From Enterprise to Makers: Driving Vision AI Innovation at the Extreme Edge

Amir Servi

Edge AI Product Manager

Sony Semiconductor Solutions

SONY

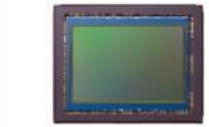
Sony Semiconductor Solutions Overview

EMBEDDED VISION SUMMIT

Gaming | Music | Movie | Music / Video Equipment | Semicon. | Finance

SONY MUSIC
SonyMusic
SONY MUSIC PUBLISHING

SONY PICTURES



ソニー生命
ソニー損保
ソニー銀行
ソニー・ライフケア

Sony Semiconductor Solutions Corporation

SONY

The Role of the Image Sensor in the AI Era

FROM Imaging TO Sensing

Higher Definition / Brighter

Pursuing images that look beautiful to the human eye

Object Recognition / Situational Awarenes

Generate data that is easy to process with AI

AITRIOS

AITRIOS™ makes edge vision AI simple

Hardware Ecosystem

AITRIOS-compatible hardware is equipped with Sony's IMX500 to enable highspeed AI processing on the chip at ultra low power.

Edge AI Sensing Platform

The platform provides tools to develop, deploy, and manage end-to-end edge AI sensing solutions.

Retail

Smart City

Factory

Logistics

Agriculture

Medical

Building

Construction

Architecture

Partner Ecosystem

Platform as a service

AITRIOS

IMX500 Edge AI

AI Model Distribution

AI Model Training

Device Management

SDK

Example of Commercialized Cases of AITRIOS

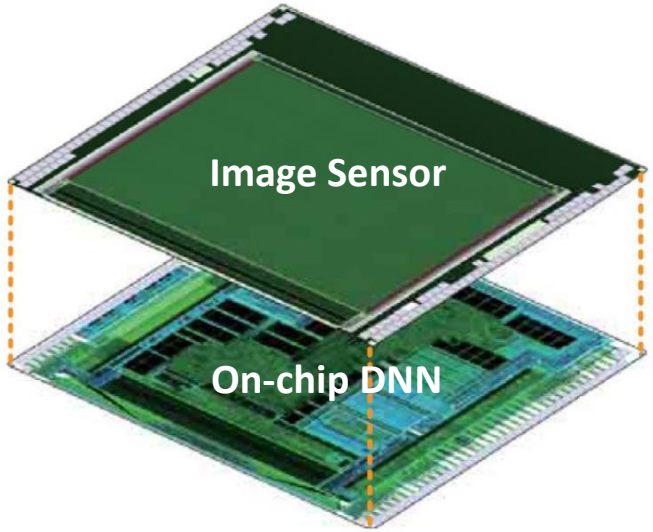
Retail | 7-Eleven ›

7-Eleven Japan utilizes a digital signage solution at 500 of their locations to develop more engaging, in-store advertisements that increase ad relevancy for their customers.

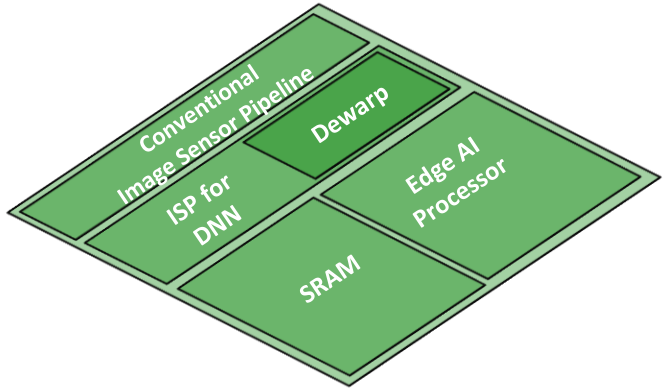
Logistics | MOVO Berth ›

AITRIOS empowers MITSUI-SOKO to improve employee timetables and reduce waiting times for drivers by detecting trucks and automating berth usage.

