



Tina_Linux_NPU 部署工具安装指导 安装说明

版本号: 1.0
发布日期: 2021.07.21

版本历史

版本号	日期	制/修订人	内容描述
1.0	2021.07.21	PDC	NPU 工具安装说明。

目 录

1 前言	1
1.1 读者对象	1
1.2 约定	1
1.2.1 符号约定	1
2 正文	2
2.1 工具列表	2
2.2 下载	2
2.3 安装	2
2.3.1 安装 IDE 仿真工具	2
2.3.2 Lincese 申请	8
2.3.3 安装 License	8
2.3.4 安装模型转换工具 acuity tools	9
2.3.5 测试 acuity tools	10
3 结束	11

插图

2-1 npu_toolchain	2
2-2 npu_ext	3
2-3 npu_install_1	3
2-4 npu_install_2	4
2-5 npu_install_3	4
2-6 npu_install_3	5
2-7 npu_install_4	5
2-8 npu_install_3	6
2-9 npu_install_6	6
2-10 npu_install_3	7
2-11 npu_ide_ok	7
2-12 npu_lic_req	8
2-13 npu_install_lic	9
2-14 npu_acuity	9

1 前言

1.1 读者对象

本文档（本指南）主要适用于以下人员：

- 技术支持工程师
- 软件开发工程师
- AI 应用案客户

1.2 约定

1.2.1 符号约定

本文中可能出现的符号如下：



警告

警告



技巧

1. 技巧
2. 小常识



说明

说明

2 正文

2.1 工具列表

```
caozilong@AwExdroid65:~/lib-ai/vis_npu/VIP_Release_20211228$ tree
├── acuity-ide-toolkits
│   └── 20211228
│       ├── Verisilicon_Tool_Acuity_Toolkit_6.0.14_Binary_Whl_src_20211228.tgz
│       └── Verisilicon_Tool_VivanteIDE_v5.5.0_CL421327_Linux_Windows_SDK_6.4.x_dev_6.4.8_21Q3_CL423100B_20211228.tgz
```

图 2-1: npu_toolchain

- Verisilicon_Tool_Acuity_Toolkit_6.0.14_Binary_Whl_src_20211228.tgz: 模型部署工具，提供了命令行和 python 脚本两种界面协助客户将模型部署到芯原 NPU 上。acuity tools 做的工作包括网络导入，优化，训练，量化以及推理。
- Verisilicon_Tool_VivanteIDE_v5.5.0_CL421327_Linux_Windows_SDK_6.4.x_dev_6.4.8_21Q3_CL423100B_20211228.tgz: IDE 工具，用于 PC 侧的模型仿真验证，以及 Profile 性能分析，比如模型带宽，帧率等等。业务流上，IDE 依赖于 acuity toolkit，比如 IDE 需要 acuity toolkit 导入过程中创建的 C CODE 工程进行仿真。工具角度，acuity toolkit 依赖于 IDE 提供的一些支持库才能运行。

2.2 下载

该软件包当前没有提供外部下载地址，需要通过流程申请获取。

2.3 安装

2.3.1 安装 IDE 仿真工具

IDE 仿真工具用于离线仿真。该 IDE 工具分为 Window 版本和 Linux 版本。为方便开发，建议安装 Linux 版本。因此，需要先准备 Ubuntu 的环境（建议使用 Ubuntu 16.04LTS、18.04LTS、20.04LTS）。这两个版本都在一个压缩包中，解压 IDE 仿真工具包：

```
$ tar xvf Verisilicon_Tool_VivanteIDE_v5.5.0_CL421327_Linux_Windows_SDK_6.4.x_dev_6.4.8_21Q3_CL423100B_20211228.tgz
```

```
caozilong@AwExdroid65:~/lib-ai/vis_npu/VIP_Release_20211228/aciuty-ide-toolkits/20211228$ tree
tree
├── doc
│   └── README
│       └── Vivante_IDE_User_Guide.pdf
├── Verisilicon_Tool_Aciuty_Toolkit_6.0.14_Binary_Whl_src_20211228.tgz
├── Verisilicon_Tool_VivanteIDE_v5.5.0_CL421327_Linux_Windows_SDK_6.4.x_dev_6.4.8_21Q3_CL423100B_20211228.tgz
├── Vivante_IDE-5.5.0_CL421327-Linux-x86_64-12-27-2021-15.31.40-plus-W-p6.4.x_dev_6.4.8_21Q3_CL423100B-Install
├── Vivante_IDE-5.5.0_CL421327-Win32-x86_64-12-27-2021-15.26.51-plus-W-p6.4.x_dev_6.4.8_21Q3_CL423100B-Setup.exe
└── 1 directory, 6 files
caozilong@AwExdroid65:~/lib-ai/vis_npu/VIP_Release_20211228/aciuty-ide-toolkits/20211228$
```

图 2-2: npu_ext

先解压，解压后可以看到 Linux 版本的安装文件：

Vivante_IDE-5.5.0_CL421327-Linux-x86_64-12-27-2021-15.31.40-plus-W-p6.4.x_dev_6.4.8_21Q3_CL423100B-Install

