

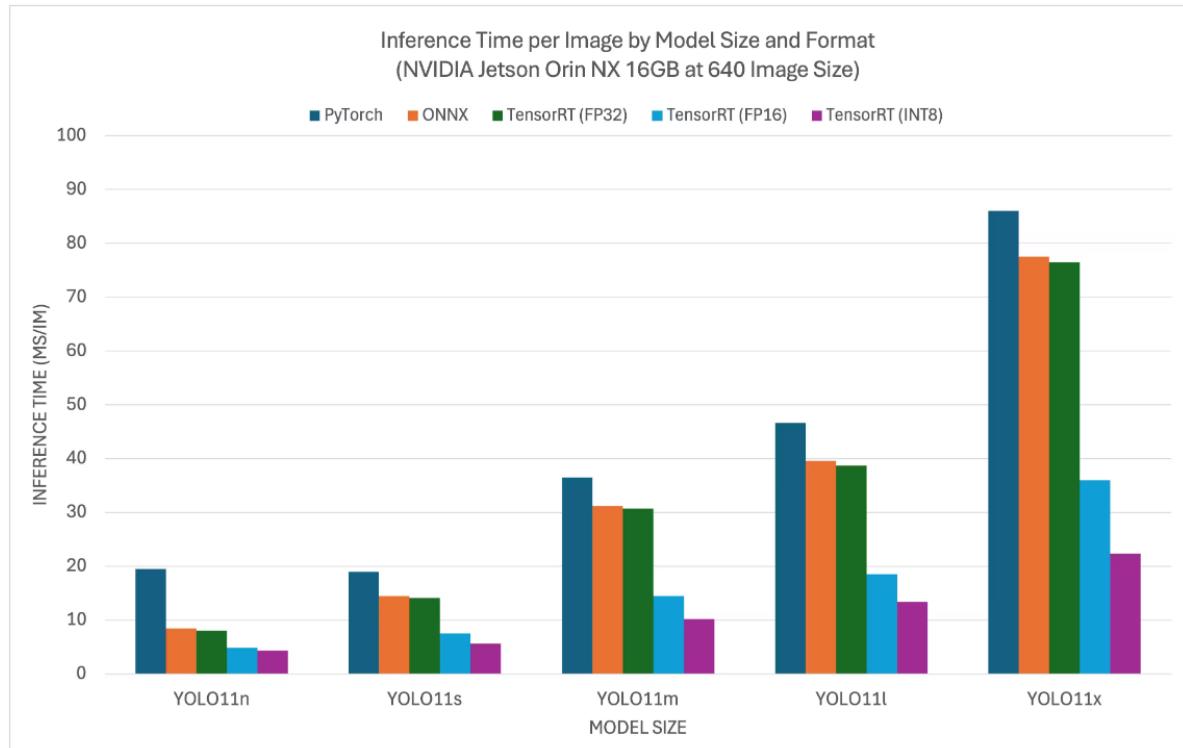
# Model Conversion

## Model Conversion

1. Jetson Orin YOLO11 (Benchmark)
2. Enable Maximum Board Performance
  - 2.1. Enable MAX Power Mode
  - 2.2. Enable Jetson Clocks
3. Model Conversion
  - 3.1. CLI: pt → onnx → engine
  - 3.2. Python: pt → onnx → engine
4. Model Prediction  
CLI Usage  
Common Issues  
ERROR: onnxslim  
References

## 1. Jetson Orin YOLO11 (Benchmark)

YOLO11 benchmark data is sourced from the Ultralytics team, testing various model formats (data for reference only)



## 2. Enable Maximum Board Performance

## 2.1. Enable MAX Power Mode

Enabling MAX Power Mode on Jetson will ensure that all CPU and GPU cores are turned on:

```
#Orin Nano  
sudo nvpmode1 -m 2  
#Orin Nx  
sudo nvpmode1 -m 0
```

## 2.2. Enable Jetson Clocks

Enabling Jetson Clocks will ensure that all CPU and GPU cores run at maximum frequency:

```
sudo jetson_clocks
```

