Computer Science and Application 计算机科学与应用, 2024, 14(4), 177-192
Published Online April 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.144088

Hans汉斯

基于混合域注意力的深度强化学习交叉口信号控制方法

李忠华[1]，何子登[2*]

[1]广东工业大学自动化学院，广东 广州
[2]广州羊城通有限公司大数据分公司，广东 广州

收稿日期：2024年3月12日；录用日期：2024年4月12日；发布日期：2024年4月19日

摘　要

针对强化学习智能体对微观交通状态感知能力有限的问题，本文提出了一种基于混合域注意力的深度强化学习交叉口信号控制算法3DQN_MDAM。首先，为减少存储开销，设计了一种轻量的混合域注意力模块(Mixed Domain Attention Module, MDAM)，仅使用少量的参数就能实现自适应地调整交通状态特征图中通道之间及空间位置之间权重的功能。然后，在现有基于双深度决斗Q网络(Double Dueling DQN, 3DQN)算法模型的基础上通过引入MDAM，使智能体自动地聚焦于对当前控制任务更为重要的交通状态信息，以增强智能体的状态感知能力。最后，利用仿真平台SUMO (Simulation of Urban Mobility)进行实验。实验结果显示，在低、中、高三种不同交通流条件下，3DQN_MDAM相比3DQN在各项指标上均得到改善，其中车辆平均等待时间分别缩短了20%、20%、17.6%。与其它常用的基准算法相比，3DQN_MDAM在各项指标上均得到最好的控制效果。

关键词

混合域注意力，深度强化学习，交通信号控制，3DQN算法，SUMO

An Intersection Signal Control Method with Deep Reinforcement Learning Based on Mixed Domain Attention

Zhonghua Li[1], Zideng He[2*]

[1]School of Automation, Guangdong University of Technology, Guangzhou Guangdong
[2]Big Data Branch, Guangzhou Yangchengtong Co., Ltd., Guangzhou Guangdong

Received: Mar. 12th, 2024; accepted: Apr. 12th, 2024; published: Apr. 19th, 2024

*通讯作者。

文章引用: 李忠华, 何子登. 基于混合域注意力的深度强化学习交叉口信号控制方法[J]. 计算机科学与应用, 2024, 14(4): 177-192. DOI: 10.12677/csa.2024.144088

Abstract

Aiming at the problem that reinforcement learning agents have limited perception of microscopic traffic conditions, an intersection signal control algorithm 3DQN_MDAM with deep reinforcement learning based on mixed domain attention is proposed. Firstly, to reduce storage overhead, a lightweight mixed domain attention module (MDAM) is designed, which can adaptively adjust the weights between channels and spatial positions in the traffic state feature map with only a small number of parameters. Then, based on the existing 3DQN algorithm model, by introducing MDAM, the agent automatically focuses on traffic status information that is more important to the current control task, in order to enhance the agent's state perception ability. Finally, experiments were conducted using the simulation platform SUMO (Simulation of Urban Mobility). The experimental results show that under three different traffic flow conditions of low, medium, and high, 3DQN_MDAM has improved in various indicators compared to 3DQN, with average waiting time of vehicles reduced by 20%, 20%, and 17.6%, respectively. Compared with other commonly used benchmark algorithms, 3DQN_MDAM achieved the best control effect on all indicators.

Keywords

Mixed Domain Attention, Deep Reinforcement Learning, Traffic Signal Control, 3DQN Algorithm, SUMO

1. 引言

随着经济的发展，汽车需求量爆发，特别是新能源汽车的推广，让各城市的汽车数量越来越多，交通拥堵问题也随之成为诸多城市的老大难问题，制约着城市的发展。解决交通拥堵问题，一方面要从道路的规划整改入手，另一方面，交叉路口的信号调度控制至关重要。

传统的交通信号灯控制方案根据历史交通需求数据预先确定相位的持续时间[1] [2]，或根据当前交通状况制定一套信号控制规则[3]，这些方法往往缺乏实时性，不能对交通需求的突发变化做出及时的响应。

强化学习(Reinforcement learning, RL)理论基于马尔可夫决策过程，非常适用于交通信号控制这样的序列决策问题。而传统的强化学习利用表格来记录每个状态－动作对的 Q 值以进行策略搜索，需要为每个可能的状态分配存储空间，但过大的状态空间在实际应用中往往会面临存储和计算资源的问题。随着物联网、5G 通信、大数据等技术的发展及其在交通领域的应用，交通数据的采集速度更快、数据种类更丰富、数据量更庞大，合理利用各种交通数据是缓解交通拥堵的关键。利用人工神经网络的深度学习可以将原始批量数据中底层不同类型的特征融合并抽象成高层特征，从而能有效处理种类丰富、数量庞大的数据[4]。深度强化学习(Deep Reinforcement Learning, DRL)是深度学习和强化学习的结合[5]，利用深度学习对环境进行状态感知，利用强化学习进行策略搜索和决策优化，在交通信号控制的研究领域得到广泛的应用。

现阶段，深度强化学习方法在单点交叉口信号控制问题上的研究已取得一定成果，但仍然存在研究的不足。现有研究多依赖于对交叉口的交通环境状态进行微观刻画，力图通过获得更加全面细致的环境状态信息以提升交通信号控制的效果。然而，一味地将这些高维的状态信息输入到神经网络中会让智能

体难以分辨出对当前控制任务更为重要的信息，削弱了智能体的状态感知与表达能力。近年来，注意力机制在计算机视觉领域得到了广泛应用，其作用在许多任务上得到验证[6]。注意力机制可自适应地在众多的输入信息中强调对当前任务更为有用的信息，降低对其它非重要信息的关注度。因此，为了解决上述问题，本文在基于 DRL 的单点交叉口信号控制方法中引入注意力机制以提高信号控制的效果。然而，现有的注意力机制大多参数量较大且较为复杂，这往往会增大交通信号控制器的存储开销且影响控制器的响应速度。为此，有必要设计一种更适用于交通信号控制场景的注意力模块。本文的主要贡献有：

1) 设计了一种轻量的混合域注意力模块 MDAM。该模块不仅能快速地构建交通状态特征图中通道之间特征信息的依赖关系，还能通过捕获特征图中各车道之间及距交叉口停止线不同距离位置之间车辆信息的相关性，高效地构建空间位置之间特征信息的依赖关系。

2) 提出了一种基于混合域注意力的深度强化学习单点交叉口控制算法 3DQN_MDAM。在现有 3DQN 算法网络结构的基础上进行改进，设计了基于 MDAM 的状态特征提取网络，智能体可以分别在通道域和空间域上自动地关注重要的交通状态特征，增强智能体对交通状态的感知与理解能力。

在模拟现实高峰车流的环境下，分别在低、中、高三种不同交通流量的条件下验证了本文所提模型的有效性，对工程应用具有指导意义。

2. 相关工作

强化学习在交通信号控制的场景中已得到了广泛的研究与应用。Abdulhai 等人[7]利用 Q-learning 算法控制具有两相位的单点交叉口信号，通过维护一张存储有每个状态－动作对的 Q 值的表格(Q 表格)以进行策略搜索。但由于实际交通环境具有复杂性和时变性的特点，当要利用更多的交通状态信息时，状态空间和动作空间的维度随之增加，此时维护一张 Q 表格需要花费更大的存储开销，且在策略搜索时消耗更长的时间。

