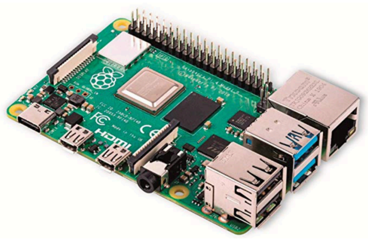
Efficient Deep Learning:
—— from 2D to 3D

Song Han (韩松)
Assistant Professor
Massachusetts Institute of Technology
songhan@mit.edu

We need AI on Edge Devices



Low Latency

Low Cost

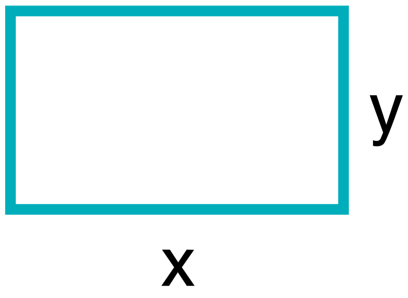
User Privacy

However, edge device has low computation power

Efficient Deep Learning on the Edge

- **Efficient 3D Algorithms:**
  - PVCNN for efficient point-cloud recognition [NeurIPS'19, spotlight]
  - TSM for efficient video recognition [ICCV'19]
- **Compression / NAS**
  - Deep Compression [NIPS'15, ICLR'16]
  - ProxylessNAS, AMC, HAQ [ICLR'19, ECCV'18, CVPR'19, oral]
  - Once-For-All (OFA) Network

From 2D to 3D Deep Learning

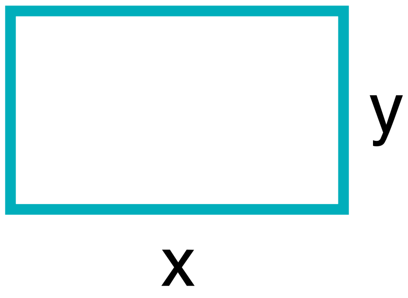


Image

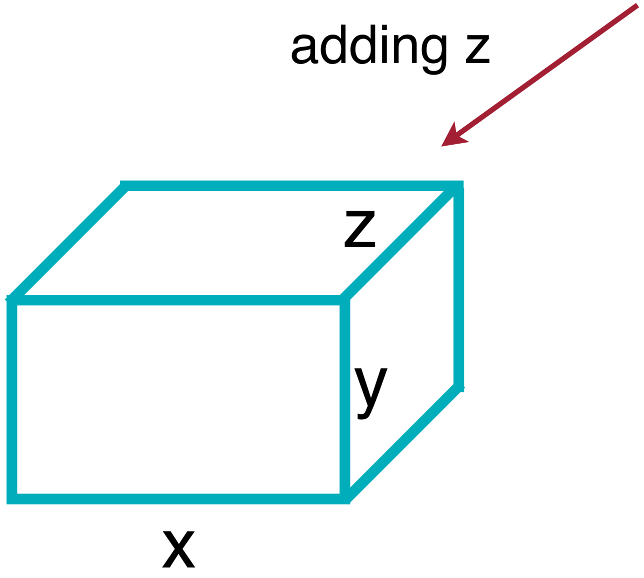
Training:
ImageNet: 1.2M images

Inference:
ResNet-50: 8G FLOPs

From 2D to 3D Deep Learning



Image

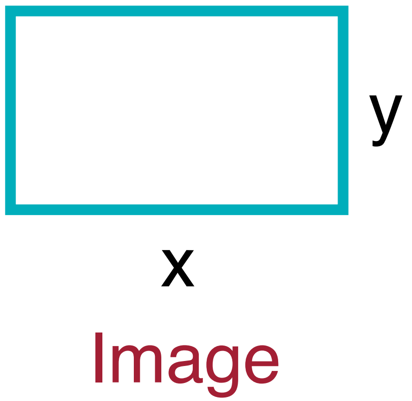


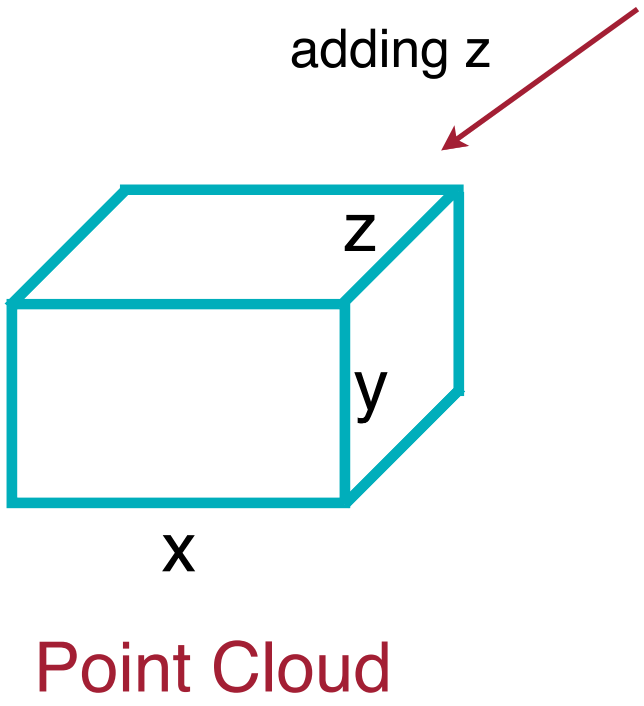
Point Cloud

Training:
ImageNet: 1.2M images

Inference:
ResNet-50: 8G FLOPs

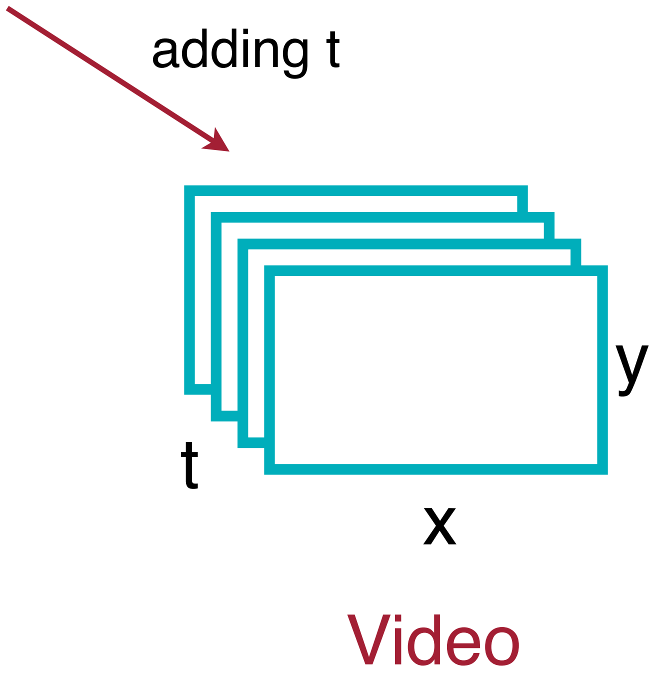
From 2D to 3D Deep Learning



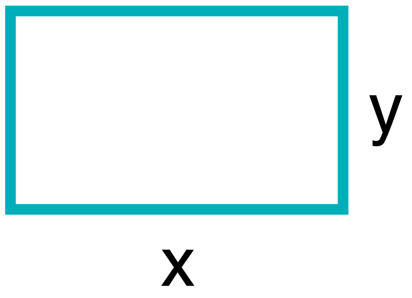


Training:
ImageNet: 1.2M images

Inference:
ResNet-50: 8G FLOPs



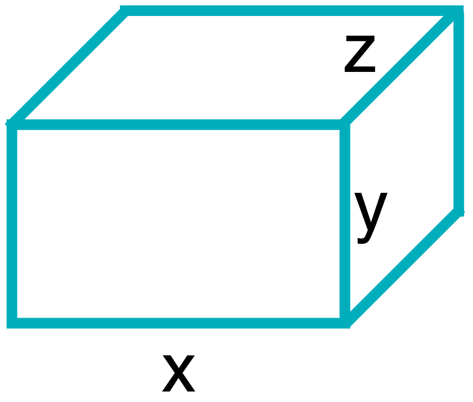
From 2D to 3D Deep Learning



Image

adding z

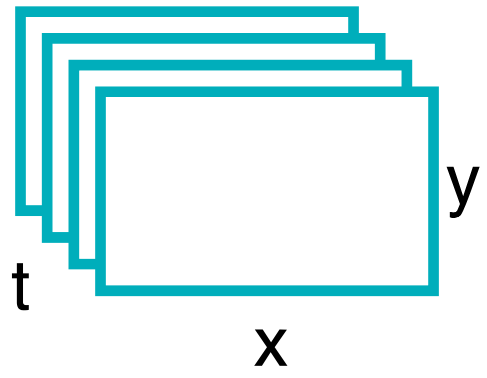
adding t



Point Cloud

Training:
ImageNet: 1.2M images
Kinetics: 63M frames, 50x

Inference:
ResNet-50: 8G FLOPs
ResNet-50-I3D: 65G FLOPs, 8x



Video

Efficient Deep Learning on the Edge

✦ **Efficient 3D Algorithms:**

- PVCNN for efficient point-cloud recognition [NeurIPS'19, spotlight]
- TSM for efficient video recognition [ICCV'19]

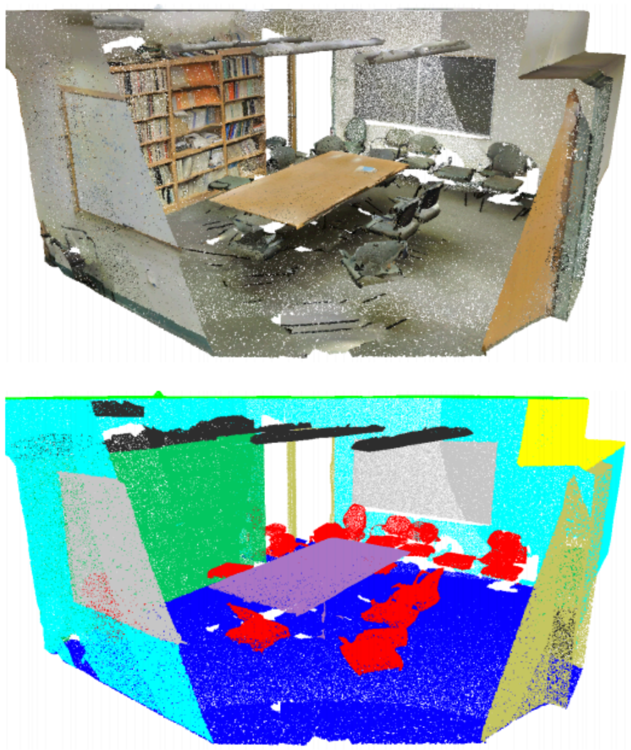
✦ **Compression / NAS**

- Deep Compression [NIPS'15, ICLR'16]
- ProxylessNAS, AMC, HAQ [ICLR'19, ECCV'18, CVPR'19, oral]
- Once-For-All (OFA) Network

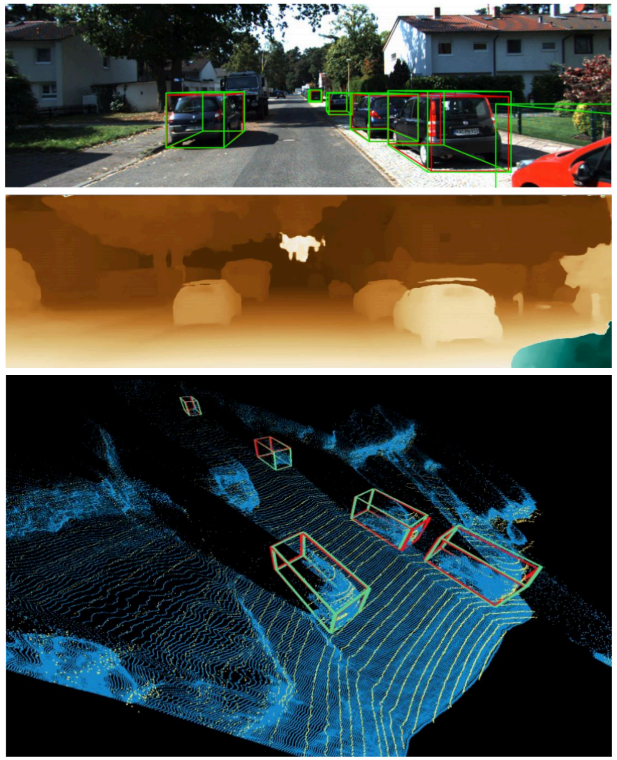
[PVCNN, NeurIPS'19]

3D Deep Learning

3D Part Segmentation
(for Robotic Systems)



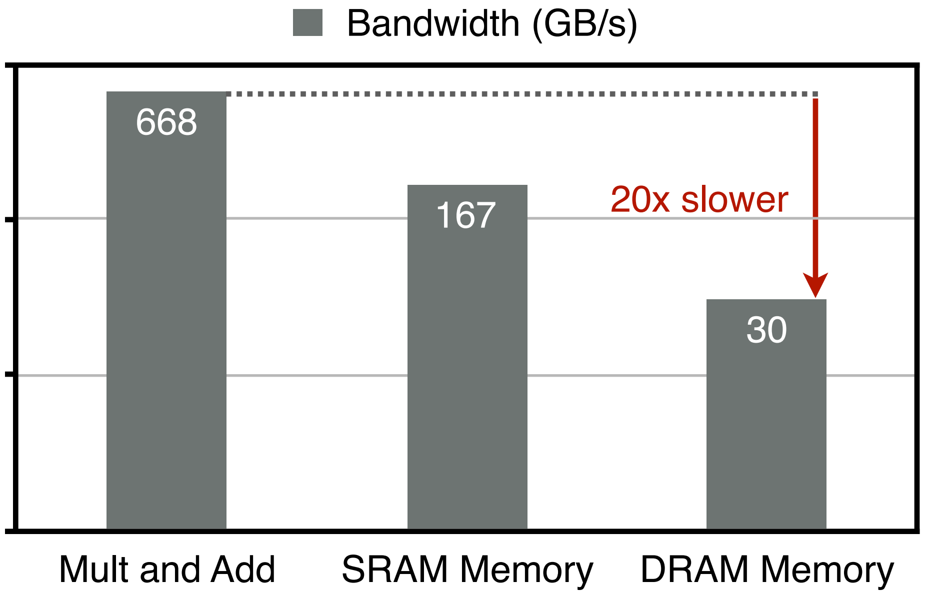
3D Semantic Segmentation
(for VR/AR Headsets)



3D Object Detection
(for Self-Driving Cars)

[PVCNN, NeurIPS'19]

Efficient 3D Deep Learning: Hardware Bottleneck



Sequential Memory Access

1	2	3	4	5	6	7	8

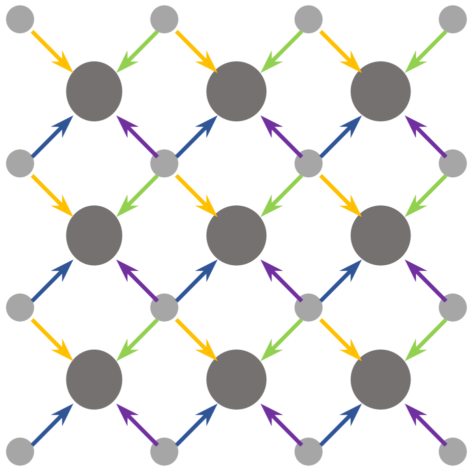
Random Memory Access

7	5	2	4	6	1	8	3

Off-chip DRAM access is much more expensive than arithmetic operation!

Random memory access is inefficient due to the potential bank conflicts!

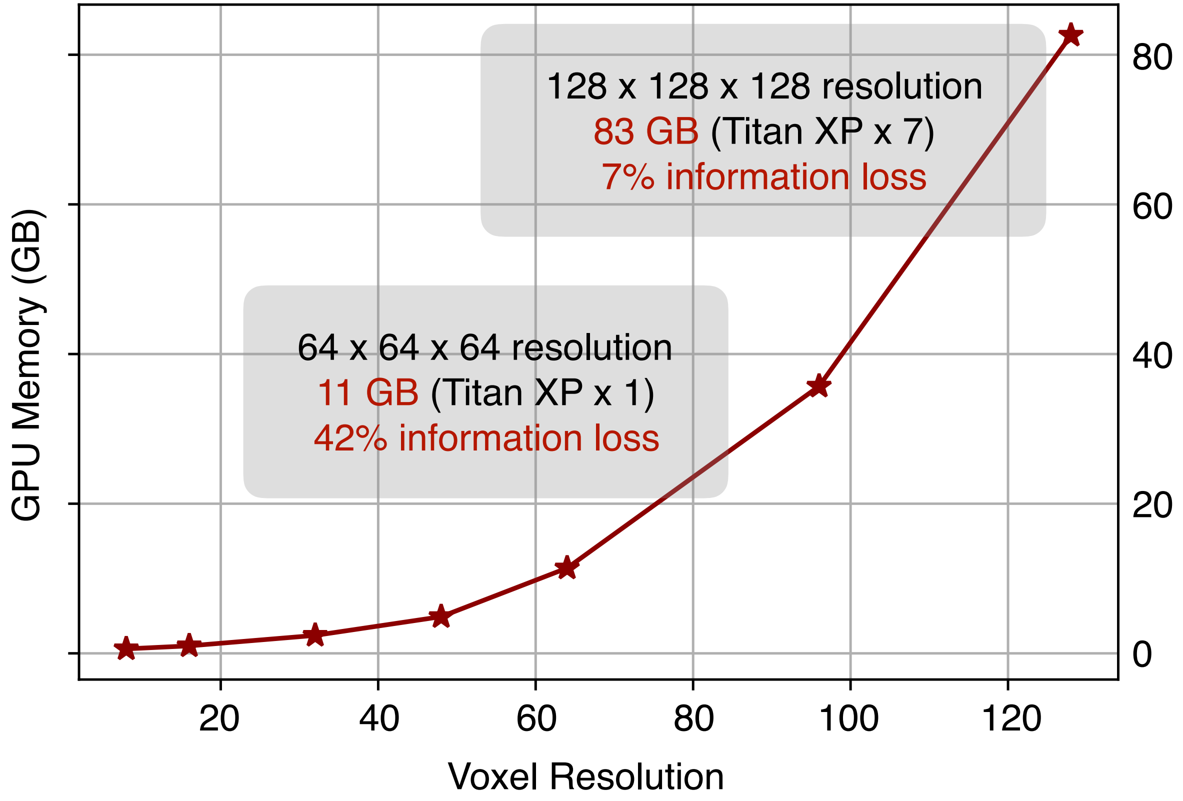
Voxel-Based Models: Cubically-Growing Memory



3D ShapeNets [CVPR'15]

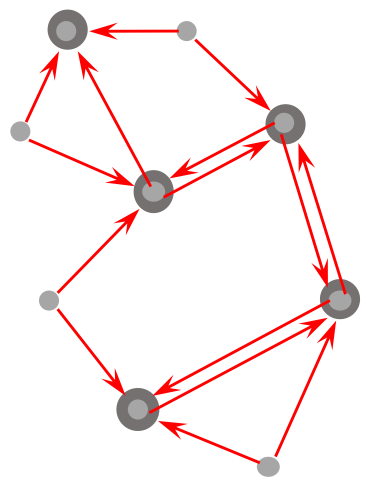
VoxNet [IROS'15]

3D U-Net [MICCAI'16]



[PVCNN, NeurIPS'19]

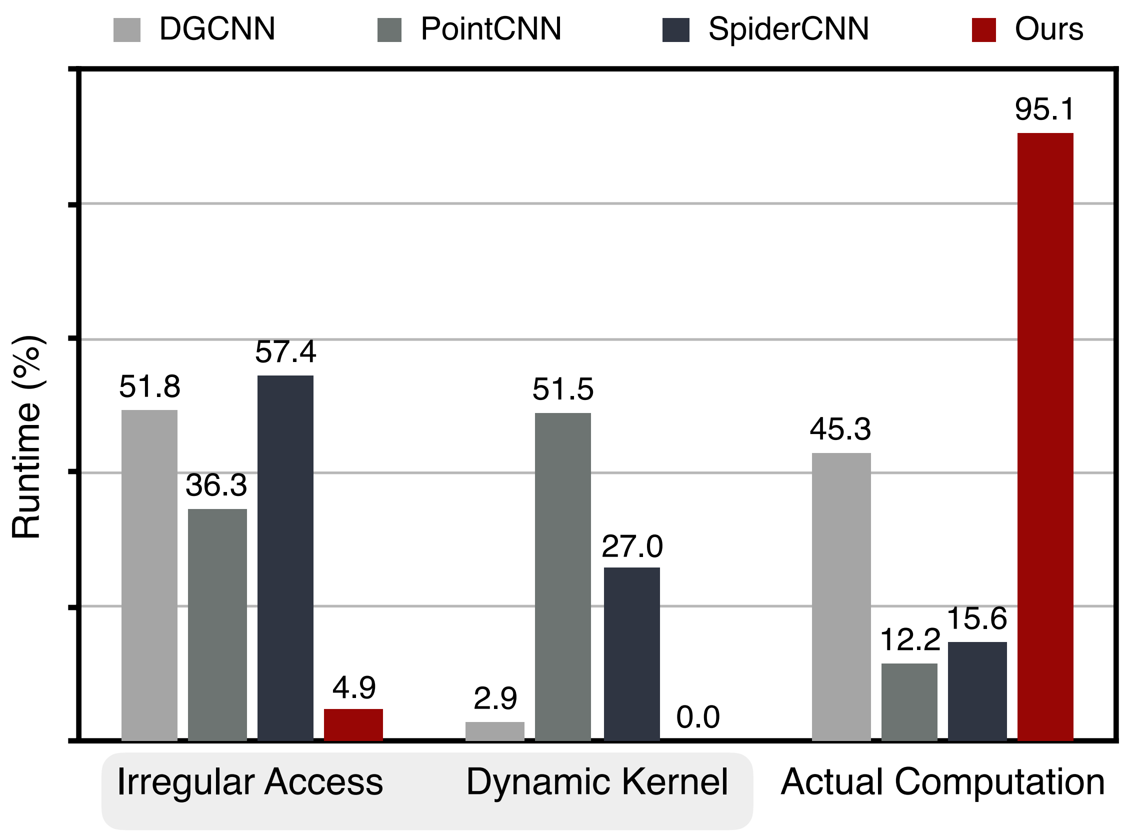
Point-Based Models: Sparsity Overheads



PointNet [CVPR'17]

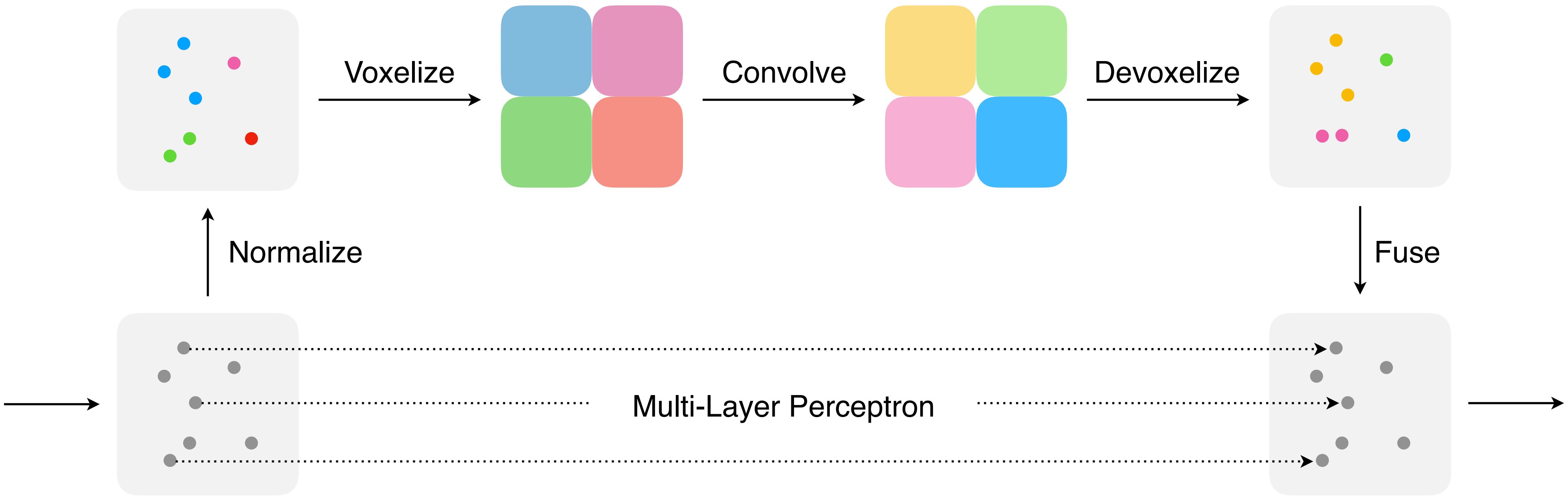
PointCNN [NeurIPS'18]

DGCNN [SIGGRAPH'19]



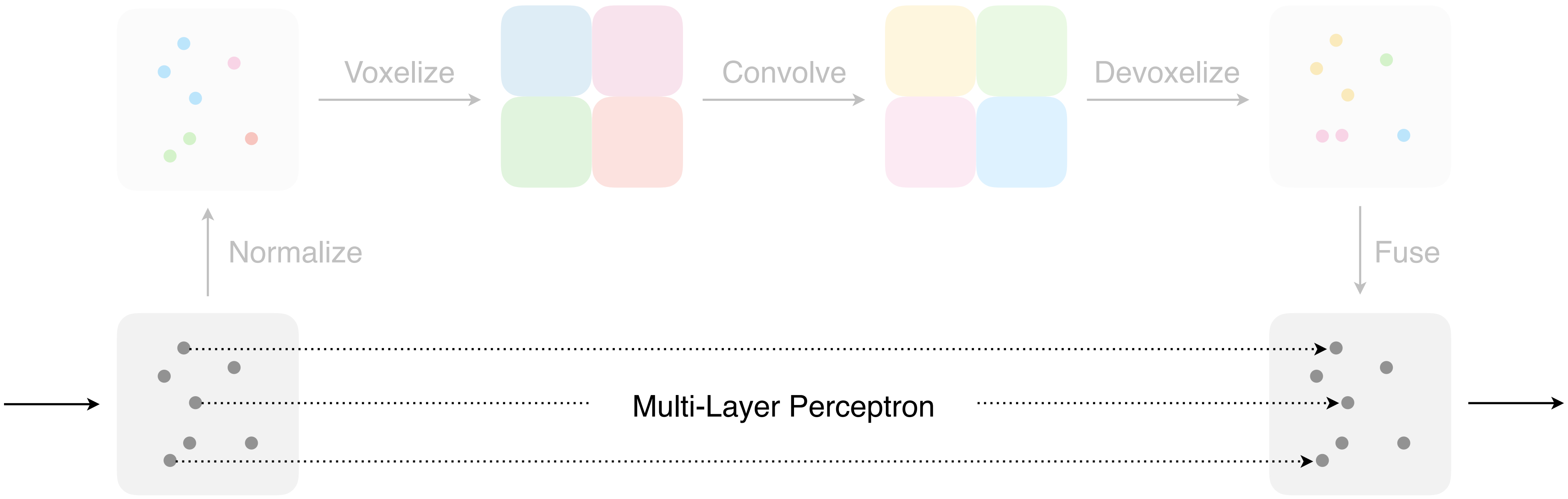
[PVCNN, NeurIPS'19]

Point-Voxel Convolution (PVConv)



[PVCNN, NeurIPS'19]

