Deep Learning On Embedded Computers

Adding Smart Sensors To Your Data Collection Pipeline

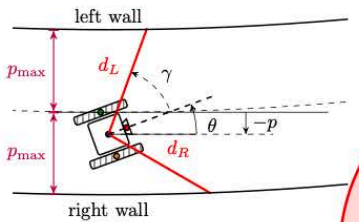
Instructor: Hasan A. Poonawala

Mechanical and Aerospace Engineering
University of Kentucky, Lexington, KY, USA

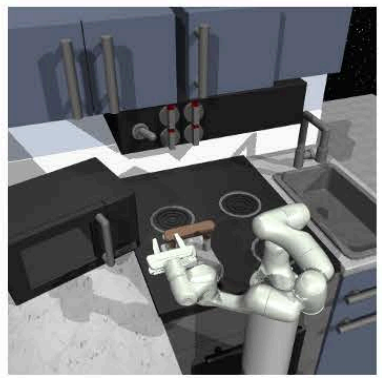
HASAN A. POONAWALA

poonawalalab.github.io

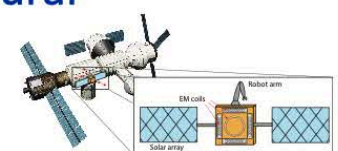
- Robotics & Automation
- Motion planning and control
- Machine Learning for control
- Biological control systems: spiking neural networks
- Sponsors:



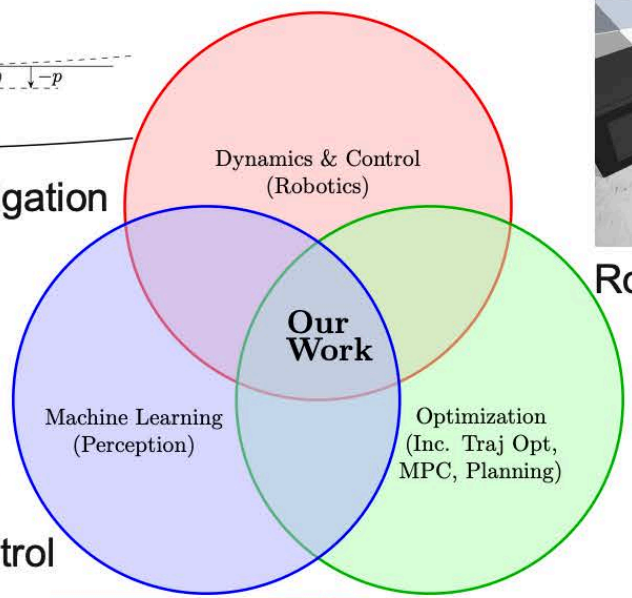
LiDAR navigation

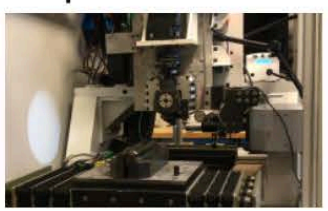


Robotic manipulation



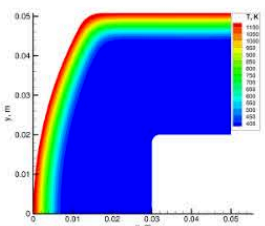
Small Sat EM control





AI-enabled machining

Physics-based ML for Additive Manufacturing and Thermofluid sim



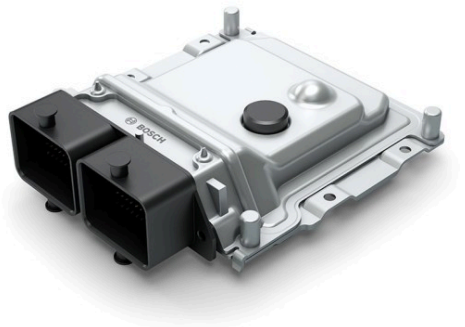
Outline

- Embedded Computers
- Edge ML Applications
  - AI Coach Device
  - Object-aware Robot Navigation
  - Perception for Robot Arm Control
- Non-NVIDIA Microprocessors

Embedded Computers

> An embedded system is a computer system built into a larger mechanical or electronic device to perform a specific function.

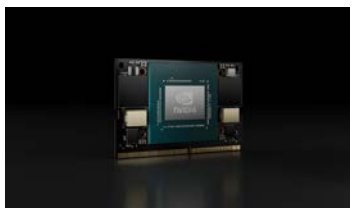
Examples



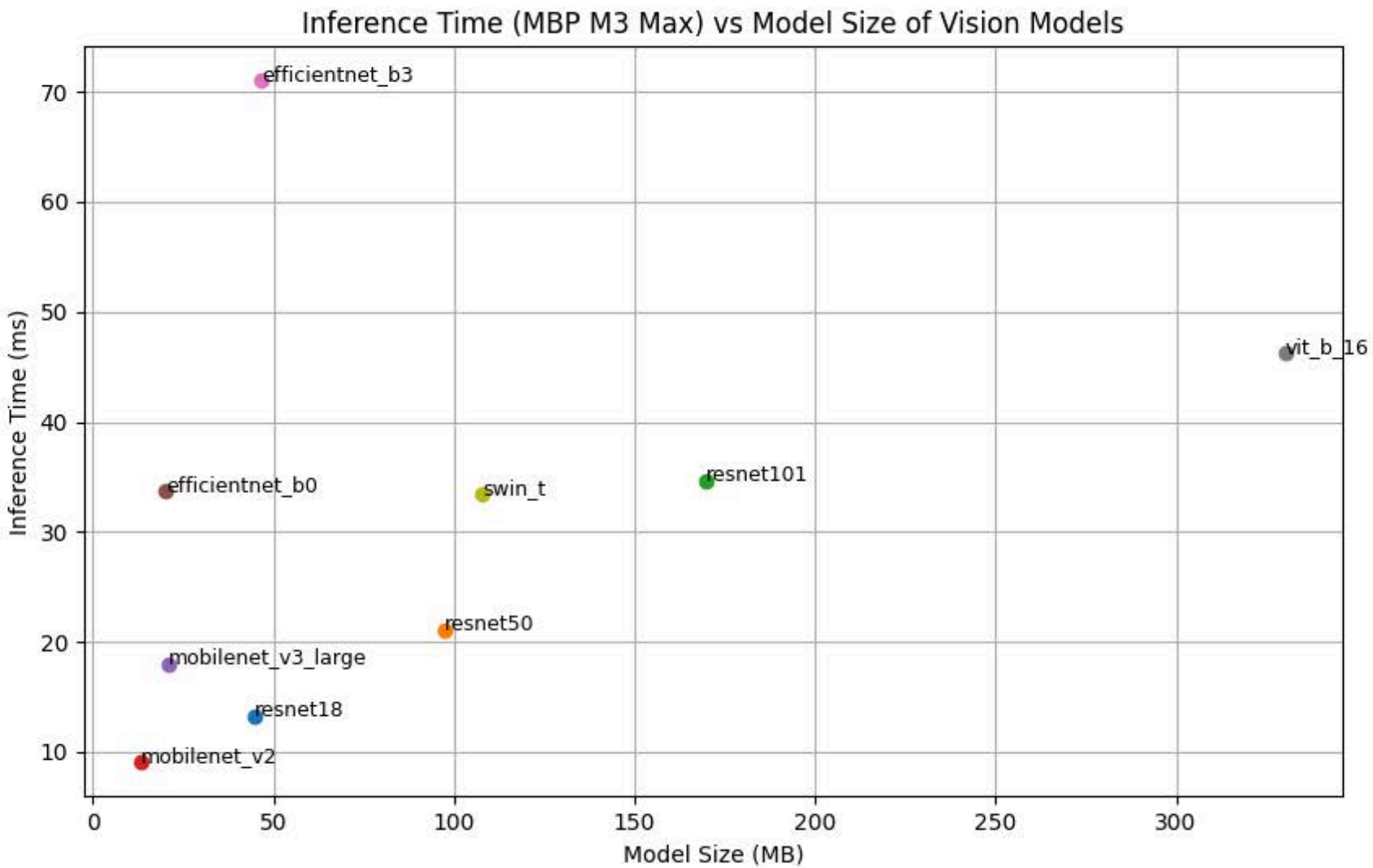
Specs

	Intel Core i9 (PC)	STM32
		
Cores	18	1 or 2
Address length	64 Bits	32 bits
Cache memory	32 MB	32 KB
RAM	16 GB (ext)	1 MB
NV Memory	1 TB (ext)	2 MB
Clock Speed	6.0 GHz	600 MHz
Cost	$300-$500	$8-$20

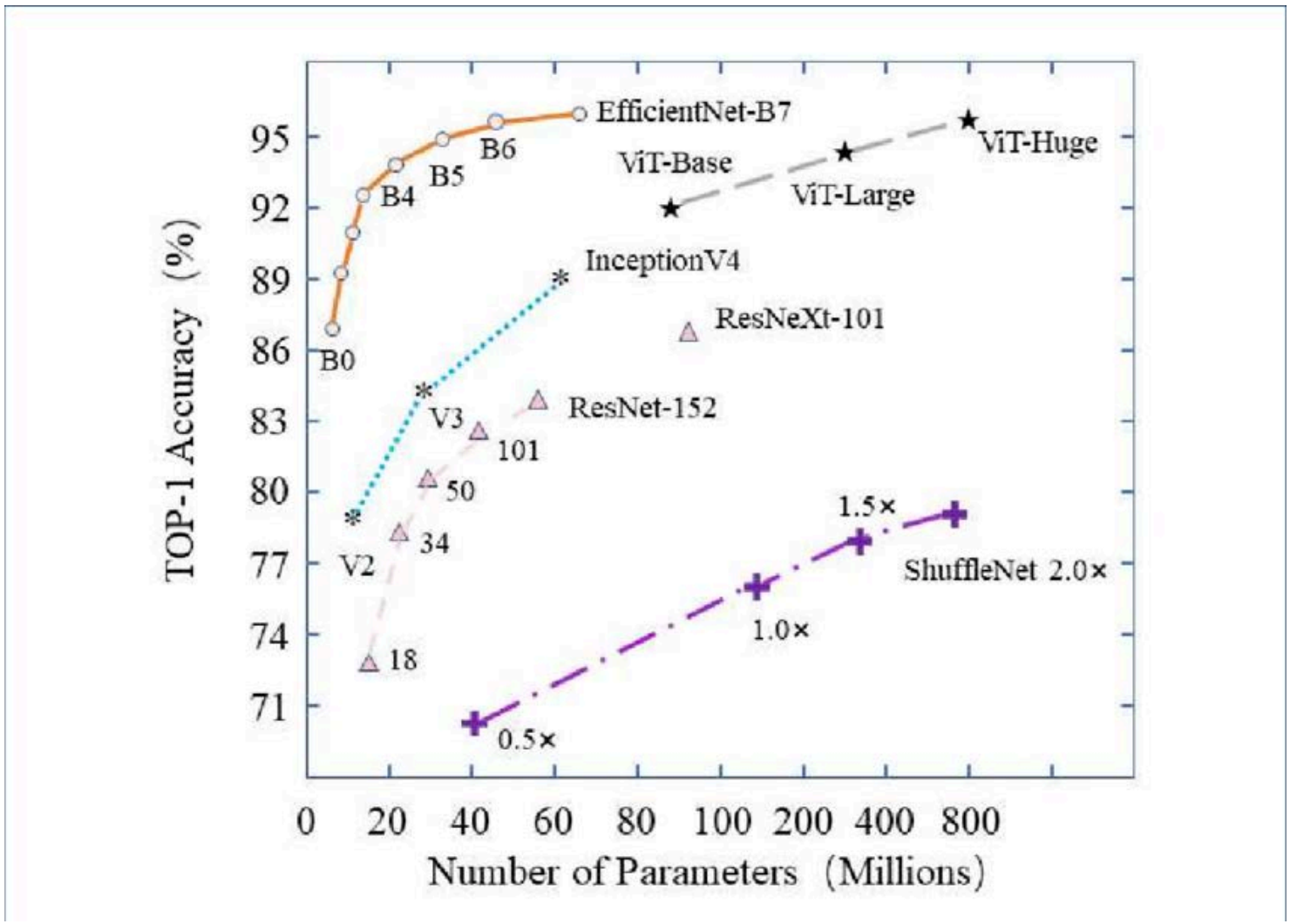
Embedded GPU Specs

	A100	RTX 4070	Jetson Nano	Jetson Orin Nano
				
CPU			ARM Cortex A57 (4)	Arm® Cortex®-A78AE
Memory	80 GB	12 GB	4 GB	8GB
Cores	6912 CUDA, 432 Tensor	5888 Shader, 184 Tensor	128 CUDA	1024 GPU, 32 Tensor
TFLOPS FP32	19.2	29	0.4	67 TOPS
Bandwidth	2000 GB/s	500 GB/s	25.6 GB/S	100 GB/s
Cost	$8000	$600	$100	$250

