从示范中学习自动驾驶汽车的驾驶风格

摘要

自动驾驶技术在近年来发展迅速，自动驾驶的普及已经成为未来必然的趋势之一。然而，为了提高用户对自动驾驶的满意度，这些车辆不仅应该是安全可靠的，还应提供舒适的用户体验。然而，对于舒适度的个体感知在用户之间可能存在显著差异。有些用户可能喜欢运动化的驾驶，具有高加速度，而其他人可能更喜欢更为轻松的风格。通常，许多参数，如加速度曲线、与其他车辆的距离、车道变道时的速度等，都能表征出人类驾驶员的驾驶风格。手动调整这些参数可能是一项繁琐且容易出错的任务。因此，文章提出了一种基于示范学习的方法，允许用户通过手动驾驶车辆来简单地展示所期望的驾驶风格。文章通过成本函数建模个体驾驶风格，并使用基于特征的逆强化学习来找到最适应观察到的风格的模型参数。一旦学到了模型，就可以用于在自动驾驶模式下高效地计算车辆的轨迹。文章展示了该方法能够使用来自真实驾驶员的数据学习成本函数并复现不同的驾驶风格。同时，我们创新性地提出了单步交替更新和多任务强化学习初始化技术来提高算法的学习效率和成功率。

关键词：驾驶风格；逆强化学习；自动驾驶；成本函数

1 引言

最近的研究表明，自动驾驶车辆的创新和投资步伐正在加快，消费者对这种车辆的想法持开放态度。在自动驾驶汽车的用户接受过程中，一些关键因素包括安全性、可靠性和舒适性。舒适性是主观的，受多种因素的影响，包括驾驶风格，即驾驶员通常的驾驶方式 [1]，这是速度、加速度、加速度变化率、与其他车辆的距离等特征之间的权衡。研究表明，驾驶风格在用户之间存在差异 [17]。为了适应不同用户的舒适性，自动驾驶汽车应根据用户的偏好调整其驾驶风格，除了保持安全性。通过改变自动驾驶车辆运动规划算法的模型参数，可以实现不同的驾驶风格。然而，由于参数数量较多可能产生对立效应，手动调整这些参数通常难以执行。如果用户偏好的所有变异都落入一小组类别，可能可以手动调整参数一次并根据其偏好类别为用户选择参数。如果不是这样，手动调整，即使可能，也可能是一个繁琐且耗时的过程。

文章提出了一种从示范学习中学习每个用户的模型参数的方法，通过观察其驾驶风格。文章假设所期望的驾驶风格最大化某种奖励的概念，即驾驶员的风格可以通过成本函数解释。挑战在于找到最佳解释观察到的风格并且在不同情况下也具有泛化能力的成本函数。文章提出了一种基于特征的逆强化学习（IRL）方法，从示范中学习驾驶风格。特征是从轨迹到实值的映射，捕捉我们想要复制的驾驶风格的重要属性。文章的模型使用成本函数，该函数是这

些特征的线性组合。学习方法的目标是找到最适应观察到的风格的特征权重。一旦模型学到了，我们可以在自动驾驶任务期间在线计算轨迹。特别是对于高速公路驾驶，捕捉速度、加速度和加速度变化率等高阶属性至关重要。研究表明，加速度和加速度变化率对乘客的舒适度有很大影响 [8]。为了捕捉这些属性，文章建议使用轨迹的连续表示。

最后，为了提高学习的效率，本文创新性的提出了单步交替更新和多任务逆强化学习初始化技术。实验证明，这两个技术可以大幅加速算法的收敛速度和成功率。

2 相关工作

在本节中，我们开始讨论成本函数学习，随后深入研究损失函数学习。值得注意的是，损失函数学习和代价函数学习有显著的相似之处。在我们的成本函数学习框架中，成本函数在一定程度上可以被视为损失函数。

