arXiv:2304.00227v1 [cs.RO] 1 Apr 2023

Tracker：采用基于模型的强化学习实现连接细型麦基宾肌肉的人类手指跟踪控制

Daichi Saito[1]*、Eri Nagatomo[1]*、Jefferson Pardomuan[1],和Hideki Koike[1]

摘要—为了采用软体手部外骨骼来支持日常生活活动，有必要利用外骨骼对手指关节进行精确控制。控制关节跟随给定轨迹的问题称为跟踪控制问题。在本研究中，我们聚焦于连接细型麦基宾肌肉的人类手指的跟踪控制问题。为了利用细型麦基宾肌肉实现精确控制，存在两个问题：一是肌肉的复杂特性，例如非线性、滞后以及真实世界中的不确定性；二是难以获得肌肉和人类手指的精确模型。为了解决这些问题，我们采用了DreamerV2，这是一种基于模型的强化学习方法，但学习得到的模型无法生成目标轨迹。因此，我们提出了Tracker，它是DreamerV2针对跟踪控制问题的扩展。在实验中，我们表明，对于模拟人类食指从掌骨到近节指骨部分的两连杆机械臂的控制，Tracker能够实现比PID低约81%的误差。Tracker在训练约60 分钟后，能够以较小误差实现对人类食指第三关节的控制。此外，在用户手指上预训练的策略取下并重新连接细型麦基宾肌肉后，通过微调该策略，仅需约15 分钟即可达到几乎相同的精度；这一时间少于首次训练所需的时间，也能达到取下肌肉前的精度。

I. 引言

外骨骼在控制人体、辅助日常生活活动（ADL）方面发挥着重要作用，例如行走 [1], 抓握和灵巧操作 [2], [3]。聚焦于手部外骨骼，已有研究提出了各种手套，用于康复、辅助残障人士完成手动任务以及灵巧技能习得。与电机等电动执行器相比，采用气动人造肌肉等软体执行器的外骨骼 [4], [5] 由于具有柔顺性和轻量化的特点，正呈现出日益增长的趋势。已有多项研究提出了使用细型麦基宾肌肉的软体手部外骨骼，用于康复 [6], 手部辅助 [7] 以及音乐技能习得 [8]。

为了让软体手部外骨骼能够支持日常生活活动并帮助使用者习得技能，必须借助软体手部外骨骼精准控制手指关节。控制关节跟随给定轨迹的问题称为跟踪控制问题。由于气动人工肌肉这类软体执行器具有高度非线性、滞后（这是一种随时间变化的特性），并且在真实世界中还存在各种不确定性，因此很难对其进行精准控制。这些不确定性源于橡胶管和纤维套管的特性、气压的复杂性以及摩擦。由于控制困难，先前的研究仅关注一些简单动作，例如张开和握紧手掌以及手指敲击。利用软体手部外骨骼控制手指关节跟踪轨迹，对于实现灵巧的手指运动至关重要，但设计相应的控制策略颇具挑战，至今仍未解决。本研究关注佩戴细型麦克宾肌肉的人类手指的跟踪控制问题（图 1）。

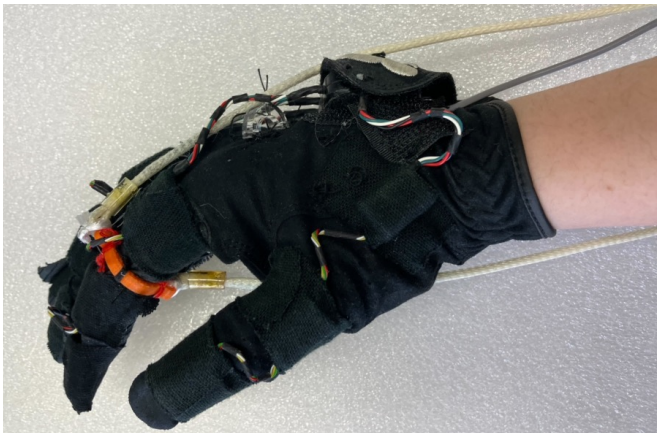


图1. 安装有细型麦基宾肌肉的人类手指。

许多以往研究将开环控制 [6],[7] 和反馈控制（如 PID 控制 [9], [10], [11],）用作气动执行器的控制方法。由于难以建立手指和气动人工肌肉的动力学模型，开环控制器无法实现精确控制。此外，PID 控制器需要手动调节参数。找到最优控制参数并不容易，因为这需要熟练的技巧。近年来，深度学习技术的进步已被用于控制气动人工肌肉。研究人员采用卷积神经网络控制织物执行器的变形 [12]。在 [12],中，卷积神经网络使用随机收集的数据集进行训练。然而，这项研究仅关注静态变形。由于无法明确设定代价函数，将这种方法用于动态控制需要海量数据。强化学习是一种能够设定称为奖励的代价函数，并将其应用于各种控制问题的技术，例如游戏、机器人抓取和操作 [13]。研究人员使用深度强化学习训练了肌肉骨骼机器人的控制策略 [14]。在 [14],中，研究人员使用了无模型强化学习。无模型

[1]作者来自日本东京的东京工业大学计算机科学系，saito.d.ah@m.titech.ac.jp

*这些作者贡献相同。

强化学习需要与环境进行大量交互，因此需要物理仿真器。我们的研究聚焦于控制附着细型麦基宾肌肉的人类手指，因此需要人类手指和细型麦基宾肌肉的精确模型。然而，这些模型难以获取。因此，需要一种数据高效的学习方法，来控制附着细型麦基宾肌肉的人类手指。

