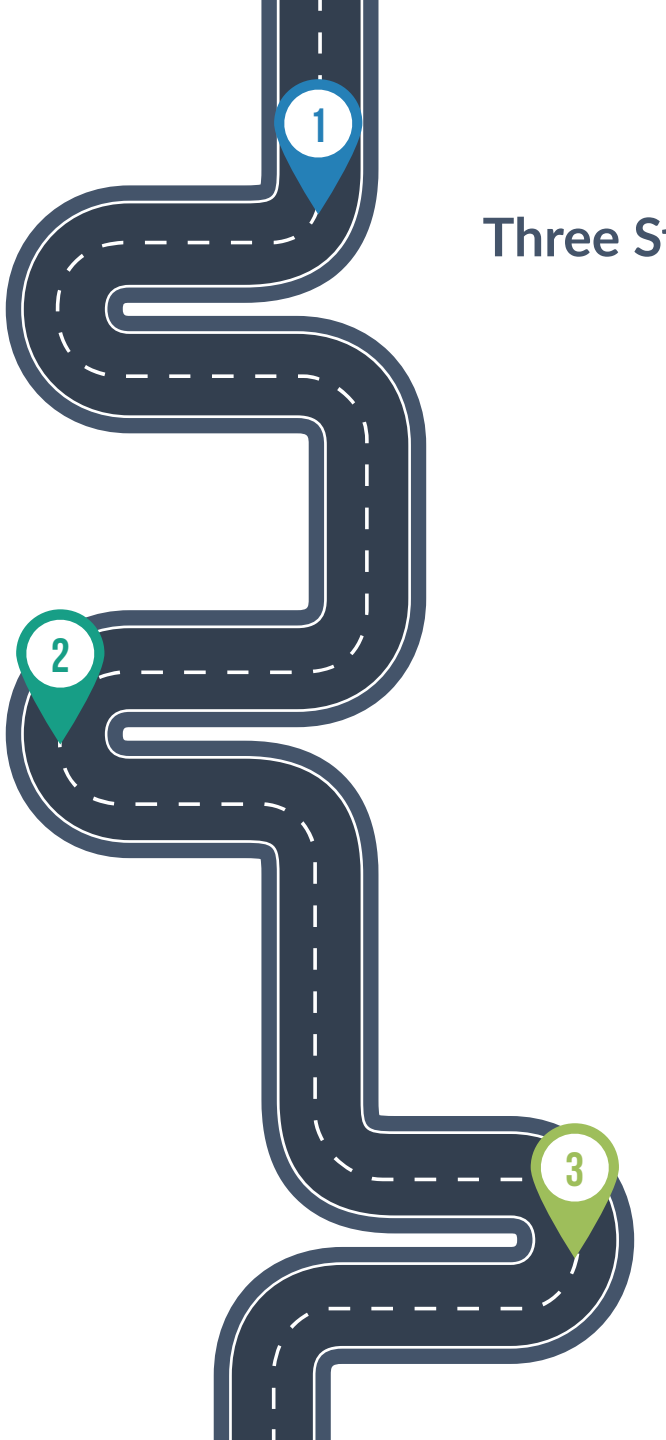
PORTING LIBREYOLOXS ON HAILO-8L CHIP

CVPR 2026 Tutorial

The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2026

Denver, CO, USA



PORTING STEPS

Three Steps to Convert from ONNX to HEF Format

STEP 01

ONNX to HAR

Converts the ONNX model into Hailo's intermediate HAR representation.

STEP 02

Quantize Model

Optimizes the model using lower-precision numerical formats for efficient edge inference.

STEP 03

Compiling HEF

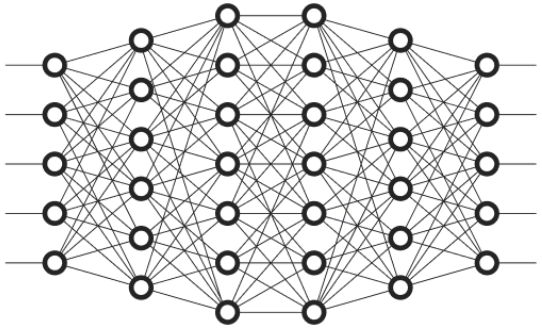
Compiles the optimized model into the deployable HEF binary format for Hailo hardware.

