arm AI

AI Virtual Tech Talks Series

Bringing Edge AI to Life

Ali Osman Örs – NXP Semiconductors
David Steele – Arcturus
July 13, 2021

Presenters

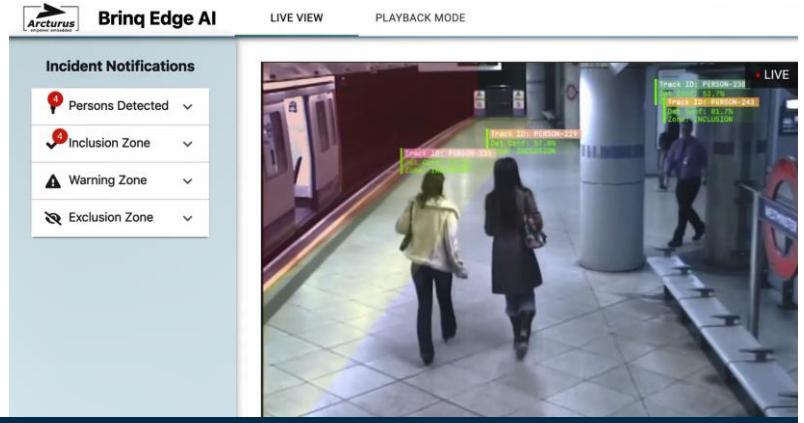
Ali Osman Örs
Director, AI ML Strategy and Technologies
NXP Semiconductors

David Steele
Director of Innovation
Arcturus

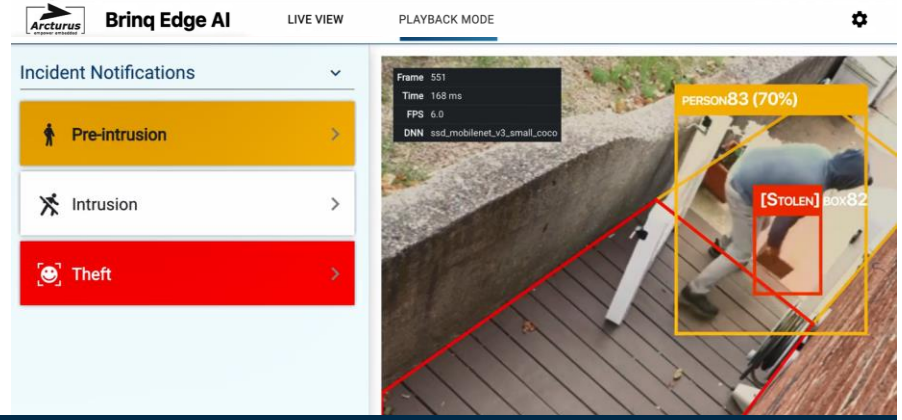
Agenda

- Introduction
- Tools and Enablement
- Building and Scaling Applications
- Optimization Techniques
- Demo
- 2x $250 Amazon Gift Card Draw

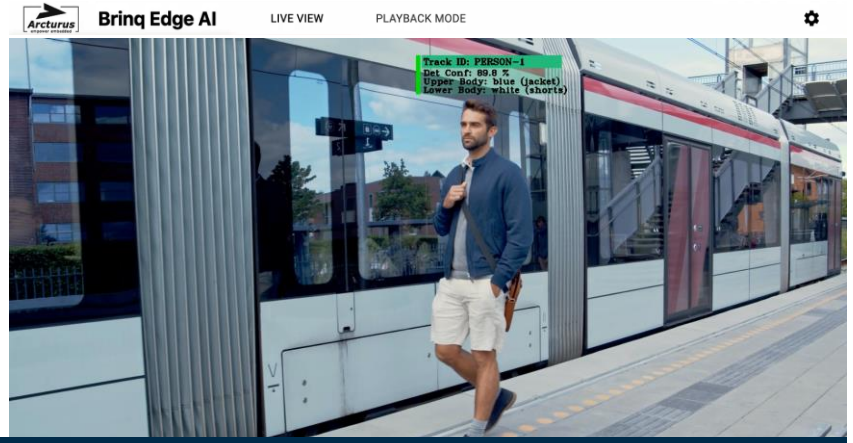
Introduction to Edge AI Applications



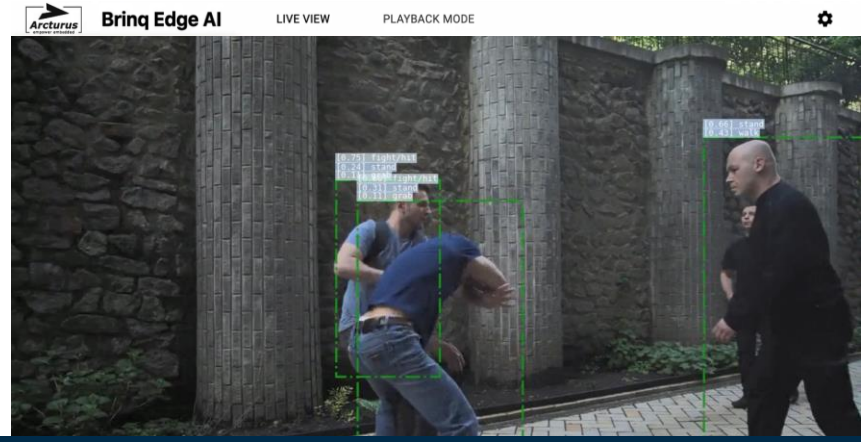
Boundary Crossing and Intrusion



Package Analytics



Characterization



Behaviour Analysis

Development Challenges

- Development paradigm
  - Experimentation and adaptation
  - Heavily reliant on optimization
  - Complex interdependencies
  - Requires broad expertise
- Starting points
  - Do not move to production seamlessly
  - Dependencies / edge runtime limitations
  - Lab vs field data
  - Missing components / specialized requirements
- Considerations
  - Move an application to the edge
  - Improve accuracy and performance
  - Develop a scalable/flexible architecture

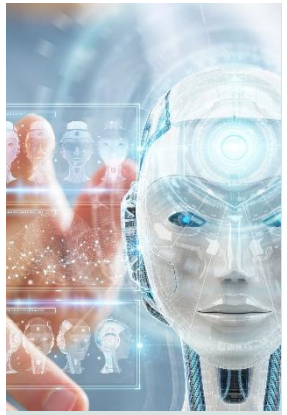
arm

Edge Enablement and Tools

Ali Osman Örs

NXP Semiconductors

NXP Broad-based Machine Learning Solutions and Support



eIQ Machine Learning SW

eIQ™ ML SW Development Environment

eIQ Toolkit with eIQ Portal GUI to:

- Import/create, convert, optimize , validate and deploy ML models
- Dataset curation tools to create new, augment, label/annotate datasets

eIQ inference with: TensorFlow Lite, TensorFlow Lite Micro, Arm NN, ONNX Runtime, Glow and DeepViewRT

Support for i.MX 8 family, i.MX RT family

Integrated into NXP development environments (MCUXpresso, Yocto/Linux)

DIY

eIQ Auto

eIQ™ Auto AI Enablement

Deep Learning toolkit for S32V processors

Auto Quality : A-SPICE qualified inference engine

Optimization: Prunes, quantizes, compresses the Neural Network

Automated neural net layer deployment to optimum available compute resource

Automotive Grade

CORAL

Third Party SW and HW

Google Coral Dev Board

i.MX 8M Mini Development Kit for Amazon® Alexa Voice Service

Au-Zone Value-add packages for NXP eIQ Toolkit

Arcturus video applications

SensiML tools for sensor analysis

…. And more

SLN-ALEXA-IOT

Turnkey Solutions

Alexa Voice Services (AVS) solution

- i.MX RT106A (kit – SLN-ALEXA-IOT)

Local voice control solution

- i.MX RT106L (kit – SLN-LOCAL-IOT)

Face & emotion recognition solution with Anti-Spoofing

- i.MX RT106F (kit – SLN-VIZN-IOT)

Fully Tested

NXP Scalable Edge Processing Continuum

Scalable/reusable software, tools and solutions

Arm Cortex-A and Cortex-M cores
APPLICATIONS PROCESSORS
S32x | i.MX | Layerscape
Voice / Audio Graphics Video ML Networking 5G Access

CROSSOVER
i.MX RT
i.MX ULP

Arm Cortex-M cores
MCUs
LPC | Kinetis | S32K
Motors, Control, Sensors, Power, Wireless Connectivity

Future i.MX 9 MPUs

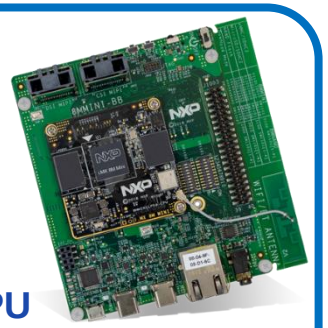
Example Applications:

- Smart Cities
- Smart Homes
- Smart Buildings
- Smart Factories

Cortex-A cores & Cortex-M cores
First use of Arm Ethos-U65 NPU

Example Applications:

- Machine vision
- Industrial computer
- Security and vision



i.MX 8M Plus MPU

Quad/Dual Cortex-A53 and M7
2.3 TOPS NPU, GPU, DSP
Camera, display, video, audio
CAN-FD, GbE, TSN, PCIe

With Industrial-grade reliability, security, longevity

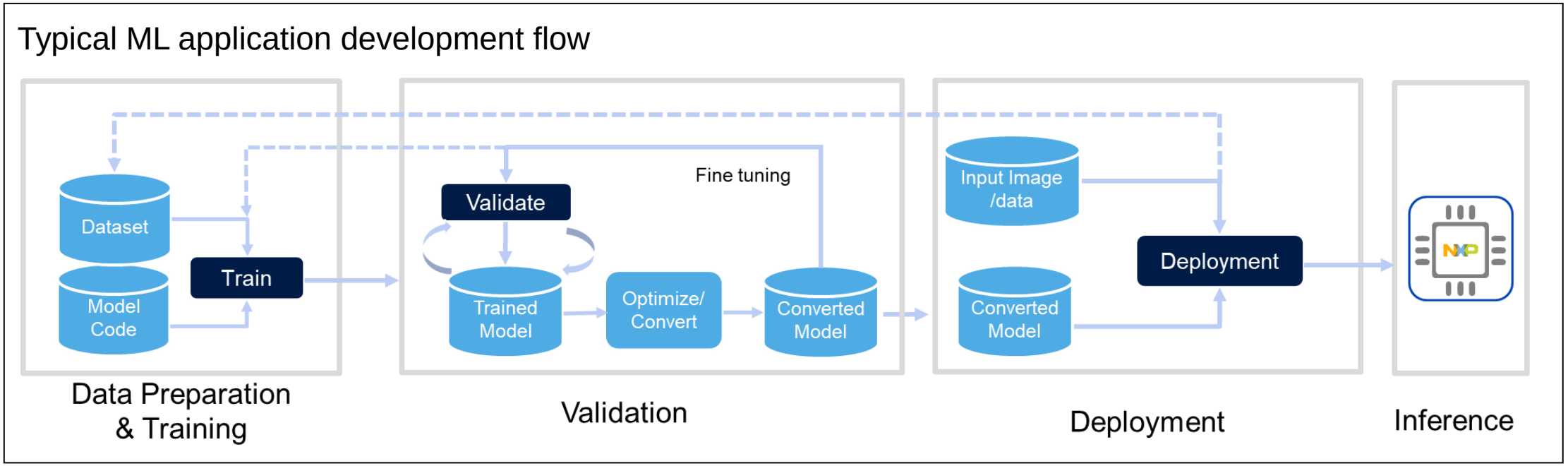
Machine learning use cases and accelerators

Cortex-M | 4x Cortex-A53 | GPU | Neural Processing Unit (NPU)

0 | 0.25 | 0.5 | 1 | 2 | 3 (TOPS)

Vision	Face and still image recognition, person detection (images)	Multi-face recognition, object detection (video)	Live video face and object recognition	Multi-object surveillance (people, cars, animals)	
Audio	Wake word, 10 Word speech, speaker recognition	Automatic speech recognition (basic command phrases)	40,000 words vocabulary, multiple speaker recognition	Speech accents interpretation	
Time Series Data	Anomaly detection (environmental sensors)	Pose estimation	Gesture recognition	Complex real-time motion analysis	

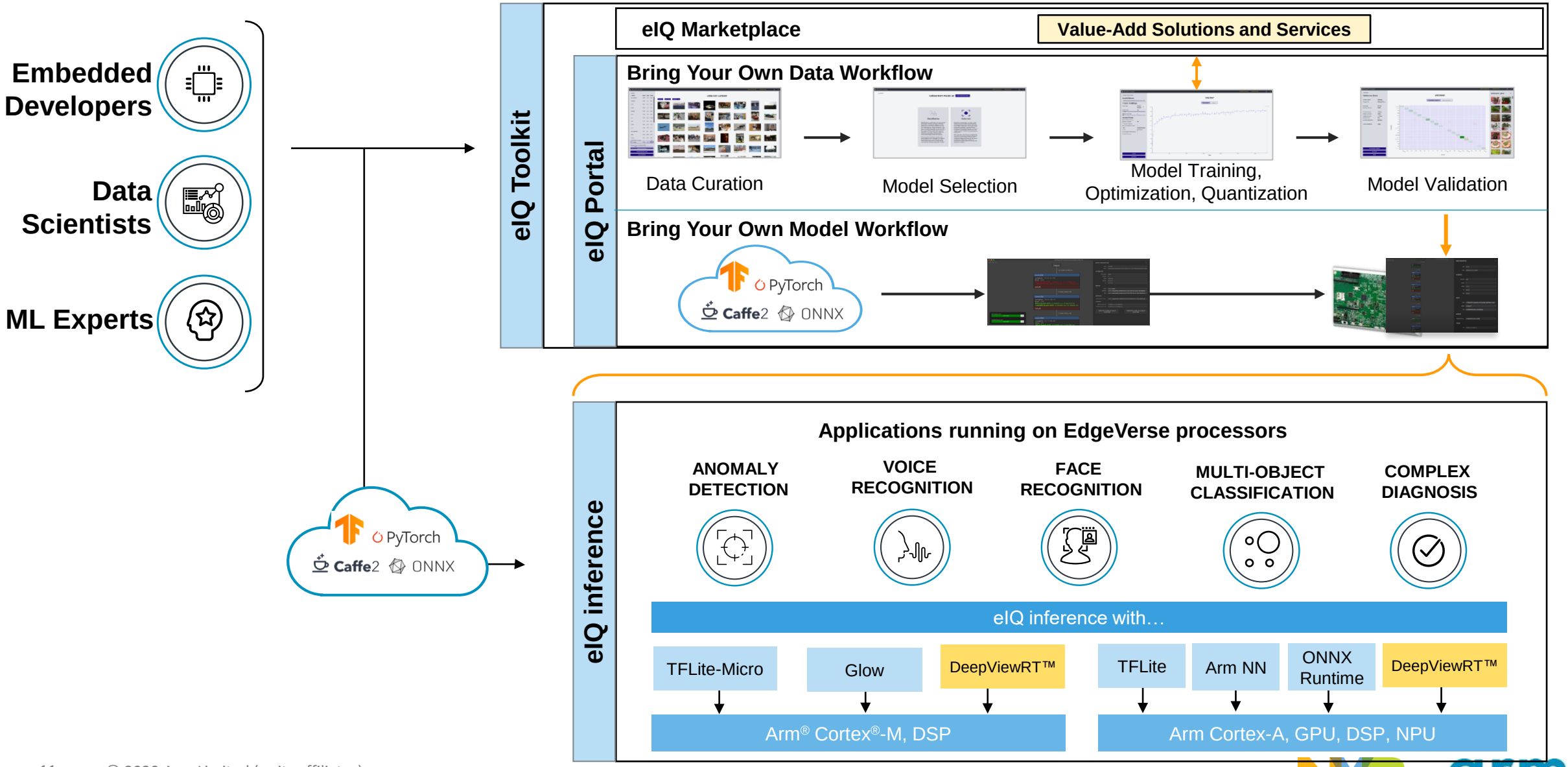
eIQ Machine Learning SW Development Environment



