


# 后端部署文档

项目名称：

图匠 —— 对话式图像创意室

开发团队：同济大学 锟斤拷

后端文件结构如下图：（本地部署无需 `weights` 文件夹）

名称	大小	类型	修改时间	属性	所有者
..					
configs		文件夹	2024/6/5, 14:44	drwx...	root
img_tmp		文件夹	2024/6/2, 23:26	drwx...	root
models		文件夹	2024/6/2, 22:03	drwx...	root
simple_image_process		文件夹	2024/6/1, 22:39	drwx...	root
weights		文件夹	2024/6/2, 23:13	drwx...	root
__pycache__		文件夹	2024/6/5, 17:10	drwx...	root
app.py	201KB	Python 源...	2024/6/5, 15:00	-rw-r...	root
clean_split.py	2KB	Python 源...	2024/6/2, 22:58	-rw-r...	root
config.py	6KB	Python 源...	2024/6/2, 22:58	-rw-r...	root
config.yml	599 Bytes	Yaml 源文件	2024/6/2, 22:58	-rw-r...	root
data.db	196KB	SQLite	2024/6/5, 20:50	-rw-r...	root
dataset.py	2KB	Python 源...	2024/6/2, 22:58	-rw-r...	root
IAA_main.py	9KB	Python 源...	2024/6/2, 22:59	-rw-r...	root
option.py	1KB	Python 源...	2024/6/2, 22:58	-rw-r...	root
P2P.py	1KB	Python 源...	2024/6/5, 16:59	-rw-r...	root
requirements.txt	545 Bytes	文本文档	2024/6/5, 16:56	-rw-r...	root
search_space.json	136 Bytes	JSON 源文件	2024/6/2, 22:58	-rw-r...	root
startWSGL.ini	290 Bytes	配置设置	2024/6/5, 19:43	-rw-r...	root
util.py	1KB	Python 源...	2024/6/2, 22:58	-rw-r...	root

注：服务端部署和本地部署的区别在于是否使用 `huggingface` 在线下载模型权重 / `ip`与端口设置

## 本地部署

### 配置要求

请使用带 GPU 的电脑进行配置，以下是锟斤拷团队本地部署成功运行的配置：

配置	参数
操作系统	Windows 11
GPU	NVIDIA GeForce RTX 4050
CPU	12th Gen Intel(R) Core(TM) i7-12650H
CUDA	12.1.112
CUDNN	8.9.3
Python	3.8.19
Pytorch	2.2.2

## 安装环境

在后端项目文件夹 `backend` 运行以下指令：

```
1 pip install -r requirements.txt
```

即可一键安装所有需要的库。

## 运行后端

本地部署由于网络环境较为灵活，建议使用 huggingface 的在线模型权重下载，将项目文件夹中 `P2P.py` 中的 `model_id` 修改如下：

```
1 model_id = "timbrooks/instruct-pix2pix"
```

由于需要从来自美国的 hugging-face.org 网站上下载大文件，首次运行时需要合理配置本机网络环境。模型文件将被默认下载至 `C:/.cache/` 文件夹中。

此外，可以手动在 `app.py` 中调整打开的端口：

```
5780 if __name__ == '__main__':
5781     config = dict(
5782         host='0.0.0.0',
5783         port=3306,
5784         debug=True,
5785         allow_unsafe_werkzeug=True
5786     )
5787     socketio.run(app, **config)
```

并于前端 `fontend` 文件夹中更新 `main.ts` 中连接的后端 ip 和端口号即可：

```
19 // 设置后端服务器的基准 URL
20 // axios.defaults.baseURL = 'http://127.0.0.1:8080';
21 axios.defaults.baseURL = 'http://123.60.90.34:3306';
```

设置完毕后，运行 `app.py` 即可打开后端：

```
1 python app.py
```

```
(Chat) D:\Code\SE\Final2\kunjinkao\backend>python app.py
* Serving Flask app 'app'
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on all addresses (0.0.0.0)
* Running on http://127.0.0.1:3306
* Running on http://100.78.19.205:3306
Press CTRL+C to quit
* Restarting with stat
* Debugger is active!
* Debugger PIN: 317-208-838
```

## 云端部署

### 配置要求

选择带 GPU 的服务器即可，锟斤拷团队部署于华为云 ECS 弹性云服务器：

云服务器信息

ID	7dcefa4-8de0-44a6-bfc1-2a02d2117dd9
名称	Backend
描述	图匠后端服务器（带T4 GPU）
区域	华东-上海一
可用区	可用区1
规格	GPU加速型   8vCPUs   32GiB   p2.xlarge.4
	GPU显卡: 1 * NVIDIA T4 / 1 * 16G
镜像	<a href="#">后端镜像-未配置完毕</a>   私有镜像
	版本: Ubuntu 22.04 server 64bit
虚拟私有云	<a href="#">vpc-default</a>
全弹性公网IP	-- <a href="#">绑定</a>

计费模式	按需计费
创建时间	2024/05/24 02:04:50 GMT+08:00
启动时间	2024/05/24 02:05:08 GMT+08:00
定时删除时间	-- <a href="#">修改</a>

运行中 | 监控 监控中 | 主机安全 关闭 [开启防护](#)

▼ 云硬盘

系统盘  
[Backend-volume-0000](#) 通用型SSD | 70 GiB [扩容](#)

▼ 网卡

主网卡  
[subnet-default](#) 192.168.0.191 | [123.60.90.34](#)

