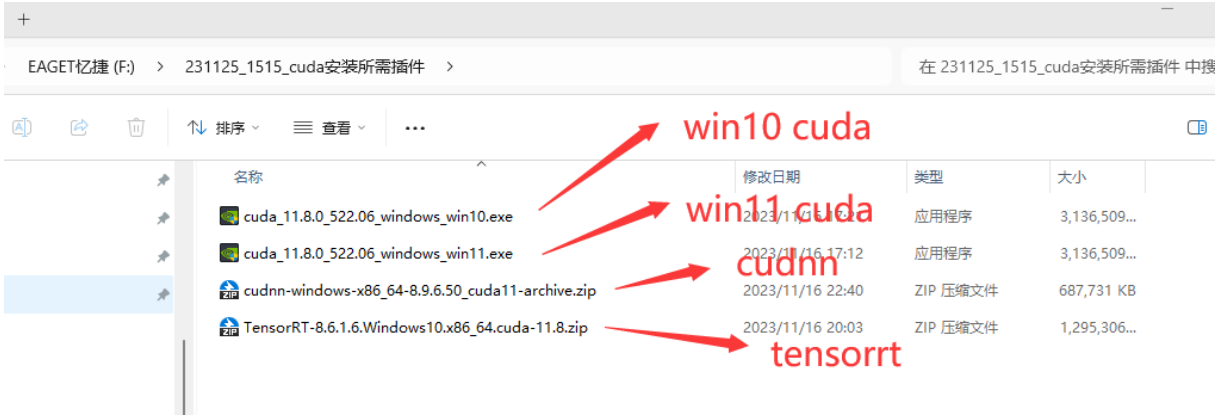
231125_tenserrt 环境安装

范仁义

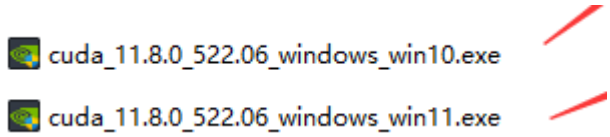
一、所需文件

1、所需文件



二、cuda 安装教程

1、cuda 安装



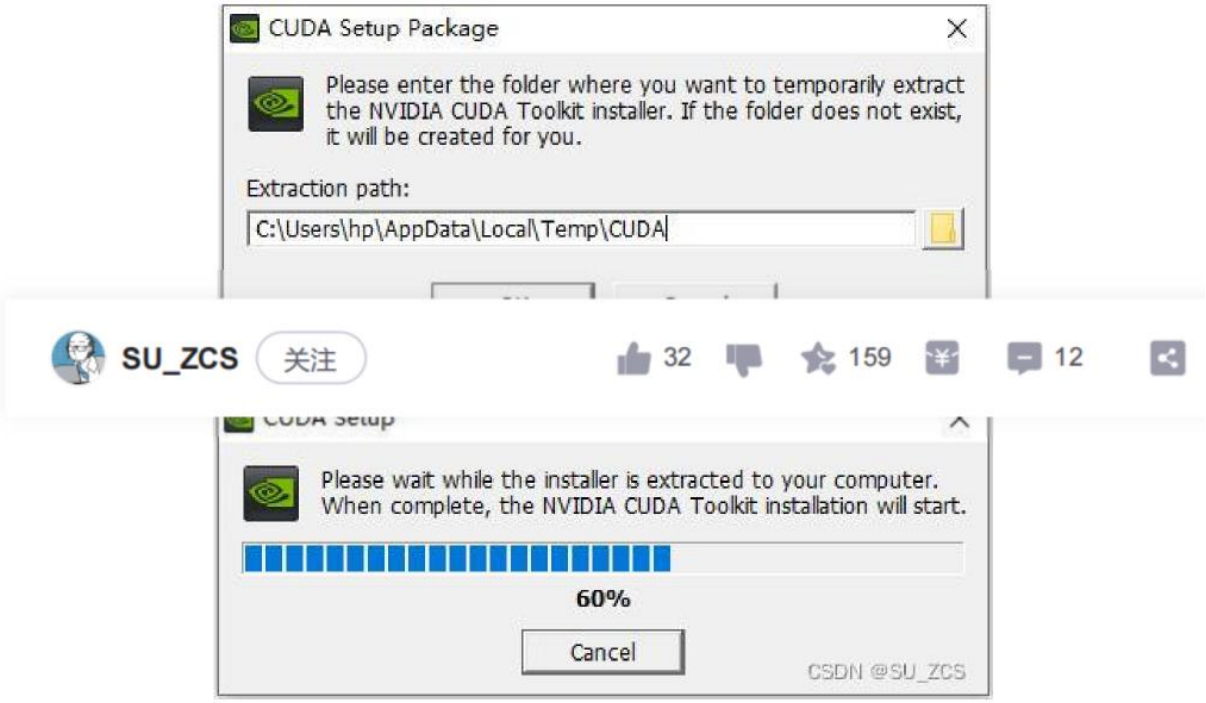
点击上面的 cuda.exe，win10 就装 win10 的，win11 就装 win11 的

