农业机械学报
*Transactions of the Chinese Society for Agricultural Machinery*
ISSN 1000-1298,CN 11-1964/S

# 《农业机械学报》网络首发论文

题目：基于分步迁移策略的苹果采摘机械臂的轨迹规划方法
作者：郑嫦娥，高坡，GAN Hao，田野，赵燕东
收稿日期：2020-08-09
网络首发日期：2020-10-14
引用格式：郑嫦娥，高坡，GAN Hao，田野，赵燕东．基于分步迁移策略的苹果采摘机械臂的轨迹规划方法．农业机械学报.
https://kns.cnki.net/kcms/detail/11.1964.s.20201014.1048.008.html

网络首发时间：2020-10-14 13:21:00
网络首发地址：https://kns.cnki.net/kcms/detail/11.1964.s.20201014.1048.008.html

# 基于分步迁移策略的苹果采摘机械臂的轨迹规划方法

郑嫦娥[1] 高坡[1] GAN Hao[2] 田野[1] 赵燕东[1]
（1. 北京林业大学工学院，北京 100083；
2. 田纳西大学生物系统工程及土壤科学系，田纳西 TN37996）

**摘要：**采摘轨迹规划是苹果采摘机械臂研究的重要内容之一。针对非结构化自然环境给基于深度强化学习的采摘轨迹规划带来的训练效率低的问题，提出了基于分步迁移策略的深度确定性策略梯度算法(DDPG)进行苹果采摘轨迹规划：首先，提出了基于 DDPG 的渐进空间约束分步训练策略。其次，利用迁移学习思想，将轨迹规划的最优策略由无障碍场景迁移到单一障碍场景、由单一障碍场景迁移到混杂障碍场景。最后，对多自由度苹果采摘机械臂进行了采摘轨迹规划仿真实验，结果表明分步迁移策略可以提高 DDPG 算法的训练效率与网络性能，验证本文方法的有效性。

**关键词：**苹果采摘机械臂；轨迹规划；深度确定性策略梯度算法；迁移学习
中图分类号：TP242.6　　文献标识码：A　　OSID：

# Trajectory Planning Method for Apple Picking Manipulator Based on Stepwise Migration Strategy

ZHENG Chang'e[1]　GAO Po[1]　GAN Hao[2]　TIAN Ye[1]　ZHAO Yandong[1]
(1. School of Technology, Beijing Forestry University, Beijing 100083, China
2. Department of Biosystems Engineering and Soil Science, University of Tennessee, Tennessee TN37996, USA)

Abstract: Picking trajectory planning is one of the important research aspects of apple picking manipulator. The unstructured natural environment leads to low training efficiency for picking

收稿日期：2020-08-09　修回日期：2020-10-02
基金项目：国家自然科学基金面上项目（31971668）
作者简介：郑嫦娥（1977-），女，副教授，主要从事果实采摘机械臂和林火监测研究，E-mail: zhengchange@bjfu.edu.cn.
通信作者：赵燕东（1965-），女，教授，博士生导师，主要从事生态信息智能检测与控制研究，E-mail: yandongzh@bjfu.edu.cn

trajectory planning based on deep reinforcement learning. A deep deterministic policy gradient algorithm (DDPG) based on a stepwise migration strategy was proposed for apple picking trajectory planning: firstly, a progressive spatially constrained stepwise training strategy based on DDPG was put forward to solve the problem of hard converging in natural environments. Secondly, the Transfer Learning idea was utilized to migrate the strategies obtained from the obstacle-free scenario to the simple obstacle scenario, from the simple obstacle scenario to the hybrid obstacle scenario, to accelerate the training process in an obstacle scenario from prior strategies and guide the obstacle avoidance trajectory planning of the apple picking manipulator. Finally, the simulation experiments on the multi-degree-of-freedom apple picking manipulator for picking trajectory planning were carried out, and the results showed that the stepwise migration strategy can improve the training efficiency and network performance of the DDPG algorithm. It validated that the trajectory planning method for apple picking manipulator based on stepwise migration strategy was feasible.
Keywords: apple picking manipulator; trajectory planning; DDPG; migration learning

# 0 引言

在果园果实的采摘中，以多自由度机械臂作为采摘装置，通过果实识别与轨迹规划进行果实的自动采摘，是农业现代化的需求。不同于工业机械臂的结构化工作环境，果实采摘是在非结构化的自然环境中进行，自然生长的枝干以及未成熟果实等障碍物都给机械臂的采摘带来了困难。因此，非结构化自然环境下的采摘轨迹动态规划是果实采摘机械臂的主要研究内容之一[1-3]，本文以果园乔砧大冠稀植的苹果为研究对象进行采摘机械臂轨迹规划方法的研究。

多自由度机械臂采摘轨迹规划在多维状态空间中进行，以采摘苹果为目标，在避障的前提下，规划出一条最佳的采摘轨迹。对于轨迹规划，研究人员已经提出了多种规划算法，如A*算法[4-5]、蚁群算法[6-8]、栅格法[9]、人工势场法[10-11]等。但这些算法大多依赖于机械臂和环境的实时建模，计算复杂度随机械臂自由度的增加呈指数级增加；而由于采摘环境的多变性，很难对环境进行精确建模。深度强化学习是一种在与环境发生交互的过程中，通过奖惩函数来进行自我学习推理，最终在自我探索的过程中解决问题的办法[12]。由于深度强化学习不需要进行环境建模，因此在复杂的采摘环境中，利用深度强化学习求解多自由度采摘机械臂的轨迹具有更好的鲁棒性[13-15]。

采摘过程中，采摘机械臂的运动可以描述为高维空间中连续的状态-动作模型，而在深度强化学习中深度确定性策略梯度算法(Deep deterministic policy gradient，DDPG)可以用于连续行为的控制。但是，非结构化自然环境中采摘目标位置的复杂性和无序性，使得 DDPG

算法在训练过程中网络收敛难度大，存在较多无效搜索，样本采样效率低，有效奖励稀疏，使得训练时间过长。XIE 等[16]为了提高基于深度强化学习(Deep reinforcement learning，DRL)的机器人轨迹规划方法在有障碍物的非结构化工作环境中的网络训练效率，根据奖励塑形的思想，提出了一种新的密集奖励函数。该函数包括方位奖励函数和子任务级的奖励函数，方位奖励函数提高了局部轨迹规划的效率，子任务级的奖励函数减少了全局上的无效探索。