执行该文件开始安装，会弹出安装向导，选择“Yes”。

```
$ ./Vivante_IDE-5.5.0_CL421327-Linux-x86_64-12-27-2021-15.31.40-plus-W-p6.4.x_dev_6.4.8_21Q3_CL423100B-Install
```

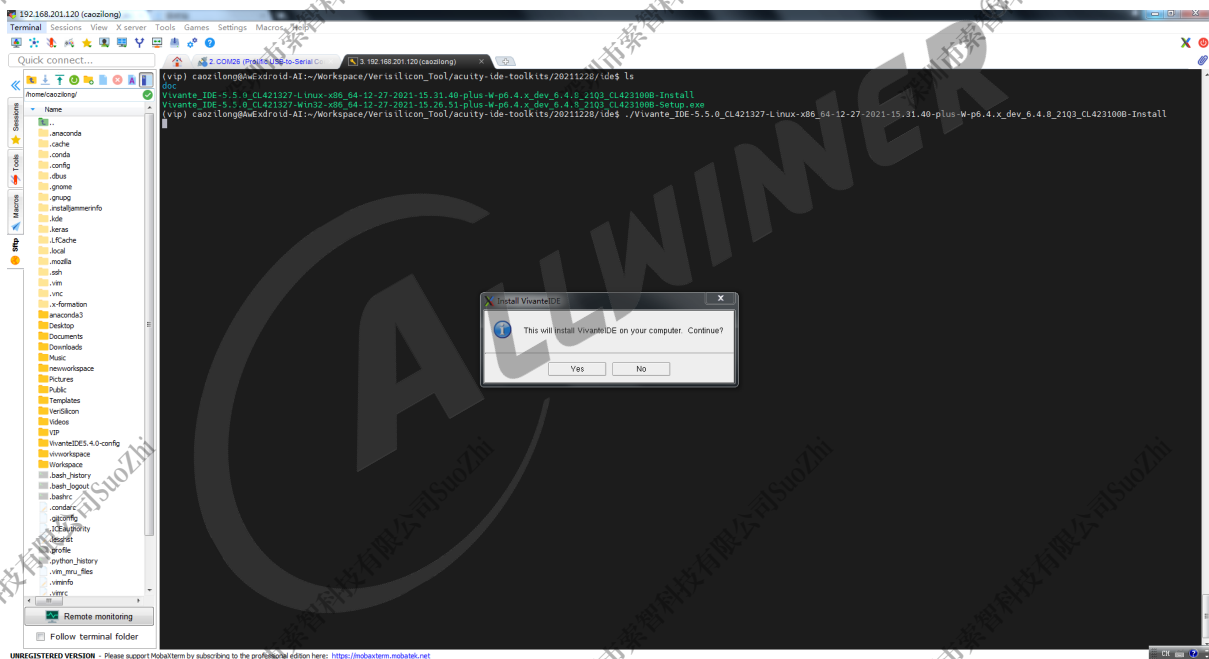


图 2-3: npu_install_1

根据安装向导，逐步安装。

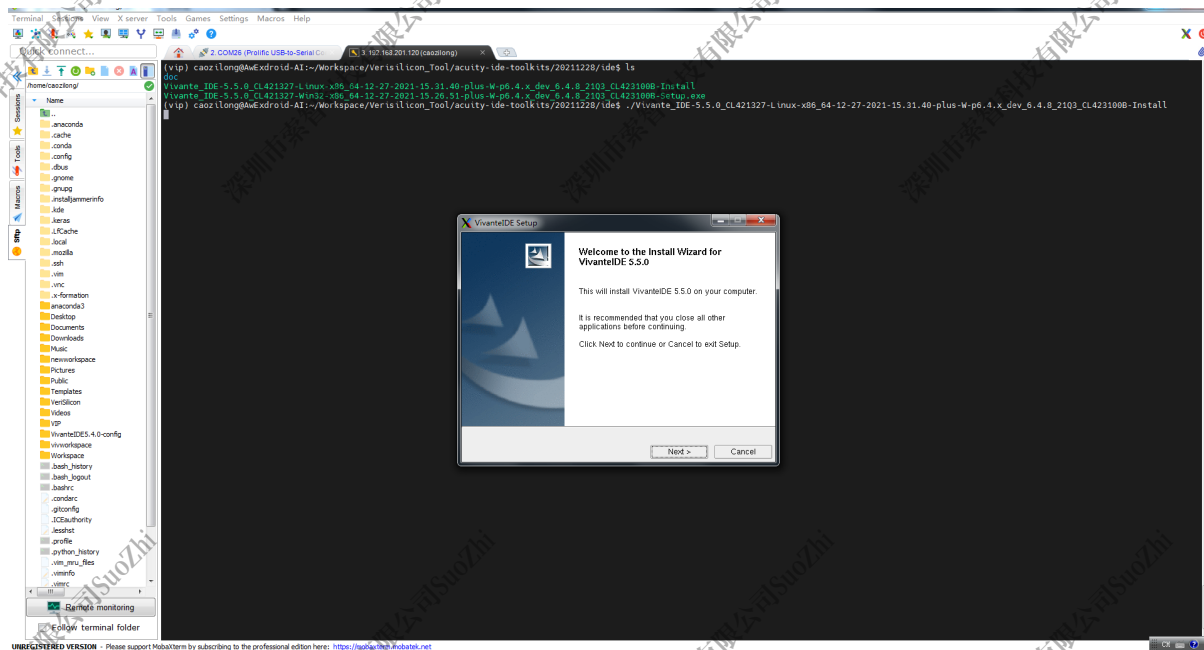


图 2-4: npu_install_2