▼ 安全组

[Sys-WebServer](#)  
[Sys-FullAccess](#)  
[default](#)

▼ 弹性公网IP

[123.60.90.34](#) | 4 Mbit/s

带有一张 T4 显卡用于指令修改图像模型的在线推理。

## 安装环境

### 安装英伟达驱动

首先检查是否存在驱动：

```
1 nvidia-smi
```

若报错，则不存在驱动，使用以下指令查看显卡型号：

```
1 lspci | grep NVIDIA
```

查看对应的显卡驱动：<https://www.nvidia.cn/geforce/drivers/>

安装驱动文件，并复制到 `/tmp` 目录下，安装驱动：

```
1 sudo sh NVIDIA-Linux-文件名.run -no-x-check -no-nouveau-check -no-opengl-files
```

### 安装 Anaconda

直接使用 `wget` 指令在终端下载，参考官网指令：<https://www.anaconda.com/products/distribution>

下载为 `.sh` 文件，使用以下指令修改为可执行：

```
1 sudo chmod -R 777 Anaconda3-2022.05-Linux-x86_64.sh
```

直接执行 `.sh` 文件即可完成下载。

### 安装 cuda/cudnn

验证是否安装cuda：

```
1 nvcc -V
```

若未安装，参考 [https://docs.nvidia.com/deeplearning/triton-inference-server/release-notes/rel\\_22-06.html#rel\\_22-06](https://docs.nvidia.com/deeplearning/triton-inference-server/release-notes/rel_22-06.html#rel_22-06) 查找对应型号。

CUDA 下载链接: <https://developer.nvidia.com/cuda-toolkit-archive>

CUDNN 下载链接: <https://developer.nvidia.com/cudnn>

安装依赖库:

```
1 sudo apt-get install freeglut3-dev build-essential libx11-dev libxmu-dev libxi-dev  
libgl1-mesa-glx libglu1-mesa libglu1-mesa-dev
```

CUDA安装:

```
1 sudo sh CUDA文件名.run
```

解压 CUDNN 压缩包:

```
1 tar -xzf cudnn-11.0-linux-x64-v8.0.4.30.tgz
```

当前目录下输入以下命令进行安装:

```
1 sudo cp cuda/include/cudnn*.h /usr/local/cuda/include  
2 sudo cp cuda/lib64/libcudnn* /usr/local/cuda/lib64  
3 sudo chmod a+r /usr/local/cuda/include/cudnn.h /usr/local/cuda/lib64/libcudnn*
```

## 安装 pytorch

版本对应参考官网: <https://pytorch.org/>

下载对应版本的 pytorch:

```
1 conda install pytorch==x.x.x torchvision==x.x.x torchaudio==x.x.x cudatoolkit=x.x -  
c pytorch
```

## 其他环境部署

参考《本地部署 - 安装环境》即可。

## 模型权重加载

由于服务器端对 huggingface 网络获取较差, 直接获取模型权重文件耗时较长, 因此采用本地下载传入的方式。

在服务器端建立文件夹 `P2P_Weights`, 前往网站 [timbrooks/instruct-pix2pix · Hugging Face](#) 下载以下全部内容:

main

instruct-pix2pix

5 contributors

History: 18 commits

<div>patrickvonplaten</div>	Fix deprecated float16/fp16 variant loading through new 'version' API. (#17)	31519b5	11 months ago
<div>feature_extractor</div>	add model		over 1 year ago
<div>safety_checker</div>	Fix deprecated float16/fp16 variant loading through new ...		11 months ago
<div>scheduler</div>	Update scheduler/scheduler_config.json		over 1 year ago
<div>text_encoder</div>	Fix deprecated float16/fp16 variant loading through new ...		11 months ago
<div>tokenizer</div>	add model		over 1 year ago
<div>unet</div>	Fix deprecated float16/fp16 variant loading through new ...		11 months ago
<div>vae</div>	Fix deprecated float16/fp16 variant loading through new ...		11 months ago
<div>.gitattributes</div>	1.48 kB	<div></div> initial commit	over 1 year ago
<div>README.md</div>	1.29 kB	<div></div> Update README.md (#15)	12 months ago
<div>instruct-pix2pix-00-22000.ckpt</div>	<div><div></div><div>pickle</div></div> 7.7 GB	<div></div> <div>LFS</div> <div></div> uP	over 1 year ago
<div>instruct-pix2pix-00-22000.safetensors</div>	<div><div></div><div></div></div> 7.7 GB	<div></div> <div>LFS</div> <div></div> Adding 'safetensors' variant of this model (#1)	over 1 year ago
<div>model_index.json</div>	616 Bytes	<div></div> Update model_index.json	over 1 year ago

将所有下载权重文件上传至 `P2P_Weights` 文件夹，并修改 `P2P.py` 内 `model_id` 为完整路径：

```

9     # model_id = "timbrooks/instruct-pix2pix"
10    model_id = "/root/Code/Models/P2P/weights"
11    pipe = StableDiffusionInstructPix2PixPipeline.from_pretrained(
12        model_id, torch_dtype=torch.float16, safety_checker=None)
13    pipe.to("cuda")
14    pipe.scheduler = EulerAncestralDiscreteScheduler.from_config(
15        pipe.scheduler.config)

```

## 生产环境部署

本团队参考 [部署到生产环境 — Flask 文档 \(2.0.x\) \(flask-zh.readthedocs.io\)](#)

将项目使用 WSGI 生产服务器 `waitress` 运行。