HAILO-AI DATAFLOW COMPILER

A Toolchain to Compile, Optimize Models for HailoAI Chips

Trained Network Model

ML Frameworks



TensorFlow TensorFlow Lite

Keras PyTorch ONNX

1

Hailo Dataflow Compiler

Model Parser

Model Optimizer

Resource Allocator

Compiler

Emulator

Profiler

2

Hailo Device

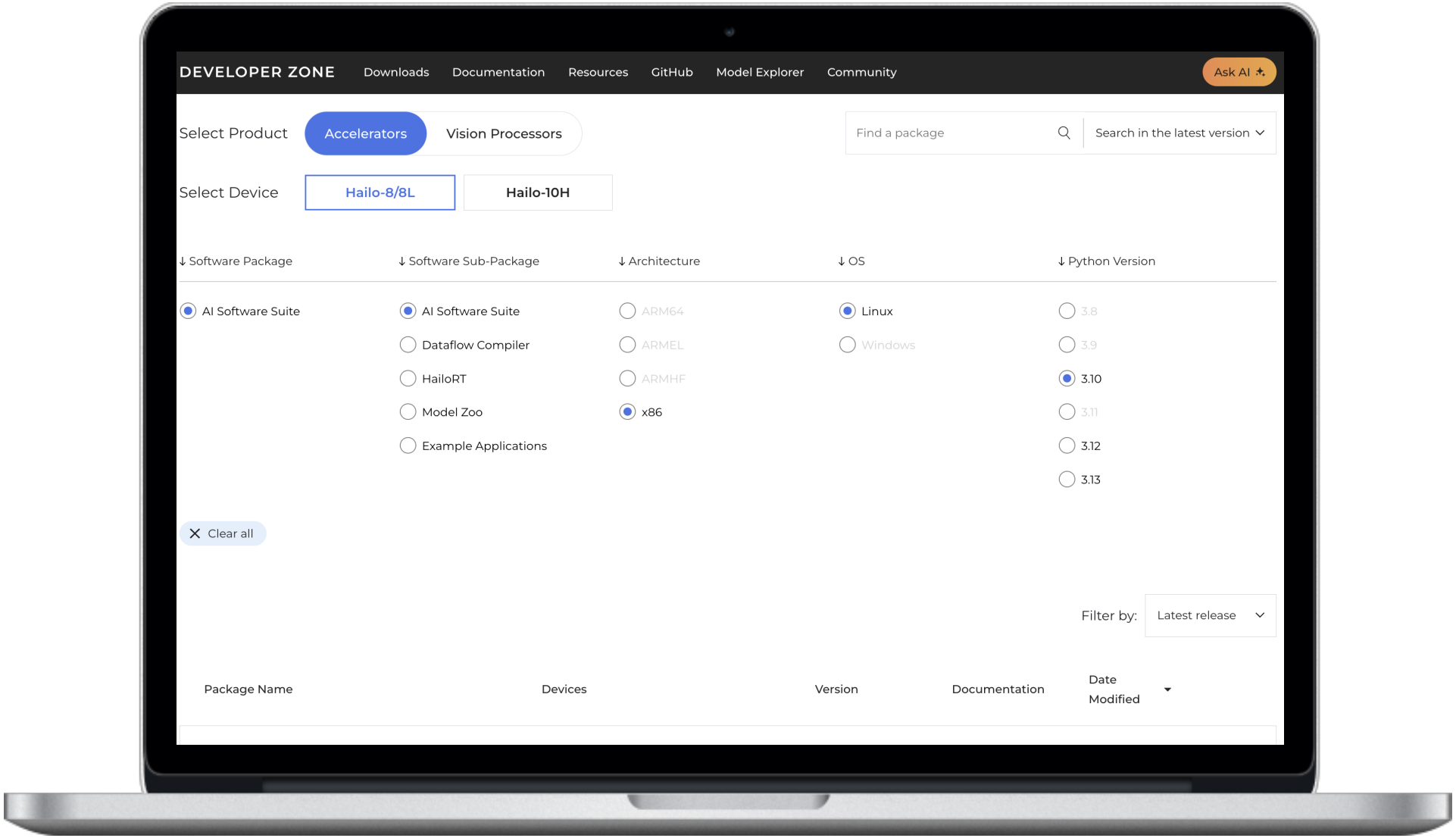
HAILO

3

Binary File

HAILO AI SOFTWARE SUITE DOCKER IMAGE

The Fastest Way to Get Started is the Prebuilt Hailo AI Software Suite Docker Image



PARSING ONNX TO HAR

Convert ONNX to Hailo's Archive Format

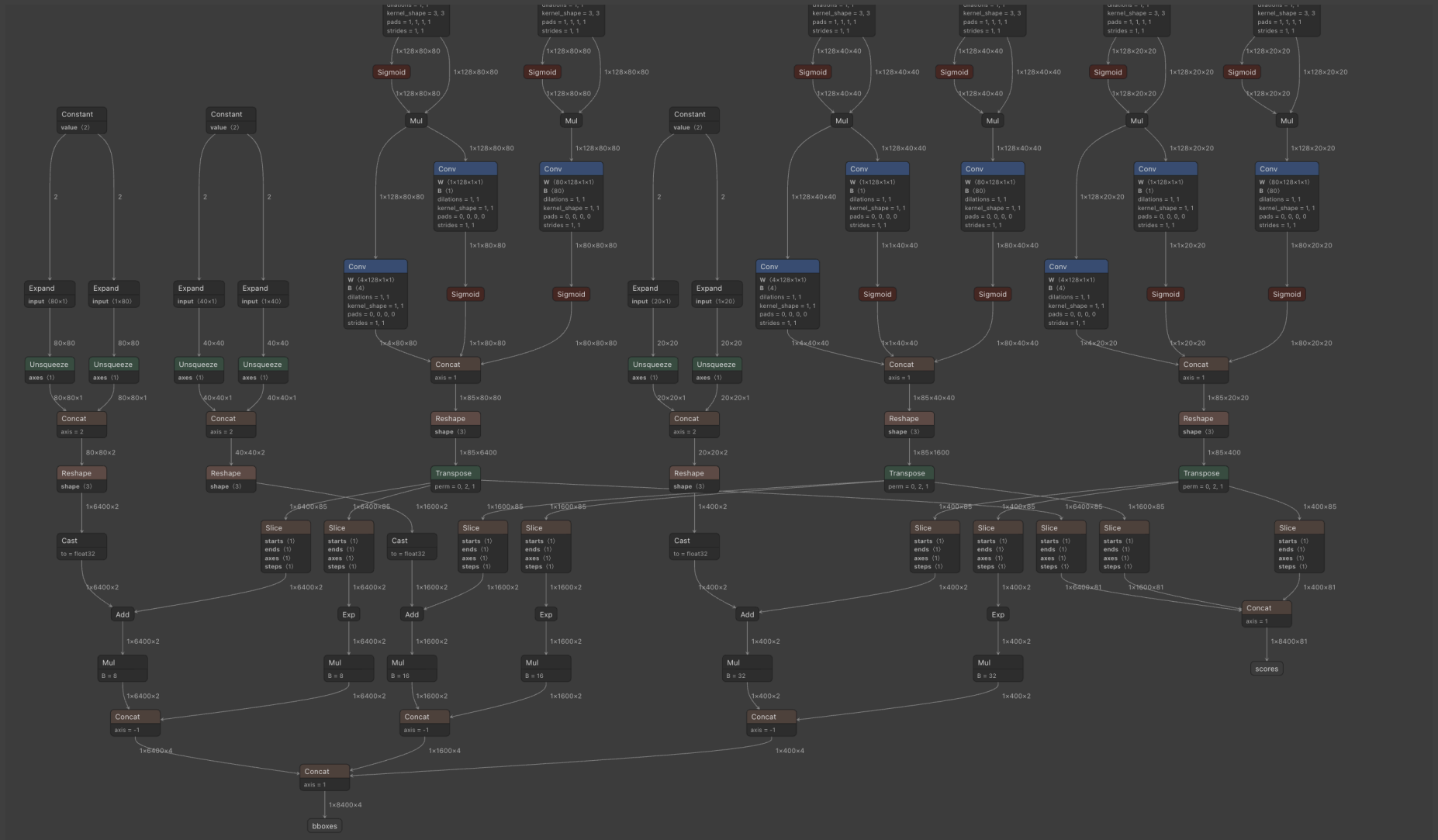
```
$ hailo parser onnx LibreYOLOXs.onnx --hw-arch hailo8l --har-path LibreYOLOXs.har
```