同时，DDPG 算法在训练时，算法初始参数是随机的且智能体的行为没有先验知识的指导，非结构化自然环境中障碍的复杂性使得随机初始化参数带来的训练速度低、收敛难的问题更为突出。针对此问题，迁移学习显示出了巨大的潜力，其可以从过去学习的相关任务中获得知识来加速训练过程[17]。胡晓东等[18]利用深度强化学习算法求解动态环境下空间机器人的路径规划问题时，设计了一种适应动态环境的快速路径规划器。首先在静态环境下对网络模型进行预训练，然后将静态模型的网络参数迁移到动态模型中，再经过动态环境下的训练进行参数微调。实验结果表明，该方法在保证规划路径准确率的前提下，显著提高了训练速度。为了解决 DDPG 算法在训练机器人任务规划中存在时间长、收敛慢的问题，陈建华等[19]利用仿真环境中训练好的基于迁移学习思想将 NAO 机器人右臂位姿规划策略在实际 NAO 机器人上进行了不同目标物体在不同位姿下的规划抓取实验。

本文将深度强化学习的方法应用于多自由度采摘机械臂的轨迹规划中。针对深度强化学习在非结构化自然环境中训练效率低的问题，提出了两种解决方法。首先，针对采摘目标位置无序性引起收敛困难的问题，提出一种渐进空间约束的分步训练策略。其次，针对果实障碍和枝干障碍的复杂性引起收敛困难的问题，提出基于迁移学习的 DDPG 算法(TL+DDPG)。最后在仿真实验中验证两种方法的有效性。

# 1 原理

## 1.1 采摘轨迹规划

### 1.1.1 DDPG 算法

DDPG 算法是基于 AC 策略梯度架构的深度强化学习算法，既有策略网络也有价值网络[20]。图 1 为 DDPG 算法的网络结构图，它借鉴了 Double DQN 的思想，拥有 4 个神经网络，分别为：actor 网络、actor target 网络、critic 网络和 critic target 网络。

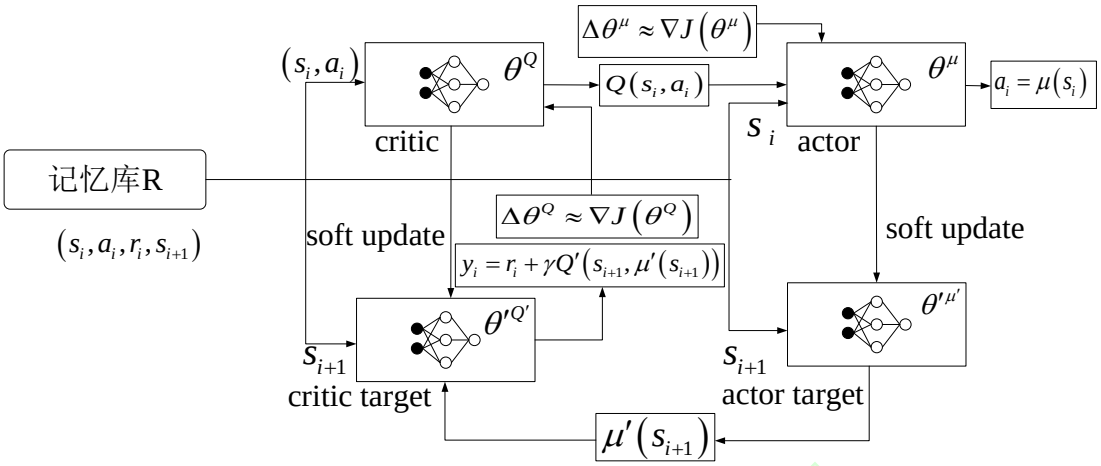



图 1 DDPG 算法网络结构图

Fig.1 Network diagram of DDPG algorithm

网络以采摘机械臂的当前状态 $s_i$ 为输入，其中 $s_i$ 包括机械臂各关节的角度、角速度等信息，以采摘机械臂的关节动作值 $a_i$ 为输出，环境根据机械臂末端当前位置 $e$ 与目标位置 $y$ 的相对距离反馈即时奖励 $r_i$。机械臂通过不断地与环境进行交互，执行相应动作，从而完成对采摘机械臂的轨迹规划。当前动作在以下情况会终止：①采摘机械臂末端到达目标点。②机械臂碰到障碍或者与环境交互的步数到达上限。采摘机械臂的轨迹规划算法如下：

（1）初始化机械臂姿态、记忆库 R、仿真环境；

（2）初始化 critic 和 actor 网络参数 $\theta^{Q}$、$\theta^{\mu}$；

（3）对每个步骤循环执行：

①获取采摘机械臂的当前状态 $s_i$；

②网络输入当前状态 $s_i$，输出机械臂的关节动作值 $a_i = \mu\left(s_i \mid \theta^{\mu}\right)$；

③机械臂执行动作 $a_i$，返回奖励 $r_i$，并获取采摘机械臂的新状态 $s_{i+1}$；

④将样本 $\left(s_i, a_i, r_i, s_{i+1}\right)$ 存入记忆库 R 中；

⑤从 R 中随机采样 64 个训练样本 $\left(s_i, a_i, r_i, s_{i+1}\right)$，分别更新 actor、critic 网络参数 $\theta^{\mu}$ 和 $\theta^{Q}$；

⑥每隔 100 步，更新 actor target、critic target 网络参数：

$$\theta'^{\mu'} \leftarrow \tau\theta^{\mu} + \left(1-\tau\right)\theta'^{\mu'}$$

$$\theta'^{Q'} \leftarrow \tau\theta^{Q} + \left(1-\tau\right)\theta'^{Q'}$$

$\tau$ 为更新系数；

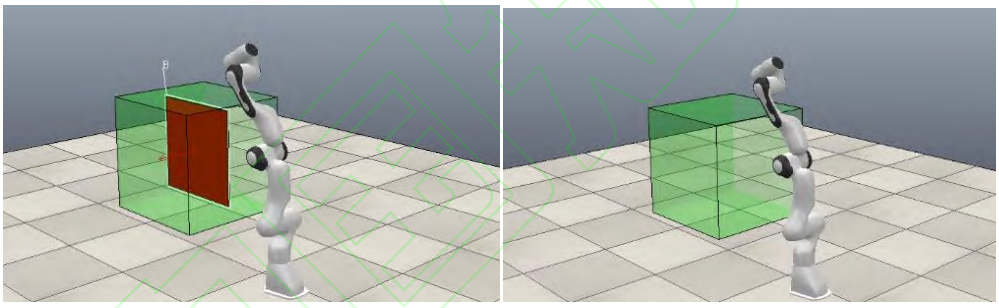
⑦如果 $s_{i+1}$ 为终止状态，则当前迭代结束，否则转到步骤②，

结束循环。

### 1.1.2 基于 DDPG 算法的渐进空间约束分步训练策略

基于 DDPG 算法的果实采摘轨迹规划中存在的主要问题是，采摘目标位置的复杂性和无序性，使得训练过程中网络收敛难度大，导致训练效率低。通过引入平面约束，降低环境复杂度，可以有效减小网络维度，加快模型学习速度。引入平面约束前后，DDPG 算法的网络模型结构保持一致，使得基于平面约束的模型训练参数可以为不引入平面约束的采摘轨迹规划模型提供有效的初始化参数，在缩短训练时间的同时增加模型的有效性。因此，为了加速训练过程，提高训练效率，本文提出一种基于 DDPG 算法的渐进空间约束分步训练策略。

