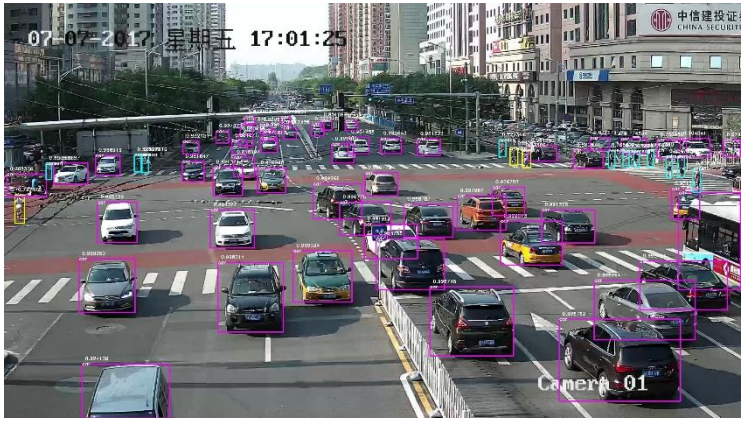
Achieving 15 TOPS/s Equivalent Performance in Less Than 10 W Using Neural Network Pruning on Xilinx Zynq

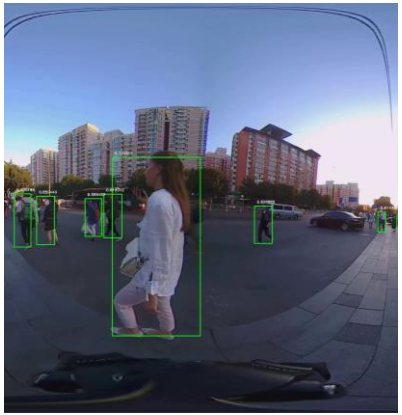
Nick Ni

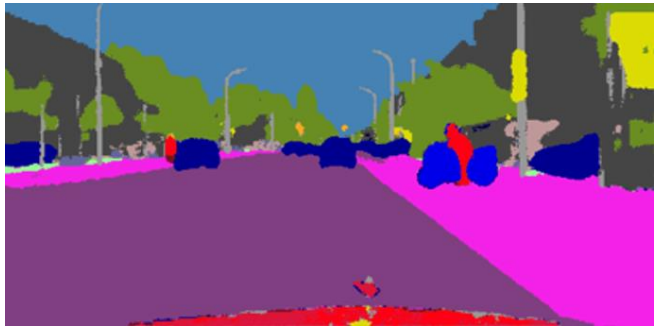
May 23, 2018

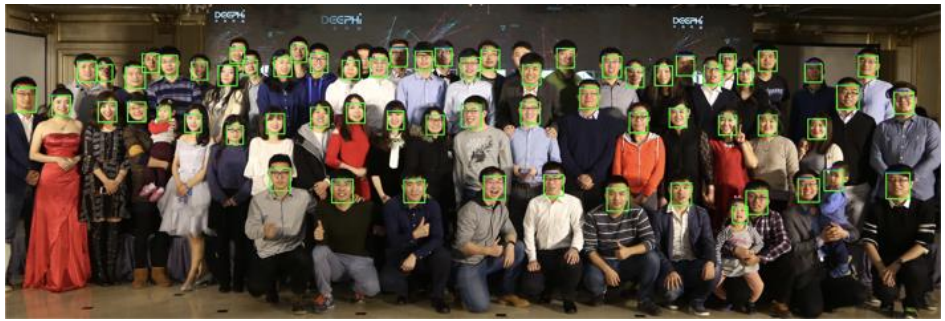
Wide Range of Rapidly Changing Vision Guided Systems

Embedded Vision Systems → Vision Guided Autonomous Systems

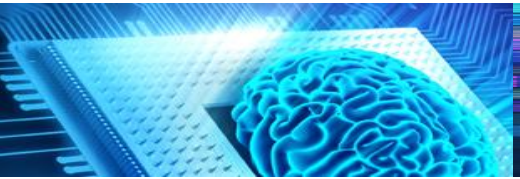








Mandates: From Embedded Vision to Autonomous Systems



Intelligent and Immediate Response with Efficiency

Flexibility to Upgrade to Latest Algorithms & Sensors

Always Connected to Other Machines and the Cloud

Xilinx Unique Application Advantages

Optimized from Sensors to <8-bit Inference & Control

Reconfigurable for Latest Networks & Sensors

Any-to-Any Connectivity

Xilinx Unique Application Advantages

Responsive

Optimized from Sensors to <8-bit Inference & Control

Reconfigurable

Reconfigurable for Latest Networks & Sensors

Connected

Any-to-Any Connectivity

Barrier to Broad Adoption:

Software Defined Programming, Libraries and Frameworks

Software development environment for Embedded Vision

reVISION Stack

Frameworks & Libraries

Application specific

DEEPHi DNNDK

Machine Learning

OpenCV

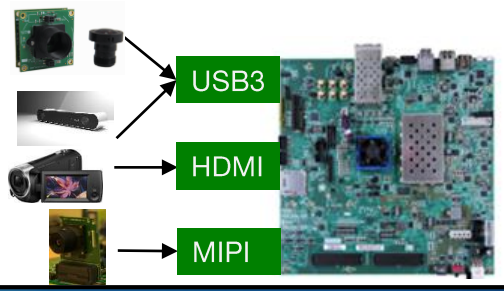
Foundational

Math.h, BLAS…

C/C++ Development tools

SDSoC™ Environment

Ready-to-develop boards



ML inference flow on Xilinx Zynq

Efficient Platform Stack for Deep Learning

Deep Learning

Caffe TensorFlow dmlc mxnet

DL Framework

Compressio
Pruning | Quantization

Compilation
Compiler | Assembler

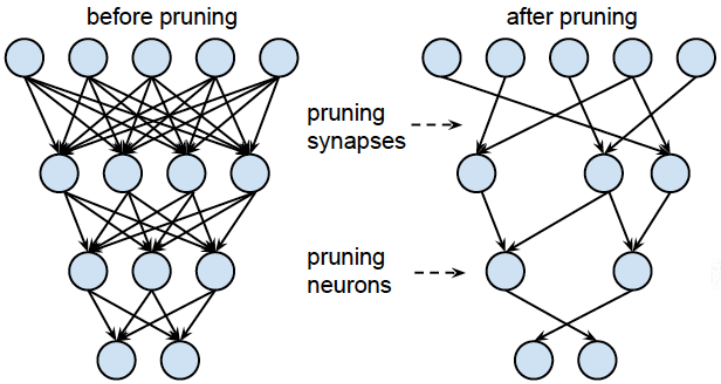
Runtime
Core API | Loader
Driver | Profiler