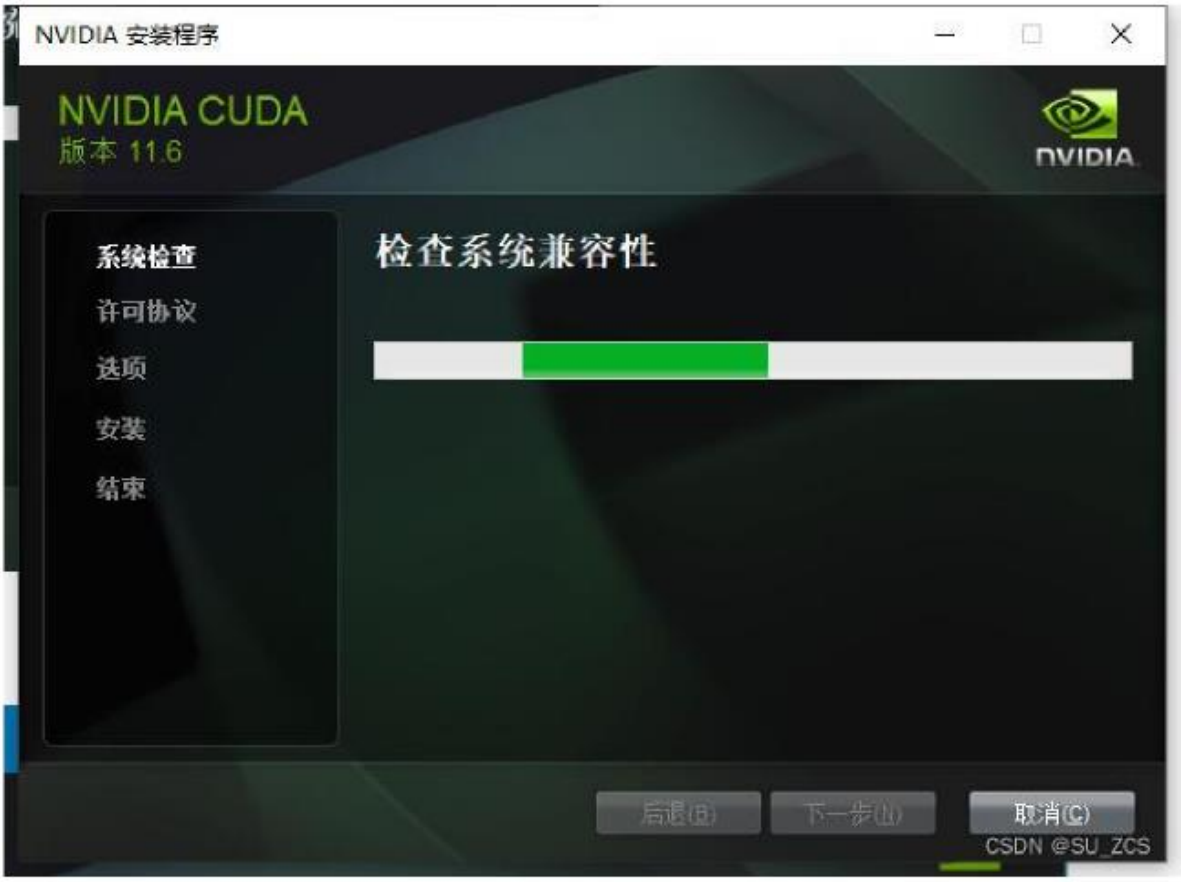
基本一步步默认安装即可

以下图是从别的位置截取到的图：

1.3 cuda toolkit安装

双击exe文件进行安装即可

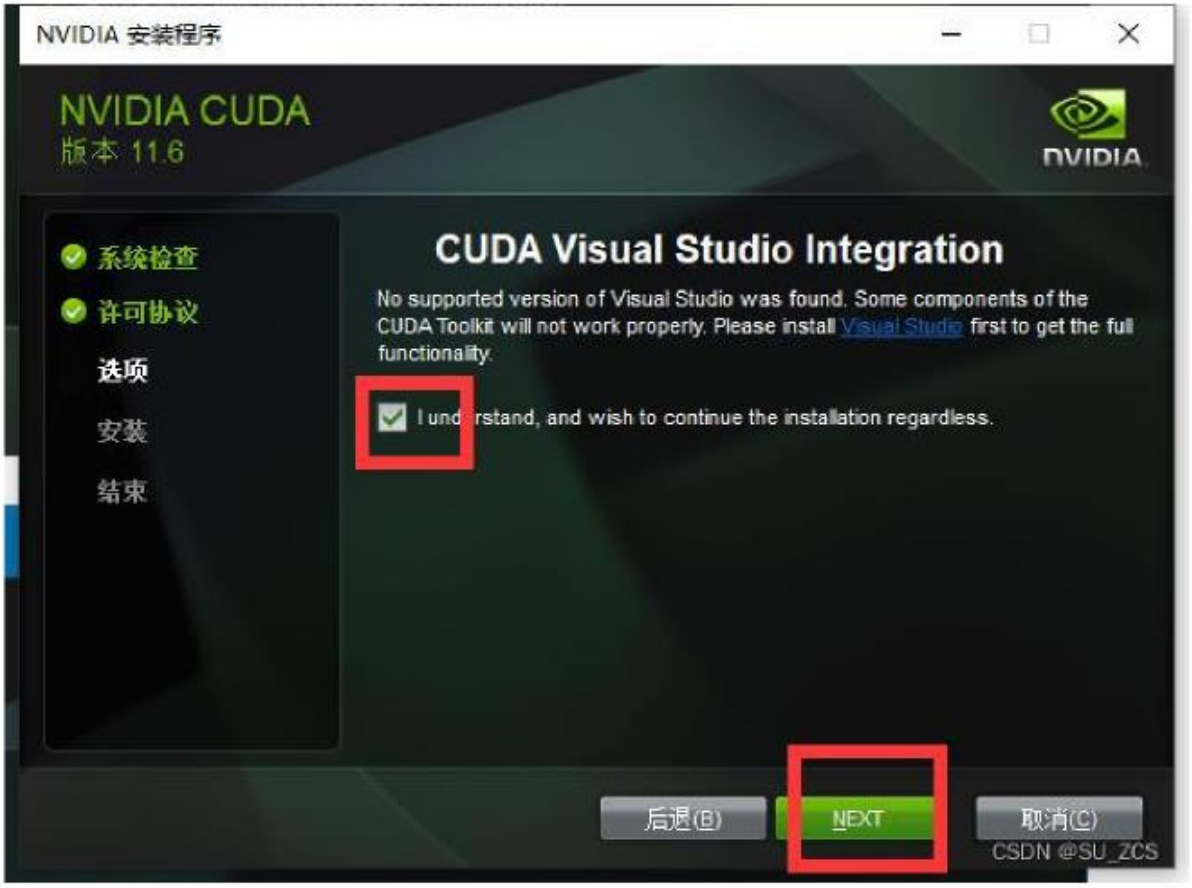


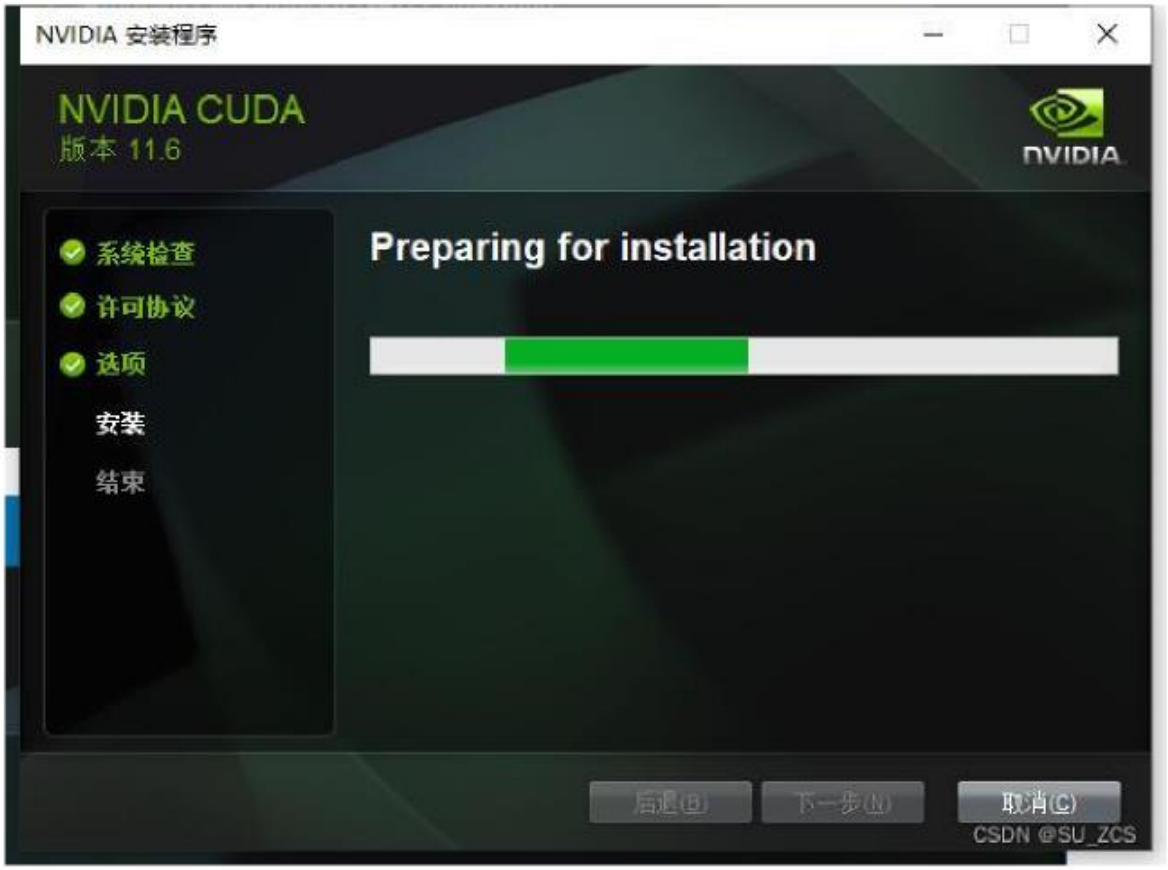

NVIDIA 安装程序
NVIDIA CUDA
版本 11.6
系统检查
许可协议
选项
安装
结束
检查系统兼容性
后退(B)
下一步(N)
取消(C)
CSDN @SU_ZCS

