语义分割和强化学习在自动驾驶中的应用

摘　要

自动驾驶是一个最近在产业界炙手可热的关键词。无论是与人工智能相关的顶级会议还是各大造车厂商甚至各大投资商都为这个成长初期的蓝海产业下注了美好的未来。然而，自动驾驶技术依然面临着诸多挑战，其中最重要的挑战之一是分析感知数据，以实现安全可靠的行驶控制。本文探讨了语义分割和强化学习在自动驾驶中的应用。语义分割是一种有效的技术，可以将汽车周围环境中的每个物体进行像素级分割，并可以让它们更加详细地表征。此外，强化学习可以确定汽车的最优行驶控制动作，以实现最大的安全性。具体而言，无人驾驶汽车可以通过学习惩罚和奖励系统，从而确定在给定环境中的最优行动，最大限度地减少碰撞风险。本文通过总结高级感知技术在自动驾驶中的应用，以期能更好地应用这些技术，从而构建更加安全可靠的自动驾驶系统。

针对语义分割，本文首先对语义分割的概念进行了介绍，指出了语义分割是讲图像进行分类标签的过程。其次本文提出了语义分割的四个主要方向，半监督语义分割、实例语义分割、视频语义分割和实时语义分割，并对这四个方向进行了优缺点的分析，提出了自动驾驶方向的研究更适合使用实时语义分割。在此基础上，本文对实时语义分割所使用的模型结构进行了介绍，主要可以分为双分支结构 ，编码器-解码器结构和多分支结构这三种。本文介绍了双分支结构中的 BiseNet，BiseNet V2，Fast-SCNN，编码器-解码器结构的 ENet，MSCFNet 还有多分支结构的 ICNet，DFANet。本文还总结了实时语义分割在自动驾驶中的应用中，对实时语义分割做出的改进，且给出了多种语义分割结构的代码。

针对强化学习，本文首先介绍了强化学习的概念，给出了强化学习算法的分类，并分别介绍了基于模型（Model-Based）和无模型的 （Model-Free）学习方法，基于价值（Value-Based）和基于策略的（Policy-Based）学习方法（或两者相结合的 Actor-Critic 学习方法），蒙特卡罗（Monte Carlo）和时间差分（Temporal-Difference）六种强化学习算法。在此基础上又介绍了深度学习，深度机器学习和 DQN,DDPG,PPO,SAC 四种深度神经网络。

针对强化学习在自动驾驶中的应用，本文提出状态、动作空间、奖励的设计至关重要，并给出了一些可以用于设计状态的指标。此外，本文还介绍了运动规划和轨迹优化，模拟器和场景生成工具还有示例学习（LfD）和逆强化学习（IRL）的自动驾驶应用。

最后，本文给出了强化学习在自动驾驶上应用的挑战和未来展望。本文分别提出了本文分别提出了强化学习系统验证，连接模拟-现实的缺口，内在的奖励函数，多智能体强化学习四种改进的方向，并指出了采样效率和模仿的探索问题两个影响强化学习效率的问题，并提出了可能的解决方案。

关键词：自动驾驶　强化学习　计算机视觉　图像分割

一、语义分割

1.1 语义分割简介

语义分割 (Semantic segmentation) 是指将图像中的每个像素链接到类标签的过程。这些标签可能包括人、车、花、家具等。我们可以将语义分割视为像素级别的图像分类。它的一些主要应用是自动驾驶汽车、人机交互、机器人技术和照片编辑/创意工具。例如，语义分割在自动驾驶汽车和机器人技术中非常重要，因为模型理解其运行环境中的上下文非常重要。但在自动驾驶等低延迟操作应用中，由于它们的计算成本相对有限，它需要在精确的时间间隔内做出正确的决定。因此构建一个内存小、推理速度快、精度高的轻量化网络设计，以实现高效的体系结构，从而能够以适当的精度来实现执行。

目前，语义分割已进入与深度学习相融合的阶段。在深度学习阶段，语义分割主要通过卷积神经网络(CNN)实现。在现有的图像语义分割中，主要的方向可以分为四类，分别为：半监督语义分割、实例语义分割、视频语义分割和实时语义分割，其中实时语义分割因其速度快且准确率高，适合在动驾驶领域广泛应用。

方向	文章	优点	缺点
半监督语义分割 (semi-supervised semantic segmentation)	[1]	概述了流行模型的优缺点、测量工具和潜在发展	相关场景的分析不准确
实例语义分割 (instance segmentation)	[2]	讨论研究的现状，以及算法应用领域的优点和缺点	未涵盖任何应用场景
	[3]	讨论了评估指标、数据来源和当前的传统认知	没有对过去几年的关键概念和主流模型架构中的缺陷进行分析
	[4]	简要总结了基本主流观念和预测	使用场景不够充分
视频语义分割 (video semantic segmentation)	[5]	通过对各种数据集和技术的广泛研究，使读者对语义分割所需的深度学习有了基本的了解	缺少精炼的综合分析
实时语义分割(real-time semantic segmentation)	[6]	总结了目前的观点、模型和趋势	没有硬件对实时语义分割的影响的总结

表 1 语义分割分支方向相关的综述

Table 1 A review related to semantic segmentation branching direction

1.2 近年来主流发展趋势

2016 年，研究学者首次提出了关于实时语义分割的模型 ENet[7]，Enet 模型打开了实时语义分割的大门，它是实时语义分割的里程碑，为后续的发展奠定了基础。BiseNet 双分支模型、ICNet 多分支模型让实时语义分割迅速发展。在往后的研究过程中，提出的新兴模型基本都是基于 ENet、BiseNet[8]、ICNet[9]这三种模型思想衍变而成的。近年

来主流的实时语义分割模型主要分为双分支结构 (Bilateral network)，编码器-解码器结构 (Encoder-decoder network) 和多分支结构 (Multi-branch structure)。

Networks	Common structures
Bilateral network	Bisenet, Bisenet V2, Fast-SCNN
Encoder-decoder network	ENet, MSCFNet
Multi-branch structure	ICNet, DFANet

表 2 基于实时语义分割模型特征总结

Table 2 Summary of features based on real-time semantic segmentation model

1.2.1 双分支结构

实时语义分割需要丰富的空间信息和感受野来进行感知，因此研究学者利用双分支结构的一个分支捕捉空间的细节并生成高分辨率的特征表示，而另一个分支获得高级语义上下文信息。并且双分支结构解决了准确性和推理速度的权衡难题，在二者之间取得了平衡。与编码器-解码器架构不同的是，双分支结构还保留了下采样操作后丢失的部分信息。总的来说，此方法依赖一个轻量级的体系结构，该结构结合低层和高层从而提高了效率。

1.2.1.1 BiseNet

双分支结构中最具代表性的网络模型是 BiseNet。如图所示，BiseNet 中的一个分支利用一条步数小的空间路径来保存空间信息，以此来生成高分辨率特征。另一个分支采用快速下采样方法获得足够多的感受野，同时引入一个独特的特征融合模块，用于过滤每个阶段特征的注意细化模块。BiseNet 通过这种方法避免了上采样的繁琐操作，有效