The parser **auto-detects YOLOx architecture** and warns if a built-in NMS processor is found.
NMS runs on the CPU by default. If output layers need manual mapping, we use the visualizer next.

```
[info] NMS structure of yolox (or equivalent architecture) was detected.
[info] In order to use HailoRT post-processing capabilities, these end node names should be used: /head/reg_preds.0/Conv /
head/Sigmoid_1 /head/Sigmoid /head/Sigmoid_3 /head/reg_preds.1/Conv /head/Sigmoid_2 /head/reg_preds.2/Conv /head/Sigmoid_4
 /head/Sigmoid_5.
[info] Start nodes mapped from original model: 'images': 'LibreYOLOXs/input_layer1'.
[info] End nodes mapped from original model: 'bboxes', 'scores'.
[info] Translation completed on ONNX model LibreYOLOXs (completion time: 00:00:02.50)
Would you like to parse the model again with the mentioned end nodes and add nms postprocess command to the model script?
(y/n)
```

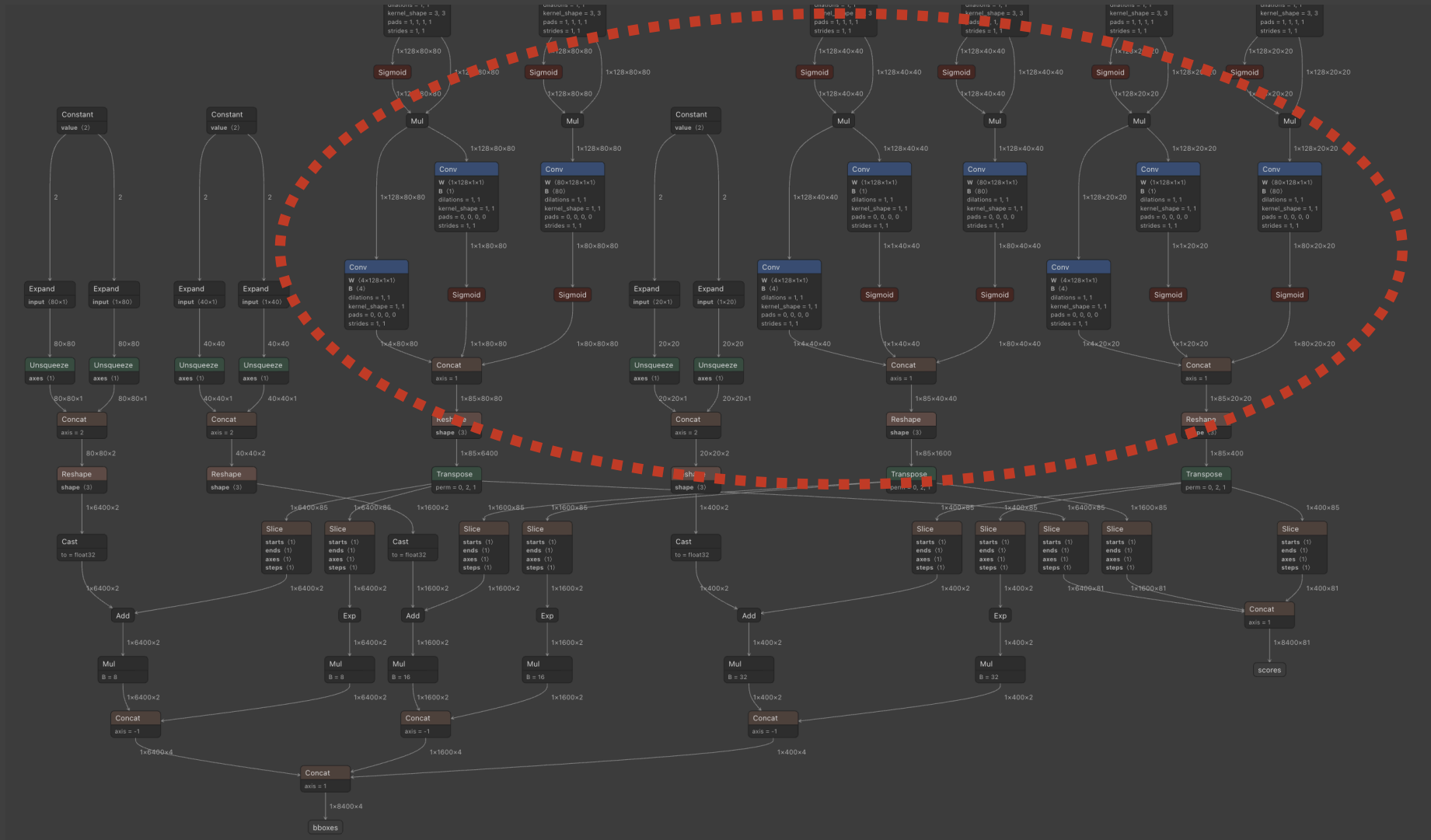
AUTO NMS PROCESSING

Usually all YOLO model outputs are fused from different levels to get two outputs boxes, scores