Inference Time



Accuracy



Source: MulTiNet: Multimodal Neural Networks for Glaucoma Based on Transfer Learning

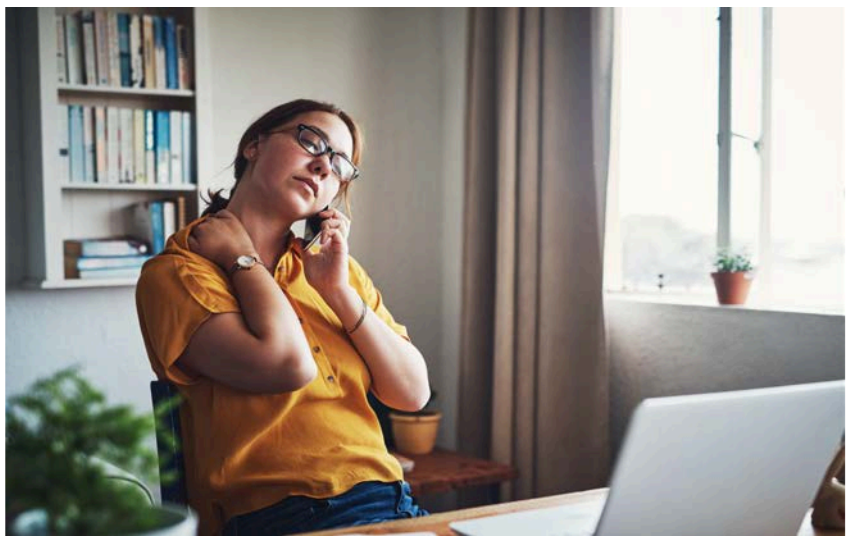
AI Coach

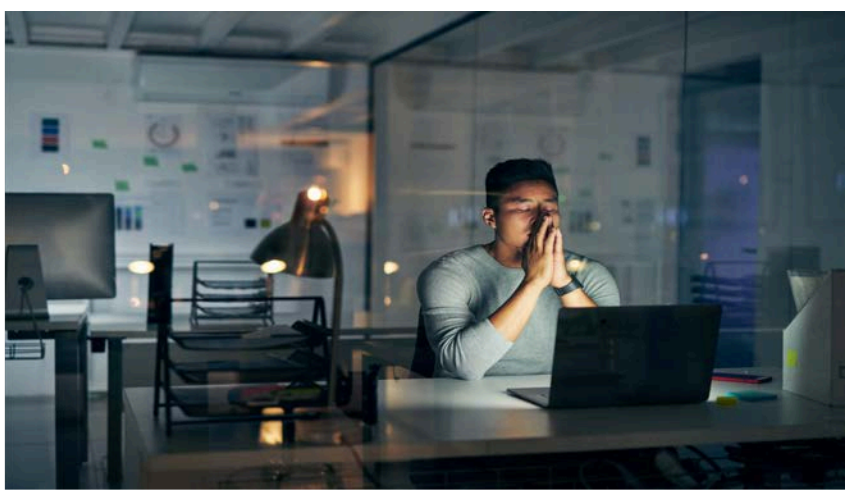
Background

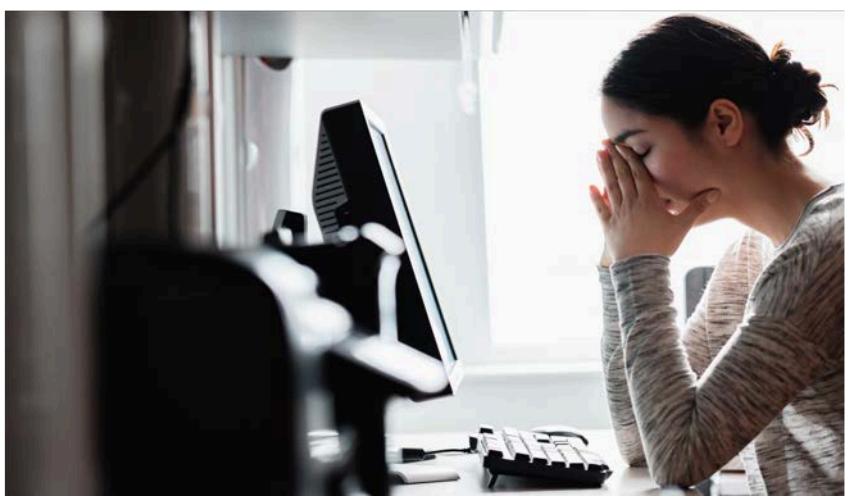
- 2020 UK Igniting Research Grants pitch:

> Use new GPU-enabled embedded systems to provide real-time feedback on body movement