基于模型的强化学习是一种数据高效的方法，可用于学习控制策略 [13]。基于模型的强化学习是一种从环境中的经验中学习环境模型，并根据在已学习环境模型中的经验训练策略的技术。近年来，世界模型 [15] 已被用于学习精确的环境模型 [16]。DayDreamer [17] 表明，可以使用基于模型的强化学习 DreamerV2 [16] 来学习实体机器人控制策略。受 [17], 启发，我们展示了基于模型的强化学习对于学习连接细型麦克基本人工肌肉的人类手指的控制策略是有效的。在 DayDreamer 中，基于模型的强化学习对软体执行器控制的有效性尚未得到研究。一般而言，由于软体执行器的特性复杂，其控制比硬致动器更困难。世界模型无法生成用于跟踪的目标轨迹，因为目标轨迹由环境之外给定，而非由环境给定。请注意，如果将轨迹作为世界模型的输入，就能够生成轨迹，但无法保证生成的轨迹就是预期轨迹。因此，仅采用 DreamerV2 无法学习跟踪控制策略。为此，我们提出了 Tracker，一种针对跟踪控制问题的基于模型的强化学习方法（图 2）。Tracker 是 DreamerV2 在跟踪控制问题上的扩展。Tracker 通过与世界模型交互来训练跟踪控制策略，而该世界模型是根据环境中的经验学习得到的。策略根据世界模型生成的观测以及从预先收集的轨迹数据集中随机选取的目标轨迹来决定动作。在执行过程中，跟踪控制策略根据对人类手指的观测和目标轨迹来决定动作。我们

可以在关节运动范围内选择任意目标轨迹。

本研究聚焦于利用基于模型的强化学习，对连接细型麦基宾肌肉的人类手指进行跟踪控制的问题。本研究的贡献如下：

- 我们研究了基于模型的强化学习在控制连接细型麦克基本人工肌肉的人类手指方面的有效性。
- 我们将基于模型的强化学习扩展至跟踪控制问题。
- 我们通过实验证明，利用扩展后的基于模型的强化学习，可以学习使用细型麦基宾肌肉的跟踪控制策略。

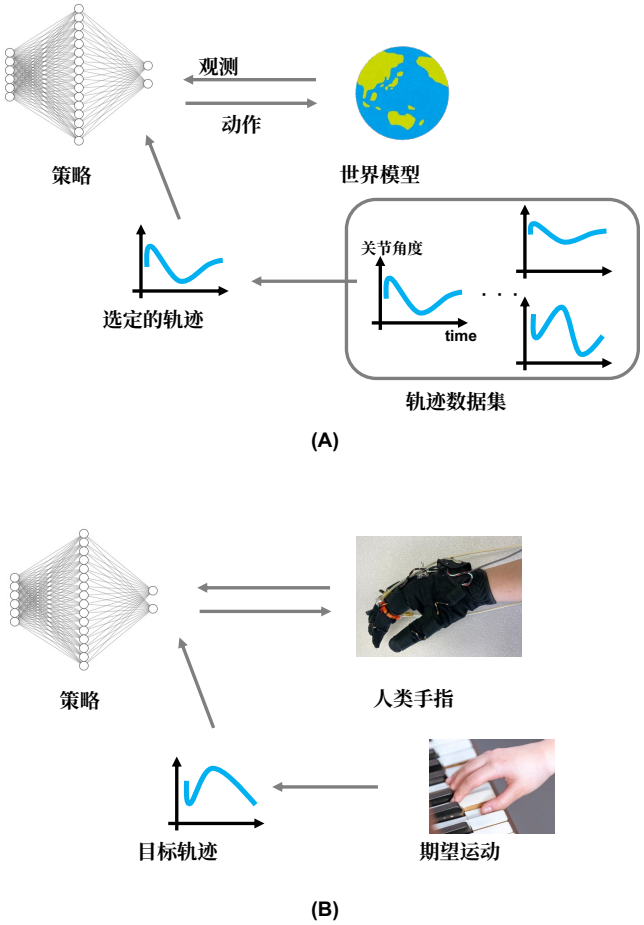


图 2. Tracker 概览。(A) 为训练阶段示意图，(B) 为执行阶段示意图。(A) Tracker 通过与世界模型交互来训练跟踪控制策略。该策略根据世界模型生成的观测和目标轨迹决定动作。目标轨迹从预先收集的轨迹数据集中随机选取。(B) 在执行过程中，跟踪控制策略根据人类手指的观测和目标轨迹决定动作。目标轨迹是用户的期望运动（例如，本图中的钢琴演奏任务）。

II. 相关工作

先前的研究使用开环控制器控制气动人工肌肉 [6], [7]。开环控制器易于实现，但仅适用于简单运动（例如，手指的张合和手指敲击）。由于执行器具有复杂特性（例如高度非线性和滞后），难以建立其动力学模型，因此很难采用开环控制器对气动人工肌肉进行跟踪控制。反馈控制器（例如 PID控制器）是一种无模型控制方法。许多先前的研究使用 PID控制器控制气动人工肌肉 [9], [10], [11], [18]。PID控制器是一种通用控制器，但需要熟练的技术来寻找最优控制参数。已有研究提出了用于补偿气动人工肌肉响应迟缓的控制方法 [19], [20]。然而，这些方法需要对执行器进行精确建模。

深度神经网络已被用于控制气动人工肌肉。使用深度神经网络的控制器无需手动进行参数调节。已有研究采用卷积神经网络控制织物执行器的变形 [12]。该研究关注的是静态变形，而非动态变形。在本研究中，这种方法被称为“直接回归”。直接回归无法显式设定代价函数，因此在动态控制中需要大量数据。已有研究将直接回归与循环神经网络结合，用于柔性机械臂的动态控制 [21],然而， [21]该研究关注的是简单的动态运动。

强化学习是一种可以设定代价函数的技术，该代价函数称为奖励。强化学习分为两种类型：无模型强化学习和基于模型的强化学习。无模型强化学习需要与环境进行大量交互，因此通常需要模拟器。事实上， [14] 在模拟器中训练了肌骨机器人控制策略。此外，还应考虑仿真到现实问题 [22] 。基于模型的强化学习所需的交互量少于无模型强化学习。在本研究中，我们采用基于模型的强化学习，因为难以获取人类手指和气动人工肌肉的精确模型。基于模型的强化学习方法有多种 [15], [23], [16]。利用基于模型的强化学习，可以学习实体机器人控制策略 [17]。 [17] 表明，基于模型的强化学习能够在真实世界中进行在线学习。然而，尚无研究证明基于模型的强化学习能够有效学习软体机器人的控制。通常，软体机器人的控制比刚性机器人的控制更加困难。在本研究中，我们考察基于模型的强化学习对细型麦克基本人工肌肉这一软体执行器进行控制的有效性。