IMX500 Overview



1 chip solution with stacked die technology



Item	Specification
Pixels	12.4 MP; 4056 (H) x 3040 (V); 1.55 um 1/2.3 inch; 2x2 binning (3MP, 3.1 um)
FPS with DNN	Typically, 30 FPS (DNN model dependent)
AI Engine	Edge AI Processor (Sony home-grown) Typically activating AI inference adds <100 mW to power
Features	Compact ISP for sensing 8 MB exclusively for AI weights + Feature Maps
DNN Inference	Up to 640x640 RGB or 1K Monochrome Mobilenet SSD @ ~10 ms including post processing NanoDet-Plus @ ~18 ms including post processing
Numerical Formats	2/4/8/12/16 bit tensors Support for non-linear (user defined) representations Lossless calculation of layers and fused layers

Address privacy concerns

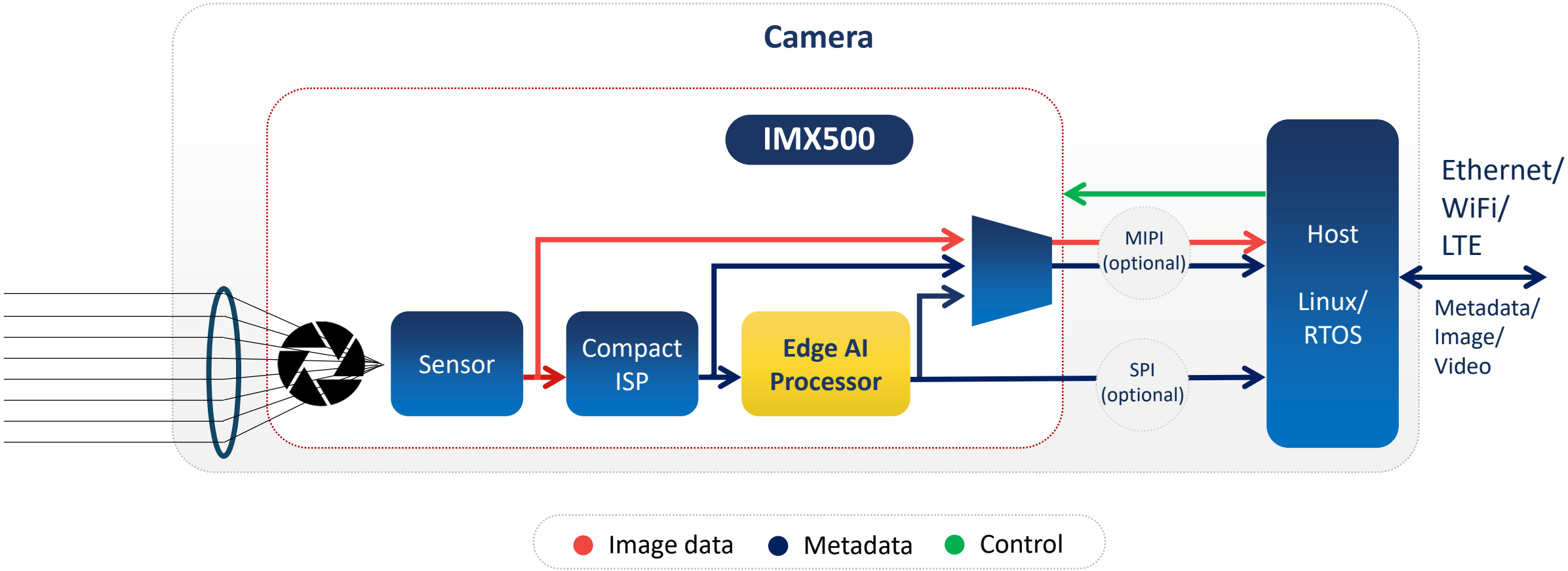
Cost-effective

Optimized processing & low power

One-chip solution

SONY

Typical Signal Path – Photon to Application



Instead of Generating Frames – Tell The Whole Story

Metadata enabled AI inference as an alternative to streaming images

IMX500-enabled edge device transmits metadata to represent the scene in contrast to a conventional smart camera with edge server which transmits actual images and video.

SONY

AITRIOS IMX500 Enabled Edge Devices

Rayprus
CSV26 (T3P)

Network: Ethernet

Power: PoE

Uses: Indoor, Retail, Factory, Logistics...

Sony Semiconductor Solutions
AIH-IVRW2 (T3Ws)