## 3. Model Conversion

According to the test parameters provided by the Ultralytics team for different format models, we can find that using TensorRT provides the best inference performance!

```
First time using YOLO11 export mode will automatically install some dependencies,  
please wait for the automatic completion!
```

### 3.1. CLI: pt → onnx → engine

Convert PyTorch format model to TensorRT: The conversion process will automatically generate ONNX model

```
cd /home/jetson/ultralytics/ultralytics
```

```
yolo export model=yolo11n.pt format=engine  
# yolo export model=yolo11n-seg.pt format=engine  
# yolo export model=yolo11n-pose.pt format=engine  
# yolo export model=yolo11n-cls.pt format=engine  
# yolo export model=yolo11n-obb.pt format=engine
```

```

Activities Files Dec 31 12:25
jetson@yahboom: ~/ultralytics/ultralytics
jetson@yahboom:~/ultralytics$ cd /home/jetson/ultralytics/ultralytics
jetson@yahboom:~/ultralytics$ yolo export model=yolo11n.pt format=engine
WARNING TensorRT requires GPU export, automatically assigning device=0
Ultralytics 8.3.55 Python-3.10.12 torch-2.5.0a0+872d972e41.nv24.08 CUDA:0 (Orin, 7620MiB)
YOLO11n summary (fused): 238 layers, 2,616,248 parameters, 0 gradients, 6.5 GFLOPs
PyTorch: starting from 'yolo11n.pt' with input shape (1, 3, 640, 640) BCHW and output shape(s) (1, 84, 840
0) (5.4 MB)

ONNX: starting export with onnx 1.17.0 opset 19...
ONNX: slimming with onnxslim 0.1.45...
ONNX: export success ✓ 2.zs, saved as 'yolo11n.onnx' (10.2 MB)

TensorRT: starting export with TensorRT 10.3.0...
[12/31/2024-12:21:22] [TRT] [I] [MemUsageChange] Init CUDA:0
[12/31/2024-12:21:24] [TRT] [I] [MemUsageChange] Init builder: 1657, GPU 5196 (MiB)
[12/31/2024-12:21:24] [TRT] [I] -----
[12/31/2024-12:21:24] [TRT] [I] Input filename: yolo11n.onnx
[12/31/2024-12:21:24] [TRT] [I] ONNX IR version: 0.0.9
[12/31/2024-12:21:24] [TRT] [I] Opset version: 19
[12/31/2024-12:21:24] [TRT] [I] Producer name: pytorch
[12/31/2024-12:21:24] [TRT] [I] Producer version: 2.5.0
[12/31/2024-12:21:24] [TRT] [I] Domain:
[12/31/2024-12:21:24] [TRT] [I] Model version: 0
[12/31/2024-12:21:24] [TRT] [I] Doc string:
[12/31/2024-12:21:24] [TRT] [I] -----
TensorRT: input "images" with shape(1, 3, 640, 640) DataType: kFloat32
TensorRT: output "output0" with shape(1, 84, 8400) DataType: kFloat32
TensorRT: building FP32 engine as yolo11n.engine
[12/31/2024-12:21:25] [TRT] [I] Local timing cache in use. Results will be stored.
[12/31/2024-12:24:33] [TRT] [I] Detected 1 inputs and 1 outputs
[12/31/2024-12:24:35] [TRT] [I] Total Host Persistent Memory: 2764864 bytes
[12/31/2024-12:24:35] [TRT] [I] Total Device Persistent Memory: 2764864 bytes
[12/31/2024-12:24:35] [TRT] [I] Total Scratch Memory: 2764864 bytes
[12/31/2024-12:24:35] [TRT] [I] [BlockAssignment] Started assignment to complete.
[12/31/2024-12:24:35] [TRT] [I] [BlockAssignment] Algorithm took 216 nodes requiring 19252224 bytes.

Recent
Starred
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
Trash
Other Locations
assets
cfg
data
engine
hub
models
nn
output
runs
solutions
trackers
utils
videos
yahboom_demo
__init__.py
yolo11n.engine
yolo11n.onnx
yolo11n.pt
yolo11n-cls.pt
yolo11n-obb.pt
yolo11n-pose.pt
yolo11n-seg.pt