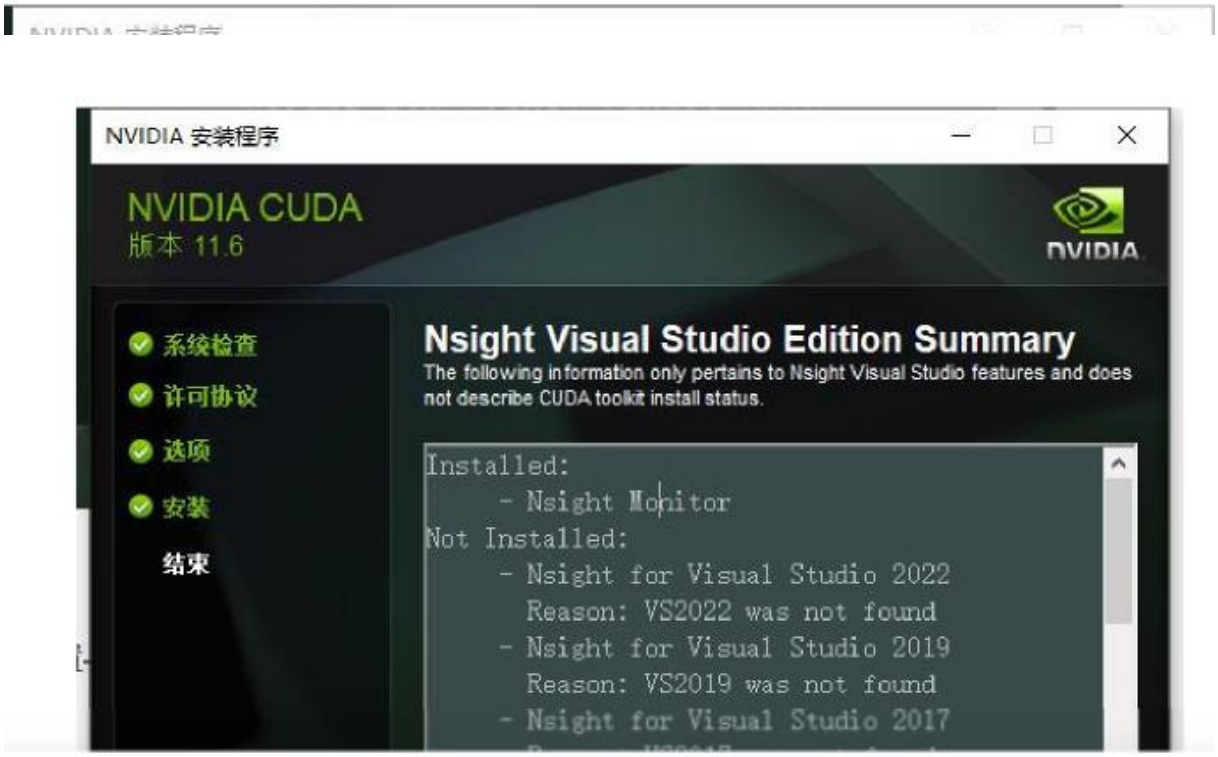

NVIDIA 安装程序
NVIDIA CUDA
版本 11.6
系统检查
许可协议
选项
安装
结束
NVIDIA 软件许可协议
请仔细阅读以下 NVIDIA 软件许可协议。
to the
NVIDIA CUDA Toolkit, the NVIDIA CUDA Samples, the
NVIDIA
Display Driver, NVIDIA Nsight tools (Visual Studio
Edition),
and the associated documentation on CUDA APIs,
programming
model and development tools. If you do not agree with
the
terms and conditions of the license agreement, then
如果您接受此协议条款，请单击同意并继续。
同意并继续(A)
取消(C)
CSDN @SU_ZCS


NVIDIA 安装程序
NVIDIA CUDA
版本 11.6
直接选精简，安装即可
NVIDIA
系统检查
许可协议
选项
安装
结束
安装选项
精简(E)（推荐）
Installs all CUDA components and overwrites current Display Driver.
自定义(C)（高级）
Allows you to select the components you want to install.
备注：在安装过程中可能会出现闪烁。
后退(B)
下一步(N)
取消(C)
CSDN @SU_ZCS


NVIDIA 安装程序
NVIDIA CUDA
版本 11.6
不要修改默认路径
NVIDIA
系统检查
许可协议
选项
安装
结束
选择安装位置
CUDA Documentation
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.6
Samples
C:\ProgramData\NVIDIA Corporation\CUDA Samples\v11.6
CUDA Development
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.6
浏览...
浏览...
浏览...
后退(B)
下一步(N)
取消(C)
CSDN @SU_ZCS


NVIDIA 安装程序
NVIDIA CUDA
版本 11.6
NVIDIA
系统检查
许可协议
选项
安装
结束
CUDA Visual Studio Integration
No supported version of Visual Studio was found. Some components of the CUDA Toolkit will not work properly. Please install Visual Studio first to get the full functionality.
I understand, and wish to continue the installation regardless.
后退(B)
NEXT
取消(C)
CSDN @SU_ZCS


NVIDIA 安装程序
NVIDIA CUDA
版本 11.6
系统检查
许可协议
选项
安装
结束
Preparing for installation
后退(B)
下一步(N)
取消(C)
CSDN @SU_ZCS

