

分类号\_\_\_\_\_

密级\_\_\_\_\_

UDC <sup>注1</sup> \_\_\_\_\_



南京理工大学  
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

# 硕士专业学位论文

## 基于强化学习的区域信号控制方法

(题名和副题名)

万超

(作者姓名)

指导教师姓名 张伟斌 教授

学 位 类 别 工程硕士

专 业 名 称 光电信息工程

研 究 方 向 自适应交通信号控制

论 文 提 交 时 间 2025 年 3 月

注 1：注明《国际十进分类法 UDC》的类号。

### 声 明

本学位论文是我在导师的指导下取得的研究成果，尽我所知，在本学位论文中，除了加以标注和致谢的部分外，不包含其他人已经发表或公布过的研究成果，也不包含我为获得任何教育机构的学位或学历而使用过的材料。与我一同工作的同事对本学位论文做出的贡献均已在论文中作了明确的说明。

研究生签名：\_\_\_\_\_

年 月 日

### 学位论文使用授权声明

南京理工大学有权保存本学位论文的电子和纸质文档，可以借阅或上网公布本学位论文的部分或全部内容，可以向有关部门或机构递交并授权其保存、借阅或上网公布本学位论文的部分或全部内容。对于保密论文，按保密的有关规定和程序处理。

研究生签名：\_\_\_\_\_

年 月 日

## 摘要

现代交通系统因其动态变化和随机波动，使得传统信号控制技术难以满足日益复杂的出行需求。近年来，强化学习技术作为一种处理最优控制问题的有效方法，逐渐被引入交通信号控制领域。强化学习通过智能体与环境的交互，基于交互过程中积累的经验进行学习，并最大化预设的奖励函数。在交通系统中，奖励函数通常与交通流参数密切相关。然而，将强化学习应用于交通信号控制仍面临诸多挑战，例如状态形式设计、非平稳性处理、计算资源分配、相邻交叉口合作，以及未知状态应对等。本文针对上述问题，对强化学习技术在交通信号控制领域的应用进行了深入研究。

(1) 提出了一种基于互模拟度量的交通场景下异构观测间分布距离度量算法。合适的状态表示是强化学习算法实现的关键，但状态设计面临显著挑战：过于简单状态形式可能导致关键信息的丢失，而复杂的状态形式则可能引发维度爆炸，进而消耗大量计算资源。本文提出的算法通过同步发生的马尔可夫过程共享相同的奖励值，简化了互模拟度量的计算。同时，提出一个浅层神经网络模型，用以弥补异构观测间形式上的差异，从而计算得到最优的状态表示。基于该状态表示进行强化学习模型训练，与其它观测形式对比，结果显示旅行时间缩短 30%以上，整体交叉口车辆吞吐量提升 20%左右。

(2) 提出了一种有偏好状态编码方法。随着强化学习模型需要控制的区域不断扩大，状态空间也随之增长，进而可能引发维度爆炸问题。维度爆炸不仅增加了计算复杂度，还可能导致模型性能下降。本文提出的方法利用先验知识，确定模型对状态空间中不同部分的偏好，并根据偏好对不同部分的状态空间进行不同比例的压缩。实验结果表明，所提出的有偏好状态编码方法能够有效缓解交通环境动态变化的影响。

(3) 提出了一种时空依赖关系捕捉方法。在去中心化多智能体强化学习框架中，本地交叉口智能体的可观测范围有限，导致模型难以获得全局最优解。本文提出的方法从两个方面进行优化：一方面，通过聚合交通历史状态信息，刻画交通系统的时序特征；另一方面，允许本地交叉口的智能体获取一阶相邻交叉口的信息，从而描述相邻交叉口之间车辆的流动等空间特征。通过从时间和空间两个维度入手，有效缓解了去中心化框架中部分可观测性导致的非平稳性问题，同时弥补了有偏好状态编码中压缩状态空间引起的信息损失。

(4) 提出了一种基于双线性池化的特征融合模块。在经过有偏好状态编码和时空依赖关系捕捉模块处理后，组成状态的观测形式仍然是独立的。现有研究通常采用全连接层聚合或直接拼接的方式来生成特征，忽略了交通观测之间的隐藏关系。本文提出的特征融合模块通过计算观测之间的互信息，得到了用于评估和预测的特征。不仅提取出了观测之间的潜在关系，还提高了特征的稳定性和预测精度。

**关键词:** 智能交通, 交通信号控制, 强化学习, 深度强化学习, 多智能体强化学习, 深度 Q 学习, PPO 算法

## Abstract

The dynamic changes and random fluctuations in modern transportation systems make it challenging for traditional signal control technologies to meet the increasingly complex travel demands. In recent years, reinforcement learning has gradually been introduced into traffic signal control as an effective method for solving optimal control problems. Reinforcement learning learns through the interaction between agents and the environment, accumulating experience and maximizing the predefined reward function. However, applying reinforcement learning to traffic signal control still faces many challenges, such as state representation design, handling non-stationarity, computational resource allocation, cooperation between adjacent intersections, and addressing unknown states. This thesis provides an in-depth study of the application of reinforcement learning in traffic signal control, addressing the challenges mentioned above.

(1) This thesis proposes a distribution distance measurement algorithm for heterogeneous observations in traffic scenarios, based on mutual simulation metrics. An appropriate state representation is essential for the implementation of reinforcement learning algorithms. However, the design of the state faces significant challenges: overly simplistic state forms may result in the loss of critical information, while overly complex state forms may lead to dimensionality explosion, thus consuming excessive computational resources. The algorithm proposed in this thesis simplifies the computation of mutual simulation metrics by utilizing a Markov process with synchronized events that share the same reward value. Additionally, a shallow neural network model is introduced to address the form differences between heterogeneous observations, calculating the optimal state representation. Using this state representation for reinforcement learning model training and comparing it with other observation forms, the travel time is reduced by over 30%, and the overall intersection vehicle throughput increases by approximately 20%.

(2) This thesis proposes a preference-based state encoding method. As the area controlled by the reinforcement learning model expands, the state space also grows, potentially leading to the issue of dimensionality explosion. Dimensionality explosion not only increases computational complexity but also degrades model performance. This method leverages prior knowledge to determine the model's preference for different parts of the state space and performs compression of the state space in varying proportions based on these preferences.

Experimental results show that the proposed preference-based state encoding method can effectively mitigate the impact of dynamic changes in the traffic environment.

(3) This thesis proposes a method for capturing spatiotemporal dependencies. In decentralized multi-agent reinforcement learning frameworks, the limited observation range of local intersection agents makes it difficult for the model to achieve a global optimal solution. The proposed method optimizes the model in two ways: First, it aggregates historical traffic state information to capture the temporal characteristics of the traffic system. Second, it allows local intersection agents to access information from first-order neighboring intersections, enabling the model to capture spatial features such as vehicle flow between adjacent intersections. By addressing both temporal and spatial dimensions, the method effectively mitigates non-stationarity issues caused by partial observability in the decentralized framework, while also compensating for information loss caused by state space compression in the preference-based state encoding method.

(4) This thesis proposes a feature fusion module based on bilinear pooling. After the processing by the preference-based state encoding and spatiotemporal dependency capturing modules, the observation representations that constitute the state remain independent. Existing studies typically use fully connected layer aggregation or direct concatenation to generate features for prediction and evaluation, which overlook the hidden relationships between traffic observations. The feature fusion module proposed in this thesis calculates the mutual information between observations to obtain features for evaluation and prediction. This method not only extracts the latent relationships between observations but also enhances feature stability and improves prediction accuracy.

**Key word:** Intelligent transportation, Traffic signal control, Reinforcement learning, Deep reinforcement learning, Deep Q-learning, Bisimulation metrics, Bilinear pooling

# 目 录

|   |             |
|---|-------------|
| <b>摘 要</b> .....  | <b>I</b>    |
| <b>Abstract</b> .....   | <b>III</b>  |
| <b>目 录</b> .....  | <b>V</b>    |
| <b>缩略语表</b> .....   | <b>VIII</b> |
| <b>1 绪论</b> .....   | <b>1</b>    |
| 1.1 研究背景及意义 .....   | 1           |
| 1.2 国内外研究现状 .....   | 3           |
| 1.3 研究内容和章节安排 .....   | 6           |
| 1.3.1 研究内容 .....  | 6           |
| 1.3.2 章节安排 .....  | 7           |
| 1.4 本章小结 .....  | 8           |
| <b>2 深度强化学习以及信号控制框架</b> .....                                     | <b>9</b>    |
| 2.1 强化学习 .....  | 9           |
| 2.1.1 马尔科夫决策过程 .....  | 10          |
| 2.1.2 传统强化学习算法: Q-learning .....                                  | 11          |
| 2.2 深度强化学习 .....  | 12          |
| 2.2.1 深度学习 .....  | 12          |
| 2.2.2 最大池化 .....  | 14          |
| 2.2.3 矩归一化 .....  | 14          |
| 2.2.4 L2 归一化 .....  | 14          |
| 2.2.5 深度 Q 学习 .....   | 15          |
| 2.2.6 策略梯度方法 .....  | 16          |
| 2.2.7 演员-评论家结构的强化学习算法 .....                                       | 17          |
| 2.3 多智能体强化学习 .....  | 18          |
| 2.3.1 多智能体框架 .....  | 18          |
| 2.3.2 集中式控制和去中心化控制 .....  | 18          |
| 2.3.3 部分可观测性 .....  | 19          |
| 2.3.4 独立 Q 学习 (Independent Q-learning, IQL) .....                 | 19          |
| 2.3.5 QMIX 算法 .....   | 19          |
| 2.3.6 基于通信的多智能体强化学习 (Communication-Based Multi-Agent, COMA) ..... | 20          |

|                                     |           |
|-------------------------------------|-----------|
| 2.4 城市交通系统.....                     | 20        |
| 2.4.1 道路结构和交通运动.....                | 20        |
| 2.4.2 信号灯 .....                     | 21        |
| 2.4.3 道路、车辆相关参数.....                | 21        |
| 2.4.4 其它常见参数.....                   | 21        |
| 2.5 本章小结.....                       | 22        |
| <b>3 交通场景中异构观测间分布距离计算 .....</b>     | <b>23</b> |
| 3.1 问题描述.....                       | 23        |
| 3.2 主要贡献.....                       | 24        |
| 3.3 基于分布距离的观测距离计算.....              | 25        |
| 3.3.1 传统的同构观测之间的距离度量.....           | 25        |
| 3.3.2 异构观测之间的距离度量.....              | 25        |
| 3.4 实验设置.....                       | 27        |
| 3.4.1 实验数据 .....                    | 27        |
| 3.4.2 模型设置 .....                    | 27        |
| 3.5 实验结果.....                       | 29        |
| 3.5.1 异构与同构的单交叉口的瓦氏距离度量.....        | 29        |
| 3.5.2 各种观测形式在强化学习模型中的表现.....        | 30        |
| 3.5.3 路网层面各个观测形式之间的瓦氏距离度量.....      | 32        |
| 3.6 本章小结.....                       | 35        |
| <b>4 基于有偏好状态编码的区域交通信号控制算法 .....</b> | <b>36</b> |
| 4.1 问题描述.....                       | 36        |
| 4.2 主要贡献.....                       | 37        |
| 4.3 强化学习模型设置.....                   | 38        |
| 4.4 有偏好状态编码.....                    | 39        |
| 4.5 时空模型.....                       | 43        |
| 4.6 特征融合 .....                      | 45        |
| 4.7 实验设置 .....                      | 46        |
| 4.8 实验结果以及分析 .....                  | 49        |
| 4.8.1 杭州路网的实验结果以及分析 .....           | 49        |
| 4.8.2 凤林路网的实验结果以及分析 .....           | 50        |
| 4.8.3 复杂区域的实验结果以及分析 .....           | 51        |
| 4.8.4 实验结果总结 .....                  | 55        |

|                      |           |
|----------------------|-----------|
| 4.9 消融实验.....        | 57        |
| 4.10 关于模型泛化性的讨论..... | 62        |
| 4.11 本章小节.....       | 62        |
| <b>5 总结与展望.....</b>  | <b>63</b> |
| 5.1 全文总结.....        | 63        |
| 5.2 研究展望.....        | 64        |
| <b>致    谢.....</b>   | <b>65</b> |
| <b>参考文献.....</b>     | <b>66</b> |
| <b>附录.....</b>       | <b>76</b> |

## 缩略语表

|        |  |               |
|--------|--|---------------|
| RL     | Reinforcement Learning                       | 强化学习          |
| TSC    | Traffic Signal Control                       | 交通信号控制        |
| ATSC   | Adaptive Traffic Signal Control              | 自适应交通信号控制     |
| DRL    | Deep Reinforcement Learning                  | 深度强化学习        |
| MARL   | Multi-agent Reinforcement Learning           | 多智能体强化学习      |
| DQN    | Deep Q-network                               | 深度 Q 网络       |
| LSTM   | Long Short-Term Memory                       | 长短期记忆         |
| PPO    | Proximal Policy Optimization                 | 近端策略优化        |
| MDP    | Markov Decision Process                      | 马尔可夫决策过程      |
| OOD    | Out of Distribution                          | 超出分布          |
| iid    | independent identically distribution         | 独立同分布         |
| ST     | Spatio-Temporal                              | 时空            |
| MCTS   | Monte Carlo Tree Search                      | 蒙特卡洛树搜索       |
| GESA   | General Scenario-Agnostic                    | 一般场景-不可知场景    |
| A2C    | Advantage Actor-Critic                       | 基于优势的演员评论家算法  |
| GAE    | Generalized Advantage Estimation             | 广义优势估计        |
| POMDPs | Partially Observable Markov Decision Process | 部分可观测马尔可夫决策过程 |
| DRQN   | Deep Recurrent Q learning                    | 深度循环 Q 学习     |
| IQL    | Independent Q-Learning                       | 独立 Q 学习       |
| IPPO   | Independent Proximal Policy Optimization     | 独立近端策略优化      |

# 1 绪论

本章首先从研究的背景和意义出发，阐述了交通拥堵的成因及其带来的危害，并说明了采用交通信号控制技术缓解交通拥堵的现实意义。随后，从传统信号控制和基于强化学习的信号控制两个方面，回顾了国内外信号控制技术的研究现状。在回顾过程中，分析并指出了现有研究存在的不足。最后，本章介绍了研究以及所做出的贡献，并说明了本文的结构安排。

## 1.1 研究背景及意义

随着经济的快速发展和城市化进程的加快，机动车数量急剧增加。根据公安部统计，截至 2024 年 5 月底，全国机动车保有量达到 4.4 亿辆，其中汽车保有量达到 3.4 亿辆，驾驶人达 5.3 亿人<sup>[1]</sup>。过去 10 年，汽车年均上牌量超过 2000 万，新领证驾驶人超过 2800 万，总量和增量均位居世界第一<sup>[2]</sup>。然而，机动车数量的快速增长对城市交通系统提出了严峻挑战。交通拥堵、资源消耗和环境污染等问题日益突出，亟需有效的交通管理手段来应对这一挑战。

随着城市的扩张，居民的出行需求不断增加，城市交通系统面临着如何与城市空间协调发展、如何提升系统运行效益以及如何提升现代化治理能力等诸多挑战。在城市空间方面，对现有交通环境进行物理结构上的调整通常会导致短时间小范围的交通系统失效。因此，除非万不得已，通常不会轻易进行结构上的调整<sup>[3]</sup>；而运行效益上，根据百度地图 2023 年中国城市交通报告<sup>[4]</sup>显示，在通勤耗时前十的城市中，有八个城市平均通勤耗时相比于前一年有所上升。具体数据如表 1.1 所示。各类交通问题亟需解决，推动着人们不断探索改善交通环境的新方法。

城市道路交通系统是交通系统的重要组成部分，而交叉口作为管控城市交通的关键节点，其信号控制方案的优劣直接影响交通系统的运行效率。非最优的交通信号控制方案是引发各类交通问题的主要原因之一<sup>[5]</sup>。据相关研究统计<sup>[6]</sup>，由信号控制引起的交通延误约占全球交通延误的 10%。交通延误不仅会导致通勤时间增加，还会引起额外的燃油消耗，进而导致二氧化碳排放增加<sup>[7]</sup>。根据国际公路运输联盟（International Road Transport Union, IRU）的调查数据<sup>[8]</sup>，交通拥堵时的车辆频繁启停和低速行驶导致发动机效率显著下降，使得油耗与碳排放相比于畅通交通水平增加 300%。此外，非最优的信号控制方案还可能间接引发交通事故<sup>[9]</sup>。针对信号控制问题的研究早已广泛开展，早在 20 世纪中叶就有相关研究。一方面，传统的信号控制方案虽然在一定程度上可以提升交叉口的通行效率，但是难以应对突发变化<sup>[10-11]</sup>；另一方面，当前交通环境与过去相比，道路结构和车流组成等都变得更为复杂，单纯的优化某个交叉口的信号配时往往不

能达到预期效果，甚至可能引发其它交叉口的拥堵<sup>[12]</sup>。综上所述，利用现代化技术提升交通治理能力已迫在眉睫。

表 1.1 城市通勤时长及变化

| 2023 年排名 | 城市 | 2023 年平均通勤时长(min) | 通勤时长同比 2022 年 |
|----------|----|-------------------|---------------|
| 1        | 北京 | 44.47             | ↑3.91%        |
| 2        | 上海 | 39.60             | ↑11.06%       |
| 3        | 南京 | 37.40             | ↑5.95%        |
| 4        | 天津 | 37.11             | ↓1.10%        |
| 5        | 大连 | 36.22             | ↑3.00%        |
| 6        | 成都 | 35.88             | ↓0.62%        |
| 7        | 武汉 | 35.86             | ↑1.68%        |
| 8        | 深圳 | 35.40             | ↑3.65%        |
| 9        | 沈阳 | 35.27             | ↑2.21%        |
| 10       | 广州 | 35.25             | ↑4.44%        |

深度强化学习技术（Deep Reinforcement Learning, DRL）的出现为解决交叉口的信号控制问题提供了新的思路。与传统的控制方法（如固定配时控制和感应式控制）不同，DRL 能够根据当前交通环境动态调整信号配时方案，并实现与邻接交叉口的协同合作，从而在路网级别上实现系统最优目标<sup>[13]</sup>。因此，近些年来，越来越多的研究开始尝试将信号控制转换为一个强化学习问题并进行求解。

新的解决方案的产生是城市发展的必然结果。传统的固定式配时<sup>[14]</sup>最早出现，直至今日仍在沿用。它仅依赖历史数据进行配时，且一旦下发到信号灯后，无需再进行任何人工干预。然而，这种固定配时方案难以适应现代复杂的交通环境。为了处理变化的交通环境，感应式配时<sup>[15]</sup>出现。通过感应元件实时获得当前交通状态，并相应地调整配时，感应式配时在一定程度提升了信号控制的灵活性。然而，它仍难以实现不同交叉口之间的协同合作。不同发展阶段对交叉口信号配时算法提出了不同的需求，如何实现更加智能、更系统的控制效果，已成为提升现代化交通治理能力的核心问题。

DRL 在控制问题和组合博弈领域已表现出令人惊叹的效果<sup>[16-18]</sup>。从早期的 AlphaGo<sup>[19]</sup>到最新的具身智能<sup>[20]</sup>以及大语言模型<sup>[21]</sup>，DRL 在其中发挥了重要作用。概括而言，DRL 通过最大化目标函数实现优化。这一特性使其与信号控制等优化问题的目标高度契合，为解决复杂的交通信号控制问题提供了强有力的工具。

信号控制问题的目的是长期优化交通状况。这不仅要求解决方案能够应对当前的交通状况，还需要优化长期的交通流畅度和减少拥堵。而 DRL 正是通过最大化累积奖励（如减少等待时间、提高通行效率等）来优化长期目标。同时，动态环境适应性也是 DRL 可以被用于求解信号控制问题的一个原因，常见的控制系统难以实时响应交通变化，而 DRL 能够通过与环境的交互，动态调整信号灯的控制策略，以适应不断变化的交通状况。

更进一步的，DRL 可以同时优化多个目标<sup>[22]</sup>，如减少车辆等待时间、提高通行效率、减少排放等。这些需要优化的目标都可以在奖励函数中显式的体现出来。除此之外，在路网层面，DRL 可以使用分布式系统<sup>[23]</sup>，通过局部决策和全局协调来优化整个复杂交通网络的性能。这是其它控制算法难以做到的。此外，考虑到交通系统作为复杂系统，其变化多样且动态性强，配时系统必须具备实时性。在有限的计算资源下，需要在资源消耗与系统性能之间找到平衡。

本论文从强化学习算法出发，以交叉口交通协同控制问题为研究主题，研究路网层面的交叉口信号控制以及不同交叉口之间的协同优化，为路网信号控制问题提供研究思路。具体的研究意义在于：

1. 在理论意义方面，本文深入探讨了强化学习框架在信号控制问题中的适配性。首先，通过对强化学习基本原理的分析，明确了其在动态环境中的适应性和学习能力，为信号控制这一复杂且动态变化的系统提供了理论支撑。其次，本文详细阐述了强化学习在信号控制中的应用机制，包括状态空间、动作空间和奖励函数的定义，确保了强化学习算法能够有效地与信号控制系统进行交互和优化。
2. 在实践意义方面，本文不仅关注强化学习算法在信号控制中的理论适配性，还深入探讨了其在实际部署中的可行性和效率。特别是在计算资源的优化与管理方面，本文进行了深入分析，并提出了一系列优化措施，以确保强化学习算法在有限计算资源环境下的高效运行能力。这些优化措施不仅有助于提高信号控制方案的实践应用效果，还为未来智能交通系统的建设提供了重要的技术参考。

## 1.2 国内外研究现状

在本小节，将回顾国内外在解决信号控制问题时的一些关键技术发展。尽管将 DRL 应用于信号控制问题已不是一个新鲜话题，但即便存在众多有效的算法，实际部署到现实世界中的研究却相对较少。一方面，DRL 通过与环境的不断交互积累；另一方面累计经验，城市交通系统极其复杂，导致 DRL 模型在求解时需要更多的计算资源。而交通信号控制对实时性要求极高，因此本小节还回顾了 DRL 中优化计算效率的方法。这些方法主要集中在减少不必要的探索空间，从而有效缩短模型探索到最优化所需的时间，为本文进行提高计算效率提供了研究思路。

1868 年，世界第一盏交通信号灯在英国诞生。到了 1914 年，第一盏电子交通信号灯出现在美国的一个路口。三年之后，出现了手动控制的连动式信号系统。很快人们就注意到，手动控制也已经不能满足实际的出行需要，交通信号控制作为一类研究登上舞台。Webster 在 1958 年提出了 Webster 算法<sup>[14]</sup>，基于历史数据和专家经验计算出一组固定的配时方案。这对于稳定的交通模式是有效的，但在适应动态交通条件方面存在不足。

随着交通情况的日益复杂以及其不可预测性, Webster 算法的效果难以满足需要。因此,一些能够实时反馈交通流状况的算法被发明,如 SOTL 和 SCATS<sup>[15,24-25]</sup>。通过结合感应设备,这些算法实现了配时方案的局部实时优化。然而,随着信号配时需求范围的扩大,交叉口间对协调优化的需求日益明显,这些算法不能再满足出行的需要。面对这些挑战,将 DRL 应用于交通信号控制问题为自适应和协作信号优化提供了有前景的方法。

强化学习 (Reinforcement Learning, RL) 早在上世纪就已经出现,但其难以应对现实任务。这是因为传统 RL 在处理高维状态空间时面临维度灾难,难以有效建模。而深度神经网络的出现帮助强化学习算法有效解决了难以拟合值函数的问题。深度强化学习的第一次出现,即深度 Q 网络<sup>[17]</sup> (Deep Q-network, DQN), 它使用深度神经网络估计 Q 值函数。并且将模型成功的应用在了游戏任务上。此后 DRL 开始广泛应用在各种简单任务上。后来也有人开始将 DRL 应用在交通信号控制这一复杂任务上。

早先的研究主要集中在单交叉口<sup>[26-28]</sup>, 完全贪心式的优化本地交叉口。然而,研究人员很快发现,这种贪心式的信号配时优化虽然在当前交叉口表现良好,但忽略了交通是一个统一的整体,交叉口的行为会对相邻交叉口的交通状态产生影响。因此,研究视野逐渐转向多交叉口之间的合作。楚天舒等人<sup>[29]</sup>提出的基于 A2C (Advantage Actor-Critic) 算法的多交叉口信号控制算法产生了极大的影响。他们摒弃了传统的集中式训练框架,为每个交叉口分配了一个智能体,并通过扩大局部智能体的可观测范围降低了学习难度。然而,他们并未深入研究模型推理时的时间和内存消耗,也未充分探讨不同交叉口之间的合作机制。除此之外,如何使得训练出的模型更具有鲁棒性也是一个难点。之后,郑冠杰等人提出一系列模型<sup>[30-34]</sup>希望解决不同交叉口之间的合作问题。这一系列模型以优化不同交叉口之间的合作为主线,不断深入。FRAP<sup>[33]</sup>首先提出了交叉口内相位的竞争得分的概念,并将不同相位的得分量化。CoLight<sup>[30]</sup>使用图注意力机制,通过聚合相邻交叉口的状态信息扩大可观测范围,提升了学习能力。而 PressLight<sup>[31]</sup>则是参考了 Varaiya 等人的工作<sup>[35]</sup>,提出了一种新的状态表示。尽管上述工作在路网规模上有所突破,但遗憾的是,这些研究大多忽略了城市交通的复杂性,所选路网中的交叉口多为同构结构。随着研究的深入,研究人员开始关注模型的鲁棒性和可迁移性。郑冠杰等人最初的工作虽然有效解决了部分信号控制问题,但其模型难以迁移到其它路网。为解决这一问题,姜浩源等人在注意到这一情况后提出了 GESA (General Scenario-Agnostic) 模块<sup>[36]</sup>。该模块将不同的路口映射成同一的结构,并据此采用了多场景大规模协同训练,产生了较为通用的交通信号控制算法,在解决异构交叉口间学习经验共享这一问题上迈出了重要一步。后续也有工作<sup>[37-39]</sup>在此基础上不断深入。

上述研究解决了部分不同交叉口之间的合作问题,异构交叉口之间的经验迁移问题,但并未考虑到实际信号控制算法部署时的资源消耗问题。在实际应用场景中,以上算法

是否能满足实时性的需要，或者说满足实时性需要的算力依然是未知数。潘刚等人意识到算法实际部署时的算力需求，提出了 TinyLight<sup>[40]</sup>。该模型通过构建一个超图并且候选特征与一组轻量级网络相关联，并且使用了一种新颖的熵最小化目标函数消除超图的边缘来达到节省计算资源的目的。该模型可以在一块 2KB RAM 和 32KB ROM 的微处理器上运行。 $\Pi$ -Light<sup>[41]</sup>在此基础上定义了一种领域特定语言和构建程序的转换规则，并利用蒙特卡洛树搜索 (Monte Carlo Tree Search, MCTS) 在离散空间中找到最优程序。此外，该研究还分析了如何将学习到的程序策略直接部署在资源极其有限的边缘设备上，尽可能的减少了计算资源的消耗。尽管这两个工作确实减少了算法部署时的资源消耗，但它们的优化策略独立于强化学习算法之外，未能与强化学习模型深度融合。

在 DRL 自身的框架中，存在一些起到减少计算资源消耗的方法。交通环境的高度复杂性导致需要探索的状态空间变得极其庞大，因此减少需要探索的状态空间是优化资源消耗的有效途径。然而，目前针对交通环境中状态空间引起的资源消耗问题的研究较少。张良在研究中提出了一种基于先验知识的状态表示<sup>[42]</sup>并通过实验验证了该表示的有效性。但是基于先验知识的状态表示方法在实际应用中可能面临知识获取的困难，尤其是在动态变化的交通环境中，先验知识的更新和维护成本较高。同时，这种方法的泛化能力可能受到限制，难以适应不同交通场景的需求。张艾米的研究通过直接过滤任务无关信息有效减少了状态空间，但其依赖于特定奖励函数的设计，导致该架构在应对不同任务时的灵活性受限。这种依赖性使得该方法在面对复杂多变的任务时，难以适应新的环境和需求。此外，过滤机制的精确性直接影响智能体的学习效果，若过滤不当，可能导致关键信息的丢失，进而影响任务的完成质量。Himanshu Sahni<sup>[43]</sup>则更直接地将策略分解到更小、更简单的领域，通过设计一系列较小的子任务并对其进行有效分解，引导智能体实现目标。然而，这种分解方法在面对高度复杂或动态变化的任务时，难以保证子任务之间的协调性和整体策略的一致性。此外，在基于模型的强化学习中，可以通过计算状态相似性来最小化探索空间。例如，较为严谨的互模拟度量<sup>[44]</sup>提供了状态空间分解公式，利用状态转移概率和奖励值函数的相似性进行计算，尽管其仅适用于形式单一的状态空间，但对于简单任务而言，可以显著减少计算资源消耗。本文研究的基于深度强化学习的信号控制问题范围如图 1.1。