分步训练的策略是：与直接求解轨迹规划不同，该方法通过引入空间约束，如图 2a 所示，简化求解过程，渐进获得最终规划轨迹，其实现过程如图 3 所示。首先通过施加平面约束，将轨迹规划限定在采摘平面上[21]，通过对网络进行训练，并得到该平面约束下的最优模型参数。图 2a 中的红色平面设定为目标果实所在的采摘平面；其次，在实际采摘环境中，进一步将平面约束下获得的网络进行训练，对网络参数进行微调，从而加速实际采摘场景的训练速度。图 2b 中绿框为采摘机械臂的实际采摘空间，目标果实可以出现在绿框中的任意位置。



(a) 引入空间约束的采摘平面　　(b) 三维采摘空间

图 2 渐进空间约束分步训练场景

Fig.2 Progressive spatially constrained stepwise training scene

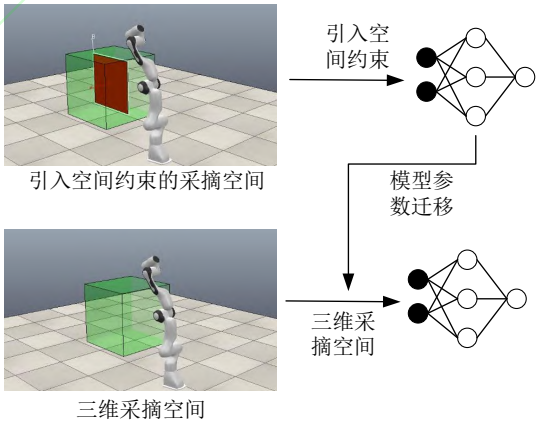



图 3 渐进空间约束分步训练策略流程图

Fig.3 Flow chart of the progressive spatially constrained stepwise training strategy

### 1.1.3 基于迁移学习的 DDPG 算法

1.1.2 节主要研究在无障碍环境下，多自由度机械臂如何通过分步训练，快速地完成三维空间中的轨迹规划任务。而在采摘环境有障碍场景中，不仅需要考虑目标的位置还要避开障碍，以保护果实和机械臂的安全[11]。因此，本节针对非结构化自然环境中的复杂障碍所带来的训练时间长的问题，利用迁移学习思想将无障碍场景下学习到的最优策略向单一障碍场景进行迁移，并将单一障碍场景学习到的策略迁移用于指导混杂障碍场景下的轨迹规划任务，流程图如图 4 所示。在本文所研究的采摘场景下，以无障碍和单一障碍场景下采摘机械臂的轨迹规划为源域，将它的轨迹规划策略迁移到目标域，目标域分别为单一和混杂障碍场景下采摘机械臂的轨迹规划。

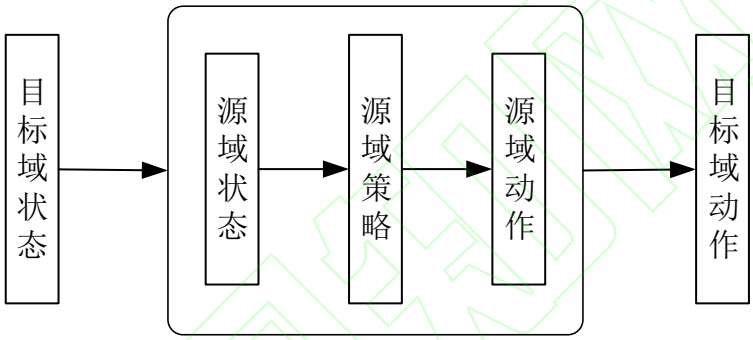



图 4 策略迁移流程图

Fig.4 Flow chart of strategy migration

进行策略迁移首先需要完成状态和动作信息从源域到目标域的映射，即完成 $s_{source}=\eta\left(s_{target}\right)$ 和 $a_{target}=\varphi\left(a_{source}\right)$ 的过程，其中 $\eta$ 和 $\varphi$ 分别为状态和动作信息的映射函数。$s_{source}$ 表示源域中的状态信息，$a_{source}$ 表示源域中的动作信息，$s_{target}$ 表示目标域中的状态信息，$a_{target}$ 表示目标域中的动作信息。具体状态和动作信息如表 1 所示。

表 1 采摘机械臂轨迹规划的状态和动作信息

Tab.1 State and action information of picking manipulator trajectory planning

| 无障碍场景(源域) | | 单一障碍和混杂障碍场景(目标域) | |
|---|---|---|---|
| 状态 $s_{source}$ | 动作 $a_{source}$ | 状态 $s_{target}$ | 动作 $a_{target}$ |
| 目标位置 | 关节 1 | 目标位置 | 关节 1 |
| 机械手末端位置 | 关节 2 | 机械手末端位置 | 关节 2 |
| 关节角度 | 关节 3 | 关节角度 | 关节 3 |
| 关节角速度 | 关节 4 | 关节角速度 | 关节 4 |

| | | |
|---|---|---|
| 关节 5 | 障碍物位置 | 关节 5 |
| 关节 6 | 末端与目标的距离 | 关节 6 |
| 关节 7 | 末端与障碍的距离 | 关节 7 |

目标域中的状态信息与源域中的状态信息的映射关系为

$$s_{source}^{(i)} = \sum_{j=1}^{J} w_{ij} \times s_{target}^{(j)} \tag{1}$$

式中 $w_{ij}$ ——状态与动作转换系数

J——目标域中状态信息的个数

$s_{source}^{(i)}$ ——源域中第 i 个状态信息

$s_{target}^{(j)}$ ——目标域中第 j 个状态信息

其中，当目标域中的状态信息与源域中的状态信息相互对应时，$w_{ij}$ 取 1；当目标域中的状态信息与源域中的状态信息不一致时，则 $w_{ij}$ 取 0。对于源域中的动作向目标域中的动作映射时同理。这样就解决了源域中的状态和动作到目标域中的状态和动作的映射。

在完成状态和动作信息之间的映射后，还需要解决状态值函数的迁移，因为状态值函数在网络参数更新中起到决定误差的作用。策略迁移后的整体状态值函数为

$$Q(s,a) = Q_{source}\left(\eta\left(s_{target}\right), a_{source}\right) + Q_{target}\left(s_{target}, \varphi\left(a_{source}\right)\right) \tag{2}$$

