强化学习在自动驾驶中的应用综述

黄施捷, 张桐榆, 王伟燕, 戴丹阳, 罗宇轩

Abstract

近年来，随着人工智能算法的发展，其在自动驾驶汽车方面的应用已经成为一个热门的研究领域。本文将近几年基于深度学习和强化学习，尤其是基于深度强化学习的用于解决自动驾驶车辆主要功能的方法进行概述，主要包括决策、规划、控制、感知和社会行为。另外，我们实现了基于深度强化学习算法 DQN 的自动驾驶泊车的功能，并在第六章进行介绍。最后我们提出了开放问题和未来方向的讨论。

keyword

强化学习、深度强化学习、自动驾驶、深度学习、自动驾驶泊车

*联系我们: https://github.com/Huang-Shijie-SDUWH/Reinforcement_Learning_Auto_Parking

目录

1. 简介

随着经济和社会不断进步发展，人们对于自动驾驶汽车的要求也越来越高。如今，已有已经能够在城市道路和高速公路上行驶的自动驾驶车辆，它通过地图数据与全球定位系统 (Global Positioning System,GPS) 定位信号或者车载摄像头来获取车辆位置，通过识别道路上的路面标记、交通标志以及交通信号灯来作出正确的决策。但在一些地下停车场、小区车道等路况复杂的空间场景，GPS 信号较弱，同时缺乏路面标记以及交通辅助信息，自动驾驶车辆往往难以应对此类场景。

传统的自动驾驶系统在设计的过程中被分解为多个子系统，通过子系统之间的相互配合来完成自动驾驶任务，并在一些复杂场景中设计大量的子模块辅助车辆进行自动驾驶，这样的设计使得自动驾驶技术非常复杂，维护成本高昂。近些年，人工智能技术发展迅猛，尤其是深度学习和强化学习展现出了巨大的潜力。随着深度学习方法在决策支持、图像处理等领域大放异彩，自动驾驶各模块中也越来越多地应用了深度学习算法，但深度学习算法具有一定局限性，强化学习正好补足了这些短板。强化学习分为基于值函数的强化学习方法和基于策略梯度的强化学习方法。它是一种学习、预测、决策的方法框架,也是一种致力于实现通用智能解决复杂问题的方式。但是传统的强化学习方法在一些奖励稀疏问题上表现较差，针对该问

题,一些研究人员提出使用深度强化学习的方法解决。

强化学习在自动驾驶领域也有大量的应用，在驾驶车辆的过程中，驾驶员需要时刻注意车辆周围的环境情况，不断根据周围环境的变化作出决策，而深度强化学习技术能解决端到端的感知与决策问题，越来越多的学者开将深度强化学习应用在自动驾驶领域。

本文从 DL 和 RL 在自动驾驶领域的应用出发,对近年来涉及基于深度学习的自动驾驶方法及基于强化学习的自动驾驶方法的文献进行了总结，我们在 Elallid 等 [1] 这篇深度学习和强化学习在自动驾驶中应用的综述基础上，查阅了近几年的其他相关文献，并将本文的侧重点在强化学习的自动驾驶应用的介绍以及未来方向的探讨，在最后简单介绍了我们所做的基于深度 Q 网络的自动驾驶泊车方法的实现。

本片综述的组织如下：第二章介绍了自动驾驶的基本原理及其系统；第三章简单介绍了深度学习和基于深度学习的自动驾驶方法，概括了近五年的主要文献；第四章介绍了强化学习的发展历史，原理以及深度强化学习的常用算法；第五章对近五年的基于强化学习的自动驾驶方法进行概述；第六章介绍了我们所做的基于 DQN 的自动驾驶泊车方法的实现；第七章介绍了开放的研究问题和未来的研究方向；第八章总结了本文内容。

2. 自动驾驶汽车

自动驾驶就是车辆在无驾驶员操作的情况下自行实现驾驶。自动驾驶有多种发展路径，单车智能、车路协同、联网云控等。车路协同是依靠车-车，车-路动态信息的实时交互实现自动驾驶。联网云控更注重通过云端的控制实现自动驾驶。本章节将介绍单车智能。

自动驾驶基本原理概述

单车智能实现的基本原理是通过传感器实时感知到车辆及周边环境的情况，再通过智能系统进行规划决策，最后通过控制系统执行驾驶操作。

主要分为三个环节：

- 感知：车辆自身以及环境信息的采集与处理，包括视频信息、GPS 信息、车辆姿态、加速度信息等。
- 决策：依据感知到的情况，进行决策判断，确定适当的工作模型，制定适当的控制策略，代替人类做出驾驶决策。
- 控制：系统做出决策后，自动对车辆进行相应的操作执行。类比人类进行的方向盘以及油门、刹车的操作。系统通过线控系统将控制命令传递到底层模块执行对应操作任务。

硬件系统

硬件系统在各层都有。

感知层主要是为自动驾驶系统获取外部行驶道路环境数据并帮助系统进行车辆定位。当前无人驾驶系统中代表性的传感器有摄像头、激光雷达、毫米波雷达、超声波雷达、GNSS/IMU 等。

决策层需要自动驾驶芯片流畅地处理这些数据才能保证系统及时作出正确的决策，从而控制车辆自动行驶并确保安全。

控制层则相对简单，主要是线控。线控就是用线（电信号）的形式来取代机械、液压或气动等形式的连接，实现电子控制。

软件系统

自动驾驶在软件方面需要具备的能力如下：

- 地图引擎（Map）：提供道路、周边建筑等地图信息，高精地图还包含全局车道、曲率、坡度、红绿灯、护栏情况等等信息。
- 高精定位（Localization）：定位是一个重要模块，L3 及以上自动驾驶场景需要高精定位，是车辆信息感知的一个重要元素。
- 感知（Perception）：感知模块接受并处理传感器信息，从而识别自车以及周边的情况。
- 预测（Prediction）：预测模块主要用于预测感知到的障碍物的运动轨迹。
- 规划（Planning）：根据感知到的信息，规划出一条到达目的地的行进路线，而且还需要规划出未来一段时间内，每一时刻所在位置的精细轨迹和自车状态。
- 控制（Control）：如字面意思，通过指令控制车辆硬件进行操作。
- 交互界面（HMI）：人类在中控屏幕上看到的人机交互模块。
- 实时操作系统（RTOS）：Real Time Operation System 根据感知的数据信息，及时进行计算和分析并执行相应的控制操作。

整个架构的数据流向图如图 1。

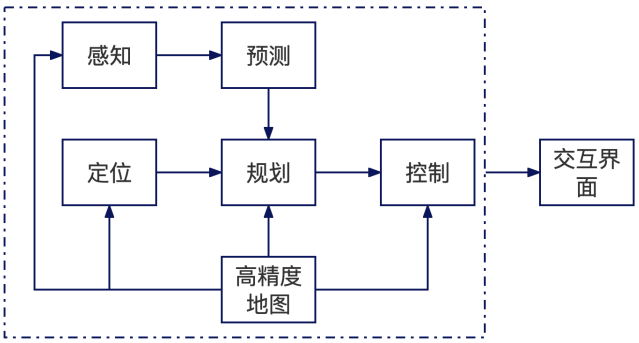

Figure 1. 自动驾驶数据流向

自动驾驶分级

NHTSA 将自动驾驶分为 5 类，自动驾驶分级如图 2。

- L1 表示车辆可以自动完成横向或纵向操控中的一项，其余所有工作仍然需要人类来完成。实现功能如 ACC(自适应巡航控制)、AEB(自动紧急制动)、LKA(车道保持辅助) 等。
- L2 是对横向和纵向多项操作同时进行控制。实现如超级巡航系统、APA(自动泊车)、TJA(交通拥堵辅助系统)、HWA(高速公路辅助驾驶) 等功能。
- L3 的驾驶主体切换成了系统，驾驶员只是支援角色。在 L3 功能开启时，系统完全负责操控和环境监测。实现功能如 HWP(高速公路自动驾驶)、TJP(交通拥堵自动驾驶) 等。
- L4 与前序过程的关键差异在于，系统不再需要人类的支援。在限定道路和环境情况下，系统可以完全负责操控和环境监测。
- L5 级时，道路和环境将不再是限制。系统将在所有情况下实现自动驾驶。到了这个阶段，方向盘、刹车、油门这些操控装置已经不再必须。

3. 深度学习与自动驾驶

3.1 深度学习简介

深度学习 (Deep Learning,DL) 是机器学习 (Machine Learning,ML) 的最重要一个分支，源于人工神经网络的研究，大多数深度学习方法使用神经网络架构，因此深度学习又称深度神经网络模型，它以神经网络的形式使用多层算法。通过不同的网络层来分析输入数据，每一层都定义数据中的特定要素和模式。深度学习有广泛的应用，从目标检测、语义分割等到自动驾驶汽车。

目前，深度学习已经成为自动驾驶的主要分支领域。在深度学习可以解决的关键问题中，包括检测和定位图像和视频中的物体，使自动驾驶汽车能够识别周围环境。在物体检测中使用的深度学习方法是卷积神经网络 (Convolutional Neural Network,CNN)，为自动驾驶汽车提供可操作的信息，即检测和分类物体 (例如，车道、交通灯、行人、交叉线和交通标志)。卷积神经网络在图像分类、目标检测和语义分割方面取得了很好的效果。它具有三个基本特征，即卷积层、池化层和全连接层。卷积层由过滤器组成，用于提取图像或视频中的主要视觉特征。这些过滤器将图像的大小转化为小的多阵列，并输入全连接层，输出层预测图像类别。

另一个适合处理数据序列的深度学习模型是 LSTM；LSTM 网络是一种循环神经网络 (Recurrent Neural Network, RNN)。LSTM 网络使用反馈连接进行序列和模式识别，并使用输入、输出和遗忘门。因此，它记住了前一个时间步骤计算的输出，并根据当前的输入提供输出。LSTM 网络已被应用于不同的自主驾驶任务，如运动规划、决策和车辆控制。因此，这种神经网络 (Neural Network,NN) 架构能够根据自动驾驶汽车过去的行动来预测当前的行动。

深度学习的主要优势是能够处理来自自动驾驶汽车上的摄像头的非结构化数据。一个深度学习网络可以随着时间的推移，从大量的图像和视频实例中学习。这些庞大的数据可能需要 10 天的时间在一台独立的计算机上进行训练。然而，鉴于深度学习网络的性质，GPU 可以用来大大减少训练过程。然而，它的局限性越来越明显，其中包括容易受到对抗性例子的影响，在这种情况下，数据的呈现可能会导致学习的模型更容易犯错。

