1. 实验名称及目的

基于强化学习的最优控制实验：利用基于模型的强化学习方法，使用了近似策略迭代算法，将最优控制问题分解为两个阶段:近似策略评估和策略提升。在近似策略评估阶段，使用一个线性结构的逼近器来近似值函数，并利用系统模型和贝尔曼方程来更新逼近器的参数。在策略提升阶段，使用一个线性结构的逼近器来近似最优控制策略，并利用值函数和系统模型来更新逼近器的参数。这两个阶段交替进行，直到收敛到一个近似最优解。基于此，本实验首先基于扩张状态观测器观测飞行器模型的不确定性，并对其进行补偿，然后利用基于模型的强化学习最优控制方法近似被补偿系统的最优价值函数，进而确定最优控制律，接着基于控制屏障函数对最优控制律设计安全反馈项，保证闭环系统安全集的正向不变性。

2. 实验原理

主要思想：根据最优控制理论，带有成本函数的系统的最优控制问题可以转化为求解以下的哈密顿-雅可比-贝尔曼（HJB）方程：

$$V^{*'}(x)[Ax+B\left(f_0(x)+b_0(x)u_0^*(x)\right)]$$
$$+Q(x)+u_0^{*\mathrm{T}}(x)Ru_0^*(x)=0,$$

其中$V^*\in\mathcal{C}(\mathbb{R}^n,\mathbb{R}_{\geq 0})$，$V^*(0)=0$，是最优值函数。最优控制策略可以从最优值函数中确定为

$$u_0^*(x)=-\frac{1}{2}R^{-1}b_0^{\mathrm{T}}(x)B^{\mathrm{T}}V_x^{*\mathrm{T}}(x)$$

通常，HJB 方程的解析解很难获得。然而，最优值函数$V^*(x)$和最优控制策略$u_0(x)$可以通过基于演员-评论家神经网络（网络）的方法近似得到。对于任意给定的紧凑集$\mathcal{X}\subset\mathbb{R}^n$和正常数$\bar{\kappa}$，最优值函数$V^*(x)$可以表示为

$$V^*(x)=W^{\mathrm{T}}\phi(x)+\kappa(x)$$

其中$\phi:\mathcal{X}\to\mathbb{R}^l$是一个连续可微的激活函数，$W\in\mathbb{R}^l$是理想的权重向量，$l\in\mathbb{N}$是神经元的数量，$\kappa:\mathbb{R}^n\to\mathbb{R}$是近似误差函数，满足$|\kappa(x)|\leq\bar{\kappa}$和$|\kappa_x(x)|\leq\bar{\kappa},\forall x\in\mathcal{R}$。从值函数的网络表示可以得出最优控制策略可以获得为

$$u_0^*(x)=-\frac{1}{2}R^{-1}b_0^{\mathrm{T}}(x)B^{\mathrm{T}}\left(\phi'^{\mathrm{T}}(x)W+\kappa_x^{\mathrm{T}}(x)\right)$$

因此，考虑到 ESO 提供了状态 x 的估计，最优值函数 $V^*(x)$ 和最优控制策略 $u_0(x)$ 的网络基础近似由以下公式给出

$$\hat{V}\left(\overline{x},\hat{W}_c\right)=\hat{W}_c^{\mathrm{T}}\phi(\overline{x})$$

$$\hat{u}_0\left(\overline{x},\hat{W}_a\right)=-\frac{1}{2}R^{-1}b_0^{\mathrm{T}}(\overline{x})B^{\mathrm{T}}\phi_x^{\mathrm{T}}(\overline{x})\hat{W}_a$$

其中 $\hat{W}_c,\hat{W}_a\in\mathbb{R}^l$ 是 W 的估计，分别为评论家和演员网络的权重。通过计算出近似的瞬时 Bellman 误差 $\delta_t:\mathbb{R}^n\times\mathbb{R}^l\times\mathbb{R}^l\to\mathbb{R}$，

$$\begin{aligned}\delta_t&\triangleq\delta\left(\overline{x},\hat{W}_c,\hat{W}_a\right)\\&\triangleq\hat{V}_x\left(\overline{x},\hat{W}_c\right)\left[A\overline{x}+B\left(f_0(\overline{x})+b_0(\overline{x})\hat{u}_0\left(\overline{x},\hat{W}_a\right)\right)\right]\\&\quad+Q(\overline{x})+\hat{u}_0^{\mathrm{T}}\left(\overline{x},\hat{W}_a\right)R\hat{u}_0\left(\overline{x},\hat{W}_a\right).\end{aligned}$$

在本文中，我们通过模拟经验实现了基于ESO的学习策略，使用估计状态 $\hat{x}$，$f_0(\cdot)$ 和 $b_0(\cdot)$ 的知识将 Bellman 误差推导到预定义的点集 $x^i\in\mathbb{R}^n,i=1,\ldots,N$。推导到点 x^i 的近似 Bellman 误差表示为

$$\delta_i\triangleq\delta\left(x^i,\hat{W}_c,\hat{W}_a\right)$$

然后，演员-评论家网络使用 Bellman 误差 δ_t 和 δ_i 更新估计值 $\hat{W}_c$ 和 $\hat{W}_a$。评论家网络的最小二乘更新法则表示为

$$\dot{\hat{W}}_c=-q_{c1}\Gamma\frac{\mu}{\rho}\delta_t-\frac{q_{c2}}{N}\Gamma\sum_{N}^{i=1}\frac{\mu_i}{\rho_i}\delta_i,$$

$$\dot{\Gamma}=\left(\beta\Gamma-q_{c1}\frac{\Gamma\mu\mu^{\mathrm{T}}\Gamma}{\rho^2}\right)\mathbf{1}_{\{\|\Gamma\|\le\varsigma_1,\|\Gamma(0)\|\le\varsigma_1\}},$$

其中 $\Gamma:\mathbb{R}\ge0\to\mathbb{R}^{l\times l}$ 是一个时间变化的最小二乘增益矩阵，$\varsigma_1>0$ 是饱和常数，$\lambda_{c1},\lambda_{c2}>0$ 是常数适应增益，$\beta>0$ 是常数遗忘因子，以及

$$\mu=\phi'(\overline{x})\left[A\overline{x}+B\left(f_0(\overline{x})+b_0(\overline{x})\hat{u}_0\left(\overline{x},\hat{W}_a\right)\right)\right],$$

$$\mu_i=\phi'\left(x^i\right)\left[Ax^i+B\left(f_0\left(x^i\right)+b_0\left(x^i\right)\hat{u}_0\left(x^i,\hat{W}_a\right)\right)\right],$$

$$\rho=1+\lambda\mu^{\mathrm{T}}\Gamma\mu,$$

$$\rho_i=1+\lambda\mu_i^{\mathrm{T}}\Gamma\mu_i,$$

其中 $\lambda>0$ 是常数归一化增益。该更新法则表示时间变化的增益矩阵 Γ 是有界的，边界条件如下：

$$\varsigma_0 I \le \Gamma(t) \le \varsigma_1 I, t \ge 0,$$

其中 I 是单位矩阵。根据后续的稳定性分析，演员网络的更新法则表示为

$$\dot{\hat{W}}_a = -q_{a1}\left(\hat{W}_a - \hat{W}_c\right) - q_{a2}\hat{W}_a + \frac{q_{c1}\psi_t^T \hat{W}_a \mu^T}{4\rho} + \frac{1}{N}\sum_{i=1}^{N}\frac{q_{c2}\psi_i^T \hat{W}_a \mu_i^T}{4\rho_i}$$