式中 $Q(s,a)$ ——整体状态值函数

$Q_{source}$ ——源域的状态值函数

$Q_{target}$ ——目标域的状态值函数

模型网络参数更新时，对于源域的状态值函数 $Q_{source}\left(\eta\left(s_{target}\right), a_{source}\right)$ 的神经网络参数不需要更新，需要更新的是目标域中的状态值函数 $Q_{target}\left(s_{target}, \varphi\left(a_{source}\right)\right)$ 的神经网络参数。图 5 是基于迁移学习的 DDPG 算法的参数更新的整体框图。

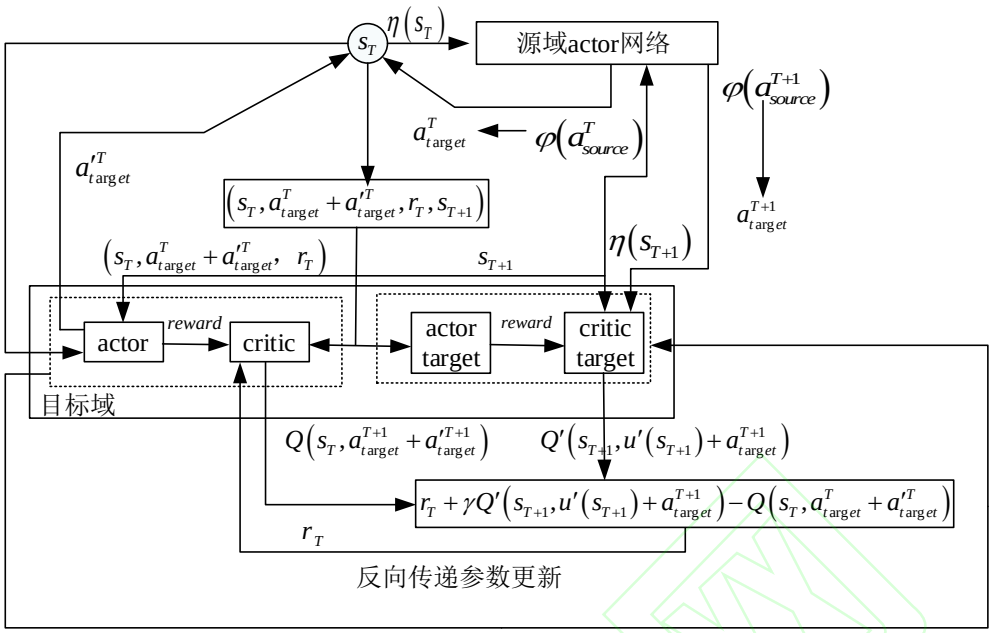



图 5 基于迁移学习的 DDPG 算法的参数更新示意图

Fig.5 Parameter update diagram of the DDPG algorithm based on transfer learning

# 2 实验与分析

利用 Cinema 4D 和 CoppeliaSim 软件搭建仿真采摘环境进行多自由度采摘机械臂的运动仿真测试，如图 6 所示。本实验在 Ubuntu16.04 操作系统平台上完成，其主要硬件配置为 Intel(R) Core(TM) i7 处理器、NVidia GTX 1060 显卡、16GB 内存。

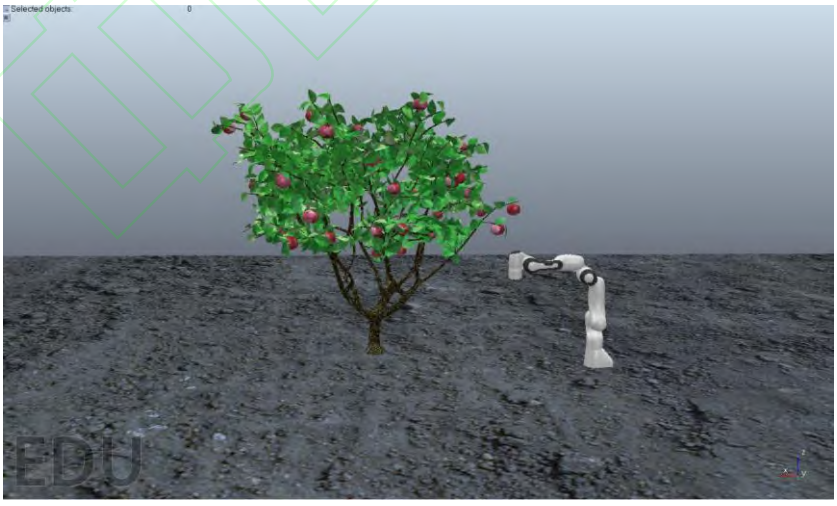


图 6 采摘场景仿真

Fig.6 Picking scene simulation

## 2.1 机械臂运动学模型与障碍模型

### 2.1.1 机械臂的运动学模型

仿真实验使用的是 Franka 7-DOF 机械臂，图 7 为机械臂的整体结构示意图。所有关节都是转动关节，关节 7 连接末端执行手爪以抓取目标。其机械臂关节参数信息如表 2 所示。

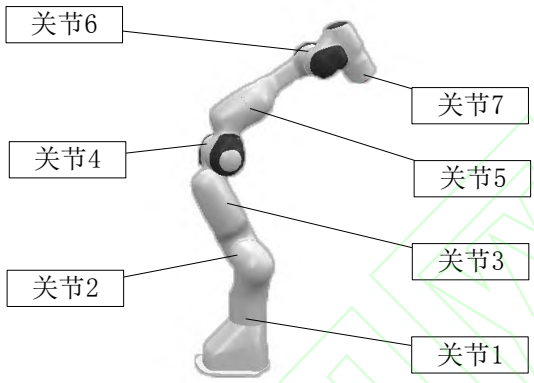



图 7 采摘机械臂的整体结构示意图

Fig.7 Schematic diagram of picking manipulator structure

表 2 机械臂关节角参数信息(°)