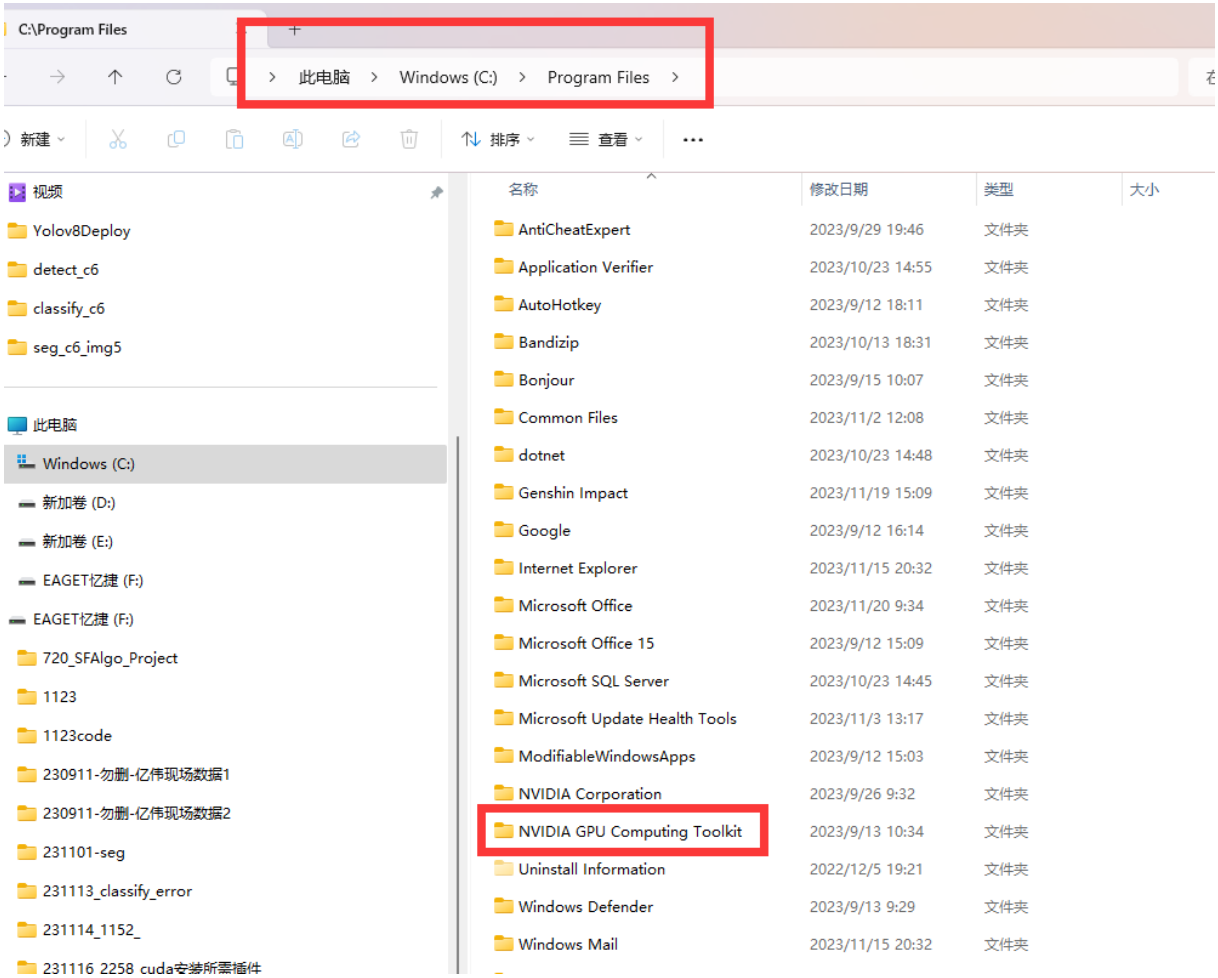

NVIDIA 安装程序
NVIDIA CUDA
版本 11.6
系统检查
许可协议
选项
安装
结束
Nsight Visual Studio Edition Summary
The following information only pertains to Nsight Visual Studio features and does not describe CUDA toolkit install status.
Installed:
- Nsight Monitor
Not Installed:
- Nsight for Visual Studio 2022
Reason: VS2022 was not found
- Nsight for Visual Studio 2019
Reason: VS2019 was not found
- Nsight for Visual Studio 2017

2、cudnn 安装

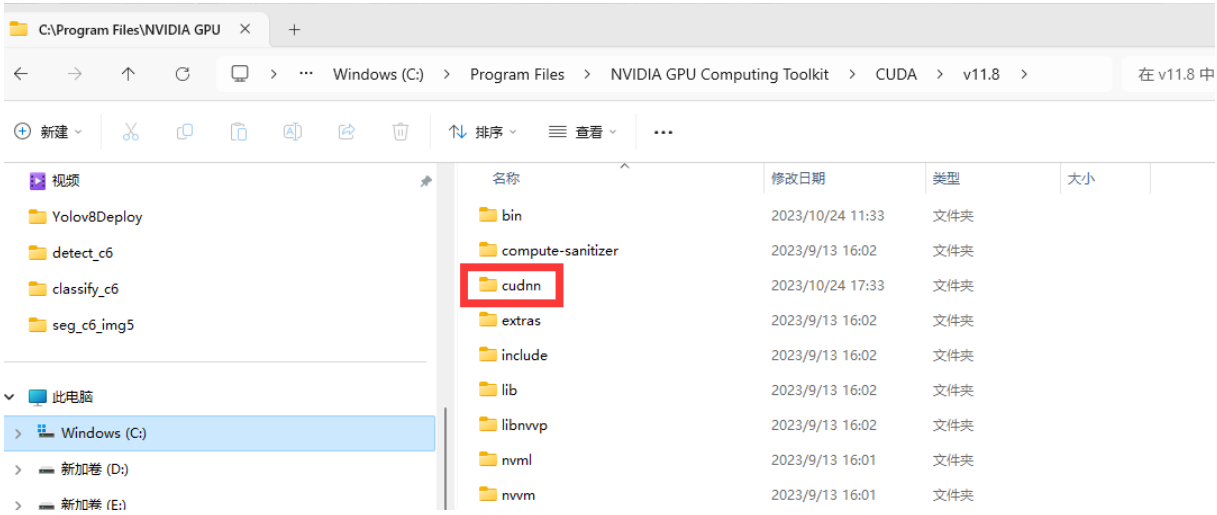
cudnn-windows-x86_64-8.9.6.50_cuda11-archive.zip