III. 方法

本研究聚焦于由细型麦基宾肌肉驱动的人类手指的跟踪控制问题。要实现精确的跟踪控制，需要解决以下两个问题：

- 细型麦基宾肌肉具有复杂特性，例如非线性、滞后（这是一种随时间变化的特性），以及真实世界中的不确定性
- 难以获取细型麦基宾肌肉和人类手指的精确模型

为解决前一个问题，我们采用基于深度神经网络的控制策略，以应对非线性、滞后现象和各种不确定性。为解决后一个问题，由于训练控制策略时必须使用真实的人类手指和执行器，因此我们需要一种数据高效学习方法。于是，我们采用基于模型的强化学习，这是一种数据高效的强化学习。受先前研究的启发 [17], ，我们使用 DreamerV2 [23],，它是一种基于世界模型的强化学习方法。然而，目标轨迹由环境之外给定，而非由环境生成；因此，无法通过世界模型生成该轨迹。因此，我们提出了 Tracker，它是针对跟踪控制问题对 DreamerV2 的扩展。

本节介绍 DreamerV2 和 Tracker，这是一种针对跟踪控制问题的基于模型的强化学习方法。我们还将介绍观测以及轨迹（即目标关节角度）、动作和奖励。

A. 预备知识：*DreamerV2*

Tracker 基于 DreamerV2，后者由递归状态空间模型（RSSM） [24], （世界模型的组成部分）、执行器和评论家构成。RSSM 通过过往经验数据集进行训练。经验数据集通过执行器与环境交互而不断扩充。执行器和评论家则利用 RSSM 生成的经验序列进行训练。

具体而言，DreamerV2 由六个组件构成：RSSM（递归模型、表征模型、转移模型）、观测预测器、奖励预测器、执行器和评论家。这些组件通常用以下方程表示：

$$\text{Recurrent model: } h_t = f_\phi(h_{t-1}, z_{t-1}, a_{t-1}) \tag{1}$$

$$\text{Representation model: } z_t \sim q_\phi(z_t|h_t, x_t) \tag{2}$$

$$\text{Transition model: } \hat{z}_t \sim p_\phi(\hat{z}_t|h_t) \tag{3}$$

$$\text{Observation predictor: } \hat{x}_t \sim p_\phi(\hat{x}_t|h_t, z_t) \tag{4}$$

$$\text{Reward predictor: } \hat{r}_t \sim p_\phi(\hat{r}_t|h_t, z_t) \tag{5}$$

$$\text{Discount predictor: } \hat{\gamma}_t \sim p_\phi(\hat{\gamma}_t|h_t, z_t) \tag{6}$$

$$\text{Actor: } \hat{a}_t \sim p_\psi(\hat{a}_t|\hat{z}_t) \tag{7}$$

$$\text{Critic: } v_\xi(\hat{z}_t) \approx E_{p_\phi, p_\psi}\left[\sum_{i \geq t} \hat{\gamma}^{i-t} \hat{r}_i\right] \tag{8}$$

z_t 和 h_t 分别表示随机潜变量和确定性潜变量。 a_t 表示动作， o_t 表示来自环境的观测， γ_t 表示回合结束的概率。执行器和评论家根据随机潜变量z_t 和奖励 r_t进行学习。

目标轨迹由环境外部提供，而非由环境生成；因此，RSSM 无法生成该轨迹。请注意，如果将轨迹纳入观测中，模型便可以生成轨迹，但无法保证生成的轨迹就是预期轨迹。当关节运动范围受限，或一个关节的运动与另一个关节的运动相关联时，这就会成为问题。因此，我们提出 Tracker，一种用于轨迹跟踪控制问题的基于模型的强化学习方法。

B. Tracker

Tracker 是 DreamerV2 在跟踪控制问题上的扩展（图 3）。目标轨迹会影响奖励和动作，但不会影响观测。因此，我们使用目标

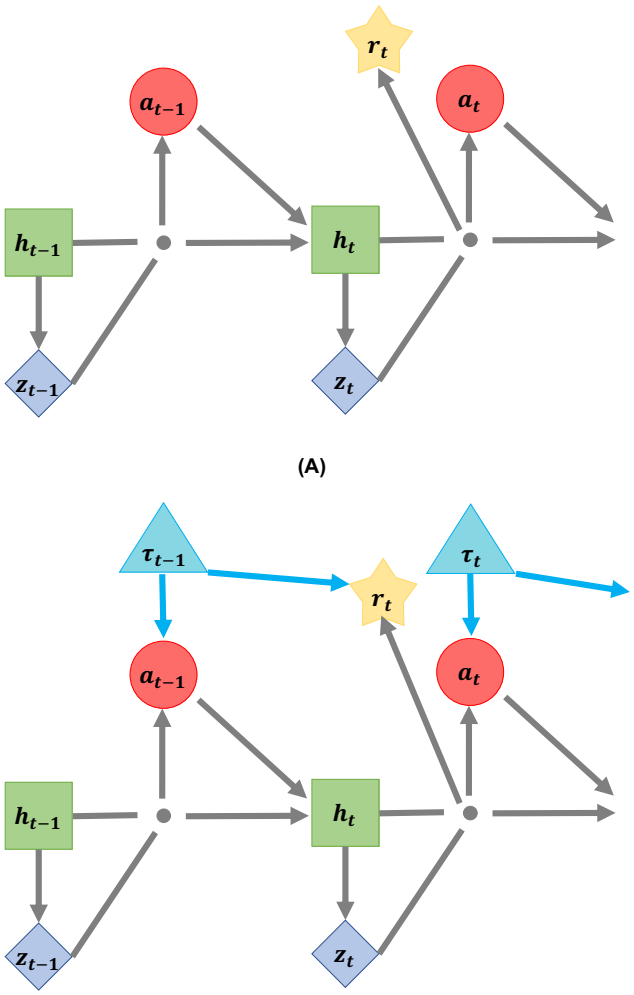


图3. 图模型与（A）DreamerV2和（B）Tracker的比较。DreamerV2与Tracker之间两个潜变量的转移 h_t, z_t 相同。Tracker中的动作 a_{t-1} 和奖励 r_t 以目标轨迹 τ_{t-1} 为条件。

