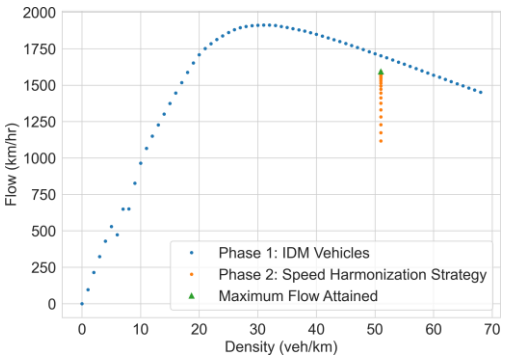


(c) 平均速度 - 网联自动驾驶车辆

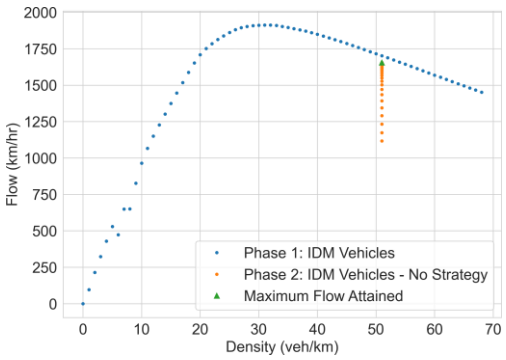


(d) 平均速度 - 智能驾驶员模型

图 12：网联自动驾驶汽车与智能驾驶员模型的比较



(a) 平均流量 - 速度协调策略



(b) 平均流量 - 智能驾驶员模型

图 13：速度协调与智能驾驶员模型对比

5 结论与未来工作

本研究提出了一个框架，主要目标是实现交通整形和交通拥堵缓解。为此，该框架利用深度强化学习和编队策略，优化环形道路环境中网联自动驾驶车辆（CAVs）的行为。CAVs旨在控制交通流，从而改善整体交通状况。我们选择双重深度Q网络（DDQN）来训练智能体（即CAVs）执行当前任务。该网络的输入状态是系统在每个时间步的平均速度，对应的可能输出动作为加速度/减速度值，包括 -1、0 和 1 米/秒^2，供智能体选择。采用这一离散动作空间的所提方法展现出显著改进，预计在连续动作空间下性能会更好（目前这一方向留待未来研究）。该问题被视为集中式训练、集中式执行任务：集中式训练是指使用交通流平均速度训练单一模型；集中式执行则是指模型发出一个动作，由环形道路上的所有CAV共同执行。这是该环境中可采用的更复杂的分散式训练、分散式执行方法的简化版本。通过环形道路仿真，我们展示了该框架在交通流达到超过临界密度的状态后显著改善交通流的能力，也表明CAV能够整形交通流并实现目标流量水平（前提是CAV的MPR达到最低要求）。我们还将该框架与常见的拥堵管理策略进行比较，并展示了其优越性。该框架假设CAV具备一定的感知能力，能够推断交通流的平均速度；鉴于CAV配备全球定位系统和雷达等常见传感器，这是一项相对简单的任务。未来将探索分散式训练-分散式执行，以及集中式训练与分散式执行的情形，以实现交通整形。还需要探索不同深度强化学习模型及其性能，并研究更新后的状态和动作定义。

6 致谢

本材料　基于美国国家科学基金会第 204793 号拨款支持的研究工作　　7.

7 作者贡献声明

作者确认，在研究构思与设计、结果分析与解释以及论文撰写方面贡献相等。所有作者均审阅了研究结果，并批准了论文的最终版本。

参考文献

[1] Hu X Zhou H. Wang J, Chi L. 考虑碳排放成本的城市交通拥堵定价模型。《可持续发展》，6(2):676–691, 2014年。[2] Sukumar Kalmanje Kara M. Kockelman. 基于积分的拥堵收费：一项政策提案及公众响应。《交通研究A辑：政策与实践》，39(7-9):671–690, 2005年。[3] Juan C. Muñoz Yafeng Yin J orge A. Laval, Hyun W. Cho. 收费设施的实时拥堵收费策略。《交通研究B辑：方法论》，71:19–31, 2015年。[4] Steven E. Shladover Hesham A. Rakha Xiao-Yun Lu Ramanujan Jagannathan Daniel J. Dailey Jiaqi Ma, Xi- aopeng Li. 高速公路速度协调。《IEEE智能车辆汇刊》，1(1):78–89, 2016年。[5] Petros Yuan, Tianchen Ioannou. 面对不确定性时用于缓解交通拥堵的可变限速控制。交通研究所工作论文系列 qt2sz308jc，交通研究所、加州大学戴维斯分校，2022年。[6] Hamdar SH Talebpour A, Mahmassani HS. 速度协调：拥堵条件下的有效性评估。《交通研究记录》，第69–79页，2013年。[7] Mahmassani SH Elfar A, Talebpour Alireza. 网联环境下的预测性速度协调：一种机器学习方法。美国交通研究委员会，2019年。[8] Kiron Chatterjee Mike Mcdonald. 使用可变信息标志发布动态交通信息的有效性：来自欧洲城市实地试验的证据。《交通评论》，第559–585页，2004年。[9] Les Jacobsen Nick Thompson Arianne Mizuta, Kim Roberts. 匝道控制：经实践证明且经济有效的运营策略——入门指南。技术报告，美国联邦公路管理局运营办公室，2014年。