3.2 基于深度学习的自动驾驶方法

在本节中，我们将从感知、规划、决策和控制等方面简单介绍基于深度学习的自动驾驶方法。

自动驾驶分级							
分级	NHHTSA	L0	L1	L2	L3	L4	
	SAE	L0	L1	L2	L3	L4	L5
SAE		无自动化	驾驶支持	部分自动化	条件自动化	高度自动化	完全自动化
定义		人类驾驶员全权驾驶汽车，在行驶过程可以得到警告	通过驾驶环境对方向盘和加速减速中一项提供支持，其余有人类做	通过驾驶环境对方向盘和加速减速中多项提供支持，其余有人类做	由无人驾驶系统完成所有的操作，根据系统要求，人类提供适当应答	由无人驾驶系统完成所有的操作，根据系统要求，人类不一定提供适当应答；限定道路和环境条件	由无人驾驶系统完成所有的操作，可能的条件下，人类接管；不限定道路和环境条件
主体	驾驶操作	人类	系统				
	周边监控	人类			系统		
	支援	人类				系统	
	系统作用	无	部分				全部

Figure 2. 自动驾驶分级表

感知

深度卷积神经网络 (CNN) 经过前期的数据训练以及特征提取，通过使用不同的传感器，如激光雷达，给出复杂环境中场景物体类别等信息，能够帮助自动驾驶汽车对周边的环境态势的理解达到显著效果 [2]。

如图 3 展示了本节所提及的文献。

道路场景

在准确识别和提取道路信息方面，Balado 等 (2019) [3] 使用移动激光扫描 (MLS) 获得信息，基于点云 (Point-Net) 和语义分割 (SS) 的方法来识别城市道路的元素；在识别道路的几何结构方面，Laddha 等 (2016) [4] 使用基于 CNN 的一种监督学习和无监督学习的混合算法来识别道路几何结构，但仍受天气条件的限制；在识别道路的拓扑结构方面，Yan 等 (2020) [5] 提出了一种基于激光雷达数据的多任务道路感知网络 (LMRoadNet) 方法；在识别道路拐角方面，Bolte 等 (2019) [6] 基于 CNN 提出在自动驾驶汽车的视频信号中检测的方法；在对交通标志的检测方面，Sajjad 等 (2020) [7] 提出了一种基于视觉传感器识别各种交通标志，并使用超声波传感器躲避障碍物的方法。

物体检测

在识别行人和其他车辆方面，Wang 等 (2020) [8] 提出了使用 CNN 和融合网络 (FoFNet) 的端到端的 3D 物体检测方法；在识别与物体的距离方面，Chen 等 (2018) [9] 提出了多任务组合策略 (CP-MTL) 算法；在夜间检测方面，Li 等 (2021) [10] 利用 CNN 来改善低光图像识别，可以帮助自动驾驶汽车在没有路灯的农村识别道路；在三维物体检测方面，Chen 等 (2017) [11] 通过使用激光雷达点云和 RGB 图像提出基于多视角三维网络 (MV3D) 的检测方法，优于只依靠传感器的方法；在弱势行人的安全方面，Zhao 等 (2020) [12] 基于内部级联网络 (INCHet) 提出行人位置感知网络 (P-LPN) 的方法，但未考略到行人的行动趋势。

规划

规划是自动驾驶汽车感知后的下一个主要任务，随着深度学习算法的发展，其应用能够帮助自动驾驶汽车在决策方面更为有效。

如图 4 展示了本节所提及的文献。

运动规划

在规划转向角方面，由于现实交通情况的复杂性，建立起一个通用的运动规划系统是复杂的，尤其是在实时条件下，目前对于其他道路使用者的行为仍然是

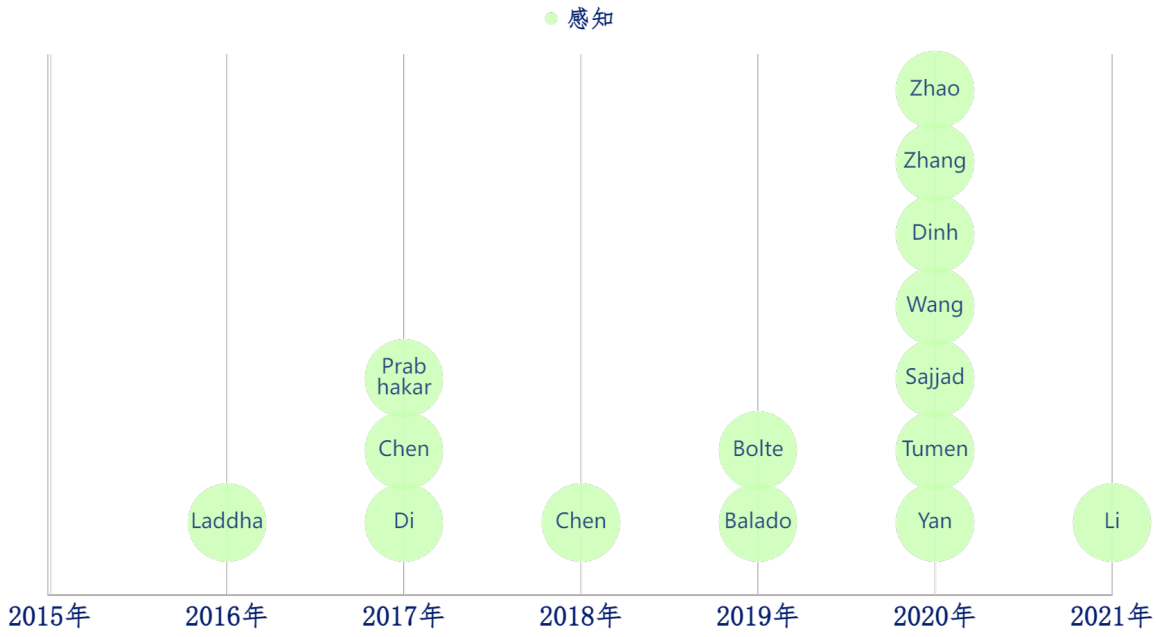

Figure 3. 感知

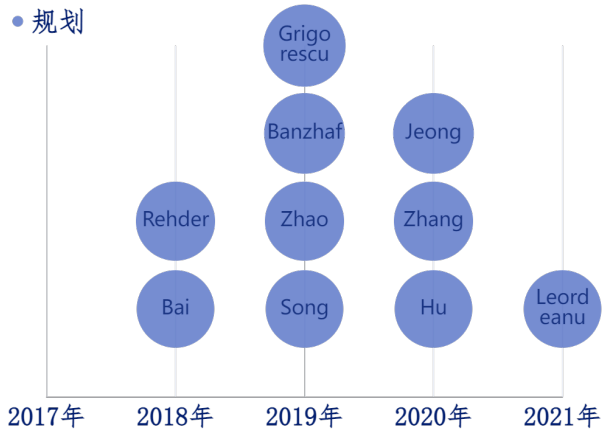

Figure 4. 规划

不可测的，没有一个良好的模型能够对此建模。Bai 等 (2018) [13] 结合 CNN 和 LSTM 来提出空间-时间信息来实时规划转向角。Song 等 (2018) [14] 使用从驾驶模拟器收集的数据来规划转向角；在同时规划多个运动指令方面，Hu 等 (2020) [15] 提出了基于 CNN 和 LSTM 的深度级联网络来同时规划转向角、加速度和刹车。

轨迹规划

自动驾驶汽车要分析有传感器收集的数据以躲避障碍物来安全导航到目的地。Zhao 等 (2019) [13] 提出学习卡尔曼网络 (LKN) 来规划汽车的轨迹。Banzhaf 等 (2019) [16] 等提出基于 CNN 的规划车辆轨迹的方法，考虑到了静态障碍物和非结构化道路；Grigorescu 等 (2019) [17] 提出了融合激光雷达和雷达数据来规划最佳路径的方法；Zhang 等 (2020) [18] 通过使用模仿学习 (Imitation Learning,IL) 探索人类驾驶员在环境中的经验特征，利用 CNN 来处理环境信息并识别周围车辆状态；Leordeanu 等 (2021) [19] 根据视频和最终目的地来规划轨迹，可以适应不同天气条件；Jeong 等 (2020) [20] 提出基于 LSTM 和 RNN 的规划算法来减少交叉口的碰撞；

决策

目前，深度学习算法已经显示出了在复杂环境中实时决策的能力。

如图 5 展示了本节所提及的文献。

Li 等 (2018) [21] 提出了基于 CNN 提取道路场景图象来模仿人类驾驶员进行正确的决策；Gallardo 等 (2017) [22] 提出了基于 Alexnet 架构的环境来为自动驾驶汽车做出决策的方法；Xie 等 (2019) [23] 提出了一种基于 LSTM 的车道变化的决策方法来模拟自动驾驶汽车和周围车辆的相互作用；Liu 等 (2019) [24] 提出了基于 DNN 的利用驾驶员的过往经验来实现变道决策的方法；Strickland 等 (2018) [25] 提出了基于

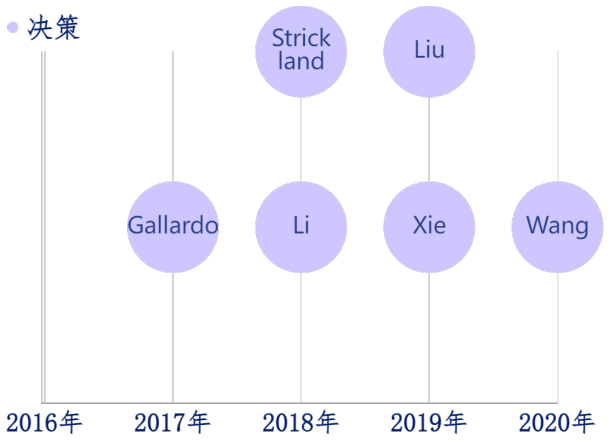

Figure 5. 决策

贝叶斯 Conv-LSTM 的方法来处理数据并避免碰撞的方法；Wang 等 (2020) [26] 提出了基于 R-CNN 的在达到环岛处的决策方法。

控制

自动驾驶汽车的控制负责修正规划和决策时产生的错误，稳定和引导车辆沿道路行驶。传统方法的控制器无法处理复杂的情况，深度学习算法的出现弥补了传统方法的缺陷。

如图 6 展示了本节所提及的文献。

横向控制

横向控制主要通过调整转向角来实现对航向的控制，并纠正驾驶过程中可能积累的任何误差。

Rausch 等 (2017) [27] 提出了一种基于 CNN 的横向控制方法，使用从自动驾驶汽车的摄像头中收集的数据进行训练，但只能双车道道路进行了评估；Sharma 等 (2018) [28] 提出了基于 CNN 和 TORCS(一个赛车模拟器) 的不同道路的横向控制方法，使用从传感器收集的数据在 TORCS 的两个轨道上进行了模拟；Lee 等 (2020) [29] 提出了基于 CNN 提取特征，LSTM 控制转向角的横向控制方法；Maqueda 等 (2018) [30] 利用 CNN 来训练来自事件相机的数据，解决了在白天和夜晚的横向控制问题；Mújica-Vargas 等 (2020) [31] 提出了基于 CNN 和 RNN 的增强的横向控制方法，但未考虑到障碍物和其他道路使用者；