轨迹 τ 对奖励预测器、执行器和评论家进行条件化。奖励预测器、执行器和评论家由以下方程表示：

$$\text{Reward predictor: } \hat{r}_t \sim p_\phi(\hat{r}_t|h_t, z_t, \tau_{t-1}) \tag{9}$$

$$\text{Actor: } \hat{a}_t \sim p_\psi(\hat{a}_t|\hat{z}_t, \tau_t) \tag{10}$$

$$\text{Critic: } v_\xi(\hat{z}_t, \tau_t) \approx E_{p_\phi, p_\psi}\left[\sum_{i\geq t} \hat{\gamma}^{i-t}\hat{r}_i\right] \tag{11}$$

执行器和评论家的输入相同（$\hat{z}_t$ 和 τ_t）；因此，可以按照传统方式训练这两个模型。

执行器和评论家通过长时域配对数据进行训练，该数据由目标轨迹和 RSSM 生成的序列组成。在执行器和评论家的学习过程中，从满足关节活动范围要求的轨迹数据集中随机选取目标轨迹。我们希望控制策略学习的任何轨迹都可以纳入轨迹数据集。

C. 观测、动作、轨迹、奖励

1）观测：观测 o_t 包括每个关节的角度和角速度，以及每个细型麦基宾肌肉的气压。设 n 为关节数量，m 为执行器数量，则 $o_t \in \mathbb{R}^{2n+m}$。

2) 动作：动作 a_t 包括每个细型麦基宾肌肉的气压；然后，$a_t \in \mathbb{R}^m$。

3) 轨迹：轨迹 τ_t 包括每个关节未来 l 步以内的角度和角速度（$\tau_t \in \mathbb{R}^{2ln}$）。通过提供未来轨迹，可以确定更具前瞻性的动作。与经典反馈控制器相比，基于深度神经网络的控制器的优势在于能够纳入误差以外的特征。

4) 奖励：奖励 r_t 由以下公式给出：

$$r_t = -\sum_{i=1}^{n}\{(\frac{\|\theta_{i,t} - \hat{\theta}_{i,t}\|}{\alpha})^2 + \frac{I_{i,t}}{\beta}\} - \sum_{i=1}^{m}(\frac{p_{i,t}}{\gamma})^2 \tag{12}$$

$\hat{\theta}_{i,t}$ 和 $\theta_{i,t}$ 分别是 i关节的可观测角度和目标角度。$\hat{I}_{i,t}$ 当$\theta_{i,t} \times \theta_{i,t} < 0$时等于 1，否则为 0。$\theta_{i,t}$ 是 i关节的角速度，$\theta_{i,t}$是 i关节的目标角速度。$p_{i,t}$ 表示 i执行器中的气压。第一项用于跟踪控制。为了实现平滑运动，加入了$I_{i,t}$，该项表示角速度。为避免过大的气压给关节和执行器带来负担，加入了第二项。

D. 实现细节

RSSM、观测预测器、奖励预测器、折扣预测器、执行器和评论家均实现为神经网络。循环模型使用全连接层和门控循环单元（GRU） [25]。其他组件仅使用全连接层。我们对所有组件的激活函数均使用指数线性单元（ELU） [26] 。我们使用Adam优化器 [27]。

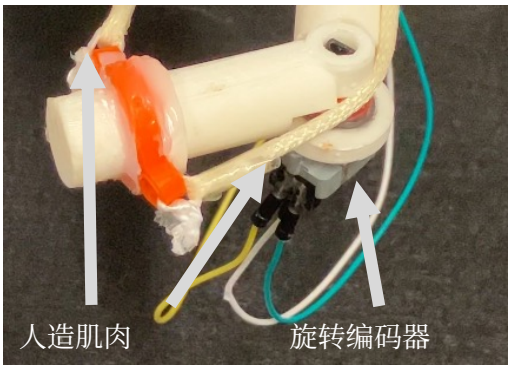
Tracker 中包含三个过程：世界模型学习、演员-评论家学习，以及从真实世界进行数据收集。同步运行这些过程会导致学习速度缓慢，因为在真实世界中收集数据需要很长时间。这会给用户带来负担。因此，我们参考 [28] 实现了异步数据收集，以加快训练速度。

IV. 实验一

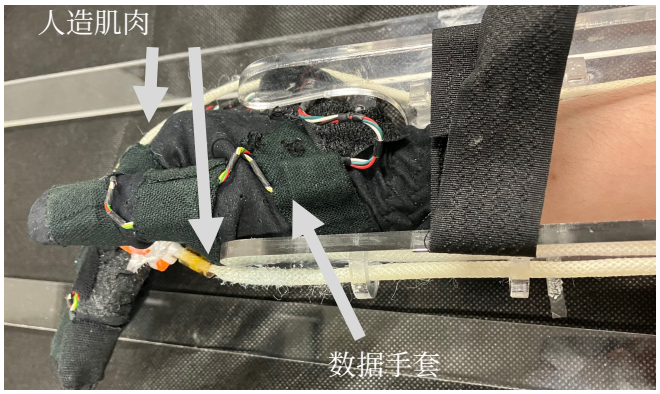
为了确认我们的控制策略 Tracker 是否能够应用于薄型人工肌肉，并检验我们的策略在跟踪控制问题上的性能，我们在一个模拟人类手指的两连杆机械臂上训练了该策略，并将 Tracker 与 PID控制器进行了比较。