地提高了精度与计算开销。

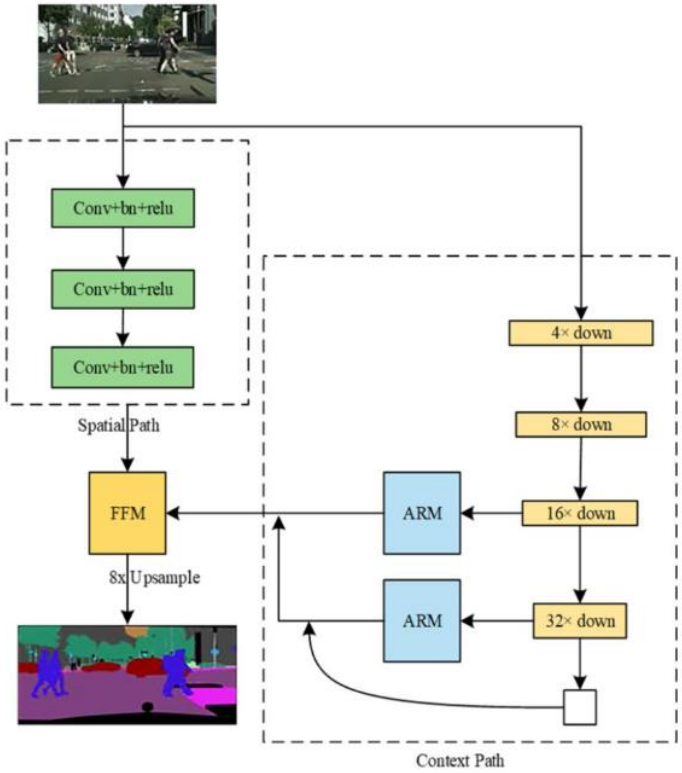

图 1 BiseNet 的双分支结构

Fig.1 BiseNet' s two-branch structure

1.2.1.2 BiseNet V2

为了处理并行分支间的通信，BiseNet V2[10]在推理速度和精度做出了一定的均衡处理。研究学者提出引导聚合层这一模块，引导聚合层利用双向聚合的方法来加强细节与语义分支之间的通信。并提出了 Booster Training Strategy 策略，该策略主要用于插入语义信息中的不同分支，以提高分割性能。

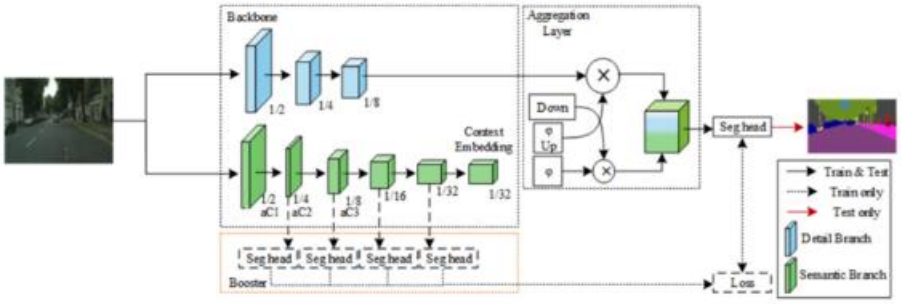

图 2 BiseNet V2 的模型图

Fig.2 A description of BiseNet V2

1.2.1.3 Fast-SCNN

Fast-SCNN[11]在现有的双边基础上，通过“深网络+低分辨率输入”与“浅网络+高分辨率输入”的组合来控制计算复杂度。此结构还引入了“学习下采样”模块，可以同时计算多个分辨率分支下的低层特征，通过全局特征提取器来捕获全局上下文信息，便于进行分割。此外，还利用了深度可分离卷积和剩余瓶颈块来提高运算速度，以此减少计算开销。最后通过特征融合模块对高分辨率和低分辨率进行融合操作。

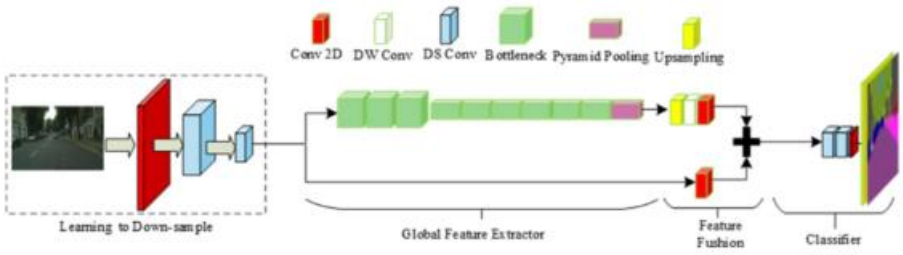

图 3 Fast-SCNN 的框架
Fig.3 The framework of Fast-SCNN

1.2.2 编码器-解码器结构

1.2.2.1 ENet

在移动应用设备中，执行像素级语义分割的能力至关重要，针对这一难点，ENet 模型大体遵循 UNet 的设计理念。但是 ENet 却是不对称的编码器与解码器，小解码器可用于降低内存成本与提高计算速度。ENet 是第一个以实时性能为目标的语义图像分割模型，也是后续研究的里程碑，ENet 的结构如表所示。

Layer	Type	Output Channel	Output Resolution
1	Downsampling block	16	512×256
2	Downsampling block	64	256×128
3-5	3×Non-bt-1D	128	128×64
5-7	2×Conv-module	64	256×128
8	Downsampling block	128	128×64
9	Non-bt-1d (dilated 2)	128	128×64
10	Non-bt-1d (dilated 4)	128	128×64
11	Non-bt-1d (dilated 8)	128	128×64
12	Non-bt-1d (dilated 16)	128	128×64
13	Conv-module (dilated 2)	128	128×64
14	Conv-module (dilated 4)	128	128×64
15	Conv-module (dilated 8)	128	128×64
16	Conv-module (dilated 16)	128	128×64
17	Deconvolution (upsampling)	64	256×128
18-19	2×Non-bt-1D	64	256×128
20	Deconvolution (upsampling)	16	512×256

21-22	2×Non-bt-1D	16	512×256
23	Deconvolution (upsampling)	C	1024×512

表 3 ENet 的网络结构

Table 3 Network architecture of ENet

1.2.2.2 MSCFNet

MSCFNet[12]是一种非对称式结构，编码器由分解卷积块与扩张卷积的非对称残差块组成，解码器使用反卷积代替高昂的计算成本。与其他解码器相对比，保证了计算的简效性和解码信息的最大恢复性。它的分支结构为了抓取多尺度上下文信息，在网络的不同阶段安置高效注意力模块，在最后可将它们结合在一起，增强特征与提高分割精度。MSCFNet 的结构如图所示。

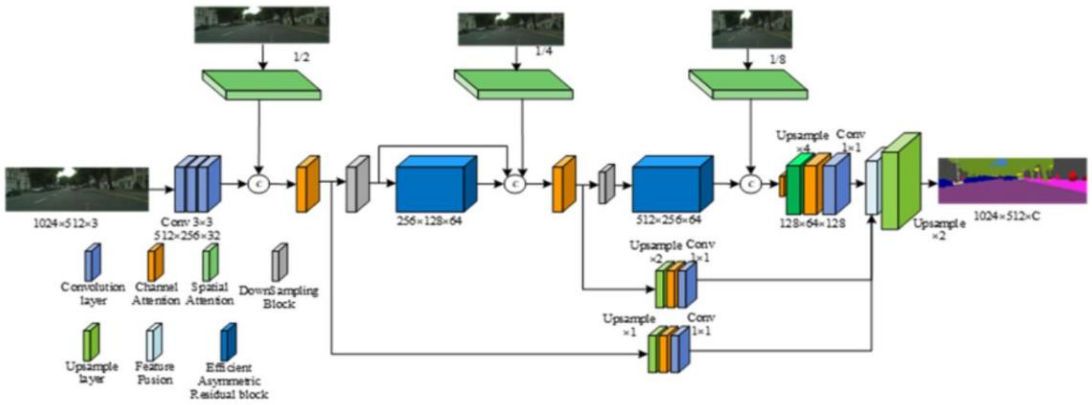

图 4 MSCFNet 的程序结构

Fig.4 Program structure of MSCFNet

1.2.3 多分支架构

1.2.3.1 ICNet

为了平衡精度和速度，ICNet 提出一种多分辨率下保存操作的框架，低分辨率利用 PSPNet 网络的思想，在低分辨率路径中提取语义信息，在中低分辨率之间共享参数，在高分辨率路径提取细节信息。最后通过级联特征融合单元进行融合。ICNet 的结构如图所示。