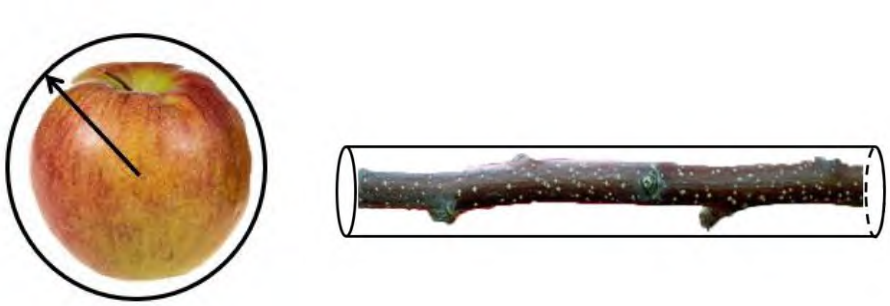
Tab.2 Parameter information of the manipulator joint angle(°)

| 参数 | 关节 1 | 关节 2 | 关节 3 | 关节 4 | 关节 5 | 关节 6 | 关节 7 |
|---|---|---|---|---|---|---|---|
| 初始值 | 0 | 0 | 90 | -90 | 90 | 0 | 0 |
| 范围 | -166~166 | -101~101 | -166~166 | -176~-4 | -166~166 | -1~215 | -166~166 |

### 2.1.2 障碍模型简化

在苹果采摘过程中，不同栽培方式下所遇到的障碍主要是枝干、树叶以及非目标果实，由于叶子柔曲，对机械臂采摘作业的影响很小，可以忽略不计，所以主要考虑枝干障碍和非目标果实障碍。

针对本文障碍物的外形特点，使用包络法对障碍物进行近似描述[22-24]。如图 8 所示，用球体表示非目标果实障碍，圆柱体表示枝干障碍。



(a) 非目标果实障碍　　(b) 枝干障碍

图 8 障碍模型

Fig.8 Obstacle models

如图 8 所示，使用包络法进行建模虽然扩大了障碍区域，但简化了计算，提高了可靠性，有效地提高了轨迹规划效率，同时也保证了机械臂和果树的安全性。

## 2.2 分步训练策略的实验与分析

DDPG 算法的网络参数如表 3 所示。

表 3 DDPG 算法的网络参数

Tab.3 Parameters for DDPG algorithm

| 网络 | 学习率 | 优化器 | 网络节点数 |
|---|---|---|---|
| actor | 0.0005 | Adam | 8×128×64×7 |
| critic | 0.001 | Adam | 8×128×64×1 |
| actor target | 0.001 | | 8×128×64×7 |
| critic target | 0.001 | | 8×128×64×1 |

仿真环境中设定的采摘空间是以(0.25m, 0m, 1.002m)为中心，尺寸为 0.5m×0.8m×0.5m 的立体空间，如图 9 所示。由于苹果生长期间果农通常以间距 0.2、 0.25、0.3m 进行疏花疏果操作[25]，以保证苹果品质与产量。本文考虑苹果结果间距以及采摘空间大小，以 0.2m 为间距沿 $y$ 方向在 0~0.4m 范围内均匀引入 3 个约束平面(平面 1、平面 2、平面 3)作为采摘平面进行对照实验，以观察不同约束平面对空间范围内轨迹规划的影响。3 个平面具体位置为：平面 1(蓝色)方程为 $y=0\left(0\le x\le 0.5,\ 0.752\le z\le 1.252\right)$，平面 2(绿色)方程为 $y=0.2\left(0\le x\le 0.5,\ 0.752\le z\le 1.252\right)$，平面 3(红色)方程为 $y=0.4\left(0\le x\le 0.5,\ 0.752\le z\le 1.252\right)$。

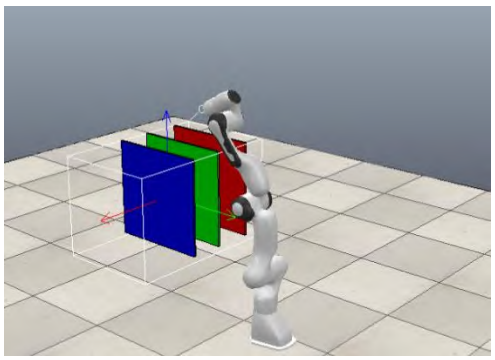
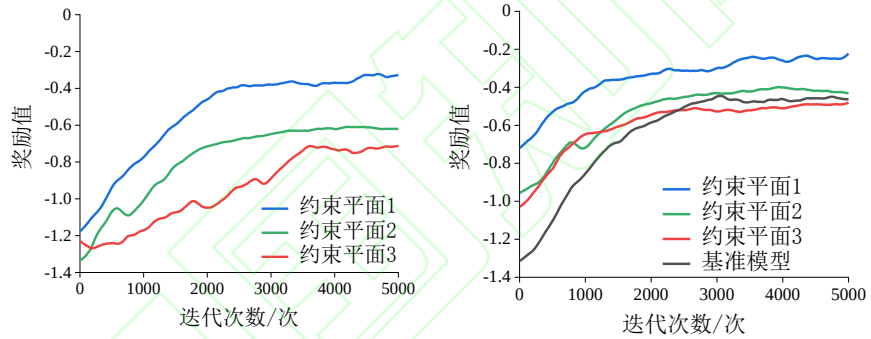


图 9 采摘空间示意图

Fig.9 Simulation scene

按照渐进空间约束分步训练策略，分别在 3 个采摘平面上进行训练，在 3 组网络模型的基础上，进一步在实际采摘环境下进行轨迹规划网络的训练，观察不同位置的采摘平面对实际环境的泛化性。共进行 5000 次迭代训练，图 10 为训练期间奖励值的变化情况，图 10a 为在 3 个采摘平面上进行预训练的结果，图 10b 为利用相应预训练网络参数在三维空间中训练的结果。




(a) 引入空间约束的奖励值变化曲线　　(b) 三维采摘空间下奖励值变化曲线

图 10 奖励值变化曲线

Fig.10 Reward value change curves

由图 10a 可知，随着迭代次数的增加，采摘机械臂所获奖励逐渐变大，最终达到收敛状态。由图 10b 可知，随着迭代次数的增加，采摘机械臂所获奖励逐渐变大，最终达到收敛状态。

从图 10b 可以看出，训练开始时基准实验的曲线奖励值起点在-1.30 附近，而经过预训练的奖励曲线的奖励值起点在-1.00 附近，这表明经过预训练，采摘机械臂的动作策略获得了一些先验知识，具有较好的初始假设，减少了无效探索，相对于随机初始化性能有较为明显的提升。表 4 统计了迭代中 4000~5000 次的奖励值均值以及训练期间收敛所用回合数，其中基准模型为直接在三维空间中训练所得模型。

表 4 训练结果对比

Tab.4 Comparison of training results

| | 基准模型 | 基于约束平面 1 的模型 | 基于约束平面 2 的模型 | 基于约束平面 3 的模型 |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| 奖励值初值 | -1.30 | -0.70 | -0.95 | -1.06 |
| 奖励值均值 | -0.46 | -0.25 | -0.42 | -0.49 |
| 收敛所需迭代次数 | 3000 | 1100 | 1800 | 2000 |