2.1 成本函数学习

[19] 提出了最大熵逆强化学习 (MaxEnt-IRL) 方法，该方法有效地解决了先前方法中存在的模糊性，并应用于路径偏好建模问题。本文 [9] 针对自然公路场景手工设计成本函数，并使用 IRL 成功从演示中学习了驾驶风格对应的线性权重。该系统以 5hz 的频率连续计算成本最低的轨迹，并在真实的仿真环境中使用该轨迹来控制汽车。[18] 提出了基于采样的最大熵逆强化学习 (SMIRL) 方法，该方法通过引入有效的轨迹采样器直接学习连续域的代价函数。[7] 使用多项式轨迹采样器生成考虑高阶意图的候选轨迹。通过 MaxEnt-IRL 学习成本函数，对个性化驾驶行为进行建模。[16] 引入了一个结合了行为和局部运动规划的通用规划器，并使用 MaxEnt IRL 来学习超过人类专家调整水平的奖励函数。[5] 使用行为生成模块来生成各种候选行为。随后，它通过利用每个候选人的行为来预测其他代理人的未来轨迹。最后，使用 MaxEnt IRL 方法学习成本函数来评估候选计划的有效性。[6] 同时学习预测模型，规划模块和可解释的线性成本函数从人类示范数据。由于跨多个模块的优化目标的一致性，它比单一训练方法具有更好的性能。这些研究采用了 MaxEnt IRL 方法来学习成本函数的线性权重，从而完成了从演示中学习的任务。尽管该方法在自动驾驶领域得到了广泛的应用，但其代价函数无法进行快速学习和更新限制了成本函数应用的进一步发展。本文提出了单步交替更新和多任务逆强化学习初始化来提高收敛的效率。

2.2 损失函数学习

[3, 11] 采用进化算法搜索损失函数的结构。[4] 提出了一个自动辅助损失搜索 (A2LS) 框架，该框架自动搜索强化学习 (RL) 中表现最好的辅助损失函数。[15] 提出了一个新的元学习框架。首先，采用基于进化的方法在原始数学运算空间中搜索一组符号损失函数。随后，通过端到端基于梯度的训练过程，对发现的损失函数集进行参数化和优化。[12] 率先开发了由强化学习驱动的控制器模型，以生成损失函数。此外，还实现了迭代和改进的调谐优化计划，以更新控制器模型和推荐模型的参数。这些工作虽然在学习损失函数方面取得了进展，但与手工制作的功能结构相比，搜索的损失函数结构缺乏可解释性——这是增强自动驾驶信心的关键因素。此外，这些方法通常需要大量的时间来搜索或生成可行的结构，在效率方面比本文的方法要低很多。

3 本文方法

3.1 本文方法概述

在本节中，我们概述了在高速公路驾驶的背景下，从演示中学习驾驶风格的方法。首先介绍数据集的划分和处理，接着介绍如何参数化轨迹，最后如何基于特征的方法进行优化。

3.2 轨迹类型与风格的划分

由于不同轨迹类型和轨迹风格所对应的成本函数权值不同，我们初步将规划轨迹划分为 10 种行为类型: 静止、直线、直线向左、直线向右、轻微左转弯、急剧左转弯、轻微右转弯、急剧右转弯、左 u 型转弯、右 u 型转弯。轻微右转，急右转，左掉头，右掉头。在构建数据集时，我们排除了对应于平稳、左 u 型和右 u 型类型的数据。随后，我们根据其余七种类型的轨迹特征对其进行风格聚类。在我们的实验中，我们将每种行为类型细分为五种不同的类型。对于轨迹特征设计，我们参考 [9] 设计了五个基本轨迹特征: 加速度、法向加速度、加力、法向加力和曲率。此外，我们利用 [6] 中提出的方法来制定与特定场景相关的三个轨迹特征，包括速度效率、道路偏移和安全性。

3.3 样条曲线

我们对轨迹进行等时间采样，将其维数从无限降至有限。同时，为了增强表示能力，我们将轨迹分割成多个连接的片段:

$$r_i : [t_i, t_{i+1}] \to \mathcal{R}^2 \tag{1}$$

其中 $0 \leq i \leq S$ 表示具有 S 段的样条。

对于每个部分，我们采用五次样条曲线来表征车辆轨迹在 x 和 y 维度上的时间演变。五次样条的具体公式定义如下:

$$r(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5 \tag{2}$$

由于关系 $s = r, v = \dot{r}, a = \ddot{r}$，并且已知 x 和 y 方向上的初始位置和速度，因此第一段的优化只涉及四个参数 $a_i^0 (i = 2, 3, 4, 5)$。对于从第二段到最后一段的串联轨迹，前一段的最终状态成为后一段的初始状态，每一段都需要优化 $a_i^{-0} (i = 3, 4, 5)$。为了简单起见，在我们的实验中，我们将片段的数量设置为 1，在本例中意味着 $\tau = r_0$。

