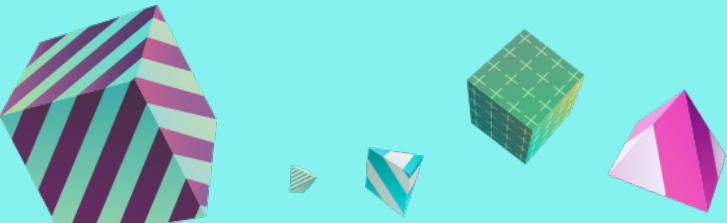


QNAP

Accelerate AI Inference for Computer Vision with OpenVINO™

Workflow Consolidation Tool





Agenda



- **How to Accelerate Deep Learning Inference:**

Using Intel® OpenVINO™ Toolkit

- **QTS Application Introduction:**

OpenVINO™ Workflow Consolidation Tool

- **Demo:**

Car detection and fruit detection



OpenVINO™



How to accelerate Deep Learning inference: Using Intel® OpenVINO™ Toolkit

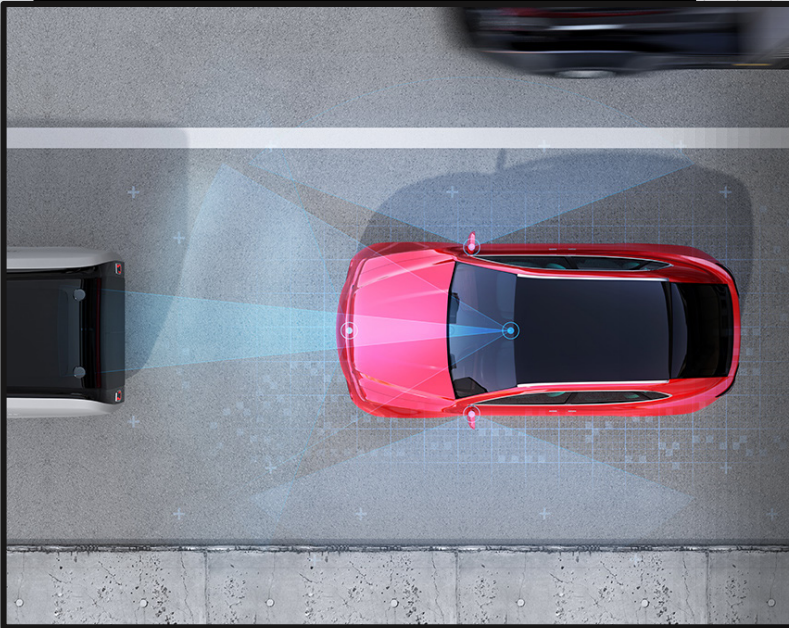


AI makes our life easier



Computer Vision (Image & Video)

Autonomous Car



Smart Camera



Healthcare



AI, ML, DL, CNN...



Machine Learning (ML):

Machine Learning is the practice of giving a computer a set of rules and tasks, then letting it figure out a way to complete those tasks.

Deep Learning (DL):

A subset of machine learning which makes the computation of multi-layer neural networks feasible.

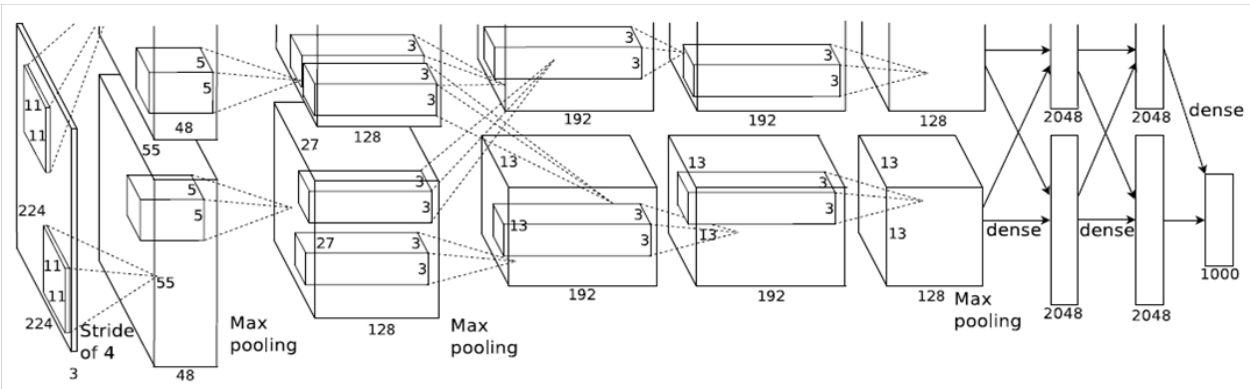
Convolutional Neural Network(CNN):

Deep learning topology particularly effective at **image/video classification** and **recognition**.

Why is CNN more commonly used for computer vision?



- **Convolutional Neural Network (CNN)** are using complex matrix calculation to label the data feature.
- The video and image are composed by the matrix with **Red, Green, Blue** parameter and pixel.



CNN Topology

(Model Sample : AlexNet)

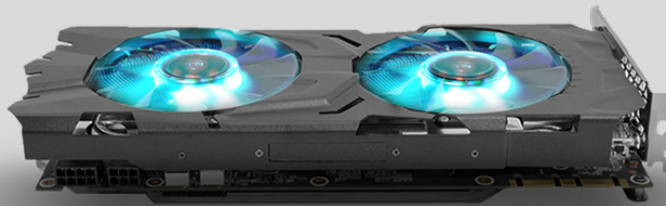
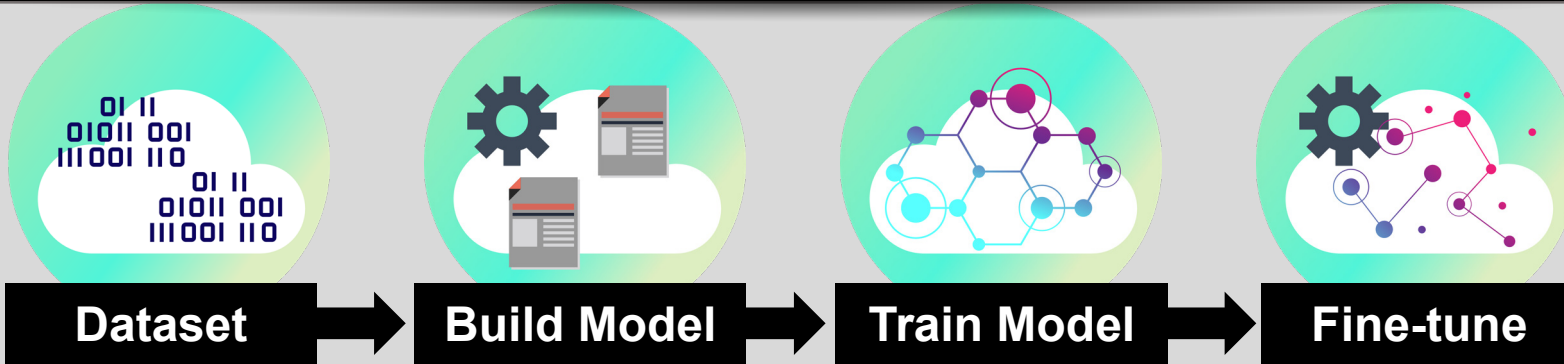


