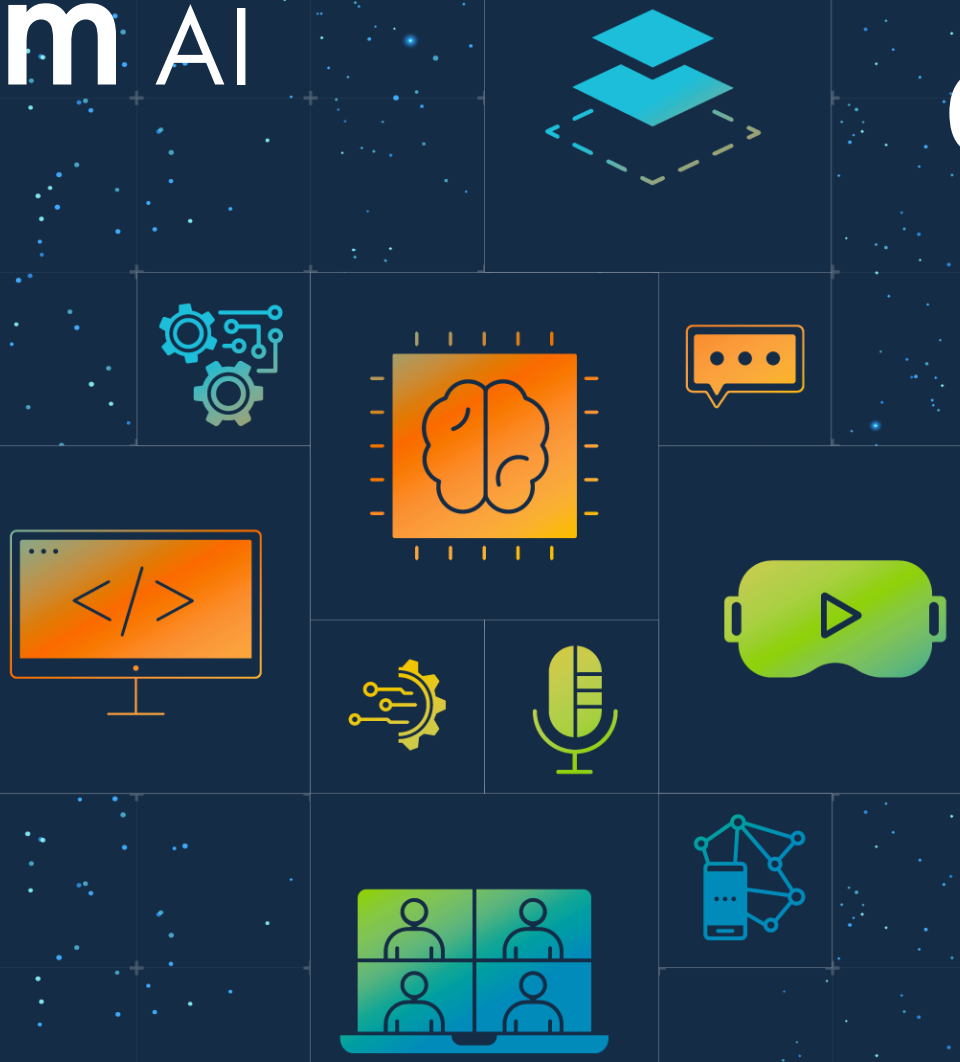
arm AI



Object Detection with Arm's Ethos-U55

Emza & Alif

Henrik Flodell & Eitan Weintraub

17 May 2022

arm AI

Welcome!

Tweet us: @ArmSoftwareDev -> #AIVTT

Check out our Arm Software Developers YouTube channel

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

Our upcoming Arm AI Tech Talks

Date	Title	Host
17th May	Object Detection with Arm's Ethos-U55	Emza & Alif
31st May	Advancing computer vision on the edge with different ML approaches	Plumerai, Deeplite, Roviero
14th June	How to run object detection on Arm Cortex-M7 processors	Edge Impulse
28th June	A Hardware-aware Approach for Designing Neural Models	Nota.ai

Visit: www.arm.com/techtalks

Presenters

Henrik Flodell

Marketing Director @Alif Semiconductor

Eitan Weintraub

Tech lead, Machine Learning engineer @Emza visual sense

Agenda

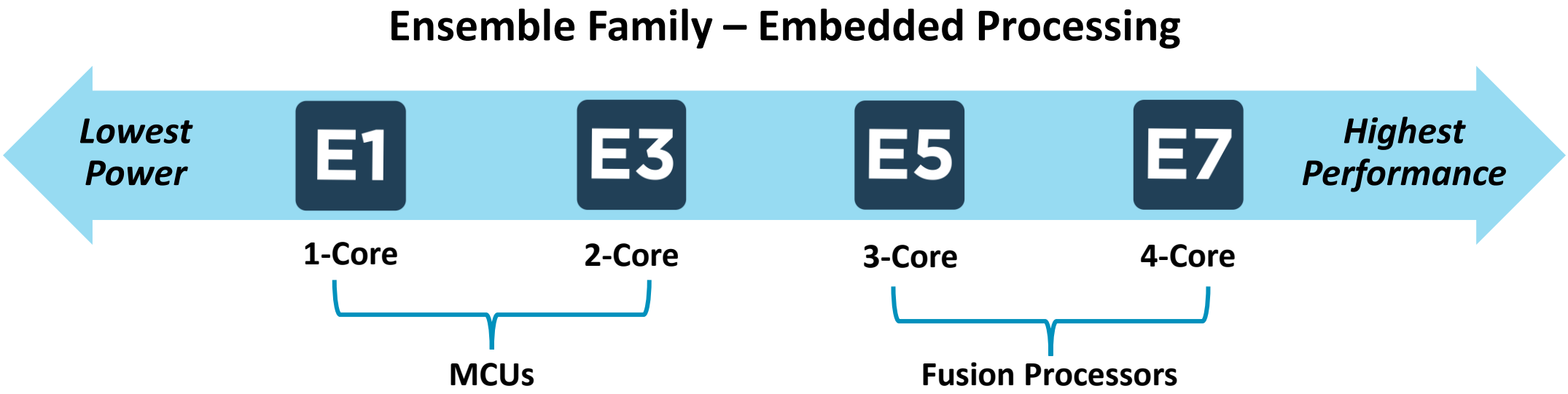
- Alif Semi & Ethos U55
- Model architecture
- Datasets
- Training parameters and results
- Model conversion
- Deploy on FVP and time measurements
- Deploy on Alif EVB and time measurements
- Summary