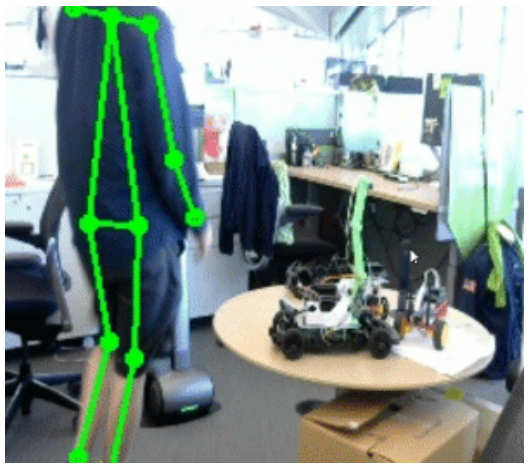
Background



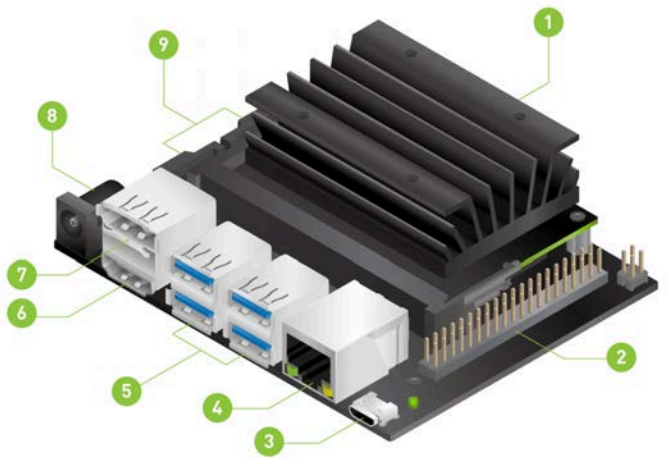




TRT-Pose



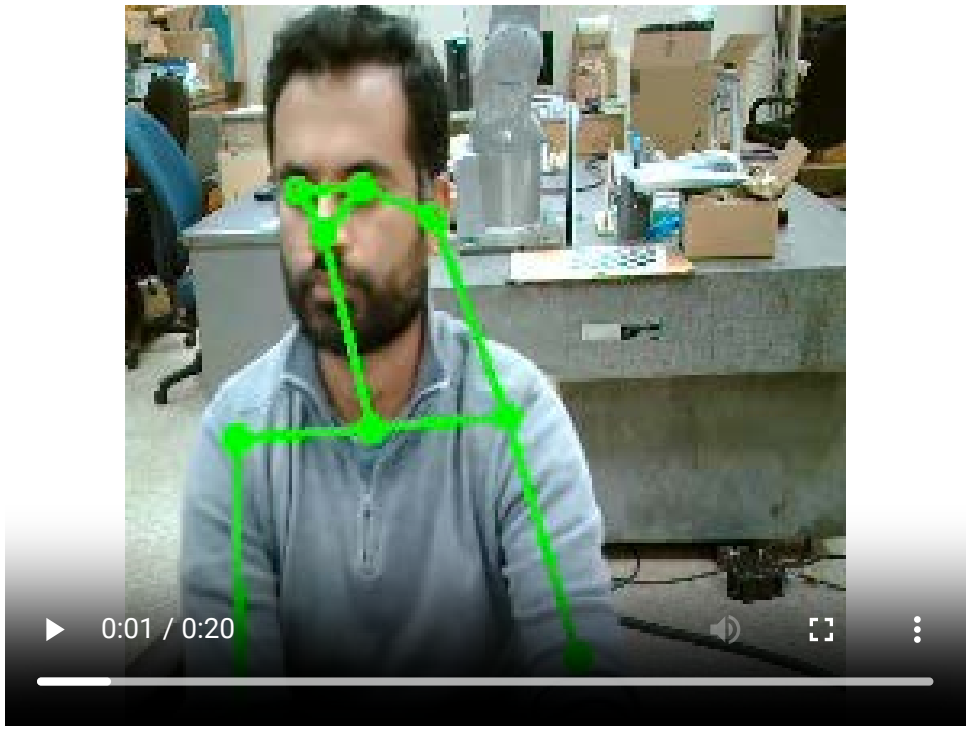
Demo

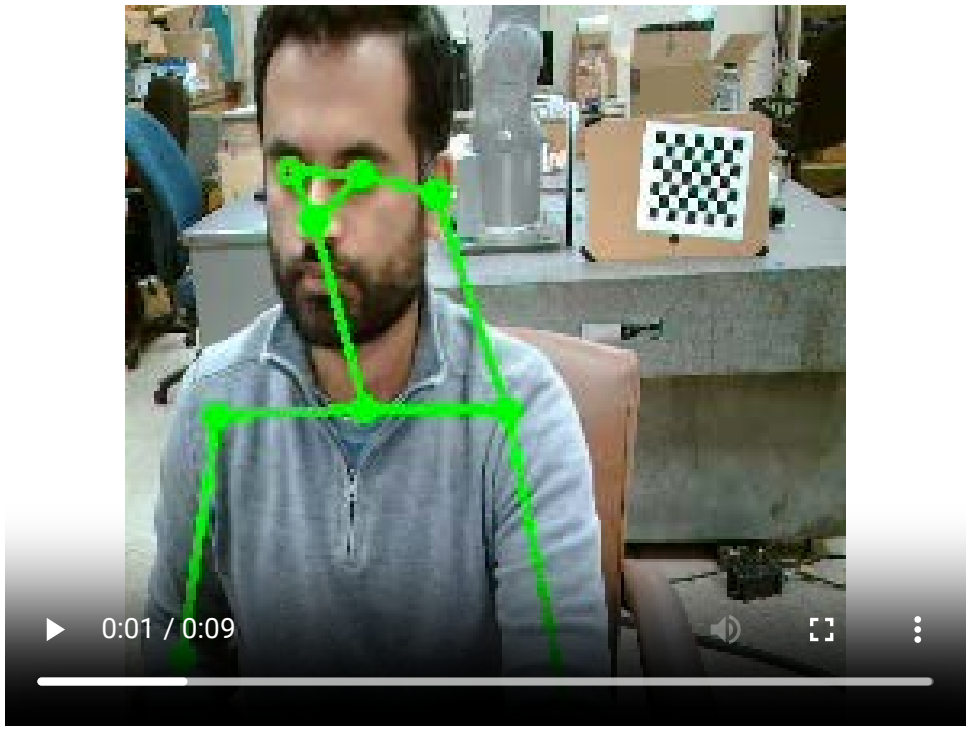


Jetson Nano

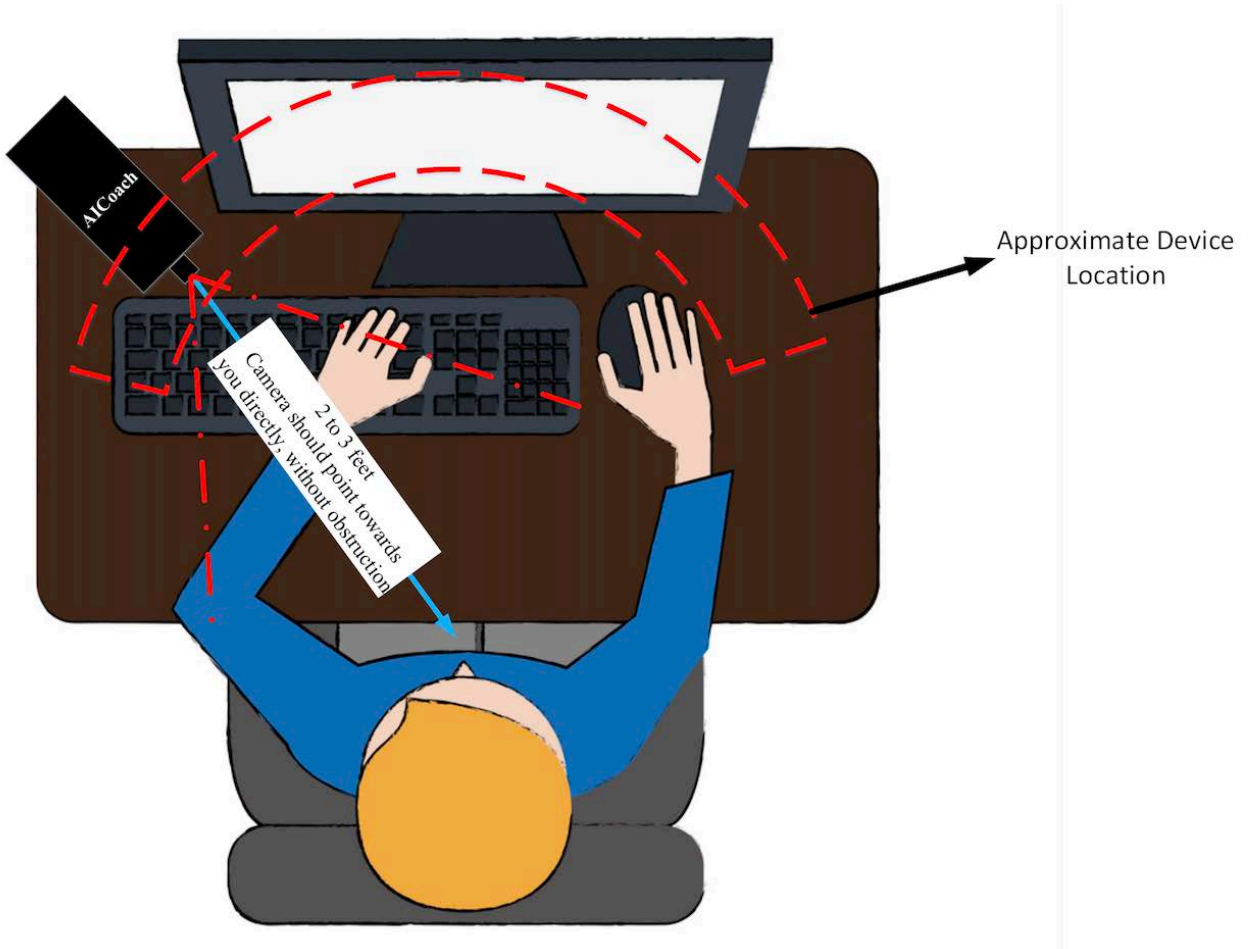
- TRT: TensorRT, NVIDIA's high-performance deep learning inference library
- TRT Pose: optimized deep learning model for real-time pose estimation on NVIDIA Jetson devices
- 80 MB Model

TRT Pose



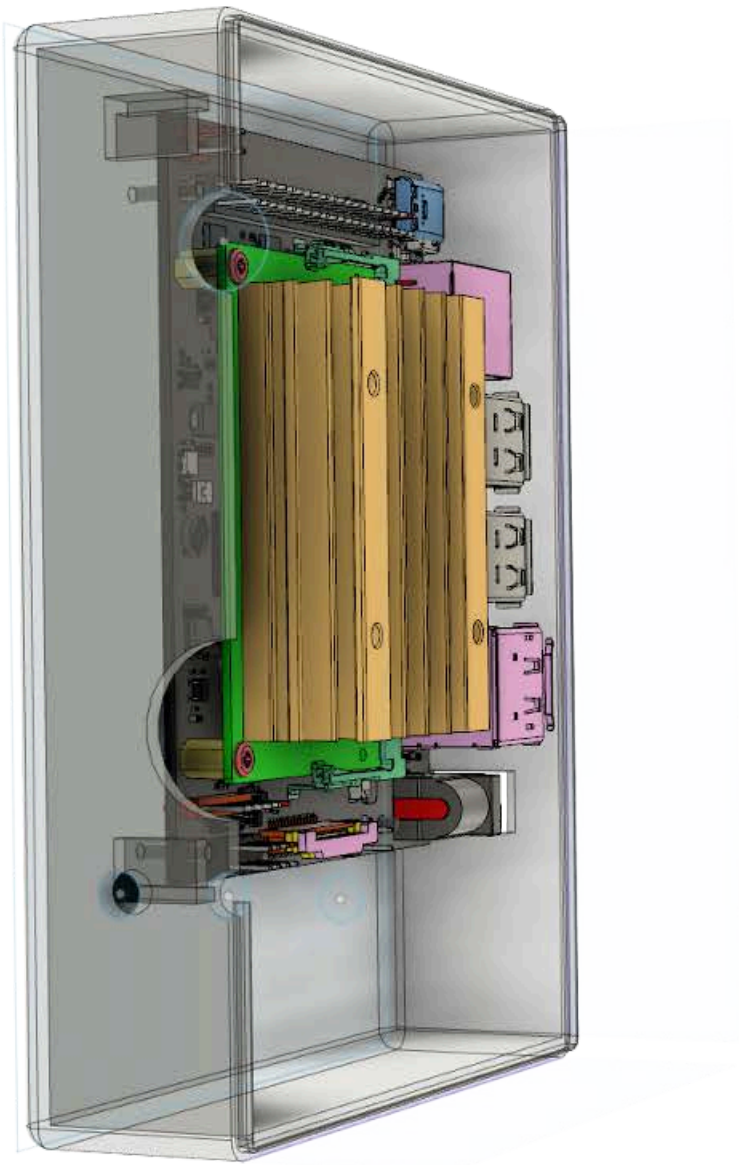


Research Goal

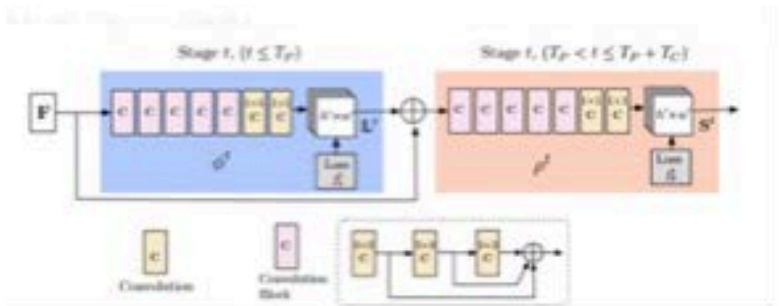


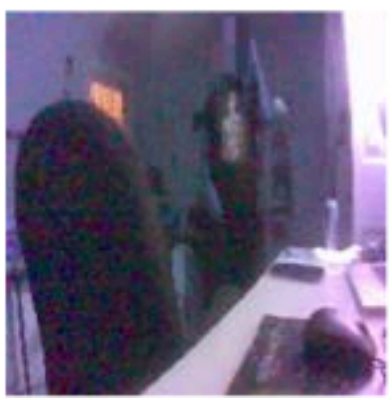
Observe study participants in their home while they work at a desk.

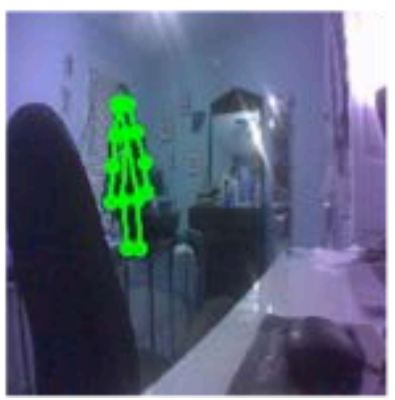
AI Coach Device

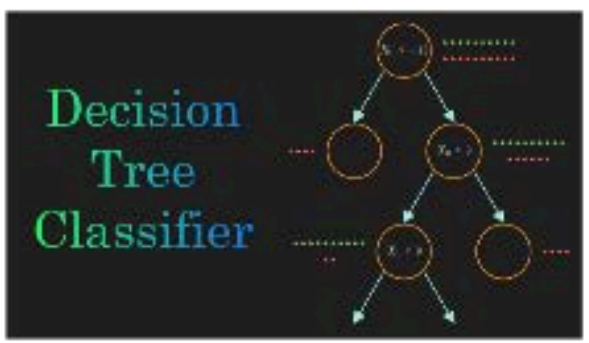


Sit vs Stand vs DK Classification



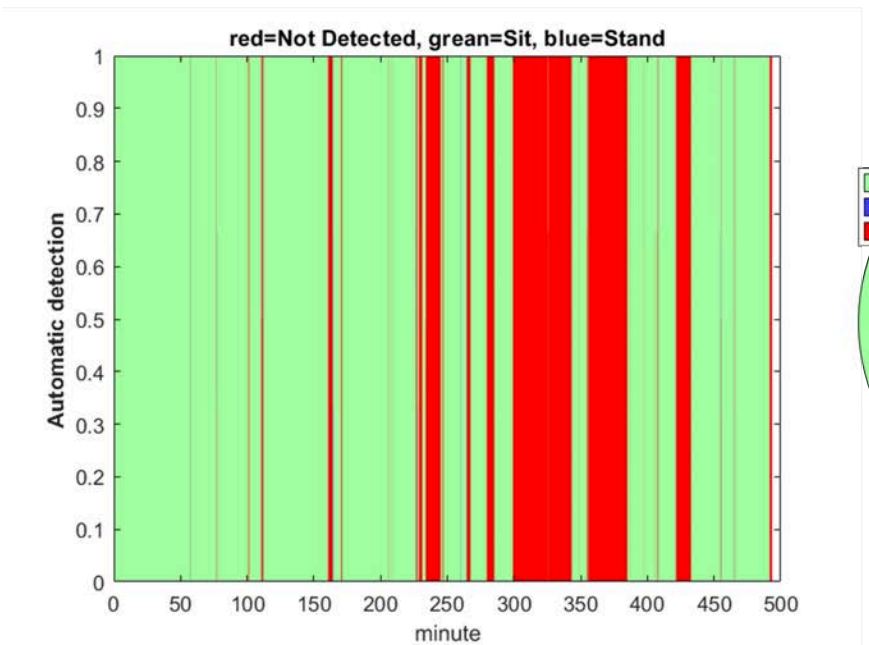


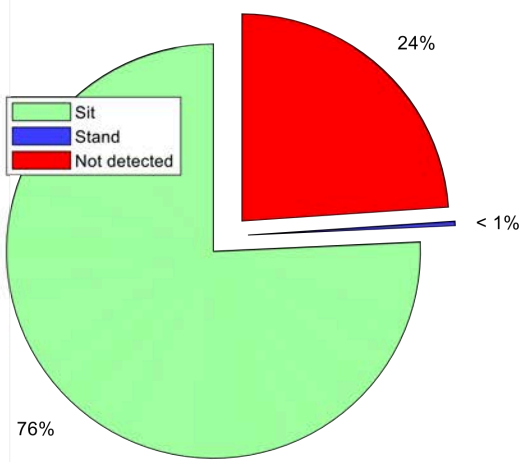




Example Data 1

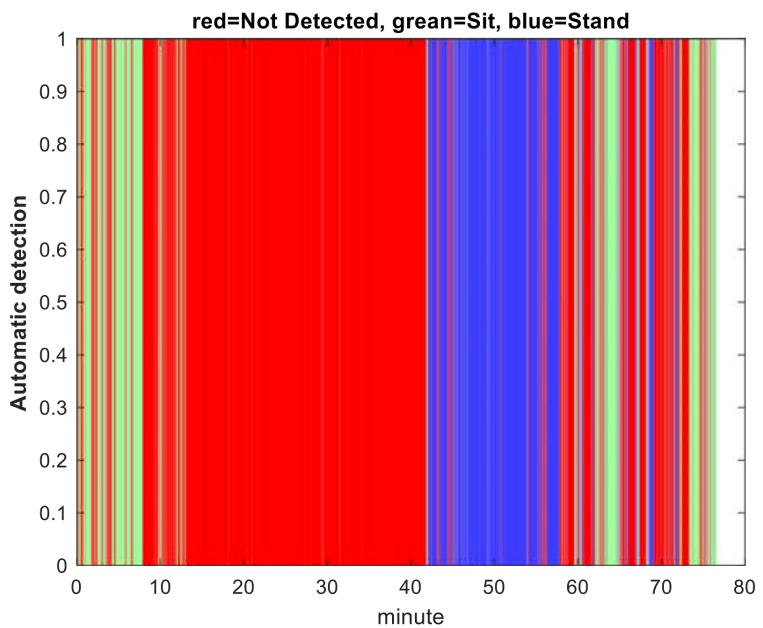
Type	Duration (minutes)
Total	480
Seated	360
Standing	5
Not visible	115