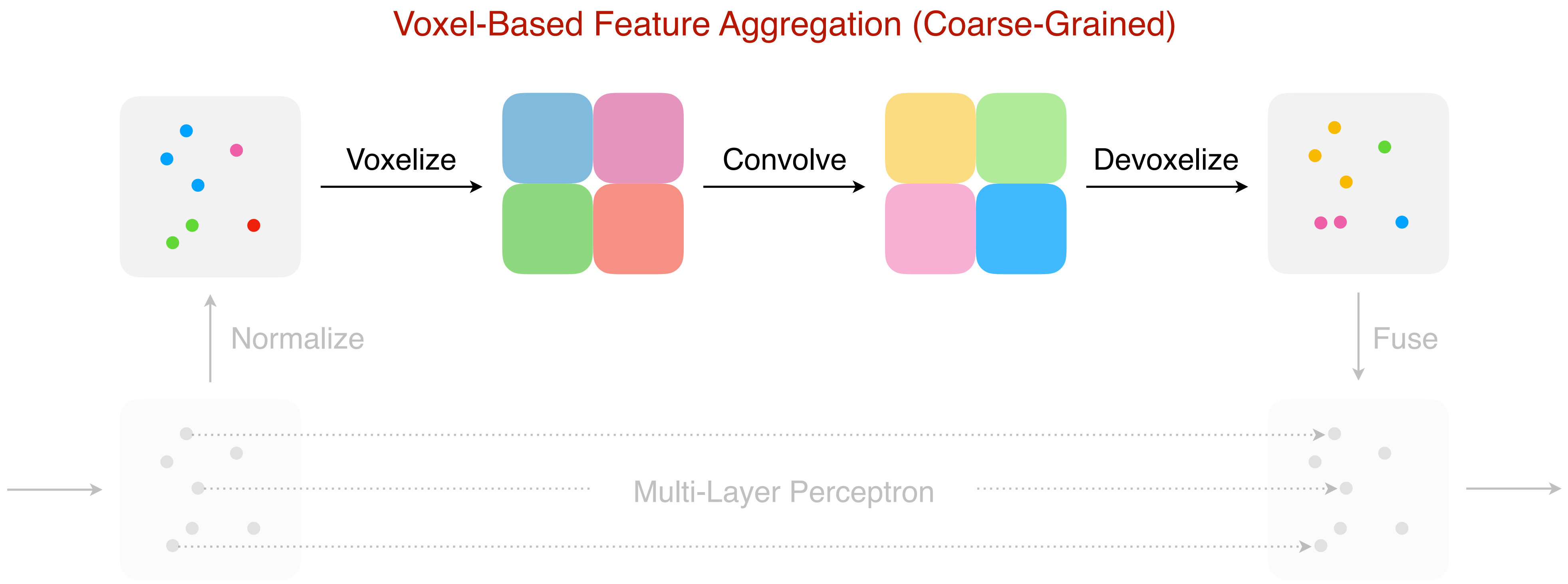
Point-Voxel Convolution (PVConv)



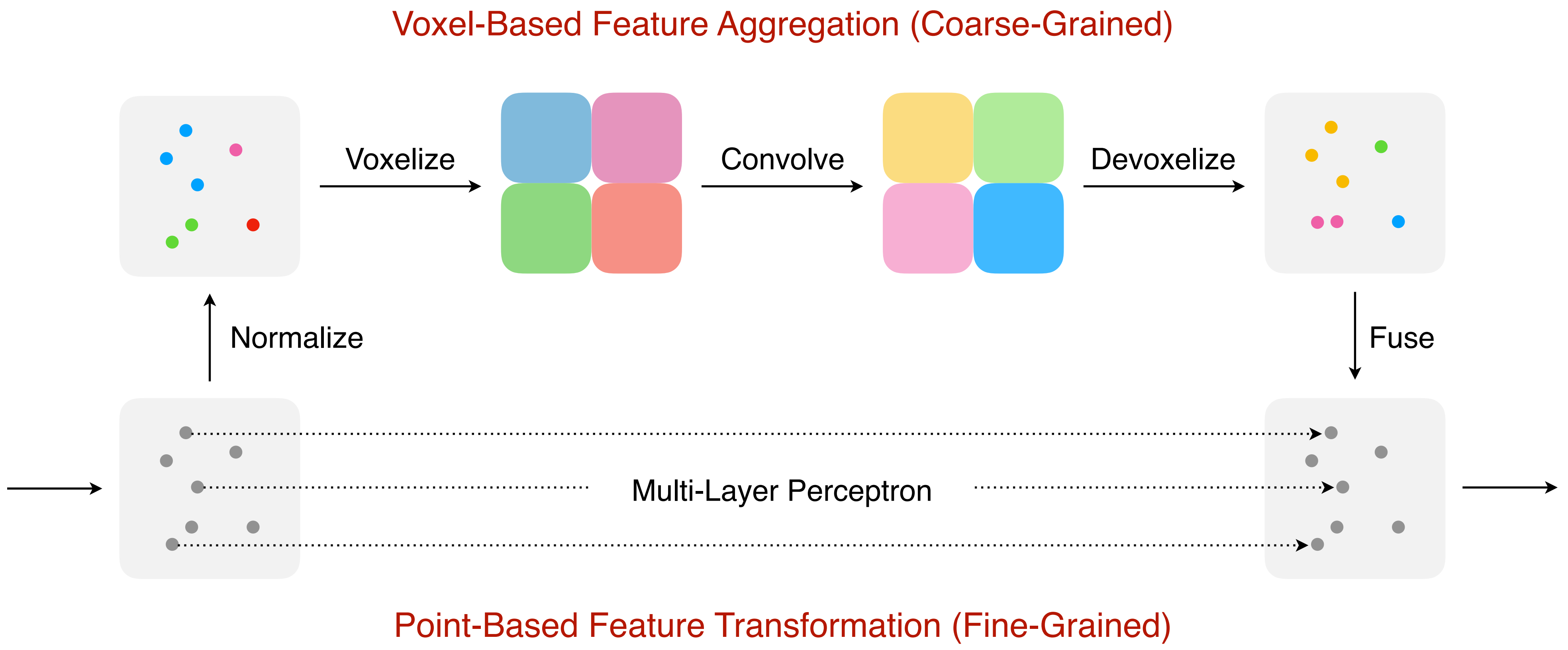
Point-Based Feature Transformation (Fine-Grained)

[PVCNN, NeurIPS'19]

Point-Voxel Convolution (PVConv)



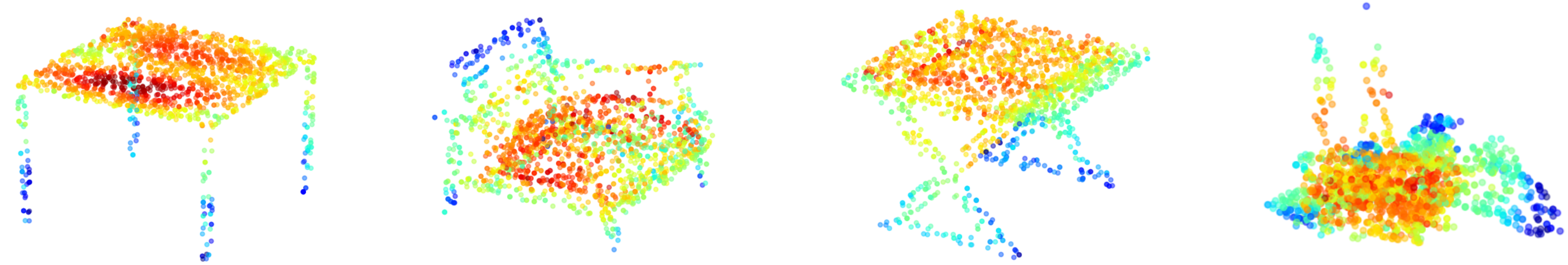
Point-Voxel Convolution (PVConv)



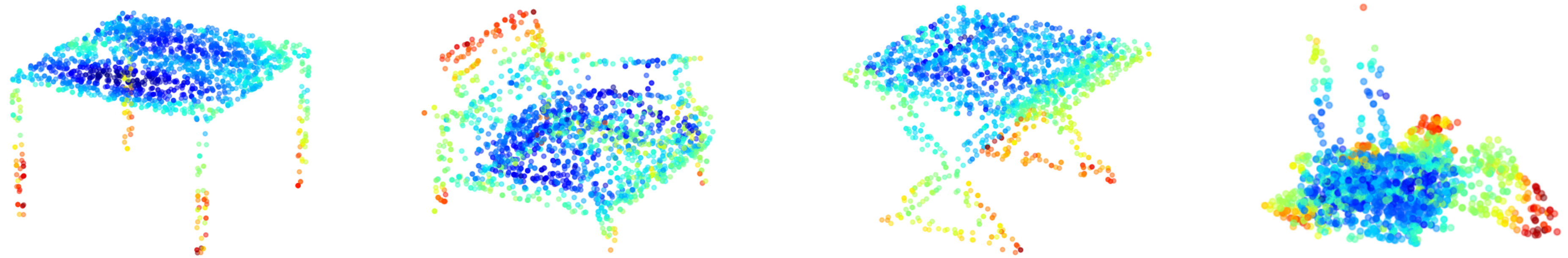
[PVCNN, NeurIPS'19]

Point-Voxel Convolution (PVConv)

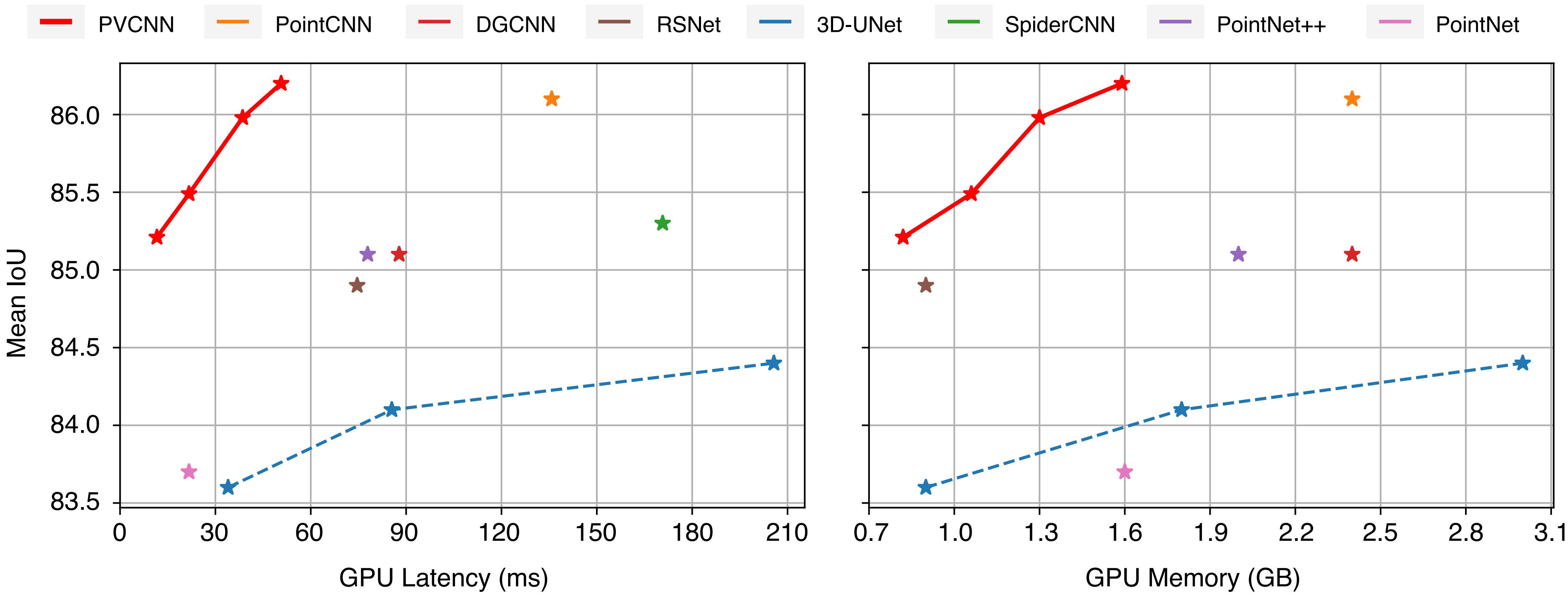
Features from Voxel-Based Branch:



Features from Point-Based Branch:

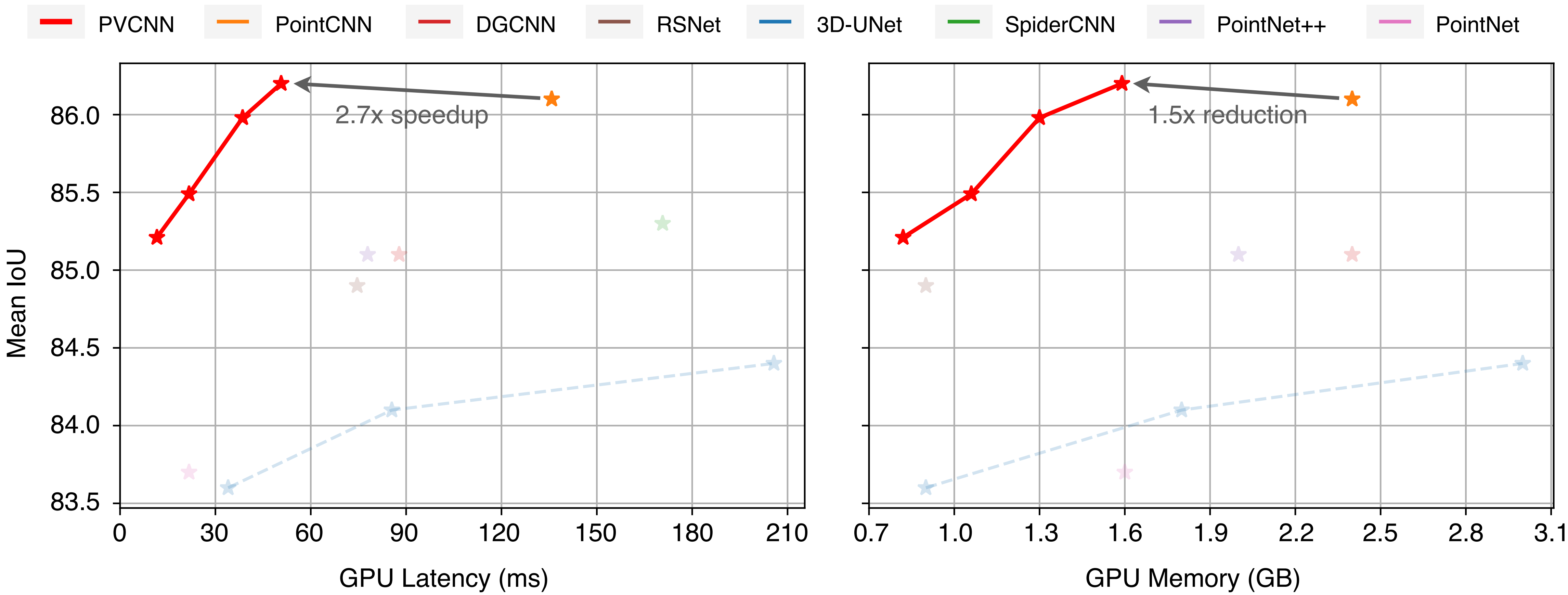


Results: 3D Part Segmentation (ShapeNet)



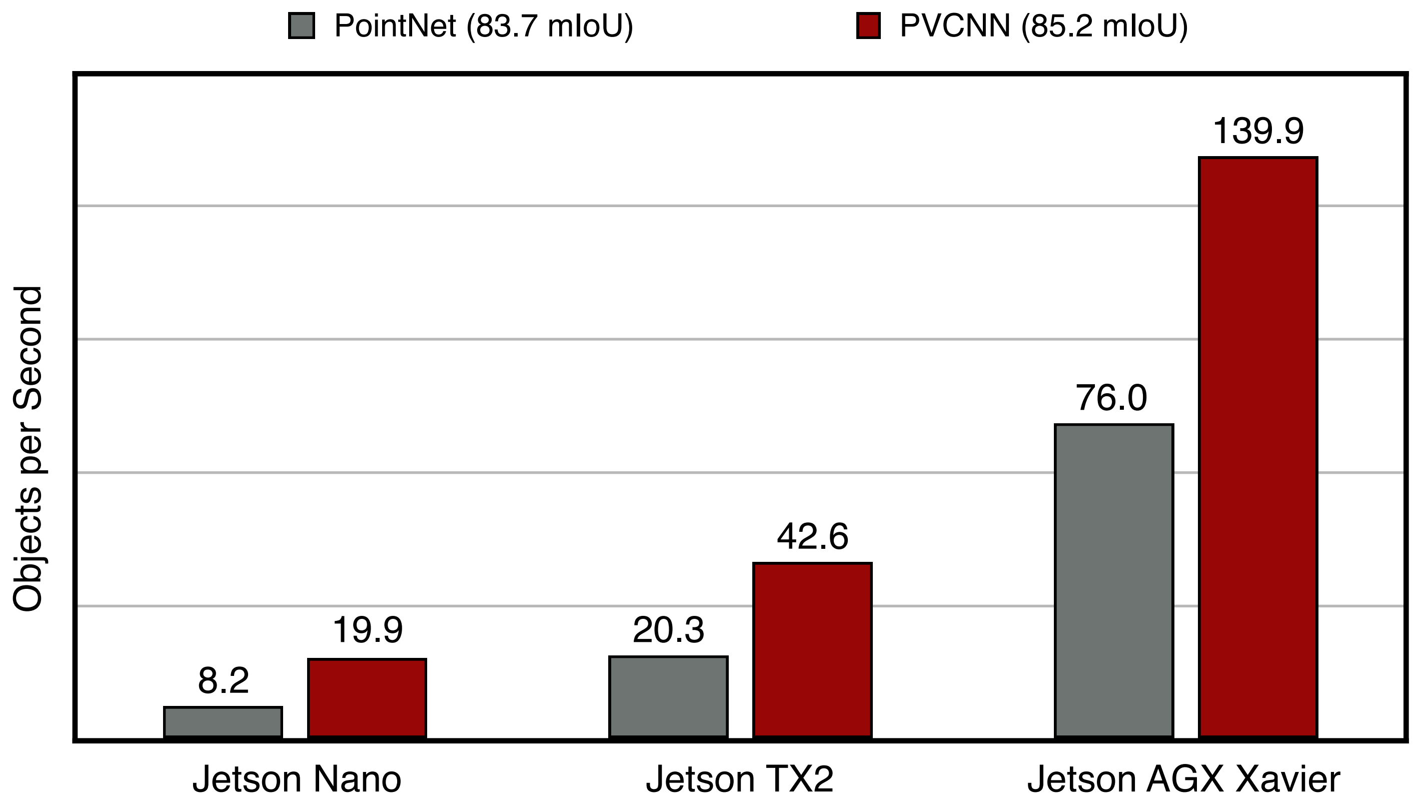
[PVCNN, NeurIPS'19]

Results: 3D Part Segmentation (ShapeNet)



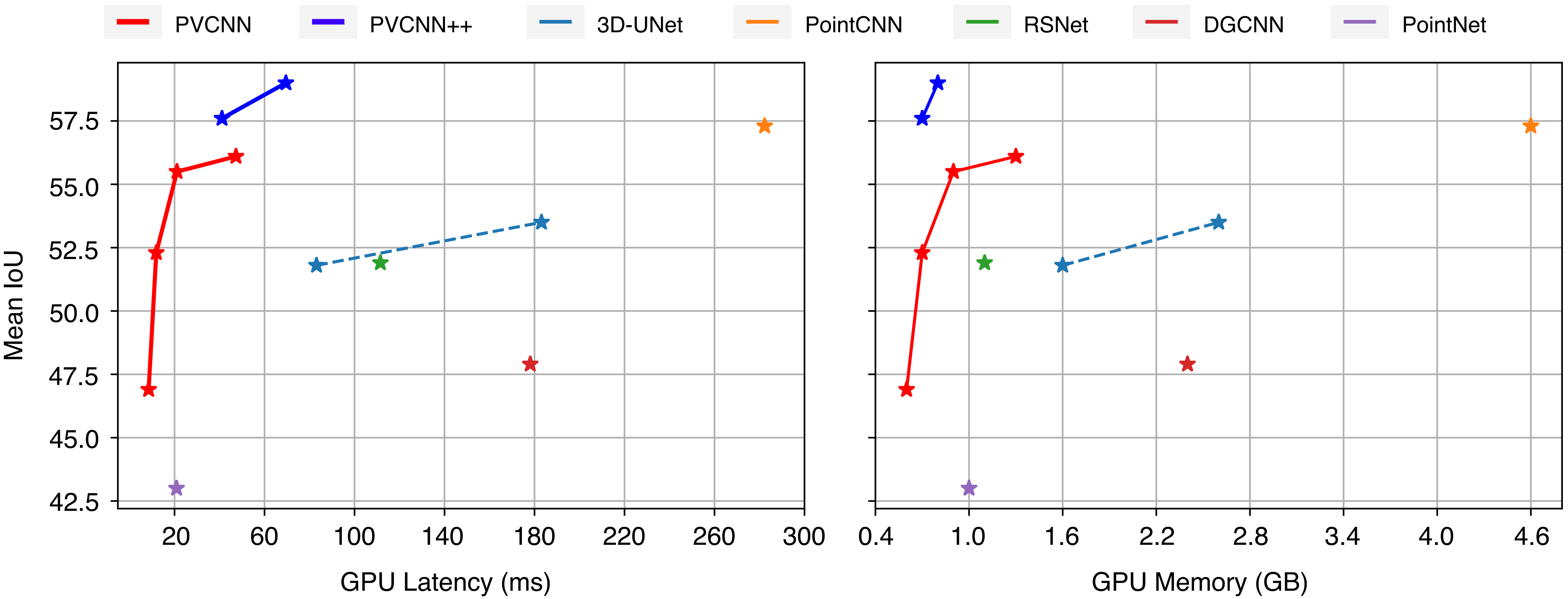
[PVCNN, NeurIPS'19]

Results: 3D Part Segmentation (ShapeNet)

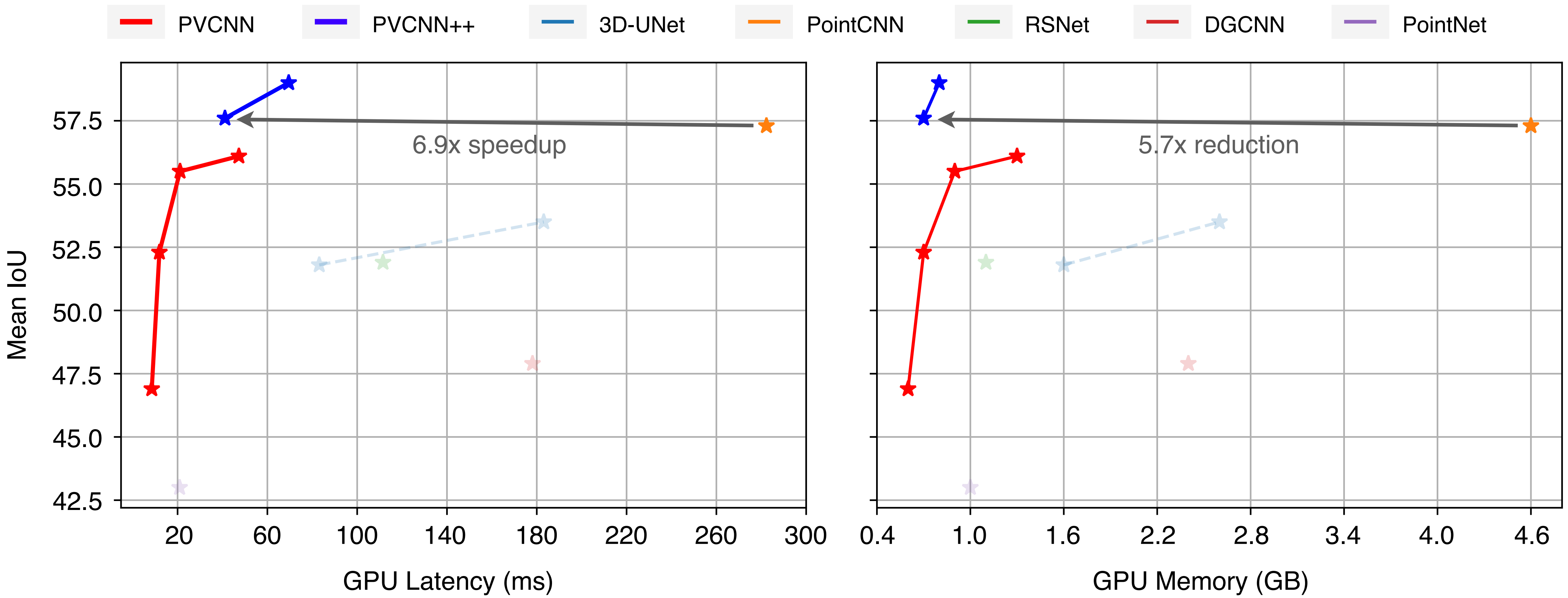


[PVCNN, NeurIPS'19]

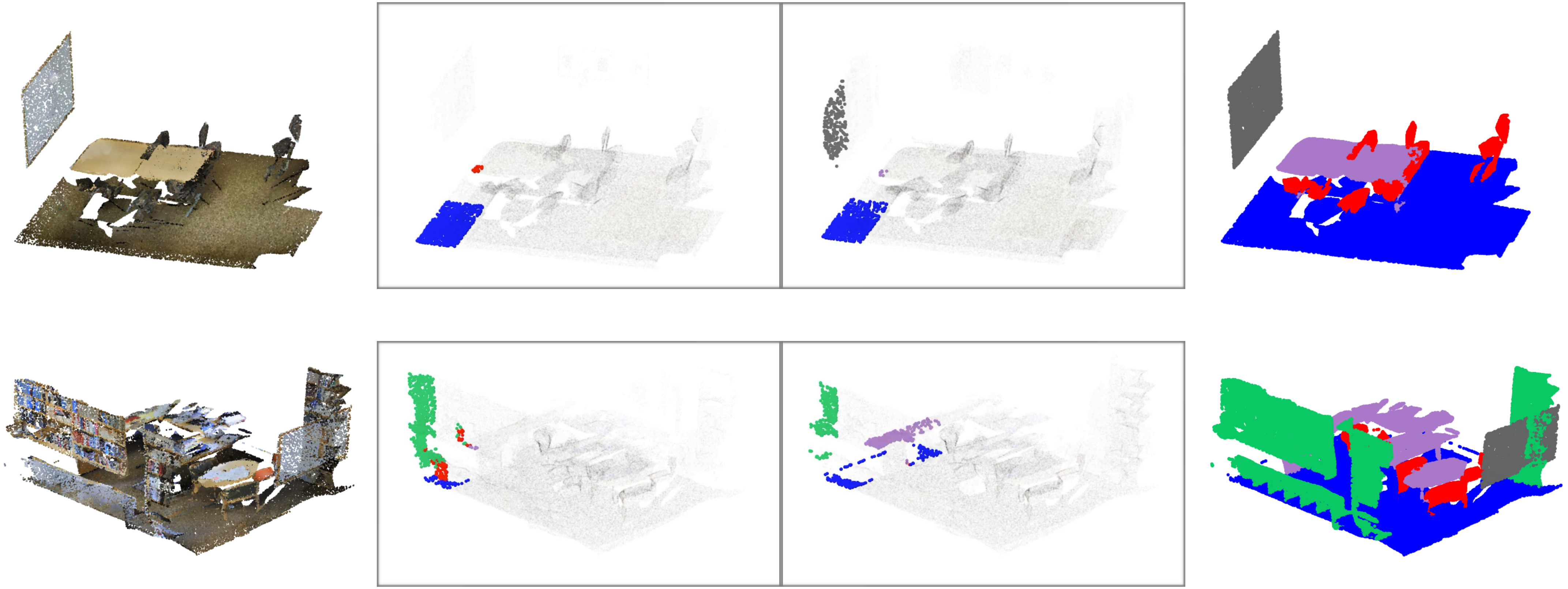
Results: 3D Semantic Segmentation (S3DIS)



Results: 3D Semantic Segmentation (S3DIS)



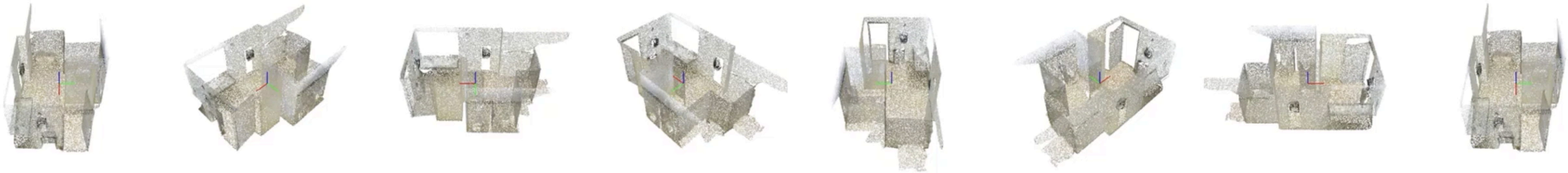
Results: 3D Semantic Segmentation (S3DIS)



Input Scene PointNet PVCNN (1.8x faster) Ground Truth

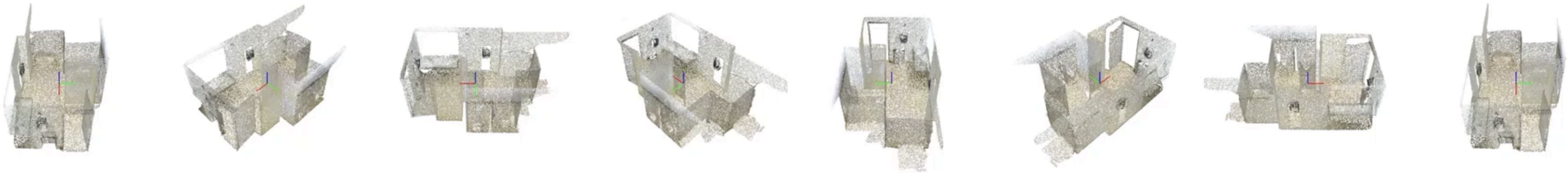
PointCNN (mIoU = **57.3**)