A. 设置

1）硬件设置：本实验在一台两连杆机械臂上进行，该机械臂由两根连杆和一个关节（$n = 1$）组成，模拟了人类食指从掌骨到近节指骨的部分结构（图 4-A）。两根连杆通过旋转编码器连接，以获取



(A)



(B)

图 4. (A)：模仿人类手指的两连杆机械臂。(B)：连接了人造肌肉的人类手指。用两块丙烯酸板固定手部。

关节角度。关节角度被发送至微控制器 ESP32-WROOM-32[1]，，用于关节角度反馈。

我们使用了来自 s-muscle[2] 的细型 McKibben 人造肌肉，直径为 2.5mm。在第一连杆的末端安装了一个环，并将两个人造肌肉（$m = 2$）以拮抗状态连接到该环上。每 100ms 确定一次动作（气压）。

PID控制器的参数通过齐格勒-尼科尔斯法确定，然后通过若干次实验进行手动微调。

2）训练设置：RSSM生成的经验序列长度设为50。轨迹包括未来最多3步的角度和角速度（$l = 3$）。奖励参数设置如下：

$$\alpha = 40, \beta = 1000, \gamma = 50000 \tag{13}$$

奖励的参数设置如下：首先，计算参数，将每一项的奖励归一化至 -1 到1的范围内。随后，调整参数以控制各项的影响。

根据多次实验，世界模型、行动者模型和评论者模型的学习率分别被

[1]https://www.espressif.com/en/products/modules/esp32
2https://www.s-muscle.com/

表I Tracker和PID控制器对10 条相同目标轨迹（包括线性和非线性）的跟踪结果

	Tracker	PID控制器
MSE	86 ± 36	461 ± 433

设定为 4×10^{-4}、4×10^{-4} 和 2×10^{-4}。

在本实验的训练阶段，我们使用了在 $-25°$ 至$40°$范围内移动随机值的轨迹。训练持续进行，直到评估期间的总奖励大多达到最大奖励，且Tracker的损失收敛为止。在本实验中，训练大约耗时40分钟。

B. 跟踪控制结果

我们使用了十条轨迹作为评估轨迹：四条轨迹在每个时间步以不同范围内的随机值移动，两条轨迹以不同的恒定速度移动，两条轨迹每20个时间步以不同的值移动，以及两条以不同速度变化的正弦曲线。

Tracker和PID控制器的均方误差（MSE）如表I所示。均方误差通过以下公式计算：

$$\frac{1}{T}\sum_{t=1}^{T}(\theta_t - \hat{\theta}_t)^2 \tag{14}$$

$\hat{\theta}_t$ 和 θ_t 分别表示时间步t处的可观测角度和目标角度。 T 为评估轨迹的长度。Tracker的均方误差明显小于PID控制器的均方误差。

MSE差异最大的轨迹是以恒定速度 $\pm 10°$移动的轨迹，其中我们的控制器和PID控制器的MSE分别为102和1440（图5-a）。相比之下，差异最差的轨迹是每20步移动 $-5°$的轨迹，其中我们的控制器和PID控制器的MSE分别为58和48（图5-b）。由于训练过程中未学习阶跃移动的轨迹，Tracker无法控制关节。因此，对于这条阶跃轨迹，PID控制器能够取得更好的结果。所提出的控制器的MSE标准差小于PID控制器（表I）。这表明，与PID控制器相比，我们的控制器有潜力跟踪更广泛的轨迹。总而言之，该实验结果表明，Tracker能够比PID控制器更准确地跟踪更多轨迹。

V. 实验二

为了检验能否用细型麦基宾肌肉控制人类手指，我们使用控制器控制人类食指的第三关节。如果将细型麦基宾肌肉取下后再次安装到人类手指上，由于肌肉的状态和位置略有变化，我们就需要重新训练策略。从头开始训练策略既耗时又费力。在

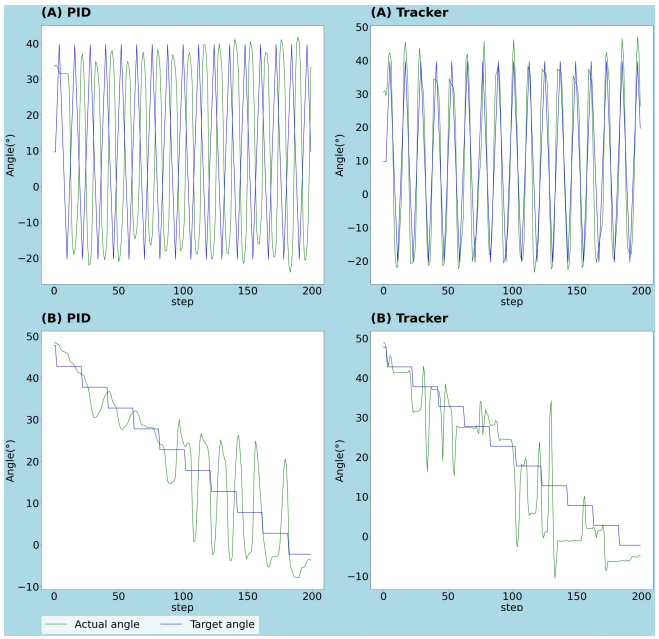


图5。（A）在评估轨迹中，我们的控制器、Tracker 和 PID 之间均方误差差异最大的轨迹（即以恒定速度 $\pm 10°$ 移动的轨迹）。（B）在评估轨迹中，均方误差差异最小的轨迹（即每隔 20 步移动 $-5°$ 的轨迹）。左侧为 PID控制器的结果，右侧为我们的控制器的结果。

这种情况下，微调策略通常有助于降低训练成本。因此，我们也验证了能否利用训练好的模型进行微调。

A. 设置

1) 硬件设置：本实验针对连接有细型麦基宾肌肉的人类食指第三关节（$n = 1$）进行（图4-B）。使用 Spice[3] 的 CobraGloves 获取手指弯曲角度。数据手套包含 16 个惯性测量单元传感器。肌肉与 IV 中相同，并处于拮抗状态。此外，为减少腕部角度变化所导致的重力影响，使用丙烯酸板固定腕部的位置和朝向。我们将手指完全伸直时的关节角度定义为 $0°$，并将手指屈曲时的角度定义为正角度。