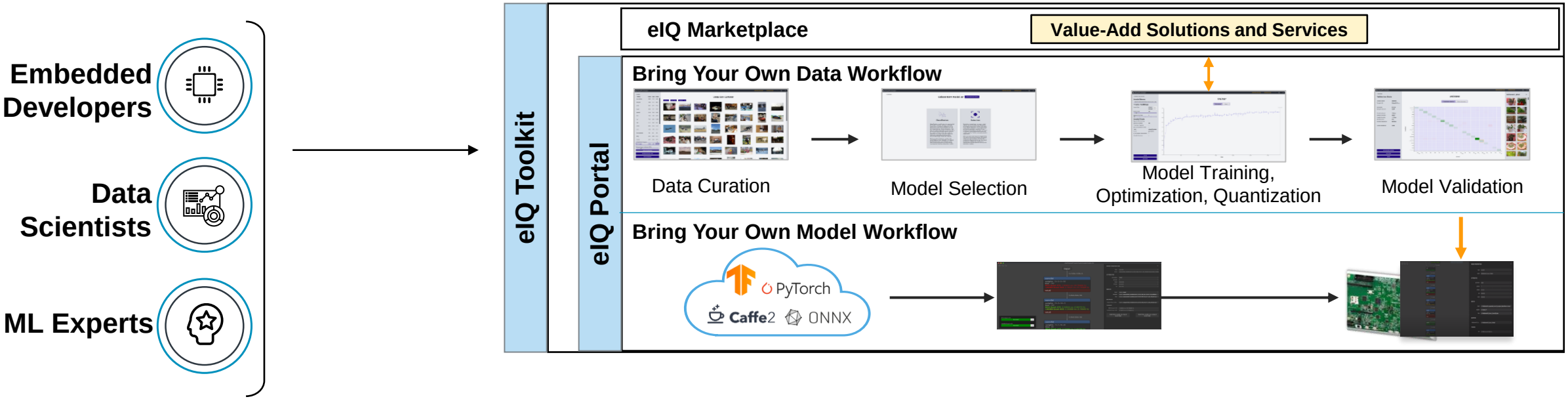
NXP's **eIQ ML Software** provides a collection of development tools, utilities and libraries for building ML applications using NXP MCUs and applications processors (MPUs).

eIQ ML software can be leveraged as part of a user's existing flow or can be used for the complete flow depending on the ML application targeted.

eIQ™ ML SW Development Environment



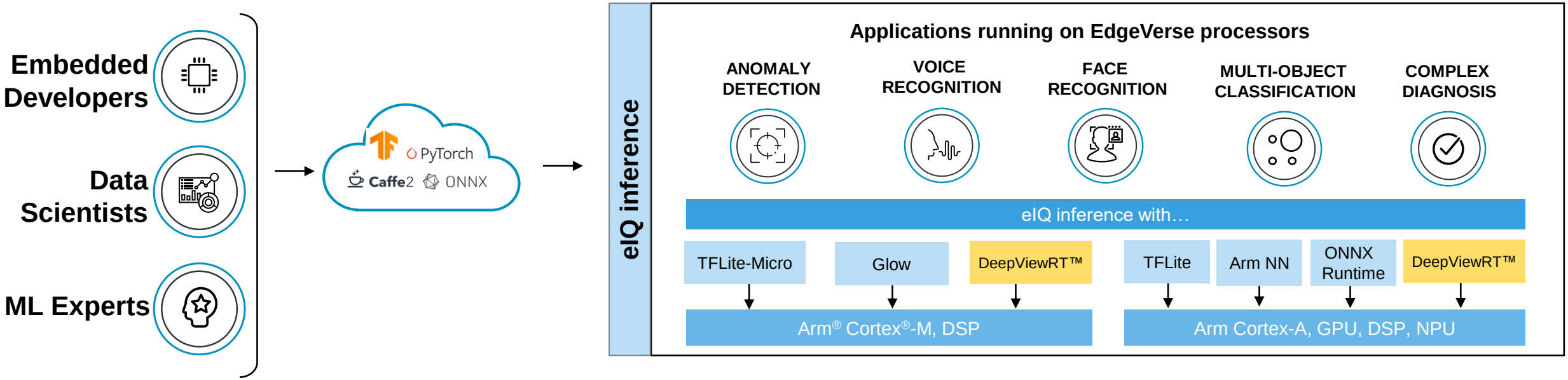
eIQ Toolkit and eIQ Portal



eIQ Toolkit

- Eases the ML development experience with the **eIQ Portal** as well as with command line host tools like; Glow tools, TensorFlow, PyTorch and other third-party tools
- Enables graph-level profiling capability with runtime insights to help optimize neural network architectures for execution on target NXP processors
- Includes ML application examples
- **eIQ Portal** intuitive graphical user interface (GUI) that simplifies ML development:
  - Creates, optimizes, debugs, converts, and exports ML models
  - Imports datasets and models, rapidly trains and deploys neural network models and ML workloads
  - Output seamlessly feeds into DeepViewRT, TensorFlow Lite, TensorFlow Lite Micro, Glow, Arm NN, and ONNX Runtime inference engines
  - Can import models from TensorFlow and PyTorch ML frameworks
  - Includes object detection and image classification models for computer vision applications
- **eIQ Marketplace** offers value-add solutions, professional support and design services from trusted eco-system partners and NXP
- Delivered with a single click from the eIQ Toolkit at www.nxp.com/eIQ

eIQ inference



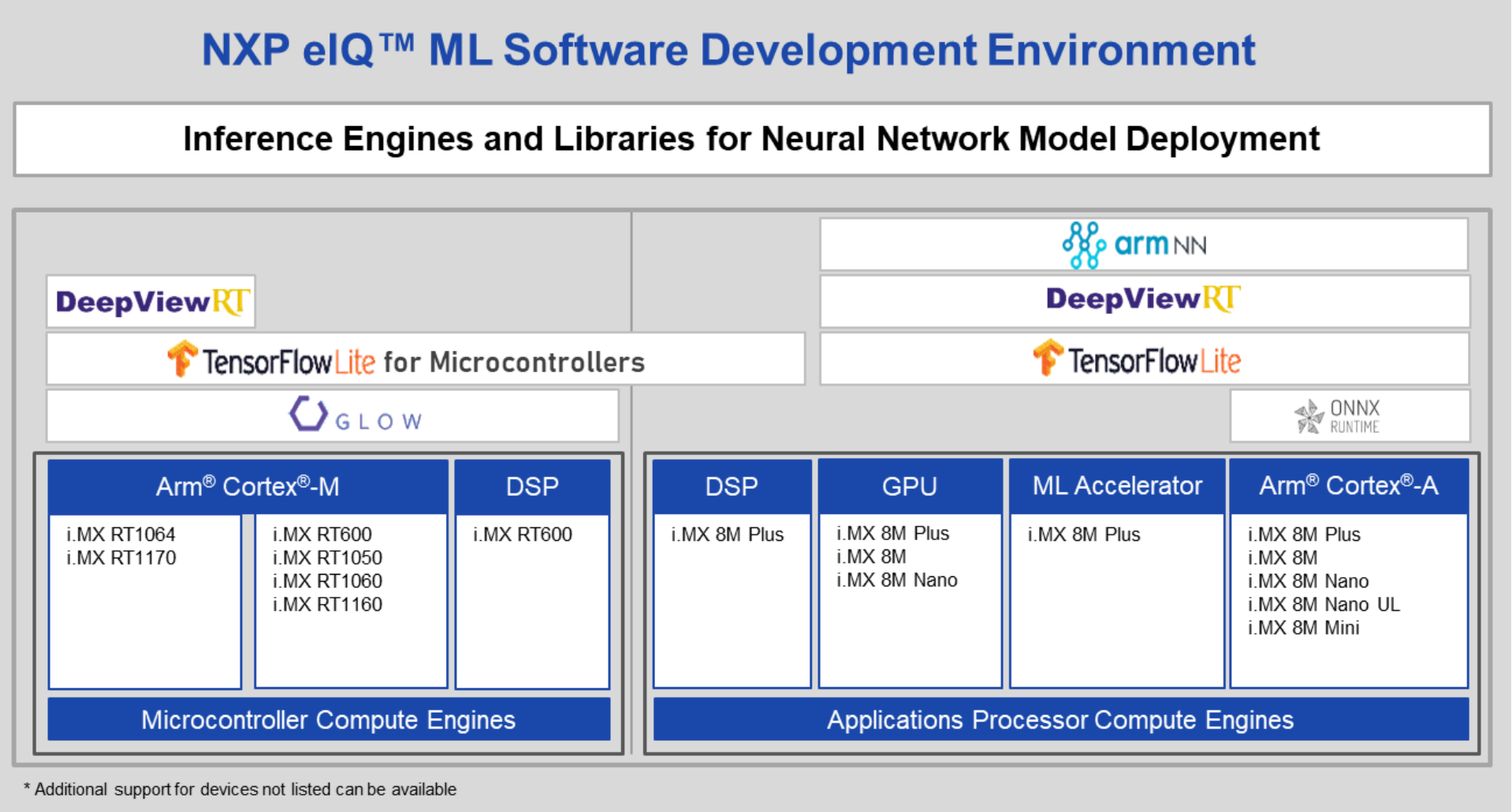
eIQ inference with TensorFlow Lite, ARM NN, Glow and ONNX Runtime

- Optimized support for open community-based engines on NXP target devices

eIQ inference with DeepViewRT runtime

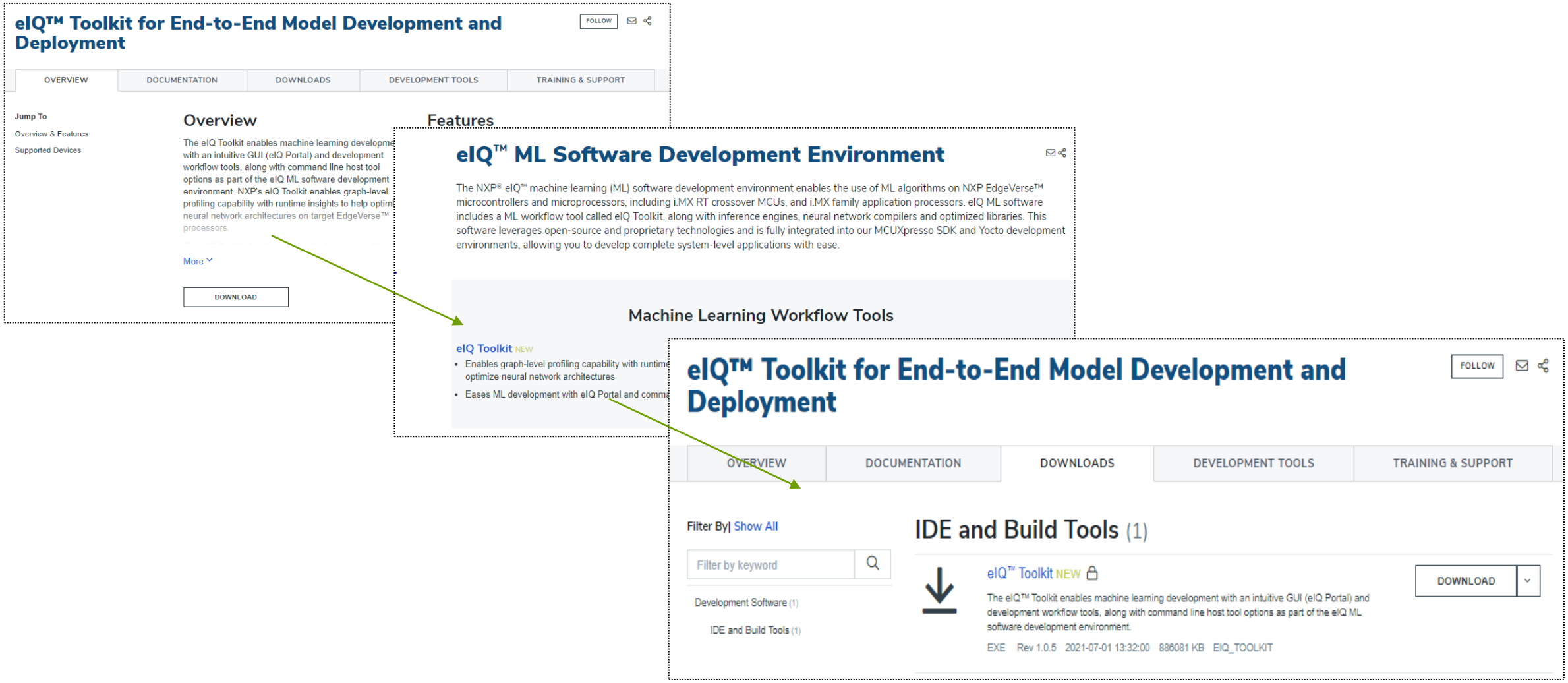
- Platform-optimized, proprietary runtime inference engine that scales across a wide range of NXP devices and neural network compute engines
- Stable and longer-term maintained solution to complement the open community-based inference solutions.
- Supports EdgeVerse™ processors, including the i.MX RT crossover MCUs (Arm® Cortex®-M cores), i.MX applications processors (Cortex-A and Cortex-M cores, dedicated Neural Processing Units (NPU) and GPUs)
- Enables compact code size for resource-constrained devices with ease of analysis and fine-tuning of model performance using the eIQ Toolkit
- Delivered via NXP standard Yocto BSP release for Linux® OS-based development, and MCUXpresso SDK release for embedded RTOS-enabled MCU development
- eIQ inference with DeepViewRT runtime is provided free-of-charge to NXP customers as part of our ML enablement