其中 $q_{a1}, q_{a2} > 0$ 是常数适应增益，

$$\psi_t \triangleq \phi'(\overline{x}) B b_0(\overline{x}) R^{-1} b_0^{\mathrm{T}}(\overline{x}) B^{\mathrm{T}} \phi'^{\mathrm{T}}(\overline{x}),$$
$$\psi_i \triangleq \phi'\left(x^i\right) B b_0\left(x^i\right) R^{-1} b_0^{\mathrm{T}}\left(x^i\right) B^{\mathrm{T}} \phi'^{\mathrm{T}}\left(x^i\right).$$

基于上述演员-评论家神经网络的策略，可以开发出一个自适应的最优控制策略来解决复杂的非线性系统的最优控制问题。通过实时更新评论家和演员神经网络的权重，可以实现对系统的实时控制，并进一步改善系统的性能。

基于 ESO 的输出，我们可以得到原闭环系统的最优控制策略为：

$$u^* = \hat{u}_0\left(\overline{x}, \hat{W}_a\right) - \frac{\overline{x}_{n+1}}{b_0(\overline{x})}.$$

3. 实验效果

分别设计了 SITL，HITL，实飞实验验证了强化学习的最优控制的效果。

4. 文件目录

文件夹/文件名称	说明
SITL	包含 SITL 所需的 slx 文件和参数文件
HITL	包含 HITL 和实飞所需的 slx 文件和参数文件

5. 运行环境

6. 运行环境

序号	软件要求	硬件要求	
		名称	数量
1	Windows 10 及以上版本	笔记本/台式电脑①	1
2	RflySim 平台完整版及以上版本	Pixhawk 6C 或 Pixhawk 6C mini②	1
		遥控器③	1
		遥控器接收器	1
		数据线、杜邦线等	若干

①：推荐配置请见：https://doc.rflysim.com

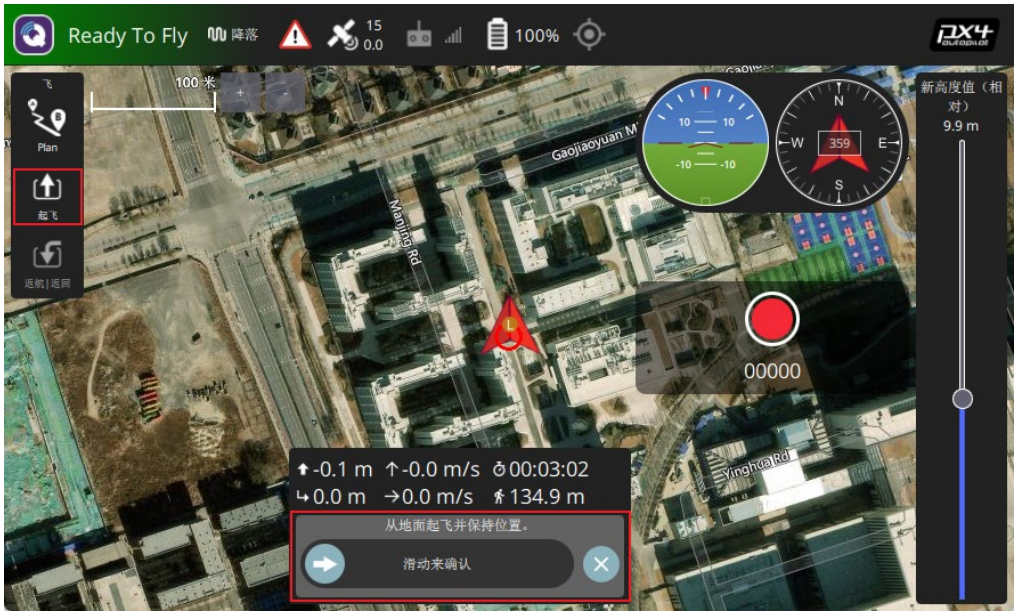
②：须保证平台安装时的编译命令为：px4_fmu-v6c_default，固件版本为：1.13.3。其他配套飞控请见：http://doc.rflysim.com

③：本实验演示所使用的遥控器为：天地飞ET10、配套接收器为：WFLY RF209S。遥控器相关配置见：..\e11_RC-Config\Readme.pdf

7. SITL 实验步骤

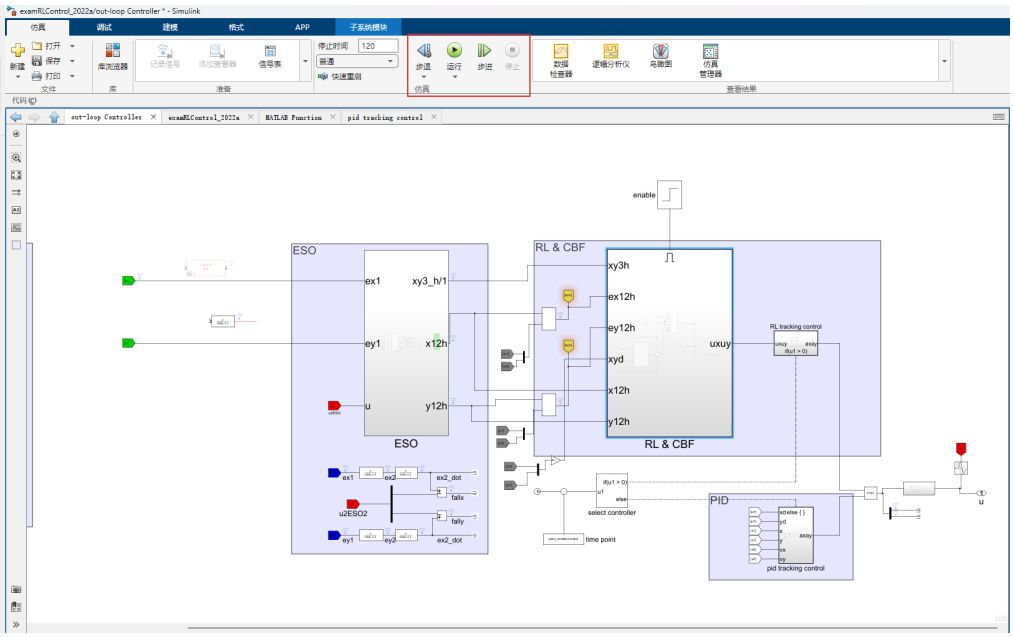
Step 1：

管理员运行"*\桌面\RflyTools\SITLRun.lnk"或"*\PX4PSP\RflySimAPIs\HITLRun.bat"文件，在弹出的 CMD 对话框中输入仿真数量：1，即可自动启动 RflySim3D、CopterSim、QGroundControl 软件，等待 CopterSim 的状态框中显示：PX4: GPS 3D fixed & EKF initialization finished。在 QGC 中点击起飞，滑动解锁。