```

### 3.2. Python: pt → onnx → engine

Convert PyTorch format model to TensorRT: The conversion process will automatically generate ONNX model

```
cd /home/jetson/ultralytics/ultralytics/yahboom_demo
```

```
python3 model_pt_onnx_engine.py
```

```

from ultralytics import YOLO

# Load a YOLOv11n PyTorch model
# model = YOLO("/home/jetson/ultralytics/ultralytics/yolo11n.pt")
model = YOLO("/home/jetson/ultralytics/ultralytics/yolo11n-seg.pt")
# model = YOLO("/home/jetson/ultralytics/ultralytics/yolo11n-pose.pt")
# model = YOLO("/home/jetson/ultralytics/ultralytics/yolo11n-cls.pt")
# model = YOLO("/home/jetson/ultralytics/ultralytics/yolo11n-obb.pt")

# Export the model to TensorRT
model.export(format="engine")

```

Note: The converted model files are located at the same location as the original model files

```

Activities Files Dec 31 13:05 MAXN en NVIDIA Jetson Community Pro...
jetson@yahboom:~$ cd /home/jetson/ultralytics/ultralytics/yahboom_demo
jetson@yahboom:~/ultralytics/ultralytics/yahboom_demo$ python3 model_pt_onnx_engine.py
WARNING:TensorRT requires GPU export, automatically assigning device=0
Ultralytics 8.3.55 Python-3.10.12 torch-2.5.0a0+872d972e41.nv24.08 CUDA:0 (Orin, 7620MiB)
YOLOv8n-seg summary (fused): 265 layers, 2,868,664 parameters, 8 gradients, 10.4 GFLOPS

PyTorch: starting from '/home/jetson/ultralytics/ultralytics/yolo1n-seg.pt' with input shape (1, 3, 640, 640) BCHW and output shape(s) ((1, 116, 8400), (1, 32, 160, 160)) (5.9 MB)

ONNX: starting export with onnx 1.17.0 opset 19...
ONNX: slimming with onnxslim 0.1.45...
ONNX: export success ✓ 2.0s, saved as '/home/jetson/ultralytics/ultralytics/yolo1n-seg.onnx' (11.2 MB)

TensorRT: starting export with TensorRT 10.3.0...
[12/31/2024-13:01:15] [TRT] [I] [MemUsageChange] Init CUDA: 719 (MiB)
[12/31/2024-13:01:17] [TRT] [I] [MemUsageChange] Init builder 3, now: CPU 1743, GPU 5314 (MiB)
[12/31/2024-13:01:17] [TRT] [I] -----
[12/31/2024-13:01:17] [TRT] [I] Input filename: /home/jetson/ultralytics/yolo1n-seg.onnx
[12/31/2024-13:01:17] [TRT] [I] ONNX IR version: 0.0.9
[12/31/2024-13:01:17] [TRT] [I] Opset version: 19
[12/31/2024-13:01:17] [TRT] [I] Producer name: pytorch
[12/31/2024-13:01:17] [TRT] [I] Producer version: 2.5.0
[12/31/2024-13:01:17] [TRT] [I] Domain:
[12/31/2024-13:01:17] [TRT] [I] Model version: 0
[12/31/2024-13:01:17] [TRT] [I] Doc string:
[12/31/2024-13:01:17] [TRT] [I] -----
TensorRT: input "images" with shape(1, 3, 640, 640) DataType: TensorType::FLOAT
TensorRT: output "output0" with shape(1, 116, 8400) DataType: TensorType::FLOAT
TensorRT: output "output1" with shape(1, 32, 160, 160) DataType: TensorType::FLOAT
TensorRT: building FP32 engine as /home/jetson/ultralytics/yolo1n-seg.engine
[12/31/2024-13:01:18] [TRT] [I] Local timing cache in use. Results will not be stored.