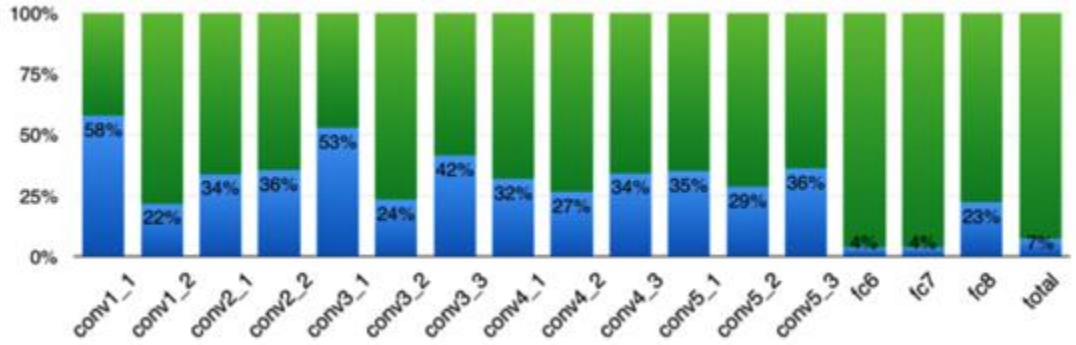
DEEPHi DNNDK

XILINX

Xilinx Zynq-7000 or MPSoC

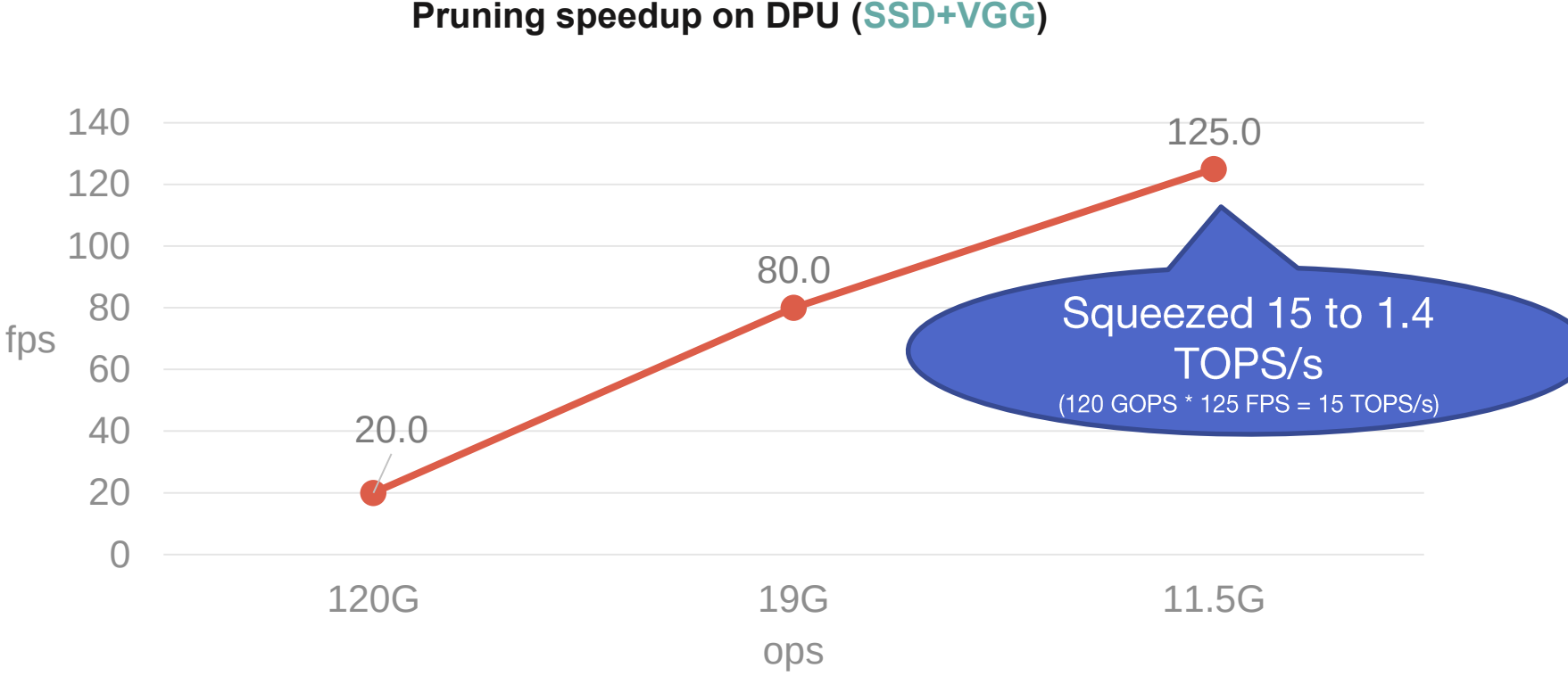
Network Pruning: Achieve same with less





30x – 50x compression rate without hurting accuracy
e.g. AlexNet

Performance Boost Using Pruning