纵向控制

纵向控制主要通过控制汽车的刹车、油门来实现对汽车的速度控制。

Mohseni 等 (2018) [32] 提出了基于 DNN 的模型预测控制 (MPC) 方法来调节自动驾驶汽车的速度并避免障碍物，实现纵向控制；Szilassy 等 (2019) [33] 提出了基于 DNN 的控制路口处的自动驾驶汽车速度，但未考虑到路口的其他因素；Al-Sharman 等 (2020) [34] 提出了基于 DNN 的刹车控制方法，使用从真实车辆中收集的数据以模仿人类驾驶员。

横向和纵向控制

深度学习算法也可以同时应用的横向和纵向控制中。Chen 等 (2015) [35] 提出了基于 CNN 的同时控制转向角、加速度和刹车的方法，能够使自动驾驶汽车在不同场景下行驶；Devineau 等 (2018) [36] 提出了基于 CNN 的多层感知 (MLP) 方法，训练自动驾驶汽车在包括从长直线行驶到弯道行驶转换等挑战性场景下进行控制；Sharma 等 (2019) [37] 提出了基于 CNN 的同时控制速度和转向角的方法，使用 TORCS 收集包括速度、转向角、油门和刹车位置的道路图像进行模拟；Xing 等 (2020) [38] 提出了基于 CNN 和 RNN 的在高速公路上的控制方法，但未但交通密集的条件下测试。

4. 强化学习

4.1 强化学习发展历史

强化学习是动物心理学、最优控制理论和时序差分学习等学科交叉的产物 [39]。强化学习的“试错”思想源于动物心理学家对试错行为的研究，最早可追溯到巴甫洛夫的条件反射实验。1953 年，美国应用数学家 Richard Bellman 提出求解最优控制的动态规划 (dynamic programming,DP) 的数学理论和方法，后来该方法衍生出了强化学习试错迭代求解的机制。1957 年，Bellman 提出了马尔可夫决策过程 (Markov Decision Process,MDP)，后来马尔可夫决策过程也成为定义强化学习问题的最普遍形式。时序差分学习 (temporal-difference learning,TDL) 是动态规划和蒙特卡洛方法结合的产物。1959 年 Samuel 首次提出并实现一个包含时序差分思想的学习算法。1989 年 Watkins 在他的博士论文最早提出 Q 学习算法，这项工作正式标志着强化学习的诞生，该算法通过优化累积未来奖励信号学习最优策略。1992 年，Watkins 和 Dayan 给出了 Q 学习算法收敛性的证明。2013 年，来

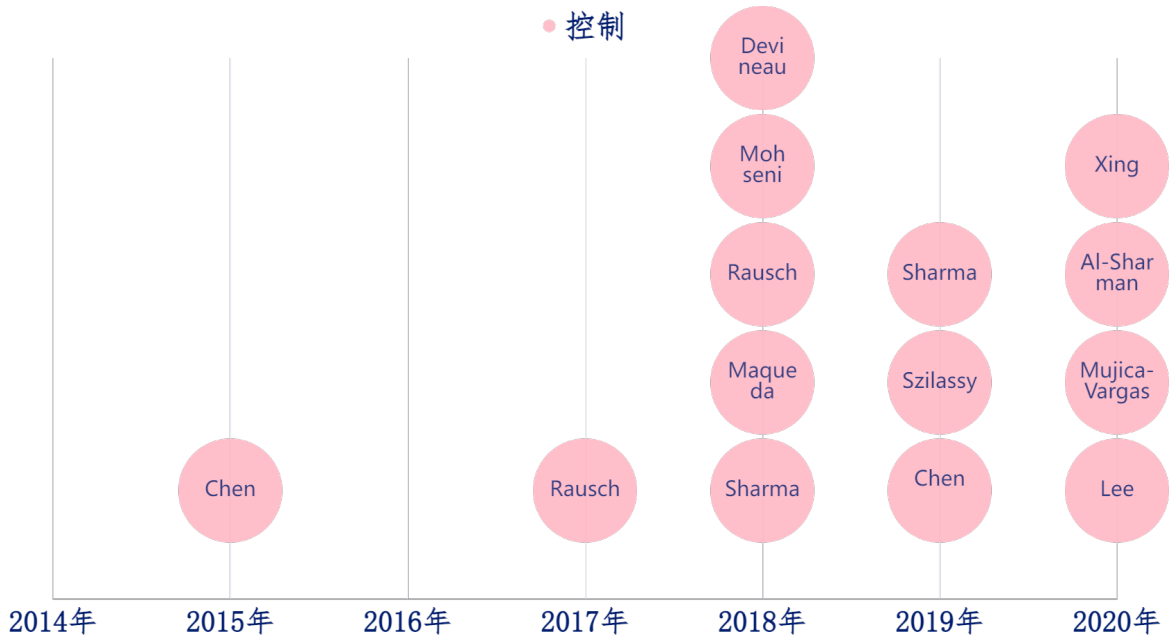

Figure 6. 控制

自 DeepMind 的 Mnih 等人利用深度学习网络 (CNNs) 直接从高维度的感应器输入 (sensory inputs) 提取有效特征，并利用 Q-Learning 学习主体的最优策略，提出了结合深度学习的深度 Q 学习 (DQL)。表 1 总结了强化学习发展历程中的若干重要事件。

Table 1. 强化学习发展历史

年份	提出者	事件
1953	Bellman [40]	动态规划
1957	Bellman [41]	马尔可夫决策过程
1977	Werbos [42]	自适应动态规划
1988	Sutton [43]	时序差分算法
1989	Watkins [44]	Q 学习算法
1994	Rummery [45]	SARSA 算法
2013	Mnih [46]	DQN 算法
2014	Silver [47]	DPG 算法
2015	Schulman [48]	TRPO 算法
2015	Lillicrap [49]	DDPG 算法
2016	Mnih [50]	A3C 算法
2016	Huang [51]	AlphaGO 围棋机器人
2017	Schulman [48]	PPO 算法

4.2 强化学习

强化学习 (Reinforcement Leaning,RL) 与监督学习 (Supervised Learning,SL)、非监督学习 (Unsupervised Learning,UL) 构成机器学习的三个种类。强化学习又称增强学习、再励学习，是一种通过不断试错 (Trial and Error) 来学习的方法，思想是通过奖励或惩罚智能体 (Agent) 来使其未来更容易重复或放弃某一动作。强化学习的主要角色是智能体和环境 (Environment)，环境是智能体之外一切组成和与之互动的事物组成的世界。如图 7: 智能体处在一个环境中，每个状态为智能体对当前环境的感知；智能体只能通过动作来影响环境，当智能体执行一个动作后，会使得环境按某种概率转移到另一个状态；同时，环境会根据潜在的奖赏函数反馈给智能体一个奖赏。

马尔可夫决策过程 强化学习智能体与环境的交互过程可以看作一个马尔可夫决策过程 (Markov Decision Process,MDP)。马尔可夫决策过程中的下一时刻状态 s_{t+1} 只取决于当前状态 s_t 和动作 a_t，即:

$$p(s_{t+1}|s_t,a_t,\ldots,s_0,a_0) = p(s_{t+1}|s_t,a_t)$$

马尔可夫决策过程可以定义为一个四元组 (S,A,R,P,γ)，其中:

1.S 代表所有有效状态的集合;

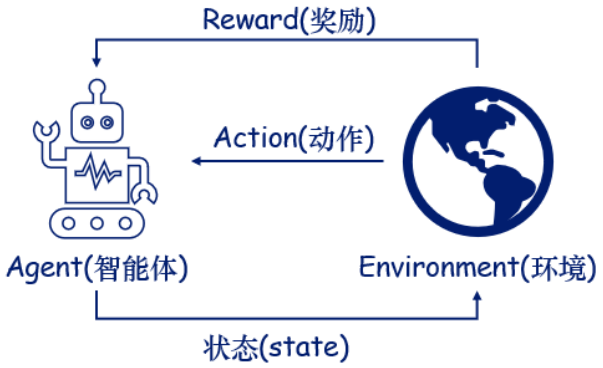

Figure 7. 强化学习过程

2.A 代表所有有效动作的集合；

3.$R: S \times A \times S \to R$ 为奖励函数，$r_t = R(s_t, a_t, s_{t+1})$；

4.$P: S \times A \to P(S)$ 代表状态转移概率函数。$P(S_{t+1}|s_t, a_t)$ 代表在状态 s_t 下执行动作 a_t 转移到状态 s_{t+1} 的概率；

5.γ 是折扣因子，是计算累积回报的参数之一。

策略 (Policy) 在强化学习中，策略即决策函数，策略 $\pi: S \to A$ 是状态集合到动作集合的一个映射，表示智能体在状态 s_t 选择动作 a_t 的规则。强化学习策略分为确定性策略和随机性策略，确定性策略输出的是智能体在某状态下应该执行的操作，随机性策略输出的是某个状态的所有可能动作的概率分布。