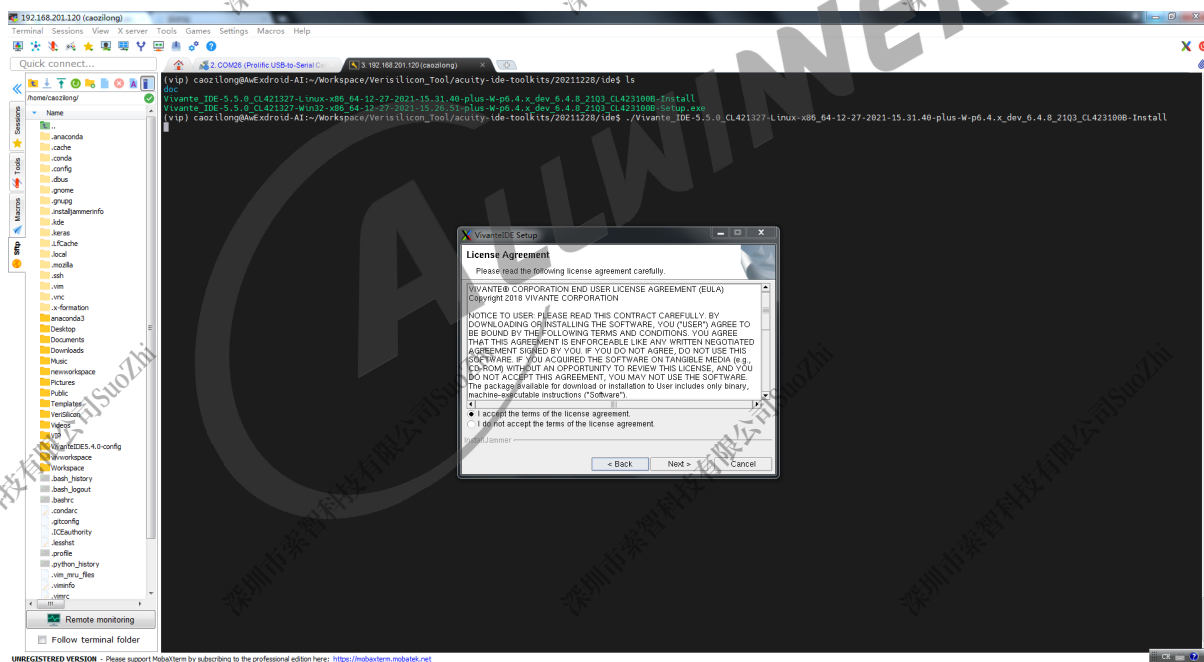


图 2-5: npu_install_3

自己指定安装的目录。

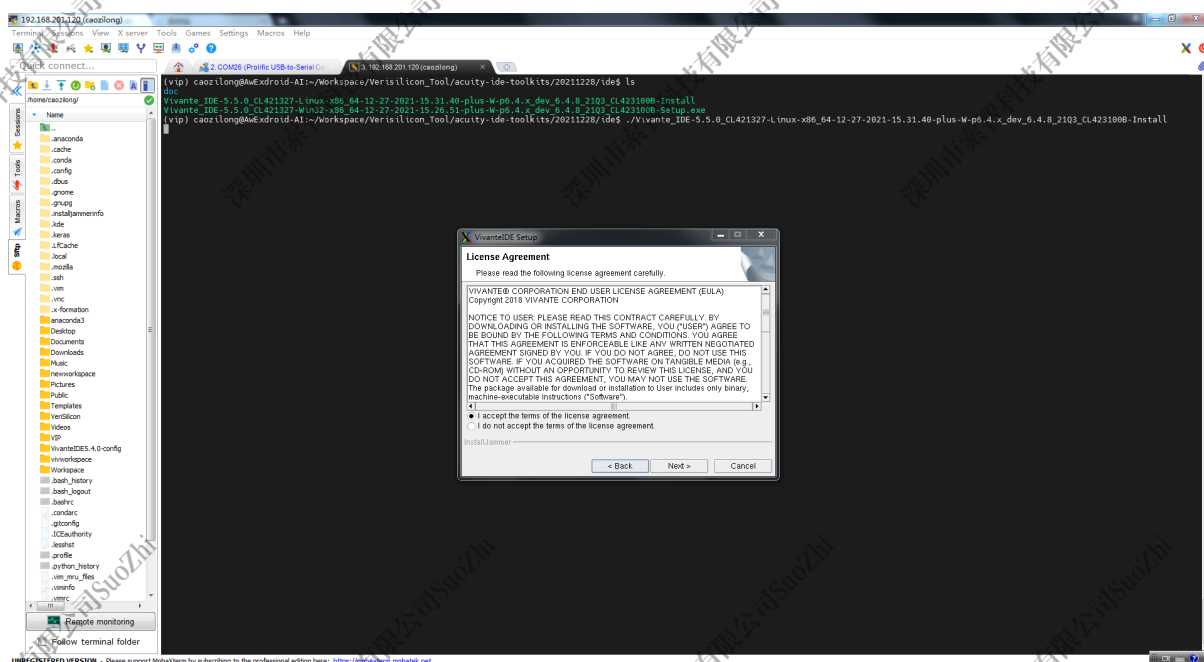


图 2-6: npu_install_3

License file 暂时可不指定。正式使用时，需要申请 License。

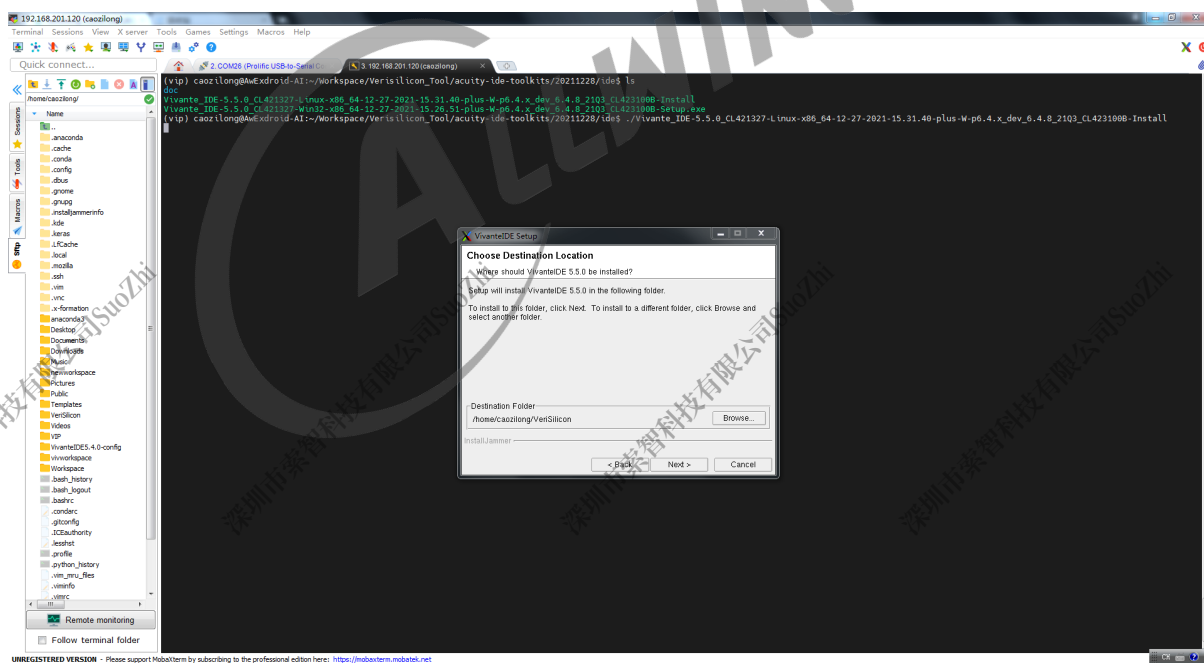


图 2-7: npu_install_4

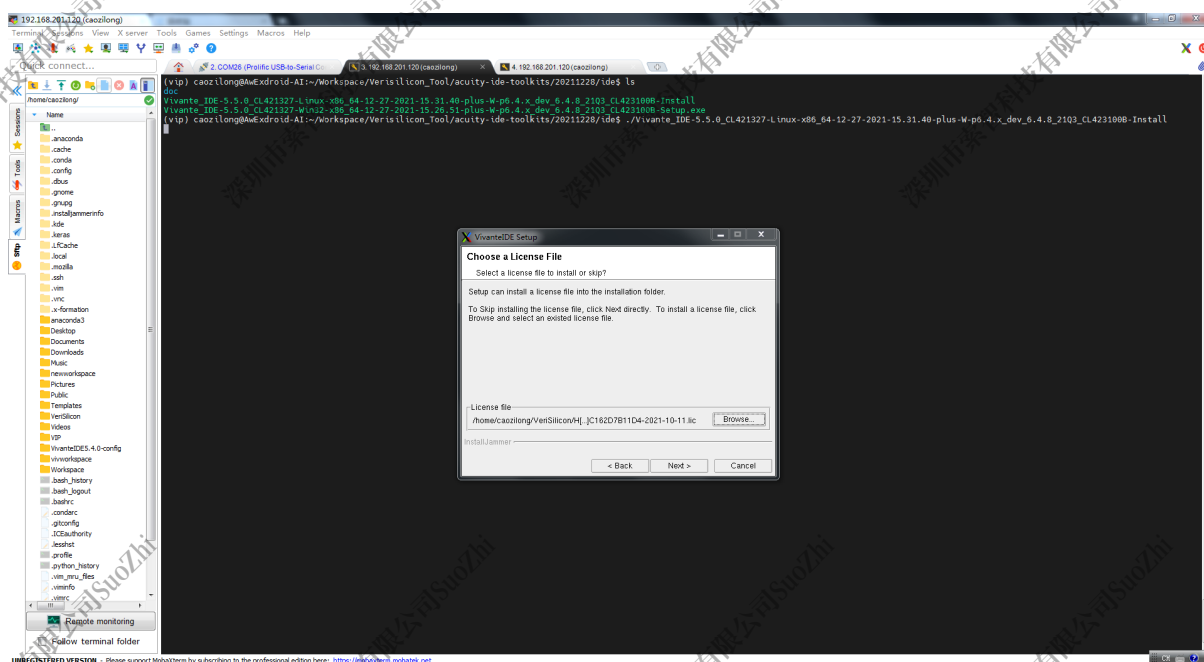