深度强化学习使用神经网络拟合状态－动作对的 Q 值，较好地缓解了状态空间维度迅速增加所带来的问题。基于深度强化学习的交通信号控制问题已成为当今智能交通领域研究的热点。Li 等人[8]提出了一种基于 DQN 算法的交叉口信号控制方法，用深度神经网络去拟合强化学习的 Q 函数，使智能体学会根据交叉口的交通状态选择对应的控制策略，这是较早利用 DRL 控制单点交叉口信号的工作之一。在基于 DRL 的交叉口信号控制研究中，既有宏观的状态表示，其具有对交通流全局信息的估计，如车道上车辆的排队长度[9]、车辆的数量[10]；也有微观的状态表示，其具有针对单独车辆的原始数据的描述，如车辆的速度[11]、车辆的位置[11] [12]。Yi 等人[13]提出了一种用于交通信号控制的时空深度强化学习模型，其将交叉口各车道上的车辆数量、当前信号灯相位及相位的持续时间作为状态输入。为了全面微观地捕获交叉口的实时信息，一种离散交通状态编码(Discrete Traffic State Code, DTSE)技术被广泛使用[14] [15] [16] [17]，通过对交叉口各进口道路进行离散化状态编码，再利用卷积神经网络从原始的交通数据中提取交叉口实时的交通状态特征信息。Liang 等人[17]将交叉路口划分为网格，获取各网格上车辆的速度与位置信息，并利用 Dueling DQN 算法和 Double DQN 的思想，进行交通信号控制。An 等人[18]将交叉口的图像作为强化学习智能体的状态输入，以更加全面地获取道路交通状态信息。Hua 等人[19]将交叉口车辆排队长度、车辆数量、车辆位置图像等信息作为智能体的状态输入，同时兼顾了微观和宏观状态特征。

上述这些研究中较多用到交通状态的微观表示。相比之下，微观状态表示相对于宏观状态表示具有较好的性能增益[20]。然而，在众多微观的交通状态信息中，并非所有的信息对智能体的优化控制有着同等的重要性，一味地将这些高维的状态信息输入到神经网络中往往会让智能体忽略了其中重要的信息转而关注了非重要信息，这对智能体的控制任务是不利的，以往的研究很少关注这个问题。任等人[21]将交

叉口各进口道划分为长度均匀的网格，网格中的值为车辆数量，再将网格展平为 80 个通道，然后经过通道注意力机制处理，最后通过实验验证了其方法的有效性。然而，在应对微观的交通状态信息时，仅用通道注意力往往不能让智能体充分地学习关注重要的状态信息。本文提出了基于混合域注意力的深度强化学习交通信号控制方法，通过引入一种结合了通道注意力和空间注意力机制的混合域注意力模块 MDAM，使智能体更能充分地关注重要的交通状态信息。

3. 强化学习的要素定义

基于强化学习的交通信号控制问题可以描述为智能体如何根据交叉口环境输入的交通状态信息，通过最大化环境反馈的奖励，学习一个神经网络模型，输出最优的交通信号控制策略，以最小化所有车辆在交叉口上的行驶时间。本节将对交通信号控制中强化学习的三要素：状态、动作、奖励进行定义。

3.1. 状态表示

为了较为全面准确地刻画交通状态，验证所提方法的有效性，本文采用离散交通状态编码(DTSE)技术[14]进行状态表示。DTSE 是一种类似图像的、微观的交叉口道路状态表示方法。相较于以交叉口图像信息作为状态表示，DTSE 能大大降低状态空间维度，减少模型的计算量，从而加快信号控制器的响应速度。如图 1 所示，本文的状态表示过程可大致分为“网格化”、“编码”、“堆叠”三个阶段。

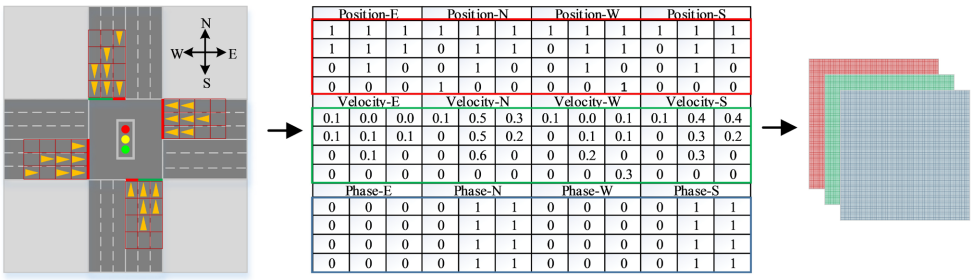


Figure 1. The schematic diagram of traffic state representation process
图 1. 交通状态表示过程示意图

在“网格化”阶段，从交叉口停止线到小于或等于相应进口道长度的检测范围 L 的路段上，将其划分为具有固定长度 l 的单元格，每个单元格长度略大于车辆长度，这样确保每个单元格内只能容纳一辆车，以跟踪个体车辆的状态信息，然后将各方向车道上的网格进行拼接。

在“编码”阶段，针对车辆的位置、速度以及当前交叉口的相位这三个刻画交通状态的变量分别提取三个矩阵。对于每一个单元格，在采样时间步 t，如果有车辆存在，则设置对应单元格的值为 1，否则为 0，形成位置矩阵；同理，设置对应单元格的值为归一化速度，形成速度矩阵；如果控制当前车道的交通灯为绿色放行状态，则当前车道上所有的单元格的值设置为 1，否则为 0，形成相位矩阵。

在“堆叠”阶段，将上述三个特征矩阵堆叠形成形状大小为 $H \times W \times C$ 的交通状态特征图，其中 $H = L/l$ 为每条车道的单元格数目，W 为交叉口各方向进口道的车道总数，C 表示刻画交通状态的变量数目。

3.2. 动作表示

智能体的动作空间由交叉口采用的相位方案决定。本文采用具有四相位的信号控制方案，四个相位分别表示南北方向直行和右转(S→NE, N→SW)、南北方向左转(S→W, N→E)、东西方向直行和右转(E→WN, W→ES)、东西方向左转(E→S, W→N)。此时动作空间表示为

$$A=\left\{\left(\mathrm{S}\rightarrow\mathrm{NE},\mathrm{N}\rightarrow\mathrm{SW}\right),\left(\mathrm{S}\rightarrow\mathrm{W},\mathrm{N}\rightarrow\mathrm{E}\right),\left(\mathrm{E}\rightarrow\mathrm{WN},\mathrm{W}\rightarrow\mathrm{ES}\right)\right\}$$

智能体每次执行动作时，从动作空间中选择一个相位，该相位的持续时间为 T_g 秒。如果智能体在下一个控制步执行的动作与当前相同，则继续保持当前相位 T_g 秒，否则需执行黄灯相位 T_y 秒，以保证行车安全。

3.3. 奖励函数

在基于 DRL 的交通信号控制问题中，通常旨在最小化所有车辆在交叉口上的行驶时间。然而，这很难被直接优化，因为以所有车辆行驶时间定义的奖励是一种延迟奖励，智能体在延迟奖励面前难以学习[22]。因此，奖励函数通常由车辆等待时间、车辆排队长度、车辆数量等指标来定义，也有由这些指标进行加权组合的定义，以间接优化所有车辆在交叉口的行驶时间。指标的加权组合虽然能综合优化各指标，但同时引入了过多的超参数。这些超参数要凭经验设置，无疑增大了实际操作的难度。本文采用相邻采样时间步的所有车辆排队长度之差来定义奖励函数。

$$r_t = q_t - q_{t+1} \tag{1}$$

式(1)中，r_t 为第 t 采样时间步的即时奖励，q_t 为第 t 采样时间步交叉口各车道车辆排队长度的总和。$r_t>0$ 时，表明当前采样时间步交叉口排队长度比上一个采样时间步少，交通拥堵状况得到改善，反之亦然。智能体通过最大化奖励，优化自身的交通信号控制策略，以改善交叉口的拥堵状况。

4. 算法

本文设计了一种混合域注意力模块 MDAM，并将此模块引入到现有的单点交叉口信号控制算法模型中，提出了一种基于混合域注意力的深度强化学习交叉口信号控制算法 3DQN_MDAM。本节将分别介绍 MDAM 和 3DQN_MDAM 的网络结构及原理，以及 3DQN_MDAM 的训练流程。