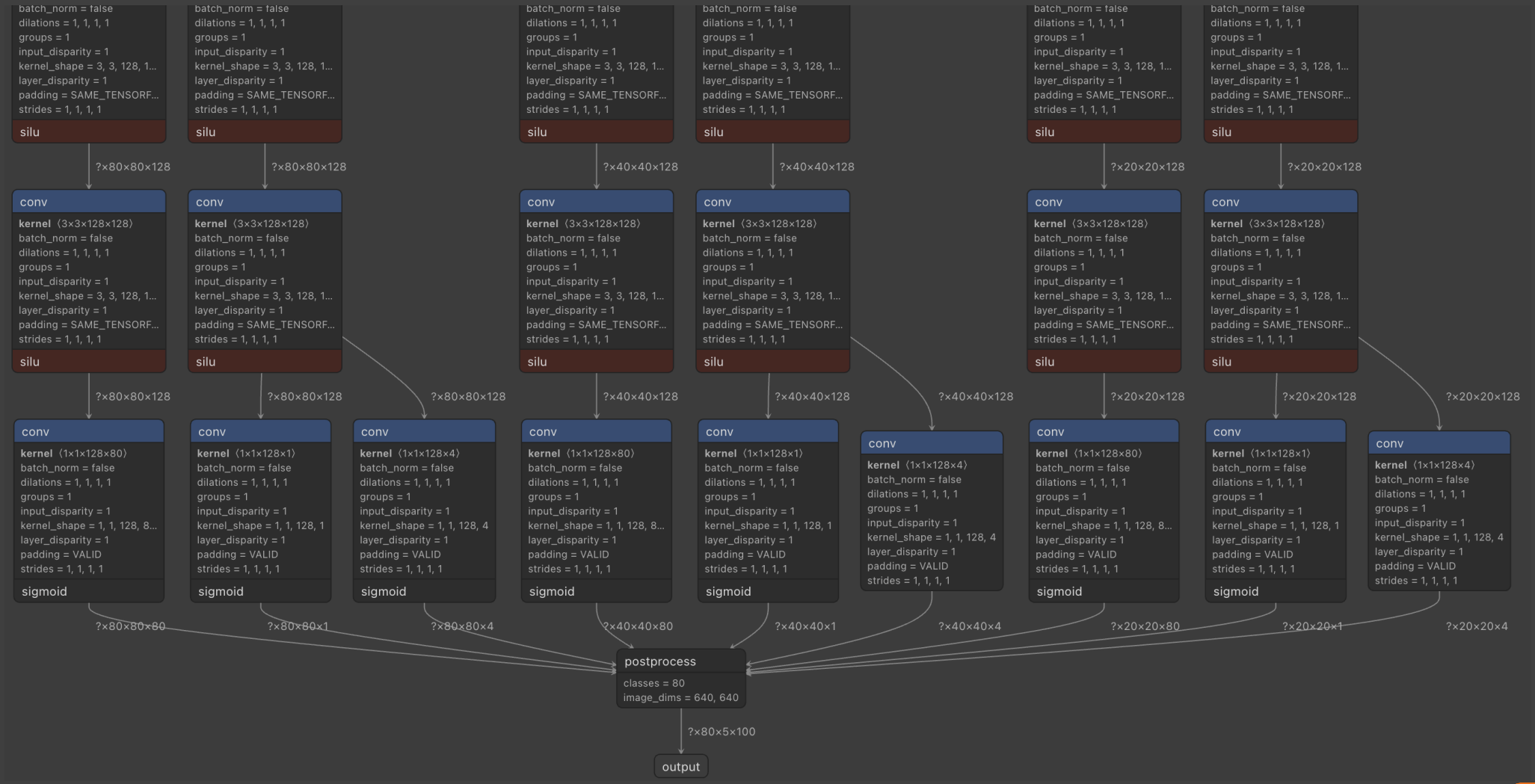
AUTO NMS PROCESSING

Usually all YOLO model outputs are fused from different levels to get two outputs boxes, scores