3.4 优化过程

在前向过程中，我们使用可微非线性优化来求解五次多项式的参数。具体而言，其优化目标如下:

$$\mathbf{a}^* = \arg\min_{\mathbf{a}} \ \frac{1}{2} \sum_i \| \theta_f^i f_{\theta_p^i}^i (\tau_{\mathbf{a}}) \|^2 \tag{3}$$

其中 $f_{\theta_p^i}^i (\tau_{\mathbf{a}})$ 表示轨迹的特征函数 $\tau_{\mathbf{a}}$, $\theta_f^i (i = 1, ..., N_f)$ 表示第 i 个特征函数的标度权值。这些轨迹特征是为自动驾驶规划任务量身定制的，包括旅行效率、乘坐舒适性、车道偏离和安全性等因素。这些特性函数的详细信息可以在 [6] 找到。对于具体实现，我们利用 Theseus [14] 来促进成本函数的自动微分，而 PyTorch [13] 用于梯度反向传播。

在反向过程中，MaxEnt IRL [19] 旨在从人类行为的演示中恢复潜在的奖励功能。通过恢复所有轨迹上的分布，可以解决多解的模糊性和专家行为的随机性。本质上，根据最大熵原则，确定候选行为 (轨迹) 的概率分布如下:

$$P(\tau|\theta) = \frac{\exp(-c(\tau|\theta))}{\sum_{\tau}\exp(-c(\tau|\theta))} \tag{4}$$

其中 $c(\tau|\theta)$ 表示成本函数，等于 $\sum_i \theta_f^i f_{\theta_p^i}^i(\tau)$。

该算法的对数似然的梯度可计算如下:

$$\nabla_\theta \mathcal{L}_\tau^{IRL}(\theta) = \frac{\partial c_\theta}{\partial \theta}[\mathbb{E}[f_\tau] - f_{D_\tau}] \tag{5}$$

4 复现细节

在本节中，我们首先说明开源代码引用情况和实验数据集的构建过程，并说明提供所使用的实验设备的详细说明。随后，我们详细阐述了我们的创新点。最后，我们根据实验结果讨论了算法的有效性。

4.1 开源代码引用情况

该论文没有可参考的相关源代码。在复现过程中引用了一小部分 [6] 中的代码，主要引用了两部分的代码，第一部分是对数据集进行预处理的代码。另一部分是引用了 Theseus [14] 的可差分的成本函数学习框架代码。引用的代码占比关键代码不超过 20%。

4.2 数据集处理

我们利用 *waymo 开放运动数据集 (WOMD)* [2] 构建任务数据集，这是一个专注于城市驾驶场景的大规模真实驾驶数据集。我们取一小部分数据，将其处理成以 10Hz 频率采样的 5 秒长的数据帧。我们获得了 22189 个数据帧样本作为训练集的前驱，同时获得了 18004 个数据帧样本作为测试集的前驱。随后，我们对前驱训练集和测试集内的行为类型进行分类，然后对每种行为类型进行聚类以形成任务数据集。最终，我们得到了 35 个训练任务集和 35 个测试任务集，每个任务集在任务数据集中的最大长度为 48。

4.3 实验设备

我们的设备是一台配备 8 核英特尔 (R) I5-9300H (2.40 GHz) CPU 的笔记本电脑。显卡是 GeForce GTX 1650，具有 4GB 的显存。所有的训练任务，以及运行时间和计算效率评估，都是在这个特定的设备上进行的。

4.4 创新点

本文提出了两个改进的创新点。首先在训练过程中，相比原来的回合更新，我们使用单步交替更新方法来学习成本函数的参数。

$$\theta \leftarrow \theta - \alpha_{irl}\nabla_\theta \mathcal{L}_\tau^{IRL}(\theta) \tag{6}$$

$$\mathbf{a} \leftarrow \mathbf{a} - \alpha_{opt} \nabla_{\mathbf{a}} \mathcal{L}_{\theta}^{OPT}(\mathbf{a}) \tag{7}$$