4.1. 混合域注意力模块 MDAM

根据车辆在交叉口分布的特点及交叉口信号控制任务中对控制器实时响应的需求，加之文献[23] [24] [25] [26]的启发，本文设计了一种轻量化的混合域注意力模块。该模块由通道域注意力和空间域注意力两个子模块串联构成，整体结构如图 2 所示。

在通道域注意力子模块中，并行进行全局平均池化(global average pooling)和全局最大池化(global max pooling)，以减少仅利用单一池化造成的信息丢失[23]。它们分别生成一个描述通道域信息的通道特征描述子，根据文献[24]，此过程的计算表达式为

$$z_{avg(c)}^{channel} = \frac{1}{HW}\sum_{j=1}^{H}\sum_{i=1}^{W} x_c\left(i,j\right) \tag{2}$$

$$z_{max(c)}^{channel} = \max_{0<i\le W,0<j\le H} x_c\left(i,j\right) \tag{3}$$

式(2)、(3)中，$z_{avg(c)}^{channel}$ 和 $z_{max(c)}^{channel}$ 分别表示第 c 通道的平均池化和最大池化特征描述子。H 和 W 表示输入特征图 $F=\left[x_1,\cdots,x_c,\cdots,x_C\right]$ $\left(F\in\mathbb{R}^{H\times W\times C}\right)$的长和宽。

将上述的平均池化和最大池化特征描述子分别输入一个共享的一维卷积(Conv1d)，利用快速的一维卷积实现不同通道之间的信息交互[25]，再将此时得到的两个 $1\times1\times C$ 的特征图的信息通过逐元素相加的方式融合，最后通过归一化得到通道注意力系数，此过程的计算表达式为

$$A_c = \sigma\left(f_{conv1d}^{k}\left(z_{avg}^{channel}\right)+f_{conv1d}^{k}\left(z_{max}^{channel}\right)\right) \tag{4}$$

式(4)中，$A_c \in \mathbb{R}^{1\times1\times C}$ 为通道注意力系数张量，σ 表示 *sigmoid* 激活函数，f_{conv1d}^{k} 表示卷积核个数为 k 的一维卷积操作(本文 k 取 3)，$z_{avg}^{channel} \in \mathbb{R}^{1\times1\times C}$ 和 $z_{max}^{channel} \in \mathbb{R}^{1\times1\times C}$ 分别表示由通道特征描述子构成的平均池化和最大池化通道特征描述张量。

通道域注意力子模块的最终输出为

$$F' = F \otimes A_c \tag{5}$$

式(5)中，$F' \in \mathbb{R}^{H\times W\times C}$ 表示从通道域上重构特征的特征图，$\otimes$表示逐元素相乘。

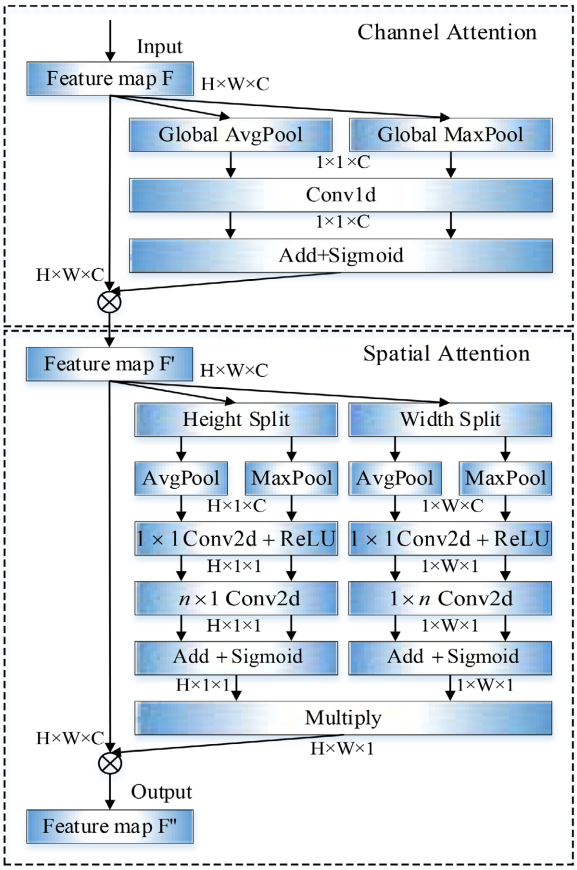


Figure 2. The mixed domain attention module (MDAM)
图 2. 混合域注意力模块(MDAM)

空间注意力通常是对空间上每个位置的特征信息进行描述，以此捕捉位置之间特征信息的相关性，从而构建空间位置上的注意力系数[23]。然而，描述交叉口车辆状态信息的位置矩阵和速度矩阵是稀疏的[27]，通过描述稀疏矩阵上每个位置(网格)的特征信息来构建空间注意力系数并不高效。考虑到车辆在交叉口上的分布具有“条状”的特点，同时受到文献[26]的启发，本文通过捕捉各车道之间及距交叉口停止线不同距离位置之间交通状态特征信息的相关性来构建交叉口不同网格位置的空间注意力系数。为此，在本文的空间域注意力子模块中，首先将输入特征图 F' 分别沿着高度和宽度方向划分成 H 块形状大小为 $1\times W\times C$ 和 W 块形状大小为 $H\times1\times C$ 的“薄片”，然后分别在各自划分方向上学习每块“薄片”特征信息的相关性系数，最后，空间每个位置的注意力系数则是该位置所在的两个划分方向“薄片”特征信息的相关性系数的乘积。具体地，在每块“薄片”的每个通道上并行进行平均池化和最大池化，生成描述每块“薄片”中每个通道特征信息的特征描述子，根据文献[26]，此过程的计算表

达式为

$$z_{avg(h,c)}^{height}=\frac{1}{W}\sum_{i=1}^{W}y_c(h,i) \tag{6}$$

$$z_{max(h,c)}^{height}=\max_{0<i\le W}y_c(h,i) \tag{7}$$

$$z_{avg(w,c)}^{width}=\frac{1}{H}\sum_{j=1}^{H}y_c(j,w) \tag{8}$$

$$z_{max(w,c)}^{width}=\max_{0<j\le H}y_c(j,w) \tag{9}$$

式(6)~(9)中，$z_{avg(h,c)}^{height}$ 和 $z_{max(h,c)}^{height}$ 分别表示沿着高度方向划分的第 h 块“薄片”中第 c 通道的平均池化和最大池化特征描述子，$z_{avg(w,c)}^{width}$ 和 $z_{max(w,c)}^{width}$ 分别表示沿着宽度方向划分的第 w 块“薄片”中第 c 通道的平均池化和最大池化特征描述子，H 和 W 表示输入特征图 $F'=[y_1,\cdots,y_c,\cdots,y_C]$ ($F'\in\mathbb{R}^{H\times W\times C}$)的高和宽。

各“薄片”中各通道的平均池化和最大池化特征描述子经过共享的卷积核为 1 × 1 的二维卷积进行降维，同时将各通道特征描述子的信息进行融合，形成了描述各“薄片”特征信息的特征描述子。此过程的计算表达式为

$$\overline{z}_{avg}^{height}=\delta\left(f_{conv2d}^{1\times1}\left(z_{avg}^{height}\right)\right) \tag{10}$$

$$\overline{z}_{max}^{height}=\delta\left(f_{conv2d}^{1\times1}\left(z_{max}^{height}\right)\right) \tag{11}$$

$$\overline{z}_{avg}^{width}=\delta\left(f_{conv2d}^{1\times1}\left(z_{avg}^{width}\right)\right) \tag{12}$$

$$\overline{z}_{max}^{width}=\delta\left(f_{conv2d}^{1\times1}\left(z_{max}^{width}\right)\right) \tag{13}$$