We Make Devices Intelligent

- Founded in January of 2019
- Global team
  - US, India & Singapore locations

- Our Ensemble MCUs and Fusion Processors contain:
  - **Scalable Processing** – First silicon in the market using Cortex-M55
  - **Battery Friendly** – Architected for Lowest Power Consumption
  - **Strong Security** – Built in from the Ground Up
  - **Edge AI Enabled** – First silicon in the market with (dual) Ethos-U55 microNPU

Alif Ensemble Product Family

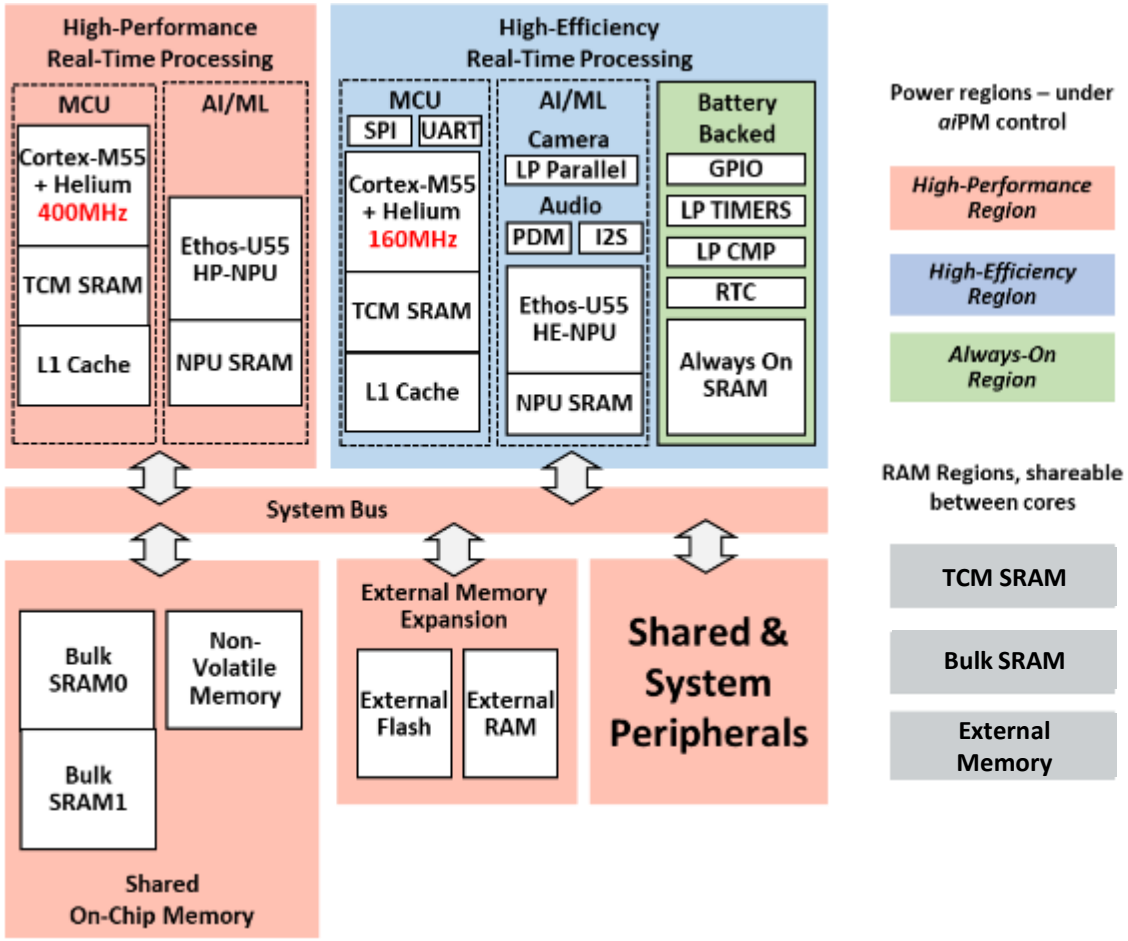


Scalable performance

Processing Combinations	E1 Single-Core MCU	E3 Dual-Core MCU		E5 Triple-Core Fusion Processor		E7 Quad-Core Fusion Processor	
Real-Time MCU Core	Cortex-M55 160 MHz	Cortex-M55 160 MHz	Cortex-M55 400 MHz	Cortex-M55 160 MHz	Cortex-M55 400 MHz	Cortex-M55 160 MHz	Cortex-M55 400 MHz
microNPU AI/ML Accelerator	Ethos-U55 128 MAC/c	Ethos-U55 128 MAC/c	Ethos-U55 256 MAC/c	Ethos-U55 128 MAC/c	Ethos-U55 256 MAC/c	Ethos-U55 128 MAC/c	Ethos-U55 256 MAC/c
Application MPU Core				Cortex-A32 800 MHz		Cortex-A32 800 MHz	Cortex-A32 800 MHz

arm AI

AI-Powered Environment sniffing engine, architected for flexibility, and low-power operation