Computer Vision

What's different between Training and Inference

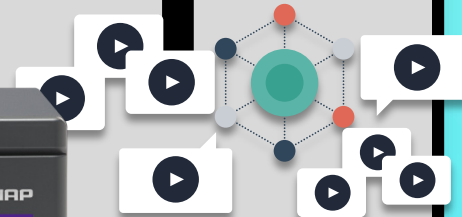


Training



NVIDIA

Inference

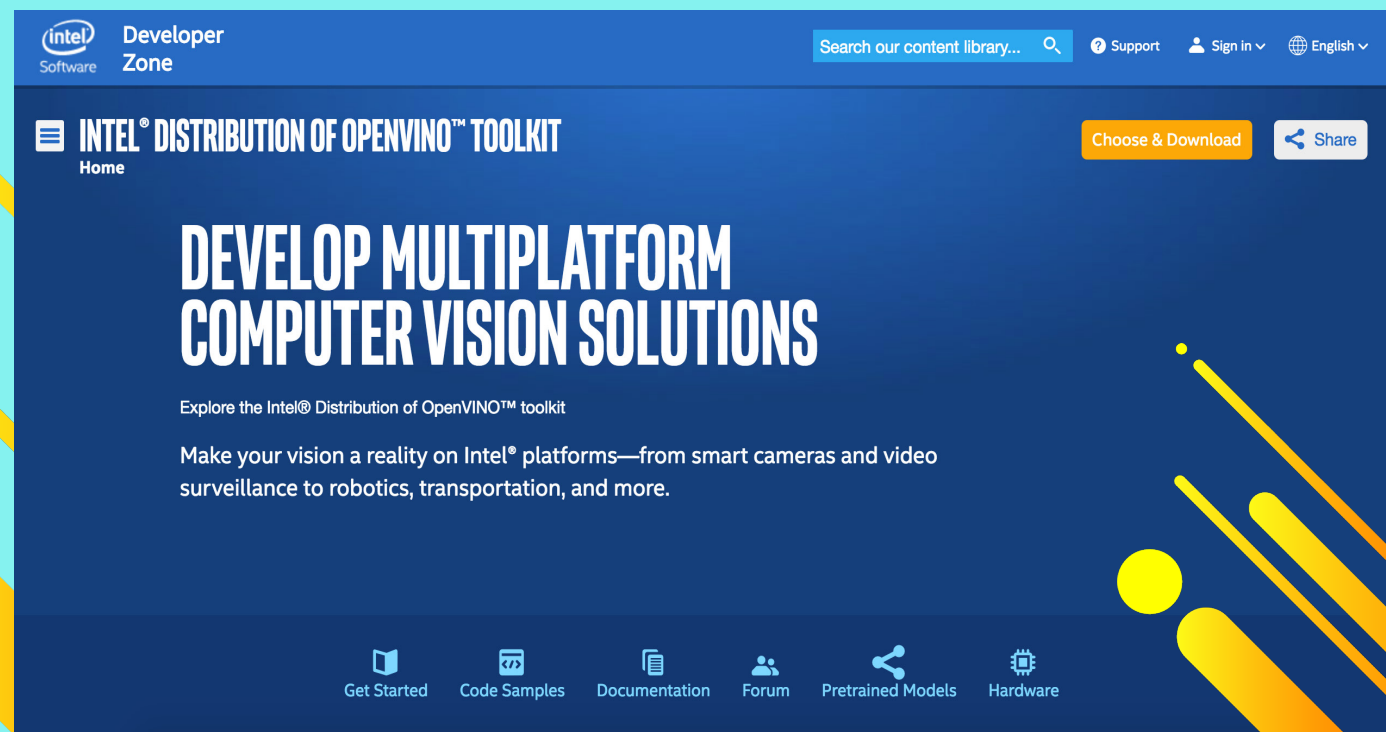


Edge Data

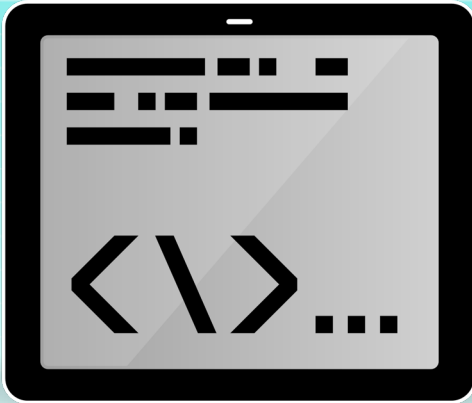
OpenVINO™ Toolkit - Accelerating the deep learning inference for computer vision



- Intel® OpenVINO™ toolkit (**Open Visual Inference and Neural network Optimization**) is **free software** that helps developers and data scientists speed up computer vision workloads, streamline deep learning inference and deployments.
- Enable easy, heterogeneous execution across Intel® platforms **from edge to cloud**. It helps to:
 - Increase deep learning workload performance
 - Unleash convolutional neural network (CNN)-based deep learning inference
 - Accelerate development



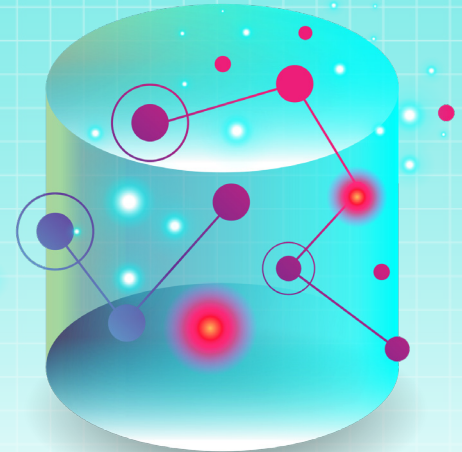
Features



**Intel® Deep Learning
Deployment Toolkit**
(Model Optimizer and Inference Engine)



**Optimized computer
vision libraries**
(OpenCV & OpenVX)



**Pre-trained Deep
Learning Model**

How does OpenVINO™ Toolkit work?



Model Optimizer

What it is: preparation step -> imports trained models

Why important:

Optimizes for performance/space with conservative topology transformations; biggest boost is from conversion to data types matching hardware.