图 2-8: npu_install_3

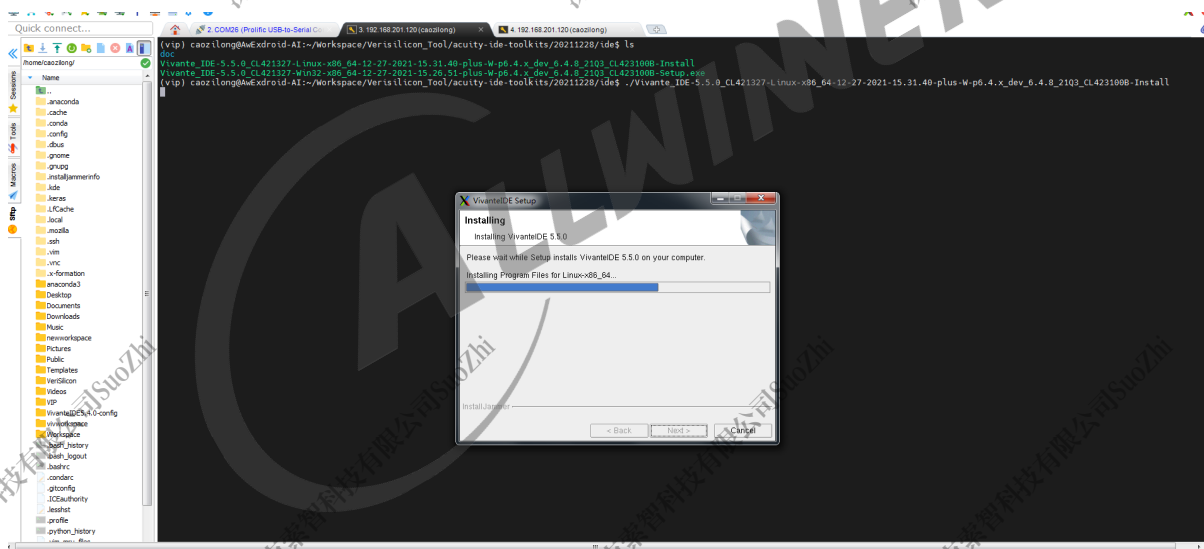


图 2-9: npu_install_6

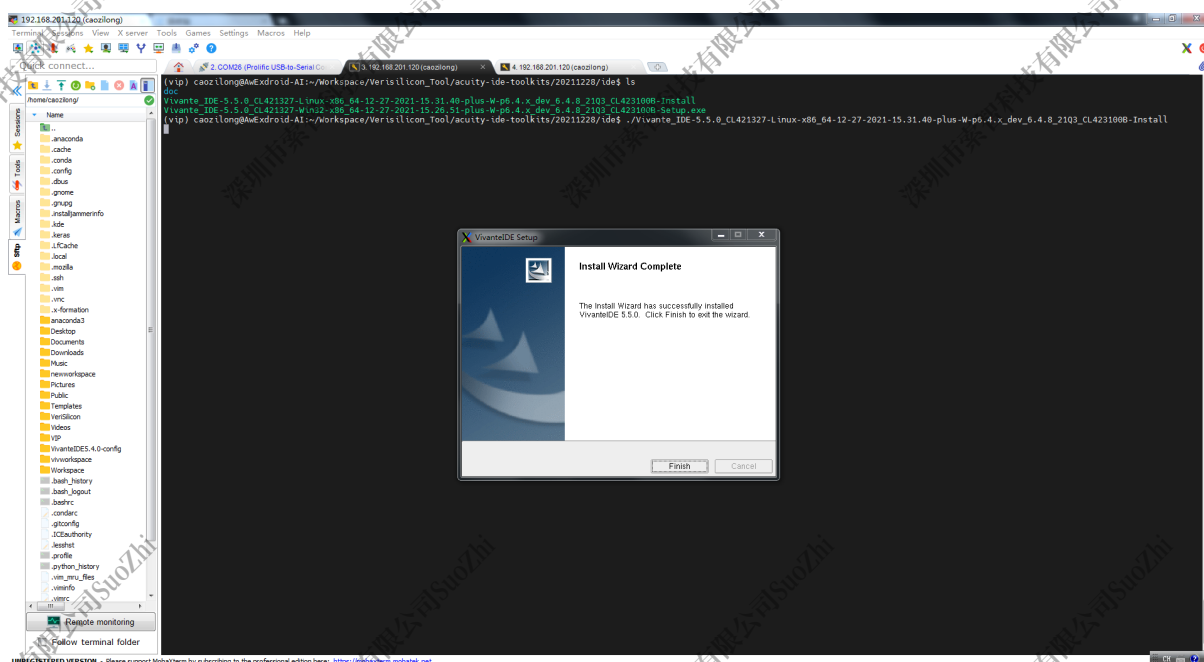


图 2-10: npu_install_3

安装完成后，在控制台中，执行`path/to/VivanteIDE5.4.0/ide/vivanteide5.4.0`打开 IDE 软件，界面如下：

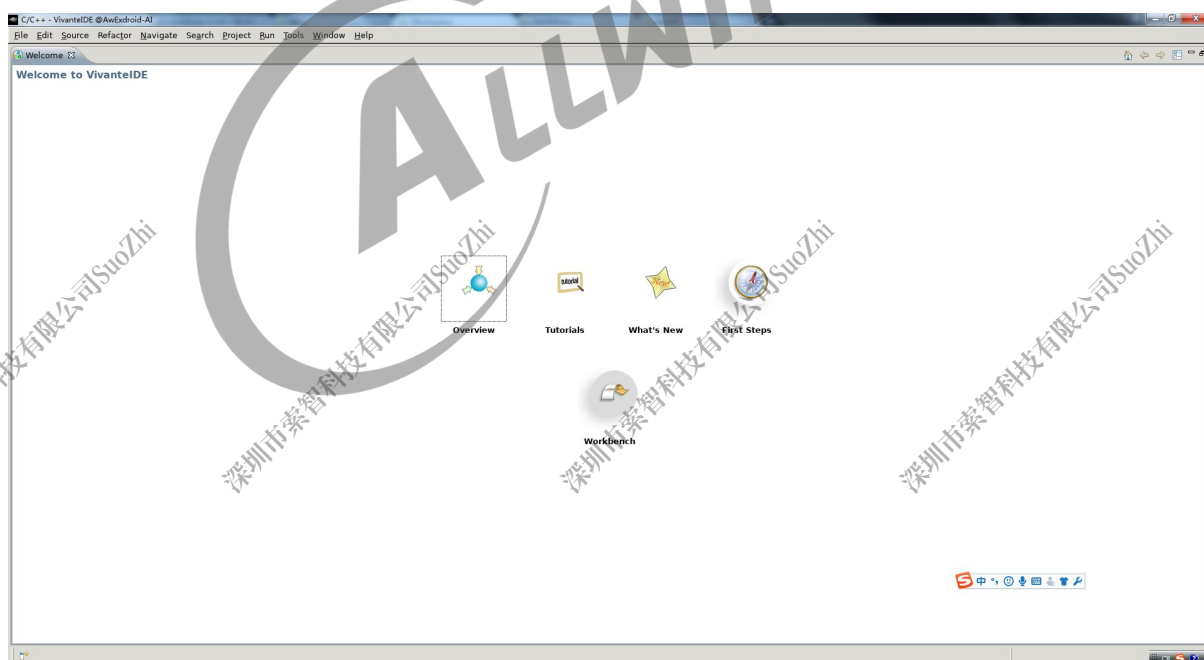


图 2-11: npu_ide_ok

关于 IDE 的使用，可以参考文档 `VivanteIDE_User_Guide.pdf`。

2.3.2 Lincese 申请

仿真工具 IDE 需要 Lincese 才能使用全部的功能 (acuity tools 不需要安装 license)，到芯原官网，填写必要的信息，申请一个可用的 License，合法的 License 会通过邮件发送到你的邮箱，根据之前的经验，周期是非常短的。

License 申请地址：<https://www.verisilicon.com/cn/VIPAcuityIDELicenseRequest>

申请页面说明：



图 2-12: npu_lic_req

填写相关信息，包括服务器中的 MAC 地址，等待邮件发送 HostID-*.lic 文件，需要注意的是，license 貌似只和你这里所填写的信息有关，和工具版本没多大关系，这次升级工具使用的 license 仍然是 5.4.0 IDE 所使用的同一个 license 文件。

2.3.3 安装 License

在安装过程中，遇到安装 License 页时，可以先直接跳过这一步，等安装过程结束后，在 IDE 环境下通过 Help->Install License 入口安装 Lincese 文件。