- High-Efficiency region continuously senses environment for changes
  - Dedicated low-power sensing peripherals
  - 18 µA/MHz in Active mode
  - ~700 nA in standby
- High-Performance region wakes fast, and executes sophisticated models quickly
  - 400 µS wakeup after first boot
- Flexible RAM regions & peripherals, shareable between all cores
  - Including TCM snooping

arm AI

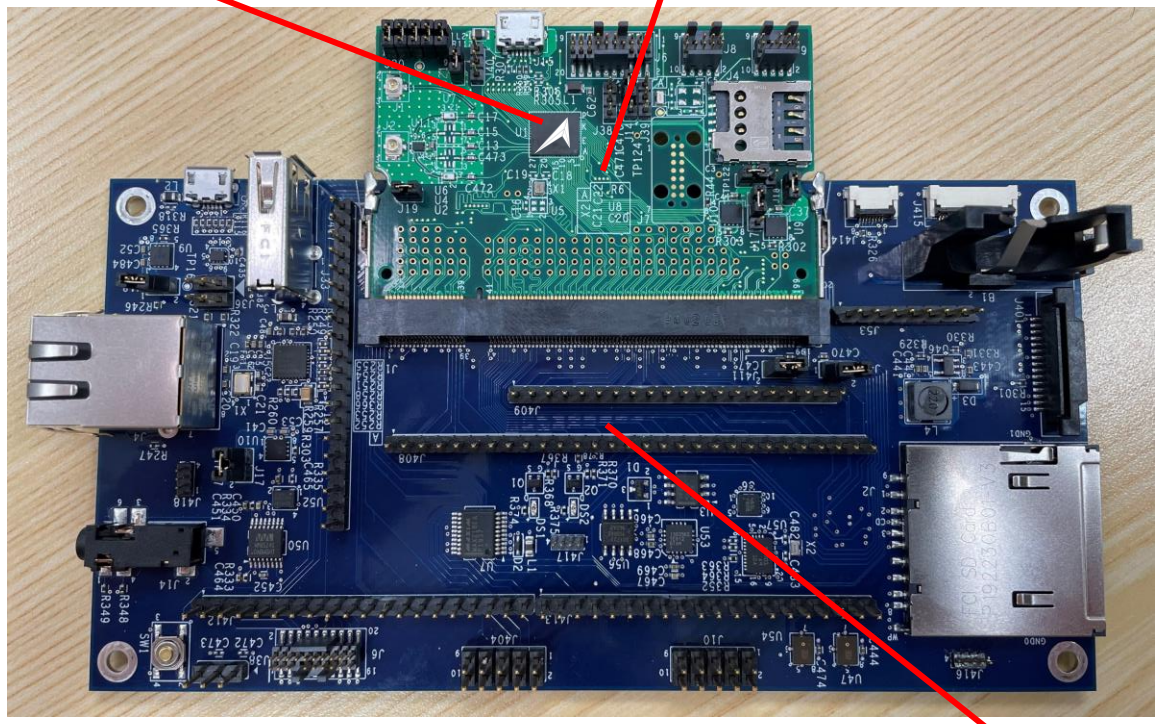
Development System

Available Now For E7, E5, E3, E1 Series devices

For more information & ordering

www.alifsemi.com

contact@alifsemi.com



arm AI

About Emza visual-sense

- Pioneers of tiny edge AI-based vision solutions since 2016
- Enabling mass market deployment with optimized power, size, and cost
- WiseEye ULP vision solution is shipping with Dell laptops
- Emza partnered with Arm to expand tinyML CV solutions to new markets

https://www.emza-vs.com/

Objective

Demonstrate real application on target HW using Arm's Cortex-M55 and Ethos-U55 architecture

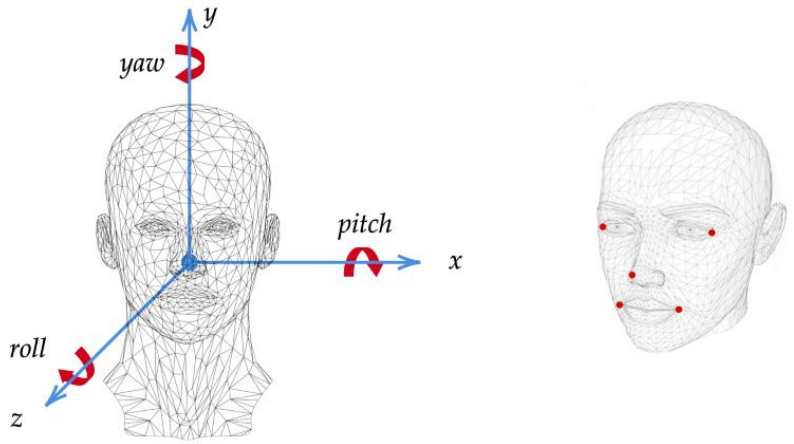
The development process:

Application

Face detection with additional facial information (pose, landmarks) up to 2m

Running on Alif E1/E3 (M55+U55) for low power

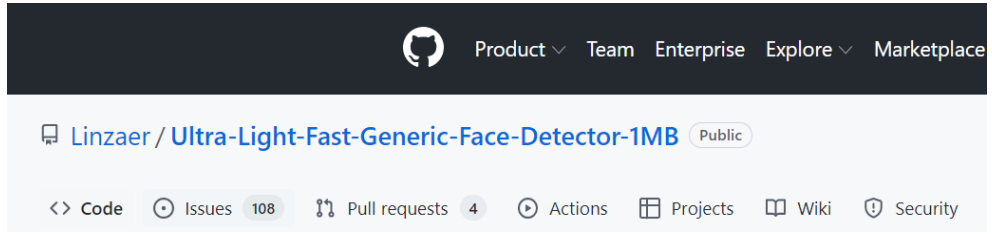
applications



Potential use cases:

- User presence and context awareness for smart devices (notebooks, PCs, smart TV)
- Human detection for home security (video doorbell, smart cameras)

Model architecture

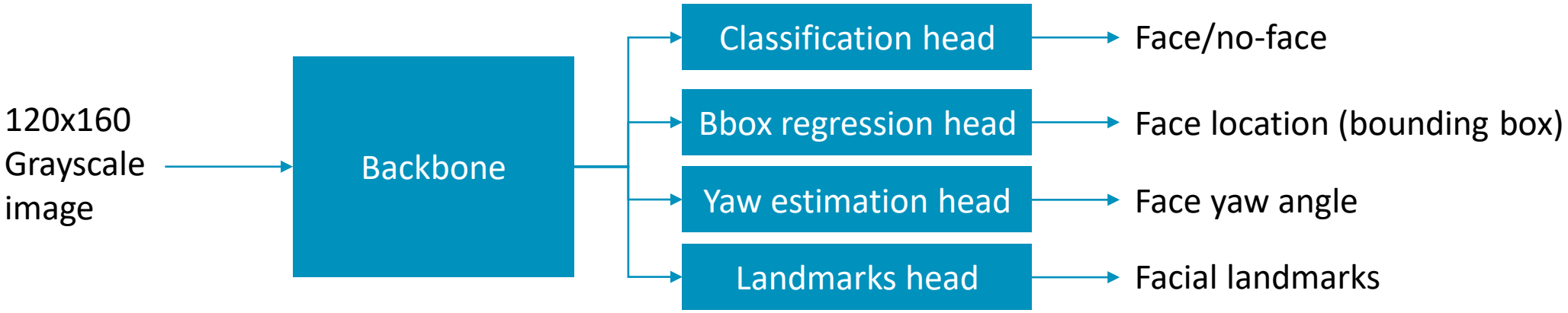


Baseline model - SSD from open source 120x160x3 - repo

Customization: - Modifying input from color RGB to grayscale

- Adding a regression head for yaw detection

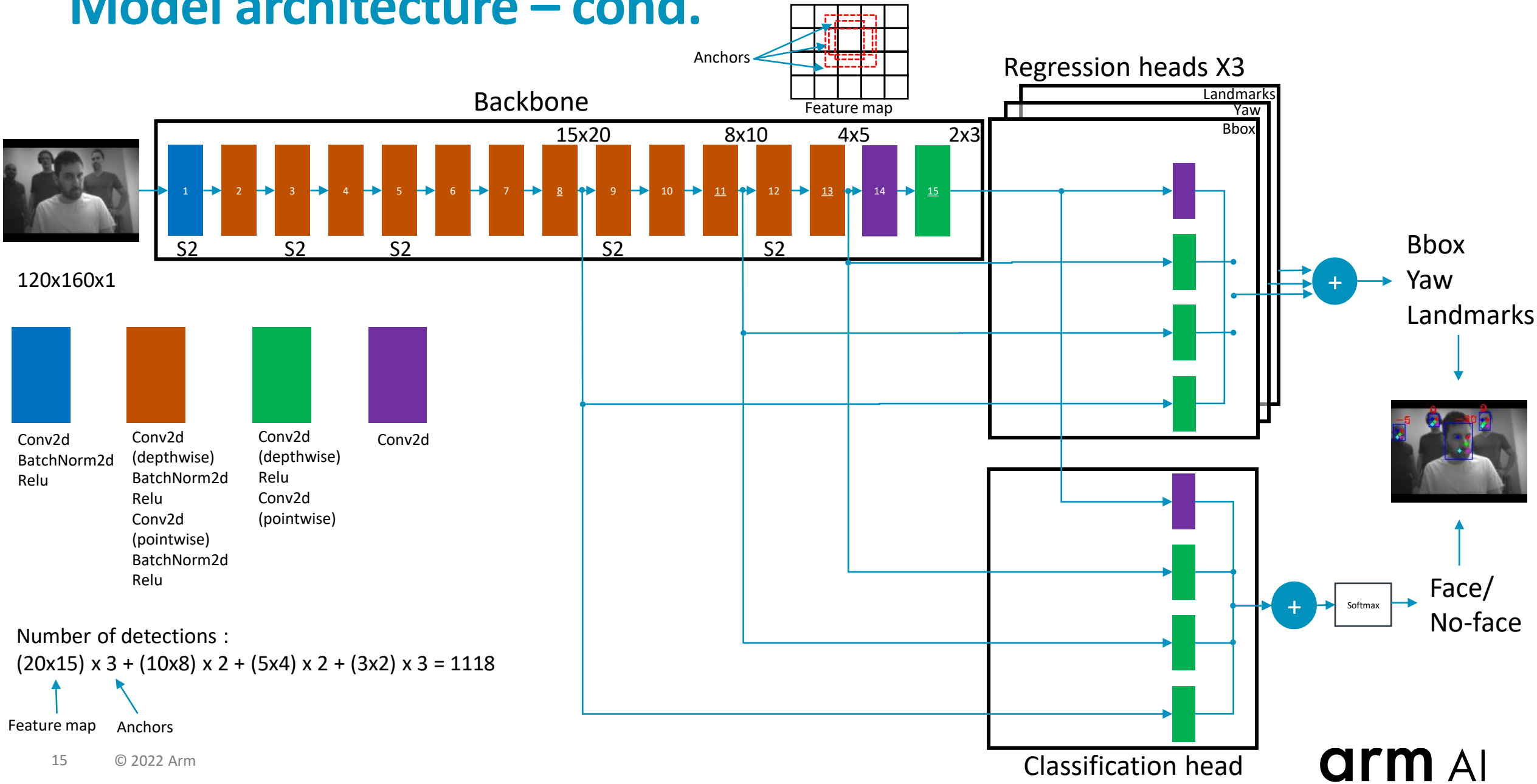
- Adding a regression head for landmarks detection