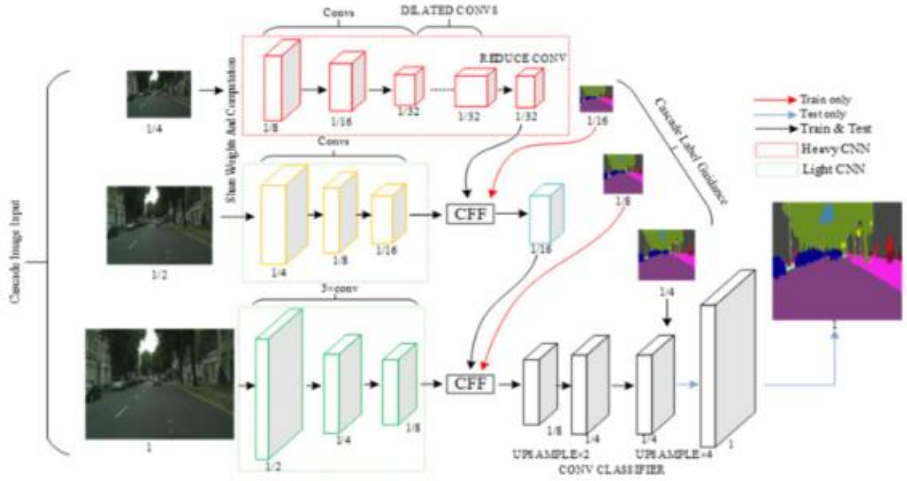

图 5 ICNet 模型的结构

Fig.5 Network architecture of ICNet

1.2.3 DFANet

DFANet[13]利用几个相互连接的编码路径，将高级上下文添加到编码特性中，然后通过子网、子网级联和多尺度特征传播，在很大程度上减少了参数量，且可以获得足够多的感受野与模型学习能力。图为 DFANet 模型的构造图

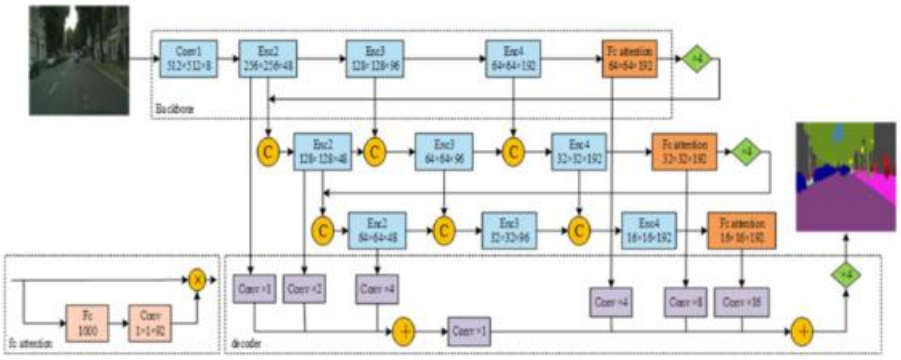

图 6 深度特征聚合网络概述

Fig.6 Overview of the Deep Feature Aggregation Network

1.3 实时语义分割在自动驾驶中的应用场景

在自动驾驶领域中，传入的数据必须被迅速分析并处理，以促进对外部的迅速反应。由于近年来实时语义分割的快速发展，此思想在自动驾驶等领域中得到了广泛的应用。这节将集中讨论实时语义分割应用在某些应用层面。

1.3.1 无人驾驶

自动驾驶必须使用实时语义分割作为视觉效果感知的手段，这样可以提高整个自动驾驶系统的处理速度，减少事故数量。在硬件方面，主要设计了一个用于三维 LIDAR[14]语义分割的二维网络，在此网络的基础上对三维 LIDAR 语义分割算法，通过使用后处理 ABD 滤波器来减少二维网络的分类误差。最后，采用比传感器推理速度更快的三维激光雷达数据语义

分割方法，减少错误的传播。

1.3.2 拍摄图像

目前，实时语义分割提出的模型存在推理速度慢、轮廓信息缺失和语义信息不充足的问题。这些难点让摄像机拍摄的图像无法进行实时语义分割。因此，模型首先使用 MobileNetV2 作为骨干网络[15]，用于提高实时性推理速度。然后提出交叉注意力混合机制来解决轮廓信息缺失的缺点。最后金字塔注意力机制用于解决 CNN 无法捕获长范围语义信息的局限性。

1.3.3 边界感知

为了提高边界感知的能力，在 BiseNet V2 的基础上，BASeNet[16]在语义分支与细节分支进行了一定的改进。为了尽可能利用语义边界信息，语义分支提出 LRA(Lightweight region adaptive)轻量级区域自适应模块。在细节分支上，使用 BA(Boundary-aware)边界感知模块来更好地利用细节边界信息。在模型的最后添加了 EDASPP(Efficient distinctive atrous spatial pyramid pooling)模块，进一步增强了语义与细节的边界特征。

Network Model	Source Code Address
ENet	https://github.com/iArunava/ENet-Real-Time-Semantic-Segmentation
BiseNet	https://github.com/osmr/imgclsmob
BiseNet V2	https://github.com/CoinCheung/BiSeNet
ICNet	https://github.com/hszhao/ICNet
Fast-SCNN	https://github.com/Tramac/Fast-SCNN-pytorch
DFANet	https://github.com/huaifeng1993/DFANet

表 4 实时语义分割模型源代码链接

Table 4 Real-time semantic segmentation model source code link

二、强化学习

2.1 概念

强化学习是以马尔可夫决策过程为基础的理论框架，阐述了在解决动态决策问题中智能体与环的交互过程。强化学习可通过其 5 个主要要素表示 成为 5 元组<S ,A, P, R,γ>，其中 S 表示状态集合，A 定义智能体可采取的动作集合，P 为状态转移矩阵，刻画环境状态的动态变化方式，R 是智能体采取动作后获得的奖励集合，γ(0 ⩽ γ ⩽ 1)表示未来奖励对当前累计奖励的折扣率。强化学习将决策主体建模成能与环境进行动态交互和学习的智能体。
在时刻 t = 1, 2, · · · , T 时，当智能体采取动作a_t，环境会以概率 p(s_t+1|(s_t，a_t) 从当前状态s_t转移到下一个状态s_{t+1}，此时智能体获得奖励r_t。强化学习赋予机器自我不断学习的能力，在不断的与陌生环境的交互过程中，调整或者改变策略从而从环境中获取最大的收益，原理如图 7 所示。