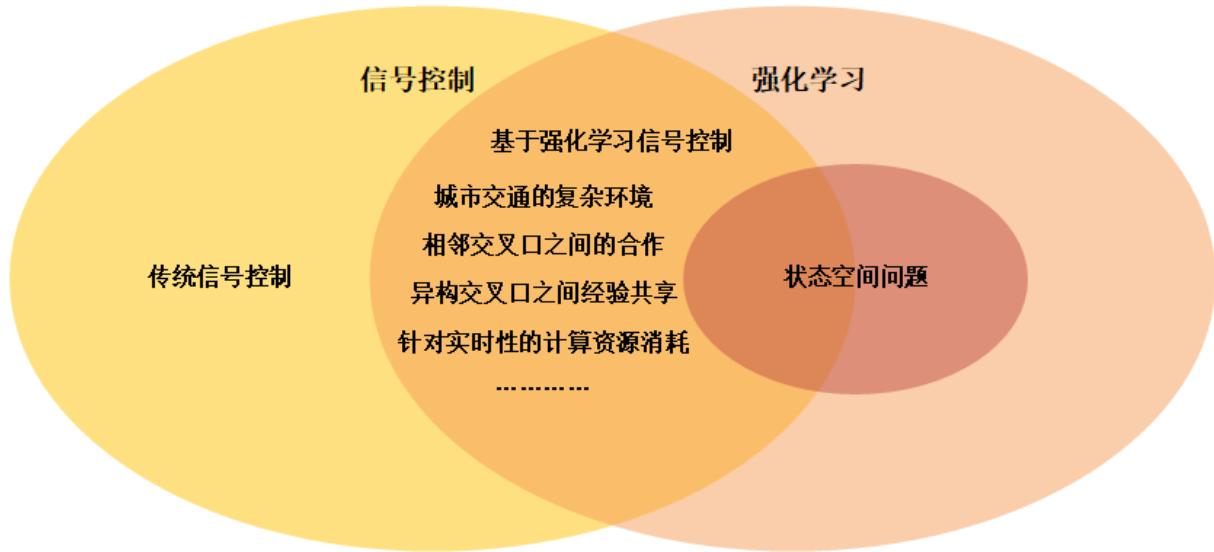


图 1.1 基于强化学习的信号控制问题框架

### 1.3 研究内容和章节安排

#### 1.3.1 研究内容

本文研究了基于强化学习的信号控制问题解决方案。以强化学习，深度强化学习以及多智能体强化学习等理论方法为指导，同时考虑在求解信号控制问题时的资源消耗。以求在最小的计算资源消耗下解决信号控制问题。本文的主要研究内容包括以下几个方面：

1. 提出了一种异构观测分布距离计算模型。强化学习框架中的智能体探测到的交通状态依赖于人工设计。不同的观测形式之间可能存在信息冗余，单一的观测形式又会出现丢失关键的信息。为此，本文提出了一种计算异构观测分布距离的模型。分布距离较远的观测之间重叠的信息更少，将其组合起来可以获得更大的信息熵。同时，减少冗余信息有助于降低计算资源的消耗，更小的状态空间也能帮助智能体更快地找到最优轨迹。
2. 验证了提出的异构观测分布距离模型的有效性，并将其应用于路网层面的信号控制问题。此外，本文还提出了一种状态压缩函数，利用传统信号控制算法之间的对比，得到最优模型对状态空间不同部分之间的偏好，用于对状态进行压缩。基于交通先验知识，确定了基准情况，即在低流量情况下可以减少对交通的干预。同时，增大了高流量情况在状态空间中的占比，从而减少了智能体需要探索的状态空间，进一步优化了计算资源的利用率。
3. 引入了一种多模态特征融合模型。传统的特征融合方法，如简单的拼接或全连接层，难以有效融合不同特征之间的信息。为此，本文采用了基于双线性池化的特征融合

方案。在获得时空特征后，通过双线性池化进行融合，得到不同形式观测之间隐藏的互信息，从而得到对模型训练更有利的特征表示。

4. 提出了一种邻接交叉口合作模块。传统的信号控制方案大多采用贪心式的单交叉口优化方法，然而这种策略已被证明是次优的。为此，本文调整了当前交叉口智能体的可见范围，并通过相邻交叉口的状态聚合来调整可获得信息范围。实验结果表明，提出的模型在优化效果上表现显著，同时性能优于传统方法和现有的多智能体强化学习（Multi-agent reinforcement learning, MARL）方法。

### 1.3.2 章节安排

本论文共分为五章，其组织结构如下：

第一章论述了本文的研究背景以及意义；之后对国内外相关研究进行文献总结与现状分析，提出本文的研究内容。并给出本文的组织结构。

第二章介绍本文的相关背景知识。从原始的强化学习出发，介绍了强化学习中的相关概念，指出传统强化学习算法中的不足。再介绍深度学习相关的知识，二者结合形成深度强化学习的框架，作为本文处理信号控制问题的方法核心。除此之外，本文的研究中心在于交通信号控制问题，因此在第二章，同步介绍了城市交通中的一些相关概念，包括评价指标，交叉口结构等。为后文具体在有限资源的条件下解决信号控制问题做铺垫。

第三章将互模拟度量转换到异构观测形式中。首先统计了近些年来有影响力的论文中使用的状态表示。之后介绍了计算异构观测分布距离的框架，并基于该框架计算了所统计的观测之间的距离。不仅是在常见的标准交叉口中，还在非标准交叉口以及路网层面，验证了提出的模型。并且在计算得到最优表示以后，通过实际的强化学习算法的应用验证了提出的模型的有效性。为第四章在优先资源的情况下解决路网层面的信号控制问题打下基础。

第四章在路网层面训练信号控制模型。首先指出了所使用基线算法的不足，并通过对不同表现信号控制算法的可访问状态空间分布，本文提出了有偏好状态编码，以弥补基线算法的不足。此外，为了应对去中心化框架中部分可观测性带来的额外非平稳性影响，本文提出了时空依赖关系捕捉模型，通过聚合一阶邻接节点的信息来扩大可观测范围。随后，为融合第三章中组成状态的观测信息，本文引入了基于双线性池化的特征融合方式，聚合不同形式交通观测间的互信息，从而得到了更优的特征。在多个场景中进行了测试。结果表明，本文提出的模型在交通参数优化方面表现优于其它对比算法。最后，通过消融实验验证了所提出模型应对环境非平稳性的能力。

第五章总结了全文的内容，分析当前研究的不足之处，并对未来的研方向提出展望。

## 1.4 本章小结

本章介绍了课题的研究背景和意义。介绍了强化学习和强化学习在自适应交通信号控制中的研究成果。在此基础上，确定了具体的研究内容和章节安排。

## 2 深度强化学习以及信号控制框架

本章介绍本论文所涉及的相关背景知识，主要包括强化学习，深度学习，深度强化学习以及城市交通系统。

### 2.1 强化学习

强化学习是一种通过与环境交互来学习最优决策策略的机器学习范式<sup>[45]</sup>。相比于其它机器学习方法，强化学习更加侧重于以交互目标为导向进行学习。在这种范式中，智能体通过采取行动并观察结果来逐步改进其行为，目标是最大化累积奖励。智能体在每个时间步长接收来自环境的反馈，并利用这些反馈来迭代更新其动作策略，以达到最佳控制策略。强化学习的核心在于从环境经验中学习，表现出一种反复试验的学习方式。智能体在没有先验知识的情况下，通过最大化定义的数字奖励（或最小化惩罚）来学习如何在环境中采取行动。此外，智能体需要有效地权衡探索和利用之间的关系，以充分了解其状态空间并找到最优的动作序列<sup>[46]</sup>。

图 2.1 展示了强化学习与简单交通系统之间的交互。在时刻  $t$ ，交通环境的状态  $s_t$  被智能体观测到。智能体根据当前策略  $\pi_\theta(a_t|s_t)$  计算出最优动作  $a_t$ ，并且依据探索或是利用的概率将动作下发给交通信号灯。在动作执行结束后接收到奖励  $r_t$ ，并且转移到下一个状态  $s_{t+1}$ 。每执行一次上述交互，就会得到一条轨迹，同时智能体也会加深对于环境的了解。

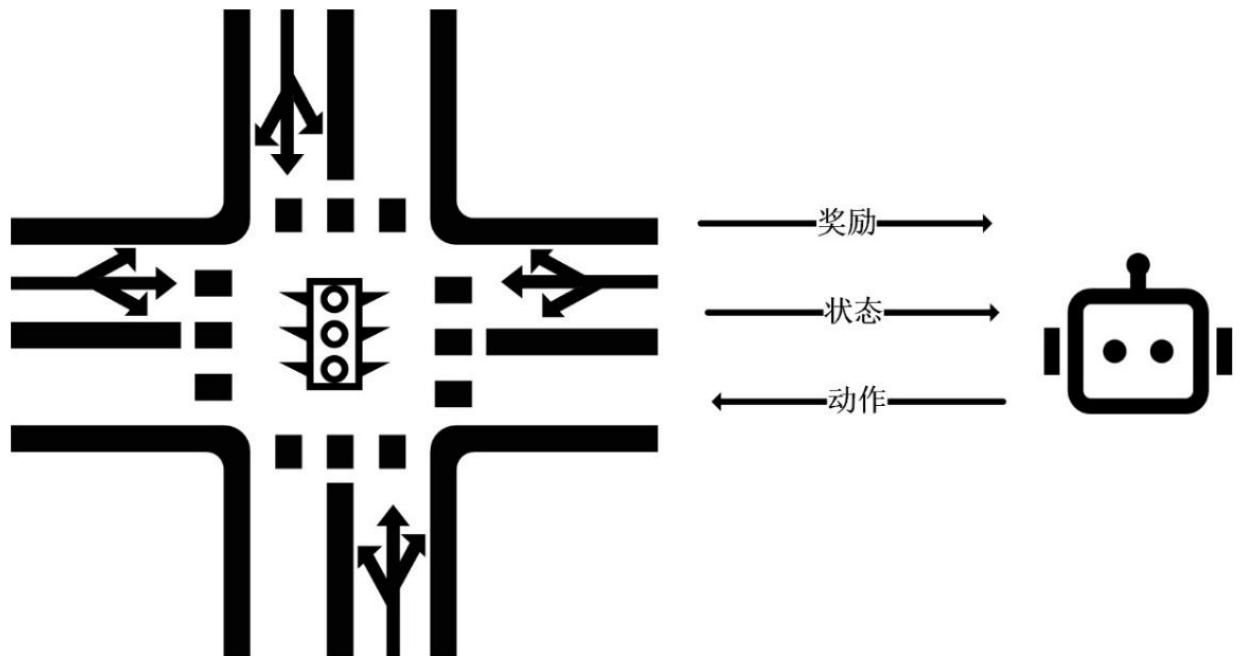


图 2.1 强化学习与环境的交互示意

### 2.1.1 马尔科夫决策过程

在强化学习范式中，智能体通过与环境的不断交互最后得到一个策略，并同时最大化奖励函数。通常来说，一个能被强化学习算法解决的问题可以被建模为一个马尔可夫决策过程<sup>[47]</sup>（Markov Decision Process, MDP）。在上一小节已经描述了智能体与环境之间的单次交互。这种交互就可以被抽象为一个 MDP。

用一个五元组<sup>[48]</sup>  $\langle S, A, R, \gamma, P \rangle$  描述该 MDP。它包含状态集合  $S$ ，由环境中所有可能产生的状态  $s_t$  组成；动作集合  $A$ ，由环境中所有合法的可以执行的动作  $a_t$  组成；奖励函数  $R$ ，在实际交互中，奖励函数通常是人工设计的，根据智能体的策略  $\pi_\theta(a_t | s_t)$  执行一次动作时，环境同步生成一个奖励值  $r$ ， $r$  一般是一个标量； $\gamma$  是奖励折扣因子，描述未来收益的现值，值在 0-1 之间；转移概率  $P$ ，描述在执行一个动作执行后，转移到下一个状态  $s_{t+1}$  的概率分布。一次轨迹的产生可以用  $S \times A \rightarrow S$  描述。对于一系列连续产生的 MDP 过程，它有公式(2.1)描述的性质：

$$P(s_{t+1} | s_t, a_t, r_t, s_{t-1}, a_{t-1}, \dots, r_1, s_0, a_0) = P(s_{t+1} | s_t, a_t) \quad (2.1)$$

表示转移函数仅依赖于当前状态  $s_t$  和采取的行动  $a_t$ 。

策略  $\pi_\theta(a_t | s_t)$  的目标是最大化累计奖励预期  $E[R_t | s, \pi]$ ，其中  $s$  表示环境最开始时的初始状态， $R_t$  的定义为如式(2.2)。

$$R_t = \sum_{i=0}^{T-1} \gamma^i r_{t+i} \quad (2.2)$$

其中， $r_{t+i}$  是每个时间步的奖励。几乎所有的强化学习算法都涉及价值函数的计算。价值函数是状态的函数，用来评估当前智能体在给定状态下有多好。把策略  $\pi$  下状态  $s$  的价值函数记为  $v_\pi(s)$ ，表示为公式(2.3)。

$$v_\pi(s, a) = E_\pi[R_t | S_t = s] \quad (2.3)$$

类似的，把策略  $\pi$  下状态  $s$  时采取动作  $a$  的价值记为  $q_\pi(s, a)$ ，表示为公式(2.4)。这就是根据策略  $\pi$ ，从状态  $s$  开始，执行动作  $a$  之后，所有可能的决策序列的期望回报。

$$q_\pi(s, a) = E_\pi[R_t | S_t = s, A_t = a] \quad (2.4)$$

称  $q_\pi$  为策略  $\pi$  的动作价值函数。

此外，强化学习模型可以根据转换概率矩阵  $P$  学习方式分为两类<sup>[45]</sup>，一类是基于模型（Model-based）的强化学习方法，需要学习状态之间的转换概率矩阵  $P$ ；另一类是无模型（Model-free）的强化学习算法，仅利用在交互过程中产生的轨迹进行学习，不需要显式的学习转换概率矩阵  $P$ ，也是目前较为主流的强化学习方法。在交通系统中，车辆与车辆之间的交互，车辆与信号灯之间的交互非常复杂，导致转换概率矩阵  $P$  难以得到。

因此，无模型的强化学习方法成为本文的选择。无模型的强化学习也可分为两类，基于值（Value-based）的方法与基于策略（Policy-based）的方法。在结合这两种方法以后，基于演员评论家的强化学习算法也被提出。

### 2.1.2 传统强化学习算法：Q-learning

强化学习算法根据与环境互动时的智能体可以被分为离轨和同轨策略。同轨策略中，用于生成采样数据的策略和用于实际决策的待评估和改进的策略是相同的，而离轨策略则是不同的。离轨策略下的时序差分控制算法的提出是强化学习早期的一个重要突破。这一算法被称为 Q 学习<sup>[49]</sup>，其改进方式如公式(2.5)所示。

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \max Q(S_{t+1}, a) - Q(S_t, a)] \quad (2.5)$$

此处，待学习的动作价值函数  $Q$  采用了对最优动作价值函数  $q^*$  的直接近似作为学习目标。与智能体决策序列轨迹的行动策略是什么无关。只需要所有的“状态-动作”二元组以及对应的价值可以被持续更新，整个学习过程就能被正确的收敛。 $Q$  学习算法的流程如下：

---

**Q 学习算法，用于预测  $\pi \approx \pi^*$**

---

算法参数：步长  $\alpha \in (0, 1]$ ，很小的  $\epsilon > 0$

初始化  $Q(s, a)$ ，其中  $Q(\text{终止状态}, \cdot) = 0$

对每幕：

    初始化  $S$

    对幕中的每一步循环：

        使用从  $Q$  得到的策略（例如  $\epsilon$ -贪心），在  $S$  处选择  $A$

        执行  $A$ ，观察到  $R, S'$

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \max Q(S_{t+1}, a) - Q(S_t, a)]$

$S \leftarrow S'$

    直到  $S$  是终止状态

---

$Q$  学习适用于离散状态和动作空间。在实际使用中，往往用一张表结构存储  $Q(S_t, A_t)$  信息。在经过足够多的回合后， $Q$  值表的  $Q(S_t, A_t)$  值会收敛到一个稳定值，表示最优  $Q$  值。在这个  $Q$  值表的基础上，可以定义一个策略，使得在每个状态下选择的动作都可以最大化累计奖励。最优策略  $\pi^*$  表示如式(2.6)。

$$\pi^*(s) = \arg \max_a Q(s, a) \quad (2.6)$$

显而易见的，在  $Q$  学习的过程中有一个强假设，即每个“状态-动作”二元组都可以被尽可能多（甚至是无穷次）的访问到<sup>[50]</sup>。这通过一个探索策略（例如  $\epsilon$ -贪心策略）来实

现。在训练初期，使用随机策略保证每个“状态-动作”二元组的访问。随着 Q 表不断更新，Q 值趋向于稳定，策略逐渐向贪婪过度。

## 2.2 深度强化学习

基于表格的学习方式在环境的状态和动作都是离散的，并且空间都比较小的情况下是适用的。比如常见的迷宫问题<sup>[51]</sup>和悬崖问题。当状态或者动作数量非常巨大时，这种做法不再适用。例如状态是一张 RGB 图像时，假设图像大小是  $210 \times 160 \times 3$ ，此时一共有  $256^{210 \times 160 \times 3}$  种状态。在计算机中存储这个数量级的 Q 值表显然是不切实际的。有些时候所需要解决的问题，其状态空间或者动作空间甚至是连续的，此时就有无限个“状态-动作”二元组<sup>[52]</sup>。再使用 Q 表的形式不可能解决问题。对于此类情况，使用函数拟合 Q 值就成为一种解决途径。本节介绍以深度学习作为函数拟合器的深度强化学习技术。

### 2.2.1 深度学习

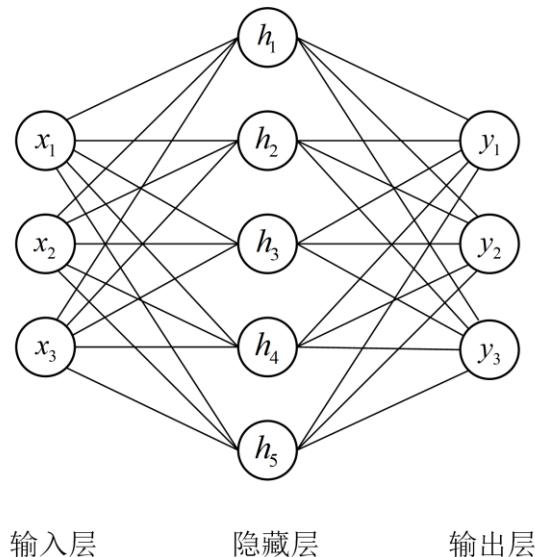


图 2.2 浅层神经网络架构示意

深度学习是机器学习的子领域，专注于使用多层神经网络来解决复杂的问题。这些神经网络通常被称为“深度”网络，因为它们由多个层次组成，每一层都负责从数据中提取不同层次的特征。深度学习的核心思想是通过这些层次的组合，自动地从原始数据中学习到有用的表示，从而实现对复杂任务的建模。深度学习的基本步骤是：

1. 前向传播：输入数据通过网络的每一层，逐层计算输出，直到最终输出层。每一层的输出作为下一层的输入，最终得到网络的预测结果。
2. 计算损失：使用损失函数（如均方误差、交叉熵等）计算预测结果与真实标签之间的差异，得到损失值。
3. 反向传播：从输出层开始，逐层计算损失函数对每一层参数的梯度。利用链式法则，

将误差从输出层反向传播到输入层。

4. 参数更新：使用计算得到的梯度，通过梯度下降法或其它优化算法（如 Adam<sup>[53]</sup>、 RMSprop<sup>[54]</sup>等）更新网络参数，使得损失函数最小化。

假设有一个简单的多层神经网络，如图 2.2 所示，包含输入层，隐藏层和输出层。前向传播过程中输出层到隐藏层可以被表示为：

$$z^{(1)} = W^{(1)}x + b^{(1)}a^{(1)} = \sigma(z^{(1)}) \quad (2.7)$$

其中， $x$  表示输入数据， $W^{(l)}$  表示第  $l$  层的权重矩阵， $b^{(l)}$  表示第  $l$  层的偏置向量， $z^{(l)}$  表示第  $l$  层线性组合的结果，即  $z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)}$ ， $a^{(l)}$  表示第  $l$  层的激活值，即  $a^{(l)} = \sigma(z^{(l)})$ ，其中  $\sigma$  是激活函数。数据从隐藏层到输出层这一过程被表示为：

$$z^{(2)} = W^{(2)}a^{(1)} + b^{(2)} \quad (2.8)$$

$$\hat{y} = \sigma(z^{(2)}) \quad (2.9)$$

$\hat{y}$  表示网络预测输出，对应的，用  $y$  表示真实标签。假设用均方误差作为损失函数，用  $L$  表示，此时有：

$$L = \frac{1}{2} \|y - \hat{y}\|^2 \quad (2.10)$$

在反向传播过程中，权重和偏置的梯度为：

$$\frac{\partial L}{\partial W^{(2)}} = \frac{\partial L}{\partial z^{(2)}} \cdot a^{(1)} \quad (2.11)$$

$$\frac{\partial L}{\partial b^{(2)}} = \frac{\partial L}{\partial z^{(2)}} \quad (2.12)$$

$$\frac{\partial L}{\partial W^{(1)}} = \frac{\partial L}{\partial z^{(1)}} \cdot x \quad (2.13)$$

$$\frac{\partial L}{\partial b^{(1)}} = \frac{\partial L}{\partial z^{(1)}} \quad (2.14)$$

采用梯度下降法更新参数时，学习率定义为  $\eta$ ，表示为：

$$W^{(l)} \leftarrow W^{(l)} - \eta \frac{\partial L}{\partial W^{(l)}} \quad (2.15)$$

$$b^{(l)} \leftarrow b^{(l)} - \eta \frac{\partial L}{\partial b^{(l)}} \quad (2.16)$$

至此，完成了具有一层隐藏层的简单神经网络的整个步骤。深度学习模型的核心思想是利用多层神经网络来逼近复杂的函数。每一层神经网络可以看作是一个简单的非线性函数，通过多层的组合，整个网络能够逼近任意复杂的函数。同时，利用反向传播算法自动更新网络参数，从而避免了手动计算复杂函数的需求，极大简化了求解过程。强化学习技术与深度学习的结合，利用深度神经网络能够处理复杂的环境和状态表示，突破了传统表格型方法的局限性，推动了深度强化学习的发展。

## 2.2.2 最大池化

最大池化(Max Pooling)是一种常见下采样操作，能够有效减少特征图的空间维度，从而降低后续层的计算负担<sup>[55]</sup>。此外，最大池化还具有一定的平移不变性，使得网络对输入图像中的微小位移更加鲁棒，增强了模型的泛化能力。图 2.3 是最大池化示意。

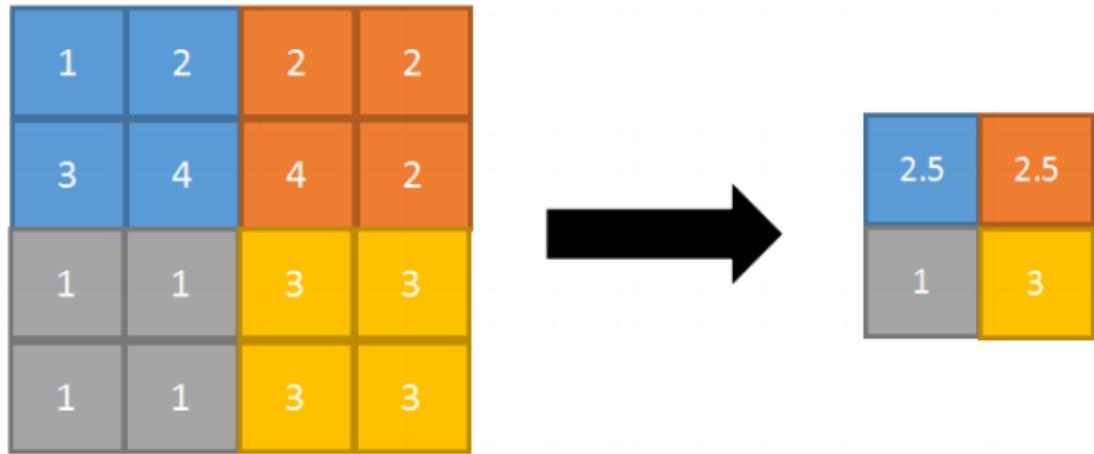


图 2.3 最大池化示意图

## 2.2.3 矩归一化

矩归一化(Moment Normalization)是一种数据预处理技术<sup>[56]</sup>。通常用于将数据集的统计特性(如均值和方差)归一化到一个标准范围内。矩归一化的主要目的是消除数据中的量纲差异，使得不同特征之间具有可比性，从而提高机器学习模型的性能。对于数据中的每个特征  $x$ ，计算矩阵归一化的公式为式(2.17)。

$$x_{norm} = \frac{x - u}{\delta} \quad (2.17)$$

$x$  是原始数据中的一个值， $u$  是该特征的均值， $\delta$  是该特征的标准差。

## 2.2.4 L2 归一化

L2 归一化也称为欧几里得归一化，用于将数据向量长度缩放到单位长度，从而消除量纲差异，提高模型性能和稳定性<sup>[57]</sup>。从实现上，每个元素会除以其 L2 范数。对于一个数据向量  $x = [x_1, x_2, \dots, x_n]$ ，L2 归一化公式如式(2.18)。

$$x_{norm} = \frac{x}{\|x\|_2} \quad (2.18)$$

$\|x\|_2$  是向量  $x$  的欧几里得范数，计算公式如式(2.19)。

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} \quad (2.19)$$

## 2.2.5 深度 Q 学习

有了深度学习，现在可以处理强化学习中由状态和动作空间过大带来的问题。Q 学习与深度学习结合的算法被称为深度 Q 学习<sup>[58]</sup>。在这种情况下，Q 值不再需要通过 Q 值表进行更新，而是通过神经网络学习一组权重模型作为替代。这种方法减少了存储需求，能够处理连续和高维的状态和动作空间，极大地扩展了 Q 学习的应用范围。

回顾传统 Q 学习的更新公式(2.5)，采取时序差分学习(Temporal Difference Learning)，通过计算  $R_{t+1} + \gamma \max Q(S_{t+1}, a)$  增量式的更新 Q 值，也就是说希望使得 Q 值向  $R_{t+1} + \gamma \max Q(S_{t+1}, a)$  的值靠近。很自然的，可以将深度 Q 学习中的神经网络的损失函数构造为均方误差的形式如式(2.20)。

$$L(\theta) = \frac{1}{2N} \sum_{(s, a, r, s') \in \mathcal{B}} (r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta))^2 \quad (2.20)$$

其中， $\theta$  是神经网络的参数， $\theta^-$  是目标网络的参数。通过最小化这一损失函数，神经网络可以逐步优化其权重，从而更准确的更新 Q 值。

有监督学习中，一般假设数据是独立同分布(independent identically distribution, iid)的，每次训练神经网络时从训练数据中随机抽样多个数据来进行更新。在 DQN 中，人们通常会维护一个经验池(experience replay)，表示为  $\mathcal{B}$ ，将智能体与环境交互得到的轨迹数据存入经验池中，使之满足 iid 假设，提高样本利用效率。

但注意到， $R_{t+1} + \gamma \max Q(S_{t+1}, a)$  中本身就有神经网络输出的 Q 值。在更新参数的同时，目标也在不断的改变，这使得神经网络变得不那么稳定，为了解决这一问题，目标网络(target network)被提出。主要思想是在训练过程中先将计算  $R_{t+1} + \gamma \max Q(S_{t+1}, a)$  中 Q 值的网络固定住，延后其更新。具有目标网络的更新过程如下：

---

### 采用目标网络的深度 Q 学习算法

---

初始化经验回放池  $\mathcal{R}$ ，用随机的网络参数  $w$  初始化网络，复制相同的参数来初始化目标网络

episode  $e = 1 \rightarrow E$

选择环境初始状态  $s_0$

时间步  $t = 1 \rightarrow T$

根据当前网络以  $\epsilon$  贪婪策略选择动作  $a_t$

执行动作  $a_t$ ，获得回报  $r_t$ ，状态变为  $s_{t+1}$ ，并将  $(s_t, a_t, r_t, s_{t+1})$  存储在回放池  $\mathcal{B}$  中

若  $\mathcal{B}$  中数据足够，从  $\mathcal{B}$  中采样  $N$  个数据  $\{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^N$

对每个数据，用目标网络计算  $y_i = r_i + \gamma \max_a Q^\omega(s_{i+1}, a)$

最小化目标函数损失  $L = \frac{1}{N} \sum_{i=1}^N (y_i - Q(s_i, a_i))^2$ ，以此更新当前网络  $Q(s, a)$

每  $C$  个时间步同步更新一次目标网络

---

## 2.2.6 策略梯度方法

无论是 Q 学习还是深度 Q 学习都是基于动作价值函数的方法，先得到动作价值函数，再根据动作价值函数去选择动作。在本节，策略将被参数化从而直接学习，动作也将不再完全依赖动作价值函数。用  $\theta \in \mathbb{R}^d$  表示策略的参数向量，策略参数的学习方法往往基于某种性能度量  $J(\theta)$  的梯度，这些梯度是标量  $J(\theta)$  对策略参数的梯度<sup>[59]</sup>。这些学习方法的目的是最大化性能指标，因此它们的更新近似于  $J$  的梯度上升，表示为式(2.21)。

$$\theta_{t+1} = \theta_t + \alpha \nabla J(\theta_t) \quad (2.21)$$

其中， $J(\theta_t)$  的期望是性能指标对它参数  $\theta_t$  的梯度的近似。在函数逼近的情况下，通过调整策略参数达到性能改善的目的实际上相当具有挑战性。性能依赖于动作的选择，也依赖于选择动作时所处的状态，而二者都会被策略参数影响。此处，策略梯度定理<sup>[60]</sup>给出了策略对状态分布的影响：

$$\nabla J(\theta) \propto \sum_s u(s) \sum_a q_\pi(s, a) \nabla \pi(a | s, \theta) \quad (2.22)$$