式(10)~(13)中，$\overline{z}_{avg}^{height}$、$\overline{z}_{max}^{height}$、$\overline{z}_{avg}^{width}$、$\overline{z}_{max}^{width}$ ($\overline{z}_{avg}^{height}$、$\overline{z}_{max}^{height}\in\mathbb{R}^{H\times1\times1}$，$\overline{z}_{avg}^{width}$、$\overline{z}_{max}^{width}\in\mathbb{R}^{1\times W\times1}$)分别表示由沿着高度和宽度方向划分的各“薄片”特征描述子构成的平均池化和最大池化特征描述张量，z_{avg}^{height}、z_{max}^{height}、z_{avg}^{width}、z_{max}^{width} (z_{avg}^{height}、$z_{max}^{height}\in\mathbb{R}^{H\times1\times C}$，$z_{avg}^{width}$、$z_{max}^{width}\in\mathbb{R}^{1\times W\times C}$)分别是由 $z_{avg(h,c)}^{height}$、$z_{max(h,c)}^{height}$、$z_{avg(w,c)}^{width}$、$z_{max(w,c)}^{width}$ 构成的张量，$f_{conv2d}^{1\times1}$ 表示卷积核为 1 × 1 的二维卷积操作(不同划分方向不共享卷积参数)，δ 表示 *Relu* 激活函数。

利用两个轻量的“条状”卷积分别对两个划分方向上各“薄片”的特征描述子进行相关性建模，以捕获“薄片”之间特征信息的依赖关系，由平均池化和最大池化所描述的特征信息通过逐元素相加的方式融合，此过程的计算表达式为

$$A_h=\sigma\left(f_{conv2d}^{n\times1}\left(\overline{z}_{avg}^{height}\right)+f_{conv2d}^{n\times1}\left(\overline{z}_{max}^{height}\right)\right) \tag{14}$$

$$A_w=\sigma\left(f_{conv2d}^{1\times n}\left(\overline{z}_{avg}^{width}\right)+f_{conv2d}^{1\times n}\left(\overline{z}_{max}^{width}\right)\right) \tag{15}$$

式(14)、(15)中，$A_h\in\mathbb{R}^{H\times1\times1}$ 和 $A_w\in\mathbb{R}^{1\times W\times1}$ 分别表示沿着高度和宽度方向划分的“薄片”特征信息的相关性系数张量，$f_{conv2d}^{n\times1}$ 和 $f_{conv2d}^{1\times n}$ 分别表示卷积核为 $n\times1$ 和 $1\times n$ 的二维卷积(本文 n 取 5)，σ 表示 *sigmoid* 激活函数。

此时，空间位置上的注意力系数可通过式(16)计算，其计算表达式为

$$A_s=A_hA_w \tag{16}$$

式(16)中，$A_s\in\mathbb{R}^{H\times W\times1}$ 表示空间注意力系数矩阵。

空间域注意力子模块的最终输出为

$$F''=F'\otimes A_s \tag{17}$$

式(17)中，$F'' \in \mathbb{R}^{H\times W\times C}$ 表示从空间域上重构特征的特征图，$\otimes$表示逐元素相乘。

4.2. 算法 3DQN_MDAM 的网络结构

3DQN_MDAM 的网络结构如图 3 所示，其由基于 MDAM 的状态特征提取网络和基于 Dueling DQN 的状态－动作价值评估网络构成。状态特征提取网络负责提取交叉口车辆的实时状态特征信息，状态－动作价值评估网络则评估在当前状态下动作空间中所有动作的价值，通过贪婪地选择价值最大的动作作为当前状态的输出。

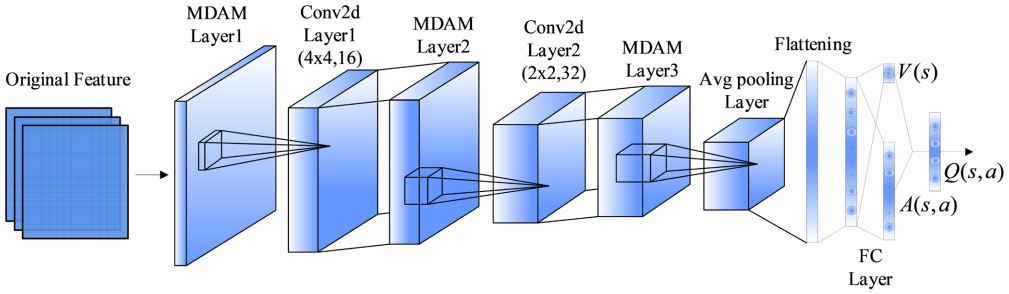


Figure 3. The structure of the 3DQN_MDAM network
图 3. 3DQN_MDAM 网络结构

4.2.1. 基于 MDAM 的状态特征提取网络

基于 MDAM 的状态特征提取主要通过注意力层 MDAM Layer 和卷积层 Conv2d Layer 实现。在状态特征提取网络中，原始交通状态特征图通过注意力层 MDAM Layer1 对原始特征分别在通道域和空间域上进行重构。一方面，利用 MDAM Layer1 的通道注意力子模块可以权衡原始状态特征图中位置、速度、相位三种不同特征信息的输入，捕捉通道之间特征信息的依赖关系；另一方面，利用 MDAM Layer1 的空间注意力子模块则可以学习原始状态特征图中各车道之间及距交叉口停止线不同距离位置之间车辆信息的相关性，捕捉不同网格位置特征信息的依赖关系。MDAM Layer1 输出的特征图利用卷积层 Conv2d Layer1，对重构后的特征进行特征提取。然后，再先后通过注意力层 MDAM Layer2、卷积层 Conv2d Layer2、注意力层 MDAM Layer3，对更深层抽象的特征分别进行重构、提取、重构。原始交通状态特征经过多次的重构与提取，可以增强网络的表达能力，使智能体自适应的关注对当前控制任务更为关键的特征信息。接着，利用平均池化层 Avg pooling Layer 进行下采样，在保留重构后的特征信息的同时，可有效减少模型参数量并提高模型的泛化能力。最后，经过 Flattening 对最终输出的高层特征进行展平得到形状大小为 M 的抽象特征。该特征提取网络从输入到输出的映射关系可表达为

$$\hat{s}_t = f_{MDAM\_Conv}\left(s_t;\theta\right) \tag{18}$$

式(18)中，$s_t \in \mathbb{R}^{H\times W\times C}$ 为采样时间步 t 时的原始状态，$\hat{s}_t \in \mathbb{R}^{M}$ 为原始状态经特征提取网络后得到的隐藏状态，θ 为特征提取网络的参数。

4.2.2. 基于 Dueling DQN 的状态－动作价值评估网络

状态－动作价值评估采用 Dueling DQN 技术[28]，可以更好地估计不同动作对于状态的贡献，加速收敛，提升学习效果。在基于 Dueling DQN 的评估网络中，全连接层 FC Layer 将特征提取网络提取的抽象特征信息分为状态价值流和动作优势流两条支路，分别输出一个标量 $V(s)$ (状态价值函数)和一个 $|A|$ 维的向量 $\bar{A}(s,a)$ (动作优势函数)。根据公式(19)将两条支路信息聚合得到状态－动作价值函数(Q 函数)。

$$Q\left(\hat{s}_t,a_t;\beta,\omega\right)=V\left(\hat{s}_t;\beta\right)+\bar{A}\left(\hat{s}_t,a_t;\omega\right)-\frac{1}{|A|}\sum_{a'}\bar{A}\left(\hat{s}_t,a';\omega\right) \tag{19}$$