Total params: 355,562
Params size (MB): 1.36

arm AI

Model architecture – cond.



Number of detections :

(20x15) x 3 + (10x8) x 2 + (5x4) x 2 + (3x2) x 3 = 1118

Feature map Anchors

arm AI

Datasets

Used datasets: Wider, Yale, CelebA and Emza proprietary
Adding yaw and landmarks annotation using SOTA models

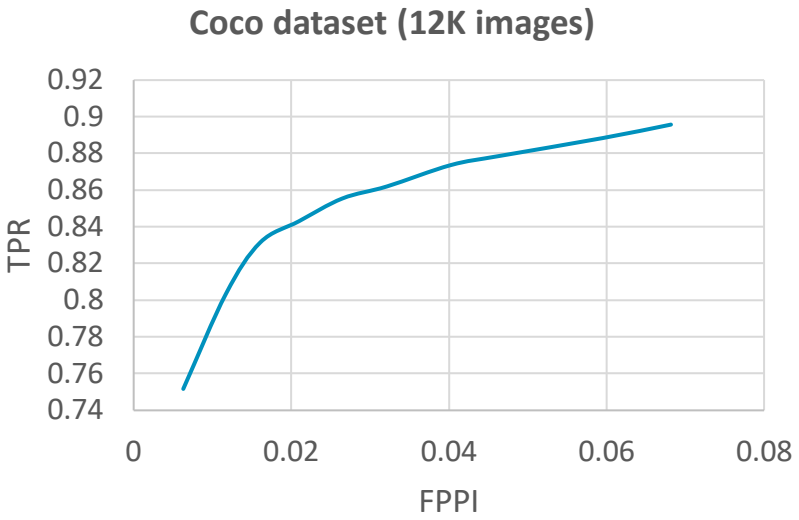
RetinaFace repo

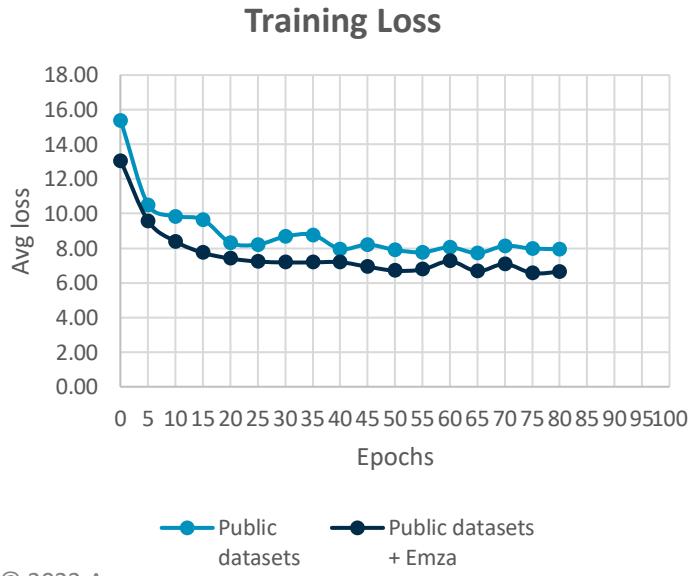
WHENet repo

Training parameters

- Training from scratch – grayscale backbone
- Input resolution – 120x160x1 (distance vs memory)
- Datasets:
  - Choose relevant scene for the application
  - Image quality should be the same as the application image sensor
- Augmentations
- Loss function – classification: cross entropy, regression: smooth L1