Latency: **1129.2** ms / scene



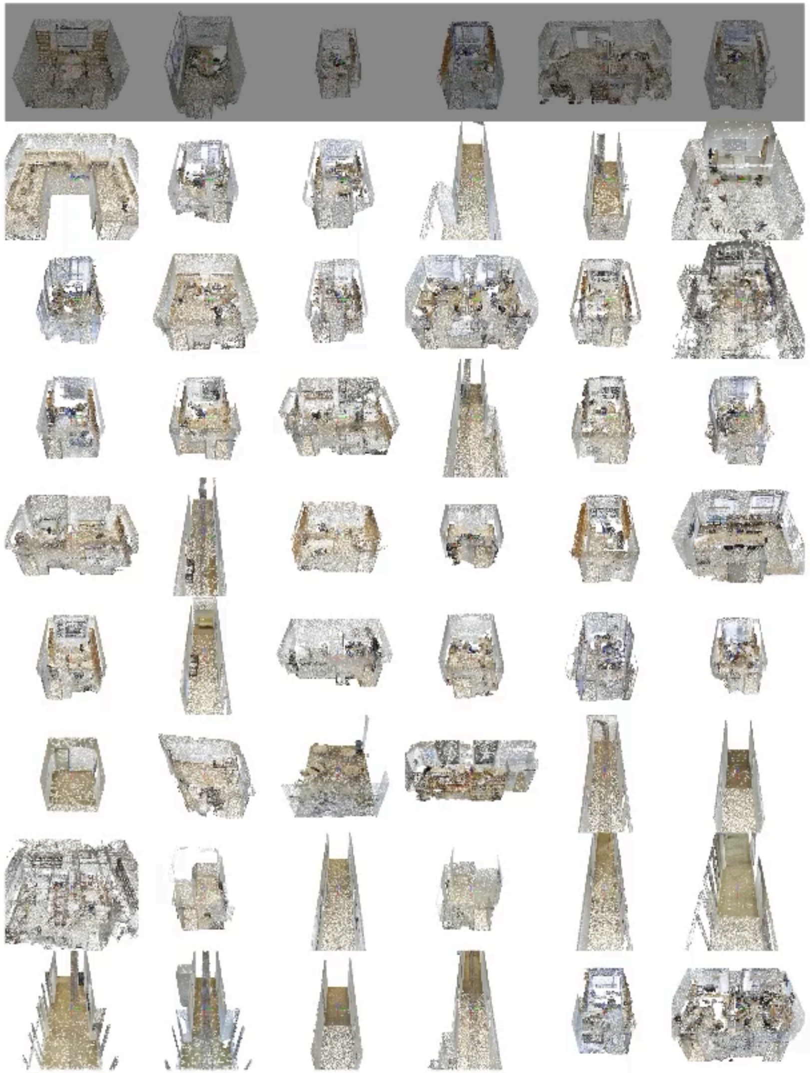
PVCNN (mIoU = **59.0**)

Latency: **278.0** ms / scene



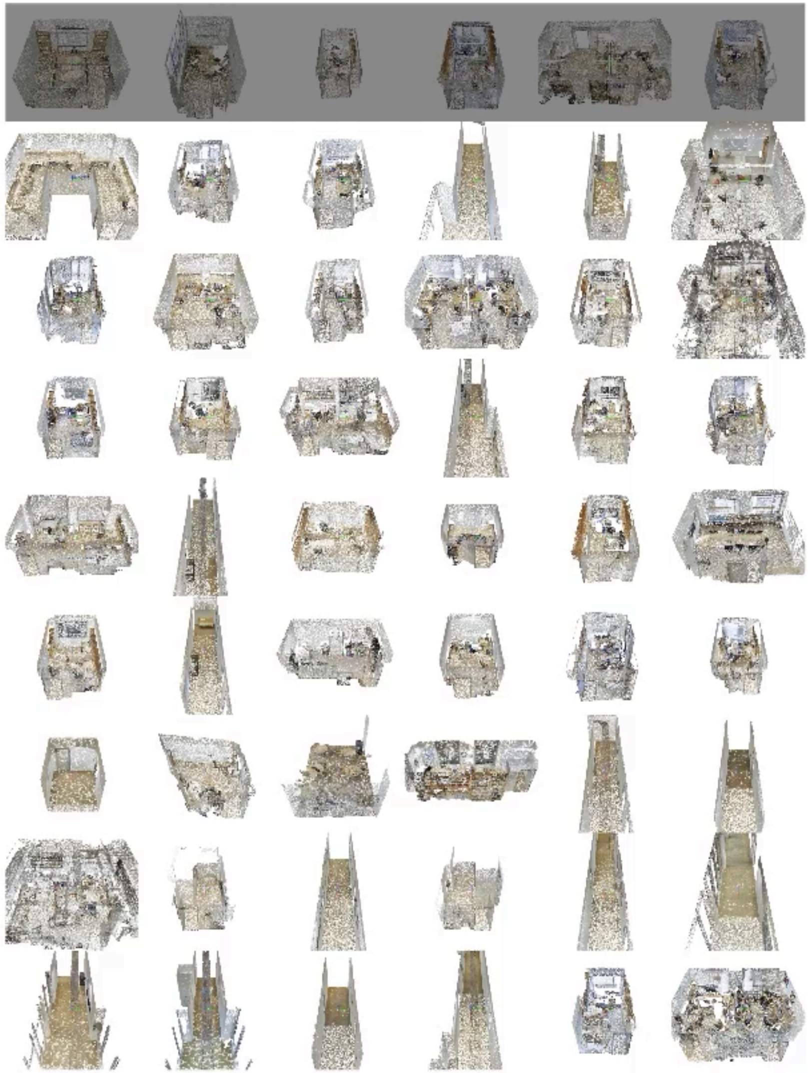
PointNet

Throughput: **12 scenes / sec**



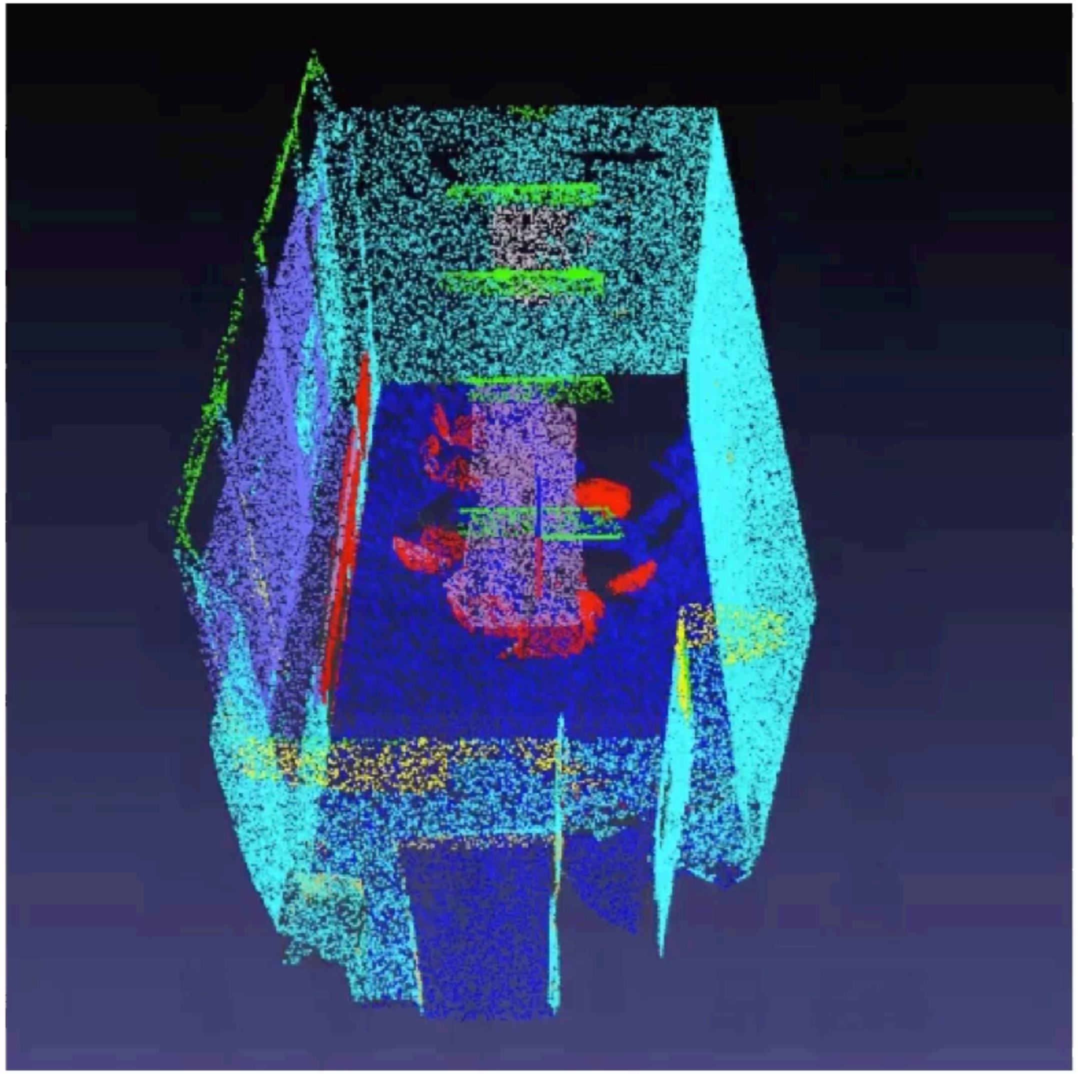
0.25 PVCNN (Ours)

Throughput: **21 scenes / sec**



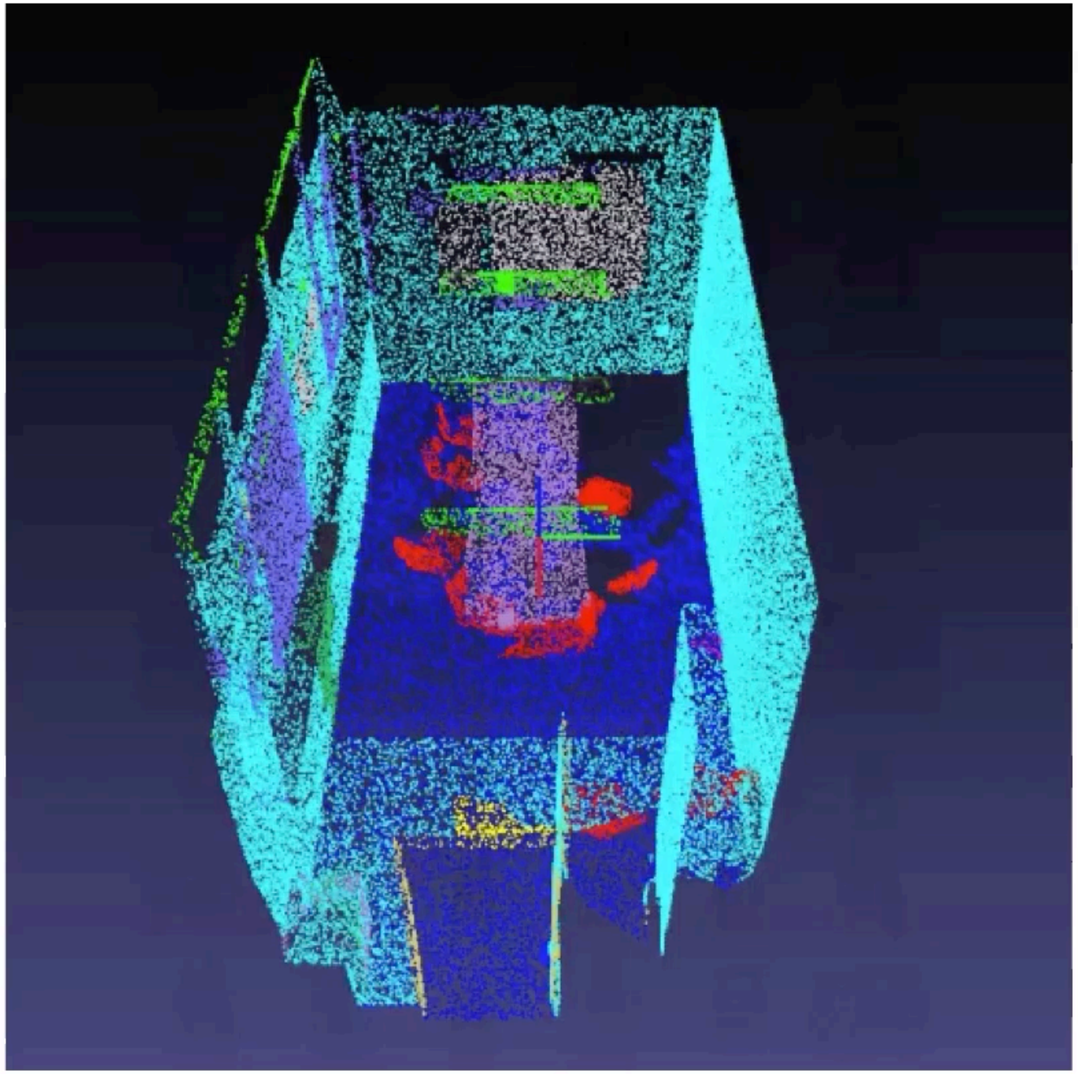
PointNet

S3DIS Area 5 IoU = **43.0**%



0.25 PVCNN (Ours)

S3DIS Area 5 IoU = **52.3**%



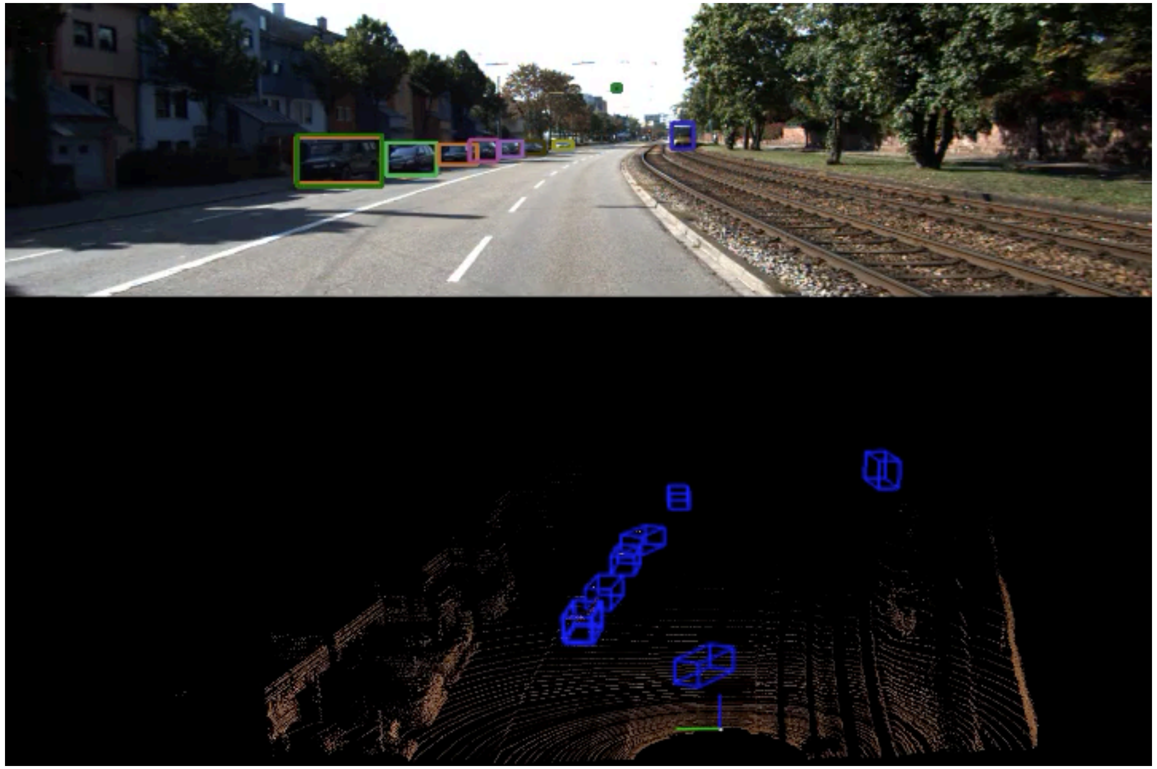
[PVCNN, NeurIPS’19]

Results: 3D Object Detection (KITTI)

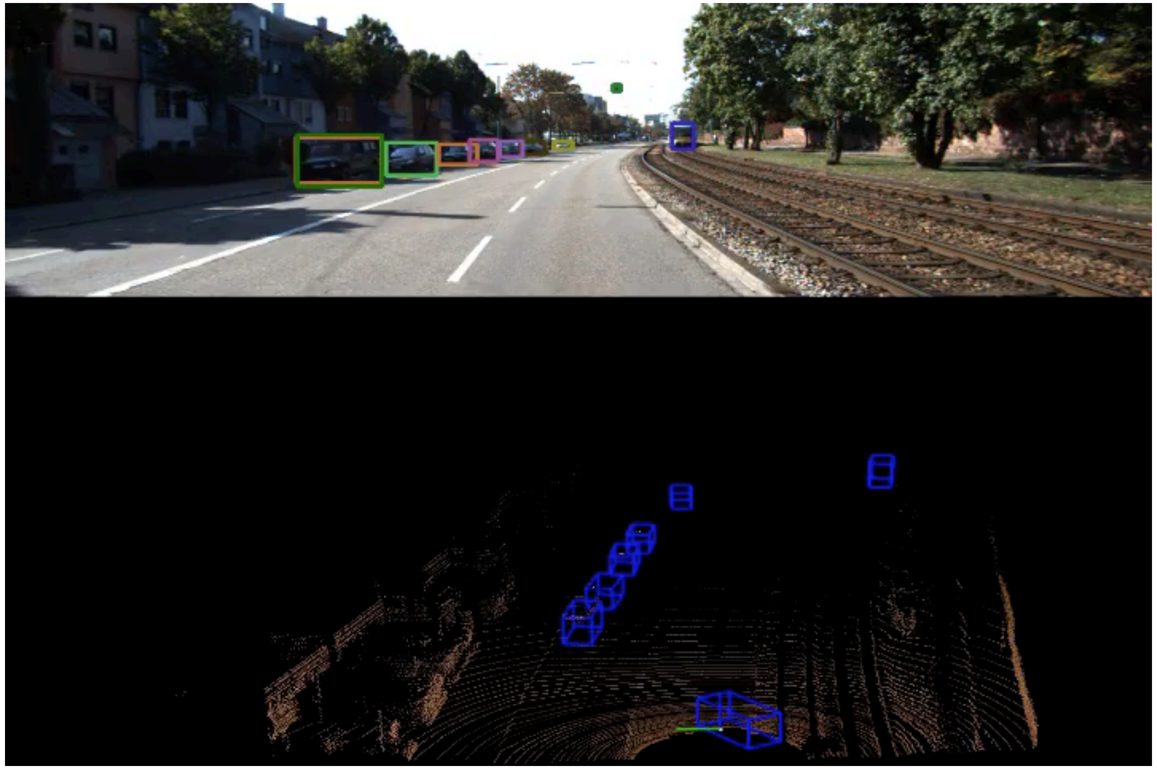
	GPU Latency	GPU Memory	Pedestrian	Cyclist	Car
F-PointNet++	105.2 ms	2.0 GB	61.6	62.4	72.8
PVCNN (efficient)	58.9 ms (1.8x)	1.4 GB (1.4x)	60.7 (-0.9)	63.6 (+1.2)	73.0 (+0.2)
PVCNN (complete)	69.6 ms (1.5x)	1.4 GB (1.4x)	64.9 (+3.3)	65.9 (+3.5)	73.1 (+0.3)
	Faster	Lower	More Accurate		

[PVCNN, NeurIPS'19]

Results: 3D Object Detection (KITTI)



F-PointNet++
(10 FPS)

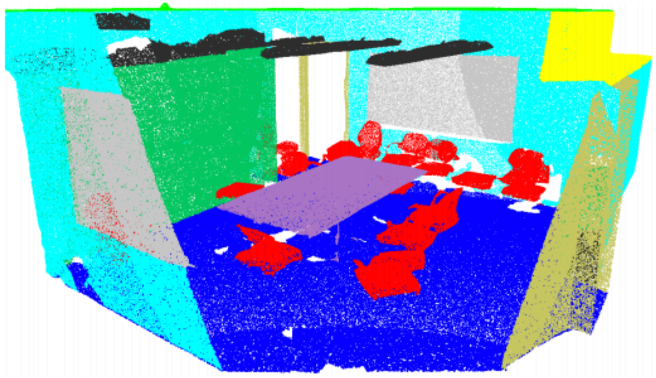


PVCNN
(17 FPS, 1.8x faster)

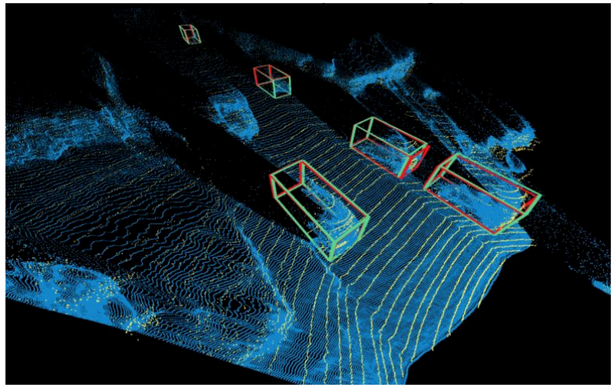
[PVCNN, NeurIPS'19]

Point-Voxel CNN for Efficient 3D Deep Learning

2.7x measured speedup
1.5x memory reduction



6.9x measured speedup
5.7x memory reduction



1.8x measured speedup
1.4x memory reduction

Gold Medal in Lyft Challenge on 3D Object Detection for Autonomous Vehicles

Project Page: http://pvcnn.mit.edu

GitHub: https://github.com/mit-han-lab/pvcnn

Efficient Deep Learning on the Edge

- **Efficient 3D Algorithms:**
  - PVCNN for efficient point-cloud recognition [NeurIPS'19, spotlight]
  - TSM for efficient video recognition [ICCV'19]
- **Compression / NAS**
  - Deep Compression [NIPS'15, ICLR'16]
  - ProxylessNAS, AMC, HAQ [ICLR'19, ECCV'18, CVPR'19, oral]
  - Once-For-All (OFA) Network

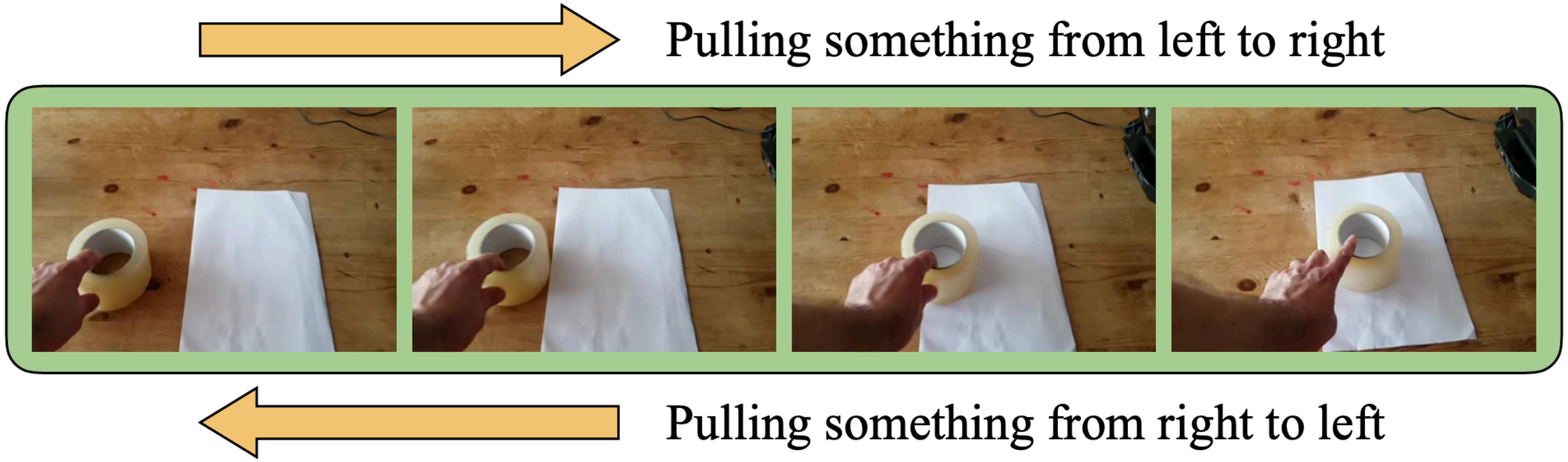
Background

- Videos are growing explosively: 10^5 hours of videos are uploaded to YouTube/day
- Efficient Video processing is essential for both Cloud and Edge (e.g., hospitals)

YouTube Uploads: > 300 Hours of Video per Minute

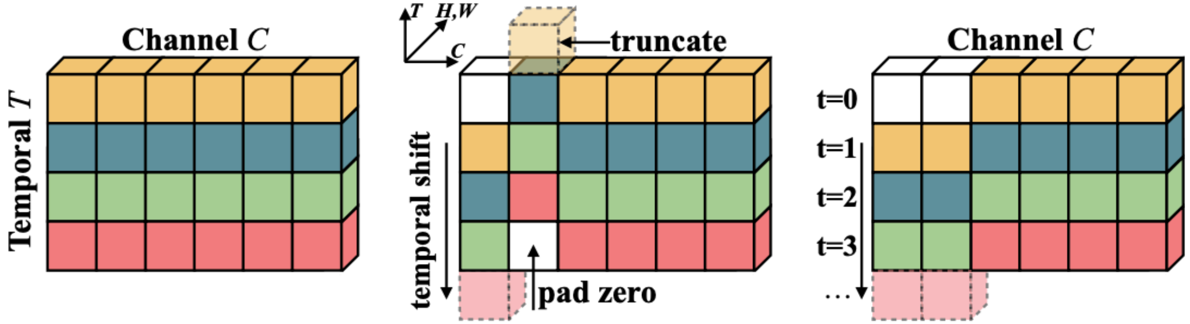
Overview

- Efficient spatial-temporal modeling is important for video understanding
- **2D CNN** is more efficient, but it cannot handle temporal modeling
- **3D CNN** can perform joint spatial-temporal feature learning, but it is computationally expensive
- We aim to **achieve 3D CNN performance at 2D complexity**



Temporal Shift Module (TSM)

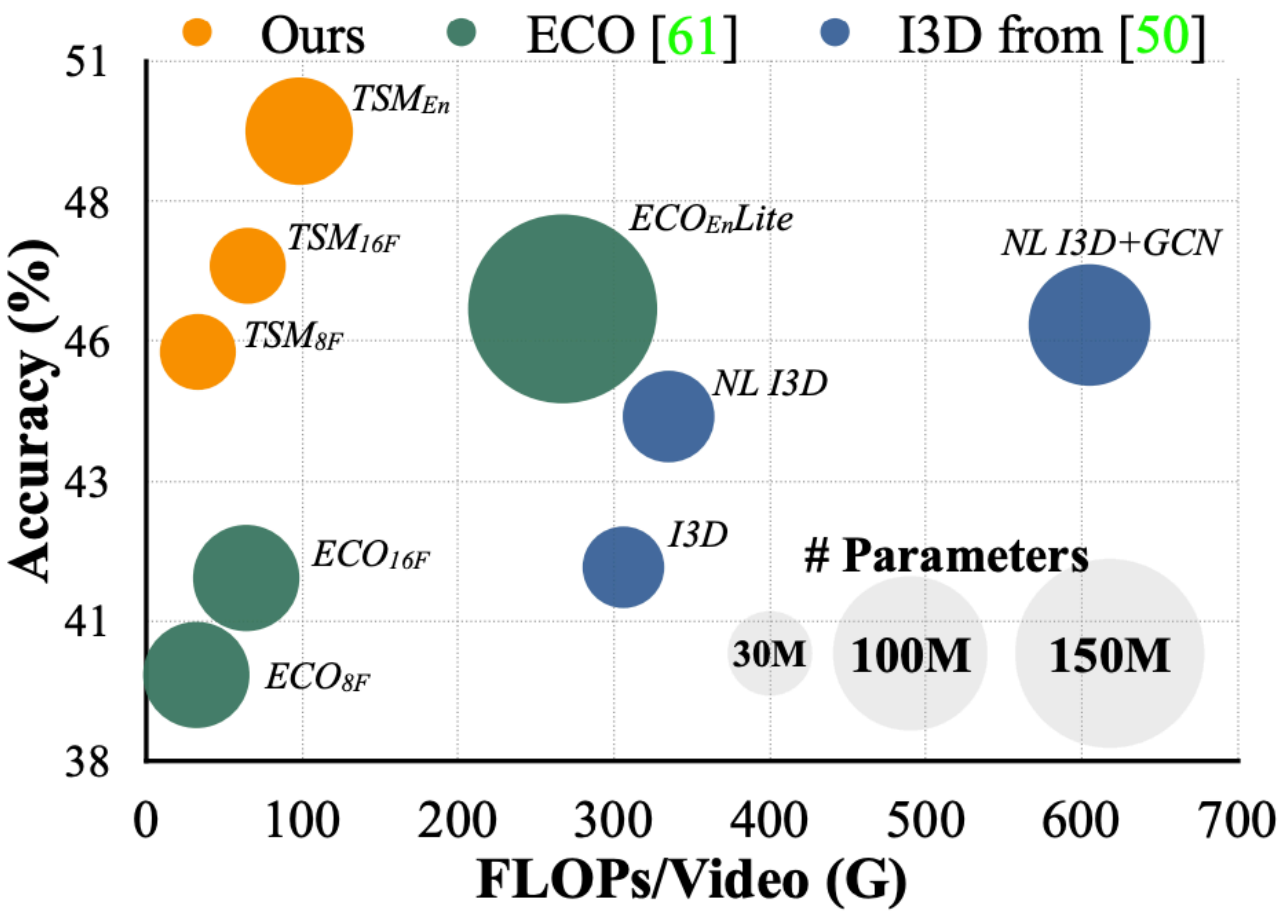
- **Bi-directional** TSM shifts part of the channels along the temporal dimension to facilitate information exchange among neighboring frames
- **Uni-directional** TSM shifts channels from past to future for **online** video understanding.
- It can be inserted into off-the-shelf 2D CNN to enable temporal modeling at the cost of *zero FLOPs* and *zero parameters*



(a) The original tensor without shift. (b) Offline temporal shift (bi-direction). (c) Online temporal shift (uni-direction).