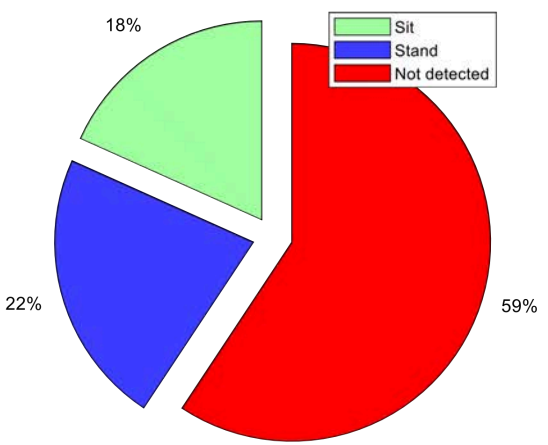




Example Data 2

Type	Duration (minutes)
Total	75
Seated	14
Standing	17
Not visible	44





Highlights

Extended remote data collection

Collected data on a ~$200 device over weeks in participant's homes

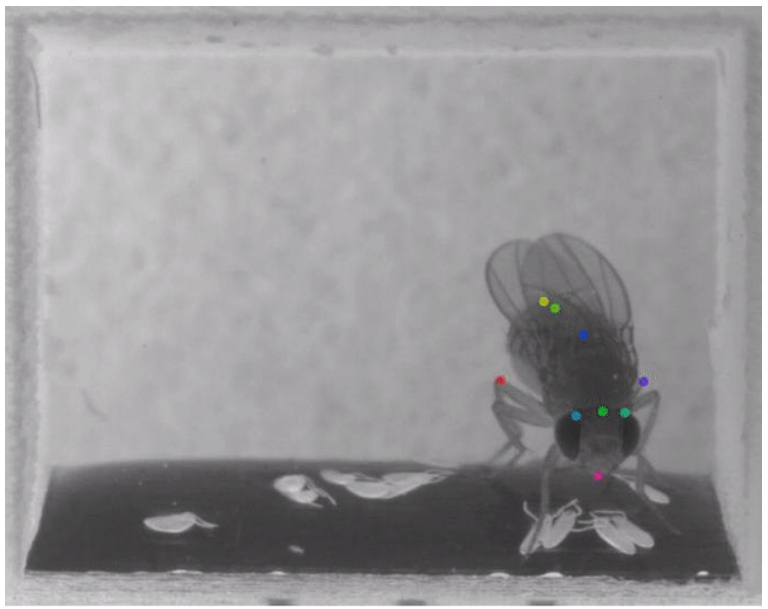
No Wi-Fi

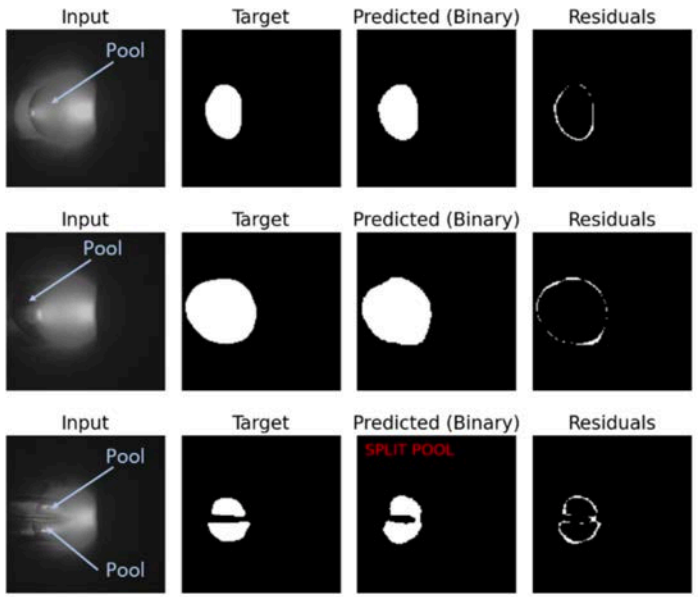
Avoids privacy and cybersecurity concerns (and complex IRB)

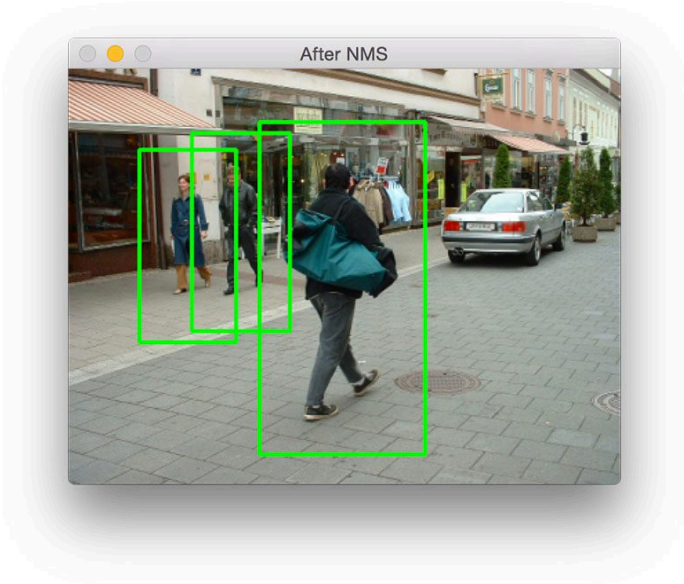
Better Models/Devices

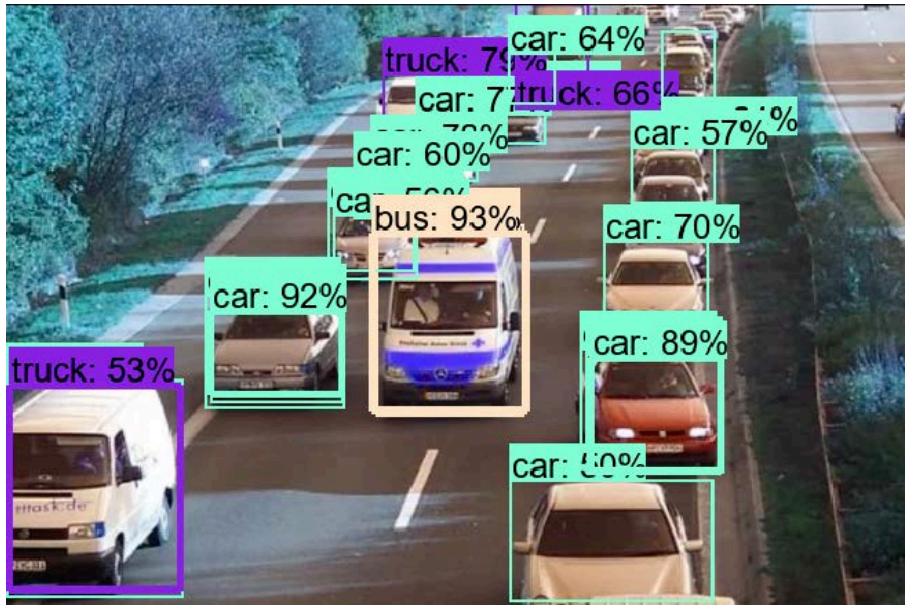
AI Models and embedded ML devices have advanced a lot since 2021.