Network: Wi-Fi

Power: USB (Type-C), or AC adapter

Uses: Indoor, Retail, Logistics, Facilities…

LUCID Vision Labs, Inc.
TRITON SMART

Network: 1GBASE-T, M12, PoE

Power: PoE, or 12-24 VDC external

Uses: Machine vision, C mount, Factory Automation

Leopard Imaging
GS500

Network: PoE/ Cellular

Power: PoE, or 12-24 VDC external

Uses: Smart City, Pedestrian ID, Car Count, Traffic & License plate ID

Sony & Raspberry Pi
AI Camera Kit

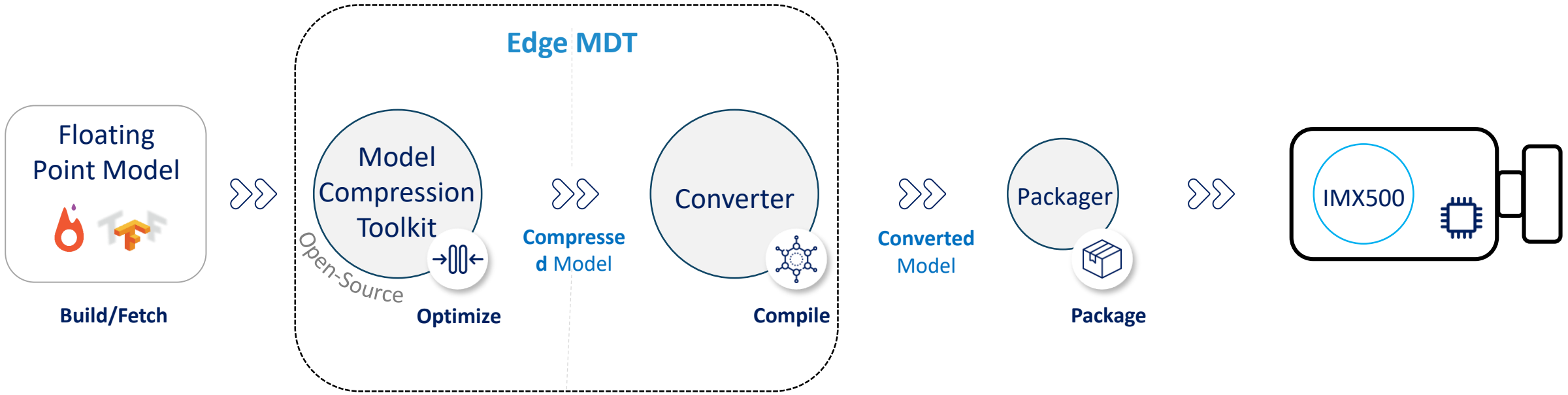
Network: Based on Raspberry Pi

Power: Based on Raspberry Pi

Uses: Development kit, IoT, vision AI creation…

SONY

Edge AI Model Development Toolkit



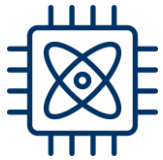
*low-code / no-code solutions within the AITRIOS platform

Model Compression Toolkit

Bridging the Gap Between AI Research to Real-Time Edge

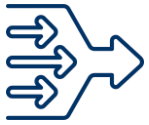
Open-sourced python library for Optimizing PyTorch & TensorFlow models