在实验中，我们观察到，与之前的方法相比，单步交替更新方法表现出更直接的收敛性，并且显著提高了收敛速度。

此外，我们采用多任务逆强化学习算法对成本函数进行参数初始化。对于各种类型的演示，知识共享是可行的。为了促进这种知识共享，我们引入了用于参数初始化的多任务逆强化学习算法如下:

$$\nabla_{\theta} \mathcal{L}_{\tau}^{init}(\theta) = \frac{\partial c_{\theta}}{\partial \theta}[\mathbb{E}[f_{\tau}] - f_{D_{\tau}}] + \frac{\partial c_{\theta}}{\partial \theta}[\mathbb{E}[\tau] - D_{\tau}] \tag{8}$$

其中 D_{τ} 表示示范轨迹，τ 表示政策轨迹，$f_{D_{\tau}}$ 表示示范轨迹特征，f_{τ} 表示政策轨迹特征。有关轨迹特征设计的全面说明，请参阅轨迹类型与风格的划分小节。值得注意的是，第一个组件来自 MaxEnt IRL 算法 [19]，第二个组件来自行为克隆算法 [10]。实验结果表明，与单独使用相比，这两种成分的联合利用产生了更好的性能。

我们使用在8中描述的梯度进行更新，并采用基于监督学习的训练方法来统一训练相同行为类型的任务。我们使用获得的结果作为初始化参数，并通过实验验证了该方法的有效性。

5 实验结果分析

algorithm	initialization	training time/step	testing time/step	raw irl loss	learned irl loss	success rate	learning speed
LfD	None	301.9/497.7	3.95/7.4	0.24771	0.07013/1.0x	71.43%	1.0x
LfD	MT-IRL	289.1/479.4	4.94/9.4	0.06568	0.00549/0.078x	97.14%	1.04x
Single-Step Update	None	103.8/169.8	3.08/5.5	0.25486	0.02866/0.409x	80.00%	2.91x
Single-Step Update	MT-IRL	92.1/207.3	3.69/9.3	0.06568	0.00554/0.079x	97.14%	3.28x

表 1. 统计结果。我们将每一个算法在 35 个测试任务集上进行训练，并统计平均表现展示在上表中。

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。我们首先在 35 个测试任务上进行训练，得到统计结果展示在表格1中。我们给出了每种算法在 35 个学习任务中的平均性能。“training time” 列表示训练所需的时间 (单位为秒)，“raw irl loss” 表示初始化代价函数的性能，“learned irl loss” 表示学习后代价函数的性能度量。从表格数据来看，采用单步交替更新的学习速度比原来的回合更新要更快，同时采用多任务逆强化学习初始化设置的算法对算法也有加速效果，同时在学习成功率上大幅提升。同时采用单步交替更新和多任务逆强化学习初始化设置的算法表现最好，在减少学习的 irl 误差的同时，还比较大程度上加速了整个学习过程，同时保持较高的学习成功率。

如图1和图2所示，我们直观地展示了成本函数学习的结果。图中，“raw traj” 表示初始输入轨迹，“before traj” 表示经过未学习成本函数后生成的轨迹，“expert traj” 表示 ground truth 专家轨迹。从图中可以明显看出，位置、速度和加速度曲线都接近演示风格。此外，特征条形图表明了学习轨迹与专家在关键特征上的总体一致。然而，由于使用最大熵特征匹配的学习方法，系统只能捕获近似的特征样式，而不能精确地复制特定的轨迹路径。