将如下文件复制到 cuda 的安装目录，默认是

C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.8

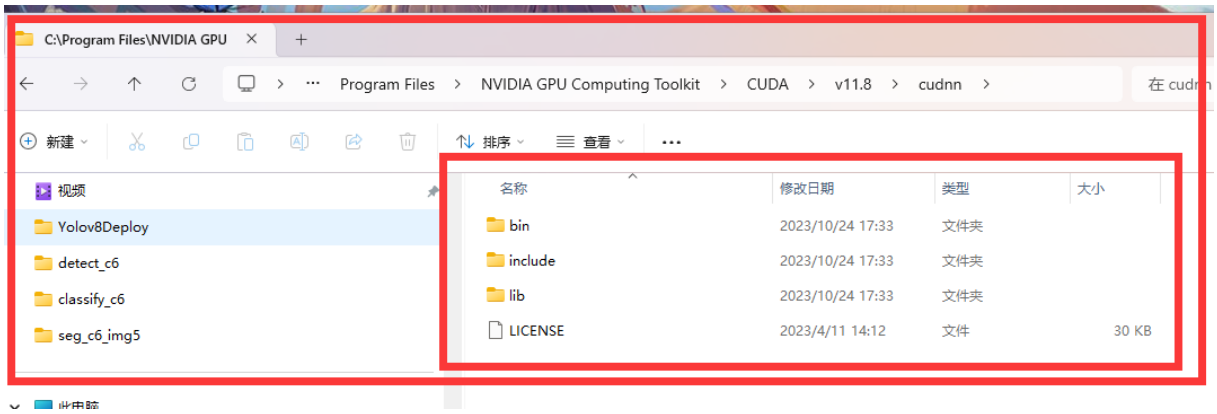


解压这个文件，并且重命名为 cudnn



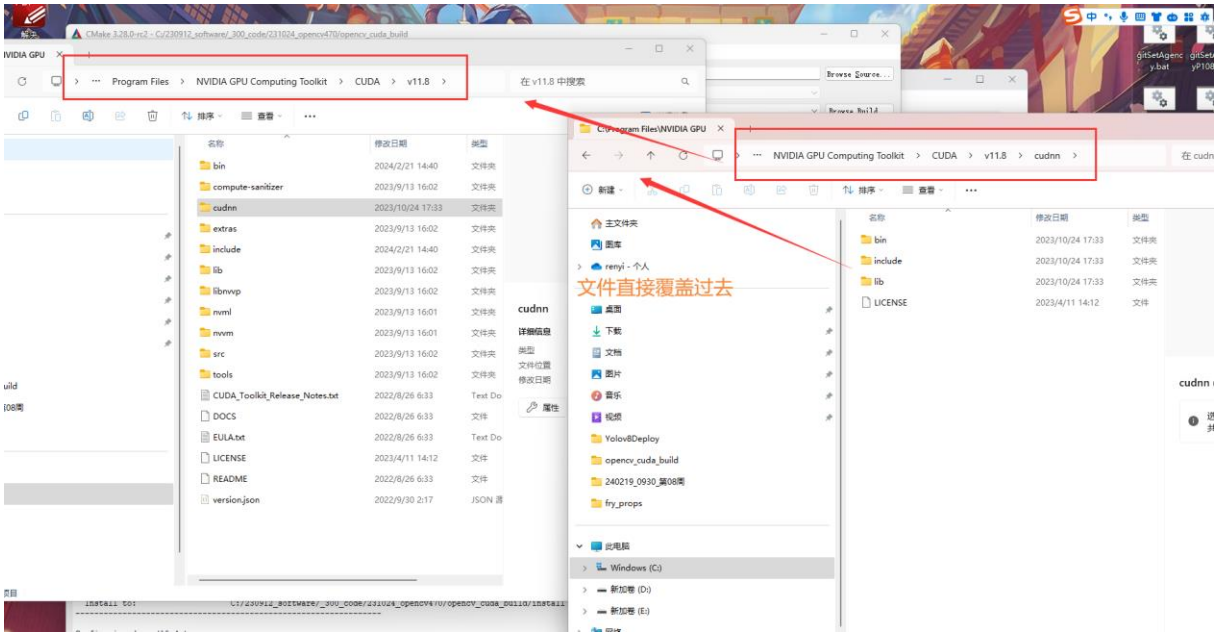
cudnn 下是这样的

注意：不要在解压的时候多增加一层目录了



特别重要：再将 cudnn 里面的内容直接覆盖到 cuda 目录

就是直接覆盖

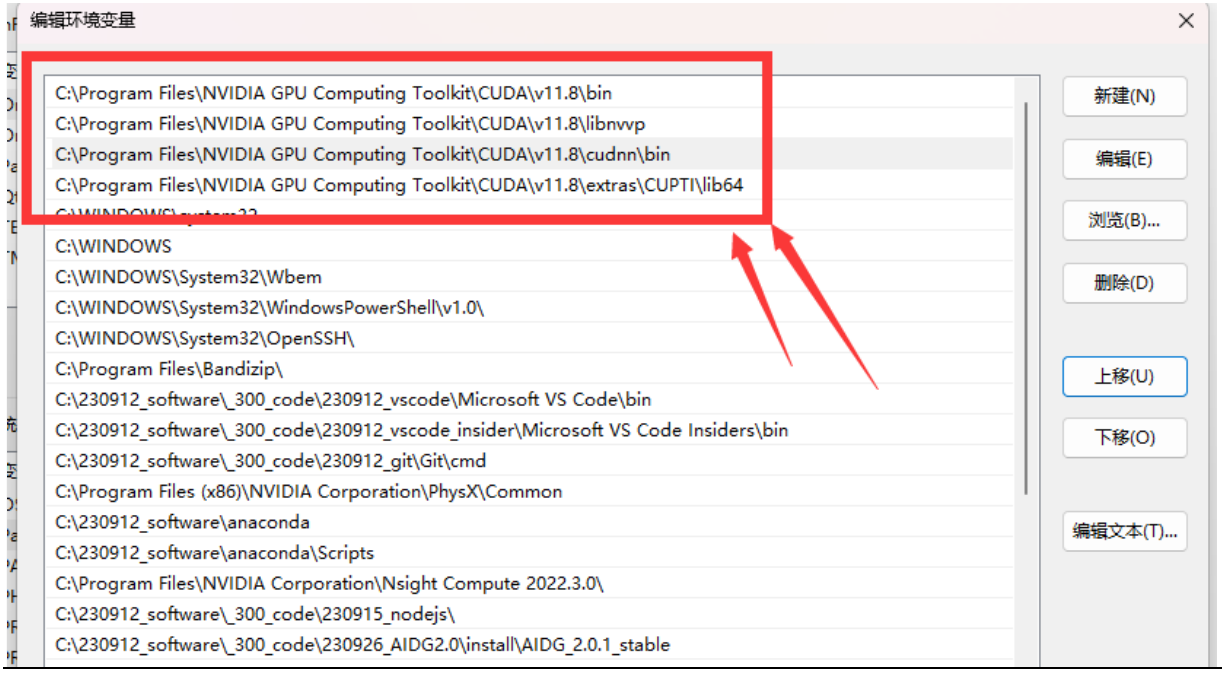


3、配置系统变量

将如下路径加入到系统变量

C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.8\bin
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.8\libnvvp
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.8\cudnn\bin

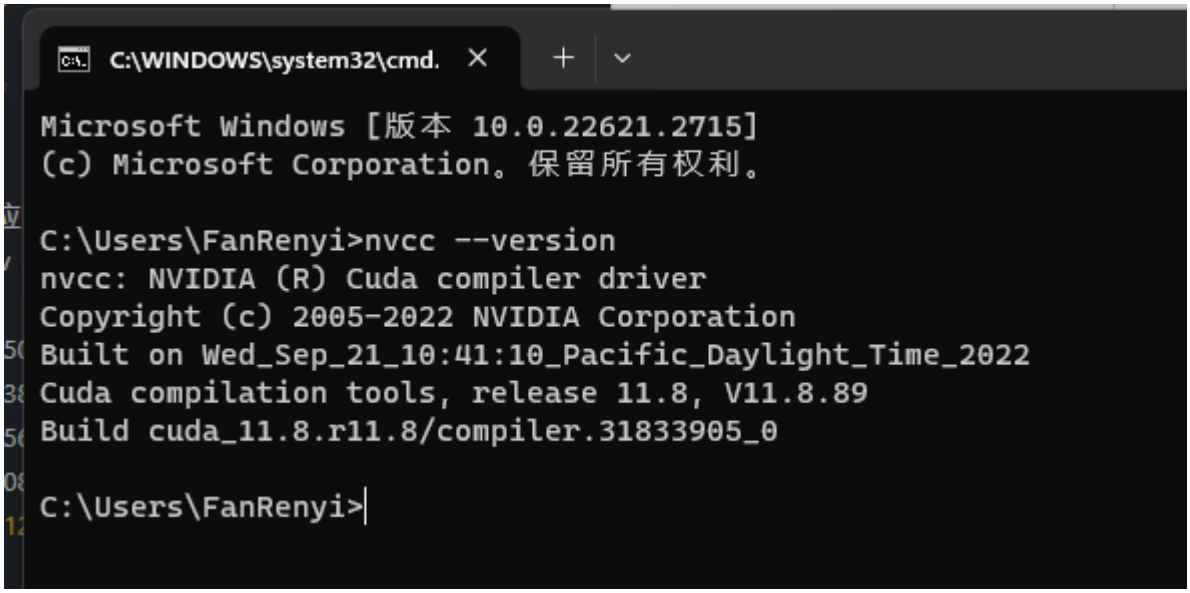
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.8\extras\CUPTI\lib64



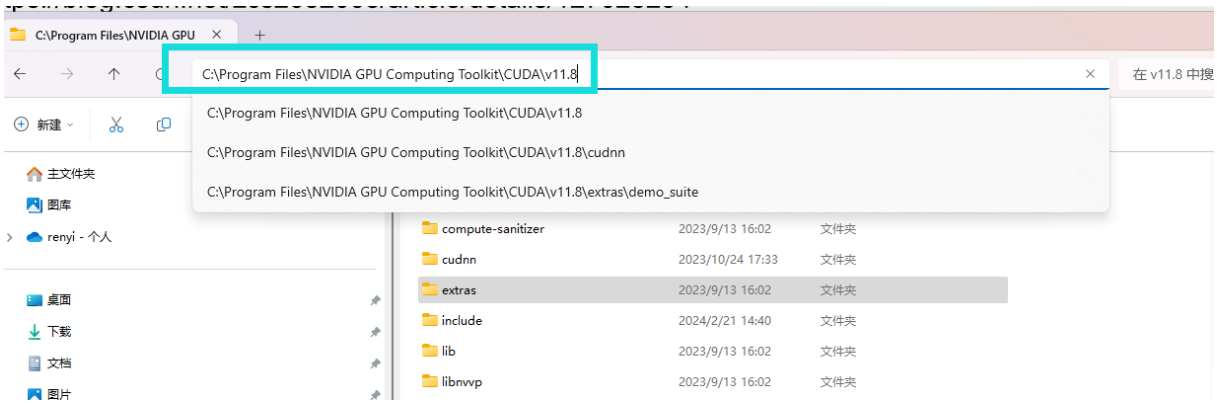
4、验证 cuda 是否安装成功

命令行输入：

```
nvcc --version
或者
nvcc -V
```



或者进入 cuda 的安装目录:



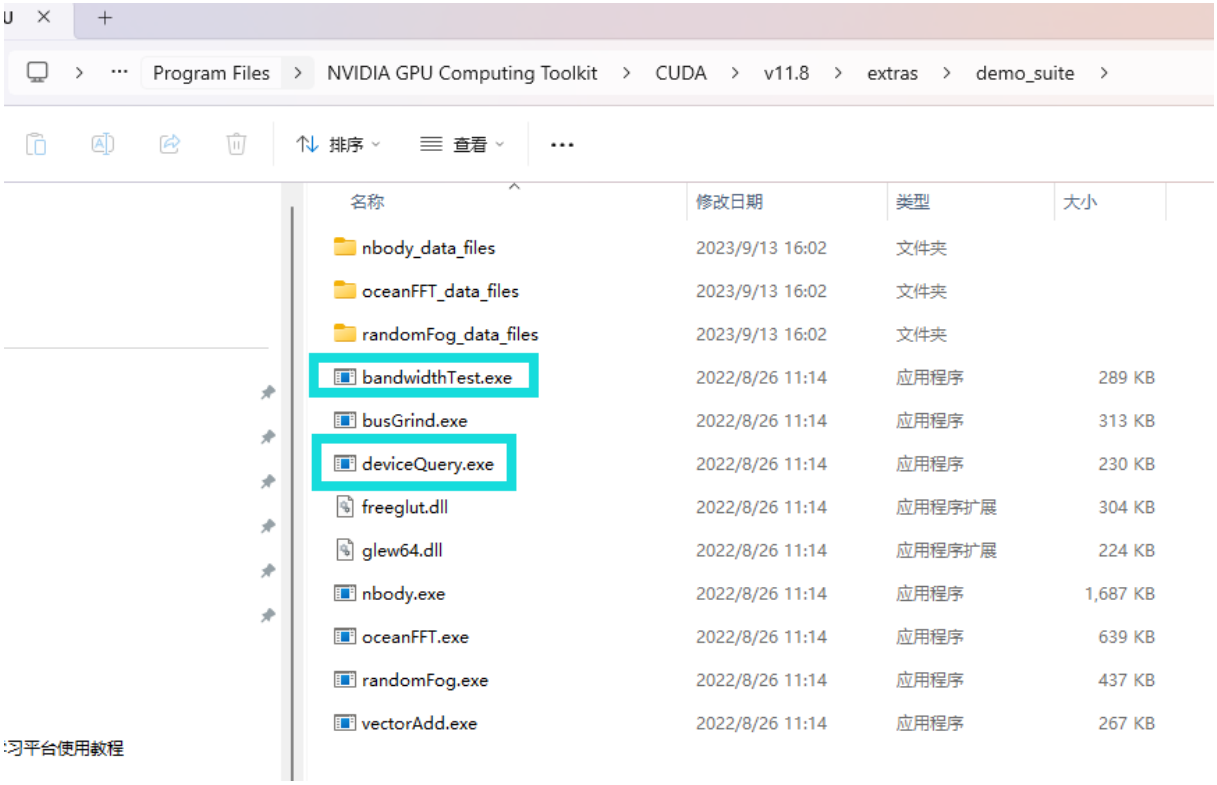
安装的默认地址都在这个目录

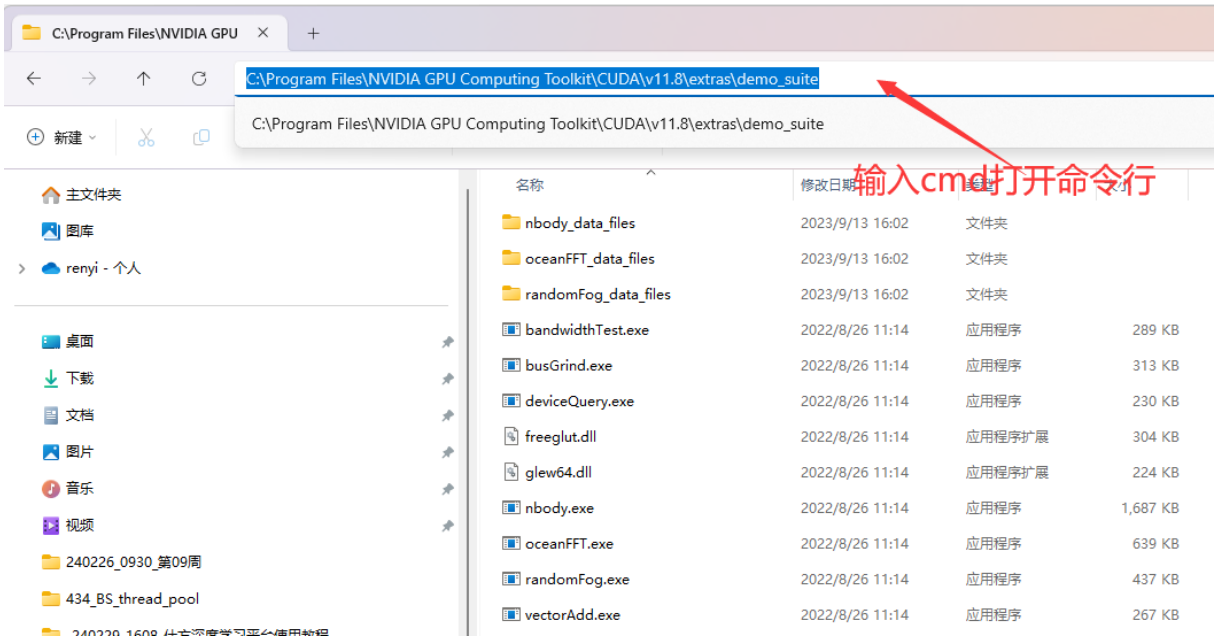
```
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.8
```

进入

```
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.8\extras\demo_suite
```