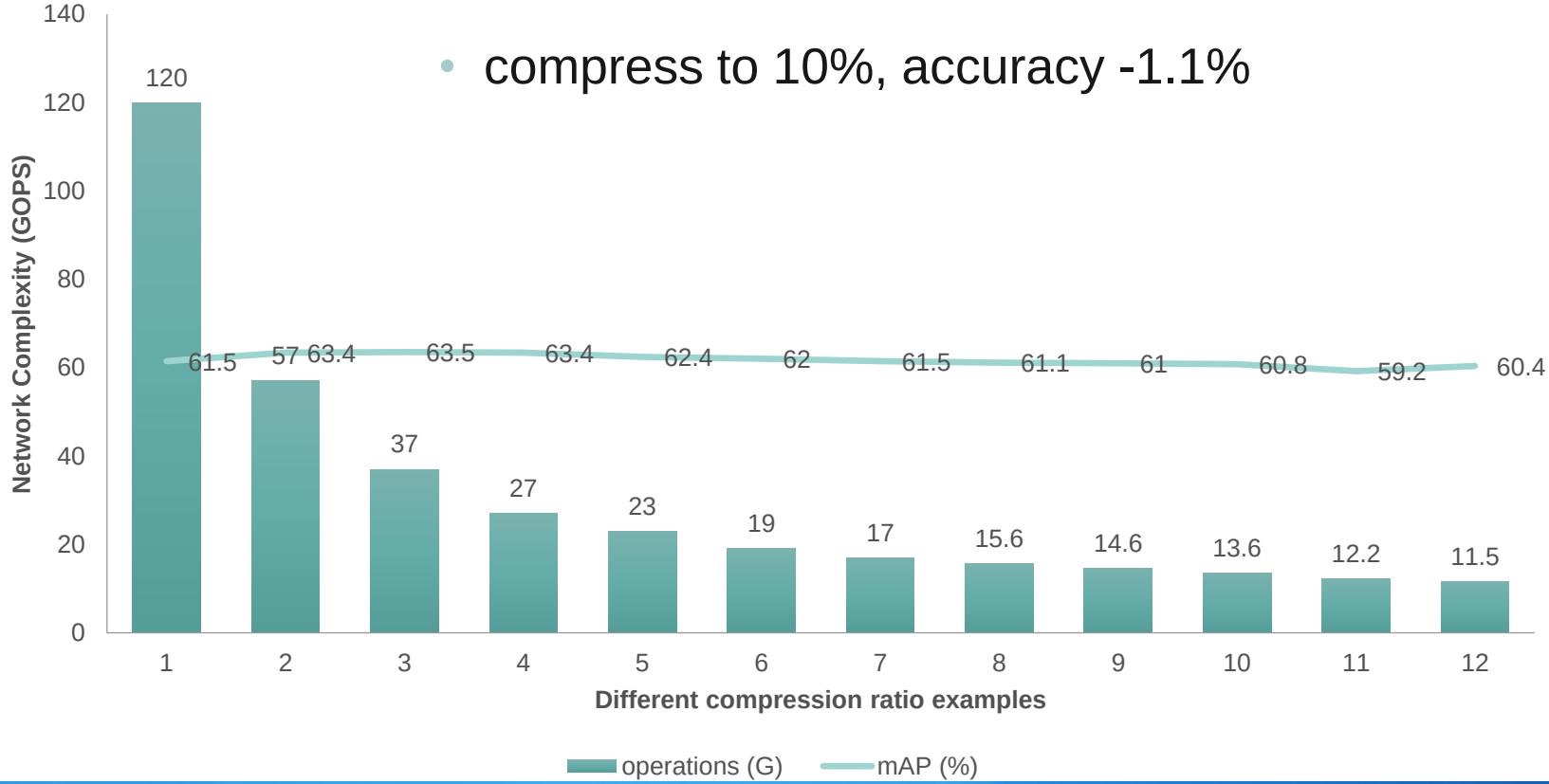



新人民医院往三转盘方向
2017-07-17 11:38:49
新人民医院往三转盘方向
2017-07-17 11:40:15
新人民医院往三转盘方向
2017-07-17 15:30:37
新人民医院往三转盘方向
2017-07-17 11:30:46


新人民医院往三转盘方向
2017-07-17 11:38:49
新人民医院往三转盘方向
2017-07-17 11:40:15
新人民医院往三转盘方向
2017-07-17 15:30:37
新人民医院往三转盘方向
2017-07-17 11:30:46