DL for Smart Sensors and Data



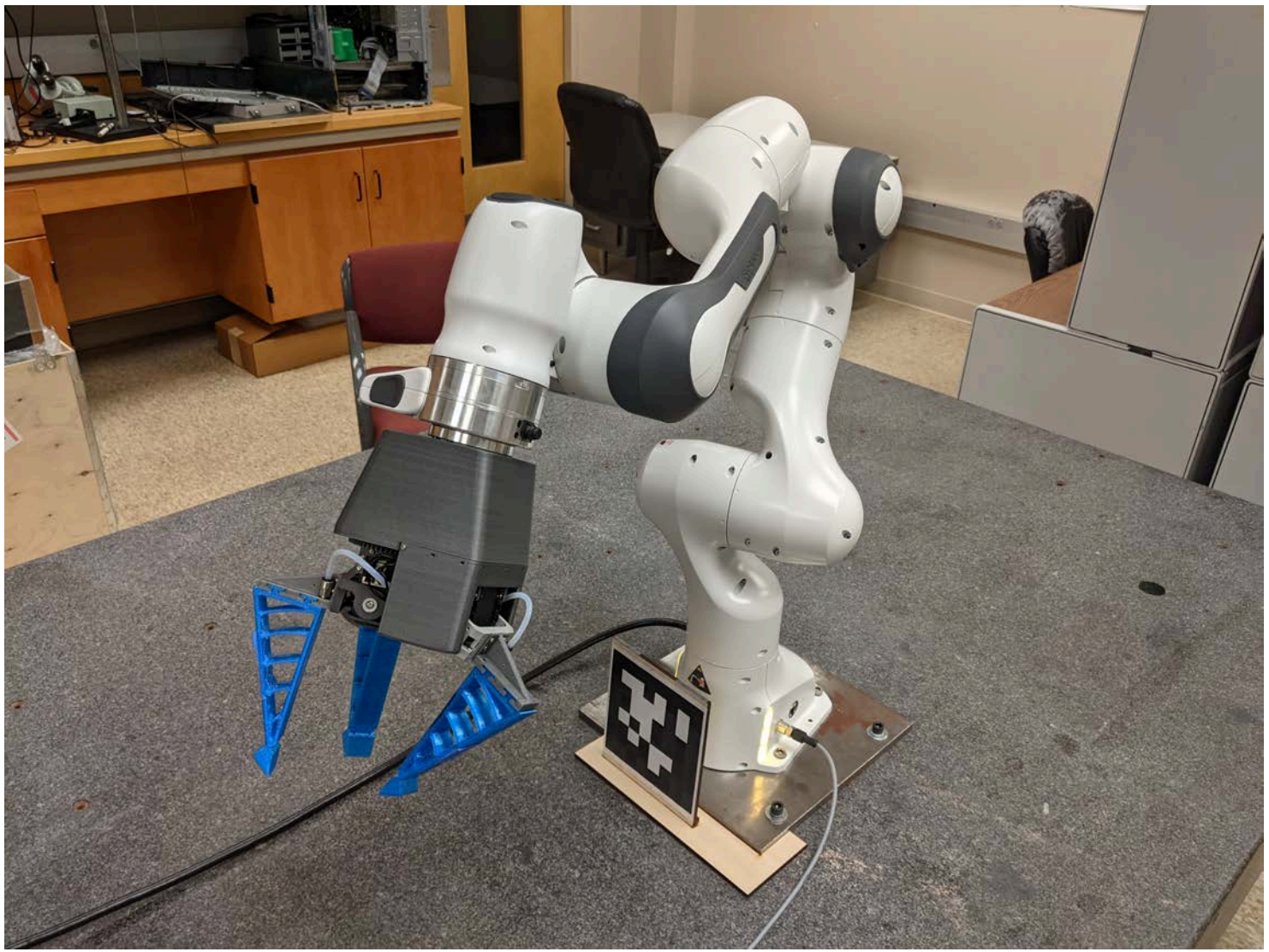



After NMS

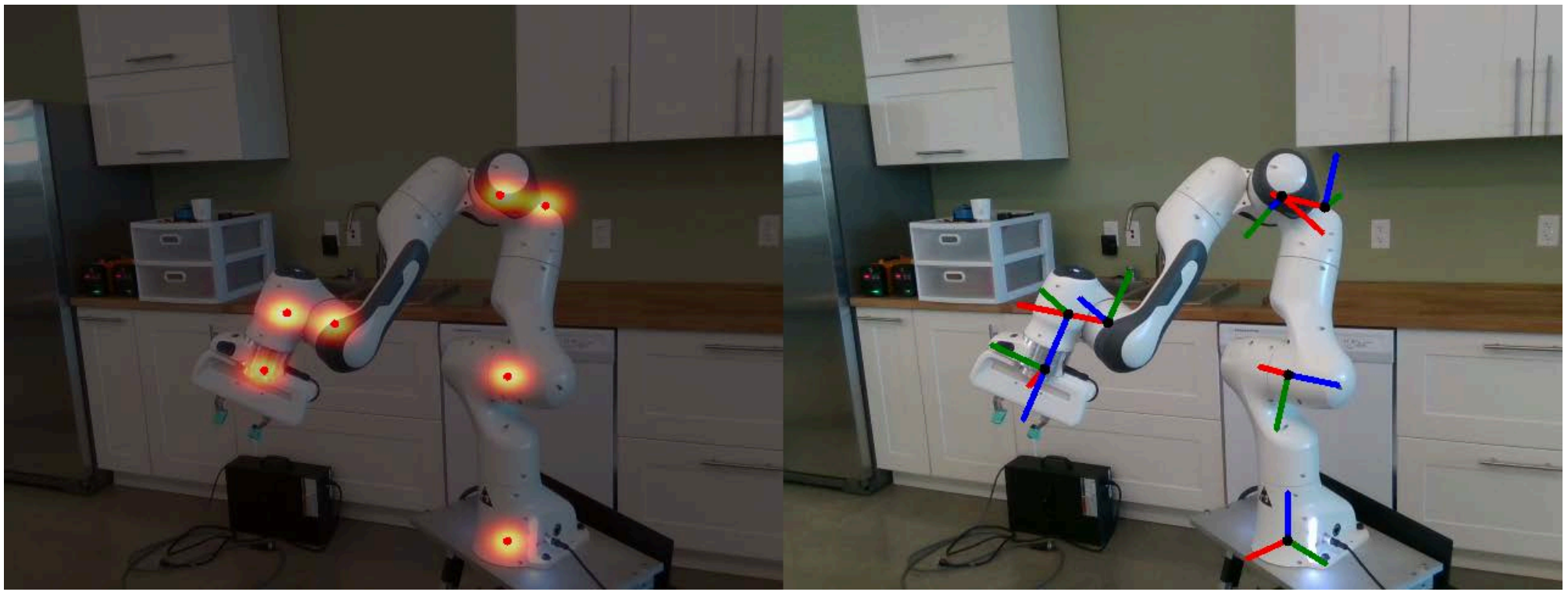

car: 64%
truck: 66%
car: 57%
car: 60%
bus: 93%
car: 70%
car: 92%
car: 89%
truck: 53%

Tracking Specific Objects

Tracking Robot Arms

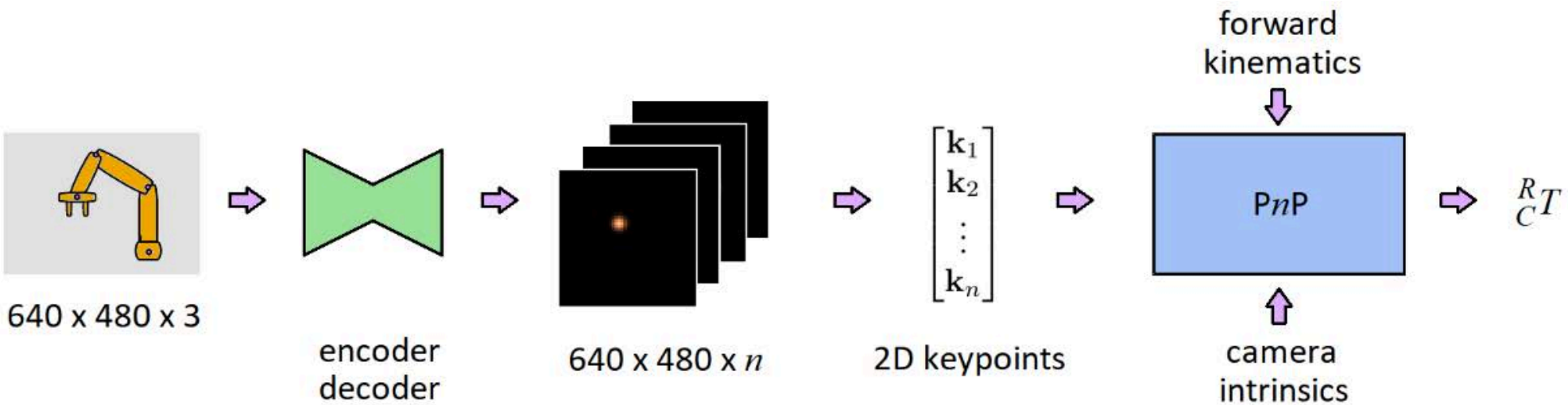


DREAM Model



https://sim2realai.github.io/assets/img/2020-06-29/Lee_etal_2020_dream_panda_reaching_frame.png

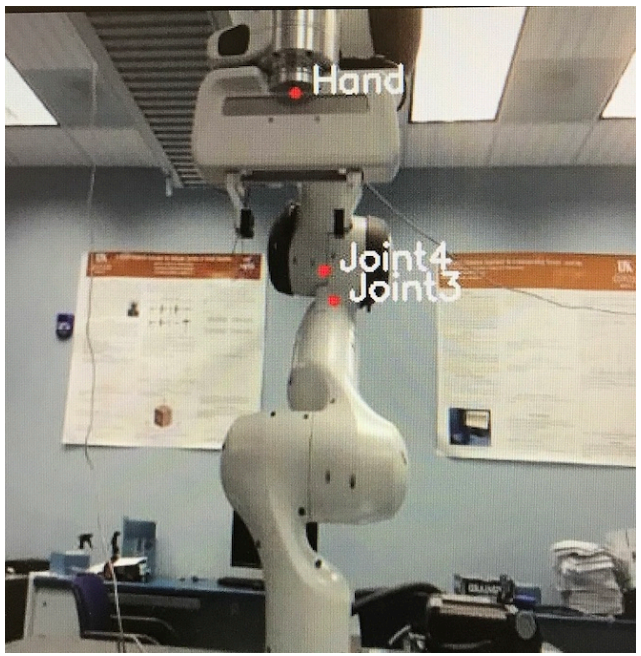
DREAM Model

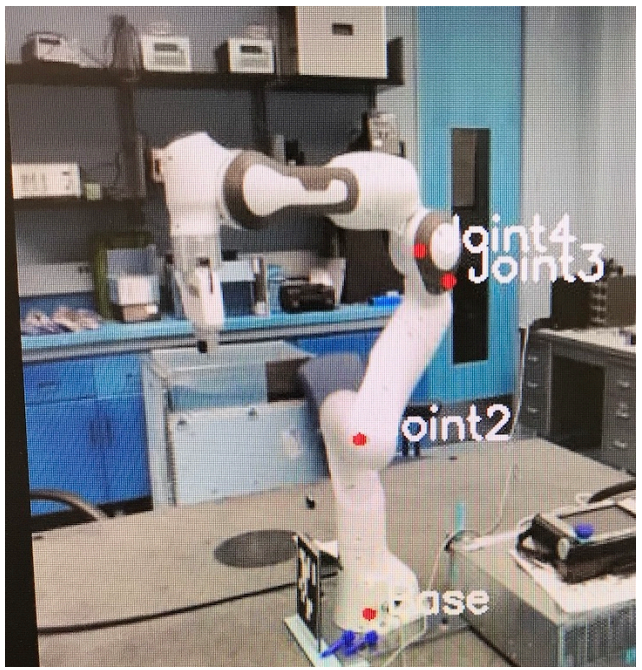


https://sim2realai.github.io/assets/img/2020-06-29/Lee_etal_2020_dream_pipeline.png

- 80 - 200 MB model depending on the backbone (VGG-X vs ResNet-X)

DREAM On Jetson Nanos





Tracking Manipulation Objects

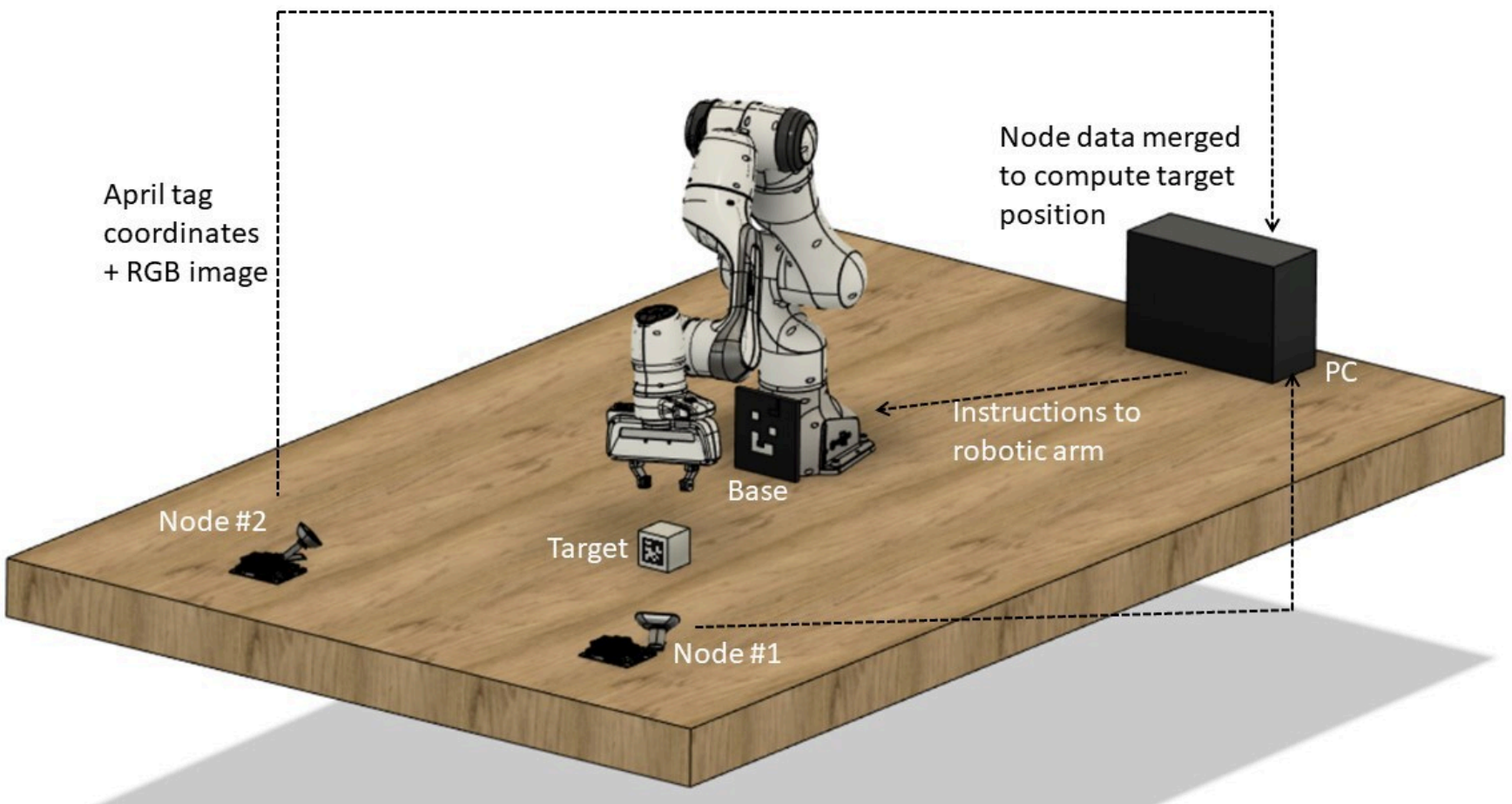
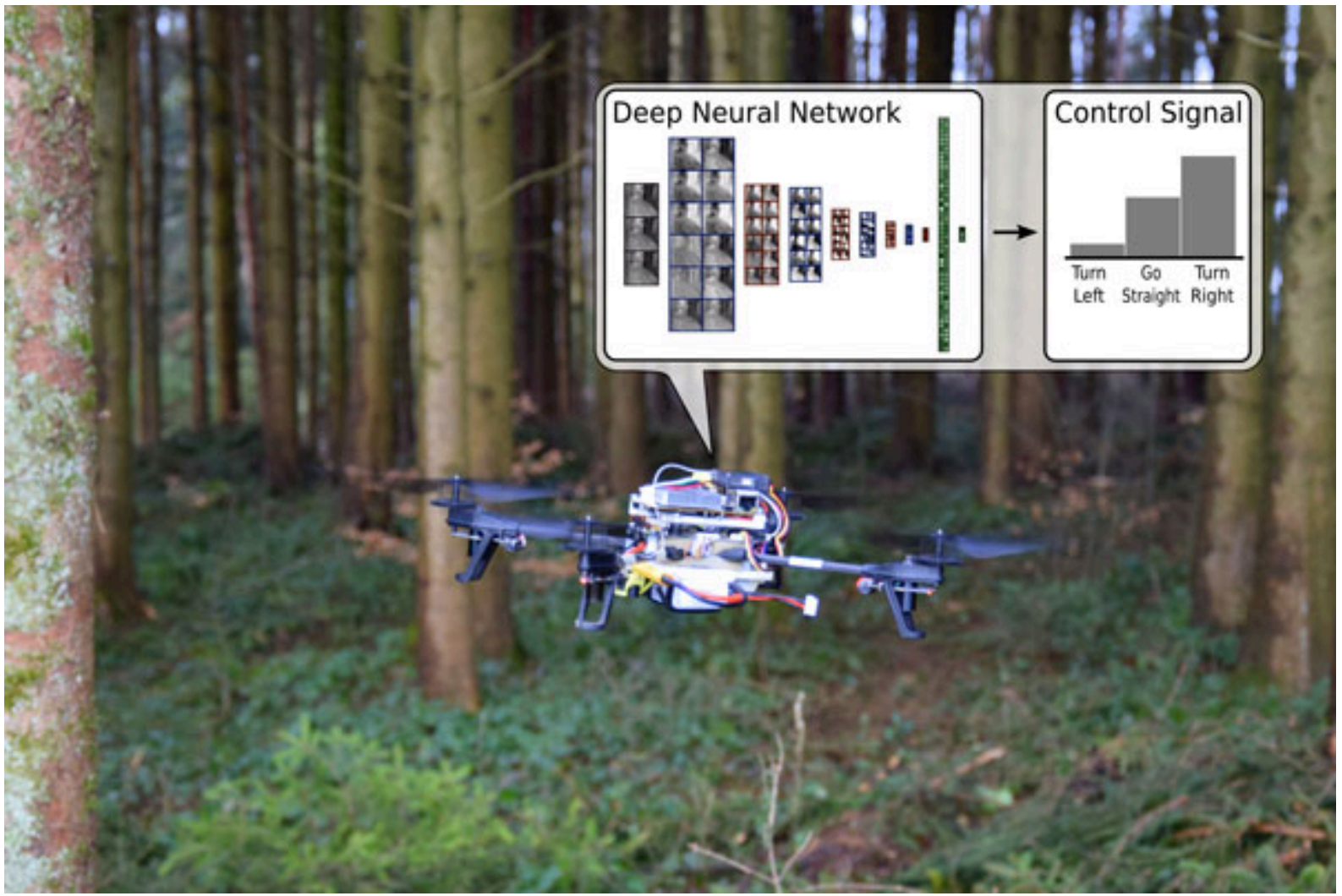