式(19)中，β、ω 分别表示状态价值流和动作优势流的参数，a_t、$a'\in A$ 表示动作。

4.3. 算法 3DQN_MDAM 的训练流程

为了缓解训练过程中状态－动作价值(Q 值)被高估的问题，还采用了 Double DQN 技术[29]，通过引入目标 Q 值网络，将动作选择与动作 Q 值计算这两个过程解耦，使这两个过程分别由当前 Q 网络和目标 Q 网络完成，大大降低在训练过程中 Q 值被高估的概率。此时，根据文献[21] [30]，3DQN_MDAM 的损失函数为

$$L\left(\Theta\right)=E_{\left(s_t,a_t,r_t,s_{t+1}\right)\sim B}\left[\left(r_t+\gamma Q\left(s_{t+1},a^*;\bar{\Theta}\right)-Q\left(s_t,a_t;\Theta\right)\right)^2\right] \tag{20}$$

式(20)中，$\bar{\Theta}=\left\{\bar{\theta},\bar{\beta},\bar{\omega}\right\}$ 和 $\Theta=\left\{\theta,\beta,\omega\right\}$ 分别表示目标网络和当前网络的参数，r_t 为即时奖励值，γ 为折扣系数，s_t、s_{t+1} 分别表示当前采样时间步和下一采样时间步的原始状态，$a^*=\arg\max_{a'}Q\left(s_{t+1},a';\Theta\right)$ 表示当前网络在状态 s_{t+1} 下选择最大 Q 值所对应的动作。

参考[21]，得到如图 4 所示的 3DQN_MDAM 算法的训练流程。模型的训练与环境的交互交替进行。具体地，智能体每次与环境完成一个回合的交互后，再利用交互得到的数据对模型的参数进行 N 次的训练迭代，如此反复，保证经验回放缓冲区里的数据得到更新，并使模型得到充分的训练。在训练的过程中，3DQN_MDAM 从经验回放缓冲区 B 中随机抽取小批量经验数据进行训练，利用损失函数对 Θ 求梯度以更新 Q 网络的参数。

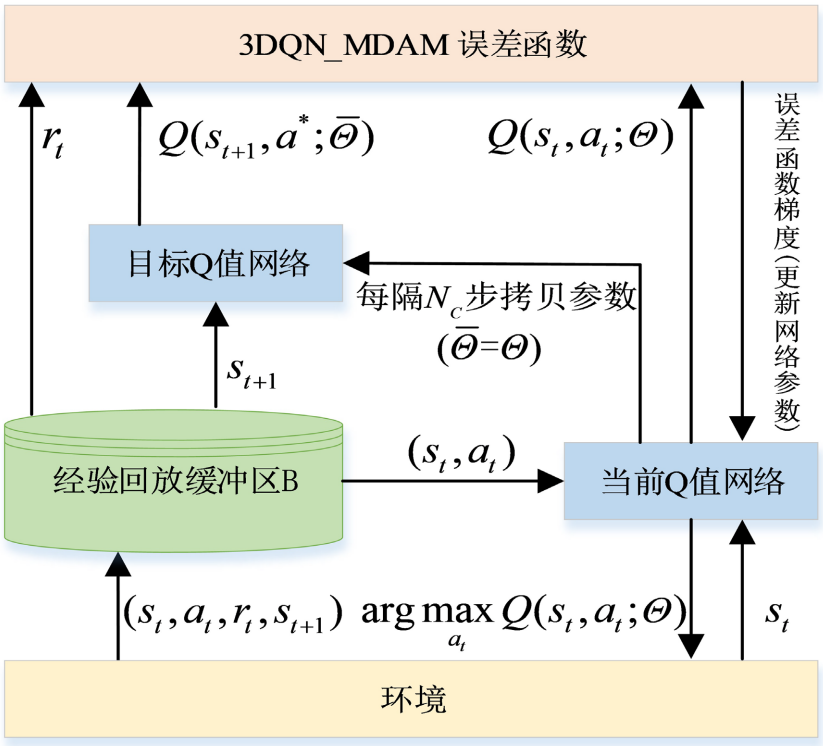


Figure 4. The training process of 3DQN_MDAM algorithm
图 4. 3DQN_MDAM 算法训练流程

3DQN_MDAM 详细的训练过程见表 1。

Table 1. The training process of 3DQN_MDAM algorithm
表 1. 3DQN_MDAM 算法训练流程

算法 1　3DQN_MDAM 算法训练流程
输入：采样时间步 t 交叉口的原始交通状态 s_t
初始化：当前 Q 网络参数 $\Theta=\{\theta,\beta,\omega\}$、目标 Q 网络参数 $\overline{\Theta}=\{\overline{\theta},\overline{\beta},\overline{\omega}\}$ （$\overline{\Theta}=\Theta$）；
for $e\in\{1,2,\cdots,episodes\}$ do
重置交叉口环境；
for $t\in\{1,2,\cdots,T/T_g\}$ do
智能体观测交叉口环境获得当前状态 s_t；
s_t经过嵌有 MDAM 的特征提取网络以构建状态特征图的通道依赖和空间依赖关系，利用式(18)获得隐藏状态 $\hat{s}_t$；
隐藏状态 $\hat{s}_t$ 经过策略评估网络，利用式(19)计算当前 Q 值 Q_t；
以 ε 的概率从动作空间 A 中随机选择一个动作 a_t，否则取 $a_t=\arg\max_{a'}Q_t$；
智能体执行当前动作 a_t 后，进入下一个状态 s_{t+1}，并由式(1)计算环境当前反馈的奖励 r_t，将序列 (s_t,a_t,r_t,s_{t+1}) 存放在经验回放缓冲区 B 中；
if $\lvert B\rvert\geq N_B$ then 溢出旧的经验数据；
end for
for $n\in\{1,2,\cdots,N\}$ do
从缓冲区 B 中随机抽取一小批量经验数据，利用式(20)计算损失函数 $L(\Theta)$；
更新当前 Q 网络参数 $\Theta=\Theta-\alpha\times\nabla_{\Theta}L(\Theta)$；
if $n\%N_c==0$ then $\overline{\Theta}=\Theta$；
end for
end for
输出：更新后的 Q 网络参数 Θ

5. 实验设置与结果分析

本节首先介绍实验设置，包括交叉口设置、车流生成设置、实验参数设置，然后介绍对比算法和算法的评价指标，最后在单点交叉口场景不同交通流量下与基准算法进行对比，以验证本文方法的有效性。

5.1. 实验设置

5.1.1. 交叉口设置

本文利用 SUMO 交通仿真软件搭建单点交叉口实验环境。交叉口由东南西北四个朝向的道路连接组成，四条道路长度均为 750 米，每条道路分别有 4 条进车道和 4 条出车道，最左侧为左转车道，最右侧为直行右转混合车道，中间两条为直行车道。

5.1.2. 车流生成设置

为了模拟现实交通流的高峰情况，本文设定车辆的产生服从 Weibull 分布，其概率密度函数为

$$f(x;\lambda,k)=\begin{cases}\dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1}\mathrm{e}^{-(x/\lambda)^k}, & x\geq 0\\ 0, & x<0\end{cases} \tag{21}$$

式(21)中，x 为随机变量，λ 为比例参数，k 为形状参数，本文取 $\lambda=1, k=2$。

低、中、高三种不同车流条件下，由 Weibull 产生的车流分布如图 5 所示。随着时间的推移，车辆产生的频率快速上升，到达高峰之后频率逐渐降低。不失一般性，我们将入口道路车辆的直行概率设为75%，将左转和右转概率均设为 12.5%。