NXP eIQ™ ML Software Development Environment

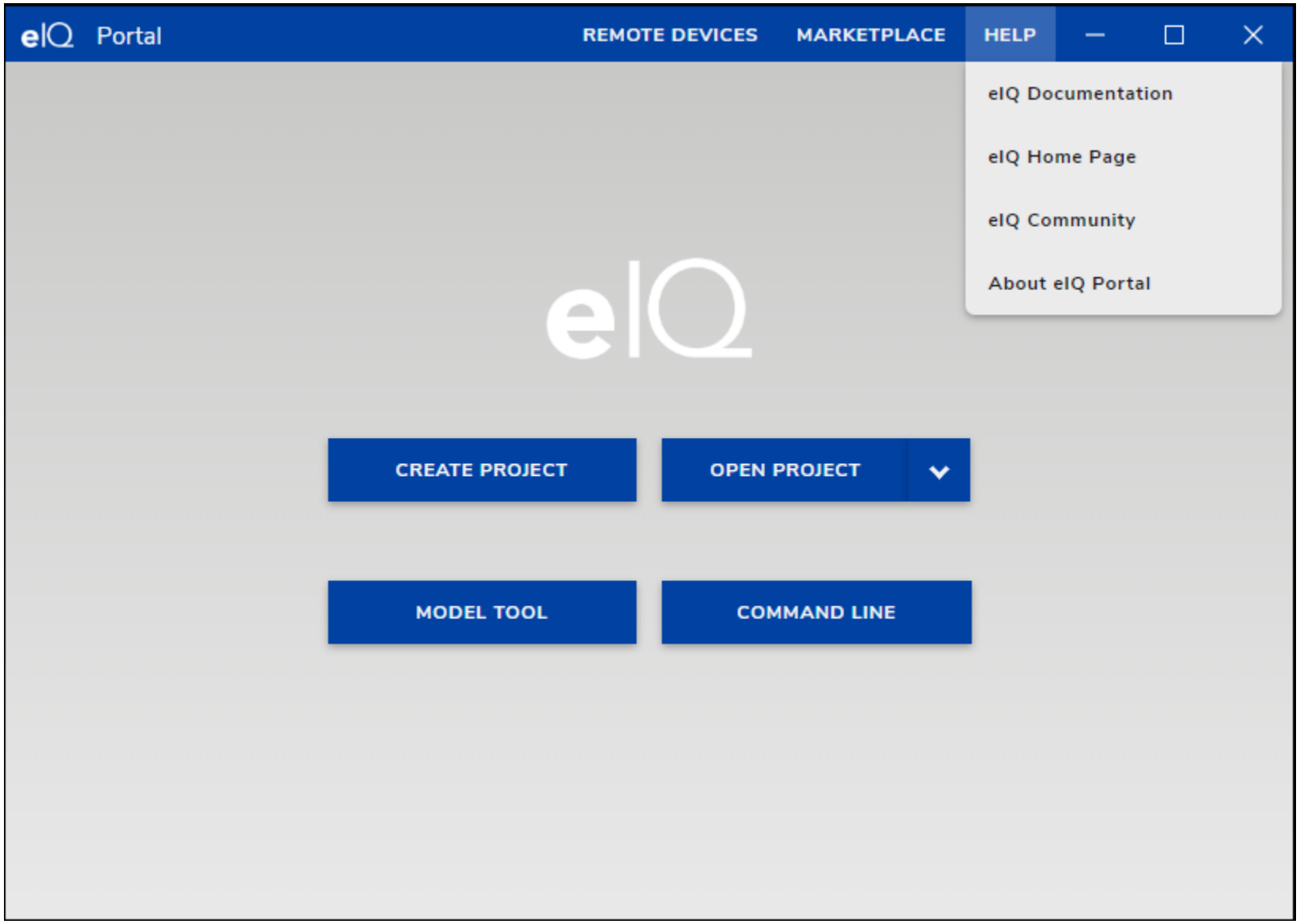


eIQ Toolkit availability

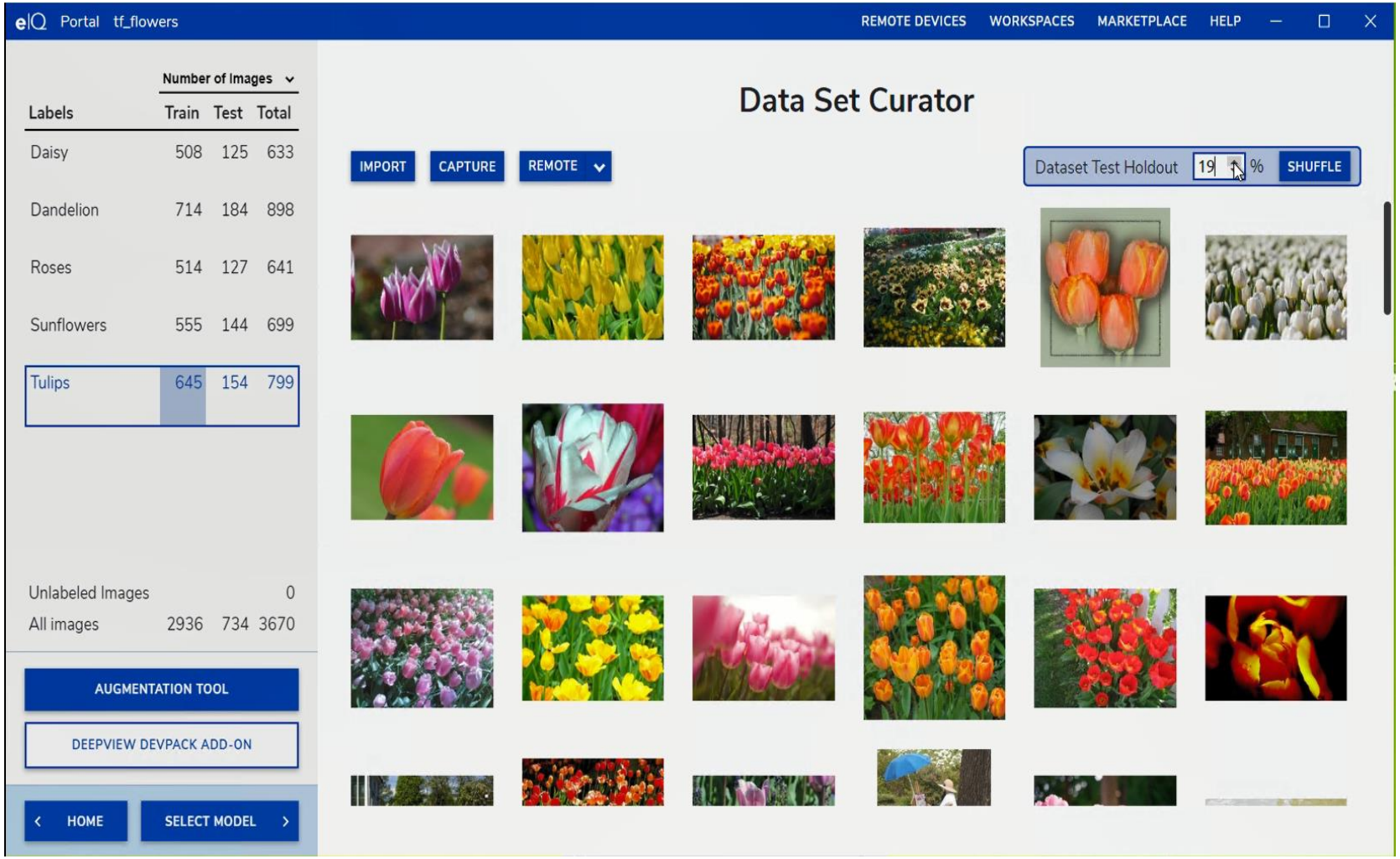
Registered users can download eIQ Toolkit from: http://www.nxp.com/eiq



eIQ Portal

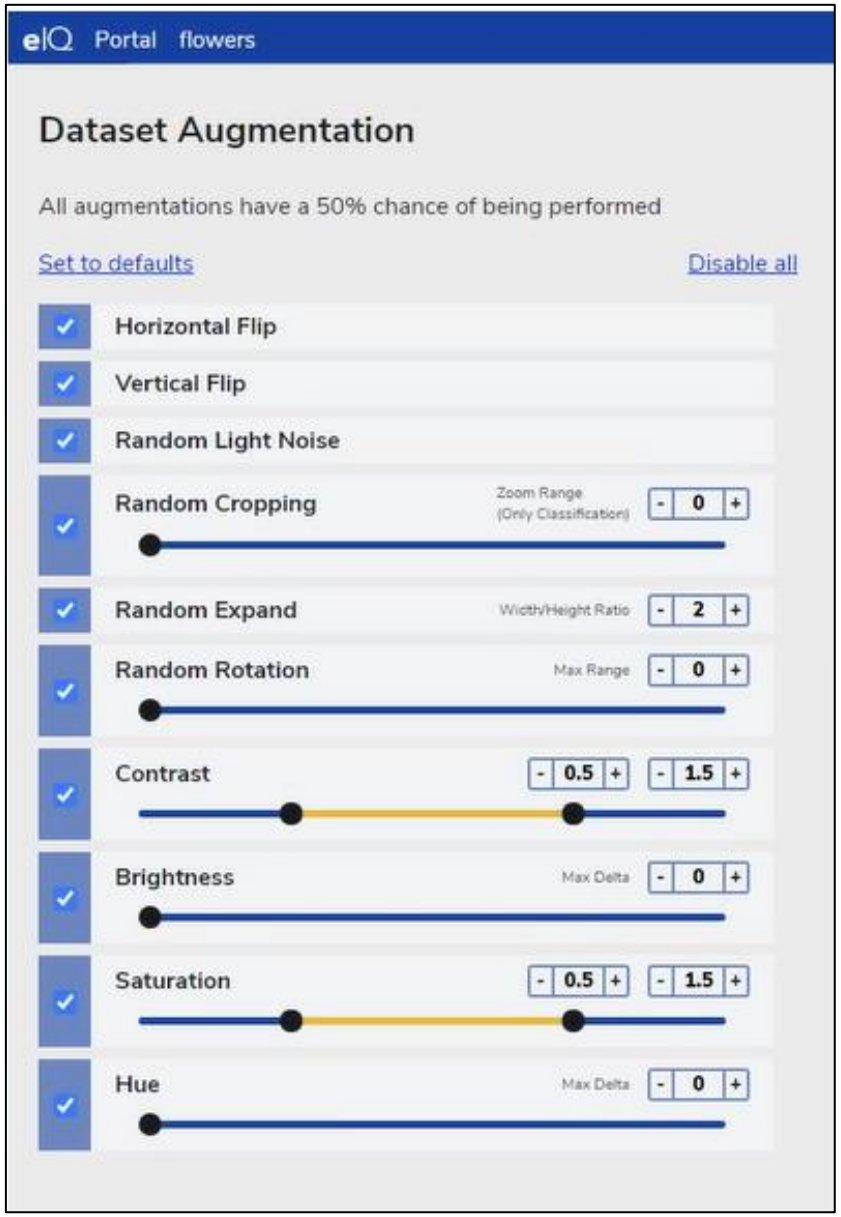


Data Curation



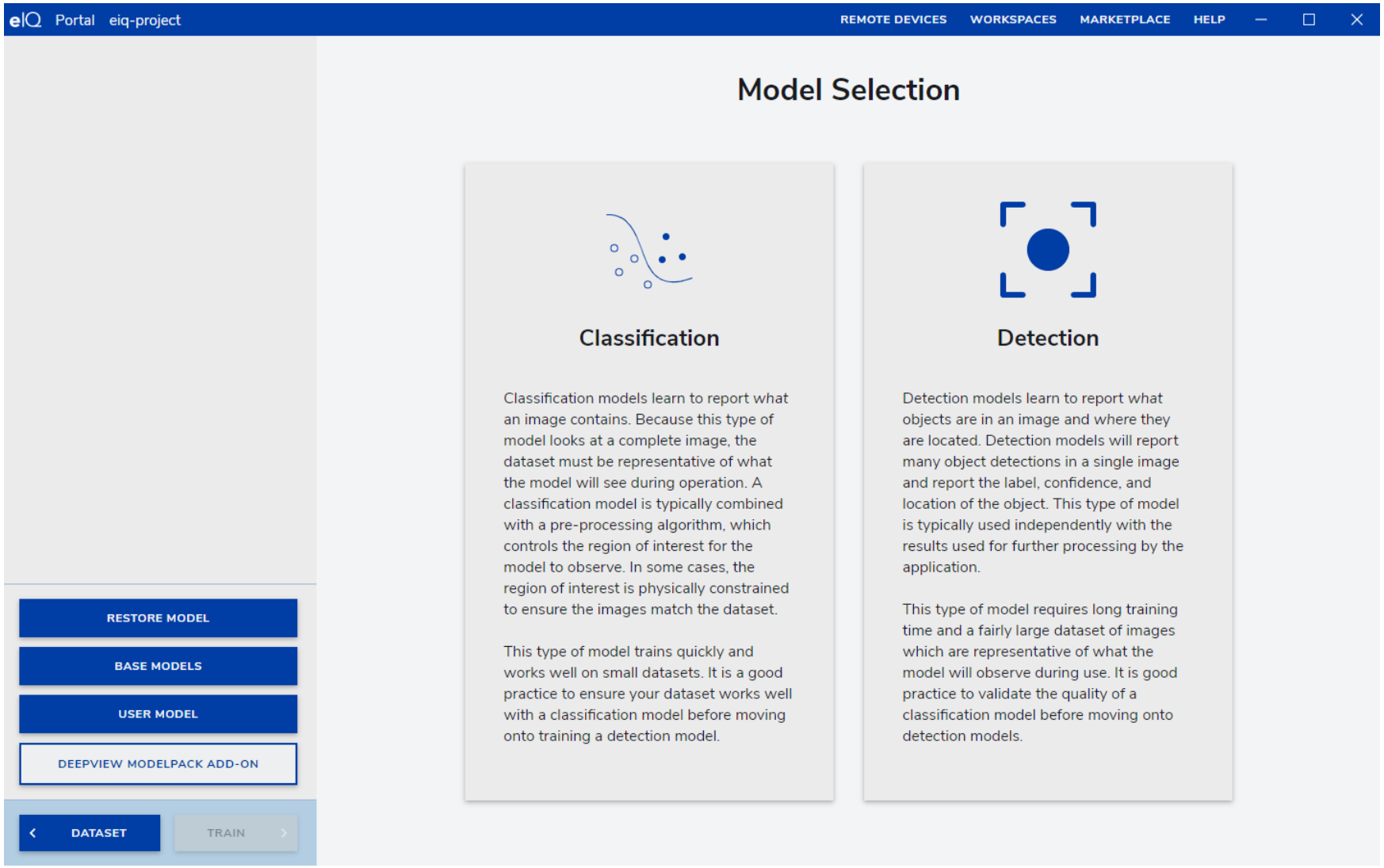
- Capture and annotate images for model training and validation
- Import datasets from public or user-defined formats
- Distribute data for training and testing