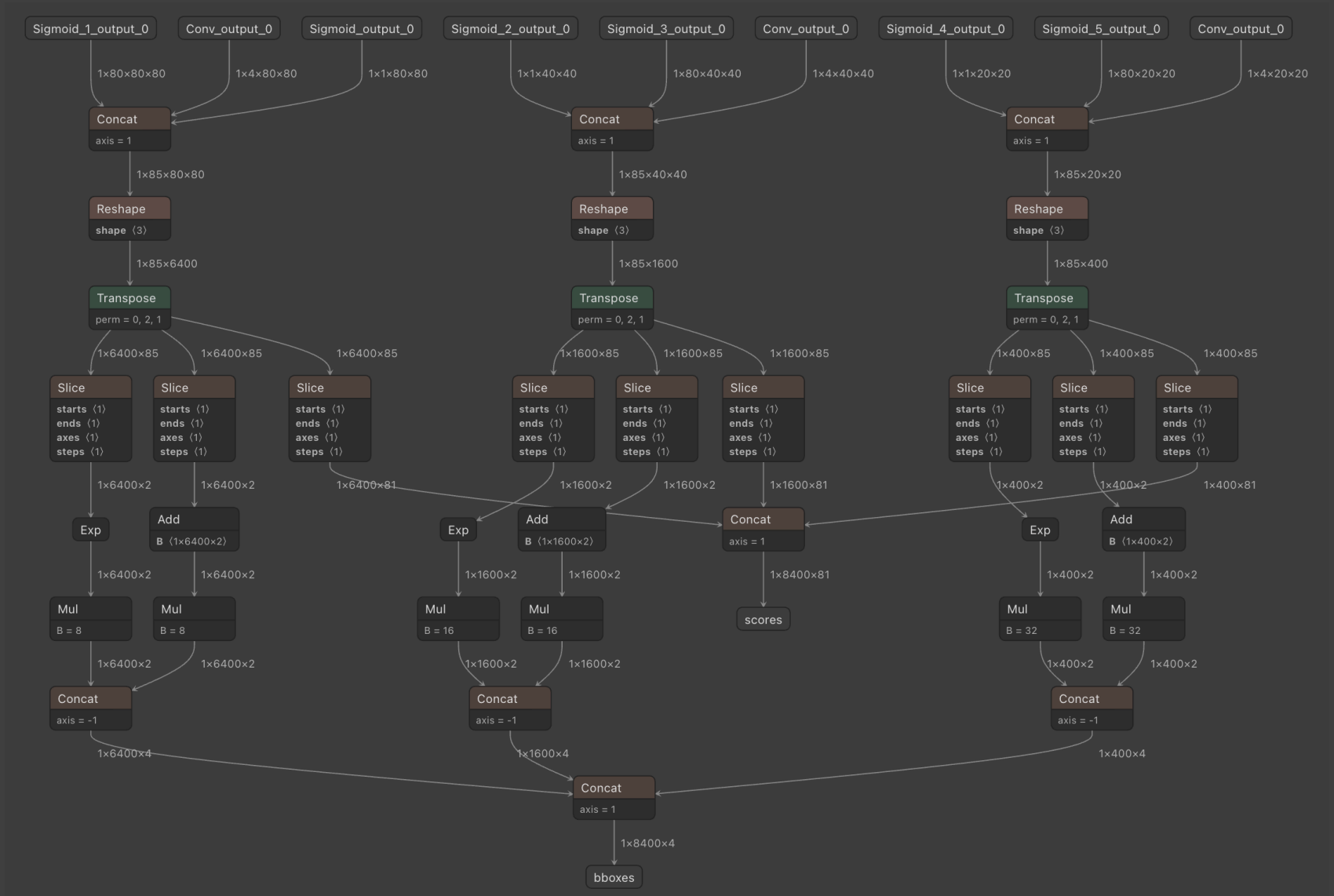
AUTO NMS PROCESSING

HAR Output Format



AUTO NMS PROCESSING

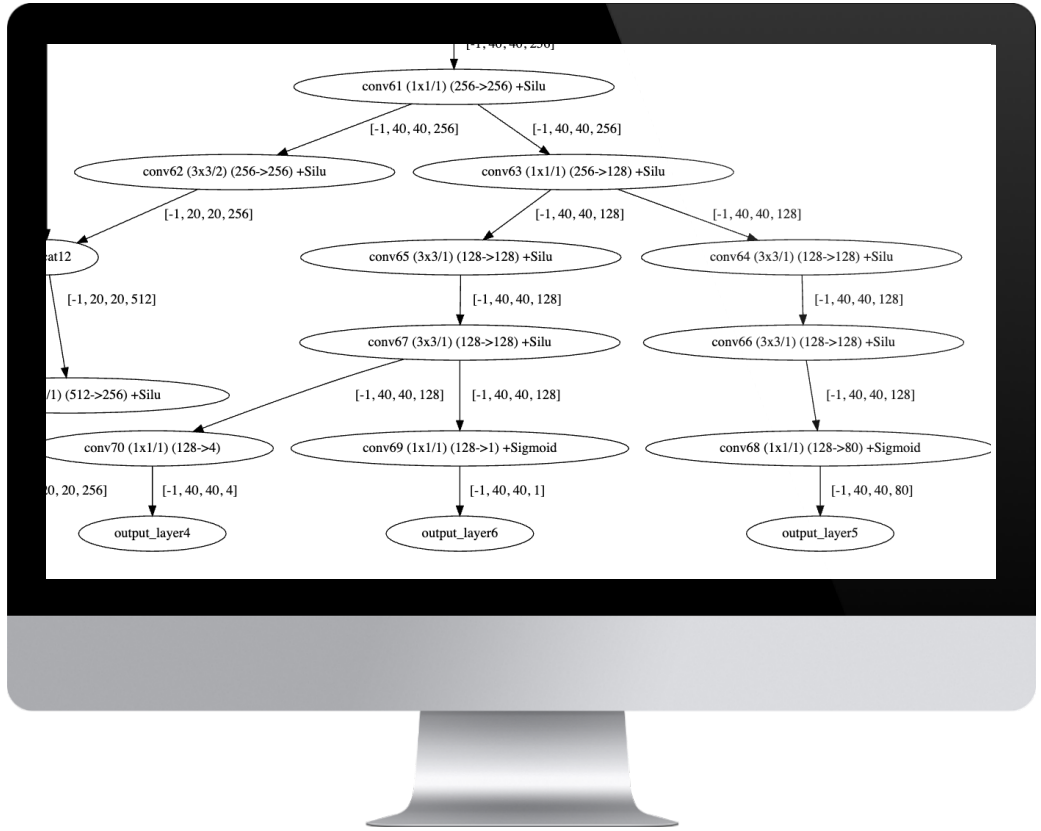
Postprocessing



MAP OUTPUT LAYERS

Visualization

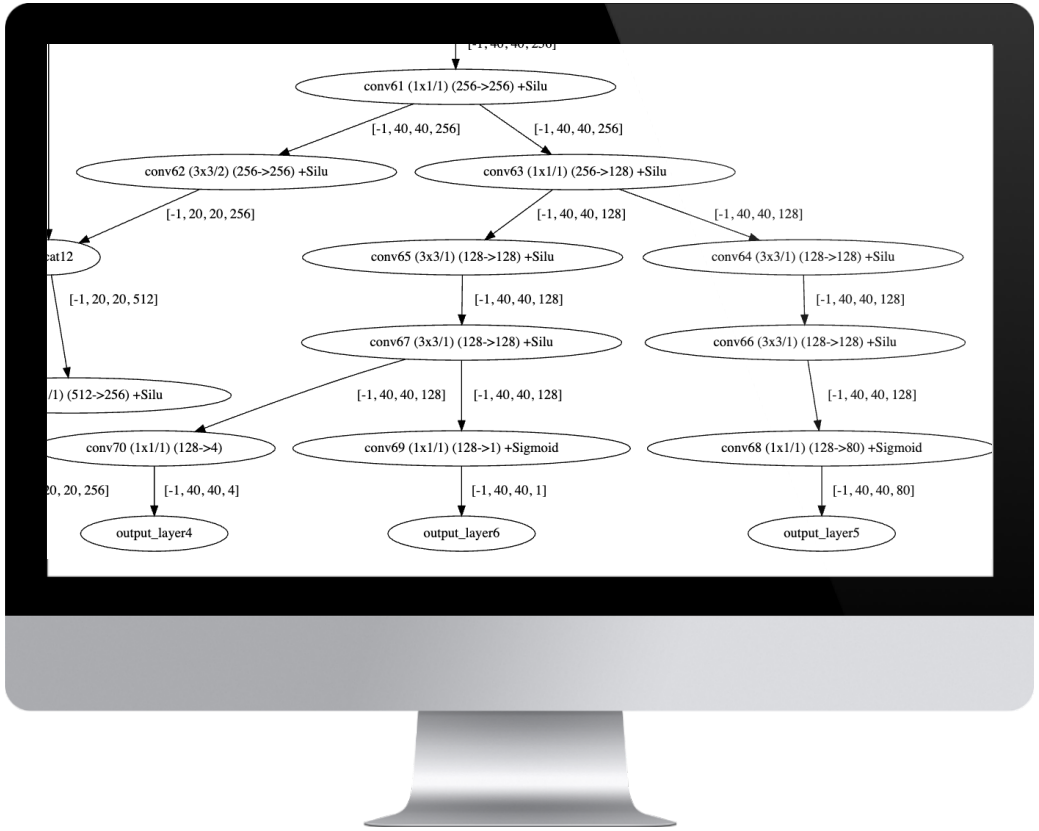