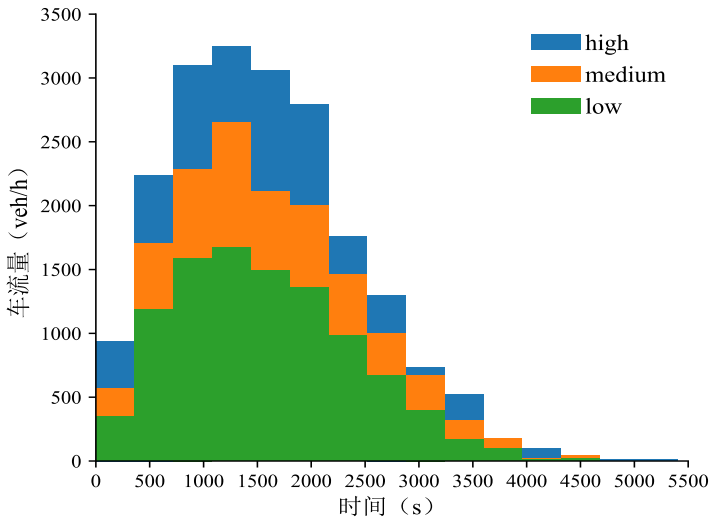


Figure 5. The histogram of traffic flow distribution
图 5. 车流分布直方图

5.1.3. 实验参数设置

本文采用 Adam 优化器进行模型训练。为让智能体在训练过程中得到充分的探索，本文分别设置了最大探索率 $\varepsilon_{\max}$、最小探索率 $\varepsilon_{\min}$。探索率 ε 随着训练回合数 e 的增加而减小，其计算表达式为

$$\varepsilon=\max\left(0.95^{e}\varepsilon_{\max},\varepsilon_{\min}\right) \tag{22}$$

实验相关的参数设置见表 2。

Table 2. Experimental parameter settings
表 2. 实验参数设置

参数	取值
训练总回合数 *episodes*	100
每回合训练迭代次数 N	800
每回合仿真时长 T	5400 s
学习率 α	0.001
最大、最小探索率 ε_{max}、ε_{min}	0.8、0.1
批量大小	128
折扣系数 γ	0.75
经验回放缓冲区最大容量 N_B	50,000
目标网络更新频率 N_c	5
绿灯相位 T_g、黄灯相位 T_y	10 s、3 s
网格长度 l、检测范围 L	7 m、280 m

5.2. 评价指标

本文采用以下四项指标来评价算法的性能：

a) 平均等待时间(Average Waiting Time, AWT)：所有车辆在交叉口道路上等待时间的平均值。

b) 平均旅行时间(Average Traveling Time, ATT)：所有车辆从驶入交叉口进口道到驶离交叉口出口道所用时间的平均值。

c) 平均等待次数(Average Waiting Count, AWC)：所有车辆在交叉口道路上等待次数的平均值。

d) 平均氮氧化物排放(NOx)：所有车辆从驶入交叉口进口道到驶离交叉口出口道所排放氮氧化物的平均值。

5.3. 对比算法

为了验证所提模型的有效性，本文使用以下三种方法作为基准方法。

a) 固定配时控制(Fixed Time) [2]：它预先确定相位的顺序和持续时间，由于其具备实现简单、稳定可靠的特点被广泛应用于现实交通信号控制的场景中。

b) 最大压力控制(Max Pressure) [31]：它是通过贪婪地选择具有最大压力的相位来优化交通灯配时的信号控制方法。

c) 基于 3DQN 的交通信号控制(3DQN) [17]：它使用与本文所提 3DQN_MDAM 模型相同的强化学习算法 3DQN，唯一的不同在于其不引入混合域注意力模块。

此外，本文还比较了所提出的 3DQN_MDAM 算法模型的以下变型：

d) 3DQN_MDAM_C：它不使用 MDAM 模块中的空间域注意力子模块，而仅使用通道域注意力子模块学习交通状态特征图中通道间的依赖关系。

e) 3DQN_MDAM_S：它不使用 MDAM 模块中的通道域注意力子模块，而仅使用空间域注意力子模块学习交通状态特征图中空间位置的相关性。

5.4. 实验结果与分析

高车流量条件下，3DQN_MDAM 与 3DQN 在训练过程中的累积奖励变化曲线如图 6(a)所示，车辆平均等待时间变化曲线如图 6(b)所示。从这两幅图可以看出，随着模型训练回合的增加，智能体通过回放缓冲区中的经验数据多次迭代更新其参数，最终累积奖励曲线和平均等待时间曲线收敛到一定水平。由于不同的训练回合使用的随机种子不同，以及在训练后期智能体还保留着 10%的探索率，累积奖励曲线和平均等待时间曲线收敛后还会存在小范围内的波动。图 6(c)和图 6(d)分别是由图 6(a)和图 6(b)的曲线经过滑动平均处理后得到的较为平滑的曲线。从这两幅图可以较为直观地看出，3DQN_MDAM 的累积奖励曲线和平均等待时间曲线均比 3DQN 的曲线收敛到更优的水平，这表明 3DQN_MDAM 的性能更好。

图 7(a)~(c)分别展示了在低、中、高三种车流量条件下，3DQN_MDAM 与基准算法在 20 组测试中的车辆平均等待时间对比结果。结果显示，与传统的 FixedTime 相比，基于强化学习的信号控制能更显著地缩短车辆在交叉口的等待时间；与 3DQN 和 MaxPressure 相比，3DQN_MDAM 的信号控制效果更好。

上述通过实验结果曲线图的定性分析表明了本文所提方法的有效性，实验结果的定量分析如表 3 所示。与 3DQN 相比，3DQN_MDAM 在低、中、高三种不同交通流量条件下车辆平均等待时间分别减少了 20%、20%、17.6%；与 MaxPressure 相比，3DQN_MDAM 在低、中、高流量条件下车辆的平均等待时间分别减少了 20%、15.6%、3.7%；在另外三项指标中，3DQN_MDAM 相比基准算法有明显的改善，三项指标均为最优。

