Lecture 4

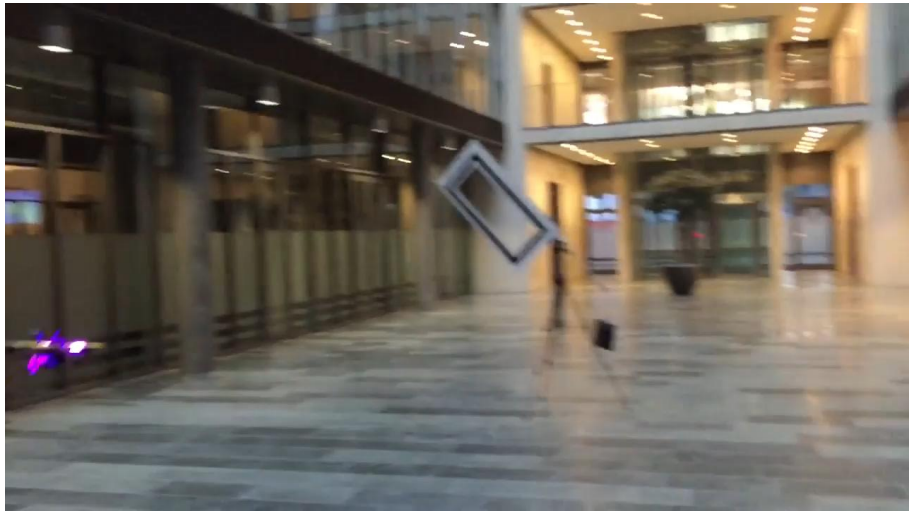
RL在无人机中的应用

高飞

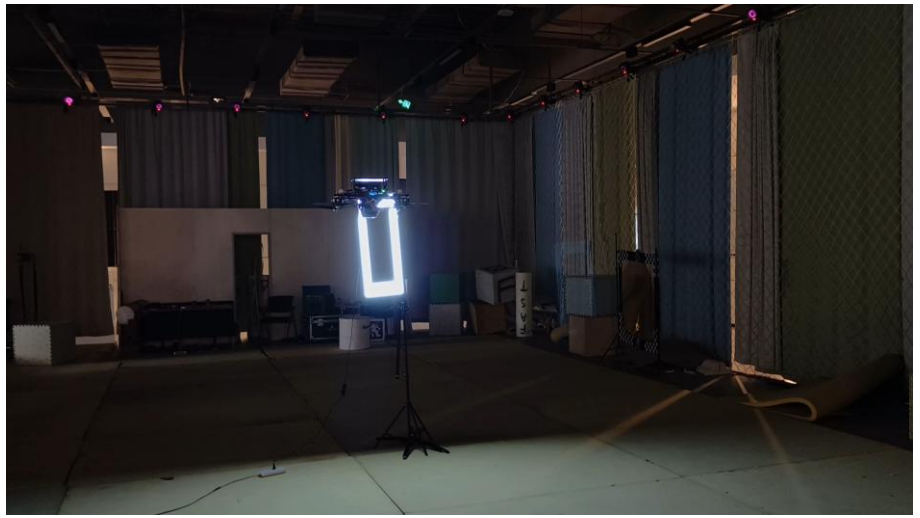
浙江大学 控制学院

强化学习应用：大机动飞行

熟练的人类飞手难以在不尝试的情况下成功

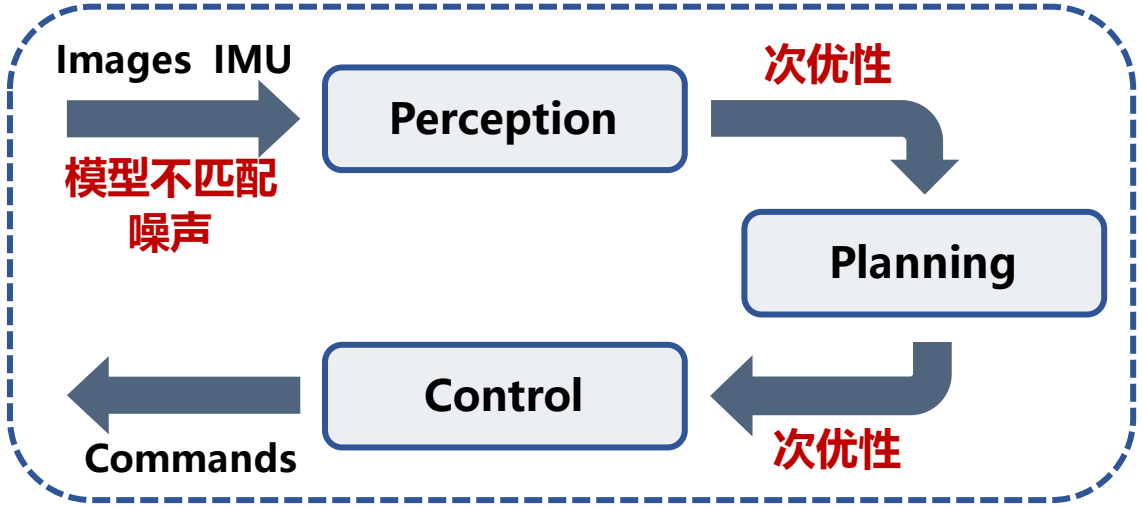


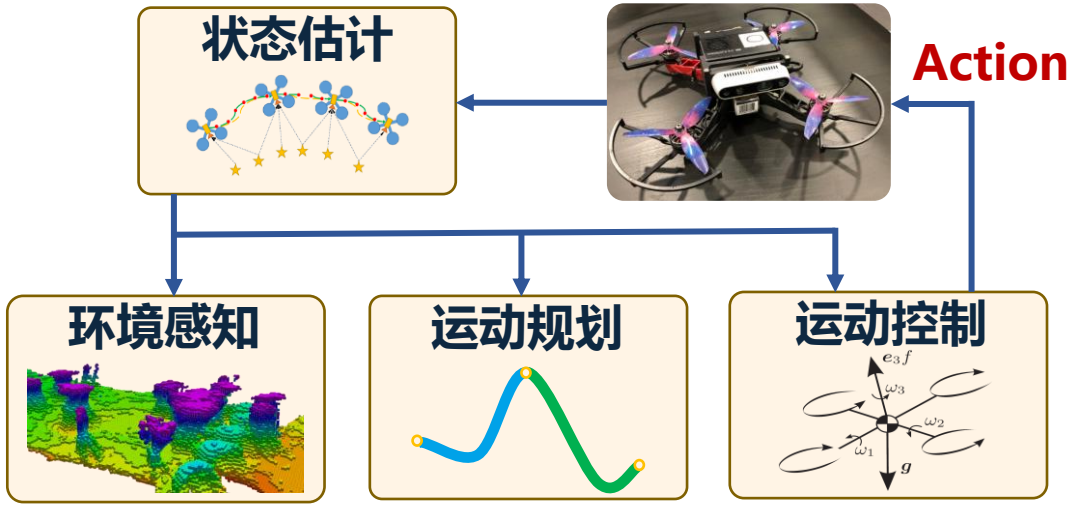
端到端方法拥有超过99%的成功率



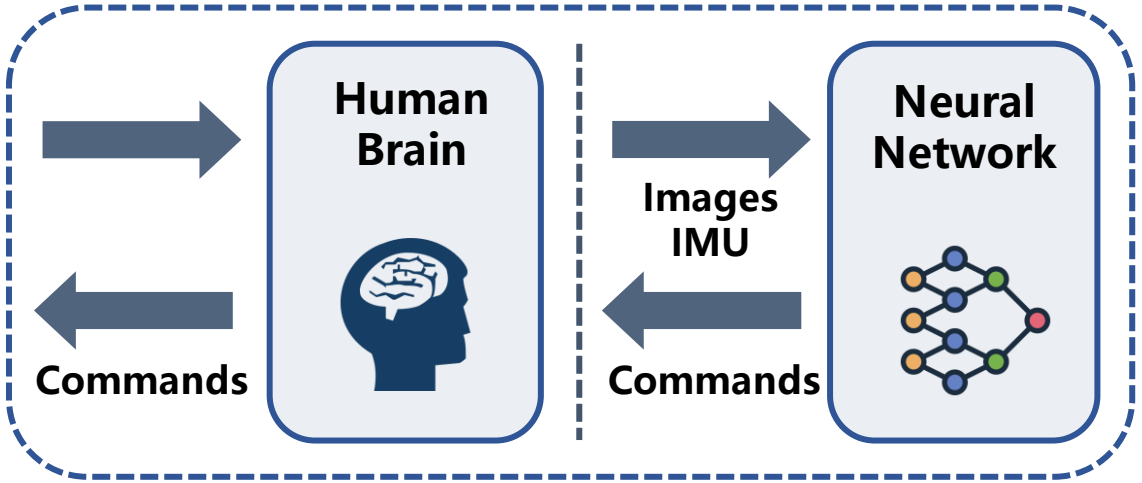
端到端方法与传统方法对比

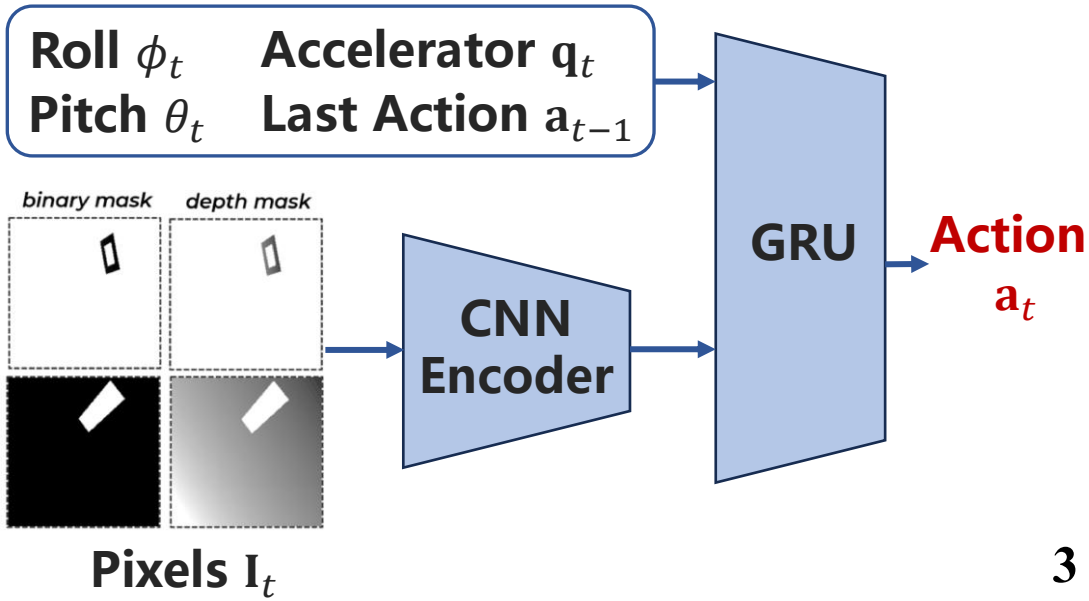
传统流程





端到端流程*

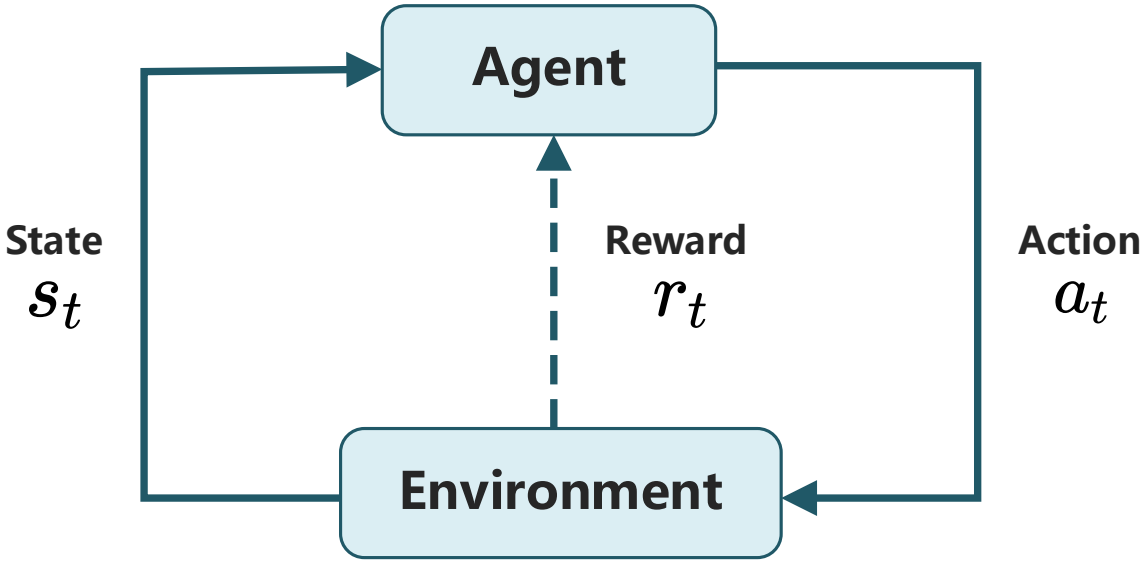




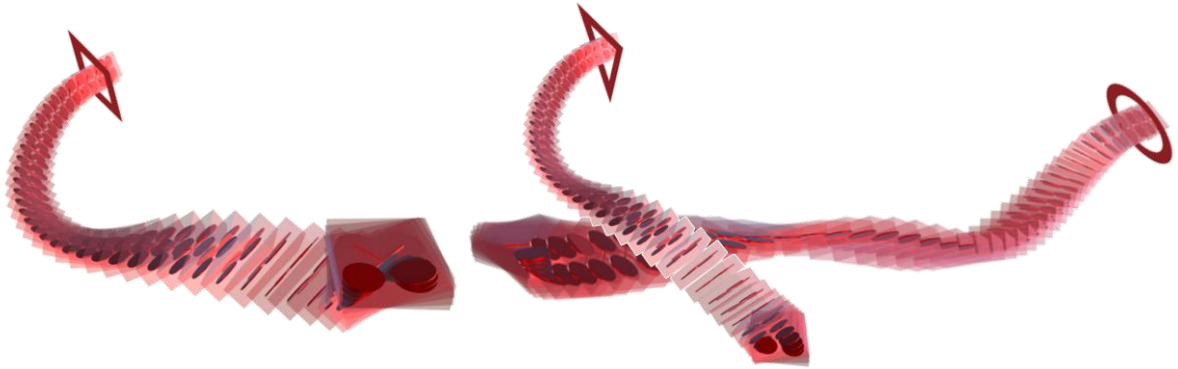
* 以P2中端到端大机动飞行为例

强化学习原理

强化学习（Reinforcement Learning, RL）不从一个固定的数据集中学习，
而是从与环境的主动**交互**中学习一个能最大化交互过程中所获**奖励**的最优**策略**。



$$\pi^\star = \max_{\pi} J(\pi)$$ 在策略 π 下总收益的期望



要素	在大机动飞行中对应
Agent	无人机及其**神经网络构成的策略 π**
State	对框和自身状态的观测
Reward	避障奖励，平滑运动奖励，…
Action	无人机总推力与角速度（body rate）

策略如何产生动作*

在使用 θ 参数化的**神经网络策略** π_θ 下，动作 a 被建模为**多元正态分布**，给定状态 s，

$$a \sim \mathcal{N}(\boldsymbol{\mu}_\theta(s), \boldsymbol{\sigma}_\theta(s)), \text{ where } a = (T, p, q, r) \in \mathbb{R}^4$$