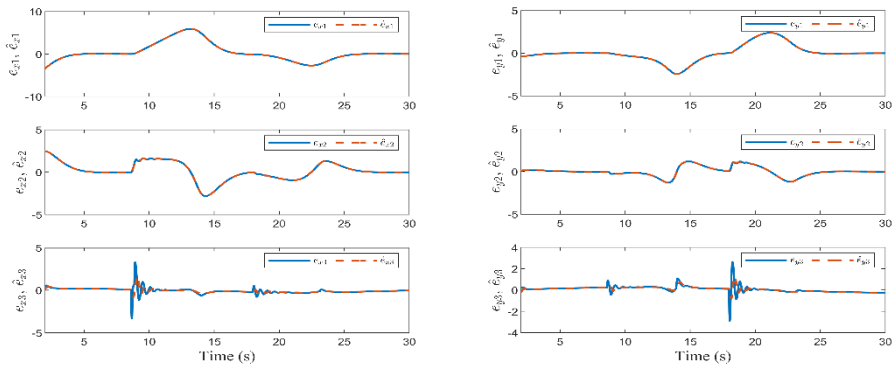
Step 2：

用 MATLAB 打开 SITL\main_para.m 运行，然后运行 SITL\RL_SITL.slx.slx 文件。

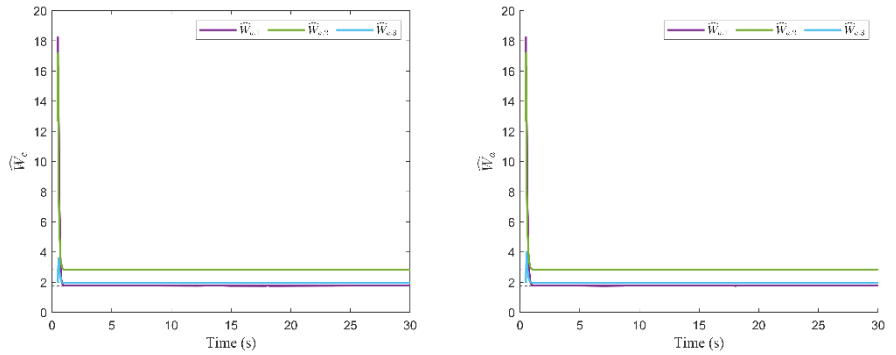


观察无人机跟踪圆形轨迹一段时间后，打开数据检查器。

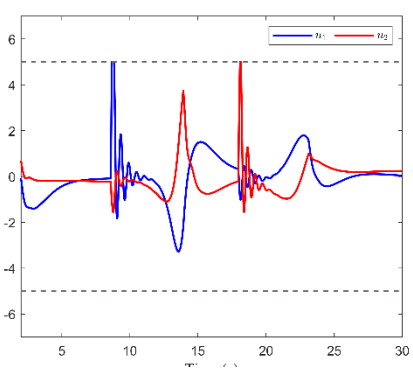
仿真结果如下图所示，



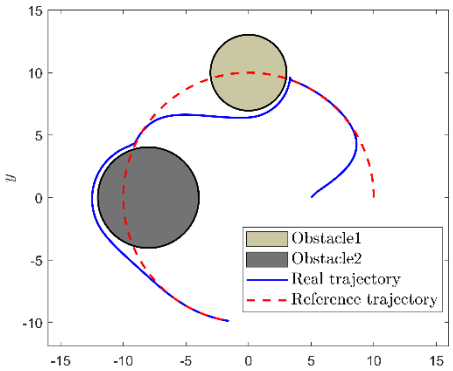
x/y 方向跟踪误差状态轨迹与 ESO 估计状态轨迹对比



x/y 方向演员网络权重估计参数曲线



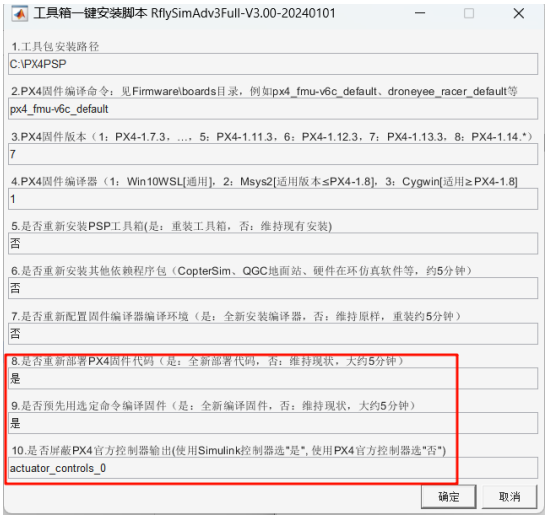
控制信号轨迹图



无人机飞行轨迹俯视图

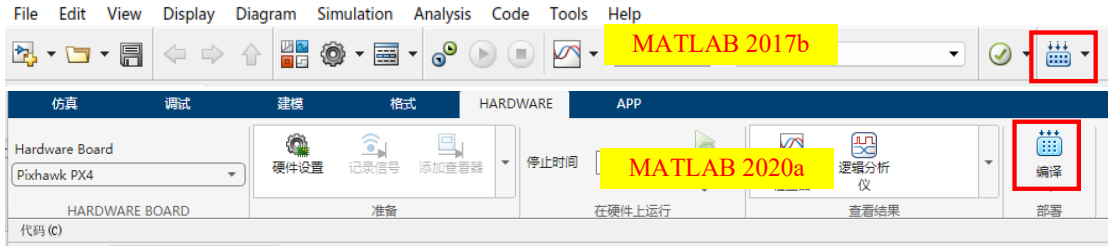
8. HITL 实验步骤

在进行硬件在环仿真前，请保证平台安装设置如下，第 10 项中选择为：actuator_controls_0，即屏蔽 PX4 软件中的 actuator_controls_0 消息输出。



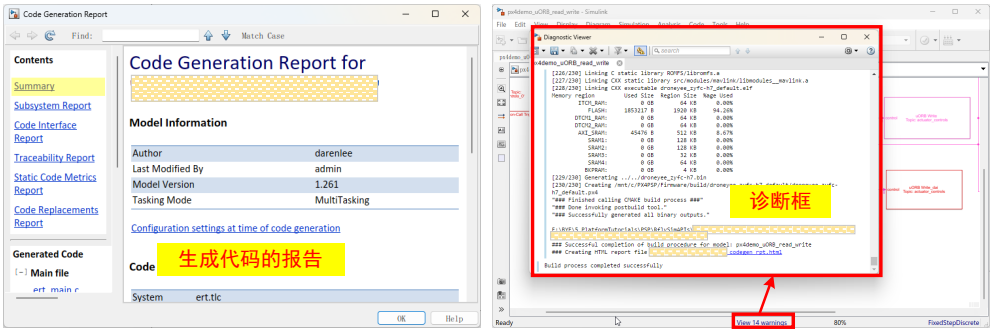
Step 1：

打开 MATLAB 运行 HITL&FLY\main_para.m 文件后，打开 HITL&FLY\RL_HITL_FLY.slx 文件，在 Simulink 中，点击编译命令。



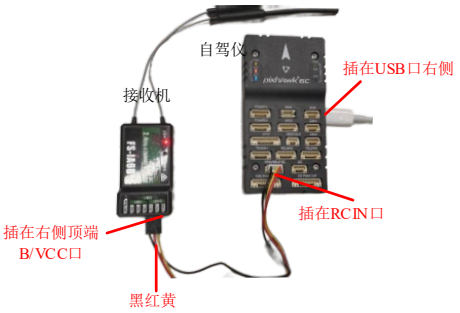
Step 2：

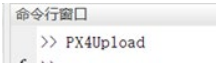
在 Simulink 的下方点击 View diagnostics 指令，即可弹出诊断对话框，可查看编译过程。在诊断框中弹出 Build process completed successfully，即可表示编译成功，左图侧为生成的编译报告。



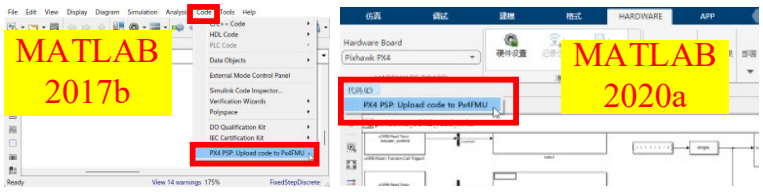
Step 3：

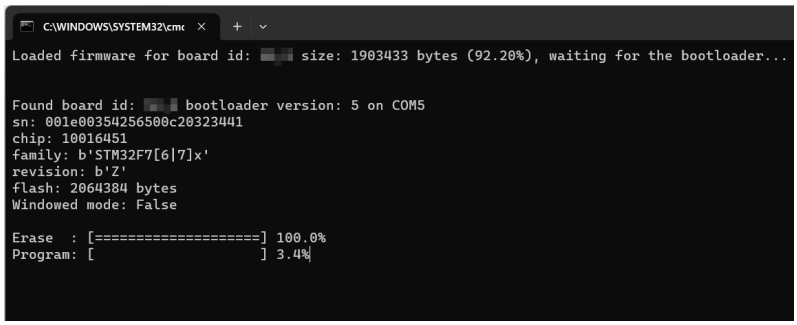
用 USB 数据线链接飞控与电脑。在 MATLAB 命令行窗口输入：PX4Upload 并运行，弹出 CMD 对话框，显示正在上传固件至飞控中，等待上传成功。