分布  $u$  是策略  $\pi$  下的同轨策略分布。策略梯度定理右边将目标策略下每个状态出现的频率作为加权系数，如果按策略  $\pi$  执行，那么状态将按比例出现，右边可以被化为：

$$\mathbb{E}_\pi [\sum_a q_\pi(S_t, a) \nabla \pi(a | S_t, \theta)] \quad (2.23)$$

之后，在公式中引入动作价值期望  $A_t$ ，把所有对随机变量的可能取值的求和运算替换为对  $\pi$  的期望，然后对期望进行采样，过程如下：

$$\begin{aligned} \nabla J(\theta) &= \mathbb{E}_\pi (\nabla_\theta q_\pi(S_t, a) \nabla_\theta \pi(a | S_t, \theta)) \\ &= \mathbb{E}_\pi \left( \sum_a \pi(a | S_t, \theta) q_\pi(S_t, a) \frac{\nabla_\theta \pi(a | S_t, \theta)}{\pi(a | S_t, \theta)} \right) \\ &= \mathbb{E}_\pi \left( q_\pi(S_t, A_t) \frac{\nabla_\theta \pi(A_t | S_t, \theta)}{\pi(A_t | S_t, \theta)} \right) \\ &= \mathbb{E}_\pi \left( G_t \frac{\nabla_\theta \pi(A_t | S_t, \theta)}{\pi(A_t | S_t, \theta)} \right) \end{aligned} \quad (2.24)$$

这个量可以通过采样计算得到，它的期望等于真实的梯度，替换随机梯度上升算法更新中的式(2.21)中的  $\nabla J(\theta)$ ，得到 REINFORCE 算法<sup>[60]</sup>更新过程如式(2.25)。

$$\theta_{t+1} = \theta_t + \alpha G_t \frac{\nabla_\theta \pi(A_t | S_t, \theta)}{\pi(A_t | S_t, \theta)} \quad (2.25)$$

这是一个策略梯度算法更新的基本框架，每个增量更新都正比于回报  $G_t$  和一个向量的乘积，该向量是选取动作的概率的梯度除以这个概率本身。同时注意到，由于  $G_t$  的存在，只有在分幕式情形下才能很好的使用该算法，从这个角度说，REINFORCE 是一个蒙特卡洛算法。

### 2.2.7 演员-评论家结构的强化学习算法

前文介绍了基于值的和基于策略梯度的强化学习算法。将二者结合起来，同时学习策略和价值函数，这样的方法被称为演员-评论家算法。“演员”是指学习到的策略，用于进行决策；“评论家”指学习到的价值函数，用于评估状态的价值。演员评论家结构也被称为 AC (Actor-Critic, AC) 结构。

近端策略优化 (Proximal Policy Optimization, PPO) 是典型的 AC 结构的强化学习算法<sup>[61]</sup>。其核心思想是限制策略更新的幅度保证策略更新的稳定性。具体来说，PPO 的优化目标表示如式(2.26)。

$$\max_{\theta} \mathbb{E}_{(s_t, a_t) \sim \pi_{\theta_{\text{old}}}} \left\{ \min \left[ r_t(\theta) \hat{A}(s_t, a_t), \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) \cdot \hat{A}(s_t, a_t) \right] \right\} \quad (2.26)$$

其中，clip 用于裁剪梯度，保证策略更新的稳定性。除了直接使用 clip 裁剪梯度之外，还可以使用 KL 散度 (Kullback-Leibler divergence) 做显式惩罚。此时式(2.26)重写为：

$$\max_{\theta} \mathbb{E}_{(s_t, a_t) \sim \pi_{\theta_{\text{old}}}} \left\{ \min \left[ r_t(\theta) \hat{A}(s_t, a_t) - \beta \text{KL}[\pi_{\theta_{\text{old}}}(\cdot | s_t), \pi_{\theta}(\cdot | s_t)] \right] \right\} \quad (2.27)$$

相比来说，使用 KL 散度作为乘法项时会带来额外的计算开销；而直接裁剪梯度是一种硬性限制，灵活性更低。而  $\hat{A}(s_t, a_t)$  被称为优势函数，是一种用于强化学习中的优势估计方法，衡量了某个状态下采取某个动作相对于平均表现得优劣程度。具体来说，优势函数定义如式(2.28)。

$$\hat{A}(s_t, a_t) = Q(s, a) - V(s) \quad (2.28)$$

$Q$  和  $V$  的意义参照式(2.3)与(2.4)。优势函数在 PPO 中非常重要，它帮助区分哪些动作可以获得更大的累计奖励，哪些动作比平均动作表现更差。但实际应用中，直接使用优势函数可能会导致高方差的问题，因此 Schulman 等人提出 GAE<sup>[62]</sup> (Generalized Advantage Estimation, GAE)，表示为式(2.29)。

$$\hat{A}_t^{\text{GAE}(\gamma, \lambda)} = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l}^V \quad (2.29)$$

其中：

$$\delta_{t+l}^V = r_{t+l} + \gamma V(s_{t+l+1}) - V(s_{t+l}) \quad (2.30)$$

通过调整衰减因子  $\gamma$  和权衡参数  $\lambda$  来平衡估计的偏差和方差。

公式(2.26)以及(2.27)中的  $r_{\theta}$  是重要性采样 (Importance Sampling) 比率，用以衡量新旧策略之间的差异性。用新旧策略表示为式(2.31)。

$$r_t(\theta) = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} \quad (2.31)$$

用以调整旧策略下生成数据在新策略下的权重。

从实现上来说，AC 结构的算法通常有两种大体上的架构。演员和评论家共享部分网络参数，这部分网络往往和输入紧密相连，在反向传播时，使用一个融合的损失函数。这种架构被证明了会损失一些神经网络的表达能力。除此之外，演员和评论家各自使用完全独立的神经网络也是一种常见架构。

## 2.3 多智能体强化学习

在现实环境中，单交叉口通常不作为单独的研究对象，更希望在路网规模上解决信号控制问题。前文的深度强化学习算法确实可以较为良好的解决单交叉口的信号控制问题。但当规模不断提升以后，贪心式的优化单交叉口的通行能力不能满足实际通行的需要。尝试将整个路网视为一个完全紧凑的系统也是不可取的，单智能体难以处理随着交叉口数量增加而不断增加的状态空间和动作空间。为了解决大规模路网的信号控制问题，引入多智能体强化学习。

### 2.3.1 多智能体框架

使用一个八元组定义多智能体框架为  $G = \{S, U, P, r, Z, O, n, \theta\}$ 。其中动作  $a \in A \equiv 1, \dots, n$ ，表示  $n$  个智能体的动作空间；环境的状态表示为  $s$ ， $s \in S$  在每个做决策的时间步，每个智能体采取一个动作  $u^a \in U$ ，形成一个联合动作；根据状态转移函数  $P : S \times U \times S \rightarrow R$ ，由环境自行生成下一个状态  $s_{next}$ ； $\theta$  的定义与单智能体系统中相同，同为奖励折扣因子。此外，对于多智能体强化学习，全局奖励往往更受关注，并且它往往不是简单的每个智能体奖励的数值和。

### 2.3.2 集中式控制和去中心化控制

在环境完全可观察的设置中，有研究尝试过上文提及的全局只是用一个智能体<sup>[63]</sup>。该智能体作为集中控制器尝试学习一个最优策略  $\pi^c(u|s_t)$  去最大化全局奖励。但这种方式的特性使得其不适用于交通信号控制任务。一方面，集中控制只有一个智能体，这就要求信息需要汇集到一处，而交通信息有时不能是全透明的，此时就产生了信息安全问题；另一方面，即便解决了信息安全问题，即交通信息可以进行传递，但信息传递的速度能否满足信号控制的实时性能要求还不得而知。除此之外，当信号控制的范围不断扩大，单智能体强化学习本身难以在如此巨大的状态空间中求解。

去中心化控制在形式上更加契合现代信号控制的需要<sup>[64]</sup>。因此本文基于去中心化控制开展研究。在这种情况下，局部或者每个交叉口有一个强化学习智能体，维持一个本地策略  $\pi^l(u^a|s_t)$ ，分解了联合动作概率分布为式(2.32)。

$$P(u|s_t) = \prod_a \pi^a(u^a|s_t) \quad (2.32)$$

形式上的分解一定程度缓解了状态空间爆炸问题。这也意味着每个智能体可以独立进行决策，而不再需要聚合远处交叉口的交通信息。

### 2.3.3 部分可观测性

在集中式控制中已经说明，全局信息聚合到一起时往往伴随着信息安全问题和传输延时问题。因此在去中心化控制中，明显的每个智能体不可能都观测到全局信息<sup>[65]</sup>。这种性质被称为部分可观测性，有这种性质发生的马尔可夫过程又可以被称为部分可观察的马尔可夫决策过程（Partially Observable Markov Decision Process, POMDPs）。在这种情况下，每个智能体只能观察到有限的部分，所有智能体观测的并集是全局的状态，表现为式(2.33)。

$$\bigcup_{i=1}^N O_i = S \quad (2.33)$$

部分可观测性导致智能体之间所拥有的信息不对等，因此不同智能体之间的合作就显得越发重要。完全贪心式的优化时，各个智能体之间会产生竞争，而实际上希望所有智能体能达成合作的目的。

### 2.3.4 独立 Q 学习（Independent Q-learning, IQL）

IQL 是 Q 学习在多智能体领域的扩展<sup>[66]</sup>，每个智能体去中心化的学习一个 Q 函数。IQL 扩展后，所有智能体都在同时更新，环境会变得极度不平稳。但在实际实验结果中，IQL 有时可以表现出良好的优化效果。

为缓解扩展后的非平稳性，深度循环 Q 网络（Deep Recurrent Q learning, DRQN）被提出<sup>[67]</sup>。它尝试用循环神经网络捕捉历史信息中的隐藏动态特性，代替空间上的部分可观测性。聚合历史信息也是解决部分可观测性的一种可能途径。

### 2.3.5 QMIX 算法

部分可观测性导致本地智能体无法了解到全局信息，因此也不知道自身在全局奖励中的占比。QMIX 算法<sup>[68]</sup>使用一种混合网络来协调多个智能体的 Q 值从而实现最优策略的学习。具体来说，全局 Q 值  $Q_t$  是通过混合网络计算得到的，表示为式(2.34)。

$$Q_t = Mix(Q_1, Q_2, \dots, Q_n) \quad (2.34)$$

混合网络直接输出全局 Q 值。QMIX 算法要求混合网络权重是正的，以确保全局 Q 值是本地 Q 值的单调函数，这要求混合网络权重 W 和偏置 b 满足式(2.35)的要求。

$$\frac{\partial Q_{\text{tot}}}{\partial Q_i} \geq 0 \quad \forall i \quad (2.35)$$

但在实际训练过程中，QMIX 算法还是依赖于全局信息，在决策时可以只依赖于本地信息。因此信息安全方面还是存在问题。

### 2.3.6 基于通信的多智能体强化学习（Communication-Based Multi-Agent, COMA）

虽然交叉口之间的信息的交互可能是危险的，此处仍介绍一个基于通信的多智能体强化学习 COMA 作为了解。在多智能体环境中，多个智能体同时行动，全局奖励是集体行为的结果。如何将全局奖励合理分配给每个智能体，以指导其策略优化，是一个关键挑战。在 COMA<sup>[69]</sup>的神经网络架构中，每个智能体有一个通信网络进行显式通信，用于发送和接收信息。通信网络可以是简单的消息传递机制，也可以是复杂的神经网络；一个策略网络，它以智能体的局部观测和接收到的通信信息作为输入，输出动作的概率分布；一个价值网络，用于估计当前状态的预期回报。价值网络的输入通常包括局部观测和接收到的通信信息。

## 2.4 城市交通系统

本文的研究目的在于从路网层面解决部分信号控制问题中的难题，本节介绍交通相关的背景知识<sup>[70-71]</sup>。

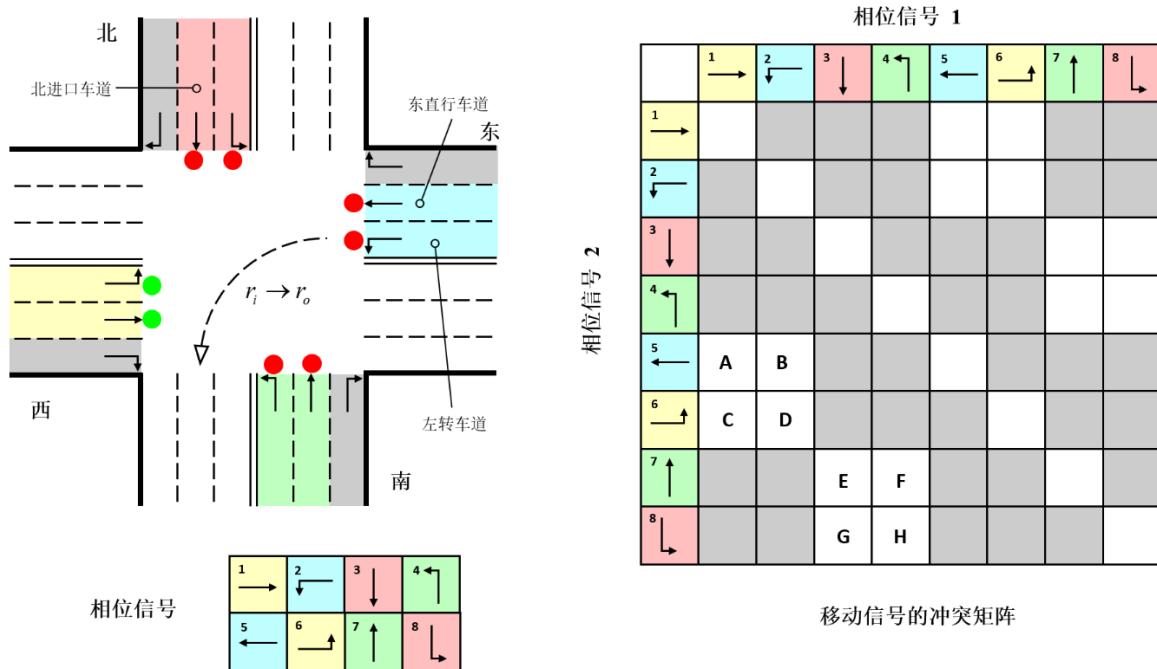


图 2.4 信号灯相位的定义

### 2.4.1 道路结构和交通运动

1. 路段：路段末端交汇形成交叉口。在交叉口中有两种路段，进口路段和出口路段。车辆从进口路段驶入交叉口，从出口路段驶出交叉口。图 2.4 展示了一个传统标准交叉

口，有四条进口路段和四个出口路段。

2. 车道：一个路段由多条车道组成。与路段类似，交叉口结构中有两种车道，进口车道和出口车道。不同路段可能具有不同数量的车道数量。车辆从进口车道驶入交叉口，经过交汇处后通过出口车道驶出交叉口。
3. 交通运动：交通运动指的是车辆从一个车道输入到从另一个车道离开交叉口的整个过程。图 2.4 中定义了一个交通运动  $r_i \rightarrow r_o$ ，表示车辆从进口道  $r_i$  驶入交叉口，从  $r_o$  驶出交叉口。通常交通运动有三种，左转、右转与直行。非标准交叉口中可能不具有所有的交通运动类型。掉头这种特殊形式不在考虑中。

#### 2.4.2 信号灯

1. 交通信号：交通信号通过交通运动来定义。绿灯代表着对应的运动是允许的，红灯代表对应的交通运动是禁止的。对于图 2.4 所示的标准交叉口，右转运动可以无视信号灯，同时该交叉口有八种交通运动信号。不同方向的来车进行的左转或者右转等视为不等的交通运行。
2. 相位：相位由多个交通信号组合起来。图 2.4 也展示了不同运动信号之间的冲突矩阵。矩阵中灰色各自表示对应的两个运动信号是冲突的，不能在一个相位中同时变为绿色。白色各自表示对应的运动信号是无关的，因此可以在同一个相位中同时变成绿的。所有的没有冲突的相位信号组合在一起可以生成八种合法相位。
3. 相序：预设的相位安排，也是固定配时的形式。可以用相位序号和对应的持续时间来表示，类似：

$$(p_1, t_1)(p_2, t_2) \dots (p_i, t_i) \dots$$

$p_i$  表示相位序号，或是具体的某个相位的放行方案。 $t_i$  表示该相位的持续时间。

#### 2.4.3 道路、车辆相关参数

车道上车辆表现出的一些交通流参数是实际优化中比较看重的。例如停车次数，吞吐量，旅行时间和排队长度。旅行时间是车辆从驶入路网到驶出路网的差值，也是强化学习过程中常见的优化目标之一<sup>[72-74]</sup>。吞吐量则是指指定时间内驶出交叉口的车辆数目，同样是常见的优化目标<sup>[75-77]</sup>。

#### 2.4.4 其它常见参数

在交通中，车辆并非唯一的参与者，行人同样是重要参与者<sup>[78]</sup>。信号灯设置也需要考虑到行人的出行需要，因此在实际决策时，需要考虑到行人过马路的时间设置最小绿灯时间。除此之外，当前相位结束时，交叉口内往往还有一部分车辆。此时如果

直接跳转到其它相位方案可能会产生冲突，因此在一个相位结束后，往往会衔接一段时间的专门用于清空交叉内车辆口的相位。

## 2.5 本章小结

本章从强化学习出发，首先介绍了传统强化学习涉及的一些概念。在指出传统强化学习的一些限制后，引入了深度学习，借由深度学习的拟合能力描述了强化学习到深度强化学习的转换，解决了由状态空间和动作空间引起的一些问题。接下来由单交叉口扩展到路网，将单智能体强化学习扩展到多智能体强化学习。在梳理深度强化学习的脉络时，同时介绍了对应的代表算法。最后，回到本文希望解决的信号控制问题上，给出了交通中常见的一些定义与概念。为本文接下来解决信号控制提供了基础。

### 3 交通场景中异构观测间分布距离计算

在运用强化学习框架解决信号控制问题时，需要优先考虑状态、奖励和动作的设计。状态应反映交叉口的交通状况，但具体应选用哪些参数来刻画交通状况仍需深入探讨。本章通过互模拟度量的扩展，计算不同形式观测间的距离，得到最优状态形式。

#### 3.1 问题描述

假设有一个交叉口，交叉口的结构表示如图 3.1。

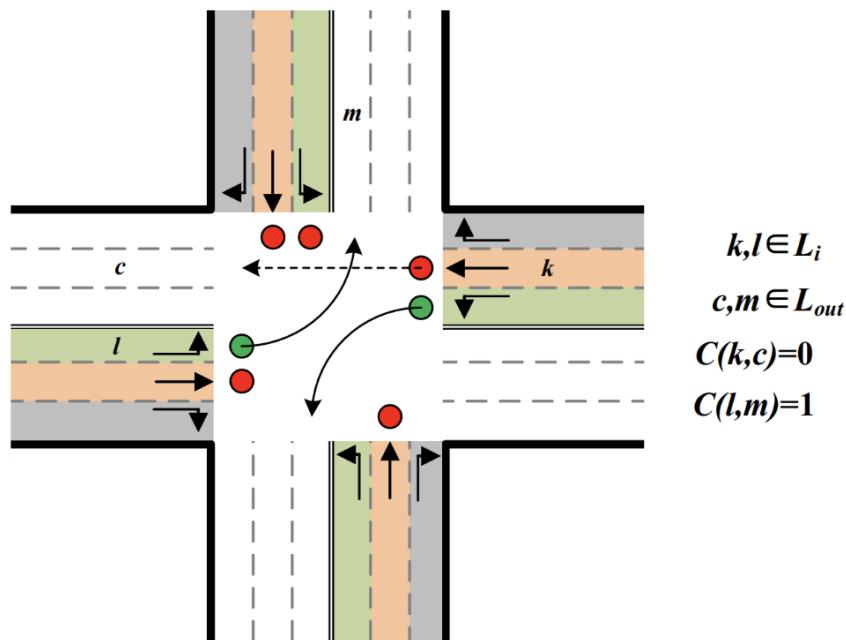


图 3.1 单交叉口结构

交叉口有两种车道，进口车道( $l \in L_i$ )与出口车道( $l \in L_o$ )。两种车道组合起来形成的交通结构称为连接器  $c$ 。连接器  $c(l, m)$  表示车辆可以从车道  $l$  进入交叉口，然后从车道  $m$  离开。将一个表示连接器车辆数目的值与连接器绑定起来，用  $x(l, m)(t)$  表示在时间段  $t$  内连接器车辆数目。现在，信号控制方案清空整个交叉口车辆的能力可以被表示式(3.1)。

$$\begin{aligned} x(l, m)(t+1) &= x(l, m)(t) - [C(l, m)(t+1)S(l, m)(t) \wedge x(l, m)(t)] \\ &\quad + d(l, m)(t+1), l \in L_i, m \in L_o \end{aligned} \quad (3.1)$$

$y \wedge z = \min(y, z)$ ，当信号灯允许连接器  $C(l, m)$  放行时， $S(l, m) = 1$ ，禁止车辆从该连接器驶出交叉口时  $S(l, m) = 0$ 。 $c(l, m)(t)$  表示可能通过连接器  $C(l, m)$  离开交叉口的车辆数目。 $d(l, m)(t+1)$  表示从交叉口外进入当前交叉口的车辆数目，对于单交叉口来说，当前交叉口的车辆输入来源于仿真软件给定的输入；对于复杂的路网来说，交叉口内的车辆输入一方面来源于仿真软件给定输入，另一方面来源于相邻交叉口之间的车辆流动。此

时，在固定时间段内，可以将模型清空交叉口内车辆的能力表示为式(3.2)。在最小化此式的过程中，交叉口内的车辆被不断清空以减少拥堵。

$$\min \left( \sum_{t>0} \sum_{l \in L_i, m \in L_o} x(l, m)(t+1) - x(l, m)(t) \right) \quad (3.2)$$

此时已经得到了信号控制问题的优化目标。此时需要将优化框架带入到强化学习框架中。强化学习希望学习到一个策略  $\pi_\theta(a_t | s_t)$ ，以最大化或者最小化优化目标。在策略表示中，模型需要基于当前状态  $s_t$  选择一个动作  $a_t$ 。然而在实际的算法中，状态虽然反应了环境的信息，但实际上智能体仅能获取到部分观测。

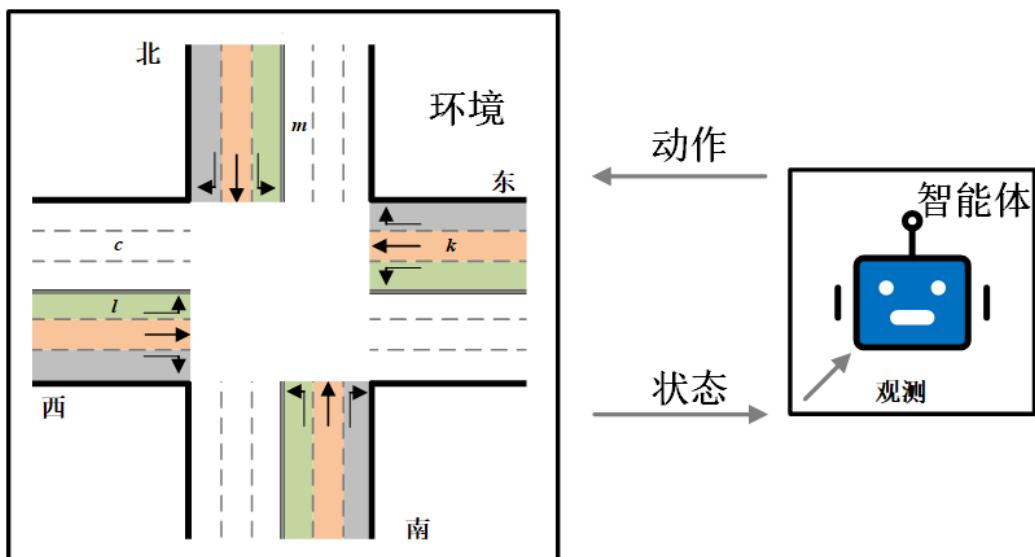


图 3.2 状态到观测的转变

全量的信息使得状态空间变大，这会导致智能体难探索到最优优化路径。因此，本章想要解决的问题是，在算法本身不变的情况下，通过调整智能体所得到的观测的形式达到更好的优化目的。

## 3.2 主要贡献

传统的信号控制算法在适应日益复杂的交通环境方面正变得越来越具有挑战性。因此，强化学习（RL）作为一种可行的方法，逐渐被用于解决信号控制问题。在开发基于强化学习的算法时，需要从设计状态、动作和奖励的形式开始。在算力的局限下，模型必须在巨大的状态空间中找出最优轨迹。状态空间的表示也因此成为一个难题。在本章中，提出了一种基于分布距离的模块，该模块能够计算不同观测形式之间的距离。选取一种常见的观测表示作为基准，并计算其它观测形式与该基准的分布距离。通过组合距离最大的几种观测形式，得到了信息熵最大的状态表示。此外，还验证了该状态表示的优越性。

### 3.3 基于分布距离的观测距离计算

#### 3.3.1 传统的同构观测之间的距离度量

早先已经有研究将单一的状态空间分为多个子状态空间以减少计算复杂度<sup>[44,79-80]</sup>。注意到这些研究的状态空间表示往往是单一的，即智能体只能从环境中获取某一类指定的信息，而且难以改变。但是该类研究为计算异构观测之间的相似性提供了思路。在信号控制问题中，假设整个状态空间为  $S$ 。一般来说，异构观测之间的相似性是无法计算的，因为从实际表示形式上就没有统一性。例如车道上平均车辆速度与车道的空间占有率，这两个属性完全无法计算相似性。

假设某个观测的空间表示为  $s$ ,  $s$  是  $S$  的非空子集。假设存在一个最优状态空间表示，表示为  $s_i$ , 则式(3.3)是成立的<sup>[81]</sup>:

$$\forall s_j \in S \setminus s_i, \forall s_j' \in s_j, \exists s_i' \in s_i : \pi(s_i') = \pi(s_j') \quad (3.3)$$

$S \setminus s_i$  表示空间划分。式(3.3)可以被改为状态转移函数和奖励函数的表示为式(3.4)和(3.5)的组合。

$$\mathcal{R}(s_i', a) = \mathcal{R}(s_j', a) \quad \forall a \in \mathcal{A} \quad (3.4)$$

$$\mathcal{P}(G|s_i', a) = \mathcal{P}(G|s_j', a) \quad \forall a \in \mathcal{A}, \quad \forall G \in \mathcal{S} \quad (3.5)$$

根据式(3.4)以及式(3.5)可以得到合并两个相似状态空间的标准：在执行同一动作后，可以得到相近的奖励值；具有相似的状态转移函数。严格坚持最优表示是不切实际的，因为智能体对环境的任何变化都十分敏感。因此，通过定义一个度量  $d$  弱化上式的严格性。定义一个常数  $c \in [0,1]$ ， $d$  的定义为式(3.6)。

$$\begin{aligned} d(s_i', s_j') &= \max_{a \in A} (1 - c) \cdot |\mathcal{R}(s_i', a) - \mathcal{R}(s_j', a)| \\ &\quad + c \cdot W_1(\mathcal{P}(G|s_i', a), \mathcal{P}(G|s_j', a); d) \end{aligned} \quad (3.6)$$

$W_1$  是 1<sup>th</sup> 瓦氏度量<sup>[82]</sup>，表示为  $W_1(P_i, P_j; d) = \left( \inf_{\gamma' \in \Gamma(P_i, P_j)} \int_{S \times S} d(s_i, s_j) \gamma'(s_i, s_j) \right)$ 。公式又被称

为互模拟度量<sup>[83]</sup>，用作计算同构观测空间的相似性。

#### 3.3.2 异构观测之间的距离度量

前文已经给出了同构观测之间的距离度量。本节将其扩展到异构观测之间的距离度量。以任意两种不同类型的观测形式为例子，每一种观测形式有其自己的观测空间，假设两种观测空间为  $o_i$  与  $o_j$ 。假如发生了一次状态转移，可以从中得到以下的马尔可夫过程：

$$\{o_i^{'}, a, o_i^{''}, r\}$$

$$\{o_j^{'}, a, o_j^{''}, r\}$$

其中  $a$  表示动作，重写上式为式(3.7)。

$$\{(o_i^{'}, o_j^{'}) \times a \rightarrow (o_i^{''}, o_j^{''})\} \quad (3.7)$$

假设以上两种观测空间是近似的，那么存在一个函数  $f$ ，对于所有马尔可夫过程，有式(3.8)所示的属性。

$$f(o_j^{'}) = o_i^{'} + \varepsilon' \quad (3.8)$$

当  $\varepsilon=0$  时，两种观测形式可以被认为是完全等价的。但是注意到，式(3.8)其实并未考虑状态转移概率。将转移概率引入式(3.7)中。首先，因为是同时发生的状态转移，奖励值相同，因此公式(3.7)中  $|\mathcal{R}(s_i^{'}, a) - \mathcal{R}(s_j^{'}, a)| = 0$ 。得到式(3.9)。

$$d(s_i^{'}, s_j^{'}) = c \cdot W_1(\mathcal{P}(G|s_i^{'}, a), \mathcal{P}(G|s_j^{'}, a); d) \quad (3.9)$$

所以要计算的只有两种形式观测的 1<sup>th</sup> 瓦氏度量。而 MDP 同时发生，因此直接度量一个 MDP 形式上的相似性。表示为式(3.10)。

$$d(f_i(o_i^{'}) \rightarrow o_i^{''}), f_j(o_j^{'}) \rightarrow o_j^{''}) \quad (3.10)$$