总推力 thrust

角速度 body rate

$\boldsymbol{\mu}_\theta$ 由神经网络前向传播状态 s 获得，$\boldsymbol{\sigma}_\theta$是全局与状态 s 相独立的可被学习的参数，它们一同构成了策略 π_θ ，

$$\pi_\theta(a|s) = p_\theta(a_t = a|s_t = s)$$

agent的动作 a 通过**采样**获得。

* 本节案例为随机性的策略，对应的，还存在确定性策略。在策略中引入随机性有许多好处，例如能够更好地探索环境。

策略梯度的整体思路

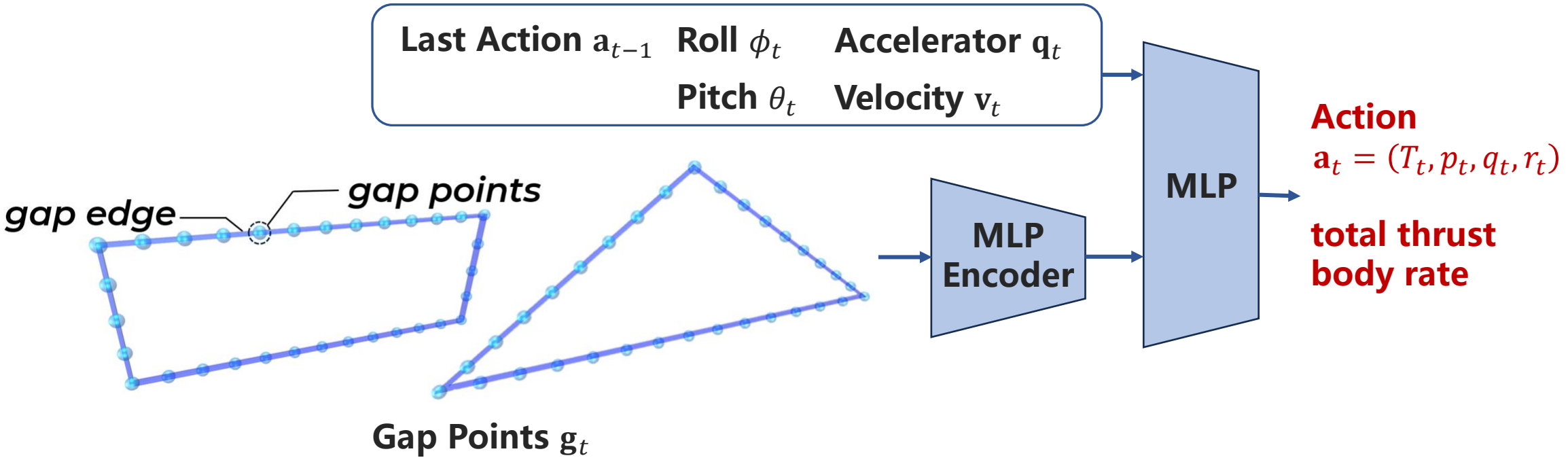
构造优化问题：

$$\max_{\theta} J(\theta)$$ ① 如何定义?

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\theta_t)$$ ② 如何求解?

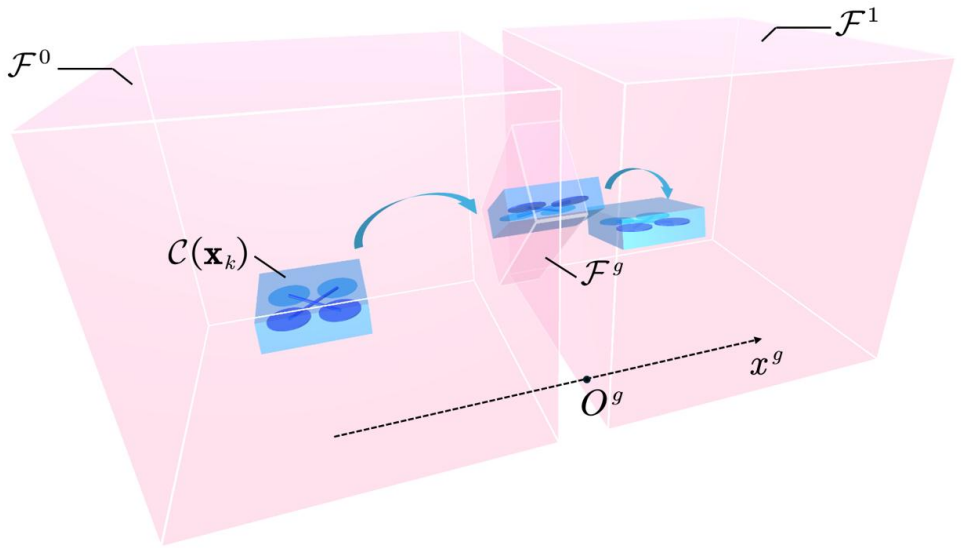
强化学习案例：大机动飞行

定义观测空间与动作空间



设计穿缝问题的奖励函数

① 正在穿缝时的奖励 traversing reward



x^g 是垂直于间隙平面的轴

x_t^g 表示 t 时刻飞机在 x^g 上的坐标

飞机质心　间隙中心　　飞机凸包属于间隙空间

$$\mathbb{I}\left[|x_t^g| \leq l^{\mathcal{C}} \text{ and } \|\mathbf{p}_t - \mathbf{p}^g\| \leq d \text{ and } \mathcal{C}(\mathbf{x}_k) \in \mathcal{F}\right] \cdot (x_t^g - x_{t-1}^g), \tag{2}$$

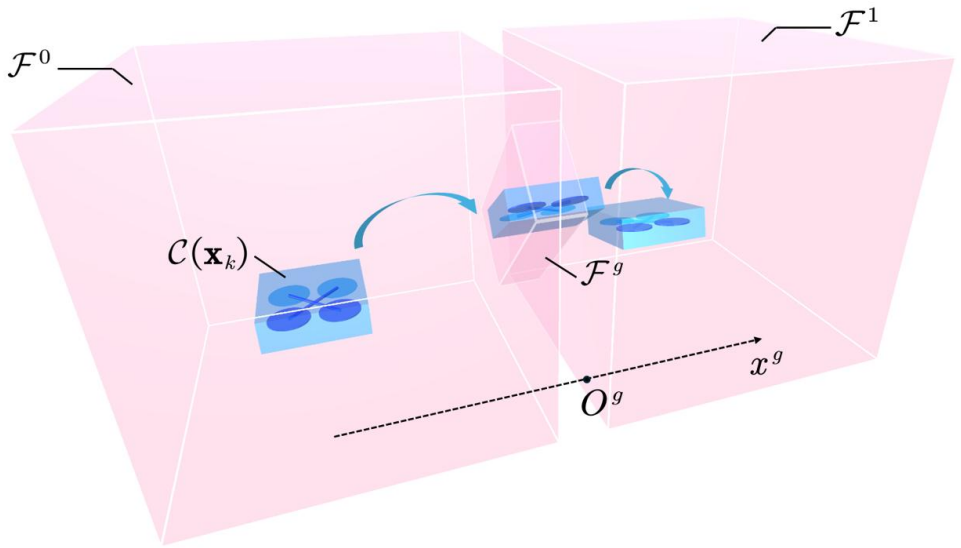
$l^{\mathcal{C}}, d$　阈值

$\mathbb{I}[\cdot]$　指示函数(indicator function)
满足条件时为1，不满足为0

强化学习案例：大机动飞行

设计穿缝问题的奖励函数

② 鼓励穿缝奖励 shaping reward



x^g 是垂直于间隙平面的轴

x_t^g 表示 t 时刻飞机在 x^g 上的坐标

飞机质心　间隙中心

$$\mathbb{I}\left[|x_t^g| > l^{\mathcal{C}}\right] \cdot \left(\|\mathbf{p}_{t-1} - \mathbf{p}^g\| - \|\mathbf{p}_t - \mathbf{p}^g\|\right). \qquad (3)$$

上一时刻距离　　此时刻距离

$l^{\mathcal{C}}$　阈值

$\mathbb{I}[\cdot]$　指示函数(indicator function)
满足条件时为1，不满足为0

强化学习案例：大机动飞行

设计穿缝问题的奖励函数

③ 不稳定动作惩罚 jerky motion penalties

$$-(\lambda_{\text{mag}} \cdot \|\mathbf{a}_t\| + \lambda_{\text{var}} \cdot \|\mathbf{a}_t - \mathbf{a}_{t-1}\|), \qquad (4)$$