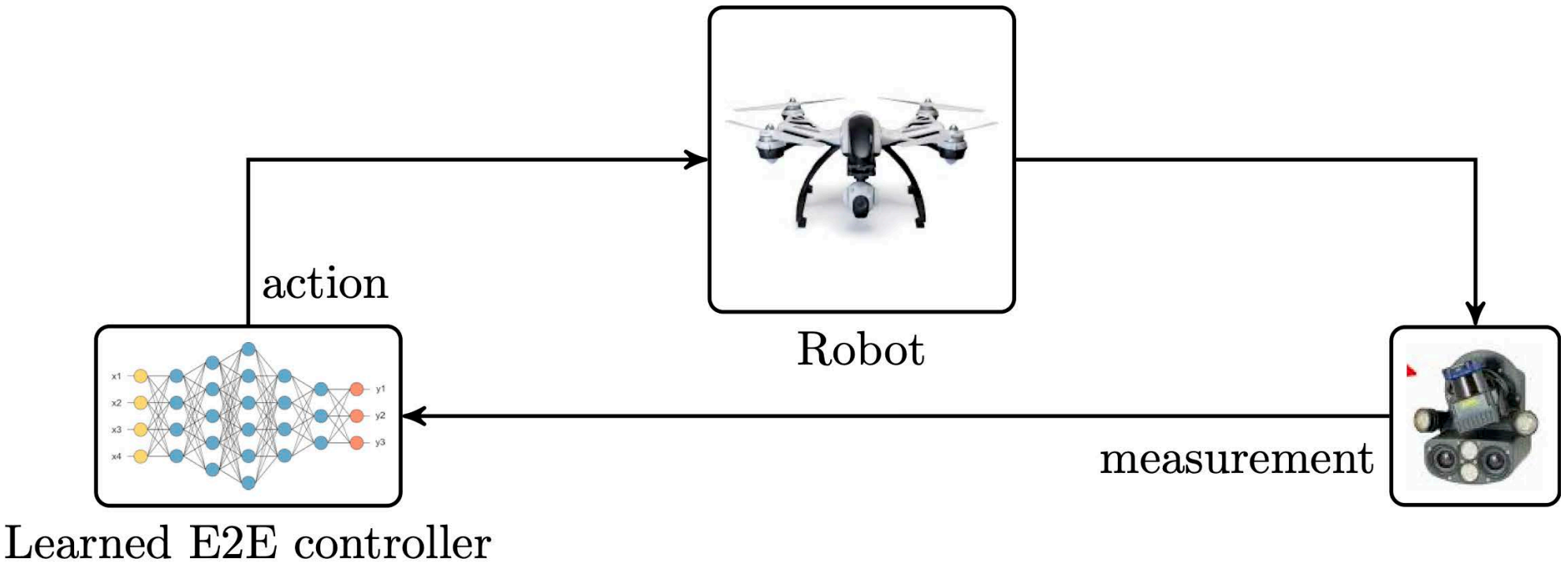
Image credit: Varun Hariprasad

Object-aware Navigation

Navigation using Deep Learning



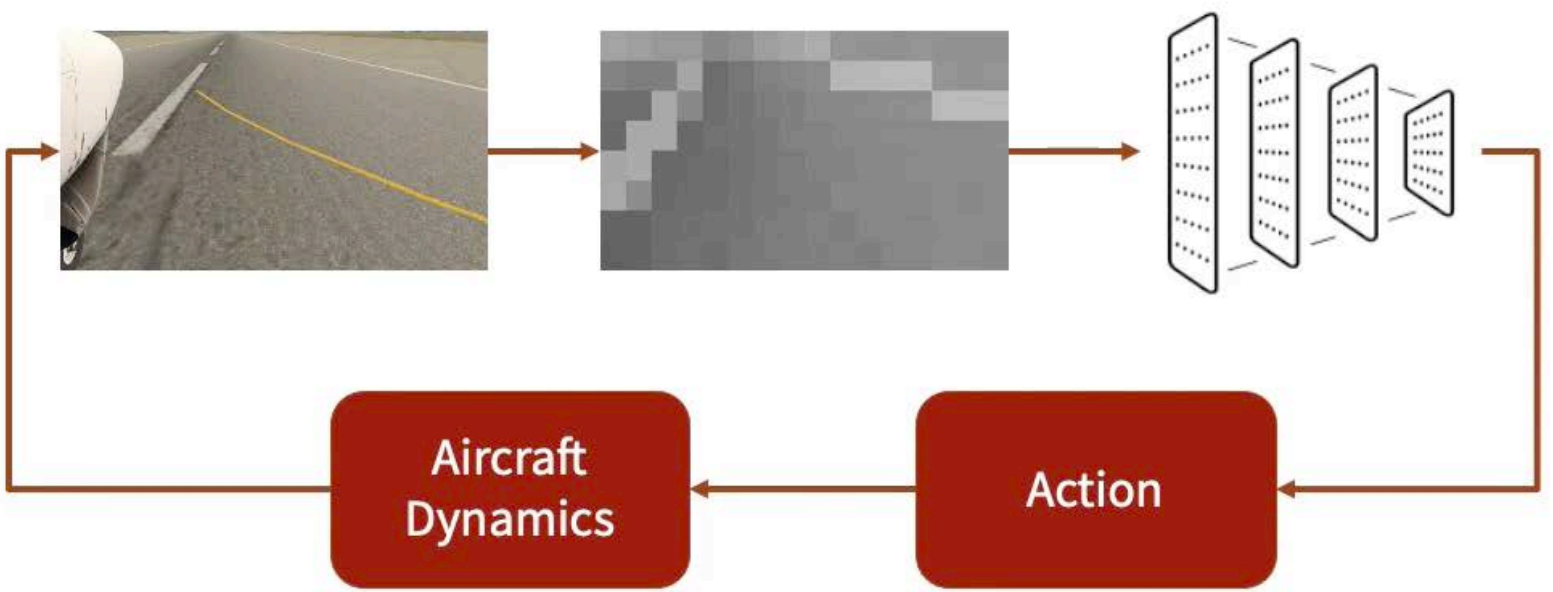
End-to-end Neural Net Control



Example: Aircraft Taxiing

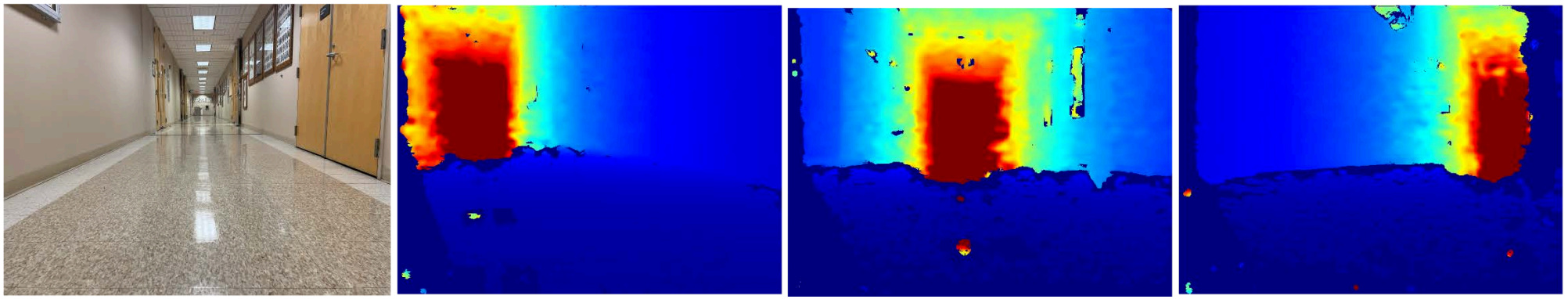
Image-based Neural Network Controllers

Image-based neural network controllers will allow for low-cost and flexible control systems in a variety of applications.



