**说 明 书**

**基于Deepracer的联合强化学习sim2real系统**

**技术领域**

1. 本发明涉及的是一种无人车技术领域的方法，具体完成了一个以深度强化学习在自动驾驶场景的应用为基础的多小车真实环境车道保持系统。一方面，将模拟环境预训练得的模型通过本地化训练的方式来进行微调，以更好的适应物理环境，从而提升部署效果；另一方面，将小车收集到的数据存于本地进行训练，只与服务器定时交互模型参数，保证了用户数据的安全性和隐私性。同时，联合学习使多小车并行共同探索相似或不同的环境成为可能，直击目前无人驾驶的痛点，一定程度上提升了sim2real的效率。

**背景技术**

1. 目前，5G乃至6G网络与端设备更强大算力的双重影响下，越来越多的研究聚焦于端设备上的分布式训练，以提升模型效果与充分利用端设备算力。同时，自动驾驶作为一个极具前景的研究课题，近两年的论文数量也骤增，其中，深度强化学习是一个倍受关注的热点领域。然而，无人车作为端智能设备的一种，如果施加传统的分布式训练，会面临着诸多问题。
2. 其一，由于强化学习本身的训练特性，如果使用同策略（on-policy）算法进行训练，模型的参数如何在各个端设备间共享便成为问题；其二，无人车上较为多而复杂的传感器产生的数据，如果使用异策略（off-policy）直接将数据进行端到端传输，将无法应对可能存在的低带宽高延迟问题；其三，在机器人领域长久以来存在着sim2real困境，即模拟环境训得的模型难以在现实环境中获得同样出色的效果，而通过少量的本地化训练以微调可以一定程度上解决这一问题；其四，车联网时代，传感器数据作为可能会泄露用户行程的私人数据变得极其敏感，如何在保证数据安全的基础上实现模型的共享与更新迭代也成为热点问题。
3. 联合学习，或称为联邦学习（Federated Learning），是最早在2015年被提出，后由Google2017年在AISTATS上确立，并提出了首个联邦学习模式以及联合均值算法FedAvg，但没有提供理论证明。后2018年由Stich提供了对于数据独立同分布下FedAvg算法收敛的理论证明，又由Li等人与Khaled等人先后证明了非数据独立同分布下FedAvg算法依然收敛。主要就是将初始化模型下载到各终端，然后根据其本身的数据更新模型参数，不同终端随之产生不同的更新结果，送到云端进行聚合之后，汇总后的模型参数将作为下一次更新的初始参数，一直迭代到收敛。它的诞生是基于目前越发受到重视的数据隐私保护（即“数据孤岛”问题）和AI安全性问题，通过避免了在训练时数据的直接传输，而改为传输模型参数本身，有效避免了中心训练服务器直接获取原始数据；通过中心服务器使用以联合均值为代表的联合学习算法，将多个站点上传的模型参数进行有效加权平均，从而实现了多站点共享中心模型的目的，一方面提升了训练的效率，另一方面也使多参与方在不共享数据的情形下共同训练获得高效模型，符合我们对未来人工智能的展望。然而，目前还有数据Non-iid问题、数据隐私问题、模型鲁棒性问题、传输过程开销问题、训练过程优化问题甚至是边缘设备硬件优化问题，还有很多问题亟待解决；同时，目前联邦学习与强化学习结合得还比较少，未能有很大的应用空间，还值得多多挖掘。

[0005] 联合强化学习是联合学习和强化学习在算法和系统层面的结合。目前分布式强化学习的相关研究已有很多。然而，大多分布式强化学习框架是不同服务器节点之间的协同学习，并不适配于目前的端设备。而随着联合学习的提出，越来越多的学者将视线转向了联合强化学习，试图研究这两者结合的优势。但尽管联合学习和强化学习在各自的领域都有大量的文献和相关研究进展，但将两者结合的模型和应用还较为少见，属于比较SOTA的领域。例如，2019年，Boyi Liu提出了一个应用于云机器人系统的终生联邦强化学习框架LFRL（Lifelong Federated Reinforcement Learning），通过不断收集不同环境下所训练得模型的参数，持续在云端共同训练一个机器人导航模型。然而其采用的模式仍然是模拟环境下训练物理环境中验证，并没有形成一个端设备通用的训练框架，且其代码尚未开源供研究。联合学习领域的论文目前以数据安全为主，深度强化学习的论文主要以不同问题场景设计的学习算法改进为主。目前，Xinle Liang等人来自微众银行的团队，先是分析将联合迁移强化学习用于自动驾驶问题的理论可能性，后又在jetson RT设备上加以验证，证明了在简单环境的避障问题中联合学习的性能。然而，该文章中，仅仅关注这一模型中引入预训练过程所提升的性能，没有对可能出现的sim2real细节问题与选择算法的影响进行探讨。其余只有将联合强化学习用于游戏中NPC个性化的研究与控制IoT设备上。因此，在强化学习与机器人领域方面，尤其是sim2real领域的应用，这一模型的研究目前还处于空白。