Inference Engine

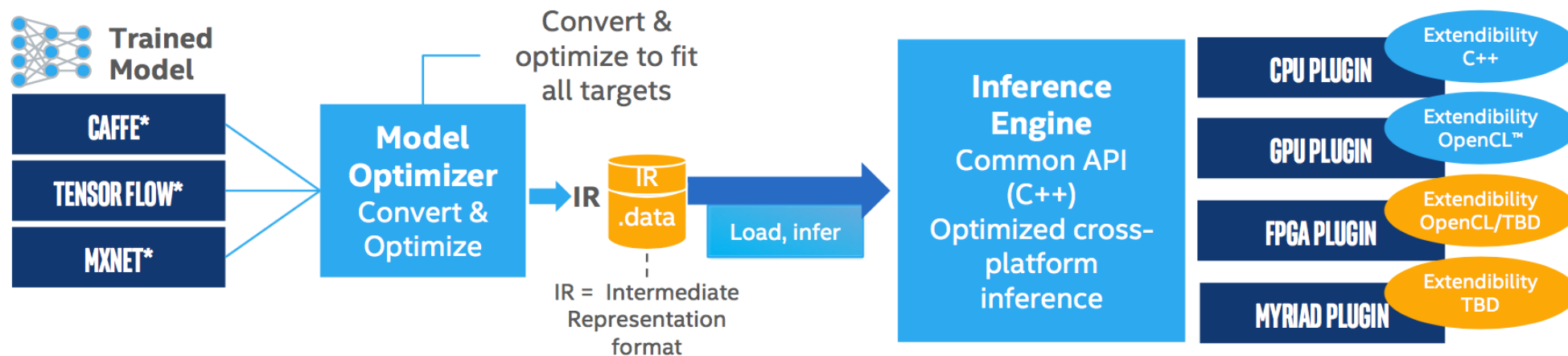


What it is: High-level inference API

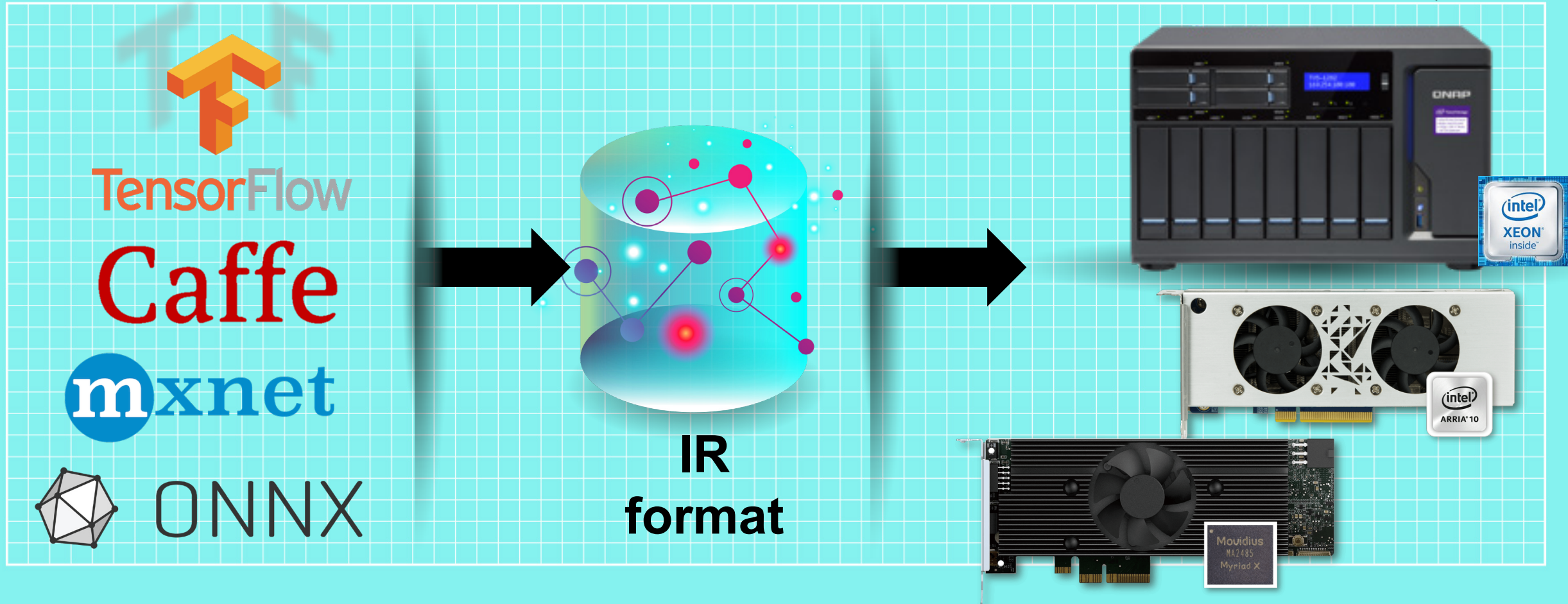
Why important:

Interface is implemented as dynamically loaded plugins for each hardware type. Delivers best performance for each type without requiring users to implement and maintain multiple code pathways.

Source: [Intel® OpenVINO™ Toolkit](#)



Migrate from other AI framework to accelerate the inference on Intel® platform



Open source code from Intel® – help you develop and optimize



Sample Code:

- Image Classification
- Image Segmentation
- Object Detection
- Object Detection for Single Shot Multibox Detector (SSD)
- Neural Style Transfer
- Validation Application

Pre-trained Model:

- Age - Gender
- Security barrier
- Crossroad
- Headpose
- MobilenetSSD
- Face mobilenetreduced SSD with shared weights
- Face detect with SQLight SSD
- Vehicle Attributes

Supported Sample still growing (2018 R5)



Source: [Intel® OpenVINO™ Toolkit](#)

Supported Samples

Reference this table for components that support the pretrained models.

Pretrained Model	Supported Samples	CPU	Integrated Graphics	FPGA	VPU
face-detection-adas-0001	Interactive face detection	✓	✓	✓	✓
age-gender-recognition-retail-0013	Interactive face detection	✓	✓	✓	✓
head-pose-estimation-adas-0001	Interactive face detection	✓	✓	✓	
emotions-recognition-retail-0003	Interactive face detection	✓	✓	✓	✓
facial-landmarks-35-adas-0001	Interactive face detection	✓	✓		
vehicle-license-plate-detection-barrier-0106	Security barrier camera	✓	✓	✓	✓
vehicle-attributes-recognition-barrier-0039	Security barrier camera	✓	✓	✓	✓
license-plate-recognition-barrier-0001	Security barrier camera	✓	✓	✓	✓
person-detection-retail-0001	Object detection	✓	✓		
person-vehicle-bike-detection-crossroad-0078	Crossroad camera	✓	✓	✓	✓
person-attributes-recognition-crossroad-0031	Crossroad camera	✓	✓	✓	✓

QTS Software:

OpenVINO™

Workflow Consolidation Tool



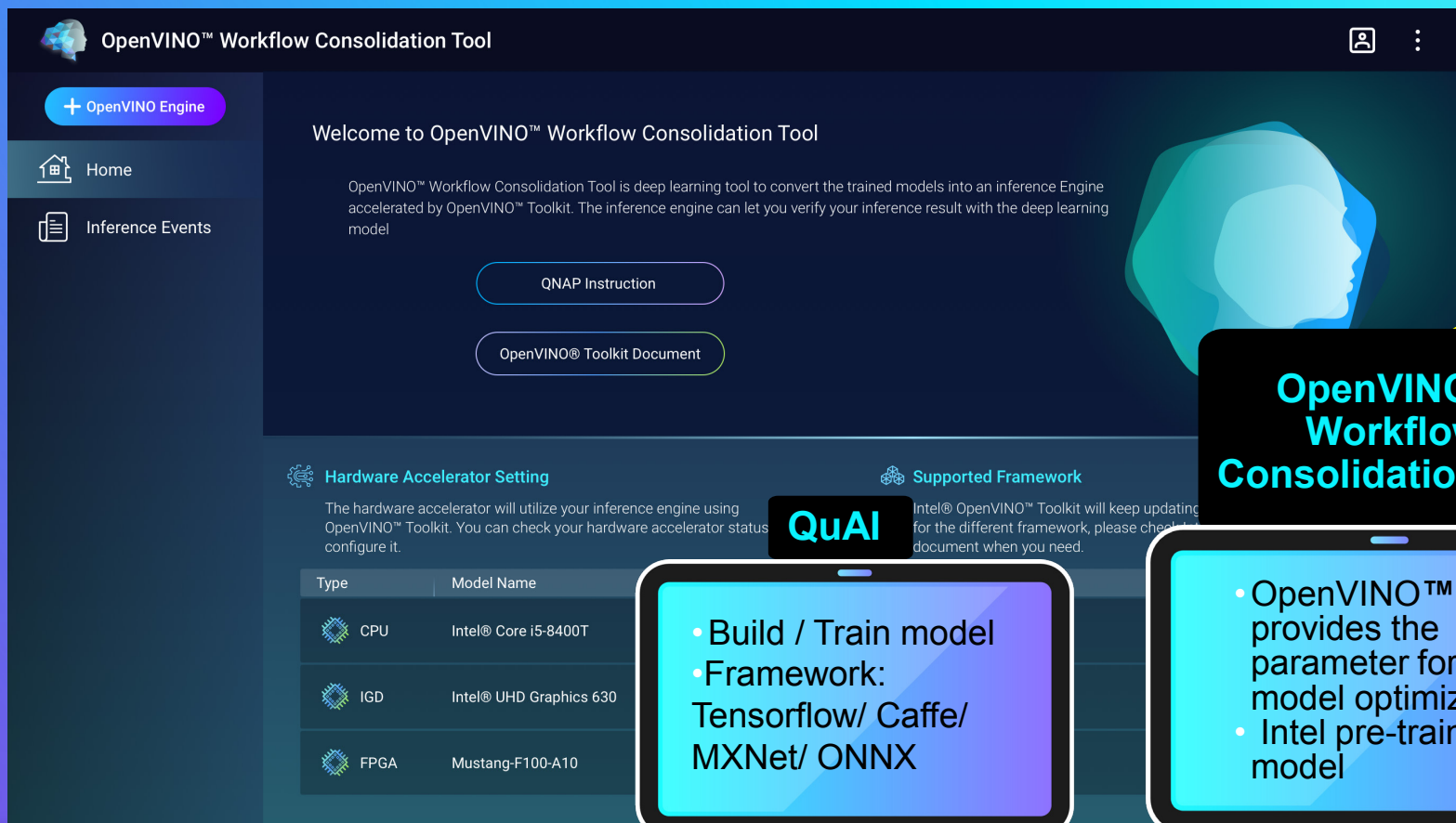
Pain points of Deep Learning and deploying OpenVINO™ Toolkit



- Steep learning curve for deep learning
- Needs lots of time to train the deep learning model
- Deep learning inference needs hardware acceleration



OpenVINO™ “Workflow Consolidation Tool” AI



Deep learning process
integrated with
OpenVINO™ toolkit

Transfer to standard
format (IR)

OpenVINO™
Workflow
Consolidation Tool

Inference
Engine

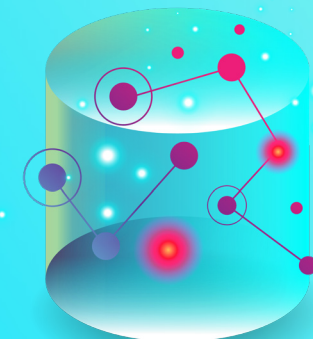
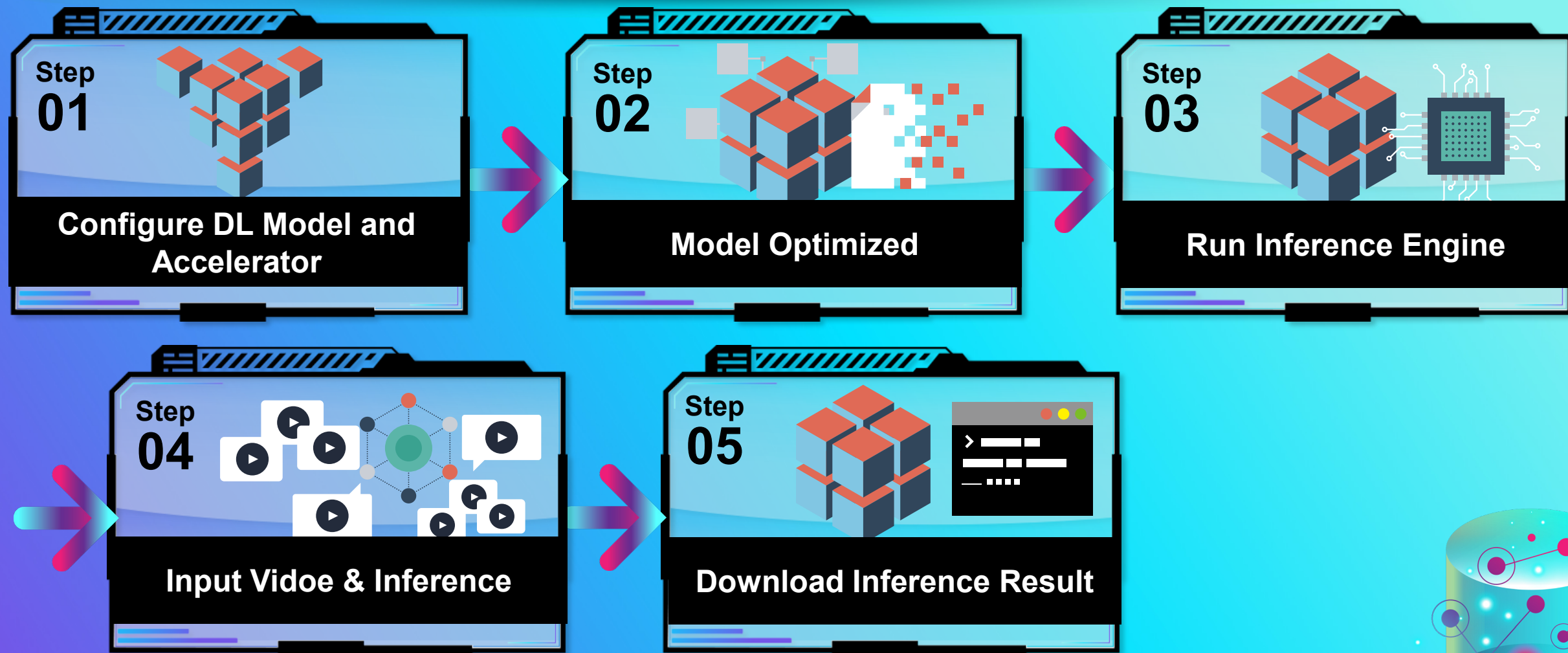
- Build / Train model
- Framework:
Tensorflow/ Caffe/
MXNet/ ONNX