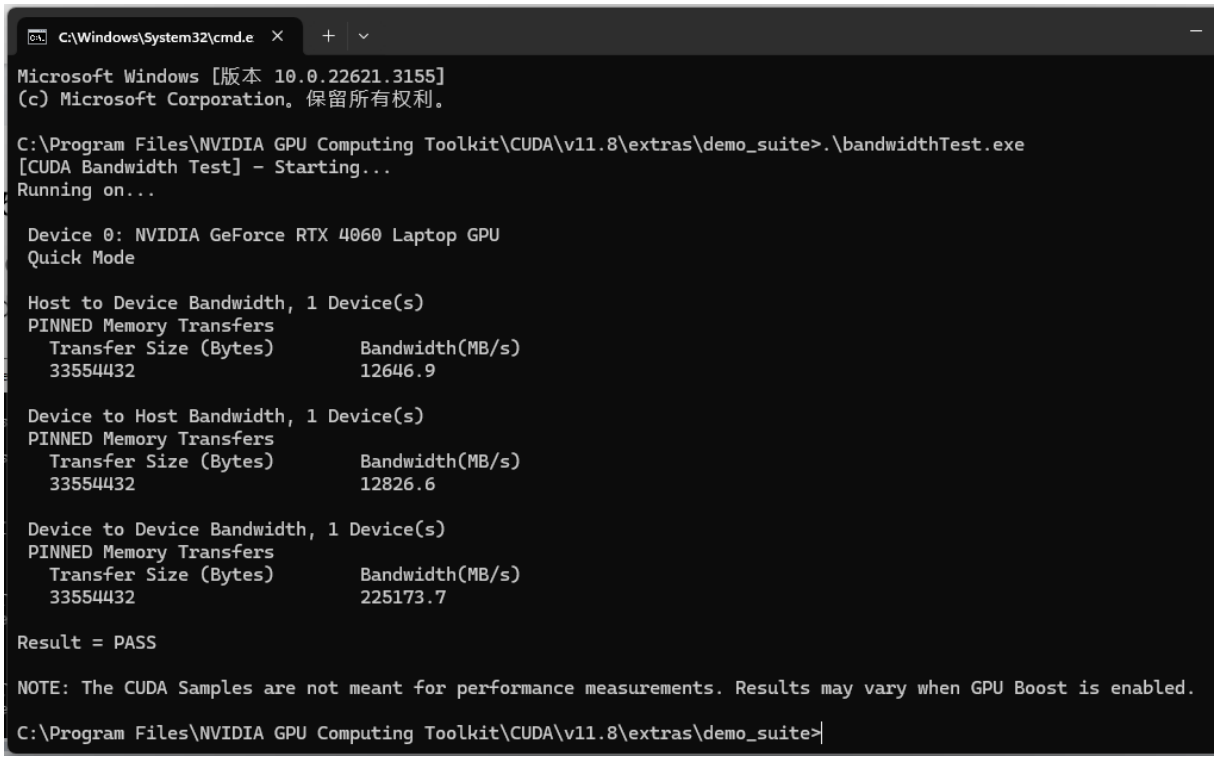
运行 bandwidthTest.exe 和 deviceQuery.exe



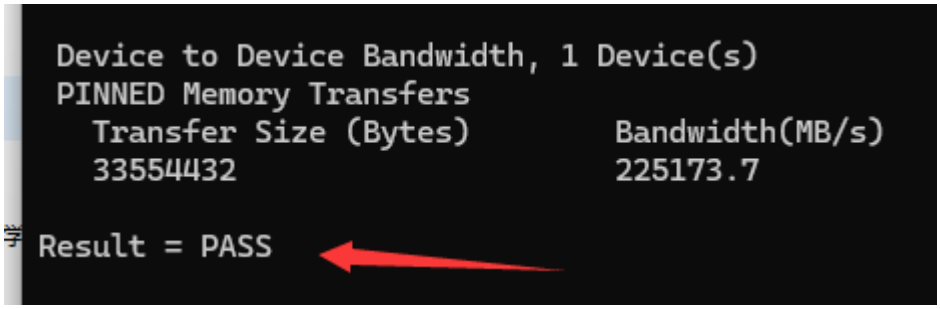


运行

```
.\bandwidthTest.exe
```



```
C:\Windows\System32\cmd.e
Microsoft Windows [版本 10.0.22621.3155]
(c) Microsoft Corporation。保留所有权利。

C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.8\extras\demo_suite>.\bandwidthTest.exe
[CUDA Bandwidth Test] - Starting...
Running on...

 Device 0: NVIDIA GeForce RTX 4060 Laptop GPU
 Quick Mode

 Host to Device Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)        Bandwidth(MB/s)
   33554432                     12646.9

 Device to Host Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)        Bandwidth(MB/s)
   33554432                     12826.6

 Device to Device Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)        Bandwidth(MB/s)
   33554432                     225173.7

Result = PASS

NOTE: The CUDA Samples are not meant for performance measurements. Results may vary when GPU Boost is enabled.

C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.8\extras\demo_suite>
```



```
Device to Device Bandwidth, 1 Device(s)
PINNED Memory Transfers
  Transfer Size (Bytes)        Bandwidth(MB/s)
  33554432                     225173.7

Result = PASS
```

运行

```
.\deviceQuery.exe
```

```
C:\Windows\System32\cmd.e

NOTE: The CUDA Samples are not meant for performance measurements. Results may vary when GPU Boost is enabled.

C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.8\extras\demo_suite>.\deviceQuery.exe
.\deviceQuery.exe Starting...

 CUDA Device Query (Runtime API) version (CUDART static linking)

Detected 1 CUDA Capable device(s)

Device 0: "NVIDIA GeForce RTX 4060 Laptop GPU"
  CUDA Driver Version / Runtime Version          12.2 / 11.8
  CUDA Capability Major/Minor version number:    8.9
  Total amount of global memory:                 8188 MBytes (8585216000 bytes)
MapSMtoCores for SM 8.9 is undefined.  Default to use 128 Cores/SM
MapSMtoCores for SM 8.9 is undefined.  Default to use 128 Cores/SM
  (24) Multiprocessors, (128) CUDA Cores/MP:     3072 CUDA Cores
  GPU Max Clock rate:                            2100 MHz (2.10 GHz)
  Memory Clock rate:                             8001 Mhz
  Memory Bus Width:                              128-bit
  L2 Cache Size:                                 33554432 bytes
  Maximum Texture Dimension Size (x,y,z)         1D=(131072), 2D=(131072, 65536), 3D=(16384, 16384, 16384)
  Maximum Layered 1D Texture Size, (num) layers  1D=(32768), 2048 layers
  Maximum Layered 2D Texture Size, (num) layers  2D=(32768, 32768), 2048 layers
  Total amount of constant memory:               zu bytes
  Total amount of shared memory per block:       zu bytes
  Total number of registers available per block: 65536
  Warp size:                                     32
  Maximum number of threads per multiprocessor:  1536
  Maximum number of threads per block:           1024
```

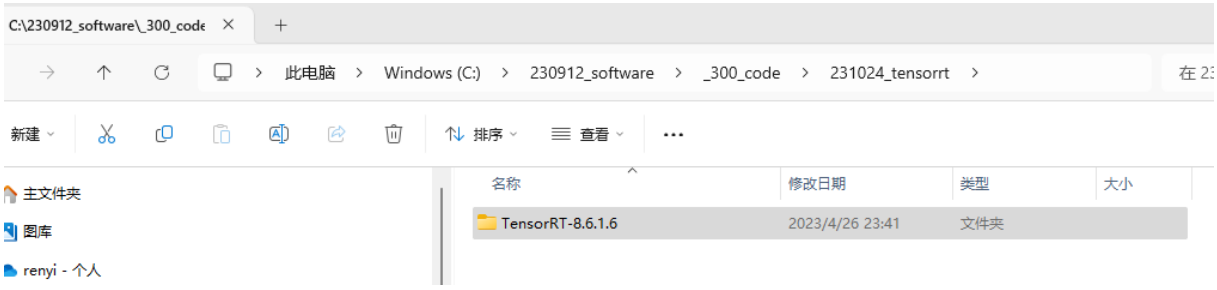
```
deviceQuery, CUDA Driver = CUDART, CUDA D
A GeForce RTX 4060 Laptop GPU
Result = PASS

C:\Program Files\NVIDIA GPU Computing Too
```