**发明内容**

1. 本发明针对现有技术存在的上述不足，提出一种提出一种基于Deepracer的联合强化学习sim2real系统，以联合学习的视野，解决了目前车联网场景下分布式训练可能出现的数据安全与隐私问题；同时，实现了相似环境下的多小车共享模型参数；最后，传递参数的方式相较于之前传递原始数据，提供了使用异策略算法时解决了端智能设备分布式训练可能遇到的带宽瓶颈问题的思路。
2. 本发明是通过以下技术方案实现的，本方法包括以下步骤：
3. 步骤1、基于PPO算法的仿真环境车道保持模型设计及实现。
4. 所述的PPO算法是指：一种策略梯度算法。如果通过传统的多步优化来优化其损失函数，虽然理论上是可行的，但是这将导致巨大的策略更新次数，从而大幅地降低计算的效率。因此PPO引入了自适应的KL惩罚系数（Adaptive KL Penalty Coefficent）与裁剪（Clipped）的小技巧，并利用重要性采样的思想进行伪离线转变。
5. 所述的仿真环境车道保持模型为：本系统使用Gazebo、ROS kinect、Gym三者结合，形成一套机器人学模拟环境，以应对本次系统的模拟实验的实现。我们考虑使用AWS提供的原始跑道环境文件，这是一个长16.64米，宽107cm的具有多种不同转弯弧度的跑道。
6. 步骤2、基于Deepracer的分布式sim2real模型设计及实现。
7. 所述的sim2real是指：在机器人领域，出于真实环境的采样频率过低和一系列安全问题，我们往往先在模拟环境中训练强化学习模型。但当实际生产落地时，我们不得不考虑从模拟到真实环境迁移和部署的问题。在这个过程中，由于模拟环境和真实物理环境的差距，导致原本训练的模型会产生一定的误差。
8. 实现方法：域随机化DR（domain randomization）；
9. 具体步骤：我们参考Sandeep论文中的DR设置，针对图像信息进行颜色随机化、阴影随机化、锐化和噪声四个层面的随机化设计。 颜色随机化：根据HSV颜色模型，对于色调（Hue）、饱和度（Saturation）、明度（Value）三个维度作颜色随机化。色调随机化：针对图片的色调进行随机化以模拟迁移环境的色调变化。经过调整之后，我们设置一个用于随机化的位于0到9之间的微调值，以针对图像色调值作轻微变化。饱和度随机化：针对图片的饱和度进行随机化以模拟迁移环境的饱和度变化。我们设置一个用于随机化的位于0到29之间的微调值，以针对图像饱和度值作轻微变化。 (3)明度随机化：针对图片的明度进行随机以模拟迁移环境的明度变化化。我们设置一个用于随机化的位于0.5到2.5之间的明度比例值，以针对图像明度值作轻微变化。阴影随机化：针对图片的光影进行随机以模拟迁移环境的光影变化。我们在图片中以随机的概率对随机选定的区域作阴影的增加。锐化：针对图片进行锐化程度的提升以增加图像的识别度。我们使用锐化内核针对相邻像素值之间的差值作针对性调整，使图片锐化程度增加。噪声：针对图片进行噪声处理，以模拟迁移环境的噪声环境。我们为图片增加椒盐噪声，即随机地将图片中的某些像素点用黑色或白色像素点来代替。
10. 步骤3、基于Deepracer的联合强化学习sim2real模型设计及实现，具体步骤包括：
11. 3.1）基于PPO算法在仿真环境中预训练车道保持模型；
12. 3.2）将预训练模型迁移到小车端进行本地化训练；
13. 3.3）使用联合均值算法，联合学习服务器对于小车上传的模型参数进行聚合；
14. 3.4）联合学习服务器下发给小车参数，进行新一轮的训练。