在强化学习中，行动轨迹 τ 是状态和动作的序列，即 $\tau = (s_0, a_0, s_1, a_1, \dots)$，如图 8。

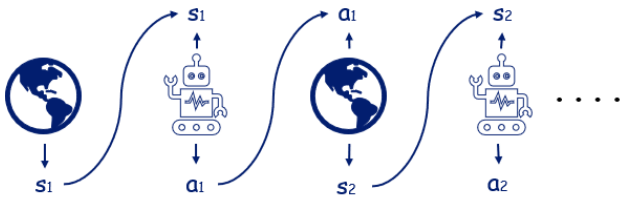

Figure 8. 行动轨迹

智能体的目标是最大化行动轨迹的累计奖励，即

$$R(\tau) = \sum_{t=0}^{T-1} r_{t+1}$$

而折扣奖励是智能体获得的随时间不同而衰减的全部奖励之和，即

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_{t+1}$$

其中 $\gamma \in [0,1]$ 为衰减率，衰减率一方面降低了未来回报的权重，另一方面也避免了无限多个奖励不收敛的问题。

强化学习的目标函数是期望回报，即

$$J(\theta) = \underset{\tau \in p_\theta(\tau)}{E}[R(\tau)] = \underset{\tau \in p_\theta(\tau)}{E}[\sum_{t=0}^{T-1} \gamma^t r_{t+1}]]$$

从状态 s 开始，按照策略 π 执行到回合结束，重复这一过程中得到的回报期望称为状态价值，$V^\pi(s)$ 为状态值函数，即

$$V^\pi(s) = \underset{\tau \sim \pi}{E}[R(\tau)|s_0 = s]$$

$V^\pi(s)$ 和 $Q^\pi(s,a)$ 都满足贝尔曼方程。贝尔曼方程的思想是：当前状态的价值等于从当前状态转换到下一状态的奖励加上下一状态的价值。根据这个思想，V^π 可以表示为

$$V^\pi = \underset{\substack{a \sim \pi \\ s' \sim P}}{E}[r(s,a) + \gamma V^\pi(s')]$$

$Q^\pi(s,a)$ 可以表示为

$$Q^\pi(s,a) = \underset{s' \sim P}{E}[r(s,a) + \gamma \underset{a' \sim \pi}{E}[Q^\pi(s',a')]]$$

4.3 深度强化学习

深度强化学习 (Deep Reinforcement Learning) 是深度学习和强化学习的结合，是一种端到端的系统。深度强化学习的基本过程与强化学习基本一致，只是对于环境中的观测状态更高维和对状态的处理方式不一样，即使用深度学习进行感知。

基于值函数的学习方法

基于值函数 (Value-based) 的学习方法中求解最优策略等价于求解最优的值函数，值函数近似指利用参数化的 $Q_\phi(s,a)$ 来近似计算值函数 $Q^\pi(s,a)$。函数 $Q_\phi(s,a)$ 在强化学习中用神经网络来表示，称为 Q 网络。

基于值函数是对策略的评估，为不断优化直至选出最优策略，一种可行的方式是依据值函数选取策略更新的方式，常见的策略有贪婪 (greedy) 策略和 $\varepsilon - greedy$ 策略。如 DQN 算法中采用了 $\varepsilon - greedy$ 策略。

$\varepsilon-greedy$ 算法中的 ε 人为定义，在每次选择动作时，随机生成一个数 $\alpha \in [0,1]$。如果 $\alpha < \varepsilon$，执行随机生成的动作，否则选择执行策略网络输出的动作。即

$$\pi(a|s) \leftarrow \begin{cases} 1-\varepsilon+\frac{\varepsilon}{|A(s)|} & if\, a = \arg\max_a Q^{\pi}(s,a) \\ \frac{\varepsilon}{|A(s)|} & if\, a \neq \arg\max_a Q^{\pi}(s,a) \end{cases}$$

基于策略梯度的学习方法

策略梯度 (Policy-gradient) 算法是最经典的基于策略优化的学习方法。策略梯度算法以累计奖励为目标函数 $J(\theta)$，使用梯度上升来优化神经网络参数 θ。目标函数 $J(\theta)$ 关于 θ 的梯度可以表示为：

$$\begin{aligned} \frac{\partial J(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta}\int p_{\theta}(\tau)R(\tau)\,\mathrm{d}\tau \\ &= E_{\tau\sim p_{\theta}(\tau)}[\sum_{t=0}^{T-1}\frac{\partial}{\partial \theta}\log \pi_{\theta}(a_t|s_t)\gamma^t R(\tau_{t:T})] \end{aligned}$$

其中 $R(\tau_{t:T})$ 是从时刻 t 开始收到的总回报。

演员-评论家学习方法

演员-评论家 (Actor-Critic) 算法，其中“演员”代表策略函数，自适应地提高，“评论家”代表值函数，对“演员”的“表演”进行评价。由于值函数可以直接估计当前状态的价值来近似真实的回报，演员-评论家算法能够在智能体每执行一步后进行策略函数和价值函数的更新，而不需要整个轨迹的样本数据。

部分深度强化学习算法的类型在表 2 中所示。

Table 2. 深度强化学习算法

算法名称	算法类型
深度 Q 网络 (DQN) [46]	基于值函数
深度确定性策略梯度算法 (DDPG) [49]	基于策略
软演员评论家算法 (SAC) [52]	基于策略
异步优势行动者评论家算法 (A3C) [50]	基于策略
分布式近似策略优化算法 (DPPO) [53]	基于策略

5. 基于强化学习的自动驾驶方法

强化学习能够使模型通过反复训练和试错来学习执行任务，本章我们将从决策、运动规划、控制和社会行为四个方面来阐述强化学习在自动驾驶中的主要应用。下图 9 是基于强化学习的自动驾驶方法的历史脉络。

5.1 决策

You 等人 (2018) [54] 将自动驾驶汽车与环境之间的交互建模为随机马尔可夫决策过程 (MDP)，并将有经验的驾驶员的驾驶风格作为学习目标。MDP 考虑到道路结构，以便融入更多不同的驾驶风格。但是，需要在不同场景中进行更多测试来验证其方法。

Hoel 等人 (2018) [55] 是基于深度强化学习的自动生成通用决策函数的方法。他们利用深度网络 DQN 来训练他们的提议并预测正确的决定。为了证明该方法的通用性，还通过在有迎面而来的车辆的道路上对超车情况进行训练，对完全相同的算法进行了测试。但是，他们的方法并不能保证在特殊情况下确保安全。

无信号交叉口是自动驾驶做出准确及时决策的最具挑战性的场景之一，提供有效的策略以安全地通过无信号的十字路口需要确定其他驾驶员的意图。Isele 等人 (2018) [56] 提出了一种基于 DQN 的方法，这能够学习在多个指标，尽管泛化能力有限，他们提出的结果优于传统方法，也为当时的学习提供了未来研究方向。

Okuyama 等人 (2018) [57] 介绍了自动驾驶汽车在仅包含车道标记和静态障碍物的简化环境中学习驾驶的仿真结果。学习将 CNN 和 RL 相结合。对于汽车前置摄像头捕获的给定街道输入图像，Deep Q Network 计算与自动驾驶汽车可用操作相对应的 Q 值 (奖励)。汽车中的自动驾驶系统强制执行具有最高奖励的动作。他们的仿真具有很高的精度，但是他们没有考虑到动态障碍。

Ye 等人 (2019) [58] 提出了一个决策训练和学习的框架，由深度强化学习 (DRL) 训练程序和高保真虚拟仿真环境组成。尽管他们在某些复杂路段取得了较好的成果，但是他们的提议需要在复杂场景中进行更多测试。

Sun 等人 (2020) [59] 提出一种基于深度确定性策略梯度算法的重型智能车辆半规则决策策略。自动驾驶的连续状态决策采用深度确定性策略梯度算法，即深度确定性策略梯度算法。网络接收自动驾驶汽车的信息状态，然后做出决定。通过仿真实验验证了算法的有效性和鲁棒性。

Duan 等人 (2020) [60] 提出了一种方法来训练自动驾驶汽车从环境中学习，以基于 MDP 将决策分为

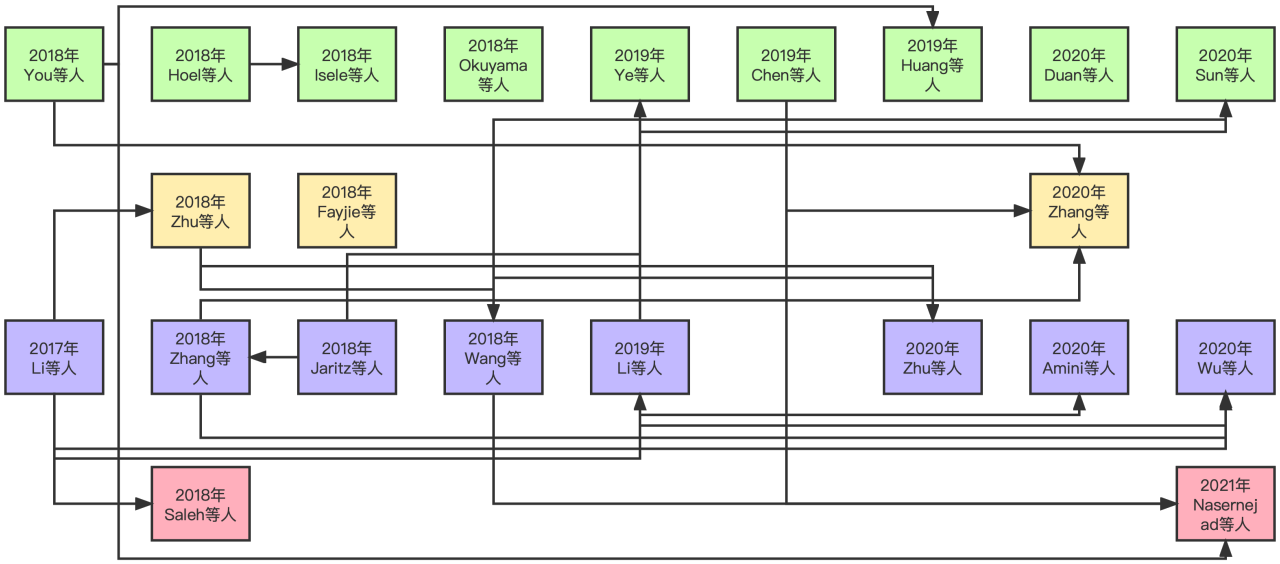

Figure 9. 基于强化学习的自动驾驶方法的历史相关论文

三个动作（包括行车道、右变道和左变道）。然而，他们只关注高速公路的情况，而没有处理其他情况（例如十字路口和城市交通）。

Chen 等人 (2019) [60] 提出了一个城市自动驾驶场景中实现无模型深度强化学习的框架。通过一个特定的输入表示，并使用视觉编码捕获低维潜在状态。他们实现了几种最先进的无模型深度 RL 算法，并提高其性能。通过使用，他们的方法能够很好地解决该环境下的应用问题。