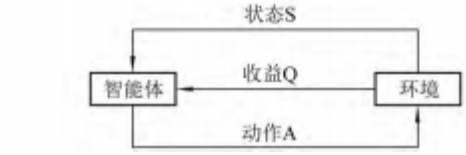

图 7 强化学习原理图

Fig.7 Schematic diagram of reinforcement learning

RL 将决策问题形式化为寻找使预期累计奖励最大化的最优策略智能体通过最大化预期累计奖励学习最优策略。

2.2 强化学习分类

强化学习算法可从多个角度进行分类，包括基于模型（Model-Based）和无模型的（Model-Free）学习方法，基于价值（Value-Based）和基于策略的（Policy-Based）学习方法（或两者相结合的 Actor-Critic 学习方法），蒙特卡罗（Monte Carlo）和时间差分（Temporal-Difference）。

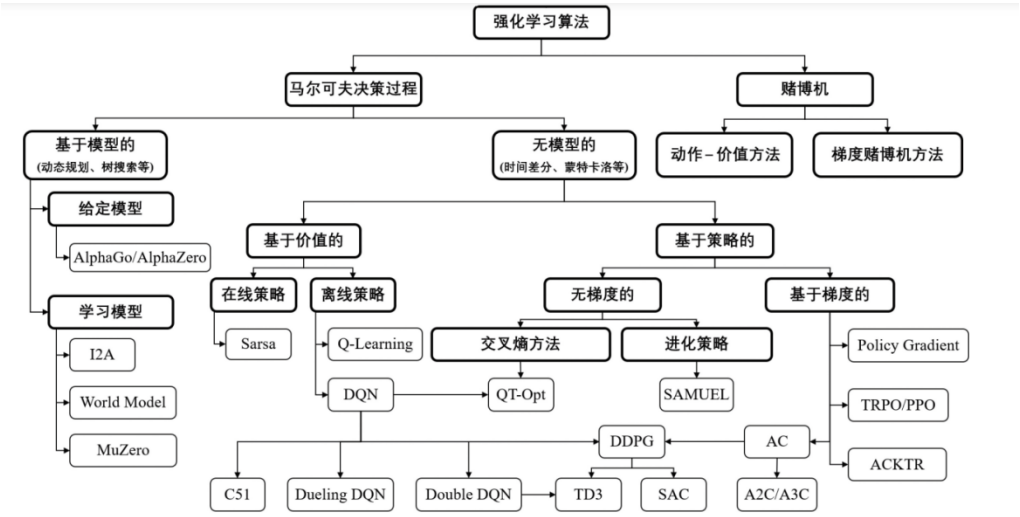

图 8 强化学习算法分类图

Fig.8 Classification diagram of reinforcement learning algorithm

2.2.1 基于模型（Model-Based）和无模型的 （Model-Free）

基于模型的方法可以分为两类：一类是给定（环境）模型（Given the Model）的方法，另一类是学习（环境）模型（Learn the Model）的方法。对于给定模型的方法， 智能体可以直接利用环境模型的奖励函数和状态转移函数。基于模型的方法的主要优点是，通过环境模型可以预测未来的状态和奖励，从而帮助智能体 进行更好的规划。基于模型的方法的缺点在于，存在或构建模型的假设过强。现实问题中 环境的动力学模型可能很复杂，甚至无法显式地表示出来，导致模型通常无法获取。另一方面，在实际应用中，学习得到的模型往往是不准确的，这给智能体训练引入了估计误差，基于带误差模型的策略的评估和提升往往会造成策略在真实环境中失效。

相较之下，无模型的方法不需要构建环境模型。智能体直接与环境交互，并基于探索得到的样本提升其策略性能。与基于模型的方法相比，无模型的方法由于不关心环境模型，无须学习环境模型，也就不存在环境拟合不准确的问题，相对更易于实现和训练。然而，无模型的方法也有其自身的问题。最常见的问题是，有时在真实环境中进行探索的代价是极高的，如巨大的时间消耗、不可逆的设备损耗及安全风险，等等。

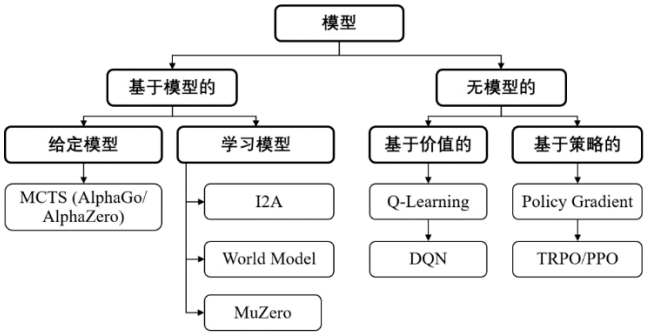

图 9 基于模型的方法和无模型的方法

Fig.9 Model-based approach and model-free approach

2.2.2 基于价值（Value-Based）和基于策略的（Policy-Based）

Policy-Based 的方法直接输出下一步动作的概率，根据概率来选取动作。但不一定概率最高就会选择该动作，还是会从整体进行考虑。适用于非连续和连续的动作。常见的方法有 policy gradients。Value-Based 的方法输出的是动作的价值，选择价值最高的动作。适用于非连续的动作。常见的方法有 Q-learning 和 Sarsa。

在基于值的 RL 方法中，智能体以 Q 值函数衡量状态s_t下动作a_t 的价值，并采用 V 值函数衡量在所有动作a_t下状态s_t的价值,RL 的目标是学习一个最佳的 Q 值函数$\mathrm{Q}*(\mathrm{s},\mathrm{a})$:

$$\mathrm{Q}(\mathrm{s},\mathrm{a}) = \max\{\mathrm{Q}(\mathrm{s},\mathrm{a})\},(s_t = \mathrm{s}, a_t = \mathrm{a})$$

基于价值的方法的优点在于采样效率相对较高，值函数估计方差小，不易陷入局部最优；缺点是它通常不能处理连续动作空间问题，且最终的策略通常为确定性策略而不是概率分布的形式。基于策略的方法直接对策略进行优化，通过对策略迭代更新，实现累积奖励最大化。与基于价值的方法相比，基于策略的方法具有策略参数化简单、收敛速度快的优点，且适用于连续或高 维的动作空间。除了基于价值的方法和基于策略的方法，还有二者的结合，如 Actor-Critic，Actor 根据概率做出动作，Critic 根据动作给出价值，从而加速学习过程。

2.2.3 蒙特卡罗（Monte Carlo）和时间差分（Temporal-Difference）

蒙特卡罗方法必须等到一条轨迹生成（真实值）后才能更新，而时间差分方法在每一步动作执行都可以通过自举法（估计值）及时更新。这种差异将使时间差分方法方法具有更大的偏差，而使蒙特卡罗方法方法具有更大的方差。

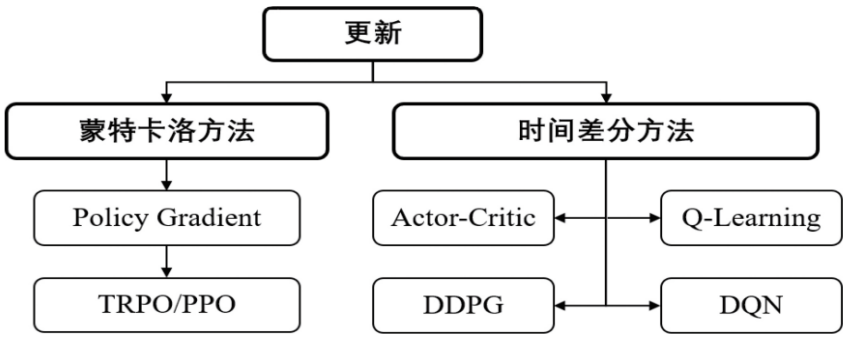

图 10 蒙特卡罗方法和时间差分方法

Fig.10 Monte Carlo method and temporal-difference method

2.3 深度机器学习

2.3.1 深度学习

深度学习是机器学习中一种基于对数据进行表征学习的算法 ,在计算机视觉 、语音识别 、自然语言处理 、音频识别与生物信息学等领域取得重要突破 并获得高度关注。典型的深度学习神经网络模型网络结构有卷积神经网络(convolutionalneuralnet- work,CNN)、深度神经网络和循环神经网络等。随 着深度学习的继续发展 ,受到猫脑视觉皮层研究中局部感知的启发 ,CNN 通过稀疏连接、权值共享和下采样等技术改进深度神经网络 ,很大程度上降低运算复杂度。