**技术效果**

1. 与现有技术相比，本发明优点包括：提出一种提出一种基于Deepracer的联合强化学习sim2real系统。以联合学习的视野，解决了目前车联网场景下分布式训练可能出现的数据安全与隐私问题；同时，实现了相似环境下的多小车共享模型参数；最后，传递参数的方式相较于之前传递原始数据，提供了使用异策略算法时解决了端智能设备分布式训练可能遇到的带宽瓶颈问题的思路。
2. **附图说明**
3. 图1为仿真环境车道保持任务的跑道示意图。
4. 图2为基于特征处理的图像处理手段实现小车位置感知的示意图。
5. 图3为本联合强化学习系统的架构示意图；
6. 图4为所使用的联合均值算法示意图。
7. 图5为实验效果示意图。

**具体实施方式**

1. 下面对本发明的实施例作详细说明，本实施例在以本发明技术方案为前提下进行实施，给出了详细的实施方式和具体的操作过程，但本发明的保护范围不限于下述的实施例。

**实施例1**

1. 本实施例包括以下步骤：
2. 第一步、在仿真环境基于PPO算法设计和实现车道保持模型。
3. 该步骤主要包含仿真环境搭建、PPO模型算法实现两个主要步骤。
4. 首先，本系统使用Gazebo、ROS kinect、Gym三者结合，形成一套机器人学模拟环境，以应对本次系统的模拟实验的实现。我们考虑使用AWS提供的原始跑道环境文件，如图1所示，这是一个长16.64米，宽107cm的具有多种不同转弯弧度的跑道。
5. 其次，PPO算法实现。
6. 第二步、基于Deepracer的分布式sim2real模型设计和实现。
7. 第三步、基于Deepracer的联合强化学习sim2real模型设计及实现。
8. **模拟实验结果**
9. 我们分别进行了基于PPO算法的仿真环境车道保持模型、基于Deepracer的分布式sim2real模型以及基于Deepracer的联合强化学习sim2real模型的设计与实现。首先，我们通过基于ROS和Gazebo平台业已搭建了服务端模拟环境预训练模型，以作为后续所使用的基础模型，我们也从模块架构和实验设计实现等角度详述了实验过程；之后我们尝试将第三章训得的模型通过物理环境再次训练的方式进行微调，并验证这种训练前后的实验结果。经证明，物理环境的本地化训练确实能让小车在真实环境的sim2real效果更好；最后，我们通过将原始模型嵌入联合学习系统，以验证联合学习的加入前后对于小车在真实环境的效果的变化。现在我们对于这三个实验的结果进行综合对比和分析，以证明本地化训练与联合学习的加入可以有效地提升整个系统的效果。
10. 我们分别对比两个跑道的原始模型与只有本地化训练，以及只有本地化训练与加入联合学习的训练时平均累计奖励值，如图5a和图5b所示。我们发现，引入本地化训练前，跑道一和跑道二的累计奖励函数几乎没有增长，而引入本地化训练后，跑道一和跑道二分别从第五轮和第八轮开始表现较优，如图5a所示。我们后又发现，跑道一引入联合学习后的奖励值函数有少许提升，但跑道二引入联合学习后奖励值函数却有少许下降，考虑是联合学习的加入使模型相对整体更拟合于跑道一的环境。于是我们最终得出结论，本地化训练和联合学习的加入,相较于原始模型，确实可以使小车sim2real效果有一定的提升，但联合学习与本地化训练的训练效果相差甚微。
11. 而从通信成本的角度来看，本次物理环境实验共获得6.27GB的传感器数据。其中每辆小车在本地训练一轮产生的传感器数据约为92.2MB。为完成两辆小车分别本地化训练10轮，将会产生1844MB左右的传感器数据。如果考虑把这部分数据传回给服务器进行训练，将给局部带宽造成很大的压力。而联合学习将传感器数据留在本地，一轮通信轮次服务器会将全局联合学习中央模型传回给两个客户端，客户端在本地进行10轮训练后，再将本地模型传给服务器端，每个模型大小仅200MB左右，四次通信共计传输800MB的模型数据，仅原数据的43.38\%。同时，仅传模型数据也有效保护了数据安全和隐私性。
12. 然而，我们需要注意的是，尽管联合学习可以使多辆小车共同并行探索相似环境效率提升、通信成本降低，但其效果与仅作本地化训练相比没有明显的优势，因此需要我们在探索速度、通信成本与训练速度之间做出取舍与权衡。
13. 最终我们获得结论，与纯模拟环境训练的模型相比，在物理环境本地化训练确实能使原模型更好地拟合物理环境，获得效果上的一定提升。同时，联合学习的引入与原本地化训练的结果相差甚小，因此需要我们在数据安全性、模型训练效果与通信成本上作出折衷。在对于数据隐私安全极为注重且通信成本较高的环境中，我们认为该联合强化学习系统可以得到较为良好的应用。**[[1]](#footnote-1)**