Supports a variety of tasks and architectures



Configurable hardware capabilities for hardware-aware quantization

Inspired by many publications on compression techniques and algorithms



Simple to use

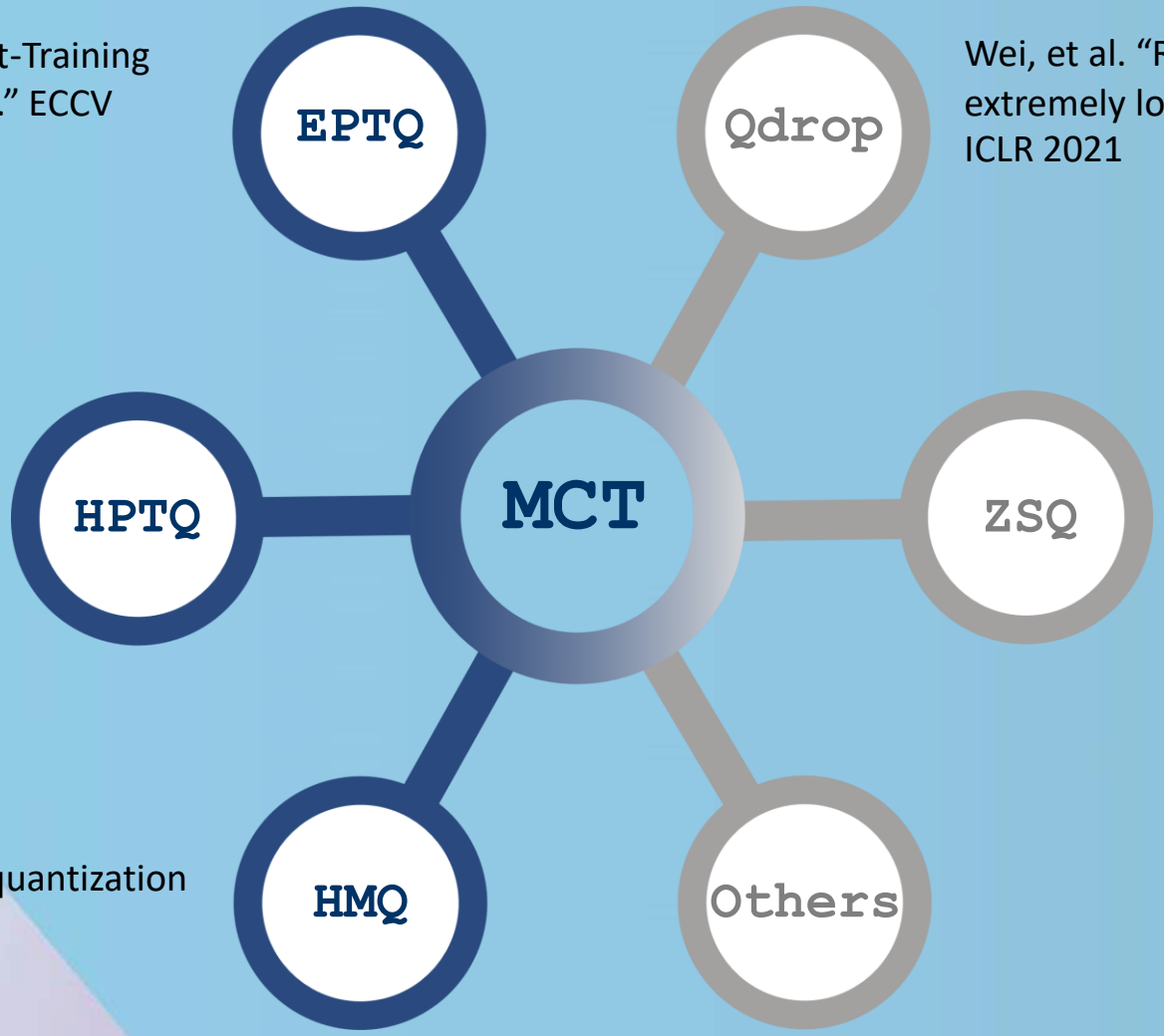
```python
import model_compression_toolkit as mct

quantized_model = mct.keras_post_training_quantization(model, dataset)
```

Gordon, et al. “EPTQ: Enhanced Post-Training Quantization via Label-Free Hessian.” ECCV 2024 Workshops

Habi, et al. “Hardware-Friendly Post Training Quantization.” Preprint 2021

Habi, et al. “HMQ: Hardware friendly mixed precision quantization block for cnns.” ECCV 2020