2.3.2 深度机器学习

由于传统的 RL 模型在处理高维数据中具有局限性， DL 可以与 RL 相结合实现更好的决策效果。深度 Q 网络(Deep Q Network, DQN)利用神经网络在高维空间学习中的优势，引入神经网络作为值 函数逼近器，计算最大化累计奖励的最优解。DQN 采用带参数 e 的神经网络估计动作价值 Q(s, a)，并基于经验回放机制进行学习，通过最小化误差损失不断逼近最优解:

$$\mathrm{L}(\mathrm{e}) = \mathrm{Es,a}\left[(\mathrm{r} + \gamma\max\{\mathrm{Q}(\mathrm{s}',\mathrm{a}';\mathrm{e})\} - \mathrm{Q}(\mathrm{s},\mathrm{a};\mathrm{e}))^2\right]$$

其中 e 表示目标网络的参数，它是周期性地从 DQN 网络中复制而来，并在一定的迭代过程中保持不变。但是传统的 DQN 模型存在高估 Q 值的问题，容易跳过最优解学习到次优解，导致模型效果不佳。为了缓解这一问题，在 DQN 的基础上引入很多变体模型，包括 Double DQN、Dueling DQN、C51 DQN、Bootstrapped DQN 和 Rainbow DQN 等。除了以上这些基于价值的深度强化学习算法，基于策略的相关算法，包括 DPG、DDPG、A3C、TRPO、PPO 和 SAC 等，也在不同领域与任务中表现出良好效果。

强化学习的基本思想是智能体在与环境的交互过程中迭代地学习最优决策。智能体接收从环境中获得的环境状态s_t,再根据该环境状态决策出动作a_t,动作a_t作用于环境后获得 奖励值r_t ；在下一时刻环境发生变化，智能体感知新的环境状态r_{t+1},再做出相应决策动作a_{t+1}。智能体的目标是要 在交互过程中学得一个最优策略，以使期望的长期累计奖励最大化。

一个智能体的强化学习过程可视为一个马尔可夫决策过程 (MDP)，由五元组$<\mathrm{S},\mathrm{A},\mathrm{r},\mathrm{P},\gamma>$表示，其中:

1．状态空间 S 表示环境状态的集合;

2．动作空间 A 表示智能体能够选择的所有动作的集合;

3．奖励函数 r: $S \times A \rightarrow R$, r(s_t, a_t)表示智能体根据状态s_t决策出动作a_t后，所获得的即时奖励值,记为r_t;

4．状态转移概率分布 P~S×A→S，P(s_t, a_t, s_{t+1})表示智能体在状态s_t下，执行决策动作a_t后,下一时刻环境转移到状态s_{t+1}的概率;

5．折扣因子γ: (0 ⩽ γ ⩽ 1)，未来奖励的折扣系数。

在大规模的状态空间中，传统的强化学习无法计算出价值函数和策略函数，而结合深度学习，则可以利用神经网络来拟合强化学习中的价值函数和策略函数，即输入是环境的状态数据，输出是价值函数值或策略函数值，如基于价值的 DQN 及其变体与基于策略的 A3C 等。

2.4 深度神经网络分类

深度学习强大的特征提取能力,结合强化学习的自主决 策能力形成深度强化学习,使强

化学习不再受数据空间维度问题,得以应用于高维、复杂的控制系统。根据优化过程中动作选取方式的不同,深度强化学习可以分为基于值的深度强化学习方法和基于策略梯度的深度强化学习方法。

2.4.1 基于值的深度强化学习方法

基于价值的深度强化学习方法通过准确估计状态-动作的价值函数，选择最大值对应的动作，隐含地获得确定性策略。深度神经网络用于逼近价值函数或动作价值函数，应用范围扩展到高维问题和连续空间问题。Watkins 等人提出的 Q-learning 算法估计 Q 值函数，在当前状态下执行一个动作，然后切换到下一个状态,智能体获得环境奖励并更新 Q 值函数。在有限的状态-动作空间中，Q-learning 算法可以收敛到最优 Q 值函数。 Mnih 等人首先提出结合深度神经网络和 Q 学习的 DQN 算法，利用卷积神经网络逼近 Q 值，然后提出利用目标网络和经验回放来稳定 DQN 的学习过程。

但是 DQN 每次更新都会最大化目标网络，导致高估动作价值函数的问题。 哈瑟尔特等采用双网络结构对当前网络选择最优动作，目标网络对选择的动作进行评价，将动作选择与策略评价分离，减少高估的可能性。 Wang 等提出了一种对抗架构 DQN 算法，直接估计状态值函数和动作优势函数，保证当前状态下每个动作的优势函数的相对顺序不变，在去除冗余自由度的同时减小 Q 值范围，并提高了算法的稳定性。Nail 等提出了一种用于深度强化学习的大规模分布式架构，充分利用计算资源。 这类算法只能处理有限的状态-动作空间问题，难以处理复杂环境，在学习过程中容易出现过拟合和收敛性差，因此适用于离散动作空间的深度强化学习。

2.4.2 基于策略梯度的深度强化学习方法

策略梯度算法将策略参数化，使用神经网络的权重参数作为值函数的参数，可以通过分析状态直接输出下一步采取各种动作的概率，然后根据相应的被选中概率。最经典的策略梯度算法 REINFORCE 采用蒙特卡洛方法计算状态值函数，逼近备选策略梯度的值函数。由于蒙特卡洛策略梯度法基于完整经验更新值函数参数，模型的学习效率较低，信赖域策略优化算法 TPRO 和在线学习的近端策略优化算法根据经验或自适应方法选择超参数，使更新步长限制在一定范围内，保证持续获取更好的策略，防止策略崩溃。

TPRO 和 PPO 算法每次更新策略都会抽取大量样本进行训练，需要大量的算力来保证算法收敛，难以应用于大规模场景下的强化学习过程。 Lillicrap 提出了深度确定性策略梯度算法 DDPG，利用非线性函数逼近值函数，使函数稳定收敛，解决了 Q 函数更新的发散问题。同时，使用经验回放机制进行批量学习，让训练过程更加稳定。 为了解决 DDPG 中 Q 值高估的问题以及超参数等参数调整的脆弱性，Fujimoto 等人提出了 TD3 算法，可以缓解动作值高估的影响，消除方差累积的问题，使得训练过程更稳定。同时避免 DDPG 中可能发生的功能故障。

与基于价值的深度强化学习方法相比，基于策略的强化学习方法具有更好的收敛性，尤其是在使用神经网络逼近函数时，可以轻松处理大量甚至连续的状态-动作空间。但其缺点是算法方差大、收敛速度慢、学习步长难以确定。