Latency and Throughput Speedup

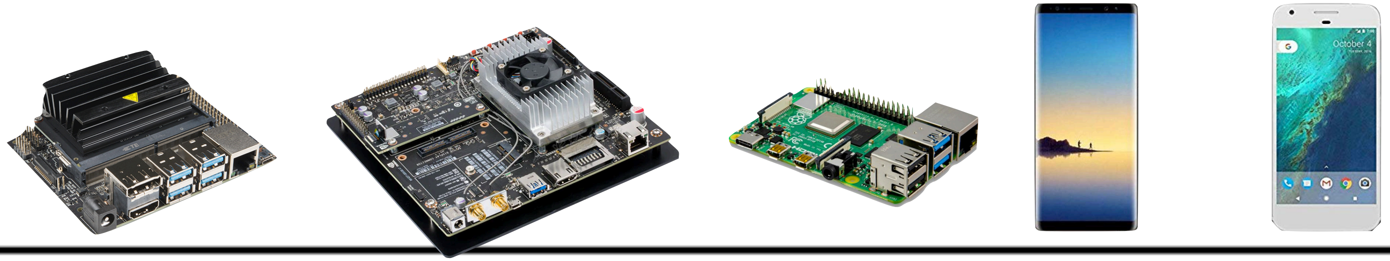
- Efficiency statistics and accuracy comparison



TSM, ICCV'19

[TSM, ICCV’19]

Scaling Down: Low-Latency Low-Power Deployment

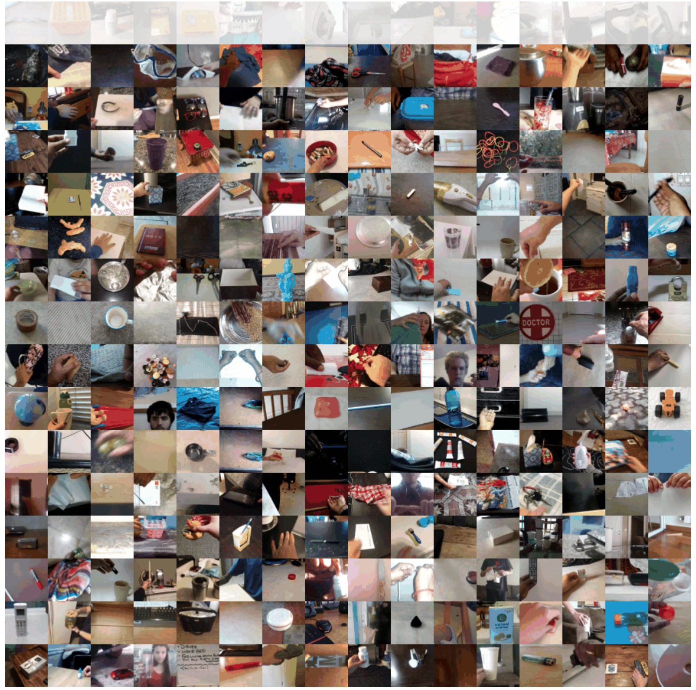


Devices	Jetson Nano		Jetson TX2		Rasp.	Note8	Pixel1
	CPU	GPU	CPU	GPU			
Latency (ms)	47.8	13.4	36.4	8.5	69.6	34.5	47.4
Power (watt)	4.8	4.5	5.6	5.8	3.8	-	-

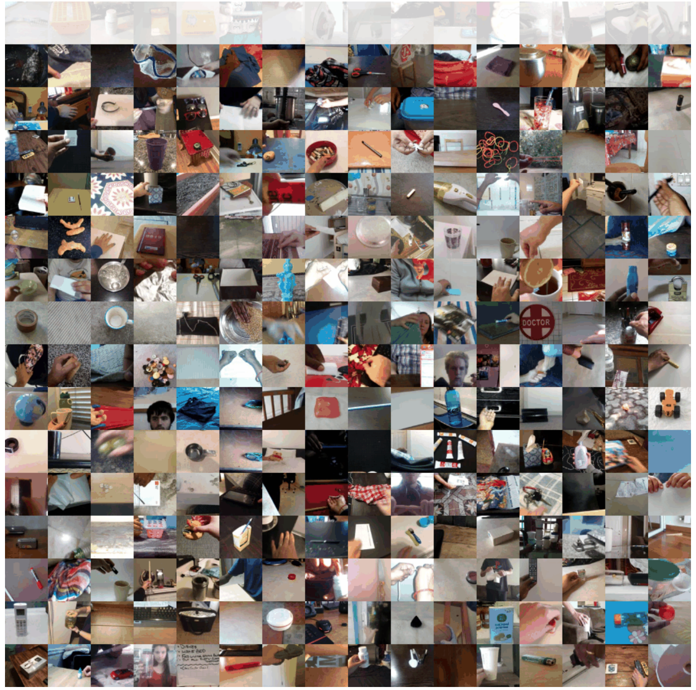
LED Bulb Level!

Accelerating Video Understanding

I3D:
Throughput: **6.1** video/s



TSM:
Throughput: **77.1** video/s



12.6x higher throughput

[TSM, ICCV’19]

Accelerating Video Understanding

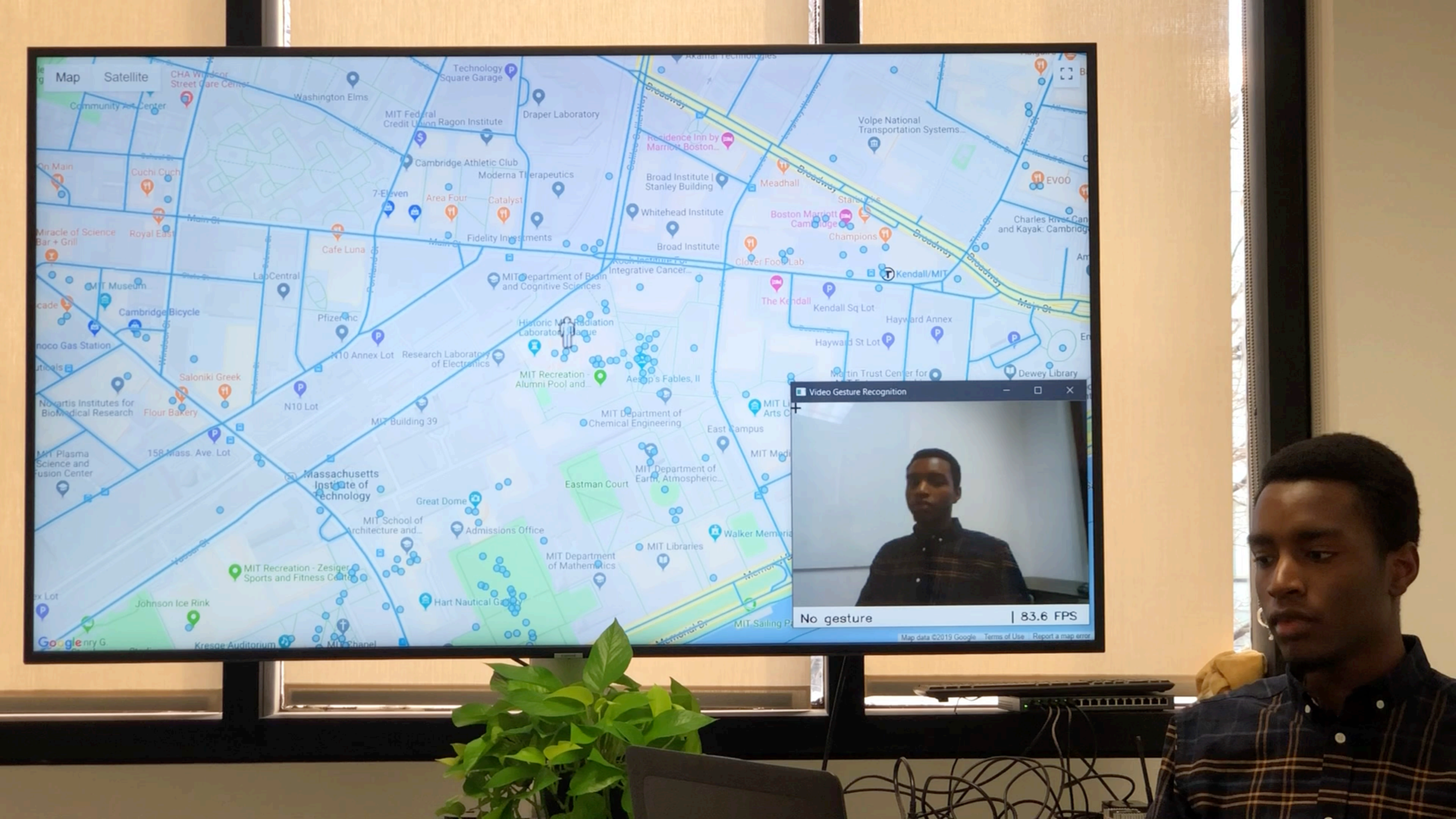
I3D:
Latency: **164.3** ms/Video

TSM:
Latency: **20.7** ms/Video

TSM, ICCV’19

Speed-up: 8x

https://youtu.be/0T6u7S_gq-4



Map
Satellite
Technology Square Garage
Washington Elms
Community Art Center
MIT Federal Credit Union
Ragon Institute
Draper Laboratory
Cambridge Athletic Club
Moderna Therapeutics
Residence Inn by Marriott Boston...
Volpe National Transportation Systems...
Broad Institute | Stanley Building
Meadhall
Whitehead Institute
Broad Institute
Boston Marriott Cambridge
Champions
EVOO
Charles River Canoe and Kayak: Cambridge
7-Eleven
Area Four
Catalyst
Cuchi Cuchi
Miracle of Science Bar + Grill
Royal East
Cafe Luna
Fidelity Investments
Main St
Koch Institute for Integrative Cancer...
Clover Food Lab
Kendall/MIT
The Kendall
Kendall Sq Lot
MIT Museum
LabCentral
MIT Department of Brain and Cognitive Sciences
Cambridge Bicycle
Pfizer Inc
Hayward Annex
Hayward St Lot
Dewey Library
N10 Annex Lot
Research Laboratory of Electronics
MIT Recreation - Alumni Pool and...
Aesop's Fables, II
Saloniki Greek
Novartis Institutes for BioMedical Research
Flour Bakery
N10 Lot
MIT Building 39
MIT Department of Chemical Engineering
East Campus
158 Mass. Ave. Lot
MIT Plasma Science and Fusion Center
Massachusetts Institute of Technology
Eastman Court
MIT Department of Earth, Atmospheric...
Great Dome
MIT School of Architecture and...
Admissions Office
Walker Memorial
MIT Libraries
MIT Department of Mathematics
MIT Recreation - Zesiger Sports and Fitness Center
Johnson Ice Rink
Hart Nautical Gallery
Kresge Auditorium
MIT Chapel
Video Gesture Recognition
No gesture
83.6 FPS
Map data ©2019 Google
Terms of Use
Report a map error

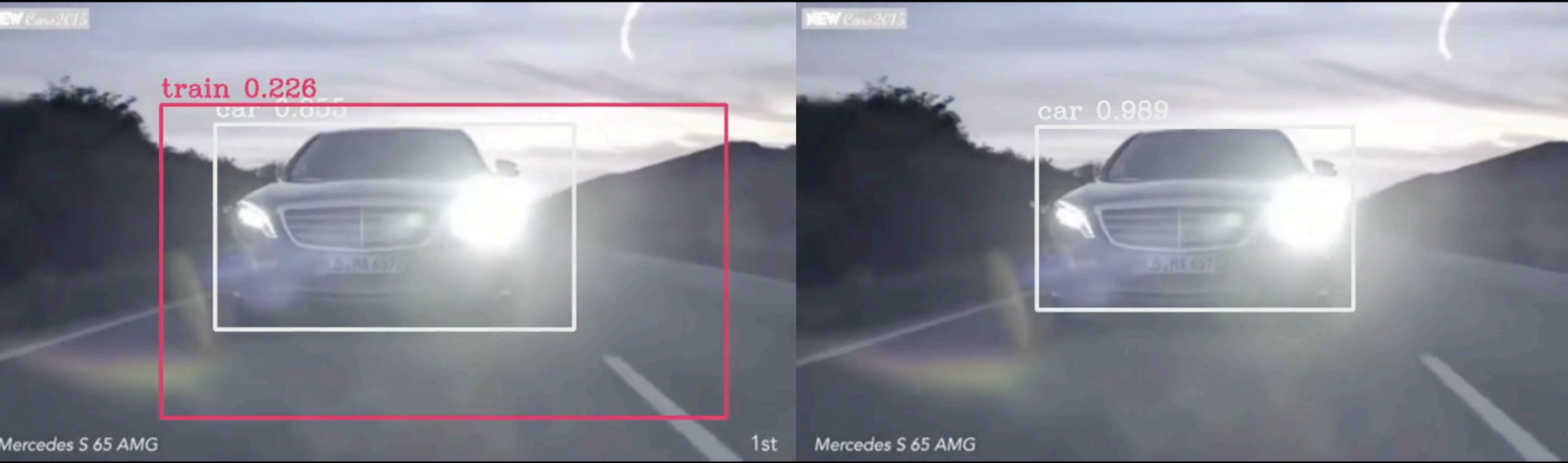
Demo: Robust Object Detection

R-FCN

TSM

Demo: Robust Object Detection

R-FCN TSM

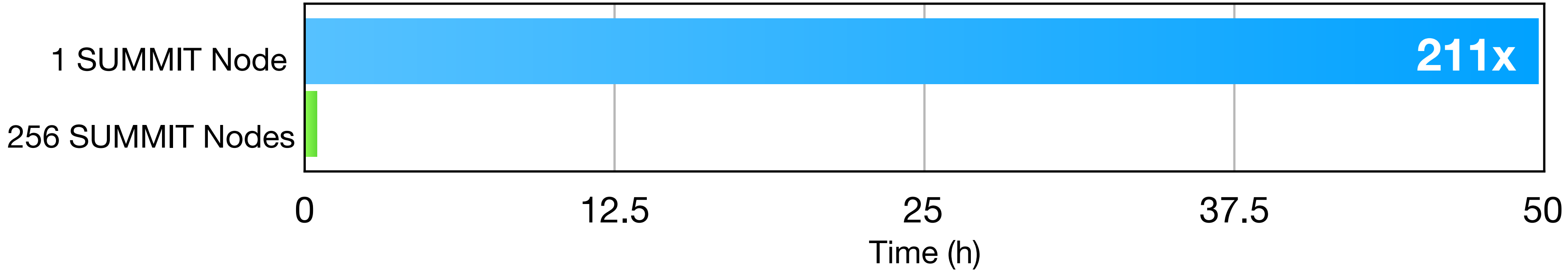


2D baseline gives false positive prediction due to the flare
TSM can correct such errors with the help of temporal information

Large-Scale Distributed Training for Videos

- Speedup video training by 200x, **from 2 days to 14minutes.**

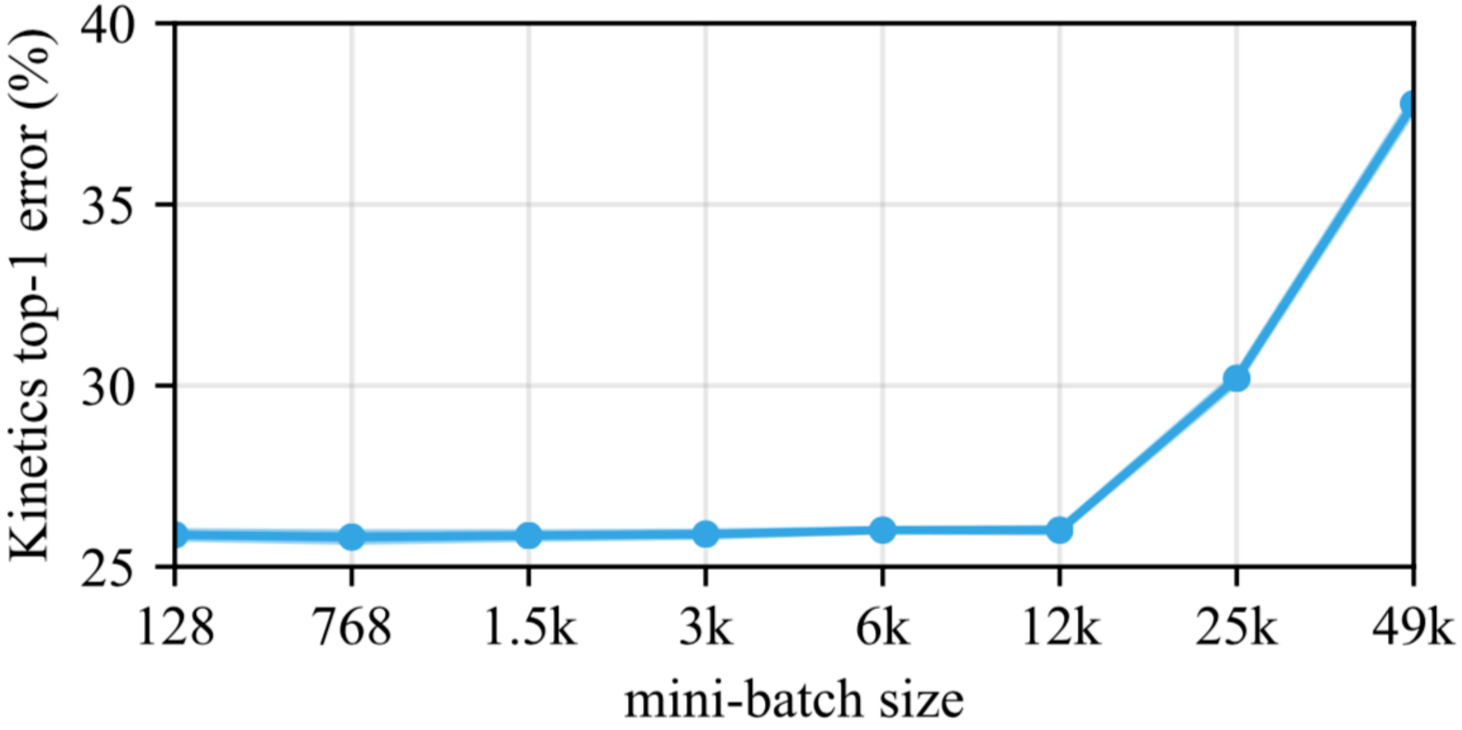
	Training Time	Accuracy	Peak GPU Performance	Speed-up
1 SUMMIT Nodes (6 GPUs)	**49h 50min**	74.1%	93TFLOP/s	Theoretical: 128x Actual: 106x
128 SUMMIT Nodes (768 GPUs)	**28min**	74.1%	12PFLOP/s	
256 SUMMIT Nodes (1536 GPUs)	**14min**	74.0%	24PFLOP/s	Theoretical: 256x Actual: 211x



Lin, Gan, Han, Training Kinetics in 15 Minutes: Large-scale Distributed Training on Videos
NeurIPS'19 workshop on System for ML

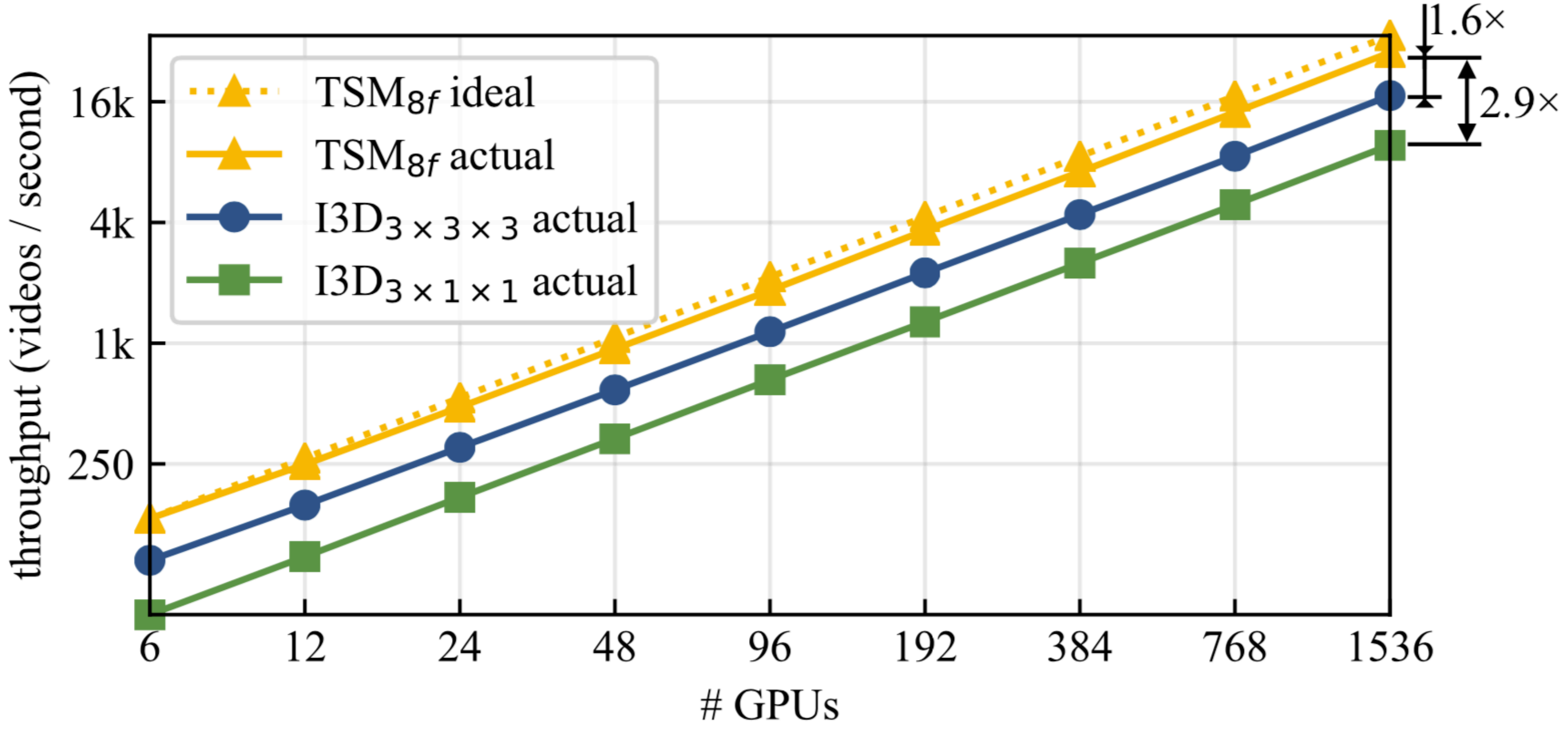
Accuracy v.s. Batch size

- The performance of TSM model does not degrade when we scale up the mini-batch size to 12k.



Scalability v.s. Model

- TSM model achieves 1.6x and 2.9x higher training throughput compared to previous I3D models



Lin, Gan, Han, Training Kinetics in 15 Minutes: Large-scale Distributed Training on Videos
NeurIPS'19 workshop on System for ML

Training Kinetics in 15 Minutes: Large-scale Distributed Training on Videos

Ji Lin[1] Chuang Gan[2] Song Han[1]

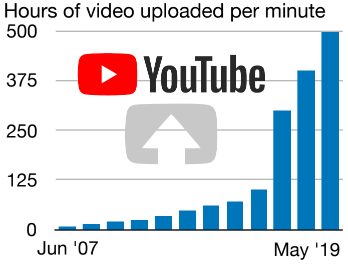
[1] MIT [2] MIT-IBM Watson AI Lab

Project Page MLSys Workshop

tsm-hanlab.mit.edu

Overview

- Video analytics (3D CNN) is heavy
  - Computation: **10x** larger than image.
  - Data I/O: **8-32x** more data per sample.
  - Networking: **1.5-3x** more parameters.
- We propose **Temporal Shift Module (TSM)** to achieve 3D performance at 2D Cost.

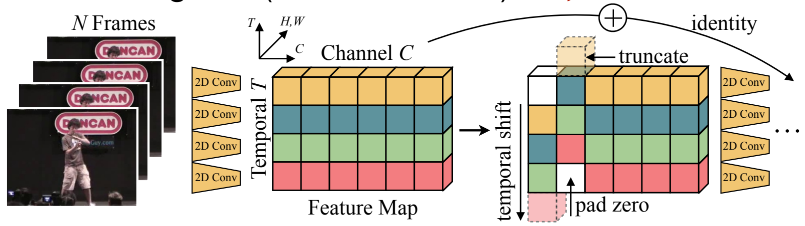


- TSM **scales up** to 1.5k GPUs, training Kinetics in **15min**.
- TSM **scales down**, running 74fps on Jetson Nano.

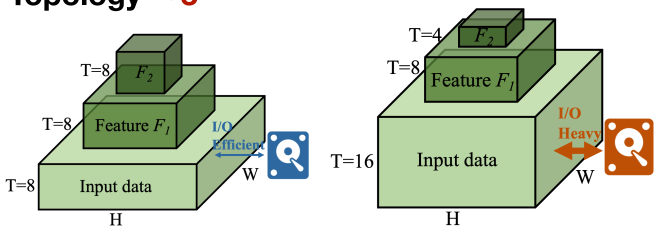
How to Design Scalable Video Models?