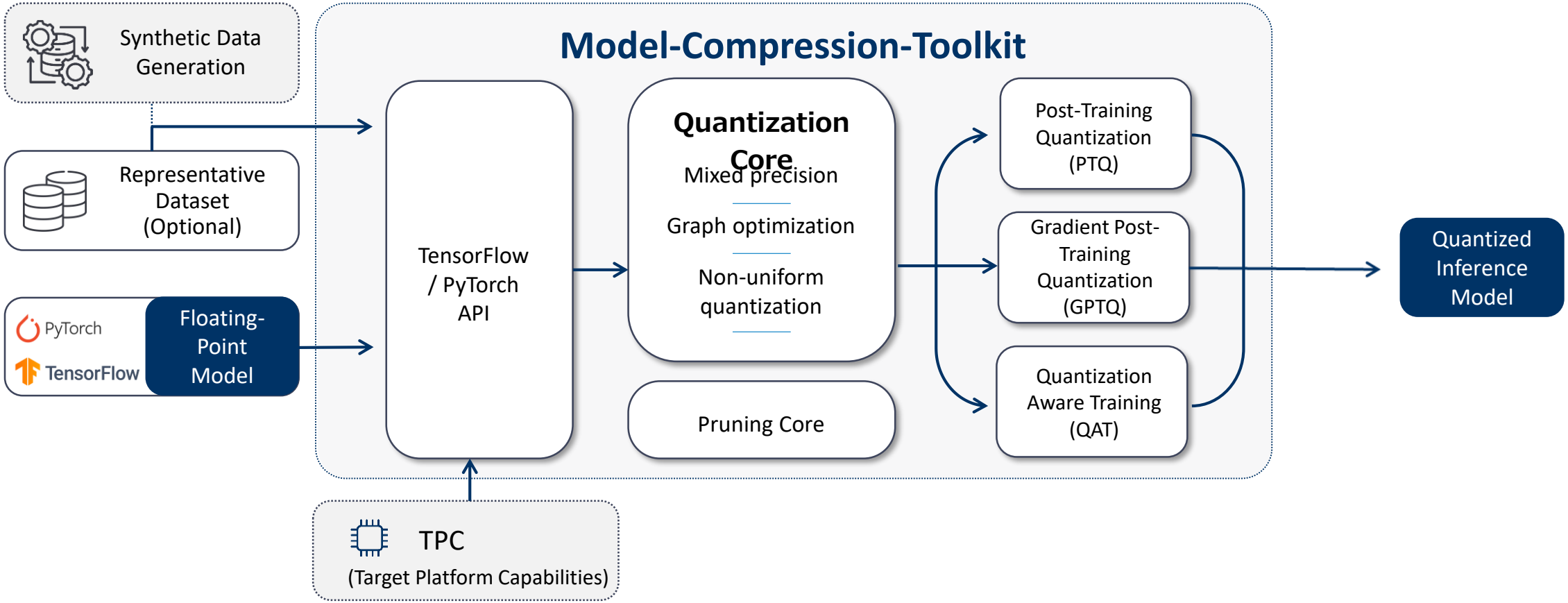


Wei, et al. “Randomly dropping quantization for extremely low-bit post-training quantization.” ICLR 2021

Jeon, et al. “Genie: Show Me the Data for Quantization.” CVPR 2023

- Sony Publications
- Other Publications

MCT Flow Diagram



Coverage

	Requirements	Comments
Architecture	**Static** models Simple Layer/Parameter compliance No structural restrictions	**Permissive** layer coverage Provisions for on-chip efficient and accurate post-processing
Input / Output	Input: Up to **640x640** RGB Or **1024x1024** Monochrome Output: Format and size can be adapted to AI task requirements including multiple output tensors	Output image is full sensor resolution regardless of AI input tensor resolution Robust task coverage
Memory	**8MB** for weights + feature maps. Firmware and graph memory stored separately	Advanced memory management and automatic compression allows very efficient memory utilization Typically models of up to 3-10 MWeights* can be supported (depending on architecture)

*if pruning is applied this refers to number of weights after pruning

EMBEDDED

Reference Neural Network Models

Classification

- **Task:** Categorize input data into predefined classes and provide a confidence score.
- **Training dataset:** `Imagenet` . Designed for use in visual object recognition research. It contains over images, making it one of the most extensive resources available for training deep learning models. of 1000 classes.