2.4.3 主要深度神经网络

2.4.3.1 DQN 算法

2013 年,Mnih 等将深度卷积神经网络和 Q-Learning 算法相结合,提出 DQN 算法,利用神经网络的强大表征能力,把强化学习中的状态作为神经网络模型的输入,输出的是每个动作对

应的 Q 值,得到将要执行的动作。通过使用神经网络的近似函数来解决大规模马尔可夫决策过程任务,使用一个参数为 θ 的动作值函数 Q(s,a,θ)来逼近最优动作值函数 Q*(s,a)。DQN 算法通过 Q-Learning 算法构建可优化的损失函数,即估计网络的输出与目标 Q 值之间平方误差的期望。在获得损失函数后,直接采用梯度下降算法对卷积神经网络模型损失函数 L(θ)的权重参数 θ 进行迭代更新,大大提升算法的稳定性。DQN 算法取得成功的一个关键是固定 Q 目标。目标网络和估计网络的权重参数不同,主要原因是 Q-Learning 算法中估计 Q 值和目标 Q 值使用相同参数模型,容易造成模型振荡和发散。因此,DQN 算法使用 2 个卷积神经网络进行学习,其中估计网络 Q(s,a,θ)用来评估当前状态-动作对的价值,目标网络 Q(s,a,θ')用来评估下一状态的状态-动作对的价值,经过多轮迭代后才将 估计网络的参数 θ 复制给目标网络中的参数 θ'。DQN 算法使用旧的网络参数 θ'评估一个经验样本中下一时间步的状态 Q 值,且只在离散的多个时间步间隔上更新旧的网络参数 θ',在一段时间内进行目标网络 Q 值不变的稳定训练,以此来降低估计 Q 值和目标 Q 值的相关性,从而使得估计误差得到更 好的控制。另一个关键是经验回放机制。为克服训练样本之间的相关性,提高算法的稳定性,DQN 使用经验回放来提高数据效率。基本过程:智能体在状态 s 下执行动作 a,到达下一时间步状态 s',获得相应的奖励 r,T 为一个布尔 值,表示状态 s'是否为终止状态,形成经验样本五元组(s,a,r,s',T),将经验样本五元组存储到经验池,当需要进行网络训练时,从经验池中随机抽取小批量的数据进行训练,若经验池已满,则第一个样本会自动被新样本替换。

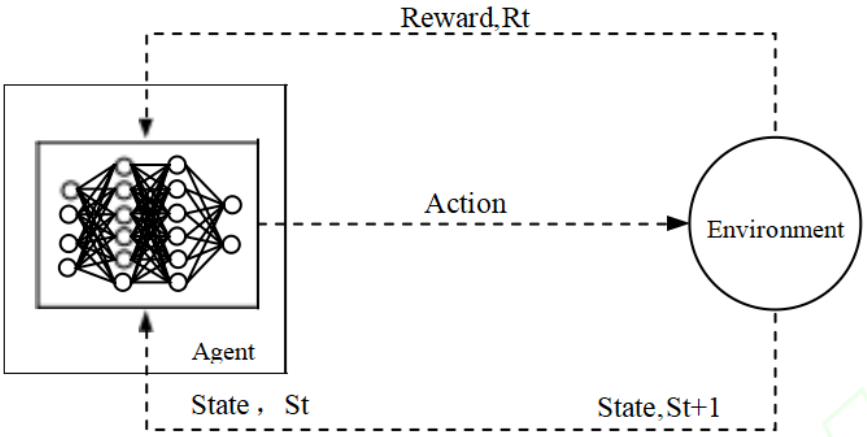

图 11 DQN 基本结构图

Fig.11 Basic structure of DQN

2.4.3.2 DDPG 算法

之前 PG 算法的思路是在累积收益(Cumulative Return)和策略之间建立一个关系函数,然后调整策略以追求更大的收益。DPG 算法可以看作是 Q-learning 的连续动作空间版本。它的思想是直接使用 critic（Q 函数）寻找可能的最优决策，然后用找到的最优决策来优化策略。Function（actor）策略调整完全依赖于 critic 而不考虑实际收益。不过，虽然 DDPG 在实验中有着非常出色的表现，但在实际训练中对各种超参数非常敏感，所以 DDPG 在各种 benchmark 上的优异表现其实是精心打造出来的，难以大量使用具体问题 .

2.4.3.3 PPO 算法

PPO 是 TRPO（Trust Region Policy Optimization）的简化版本，两者的目标都是让 PG 算法在优化过程中性能单调增加，并且增加的幅度尽可能大。 PPO 同样使用 AC 框架，但

比 DPG 更接近于传统的 PG 算法。它使用随机分布式策略函数（Stochastic Policy）采样，将得到的样本作为最终的执行动作，因此天生具有探索环境的能力，无需为了探索环境而在决策中加入扰动。 PPO 的基本思想和 PG 算法是一致的，就是直接根据策略的盈利来调整策略。PPO 的重点会放在 actor 上，critic 只是作为状态的预测器（在这个状态下获得的预期回报）。 工具和策略的调整基准在于获得的收益，而不是评论家的衍生品。 PPO 的缺陷在于依赖重要性采样的 off-policy 算法在面对过大的 policy difference 时会无能为力（被训练的 policy 与实际 policy 在与环境交互时的差异过大），并且在训练过程中需要保证训练的产生。数据的 policy 与当前的 training policy 一致，很难复用过去 policy 生成的数据。

2.4.3.4 SAC 算法

2018 年提出了一种更稳定的离线策略算法 Soft Actor-Critic (SAC)。 SAC 的前身是 Soft Q-learning，都属于最大熵强化学习的范畴。 Soft Q-learning 没有明确的策略函数，而是使用了函数 Q 的玻尔兹曼分布，在连续空间求解起来很麻烦。 所以 SAC 提出用一个 Actor 来表示策略函数来解决这个问题。 目前，在无模型强化学习算法中，SAC 是一种非常高效的算法，它学习了一个随机策略，并在许多标准环境中取得了领先的成绩。 在 SAC 算法中，我们对两个动作价值函数 Critic 和一个策略函数 Actor 进行建模。 基于 Double DQN 的思想，SAC 使用了两个 Critic 网络，但每次使用 Critic 网络时，都会选择一个价值较小的网络，从而缓解价值高估的问题。

三、强化学习在自动驾驶中的应用

强化学习在自动驾驶任务中的应用主要包括：控制器优化、路径规划和轨迹优化、运动规划和动态路径规划、复杂导航任务下的高级驾驶策略开发、快速路基于场景的策略学习，路口通行，汇入和离开车流，使用数据反向强化学习奖励用于交通参与者（如行人、车辆）意图预测，以及最终学习确保安全和评估风险的策略。

3.1 状态空间、动作空间和奖励

要想在自动驾驶中应用深度强化学习（DRL），Agent 的状态、动作空间、奖励的设计至关重要。

自动驾驶车辆中常用的状态空间特征包括：位置、车头朝向、速度、以及车辆周边障碍。为避免状态空间维度波动，一般采用以本车为原点的笛卡尔或极坐标系来构建状态空间。进一步使用环境信息，如车道数量、路径曲率、车辆的路径、碰撞时间（TTC）、以及场景信息如交通法规和信号灯位置等对车辆的当前状态进行增强。

使用摄像头、雷达的原始探测数据能够更精准的构建空间信息，而使用简化的信息可以降低状态空间的复杂度。结合以上两点，2D 鸟瞰视角（BEV），是传感器不可知的（sensor agnostic），而又接近实际场景空间。下图俯瞰图展示了占用格、历史和投射（预测）轨迹，以及场景语义信息，如交通信号灯等。

Fig. 4. Bird Eye View (BEV) 2D representation of a driving scene. Left demonstrates an occupancy grid. Right shows the combination of semantic information (traffic lights) with past (red) and projected (green) trajectories. The ego car is represented by a green rectangle in both images.