Huang 等人 (2019) [61] 提出了一种基于 DDPG 深度强化学习的端到端决策模型。先建立了一个端到端的决策模型，将驾驶状态连续映射到驾驶行为，其次在 TORCS 平台上的不同场景中训练和验证代理。结果表明，DDPG 算法可以实现端到端的自动驾驶决策。

5.2 运动规划

Zhu 等人 (2018) [62] 提出了一个基于深度强化学习的类人自动跟随规划框架，模拟了自动驾驶与试错交互的学习过程，最终获得了一个最优策略或跟车模型。如研究结果显示，一种深度确定性的策略梯度跟车模型，使用模拟速度和观察速度之间的差异作为奖励函数，可以重现类似人类的跟车行为比传统和最近的数据驱动的跟车模型具有更高的准确性，其间距验证误差低于其他模型，包括智能驱动模型、基于局部加权回归的模型和传统的基于神经网络的模型。研究表明，强化学习方法可以深入了解驾驶员行为，并有助于开发类似人类的自动驾驶算法和交通流模型。

Fayjie 等人 (2018) [63] 介绍了自动驾驶汽车的深度强化学习自主导航和避障，将深度 Q 网络应用于包括城市交通和 5 车道高速公路环境中的模拟汽车。该方法使用 CNN 提取两种类型的传感器数据的输入：车前的摄像头传感器和激光传感器。后将特征反馈到 DQN 以估计正确的动作。它还设计了一款经济高效的高速汽车原型，能够实时运行相同的算法。但是，他们的提案没有考虑到其他道路动态的个体。

即使具有从专家驾驶员的停车数据中学习的深度学习能力，人类知识也不能保证高效停车。为此，Zhang 等人 (2020) [64] 提出了一种基于模型的强化学习方法，该方法通过迭代执行数据生成、数据评估和训练网络来学习数据的停车策略。训练好的网络用于指导后续迭代中的数据生成周期。学习到的策略保证了停车过程中上述要求的多目标最优性。为了使系统独立于专家数据或先验知识，提出了一种结合蒙特卡洛树搜索和纵向和横向策略的数据生成算法通过以量产停车系统为基准的实车测试，发现所提方法具有更好的停车效率和更低的起步停车姿态要求，从而验证了算法的优越性。

5.3 控制

强化学习的目标是通过迭代方式探索环境来找到最佳控制命令，环境正在根据自动驾驶汽车的当前行为奖励他们，以纠正未来的错误。

Li 等人 (2017) [65] 提出了一种利用点云数据进行分箱拣选任务的姿态估计算法。提出了一种新颖的描述符曲线集特征（CSF）来通过该点周围的表面波动来描述该点，并且还能够评估姿势。提出了旋转匹配特征（RMF）来有效地匹配 CSF。引入了一种基于体素的姿态验证方法来评估姿态。该算法针对大量合成和真实场景进行了评估，并被证明对噪声具有鲁棒性。

Li 等人 (2019) [66] 研究了具有深度学习和强化学习方法的基于视觉的自动驾驶。该方法将基于视觉的横向控制系统分解为感知模块和控制模块。基于多任务学习神经网络的感知模块首先将驾驶员视野图像作为输入，并预测轨迹特征。基于强化学习的控制模块然后根据这些特征做出控制决策。经过训练的强化学习控制器在不同的轨道上优于线性二次调节器控制器和模型预测控制控制器。实验表明，感知模块显示出良好的性能，并且控制器能够通过视觉输入很好地控制车辆沿轨道中心行驶

Zhu 等人 (2020) [67] 提出了一种基于强化学习的汽车跟随速度控制模型。为了优化驾驶性能，通过参考人类驾驶数据并结合与安全性、效率和舒适性相关的驾驶特征，开发了奖励功能，在模型训练和测试阶段都使用安全检查策略，从而实现更快的收敛和零碰撞。仿真结果显示了他们方案在安安全、高效、舒适方面的优于前序方法。

Amini 等人 (2020) [68] 提出了一个数据驱动的模拟和训练引擎，能够仅使用稀疏奖励来学习端到端的自动驾驶汽车控制策略。通过在环境中利用真实的、人类收集的轨迹，渲染新的训练数据，允许虚拟代理沿着与道路外观和语义一致的新局部轨迹连续行驶，每个轨迹都有不同的场景视图，在模拟器中学习到的策略能够推广到以前看不见的现实世界道路并在其中导航，而无需在训练期间访问任何人工控制标签。该方法是可扩展的，利用强化学习，并广泛应用于物理世界中需要有效感知和稳健操作的情况。

Wu 等人 (2020) [69] 提出后一种更有效的深度强化学习 (DRL) 模型被开发用于差分可变速度限制 (DVSL) 控制，其中可以在车道之间施加动态和不同的速度限制。所提出的 DRL 模型使用一种新颖的 actor-critic 架构来学习连续动作空间中的大量离散速度限制。使用不同的奖励信号。所提出的基于 DRL 的 DVSL 控制器在高速公路上进行了测试，并具有模拟的循环瓶颈。仿真结果表明，基于 DRL 的 DVSL 控制策略能够提高高速公路的安全性、效率和环境友好性，从结果中观察到 DRL 代理的鲁棒性。

Zhang 等人 (2018) [70] 提出双 Q 学习的强化学习方法用于基于自然驾驶数据构建的环境来控制车辆的速度。根据直接感知方法的概念，使用一种称为集成感知方法的新方法来构建环境。模型的输入由高维数据组成，包括从视频数据处理的道路信息和从传感器处理的低维数据。该方法提高了值精度和策略质量。

Jaritz 等人 (2018) [71] 使用最新的强化学习算法进行端到端驾驶的研究，无需任何中介感知（对象识别、场景理解）。新提出的奖励和学习策略仅使用来自前置摄像头的 RGB 图像共同导致更快的收敛和更稳健的驾驶。A3C 框架用于在物理和图形逼真的拉力赛游戏中学习汽车控制，代理在具有各种道路结构（转弯、山丘）、图形（季节、位置）和物理（道路依从性）。进行了彻底的评估，并在看不见的轨道上使用合法的速度限制证明了泛化。对真实图像序列的开环测试显示了方法的一些域适应能力。

Wang 等人 (2018) [72] 提出了一种基于强化学习的方法来训练车辆代理学习自动变道行为，使其能够在各种甚至不可预见的情况下智能地进行变道。该方法将状态空间和动作空间都视为连续的，并设计了一个具有封闭式贪心策略的 Q 函数逼近器，这有助于深度 Q 学习算法的计算效率。为训练算法进行了广泛的模拟，结果表明基于强化学习的车辆代理能够学习用于车道变换操作的平稳高效的驾驶策略。

5.4 社会行为

RL 方法展示表现了实时处理城市交通中的道路使用者行为并做出正确决策的能力。

Saleh 等人 (2018) [73] 提出了一种基于逆强化学习 (IRL) 和双向循环神经网络架构 (B-LSTM) 的数据驱动框架，用于长期预测行人的轨迹。他们在交通环境中代理行为建模的真实数据集上评估了他们的框架，结果显示预测率有显著提高。

Nasernejad 等人 (2021) [74] 开发一个基于代理的框架，以真实地模拟未遂事故中的行人行为，并提高对与车辆交互中行人规避动作机制的理解。行人-车辆冲突使用马尔可夫决策过程 (MDP) 框架建模。实施连续高斯过程逆强化学习 (GP-IRL) 方法来检索行人的奖励函数并推断他们在冲突情况下的碰撞避免机制。深度强化学习 (DRL) 模型用于估计交通冲突中的最佳行人策略。结果表明，所开发的模型可以高精度地预测冲突情况下的行人轨迹及其规避动作机制。这项研究是为混合交通条件下的行人开发面向安全的微观仿真工具的关键一步。

6. 基于 DQN 的自动驾驶泊车方法实现

在本章我们将介绍我们所做的基于 DQN 的自动驾驶泊车的实现相关配置、定义及实现过程。

模拟小车环境 Webots [75]

起初我们想使用 Matlab 中的强化学习库 [76]，但是其仿真环境复杂因而改用使用 Python：面向对象编程操控小车。

强化学习库 Stable_Baseline3 [77]

强化学习方法 Deep Q-Network(DQN) [78]

Action_space

1. 向后直线行驶
2. 改变方向 90 度倒车入库

我们所使用的模拟小车如图 10 所示。

Observation

1. 小车位置 (x,y) 坐标
2. 小车朝向 orientation
3. 小车的 16 个距离传感器读数 [79]（但是加入距离传感器后，虽然 30 万个 time_steps 可以成功停车，但并没有学习到最优动作 (图 14)，原因分析见下文困难）
4. 可以加入任意传感器，如摄像机、雷达，没加入的原因是环境简单，CNN 无法提取充足有效信息，减少过拟合，加快运算。

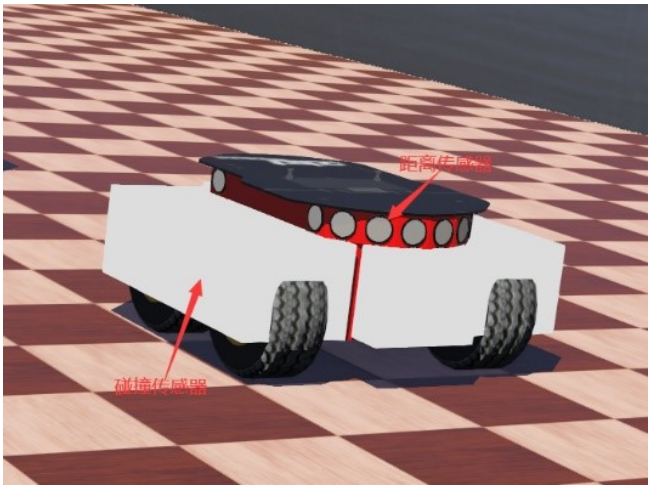

Figure 10. 所使用小车

Episode 小车成功倒车入库，或者小车进入不可能成功停车的状态。

不可能停车状态定义

1. 提前转向
2. 转向多次
3. 越界

加快训练方法 剪枝：一旦进入不可能停车的状态立刻停止该 episode。定义好环境后，使用 stable_baseline3 的代码库就可以选择强化学习算法进行训练 (函数拟合)，因此困难来到了 reward 的定义。