[10] Habib Hadj-Salem Papageorgiou, Markos and Jean-Marc Blosseville. Alinea：一种用于入口匝道控制的局部反馈控制律。《交通研究记录》，1320(1):58–67, 1991年。[11] Markos Vong Vincent Gaffney John Papamichail, Ioannis Papageorgiou. 在澳大利亚莫纳什高速公路实施的启发式匝道控制协调策略。《交通研究记录》，2178:10–20, 2010年。[12] M. Papageorgiou and A. Kotsialos. 高速公路匝道控制：综述。《IEEE智能交通系统汇刊》，3(4):271–281, 2022年。[13] Jr. E. D. Arnold. 匝道控制：文献综述。技术报告，弗吉尼亚州交通研究委员会，1998年。[14] Papageorgiou M Smaragdis E. 一系列新的局部匝道控制策略。《交通研究记录》，第74–86页，2003年。[15] Seng Desai Aniruddha Singh Jack Desai, Prajakta Loke. 基于多智能体的车辆拥堵管理。第1031–1036页，2011年。[16] Texas Transportation Institute Cambridge Systematics, Inc. 交通拥堵与可靠性：将解决方案与问题相联系。技术报告，美国联邦公路管理局运营办公室，2004年。[17] Amr Elfar. 网联自动驾驶车辆系统的预测性交通运营与控制。
https://www.proquest.com/dissertations-theses/predictive-traffic-operations-control-connected/docview/2280673633/se-2?accountid=*14553*，2019年。[18] Michael Cassidy Sudatta Mohanty, Alexey Pozdnukhov. 利用深度学习进行区域级拥堵预测与控制。《交通研究C辑：新兴技术》，116，2020年。[19] Anastasios Kouvelas Alexander Genser. 通过应用多层神经网络实现多区域城市网络的动态最优拥堵定价。《交通研究C辑：新兴技术》，134，2022年。[20] Sando Thobias Emmanuel Kidando, Moses Ren. 用贝叶斯回归方法在不确定性条件下估计用于交通流崩溃事件识别的速度阈值。《交通工程学报A辑：系统》，145:04019013。[21] H. P. Zhao H. Kim N. Ranjan, S. Bhandari and P. Khan. 基于卷积神经网络、长短期记忆网络和转置卷积神经网络的全市交通拥堵预测。《IEEE Access》，8:81606–81620。[22] A. Zhang C. Xu and Y. Chen. 基于多时段热点聚类的上海交通拥堵预测。《IEEE Access》，8:63255–63269。[23] Linghui Li Claudio Roncoli Jinda Gao Pan Liu Yu Han, Meng Wang. 一种基于物理信息强化学习的局部与协同匝道控制策略。《交通研究C辑：新兴技术》，137，2022年。[24] C. Jacob and Baher Abdulhai. 利用机器学习实现综合交通走廊控制。IEEE国际系统会议，4:3460–3465。[25] C. Urbanik T. Kohls Airton Arel, Itamar Liu. 基于强化学习的路网交通信号控制多智能体系统。《智能交通系统》，第128–135页。[26] Majid Leon-Garcia Raeis. 一种用于公平交通信号控制的深度强化学习方法。2021年。[27] Martin Greguric´, Miroslav Vujic´, Charalampos Alexopoulos, and Mladen Miletic. 深度强化学习在交通信号控制中的应用：开放交通数据的概述与影响。《应用科学》，2020年。[28] Seunghan Park Byungkyu Son Sang Ko, Byungjin Ryu. 利用网联自动驾驶车辆在高速公路车道封闭处进行速度协调与合流控制：一种强化学习方法。《IET智能交通系统》，2019年。[29] Bram Bakker Erwin Walraven, Matthijs T.J. Spaan. 交通流优化：一种强化学习方法。《人工智能工程应用》，52:203–212, 2016年。[30] Jian Wen, Jinhua Zhao, and Patrick Jaillet. 共享按需出行系统的再平衡：一种强化学习方法。载于2017年IEEE第二十届智能交通系统国际会议（ITSC），第220–225页，2017年。[31] Mofan Zhou, Yang Yu, and Xiaobo Qu. 面向信号交叉口网联自动驾驶车辆的高效驾驶策略开发：一种强化学习方法。《IEEE智能交通系统汇刊》，21(1):433–443, 2020年。[32] Y. Deng A. Jin Q. Dai T. Tan, F. Bao and J. Wang. 用于大规模交通网格信号控制的协作式深度强化学习。《IEEE Transactions on Cybernetics》，50:2687–2700, 2020年。