图 12 BEV 俯瞰图

Fig.12 BEV Aerial View

一个车辆策略需要控制一定数量不同的执行器（actuators），它们可以是连续的或离散的。连续值执行器包括转向角、油门和刹车。如挡位的执行器是离散的。为减少复杂度，并使仅适用于离散动作空间的 DRL 算法可用（如 DQN），需要将动作空间统一离散化（如将连续执行器如转向角、油门和刹车的范围等分）。由于在实践中，多数转向角接近于中心，还可以使用 log 空间离散化。然而离散化也会带来缺陷，执行器划分的步长过大会导致轨迹波动或不稳定，而步长过小会导致学习成本过高。作为替代，可以使用可处理连续值的 DRL 算法直接学习策略（如 DDPG）也可用于简化动作选取流程，其中智能体选择选项（option）而不是低级别的动作（action）。这种选项表示了一种子策略，可以扩展到多个时间步上的动作。

自动驾驶 DRL 智能体的奖励函数设计仍然没有定论。自动驾驶任务的评判标准包含如：向目标已行驶距离、本车速度、本车保持静止、与其他道路用户或场景物体的碰撞、与路肩刮擦、车道保持、行驶舒适平稳、避免急加速减速或转向，以及遵守交通规则。

TABLE I

LIST OF AD TASKS THAT REQUIRE D(RL) TO LEARN A POLICY OR BEHAVIOR

AD Task	(D)RL method & description	Improvements & Tradeoffs
Lane Keep	1. Authors [82] propose a DRL system for discrete actions (DQN) and continuous actions (DDAC) using the TORCS simulator (see Table V-C) 2. Authors [83] learn discretised and continuous policies using DQNs and Deep Deterministic Actor Critic (DDAC) to follow the lane and maximize average velocity.	1. This study concludes that using continuous actions provide smoother trajectories, though on the negative side lead to more restricted termination conditions & slower convergence time to learn. 2. Removing memory replay in DQNs help for faster convergence & better performance. The one hot encoding of action space resulted in abrupt steering control. While DDAC's continuous policy helps smooth the actions and provides better performance.
Lane Change	Authors [84] use Q-learning to learn a policy for ego-vehicle to perform no operation, lane change to left/right, accelerate/decelerate.	This approach is more robust compared to traditional approaches which consist in defining fixed way points, velocity profiles and curvature of path to be followed by the ego vehicle.
Ramp Merging	Authors [85] propose recurrent architectures namely LSTMs to model longterm dependencies for ego vehicles ramp merging into a highway.	Past history of the state information is used to perform the merge more robustly.
Overtaking	Authors [86] propose Multi-goal RL policy that is learnt by Q-Learning or Double action Q-Learning(DAQL) is employed to determine individual action decisions based on whether the other vehicle interacts with the agent for that particular goal.	Improved speed for lane keeping and overtaking with collision avoidance.
Intersections	Authors use DQN to evalute the Q-value for state-action pairs to negotiate intersection [87],	Creep-Go actions defined by authors enables the vehicle to maneuver intersections with restricted spaces and visibility more safely
Motion Planning	Authors [88] propose an improved A^* algorithm to learn a heuristic function using deep neural netowrks over image-based input obstacle map	Smooth control behavior of vehicle and better peformance compared to multi-step DQN

图 13 自动驾驶任务

Fig.13 Automatic driving task

3.2 运动规划和轨迹优化

运动规划的任务是建立目标和终点间的路径。车辆动态环境中的路径规划任务是整个自动驾驶流程的核心，完成例如通过路口、高速汇车等任务。近期的一些研究包办了真实世界多种交互驾驶场景下不同交通参与者的的行为。研究人员展示了 DRL（DDPG）在全尺寸车辆上的自动驾驶应用。在强化学习模型部署到车载电脑前，系统首先在模拟器上训练，并学会了沿路径前进，成功完成了 230 米的真实路试。基于模型的 DRL 算法被提出用于直接从原始像素输入中学习模型和策略。在 S. Chiappa 团队的研究中，深度神经网络被用于在模拟环境中数百个时间步上生成预测。强化学习同样适合控制。

3.3 模拟器和场景生成工具

自动驾驶数据集为有监督的学习提供了多种形态的特征-标签对。强化学习需要一个环境，用于发掘状态-动作对（State-Action），这个环境可以对车辆（Agent）、环境（Environment）建模，用于模拟车辆的动力学和环境交互，随机模拟车辆的实际行驶过程。下表包含了几种感知模拟器，能够模拟摄像头、激光雷达、雷达等从环境获取信息（视野）的过程。

TABLE II

SIMULATORS FOR RL APPLICATIONS IN ADVANCED DRIVING ASSISTANCE SYSTEMS (ADAS) AND AUTONOMOUS DRIVING

Simulator	Description
CARLA [78]	Urban simulator, Camera & LIDAR streams, with depth & semantic segmentation, Location information
TORCS [96]	Racing Simulator, Camera stream, agent positions, testing control policies for vehicles
AIRSIM [97]	Camera stream with depth and semantic segmentation, support for drones
GAZEBO (ROS) [98]	Multi-robot physics simulator employed for path planning & vehicle control in complex 2D & 3D maps
SUMO [99]	Macro-scale modelling of traffic in cities motion planning simulators are used
DeepDrive [100]	Driving simulator based on unreal, providing multi-camera (eight) stream with depth
Constellation [101]	NVIDIA DRIVE Constellation^{TM} simulates camera, LIDAR & radar for AD (Proprietary)
MADRaS [102]	Multi-Agent Autonomous Driving Simulator built on top of TORCS
Flow [103]	Multi-Agent Traffic Control Simulator built on top of SUMO
Highway-env [104]	A gym-based environment that provides a simulator for highway based road topologies
Carcraft	Waymo's simulation environment (Proprietary)

图 14 一些常用的模拟器

Fig.14 Some commonly used simulators

在进行实际路试之前，学习完成的驾驶策略需要在模拟环境中进行压力测试。M. Cutler 等人提出了多精度强化学习（Multi fidelity reinforcement learning, MFRL）框架，其中可以使用多个模拟器，训练验证 RL 算法、表征状态和动力学的模拟器精度逐级增加（相应成本同样增加），并使用一台远程控制车辆，以寻找真实世界中成本最低的样本，并接近最优的策略。

3.4 从示例学习（LfD）和逆强化学习（IRL）的自动驾驶应用

早期驾驶车辆 Behavior Cloning（BC）研究展示了可以从示例中学习（LfD）的智能体，试图模仿人类专家的行为。BC 是典型的有监督学习，因此，BC 难以适应新的、未遇到过的情景。M. Bojarski 提出在自动驾驶领域采用端到端的卷积神经网络。该 CNN 被训练到可以根据车辆单目前置摄像头的原始像素直接映射到转向命令。使用一个相对小的人类专家数据集，该系统学会在市区和快速路上沿车道行驶，无论是否有车道标识。该网络无需特别精细的训练，即学会了成功检测道路的图像表征。也有研究人员提出了使用 Maximum Entropy Inverse RL，将人类司机作为示例，学习舒适驾驶轨迹优化。也有人在学习拟人变道动作中，使用 DQN 作为 IRL 中的精炼步骤提取奖励。