- OpenVINO™
provides the
parameter for
model optimizer
- Intel pre-trained
model

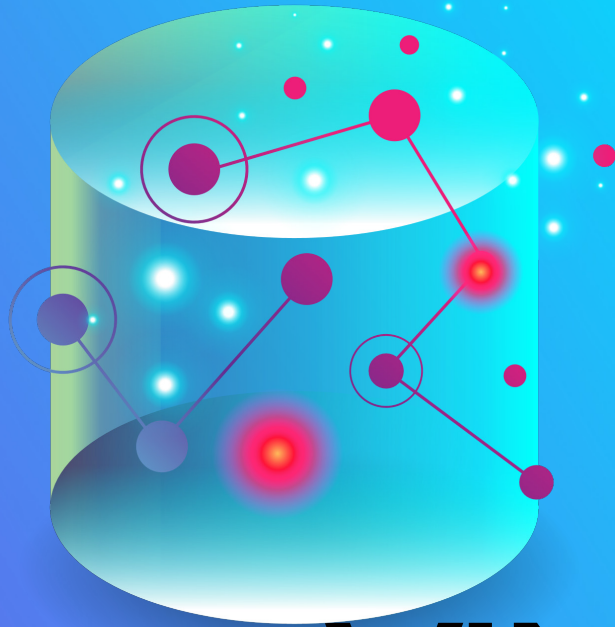
- Inference using
CPU/IGD/Intel
FPGA/Intel VPU

Pre-trained Model /
Data Label, etc.

Workflow



Features



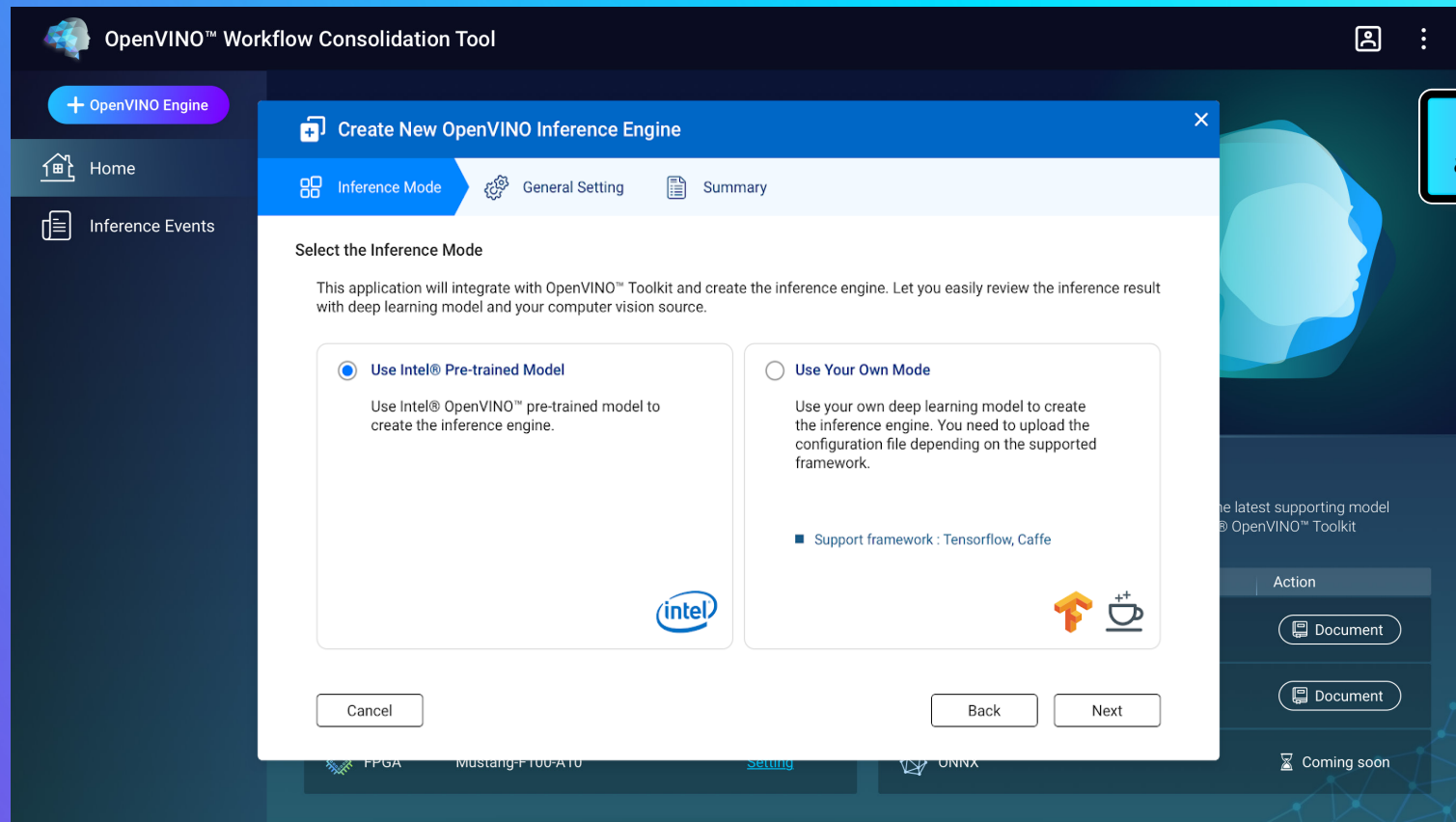
OpenVINO™

QNAP

- Include pre-trained models provided by Intel®
- Provide GUI to adapt OpenVINO™ Toolkit
- First application to support inference accelerator card - FPGA (Mustang-F100-A10)