由表 4 可知，在收敛速度方面，基于约束平面的模型收敛所需迭代次数分别为 1100、1800、2000，而基准模型经过 3000 次迭代达到收敛，基于约束平面 1、2、3 的分步训练网络收敛速度分别比基准模型提升了 63.33%、40%和 33.33%。这表明引入约束平面后，由于网络初始参数是通过预训练得到，训练初期策略的盲目性大大减少，使基于约束平面的模型减少了学习时间，在三维采摘空间上训练的收敛速度明显加快。

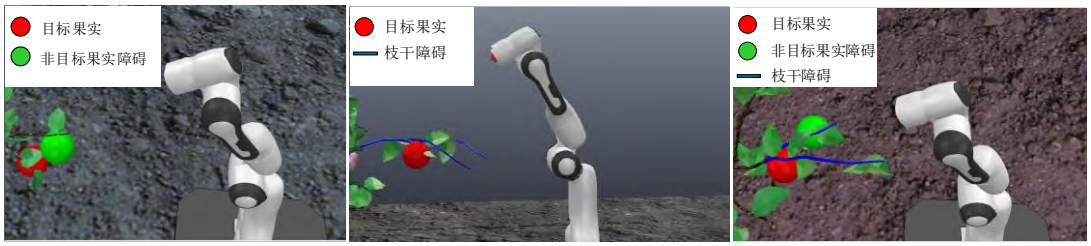
同时，由表 4 可知，基于约束平面 1 的分步训练策略在提升模型性能方面最为显著：奖励值初值为-0.70，收敛后其奖励值均值稳定在-0.25，相比于基准模型，奖励值初值和均值分别提升了 46.15%和 45.65%，表明基于约束平面 1 的训练策略在网络性能上提升效果明显。由于约束平面 1 位于采摘空间的中心位置，因此，基于约束平面 1 得到的模型相比于其他约束平面得到的模型在后续训练上其动作策略的空间泛化性和空间适应性更强。

为了测试模型的效果，本文分别统计了基准模型和基于约束平面 1 的训练模型成功采摘 100 次所需时间为 320、260s。

由以上实验结果可知，基于分步训练策略模型在收敛速度和性能上都得到了大幅度的提升，说明采摘机械臂利用渐进空间约束分步训练策略进行轨迹规划能显著加速训练过程和提升模型性能。

## 2.3 基于迁移学习的 DDPG 算法实验与分析

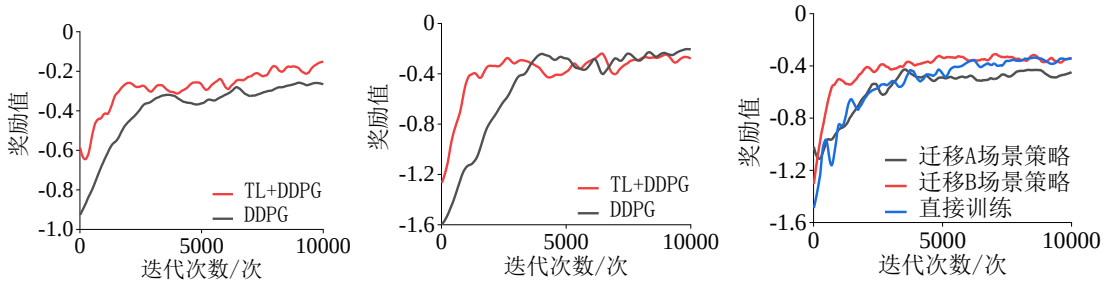
对于有障碍的场景，本文根据真实的采摘场景设计了 3 种场景，如图 11 所示，分别是模拟非目标果实障碍场景 (A 场景)、模拟枝干障碍场景 (B 场景)、模拟混杂障碍场景 (C 场景)。场景中红色苹果为目标果实，绿色苹果为非目标果实障碍，蓝色枝干为障碍。A 场景和 B 场景分别针对的是单一障碍场景，C 场景为混杂障碍场景。图 12 为不同场景下训练时奖励值的变化曲线图。




(a) A 场景　　(b) B 场景　　(c) C 场景

图 11 3 种场景示意图

Fig.11 Simulation scenes




(a) A 场景　　(b) B 场景　　(c) C 场景

图 12 不同场景下训练时的奖励值变化曲线

Fig.12 Reward value change curves

实验共执行了 10000 次迭代训练，由图(12)可以看出，随着迭代次数的增加，在各场景下采摘机械臂所获奖励逐渐变大，并最终达到收敛状态。

表 5 为 A 场景和 B 场景训练迭代 8000~10000 次的奖励值的均值以及训练期间收敛所用迭代次数。

表 5 训练结果对比

Tab.5 Comparison of experimental results

| | A 场景 | | B 场景 | |
|---|---|---|---|---|
| | TL+DDPG 算法 | DDPG 算法 | TL+DDPG 算法 | DDPG 算法 |
| 奖励值初值 | -0.62 | -0.92 | -1.35 | -1.60 |
| 奖励值均值 | -0.15 | -0.27 | -0.26 | -0.24 |
| 收敛迭代次数 | 2000 | 3500 | 2600 | 3900 |

由图 12 和表 5 可以看出，相比较采用 DDPG 算法直接训练，在 A 场景和 B 场景中基于迁移学习的 DDPG 算法训练收敛所需迭代次数从 3500 和 3900 分别缩短到 2000 和 2600，收敛速度分别提升了 42.86%和 33.33%。说明机械臂在无障碍场景下的轨迹规划策略能够为单一障碍场景的轨迹规划提供指导，可以有效缩短训练时间。

同时由表 5 可知，在 A 场景和 B 场景中基于 TL+DDPG 算法，在开始阶段奖励值初值分别为-0.62 和-1.35，比直接训练分别提升了 32.61%和 15.63%。并且，在 A 场景中该方法收敛后奖励值均值稳定在-0.15，相较于直接训练提升了 44.44%。而在 B 场景中两种方法的奖励值均值相差不大，TL+DDPG 算法的奖励值均值略低于 DDPG 算法，说明从无障碍场景向单一障碍场景进行迁移时，源任务策略在训练前期能够指导机械臂快速接近目标，该策略向较为简单的 A 场景进行避障迁移适应性强于较为复杂的 B 场景。

实际采摘环境通常存在多种障碍，为了观察采摘机械臂在面对混杂障碍时，单一障碍场景下获得的策略能否为采摘任务提供合适的指导，我们将 C 场景设计成混杂障碍场景，并分别迁移 A 场景和 B 场景的策略来指导采摘机械臂在 C 场景下进行轨迹规划任务。表 6 为