Reward 定义原则

1. 离停车位置越近,reward 越大
2. 进入停车位置后转到正确方向则给予奖励
3. 进入不可能停车状态给予惩罚
4. 停车成功给予奖励。

人为设定,和深度学习里的超参数类似,不同的 reward 会给训练结果带来相当大的差异。模型后期调整就是调整 reward 的设定。

困难 虽然根据通用近似定理，神经网络可以逼近连续函数。但是在本项目中 Q(s,a) 却是不连续的 (图 11)。因为如果小车在 A 处转向就会停车失败，在 B 处转向就会停车成功。因此相邻的状态相同的动作 Q(s,a) 差别相当大。这就会导致训练时 A 处明明不

能进行转向，但因为神经网络对此处（跳跃间断点）模拟不好，误以为可以转向（图 12）。我们意识到了这个问题，尽可能在使得 reward 关于状态连续，相比之前的训练效果得到了很好的提升（图 13）。

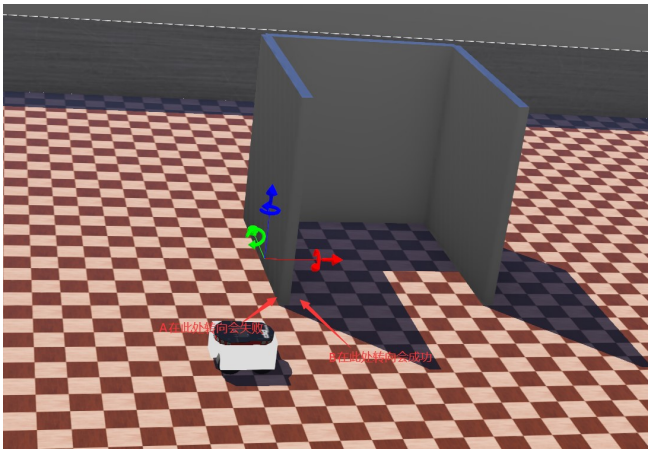

Figure 11. 小车停车 AB 点示意

后序改进思路

1. 在环境中加入其它车辆，尽可能模拟现实世界物体
2. 尽可能优化奖励函数，缩短算法收敛时间

计算机科学是一个相当美妙且实用的学科，其很多思想精髓都可以从现实生活中得到。强化学习就是一个例子：用下一状态的 Q 更新当前状态的 Q，不就是巴普洛夫的条件反射实验吗？更深一层，还有张一鸣创业初期所强调的延迟满足（虽然字节跳动产品大多是即时快乐）有了环境，reward 也会对结果带来相当大的影响，所以要按劳分配，要禁毒。如果人们研究时也会获得吸毒一样的感觉，世界将会怎样？

7. 开放问题和未来方向

尽管基于 DL 和 RL 的技术在自动驾驶解决方案中取得了显著成果，但在全自动驾驶汽车上路之前，仍有许多挑战需要克服。在本节中，我们将讨论 DL 和 RL 在自动驾驶应用的挑战、开放研究问题以及未来方向。

到目前为止，DL 方法在表征和识别交通环境中的对象方面已经显示出显着的性能。然而，大多数已发表的论文并未解决自动驾驶在各种交通类型、天气条件和照明条件下实现的问题。在各类不同现实环境背景下，缺乏能够提供真实、大规模、标准化的任务集，同时缺乏对不同气象类型、道路类型对自动驾驶影响的评估模型。

在复杂的驾驶场景中实时感知行人位置是自动驾驶相关研究中的另一个挑战。一般的 RL 方法耗时，计算量巨大，通过长时间计算通常可以获得较好的结果，但是长时间计算并不适用于自动驾驶环境，在该应用下往往需要实时的进行决策，目前 DRL 是主要的解决方式。如 Fan 等人（2022）引入 SGRL 算法，将图卷积引入到传统的 DQN，使用单智能体的训练方式提高训练效率 [80]。此外，Zhe 等人（2022）提出自动驾驶过程并行化，采用自适应更新间隔机制，也降低了延迟 [81]。

此外，由于在现实中采用大量数据实验测试的巨大消耗及其他因素，只有少数研究工作在现实世界中得到了测试。大多数现有作品都是基于模拟的。如何通过可靠的数据集，在模拟环境进行模拟，并转移至现实是另一个目前的研究热点。Paolo 等人（2022）采用 deep response 减小模拟与现实的距离 [82]。Eduardo 等人（2022）则是采用域随机化的转移策略，使用域随机化训练的策略改变车道的平均次数多，但是无法在不增加模拟器保真度的情况下完全缩小现实差距 [83]。未来对现实差距的量化和描述也很会是一个重要的研究方向。

复杂的现实环境往往不是静态的，尤其是城市内道路往来车辆的拥堵程度、意外事件的发生都会导致路线规划的改变，此时与往来车辆或行人形成动态交互就显得具有意义了。Jiseong 等人（2022）使用伪分割标签和动态校准的自动驾驶模拟到真实强化学习，模型成功地在道路上行驶 [84]。

除了在强化学习模型上进行改进和完善，提出统一的关于自动驾驶的性能的评估标准以及完善自动驾驶强化学习平台也是重要的研究学习方向。Linrui 等人（2022）在 SafeRL-Kit 中对各算法进行比较评估，以阐明了它们对安全自动驾驶的功效，但是基于视觉的 AD 的研究仍然较少 [85]。而后 Fei 等人（2022）集成模型训练和仿真评估的自动驾驶，提供具有许多真实世界场景的大规模模拟和训练能力的开放平台 [86]。

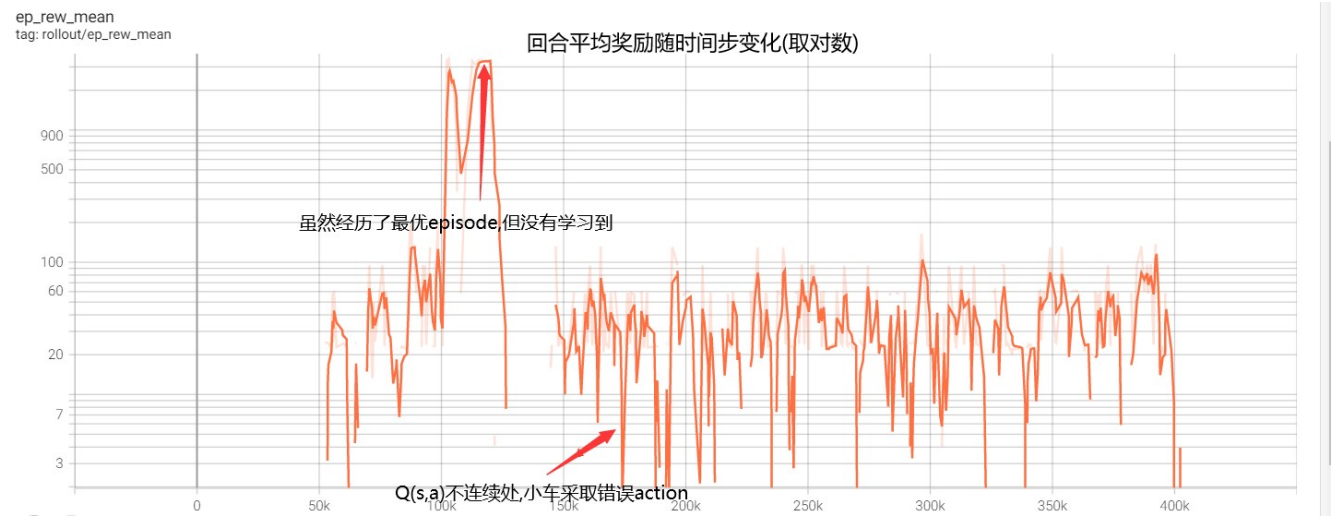

Figure 12. $Q(s,a)$ 不连续处极多，算法发散

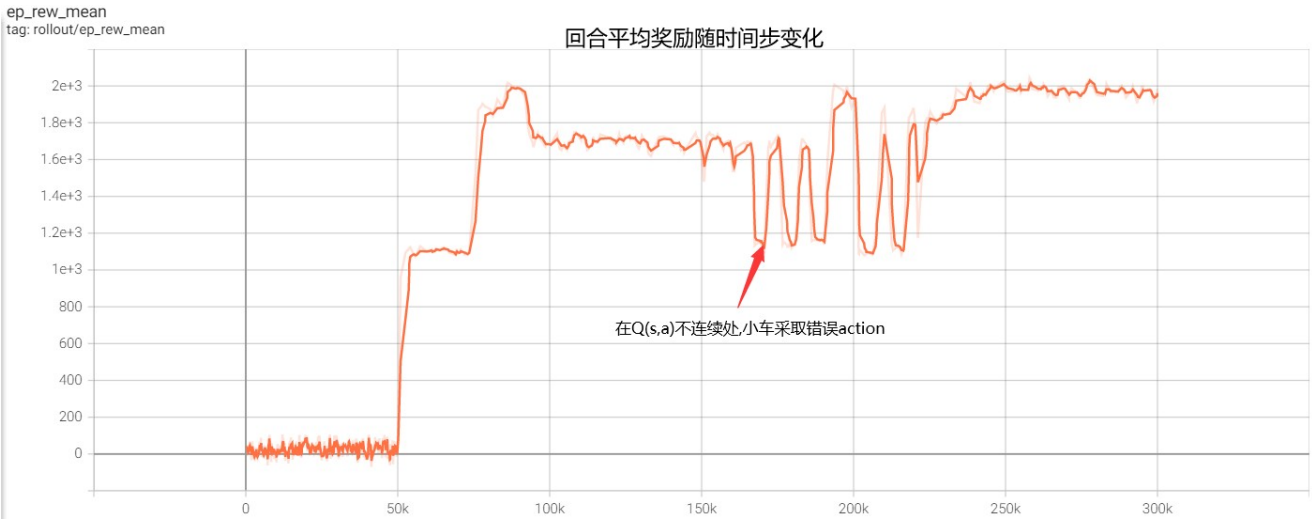

Figure 13. 在尽可能连续化 $Q(s,a)$ 后，算法收敛

8. 总结

在过去的十年里，随着实现车辆完全自动化的努力，自动驾驶车辆的研究得到了越来越多的关注。在我们的道路上引入自动驾驶车辆的目的是超越驾驶辅助技术，使车辆能够在没有人类干预的情况下自动驾驶。自动驾驶汽车可以大大减少与汽车有关的死亡和伤害，并解决各种长期存在的交通挑战，即道路拥堵、旅行延误、停车和安全。智能交通系统、计算系统和人工智能的最新进展刺激了自动驾驶车辆的发展，并为其铺平了道路。这为智能道路、智能交通安全和旅行者的舒适度开辟了新的机会。很少有研究关注 DL 和 RL 技术在解决与场景理解、运动规划、决策、车辆控制和社会行为有关的 AV 挑战方面的作用。