加速度

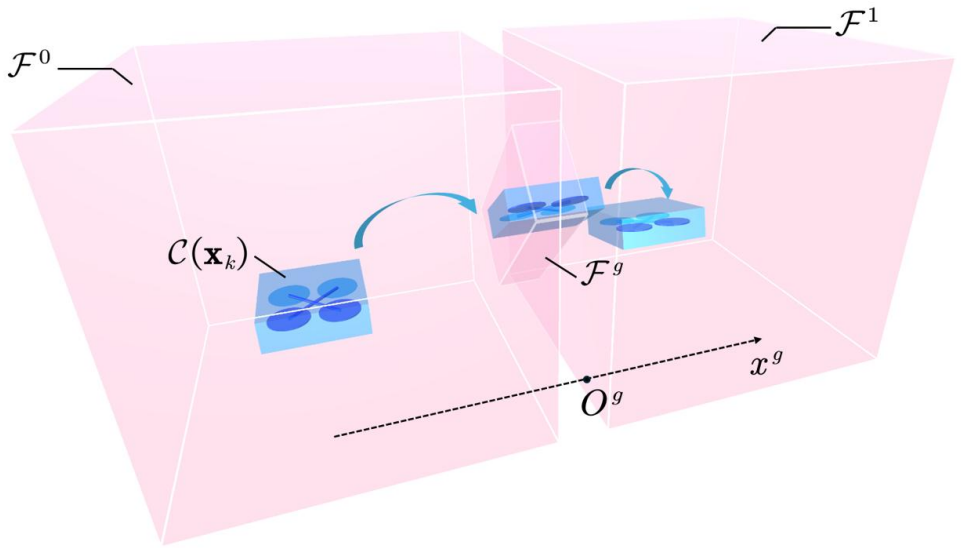
④ 最大速度约束 aggressiveness constraint

$$\mathbb{I}\left[\|\mathbf{v}_t\| \leqslant \mathrm{v}_{\max}\right] \cdot \left(-\exp\left(\|\mathbf{v}_t\| - \mathrm{v}_{\max}\right) + 1\right). \qquad (5)$$

速度

强化学习案例：大机动飞行

设计穿缝问题的终止条件



满足任意一个以下条件：

① $\mathcal{C}(x_t) \in \mathcal{F}^1$ 已完成穿缝

② $\mathcal{C}(x_t) \notin \mathcal{F}$ 飞机离开定义的世界范围

③ 飞机与间隙碰撞

④ 达到最大步长

强化学习案例：大机动飞行

问题挑战：

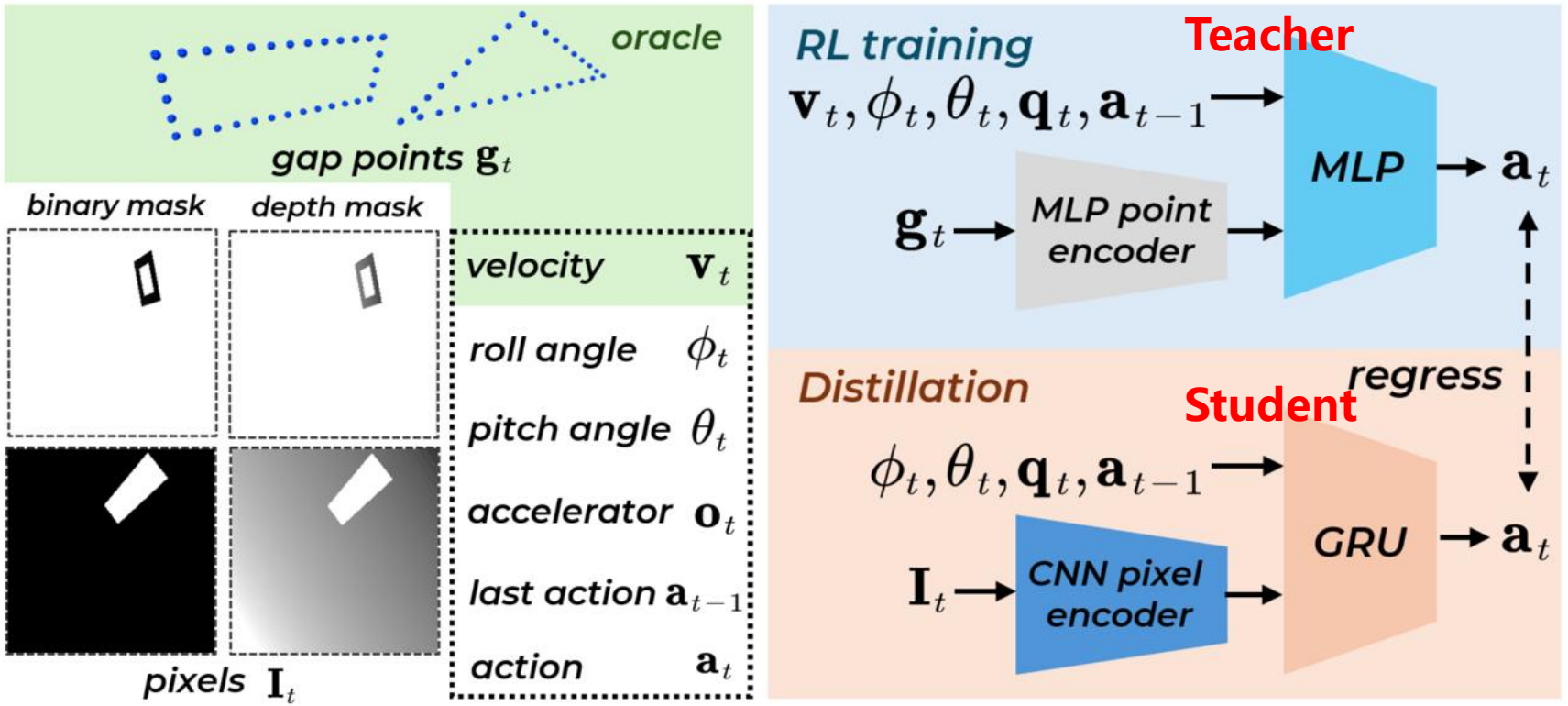
1.Q. 如何处理具备大量冗余信息的高维输入，例如图像，点云?
   A. **Teach-Student，数据提纯**

2.Q. 如何处理高速飞行时的Sim-Real Gap?例如风阻，时延?
   A. **Real-To-Sim, Domain Randomization**

3.Q. 当Good Policy在空间分布占比少时，如何保证训练稳定，有效收敛到Good Policy?
   A. 课程学习：**Curriculum for Reinforcement Learning**