Dataset Augmentation



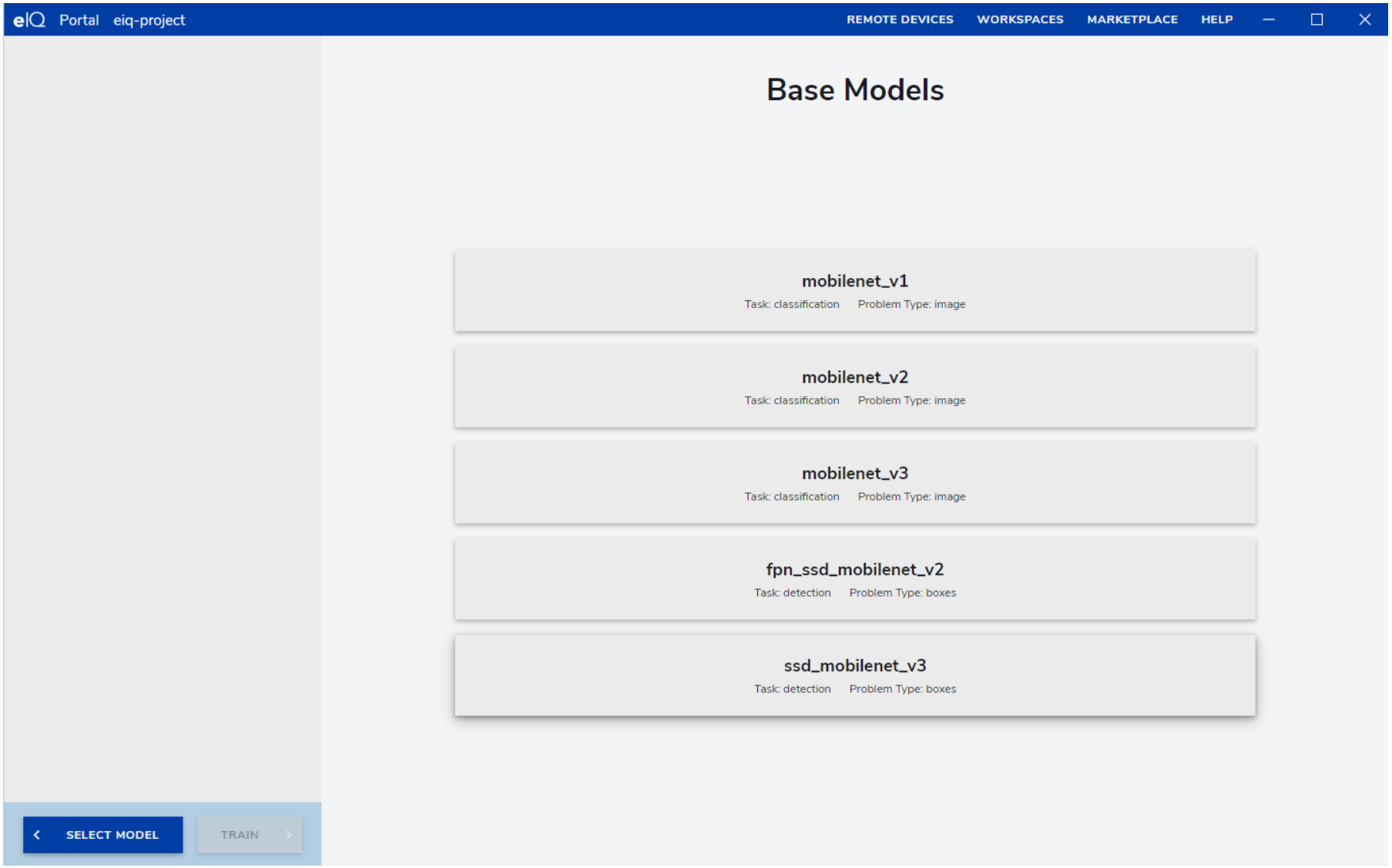
- Dataset augmentation adjusts image parameters to improve model training
- Reduce over-fitting and increase robustness to dynamic real-world environments
- Visualize how augmentation parameters affect images
- Set augmentation parameters relative to the application

MODEL TOOLs

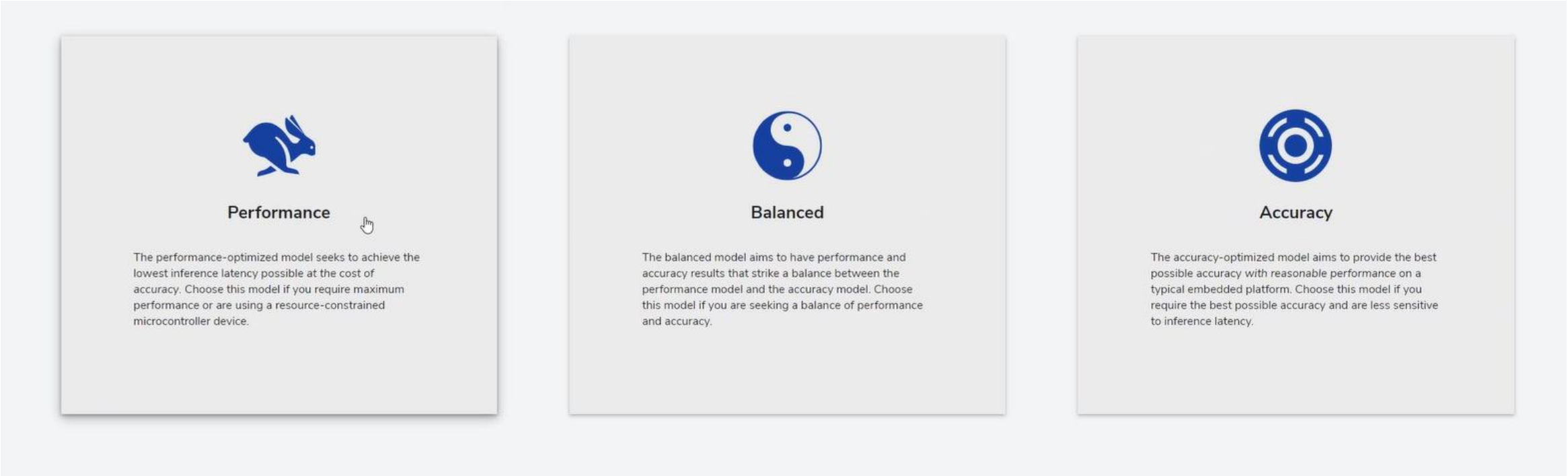


- Select the appropriate class of model:
  - Classification, Detection

MODEL TOOLs

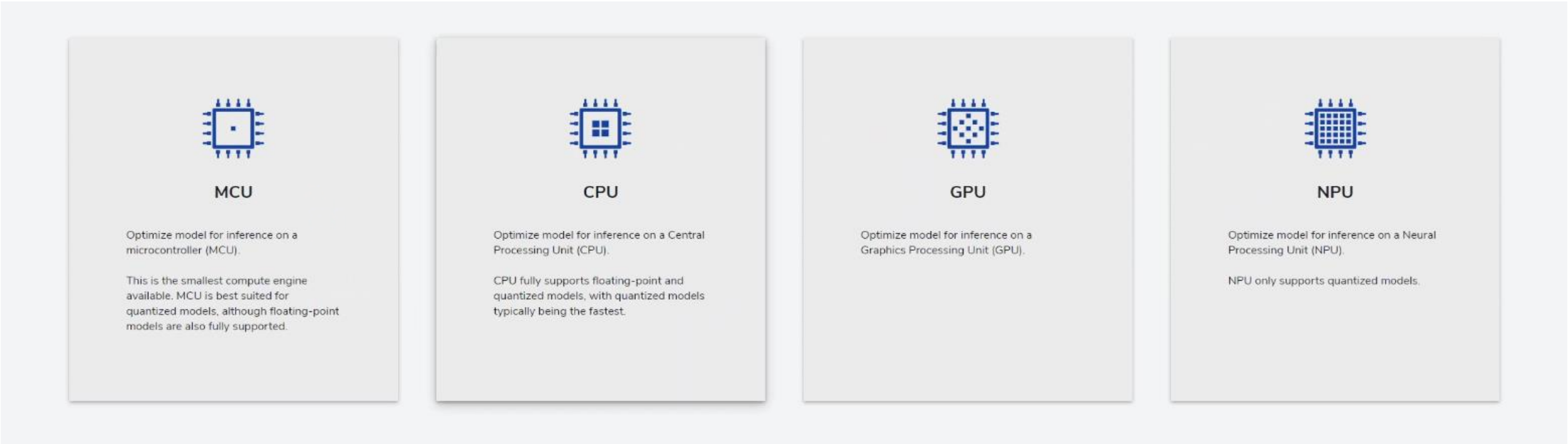


MODEL TOOLs



- Choose a model version that matches performance and accuracy requirements

MODEL Tools



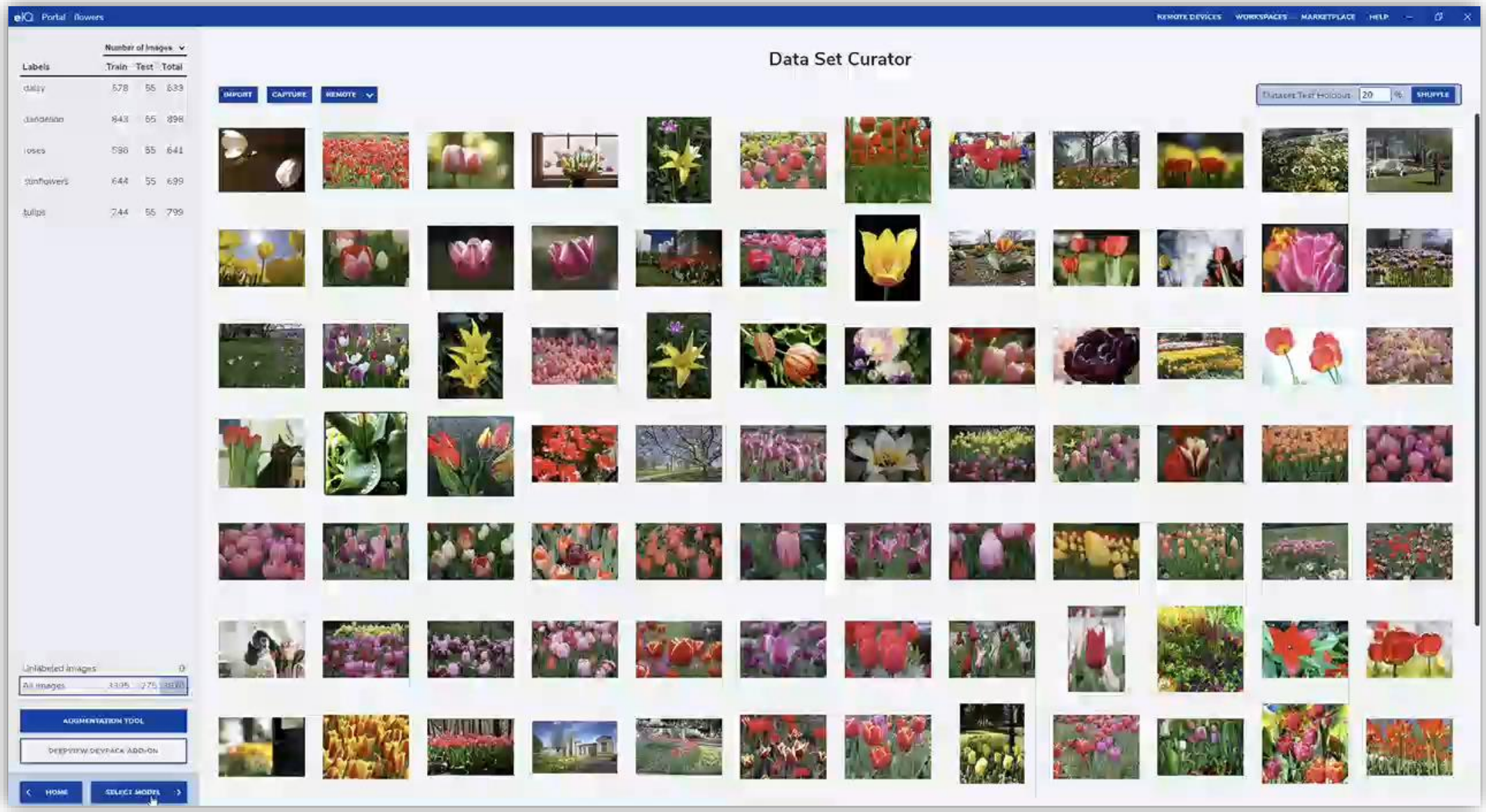
- Model selected must be correct fit for target resources
  - Application Processors or MCUs
  - Models further optimized for compute units* (MCU, CPU, GPU, NPU)

* Planned future functionality

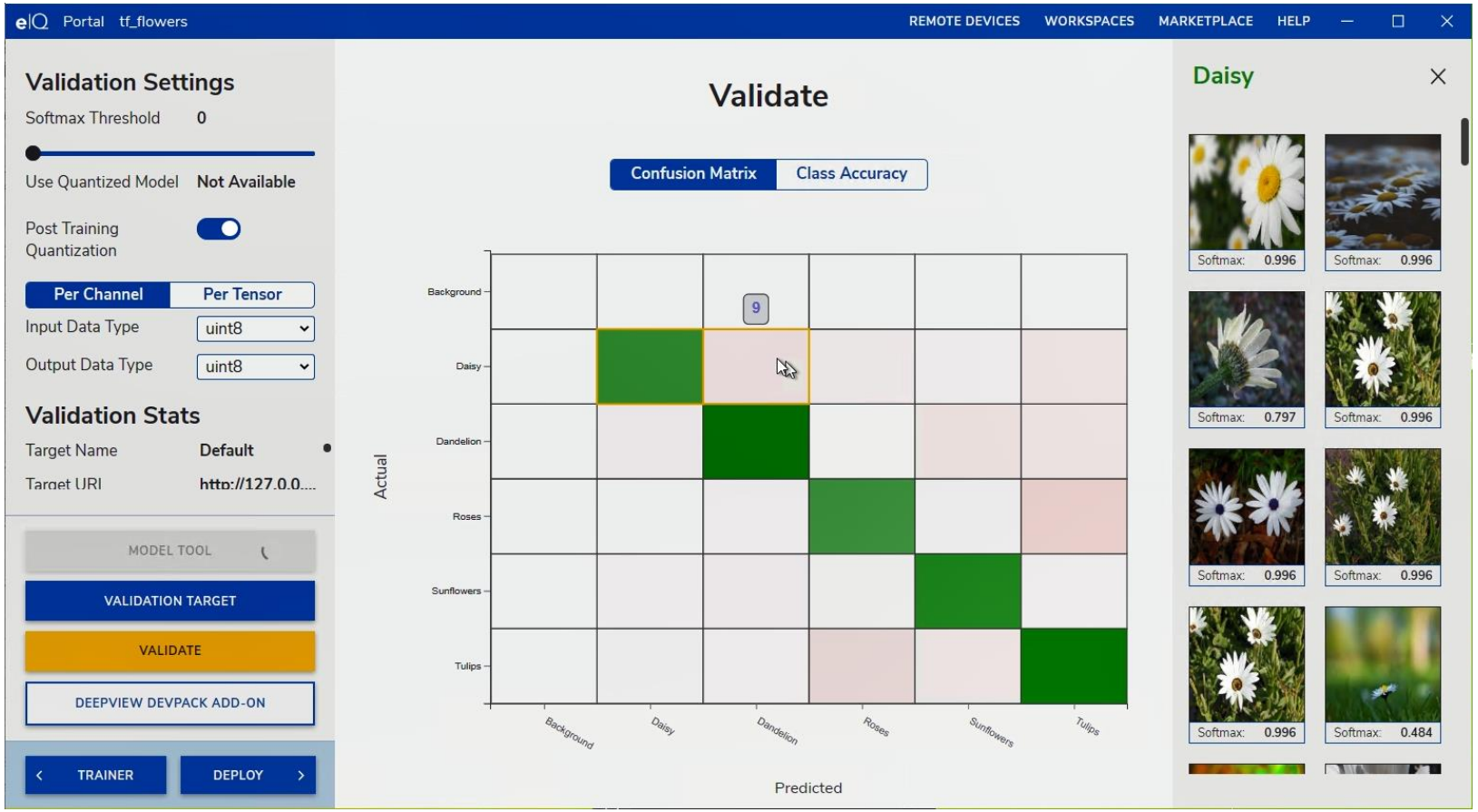
Model Training and Optimizing for BYOD and BYOM