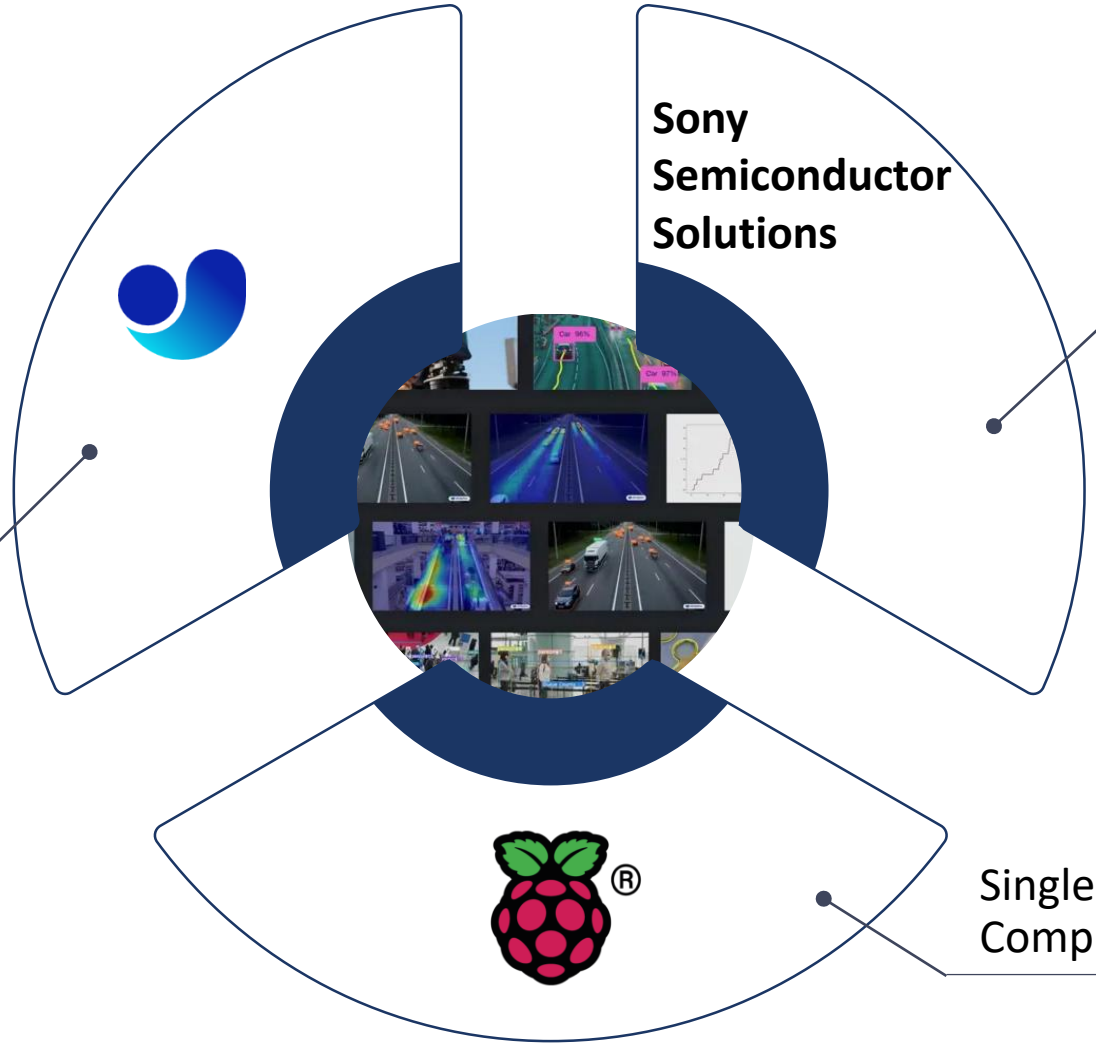
Model	[Top 1] Accuracy - Quantized(Float)	Input Resolution	Picamera2 Example Script
EfficientNet-B0	72.128 (73.876)	224x224	`python imx500_classification_demo.py --m /usr/share/imx500-models/imx500_network_efficientnet_bo.rpk`
EfficientNet Lite-0	75.252 (75.28)	224x224	`python imx500_classification_demo.py --m /usr/share/imx500-models/imx500_network_efficientnet_lite0.`
EfficientNetV2-B0	76.674 (76.424)	224x224	`python imx500_classification_demo.py --m /usr/share/imx500-models/imx500_network_efficientnetv2_b0.r`
EfficientNetV2-B1	77.032 (76.93)	240x240	`python imx500_classification_demo.py --m /usr/share/imx500-models/imx500_network_efficientnetv2_b1.r`
EfficientNetV2-B2	77.716 (77.94)	260x260	`python imx500_classification_demo.py --m /usr/share/imx500-models/imx500_network_efficientnetv2_b2.r`
MnasNet1.0	73.16 (73.078)	224x224	`python imx500_classification_demo.py --m /usr/share/imx500-models/imx500_network_mnasnet1.0.rpk`
MobileNetV2	71.572 (71.3)	224x224	`python imx500_classification_demo.py --m /usr/share/imx500-models/imx500_network_mobilenet_v2.rpk`
MobileViT-XS	72.326 (72.412)	256x256	`python imx500_classification_demo.py --m /usr/share/imx500-models/imx500_network_mobilevit_xs.rpk`
MobileViT-XXS	67.440 (67.40)	256x256	`python imx500_classification_demo.py --m /usr/share/imx500-models/imx500_network_mobilevit_xxs.rpk`
RegNetX-002	68.352 (68.20)	224x224	`python imx500_classification_demo.py --m /usr/share/imx500-`