```

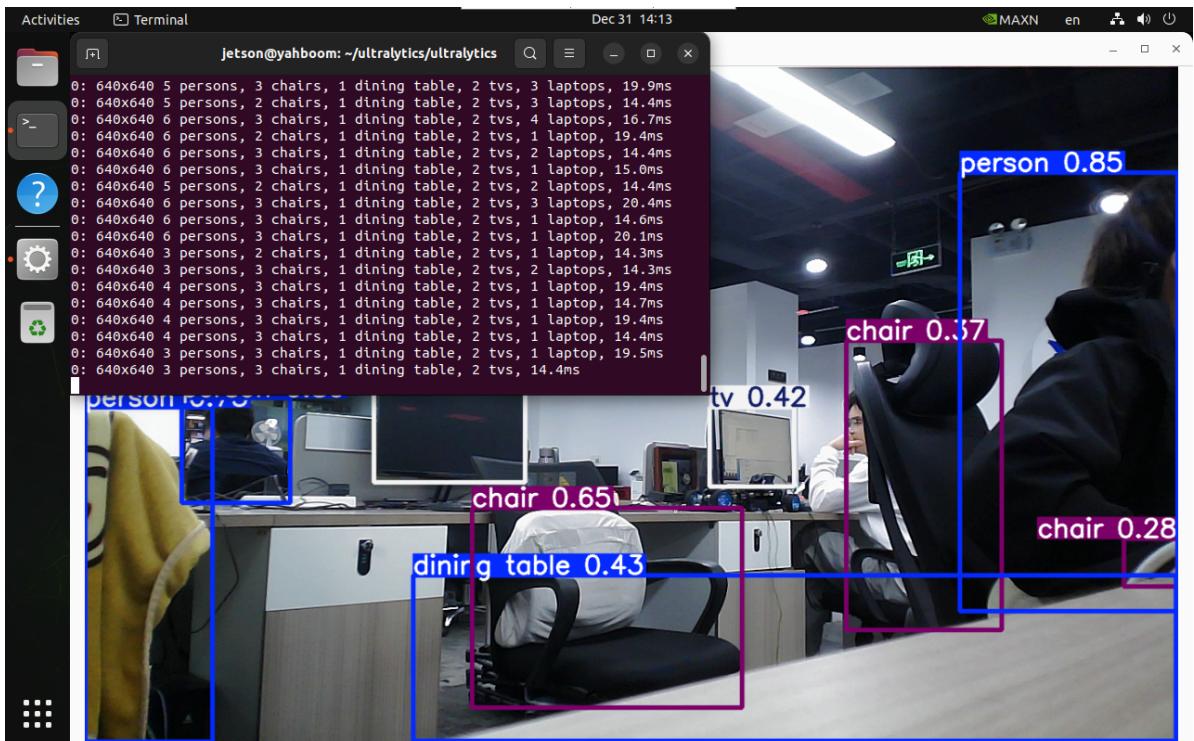
## 4. Model Prediction

### CLI Usage

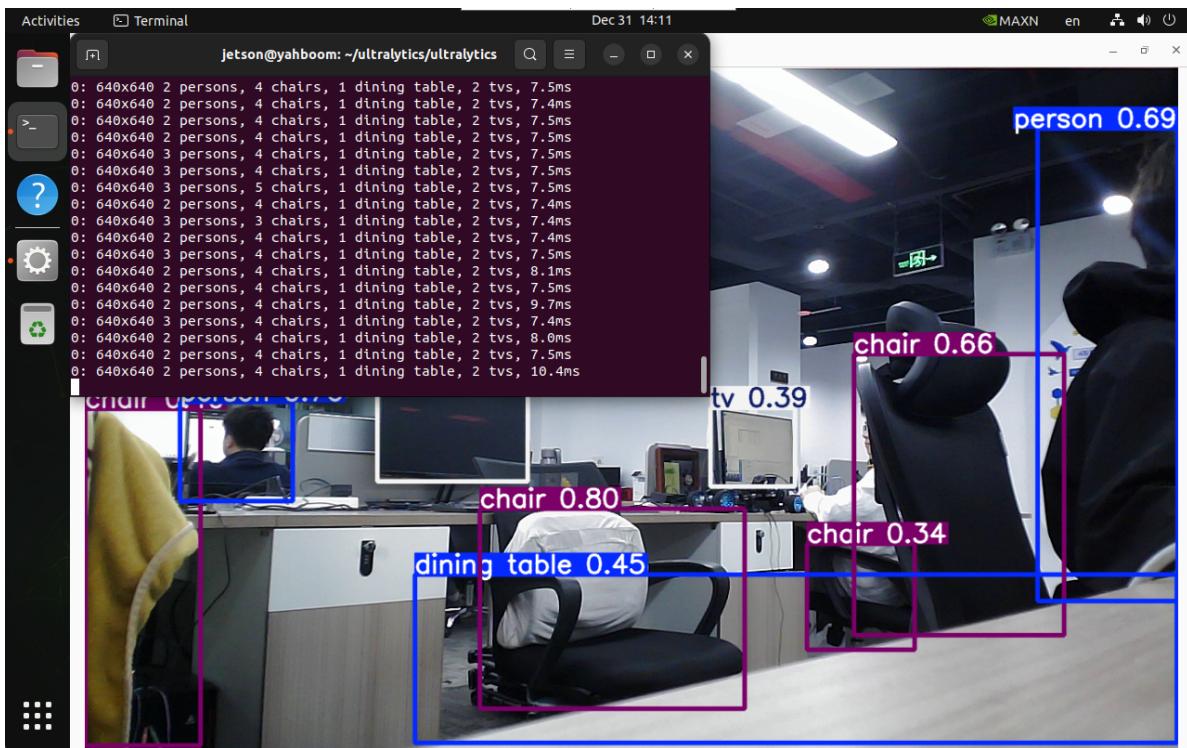
CLI currently only supports USB cameras. For CSI cameras, users can directly modify the previous Python code to call ONNX and ENGINE models!

```
cd /home/jetson/ultralytics/ultralytics
```

```
yolo predict model=yolo1n.onnx source=0 save=False show
```