**说 明 书 附 图**



图1

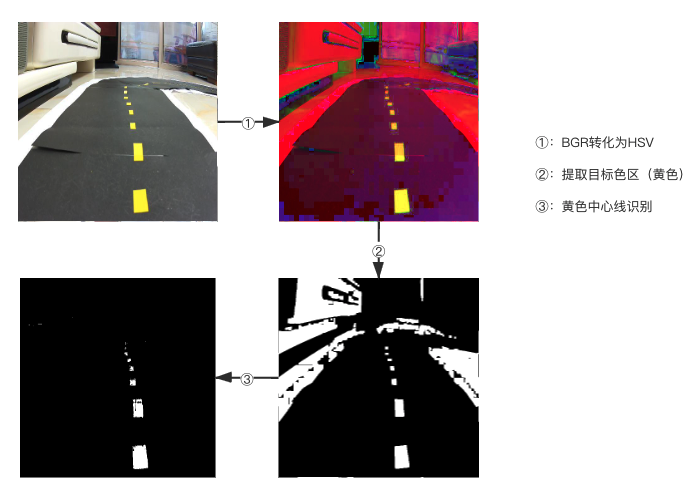


图2

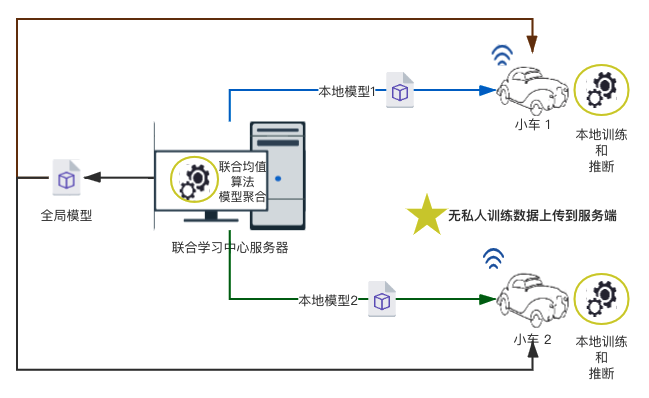


图3

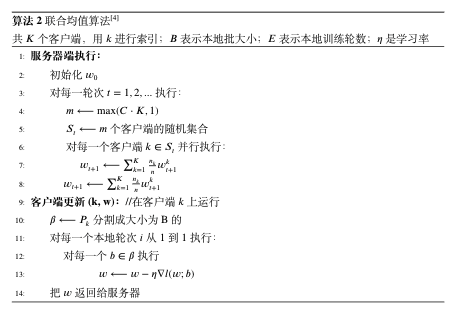
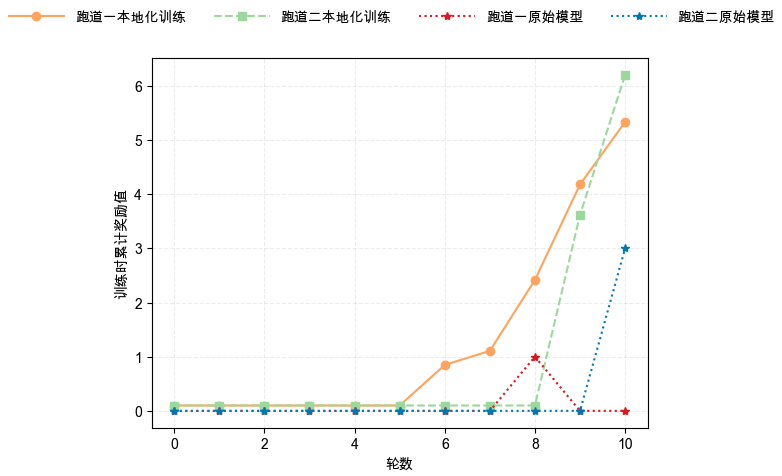
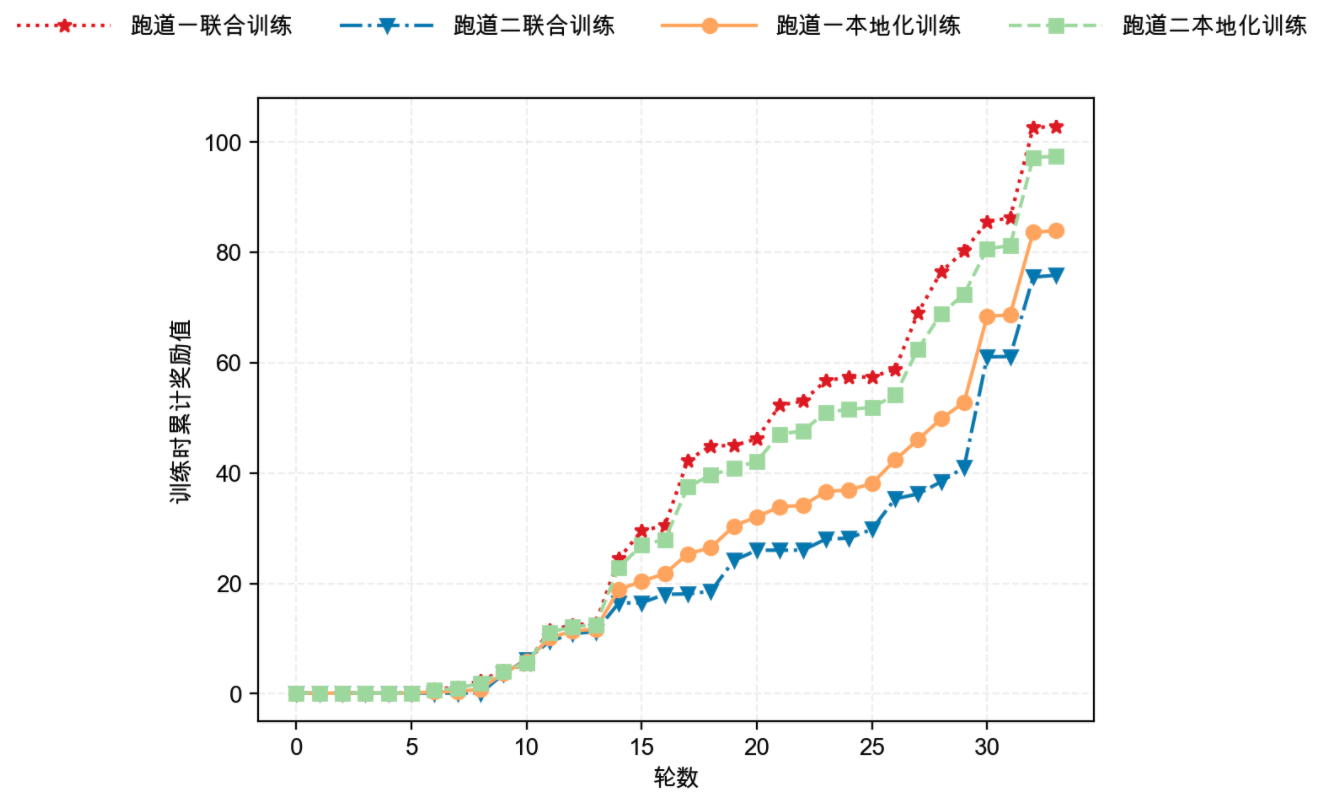


图4



a



b

图5

**权 利 要 求 书**

1、一种利用联合强化学习技术完成的自动驾驶方法，其特征在于，包括以下步骤：

步骤1、服务器端预先使用PPO模型，在模拟仿真跑道上训练车道保持模型；

步骤2、服务器端将预训练完成的PPO模型发送给小车，小车在本地物理跑道上运行该模型，并同步收集训练数据，使用PPO策略梯度下降方法离线训练，迭代C轮；

步骤3、小车将迭代了C轮的模型上传给服务器端，服务器运行联合均值算法，再将模型发回给小车，继续回到步骤2进行迭代。

所述的PPO模型是指：Proximal Policy Optimization，近端策略优化，一种离策略的主流强化学习算法。

2、根据权利要求1所述的方法，其特征是，所述的近端策略优化方法的目标函数为：. 其中，是策略参数，是基于之前t个时间步的经验期望函数，是新旧策略的概率比例值，是在第t个时间步的预估优势函数，是手动调节的超参数（通常是0.1或0.2）。其中，优势函数指的是在时间t时，执行行为a所带来的优势值。