When the pipeline can not map outputs automatically, use the Hailo Visualizer to inspect layer names:

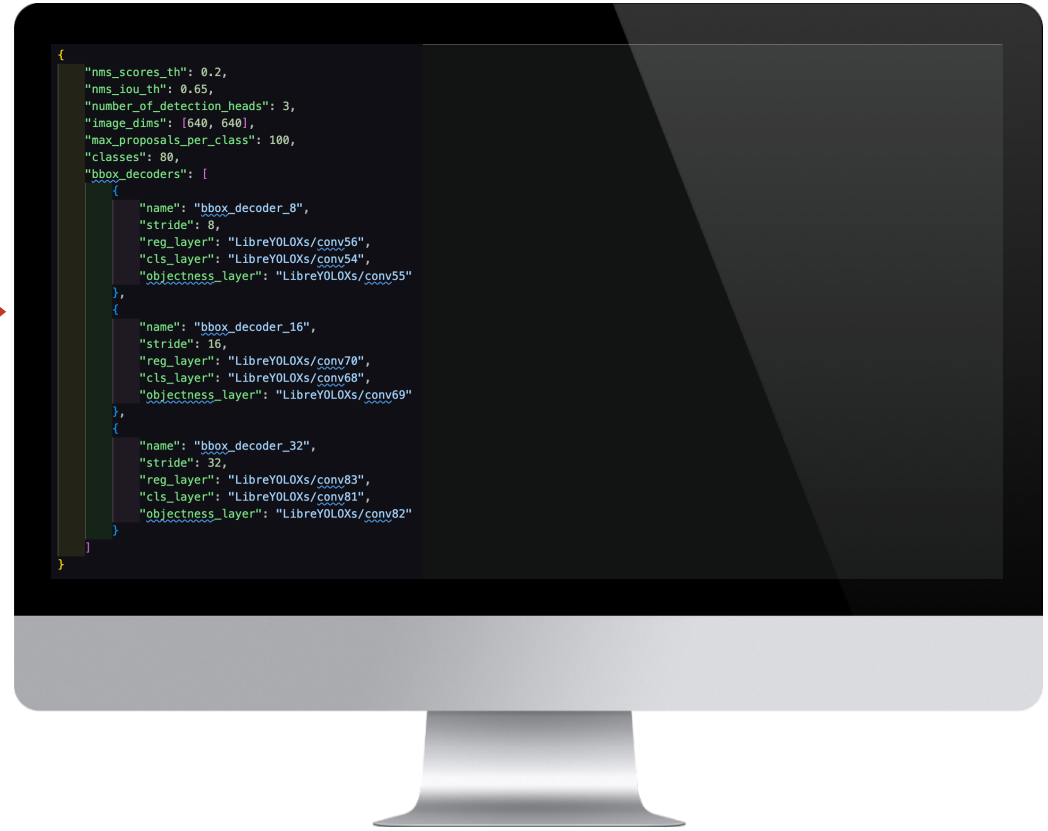


MAP OUTPUT LAYERS

Visualization

Create a JSON file to tell the compiler exactly which tensors to route:

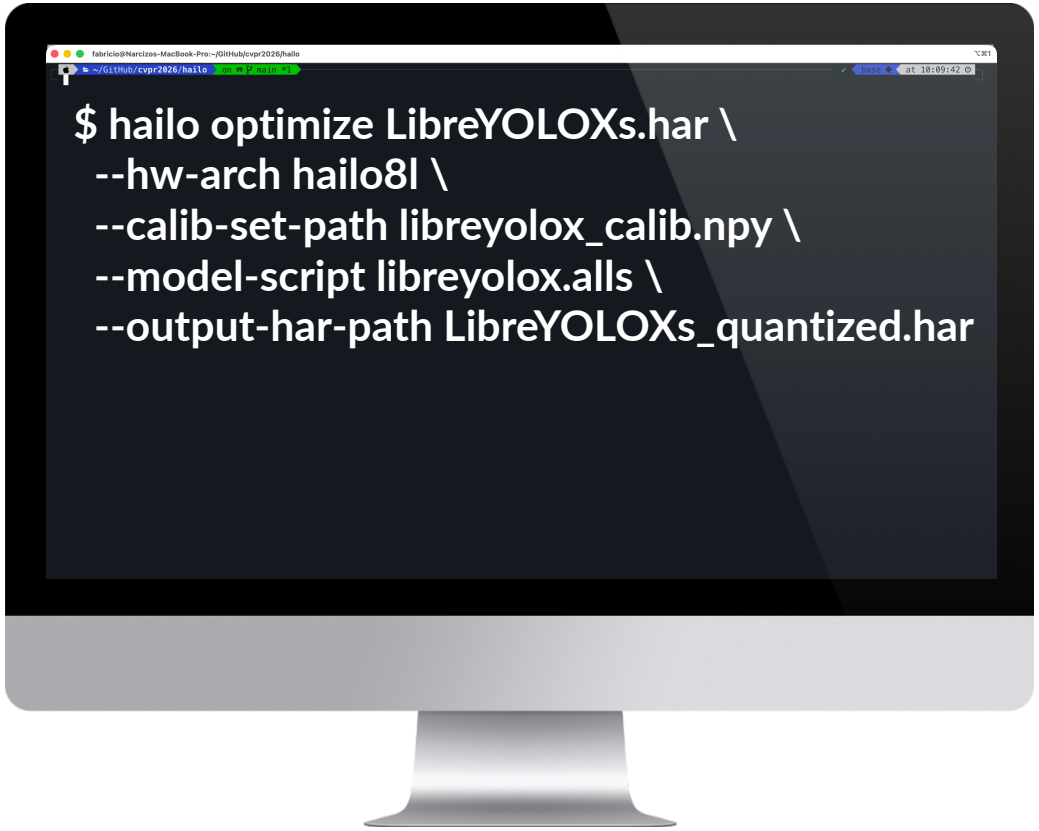


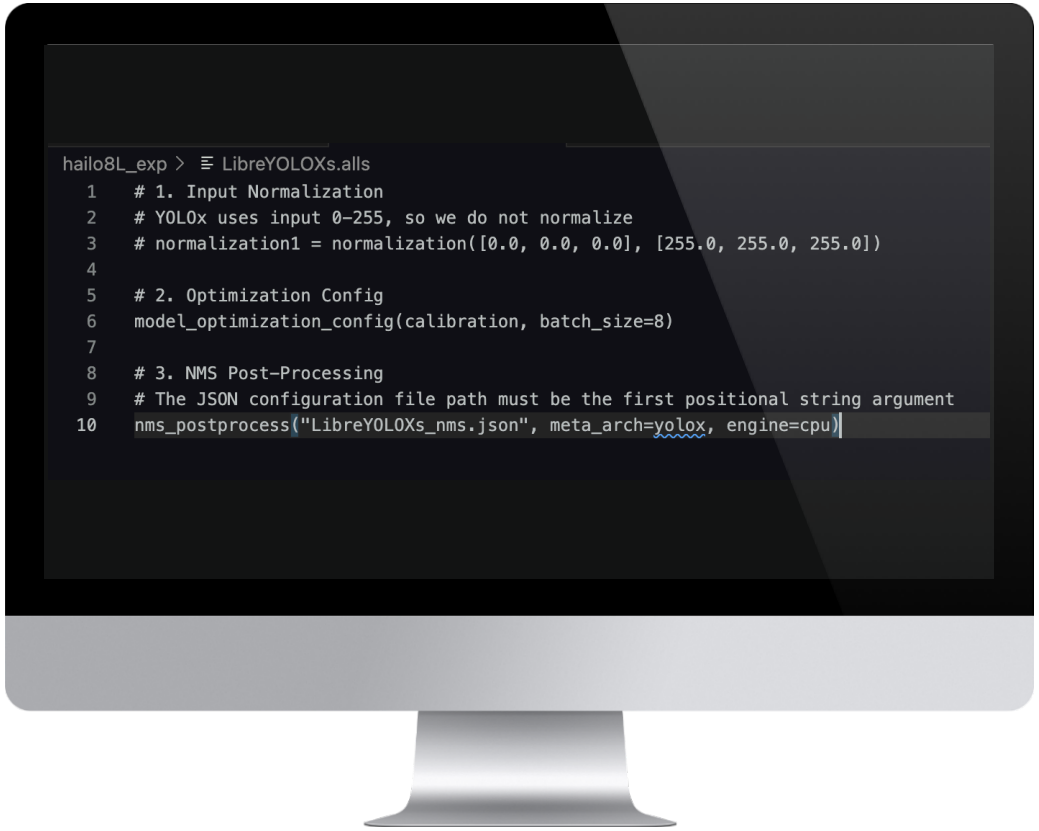


QUANTIZE HAR MODEL

Quantizing LibreYOLOXs Model with Calibration Dataset

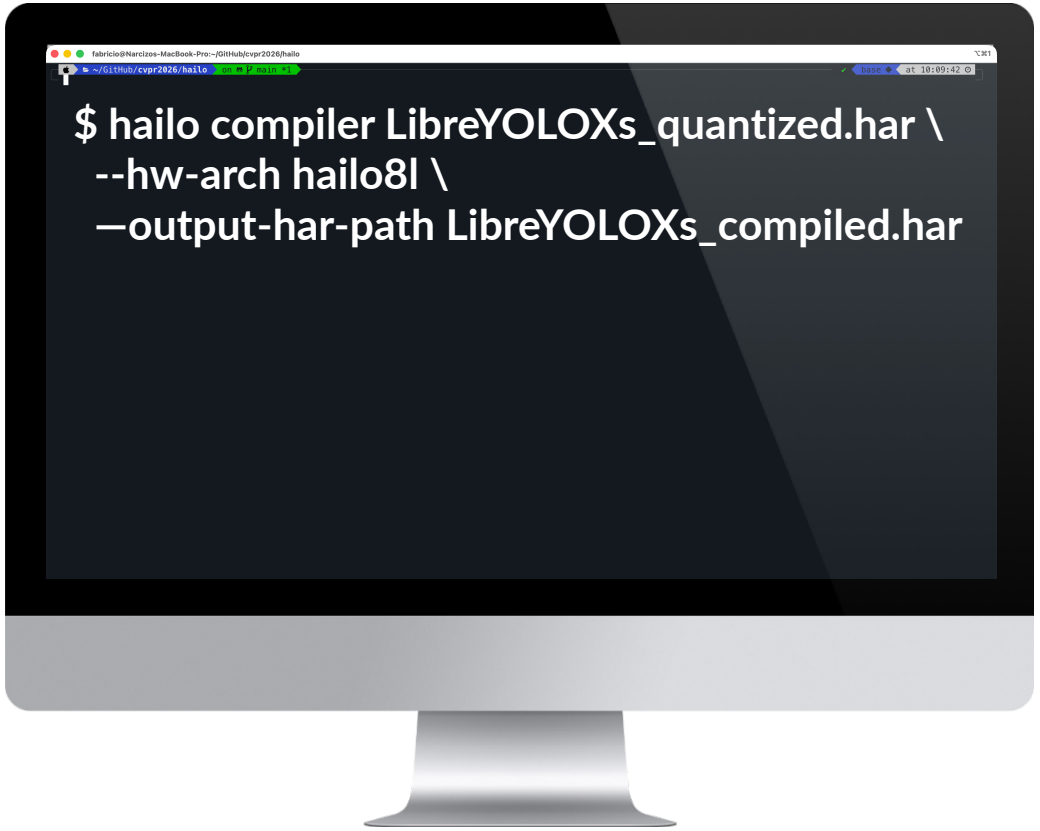
How to generate calibration dataset is provided in out CVPR repo:





COMPILE TO HEF

Generate the deployable binary



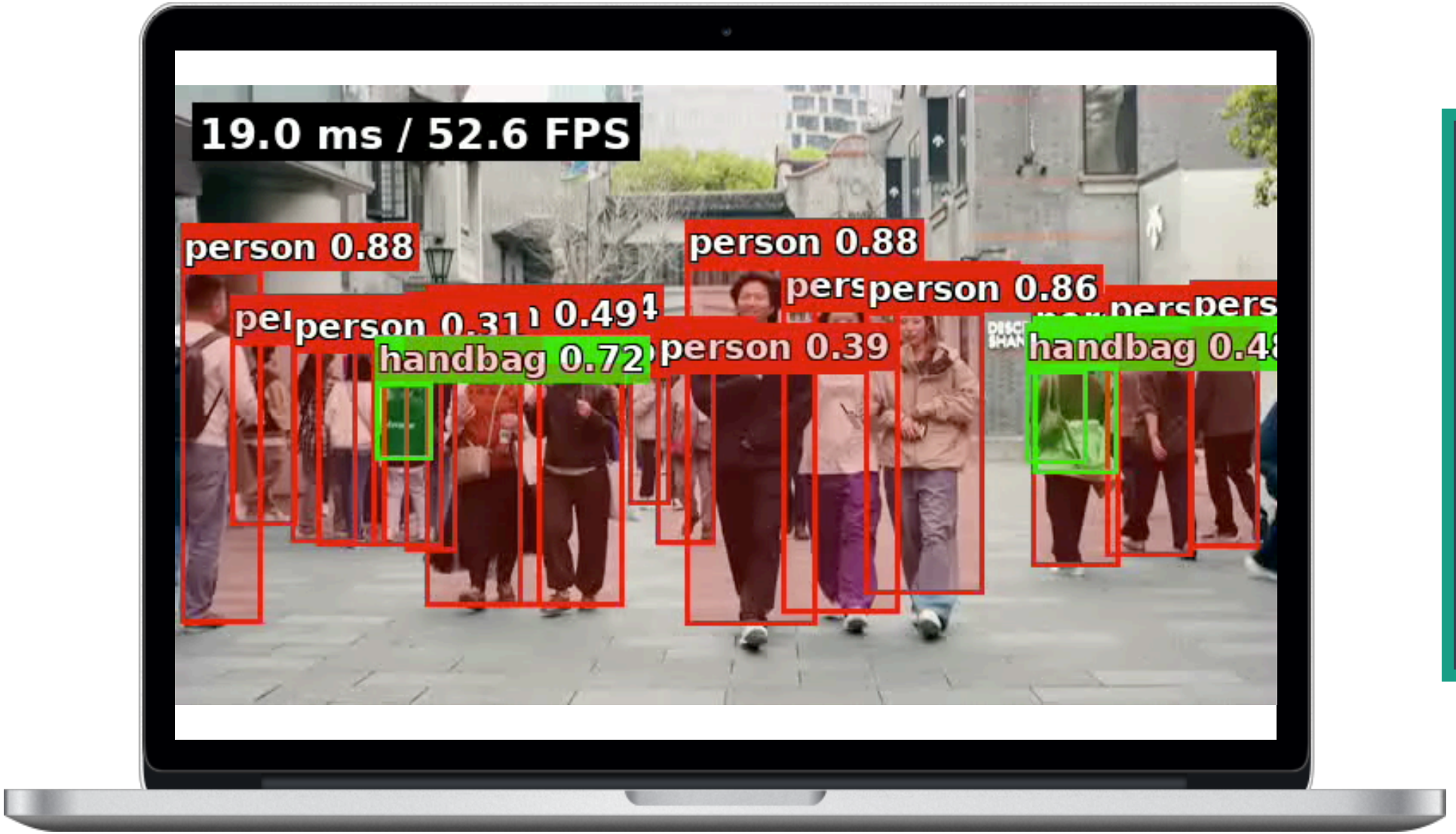
HAILO

Output (.HEF)

A Hardware-Executable Format binary ready to load onto the Hailo-8L accelerator.

DEPLOT ON HEF MODEL

Inference on Raspberry Pi 5 + Hailo-8L AI HAT



HAILO

Conclusions

Run inference with a simple Python script that loads the .HEF model and processes video frames through the Hailo-8L accelerator.

QUESTIONS & ANSWERS

BREAK (15 MINUTES)

THANK YOU!