2) 训练设置：我们使用了与 IV 中相同的奖励。我们初步收集了实际的人类手指运动，将其作为训练中使用的轨迹，并将这些轨迹裁剪至 $0°$ 到 $40°$，该范围几乎涵盖所用执行器实现的关节运动范围。轨迹数据集的采集方式确保数据集包含四种不同速度下的轨迹。在本实验中，与 IV 中一样，训练持续进行，直到评估期间的总奖励基本达到最大奖励，并且 Tracker 的损失收敛。在本实验中，完成训练大约需要 60 分钟。摘下手套并在几分钟后重新佩戴后，模型被

[3]https://mocap.jp/igs/

表 II 我们的控制器在不同速度下对三名参与者执行四种轨迹的结果。

MSE	参与者 1	参与者 2	参与者 3
首次训练	15 ± 5.4	19 ± 8.7	83 ± 11
重新佩戴后	13 ± 7.4	32 ± 8.4	229 ± 18
微调	19 ± 7.0	14 ± 4.9	67 ± 17

使用训练好的模型再次进行微调。所有人的这次训练大约耗时 15 分钟。训练期间没有人感到疼痛或不适。

3) 参与者：招募了手部尺寸适合数据手套的人员作为参与者。实验共邀请了三名参与者（平均年龄：25 岁，性别：女性）。实验期间，要求他们尽可能放松双手。

B. 实验结果

1) 跟踪控制结果：为考察 Tracker 的有效性，使用四种评估轨迹对训练好的策略进行评估：阶跃轨迹、慢速轨迹、稍快轨迹和快速轨迹。在阶跃轨迹中，关节角度逐步增大。其他轨迹中，关节角度以不同速度变化。这些轨迹未用于训练。

结果如表 II 和图 6 所示。两名参与者的均方误差较小，表明 Tracker 能够实现对人类手指关节的跟踪控制。参与者 1 和参与者 2 的均方误差差异也较小。然而，参与者 3 的均方误差较大。在训练阶段，参与者 3 的手指活动幅度往往小于其他参与者。此外，参与者 3 在下一次实验 V-B.2 中经过微调后取得了更好的结果。因此，延长研究时间或许能够改善这一情况。策略初始化是强化学习的主要问题之一 [29]。寻找良好的策略初始化方法是未来的研究方向。

快速轨迹的均方误差更大。手指移动得越快，改变关节角度所需的力就越大。因此，对于更快的轨迹，很难找到最优动作。我们需要根据运动速度定制执行器。

2) 微调结果：策略训练约 60 分钟后，拆下细型麦克基本人工肌肉并重新安装。随后对策略进行了约 15 分钟的微调。

微调后的策略的均方误差如表 I 所示。使用了与此前评估相同的轨迹。参与者 2 和 3 在微调后的均方误差比重新佩戴后的均方误差有所降低。参与者 1 的均方误差有所增加，但准确度几乎相同。结果表明，用户摘下并重新佩戴细型麦克基本人工肌肉后，通过微调基于用户手指预训练的策略，约 15 分钟即可达到几乎相同的准确度，这比首次训练所需的时间更短。

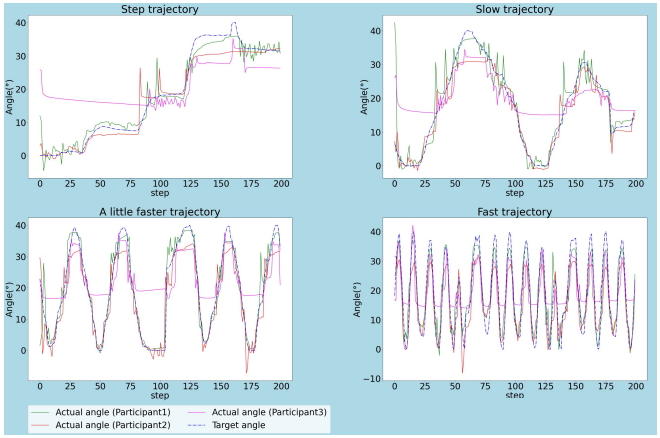


图6. 训练获得的模型在以不同速度运动时的四条评估轨迹及实际观测。绿色、红色和粉色线分别表示参与者1、2和3的实际关节角度。蓝色虚线表示目标角度。

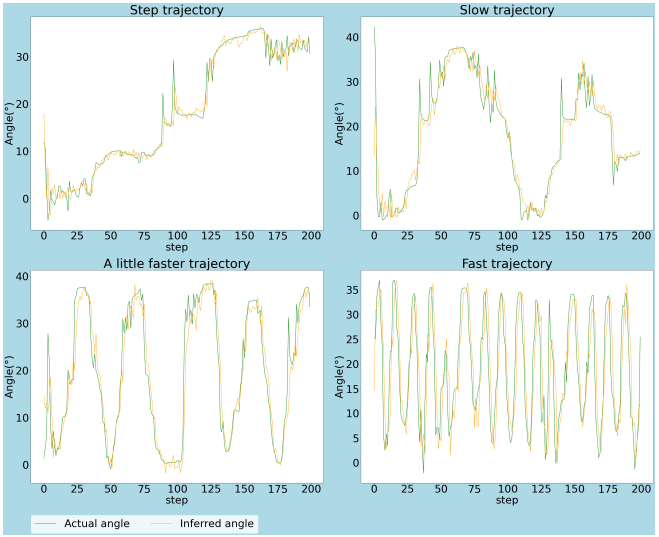


图7. 训练好的世界模型推断出的关节角度与实际关节角度。

再次佩戴细型麦克基本人工肌肉后，准确度与摘下前相同。

在本实验中，重新连接细型麦克基本人工肌肉数分钟后，对策略进行了微调。实际上，也存在需要在数天或数月后对策略进行微调的情况。在更长时间后进行微调是未来的工作。

C. 世界模型学习的结果

我们检查了世界模型是否能够被学习，这是策略训练中最重要的部分。仅使用循环模型、转移模型和执行器，通过世界模型推断关节角度。由于未使用表征模型，因此在没有实际关节角度的情况下完成了关节角度推断。参与者1在评估期间的推断关节角度和实际关节角度如图7所示。当实际关节角度减小时，推断关节角度也几乎同步减小。同样地，当实际关节角度增大时，推断关节角度也几乎同步增大。结果表明，世界模型能够大致推断由细型麦基宾执行器控制的人类手指的关节角度。世界模型的成功学习应有助于控制策略的成功学习。