- Model training should support users of all levels of experience
- Use default settings or fine-tune with hyperparameter selection

- Static model analysis for debug and bottleneck detection
- Automatic graph-level optimizations should improve performance and memory utilization without precision loss (e.g., pruning, fusing, layer folding)
- Configurable optimizations control accuracy tradeoffs (e.g., quantization, layer replacement, weight rounding)
- Quantization converts 32-bit floating point models to 8-bit integer format



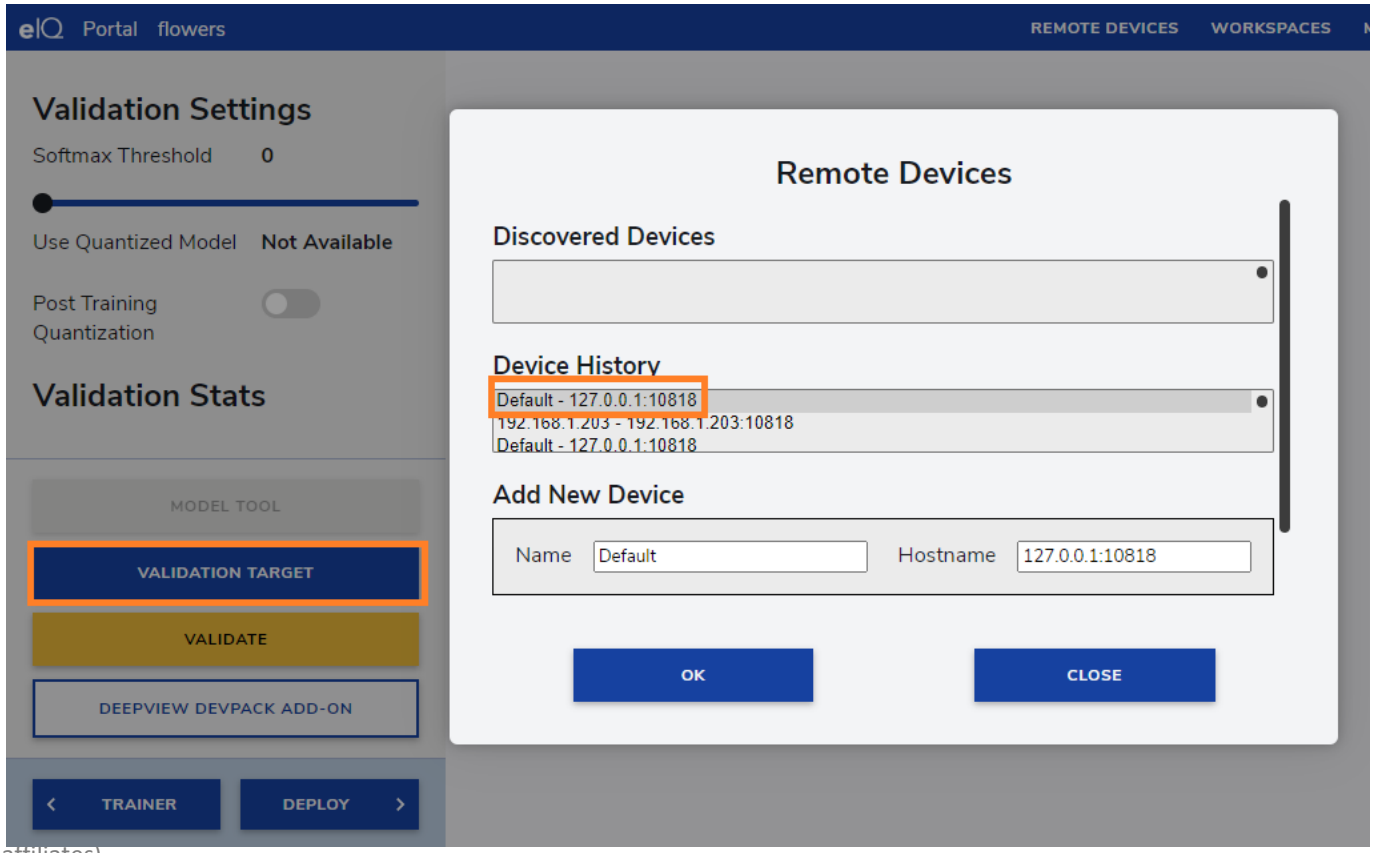
Validation to Prove and Improve Model Behavior



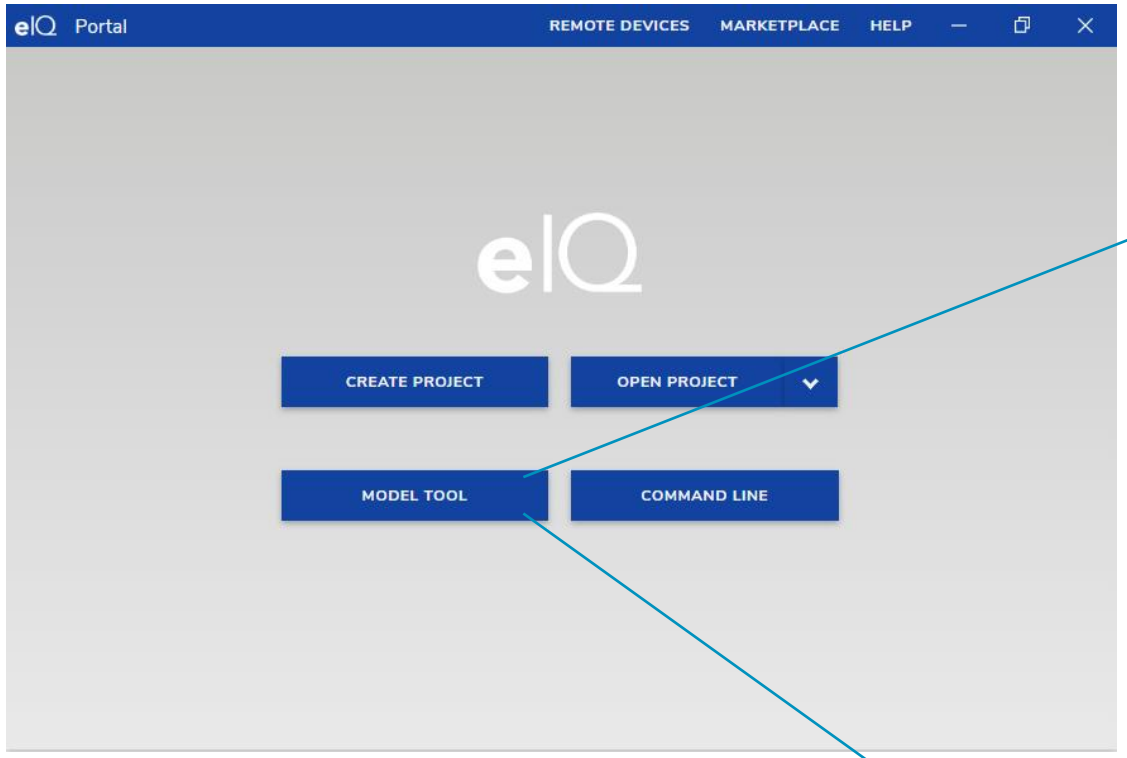
- Use validation to uncover areas that can be improved in your data set
  - Analyze and compare model accuracy running different optimizations

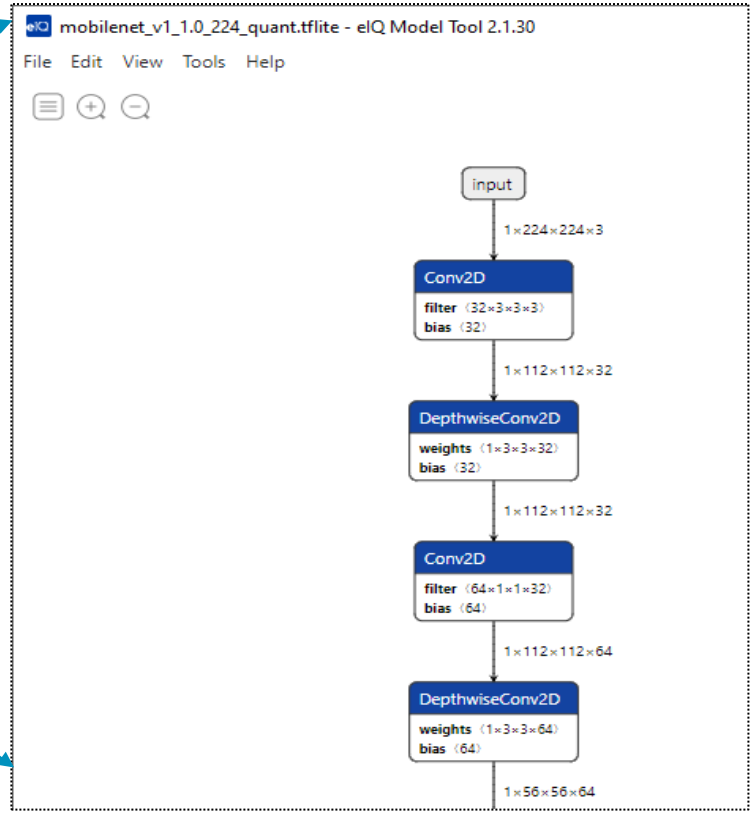
Validation target

- By default you will validate the model on your local PC
- If you have a Remote Device connected you can validate on target and get profiling information