1. **Computation efficiency**: fewer FLOPs, higher hardware utilization
2. **Networking efficiency**: fewer parameters → fewer gradients → less networking bandwidth
3. **Data loading efficiency**: fewer parameters → less disk I/O

Temporal Modeling Unit (TSM *vs* 3D Conv) **→1, 2**



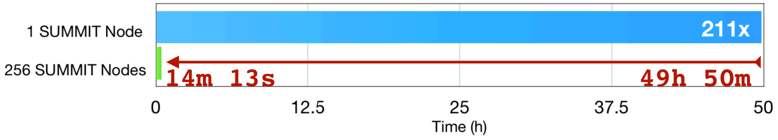
Backbone Topology →3



(a) Straight-up **(b)** Pooled-up

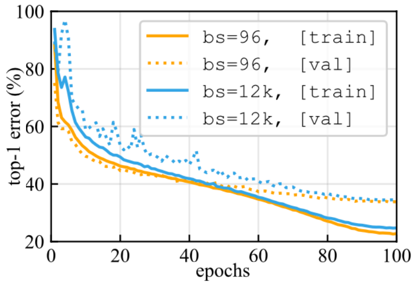
Scale Up: TSM on Summit Supercomputer

- 1,536 GPUs, Batch size: 12,288 videos/98,304 frames
- Achieving >80% scalability
- Training time (Kinetics): **2 days → 15 minutes**

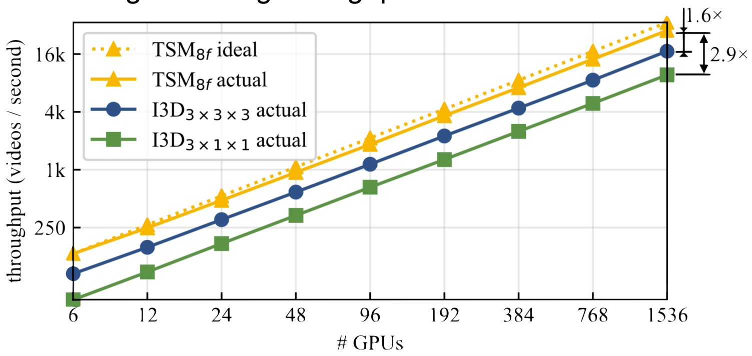


*Summit: 6*V100/node*

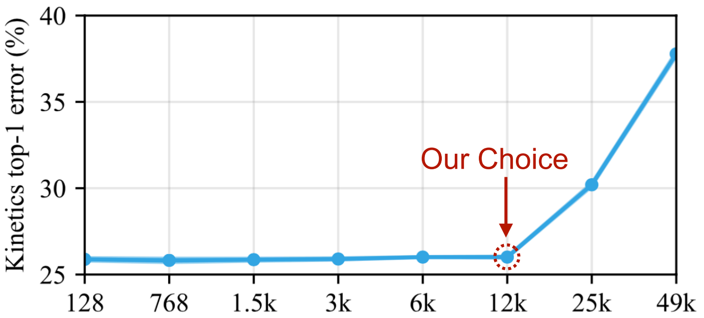
Training curves (12k)



- Better scalability than 3D CNN
- 2-3x larger training throughput on same hardware

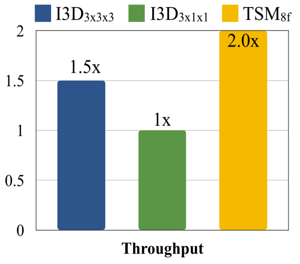


- Accuracy *v.s.* batch size. No degradation at 12k

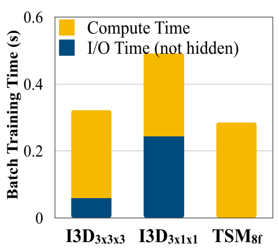


Analyzing Training Efficiency

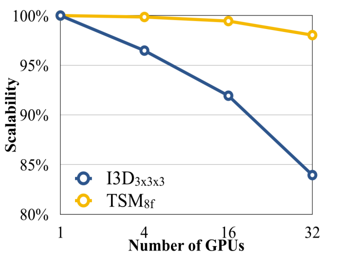
- Analyzing on AWS



(a) training throughput



(b) data I/O

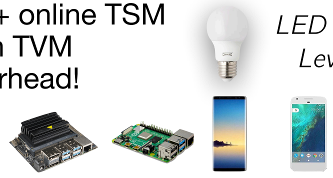


(c) Scalability/networking

Scale Down: Edge Deployment

- MobileNetV2 + online TSM
- Compiled with TVM
- Near zero overhead!

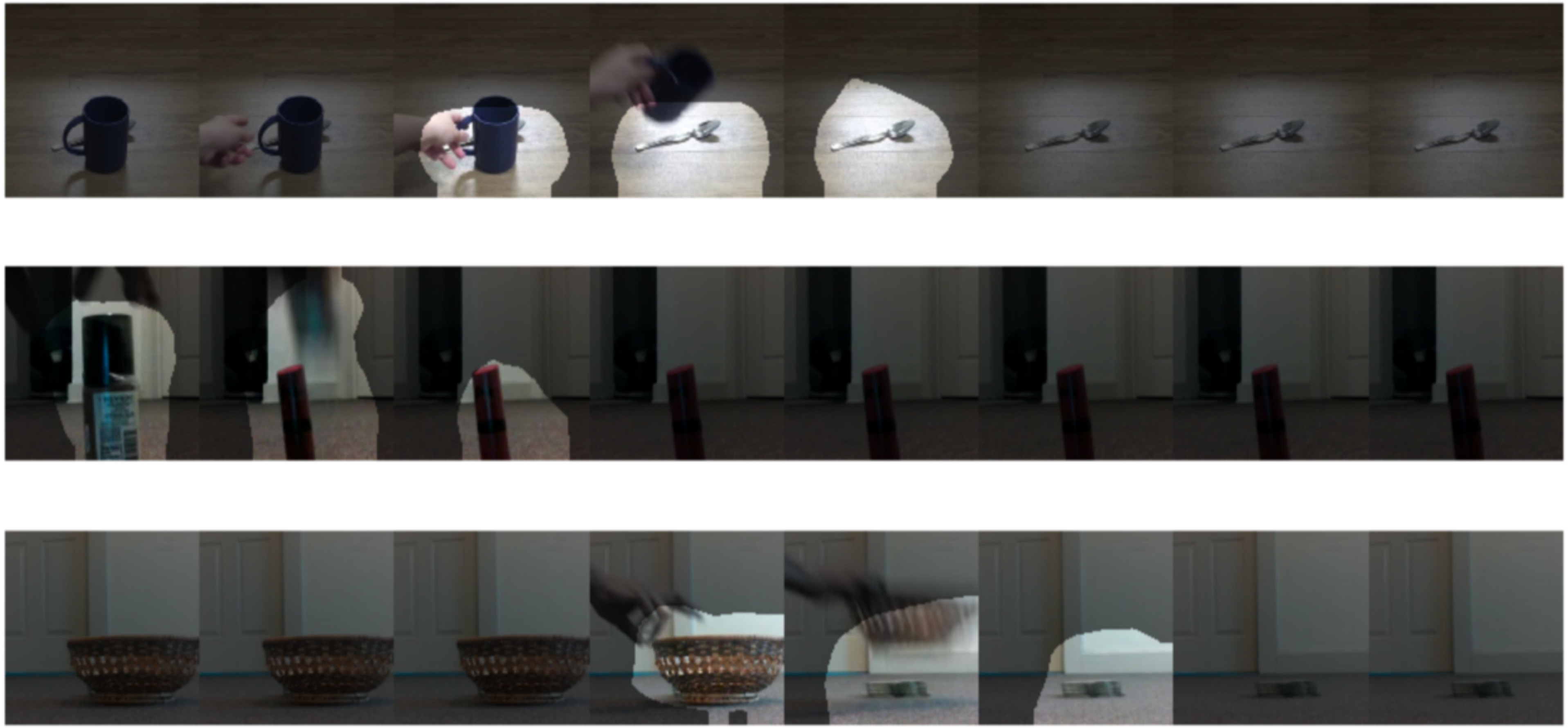
LED Bulb Level!



Devices	Jetson Nano CPU	Jetson Nano GPU	Rasp.	Note8	Pixel1
FPS	20.9	74.6	14.4	29.0	21.1
Power (watt)	4.8	4.5	3.8	-	-

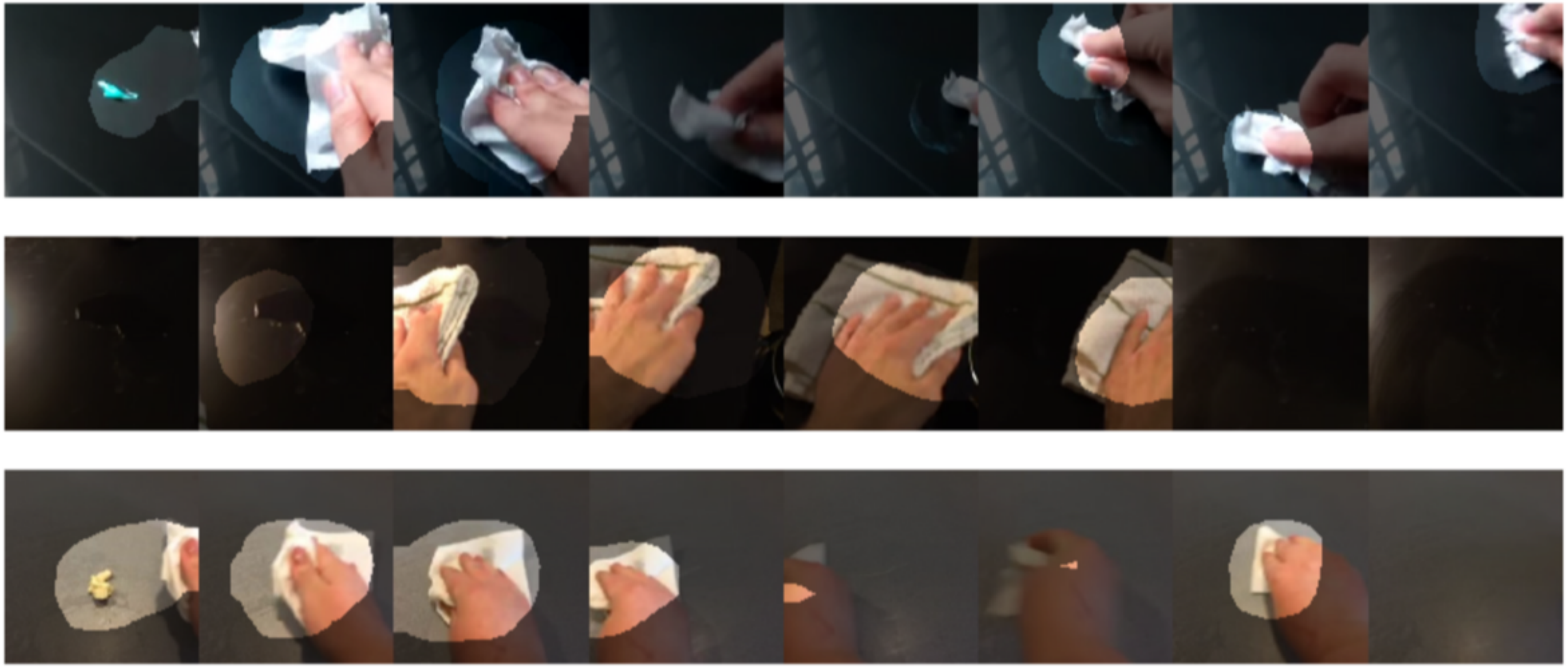
TSM Dissection: Spatial-Temporal Localization

- Each channel learns different semantics
- *Channel 5: Move something away*



TSM Dissection: Spatial-Temporal Localization

- Each channel learns different semantics
- *Channel 162: Wiping*



TSM Dissection: Spatial-Temporal Localization

- Each channel learns different semantics
- *Channel 446: Push to left*

TSM Dissection: Spatial-Temporal Localization

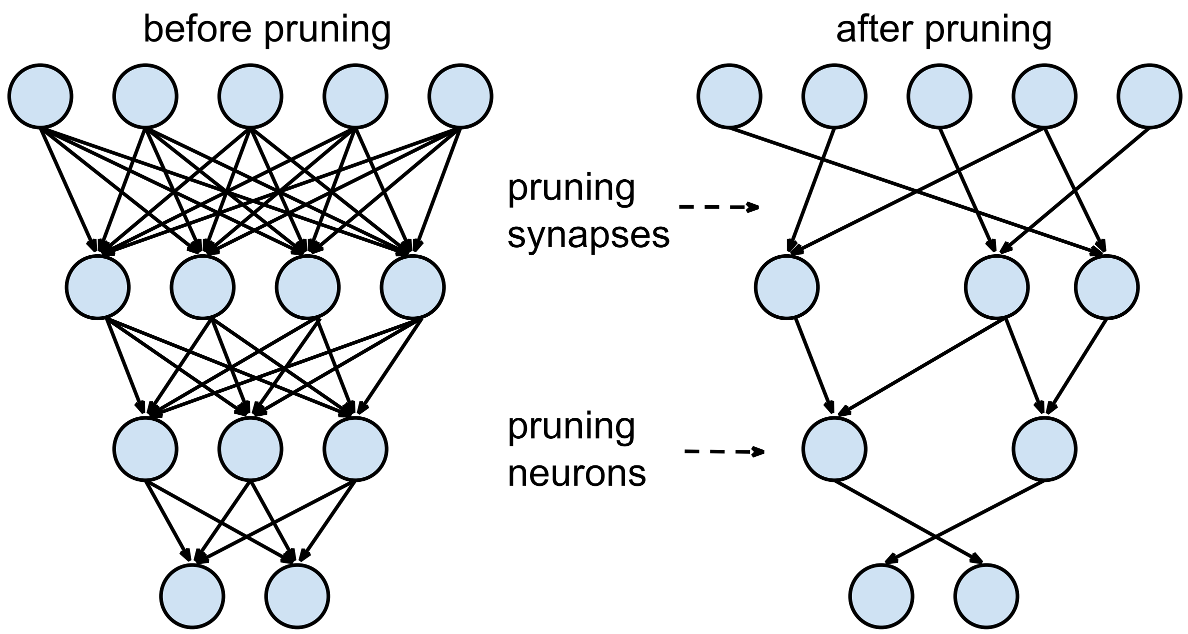
- Each channel learns different semantics
- *Channel 647: Flipping Book pages*

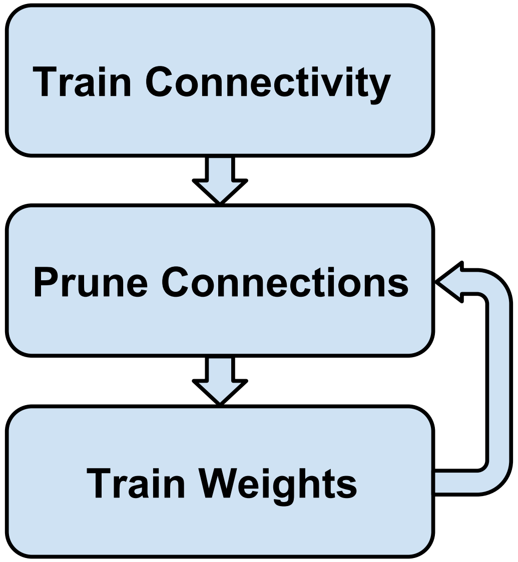
Efficient Deep Learning on the Edge

- ✦ **Efficient 3D Algorithms:**
  - PVCNN for efficient point-cloud recognition [NeurIPS'19, spotlight]
  - TSM for efficient video recognition [ICCV'19]
- ✦ **Compression / NAS**
  - Deep Compression [NIPS'15, ICLR'16]
  - ProxylessNAS, AMC, HAQ [ICLR'19, ECCV'18, CVPR'19, oral]
  - Once-For-All (OFA) Network

[NIPS'15]

Pruning

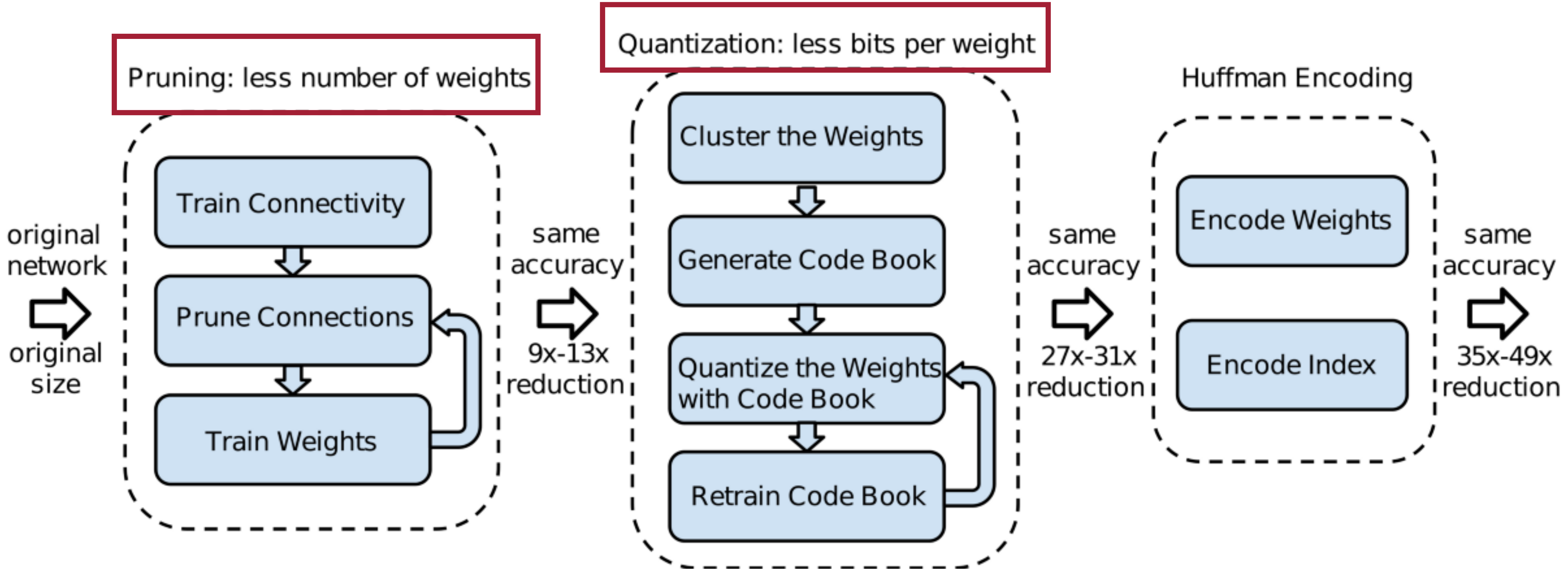




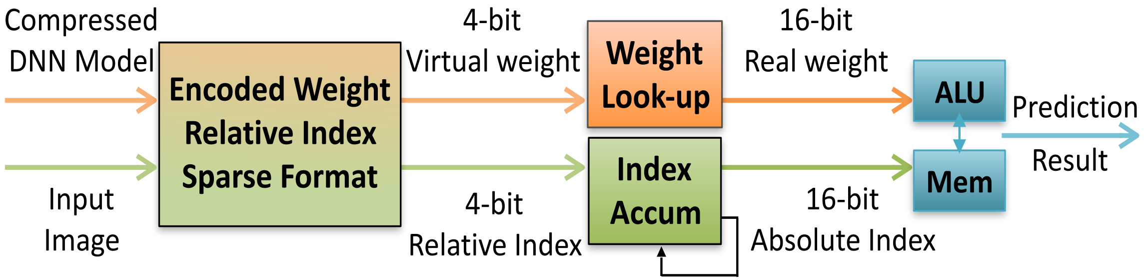
Han, Pool, Tran, Dally, [Learning both Weights and Connections for Efficient Neural Networks](), NIPS'15

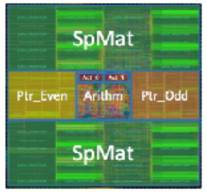
[ICLR'16]

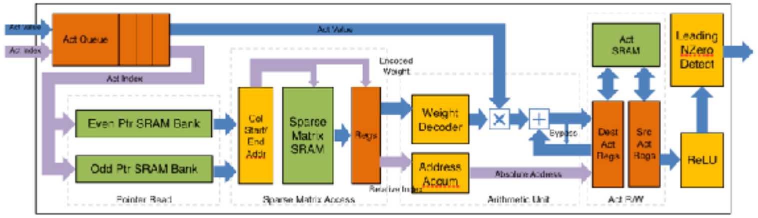
Deep Compression

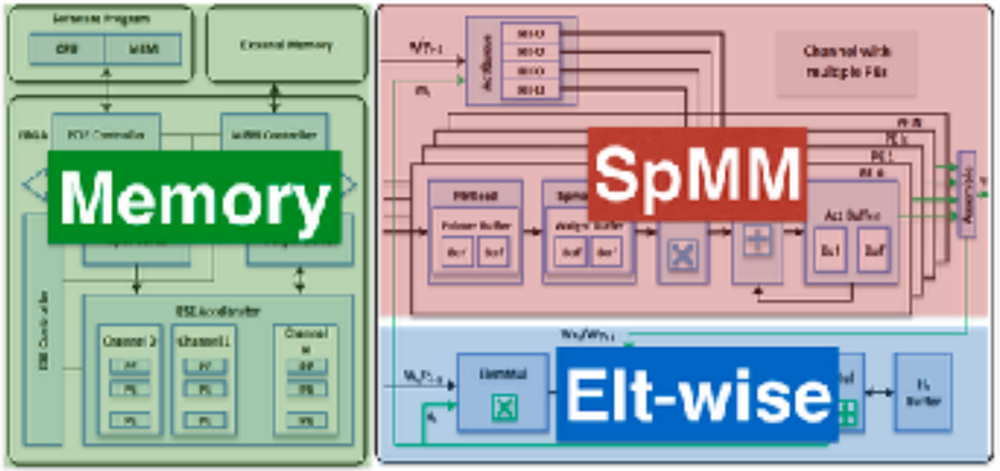


Hardware Acceleration









EIE Accelerator

Han et al [ISCA'16]

ESE Accelerator

Han et al [FPGA'17]
Best Paper Award
Available on AWS Marketplace

[EIE: Efficient Inference Engine on Compressed Deep Neural Network](#)
[ESE: Efficient Speech Recognition Engine with Sparse LSTM on FPGA](#)

Speedup Image Classification

Before Compression
30FPS

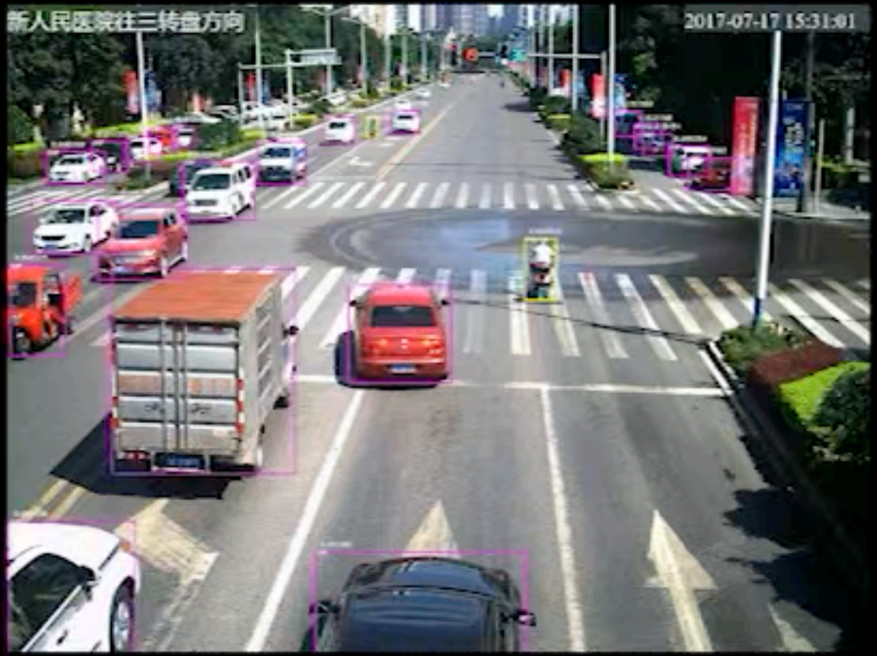
After 2.5X Compression
62FPS

DEEPHi 深鉴科技

Without Compression

4 FPS

With Compression



30 FPS

Accelerating Horse2zebra by GAN Compression

Original CycleGAN; FLOPs: 56.8G; FPS: 12.1; FID: 61.5

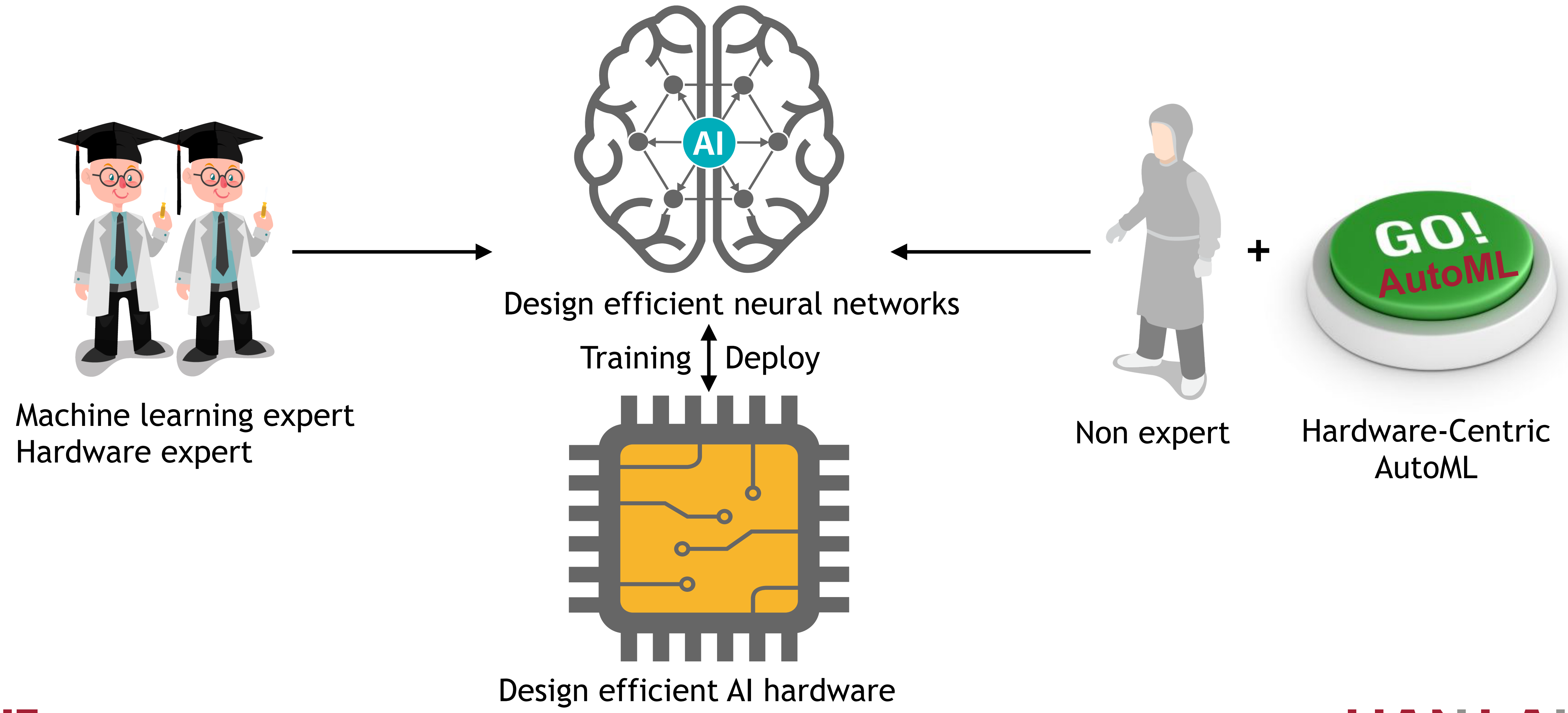
GAN Compression; FLOPs: 3.50G (16.2x); FPS: 40.0 (3.3x); FID: 53.6

Measured on NVIDIA **Jetson Xavier GPU**
Lower FID indicates better Performance.