VI. 讨论

A. 基于深度神经网络的控制器的优势

实验表明，Tracker 这一基于深度神经网络的控制器能够实现比 PID控制器更精确的控制。我们的基于深度神经网络的控制器比 PID控制器具有更多参数，因此能够更精确地学习控制指令。此外，我们的控制器不仅可以纳入误差，还可以纳入未来轨迹；因此，它能够确定前瞻性指令。与 PID控制器相比，我们的控制器更适用于各种轨迹，因为我们的控制器可以利用多条预采集轨迹进行学习，而 PID控制器的参数则是根据单条轨迹调节的。

B. 考虑人体肌肉状态以实现精确控制

在本研究中，我们未考虑人体肌肉的状态，而该状态应会影响控制指令。将人体肌肉的状态纳入观测可能会提高控制精度。基于深度神经网络的控制器可以输入肌肉状态，而PID控制器则无法做到这一点。基于深度神经网络的控制器具有整合多模态特征的优势，例如关节角度、角速度和人体肌肉状态。人体肌肉的状态可以通过电阻抗断层成像（EIT）获得 [30]。

C. 时间高效学习

我们的控制器采用基于模型的强化学习进行训练，这是一种数据高效学习方法。然而，学习控制策略耗时一小时。要为所有手指学习策略将需要更多时间，这会给用户带来负担。为了提高易用性，我们需要加快训练速度。通过在粗略模拟器中对预训练策略进行微调 [31]，或使用JAX等高性能数值计算库 [32], [33]，训练速度可能会得到提升。

VII. 结论

本研究提出了 Tracker，这是针对跟踪控制问题对基于模型的强化学习方法 DreamerV2 的扩展。实验结果表明，在两连杆机械臂的多种轨迹跟踪中，Tracker 相比 PID控制器实现了更精准、更通用的控制。此外，Tracker 还实现了对连接细型麦克基本人工肌肉的人类手指的控制。研究还表明，可以利用细型麦克基本人工肌肉学习世界模型。这一结果说明，基于世界模型的强化学习不仅可用于刚性机器人，也可用于软体机器人。

软体机器人在安全性方面适用于人机交互。这一特点与在真实人类和机器人上开展强化学习相辅相成。人手的动力学因人而异，因此我们需要针对每个人分别优化人机交互系统。解决这一问题的一种方法，是从真实人类与机器人的交互数据中学习其动力学。本研究为利用强化学习解决人与软体机器人交互问题提供了关键思路。

本研究未来的方向是将 Tracker 扩展至所有手指关节，并使用由 Tracker 控制的软体手外骨骼与物体进行交互，同时将 Tracker 应用于其他软体材料。

致谢

本研究得到了日本 JST CREST 项目（项目编号：JPMJCR17A3）的支持。

参考文献

[1] David Pinto-Fernandez、Diego Torricelli、Maria del Carmen Sanchez- Villamanan、Felix Aller、Katja Mombaur、Roberto Conti、Nicola Vi- tiello、Juan C Moreno 和 Jose Luis Pons。下肢外骨骼的性能评估：系统综述。*IEEE*神经系统与康复工程汇刊，28 1583，2020 7 1573( )：– 。[2] Muhammad Ahsan Gull、Shaoping Bai 和 Thomas Bak。上肢外骨骼设计综述。机器人学，9(1):16，2020。[3] Tiaan du Plessis、Karim Djouani 和 Christiaan Oosthuizen。用于康复和辅助的主动式手部外骨骼综述。机器人学，10(1):40，2021。[4] Koichi Suzumori。实现贾科梅蒂机器人学和软体机器人学的新型气动人工肌肉。载于制造、机器设计与摩擦学国际会议，编号15，第4–5页，2015年。[5] Shunichi Kurumaya、Hiroyuki Nabae、Gen Endo 和 Koichi Suzu- mori。细型麦基宾肌肉及多丝结构的设计。传感器与执行器*A*辑：物理，261:66–74，2017年。[6] Biyuan Wang、Nobuhiro Takahashi 和 Hideki Koike。采用人工肌肉组实现的用于手部康复的传感器手套。载于增强人类国际会议论文集，第1–4页，2020年。[7] Shoichiro Koizumi、Te-Hsin Chang、Hiroyuki Nabae、Gen Endo、Koichi Suzumori、Motoki Mita、Kimio Saitoh、Kazutoshi Hatakeyama、Satoaki Chida 和 Yoichi Shimada。用于手部辅助和康复的细型麦克基本人工肌肉软体机器人手套。载于*2020*年*IEEE/SICE*系统集成国际研讨会（*SII*），第93–98页。IEEE，2020年。[8] Nobuhiro Takahashi、Shinichi Furuya 和 Hideki Koike。具有人体解剖结构的软体外骨骼手套：灵巧手指运动和娴熟钢琴演奏的实现。*IEEE*触觉学汇刊，13(4):679–690，2020年。[9] Ahmad Athif Mohd Faudzi、Noor Hanis Izzuddin Mat Lazim 和 Koichi Suzumori。薄型软体麦基宾执行器的建模与力控制。自动化技术国际期刊，10(4):487–493，2016年。[10] Muhamad Hazwan Abdul Hafidz、Nor Mohd Haziq Norsahperi、Hong Win Soon、Mohd Najeb Jamaludin 和 Ahmad Athif Mohd Faudzi。拮抗对配置中细型麦基宾肌肉的控制。载于机器学习中的计算智能：*ICCIML 2021*精选论文集，第155–163页。Springer，2022年。[11] Nobutsuna Endo、Yuta Kizaki 和 Norihiro Kamamichi。用于机器人舌的柔性气动弯曲执行器。机器人学与机电一体化期刊，32(5):894–902，2020年。[12] Hiroki Yamaguchi、Yuki Funabora、Shinji Doki 和 Kae Doki。织物执行器的三维变形控制系统。载于*2020*年*IEEE/SICE*系统集成国际研讨会（*SII*），第253–258页。IEEE，2020年。[13] Kai Arulkumaran、Marc Peter Deisenroth、Miles Brundage 和 Anil Anthony Bharath。深度强化学习：简要综述。*IEEE*信号处理杂志，34(6):26–38，2017年。