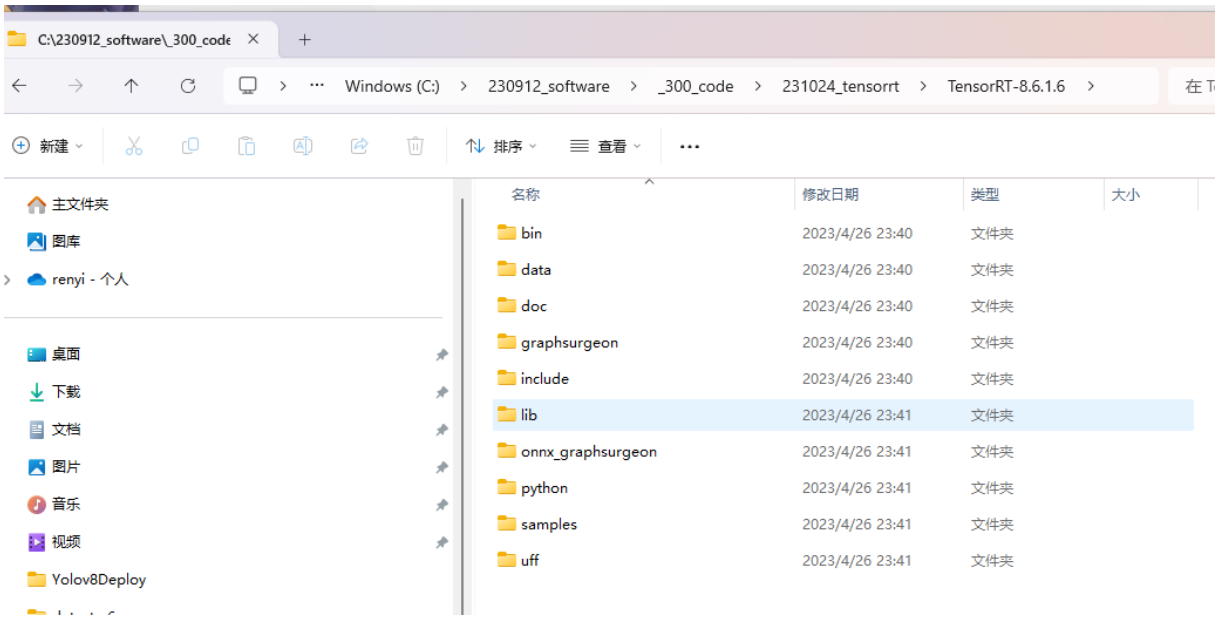
可以看到自己设备的各种信息，以及 是否成功

三、安装 tensorrt

1、解压文件

例如我放到了：

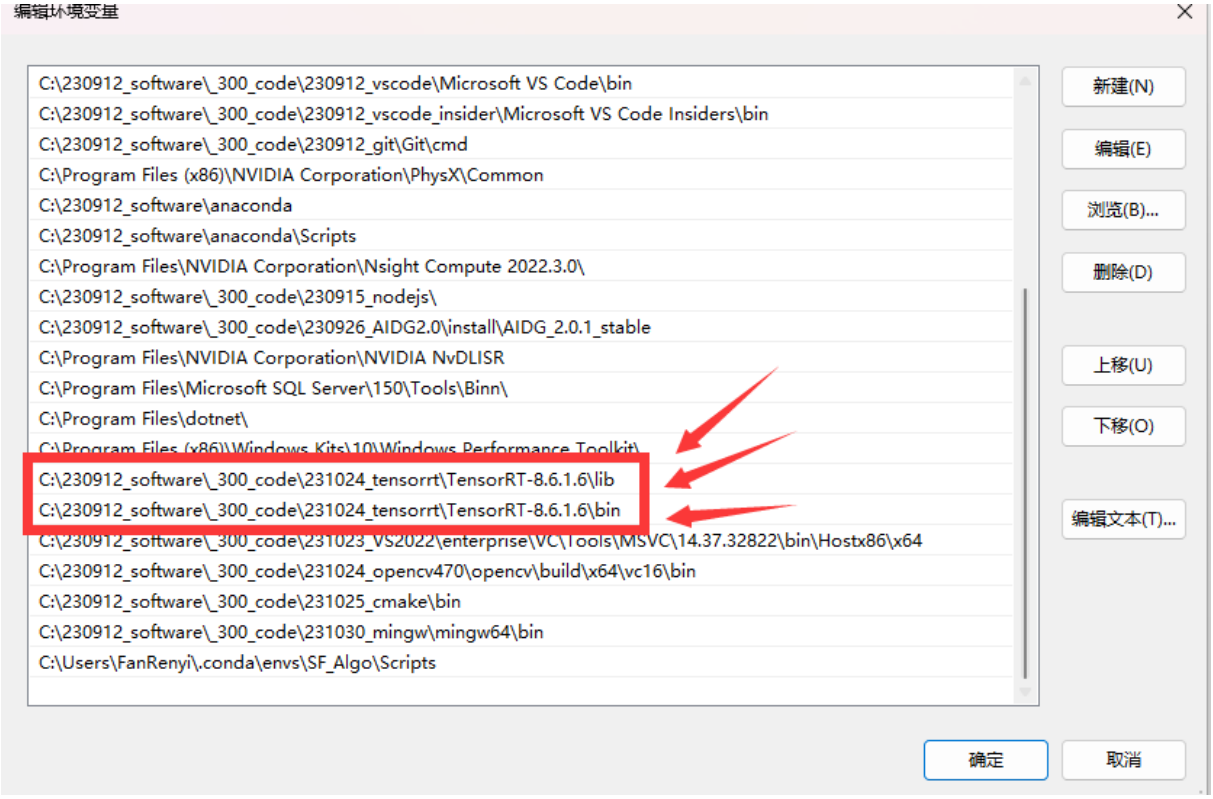




2、添加环境变量

将如下内容添加到环境变量

```
C:\230912_software\_300_code\231024_tensorrt\TensorRT-8.6.1.6\bin
C:\230912_software\_300_code\231024_tensorrt\TensorRT-8.6.1.6\lib
```



3、验证是否安装成功

命令行输入：

trtexec

```
C:\WINDOWS\system32\cmd.

Microsoft Windows [版本 10.0.22621.2715]
(c) Microsoft Corporation。保留所有权利。

C:\Users\FanRenyi>trtexec
&&&& RUNNING TensorRT.trtexec [TensorRT v8601] # trtexec
=== Model Options ===
  --uff=<file>                UFF model
  --onnx=<file>               ONNX model
  --model=<file>              Caffe model (default = no model, random weights used)
  --deploy=<file>             Caffe prototxt file
  --output=<name>[,<name>]*   Output names (it can be specified multiple times); at least one output is required for UFF
 and Caffe
  --uffInput=<name>,X,Y,Z     Input blob name and its dimensions (X,Y,Z=C,H,W), it can be specified multiple times; at l
east one is required for UFF models
  --uffNHWC                   Set if inputs are in the NHWC layout instead of NCHW (use X,Y,Z=H,W,C order in --uffInput)

=== Build Options ===
  --maxBatch                        Set max batch size and build an implicit batch engine (default = same size as --bat
ch)
                                    This option should not be used when the input model is ONNX or when dynamic shapes
are provided.
  --minShapes=spec                  Build with dynamic shapes using a profile with the min shapes provided
  --optShapes=spec                  Build with dynamic shapes using a profile with the opt shapes provided
  --maxShapes=spec                  Build with dynamic shapes using a profile with the max shapes provided
  --minShapesCalib=spec             Calibrate with dynamic shapes using a profile with the min shapes provided
  --optShapesCalib=spec             Calibrate with dynamic shapes using a profile with the opt shapes provided
  --maxShapesCalib=spec             Calibrate with dynamic shapes using a profile with the max shapes provided
                                    Note: All three of min, opt and max shapes must be supplied.
                                          However, if only opt shapes is supplied then it will be expanded so
                                          that min shapes and max shapes are set to the same values as opt shapes.
```