在本文中，我们概述了近年来旨在基于 DL 和 RL 技术解决主要自动驾驶问题的研究工作的文献。我们展望了 DL 和 RL 方法在自动驾驶汽车不同方面的潜力，并引出了在该领域可以实现的目标。我们还讨论了 RL 在自动驾驶中带来的改进，它们可以克服传统 ML 技术的局限性。另外我们还介绍了我们所做的基于 DQN 的自动驾驶泊车的方法。

最后，我们指出了现有的主要研究挑战，并确定了未来开发自动驾驶车辆的可能研究方向。

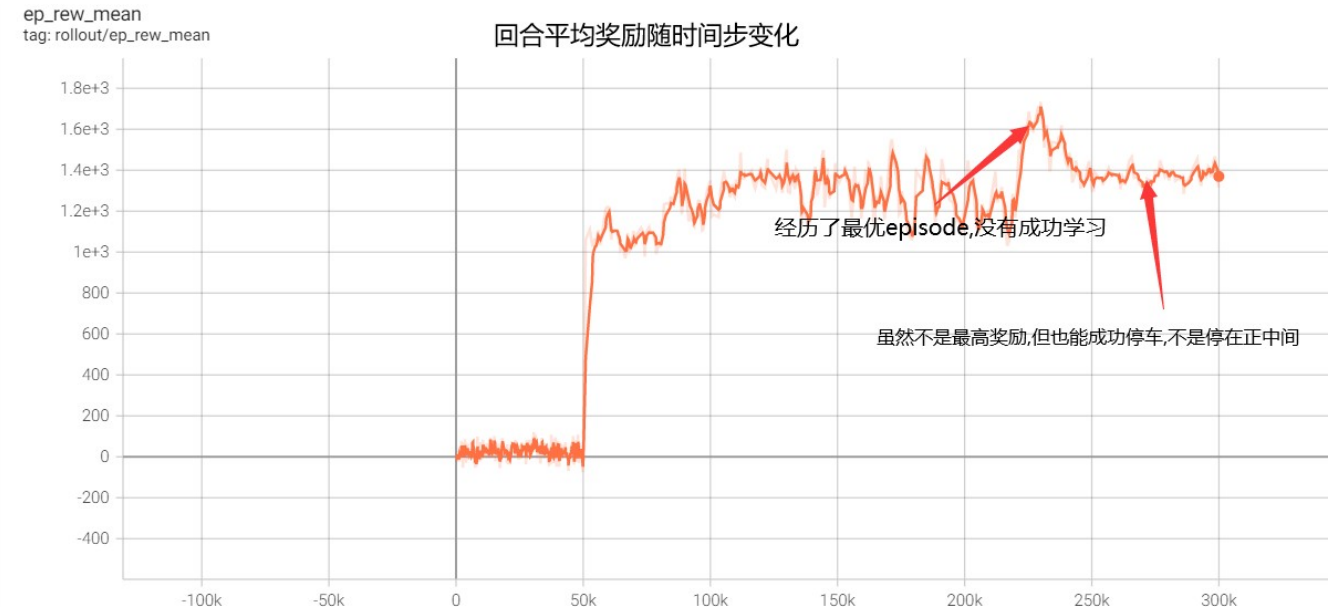

Figure 14. 加入距离传感器

References

[1] Badr Ben Elallid, Nabil Benamar, Abdelhakim Senhaji Hafid, Tajje eddine Rachidi, and N. Mrani. A comprehensive survey on the application of deep and reinforcement learning approaches in autonomous driving. *Journal of King Saud University - Computer and Information Sciences*, 2022.

[2] Duarte Fernandes, António Silva, Rafael Névoa, Cláudia Simões, Dibet Gonzalez, Miguel Guevara, Paulo Novais, João Monteiro, and Pedro Melo-Pinto. Point-cloud based 3d object detection and classification methods for self-driving applications: A survey and taxonomy. *Information Fusion*, 68:161–191, 2021.

[3] Jesús Balado, Joaquín Martínez-Sánchez, Pedro Arias, and Ana Novo. Road environment semantic segmentation with deep learning from mls point cloud data. *Sensors*, 19(16), 2019.

[4] Ankit Laddha, Mehmet Kemal Kocamaz, Luis E. Navarro-Serment, and Martial Hebert. Map-supervised road detection. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 118–123, 2016.

[5] Fuwu Yan, Kewei Wang, Bin Zou, Luqi Tang, Wenbo Li, and Chen Lv. Lidar-based multi-task road perception network for autonomous vehicles. *IEEE Access*, 8:86753–86764, 2020.

[6] Jan-Aike Bolte, Andreas Bar, Daniel Lipinski, and Tim Fingscheidt. Towards corner case detection for autonomous driving. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 438–445, 2019.

[7] Muhammad Sajjad, Muhammad Irfan, Khan Muhammad, Javier Del Ser, Javier Sanchez-Medina, Sergey Andreev, Weiping Ding, and Jong Weon Lee. An efficient and scalable simulation model for autonomous vehicles with economical hardware. 22(3):1718–1732, March 2021.

[8] Lei Wang, Xiaoyun Fan, Jiahao Chen, Jun Cheng, Jun Tan, and Xiaoliang Ma. 3d object detection based on sparse convolution neural network and feature fusion for autonomous driving in smart cities. *Sustainable Cities and Society*, 54:102002, 2020.

[9] Yaran Chen, Dongbin Zhao, Le Lv, and Qichao

Zhang. Multi-task learning for dangerous object detection in autonomous driving. *Information Sciences*, 432:559–571, 2018.

[10] Guofa Li, Yifan Yang, Xingda Qu, Dongpu Cao, and Keqiang Li. A deep learning based image enhancement approach for autonomous driving at night. *Knowledge-Based Systems*, 213:106617, 2021.

[11] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] Yi Zhao, Mingyuan Qi, Xiaohui Li, Yun Meng, Yaxin Yu, and Yuan Dong. P-lpn: Towards real time pedestrian location perception in complex driving scenes. *IEEE Access*, 8:54730–54740, 2020.

[13] Zhengwei Bai, Baigen Cai, Wei ShangGuan, and Linguo Chai. Deep learning based motion planning for autonomous vehicle using spatiotemporal lstm network. In *2018 Chinese Automation Congress (CAC)*, pages 1610–1614, 2018.

[14] Sheng Song, Xuemin Hu, Jin Yu, Liyun Bai, and Long Chen. Learning a deep motion planning model for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1137–1142, 2018.

[15] Xuemin Hu, Bo Tang, Long Chen, Sheng Song, and Xiuchi Tong. Learning a deep cascaded neural network for multiple motion commands prediction in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 22:7585–7596, 2021.

[16] Holger Banzhaf, Paul Sanzenbacher, Ulrich Baumann, and J. Marius Zöllner. Learning to predict ego-vehicle poses for sampling-based nonholonomic motion planning. *IEEE Robotics and Automation Letters*, 4(2):1053–1060, 2019.

[17] Sorin Mihai Grigorescu, Bogdan Trasnea, Liviu Marina, Andrei Vasilcoi, and Tiberiu Cocias. Neurotrajectory: A neuroevolutionary approach to local state trajectory learning for autonomous vehicles. *IEEE Robotics and Automation Letters*, 4(4):3441–3448, 2019.

[18] Yifan Zhang, Jinghuai Zhang, Jindi Zhang, Jianping Wang, Kejie Lu, and Jeff Hong. A novel learning framework for sampling-based motion planning in autonomous driving. In *AAAI*, 2020.

[19] Marius Leordeanu and Iulia Paraicu. Driven by vision: Learning navigation by visual localization and trajectory prediction. *Sensors*, 21(3), 2021.

[20] Yonghwan Jeong, Seonwook Kim, and Kyongsu Yi. Surround vehicle motion prediction using lstm-rnn for motion planning of autonomous vehicles at multi-lane turn intersections. *IEEE Open Journal of Intelligent Transportation Systems*, 1:2–14, 2020.

[21] Liangzhi Li, Kaoru Ota, and Mianxiong Dong. Humanlike driving: Empirical decision-making system for autonomous vehicles. *IEEE Transactions on Vehicular Technology*, 67(8):6814–6823, 2018.

[22] Nicolas Gallardo, Nicholas Gamez, Paul Rad, and Mo Jamshidi. Autonomous decision making for a driver-less car. In *2017 12th System of Systems Engineering Conference (SoSE)*, pages 1–6, 2017.

[23] Dong-Fan Xie, Zhe-Zhe Fang, Bin Jia, and Zhengbing He. A data-driven lane-changing model based on deep learning. *Transportation Research Part C: Emerging Technologies*, 106:41–60, 2019.

[24] Xiao Liu, Jun Liang, and Bing Xu. A deep learning method for lane changing situation assessment and decision making. *IEEE Access*, 7:133749–133759, 2019.

[25] Matthew Strickland, Dani Strickland, Andrew Cross, Brian Goss, Mina Abedi-Varnosfaderani, and Terry West. Low cost current measurement

of three phase cables. In *2018 53rd International Universities Power Engineering Conference (UPEC)*, pages 1–6, 2018.

[26] Weichao Wang, Lei Jiang, Shiran Lin, Hui Fang, and Qinggang Meng. Deep learning-based decision making for autonomous vehicle at roundabout. In *TAROS*, 2020.

[27] Viktor Rausch, Andreas Hansen, Eugen Solowjow, Chang Liu, Edwin J. Kreuzer, and J. Karl Hedrick. Learning a deep neural net policy for end-to-end control of autonomous vehicles. *2017 American Control Conference (ACC)*, pages 4914–4919, 2017.

[28] Shobit Sharma, Girma Tewolde, and Jaerock Kwon. Behavioral cloning for lateral motion control of autonomous vehicles using deep learning. In *2018 IEEE International Conference on Electro/Information Technology (EIT)*, pages 0228–0233, 2018.

[29] Myoung-jae Lee and Young-guk Ha. Autonomous driving control using end-to-end deep learning. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 470–473, 2020.

[30] Ana I. Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5419–5427, 2018.