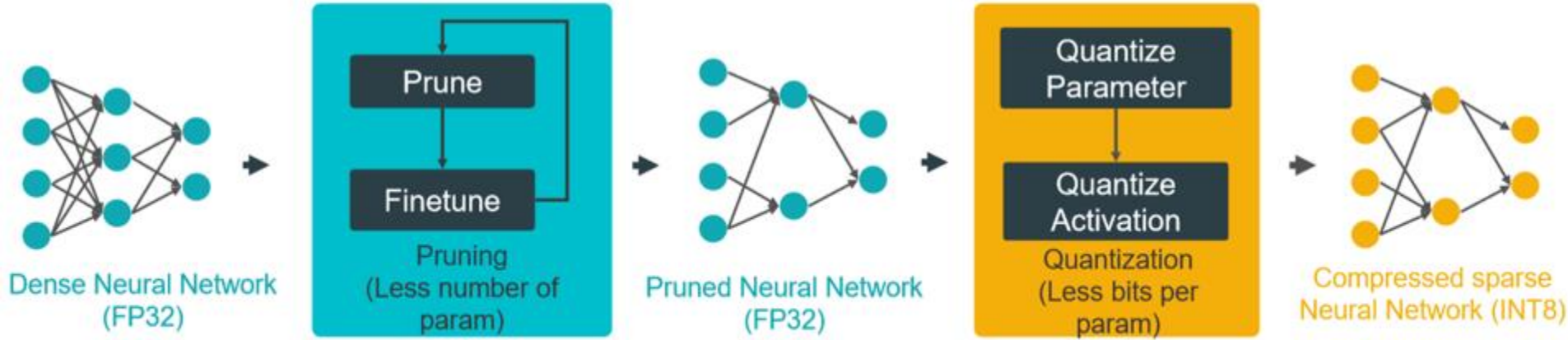
DEEPHi
深 鉴 科 技

Pruning results on SSD (Object detection CNN)



ML flow with pruning

DECENT (Deep Compression Tool)



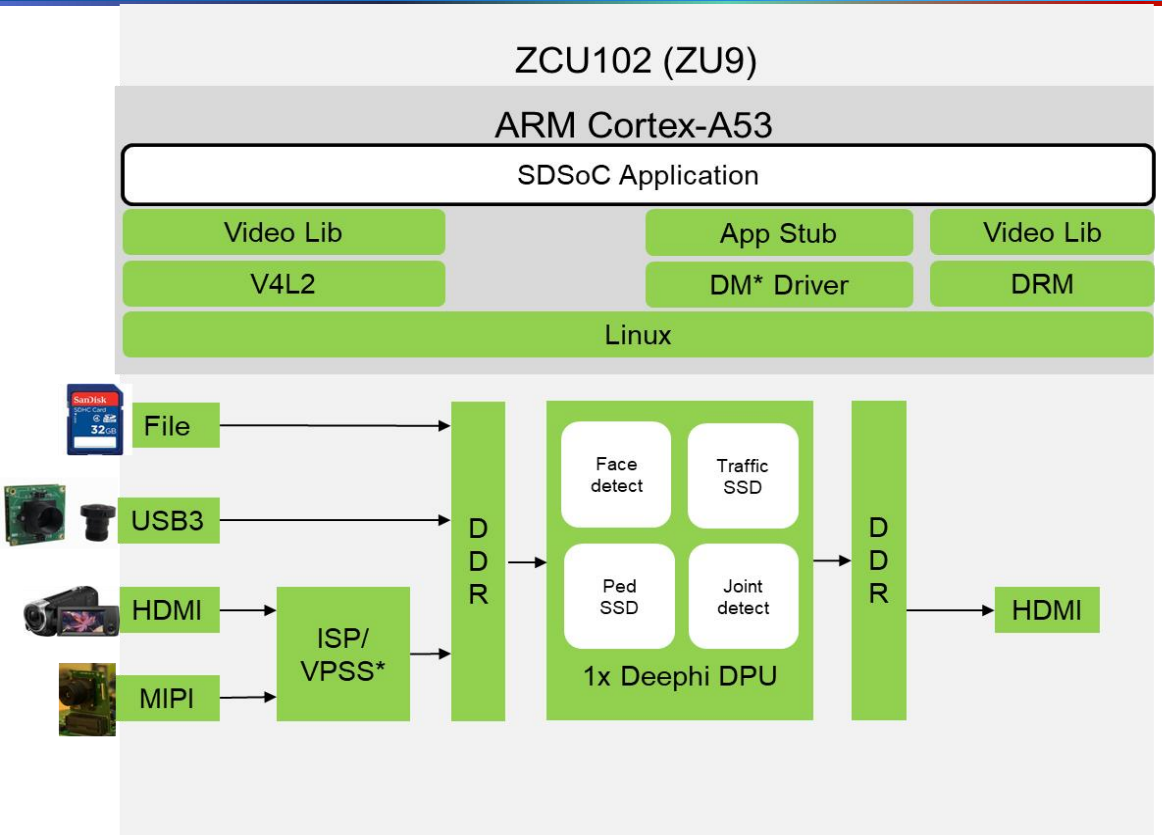
Final results and comparisons with GPU

	GOPs	ZU9	ZU5	ZU2	Z7020	Jetson TX2[1)]
Power		8W	5W	3W	2W	10W
GoogLeNet	3.2	313 img/s	156 img/s	81 img/s	38 img/s	139 img/s
GoogLeNet pruned	1.6	481 img/s	244 img/s	144 img/s	65.6 img/s	N/A
SSD(480x360)	117	20 FPS	10 FPS	4 FPS	2 FPS	10 FPS
SSD(480x360) pruned	11.6	129 FPS	65 FPS	33 FPS	14 FPS	N/A