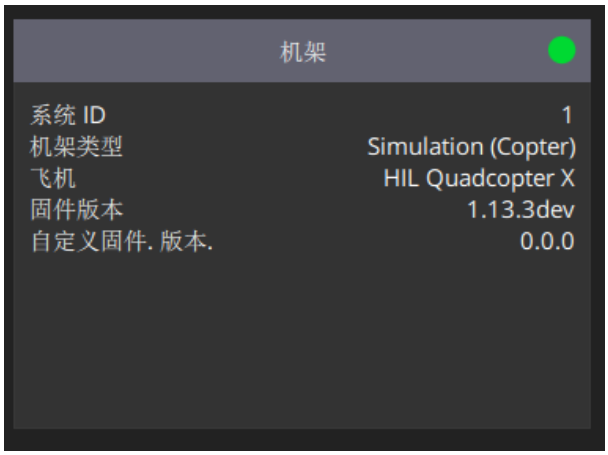
或

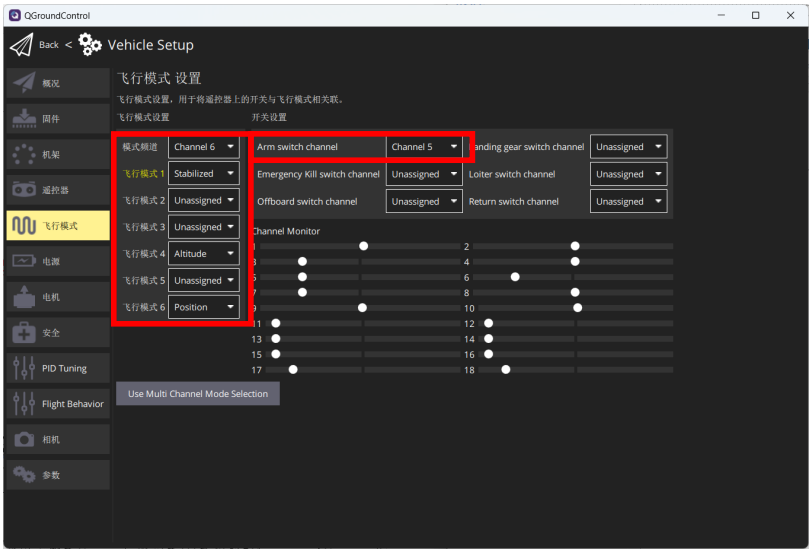




Step 4

上传成功后，打开 QGroundControl 软件，确认为如下设置：



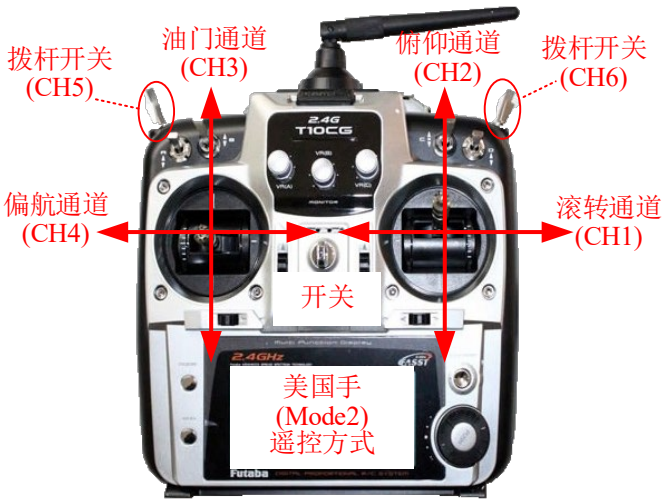


Step 5

管理员运行"*\桌面\RflyTools\SITLRun.lnk"或"*\PX4PSP\RflySimAPIs\HITLRun.bat"文件，在弹出的 CMD 对话框中输入插入的飞控 Com 端口号，即可自动启动 RflySim3D、CopterSim、QGroundControl 软件，等待 CopterSim 的状态框中显示：PX4: GPS 3D fixed & EKF initialization finished。即可在 QGroundControl 中设置飞机起飞等操作。

Step 6

遥控器的设置如下图，通过控制不同的通道即可在 RflySim3D 中观察到无人机的飞行姿态，完成硬件在环仿真。注：具体设置请见本平台的遥控器配置手册。



Step 7

拨动 CH5，解锁无人机，先用遥控器控制无人机以自稳模式或定高/定点模式起飞，然后拨动 CH7 切换至强化学习控制模式，查看控制结果，发现扩张状态观测器可以较好的观测无人机 x 方向和 y 方向上的位置和速度。

9. 官方固件实飞步骤

官方固件实飞实验运行环境			
序号	软件要求	硬件要求	
		名称	数量
1	Windows 10 及以上版本	笔记本/台式电脑①	1
2	RflySim 平台免费版及以上版本	飞思 X200 飞机②	1
	MATLAB 2017B 及以上	遥控器③	1
		数据线、杜邦线等	若干

①：推荐配置请见：https://doc.rflysim.com

②：本实验中所使用的飞机为飞思 X450 飞机的模型设计版，该飞机所搭载的飞控为 Pixhawk 6C mini，须保证平台安装时的编译命令为：px4_fmu-v6c_default，固件版本为：1.13.3。其他配套飞控请见：http://doc.rflysim.com。

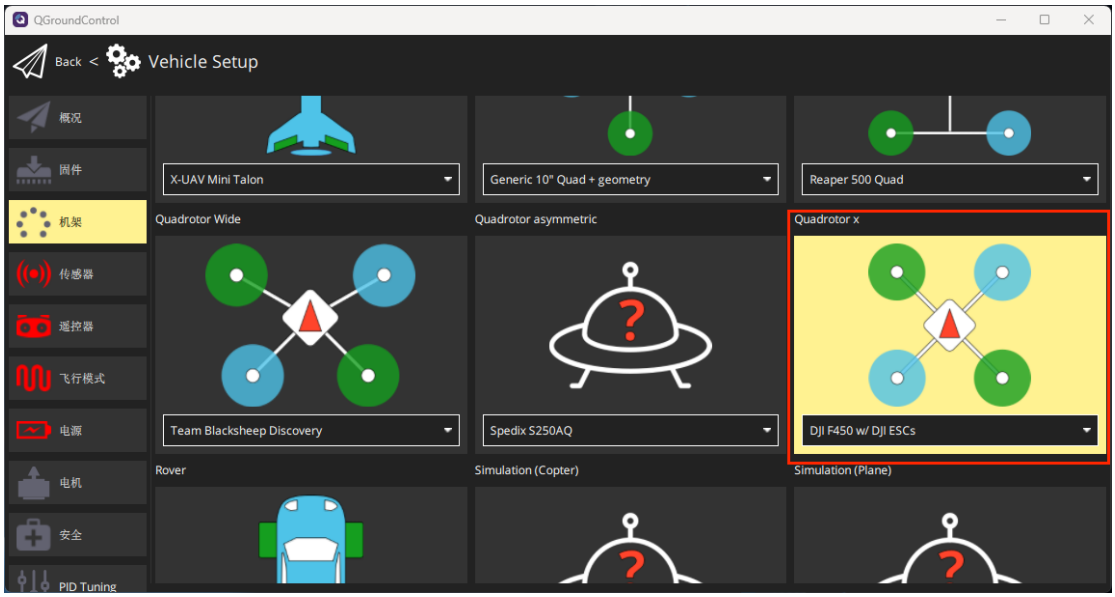
③：本实验演示所使用的遥控器为：天地飞ET10、配套接收器为：WFLY RF209S。遥控器相关配置见：..\e11_RC-Config\Readme.pdf