[14] Jianyin Fan、Jing Jin 和 Qiang Wang。基于强化学习的仿人肌肉骨骼机械臂设计与控制。载于2020年第15届IEEE工业电子与应用会议（ICIEA），第541–546页。IEEE，2020年。[15] David Ha 和 J¨urgen Schmidhuber。世界模型。arXiv预印本 arXiv:1803.10122，2018年。[16] Danijar Hafner、Timothy Lillicrap、Mohammad Norouzi 和 Jimmy Ba。利用离散世界模型掌握Atari。arXiv预印本 arXiv:2010.02193，2020年。[17] Philipp Wu、Alejandro Escontrela、Danijar Hafner、Ken Goldberg 和 Pieter Abbeel。Daydreamer：用于实体机器人学习的世界模型。arXiv预印本 arXiv:2206.14176，2022年。[18] Dieter B¨uchler、Heiko Ott 和 Jan Peters。采用气动肌肉的轻量级机械臂用于机器人学习。载于2016年IEEE国际机器人与自动化会议（ICRA），第4086–4092页。IEEE，2016年。[19] Ryuji Suzuki、Manabu Okui、Shingo Iikawa、Yasuyuki Yamada 和 Taro Nakamura。基于实验辨识模型的直纤维型人工肌肉新型前馈控制器。载于2018年IEEE国际软体机器人会议（RoboSoft），第31–38页。IEEE，2018年。[20] Lamia Melkou 和 Mustapha Hamerlain。气动人工肌肉机械臂的高阶齐次滑模控制。载于IECON 2019 IEEE工业电子学会第45届年会，第1卷，第394–399页。IEEE，2019年。[21] Kento Ka waharazuka、Toru Ogawa 和 Cota Nabeshima。使用深度循环神经网络的柔性机械臂动态任务控制方法。载于2019年IEEE/RSJ智能机器人与系统国际会议（IROS），第7695–7701页，2019年。[22] Wenshuai Zhao、Jorge Pe¯na Queralta 和 Tomi Westerlund。机器人学深度强化学习中的仿真到现实迁移：综述。载于2020年IEEE计算智能研讨会系列（SSCI），第737–744页。IEEE，2020年。[23] Danijar Hafner、Timothy Lillicrap、Jimmy Ba 和 Mohammad Norouzi。梦境控制：通过潜在想象学习行为。载于国际学习表征会议，2020年。[24] Danijar Hafner、Timothy Lillicrap、Ian Fischer、Ruben Villegas、David Ha、Honglak Lee 和 James Davidson。通过像素学习用于规划的潜在动力学。载于国际机器学习会议，第2555–2565页。PMLR，2019年。[25] Kyunghyun Cho、Bart Van Merri¨enboer、Caglar Gulcehre、Dzmitry Bahdanau、Fethi Bougares、Holger Schwenk 和 Yoshua Bengio。使用RNN编码器-解码器学习统计机器翻译的短语表示。arXiv预印本 arXiv:1406.1078，2014年。[26] Djork-Arn´e Clevert、Thomas Unterthiner 和 Sepp Hochreiter。通过指数线性单元（ELU）实现快速且准确的深度网络学习。arXiv预印本 arXiv:1511.07289，2015年。[27] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。arXiv预印本 arXiv:1412.6980，2014年。[28] Yunzhi Zhang、Ignasi Clavera、Boren Tsai 和 Pieter Abbeel。基于模型的强化学习的异步方法。载于机器人学习会议，第1338–1347页。PMLR，2020年。[29] Marcin Andrychowicz、Anton Raichuk、Piotr Sta´nczyk、Manu Orsini、Sertan Girgin、Raphael Marinier、L´eonard Hussenot、Matthieu Geist、Olivier Pietquin、Marcin Michalski 等。什么对同策略强化学习最重要？一项大规模实证研究。arXiv预印本 arXiv:2006.05990，2020年。[30] Junyi Zhu、Yuxuan Lei、Aashini Shah、Gila Schein、Hamid Ghaednia、Joseph Schwab、Casper Harteveld 和 Stefanie Mueller。Musclerehab：通过监测和可视化肌肉参与度改进无监督身体康复。载于第35届ACM用户界面软件与技术年会论文集，UIST '22，美国纽约州纽约市，2022年。计算机协会。[31] Joanne Truong、Max Rudolph、Naoki Yokoyama、Sonia Chernova、Dhruv Batra 和 Akshara Rai。重新思考仿真到现实：低保真度仿真带来更高的仿真到现实导航迁移效果。arXiv预印本 arXiv:2207.10821，2022年。[32] James Bradbury、Roy Frostig、Peter Hawkins、Matthew James John- son、Chris Leary、Dougal Maclaurin、George Necula、Adam Paszke、Jake VanderPlas、Skye Wanderman-Milne 和 Qiao Zhang。JAX：可组合的Python+NumPy程序变换，2018年。[33] Laura Smith、Ilya Kostrikov 和 Sergey Levine。漫步公园：使用无模型强化学习在20分钟内学会行走。arXiv预印本 arXiv:2208.07860，2022年。