状态转移概率也并没有显式计算。在状态转移发生时，所有形式的观测同时发生转移。将状态转移前后的观测拼接起来并进行神经网络编码，这其中包含了状态转移关系。

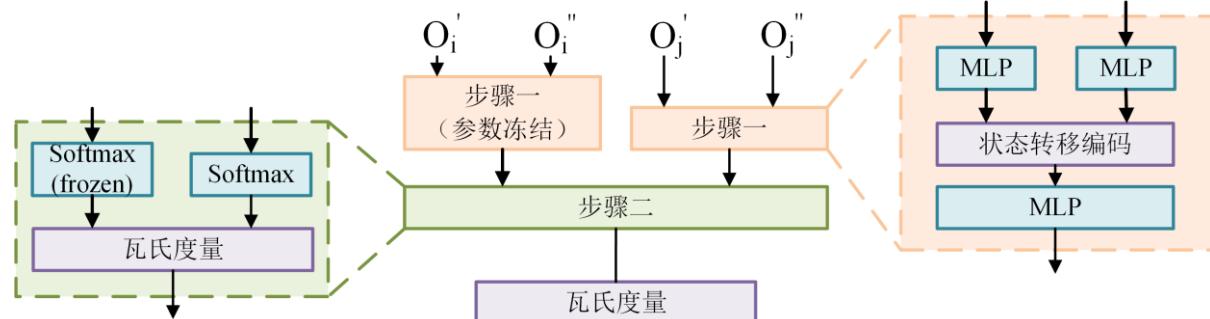


图 3.3 观测距离度量模块

接下来需要得到该函数  $f$ 。此处，使用浅层神经网络实现该映射。信号控制问题中的观测形式往往是实数值，通过神经网络学习两个观测分布之间的最小距离。浅层神经网络的拟合能力可以实现这种映射。之后用瓦氏距离作为损失函数进行反向传播。同时注意到，由于观测形式是固定的，因此损失函数绝不会减少至 0。图 3.3 表示观测距离度量模块。固定一种观测形式作为锚（这里选取为等待车辆数目  $x_w(l)$ ）。与锚相关的网络参数不会参与更新。整个网络结构用图 3.3 表示。用它作为基准，计算其余观测形式与它之间的距离。

## 3.4 实验设置

### 3.4.1 实验数据

单交叉口模型使用开源道路仿真软件 SUMO<sup>[84]</sup>进行搭建。为验证所提出模块的鲁棒性，在标准单交叉口 3.4(a)与非标准单交叉口 3.4(b)上都进行了测试。标准交叉口所有进口道都是同构的，具体渠化形式如图 3.4 车道尽头的白色指示线所示。非标准交叉口采用一个“T”字型的交叉口，东向没有进口车道和出口车道。该路口上车辆的流动特性与标准交叉口中的车流特性大不相同。由各种交叉口组合起来，形成的各异的车辆流动特性也是交通环境非平稳性来源之一。标准交叉口对应的车流数据由 SUMO 提供的 API，根据其饱和密度，直接定量生成。非标准交叉口对应的车流数据由视频采集设备收集以后转换而成。

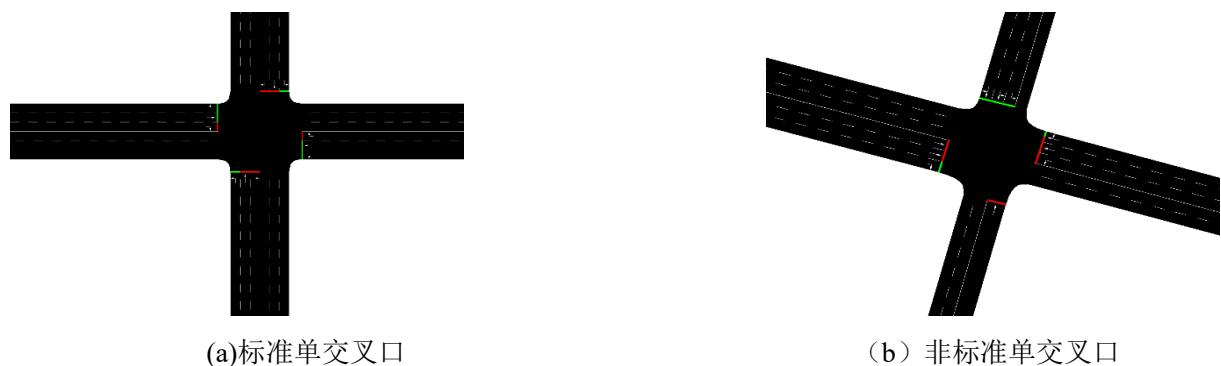


图 3.4 距离度量中使用的单交叉口

### 3.4.2 模型设置

在交通控制问题中应用强化学习的前提是相应的状态空间  $S$ ，动作空间  $A$  和奖励函数  $R$  的定义。

**动作空间  $A$  定义：**策略  $\pi_\theta(a_t | s_t)$  通过控制  $S(l, m)$  的值控制交叉口。通常来说，一个预定义的动作集合  $S_p$  会被给出，包括各种非冲突相位的组合。本章的动作被定义为从集合  $S_p$  中选择一个相位，然后设置各个连接器的  $S(l, m)$  值。此外，为了满足行人过街的需求，将最小动作持续时间设置为 15s，之后接一个 5s 的黄灯清空交叉口内的车辆。

**奖励  $R$  定义：**策略  $\pi_\theta(a_t | s_t)$  的目标是基于状态选择一个动作以最小化式(3.1)。但通常不直接使用该式作为奖励值，因为它太稀疏了，整个仿真结束以后才会获得一个标量值作为奖励，因此优化作用有限。作为替代，奖励函数被定义为累计等待时间的差值，表示为式(3.11)。

$$r(s_t, a_t) = w_t - w_{t+T} \quad (3.11)$$

$w_t$  表示在  $t$  时刻车道  $L_i$  上所有车辆平均等待时间。

状态设计是本章的重点，因此本节首先统计了在以往基于强化学习的信号控制系统中状态是怎么设计的，如表 3.1 所示。可以看到部分工作也使用了多种状态的组合。但实际上，这些工作的组合逻辑是随机的，少有理论对观测形式究竟该如何组合为状态表示进行研究。此处解释各个观测形式的实际物理意义。

- 等待车辆数目：车道上所有停止车辆的数目。用  $x_w(l)$  表示。
- 运行车辆数目：车道上所有运行车辆的数目。用  $x_r(l)$  表示。
- 车辆总数目：车道上所有车辆的数目。用  $x_t(l)$  表示。
- 信号灯相位：当前信号灯相位  $p$ 。
- 平均速度：进口车道上车辆的平均速度。用  $s(l) = \sum_{i \in [x_r(l)]} v(i) / x_r(l)$  表示。 $v(i)$  表示车辆  $i$  的速度。
- 空间占有率：车辆在车道上的空间占比。用  $x(l)/x_{max}(l)$  表示， $x_{max}(l)$  是车道  $l$  可能承载的最大车辆数目。
- 压力差：基于相位的特征，表示为  $P_i$ 。压力可以理解为车道的排队车辆数目，那么一个交叉口内的排队车辆数目表示为：

$$P_i = \left| \sum_{(l,m) \in i} (w(l,m)) \right| \quad (3.12)$$

压力差则是不同相位除了所对应的放行连接器的压力差。表示为：

$$w(l,m) = \frac{x(l)}{x_{max}(l)} - \frac{x(m)}{x_{max}(m)} \quad (3.13)$$

- 平均等待时间：排队车辆的平均等待时间，用  $w(l) = \sum_{i \in [x_s(l)]} w(i) / x_r(l)$  计算。 $w(i)$  表示车辆  $i$  的等待时间。

表 3.1 交通观测统计

| 观测的形式  | 文章                     |
|--------|------------------------|
| 平均等待时间 | [85]                   |
| 运行车辆数目 | [42][85]               |
| 等待车辆数目 | [40][72][77][86-87]    |
| 车辆数目   | [29-34][88-93]         |
| 累计延误   | [29]                   |
| 平均速度   | [72][88-93]            |
| 信号灯相位  | [30-34][72][77][86-87] |
| 排队长度   | [34][85][89][95-96]    |
| 压力     | [42][94-95]            |
| 空间占有率  | [90]                   |

从表 3.1 中可以观察到，在常见的强化学习模型中，平均等待时间、累计延误以及空间占有率三者使用的频率最小。空间占有率需要得到整个车道的车辆信息，实际中获取难度较大；累计延误是一个相对稀疏的指标，需要在车辆驶出路网以后才可以准确得到。更常用的指标有车辆数目，平均速度以及信号灯相位这三类。信号灯相位可以通过上一个决策时间点时智能体给信号灯下发的相位序号直接获得，平均速度与车辆数目同样是一个相对稀疏的指标，在实际交通环境中有一定的获取难度。但在仿真环境中，这些指标都可以通过接口较为直接的获得。有不少文献在不同种类的观测形式中交叉出现，这说明这些文章采用了多种观测形式进行组合。

## 3.5 实验结果

### 3.5.1 异构与同构的单交叉口的瓦氏距离度量

图 3.5 是标准交叉口的各个观测形式与锚（等待车辆数目）之间的瓦氏距离度量。当前信号灯相位这一观测形式并不参与计算，因为该参数往往是由智能体给出的，且往往由一个标量表示，没有计算距离度量的意义。在图 3.5 中，横轴表示距离度量模块的训练轮次。在更新过程中，采用随机的策略，使得模型可以不断探索到未知的状态，保证距离度量模块的鲁棒性。在训练快结束时，从图 3.5 可以明显发现，除了运行车辆数目外，其余形式的观测与锚点观测之间的距离都极为接近。因此可以认为，在标准交叉口中，平均等待时间与等待车辆数目这两种观测形式之间存在部分信息是完全不同的，也即  $w(l)$  和  $x_w(l)$  组合起来可以获得更大的信息量。

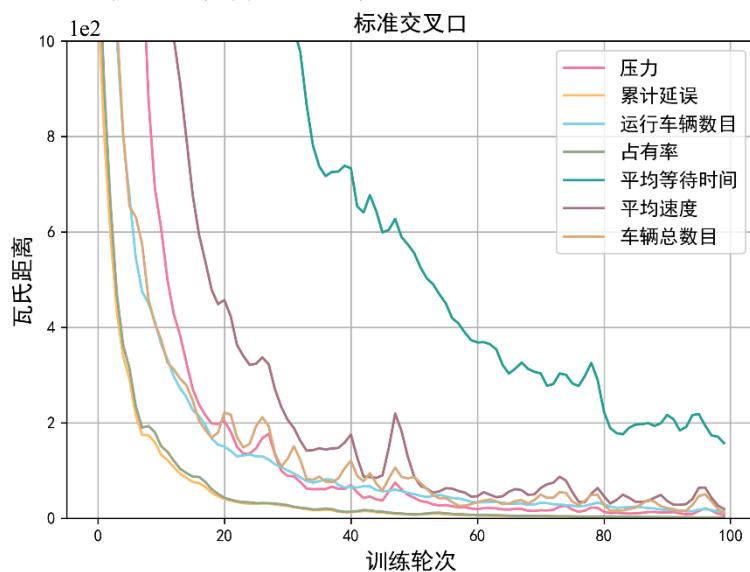


图 3.5 标准交叉口各个观测形式的瓦氏距离度量

城市交通中非标准形式的交叉口也占据了很大一部分，因此采用非标准的交叉口同样验证了距离度量模型的效果。补充了异构交叉口的距离度量结果如图 3.6。

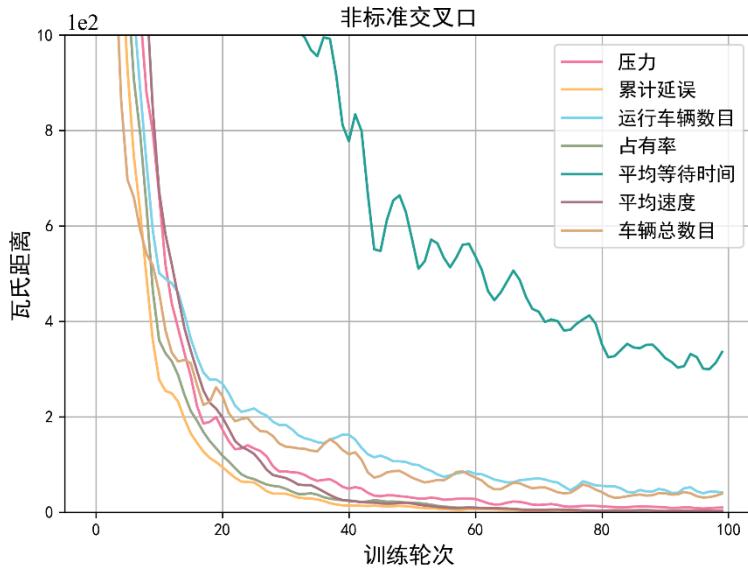


图 3.6 非标准交叉口各个观测形式的瓦氏距离度量

### 3.5.2 各种观测形式在强化学习模型中的表现

在得到最佳观测形式以后，验证该观测形式的实际优化效果。采用的算法是第二章介绍的 PPO 算法，实验中使用的路网是本章实验数据中给出的标准形式路网。车流为 2118veh/h。动作和奖励设计参考本章模型设置小节。采用旅行时间和交叉口吞吐量作为客观评价指标，总吞吐量的计算方式在问题描述中已经申明，而旅行时间的计算方式则是在路网中所有车辆的旅行时间的平均。这两个指标可以较好的刻画一个模型优化交通控制的能力<sup>[31][35][42]</sup>。对于旅行时间来说，越小越好；而对于吞吐量来说，越大越好。除此之外，累计奖励以及平均速度也同时被考虑。这二者都是越大越好。

**量纲说明：**旅行时间的单位是秒，吞吐量的单位是辆，速度的单位是米每秒。

**参数设置：**最小绿灯时间为 15s，以满足行人通行需要；黄灯时间设置为 5s，用以清空交叉口内的车辆，防止交叉口内堆积的车辆产生冲突；一幕仿真的持续时间设置为 3600s，也即一小时。学习率设置为 0.00025，同时随着训练线性衰减，衰减系数为 0.95；隐藏层大小为 128；批处理大小（batchsize）为 32；每幕仿真结束后智能体从经验池中随机采样，然后更新 8 次；同时采取梯度裁剪形式的 PPO 算法，梯度每次更新幅度限制在 0.02 以内。

除以上设置之外，每组都采取了两种形式的观测进行组合，因此需要进行特征融合。融合的方式采取全连接层直接进行融合。融合特征记为  $f(f(o^1) + f(o^2))$ ， $o^1$  和  $o^2$  分别表示两种不同形式的观测。

旅行时间和交叉口吞吐车辆的能力通常是最为关心的客观评价指标。具体的实验结果如 3.7(a)与 3.7(b)所示。随着训练轮次的递增，每种组合的信号控制能力都在提升。表现为交叉口吞吐车辆数目越来越多，同时旅行时间开始减少。随着训练推移到最后，

有两组数据的表现明显超过其余组别。一组是等待车辆数目  $w(l)$  与车辆总数  $x_t(l)$  的组合；另一组则是上一小节计算得到的平均等待时间  $x_w(l)$  与等待车辆数目  $w(l)$  的组合。这证明了提出的距离计算模型是有效的。其余组别与这两组之间的差距也可以证明，模型的性能并非完全来源于锚点观测  $w(l)$ ，在如此简单的特征融合形式下，异构观测的信息之间的融合确实给模型带来了正向收益。证明了计算得到的状态表示的正确性。

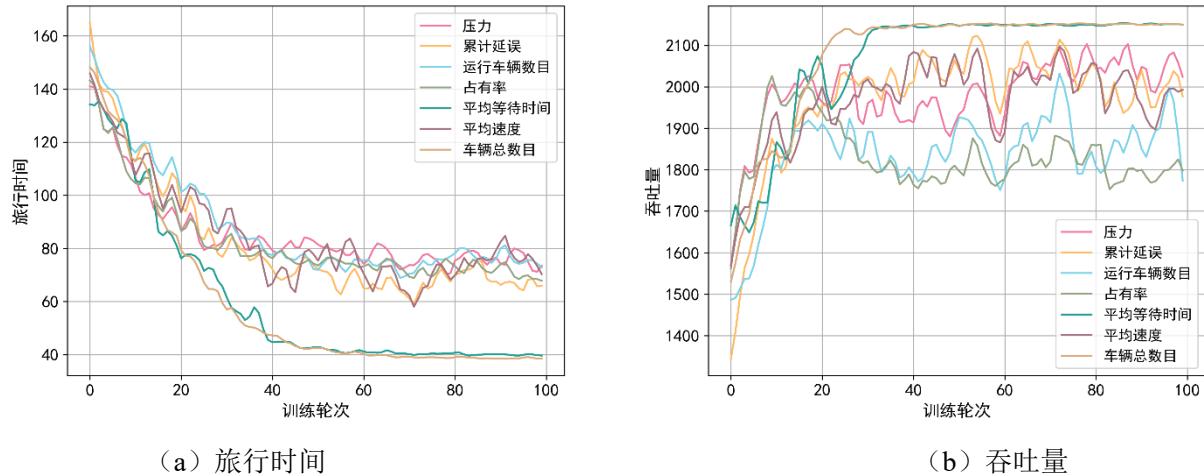


图 3.7 不同状态表示交通参数优化能力

除了吞吐量和旅行时间以外，还给出累计奖励的曲线如图 3.8。

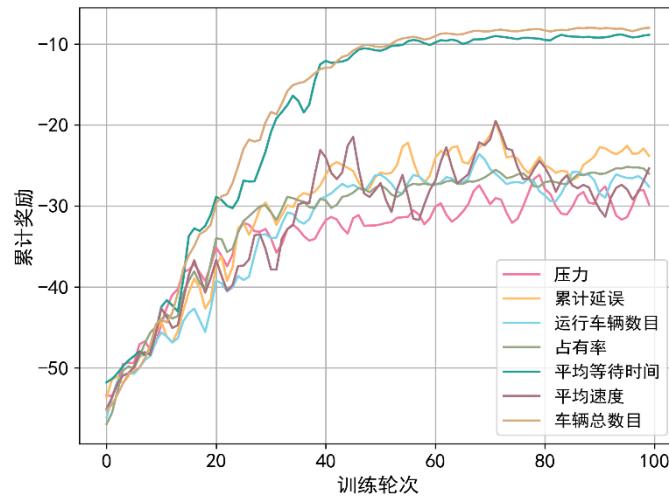


图 3.8 累计奖励

本章采用的奖励计算方式如公式(3.11)所示，基于等待时间的差值。将每个决策时间点的奖励相加，最终得到的总奖励值为不大于 0 的标量。图 3.8 展示了所有组别的奖励曲线，均呈上升趋势。其中，计算得到的最优组别奖励曲线上升速度明显快于其它组别，且训练结束时累计奖励值最大，符合模型的预期。这验证了计算得到的状态表示的有效性。此外，从收敛的平滑性来看，计算得到的组合形式状态表示也优于其它形式，进一步证明了其能够应对复杂环境的非平稳性。

除以上收敛曲线之外，给出平均速度随训练轮次的变化曲线如图 3.9。

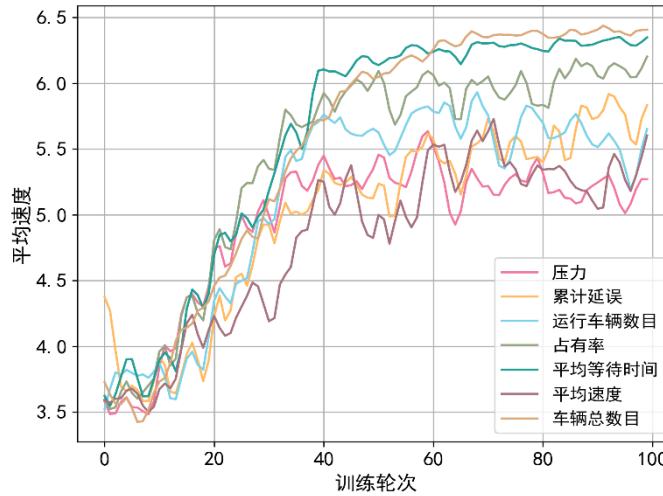


图 3.9 平均速度

平均速度与吞吐量以及旅行时间会呈现对应变化的关系，这两个参数优化的同时，平均速度也会被同步优化。总体趋势仍然符合模型的预期。

### 3.5.3 路网层面各个观测形式之间的瓦氏距离度量

本章的研究目的是在路网层面解决信号控制问题，在城市交通网络中，交叉口并非孤立存在，而是通过道路紧密相连，形成一个动态的、相互依赖的系统。因此希望该距离度量在路网层面同样具有鲁棒性。所以本小节在路网层面同样测试了该距离度量模块。首先在一个 $4 \times 4$  的只具有标准交叉口的路网进行了测试。测试使用的路网结构如图 3.10。

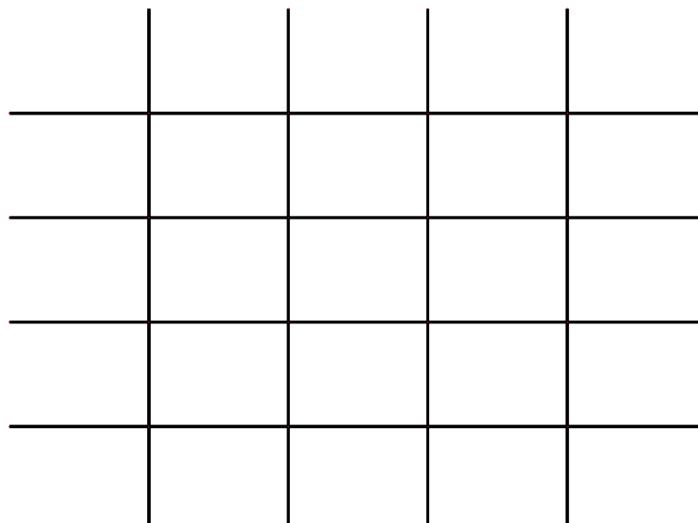


图 3.10 标准交叉口所在路网

该路网中每个交叉口的结构都与图 3.4(a)完全一致。在路网层面，各个交叉口除了仿真软件原先指定生成的车辆之外，还因为各个交叉口之间有交互，不同交叉口之间的车辆会依概率流动。从宏观方面，不同交叉口的统计特性会出现一定的区别，给出在路网层面测试观测分布距离模型的结果如图 3.11。

图 3.11 的每条曲线与 3.5 的图例相对应。可以明显观察到，每个交叉口测得的距离分布与标准单交叉口的结果一致，且均表现出明显的距离分布特征。使用神经网络对二者进行瓦氏距离计算时，难以找到一组合适的权重，使得二者间在状态转移概率矩阵固定的情况下，存在形式上转化的可能性。即平均等待时间与等待车辆数目之间的信息存在着明显的差异性。此外，观察到交叉口 1\_4 中，各个观测见分布距离计算过程中存在着明显的波动性，这是由于仿真软件在设置该车辆的起始位置时，驶入和驶出交叉口的车流之间的交互比较复杂，引起了状态转移概率矩阵的复杂性，使得该交叉口的瓦氏距离计算比较复杂。在交叉口 4\_4 上也有类似的情况发生，但是波动程度没有交叉口 1\_4 明显。在其余交叉口上，瓦氏距离收敛曲线总体还是较为平稳。

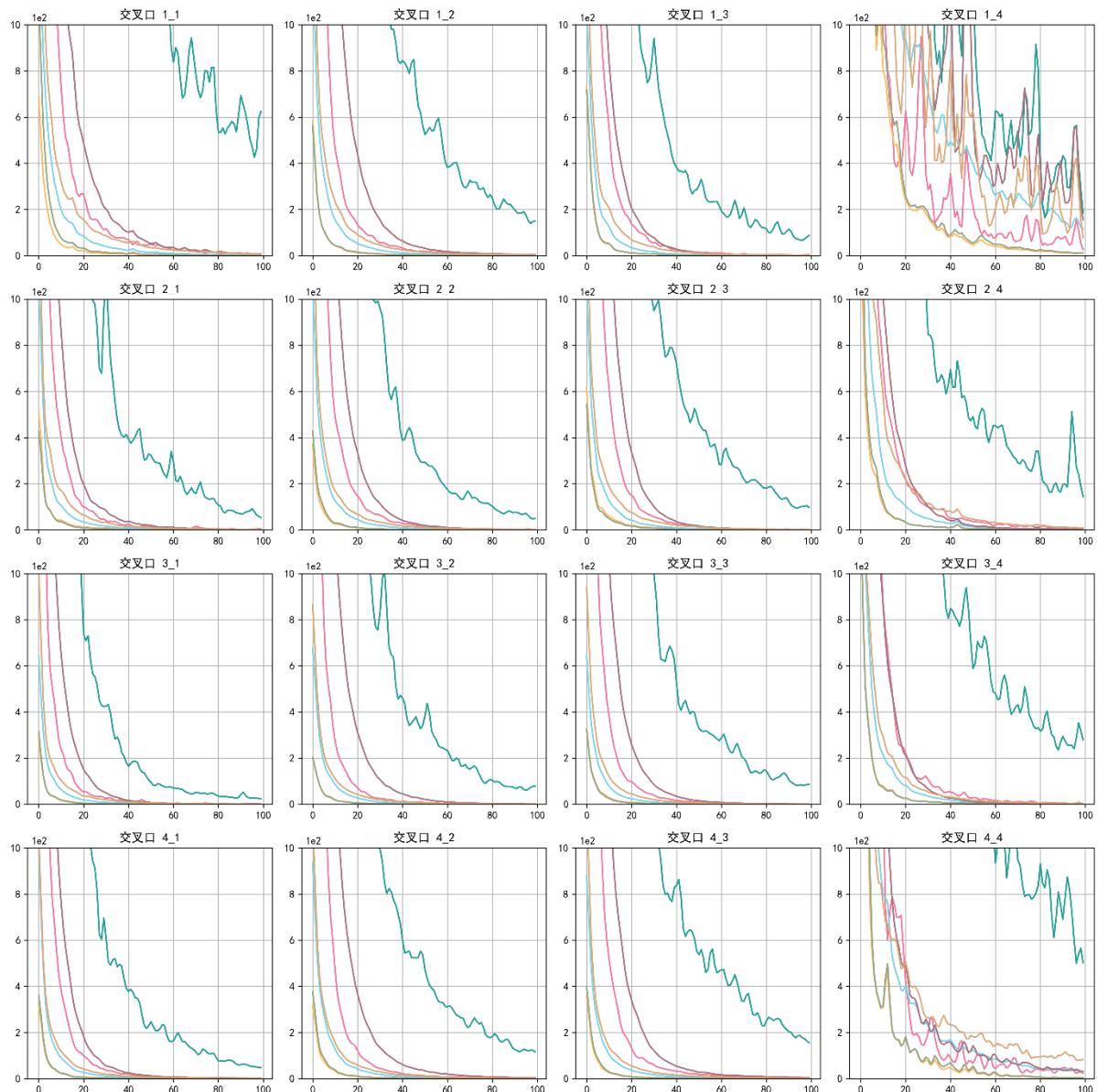


图 3.11 标准交叉口所在路网各个观测形式的瓦氏距离度量

除了只有标准交叉口的路网上测试所提出的距离度量模块，在具有多个非标准交叉口的路网中同样测试了所提出的模块，以求真实的反映城市交通的复杂情况。路网如图 3.12 所示。



图 3.12 非标准交叉口所在的路网

从中随机抽取了 9 个具有信号控制的交叉口进行测试，测试的结果如图 3.13。

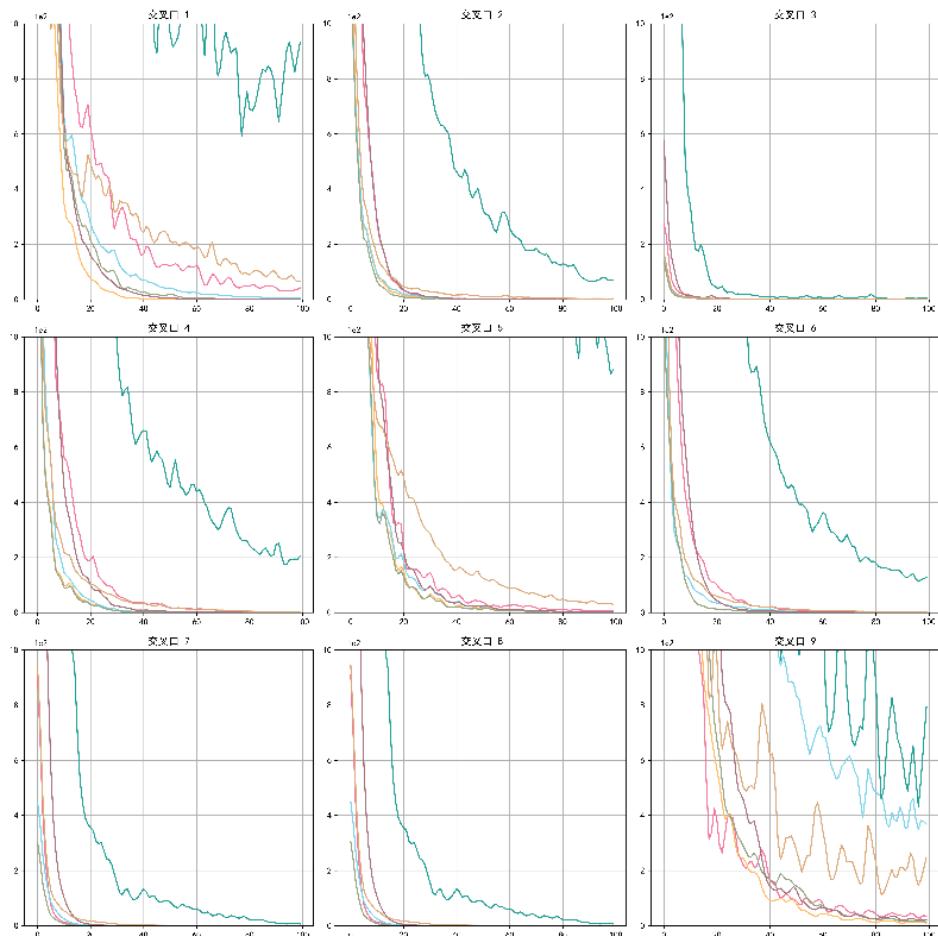


图 3.13 非标准交叉口所在路网各个观测形式的瓦氏距离度量

显而易见的，仍然是平均等待时间这一观测形式与等待车辆数目这一观测形式之间存在明显距离。同时在交叉口 1 与交叉口 9 呈现出与标准交叉口类似的波动性。在交叉口 9 中波动性更加明显，这也是由于车辆的路由引起的异常。