C 场景下应用不同策略训练迭代 8000~10000 次的奖励值均值以及训练期间收敛所用迭代次数。图 13 为混杂障碍场景下的收敛所需迭代次数。

表 6 混杂场景下不同策略训练结果对比

Tab.6 Comparison of training results in complex scenarios

| | 迁移 A 场景策略 | 迁移 B 场景策略 | DDPG 算法 |
|---|---|---|---|
| 奖励值初值 | -1.00 | -1.30 | -1.52 |
| 奖励值均值 | -0.45 | -0.35 | -0.36 |
| 收敛迭代次数 | 3600 | 2200 | 6400 |

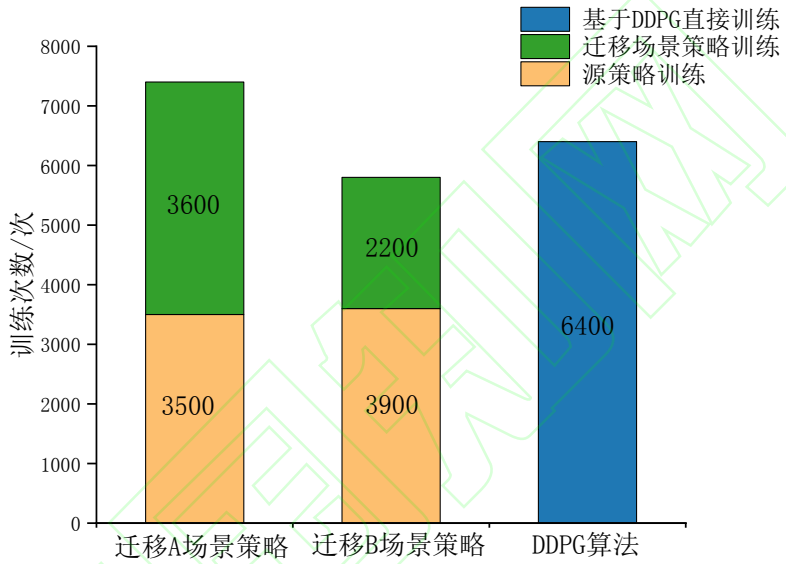


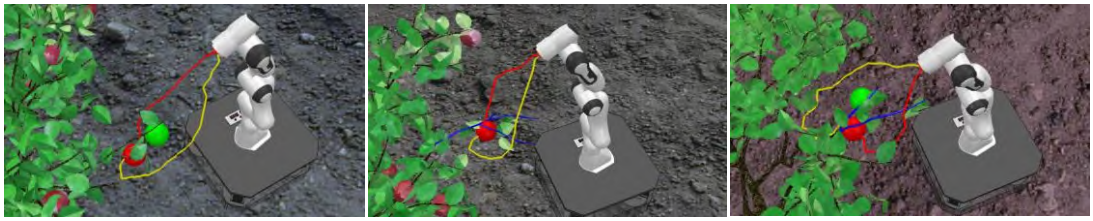

图 13 混杂障碍场景下收敛所需迭代次数

Fig.13 The number of converging rounds

由表 6 可知，迁移 A 场景和 B 场景的策略来指导采摘机械臂在 C 场景下进行轨迹规划任务，其收敛所需迭代次数分别为 3600 和 2200，相比较直接训练，收敛速度分别提升了 43.75%和 65.63%。以上结果表明基于单一障碍场景下的迁移训练相较于基于 DDPG 算法的直接训练，任务收敛速度有大幅度提升。如图 13 所示，当考虑源策略训练次数时，迁移 A 场景和 B 场景策略的总收敛迭代次数分别为 7100 和 6100，表明在混杂障碍场景中迁移 B 场景的策略更有助于提升训练速度。

同时由表 6 可知，迁移 A 场景策略和迁移 B 场景策略模型的奖励值初值分别为-1.00 和 -1.30，比 DDPG 算法分别提升了 34.21%和 14.47%；以上两种迁移方法在模型收敛后奖励值均值分别稳定在-0.45 和-0.35，相比较直接训练，迁移 B 场景策略的奖励值均值略大。这表明在苹果采摘中，从 A 场景和 B 场景向混杂障碍场景迁移时，均可以提供较好的模型初始化参数；同时，在面对混杂障碍场景时，源任务中障碍环境较为复杂，更利于提高混杂障碍

场景下的模型性能。



(a) A 场景　　(b) B 场景　　(c) C 场景

图 14 不同场景下的避障轨迹

Fig.14 Obstacle avoidance trajectories in each scenario

图 14 是在不同场景下采摘机械臂的避障轨迹。红色苹果为目标位置，蓝色树枝为障碍，绿色未成熟苹果为果实障碍。黄线为 DDPG 算法为采摘机械臂规划出的轨迹，红线为基于迁移学习的 DDPG 算法为采摘机械臂规划出的轨迹。可以看到 DDPG 算法在进行避障轨迹规划时得到的轨迹比基于迁移学习的 DDPG 算法得到的轨迹长度更长一些。

# 3 结论

采用深度强化学习方法进行采摘机械臂轨迹规划，为提高无障碍环境下模型的收敛速度和最终性能提出了渐进空间约束的分步训练策略。经过仿真实验验证，相比较直接训练，利用渐进空间约束的分步训练策略对奖励值初值最大提升幅度为 46.15%，收敛速度最大提升幅度为 63.33%。针对复杂障碍环境，提出了基于迁移学习的 DDPG 算法，将轨迹规划的最优策略由无障碍场景迁移到单一障碍场景、单一障碍场景迁移到混杂障碍场景中。经过仿真实验验证，将无障碍场景策略迁移到单一障碍场景中时奖励值初值提升幅度为 32.61%，网络收敛速度最大提升幅度为 42.86%；在将单一障碍场景策略迁移到混杂障碍场景中时奖励值初值最大提升幅度为 34.21%，对网络收敛速度最大提升幅度为 65.63%。

[参考文献]


[1] 刘晓洋, 赵德安, 贾伟宽, 等. 基于超像素特征的苹果采摘机器人果实分割方法[J/OL]. 农业机械学报, 2019, 50(11):15-23.
LIU Xiaoyang, ZHAO Dean, JIA Weikuan, et al. Fruits segmentation method based on super pixel features for apple harvesting robot[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(11):15-23. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20191102&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2019.11.002. (in Chinese)

[2] 刘继展, 朱新新, 袁妍. 枝上柑橘果实深度球截线识别方法[J/OL]. 农业机械学报, 2017, 48(10):32-39.

LIU Jizhan, ZHU Xinxin, YUAN Yan. Depth-sphere transversal method for on-branch citrus fruit recognition[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(10):32-39.
http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20171004&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2017.10.004. (in Chinese)

[3] TIEN T N, ERDAL K, JOSSE D B, et al. Task and motion planning for apple harvesting robot [J]. IFAC Proceedings Volumes, 2013, 46(18):247-252.