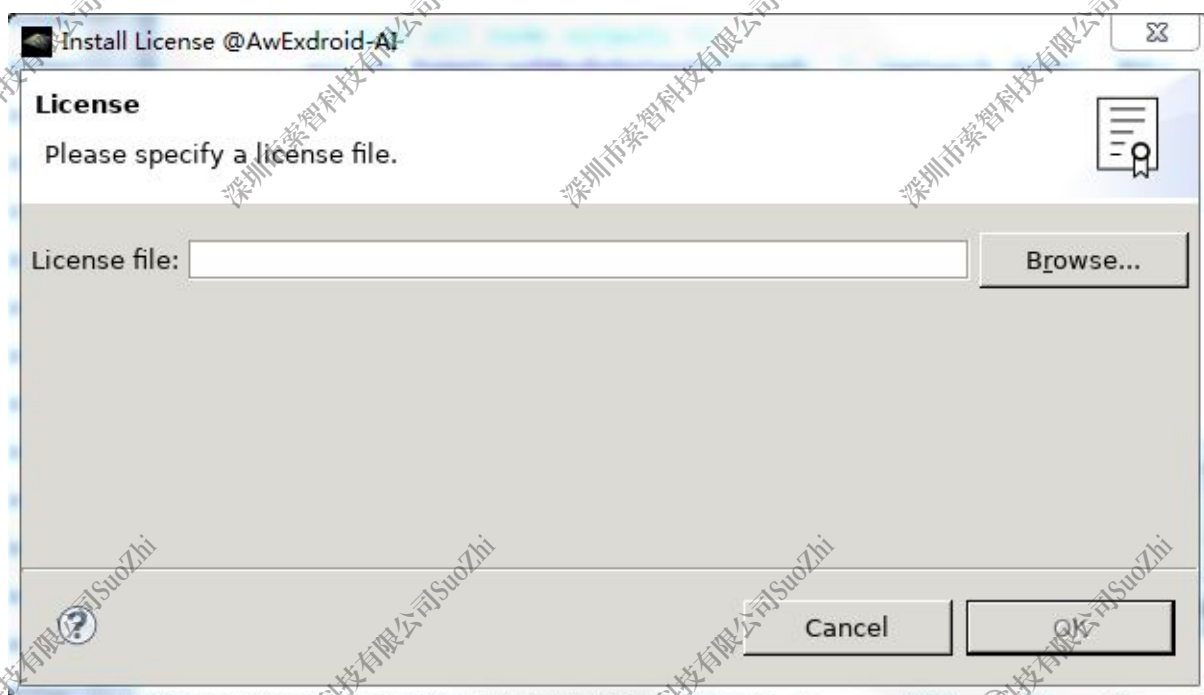


图 2-13: npu_install_lic

2.3.4 安装模型转换工具 acuity tools

先解压模型转换工具压缩包, 解压后, 得到各个版本的模型转换工具包。

```
$ tar xvf Verisilicon_Tool_Acuity_Toolkit_5.21.1_Binary_Whl_src_20210331.tgz
```

解压后, 选择 Vivante_acuity_toolkit_binary_6.0.14_20211227_ubuntu18.04.tgz 继续解压

```
$ tar xvf Vivante_acuity_toolkit_binary_6.0.14_20211227_ubuntu18.04.tgz -C /path/to/destination
```

根据需要, 当前使用的是 binary 版本, 所以选择 Vivante_acuity_toolkit_binary_6.0.14_20211227_ubuntu18.04.tgz。

关于模型转换工具的使用可以参考文档 Vivante.VIP.ACUTY.Toolkit.User.Guide-v0.92-20210922.pdf。

```
caozilong@AwExdroid65:~/lib-ai/vis_npu/VIP_Release_20211228/acuity-ide-toolkits/20211228$ tar xvf Verisilicon_Tool_Acuity_Toolkit_6.0.14_Binary_Whl_src_20211228.tgz
./
./Vivante_acuity_toolkit_whl_6.0.14_20211227_ubuntu18.04.tgz
./Vivante_acuity_toolkit_whl_6.0.14_20211227_ubuntu16.04.tgz
./Vivante_acuity_toolkit_src_6.0.14_20211227_ubuntu18.04.tgz
./Vivante_acuity_toolkit_src_6.0.14_20211227_ubuntu16.04.tgz
./Vivante_acuity_toolkit_binary_6.0.14_20211227_ubuntu18.04.tgz
./Vivante_acuity_toolkit_binary_6.0.14_20211227_ubuntu16.04.tgz
./doc/
./doc/Vivante.VIP.ACUTY.Toolkit.User.Guide-v0.92-20210922.pdf
./doc/Vivante.App.Note.ACUTY.Operation.Mapping.and.Support-V1.12-20211214.pdf
./Acuity_examples_c94812.tgz
caozilong@AwExdroid65:~/lib-ai/vis_npu/VIP_Release_20211228/acuity-ide-toolkits/20211228$
caozilong@AwExdroid65:~/lib-ai/vis_npu/VIP_Release_20211228/acuity-ide-toolkits/20211228$
caozilong@AwExdroid65:~/lib-ai/vis_npu/VIP_Release_20211228/acuity-ide-toolkits/20211228$
```