在标准的路网与复杂路网下，实验结果证明了本章所提出的距离度量模块的鲁棒性。这意味着无论是在单交叉口或者路网层面的信号控制问题中，使用平均等待时间与等待车辆数目的组合可以涵括大部分信息，且仅用两种观测形式组成状态标识也不会明显引起状态空间爆炸。而在统计的前人的工作中，少有文章使用这二者的组合。

### 3.6 本章小结

在本章中，使用神经网络减小了不同表示的观测之间的表示误差。在减小需要探索的状态空间的大小的同时保证了最大化信息量。将只能用于同构观测距离度量的互模拟度量方法扩展到了异构观测的形式上，使得计算异构观测之间的分布距离成为可能。除此之外，在单交叉口上的结果验证了距离度量模块的有效性，并且得到了最优的观测形式组合作为状态。并且利用 PPO 算法验证了各种组合在单交叉口上的表现。虽然从表现也能反推出最优组合，但在有些场景中，每个组合进行测试可能是极为消耗算力的。而提出的模型为直接求解最优观测形式组合打下了基础。更进一步的，在路网层面同样验证了所提出模块的鲁棒性。即使是在路网层面，所得到的结果仍然是稳定的。本章对基于强化学习的信号控制问题中的状态设计问题给出了一种有效的解决思路。

## 4 基于有偏好状态编码的区域交通信号控制算法

在前一章节中，提出了一种基于强化学习的信号控制问题中更为紧凑的观测形式。在模型的实际更新过程中，期望通过这种观测形式能够获得一个探索空间较小的状态空间。在本章中将这种观测形式应用于路网规模的强化学习算法训练中。此外，还提出了一种有偏好的强化学习编码方式，旨在压缩非必要的探索状态空间，从而减少求解最优路径所需的计算资源。同时，为了缓解去中心化框架中的部分可观测性带来的影响，提出了一个时空依赖关系捕捉模型，从时空和空间两方面链式扩展本地智能体的可观测范围。并且，为了获得更加准确的用于预测和评估的特征，提出一个基于双线性池化的特征融合方案。最后，将所提出的算法与近年来一些基于强化学习的信号控制算法进行了对比，验证了所提出模型的有效性。在本章，对于表示车辆的整数  $X$ ，用  $[X]$  表示对应车辆的集合。

### 4.1 问题描述

存在一个路网，路网中具有信控条件的交叉口合集表示路网为  $\mathcal{N} = \{n_1, n_2, \dots, n_k\}$ 。 $k$  表示交叉口的个数，交叉口的智能体的优化任务依然可以用公式(3.2)表示。使用  $C(l, m)$  表示可能通过  $c(l, m)$  离开交叉口的车辆数目， $d(l, m)$  表示从交叉口外进入当前交叉口的车辆数目。对于单交叉口的仿真， $[C(l, m)]$  不会对后续仿真产生任何影响，而  $[d(l, m)]$  的产生来源于仿真文件的预设。因此单交叉口的环境可以认为是平稳的，强化学习范式在这种平稳环境中可以较好的求解一个最优策略。对于多交叉口的仿真，驶入交叉口的车辆属于有两种来源，仿真软件预先设置的车流，以及相邻交叉口的驶出车辆， $[C(l, m)]$  会对后续环境产生影响。所有交叉口都会与相邻交叉口产生博弈。对于多交叉口的信号控制问题，希望最大化整个路网的交通通行能力，因此该博弈是一个合作博弈。用  $\Pi$  表示路网中  $m$  个智能体的策略集合， $j_i(\pi)$  是智能体  $i$  在策略组合  $\pi$  下的奖励函数，希望得到一个策略  $\pi^*$ ，使得式(4.1)成立。

$$j_i(\pi_i^*, \pi_{-i}^*) \geq j_i(\pi_i, \pi_{-i}^*) \quad \forall i \in N, \forall \pi_i \in \Pi \quad (4.1)$$

$\pi_{-i}^*$  表示除了  $i$  之外所有智能体的策略组合。即是得到组合策略  $\pi^*$  后，修改任何一个智能体的策略都会使得全局优化目标变差。

由于合作博弈的产生，传统的单智能体直接扩展到路网规模的范式不再适用。各个智能体之间除了环境中车辆自身的流动之外，没有任何其余交互的信息，此时所求解的策略是一个纳什均衡<sup>[97]</sup>(Nash Equilibrium)。而求解一个纳什均衡是一个 NP-完全问题，这意味着在多项式时间内找到一个精确的纳什均衡是极其复杂的。因此，即便是在本章

的去中心化强化学习范式中，也允许相邻智能体之间可以进行交流。此时纳什均衡退化为相关均衡，求解相关均衡的复杂度更低。

## 4.2 主要贡献

对于复杂交通环境来说，状态空间爆炸是一个迫切需要解决的问题。给定一个初始状态  $s_0$ ，强化学习通过找到最优轨迹来实现最大化奖励的目标。显然，状态空间越小，找到最佳轨迹所需的计算资源消耗就越少。然而，交通环境中复杂的动态特性会导致状态空间急剧膨胀，进而误导智能体在训练过程中探索最优路径。

Efficient-CoLight<sup>[42]</sup>的研究表明，良好的状态表示可以有效缓解状态空间爆炸问题。进一步的研究<sup>[98]</sup>指出，在良好的观测表示下，求解复杂度可以实现指数级下降。这意味着，良好的观测表示本身就能指向一个更小的状态空间。在第三章中，已经给出了可以得到最大信息的观测组合，避免了引入冗余的信息。对于小规模路网，这种观测形式已经足够。但是随着路网规模的不断提升，状态空间指数级增长，依然面临状态空间爆炸的问题。因此，（1）本章提出一种有偏好的状态空间编码方法，通过专家模型识别模型更新时更有价值的部分，压缩探索价值较低的部分，从而有针对性地压缩整个状态空间，加速强化学习模型的收敛。

除此之外，为了满足实时性需求并保护敏感的交通信息，本章采用了去中心化的多智能体强化学习框架。在最新的信号控制多智能体强化学习算法中<sup>[99-101]</sup>，智能体通常采用同步更新的方式。在更新过程中，对于任何一个智能体而言，相邻智能体的策略参数是不可知的。更进一步的，已知在多路口的情况下，在多路口场景中，对于每个只能观测到本地信息的智能体而言，整个环境表现出非平稳性。这种非平稳性会导致智能体学习到的策略失效。针对这个问题，（2）从时间和空间两个维度扩展了可得信息的范围。在空间上，本地智能体允许获得相邻智能体的观测信息，从而增强对局部环境的感知能力。在时间上，将经验池中之前的决策过程中产生的部分轨迹信息进行堆叠，刻画交叉口交通信息的时序特征。

获取时序特征之后，组成状态的观测在形式上还是独立的，之前的一些工作采用直接拼接或者全连接层的方式融合信息，但是这种简单的方式不能得到不同形式观测间的隐藏关系。针对这个问题，（3）引入一个基于双线性池化<sup>[102]</sup>的特征融合模块，得到不同形式观测间的互信息，用以融合异构观测的特征信息，进一步提升模型的表现。

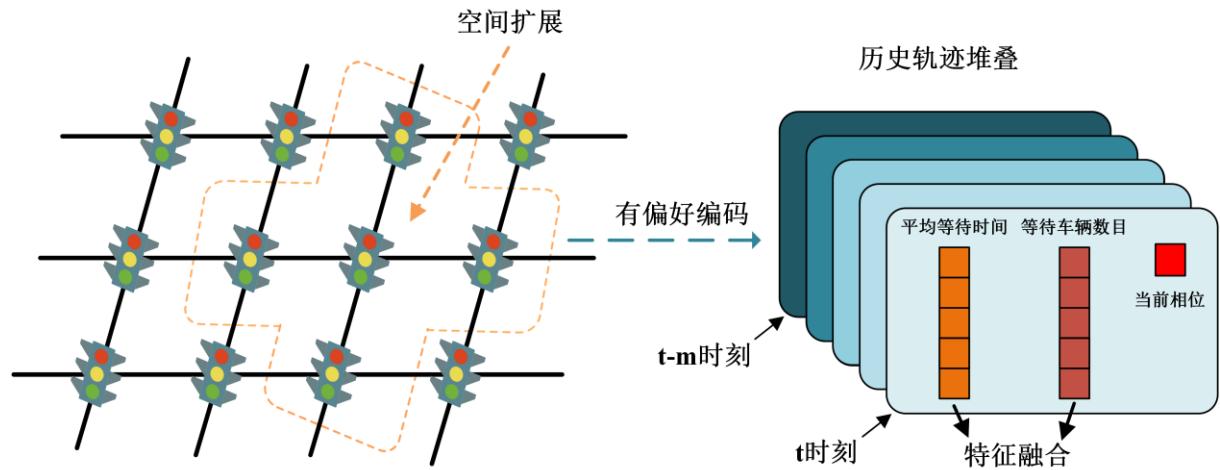


图4.1 技术要点说明

### 4.3 强化学习模型设置

在详细介绍本章所提出的算法之前，首先给出本章中所使用的强化学习模型的一些基础的设置，包括状态，奖励和动作等。

#### 4.3.1 状态设置

第三章已经详细描述了强化学习中状态的设置，本章遵循第三章最后计算得到的最优状态表示的结果。使用平均等待时间  $x_w(l)$  与等待车辆数目  $w(l)$  的组合作为状态表示。已经在第三章验证了在交叉口上该表示的优越性，不再赘述。

#### 4.3.2 动作设置

动作是环境真实世界运行的一种逻辑。智能体通过调用这种逻辑达成优化的目的。更具体的，交通环境中智能体通过对相位做出调整来与环境进行交互。改变下一相位持续时间是一种常见的动作形式，例如分层强化学习模型<sup>[103]</sup>，它假设存在一个固定相位方案，交互过程中相位顺序不变。动作集合定义为一个  $\mathbb{A} = \{15, 20, 25, 30, 35, 40, 45, 50\}$  作为相位持续时间。意味着在时间  $t$  选择了一个动作后，该智能体在时间段  $a$  内不会改变信号灯的相位。该动作在另外一个工作中<sup>[104]</sup>被改善为一种更加紧凑的形式，动作空间被整合为两个动作，保持或者改变当前相位。以此来达到切换信号灯的目的，并且相较于选择相位时间可以获得更大的优化空间。OAM<sup>[90]</sup>则是为了获得最优控制效果选择将动作分解到车道级别，参考相位的定义，每条车道有两种状态，对应两种动作：保持车道放行以及车道将要放行。在实际更新过程中先获得车道级别的动作，并考虑相位之间的冲突，再将车道级别的动作整合起来，获得完整的动作，统一下发。此外，一些工作<sup>[105-107]</sup>将动作形式改善的更加灵活，将整个相位方案的各个相位设置为一个动作，得到

动作集合  $\mathbb{A}$ 。智能体与环境交互时从中选取一个动作进行执行，相位持续时间满足行人过街需求即可，这也是本章所采取的动作。

### 4.3.3 奖励设置

奖励函数是强化学习模型的重点。在特殊场景中，比如减少碳排放、急救车等场景可能会存在特殊的奖励设计<sup>[108]</sup>。除特殊任务特殊场景外，信号控制中的优化问题设置的奖励形式大多比较类似。交叉口中的参数大多遵循可频繁测量的原则<sup>[109]</sup>。如排队长度、车辆数目、平均速度、压力差等。以及第二章提及的  $r(s_t, a_t) = w_t - w_{t+T}$ ，以累计等待时间差值作为奖励，本章也沿用这种奖励形式。

### 4.3.4 超参数设置

本章使用 PPO 算法并设置其余参数如下：奖励衰减率  $\theta$ 、批处理大小以及学习率被分别设置为 0.9, 256 以及 1e-3。每一幕仿真结束，模型更新 8 次。同时使学习率随着更新过程阶梯式衰减，学习率衰减稀疏设置为 0.99。梯度裁剪的参数设置为 0.2，也即新旧策略的比率将被设置为 [0.8, 1.2]，减缓策略出现震荡。可见历史信息步长设置为 8。

## 4.4 有偏好状态编码

对于一个单交叉口，假设其状态空间为  $n$ ，动作空间为  $m$ ，则其需要探索的空间大小为  $n^m$ 。当交叉口与相邻交叉口连接，构成路网以后，假设其交叉口都是同构的，为了满足全局最优的目的，需要探索的空间变为  $2n^{2m}$ ，呈现指数级增长的趋势。但其实对于任意一个初始状态  $s_0$ ，最优轨迹对应的路径在整个需要探索的空间中只占用很小的一部分。早先有研究尝试从加入先验知识的途径减少需要探索的状态空间<sup>[32]</sup>，它直接使用专家模型在待优化的路网上产生轨迹，并且使用产生的轨迹对演员模型进行预训练达到优化训练速度的目的。但一方面，这种方式依赖于专家模型的交互，这种交互也是消耗计算资源的；另一方面，对于复杂的交通环境，不可能每种路况都有对应的专家模型。因此，本小节从交通环境本身的特性出发，通过本身最优策略对应的最优轨迹的信息得到有偏好的状态空间部分，并据此得到了有偏好状态编码。

### 4.4.1 PPO 的策略更新

对于 PPO，它更新策略网络的方式是通过新旧策略之间的差异性，其损失函数可以用式(4.2)描述。

$$\max_{\theta} \mathbb{E}_{(s_t, a_t) \sim \pi_{\theta_{\text{old}}}} \left\{ \min \left[ r_t(\theta) \hat{A}(s_t, a_t), \text{clip} \left( r_t(\theta), 1 - \epsilon, 1 + \epsilon \right) \cdot \hat{A}(s_t, a_t) \right] \right\} \quad (4.2)$$

$$r_t(\theta) = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} \quad (4.3)$$

$$\hat{A}_t^{\text{GAE}(\gamma, \lambda)} = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l}^v \quad (4.4)$$

其中， $r_t(\theta)$  表示新策略  $\pi_\theta(a_t | s_t)$  与旧策略  $\pi_{\theta_{\text{old}}}(a_t | s_t)$  之间的差异性。 $\hat{A}_t$  用来计算在状态  $s_t$  时采取  $a_t$  的价值。通过计算使得动作  $a_t$  对应价值最大的时候，奖励函数也被最大化。

#### 4.4.2 编码策略

在视觉识别领域，针对不同的物种可能有不同的先验知识，这些知识可以帮助更准确地识别和分类目标。例如，假设希望识别鸟类，毫无疑问，天空中出现鸟类的概率远大于水中。这种先验知识可以作为模型训练和推理过程中的重要参考，从而提高识别的准确性和效率。回到信号控制问题，考虑到一个极其巨大的状态空间，在有限计算资源中访问到所有状态几乎是不可能的，这会导致模型求得的是次优解。希望从专家模型的轨迹中获得关于状态空间中更有价值的部分的特点。此处选择四种常见的传统信号控制模型：

1. 最大压力差：旨在通过最大化交通网络中的“压力”（即车辆等待时间与通行时间的差值）来提高交通效率。核心思想每次激活压力最大的相位，以最小化车辆在交叉口的等待时间，从而减少交通拥堵。
2. 固定式配时：从历史数据中，利用专家知识得到一个固定式相位方案。根据预先设定的信号配时方案来控制交通信号灯的切换。信号灯的相位和时长是固定的，不随实时交通状况的变化而调整。
3. 自组织交通信号控制：旨在通过实时监测交通流量和车辆等待时间，动态调整信号灯的相位和时长，以提高交通效率。也会计算每个车道的“压力”值，表示该车道上车辆的等待时间与通行时间的差值。压力较大的车道会获得更长的绿灯时间，以减少等待时间。
4. 随机算法：同时从历史数据中利用专家知识得到一个固定式相位方案。需要切换相位时随机选择一个相位执行。

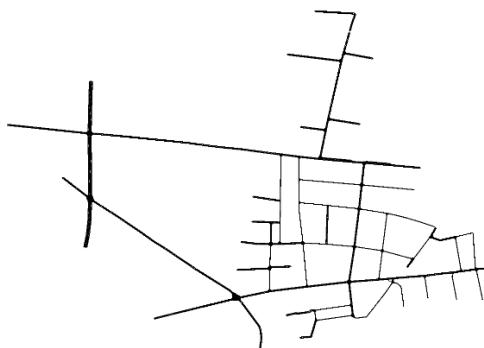


图4.2 状态分布测试路网

在如图 4.2 所示的路网中进行测试，该路网共有八个具有信号控制条件的路口。对绿灯时间有要求的算法，设置绿灯时间为 15s，黄灯时间为 5s。统计信号控制路口的交通参数（旅行时间和吞吐量）表现如表 4.1。

表 4.1 不同控制算法的交通参数表现

| 信控算法      | 旅行时间   | 吞吐量  |
|-----------|--------|------|
| 最大压力差     | 362.50 | 2728 |
| 固定式配时     | 531.74 | 2425 |
| 自组织交通信号控制 | 541.50 | 2430 |
| 随机算法      | 580.78 | 2418 |

可以明显看出，在这个路网中，最大压力差算法的表现明显优于其余算法。四种信号控制算法的效果从优到劣依次是：最大压力差>固定式配时>自组织交通信号控制>随机算法。为了得到最优模型关于状态的偏好，同时统计了优化过程中状态的分布。具体而言，在每个决策的时刻，统计得到整个路网信号交叉口内等待车辆数目  $w(l)$ ，这是一个标量。传统信号控制算法不需要进行多轮，因此在一轮仿真结束以后，可以得到模型在整轮中访问到的状态，并绘制其分布图。四种传统算法得到的状态分布如图 4.3。

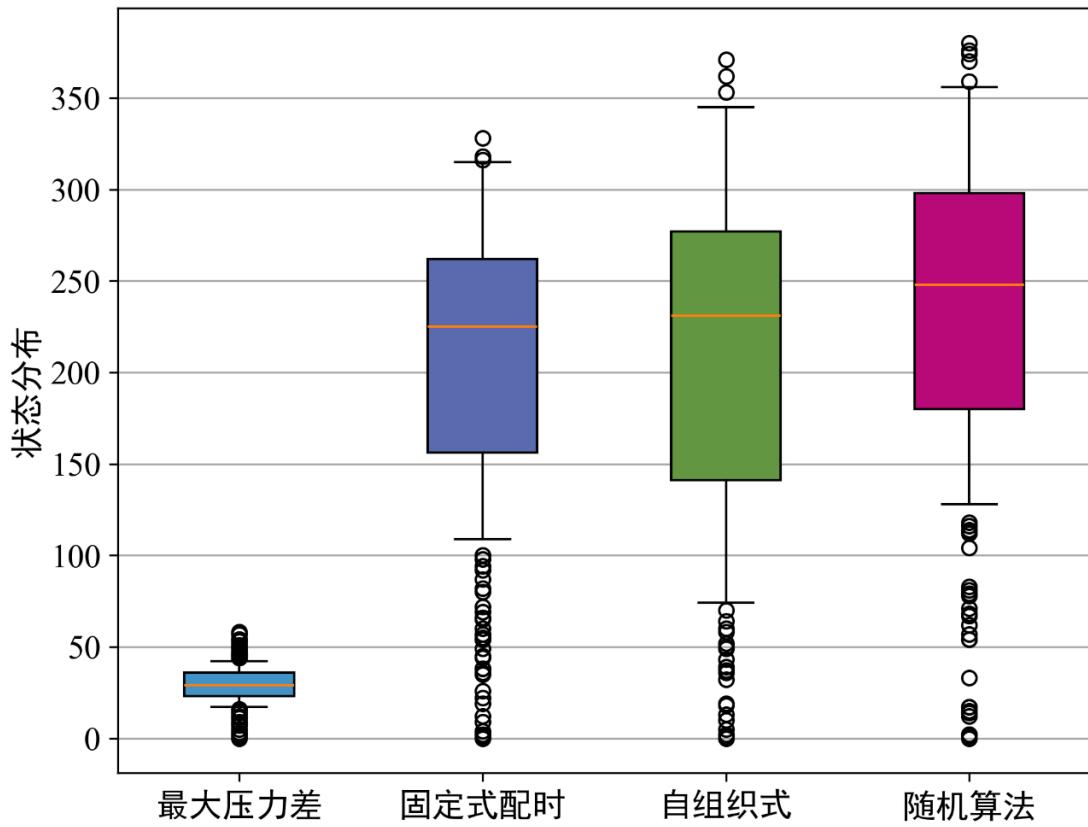


图4.3 不同算法之间的状态分布

对比四种算法的状态分布，表现最好的压力差算法其状态分布区间明显更加紧凑，而且其状态值明显聚集在值较小的区域中。其余三种算法的状态分布与其客观评价指标的参数与呈现明显的相关性。表现越好的算法，其状态分布越紧凑，并且区间会聚集在值较小的区域。因此得出一个客观事实：交通控制模型对状态值较小的区域有明显的偏好性。实际上，这也符合交通情况的需求，对于一个最优控制模型产生的效果，因其不会产生拥堵，所以其各项交通流参数都会往最优处聚集。例如速度会呈现增大的趋势，排队车辆数目会呈现减少的趋势。利用这种趋势，可知在整个状态空间中，排队车辆数目多的区域其信息的重要性是远远小于排队车辆数目少的区域。

根据以上的实验和讨论，希望获得一个函数，在智能体获得状态前对状态进行预处理，以排队车辆数目为例，使得整个状态空间中排队车辆数目多的区域的信息被压缩，排队车辆数目少的区域信息被尽可能的保留。对于其余形式的状态，也可以采取简单的方法得出状态分布，再根据状态分布得出具体的状态偏好。通过这种偏好，在模型的网络结构得到状态前对其进行预处理，就可以减小模型需要探索的状态空间的大小。

对于等待车辆数目  $w(l)$  满足编码需求的函数族可以用图 4.4 描述。

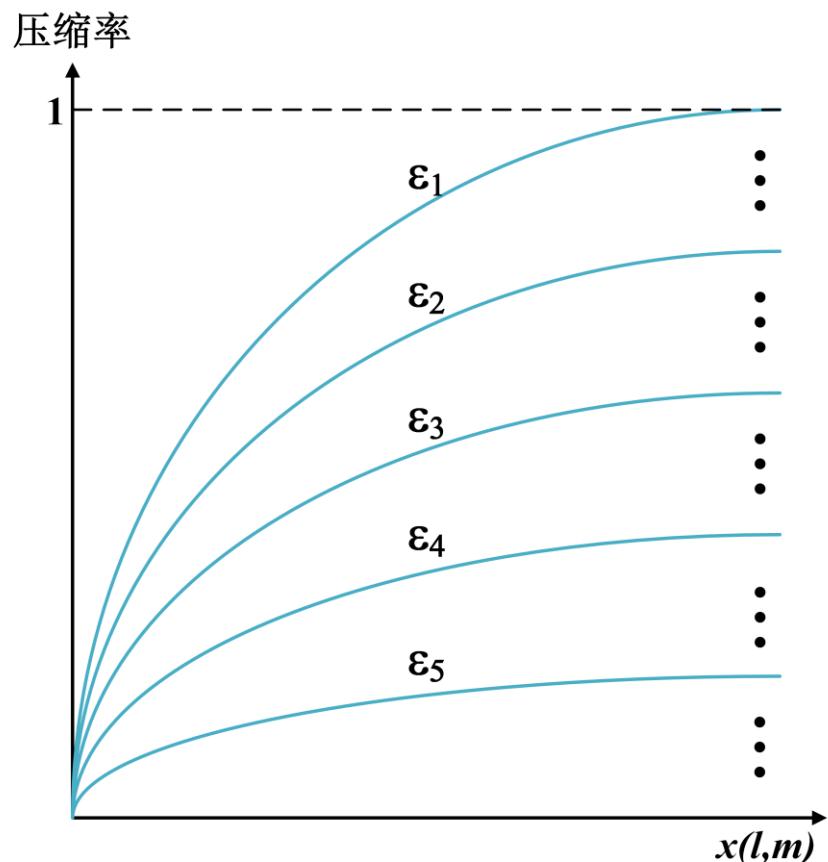


图 4.4 不同算法之间的状态分布

在图 4.3 中，横轴表示状态排队车辆数目，因此横轴较大的区域需要被压缩。纵轴表示压缩率，为 1 时表示信息被尽可能的压缩。横轴本身也可视为状态编码函数，表示 42

不压缩任何信息。图 4.3 中  $\varepsilon_1$  表示尽可能的压缩排队车辆数目较大的区域同时尽可能保留排队车辆数目较小的区域的信息。 $\varepsilon_5$  表示压缩时也尽可能保留排队车辆数目较大区域的信息。中间的压缩函数则希望在压缩空间与保留信息方面取得一个平衡。最终，采取对数函数的形式作为有偏好编码函数：

$$\mathcal{E}(s) = \lfloor (\log(x(l_1, m_1) + 1)) \rfloor, \dots, \lfloor (\log(x(l_n, m_n) + 1)) \rfloor \quad (4.5)$$

$l_1, l_2, \dots, l_n \in L_i$ ,  $m_1, m_2, \dots, m_n \in L_o$ 。对数函数的形式是统一的，避免了不同交叉口还要计算编码函数的资源消耗。

此时得到了  $w(l)$  的压缩表示。压缩状态空间可能导致信息损失，因此引入额外的时空信息来弥补信息损失。除此之外，在上一章，对于不同形式的观测采用简单的全连接度量了分布间的距离。对于同一个交叉口，不同形式的观测之间的潜在信息之间存在关联，在本章，使用一种新的特征融合模块度量这种关联。

## 4.5 时空模型

为了降低去中心化模型中部分可观测性对深度学习网络的影响，将长短期记忆网络（Long Short-Term Memory, LSTM）与评论家网络进行结合。交通流量数据通常具有明显的周期性和趋势性，本节希望通过 LSTM 捕捉到这些依赖关系。

### 4.5.1 时间依赖

LSTM 通过引入门控机制来控制信息的流动，从而解决传统长序列建模时容易出现的梯度爆炸和消失问题。以排队车辆数目  $w(t)$  为例，在 LSTM 开始时间序列建模之前，首先使用两层全连接层（Multilayer Perceptron, MLP）对输入进行处理。表示为式(4.6)。

$$x(l, m)'' = f_{c2} \left( f_{c1} (x(l, m)) \right) \quad (4.6)$$

$f_{c1}$  和  $f_{c2}$  是单层全连接层， $f_{c1}$  获得高维信息表示。PPO 算法的网络结构在代码实现大致可分为两种，演员和评论家共享网络结构，最后使用组合的损失函数；演员和评论家使用独立的网络结构，反向传播时使用自身的损失函数。共享参数可以减少模型的复杂度，使得模型可以更好捕捉数据的局部模式，提高模型的收敛速度。但参数共享可能使得模型求解的是局部最优解。当演员和评论家使用完全独立的网络时，模型的表达能力提升，与之对应的是模型复杂度上升。本章使用的是演员和评论家结构独立的 PPO 算法，但是为了减小模型复杂度，将  $f_{c2}$  的参数共享。

LSTM 主要的组件有遗忘门，输入们，单元状态和输出门。遗忘门判断从单元状态中丢弃哪些信息，计算公式如式(4.7)。

$$f_t = \sigma(W_f [x_t'', x_{t-1}^{hid}] + b_f) \quad (4.7)$$

输入门决定了哪些信息会被添加到单元状态，包含两部分，一部分决定更新的部分，另一部分生成新的候选值：

$$i_t = \sigma(W_i[x_t^{\text{hid}}, x_{t-1}^{\text{hid}}] + b_i) \quad (4.8)$$

$$\tilde{C}_t = \tanh(W_C[x_t^{\text{hid}}, x_{t-1}^{\text{hid}}] + b_C) \quad (4.9)$$

单元状态则通过遗忘门和输出门进行更新：

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (4.10)$$

输出门决定单元状态中哪些信息需要输出：

$$x^t = \sigma(W_o[O_t^{\text{hid}}, x_{t-1}^{\text{hid}}] + b_o) \quad (4.11)$$

上列公式中， $f_t$ 是遗忘门的输出，介于0和1之间； $\sigma$ 是Sigmoid激活函数； $W_f$ 是遗忘门的权重矩阵； $W_i$ 是输入门的权重矩阵； $x_{t-1}^{\text{hid}}$ 是前一时刻的隐藏状态； $x_t^{\text{hid}}$ 是神经网络的输入； $b_f$ 是遗忘门的偏置项； $C_t$ 是当前时刻的单元状态。