强化学习技巧： Teach-Student

在线蒸馏



在使用图像的导航中，直接同时学习感知和规划很难。

① **Teacher**：在知道真实环境下(例如直接输入框的角点)训练得到一个Teacher策略。

② **Student**：在线地直接把Teacher的输出和Student的输出作监督回归，即尽可能使得Student的输出和Teacher一样。

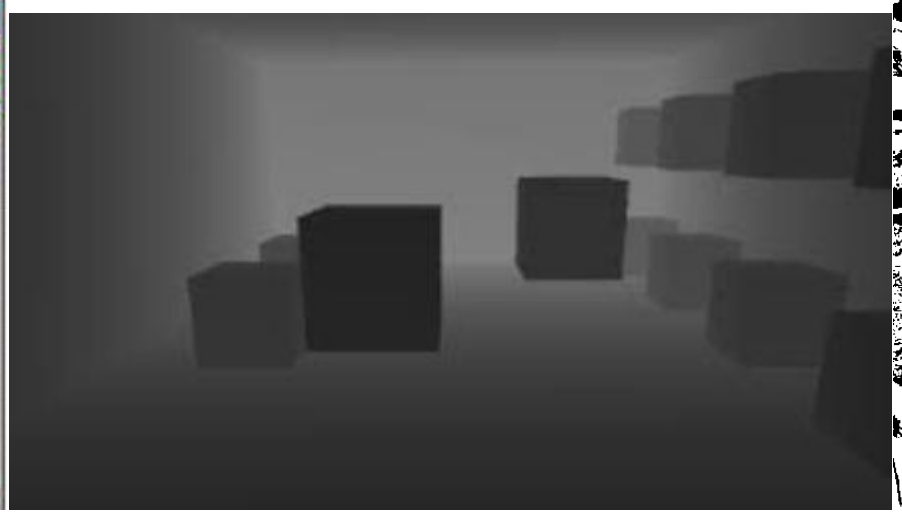
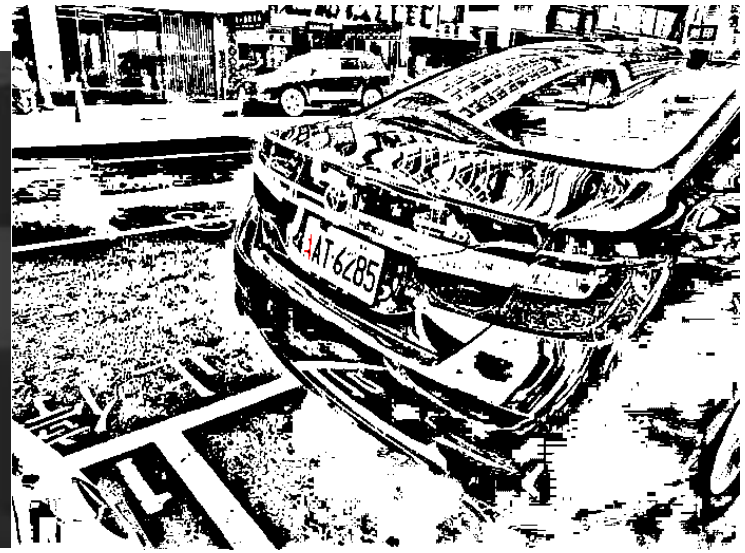
好处：

1.有效减少探索空间，样本效率大幅度提升。

2.训练更加稳定。

强化学习技巧： 数据提纯

相较于图像这样复杂的输入，使用Feature Tracker / 深度图/0-1二值化图等更加**abstract & compressed**的输入，更有利于sim-to-real和稳定训练。

Kaufmann, Elia, et al. "Deep Drone Acrobatics." (2020).
Mueller, Matthias, et al. "Driving Policy Transfer via Modularity and Abstraction." *Conference on Robot Learning*. PMLR, 2018.
Loquercio, Antonio, et al. "Learning high-speed flight in the wild." *Science Robotics* 6.59 (2021): eabg5810.

强化学习技巧： 数据提纯

点云降采样

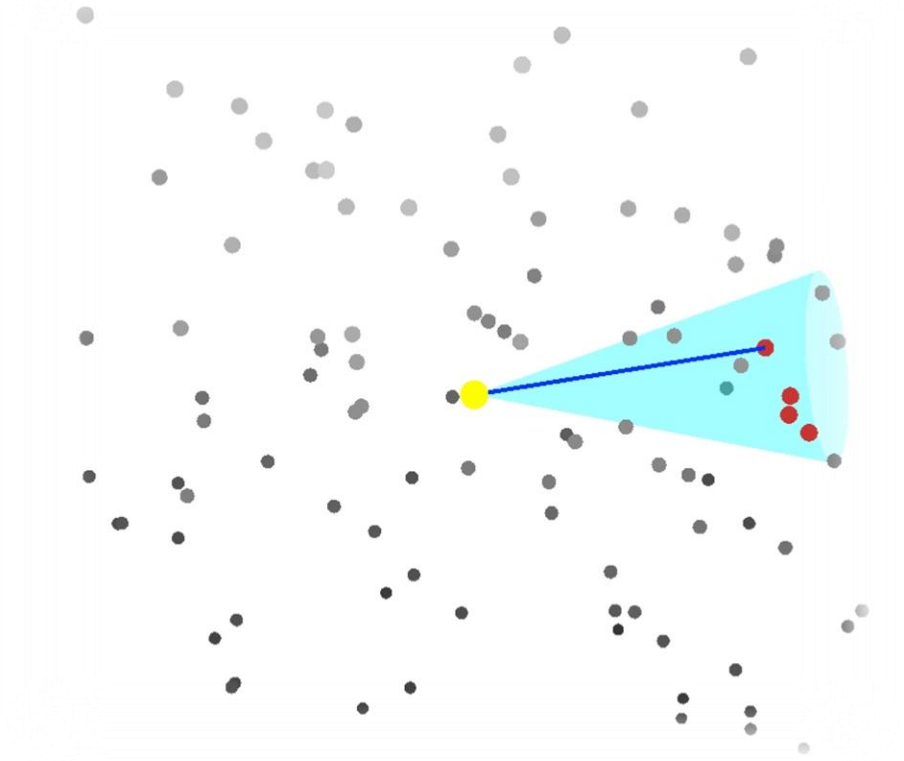
以机器人为中心，将FoV划分为3200个等角度的分区，对应4.5°的角度分辨率。若在某个分区内检测到至少一个点云，则该分区的值为离机器人最近点云的距离，实现点云的降采样表征 。