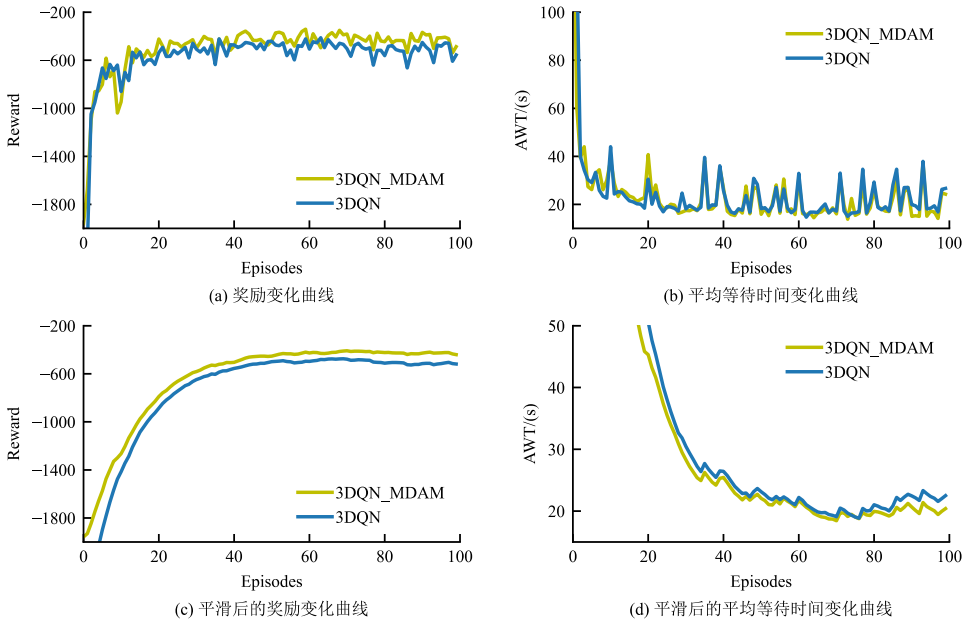
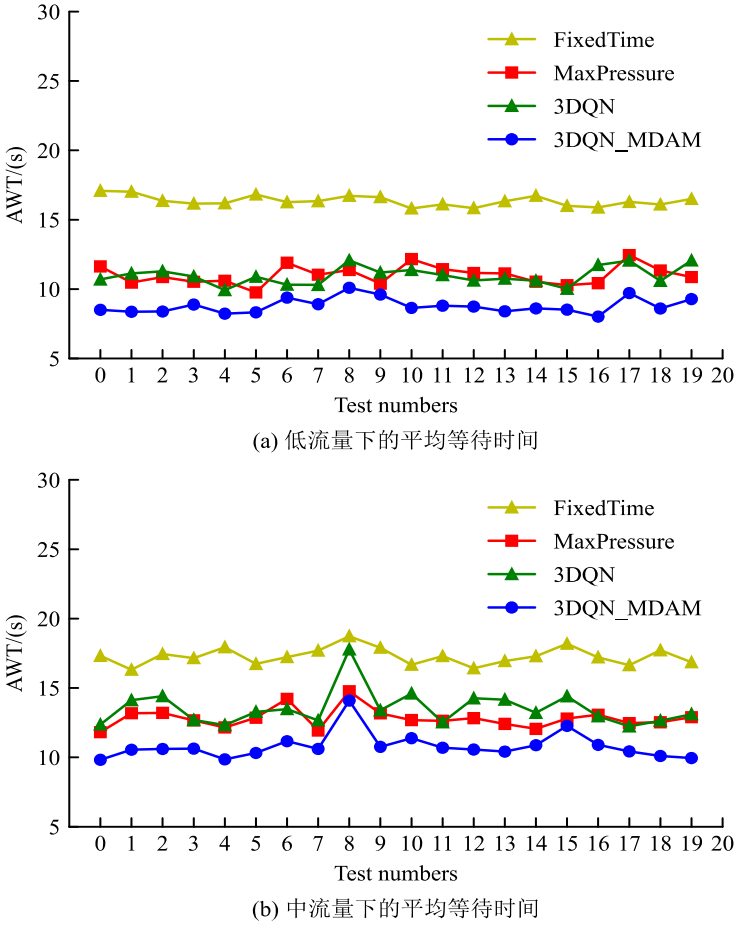
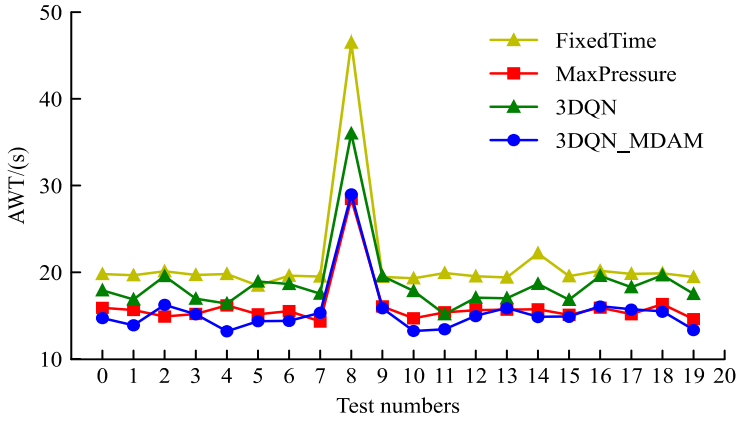


Figure 6. Comparison of reward curves and average waiting time curves of algorithms during training
图 6. 训练过程中算法的奖励及平均等待时间曲线对比





(c) 高流量下的平均等待时间

Figure 7. Comparison of average waiting time of algorithms in testing under different traffic conditions
图 7. 不同流量下算法在测试中的平均等待时间对比

Table 3. Comparison of evaluation indicators for algorithms
表 3. 算法的评价指标对比

算法	低流量				中流量				高流量			
	AWT /s	AWC	ATT /s	NOx /mg	AWT /s	AWC	ATT /s	NOx /mg	AWT /s	AWC	ATT /s	NOx /mg
FixedTime	16.4	0.73	139.2	121.7	17.3	0.76	141.1	123.8	21.1	0.86	146.5	130.2
MaxPressure	11.0	0.63	132.7	113.5	12.8	0.72	136.0	117.6	16.1	0.84	140.9	123.5
3DQN	11.0	0.57	132.0	112.4	13.5	0.64	135.9	117.1	18.8	0.76	142.9	125.5
3DQN_MDAM	**8.8**	**0.56**	**129.8**	**109.8**	**10.8**	**0.62**	**133.1**	**113.8**	**15.5**	**0.74**	**139.5**	**121.4**

5.5. 消融实验

为了进一步分析验证混合域注意力 MDAM 模块中通道域注意力子模块和空间域注意力子模块的作用，本文通过 3DQN_MDAM 的两个变型 3DQN_MDAM_C 和 3DQN_MDAM_S 进行消融实验，实验结果如表 4 所示。

Table 4. Results of ablation experiment
表 4. 消融实验结果

算法	低流量				中流量				高流量			
	AWT /s	AWC	ATT /s	NOx /mg	AWT /s	AWC	ATT /s	NOx /mg	AWT /s	AWC	ATT /s	NOx /mg
3DQN	11.0	0.57	132.0	112.4	13.5	0.64	135.9	117.1	18.8	0.76	142.9	125.5
3DQN_MDAM_C	9.6	0.58	130.8	111.1	11.8	0.63	134.1	115.0	17.0	0.74	141.0	123.3
3DQN_MDAM_S	9.5	**0.54**	130.3	110.2	11.5	**0.61**	133.6	114.2	17.1	**0.73**	141.0	123.2
3DQN_MDAM	**8.8**	0.56	**129.8**	**109.8**	**10.8**	0.62	**133.1**	**113.8**	**15.5**	0.74	**139.5**	**121.4**

从表 4 可看出，与原 3DQN 算法模型相比，本文提出的 3DQN_MDAM 及其变型 3DQN_MDAM_C、3DQN_MDAM_S 在多项指标上都得到改善，其中 3DQN_MDAM 尤为明显。这表明仅通过引入通道域注意力子模块捕捉交通状态特征图的通道依赖关系或仅通过引入空间域注意力子模块学习交通状态特征图

中空间位置的相关性都可以提升交通信号控制的效果，而同时使用这两个模块对提升交通信号控制的效果更为显著。

表 5 是引入各注意力子模块后模型参数量增加的百分数与高流量条件下车辆的平均等待时间缩短百分数之间的对比。从表 5 可看出，在原 3DQN 模型中引入各注意力子模块后，只增加了极少量的参数，车辆平均等待时间就比原来有显著的缩短。这表明本文设计的混合域注意力模块 MDAM 是轻量的，同时验证了本文所提模型 3DQN_MDAM 的有效性。

Table 5. Comparison between the percentage increase in model parameters and the percentage decrease in AWT
表 5. 模型参数增加与平均等待时间缩短的对比

算法	Parameter/%	AWT/%
3DQN_MDAM_C	0.004	9.57
3DQN_MDAM_S	0.059	9.04
3DQN_MDAM	0.063	17.6

6. 结论

本文提出了一种基于混合域注意力的深度强化学习交通信号控制模型 3DQN_MDAM 以优化单点交叉口的信号控制问题。在现有控制模型的基础上进行改进，通过设计了混合域注意力模块 MDAM，并将其嵌入到状态特征提取网络中，使智能体聚焦于对当前控制任务更为重要的信息，提升了信号控制的效果。在 SUMO 仿真平台上的实验结果表明，本文所提方法在低、中、高三种不同交通车流量的条件下均优于文中的对比方法。另外，交通信号控制器的响应速度对道路的行车安全至关重要，尤其在应对道路状态的全息感知时，控制器的运算速度面临挑战。因此，如何研究一种更轻量化、更高效的网络模型以减少控制器的运算量、加快控制器的响应速度，是本文进一步深入研究的方向。

参考文献

[1] Webster, F.V. (1958) Traffic Signal Settings. Road Research Technical Paper, 39.