## 4.5.2 空间依赖

在获得时间依赖的表示 $x^t$ 以后，希望获得邻接交叉口之间的空间依赖。车辆在路网中不断的流动构成了相邻交叉口之间的空间交互，在此节使用图注意力机制刻画这种依赖关系。以往的强化学习算法往往使用整个路网计算图注意力，这将大大损失模型的泛化能力以及计算复杂度，同时也难以满足实时性的需要。在本小节，只有一阶相邻节点被考虑进当前节点的图注意力机制中。图注意力的更新过程可以用式(4.12)。

$$x^{st} = GAT(x^t, \{x_{adj}^t\}) \quad (4.12)$$

$x_{adj}^t$ 表示一阶邻接节点的信息集合。具体来说，整个更新过程可以被表示为：

$$e_{ij} = a([Wx^t \| Wx^t]), j \in \mathcal{N}_i \quad (4.13)$$

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(e_{ij}))}{\sum_{k \in \mathcal{N}_i} \exp(\text{LeakyReLU}(e_{ik}))} \quad (4.14)$$

$$x^{st} = \sigma \left( \sum_{j \in \mathcal{N}_i} \alpha_{ij} Wx^t \right) \quad (4.15)$$

完整更新过程后，得到去中心化框架中本地交叉口的时空融合特征 $x^{st}$ 。交通系统中，各个参数之间并非是完全独立的，因此需要提取出不同观测形式之间的互信息，用作训练和推理。今年的一些基于强化学习的信号控制方法通常采用直接拼接的方式，忽略了互信息。因此，本章在下一节提出一个基于双线性池化的特征融合模块。

## 4.6 特征融合

在强化学习中，智能体只能获得部分观测。已论述过单独的一种形式的观测难以刻画整个环境的特性，复杂的观测形式又会引起状态空间爆炸。上一章已经得到了含信息量最大的观测组合，并用简单的全连接层验证了结果。但是简单的全连接层有时并不能良好的融合特征。因此本节引入基于双线性池化的特征融合形式。

在经过有偏好状态编码，全连接层刻画高维信息，时空依赖关系捕捉以后，得到特征  $x^{st}$  以及  $w^{st}$ 。同时引入当前的信号灯相位  $p$ ，由于相位在模型中往往是一个常数表示，其蕴含的信息难以被模型理解。因此本章引入其最原始的基于连接器形式的表示，即按固定顺序以 0 或者 1 表示当前连接器的禁止和放行。通常以正北方的进口道开始，然后按顺时针顺序标识进口道对应连接器的放行状态。

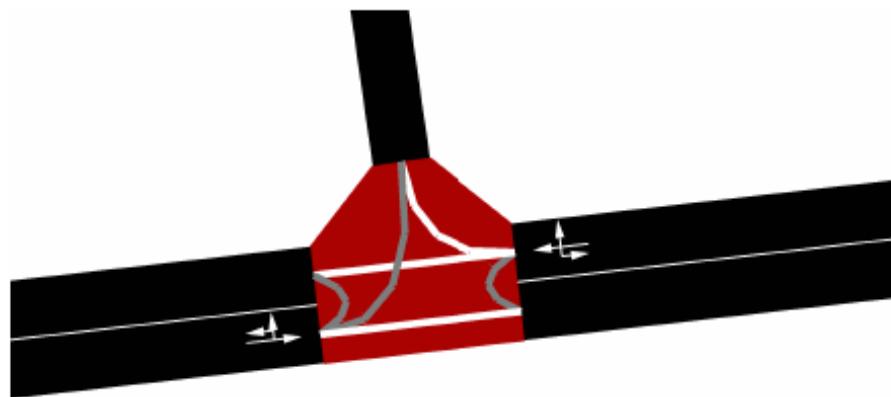


图4.5 相位状态

图 4.5 所示为一个 T 型交叉口，北方向是单向出口道，其相位状态可以表示为  $[1, 1, 0, 1, 0, 0]$ 。 $[1, 1, 0]$  表示右边的进口的相位状态， $[1, 0, 0]$  表示左边的进口的相位状态。相位状态使用单层全连接层进行处理，特征用  $p$  表示。此时，获得的特征为  $\{x^{st}, w^{st}, p\}$ 。

首先对特征  $x^{st}$  以及  $w^{st}$  进行双线性池化融合。先把两个特征进行双线性融合，得到矩阵  $b$ ，对矩阵  $b$  进行最大池化得到矩阵  $\xi$ 。将矩阵  $\xi$  张成一个张量，记为双线性张量，再对其进行矩阵归一化和 L2 归一化以后，得到特征张量  $z$ 。整个过程用公式描述为：

$$\xi(\mathcal{I}) = \sum_l b(x^{st}, w^{st}) \quad (4.16)$$

$$x = \text{vec}(\xi(\mathcal{I})) \quad (4.17)$$

$$y = \text{sign}(x) \sqrt{|x|} \quad (4.18)$$

$$z = y / \|y\|_2 \quad (4.19)$$

相位其蕴含的特征信息比较直接，因此在得到特征张量  $z$  后，将二者拼接起来作为一个特征向量。

## 4.7 实验设置

在本节，将在三种不同情况的交通环境中评估所提出的算法。超参数以及状态、动作和奖励等设置在 4.3 已经说明。评价指标使用旅行时间和路网吞吐量这两个常见指标，从两个角度分别说明提出模型的有效性。旅行时间越短越好，路网吞吐量则是越大越好。本章所有交通流参数中，旅行时间的单位是秒，路网吞吐量的单位是辆。

### 4.7.1 实验数据

在本节，基于从真实世界采集的数据绘制仿真路网，并且在仿真路网中评估提出的算法。首先是一个完全由标准交叉口构成的路网，取自杭州的真实路网。

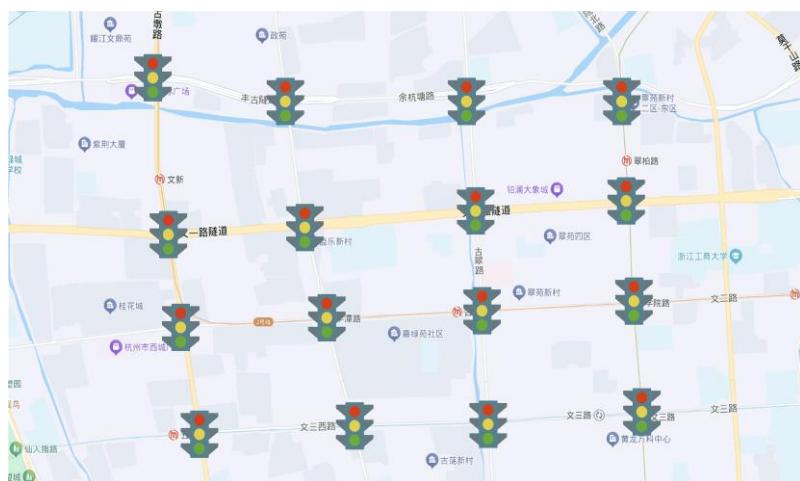


图4.6 现实世界标准交叉口结构

将其路网绘制在 SUMO 软件中，可以得到如图 4.7 所示的仿真路网。

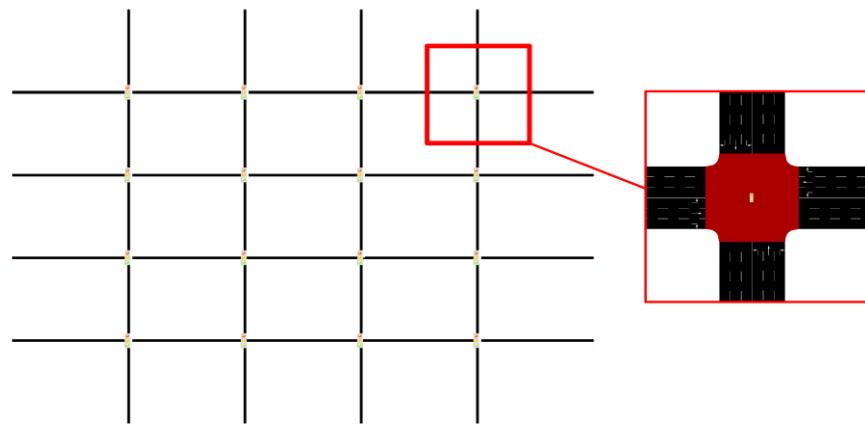


图4.7 SUMO 中绘制的标准交叉口路网

除此之外，干线作为城市交通的流量相对集中的区域，本小节的实验中也从真实世界一个城市的主干线采集了数据并且进行实验。数据来源于浙江省绍兴市凤林路，干线总长约为 3.6km。其干线结构如图 4.8。



图 4.8 现实世界干线结构

将地图上得到的干线结构数据转换为 SUMO 中的路网如图 4.9。

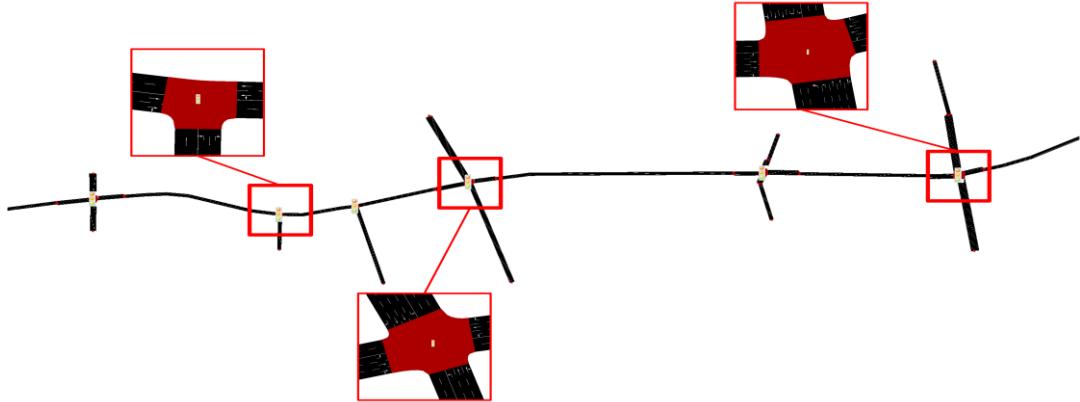


图4.9 SUMO 中绘制的主干线

上述两块路网结构都相对简单，为了应对复杂的交通情况，本章采用最近信号控制算法<sup>[106-107]</sup>中常见的一个复杂路网如图 4.10。



图4.10 SUMO 中绘制的复杂路网

该路网总占据  $51.54\text{km}^2$ , 车道总长度  $717.23\text{km}$ , 取其中的一部分, 并选择了其中 21 个交叉口进行信号控制。上述路网按顺序对其编号为路网 1(杭州), 路网 2(绍兴), 路网 3(常用复杂路网)。实验路网信息汇总如表 4.2。

表4.2 路网信息统计

| 路网  | 形式 | 三向交叉口数量 | 四向交叉口数量 | 流量(车辆数/时) |
|-----|----|---------|---------|-----------|
| 路网1 | 区域 | 0       | 16      | 2983      |
| 路网2 | 干线 | 2       | 4       | 6672      |
| 路网3 | 区域 | 17      | 4       | 4283      |

每个仿真路网的持续时间为 3600s。考虑到模型的优化能力, 在 3600s 以后仍有车辆生成, 满足优化需要。对于实验路网中需要控制的交叉口, 为了动作选择的需要, 数据中包含其初始的固定式相位方案。在仿真中, 车辆最大速度为  $16.7\text{m/s}$ , 车辆的最大加速度和减速度分别为  $2.6\text{m/s}^2$  和  $4.5\text{m/s}^2$ 。

### 4.7.2 对比方法

为验证所提出方法的有效性, 将其与如下几种算法进行了比较。

#### 传统信号控制算法

固定式配时 (Fixed Time Control, FTC) : 一组固定时间固定相位顺序的信号控制方案, 每个信号灯的控制方案来源于专家知识。

最大压力差 (Max Pressure, MP) : 通过各个相位放行的车道计算压力差, 并且设计当前的相位的压力差最大的相位。

#### 强化学习信号控制算法

MPLight<sup>[94]</sup>: 利用压力差的概念设计了强化学习智能体。通过精心设计的基于压力差的奖励机制, 个体控制代理之间可以实现隐式协调, 从而降低求解复杂度。奖励和状态都是基于压力差的。

CoLight<sup>[30]</sup>: 使用图注意力网络促进相邻智能体之间交流, 整合了相邻交叉口对本地交叉口的影响。采取去中心化的架构, 在三个大型的标准同构路网中进行训练, 并且允许智能体获取相邻交叉口的状态信息。

CosLight<sup>[99]</sup>: 提出将智能体选择合作者作为第二策略来进行学习, 并于原始策略同步更新。选择策略根据相位和交叉口级别的特征实时自适应地选择合作者。使用的算力资源略多余前三种信号控制算法。

为确保所得结果的鲁棒性, 每个算法的训练过程都会采用不同的随机种子重复 3 次, 结果取均值展示。各个算法的奖励设计不同, 因此实验结果中不会比较奖励值的大小,

转而观察客观评价指标的表现。客观评价指标（旅行时间和吞吐量等参数）将会使用仿真软件提供的接口直接输出。

## 4.8 实验结果以及分析

### 4.8.1 杭州路网的实验结果以及分析

图 4.11 展示了在杭州路网上，两种传统信号配时方案、三种对比强化学习信号控制方法以及本章提出的信号控制方案在旅行时间这一参数上的表现。传统算法在采用三次随机种子后取平均值，并以虚线绘制。可以观察到，固定式配时的表现明显劣于其它算法。这是因为固定式配时无法适应环境的动态变化，导致所有车辆在交叉口的等待时间延长，进而增加了车辆的旅行时间。相比之下，最大压力差式算法的表现已经接近基于强化学习的信号控制算法。MPLight 也是基于最大压力差设计的状态和奖励，因此二者在性能上较为接近。

本节对三种强化学习算法以及所提出的算法进行了对比。所有算法在训练初期，旅行时间均处于较高水平。随着训练回合数的增加，旅行时间这一客观指标开始下降。此外，CosLight 和 MPLight 在训练接近结束时出现了波动，而 Ours 和 CoLight 在训练初期也产生了较为明显的波动。这些波动是由环境的非平稳性引起的。相比之下，所提出的算法在整个训练过程中保持稳步下降，未出现明显波动，这表明所提出的模型在处理非平稳环境时具有显著的有效性。最终，所有基于强化学习的算法以及所提出的算法都成功收敛到一个较小的值。从图 4.11 可以明显看出，所提出的算法最终得到的旅行时间更小。从旅行时间的优化效果来看，所提出的算法  $>$  CosLight  $>$  CoLight  $>$  MPLight。

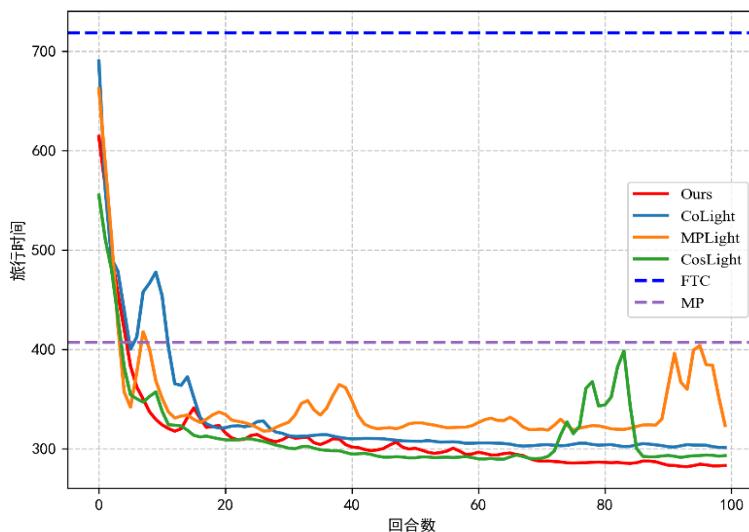


图4.11 旅行时间

接下来，对比路网中的车辆吞吐量。结果显示，固定式配时的吞吐量远小于最大压力差式的吞吐量。在对比几种强化学习算法的吞吐量表现时，可以观察到所有基于强化

学习的信号控制算法在吞吐量这一参数上均快速收敛，并维持在较高的水平。此外，吞吐量这一参数受环境非平稳性影响较小，在收敛后各算法的曲线均未产生明显波动。

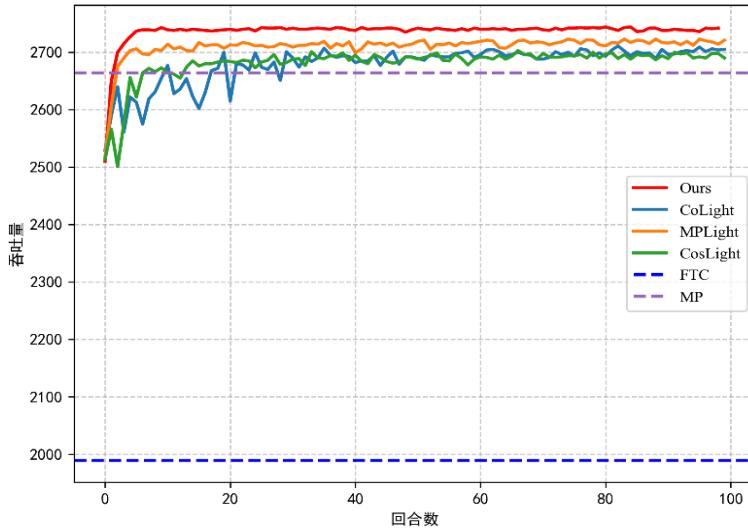


图4.12 吞吐量

综合两个结果图可以发现，旅行时间这一参数对环境的非平稳性更为敏感，能够较好地反映出环境的动态变化。相比之下，吞吐量这一参数主要反映路网的通行能力，对环境非平稳性的敏感度较低。在旅行时间参数中，CoLight 和 MPLight 在 10 回合左右，以及 CosLight 和 MPLight 在 80 回合和 90 回合左右，参数曲线均产生了明显的波动。而在吞吐量参数中，CoLight 在 10-20 回合左右也出现了类似的波动。所提出的算法在旅行时间和吞吐量参数上均表现出优于其它算法的稳定性，进一步验证了其在复杂环境中的有效性。

无论是旅行时间还是吞吐量，在初始阶段，强化学习算法通常采用随机策略进行动作选择和采样，因此这些算法在初始时处于接近的初值。在旅行时间参数中，所提出的算法稳步下降，并在 68 回合左右超过其它算法，开始达到最优效果。而在吞吐量参数中，所提出的算法收敛速度明显快于其它对比算法，在最初的 10 个回合内就接近了最优效果，并且后续一直保持最优状态。所提出的算法最终达到的最优效果优于其它算法。这不仅体现了强化学习模型的学习能力，也表明当前城市交通系统仍有很大的优化空间，尤其是通过固定式配时方案的优化效果可以看出这一点。

#### 4.8.2 凤林路网的实验结果以及分析

在城市交通中，干线是指在城市交通网络中承担主要交通流量的道路，通常具有较高的通行能力和较快的通行速度。干线在城市交通系统中扮演着至关重要的角色，连接城市的各个主要区域、交通枢纽和重要设施，确保城市内部的交通顺畅和高效。本小节采用凤林干线作为实验数据，在其上测试本章所提出的算法。给出凤林干线实验结果如表 4.3。

表4.3 凤林干线实验结果

| 算法       | 旅行时间   | 吞吐量  | 速度   | 等待时间   |
|----------|--------|------|------|--------|
| FTC      | 271.44 | 4521 | 7.10 | 131.09 |
| MP       | 257.36 | 4433 | 7.79 | 122.58 |
| MPLight  | 245.44 | 5023 | 8.22 | 103.38 |
| CoLight  | 236.72 | 5325 | 8.4  | 94.56  |
| CosLight | 237.46 | 5488 | 8.68 | 93.79  |
| 本章提出的算法  | 231.58 | 5639 | 8.79 | 90.46  |

对于干线来说，环境的非平稳性比之路网会相对较小。路网内部的交叉口中的车辆一定会流向相邻交叉口，而干线中的车辆有可能直接驶出干线，不再继续在干线中形式，因此减少了车道带来的环境的非平稳性的影响。

5 种对比算法中，优化效果相对较差的仍是两种传统算法。与区域交通场景中最大压力差式算法显著优于固定式配时算法的情况不同，在干线交通场景中，两种传统算法的表现较为接近。具体而言，在吞吐量这一关键指标上，固定式配时算法略微优于最大压力差式算法，这是因为固定式配时算法在干线这种相对规则和稳定的交通流中能够提供较为一致的信号周期，从而在一定程度上保证了交通流的连续性。而对于旅行时间和等待时间这两个指标，两种传统算法的表现则较为接近，均未能显著优化交通效率。在基于强化学习的信号控制算法中，除了 MPLight 以外，其余三种算法表现较为接近。这一现象可能与最大压力差式算法在处理干线交通场景时存在局限性有关。最大压力差式算法虽然在区域交通场景中表现优异，但在干线场景中，由于其设计初衷更多是针对局部交通压力的缓解，可能无法完美地处理干线这种长距离、连续性的交通流特性，从而导致其优化效果不如预期。另一方面，MPLight 的状态和奖励设计是基于传统的最大压力差式的，这也说明最大压力差这个形式难以适应干线的形式。本章所提出的算法在旅行时间，吞吐量，速度以及等待时间这四类指标上都优于三种对比的强化学习算法。等待时间与旅行时间以及吞吐量的优化表现都较为接近，这也一定程度说明了干线的非平稳性小于区域的非平稳性。干线交通通常具有更强的规律性和连续性，本章算法在优化过程中能够更稳定地捕捉交通流的特征，从而实现更高效的信号控制。

### 4.8.3 复杂区域的实验结果以及分析

前文已经在只具有标准交叉口的路网以及干线中测试了所提出的算法。本小节利用复杂路网如图 4.10 对所提出的算法进行测试。它包含了各种结构的交叉口，并且信控交叉口不一定是一阶相邻交叉口，因此在具体实现时会判别交叉口间的位置关系，并选择合作的交叉口进行图注意力的计算。可以从结构上明显看出，本小节路网的非平稳性远超前两个路网。

首先仍给出训练过程中旅行时间的统计图：

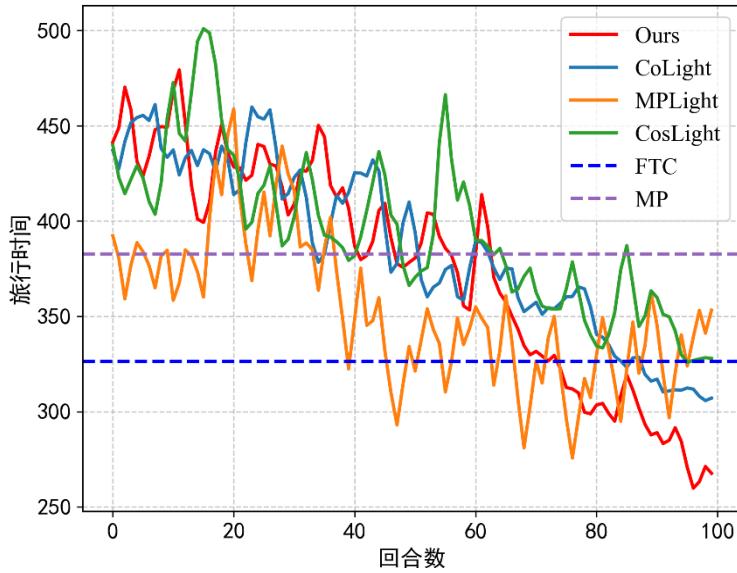


图4.13 复杂路网中的旅行时间统计

对比两种传统算法，固定式配时的优化能力超出最大压力差式配时算法 17.1%。固定配时通过历史数据获得预设的配时方案，并且对于复杂路网的优化效果超过了最大压力差式配时算法，这一方面说明固定式配时不会受到路网结构带来的复杂性的影响，另一方面也说明了从历史数据中学习而得到的模型是具有切实的优化能力的。最大压力差式的配时方案在应对复杂结构的路网时确有不足，基于压力差设计的 MPLight 表现出较强的波动性。压力差这个形式在应对复杂图结构时计算得到的结果并不准确。

给出去除最大压力差式配时，固定式配时以及 MPLight 训练曲线的旅行时间结果图观察其波动性。

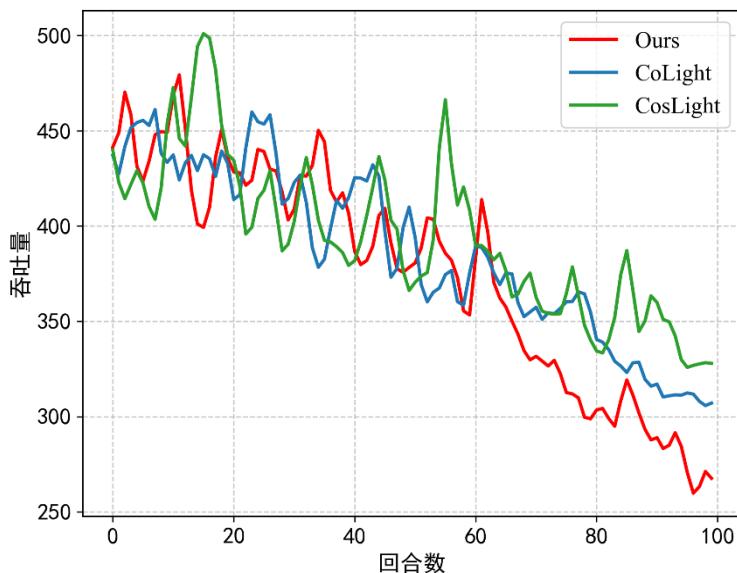


图4.14 去除三种算法后的旅行时间统计

在训练到第 65 个回合时，所提出的算法超过另外两种优化算法，并保持优势直到完整的训练过程结束。最终优化效果超过另外两种算法 15.0% 以及 22.5%。从收敛过程上来看，CosLight 的波动性最强，这与其具体实现方式有关。CosLight 是实时自适应的选择合作者，这种机制虽然能够在一定程度上适应环境的变化，但在复杂环境中，变换合作者无疑会增加环境的非平稳性，从而导致训练过程的不稳定，交通环境本身复杂性就较高，引入额外的复杂性无疑会导致模型越发难以收敛。而 CoLight 从收敛过程的稳定性来说大致与本章所提出的算法类似，较少出现明显的波动情况。CoLight 是基于值的强化学习算法，训练过程中会利用经验池。在复杂环境中，经验池能够通过存储和重用历史样本，有效平滑训练过程，减少因环境非平稳性带来的波动。这种设计使得 CoLight 在训练过程中表现出较高的稳定性，但其优化效果仍不及本章所提出的算法。本章提出的算法在优化效果超过另外两种算法后，剩余的训练过程表现的十分平稳，进一步验证了其在复杂交通场景中的鲁棒性。在训练的前半段波动性也并不明显，说明所提出的算法在应对结构复杂的路网时依旧能处理其非平稳性。

除此之外，吞吐量这一指标的训练曲线如下：

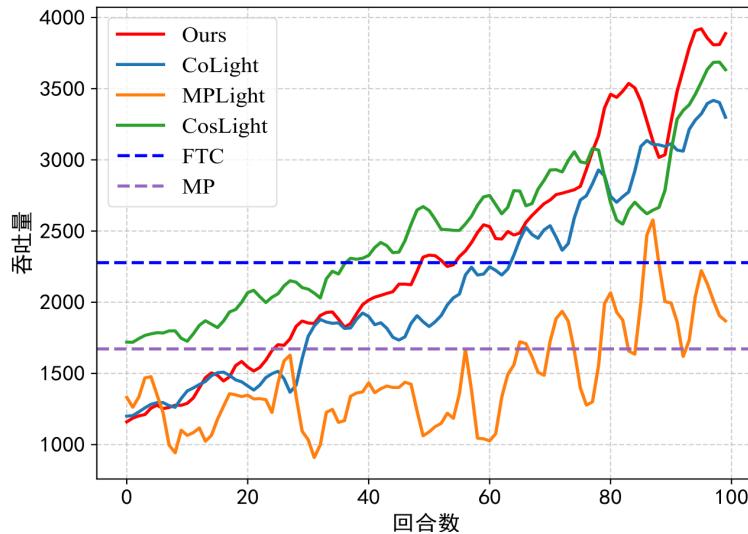


图4.15 复杂路网中的吞吐量统计

在之前的路网中，吞吐量比旅行时间更能反映环境的非平稳性，这是因为吞吐量直接受到交通流动态变化的影响，例如车辆到达率、信号控制策略以及交叉口之间的相互作用等。而在当前的复杂路网中，旅行时间则更能反映环境的非平稳性，这是因为复杂路网中的路径选择、交通拥堵以及信号延迟等因素对车辆行驶时间的影响更为显著。吞吐量这一指标上各个算法的表现与旅行时间上的表现类似。除 MPLight 之外的其余强化学习算法都并未出现明显波动。在第 80-85 回合训练时，所有强化学习算法都出现了一次明显的抖动，一方面来源于智能体在进行探索与利用，另一方面说明环境本身的随机性在这个时间步时会有一次明显的表现。