Pruning / Quantization / Deep Compression in Industry

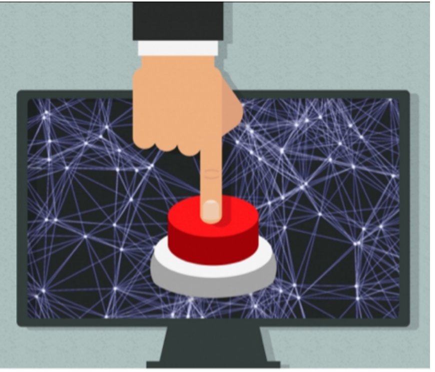
- **DeePhi Tech / Xilinx**
- **Samsung** NPU (sparsity-aware NPU in Galaxy Note10)
- **Intel NNP-I** (comp./decomp. unit support for sparse weights)
- **Qualcomm**: AIMET (a model efficiency tool) is OS soon.
- **Tensorflow / Keras**

AutoML


AI
Design efficient neural networks
Training
Deploy
Design efficient AI hardware
Machine learning expert
Hardware expert
Non expert
+
GO!
AutoML
Hardware-Centric
AutoML

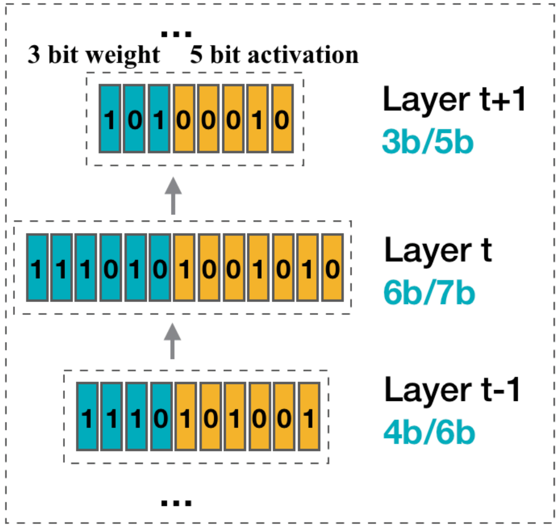
Efficient Deep Learning on the Edge

- **Efficient 3D Algorithms:**
  - PVCNN for efficient point-cloud recognition [NeurIPS'19, spotlight]
  - TSM for efficient video recognition [ICCV'19]
- **Compression / NAS**
  - Deep Compression [NIPS'15, ICLR'16]
  - **ProxylessNAS, AMC, HAQ [ICLR'19, ECCV'18, CVPR'19, oral]**
  - Once-For-All (OFA) Network



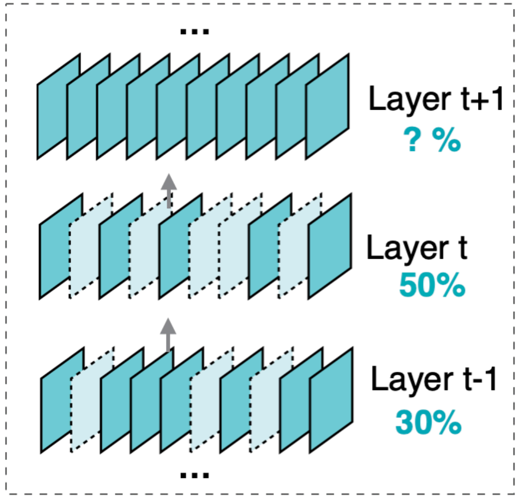
AutoML for Architecting Efficient and Specialized Neural Networks

Presented at NIPS’18 Workshop
to appear at IEEE Micro



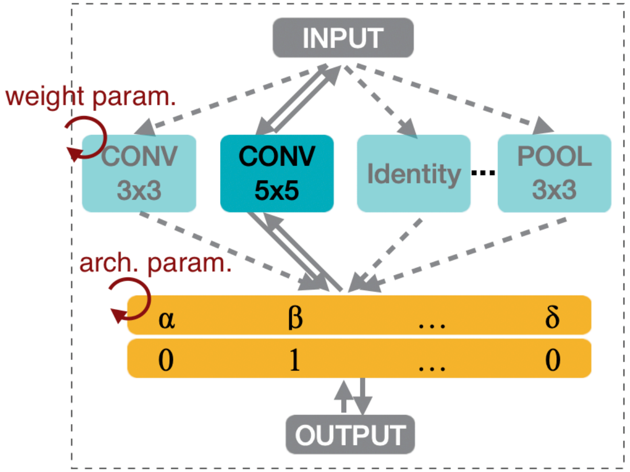
HAQ: Hardware-aware Automated Quantization

[CVPR 2019], oral



AMC: AutoML for Model Compression

[ECCV 2018]

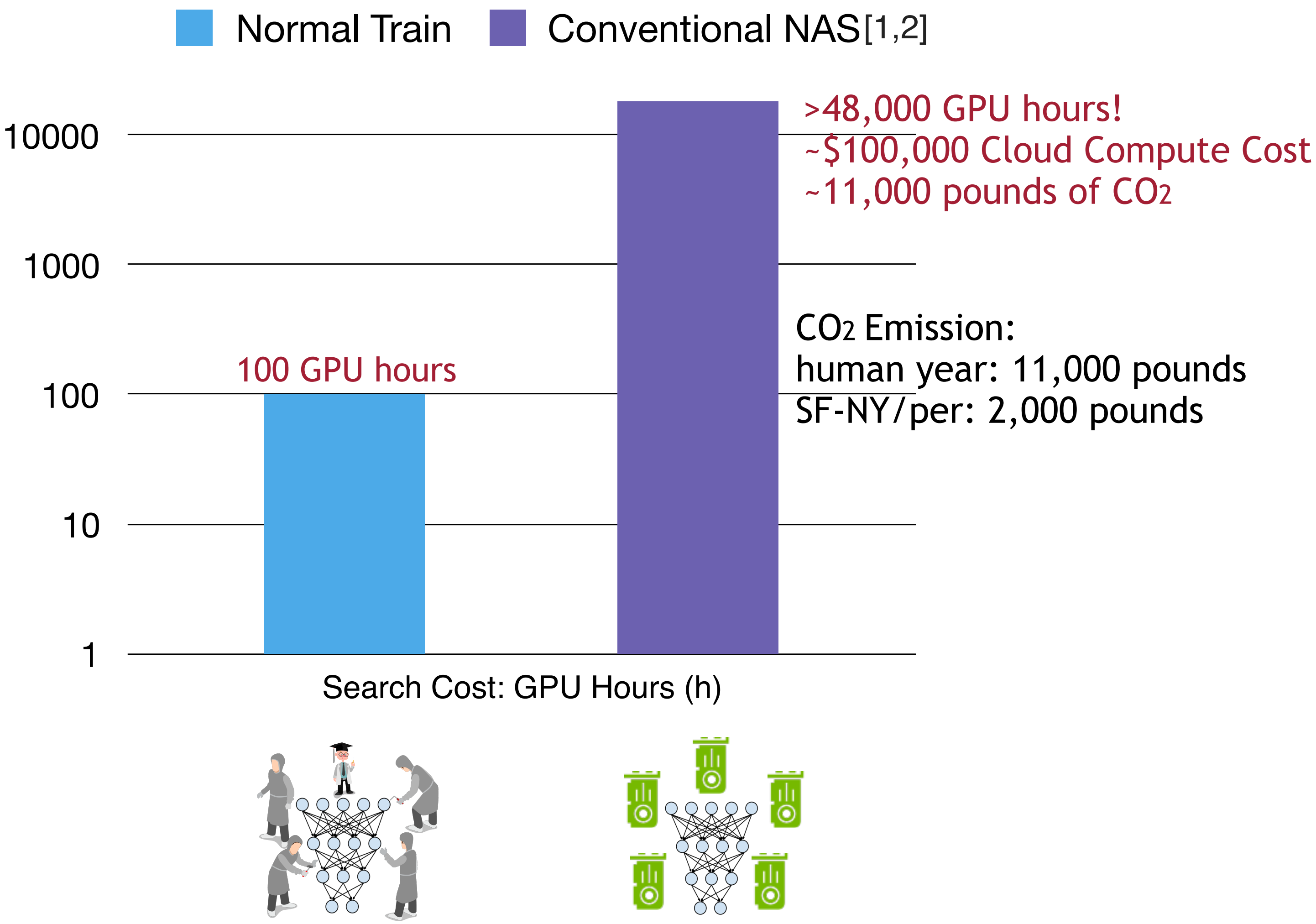


Proxyless Neural Architecture Search

[ICLR 2019]

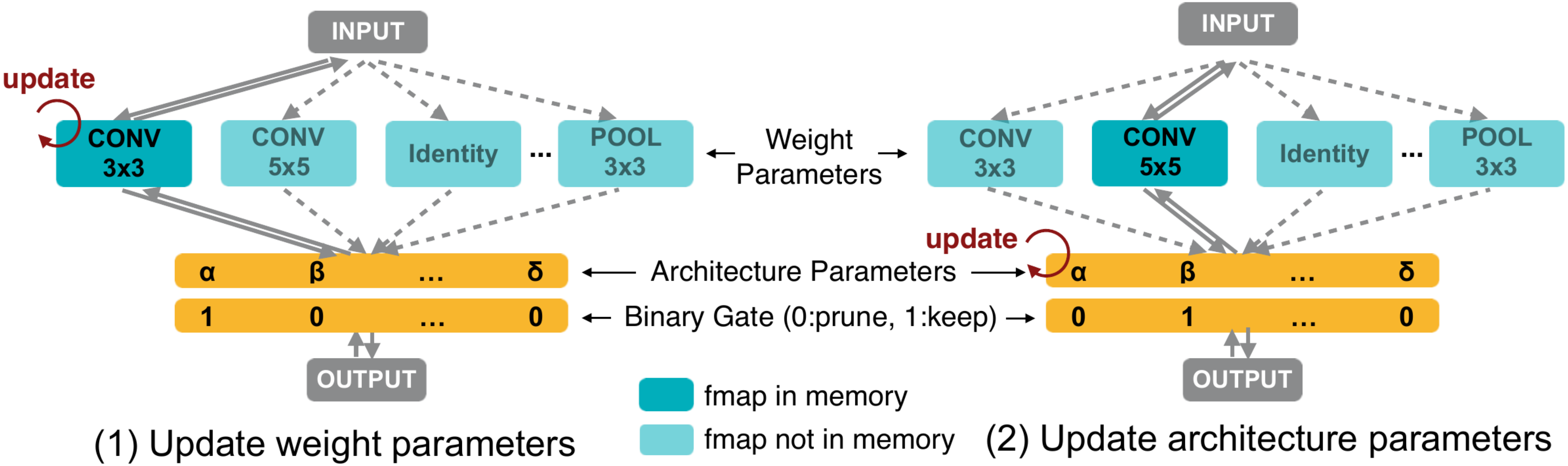
1. ProxylessNAS: automatically architect efficient neural networks
2. AMC: automatic model compression (channel pruning)
3. HAQ: automatic quantization with mixed precision

Conventional NAS: High Cost, >$100K!!!



[1] B Zoph, QV Le, "Neural Architecture Search with Reinforcement Learning"
[2] E Real, A Aggarwal, Y Huang, QV Le, "Regularized evolution for image classifier architecture search"

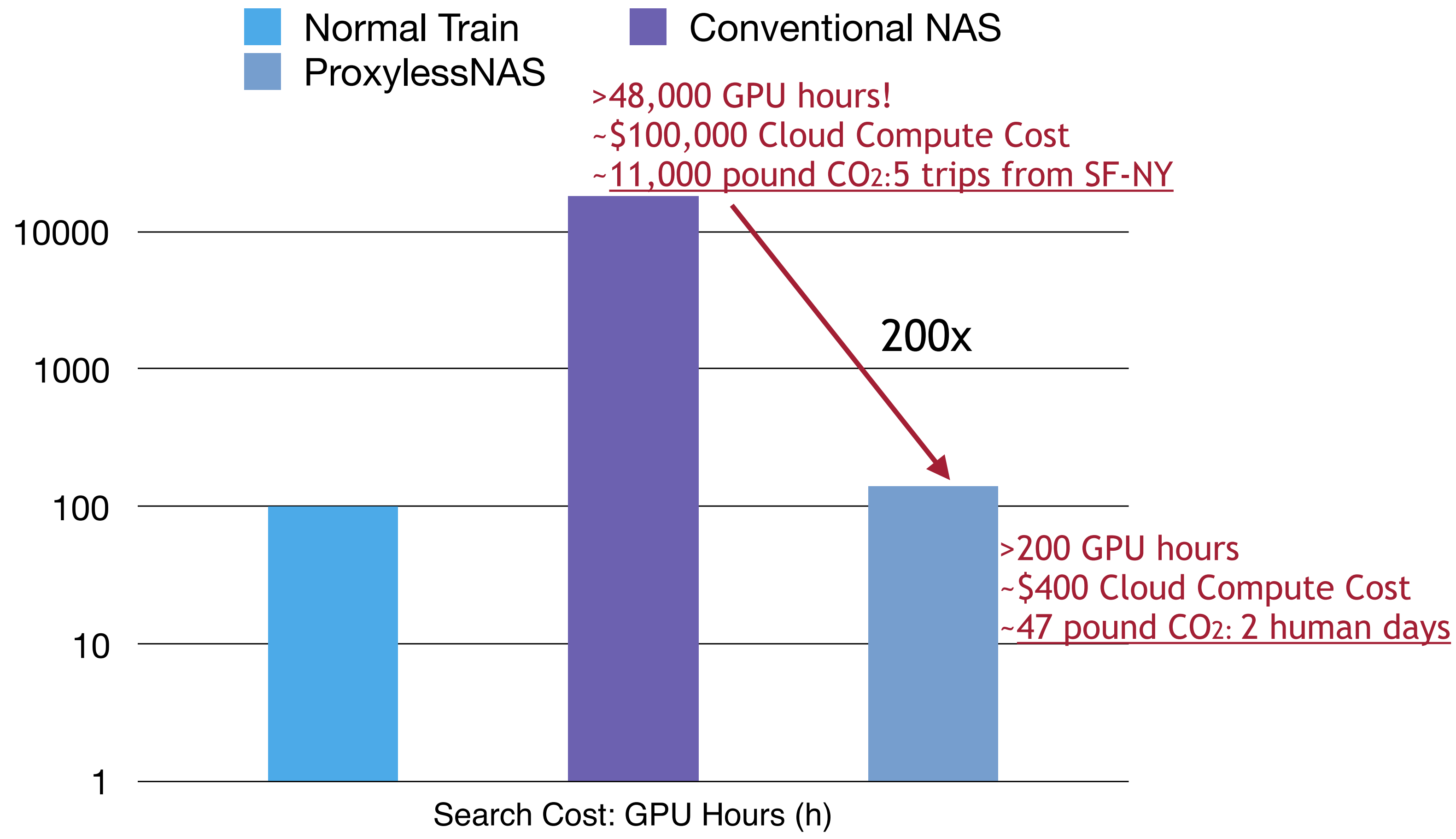
ProxylessNAS: Implementation



Only one path in GPU memory. Scalable to a large design space.

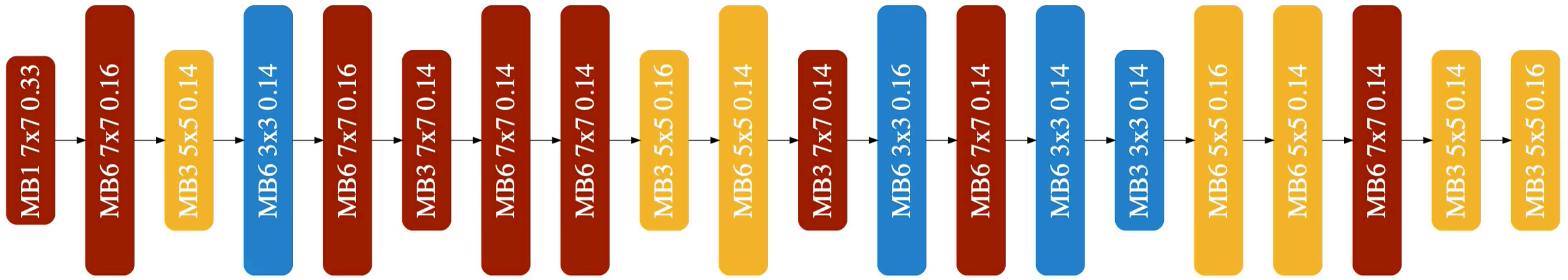
ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
https://arxiv.org/pdf/1812.00332.pdf

Ours: Efficiently Search a Model, only $400



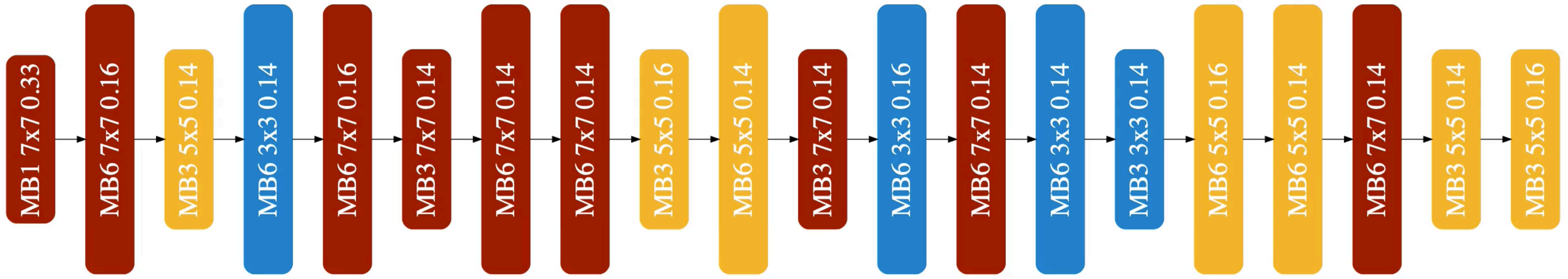
ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware, ICLR'19
https://arxiv.org/pdf/1812.00332.pdf

the Search History on Different HW



Epoch-00

The search history of finding efficient CPU model

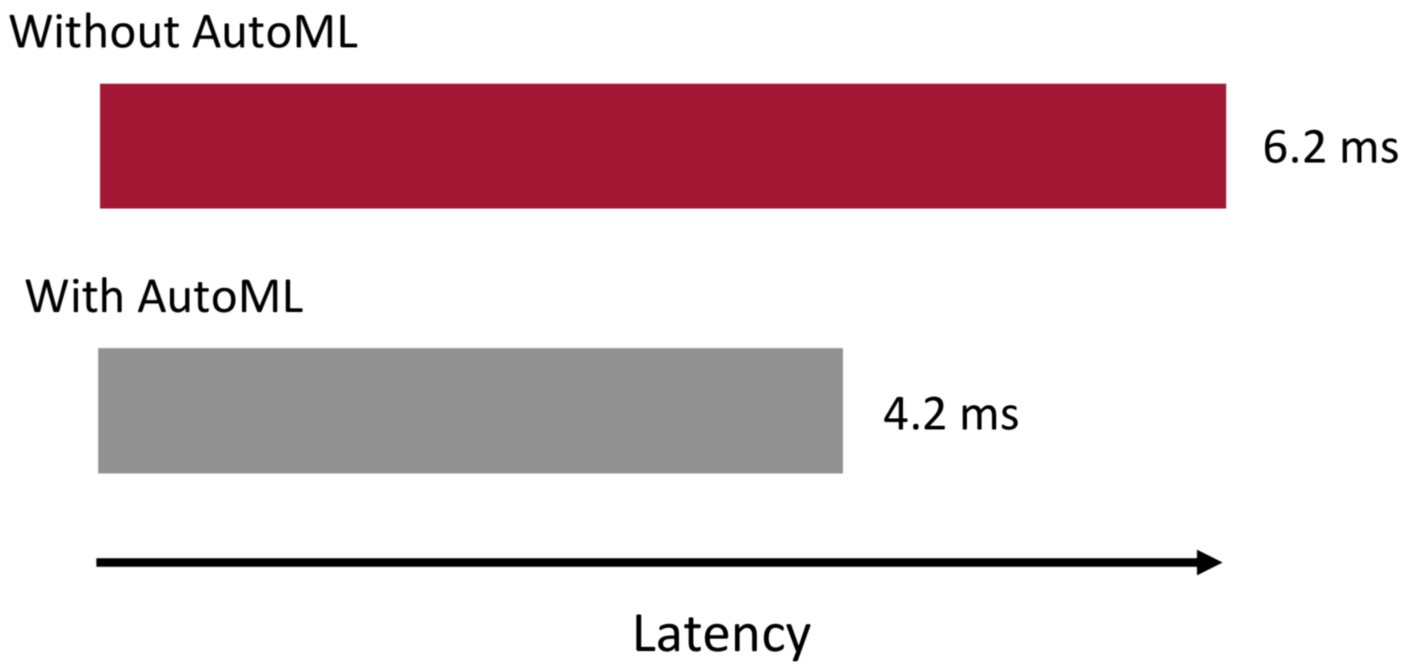


The search history of finding efficient GPU model

https://drive.google.com/file/d/1nut1owvACc9yz1ZPqcbqoJLS2XrVPp1Q/view

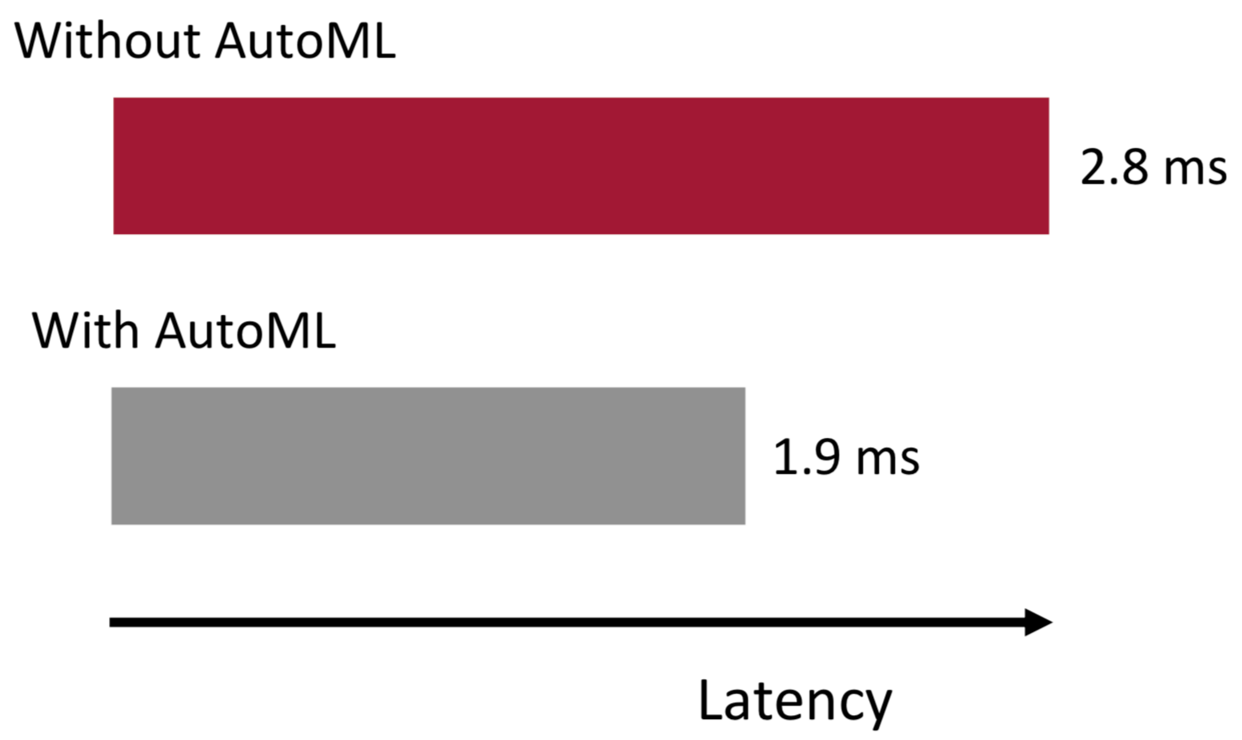
ProxylessNAS: Speedup on Xilinx ZU3 (Ultra 96)

on board measured results



ProxylessNAS: Speedup on Xilinx ZU9 (ZCU102)

on board measured results

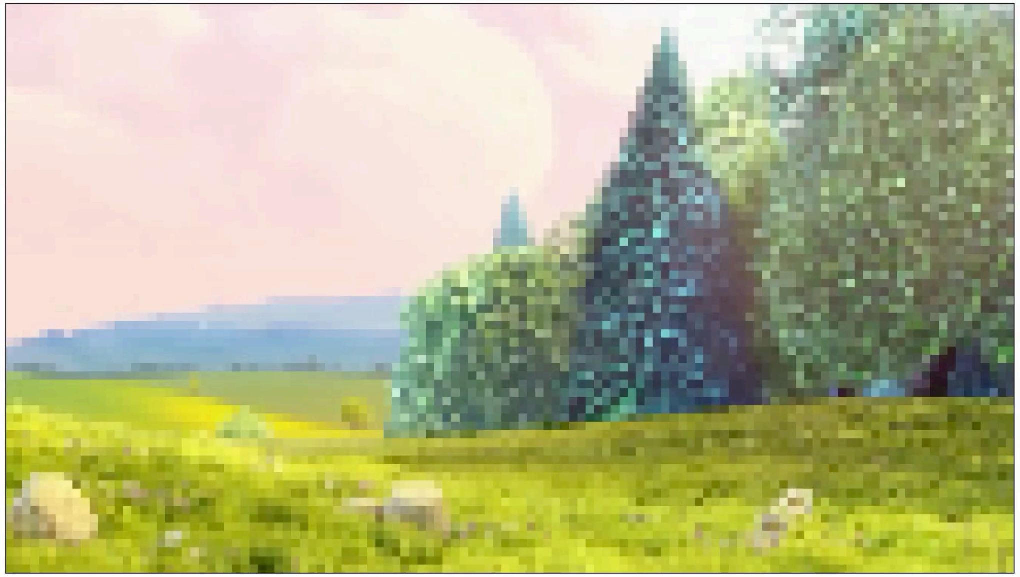


ProxylessNAS: Accelerate Super Resolution

CARN
[ECCV'18]
==>

CARN 16FPS PNSR:21.09

180=>720



AutoML
==>

Ours 41FPS PNSR:21.26

ProxylessNAS in Industry

- Amazon: landed in AutoGluon [1]
- Facebook: landed in PytorchHub [2]

[1] http://autogluon.mxnet.io.s3.amazonaws.com/tutorials/nas/enas_mnist.html
[2] https://pytorch.org/hub/pytorch_vision_proxylessnas/

mit-han-lab / ProxylessNAS

Unwatch 60 | Unstar 901 | Fork 169

Code | Pull requests 1 | Actions | Wiki | Security | Insights | Settings

[ICLR 2019] ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. https://arxiv.org/abs/1812.00332 Edit

automl specialization hardware-aware acceleration on-device-ai efficient-model Manage topics

36 commits | 1 branch | 0 packages | 0 releases | 3 contributors | Apache-2.0

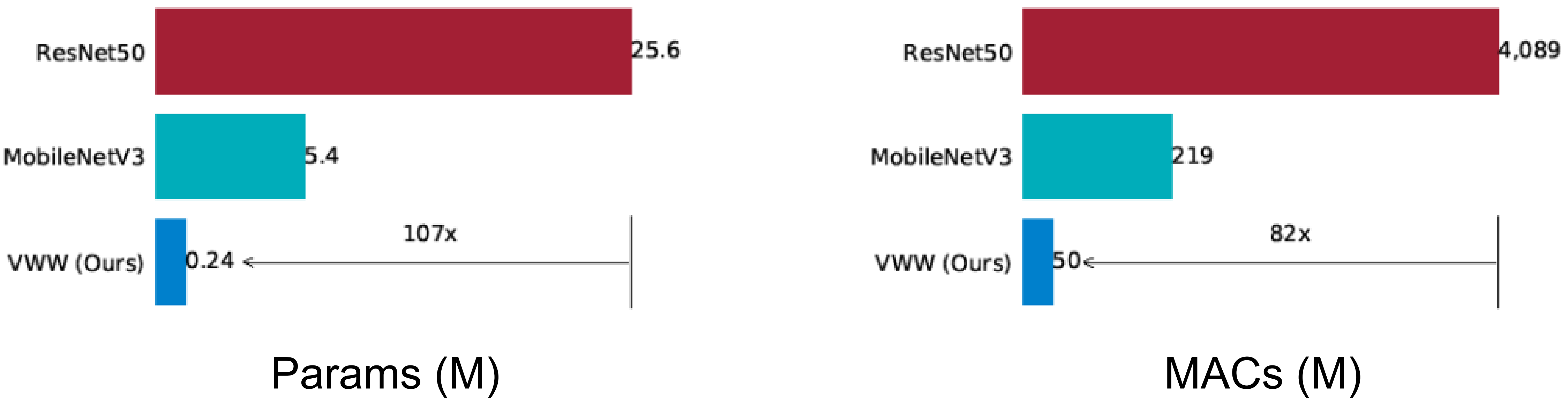
Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

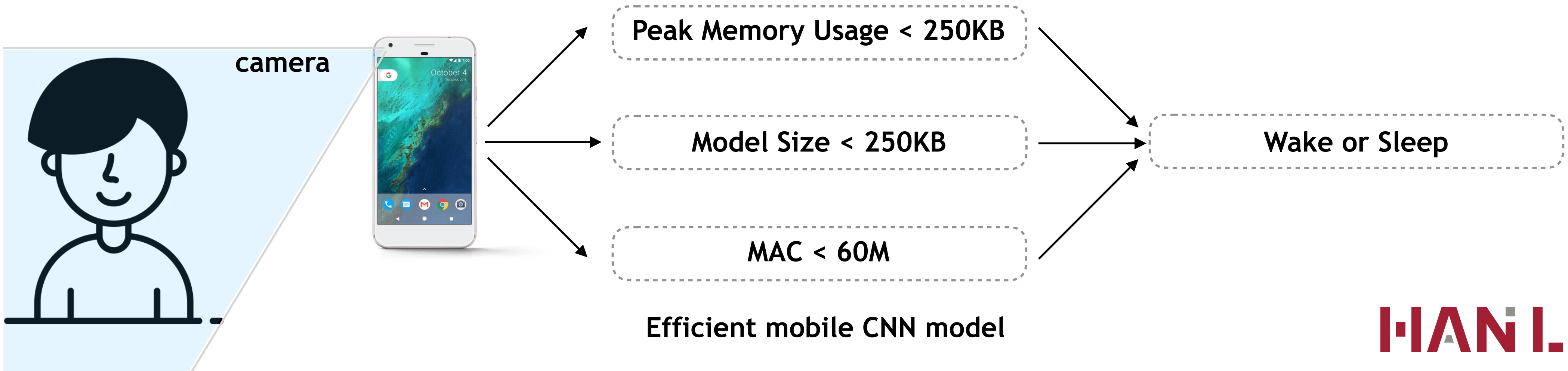
Lyken17 Update README.md — Latest commit 6ebb9b2 25 days ago

logs	Delete proxyless_mobile_10_latency.txt	7 months ago
proxyless_nas	release training code	4 months ago
proxyless_nas_tensorflow	Update tf_model_zoo.py	6 months ago
search	add search code	3 months ago
training	Update main.py	4 months ago
.gitignore	prepare conf file for pytorch hubs.	2 months ago
LICENSE	public release	last year
README.md	Update README.md	25 days ago
eval.py	reformat with pep8	8 months ago
eval_tf.py	reformat with pep8	8 months ago
hubconf.py	updated relative import	2 months ago

README.md

Visual Wake Words Challenge using ProxylessNAS





Model Size < 250 KB

Peak Memory < 250 KB

MACs < 60 M

Challenge: Deploying Deep Learning Models on Diverse Hardware Platforms and Efficiency Constraints

Diverse Mobile Platforms

Galaxy S10, 2019

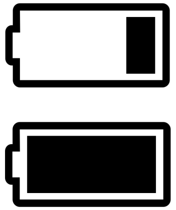
Galaxy S8, 2017

Galaxy S6, 2015

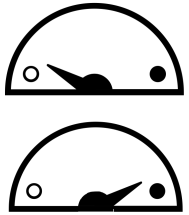
Galaxy S4, 2013

…

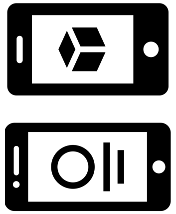
Diverse Efficiency Constraints



battery/energy



workloads/latency



application

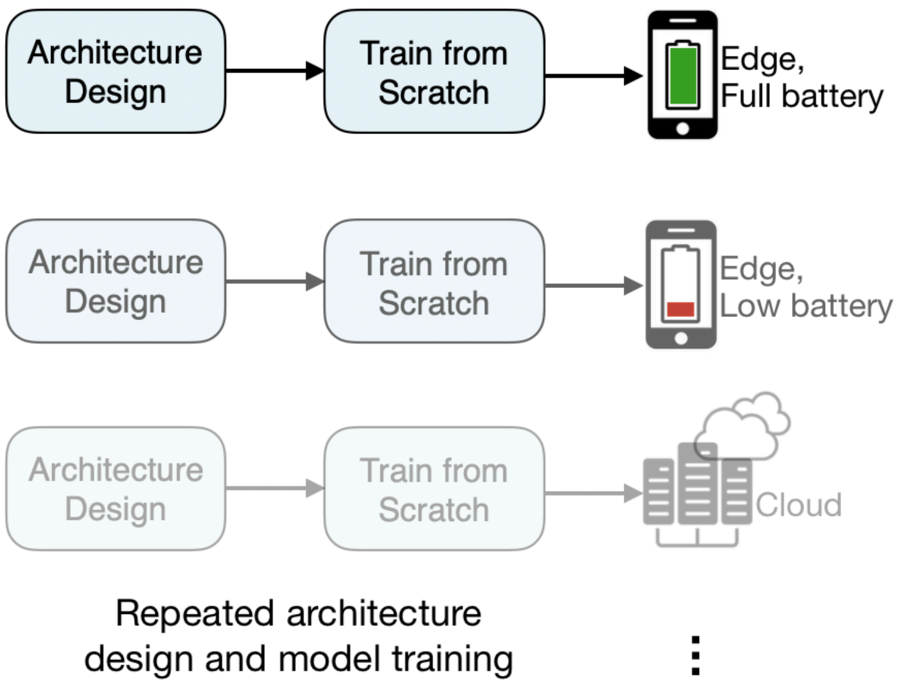
…

- A mobile Application has to support both a 2013 phone and a 2019 phone. Often a headache!