四、强化学习自动驾驶的挑战和未来展望

4.1 强化学习系统验证

RL 算法实际部署中的基础代码多样，超参数值过多且不同数据和模型间变化较大，可解释性和泛化性能差。一些学者提出在模拟器中生成罕见的挑战性驾驶场景，这种对抗性的场景是通过参数化行人和其他道路车辆行为自动生成的。将这些场景加入到模仿学习的训练集中，可以增加安全性。

4.2 连接模拟-现实的缺口

模拟-现实的转换学习是一个活跃领域，模拟是大量低成本有标注数据的来源。在自动驾驶领域，有学者 X. Pan 等人应用驾驶环境的模拟-现实转换图像训练了 A3C 智能体，然后将训练好的策略在真实驾驶数据集中评估。

4.3 采样效率

采样效率是强化学习的关键挑战之一。学习过程需要大量的样本和模拟来迭代出一个合理的策略。当有价值的经验获得成本过高或风险太大时，这一问题更加突出。在自动驾驶的案例中，由于延迟和稀疏奖励、以及在大状态空间中的观察分布不平衡，采样效率是一大难题。设计奖励（reward shaping）通过设计更频繁的奖励函数，鼓励智能体从少数样本中更快学习，使其学到中间目标。

4.4 模仿的探索问题

在示教学习中，Agent 利用确定的已知的的轨迹，但是实际情况的状态分布往往不能包含 Agent 模型在实际测试中可能遇到的所有状态。此外，模仿学习假设所有动作是独立同分布的（i.i.d.）。解决方案之一是使用数据聚集（Data Aggregation），执行训练好的策略，提取观察-动作对，再人工进行标签，再聚集到原有的人工观察-动作数据集中。这样，从示例和已训练策略收集到的训练案例进一步探索了更多有价值状态，解决了探索缺乏的问题。

4.5 内在的奖励函数

在人为设计的模拟环境中，往往 reward 比较直观且容易确定。而在真实世界自动驾驶却不同，设计一个好的奖励函数是引导 Agent 完成有效学习的关键。最常用的解决方案是设计奖励（reward shaping）。在没有外在奖励或专家演示的情形时，智能体可以采用内在奖励或内在激励机制来评价其动作好坏。

4.6 多智能体强化学习

自动驾驶本质上是多智能体任务，除了本车由智能体控制，模拟和真实世界中同样有其他参与者，如行人、自行车和其他车辆。在 MARL 领域已有一些前沿方法。

一个 MARL 可能发挥大作用的领域是高等级决策和自动驾驶车辆组团协同，如高速超车场景，或无信号灯路口同行。另一领域是开发对抗性智能体以进行自动驾驶测试策略，即

智能体在模拟器中无规律或违反交规地控制其他车辆，以暴露自动驾驶策略的缺点。最后，MARL 还将在开发自动驾驶安全策略领域发挥潜在的重要作用。

五、结论

强化学习在现实世界自动驾驶应用领域仍是一个活跃的新兴研究领域。尽管成功的商业应用还很少，有关文献也不多，而且缺乏大规模的公共数据集。自动驾驶场景包含了多个互动的智能体，它们之间需要进行协商并动态决策，这使得强化学习可以应用于该领域。然而，达到成熟的解决方案之前还有很多挑战需要解决。包括如何验证基于强化学习（RL）系统性能、模拟和现实的差距、采样效率、好的奖励函数设计以及如何将安全因素纳入到自动驾驶智能体 RL 系统的决策中等。

六、参考文献

[1]. Zhang, Man, et al. "A survey of semi-and weakly supervised semantic segmentation of images." *Artificial Intelligence Review* 53 (2020): 4259-4288.

[2]. 苏丽，孙雨鑫，and 苑守正. "基于深度学习的实例分割研究综述." *智能系统学报* 17.1 (2021): 16-31.

[3]. 李晓筱, et al. "基于深度学习的实例分割研究进展." *计算机工程与应用* 57.9 (2021): 60-67.

[4]. 王可，沈川贵, and 罗孟华. "基于深度学习的图像语义分割方法综述." *信息技术与信息化* (2022).

[5]. Garcia-Garcia, Alberto, et al. "A survey on deep learning techniques for image and video semantic segmentation." *Applied Soft Computing* 70 (2018): 41-65.

[6]. Papadeas, Ilias, et al. "Real-time semantic image segmentation with deep learning for autonomous driving: A survey." *Applied Sciences* 11.19 (2021): 8802.

[7]. Paszke, Adam, et al. "Enet: A deep neural network architecture for real-time semantic segmentation." *arXiv preprint arXiv:1606.02147* (2016).

[8]. Yu, Changqian, et al. "Bisenet: Bilateral segmentation network for real-time semantic segmentation." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[9]. Zhao, Hengshuang, et al. "Icnet for real-time semantic segmentation on high-resolution images." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[10]. Yu, Changqian, et al. "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation." *International Journal of Computer Vision* 129 (2021): 3051-3068.

[11]. Poudel, Rudra PK, Stephan Liwicki, and Roberto Cipolla. "Fast-scnn: Fast semantic segmentation network." *arXiv preprint arXiv:1902.04502* (2019).

[12]. Gao, Guangwei, et al. "Mscfnet: a lightweight network with multi-scale context fusion for real-time semantic segmentation." *IEEE Transactions on Intelligent Transportation Systems* 23.12 (2021): 25489-25499.

[13]. Li, Hanchao, et al. "Dfanet: Deep feature aggregation for real-time semantic segmentation." *Proceedings of the IEEE/CVF conference on computer vision and*

pattern recognition. 2019.

[14]. Zhang, Chris, Wenjie Luo, and Raquel Urtasun. "Efficient convolutions for real-time semantic segmentation of 3d point clouds." *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018.

[15]. WANG, Juan, et al. "Real-time semantic segmentation method based on squeezing and refining network." *Journal of Computer Applications* 42.7 (2022): 1993.

[16]. 霍占强, et al."边界感知的实时语义分割网络." *计算机工程与应用 58*.17(2022):165-173.

[17]. 孙晨,莫国美,舒坚.基于强化学习的无人机自组网路由研究综述[J/OL].计算机应用研究:1-11[2023-02-03].DOI:10.19734/j.issn.1001-3695.2022.11.0566.

[18]. 毛鹏强,谢钧,夏士明,骆西建.基于深度强化学习的无线网络边缘缓存技术综述[J].陆军工程大学学报,2022,1(06):56-64.

[19]. 张昊迪,陈振浩,陈俊扬,周熠,连德富,伍楷舜,林方真.显式知识推理和深度强化学习结合的动态决策[J/OL].软件学报:1-15[2023-02-03].DOI:10.13328/j.cnki.jos.006593.

[20]. 孙益辉,易灵芝,夏云芝,田原.基于强化学习的物联网可信信息覆盖优化算法[J].华中科技大学学报(自然科学版),2023,51(02):32-38.DOI:10.13245/j.hust.230201.

[21]. 张尊栋,刘雨珂,刘小明.深度强化学习在交通信号控制中的应用[J].自动化博览,2022,39(12):30-37.

[22]. 张倩,李天皓,白春光.基于多智能体强化学习的分层决策优化方法[J].电子科技大学学报(社科版),2022,24(06):90-96.DOI:10.14071/j.1008-8105(2022)-1056.