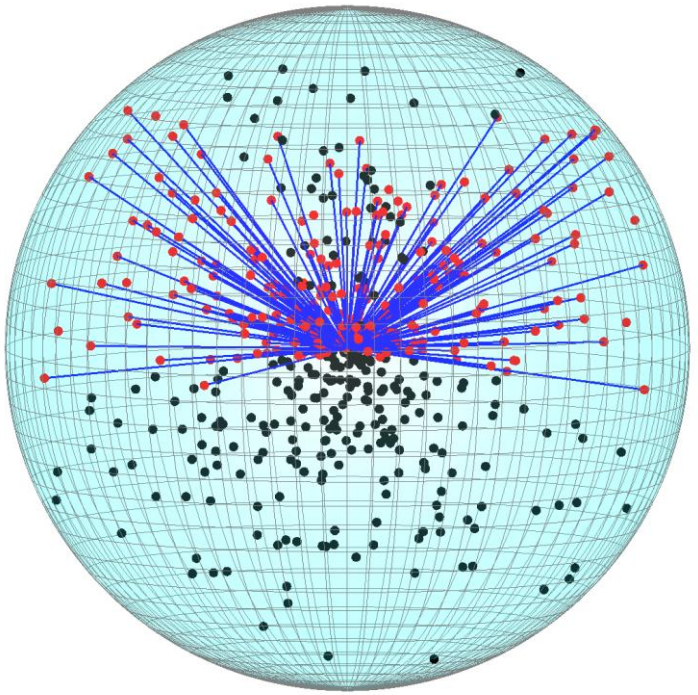
无人机

分区内点云

最近点距离

感知分区

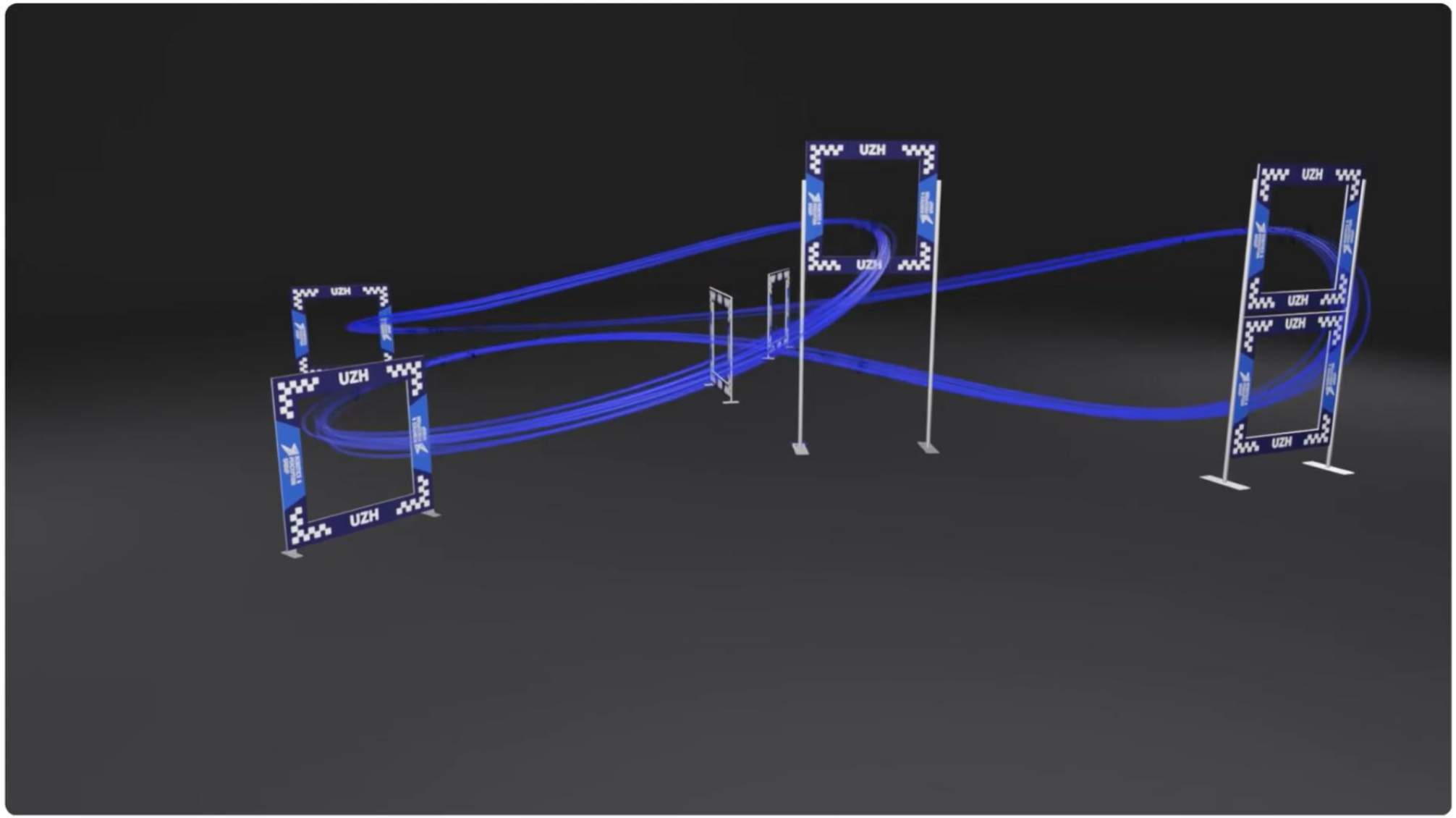




3200维距离输入

强化学习技巧： Real-To-Sim

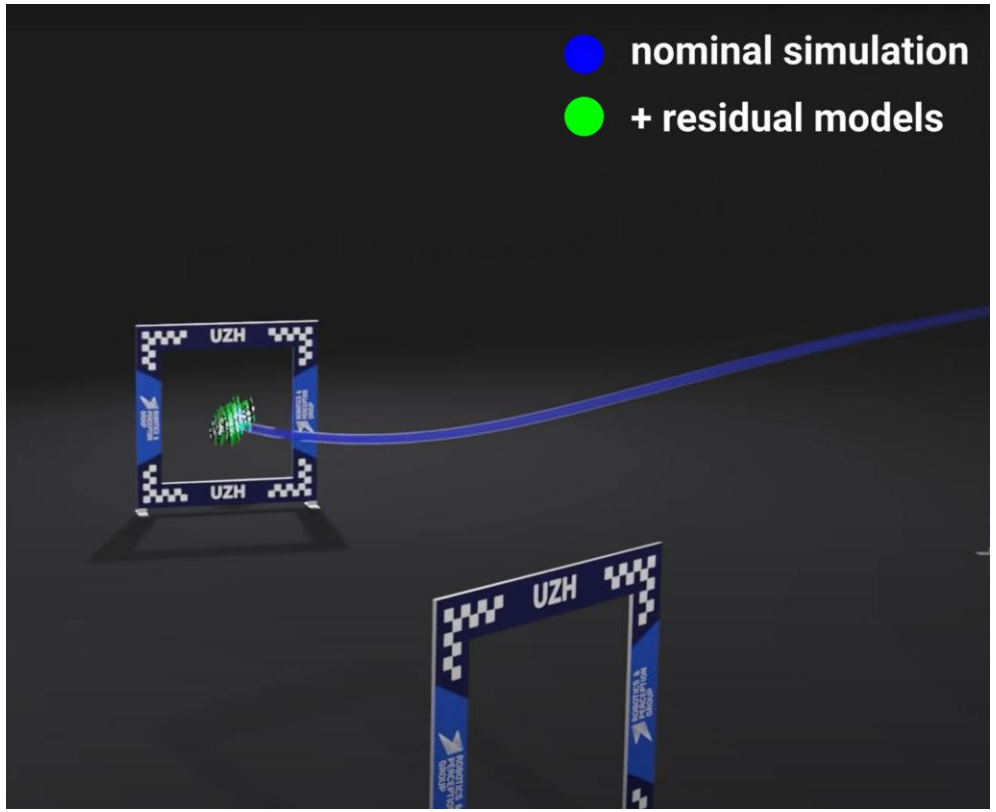
高速飞行时Sim-To-Real的Gap往往很大(风阻、VIO和框检测噪声大)

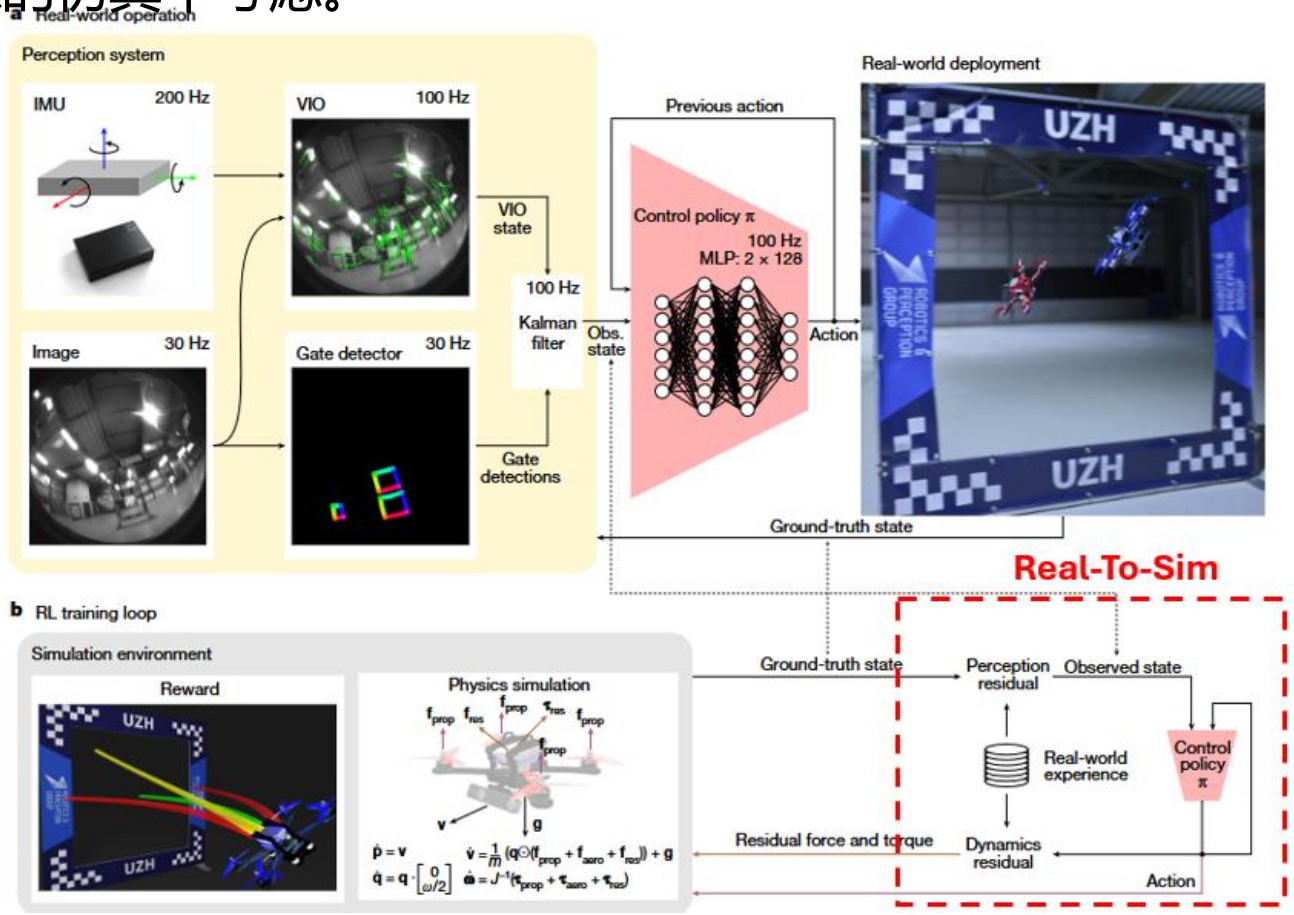


强化学习技巧： Real-To-Sim

Real-To-Sim:

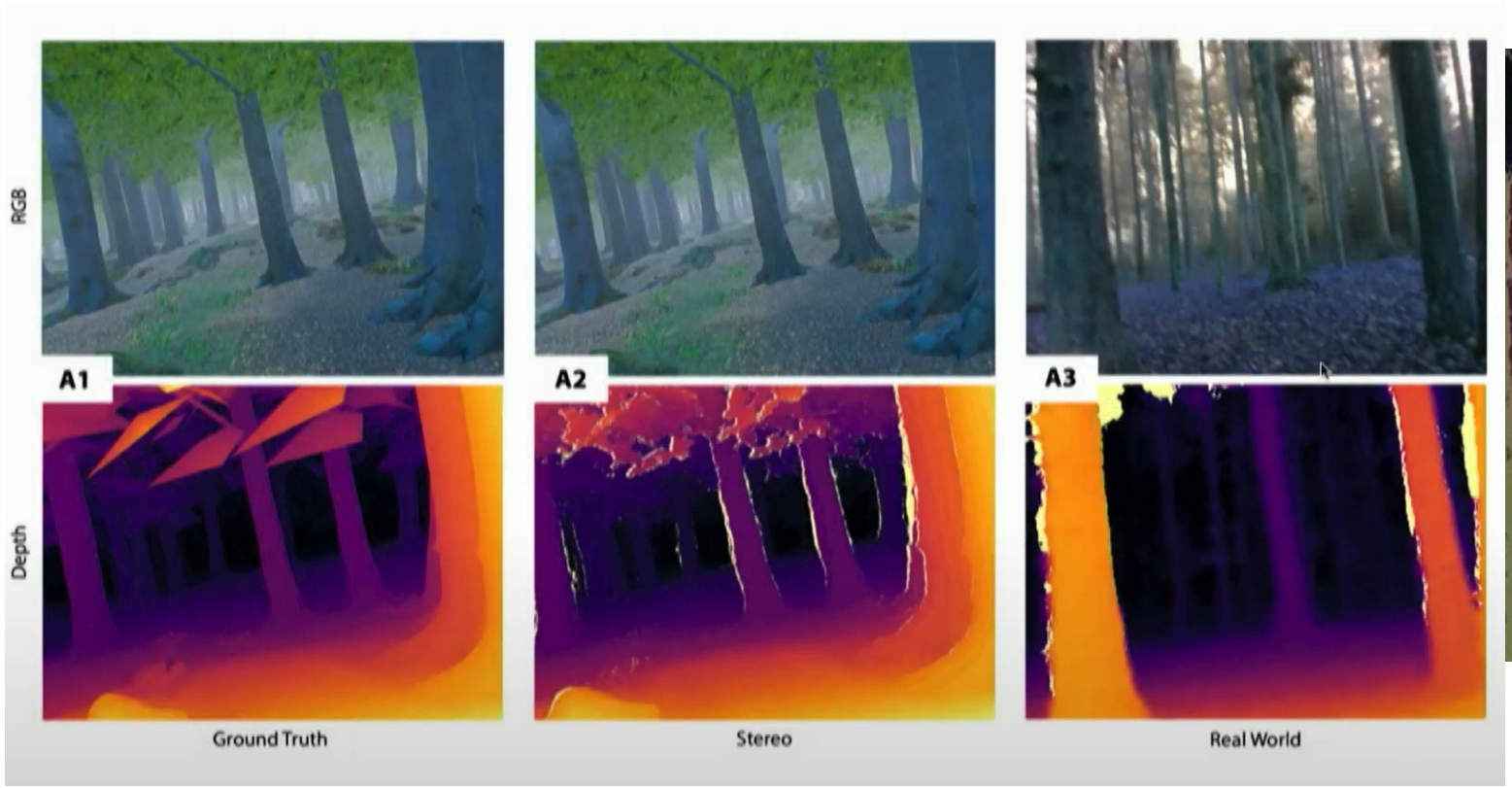
通过人类专业飞手在赛道中反复飞行采集图像和机器人状态数据，并利用这些数据拟合感知和动力学模型后将其在训练的仿真中考虑。

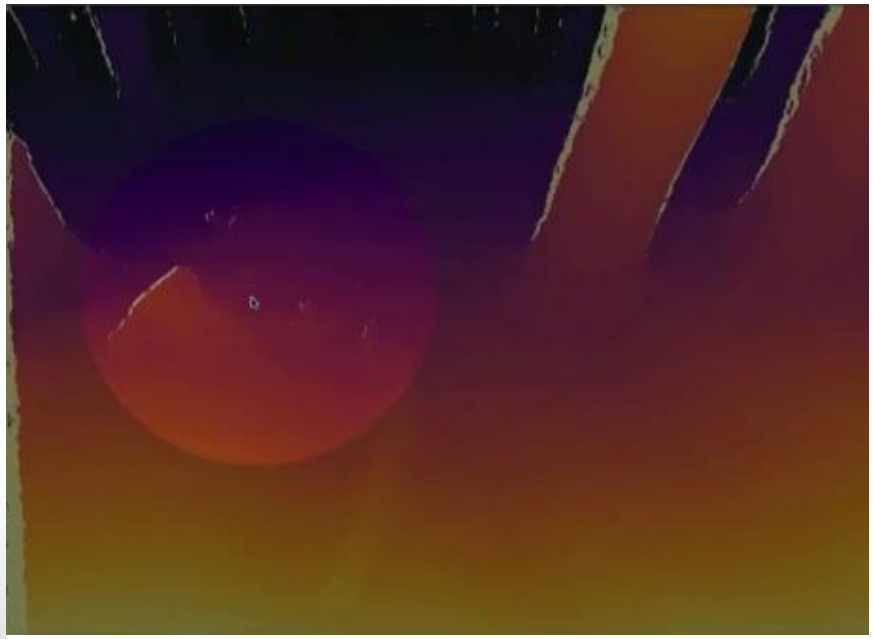




Kaufmann, Elia, et al. "Champion-level drone racing using deep reinforcement learning." *Nature* 620.7976 (2023): 982-987.

强化学习技巧： Domain Randomization

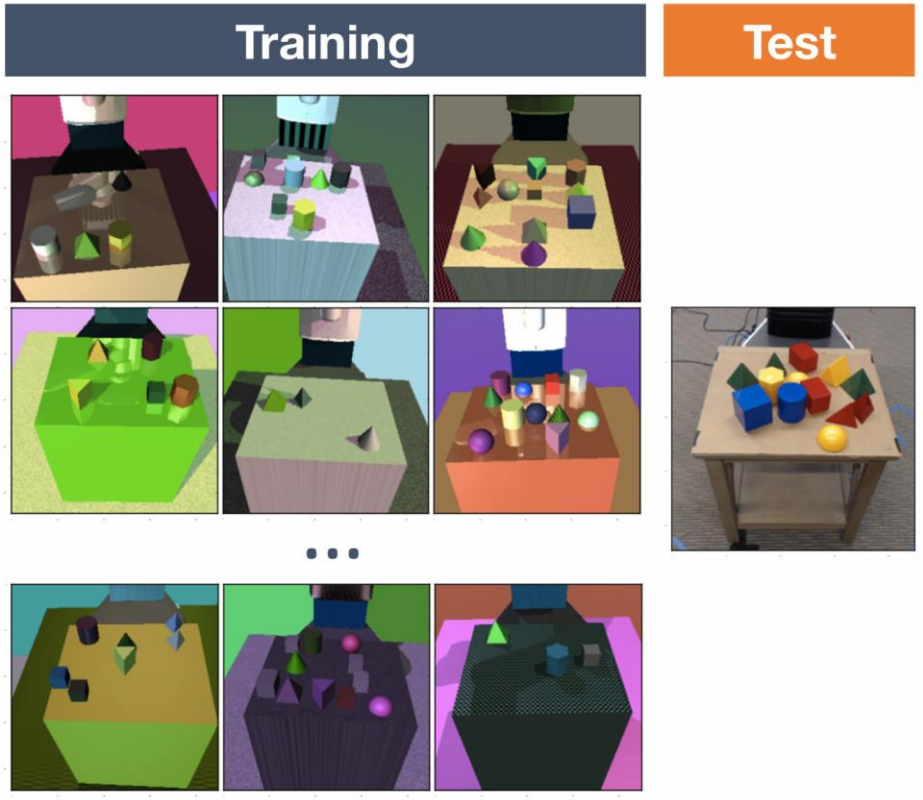




使用人为加入噪声的深度图训练

Loquercio, Antonio, et al. "Learning high-speed flight in the wild." *Science Robotics* 6.59 (2021): eabg5810.