Efficient Deep Learning on the Edge

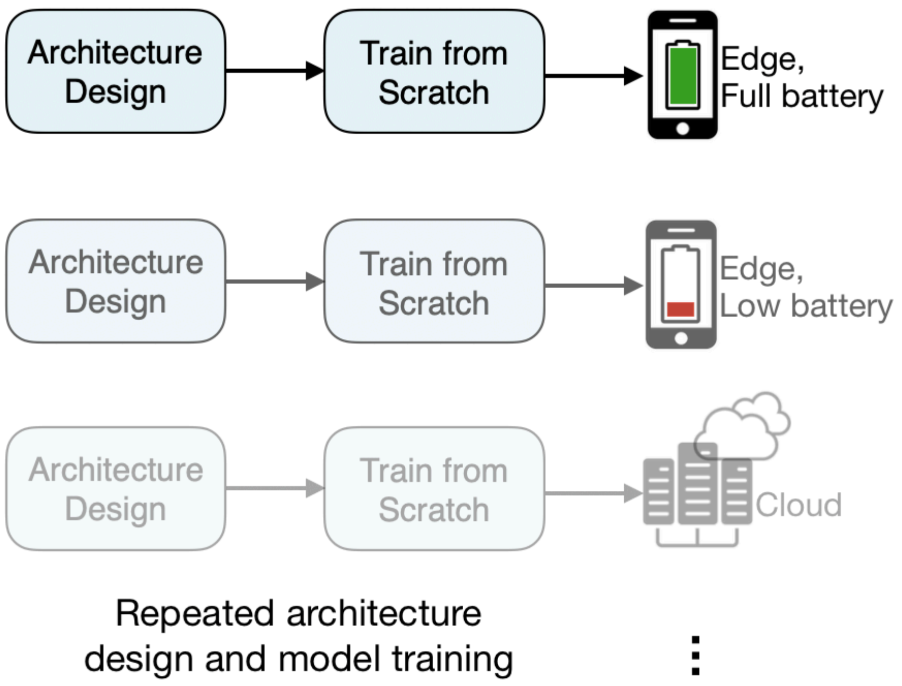
- **Efficient 3D Algorithms:**
  - PVCNN for efficient point-cloud recognition [NeurIPS'19, spotlight]
  - TSM for efficient video recognition [ICCV'19]
- **Compression / NAS**
  - Deep Compression [NIPS'15, ICLR'16]
  - ProxylessNAS, AMC, HAQ [ICLR'19, ECCV'18, CVPR'19, oral]
  - **Once-For-All (OFA) Network**

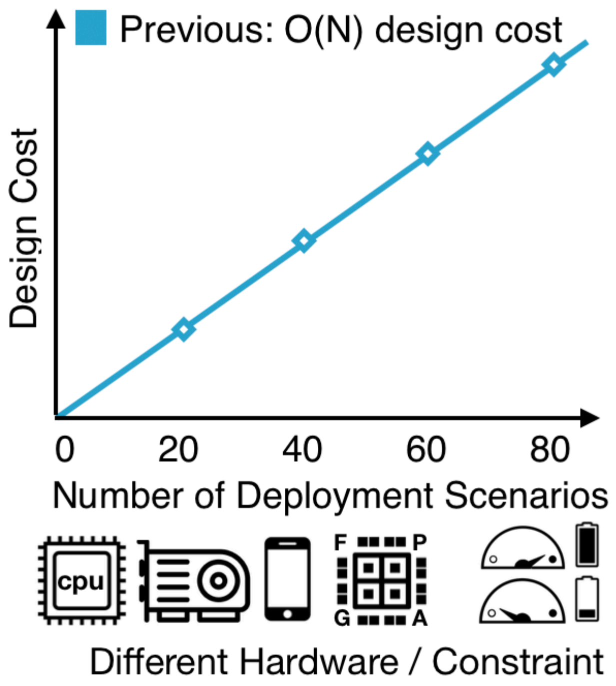
Traditional NAS Approaches: Expensive and Unscalable



- Traditional approaches **repeat** the architecture design process and **retrain** the specialized model from scratch for each case

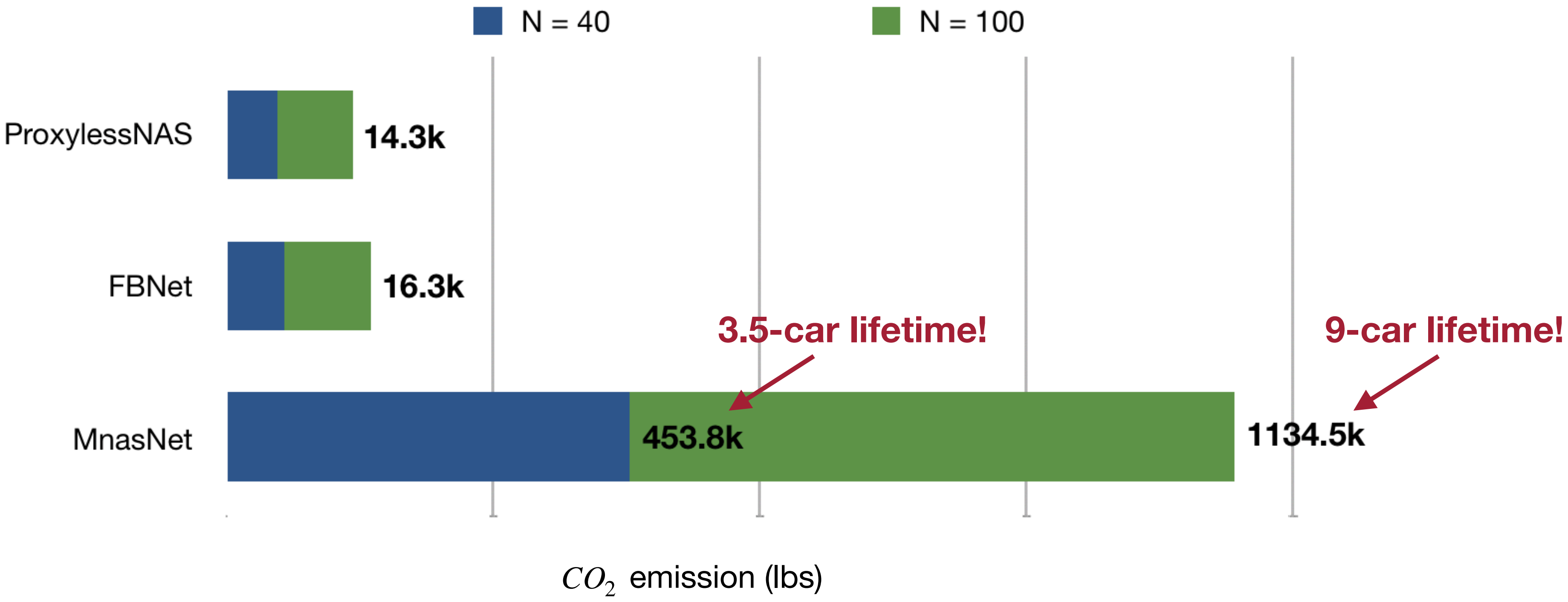
Traditional NAS Approaches: Expensive and Unscalable





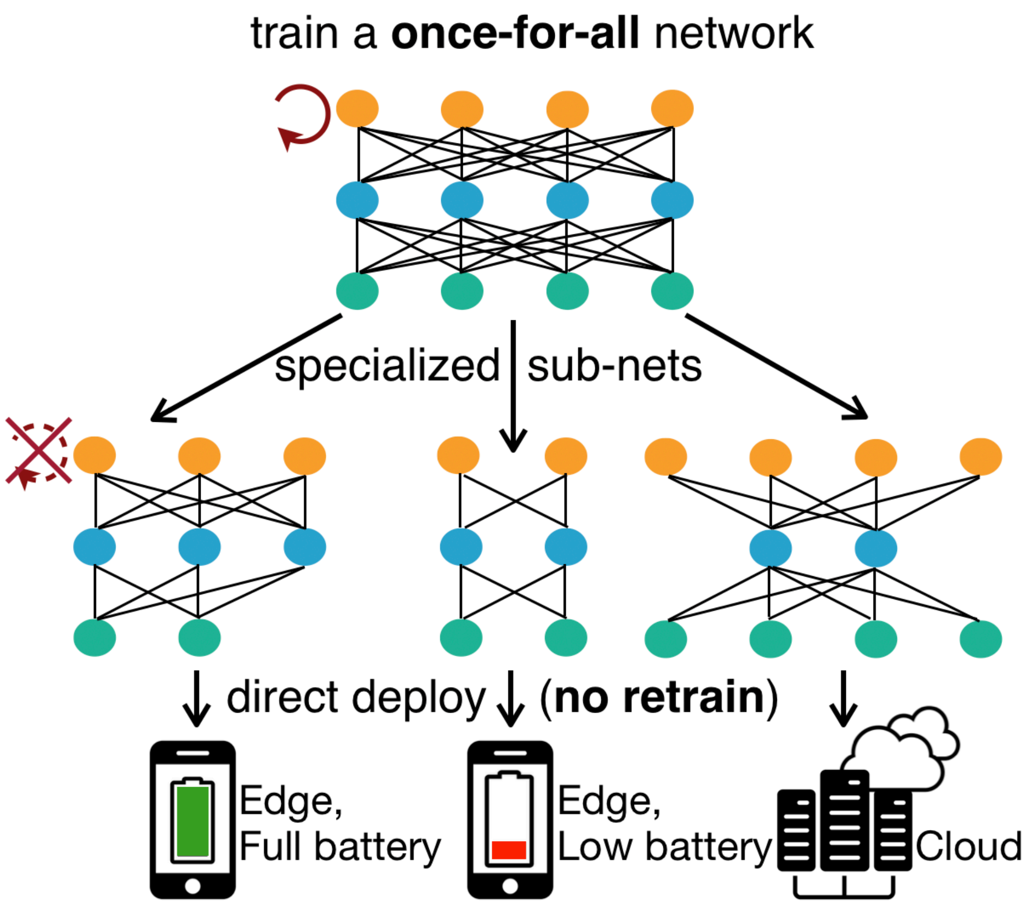
- Traditional approaches **repeat** the architecture design process and **retrain** the specialized model from scratch for each case
- The total cost **grows linearly** as the number of deployment scenarios increases

Traditional NAS Approaches: Expensive and Unscalable



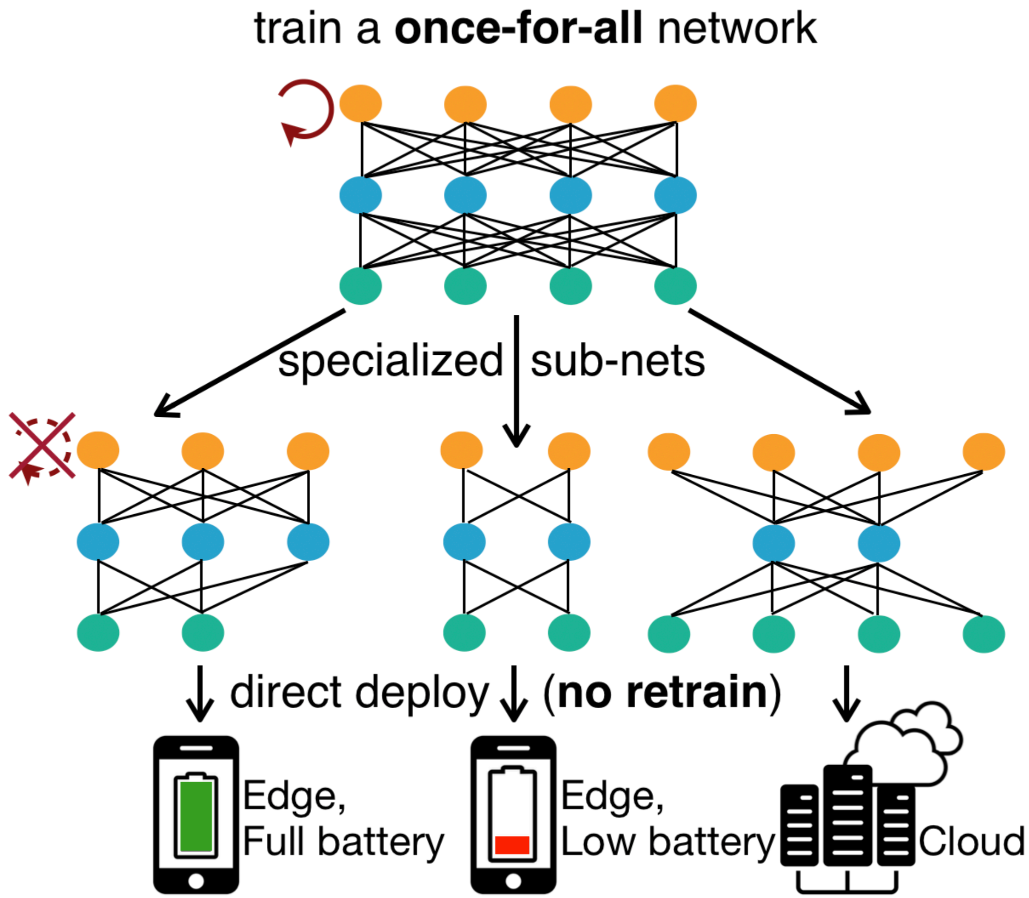
- Excessive CO_2 emission, causing severe environmental problems

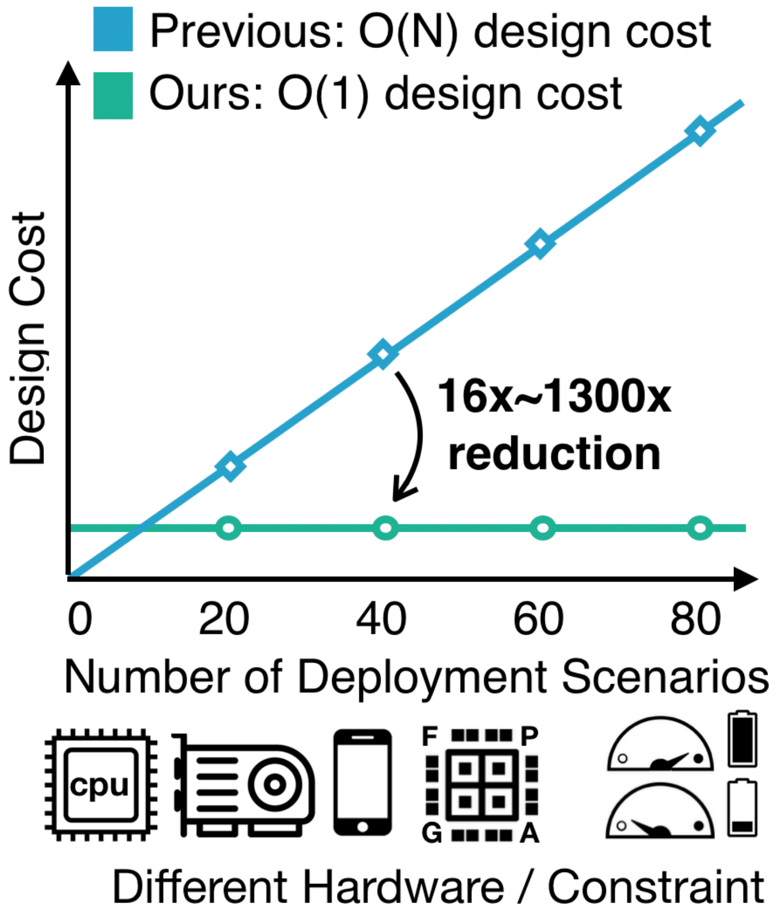
Once-for-All Network: Decouple Model Training and Architecture Design, O(1) Cost



- We introduce **Once for All (OFA)** to tackle the challenge of deep learning deployment on many hardware and constraints
- In OFA, **model training is decoupled from architecture search**
    - A **single OFA network** is trained to support **all** architectural configurations in the search space
    - Specialized sub-networks are directly derived from the OFA network without retraining

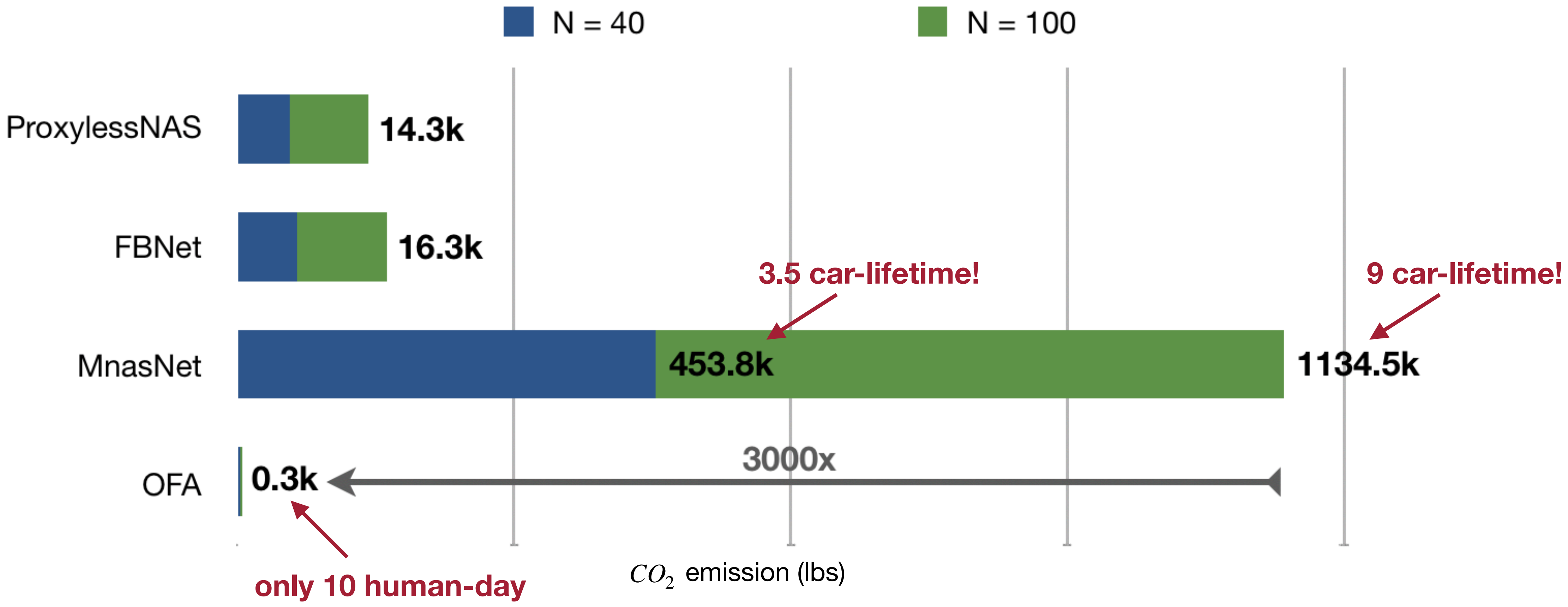
Once-for-All Network: Decouple Model Training and Architecture Design, O(1) Cost





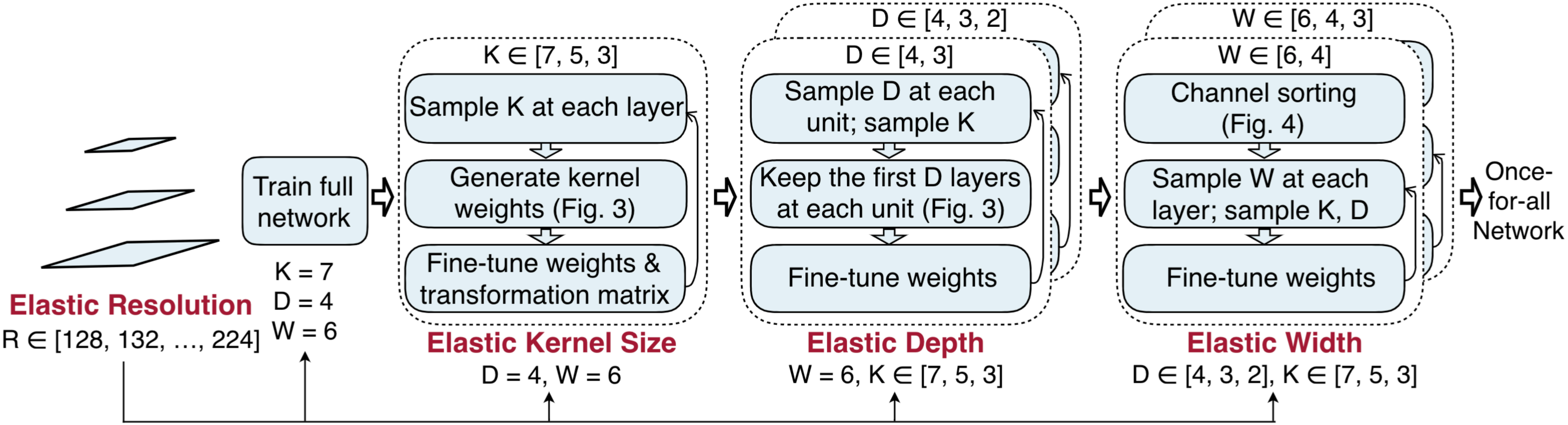
- We introduce **Once for All (OFA)** to tackle the challenge of deep learning deployment on many hardware and constraints
- In OFA, **model training is decoupled from architecture search**
  - A **single OFA network** is trained to support **all** architectural configurations in the search space
  - Specialized sub-networks are directly derived from the OFA network without retraining. O(1) cost.