- Only convolutional / feature extraction parts are calculated and compared

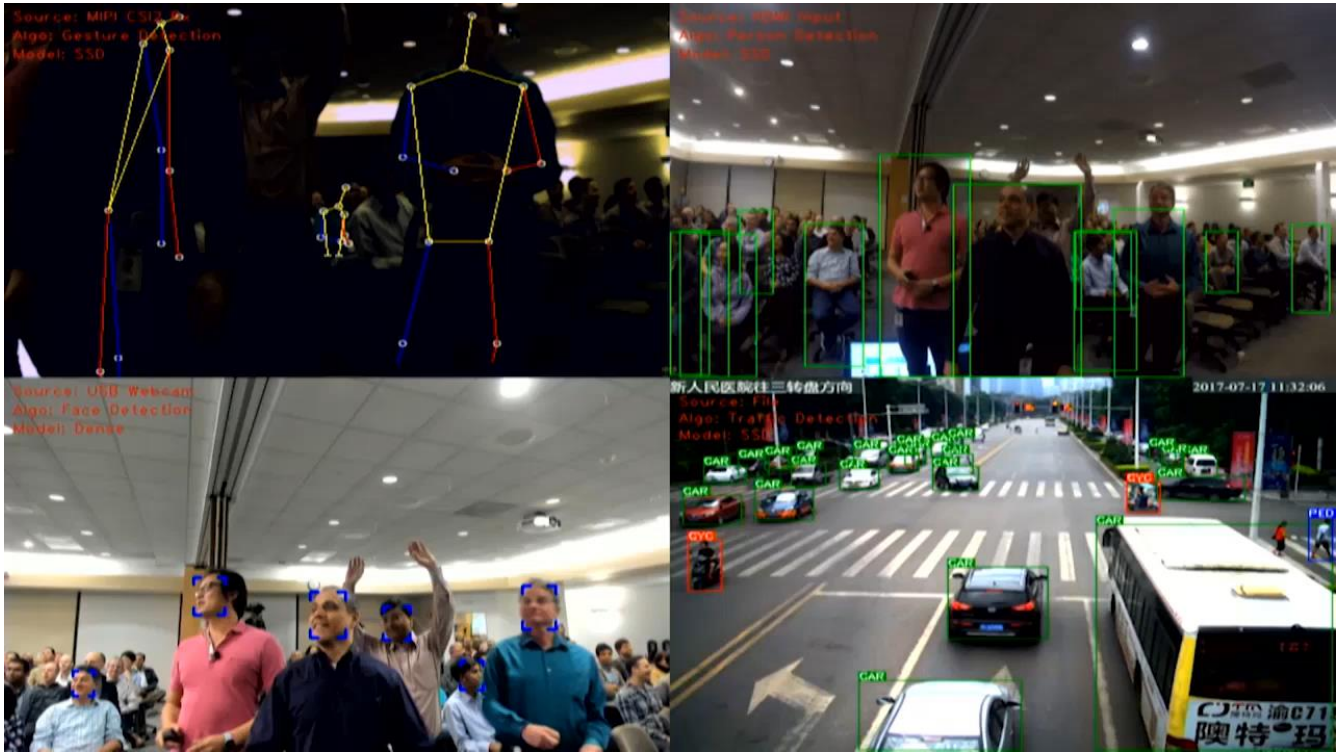
1) Jetson TX2 is tested on batch 1, with JetPack 3.1

Putting it Together: Multi-sensor + Multi CNN



- 4 CNN models
  - Face detect, Joint detect, Traffic SSD, Ped SSD
  - 14, 11, 7, 10 FPS respectively
- 3 Live inputs + file / HDMI output
- Under 10 Watts
- Available in Jun, 2018

Demo Video



Conclusion

- ML in embedded vision needs high perf, low latency and scalable power
- Network pruning can reduce the network complexity up to 10x while maintaining accuracy
- Xilinx partnered with Deephi to offer ML inference toolchain
- Now you can develop a full EV system using SW development environment, SDSoC with ML and OpenCV support
- https://www.xilinx.com/products/design-tools/embedded-vision-zone.html