3、根据权利要求1所述的方法，其特征是，所述的车道保持问题是指：研究如何让小车保持沿着车道行驶的问题。

4. 根据权利要求1所述的方法，其特征是，所述的联合均值算法是指：有固定数目K个客户端，每一个客户端有固定数目的本地数据集。在每一通信轮的开始，一个随机比例C的客户端将会被选择，并且服务端将当前的全局算法状态参数发送给这些客户端，每一个被选中的客户端将基于全局状态和本地数据集执行本地计算，并且把更新发送给服务端。服务端将基于客户端发送的更新参数进行模型的聚合操作。其中，w表示模型参数，表示学习率， 表示损失函数，b表示每一个本地批大小，K表示客户端数量，并且用k进行索引。

5、根据权利要求1所述的方法，其特征是，所述的步骤一具体包括：

（1）服务器端预先搭建ROS环境；

（2）服务器运行表演者-评论家架构的PPO模型：在每一轮迭代中，执行以下步骤：

步骤一：对于每一个表演者，在T时间步内执行旧策略；

步骤二：计算从时间1到时间T的优势函数, …, ;

步骤三：K轮迭代之后，优化

步骤四：

6、根据权利要求5所述的方法，其特征是，所述的表演者-评论家架构指的是：

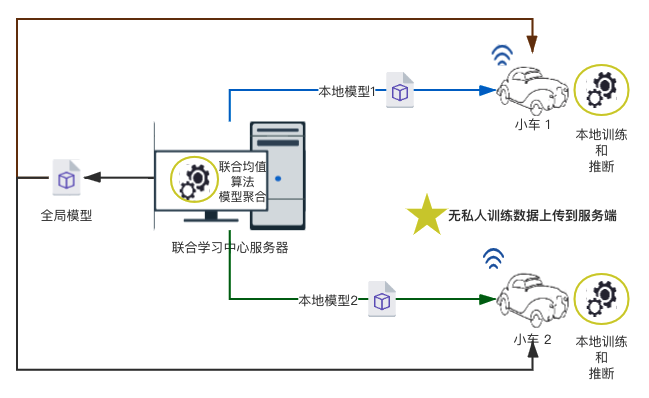
（1）表演者：表示策略函数，输入为当前状态s，输出为动作a；

（2）评论家：表示值函数，输入为当前状态s，输出该状态对应的价值v，用于评价当前策略函数的优劣；

**说 明 书 摘 要**

本文的工作以深度强化学习在自动驾驶场景的应用作为基础，在真实环境中使用两台Deepracer小车进行部署和训练，选取了车道保持这一经典问题进行验证。首先，将会在以Gazebo与ROS为基础的模拟环境上进行预训练，通过优化模型的学习算法、奖励函数和各个参数，小幅调优以最终获得适应该问题的预训练模型方案。其次，通过引入联合学习的思想，在多节点（无人车）与中心服务器共存的模式下，将传统的模拟环境训练-物理环境推断的单节点系统改造成分布式联合强化学习的多节点系统，通过在无人车本地设备上进行强化学习模型的训练和推断，使同策略算法模型实现在各个端设备之间共享模型参数。一方面，将模拟环境预训练得的模型通过本地化训练的方式来进行微调，以更好的适应物理环境，从而获得部署时更优的效果；另一方面，将小车收集到的数据存于本地进行训练，只与服务器定时交互训练得到的pb模型，保证了用户数据的安全性和隐私性。同时，联合学习使多小车并行共同探索相似或不同的环境成为可能，直击目前无人驾驶的痛点，一定程度上提升了sim2real的效率。最后，我们将横向对比原始预训练模型、通过本地化训练方式进行微调的模型与进行两轮联合学习后的模型在学习效果上的不同，来验证本联合强化学习sim2real系统引入联合学习之后的整体效果提升与折衷。本文聚焦于解决上述提到的sim2real与安全性两大问题的同时，也为其余两个问题提出解决思路。

**摘 要 附 图**



1. 更为详细的训练过程见录制的视频<https://www.bilibili.com/video/BV1Av41177fG?share_source=copy_web>. [↑](#footnote-ref-1)