Training results



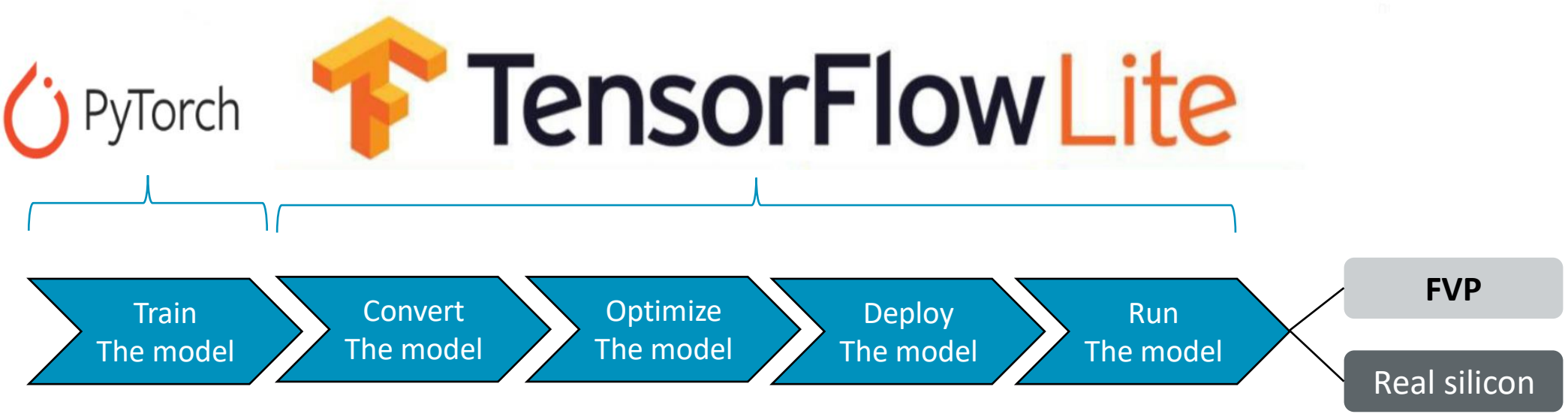


CelebA (10K images)

Landmarks NME 4.5%, Yaw MAE 6.21

arm AI

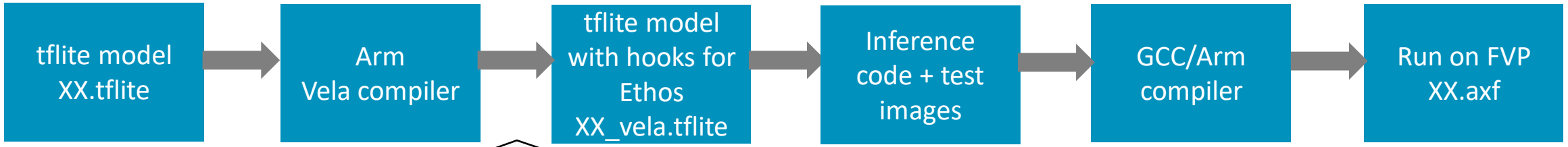
Model conversion



Total params: 355,562
Params size (MB): 1.36

Total params: 355,562
Params size (MB): 0.44

FVP deployment process




```
Network summary for ssd_slim_120x160x1_yaw_landmarks_v3_int8
Accelerator configuration               Ethos_U55_256
System configuration             Ethos_U55_High_End_Embedded
Memory mode                             Shared_Sram
Accelerator clock                              400 MHz
Design peak SRAM bandwidth                    3.20 GB/s
Design peak Off-chip Flash bandwidth          0.40 GB/s

Total SRAM used                            384.62 KiB
Total Off-chip Flash used                  443.94 KiB

106 passes fused into 2
0/231 (0.0%) operations falling back to the CPU
Total cycles                                1302546 cycles/batch

Batch Inference time                 3.26 ms,  307.09 inferences/s (batch size 1)
```