[31] Dante Mújica-Vargas, Antonio Luna-Álvarez, José de Jesús Rubio, and Blanca Esther Carvajal-Gámez. Noise gradient strategy for an enhanced hybrid convolutional-recurrent deep network to control a self-driving vehicle. *Appl. Soft Comput.*, 92:106258, 2020.

[32] Fatemeh Mohseni, Sergii Voronov, and Erik Frisk. Deep learning model predictive control for autonomous driving in unknown environments. *IFAC-PapersOnLine*, 51(22):447–452, 2018. 12th IFAC Symposium on Robot Control SYROCO 2018.

[33] Péter Szilassy, Balázs Németh, and Péter Gáspár. Design and robustness analysis of autonomous vehicles in intersections. *IFAC-PapersOnLine*, 52(8):321–326, 2019. 10th IFAC Symposium on Intelligent Autonomous Vehicles IAV 2019.

[34] Mohammad K. Al-Sharman, David Murdoch, Dongpu Cao, Chen Lv, Yahya H. Zweiri, Derek Rayside, and William W. Melek. A sensorless state estimation for a safety-oriented cyber-physical system in urban driving: Deep learning approach. *IEEE/CAA Journal of Automatica Sinica*, 8:169–178, 2021.

[35] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2722–2730, 2015.

[36] Guillaume Devineau, Philip Polack, Florent Altché, and Fabien Moutarde. Coupled longitudinal and lateral control of a vehicle using deep learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 642–649, 2018.

[37] Shobit Sharma, Girma S. Tewolde, and Jaerock Kwon. Lateral and longitudinal motion control of autonomous vehicles using deep learning. *2019 IEEE International Conference on Electro Information Technology (EIT)*, pages 1–5, 2019.

[38] Yang Xing, Chen Lv, Huaji Wang, Dongpu Cao, and Efstathios Velenis. An ensemble deep learning approach for driver lane change intention inference. *Transportation Research Part C: Emerging Technologies*, 115:102615, 2020.

[39] R.S. Sutton and A.G. Barto. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 1998.

[40] R. Bellman. An introduction to the theory of dynamic programming. *Rand Corporation Santa Monica Calif*, 1953.

[41] Richard Bellman. A markovian decision process. *Indiana University Mathematics Journal*, 6:679–684, 1957.

[42] P. J. Werbos. Advanced forecasting methods for global crisis warning and models of intelligence. *general systems yearbook*, 1977.

[43] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 2005.

[44] Chris Watkins. Learning from delayed rewards. 1989.

[45] Gavin Adrian Rummery and Mahesan Niranjan. On-line q-learning using connectionist systems. 1994.

[46] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *Computer Science*, 2013.

[47] David Silver, Guy Lever, Nicolas Manfred Otto Heess, Thomas Degris, Daan Wierstra, and Martin A. Riedmiller. Deterministic policy gradient algorithms. In *ICML*, 2014.

[48] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *CoRR*, abs/1502.05477, 2015.

[49] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2016.

[50] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, 2016.

[51] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, L. Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

[52] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *CoRR*, abs/1801.01290, 2018.

[53] Nicolas Heess, Dhruva TB, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, Tom Erez, Ziyu Wang, S. M. Ali Eslami, Martin A. Riedmiller, and David Silver. Emergence of locomotion behaviours in rich environments. *CoRR*, abs/1707.02286, 2017.

[54] Changxi You, Jianbo Lu, Dimitar Filev, and Panagiotis Tsiotras. Highway traffic modeling and decision making for autonomous vehicle using reinforcement learning. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1227–1232, 2018.

[55] Carl-Johan Hoel, Krister Wolff, and Leo Laine. Automated speed and lane change decision making using deep reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2148–2155, 2018.

[56] David Isele, Reza Rahimi, Akansel Cosgun, Kaushik Subramanian, and Kikuo Fujimura. Navigating occluded intersections with autonomous vehicles using deep reinforcement learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2034–2039, 2018.

[57] Takafumi Okuyama, Tad Gonsalves, and Jaychand Upadhay. Autonomous driving system based on deep q learnig. In *2018 International Conference on Intelligent Autonomous Systems (ICoIAS)*, pages 201–205, 2018.

[58] Yingjun Ye, Xiaohui Zhang, and Jian Sun. Automated vehicle' s behavior decision making using deep reinforcement learning and high-fidelity simulation environment. *Transportation Research Part C: Emerging Technologies*, 107:155–170, 2019.

[59] Ming Sun, Weiqiang Zhao, Guanghao Song, Zhigen Nie, Xiaojian Han, and Yang Liu. Ddpg-based decision-making strategy of adaptive cruising for heavy vehicles considering stability. *IEEE Access*, 8:59225–59246, 2020.

[60] Jianyu Chen, Bodi Yuan, and Masayoshi Tomizuka. Model-free deep reinforcement learning for urban autonomous driving. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 2765–2771, 2019.

[61] Zhiqing Huang, Ji Zhang, Rui Tian, and Yanxin Zhang. End-to-end autonomous driving decision based on deep reinforcement learning. In *2019 5th International Conference on Control, Automation and Robotics (ICCAR)*, pages 658–662, 2019.

[62] Meixin Zhu, Xuesong Wang, and Yinhai Wang. Human-like autonomous car-following model with deep reinforcement learning. *Transportation Research Part C: Emerging Technologies*, 97:348–368, 2018.

[63] Abdur R. Fayjie, Sabir Hossain, Doukhi Oualid, and Deok-Jin Lee. Driverless car: Autonomous driving using deep reinforcement learning in urban environment. In *2018 15th International Conference on Ubiquitous Robots (UR)*, pages 896–901, 2018.

[64] Jiren Zhang, Hui Chen, Shaoyu Song, and Fengwei Hu. Reinforcement learning-based motion planning for automatic parking system. *IEEE Access*, 8:154485–154501, 2020.

[65] Mingyu Li and Koichi Hashimoto. Curve set feature-based robust and fast pose estimation algorithm. *Sensors*, 17(8), 2017.

[66] Dong Li, Dongbin Zhao, Qichao Zhang, and Yaran Chen. Reinforcement learning and deep learning based lateral control for autonomous driving [application notes]. *IEEE Computational Intelligence Magazine*, 14(2):83–98, 2019.

[67] Meixin Zhu, Yinhai Wang, Ziyuan Pu, Jingyun Hu, Xuesong Wang, and Ruimin Ke. Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving. *Transportation Research Part C: Emerging Technologies*, 117:102662, 2020.

[68] Alexander Amini, Igor Gilitschenski, Jacob Phillips, Julia Moseyko, Rohan Banerjee, Sertac Karaman, and Daniela Rus. Learning robust control policies for end-to-end autonomous driving from data-driven simulation. *IEEE Robotics and Automation Letters*, 5(2):1143–1150, 2020.

[69] Yuankai Wu, Huachun Tan, Lingqiao Qin, and Bin Ran. Differential variable speed limits control for freeway recurrent bottlenecks via deep actor-critic algorithm. *Transportation Research Part C: Emerging Technologies*, 117:102649, 2020.

[70] Yi Zhang, Ping Sun, Yuhan Yin, Lin Lin, and Xuesong Wang. Human-like autonomous vehicle speed control by deep reinforcement learning with double q-learning. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1251–1256, 2018.

[71] Maximilian Jaritz, Raoul de Charette, Marin Toromanoff, Etienne Perot, and Fawzi Nashashibi. End-to-end race driving with deep reinforcement learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2070–2075, 2018.

[72] Pin Wang, Ching-Yao Chan, and Arnaud de La Fortelle. A reinforcement learning based approach for automated lane change maneuvers. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1379–1384, 2018.

[73] Khaled Saleh, Mohammed Hossny, and Saeid Nahavandi. Long-term recurrent predictive model for intent prediction of pedestrians via inverse reinforcement learning. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, 2018.

[74] Payam Nasernejad, Tarek Sayed, and Rushdi Alsaleh. Modeling pedestrian behavior in pedestrian-vehicle near misses: A continuous gaussian process inverse reinforcement learning (gp-irl) approach. *Accident Analysis Prevention*, 161:106355, 2021.

[75] Lausanne. Webot. `https://www.cyberbotics.com/` Accessed July 10, 2022.

[76] `https://ww2.mathworks.cn/campaigns/offers/reinforcement-learning-with-matlab-intro-ebook.html?s_eid=psn_22567&s_kwcid=AL!8664!88!52653856958!!146695575242&ef_id=YtgeCAAAALWUDVhI:20220804095427:s` Accessed July 10, 2022.

[77] Stable Baselines3. stablebaseline. `https://stable-baselines3.readthedocs.io/en/master/` Accessed July 10, 2022.

[78] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.

[79] `https://ww2.mathworks.cn/help/releases/R2020b/reinforcement-learning/ug/train-ppo-agent-for-automatic-parking-valet.html` Accessed July 10, 2022.

[80] Fan Yang, Xueyuan Li, Qi Liu, Zirui Li, and Xinqing Gao. Generalized single-vehicle-based graph reinforcement learning for decision-making in autonomous driving. *Sensors (Basel, Switzerland)*, 22, 2022.

[81] Zhe Huang, Adam Villaflor, Brian Yang, and Swapnil Pande. Distributed reinforcement learning for autonomous driving. 2022.

[82] Paolo Maramotti, Alessandro Paolo Capasso, Giulio Bacchiani, and Alberto Broggi. Tackling real-world autonomous driving using deep reinforcement learning. *ArXiv*, abs/2207.02162, 2022.

[83] Eduardo Candela, Leandro Parada, Luis Marques, Tiberiu-Andrei Georgescu, Y. Demiris, and Panagiotis Angeloudis. Transferring multi-agent reinforcement learning policies for autonomous driving using sim-to-real. *ArXiv*, abs/2203.11653, 2022.

[84] Jiseong Heo and Hyoung woo Lim. Sim-to-real reinforcement learning for autonomous driving using pseudosegmentation labeling and dynamic calibration. *Journal of Robotics*, 2022.

[85] Linrui Zhang, Q. Zhang, Li Shen, Bo Yuan, and Xueqian Wang. Saferl-kit: Evaluating efficient reinforcement learning methods for safe autonomous driving. *ArXiv*, abs/2206.08528, 2022.

[86] Fei Gao, Peng Geng, Jiaqi Guo, Yuanyuan Liu, Dingfeng Guo, Yabo Su, Jie Zhou, Xiao Wei, Jin Li, and Xu Liu. Apollorl: a reinforcement learning platform for autonomous driving. *ArXiv*, abs/2201.12609, 2022.