BYOM: eIQ Model tool

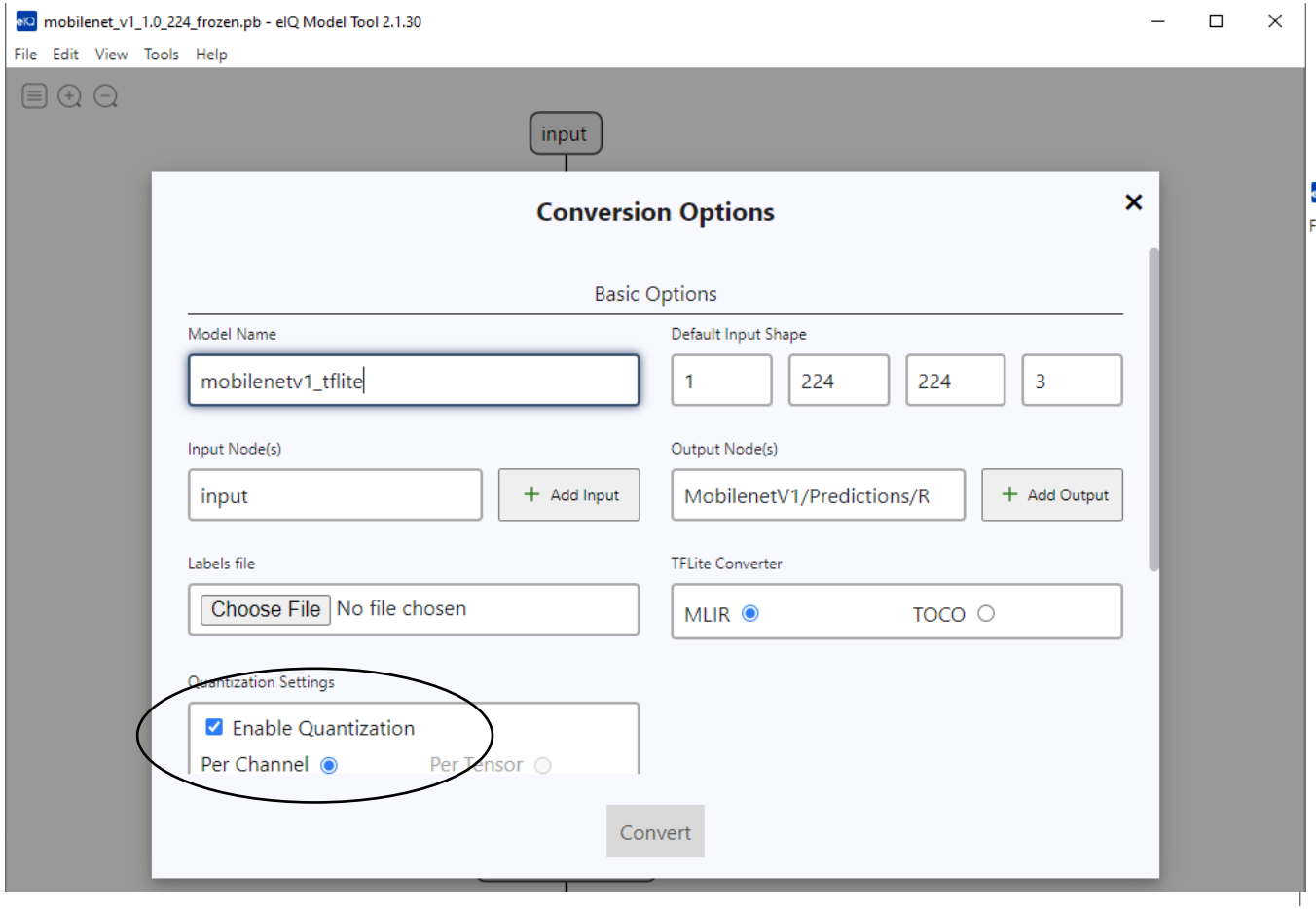


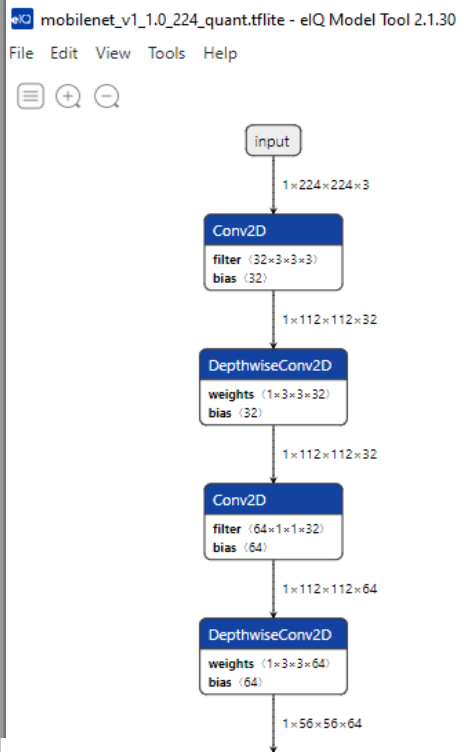


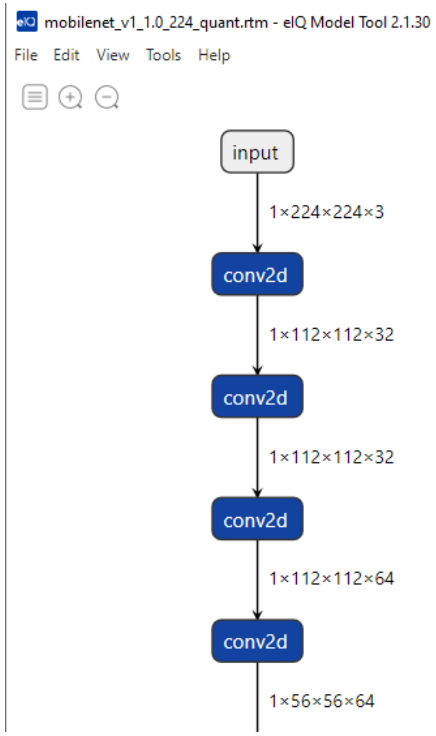
Model converter support

Source/Destination	Float			Quantized		
	DeepviewRT RTM	TensorFlow Lite	ONNX	DeepviewRT RTM	TensorFlow Lite	ONNX
TensorFlow 1.x pb	Yes	Yes	Yes	No	Yes	Yes
Saved Model (Folder/tar)	Yes	Yes	Yes	Yes	Yes	Yes
Keras (.h5)	Yes	Yes	Yes	Yes	Yes	Yes
DeepviewRT RTM	-	No	No	-	No	No
TensorFlow Lite (tflite)	Yes	-	Yes	No	No	Yes
ONNX	Yes	Yes	-	No	Yes	Yes
TensorFlow Lite Quantized	Yes	-	Yes	No	-	-
ONNX Quantized	Yes	Yes	-	No	-	-

eIQ Model Converter







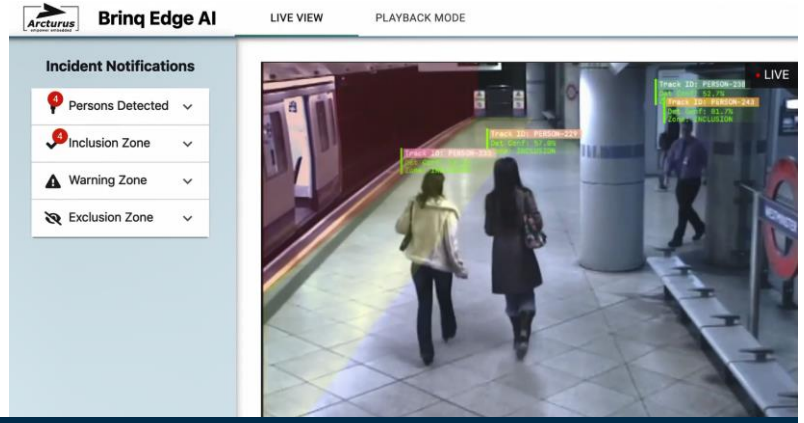
arm

Building and Scaling Applications

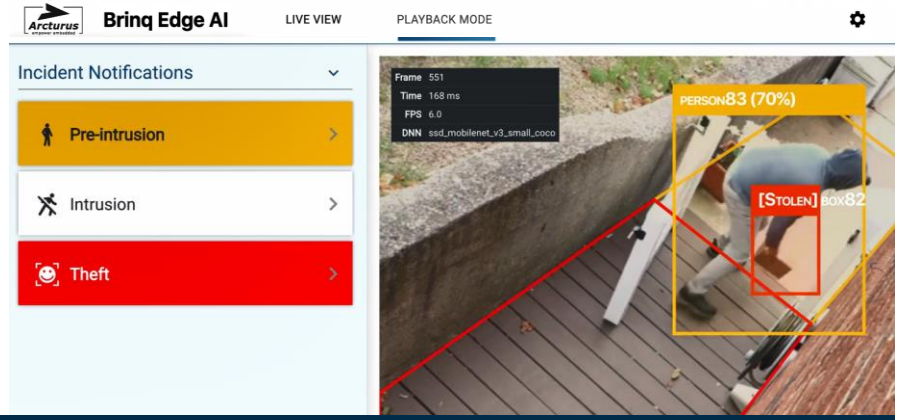
David Steele

Arcturus

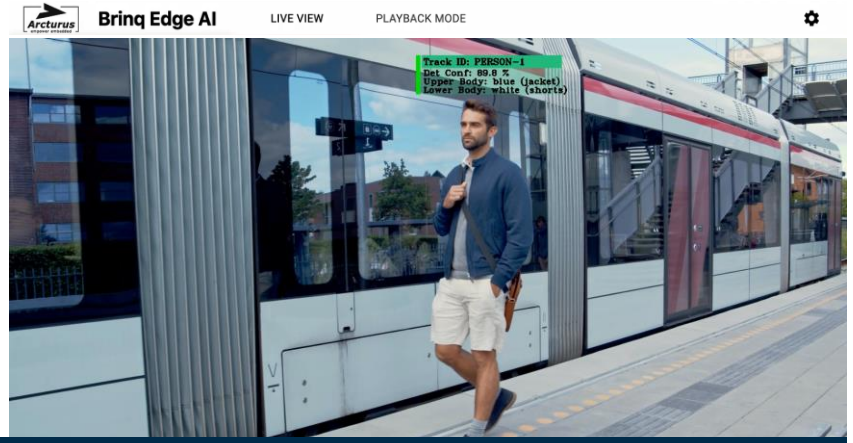
Application Examples



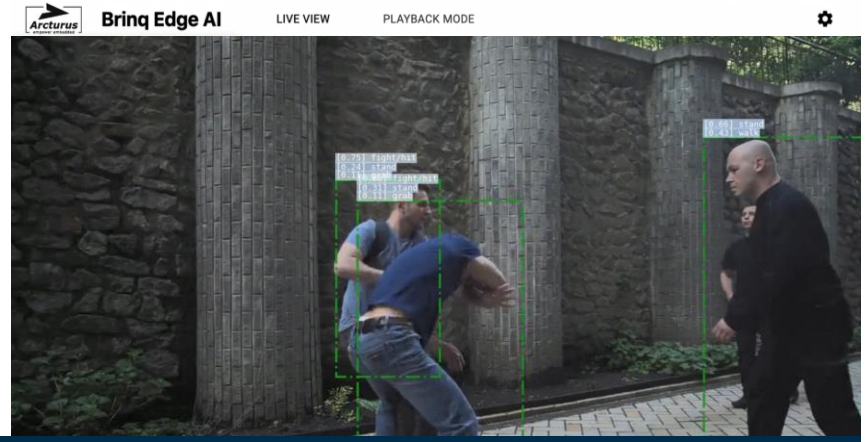
Boundary Crossing and Intrusion



Package Analytics



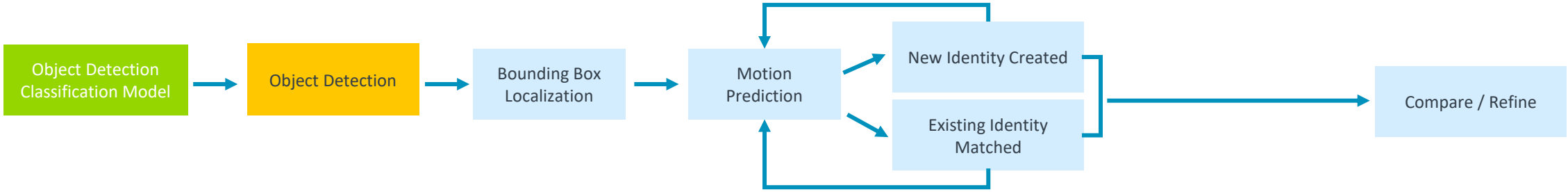
Characterization



Behaviour Analysis

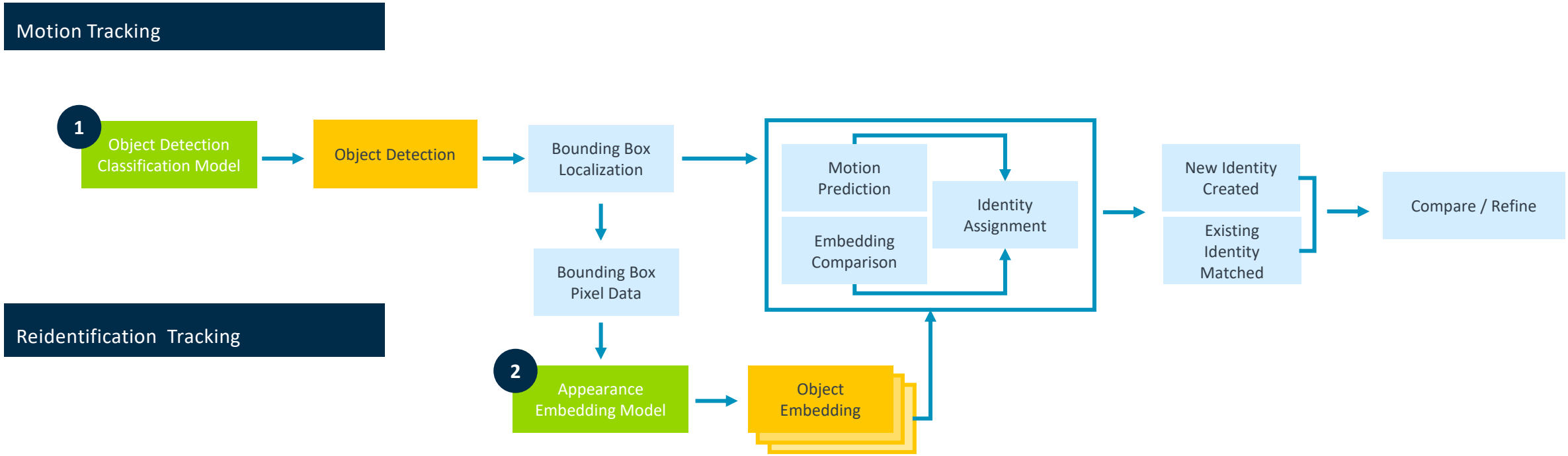
Motion Tracking

Motion Tracking



Relies on continuous detections - object cannot leave frame/FoV.

Motion and Appearance Tracking



More objects = more processing

Visual appearance reidentifies objects irrespective of time or space -but- requires generating embeddings for each object detected.

Building and Scaling Pipelines

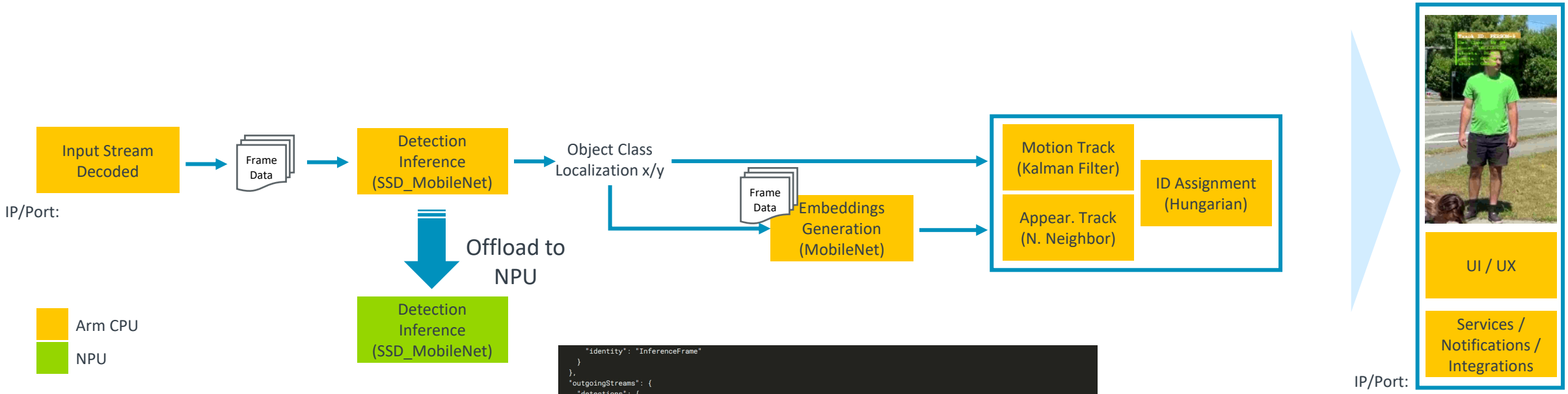
Harshad Mahadik

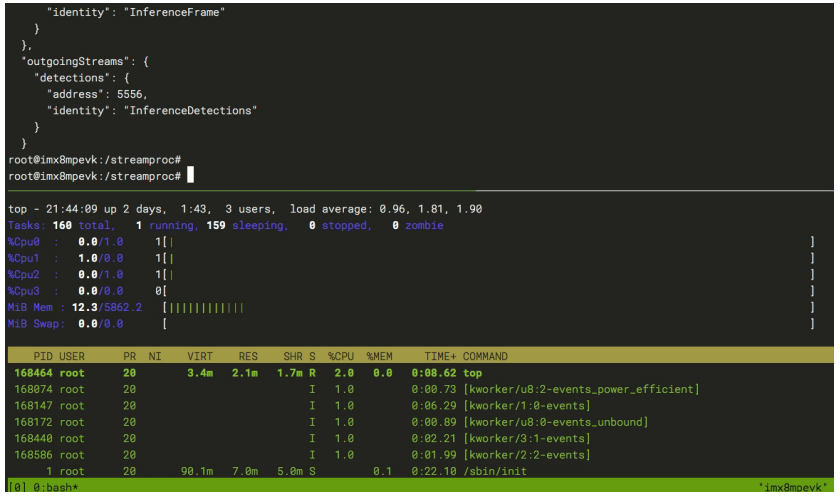
Edge AI and Vision Team Lead – Arcturus

Jonathan Rynne

Data Scientist – Arcturus

Tracking Pipeline (by resource)