强化学习技巧： Domain Randomization



"The purpose of domain randomization is to provide enough simulated variability at training time such that at test time the model is able to generalize to real-world data."

在抓取任务中随机化场景和条件：

（1）干扰物的形状和数量

（2）物体的位置和纹理

（3）场景的灯光

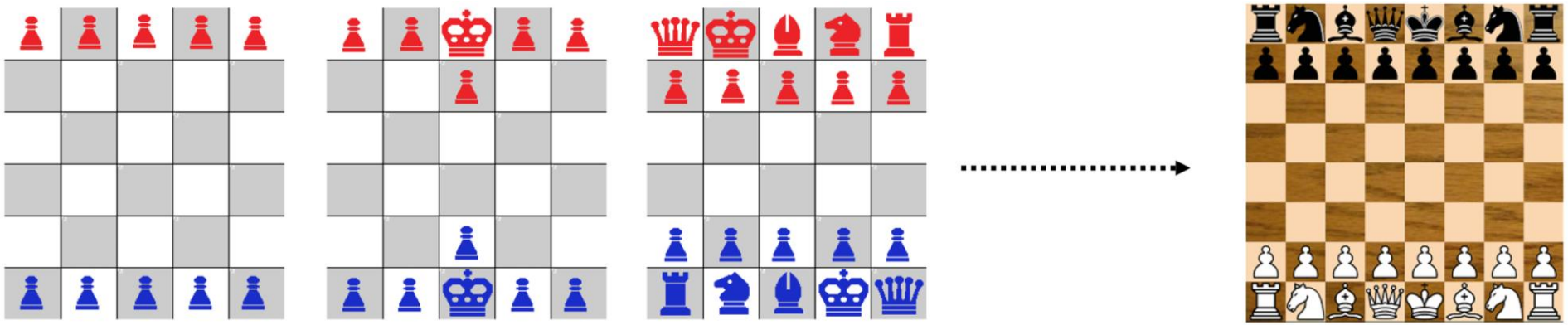
…… ……

Tobin, Josh, et al. "Domain randomization for transferring deep neural networks from simulation to the real world." 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2017.

Curriculum for Reinforcement Learning

课程学习（Curriculum Learning，CL）通过**提供一系列难度逐渐提升的学习任务**来将复杂的知识分解。课程学习能够加速模型收敛，甚至能够提升模型的最终性能。

课程学习的本质可以看作一种 continuation method 。这种方法首先优化比较smooth的问题，然后逐渐优化到不够smooth的问题。



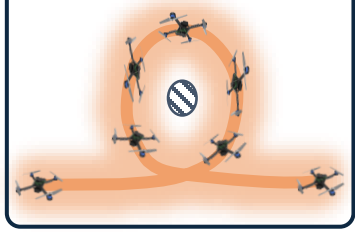
学习任务由简单到复杂

Bengio, Yoshua, et al. "Curriculum learning." *Proceedings of the 26th annual international conference on machine learning*. 2009.
Wang, Xin, Yudong Chen, and Wenwu Zhu. "A survey on curriculum learning." *IEEE transactions on pattern analysis and machine intelligence* 44.9 (2021): 4555-4576.
Narvekar, Sanmit, et al. "Curriculum learning for reinforcement learning domains: A framework and survey." *Journal of Machine Learning Research* 21.181 (2020): 1-50.

强化学习技巧：课程学习

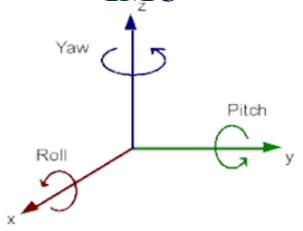
系统框架

特技意图

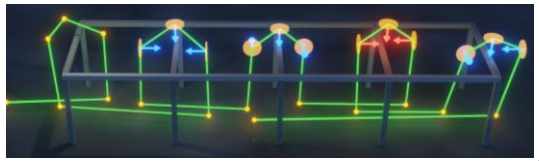


速度和姿态估计

IMU

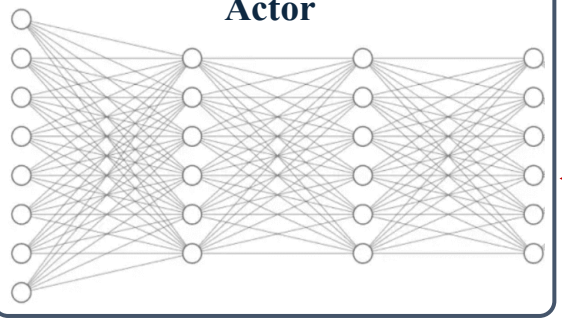


Reward反馈

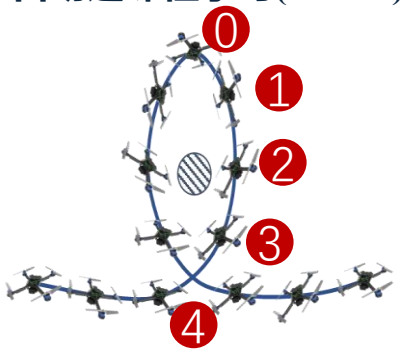


物理仿真

Actor



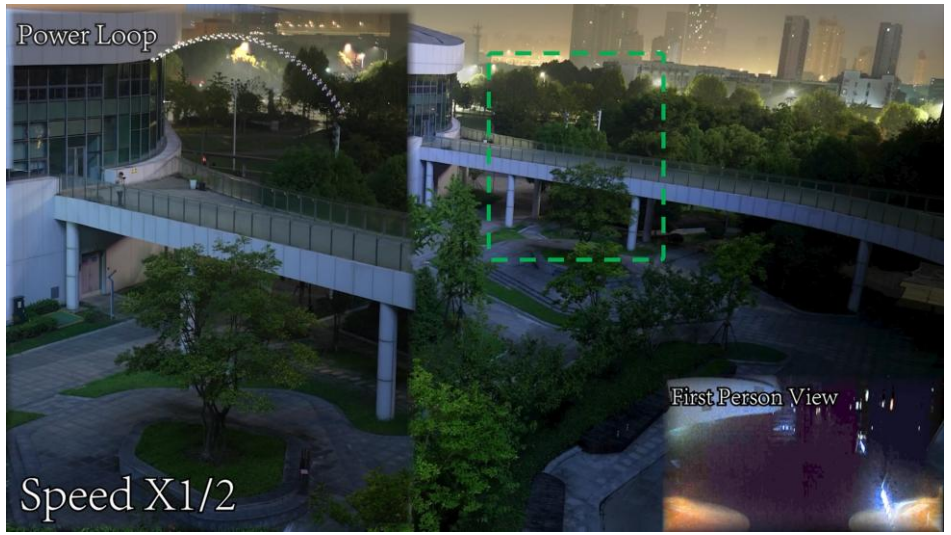
自动逆课程学习(ARCL)



从状态 0 开始学，再逐步拓展到 1 2 3

由易到难有利于网络的稳定训练

预期结果



Thanks for Listening!