[2] Miller, A.J. (1963) Settings for Fixed-Cycle Traffic Signals. *Journal of the Operational Research Society*, **14**, 373-386. https://doi.org/10.1057/jors.1963.61

[3] Cools, S.B., Gershenson, C. and D'Hooghe, B. (2013) Self-Organizing Traffic Lights: A Realistic Simulation. In: Prokopenko, M., Ed., *Advances in Applied Self-Organizing Systems*, Springer, Berlin, 45-55. https://doi.org/10.1007/978-1-4471-5113-5_3

[4] 刘志，曹诗鹏，沈阳，等. 基于改进深度强化学习方法的单交叉口信号控制[J]. 计算机科学，2020，47(12): 226-232.

[5] Arulkumaran, K., Deisenroth, M.P., Brundage, *et al*. (2017) Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine*, **34**, 26-38. https://doi.org/10.1109/MSP.2017.2743240

[6] Shi, B., Darrell, T. and Wang, X. (2023) Top-Down Visual Attention from Analysis by Synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 18-22 June 2023, 2102-2112. https://doi.org/10.1109/CVPR52729.2023.00209

[7] Abdulhai, B., Pringle, R. and Karakoulas, G.J. (2003) Reinforcement Learning for True Adaptive Traffic Signal Control. *Journal of Transportation Engineering*, **129**, 278-285. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:3(278)

[8] Li, L., Lv, Y. and Wang, F.Y. (2016) Traffic Signal Timing via Deep Reinforcement Learning. *IEEE/CAA Journal of Automatica Sinica*, **3**, 247-254. https://doi.org/10.1109/JAS.2016.7508798

[9] Touhbi, S., Babram, M.A., Nguyen-Huu, T., *et al*. (2017) Adaptive Traffic Signal Control: Exploring Reward Definition for Reinforcement Learning. *Procedia Computer Science*, **109**, 513-520. https://doi.org/10.1016/j.procs.2017.05.327

[10] Wan, C.H. and Hwang, M.C. (2018) Value-Based Deep Reinforcement Learning for Adaptive Isolated Intersection Signal Control. *IET Intelligent Transport Systems*, **12**, 1005-1010. https://doi.org/10.1049/iet-its.2018.5170

[11] Liu, S., Wu, G. and Barth, M. (2022) A Complete State Transition-Based Traffic Signal Control Using Deep Reinforcement Learning. 2022 *IEEE Conference on Technologies for Sustainability* (*SusTech*), Corona, 21-23 April 2022, 100-107. https://doi.org/10.1109/SusTech53338.2022.9794168

[12] Ye, B.L., Wu, P., Wu, W., *et al*. (2022) Q-Learning Based Traffic Signal Control Method for an Isolated Intersection. 2022 *China Automation Congress* (*CAC*), Xiamen, 25-27 November 2022, 6063-6068. https://doi.org/10.1109/CAC57257.2022.10054839

[13] Yi, C., Wu, J., Ren, Y., Ran, Y. and Lou, Y. (2022) A Spatial-Temporal Deep Reinforcement Learning Model for Large-Scale Centralized Traffic Signal Control. 2022 *IEEE* 25*th International Conference on Intelligent Transportation Systems* (*ITSC*), Macau, 8-12 October 2022, 275-280. https://doi.org/10.1109/ITSC55140.2022.9922459

[14] Genders, W. and Razavi, S. (2016) Using a Deep Reinforcement Learning Agent for Traffic Signal Control. https://arxiv.org/pdf/1611.01142.pdf

[15] Yu, P. and Luo, J. (2022) Minimize Pressure Difference Traffic Signal Control Based on Deep Reinforcement Learning. 2022 41*st Chinese Control Conference* (*CCC*), Hefei, 25-27 July 2022, 5493-5498. https://doi.org/10.23919/CCC55666.2022.9901790

[16] Haddad, T.A., Hedjazi, D. and Aouag, S. (2022) A New Deep Reinforcement Learning-Based Adaptive Traffic Light Control Approach for Isolated Intersection. 2022 5*th International Symposium on Informatics and its Applications* (*ISIA*), M'sila, 29-30 November 2022, 1-6. https://doi.org/10.1109/ISIA55826.2022.9993598

[17] Liang, X., Du, X., Wang, G. and Han, Z. (2019) A Deep Reinforcement Learning Network for Traffic Light Cycle Control. *IEEE Transactions on Vehicular Technology*, **68**, 1243-1253. https://doi.org/10.1109/TVT.2018.2890726

[18] An, Y. and Zhang, J. (2022) Traffic Signal Control Method Based on Modified Proximal Policy Optimization. 2022 10*th International Conference on Traffic and Logistic Engineering* (*ICTLE*), Macau, 12-14 August 2022, 83-88. https://doi.org/10.1109/ICTLE55577.2022.9901894

[19] Wei, H., Zheng, G., Yao, H., *et al*. (2018) Intellilight: A Reinforcement Learning Approach for Intelligent Traffic Light Control. *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, 19-23 July 2018, 2496-2505. https://doi.org/10.1145/3219819.3220096

[20] Genders, W. and Razavi, S. (2018) Evaluating Reinforcement Learning State Representations for Adaptive Traffic Signal Control. *Procedia Computer Science*, **130**, 26-33. https://doi.org/10.1016/j.procs.2018.04.008

[21] 任安妮，周大可，冯锦浩，等. 基于注意力机制的深度强化学习交通信号控制[J]. 计算机应用研究，2023，40(2): 430-434.

[22] Zhu, L., Peng, P., Lu, Z., *et al*. (2023) Metavim: Meta Variationally Intrinsic Motivated Reinforcement Learning for Decentralized Traffic Signal Control. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 11570-11584. https://doi.org/10.1109/TKDE.2022.3232711

[23] Woo, S., Park, J., Lee, J.Y., *et al*. (2018) Cbam: Convolutional Block Attention Module. In: *Proceedings of the European Conference on Computer Vision* (*ECCV*), Springer, Cham, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[24] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[25] Wang, Q., Wu, B., Zhu, P., *et al*. (2020) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 11534-11542. https://doi.org/10.1109/CVPR42600.2020.01155

[26] Hou, Q., Zhou, D. and Feng, J. (2021) Coordinate Attention for Efficient Mobile Network Design. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 19-25 June 2021, 13713-13722. https://doi.org/10.1109/CVPR46437.2021.01350

[27] 于泽，宁念文，郑燕柳，等. 深度强化学习驱动的智能交通信号控制策略综述[J]. 计算机科学，2023，50(4): 159-171.

[28] Wang, Z., Schaul, T., Hessel, M., *et al*. (2016) Dueling Network Architectures for Deep Reinforcement Learning. *International Conference on Machine Learning*, New York, 19-24 June 2016, 1995-2003.

[29] Van Hasselt, H., Guez, A. and Silver, D. (2016) Deep Reinforcement Learning with Double Q-Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **30**, 2094-2100. https://doi.org/10.1609/aaai.v30i1.10295

[30] Mnih, V., Kavukcuoglu, K., Silver, D., *et al*. (2013) Playing Atari with Deep Reinforcement Learning. https://arxiv.org/pdf/1312.5602.pdf

[31] Varaiya, P. (2013) The Max-Pressure Controller for Arbitrary Networks of Signalized Intersections. In: Ukkusuri, S.V. and Ozbay, K., Eds., *Advances in Dynamic Network Modeling in Complex Transportation Systems*, Springer, New York, 27-66. https://doi.org/10.1007/978-1-4614-6243-9_2