Tensor arena

Model weights

U55 ops delegation

arm AI

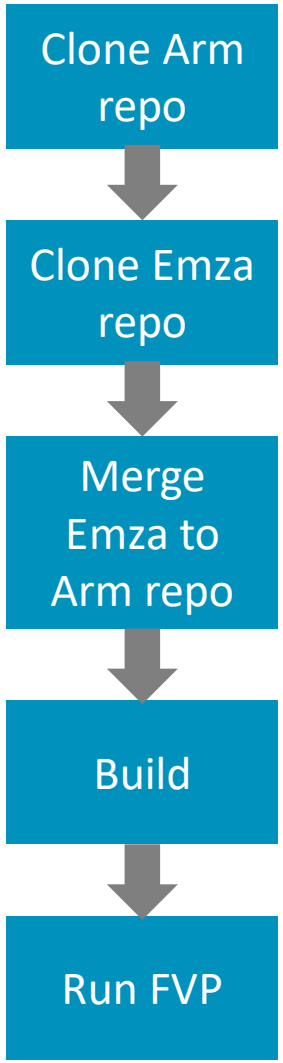
FVP deployment project

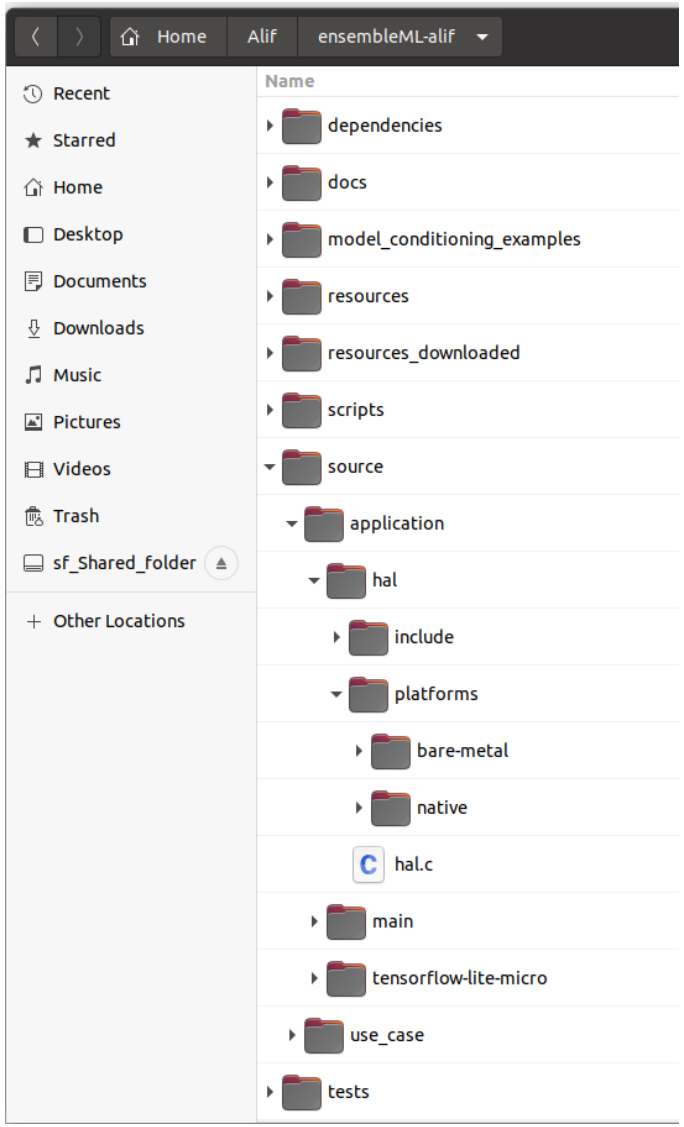
Clone Arm repo

- git checkout -b test_branch ed35a6fea4a1604db81c56fc71f7756822fcf212

Clone Emza repo

- git clone https://github.com/emza-vs/emza_yaw_landmarks_fvp.git





arm AI

FVP deployment – live demo

Deploy on Alif EVB – development process

No changes in CNN model and application

Changes in Alif board relative to FVP:

- Live camera stream
- BSP
- Display

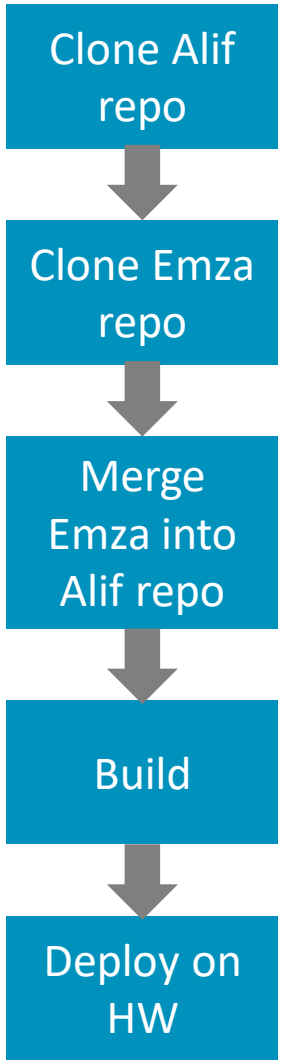
Common to FVP

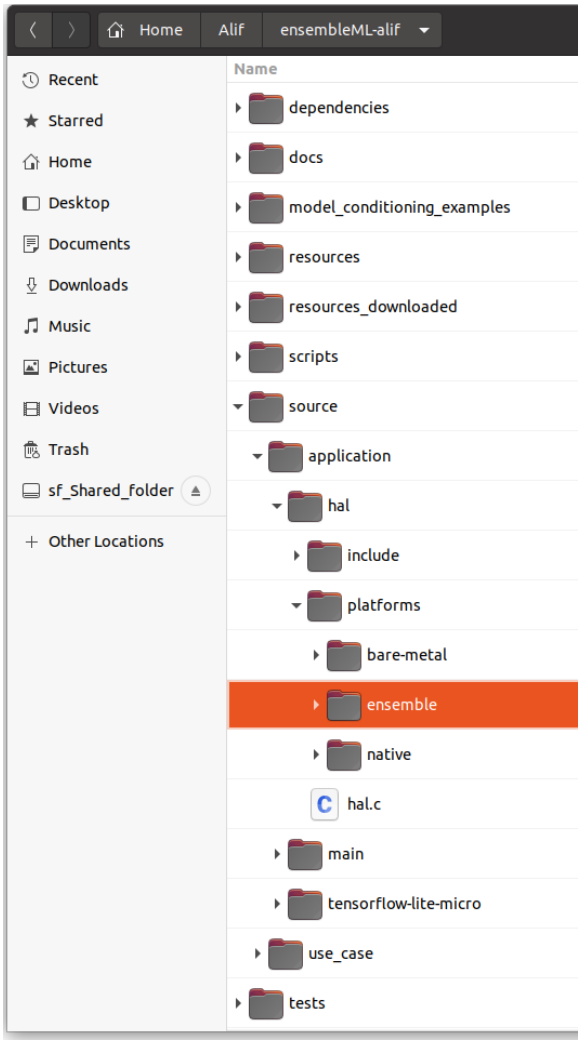
Alif BSP

Deploy on Alif EVB – project

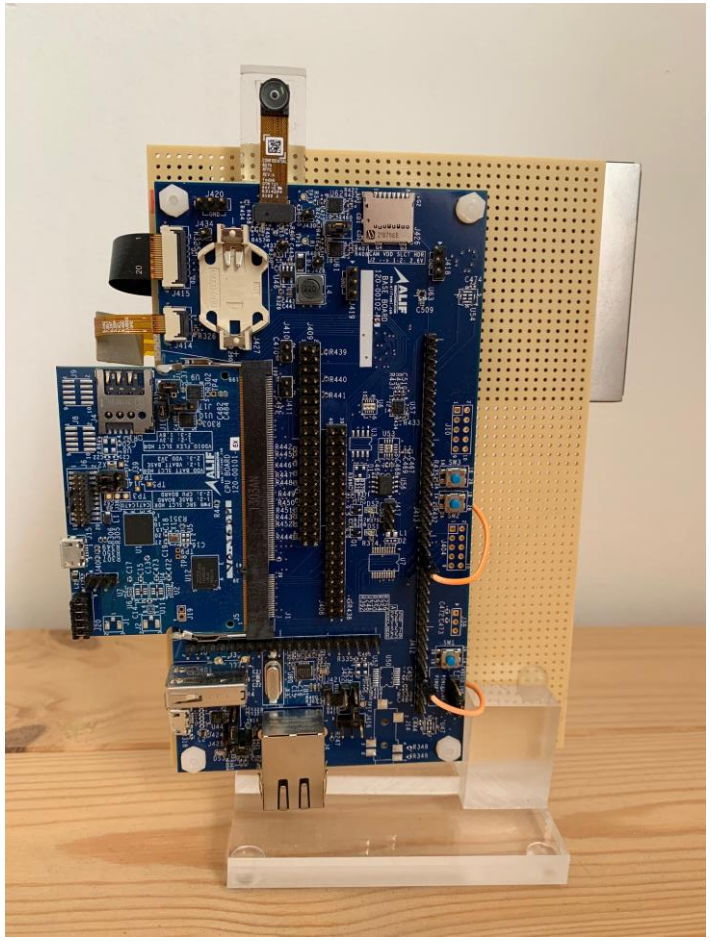
Clone Alif repo – please contact Alif

Clone Emza [repo]

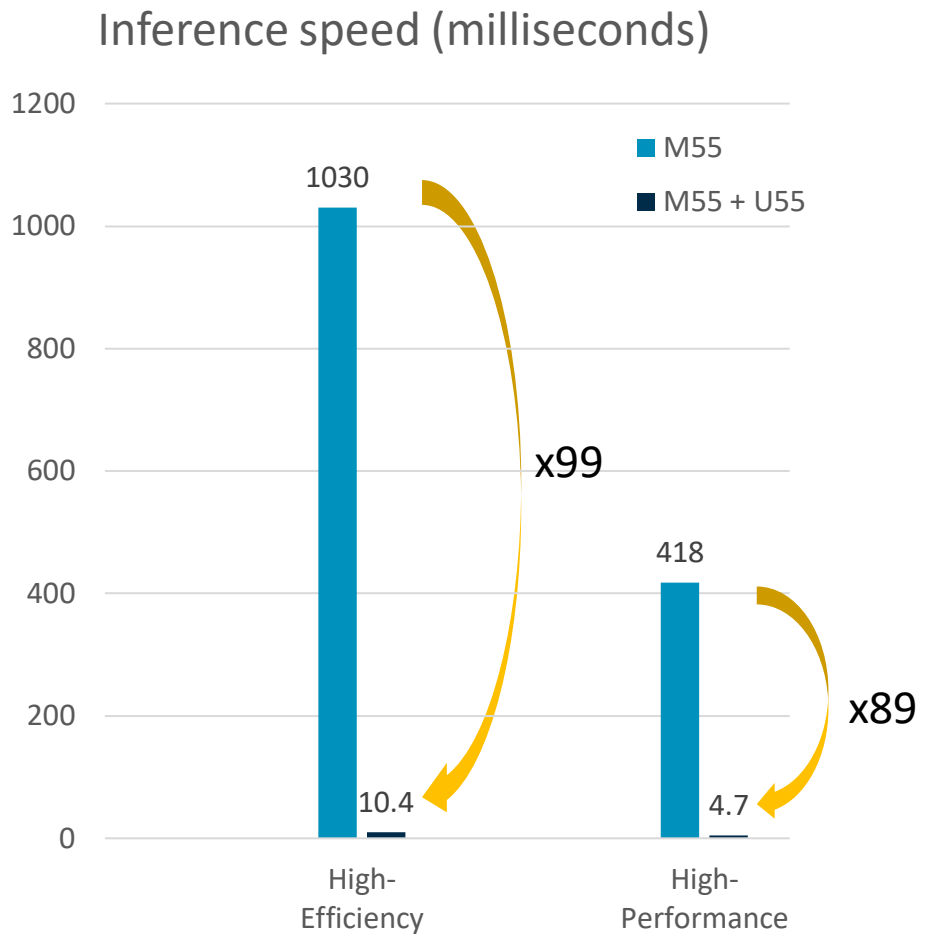




Deploy on Alif EVB – live demo



Deploy on Alif EVB – runtime measurements



arm AI

Summary

FVP enables model development before HW availability and fast deployment into real HW

U55 vs. CPU only - improvement of two orders of magnitude in inference speed

Complex models inference time in less than 5 millisecond!

We expect a wave of new tinyML vision applications leveraging the new class of MCU in the uNPU era

Resources

- https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB
- https://github.com/peteryuX/retinaface-tf2
- https://github.com/Ascend-Research/HeadPoseEstimation-WHENet
- https://review.mlplatform.org/plugins/gitiles/ml/ethos-u/ml-embedded-evaluation-kit/+/refs/tags/22.02
- https://github.com/emza-vs/emza_yaw_landmarks_fvp
- https://github.com/emza-vs/emza_yaw_landmarks_alif
- https://www.emza-vs.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

arm AI

Thank you!

Tweet us: @ArmSoftwareDev -> #AIVTT

Check out our Arm Software Developers YouTube channel

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