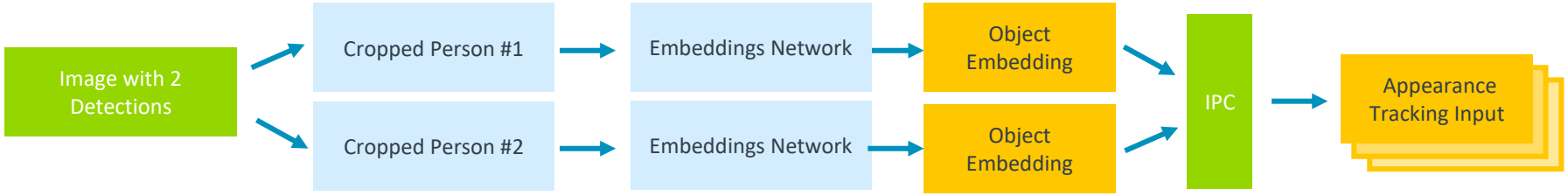




Optimization Techniques

Technique	Description	Benefit	Cost
Concurrent Processing	Allows multiple models to run at the same time	Decreases latency	Resource intensive
Batch Processing	Allow multiple images to processed at the same time	Increases throughput	Increases latency
Model Depth Reduction	Reduces the complexity of the network	Decreases inference time	Decreases accuracy
Tiling	Breaks down input image into multiple image tiles	Improves accuracy	Increases latency
Network Quantization	Reduces precision of network e.g. from Float32 to INT8	Improve performance	Decreases accuracy

Concurrent Model Processing



Pros

- Reduced latency
- Increased throughput
    - Increased FPS
- Increased analytic results

Cons

- Increased complexity
    - memory management (IPC)
- Increased memory consumption
- CPU/GPU/NPU intensive

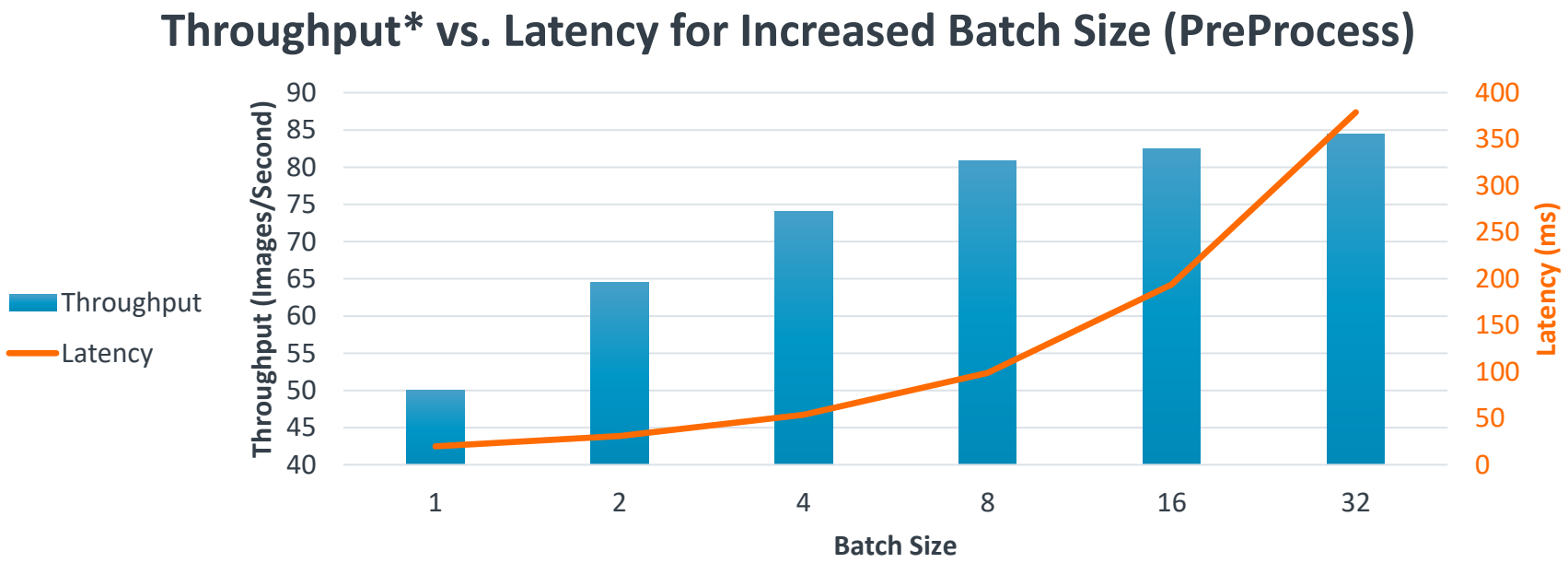
Performance Example

Mobilenet_v1_embeddings network
- 30ms inference
- 4 People Detected within Frame

Total Inference Time without Concurrent Processing
- 30ms * 4 = 120ms

Total Inference Time with Concurrent Processing
- 30ms + 30ms (overhead) = 60ms

Batch Processing/Batch Inference



Pros

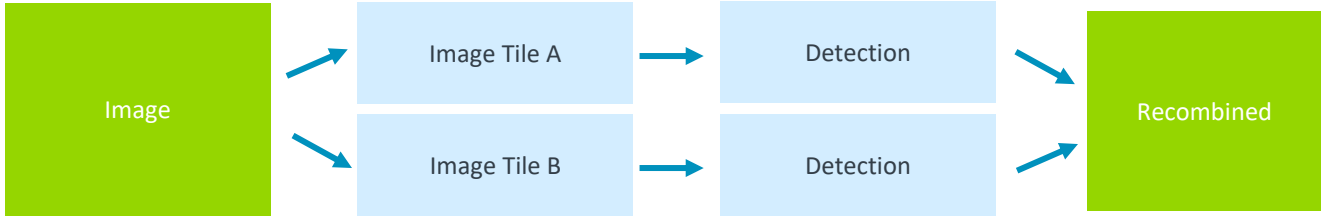
- Increased throughput
- Decreased complexity

Cons

- Increased memory consumption
- Model/backend support
- Increased latency

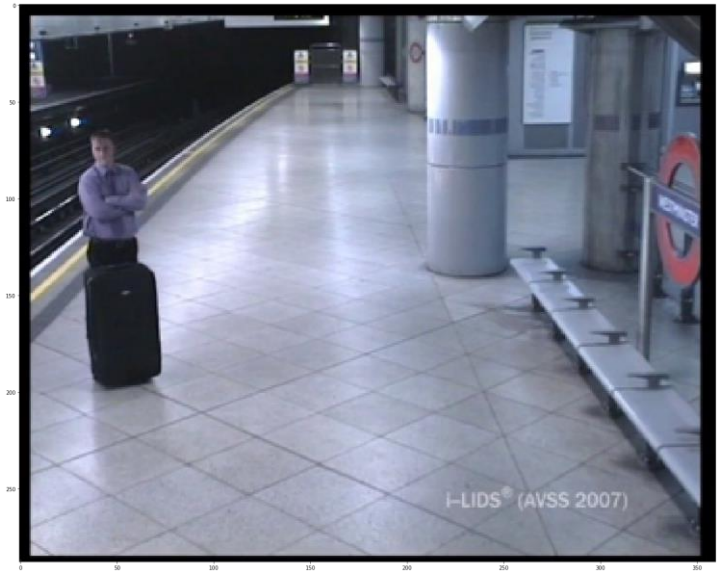
* SSD_Mobilenet_v2_300x300_quantized / i.MX 8M Plus

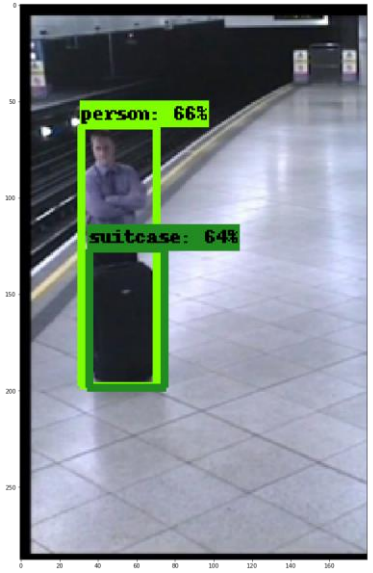
Tiling



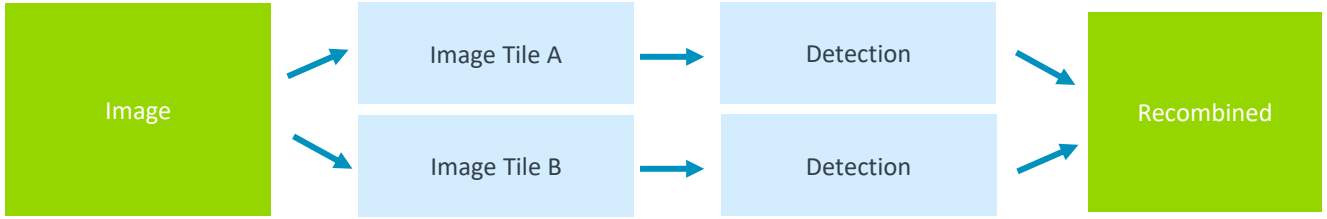
Pros

- Improve detection accuracy
- Improve Tracking
- Can be combined with batch



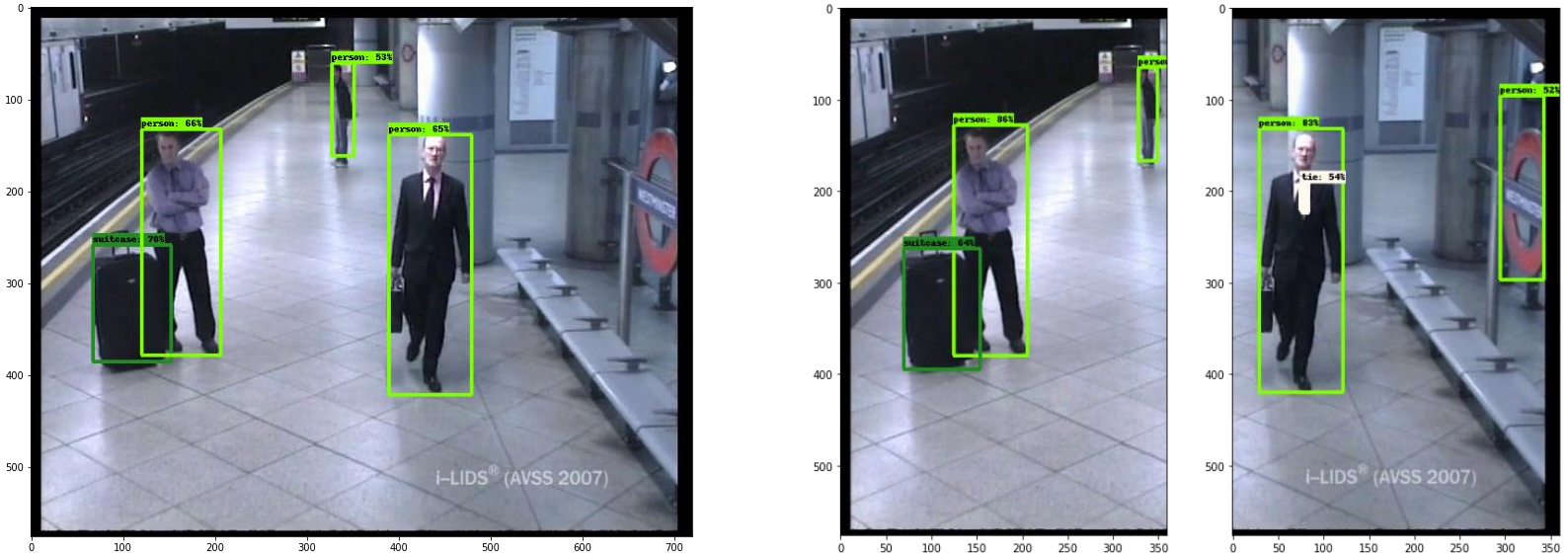


Tiling



Cons

- Increase false positives
- Add pre/post processing overhead



Model Size Depth Reduction

Definition in source code:

```
depth_multiplier: Float multiplier for the depth (number of channels)
  for all convolution ops. The value must be greater than zero. Typical
  usage will be to set this value in (0, 1) to reduce the number of
  parameters or computation cost of the model.
```

Network	Depth Multiplier	Speed* (Embed Time)	Accuracy (mAP on Market 1501)
Mobilenet v1	1.0 (100%)	51ms	63.49%
Mobilenet v1	0.75 (75%)	33ms	62.12%
Mobilenet v1	0.5 (25%)	19ms	60.21%

* Calculated by running model on the CPU (4x Arm Cortex-A53)

arm

Let’s do even more!

David Steele

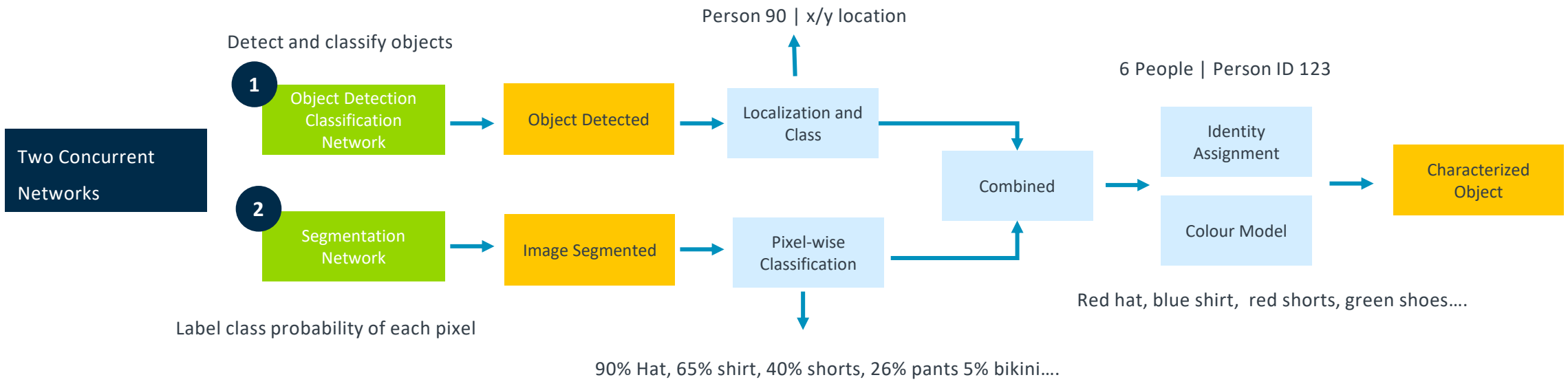
Arcturus

Characterization

Detection and Colour Modeling

Segmentation Output

Characterization



Two networks process the same frame in order to improve precision and fully characterize object

Atlas Edge AI Hardware Platform

Atlas Platform

NXP 8M Plus

Core and Memory
4x Cortex-A53 up to 1.8 GHz
Cortex-M7 up to 800 MHz
HiFi 4 DSP @ 800 MHz

Memory
6GB LP-DDR4 (32-bit)
32GB eMMC
32MB SPI NOR

AI Acceleration
NPU
up to 2.3 TOPS

Imaging and Video
2x MIPI CSI
2x ISPs up to 12MP
3D GPU, 2D GPU
Video encode/decode up to 1080p60, h.265/4,
HDMI, MIPI-DSI, LVDS