Step 1：

请扫码或点击下方二维码，将本例程文件夹下：HIL&FLY\px4_fmu-v6c_default1133.px4 (飞控固件)上传至飞控中。

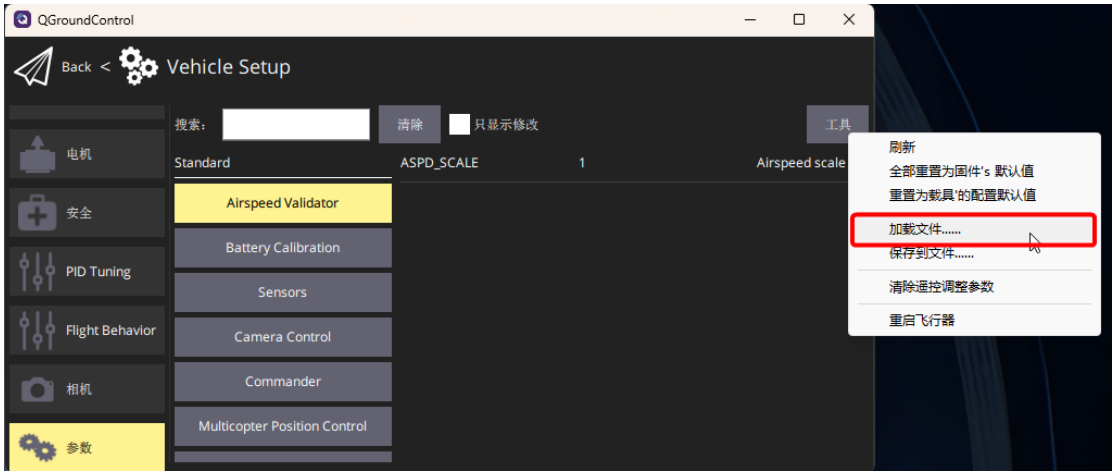
Step 2：

将飞机通过USB与电脑进行连接，打开QGC软件，设置机架为：DJI F450 w/ DJI ESCs；



Step 3：

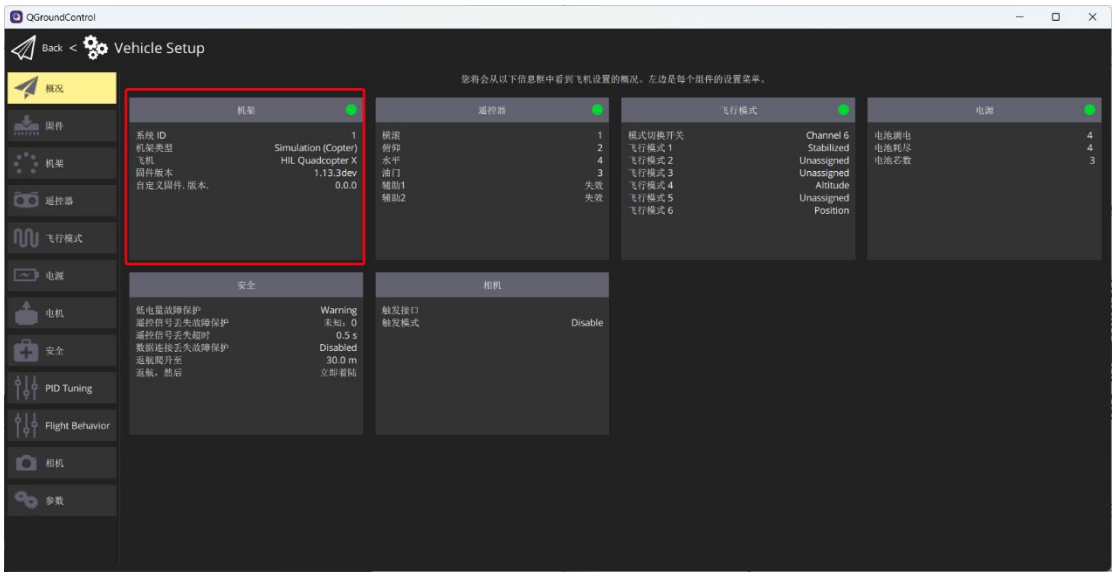
选择加载本例程文件夹下的参数文件：X450.params 文件。

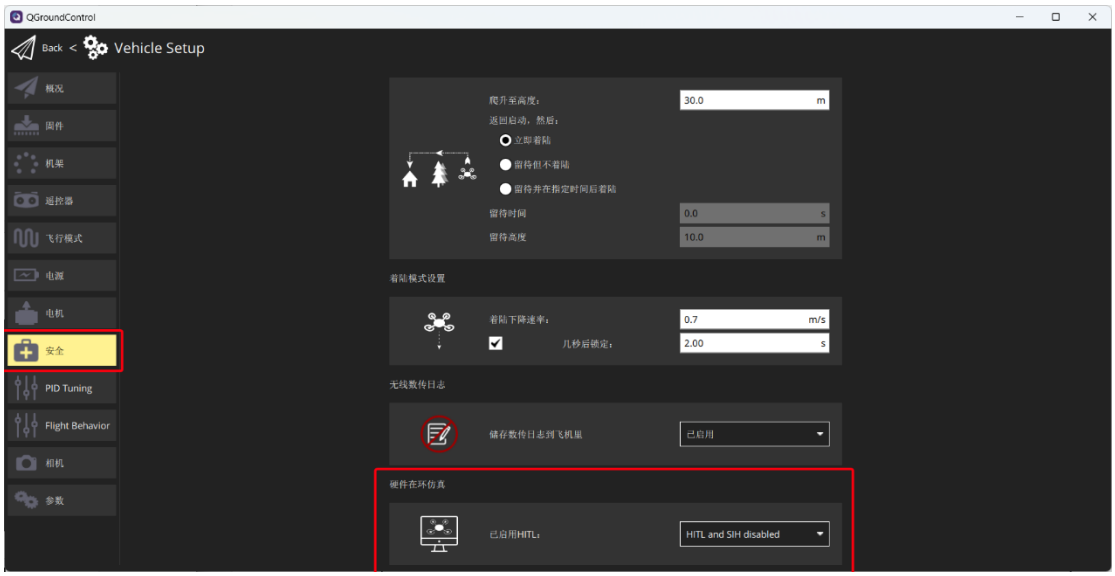


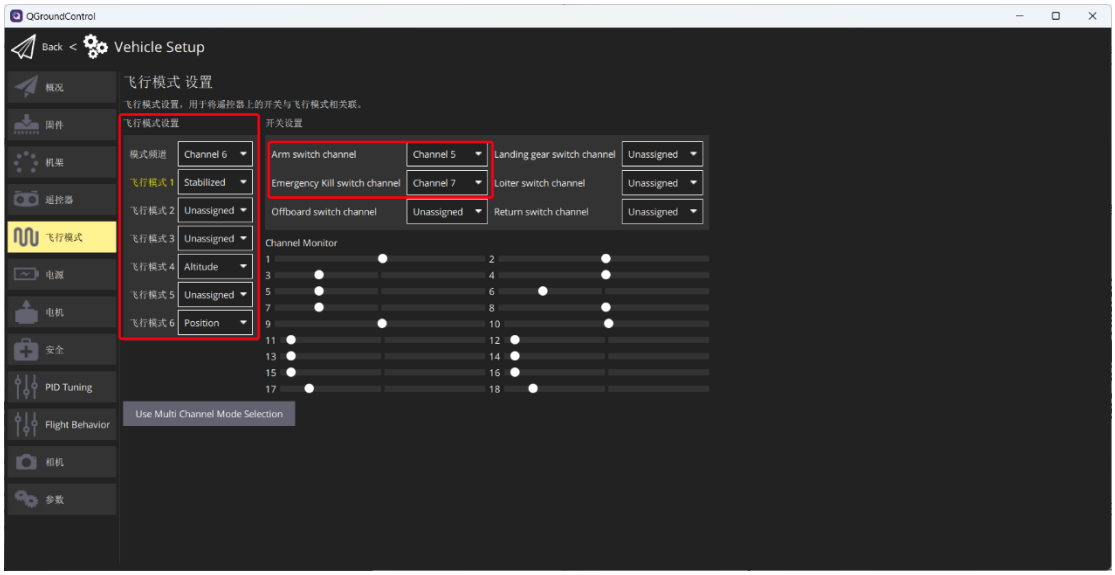
加载成功后，断开飞机，再次进行连接飞机确保所有设置均已完成。

Step 4：

打开 QGC 地面站在其中进行如下设置：



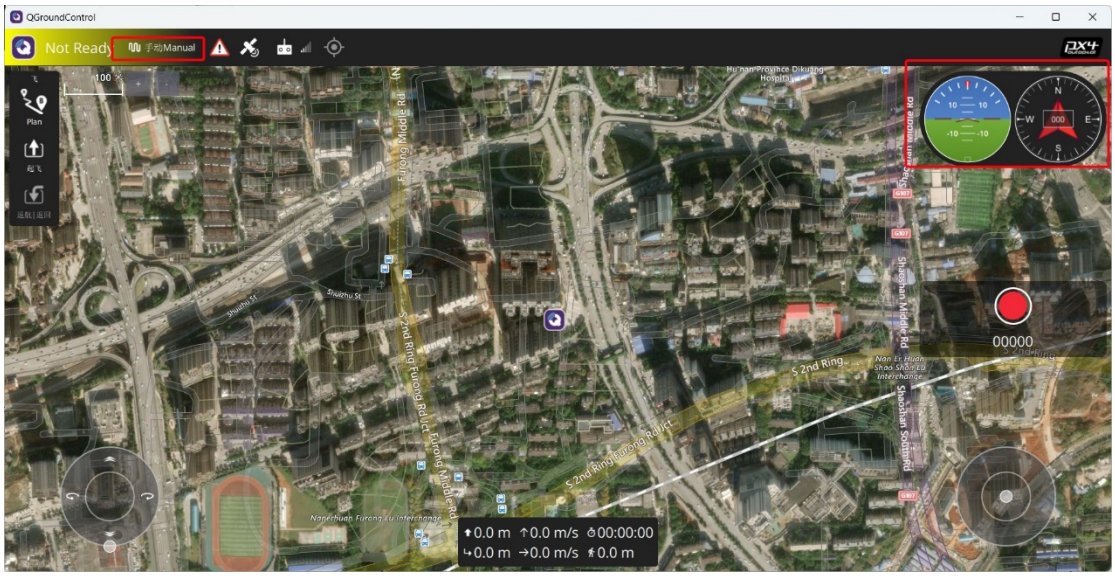




注：该飞行模式中的各通道设置须于遥控器中所设置的通道对映。

Step 5：

手动摆动飞机，查看 QGC 右上角仪表盘的显示情况，并确认飞机状态切换到手动 Manual 模式下。



Step 6：

请在指定飞场进行无人机实飞，若正常起飞，说明无人机状态良好；若未正常起飞，请检查传感器校准、参数设置等，具体请联系飞机生产厂家进行解决。**请务必保证飞机状态良好的情况下，再进行下一步操作。**

10.实飞实验步骤

官方固件实飞实验运行环境		