对比随机时刻路网中某一路口的通行情况。用以说明不同算法的实际优化能力。

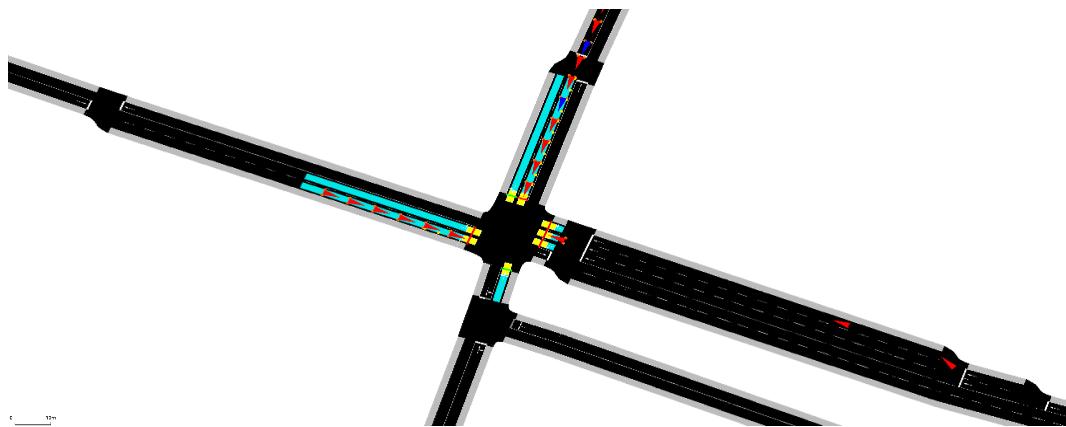


图4.16 MPLight 随机时刻路况

交叉口的东进口道以及北进口道出现明显车辆排队情况，且北方车辆明显已堆积一段时间。南方进口车道明显无车辆进入，但是 MPLight 选择放行南方进口车道以及北方右转车道，使得需要放行的车道的需求没有被及时响应。

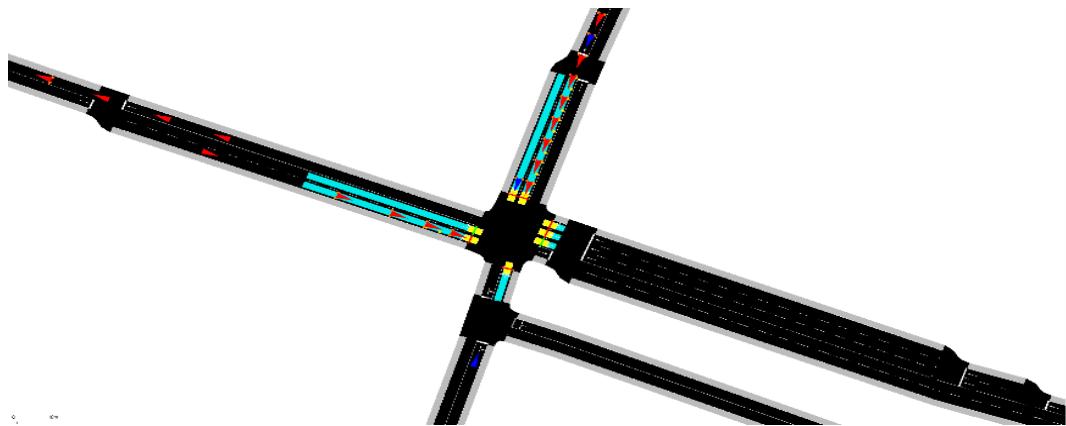


图4.17 CoLight 随机时刻路况

相比于 MPLight，CoLight 北方也有车辆堆积，但是东方进口车道车辆堆积情况减少。北方的车辆堆积可能来源于路网本身的路径信息。南方出现小流量车辆输入，但还未被检测器检测到。相对来说，这一时刻 CoLight 优化能力超过 MPLight。

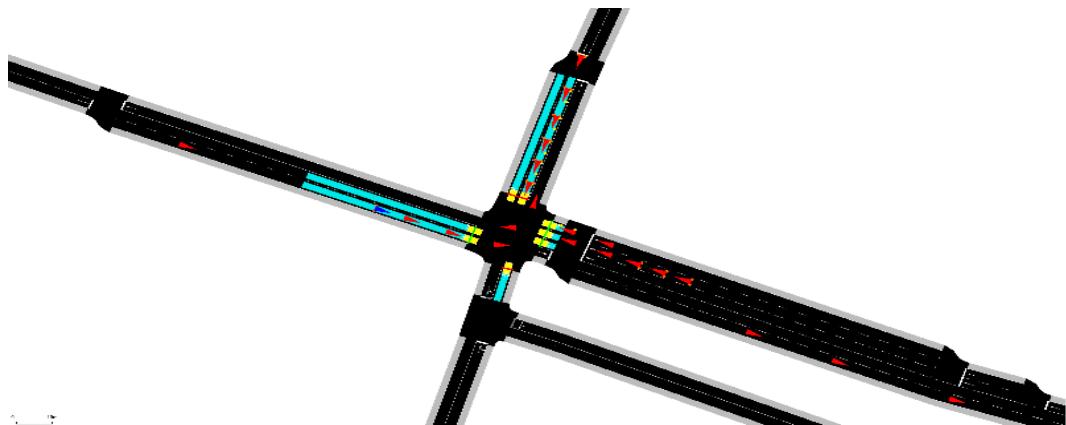


图4.18 CosLight 随机时刻路况

CosLight 控制的交叉口的东方和北方进口车道车辆堆积情况都有了极大改善，且来车较多的北方向车辆暂时未形成拥堵，没有出现放行空车道情况。北进口道的车道占有率和排队长度都优于之前的算法。南方同样没有进口车辆，这说明路网中该交叉口南方本身车辆流动就较少。

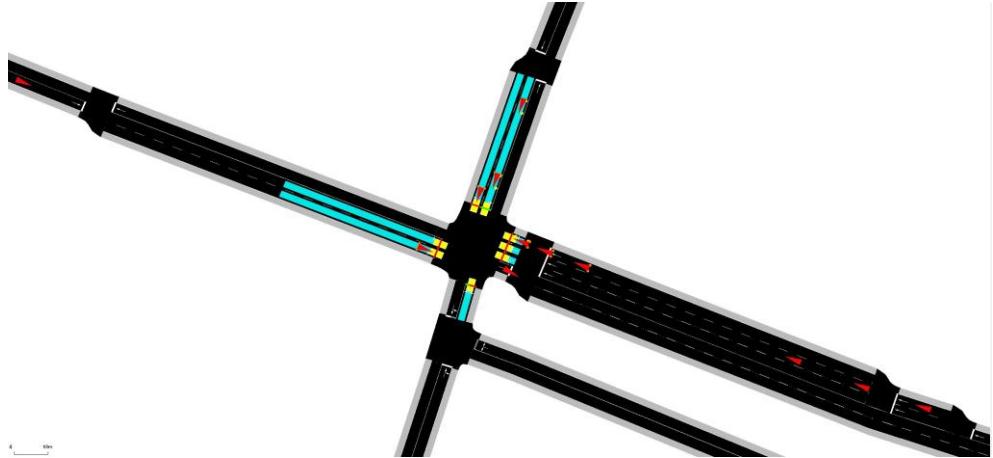


图4.19 所提出算法中随机时刻路况

图 4.19 显示所提出的算法特定时刻车道未出现明显车辆堆积，说明前文中出现的车辆堆积的情况来源于算法本身优化能力有限的问题。在此时刻之前，算法对当前交叉口做出了合理的优化。同时交叉口中明显正在行驶的车辆更多，这说明所提出的算法加强了车辆的流动性，从而提高了车辆的形式速度，缩小了旅行时间。

#### 4.8.4 实验结果总结

将前三小节实验结果汇总于此。将最优算法所得到的结果加粗，将次优算法所得到的结果用下划线标识。首先给出吞吐量这一指标汇总得到表 4.4。其中，本章所提出的算法命名为 E-ST-F。E 表示有偏好编码模块，ST 表示时空依赖关系捕捉模型，F 表示特征融合模型。

表4.4 吞吐量实验结果汇总

| 路网   | 车流量  | FTC  | MP   | MPLight     | CoLight | CosLight    | E-ST-F      |
|------|------|------|------|-------------|---------|-------------|-------------|
| 路网 1 | 2983 | 1989 | 2664 | <u>2721</u> | 2705    | 2690        | <b>2742</b> |
| 路网 2 | 6672 | 4521 | 4433 | 5023        | 5325    | <u>5488</u> | <b>5639</b> |
| 路网 3 | 4283 | 2278 | 1672 | 2531        | 3299    | <u>3633</u> | <b>3887</b> |

吞吐量的绝对值在不同路网间难以直接比较，因为各路网的规模、交通需求和结构特征存在显著差异，这使得其数值难以全面反映算法的优化效果。因此，通过计算车辆的到达率，即驶出路网的车辆数目与总车辆输入的比值，能够更直观地体现算法的实际表现。在同一路网中，到达率越高，意味着算法在提升交通效率和优化路网运行方面具有更显著的效果。计算得到车辆到达率统计结果如表 4.5。

表4.5 各个算法实际到达率

| 路网  | FTC    | MP     | MPLight       | CoLight | CosLight      | E-ST-F        |
|-----|--------|--------|---------------|---------|---------------|---------------|
| 路网1 | 66.68% | 89.31% | <u>91.22%</u> | 90.68%  | 90.18%        | <b>91.92%</b> |
| 路网2 | 67.76% | 66.44% | 75.29%        | 79.81%  | <u>82.25%</u> | <b>84.52%</b> |
| 路网3 | 53.40% | 39.04% | 59.09%        | 77.03%  | <u>84.82%</u> | <b>90.75%</b> |

可以观察到，在结构相对简单的路网1和路网2中，所提出的算法优化效果虽然最佳，但较之次优算法，性能提升有限。在路网1中超过次优算法0.7%，在路网2中超出次优算法2.27%。在结构最复杂的路网3中，超过次优算法5.93%，是同比基于强化学习的信号控制算法中优势最明显的。实验结果证明了所提出算法在应对复杂环境时的优势。类似的，给出所有实验的旅行时间统计结果如表4.6。

表4.6 各个算法旅行时间

| 路网  | FTC    | MP     | MPLight | CoLight       | CosLight      | E-ST-F        |
|-----|--------|--------|---------|---------------|---------------|---------------|
| 路网1 | 718.29 | 407.17 | 323.34  | 301.34        | <u>293.17</u> | <b>283.19</b> |
| 路网2 | 271.44 | 257.36 | 245.44  | <u>236.72</u> | 237.46        | <b>231.58</b> |
| 路网3 | 326.43 | 382.64 | 未收敛     | <u>307.10</u> | 327.92        | <b>267.63</b> |

旅行时间是一个非常复杂的指标。但是在不同的路网结构中，旅行时间的绝对值意义并不大。因此，类比吞吐量这一指标，使用固定式配时所得到的旅行时间的结果作为基线，然后进行对比，用 $tt_{fpc}$ 标识固定式配时的旅行时间， $tt_e$ 表示其余算法的旅行时间，则指标计算方式可以表示为式(4.20)。

$$\frac{tt_{fpc} - tt_e}{tt_{fpc}} \quad (4.20)$$

此处统计得到的百分数指标应该越大，说明指标对应的算法优化效果超过固定式配时越多，优化效果也就越好。所得结果如表4.7。

表4.7 各个算法旅行时间指标对比

| 路网  | MP      | MPLight | CoLight | CosLight | E-ST-F |
|-----|---------|---------|---------|----------|--------|
| 路网1 | 43.31%  | 54.98%  | 58.04%  | 59.18%   | 60.57% |
| 路网2 | 5.18%   | 9.58%   | 12.79%  | 12.52%   | 14.68% |
| 路网3 | -17.22% | /       | 5.92%   | -0.46%   | 18.01% |

在三个路网中，所有算法对于路网2的旅行时间优化能力都较小。路网2是干线，说明干线这一特殊的结构使得算法对其优化能力有限。而对于其余两个路网，路网1的旅行时间优化能力达到了60.57%，远远超过了其余几种基于强化学习的信号控制算法。再次证明了所提出算法的实际优化能力。

## 4.9 消融实验

为了更进一步的分析所提出模型的效果，在本小节将拆解所提出的各个模块，并且基于杭州的路网进行实验。所提出的模型表示为 E-ST-F 在实验中，缺少哪一种则表示该字符对应的模块没有起作用。同时引入基础的独立 PPO 模型进行对比。首先给出吞吐量这一对非平稳性敏感的指标表现：

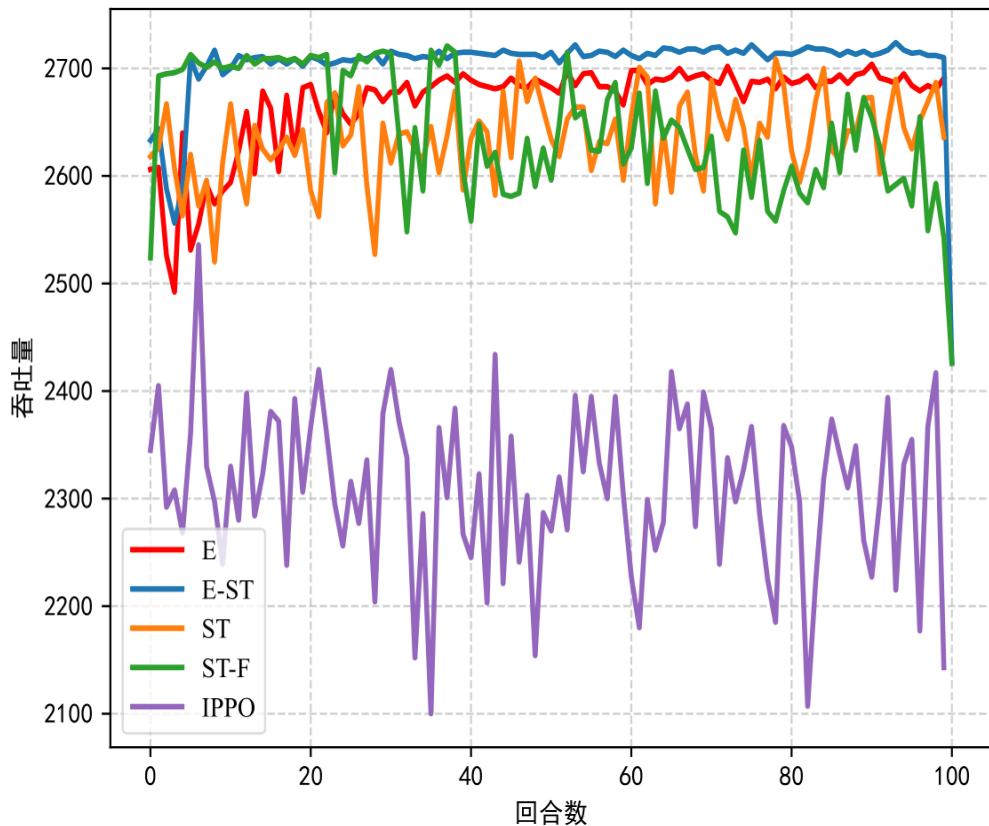


图4.20 吞吐量的消融实验

首先可以观察到，IPPO 在杭州路网上无法收敛，难以应对路网的非平稳性。整个 IPPO 算法的吞吐量曲线在 2300 左右抖动，这表明 IPPO 在处理复杂交通环境时缺乏稳定性，无法有效捕捉和适应交通流的动态变化。这一现象可能与 IPPO 的独立学习机制有关，即每个智能体单独优化策略而忽略了交叉口之间的相互依赖关系，导致其在非平稳环境中表现不佳。所提出的算法进行对比，首先是对比有无编码模块的模型，可以发现有编码模块的曲线的平稳性明显超过没有编码模块的模型，这证明了所提出的模型应对非平稳环境的能力，其能够通过提取和整合环境中的关键特征，有效降低非平稳性对算法性能的影响。除此之外，在有编码的两组实验中（E 与 E-ST），有时空依赖关系捕捉模块的模型效果又超过了仅有编码模块的模型，这一现象说明，时空依赖关系捕捉模块能够更好地建模交通环境中交叉口之间的时空关联性，从而进一步提升算法的性能。

无编码模块的模型表现出明显的波动性，并且效果也劣于有编码模块的模型。这表明在处理交通环境这一类复杂环境时，环境的非平稳性必须优先进行处理。而通过引入编码模块和时空依赖关系捕捉模块，能够有效提升算法的稳定性和性能。

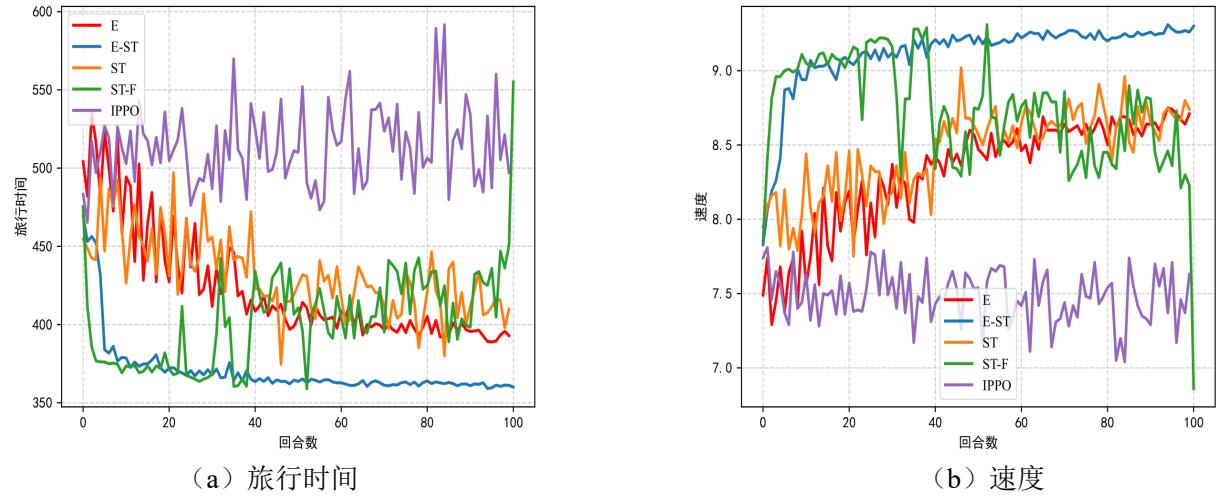


图4.21 其余参数表现

在图 4.21(a)与图 4.21(b)中，给出旅行时间与速度这两个参数的收敛曲线。IPPO 算法依旧处于剧烈震荡且无收敛表现的状态。针对所提出的模型的消融实验中，仍然是具有编码模块的模型表现最好，且同时具有编码模块和时空依赖关系捕捉模块效果最好，且收敛较为平稳。但在这两个参数中，仅仅具有编码模块的模型出现了明显的震荡，这说明时空依赖关系捕捉模块也能处理一部分环境的非平稳特性。针对这两个参数的四组消融实验中，具有时空依赖关系捕捉模块和基于双线性池化的特征融合表现出震荡，并且未能成功收敛。这证明模型抵抗非平稳性的能力大部分来源于编码模块，未编码的模型难以保证收敛。

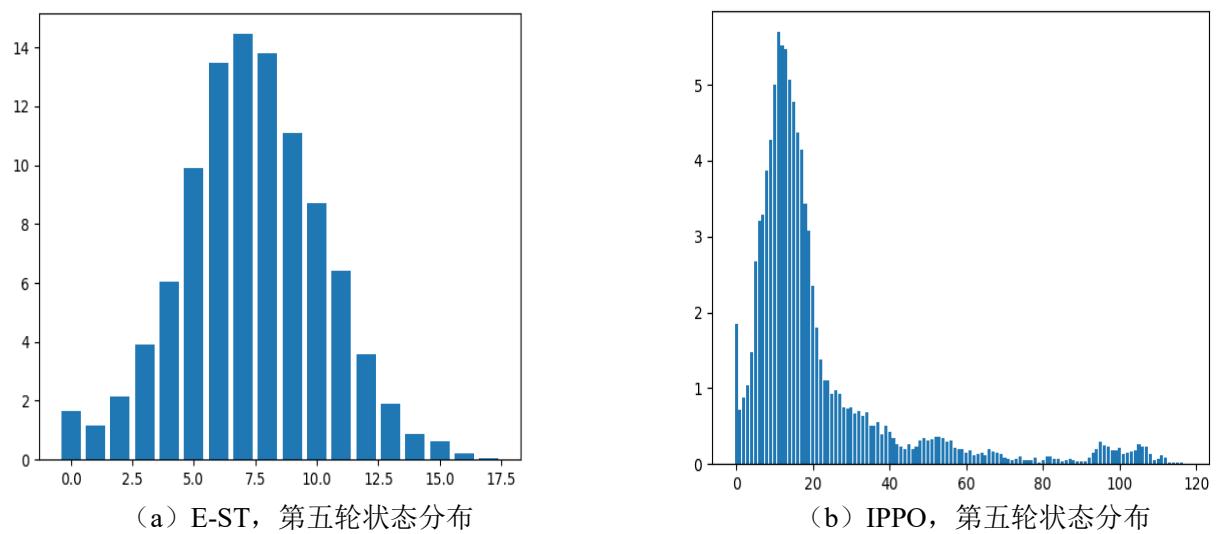


图4.22 训练过程中的状态分布

依照图 4.3 的状态分布定义，横轴表示交叉口排队车辆数目，纵轴表示该状态出现的频率，对比 E-ST 和 IPPO 算法的训练第五轮的状态分布图 4.22。可以观察到，E-ST 模型的状态分布更加紧凑，并且大多分布在排队车辆数目较少的区域，而 IPPO 算法的状态分布的较为宽广，并且状态分布大多在排队车辆数目小于 30 内。虽然状态分布大多处在奖励值较高的区域，但是其状态分布的区域太广，未经过任何处理的这种类型的状态分布是环境的非平稳性的一种体现。车辆排队数目少意味着奖励值更大。这表现出编码模型的优越性，同时可以观察到，E-ST 模型的状态分布大致呈现出正态分布的形式。聚集在奖励值较高的区域并且表现出紧凑的特性证明了编码模型缓解环境的非平稳特性的效果。

此外，给出有编码的模型和 IPPO 前三十轮训练过程中收集到的状态分布图（横向排列）。也可以明显观察到有无编码模型的表现上的差异性。有编码模型 E-ST 在前四轮时还处于探索状态，探索到的状态分布此时与 IPPO 的表现类似。但是在第五轮时，整个模型迅速收敛，与消融实验中得到的曲线结果一致。观察到 E-ST 后续收敛较为稳定，并且状态分布类似于一个正态分布。IPPO 算法前三十轮的收敛过程中，访问到的状态分布都比较广，这意味着 IPPO 未能识别出对模型更有效的区域。但随着整个收敛过程中，访问到的状态分布有缓慢的，集中的趋势。但未能表现出收敛。

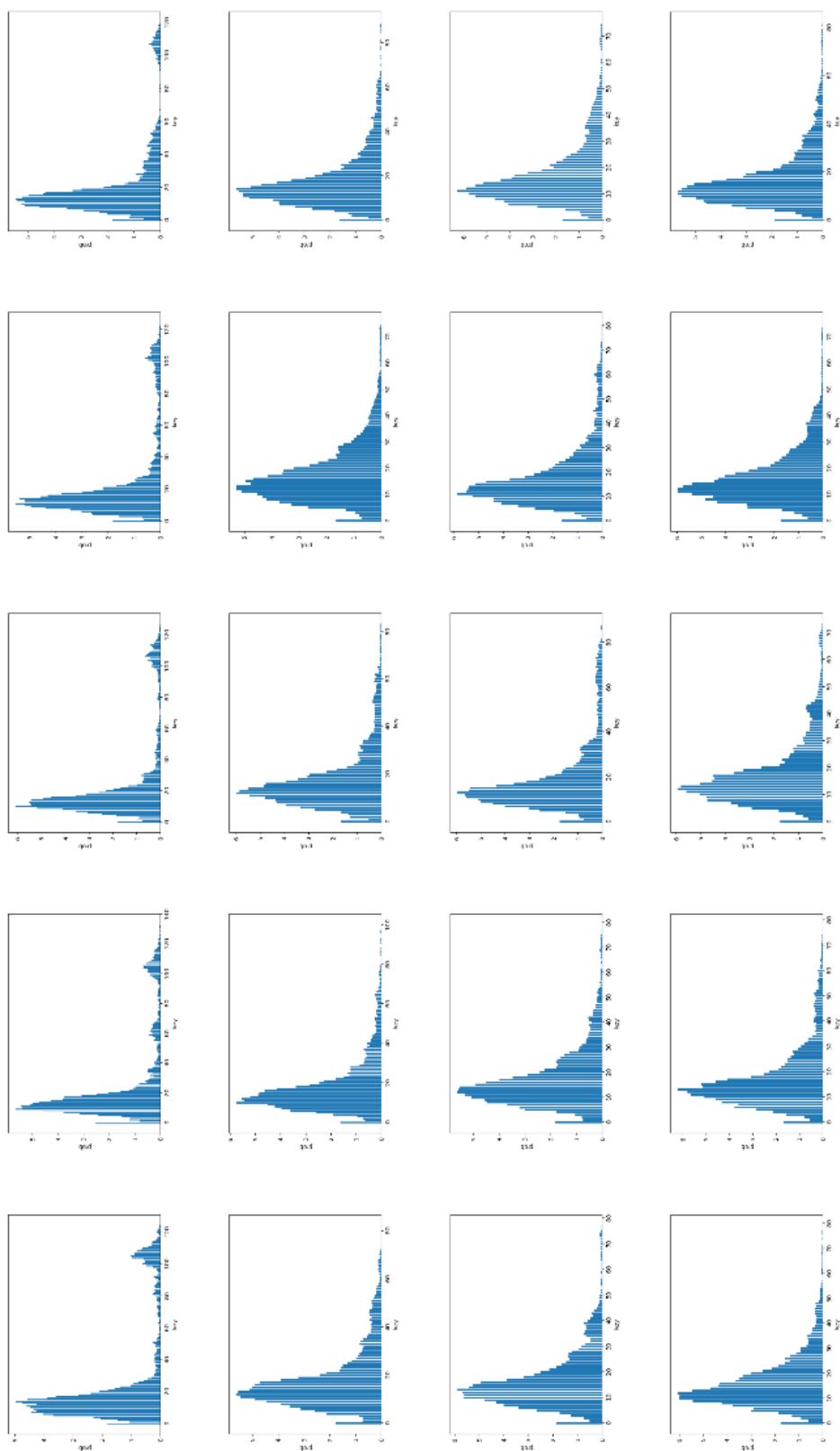


图 4.23 IPPO 前三十轮收敛性

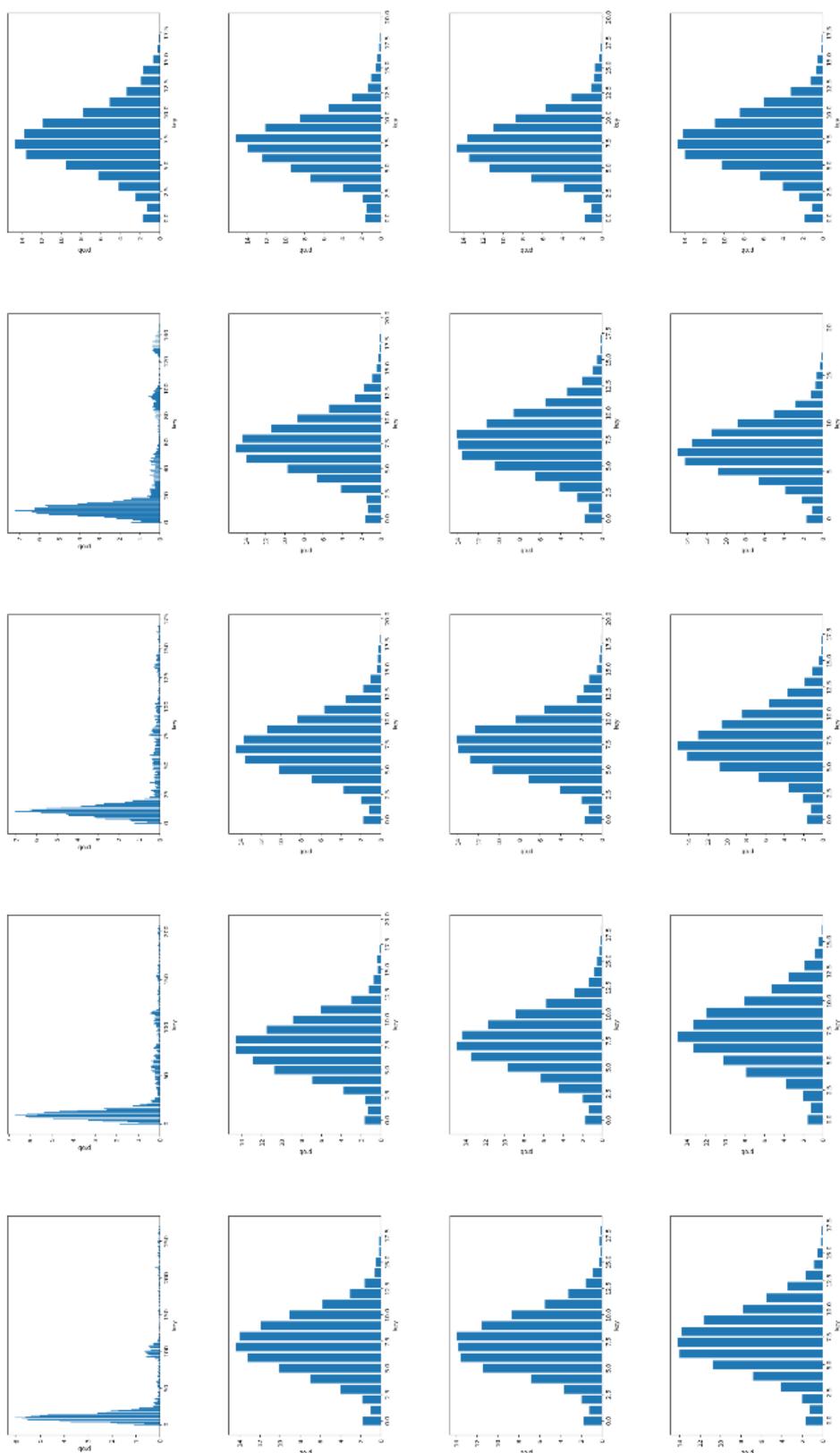


图 4.24 E-ST 前三十轮收敛性

## 4.10 关于模型泛化性的讨论

交通系统是一个有强烈非平稳性的环境。这种复杂性可以归因于其结构的复杂性以及车辆行为的复杂性。传统的固定式配时虽然优化效果不佳，但本身并没有增加系统的复杂度。但当使用在线优化方法，例如深度强化学习技术时，情况变得复杂起来。不同的配时方案会影响车辆的流动特性，这会导致智能体所探索到的状态分布变化。深度强化学习模型训练过程中的波动性也来源于此。在当前回合训练结束后，得到智能体行为策略  $\pi_\theta(a_t | s_t)$ ，该策略实际上依赖于上一个回合的智能体与环境交互采集得到的经验，其中的状态分布假设为  $S^1$ ，但当智能体实际采取策略  $\pi_\theta(a_t | s_t)$  与环境交互时，得到的状态分布  $S^2$  与  $S^1$  存在着明显的差异性， $S^1 \neq S^2$ 。这种分布上的差异性可以通过多轮仿真减缓其影响，也就是增加算力资源。但实际训练过程中，仿真不可能无限制的开展。经过  $m$  轮训练以后，可以认为得到的策略  $\pi_\theta^*(a_t | s_t)$  在状态分布  $S' = S^1 \cap S^2 \cap \dots \cap S^m$  上是最好的。