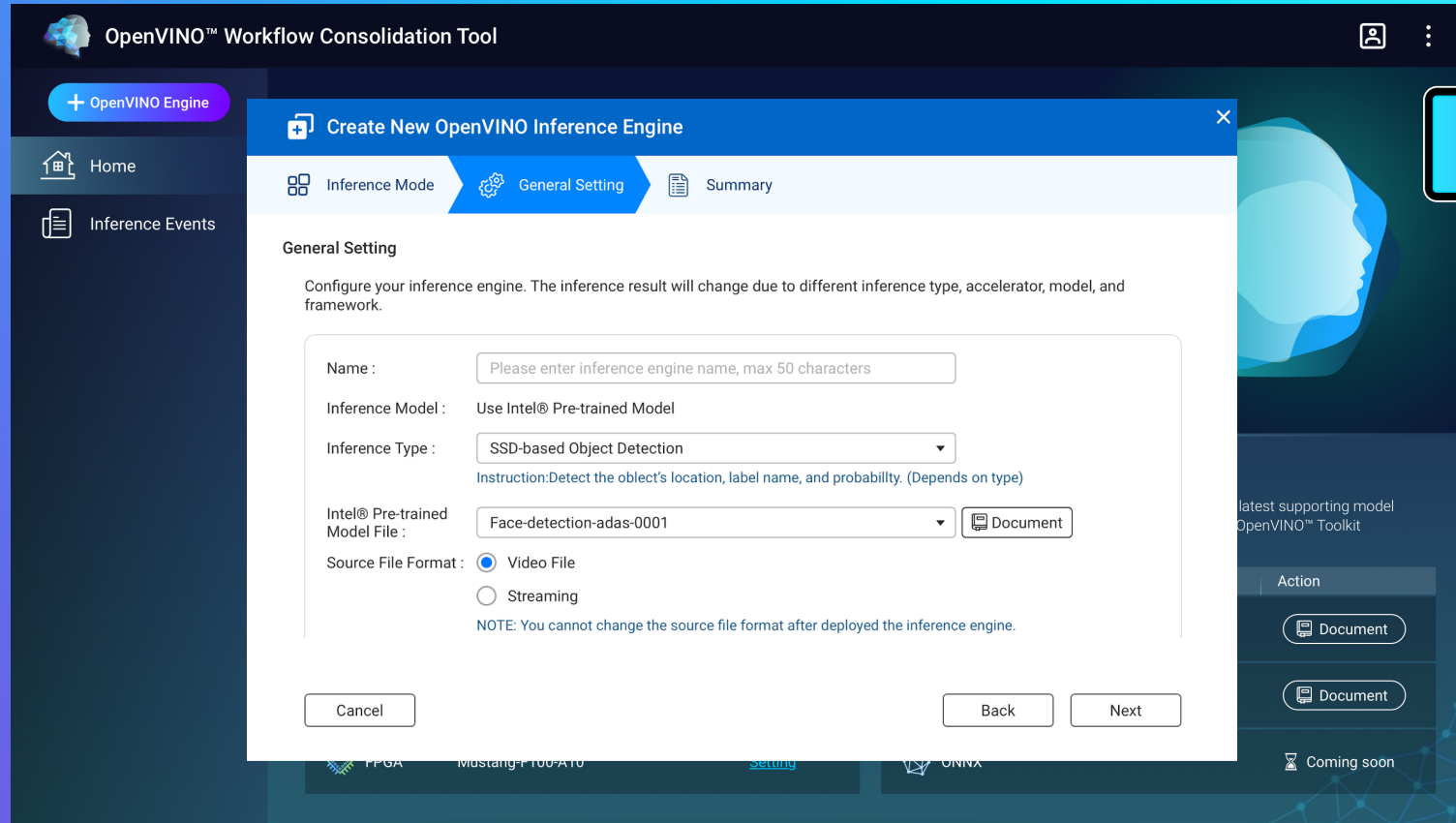
Feature 1: Support Intel® pre-trained model



Inference Mode

- Default the DL model from Intel
- Migrate other framework to inference using OpenVINO™.
- Support framework: Tensorflow, Caffe

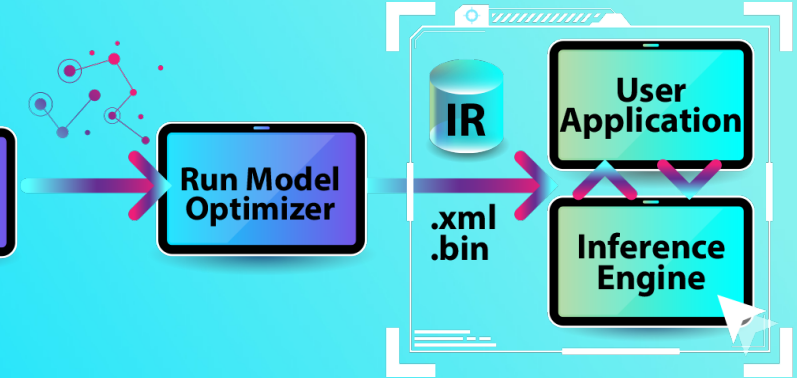
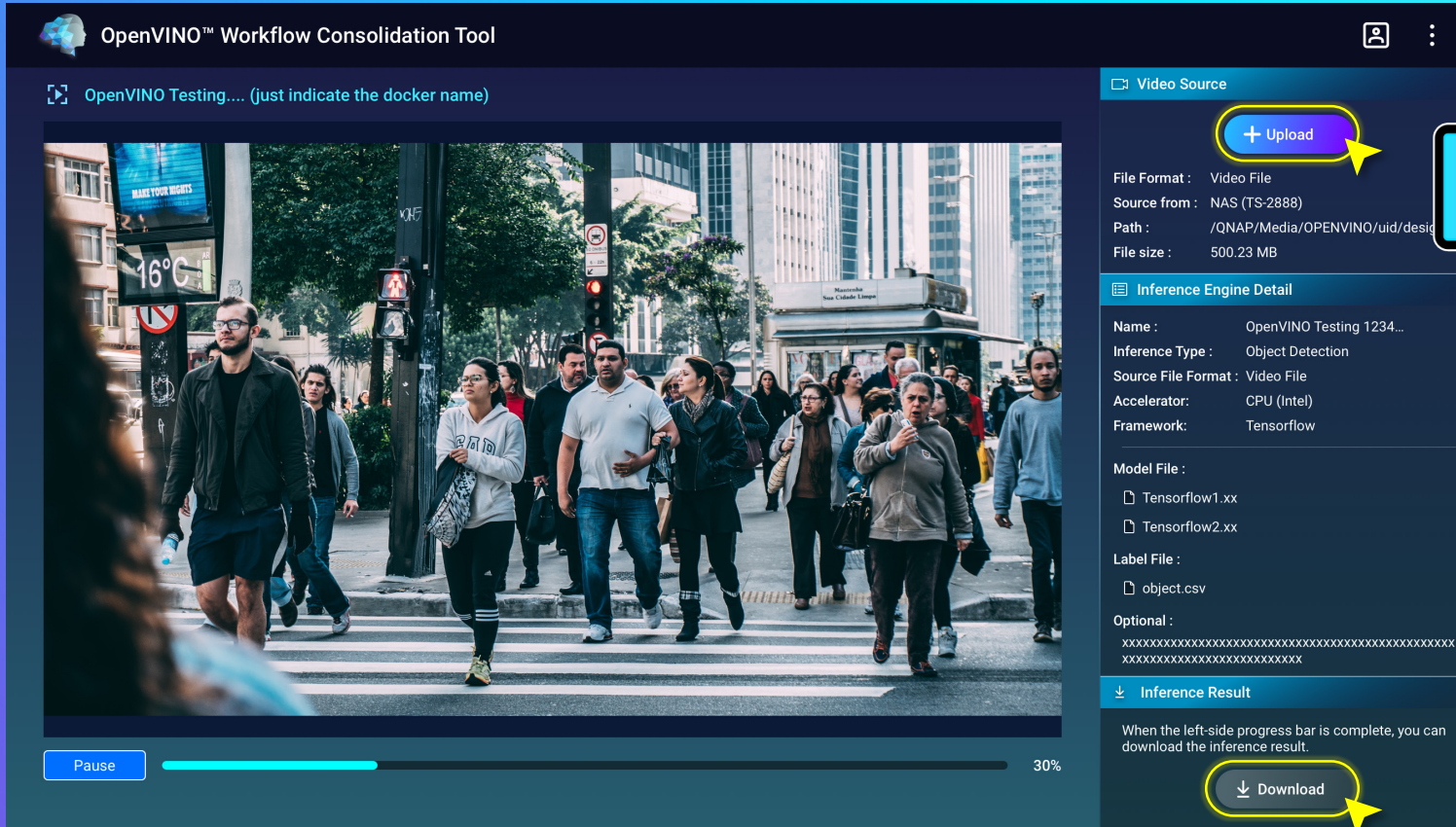
Feature 2: GUI