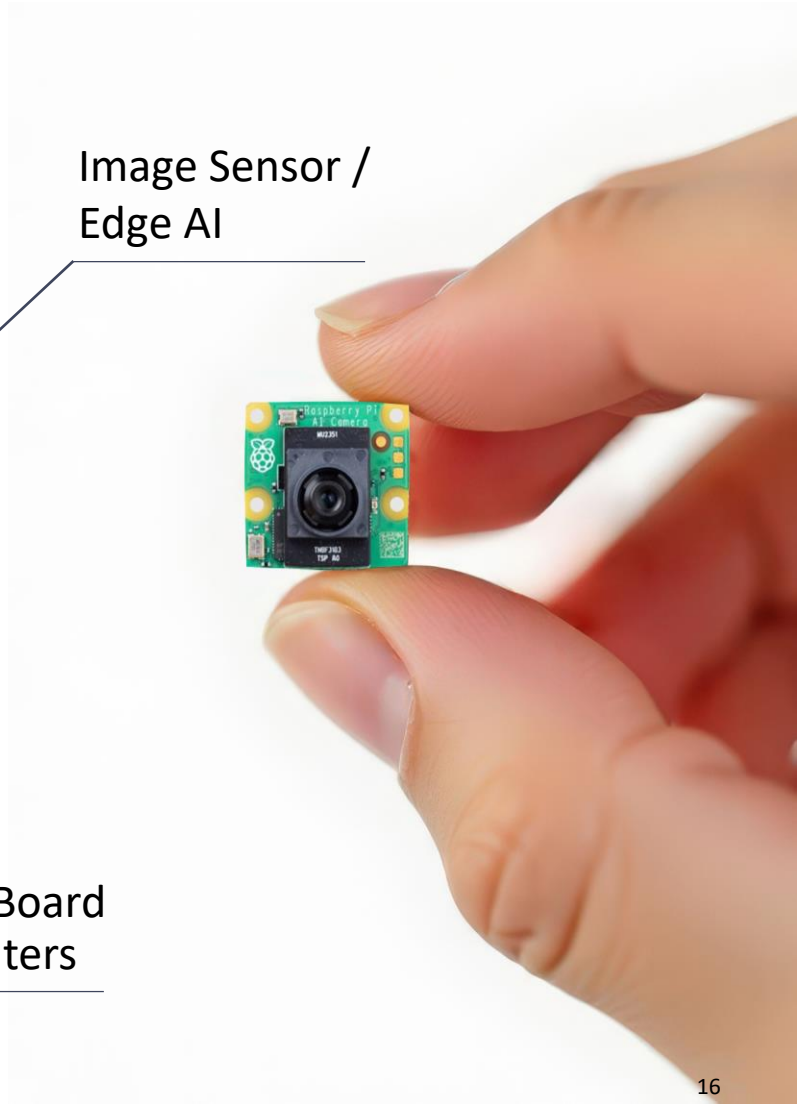
Model	Format	Status	Size of `RPK` (MB)	mAP50-95
YOLOv8n	imx	✅	3.1	0.602
YOLO11n	imx	✅	3.2	0.644

SONY

Introducing the AI Camera for developers

Joint forces with the 3 industry leaders



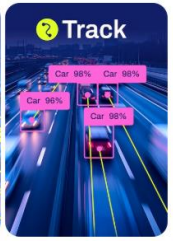


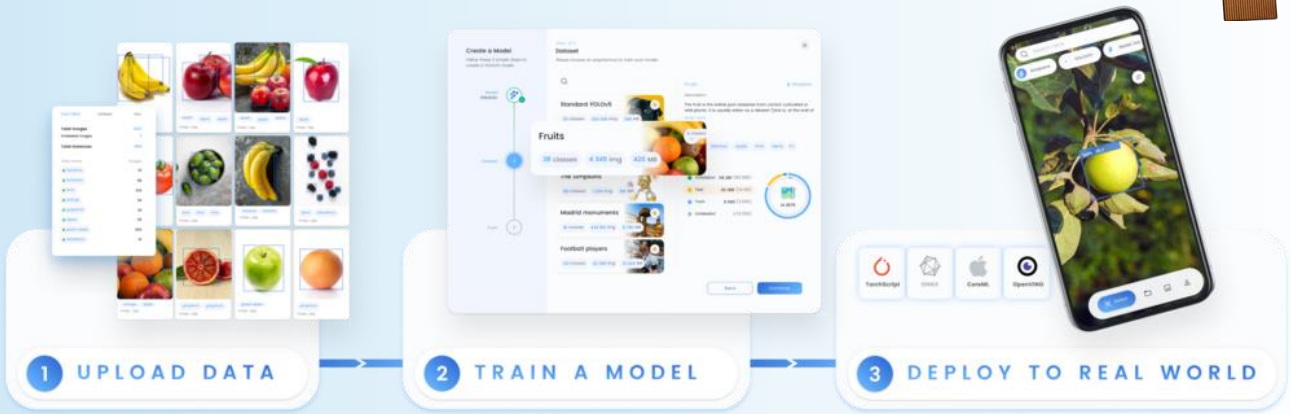
The power of Ultralytics YOLO integration with IMX500

ultralytics YOLOv8

ultralytics YOLO11

ultralytics HUB BETA





Seamlessly Integrates with the IMX500

ultralytics YOLOv8

in numbers

Object Detection Accuracy:
37.15 mAP [coco]