序号	软件要求	硬件要求

		名称	数量
1	Windows 10 及以上版本	笔记本/台式电脑①	1
2	RflySim 平台完整版及以上版本	飞思 X200 飞机②	1
	MATLAB 2017B 及以上	遥控器③	1
		数据线、杜邦线等	若干

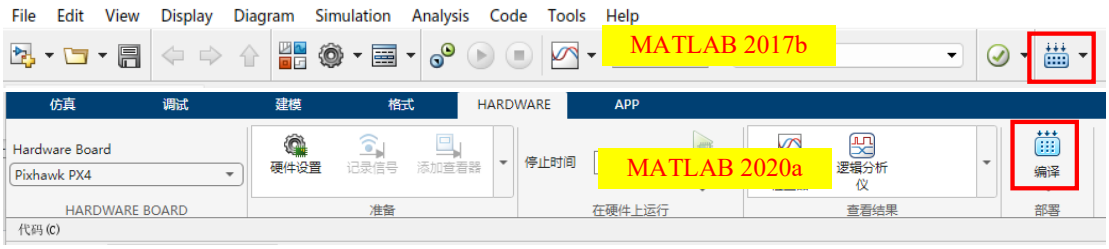
①：推荐配置请见：https://doc.rflysim.com

②：本实验中所使用的飞机为飞思 X450 飞机的模型设计版，该飞机所搭载的飞控为 Pixhawk 6C mini，须保证平台安装时的编译命令为：px4_fmu-v6c_default，固件版本为：1.13.3。其他配套飞控请见：http://doc.rflysim.com。

③：本实验演示所使用的遥控器为：天地飞 ET10、配套接收器为：WFLY RF209S。遥控器相关配置见：..\e11_RC-Config\Readme.pdf

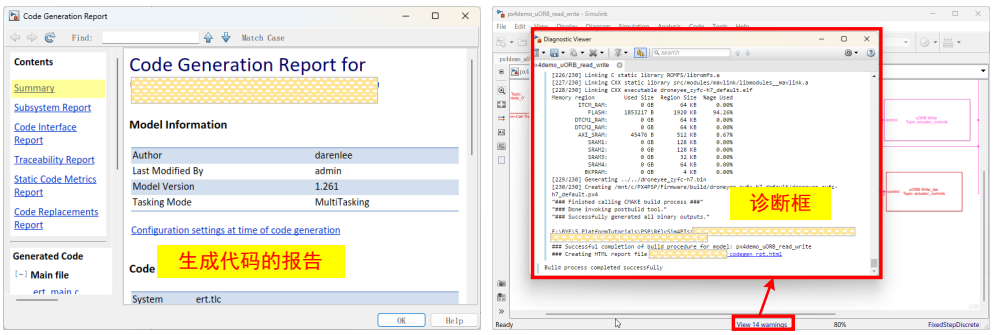
Step 1：

打开 HITL&FLY\RL_HITL_FLY.slx 文件，在 Simulink 中，点击编译命令。



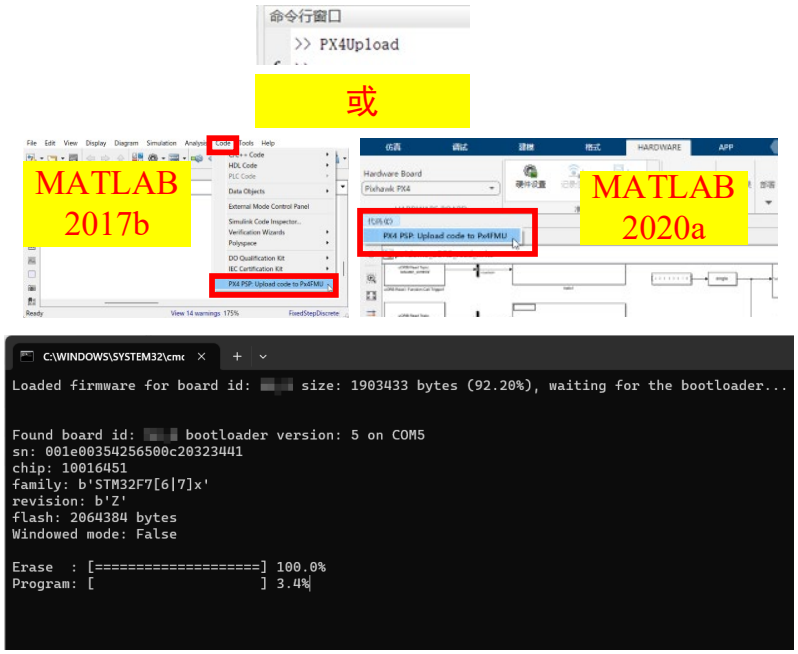
Step 2：

在 Simulink 的下方点击 View diagnostics 指令，即可弹出诊断对话框，可查看编译过程。在诊断框中弹出 Build process completed successfully，即可表示编译成功，左侧为生成的编译报告。