Connectivity
2x Gigabit Ethernet
UART, I2S, I2C, SPI
2x USB 3.0
PCIe Gen 3
SDIO
CAN-FD

Interconnect

2x M.2 CARD SLOT (M-Key)

INTERCONNECT BACKPLANE

POWER MGMNT

3.3VDC I/O

RESET/ WATCHDOG

NVIDIA

Core and Memory
4x Cortex-A57 @ 1.4 GHz)
up to 6x Carmel @ 1.1 GHz

Memory
4GB (64-bit) LP-DDR4
up to 8GB (128-bit)
16GB eMMC

AI Acceleration
GPU (472 GFLOPS)
or GPU (21 TOPs)

Imaging and Video
4x MIPI CSI up to 6
Video encode/decode up to 4Kp30 (HEVC)
HDMI, eDP

Connectivity
1x Gigabit Ethernet
USB 2.0
USB 3.0
PCIe Gen 2/3
SDIO

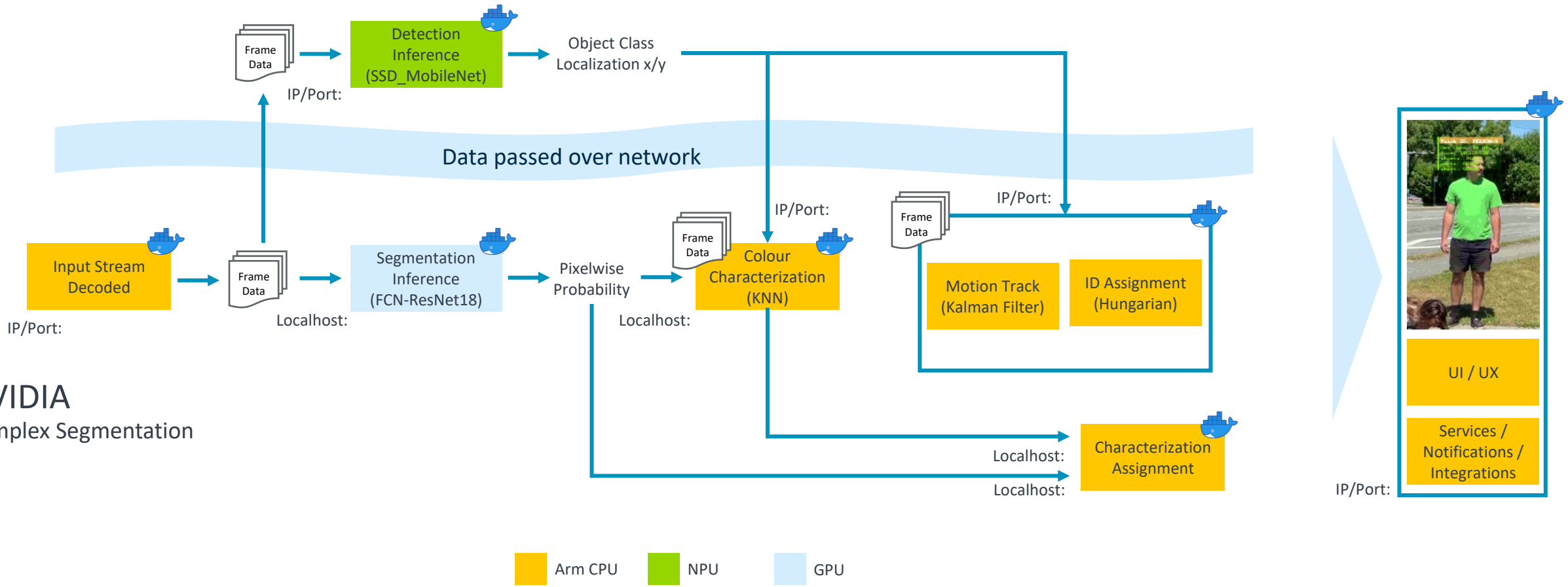
External Terminations

+12VDC | RESET | FAC. RESET | READY LED | ALARM LED | LINK / ACT LED | RJ45 ETH | USB3 (Up to 3) | HDMI (Up to 2) | I/O (4x4) | MICRO SDCARD

Characterization Pipeline (by resource)

i.MX 8M Plus

High Frame Rate Detection

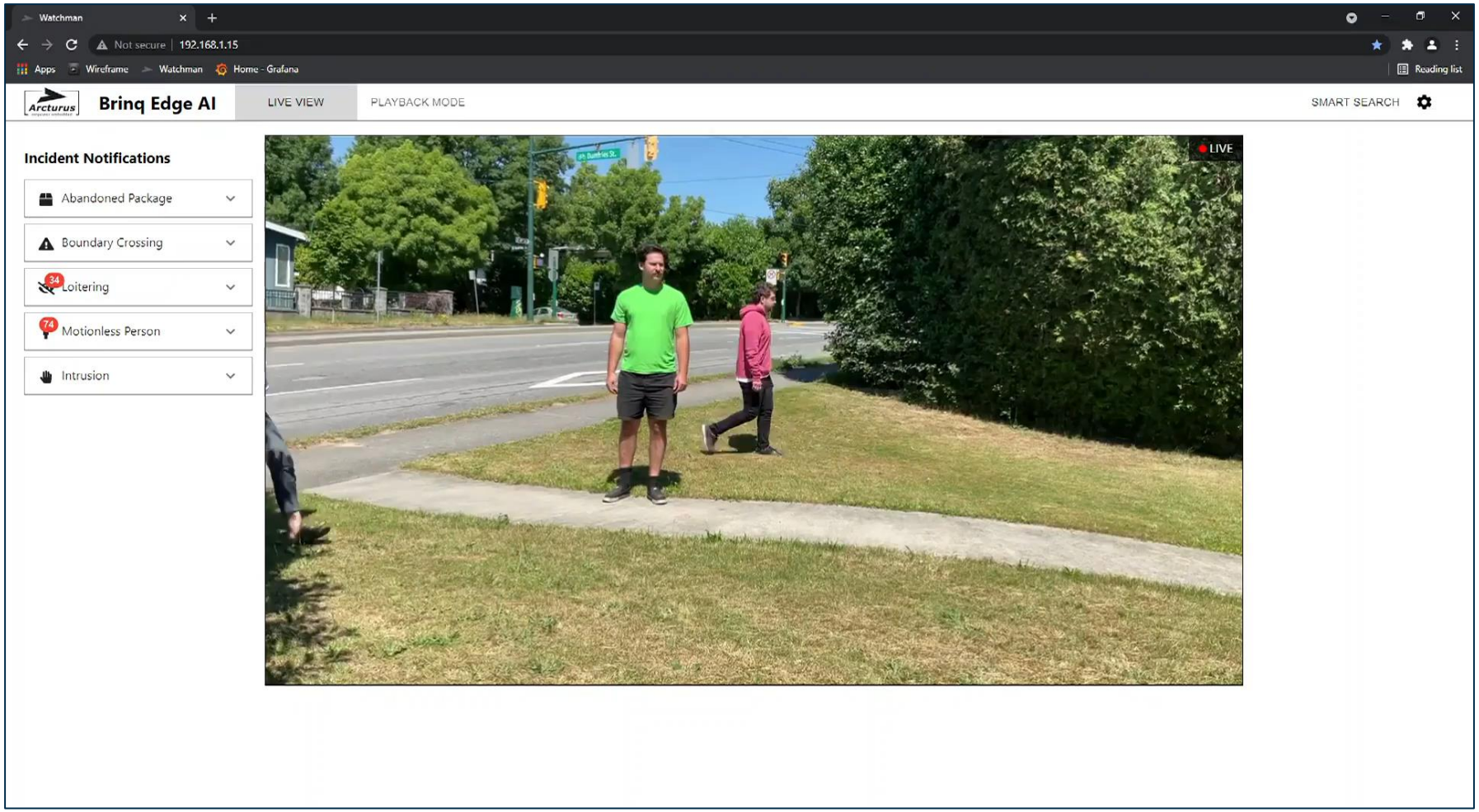


NVIDIA

Complex Segmentation

Characterization Demo

Characterization Demo



NXP ML/AI training series

- 20+ training modules
- Available at www.nxp.com/mltraining

1 MACHINE LEARNING CONCEPTS

Concepts and Introduction

2 eIQ™ SOFTWARE DEVELOPMENT ENVIRONMENT

eIQ Overview
Transfer Learning Intro & Lab
Handwritten Digit Recognition Example

3 eIQ TOOLKIT

eIQ Toolkit: How to BYOD
eIQ Toolkit: How to BYOM
eIQ Toolkit: Command Line interface

4 MACHINE LEARNING ON MCUS

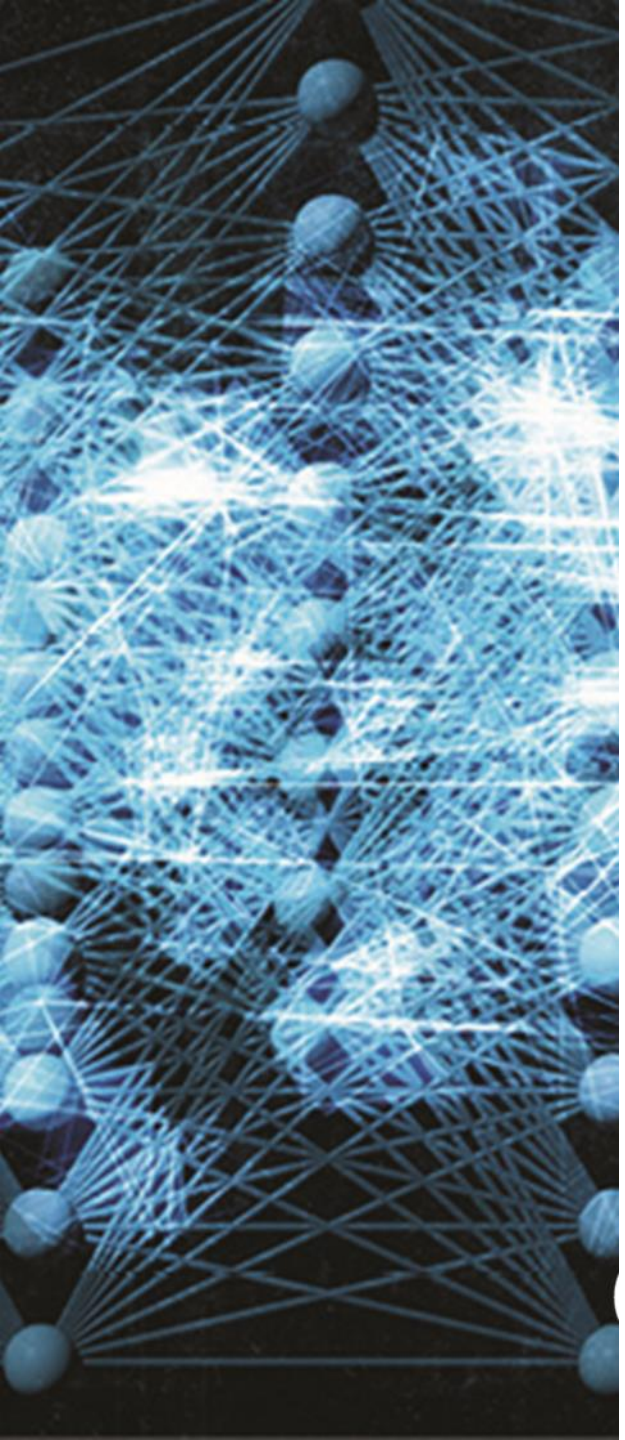
Machine Learning with i.MX RT
Get started with eIQ on i.MX RT
Create your own Model using Glow
eIQ inference with TensorFlow Lite for MCUs – Overview & Lab
eIQ inference with Glow NN Compiler – Overview & Lab
eIQ inference with DeepViewRT

5 MACHINE LEARNING ON MPUS

eIQ inference with ONNX
eIQ inference with Arm NN
eIQ inference with TensorFlow Lite (for MPUs)
TensorFlow Lite support for Android ML

6 PARTNER ML SOLUTIONS

Solutions and topics contributed by NXP eIQ partners



References and helpful links

- eIQ™ ML Software Development Environment (https://www.nxp.com/eiq)
- eIQ ML/AI Training Series (https://www.nxp.com/mltraining)
- eIQ Community (https://community.nxp.com/t5/eIQ-Machine-Learning-Software/bd-p/eiq)
- Embedded Linux for i.MX Applications Processors (https://www.nxp.com/design/software/embedded-software/i-mx-software/embedded-linux-for-i-mx-applications-processors:IMXLINUX)
- MCUXpresso Software and Tools (https://www.nxp.com/design/software/development-software/mcuxpresso-software-and-tools-:MCUXPRESSO)
- Brinq™ Edge AI and Vision Analytics (https://www.arcturusnetworks.com/brinq/)
- Brinq Edge AI for Public Safety (White paper)
- Arm AI Tech Talk – The Smart City In Motion – Intelligent Transportation Systems (webinar recording)
- Arm Dev Summit – Using Arm NN to Develop Edge AI in the Smart City (webinar recording)

Closing Remarks and Prize Draw

Thank you for participating in our Arm AI Tech Talk.

Complete your information to be eligible to win one of 2x $250 Amazon Gift cards courtesy of Arcturus and NXP.

The entry deadline is July 13 2021 at 2pm ET. Prize draw will occur on July 13th and winners will be notified by email. One entry per person, per email.

Terms and conditions - by entering you are agreeing to share your information with Arcturus Networks Inc. and NXP Semiconductors for the purpose of contacting you about this and future promotions. Refer to each company's privacy policy and terms of use for additional detail.

* Required

https://forms.gle/ciBtRgC3c76BCNVd7

arm

Thank you!

Tweet us: @ArmSoftwareDev

Check out our Arm Software Developers YouTube channel

Signup now for our next AI Virtual Tech Talk: developer.arm.com/techtalks

Attendees: don’t forget to fill out the survey to be in with a chance of winning an Arduino Nano 33 BLE board

AI Virtual Tech Talks Series

Date	Title	Host
July 13th	Bringing Edge AI to Life - from PoC to Production	Arcturus & NXP
July 20th	Easy TinyML with Arduino: taking advantage of machine learning right where things are happening	Arduino

Visit: developer.arm.com/techtalks

arm

Thank you!

Tweet us: @ArmSoftwareDev

Check out our Arm Software Developers YouTube channel

Signup now for our next AI Virtual Tech Talk: developer.arm.com/techtalks

Attendees: don’t forget to fill out the survey to be in with a chance of winning an Arduino Nano 33 BLE board

arm DevSummit

October 19-21, 2021

Where Hardware & Software **Join Forces**

Call for Papers

Learn More at devsummit.arm.com

arm AI

AI Virtual Tech Talks Series

Thank You
Danke
Merci
谢谢
ありがとう
Gracias
Kiitos
감사합니다
धन्यवाद
شكرًا
תודה