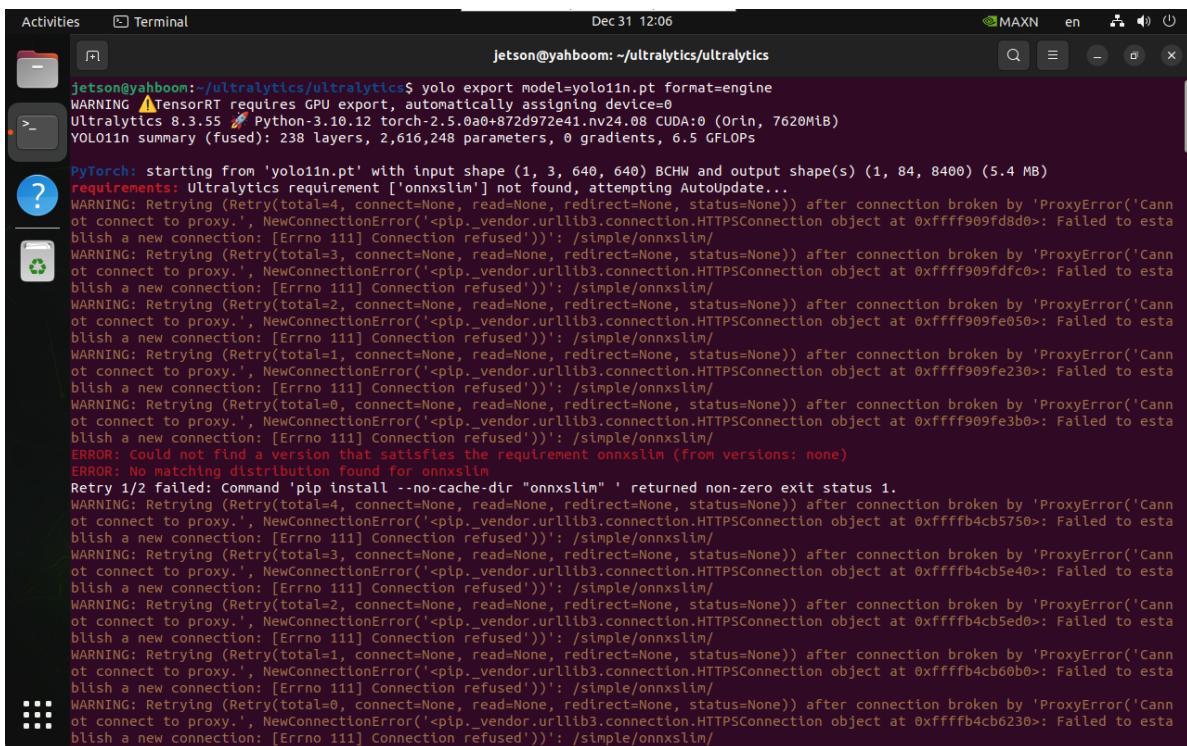


```
yolo predict model=yolo1n.engine source=0 save=False show
```