Inference Latency:
28.7 ms

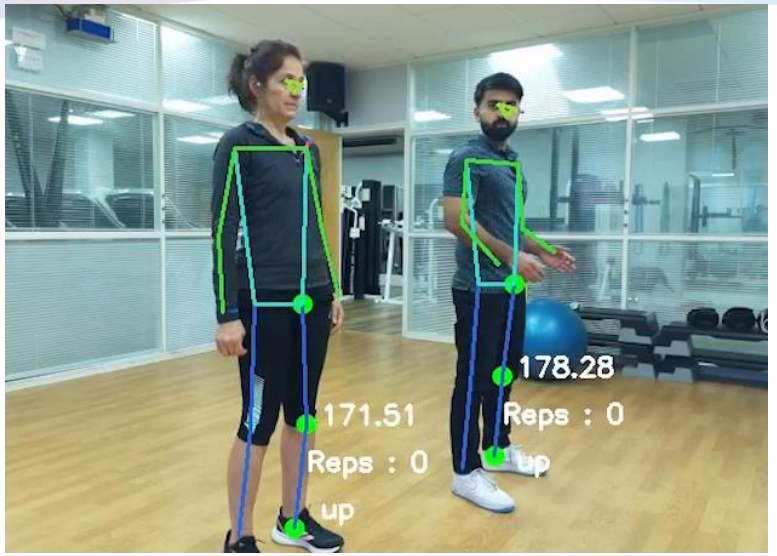
ultralytics YOLO11

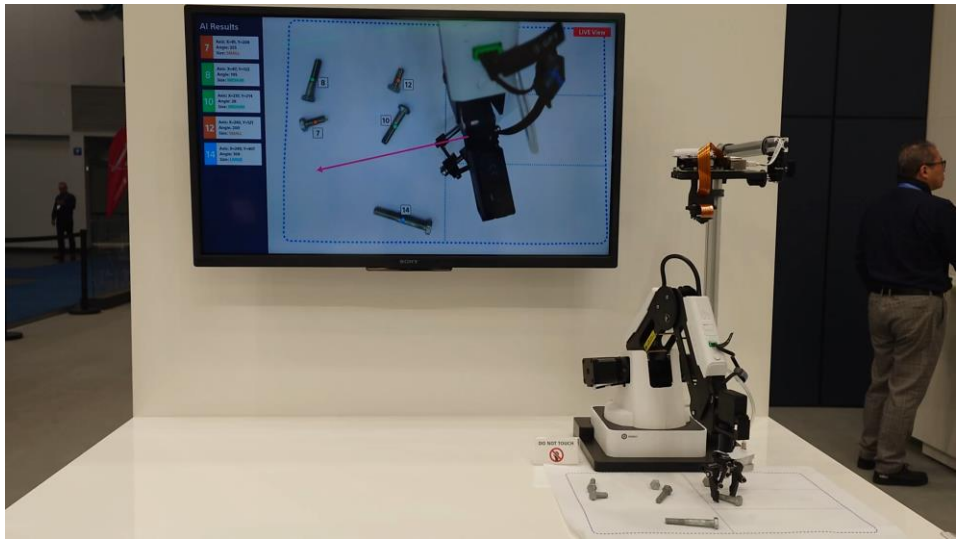
in numbers

Object Detection Accuracy:
38.9 mAP [coco]

Inference Latency:
29.95 ms

Some Project Examples





Getting Started Kit

Sony Developer Portal

Raspberry Pi AI Camera | Sony Semiconductors

RaspberryPi AI Camera

AI Camera - Raspberry Pi Documentation

Ultralytics Yolo Export to IMX500

Sony's IMX500 Export for Ultralytics YOLO11

AITRIOS - The edge AI sensing platform for vision

AITRIOS | Sony Semiconductor Solutions Group

Scan Me!

Get your own development kit

Key Takeaways

- First intelligent sensor – AI happens on the sensor!
- High performance with low power & low latency
- Privacy preserving
- Enterprise ready – AITRIOS
- Makers ready – RaspberryPi AI Camera, open source development tools
- Ultralytics export for the community and enterprise

Time for Questions

Scan Me!

Get your own development kit

Don’t forget to visit the SONY booth #309

SONY