但实际部署后，真实世界所产生的状态分布形式上是无限的，仿真中探索到的状态分布是其子集， $S' \in S$ 。此时可能产生超出分布<sup>[110]</sup> (Out-of-Distribution, OOD) 的情况，这也是深度学习领域常见的一个问题<sup>[111]</sup>。有研究尝试过用扩散模型<sup>[112]</sup>处理 OOD 的情况，但扩散模型本身就是一个较为复杂的架构，难以满足交通信号控制领域实时性的需要。

本章提出的状态编码方法则利用编码的方法提高了模型的泛化能力，实验结果也验证了所提出模型的有效性。具体而言，训练过程中探索到的状态分布为  $S'$ ，训练时使用的状态分布是编码过后  $\mathcal{E}(S')$  的。但分布  $S'$  之外的一些状态其实也可以被编码到  $\mathcal{E}(S')$  中，并且因为复杂交通系统中状态的值比较大，这种分布外所能访问到的状态的编码实际上是较大的一个区域。因此模型完成了访问状态的扩散，减缓了 OOD 情况带来的影响。

## 4.11 本章小节

在本章中，针对现有的算法在有限计算资源下可能出现 OOD 的情况，以及可能忽略了交通状态信息的时序特性和空间特性的情况，提出了 E-ST-F 模型。其使用状态编码减少了环境的非平稳性带来的影响，并且使用时空依赖关系捕捉模块增强了模型对于交通状态信息中的时空信息感知。除此之外，针对特征融合，引入了双线性池化，有助于获得一个更良好的特征表示。实验结果表明，E-ST-F 在多个指标上优于当前其余基于强化学习的信号控制算法。

## 5 总结与展望

### 5.1 全文总结

本文主要研究了采用多智能体强化学习算法解决路网层面的信号控制问题，并通过所提出的不同方法解决了现有的去中心化强化学习算法存在的一些不足。全文的总结如下：

1) 提出了一种观测形式分布距离计算模型。有效解决了基于强化学习的信号控制问题中状态形式究竟该怎么设计这个问题。它利用神经网络的有限拟合能力弥补了不同观测之间形式上的差异，并通过不同观测同时发生的马尔可夫过程对转移概率进行拟合，最终得到了基于强化学习的信号控制问题中的最优观测组合形式。并且在这个过程中，利用不同的观测同步发生的 MDP 有相同的奖励函数和状态转移概率，大大减少了计算模型中互模拟度量需要的计算资源消耗。

2) 提出了一种基于先验知识的有偏好状态编码模型。它扩大了在有限计算资源条件下智能体所能探索到的区域的占比，并且缓解了实际部署后可能出现的 OOD 的情况带来的影响。并利用先验知识判别出状态空间的哪一部分是更加值得关注的区域，并且通过有偏好状态编码放大这一区域在整个经验池中的占比，从而实现减少计算资源消耗这一目的。并且通过编码，未曾访问过的状态可以被编码为已经训练过的状态，增强了模型的外推能力。实验结果表明了编码模块可以有效环节交通环境的非平稳性带来的影响。

3) 提出了一种时空依赖关系捕捉模型。它弥补了常见的基于强化学习的信号控制模型或是忽略时序信息或是忽略空间依赖关系的情况，有助于缓解去中心化模型中的部分可观测性带来的影响。它通过长短时记忆网络和一阶相邻节点的图注意模型获得时空依赖信息。这种方式保证了在去中心化框架下所提出的算法仍能达到全局最优的效果。实验结果也证明了该模块可以缓解环境的非平稳性带来的负面影响。

4) 提出了一种基于双线性池化的特征融合模型。它帮助模型获得更加具有表达能力的时空特征。给定两个特征向量，双线性池化首先计算它们的外积，得到一个双线性特征矩阵；然后对该矩阵进行全局池化，将其转换为一个固定维度的特征向量。通过这种方式捕捉的高阶特征包含了不同交通参数之间的互信息，使得模型具备更有力的信息提取能力。

基于以上提出的四点，从真实世界获取车流，路网结构等信息进行实验。实验结果表明所提出的模型针对复杂路网的优化效果超过目前基于强化学习的信号控制算法。并且有效处理的复杂交通环境带来的状态空间爆炸问题。

## 5.2 研究展望

本章提出的模型虽然在优化效果上超过了目前基于强化学习的信号控制算法，但仍存在一些问题有待解决：

1) 现有强化学习算法难以直接迁移到真实交通环境主要面临三方面挑战。首先，仿真环境通常基于理想化假设构建，其状态空间和动作空间往往经过高度简化（如忽略天气突变、传感器误差等），导致训练出的策略在面对现实世界的长尾场景（如紧急避让、突发性道路施工）时泛化能力不足。其次，真实环境中的试错成本极高，现有算法依赖的大量探索机制（如  $\epsilon$ -greedy 策略）可能引发安全隐患，特别是在涉及人车混行或复杂路口场景时，细微的决策偏差可能引发严重后果。此外，现实系统存在难以建模的延迟特性（如控制指令传输延迟、传感器数据异步更新），这种时序错位会破坏强化学习依赖的马尔可夫假设，导致策略在实际部署时出现系统性偏差。尽管近期研究开始关注仿真到现实的域适应（Sim2Real）技术，但如何构建具有足够保真度的交通仿真器，以及如何设计兼顾安全约束与探索效率的强化学习框架，仍然是亟待突破的关键问题。

2) 在仿真实验中，本文选取了多份路网作为优化对象，以证明所提出模型的鲁棒性。然而，对于真实世界来说，实验的路网数量、交叉口的种类未能覆盖全面。不同的城市、不同的交通模式、不同的道路结构都会对模型的表现产生影响。更真实的交通数据以及情况更加多变的交叉口数据是进一步研究所需要的。因此，未来的研究应收集更多样化的真实交通数据，涵盖不同类型的路网和交叉口，以提高模型的泛化能力和鲁棒性。

3) 从对比的角度来说，本文实验只对比了吞吐量、旅行时间以及速度这三个交通流参数。然而，实际的交通环境中，还包括各种更加复杂的交通流参数，比如延误、车道占有率、燃油消耗、排放量等。这些参数对于评估交通系统的整体性能同样至关重要。考虑到这个方面，进一步的研究应对比更加全面的交通流参数，以更全面地评估模型的性能。此外，还可以考虑引入多目标优化方法，综合考虑多个交通流参数的平衡，以实现更加综合和可持续的交通管理。

综上所述，虽然本章提出的模型在仿真环境中表现较好，但在实际应用中仍需进一步验证和改进。未来的研究应着重于模型的实际部署、数据多样性的扩展以及更全面的交通流参数评估，以推动智能交通系统的发展和应用

## 致 谢

衷心感谢我的导师张伟斌教授在论文研究过程中给予的悉心指导和无私帮助。张教授渊博的学识、严谨的治学态度以及对科研工作的热情，深深感染并激励着我。在课题选择、研究方法以及论文撰写等各个环节，张教授都给予了宝贵的建议和耐心的指导，使我在学术道路上不断成长。张教授的教诲和关怀将使我受益终身，在此谨致以最诚挚的谢意。此外，感谢：

陈泽宇，管毅诚，胡一涵，李鹏飞，吴印锋，王玮，王龙，徐晓，余锡新，朱世豪。

## 参考文献

- [1] 公安部交通管理局. 全国机动车达 4.4 亿辆驾驶人达 5.32 亿人[EB/OL]. 2024. [http://www.gov.cn/lianbo/bumen/202407/content\\_6961935.htm](http://www.gov.cn/lianbo/bumen/202407/content_6961935.htm).
- [2] 公安部交通管理局. 2023 年交通运输行业发展统计公报[EB/OL]. 2024. [https://xxgk.mot.gov.cn/2020/jigou/zhghs/202406/t20240614\\_4142419.html](https://xxgk.mot.gov.cn/2020/jigou/zhghs/202406/t20240614_4142419.html)
- [3] 赵瑞东, 方创琳, 刘海猛, 等. 城市韧性研究进展与展望[J]. 地理科学进展, 2020, 39(10): 1717-1731.
- [4] 百度地图 2023 年度中国城市交通报告[EB/OL]. 2023. <https://jiaotong.baidu.com/cms/reports/traffic/2023/index.html>.
- [5] 赵靖, 陈凯佳, 周溪召. 排阵式交叉口几何设计与信号控制协同鲁棒优化[J]. 中国公路学报, 2021, 34(11): 296-305.
- [6] Collotta M, Bello L L, Pau G. A novel approach for dynamic traffic lights management based on Wireless Sensor Networks and multiple fuzzy logic controllers [J]. Expert Systems with Applications, 2015, 42(13): 5403-5415
- [7] 沈岩, 武彤冉, 闫静, 等. 基于 COPERT 模型北京市机动车大气污染物和二氧化碳排放研究[J]. 环境工程技术学报, 2021, 11(6): 1075-1082.
- [8] International Road Transport Union. Traffic Congestion Increases CO<sub>2</sub> Emissions by 300%[EB/OL]. 2014. [http://www.irtu.org/en\\_policy\\_co2\\_response\\_flowingletraffic](http://www.irtu.org/en_policy_co2_response_flowingletraffic).
- [9] 杨劲. 探究公路设计因素对交通安全的影响[J]. 现代交通与路桥建设, 2024, 3(11) : 79-81.
- [10] 王嘉文, 马万经, 杨晓光. 紧急交通流信号控制优先级划分模型[J]. 东南大学学报: 自然科学版, 2014, 44(1): 222-226.
- [11] Su H, Zhong Y D, Chow J Y, et al. EMVLight: A multi-agent reinforcement learning framework for an emergency vehicle decentralized routing and traffic signal control system[J]. Transportation Research Part C: Emerging Technologies, 2023, 146: 103955.
- [12] 项俊平. 城市道路交通信号区域均衡控制方法及应用研究[D]. 中国科学技术大学, 2018.
- [13] Wei H, Zheng G, Gayah V V, et al. A Survey on Traffic Signal Control Methods[J]. arXiv preprint arXiv:1904.08117, 2019.
- [14] Webster F. Traffic Signal Settings[J]. Road Research Technical Thesis, 1958.

- [15] Cools S B, Gershenson C, D'Hooghe B. Self-organizing traffic lights: A realistic simulation[J]. Advances in applied self-organizing systems, 2013: 45-55.
- [16] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518: 529-533.
- [17] Mnih V. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.
- [18] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575: 350-354.
- [19] Wang F y, Zhang J J, Zheng X, et al. Where does AlphaGo go: from church-turing thesis to AlphaGo thesis and beyond[J]. IEEE/CAA Journal of Automatica Sinica, 2016, 3(2): 113-120.
- [20] Savva M, Kadian A, Maksymets O, et al. Habitat: A Platform for Embodied AI Research [J]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 9338-9346.
- [21] Ray P P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope[J]. Internet of Things and CyberPhysical Systems, 2023.
- [22] Yang R, Sun X, Narasimhan K. A generalized algorithm for multi-objective reinforcement learning and policy adaptation[J]. Advances in neural information processing systems, 2019: 32-53.
- [23] Espeholt L, Soyer H, Munos R, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures[C]//International conference on machine learning. 2018: 1407-1416.
- [24] Koonce P, et al. Traffic signal timing manual[R]. United States. Federal Highway Administration, 2008.
- [25] Lowrie P. Scats, sydney coordinated adaptive traffic system: A traffic responsive method of controlling urban traffic[J]. 1990.
- [26] Lu S, Liu X, Dai S. Incremental multistep Q-learning for adaptive traffic signal control based on delay minimization strategy[C]//The 2008 7th World Congress on Intelligent Control and Automation. 2008: 2854-2858.
- [27] Chanloha P, Usaha W, Chinrungrueng J, et al. Performance Comparison between Queueing Theoretical Optimality and Q-Learning Approach for Intersection Traffi

- c Signal Control[C]// 2012 Fourth International Conference on Computational Intelligence, Modelling and Simulation, 2012: 172-177.
- [28] Joo H, Lim Y. Reinforcement Learning for Traffic Signal Timing Optimization[C]//2020 International Conference on Information Networking. 2020: 738-742
- [29] Chu T, Wang J, Codecà L, et al. Multi-Agent Deep Reinforcement Learning for LargeScale Traffic Signal Control[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(2): 1086-1095.
- [30] Wei H, Xu N, Zhang H, et al. Colight: Learning network-level cooperation for traffic signal control[C]//28th ACM international conference on information and knowledge management. 2019: 1913-1922.
- [31] Wei H, Chen C, Zheng G, et al. Presslight: Learning max pressure control to coordinate traffic signals in arterial network[C]//25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019: 1290-1298.
- [32] Xiong Y, Zheng G, Xu K, et al. Learning traffic signal control from demonstrations[C]//28th ACM international conference on information and knowledge management. 2019: 2289-2292.
- [33] Zheng G, Xiong Y, Zang X, et al. Learning phase competition for traffic signal control[C]//28th ACM international conference on information and knowledge management. 2019: 1963-1972.
- [34] Wei H, Zheng G, Yao H, et al. Intellilight: A reinforcement learning approach for intelligent traffic light control[C]//24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018: 2496-2505.
- [35] Varaiya P P. Max pressure control of a network of signalized intersections[J]. Transportation Research Part C-emerging Technologies, 2013, 36: 177-195.
- [36] Jiang H, Li Z, Li Z, et al. A General Scenario-Agnostic Reinforcement Learning for Traffic Signal Control[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(9): 11330-11344.
- [37] Jiang H, Li Z, Wei H, et al. X-Light: Cross-City Traffic Signal Control Using Transformer on Transformer as Meta Multi-Agent Reinforcement Learner[C]//the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. 2024: 94-102.

- [38] Jiang H, Xiong X, Li Z, et al. GuideLight: "Industrial Solution" Guidance for More Practical Traffic Signal Control Agents[J]. arXiv preprint arXiv:2407.10811, 2024.
- [39] Du X, Li Z, Long C, et al. FELight: Fairness-Aware Traffic Signal Control via Sample Efficient Reinforcement Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(9): 4678-4692.
- [40] Xing D, Zheng Q, Liu Q, et al. TinyLight: Adaptive Traffic Signal Control on Devices with Extremely Limited Resources[C]//the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. 2022: 3999-4005.
- [41] Gu Y, Zhang K, Liu Q, et al.  $\pi$ -light: Programmatic interpretable reinforcement learning for resource-limited traffic signal control[C]//AAAI Conference on Artificial Intelligence: vol. 38: 19. 2024: 21107-21115.
- [42] Zhang L, Wu Q, Shen J, et al. Expression might be enough: representing pressure and demand for reinforcement learning based traffic signal control[C]//International Conference on Machine Learning. 2022: 26645-26654.
- [43] Sahni H, Kumar S, Tejani F, et al. State Space Decomposition and Subgoal Creation for Transfer in Deep Reinforcement Learning[J]. arXiv preprint arXiv:1705.08997, 2017.
- [44] Castro P S. Scalable methods for computing state similarity in deterministic markov decision processes[C]//AAAI Conference on Artificial Intelligence: vol. 34: 0 6. 2020: 10069-10076
- [45] Sutton R, Barto A. Reinforcement Learning: An Introduction[J]. IEEE Transactions on Neural Networks, 1998, 9(5): 1054-1054.
- [46] March J G. Exploration and exploitation in organizational learning[J]. Organization science, 1991, 2(1): 71-87.
- [47] Kidambi R, Rajeswaran A, Netrapalli P, et al. MOReL: Model-Based Offline Reinforcement Learning[C]//Advances in Neural Information Processing Systems: vol. 33. 2020: 21810-21823.
- [48] Howard R A. Dynamic Programming and Markov Processes[J]. MIT Press google schola, 1960, 2: 39-47.
- [49] Watkins C, Dayan P. Q-learning[J]. Machine Learning, 1992, 8: 279-292
- [50] Hester T, Vecerik M, Pietquin O, et al. Deep q-learning from demonstrations[C]//AAAI conference on artificial intelligence: vol. 32: 1. 2018.

- [51] Hemmerling A. Labyrinth problems: Labyrinth-searching abilities of automata[M]. vol. 114. 1989.
- [52] Haapa-aho J, Korpela T, Björkqvist T, et al. Continuous Control Issues Concerning Operation Improvement of Small-Scale Biomass Boilers[J]. IFAC Proceedings Volumes, 2011, 44(1): 7035-7042.
- [53] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization[J]. CoRR, 2014.
- [54] Tieleman T, Hinton G. Rmsprop: Divide the gradient by a running average of its recent magnitude[J]. COURSERA Neural Networks Mach. Learn, 2012: 17-28.
- [55] Tolias G, Sicre R, Jégou H. Particular object retrieval with integral max-pooling of CNN activations[J]. CoRR, 2015.
- [56] Larionov A A, Krause A, Miller W. A standard curve-based method for relative real time PCR data processing[J]. BMC Bioinformatics, 2005, 6: 62-62.
- [57] Ng A Y. Feature selection, L 1 vs. L 2 regularization, and rotational invariance[C]//the twenty-first international conference on Machine learning. 2004: 78.
- [58] Osband I, Blundell C, Pritzel A, et al. Deep exploration via bootstrapped DQN[J]. Advances in neural information processing systems, 2016: 29-34.
- [59] Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients[C]//AAAI conference on artificial intelligence: vol. 32: 1. 2018.
- [60] Sutton R S, McAllester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation[J]. Advances in neural information processing systems, 1999, 12.
- [61] Schulman J, Wolski F, Dhariwal P, et al. Proximal Policy Optimization Algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.
- [62] Schulman J, Moritz P, Levine S, et al. High-Dimensional Continuous Control Using Generalized Advantage Estimation[J]. arXiv preprint arXiv:1506.02438, 2015.
- [63] 戚朕. 基于单智能体强化学习的交通信号控制方法研究与应用[D]. 南京理工大学, 2018.
- [64] Zhang K, Yang Z, Liu H, et al. Fully decentralized multi-agent reinforcement learning with networked agents[C]//International conference on machine learning. 2018: 5872-5881.

- [65] Zhao Q, Tong L, Swami A, et al. Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework[J]. IEEE Journal on Selected Areas in Communications, 2007, 25(3): 589-600.
- [66] Tampuu A, Matiisen T, Kodelja D, et al. Multiagent cooperation and competition with deep reinforcement learning[J]. PloS one, 2017, 12(4): e0172395.
- [67] Hausknecht M, Stone P. Deep recurrent q-learning for partially observable mdps[C]//2015 aaai fall symposium series. 2015.
- [68] Rashid T, Samvelyan M, De Witt C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[J]. Journal of Machine Learning Research, 2020, 21(178): 1-51.
- [69] Su J, Adams S C, Beling P A. Counterfactual Multi-Agent Reinforcement Learning with Graph Convolution Communication[J]. arXiv preprint arXiv:2004.00470, 2020.
- [70] El-Tantawy S, Abdulhai B, Abdelgawad H. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown Toronto[J]. IEEE transactions on Intelligent transportation systems, 2013, 14(3): 1140-1150.
- [71] Stevanovic A. Adaptive traffic control systems: domestic and foreign state of practice[M]. 2010.
- [72] Mao F, Li Z, Lin Y, et al. Mastering Arterial Traffic Signal Control With Multi-Agent Attention-Based Soft Actor-Critic Model[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(3): 3129-3144.
- [73] Ge H, Gao D, Sun L, et al. Multi-Agent Transfer Reinforcement Learning With MultiView Encoder for Adaptive Traffic Signal Control[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(8): 12572-12587.
- [74] Noaeen M, Naik A, Goodman L, et al. Reinforcement learning in urban network traffic signal control: A systematic literature review[J]. Expert Systems with Applications, 2022, 199: 116830.
- [75] Bouktif S, Cheniki A, Ouni A, et al. Deep reinforcement learning for traffic signal control with consistent state and reward design approach[J]. Knowledge-Based Systems, 2023, 267: 110440.
- [76] Kolat M, Kovári B, Bécsi T, et al. Multi-Agent Reinforcement Learning for Traffic Signal Control: A Cooperative Approach[J]. Sustainability, 2023, 15(4): 3479.

- [77] Zhang C, Tian Y, Zhang Z, et al. Neighborhood Cooperative Multiagent Reinforcement Learning for Adaptive Traffic Signal Control in Epidemic Regions[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12): 25157-25168.
- [78] 徐良杰, 王炜. 信号交叉口行人过街时间模型[J]. 交通运输工程学报, 2005, 5(1): 111-115.
- [79] Curran W J, Brys T, Taylor M E, et al. Using PCA to Efficiently Represent State Spaces [J]. arXiv preprint arXiv:1505.00322, 2015.
- [80] Ferns N, Castro P S, Precup D, et al. Methods for computing state similarity in Markov decision processes[C]//Twenty-Second Conference on Uncertainty in Artificial Intelligence. 2006: 174-181.
- [81] Ferns N, Panangaden P, Precup D. Metrics for Finite Markov Decision Processes [C]//UAI: vol. 4. 2004: 162-169.
- [82] Rubner Y, Tomasi C, Guibas L J. The Earth Mover's Distance as a Metric for Image Retrieval[J]. International Journal of Computer Vision, 2000, 40: 99-121.
- [83] Ferns N, Panangaden P, Precup D. Bisimulation metrics for continuous Markov decision processes[J]. SIAM Journal on Computing, 2011, 40(6): 1662-1714.
- [84] Behrisch M, Bieker L, Erdmann J, et al. SUMO—simulation of urban mobility: a n overview[C]//The Third International Conference on Advances in System Simulation. 2011.
- [85] Oroojlooy A, Nazari M, Hajinezhad D, et al. Attendlight: Universal attention-based reinforcement learning model for traffic signal control[J]. Advances in Neural Information Processing Systems, 2020, 33: 4079-4090.
- [86] Jiang Q, Li J, Sun W, et al. Dynamic Lane Traffic Signal Control with Group Attention and Multi-Timescale Reinforcement Learning[C]//International Joint Conference on Artificial Intelligence. 2021.
- [87] Huang X, Wu D, Jenkin M R M, et al. ModelLight: Model-Based Meta-Reinforcement Learning for Traffic Signal Control[C]//International Conference on Machine Learning. 2021.
- [88] Wang M, Wu L, Li J, et al. Traffic Signal Control with Reinforcement Learning Based on Region-Aware Cooperative Strategy[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(7): 6774-6785.

- [89] Wang Y, Xu T, Niu X, et al. STMARL: A Spatio-Temporal Multi-Agent Reinforcement Learning Approach for Cooperative Traffic Light Control[J]. IEEE Transactions on Mobile Computing, 2022, 21(6): 2228-2242.
- [90] Liang E, Su Z, Fang C, et al. OAM: An option-action reinforcement learning framework for universal multi-intersection control[C]//AAAI Conference on Artificial Intelligence: vol. 36: 4. 2022: 4550-4558.
- [91] Wu L, Wang M, Wu D, et al. Dynstgat: Dynamic spatial-temporal graph attention network for traffic signal control[J]. 2021: 2150-2159.
- [92] Xu B, Wang Y, Wang Z, et al. Hierarchically and cooperatively learning traffic signal control[C]//AAAI conference on artificial intelligence: vol. 35: 1. 2021: 669-677.
- [93] Zang X, Yao H, Zheng G, et al. Metalight: Value-based meta-reinforcement learning for traffic signal control[C]//AAAI conference on artificial intelligence: vol. 34: 01. 2020: 1153-1160.
- [94] Chen C, Wei H, Xu N, et al. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control[C]//AAAI conference on artificial intelligence: vol. 34: 04. 2020: 3414-3421.
- [95] Goel H, Zhang Y, Damani M, et al. SocialLight: Distributed Cooperation Learning towards Network-Wide Traffic Signal Control[C]//Adaptive Agents and Multi-Agent Systems. 2023: 1551-1559.
- [96] Yoon J, Ahn K, Park J, et al. Transferable traffic signal control: Reinforcement learning with graph centric state representation[J]. Transportation Research Part C-emerging Technologies, 2021, 130: 103321.
- [97] Carmona G, Podczeck K. On the existence of pure-strategy equilibria in large games [J]. Journal of Economic Theory, 2009, 144(3): 1300-1319.
- [98] Du S S, Kakade S M, Wang R, et al. Is a good representation sufficient for sample efficient reinforcement learning[J]. International Conference on Learning Representations, 2020.
- [99] Zeng J, Yu C, Yang X, et al. CityLight: A Universal Model Towards Real-world Cityscale Traffic Signal Control Coordination[J]. arXiv preprint arXiv:2406.02126, 2024.

- [100] Ruan J, Li Z, Wei H, et al. Coslight: Co-optimizing collaborator selection and decisionmaking to enhance traffic signal control[C]//The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024: 2500-2511.
- [101] Jiang H, Li Z, Wei H, et al. X-Light: Cross-City Traffic Signal Control Using Transformer on Transformer as Meta Multi-Agent Reinforcement Learner[J]. arXiv preprint arXiv:2404.12090, 2024.
- [102] Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition[C]//IEEE international conference on computer vision. 2015: 1449-1457.
- [103] Li M, Hu Z, Huang H, et al. A Hierarchical Spatio-Temporal Cooperative Reinforcement Learning Approach for Traffic Signal Control[C]//2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). 2022: 3411-3416.
- [104] Wang M, Xiong X, Kan Y, et al. UniTSA: A Universal Reinforcement Learning Framework for V2X Traffic Signal Control[J]. IEEE Transactions on Vehicular Technology, 2024, 73(10): 14354-14369.
- [105] Wang K, Shen Z, Lei Z, et al. Towards Multi-agent Reinforcement Learning based Traffic Signal Control through Spatio-temporal Hypergraphs[J]. arXiv preprint arXiv:2404.11014, 2024.
- [106] Lu J, Ruan J, Jiang H, et al. DuaLight: Enhancing Traffic Signal Control by Leveraging Scenario-Specific and Scenario-Shared Knowledge[C]//23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024. 2024: 1283-1291.
- [107] Huang X, Wu D, Boulet B. Traffic Signal Control Using Lightweight Transformers: An Offline-to-Online RL Approach[J]. ArXiv, 2023, abs/2312.07795.
- [108] Kővári B, Pelenczei B, Aradi S, et al. Reward Design for Intelligent Intersection Control to Reduce Emission[J]. IEEE Access, 2022, 10: 39691-39699.
- [109] Yang S, Yang B, Wong H S, et al. Cooperative traffic signal control using Multistep return and Off-policy Asynchronous Advantage Actor-Critic Graph algorithm [J]. Knowledge-Based Systems, 2019, 183-192.
- [110] Arjovsky M, Bottou L, Cho K, et al. Out of Distribution Generalization in Machine Learning[D]. New York University, 2020.

- [111] Liang S, Li Y, Srikant R. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks[C]/The 6th International Conference on Learning Representations, ICLR 2018.
- [112] Ada S E, Oztop E, Ugur E. Diffusion Policies for Out-of-Distribution Generalization in Offline Reinforcement Learning[J]. IEEE Robotics and Automation Letters , 2023, 9(4): 3116-3123.

## 附录

### 攻读硕士学位期间发表、录用和投稿论文列表

1. **Chao Wan**, Weiqiang Wu, Weibin Zhang, et al. State Design in Reinforcement Learning-Based Traffic Signal Control Using Similarity Metrics[C]//The 25th COTA International Conference of Transportation Professionals (CICTP 2025)(已收录)
2. Weibin Zhang, **Chao Wan** and Shoufeng Lu. State Encoding for Efficient Traffic Signal Control in High Volume[J]. Transportmetrica B: Transport Dynamics.(已投递)

### 攻读硕士学位期间受理专利列表

1. 惠铭宇, 张伟斌, 庄志洪, **万超**, 管毅诚, 王淇, 柳佳一. 基于图注意力机制和值分解强化学习的区域交通信号控制方法.申请号: CN202410237022.3 (已受理)

### 攻读硕士学位期间参与科研项目列表

1. 华为昇腾 AI 算子开发, 华为技术有限公司, 2022.9-2022.12
2. 国家自然科学基金面上项目, 基于交通因子状态网络的城市路网交通状态智能估计与预测方法研究, (No.71971116), 2022.10-2023.12