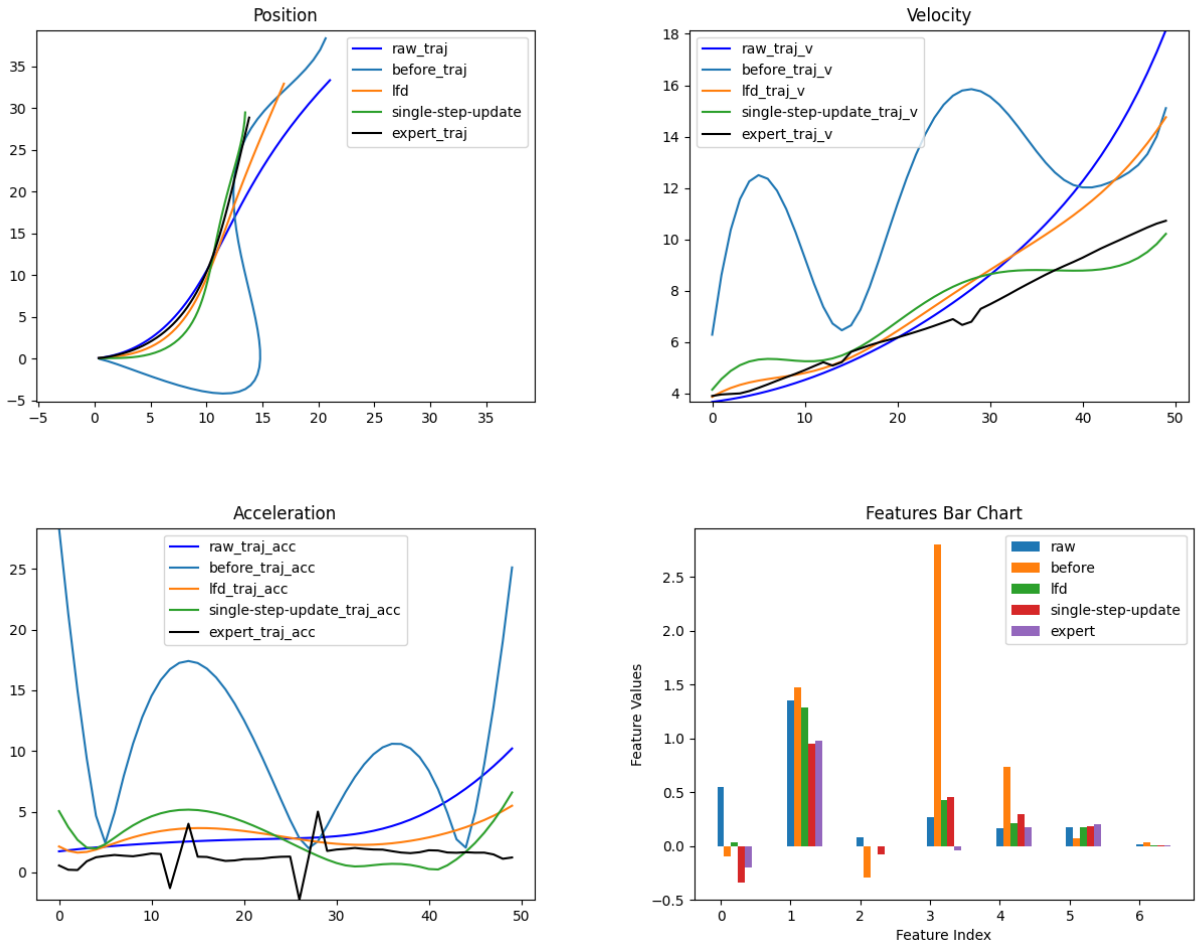

图 1. 可视化结果。我们对案例进行分析，并展示了可视化的结果。

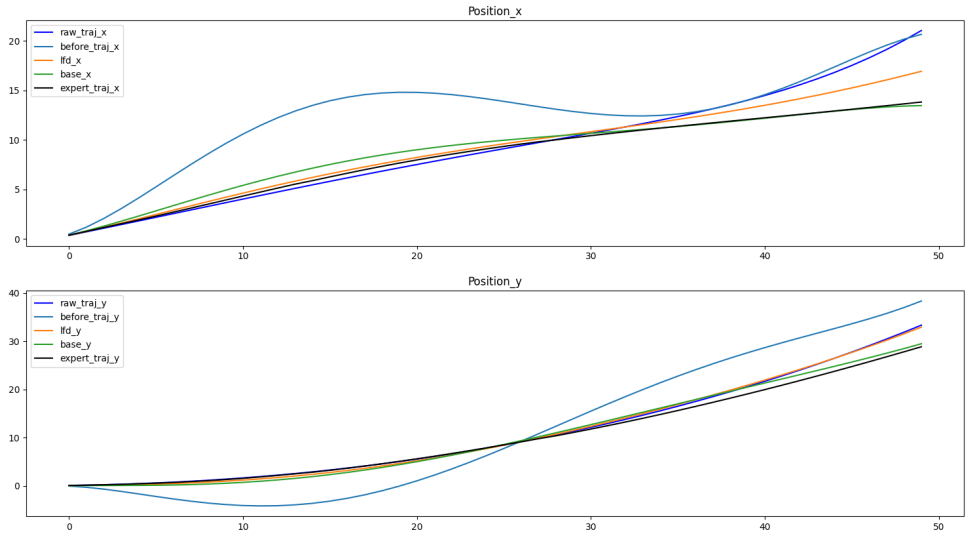

图 2. 可视化结果 2。xy 方向分解的位置偏移图。

6 总结与展望

我复现了一个基于示范学习和逆强化学习的方法，用于从驾驶员的演示中学习相应的驾驶风格。本文提到了两个创新点：采用单步交替更新的方法来加速成本函数的学习过程，以及使用多任务逆强化学习初始化来提高学习的成功率。通过在 35 个测试任务上的实验，我展示了这两个创新点的有效性，同时提供了详细的实验结果和分析。