[33] Xiaoqiang Wang、Liangjun Ke、Zhimin Qiao 和 Xinghua Chai。使用新型多智能体强化学习的大规模交通信号控制。《IEEE Transactions on Cybernetics》，51(1):174–187，2021年。[34] Tong Wang、Jiahua Cao 和 Azhar Hussain。采用基于协作群组的多智能体强化学习实现大规模场景下的自适应交通信号控制。《Transportation Research Part C-emerging Technologies》，125:103046，2021年。[35] Abdul Rahman Kreidieh、Cathy Wu 和 Alexandre M Bayen。通过深度强化学习消散闭合网络和开放网络中的走走停停波。载于2018年第21届智能交通系统国际会议（ITSC），第1475–1480页。IEEE，2018年。[36] Paul Young Joun Ha、Sikai Chen、Jiqian Dong、Runjia Du、Yujie Li 和 Samuel Labi。利用联网自动驾驶车辆和多智能体强化学习的能力缓解高速公路瓶颈拥堵。arXiv预印本，arXiv:2010.05436，2020年。[37] Meixin Zhu、Yinhai Wang、Ziyuan Pu、Jingyun Hu、Xuesong Wang 和 Ruimin Ke。基于强化学习的自动驾驶安全、高效且舒适的速度控制。《交通研究C辑：新兴技术》，117:102662，2020年。[38] Reinhard Lubashevsky、Ihor Mahnke。序参量作为不稳定交通流的附加状态变量。《Traffic and Granular Flow》*'99*，2000年。[39] Hado Van Hasselt、Arthur Guez 和 David Silver。采用双Q学习的深度强化学习。载于AAAI人工智能会议论文集，第30卷，2016年。[40] Alireza Ammourah、Rami Talebpour。自动驾驶车辆强制换道的深度强化学习方法。《交通研究记录》，2022年。[41] Richard Sutton 和 Andrew Barto。《强化学习：导论》，第2版。第2版，2018年。[42] Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Andrei Rusu、Joel Veness、Marc Bellemare、Alex Graves、Martin Riedmiller、Andreas Fidjeland、Georg Ostrovski、Stig Petersen、Charles Beattie、Amir Sadik、Ioannis Antonoglou、Helen King、Dharshan Kumaran、Daan Wierstra、Shane Legg 和 Demis Hassabis。通过深度强化学习实现人类水平控制。《Nature》，518:529–33，2015年2月。[43] Gang Chen。多智能体强化学习的新框架——通过策略蒸馏实现集中式训练与探索以及分散式执行。载于AAMAS，2020年。[44] Xueguang Lyu、Yuchen Xiao、Brett Daley 和 Chris Amato。多智能体强化学习中集中式评论家与分散式评论家的对比。ArXiv，abs/2102.04402，2021年。[45] Meng-Lin Li、Shaofei Chen 和 Jing Chen。自适应学习：一种面向协作多智能体系统的新型分散式强化学习方法。《IEEE Access》，8:99404–99421，2020年。[46] Longxin Lin。基于强化学习、规划和教学的自我改进反应式智能体。《Machine Learning》，8:293–321，2004年。[47] A. Hennecke、M. Treiber 和 D. Helbing。经验观测和微观仿真中的拥堵交通状态。《Physical Review E》，62。[48] Peter Allaby、Bruce Hellinga 和 Mara Bullock。可变限速：高速公路应用中候选控制策略的安全性和运行影响。《IEEE智能交通系统汇刊》，8(4):671–680，2007年。[49] Alireza Talebpour、Hani S Mahmassani 和 Samer H Hamdar。速度协调：拥堵条件下有效性的评估。《交通研究记录》，2391(1):69–79，2013年。[50] Bidoura Khondaker 和 Lina Kattan。可变限速：联网车辆环境中的微观分析。《交通研究C辑：新兴技术》，58:146–159，2015年。[51] Muhammet Deveci、Dragan Pamucar 和 Ilgin Gokasar。基于模糊幂海伦函数的CoCoSo方法：实时交通管理中自动驾驶车辆优势排序。《Sustainable Cities and Society》，69:102846，2021年。