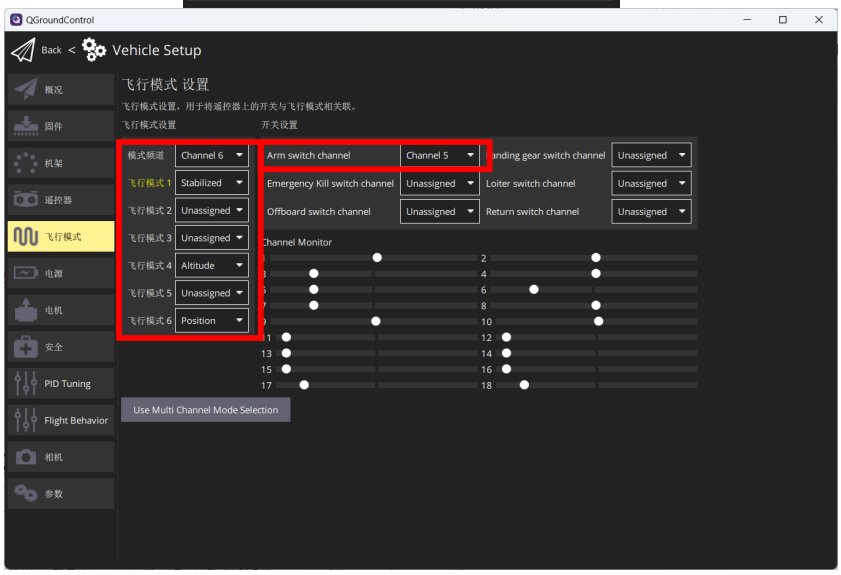
Step 3：

用 USB 数据线链接飞控(或飞机)与电脑。在 MATLAB 命令行窗口输入：PX4Upload 并运行，弹出 CMD 对话框，显示正在上传固件至飞机中，等待上传成功。



Step 4：

打开 QGroundComtrol 软件，等待飞机连接成功。确认无人机机架类型选择如下图，并设置遥控器通道如下，其中 CH5 为解锁。



Step 5:

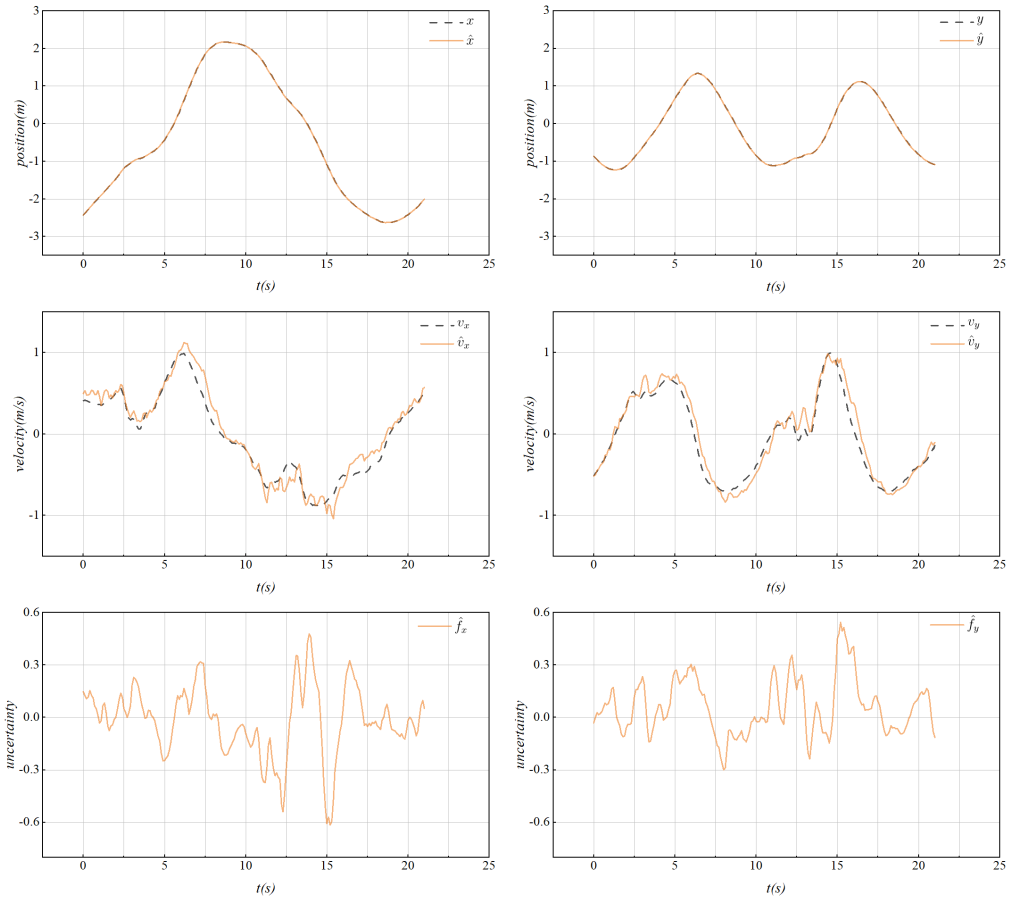
遥控器的设置如下图。注：遥控器设置中，CH5 通道需设置为二段式开关，CH6 通道设置为三段式开关。具体设置请见本平台的遥控器配置手册。

Step 6:

为确保安全，可在飞机上系上安全绳，并将安全绳的另一端固定在重物上。飞行时人在安全半径以外，在定点模式下，让油门在中位附近，即可实现定点。

打开 QGC，对无人机的各项参数进行初始化设置，如果在室内请打开 ekf 滤波器的其他传感器模式。将无人机放在试验场地，注意远离人群，保证安全。

拨动 CH5，解锁无人机，先用遥控器控制无人机以自稳模式或定高/定点模式起飞，然后拨动 CH7，切换至强化学习控制模式，查查看控制结果，发现扩张状态观测器可以较好的观测无人机 x 方向和 y 方向上的位置和速度。部分控制结果如下图所示：