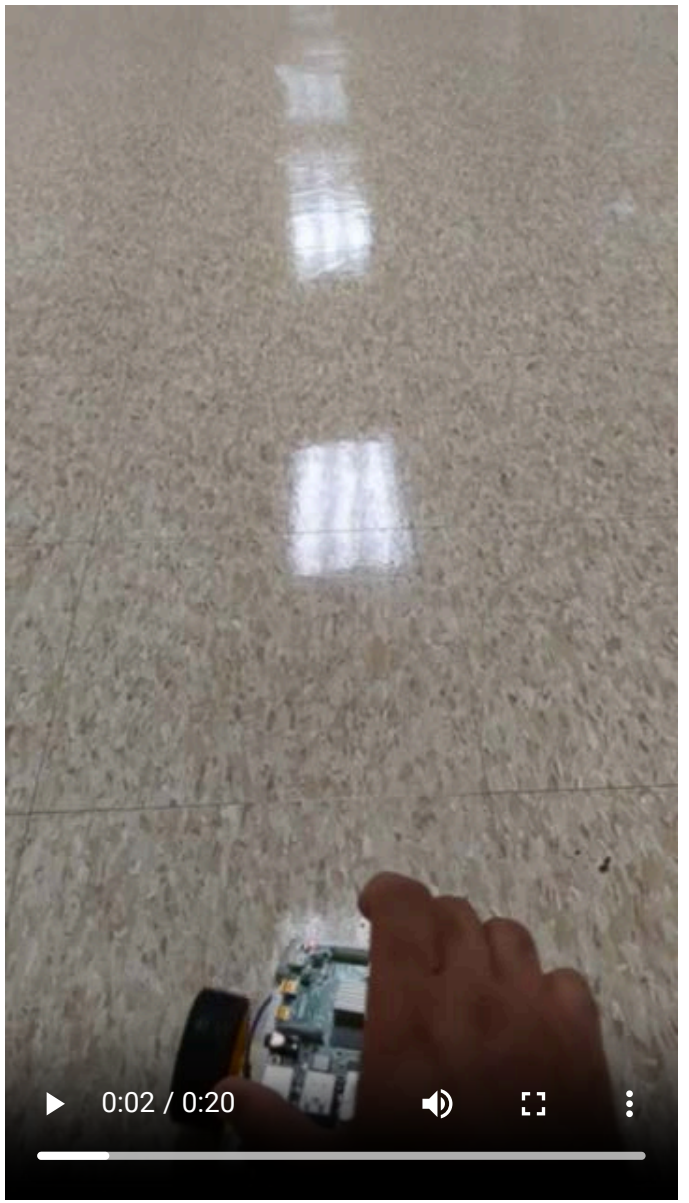
Source: slides by Sydney Katz

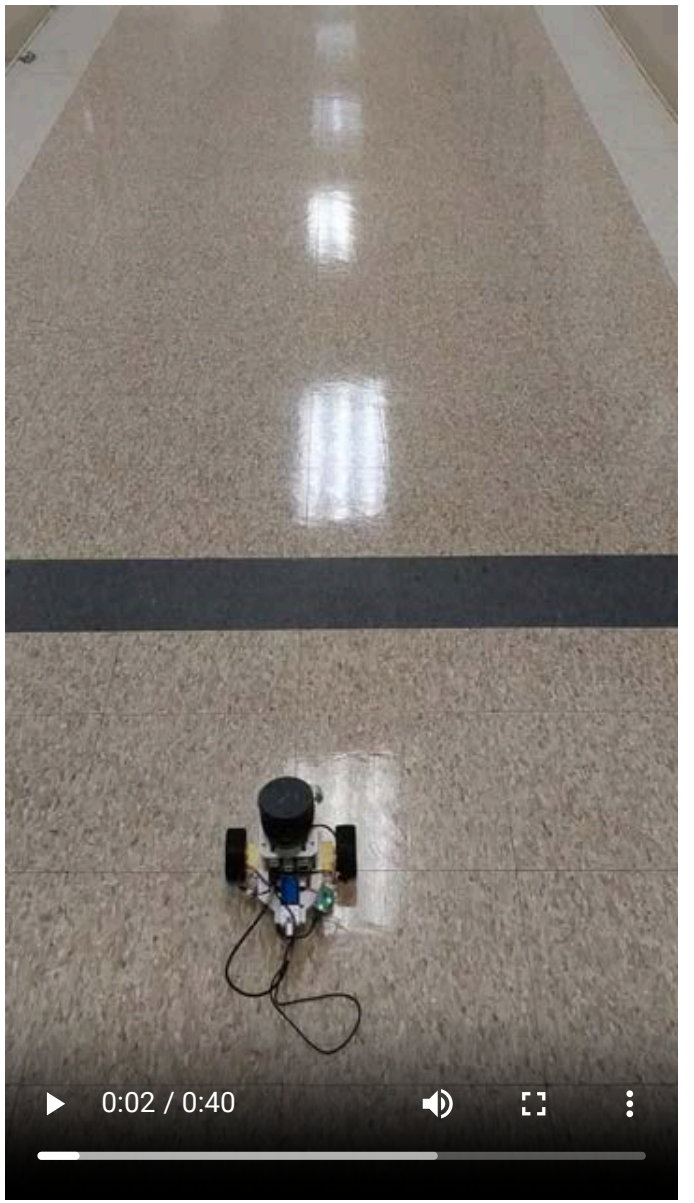
Navigation at UK



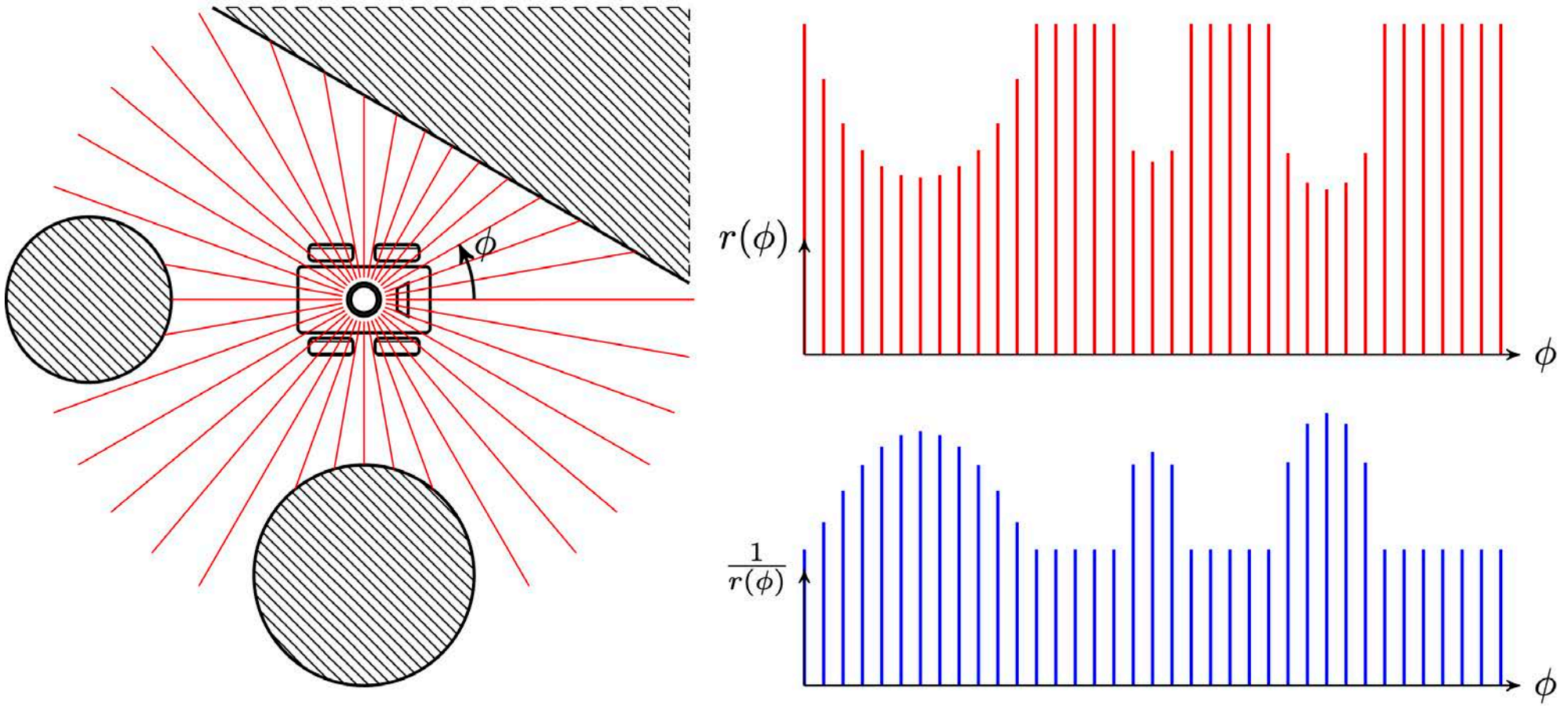
- Simple linear controller < 1 MB size
- Guaranteed to navigate hallway without crashing

Results

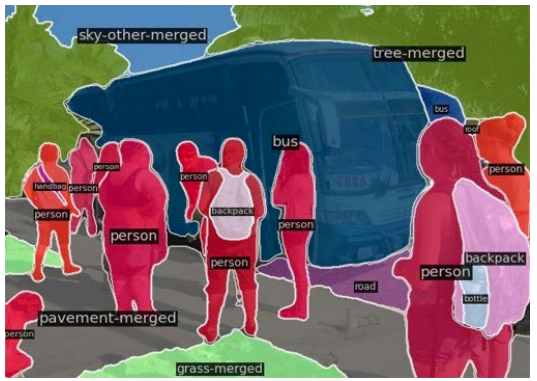




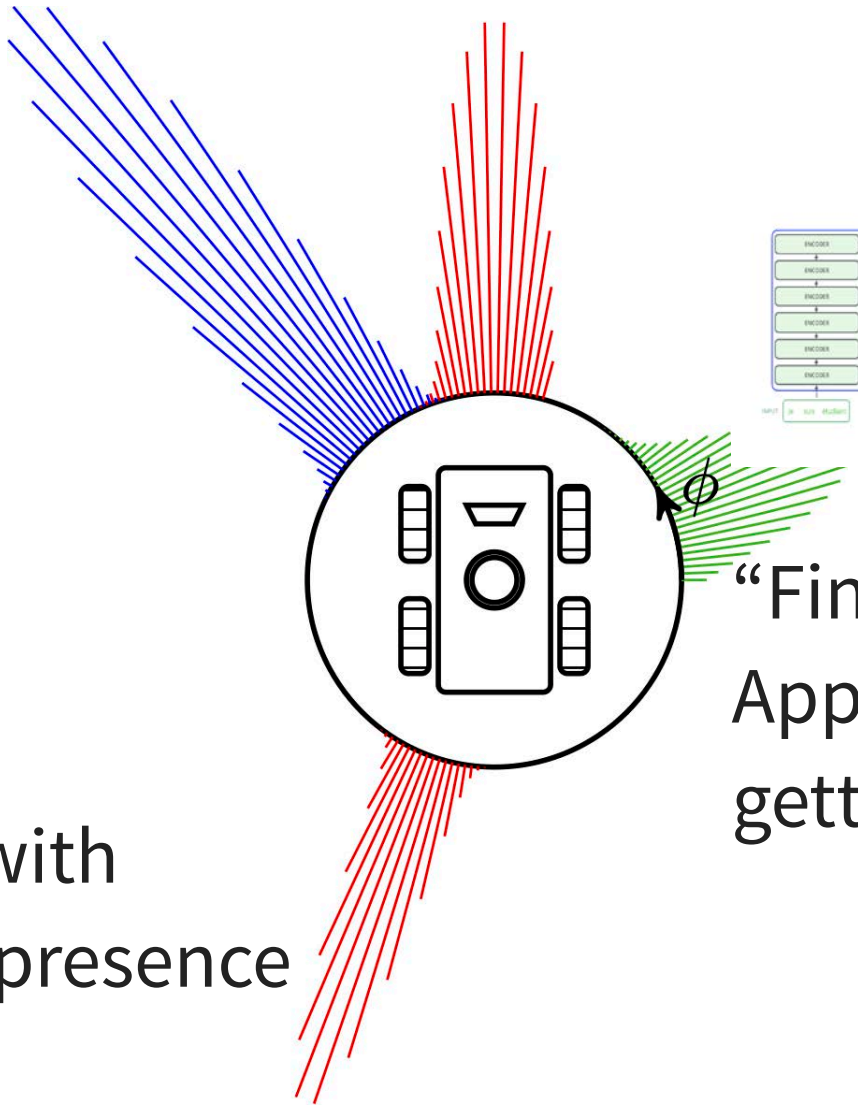
Distance-based Navigation

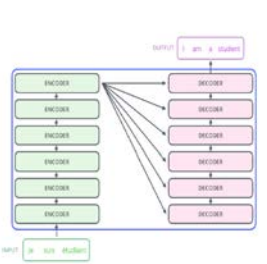


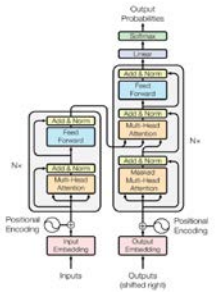
Object-aware Navigation



ϕ



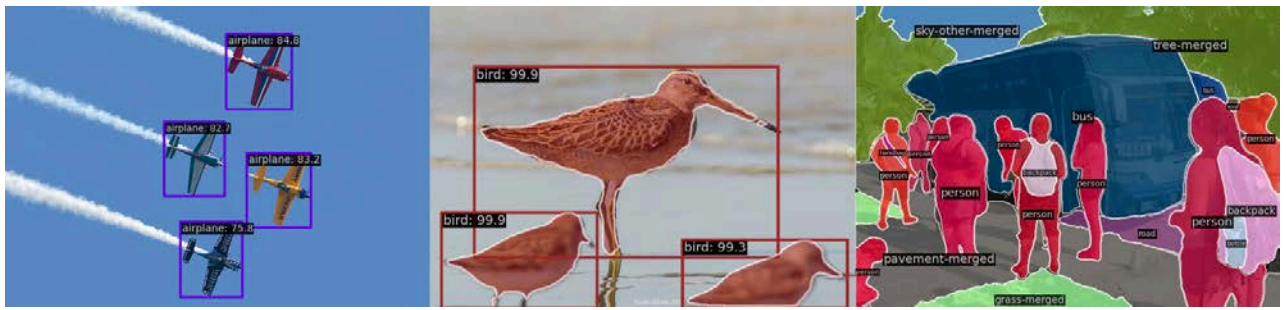




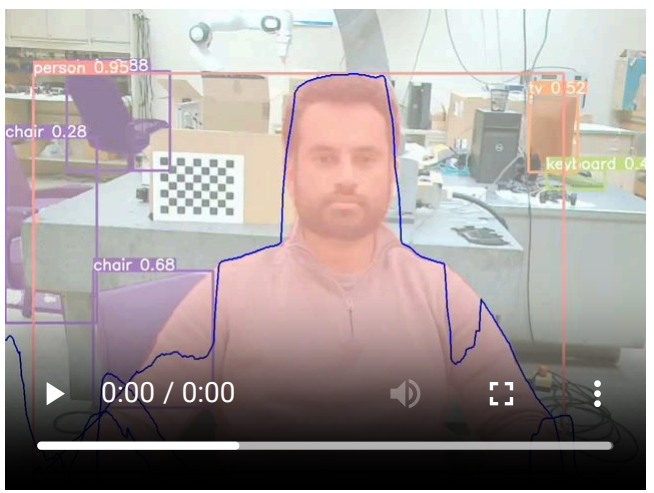
"Find bus below tree. Approach and help people getting on."