## Common Issues

### ERROR: onnxslim



Solution: Enter onnxslim installation command in terminal

```
sudo pip3 install onnxslim
```

Activities Terminal Dec 31 12:07 jetson@yahboom: ~/ultralytics/ultralytics

```
jetson@yahboom:~/ultralytics$ sudo pip3 install onnxslim
Collecting onnxslim
  Downloading onnxslim-0.1.45-py3-none-any.whl.metadata (4.2 kB)
Requirement already satisfied: onnx in /usr/local/lib/python3.10/dist-packages (from onnxslim) (1.17.0)
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages (from onnxslim) (1.13.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from onnxslim) (23.2)
Requirement already satisfied: protobuf==3.20.2 in /usr/local/lib/python3.10/dist-packages (from onnx->onnxslim) (4.25.5)
Requirement already satisfied: numpy<1.4,>=1.1.0 in /usr/local/lib/python3.10/dist-packages (from onnx->onnxslim) (1.23.0)
Requirement already satisfied: nmpmath<1.4,>=1.1.0 in /usr/local/lib/python3.10/dist-packages (from sympy->onnxslim) (1.3.0)
Downloading onnxslim-0.1.45-py3-none-any.whl (142 kB)
Installing collected packages: onnxslim
Successfully installed onnxslim-0.1.45
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager,
possibly rendering your system unusable. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv. Use the --root-user-action option if you know what you are doing and want to suppress this warning.
jetson@yahboom:~/ultralytics$ yolo export model=yolo1n-cls.pt format=engine
WARNING TensorRT requires GPU export, automatically assigning device=0
Ultralytics 8.3.55 🚀 Python-3.10.12 torch-2.5.0a0+872d972e41.nv24.08 CUDA:0 (Orin, 7620MiB)
YOLO1n-cls summary (fused): 112 layers, 2,807,024 parameters, 0 gradients, 4.2 GFLOPs

PyTorch: starting from 'yolo1n-cls.pt' with input shape (1, 3, 224, 224) BCHW and output shape(s) (1, 1000) (5.5 MB)

ONNX: starting export with onnx 1.17.0 opset 19...
ONNX: slimming with onnxslim 0.1.45...
ONNX: export success ✓ 0.9s, saved as 'yolo1n-cls.onnx' (10.8 MB)

TensorRT: starting export with TensorRT 10.3.0...
[12/31/2024-12:05:11] [TRT] [I] [MemUsageChange] Init CUDA: CPU +2, GPU +0, now: CPU 654, GPU 3655 (MiB)
[12/31/2024-12:05:13] [TRT] [I] [MemUsageChange] Init builder kernel library: CPU +927, GPU +683, now: CPU 1624, GPU 4338 (MiB)
[12/31/2024-12:05:13] [TRT] [I] -----
[12/31/2024-12:05:13] [TRT] [I] Input filename: yolo1n-cls.onnx
[12/31/2024-12:05:13] [TRT] [I] ONNX IR version: 0.0.9
[12/31/2024-12:05:13] [TRT] [I] Opset version: 19
[12/31/2024-12:05:13] [TRT] [I] Producer name: pytorch
[12/31/2024-12:05:13] [TRT] [I] Producer version: 2.5.0
[12/31/2024-12:05:13] [TRT] [I] Domain:
[12/31/2024-12:05:13] [TRT] [I] Model version: 0
[12/31/2024-12:05:13] [TRT] [I] Doc string:
[12/31/2024-12:05:13] [TRT] [I] -----
TensorRT: input "images" with shape(1, 3, 224, 224) DataType.FLOAT
```

## References

---

<https://docs.ultralytics.com/guides/nvidia-jetson/>

<https://docs.ultralytics.com/integrations/tensorrt/>