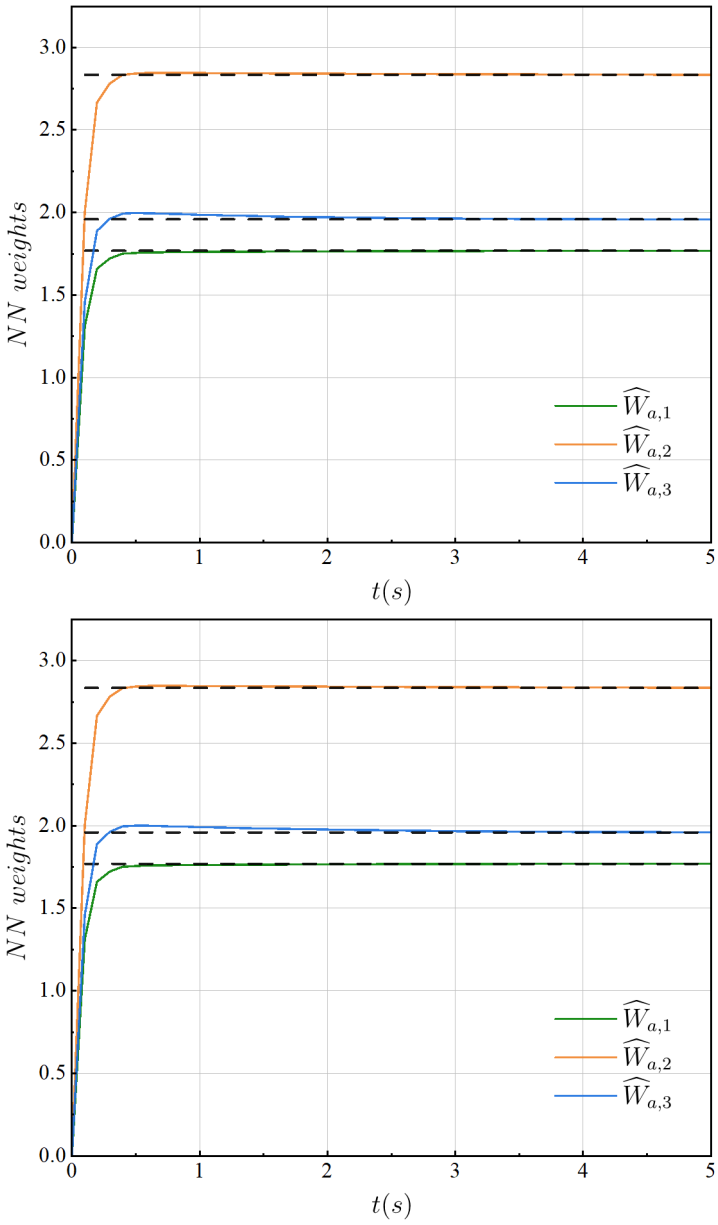

3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
$NN\ weights$
$t(s)$
$\widehat{W}_{a,1}$
$\widehat{W}_{a,2}$
$\widehat{W}_{a,3}$

图 1 神经网络权重轨迹

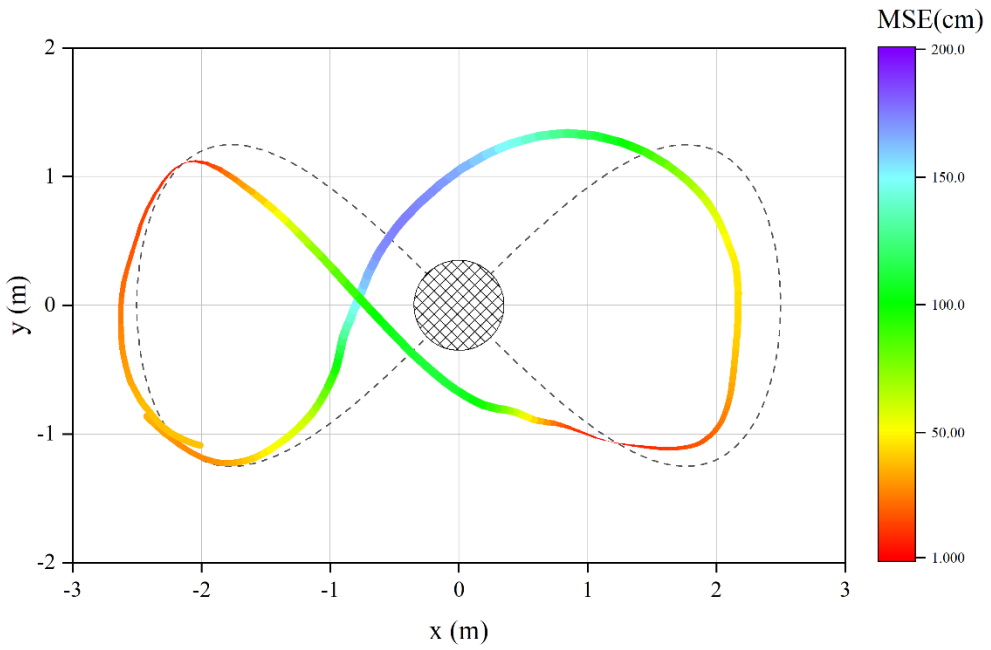


图 2 指令跟踪轨迹和实际避障轨迹

图 3 实飞逐帧合成图

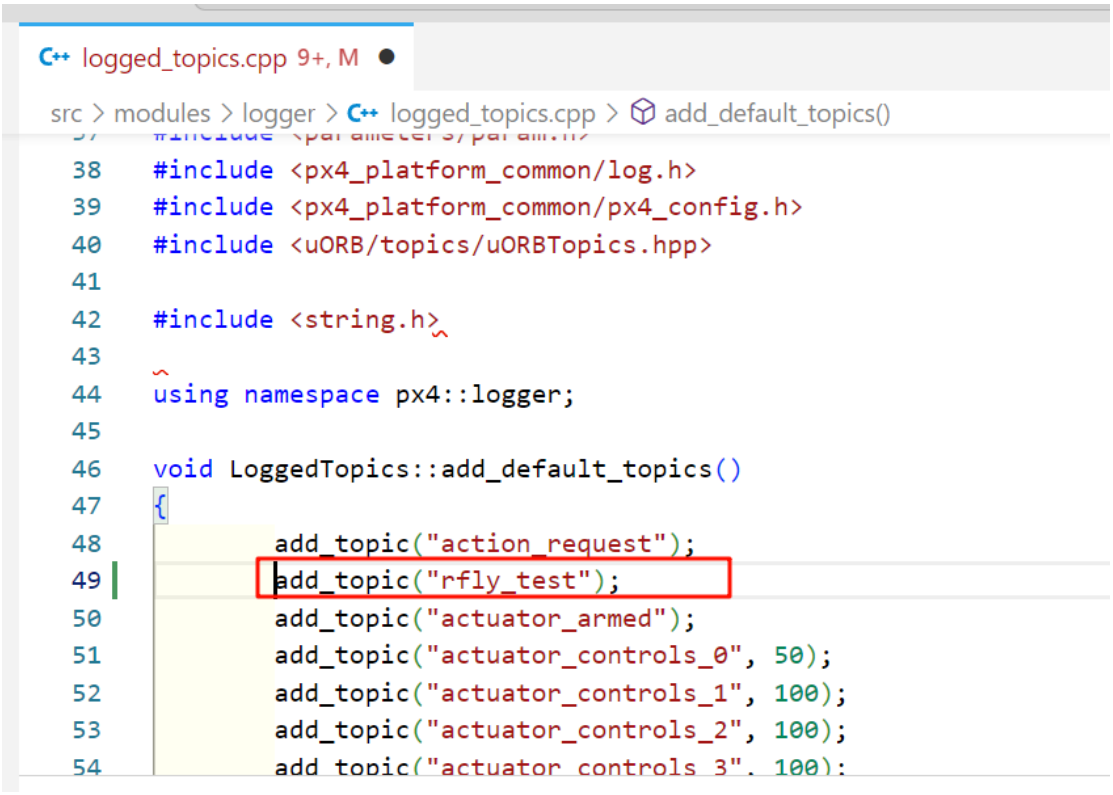
7、参考资料

[1]. 无

8、常见问题

Q1：自定义的 uORB 消息在导出的.ulg 日志文件中无法生成日志数据

A1：打开 “*:\PX4PSP\Firmware\src\modules\logger\logged_topics.c” 文件。将自定义的 uORB 消息添加到该文件中，即在第一个函数中添加代码“add_topic("rfly_test");”如下：



将该文件保存后，双击打开"*\桌面\RflyTools\Win10WSL.lnk"的 WSL 子系统，进行编译固件编译完成后，重复 Step3 烧录飞控当中，即可在 QGroundControl 导出.ulg 文件，处理后可得到自定义消息发布的数据。