图 2-14: npu_acuity

解压后, 得到 acuity-toolkit-binary-6.0.14, 为方便后面配置环境变量, 可以将该文件夹放到与 Verisilicon IDE 同级目录。

之后，编辑.bashrc，增加以下两行，导出环境变量（具体路径要依据你的安装目录）：

```
export ACUITY_PATH=/path/to/VeriSilicon/$ACUITY_TOOLS_METHOD/bin/  
export VIV_SDK=/path/to/VeriSilicon/VivanteIDE5.5.0/cmdtools/
```

之后在控制台端直接执行 `source ~/.bashrc`，安装完成。

2.3.5 测试 acuity tools

在控制台中，执行 `pegasus help` 出现以下打印即可：

```
(vip) caozilong@Axiadroid-AI:~/VeriSilicon$ pegasus help  
2022-02-15 13:04:27.950157: W tensorflow/stream_executor/platform/default/dso_loader.cc:59] Could not load dynamic library 'libcudart.so.10.1'; dlderror: libcudart.so.10.1: cannot open shared object file: No  
such file or directory; LD_LIBRARY_PATH: /home/caozilong/VeriSilicon/acuity-toolkit-binary-6.0.14/bin/acuitylib  
2022-02-15 13:04:27.950210: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if you do not have a GPU set up on your machine.  
usage: pegasus.py <command> [-args...]  
  
There are common pegasus commands used in various situations:  
  
Import models.(import)  
  caffe          Import caffe model.  
  tensorflow     Import tensorflow model.  
  tf lite        Import tf lite model.  
  darknet        Import darknet model.  
  onnx           Import onnx model.  
  pytorch        Import pytorch model.  
  keras          Import keras model.  
  
Export models.(export)  
  ovxlib         Export ovxlib code.  
  ide            Export ide code.  
  tf lite        Export tf lite model.  
  
Generate metas.(generate)  
  inputmeta      Generate input meta data.  
  postprocess-file Generate postprocess file.  
  fakedata       Generate fake data of coefficients.  
  
prune models.(prune)  
  --model        Network model file.  
  --model-data   Network coefficient file.  
  --output-data  Network coefficient file after pruning.  
If not specified, data_input file will be overwritten  
  --config-file  Prune config file containing layer_name and prune percentage.  
If file does not exist, a stub will be generated  
  --prune-percent Prune percentage of each layer, from 0.0 to 100.0  
  --prune-level   Specify the pruning granularity levels [element | vector | kernel | filter]  
  -element: pruning granularity down to individual weight element (1)  
  -vector: a vector or row of a 2D convolution kernel (Kx)  
  -kernel: 2D convolution kernel (Kx * Ky)  
  -filter: 2D convolution filter (Kx * Ky * Kz)  
  
Inference model and get result.(inference)  
  --model        Network model input file.  
  --model-data   Network coefficient input file.  
  --model-quantize Quantized tensor description file.  
  --batch-size   Batch size.  
  --iterations   Running iterations.  
  --device       Specify the compute device.  
  --with-input-meta Merge input meta into network.  
  --output-dir   Output directory of generated files.  
  --dtype        Data type used.  
  --postprocess  Postprocess task.  
  --postprocess-file Postprocess task configure file.
```

3 结束



著作权声明

版权所有 © 2022 珠海全志科技股份有限公司。保留一切权利。

本文档及内容受著作权法保护，其著作权由珠海全志科技股份有限公司（“全志”）拥有并保留一切权利。

本文档是全志的原创作品和版权财产，未经全志书面许可，任何单位和个人不得擅自摘抄、复制、修改、发表或传播本文档内容的部分或全部，且不得以任何形式传播。

商标声明

、 **全志科技** （不完全列举）均为珠海全志科技股份有限公司的商标或者注册商标。在本文档描述的产品中出现的其它商标，产品名称，和服务名称，均由其各自所有人拥有。

免责声明

您购买的产品、服务或特性应受您与珠海全志科技股份有限公司（“全志”）之间签署的商业合同和条款的约束。本文档中描述的全部或部分产品、服务或特性可能不在您所购买或使用的范围内。使用前请认真阅读合同条款和相关说明，并严格遵循本文档的使用说明。您将自行承担任何不当使用行为（包括但不限于如超压，超频，超温使用）造成的不利后果，全志概不负责。

本文档作为使用指导仅供参考。由于产品版本升级或其他原因，本文档内容有可能修改，如有变更，恕不另行通知。全志尽全力在本文档中提供准确的信息，但并不确保内容完全没有错误，因使用本文档而发生损害（包括但不限于间接的、偶然的、特殊的损失）或发生侵犯第三方权利事件，全志概不负责。本文档中的所有陈述、信息和建议并不构成任何明示或暗示的保证或承诺。

本文档未以明示或暗示或其他方式授予全志的任何专利或知识产权。在您实施方案或使用产品的过程中，可能需要获得第三方的权利许可。请您自行向第三方权利人获取相关的许可。全志不承担也不代为支付任何关于获取第三方许可的许可费或版税（专利税）。全志不对您所使用的第三方许可技术做出任何保证、赔偿或承担其他义务。