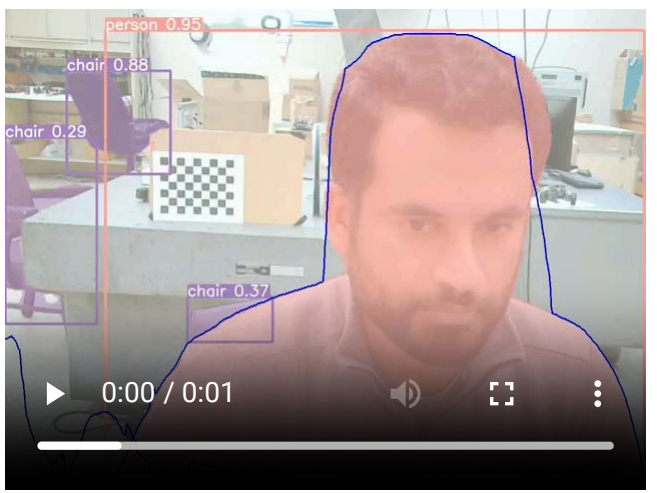
Replace $1/r(\phi)$ with measure of object presence at ϕ

Yolov9 Segmentation Model



Source: https://img.freepik.com/premium-vector/vector-realistic-washing-machine-white-3d-mockup_208581-782.jpg?semt=ais_hybrid



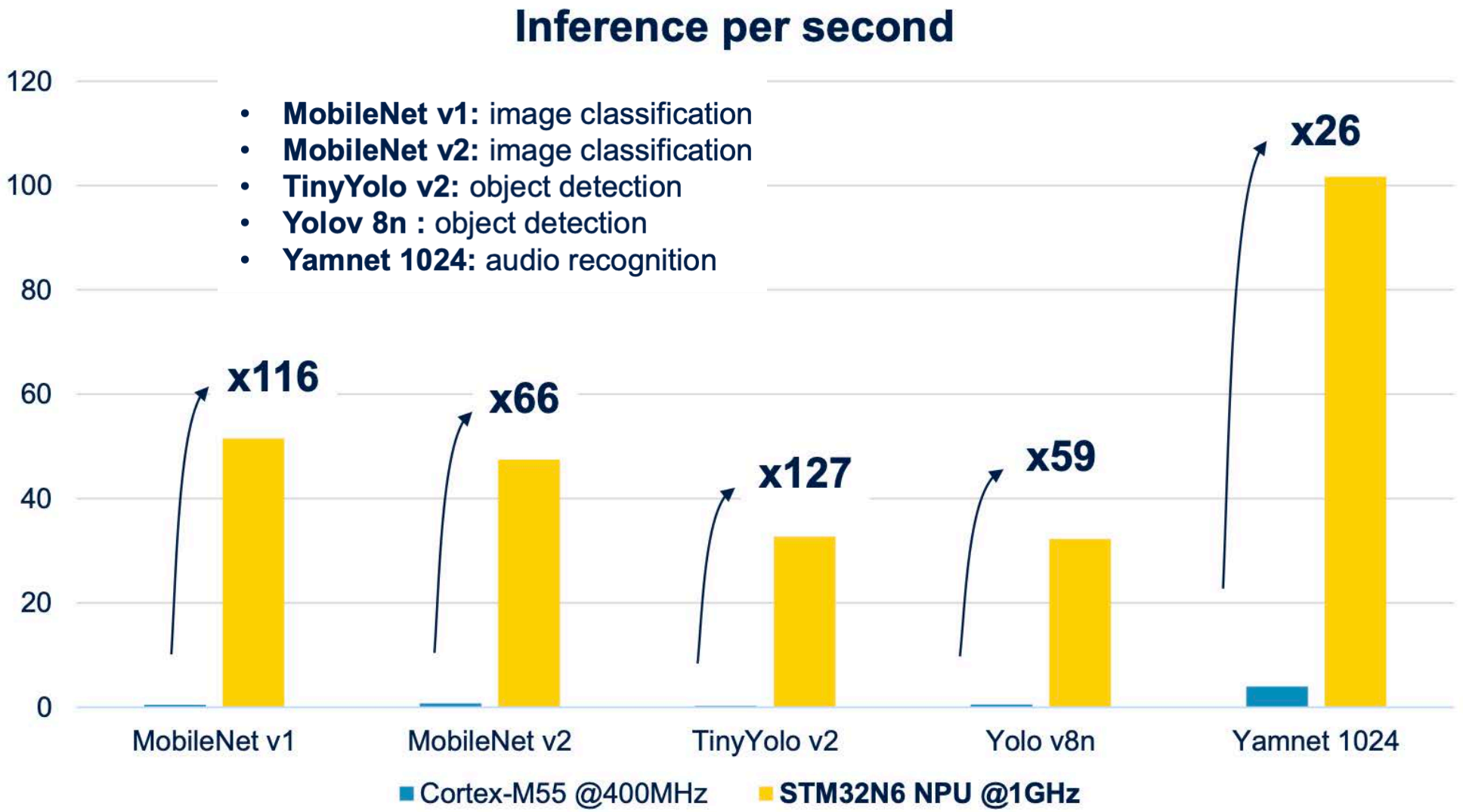


Non-NVIDIA Microprocessors

Examples

	STM32N6	ESP32s3
		
CPU	Arm® Cortex M55 (1)	Xtensa® LX7 (4)
Memory	4.2 MB	2 MB
Cores	300 MAU[1]	-
Cost	$8-$20	$2 - $10
Speed	0.6 TOPS	-
Clock Speed	1 GHz	240 MHz

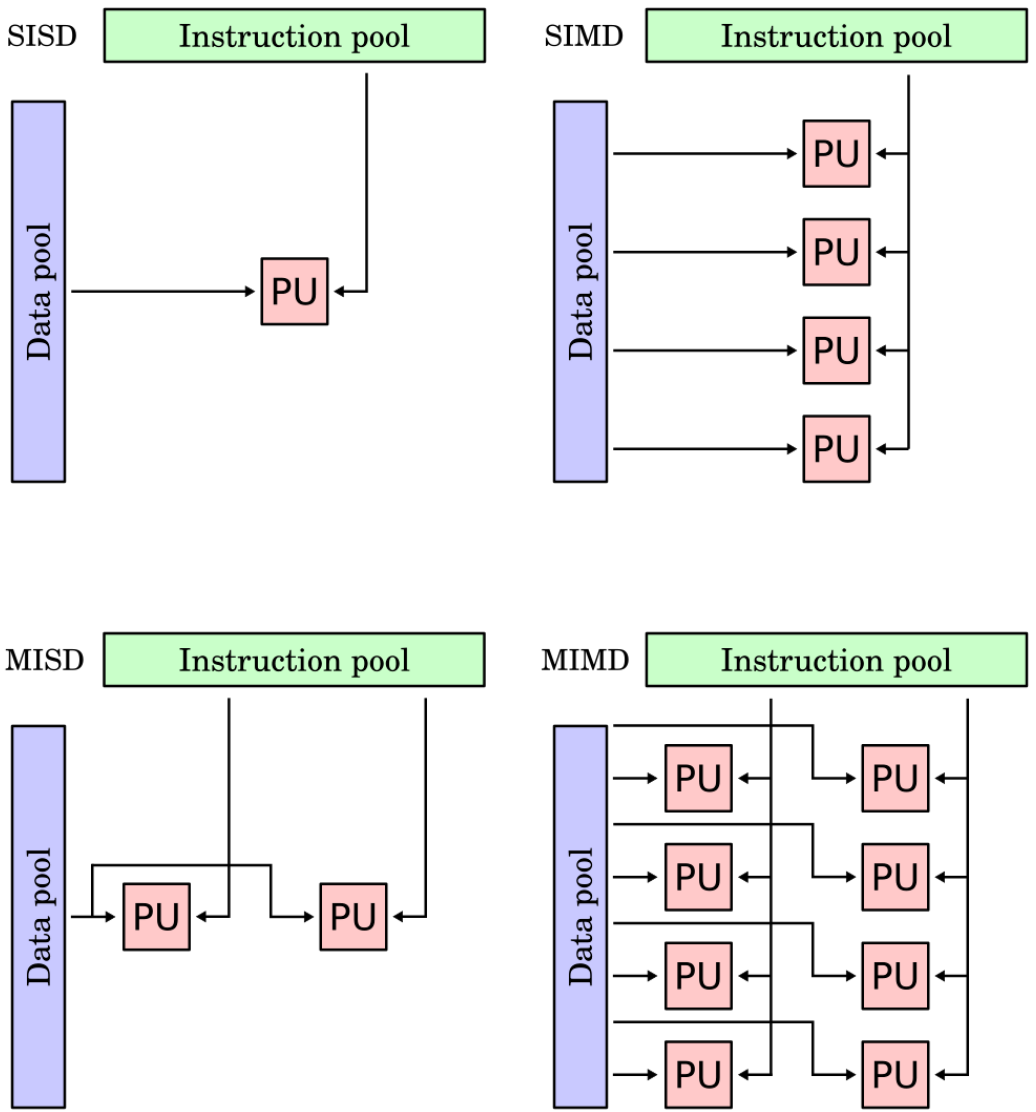
Performance: STM32N6



Performance: ESP32s3

- Custom CNN 500 KB
- Inference in ~60 milliseconds
- 93% Accuracy

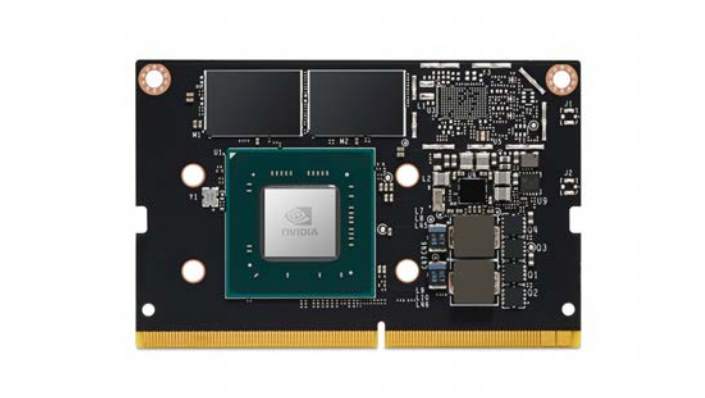
Single Instruction Multiple Data

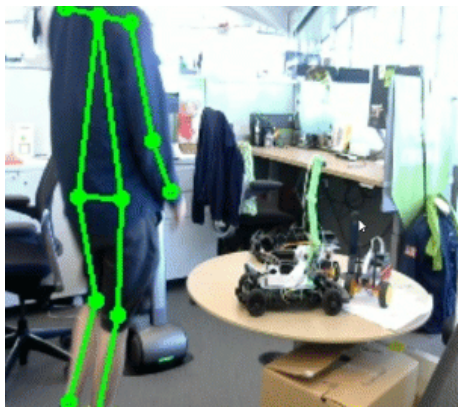


Acknowledgements

- David Yackzan (now at PikNik Robotics)
- Pouya Samanipour (Ph.D. Student)
- Varun Hariprasad (now at MIT)
- Katherine Zhang (PLD High)
- Karthika Hariprasad (PLD High)

Thank You!





Type	Duration (minutes)
Total	480
Seated	360
Standing	5
Not visible	115