然而，该模型存在一些局限性，比如该模型只是考虑基于前端规划的情况下进行运动规划风格的学习，没有考虑到行为决策层的演示偏好建模。此外，没有考虑更综合的因素，比如更复杂的特征表示、驾驶者的情感因素、用户的即时任务需求、人机交互体验等等。

参考文献

[1] James Elander, Robert West, and Davina French. Behavioral correlates of individual differences in road-traffic crash risk: An examination of methods and findings. *Psychological Bulletin*, 113(2):279, 1993.

[2] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021.

[3] Hongyang Gu, Jianmin Li, Guangyuan Fu, Chifong Wong, Xinghao Chen, and Jun Zhu. Autoloss-gms: Searching generalized margin-based softmax loss function for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4744–4753, 2022.

[4] Tairan He, Yuge Zhang, Kan Ren, Minghuan Liu, Che Wang, Weinan Zhang, Yuqing Yang, and Dongsheng Li. Reinforcement learning with automated auxiliary loss search. *Advances in Neural Information Processing Systems*, 35:1820–1834, 2022.

[5] Zhiyu Huang, Haochen Liu, Jingda Wu, and Chen Lv. Conditional predictive behavior planning with inverse reinforcement learning for human-like autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[6] Zhiyu Huang, Haochen Liu, Jingda Wu, and Chen Lv. Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[7] Zhiyu Huang, Jingda Wu, and Chen Lv. Driving behavior modeling using naturalistic human driving data with inverse reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):10239–10251, 2021.

[8] Ira D Jacobson, Larry G Richards, and A Robert Kuhlthau. Models of human comfort in vehicle environments. *Human Factors in Transport Research Edited by DJ Oborne, JA Levis*, 2, 1980.

[9] Markus Kuderer, Shilpa Gulati, and Wolfram Burgard. Learning driving styles for autonomous vehicles from demonstration. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2641–2646. IEEE, 2015.

[10] Guofa Li, Zefeng Ji, Shen Li, Xiao Luo, and Xingda Qu. Driver behavioral cloning for route following in autonomous vehicles using task knowledge distillation. *IEEE Transactions on Intelligent Vehicles*, 8(2):1025–1033, 2022.

[11] Hao Li, Tianwen Fu, Jifeng Dai, Hongsheng Li, Gao Huang, and Xizhou Zhu. Autolosszero: Searching loss functions from scratch for generic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1009–1018, 2022.

[12] Zelong Li, Jianchao Ji, Yingqiang Ge, and Yongfeng Zhang. Autolossgen: Automatic loss function generation for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1304–1315, 2022.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[14] Luis Pineda, Taosha Fan, Maurizio Monge, Shobha Venkataraman, Paloma Sodhi, Ricky TQ Chen, Joseph Ortiz, Daniel DeTone, Austin Wang, Stuart Anderson, et al. Theseus: A library for differentiable nonlinear optimization. *Advances in Neural Information Processing Systems*, 35:3801–3818, 2022.

[15] Christian Raymond, Qi Chen, and Bing Xue. Learning symbolic model-agnostic loss functions via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[16] Sascha Rosbach, Vinit James, Simon Großjohann, Silviu Homoceanu, and Stefan Roth. Driving with style: Inverse reinforcement learning in general-purpose planning for automated driving. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2658–2665. IEEE, 2019.

[17] Orit Taubman-Ben-Ari, Mario Mikulincer, and Omri Gillath. The multidimensional driving style inventory—scale construct and validation. *Accident Analysis & Prevention*, 36(3):323–332, 2004.

[18] Zheng Wu, Liting Sun, Wei Zhan, Chenyu Yang, and Masayoshi Tomizuka. Efficient sampling-based maximum entropy inverse reinforcement learning with application to autonomous driving. *IEEE Robotics and Automation Letters*, 5(4):5355–5362, 2020.

[19] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.