[4] 殷建军, 董文龙, 梁利华, 等. 复杂环境下农业机器人路径规划优化方法[J/OL]. 农业机械学报, 2019, 50(5):17-22.
YIN Jianjun, DONG Wenlong, LIANG Lihua, et al. Optimization method of agricultural robot path planning in complex environment[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(5):17-22.
http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20190502&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2019.05.002. (in Chinese)

[5] 贾庆轩，陈钢，孙汉旭，等. 基于 A*算法的空间机械臂避障路径规划[J].机械工程学报, 2010, 46(13): 109-115.
JIA Qingxuan, CHEN Gang, SUN Hanxu, et al. Path planning for space manipulator to avoid obstacle based on A* algorithm[J]. Journal of Mechanical Engineering, 2010, 46(13): 109-115. (in Chinese)

[6] 刘可, 李可, 宿磊, 等. 基于蚁群算法与参数迁移的机器人三维路径规划方法[J/OL]. 农业机械学报, 2020, 51(1):29-36.
LIU Ke, LI Ke, SU Lei, et al. Robot 3D path planning method based on ant colony algorithm and parameter transfer[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(1):29-36.
http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20200103&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2020.01.003. (in Chinese)

[7] 苑严伟, 张小超, 胡小安. 苹果采摘路径规划最优化算法与仿真实现[J]. 农业工程学报, 2009, 25(4):141-144.
YUAN Yanwei, ZHANG Xiaochao, HU Xiao'an. Algorithm for optimization of apple harvesting path and simulation[J]. Transactions of the CSAE, 2009,25(4):141－144. (in Chinese)

[8] 张强, 陈兵奎, 刘小雍, 等. 基于改进势场蚁群算法的移动机器人最优路径规划[J/OL]. 农业机械学报, 2019, 50(5):23-32,42.
ZHANG Qiang, CHEN Bingkui, LIU Xiaoyong, et al. Ant colony optimization with improved potential field heuristic for robot path planning[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(5):23-32,42.
http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20190503&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2019.05.003. (in Chinese)

[9] 王宇, 陈海涛, 李海川. 基于引力搜索算法的植保无人机三维路径规划方法[J/OL]. 农业机械学报,2018,49(2):28-33,21.
WANG Yu, CHEN Haitao, LI Haichuan.3D path planning approach based on gravitational search algorithm for sprayer UAV[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(2):28-33,21.
http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20180204&jour

nal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2018.02.004. (in Chinese)

[10] 袁朝春，翁烁丰，何友国，等. 基于改进人工势场法的路径规划决策一体化算法研究[J/OL]. 农业机械学报, 2019, 50(9):394-403.
YUAN Chaochun, WENG Shuofeng, HE Youguo, et al. Integration algorithm of path planning and decision-making based on improved artificial potential field[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(9):394-403. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20190946&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2019.09.046. (in Chinese)

[11] 姬伟，程风仪，赵德安，等. 基于改进人工势场的苹果采摘机器人机械手避障方法[J/OL]. 农业机械学报, 2013, 44(11):253-259.
JI Wei, CHENG Fengyi, ZHAO Dean, et al. Obstacle avoidance method of apple harvesting robot manipulator[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2013, 44(11):253-259. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20131143&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2013.11.043. (in Chinese)

[12] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. 2nd edition, Cambridge, MA: A Bradford Book, 2018:1-13.

[13] GU S, HOLLY E, LILLICRAP T, et al. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates[C]//2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017: 3389-3396.

[14] WEN Shuhuan, CHEN Jianhua, WANG Shen, et al. Path planning of humanoid arm based on deep deterministic policy gradient[C]// 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE, 2018: 1755-1760.

[15] KIM M S, HAN D K, PARK J H, et al. Motion planning of robot manipulators for a smoother path using a twin delayed deep deterministic policy gradient with hindsight experience replay[J]. Applied sciences, 2020, 10(2):575.

[16] XIE Jiexin, SHAO Zhenzhou, LI Yue, et al. Deep reinforcement learning with optimized reward functions for robotic trajectory planning[J]. IEEE Access, 2019, 7:105669-105679

[17] LAZARIC A. Transfer in reinforcement learning: a framework and a survey[M]//Reinforcement Learning. Springer, Berlin, Heidelberg, 2012: 143-173.

[18] 胡晓东，黄学祥，胡天健，等. 一种动态环境下空间机器人的快速路径规划方法[J]. 空间控制技术与应用, 2018, 44(5):17-24.
HU Xiaodong, HUANG Xuexiang, HU Tianjian, et al. A fast path planning method for space robot in dynamic environment[J]. Aerospace Control and Application, 2018, 44(5):17-24. (in Chinese)

[19] 陈建华. 基于深度强化学习的机械臂运动规划研究[D]. 秦皇岛：燕山大学, 2019.
CHEN Jianhua. Research on motion planning of robot arm based on deep reinforcement learning[D]. Qinhuangdao: Yanshan University, 2019. (in Chinese)

[20] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. Computer science, 2015, 8(6):A187.

[21] 顾宝兴. 智能移动式水果采摘机器人系统的研究[D]. 南京：南京农业大学, 2012.
GU Baoxing. Research on intelligent mobile fruit picking robot[D]. Nanjing: Nanjing Agricultural University, 2012. (in Chinese)

[22] 尹建军，武传宇，YANG S X，等. 番茄采摘机器人机械臂避障路径规划[J/OL]. 农业机

械学报, 2012, 43(12): 171-175,157.

YIN Jianjun, WU Chuanyu, YANG S X, et al. Obstacle-avoidance path planning of robot arm for tomato-picking robot[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2012, 43(12): 171-175,157. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20121231&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2012.12.031. (in Chinese)

[23] 蔡健荣, 赵杰文, THOMAS R, 等. 水果收获机器人避障路径规划[J]. 农业机械学报, 2007, 38(3):102-105,135.

CAI Jianrong, ZHAO Jiewen, THOMAS R, et al. Path planning of fruits harvesting robot[J]. Transactions of the Chinese Society for Agricultural Machinery, 2007, 38(3):102-105,135. (in Chinese)

[24] CAO Xiaoman, ZOU Xiangjun, JIA Chunyang, et al. RRT-based path planning for an intelligent litchi-picking manipulator[J]. Computers and Electronics in Agriculture, 2019, 156:105-118.

[25] 王毅, 李朔南. 疏花疏果间距对长富 2 苹果产量和产值的效应[J]. 中国果树, 1999(1):7-9.

WANG Yi, LI Shuonan. Effect of flower and fruit spacing on yield and value of long rich 2 apples[J]. China Fruits, 1999(1):7-9. (in Chinese)