General setting

- Inference Type
- Model File
- Source Format
- Label File
- Accelerator
- Precision (FP32, 16, 11)

Feature 2: GUI



Inference Engine

- Player interface to demo the result
- Upload the video to inference
- Download the inference result

Feature 3: First application support the Intel® inference accelerator cards



OpenVINO™ Workflow Consolidation Tool

Welcome to OpenVINO™ Workflow Consolidation Tool

OpenVINO™ Workflow Consolidation Tool is deep learning tool to convert the trained models into an inference Engine accelerated by OpenVINO™ Toolkit. The inference engine can let you verify your inference result with the deep learning model

QNAP Instruction

OpenVINO® Toolkit Document

Hardware Accelerator Setting

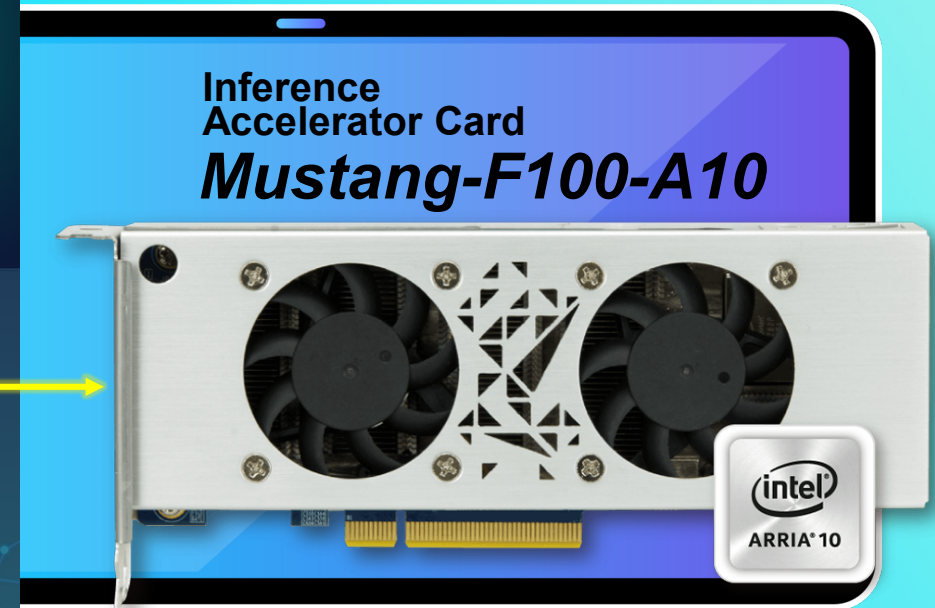
The hardware accelerator will utilize your inference engine using OpenVINO™ Toolkit. You can check your hardware accelerator status and configure it.

Type	Model Name	Status
CPU	Intel® Core i5-8400T	Working
IGD	Intel® UHD Graphics 630	None
FPGA	Mustang-F100-A10	Setting

Supported Framework

Intel® OpenVINO™ Toolkit will keep updating the latest supporting model for the different framework, please check Intel® OpenVINO™ Toolkit document when you need.

Framework	Action
Tensorflow	Document
Caffe	Document
CHNN	Coming soon



Introduction: Inference accelerator cards supported by QNAP



Accelerator CPU

Intel CPU NAS

- Intel® Xeon, Core i5, i7 (NAS with CPU)
- Power Consumption : 200 ~ 250W
- **Low Latency**
- Inference Function: Wide



Intel® CPU

Accelerator FPGA

Mustang-F100-A10

- Intel® Arria 10 GX 1150 FPGA
- Power Consumption < 60W
- **Low Latency**
- Inference Function: 2-3 function

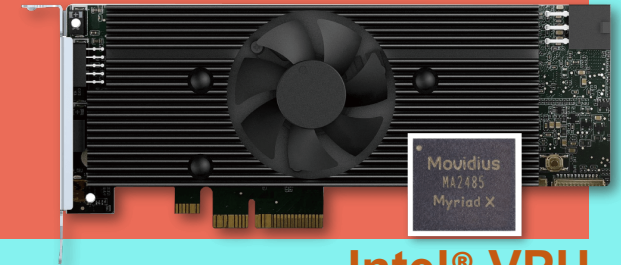


Intel® FPGA

Accelerator VPU

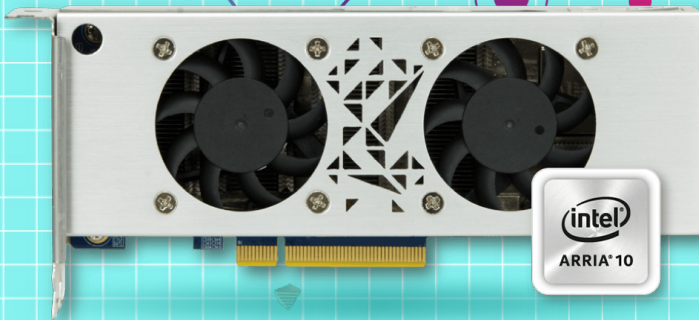
Mustang-V100-MX8

- 8 x Intel® Movidius™ Myriad™ X VPU
- Power Consumption < 30W
- **Low Latency**
- Inference Function: Focus on 1 function



Intel® VPU

Recommend equipment for DL inference



**Accelerator
Cards (Optional)**



NAS + CPU



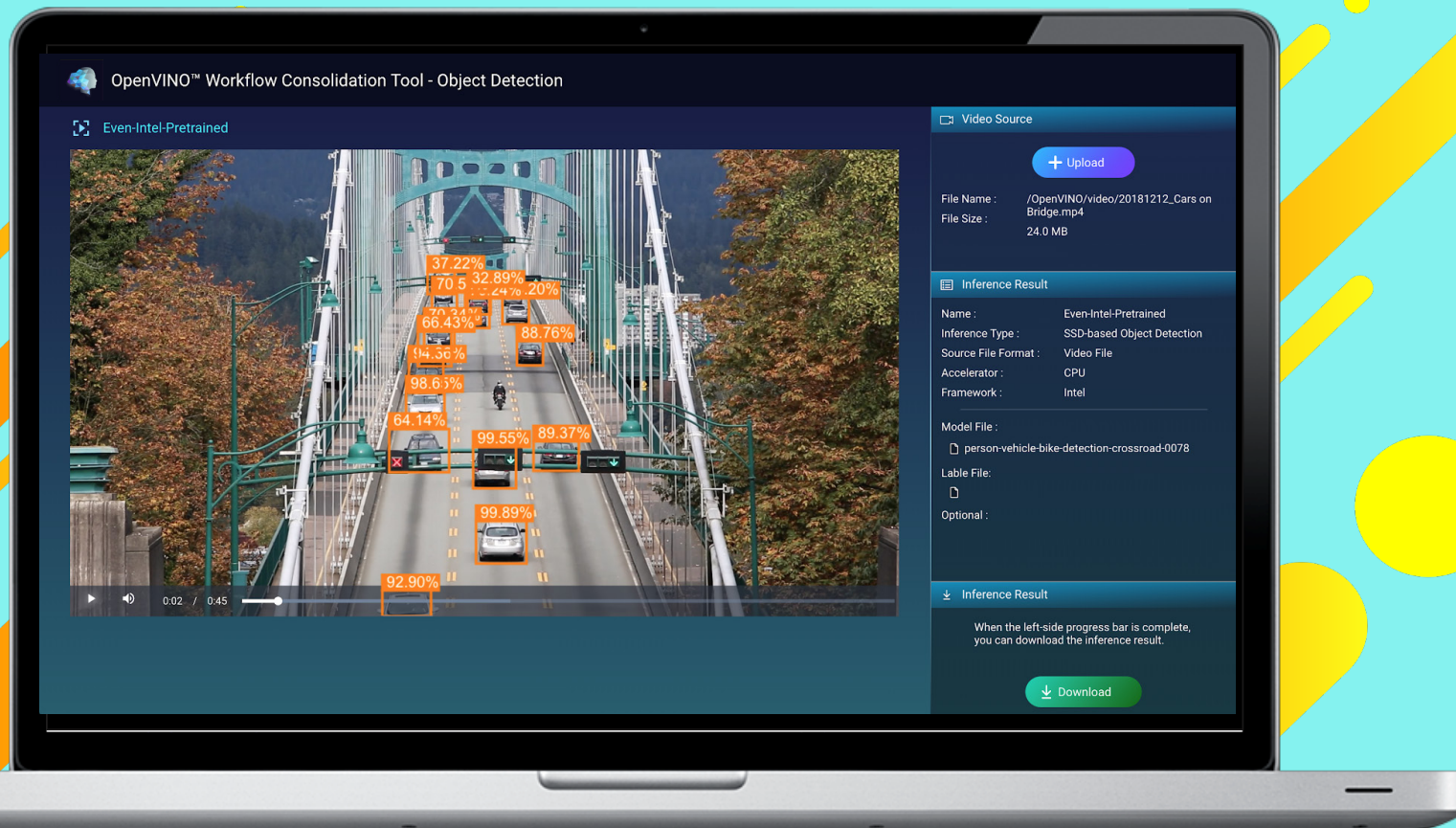
**Data & DL Model
Storage on QTS**

Demo :

Car detection and Fruit detection



Demo: Car detection (Intel® pre-trained model)



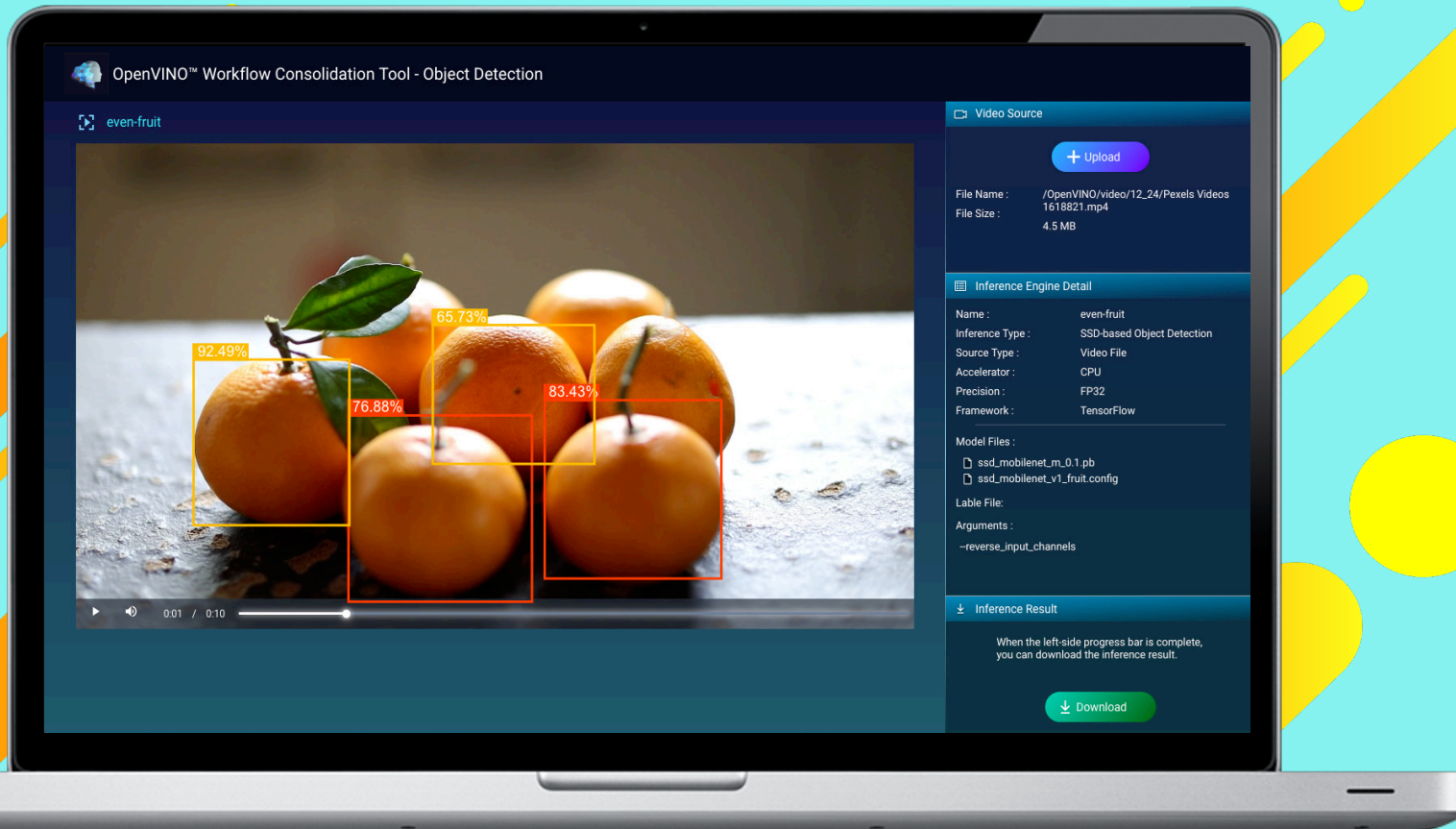
Scenario

- Monitor traffic volume

Setting

- Inference Mode: Intel re-trained model
- Sample Code: **Person-vehicle-bike-detection-Crossroad-0078**
- Accelerator Cards: CPU
- Precision: FP32

Demo: Fruit detection (Use QNAP own model)



Scenario:

- Fruit dealer collect the fruit from the farm.

Setting

Inference Mode:
Use own model

- Framework: TensorFlow
- Model: VGG-16
- Accelerator: CPU
- Precision: FP32

QNAP

is your best choice!

Thank you