Once for All: Decouple Model Training and Architecture Design, O(1) Cost

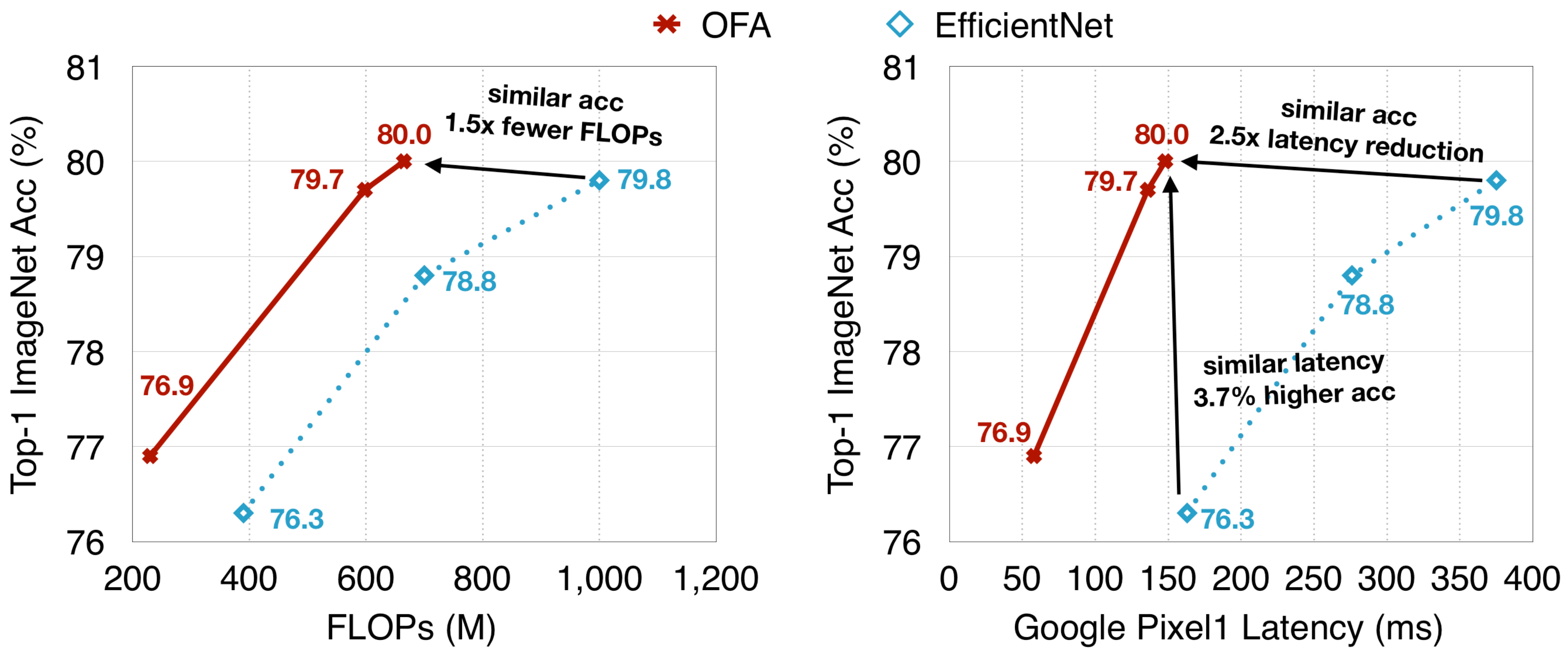


- Excessive CO_2 emission, causing severe environmental problems

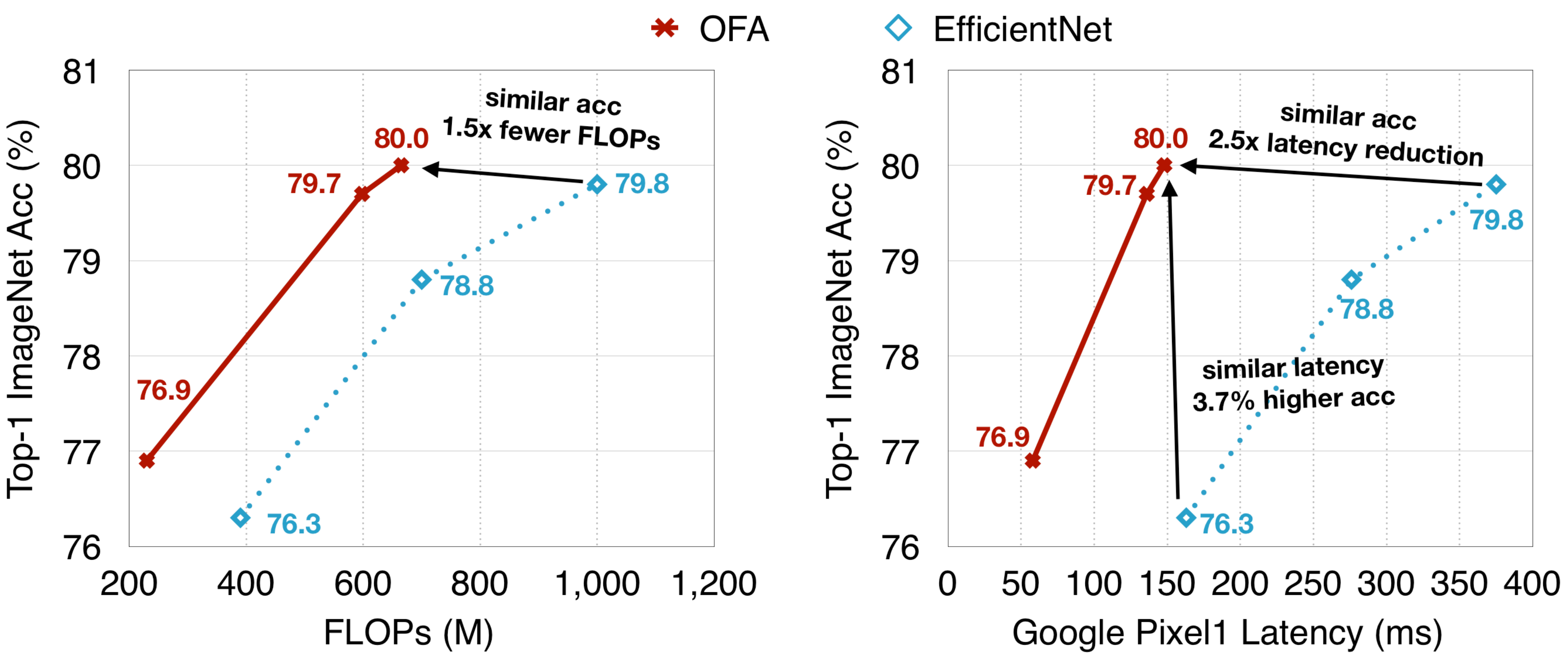
Progressive Shrinking for Training OFA Networks



OFA: 80% Top-1 Accuracy on ImageNet
Outperforms EfficientNet by a Large Margin



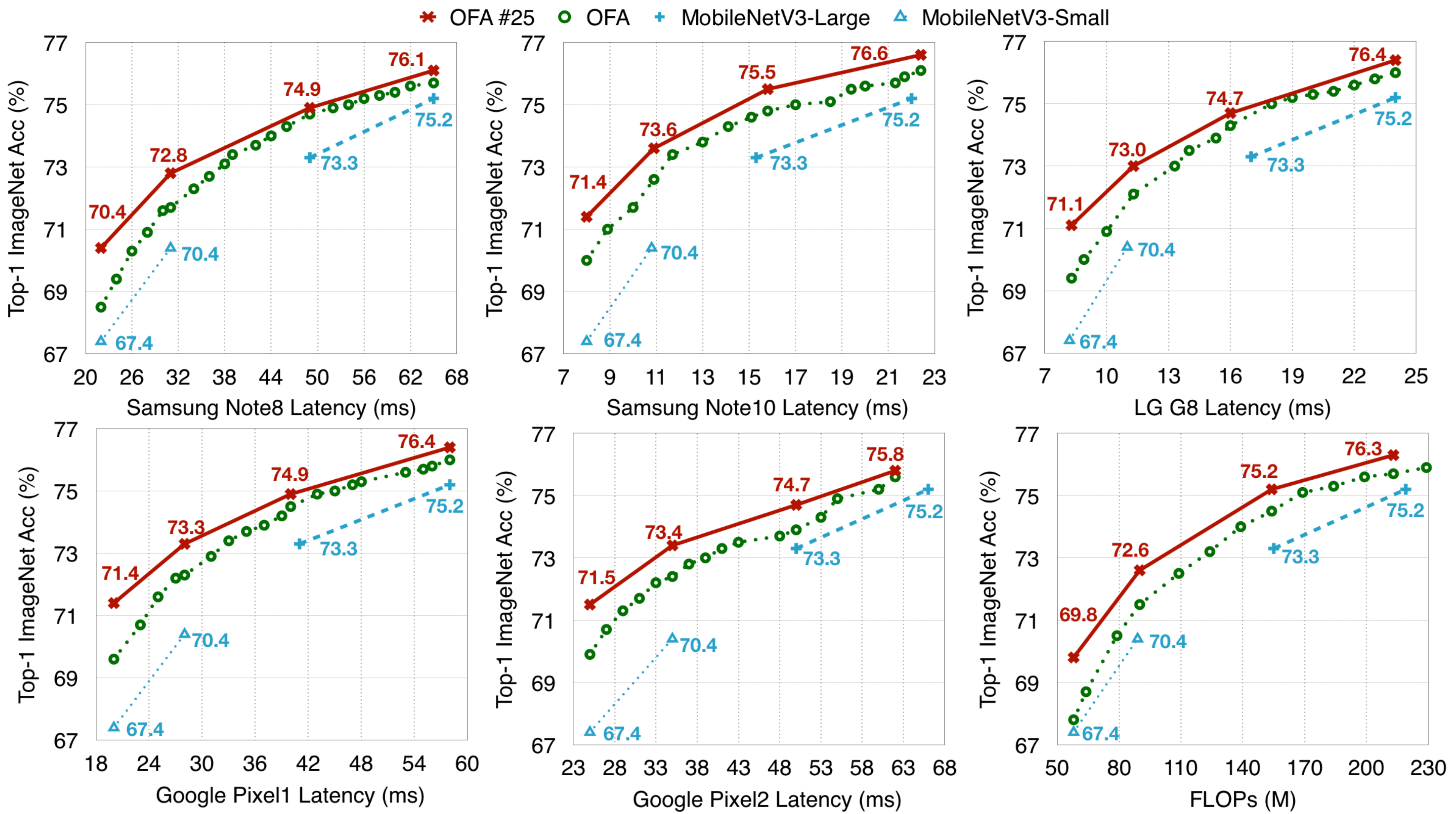
OFA: 80% Top-1 Accuracy on ImageNet
Outperforms EfficientNet by a Large Margin



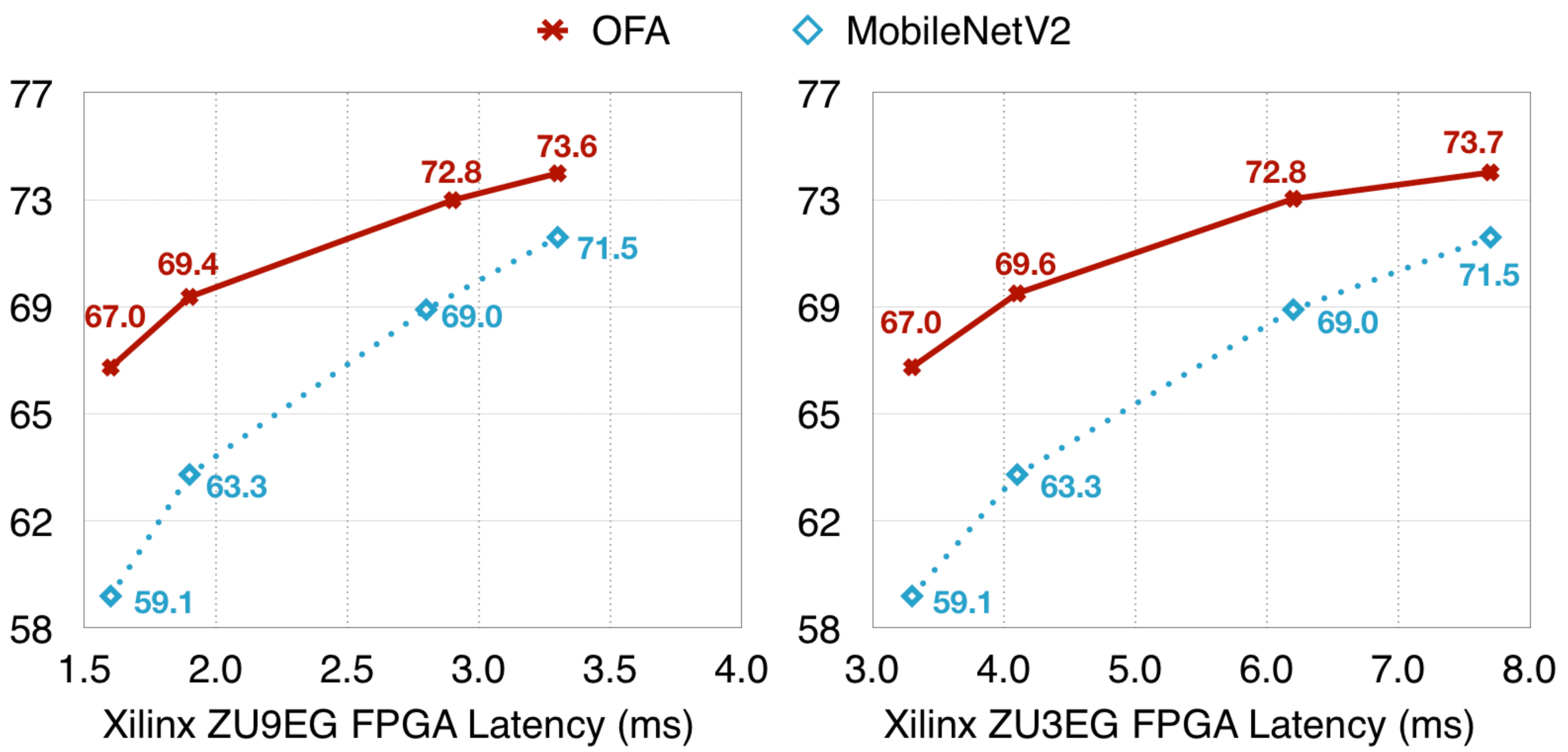
~~EfficientNet: initial learning rate 0.256 that decays by 0.97 every 2.4 epochs~~

OFA Enables Fast Specialization

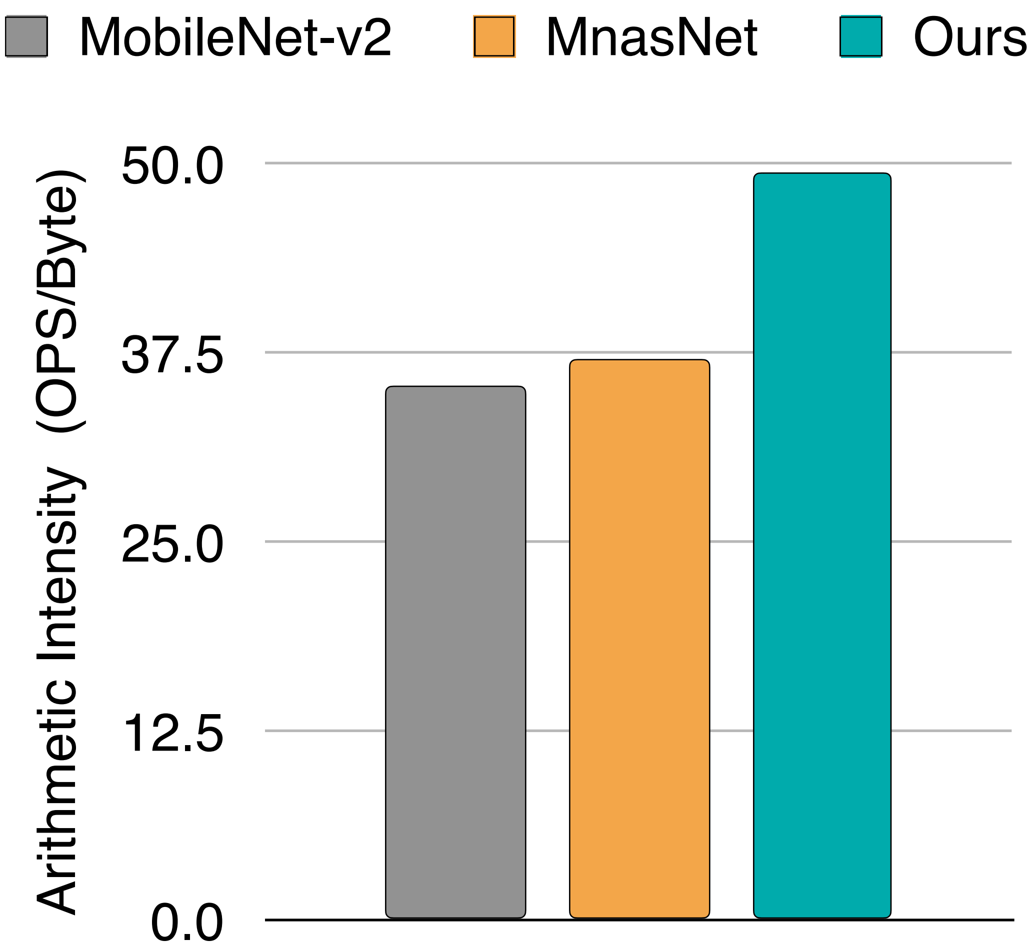
Outperforms MobileNet-v3 by a Large Margin across Many Devices



OFA for FPGA Accelerators



OFA: Higher Arithmetic Intensity on FPGA

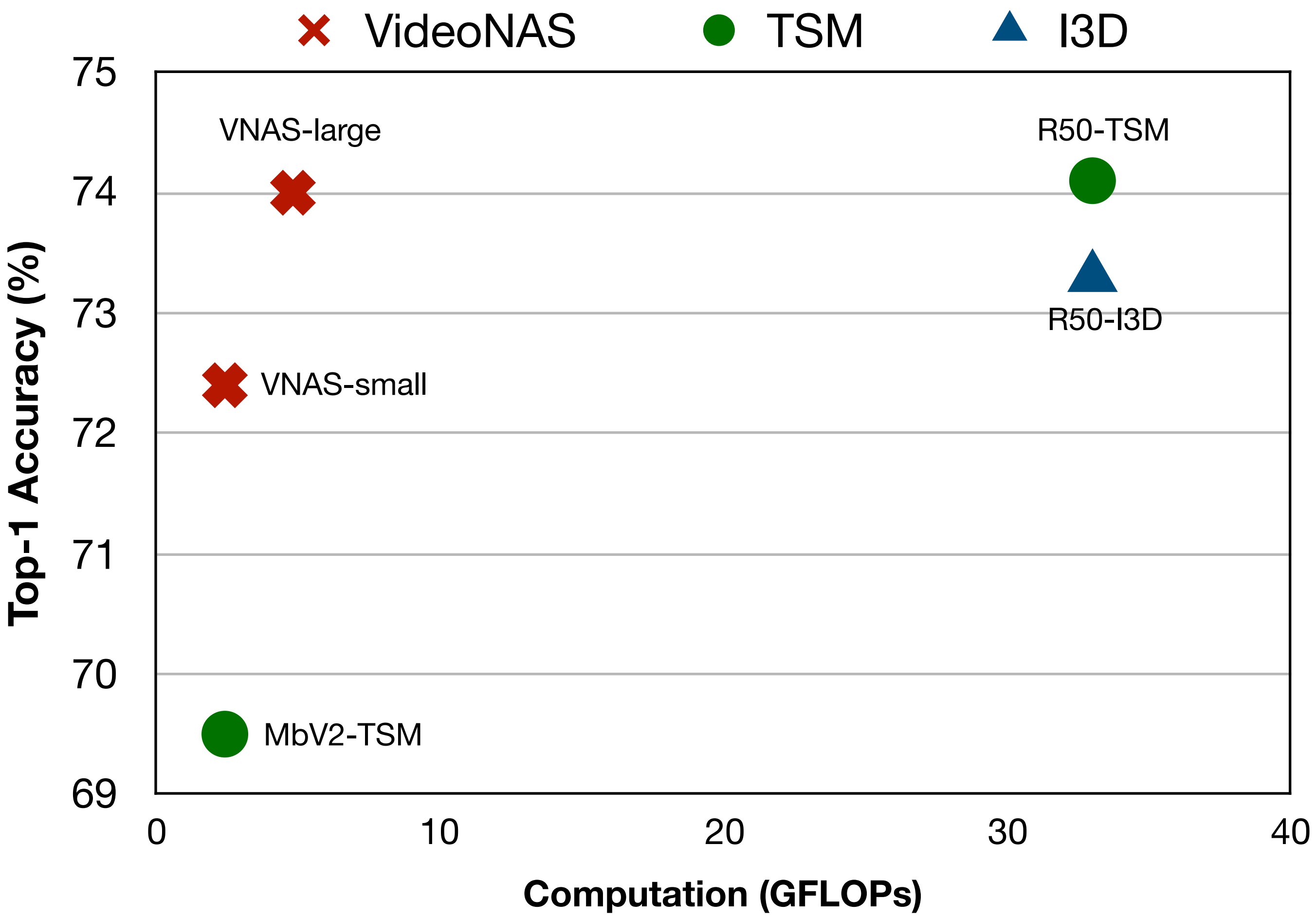


on board measured results

OFA for Video?

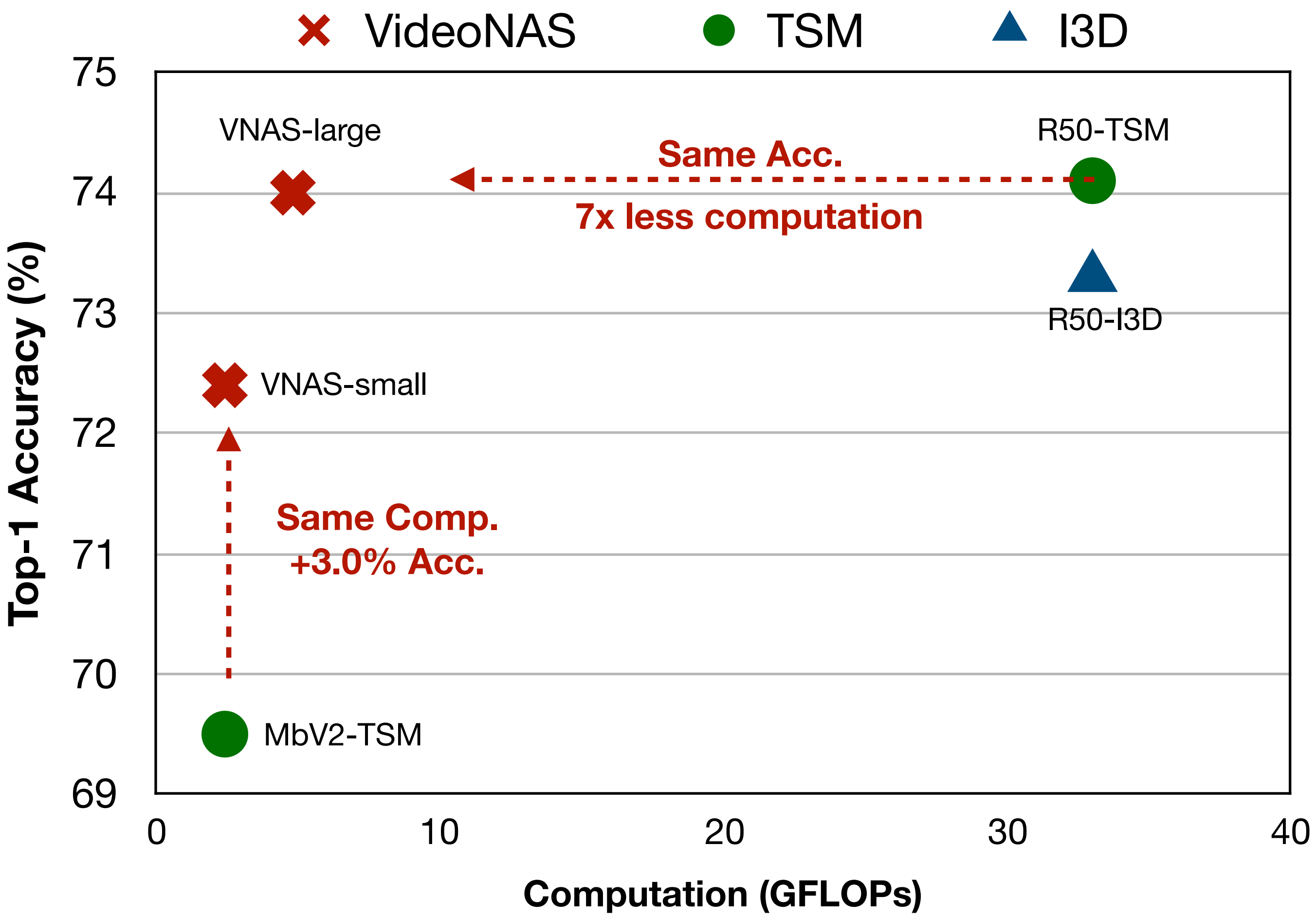
OFA for Video TSM: VideoNAS

- Experiments on Kinetics dataset (the mostly used, largest benchmark)



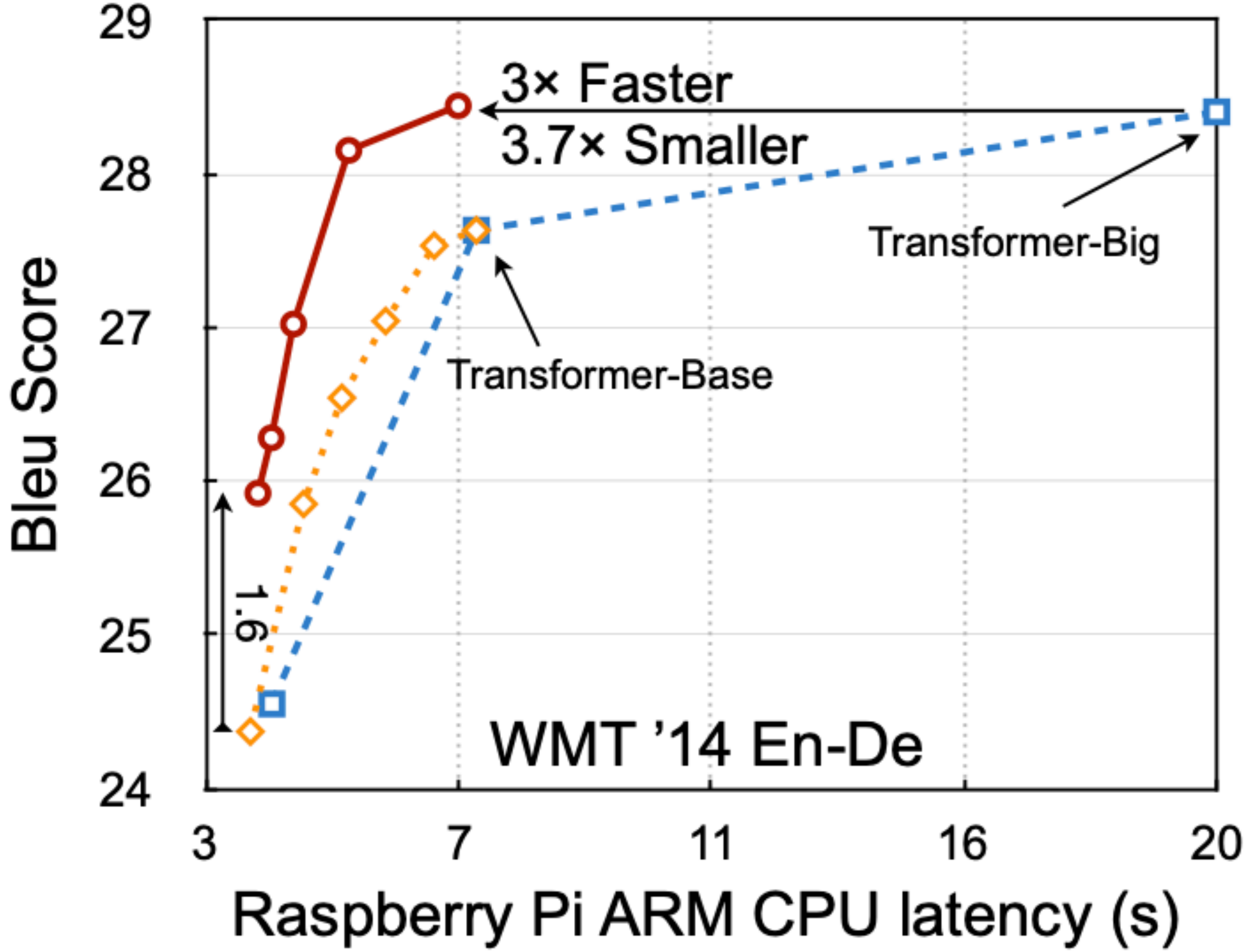
OFA for Video TSM: VideoNAS

- Experiments on Kinetics dataset (the mostly used, largest benchmark)

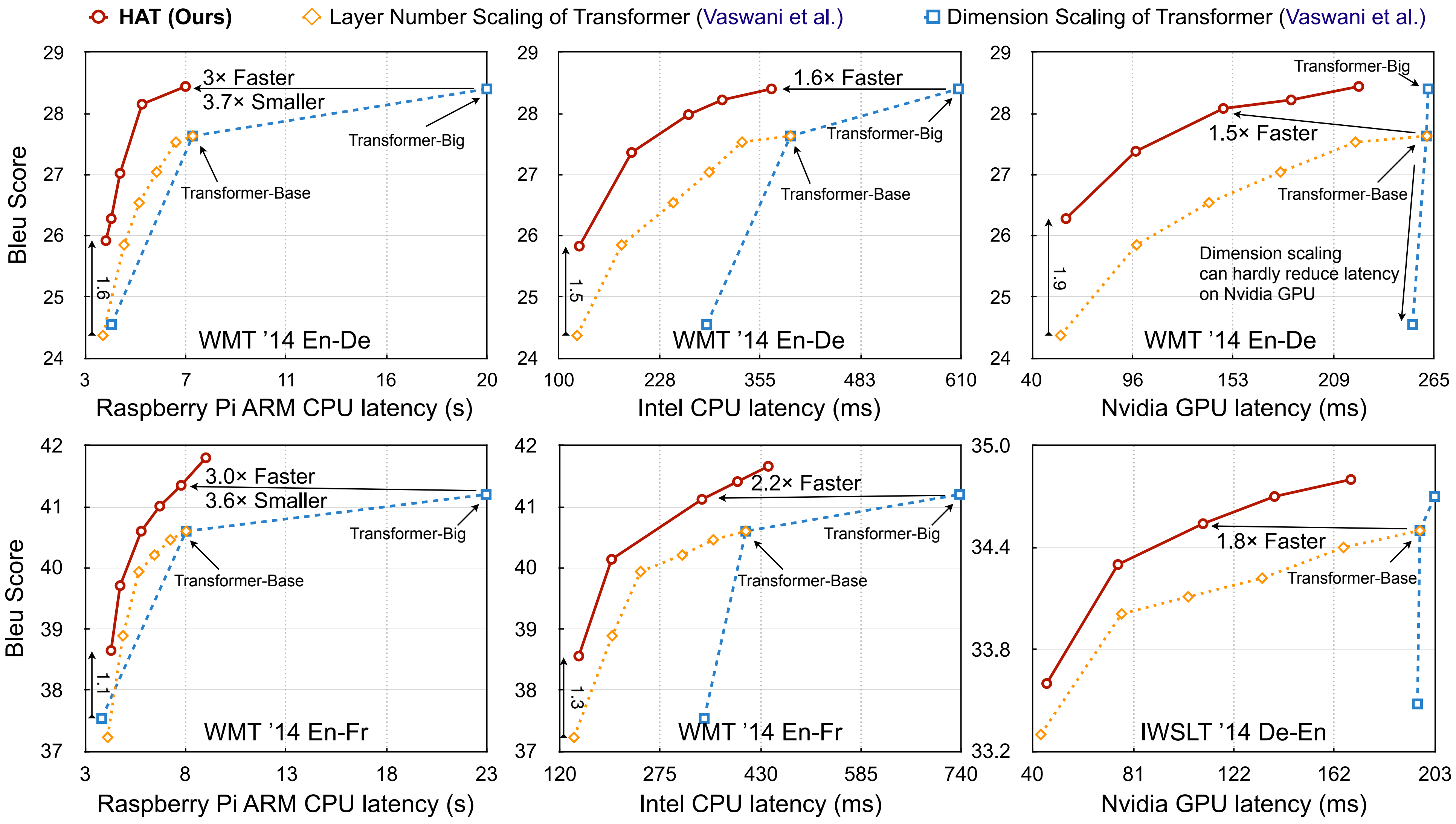


OFA for NLP?

HAT: Hardware-Aware Transformer



HAT: Hardware-Aware Transformer

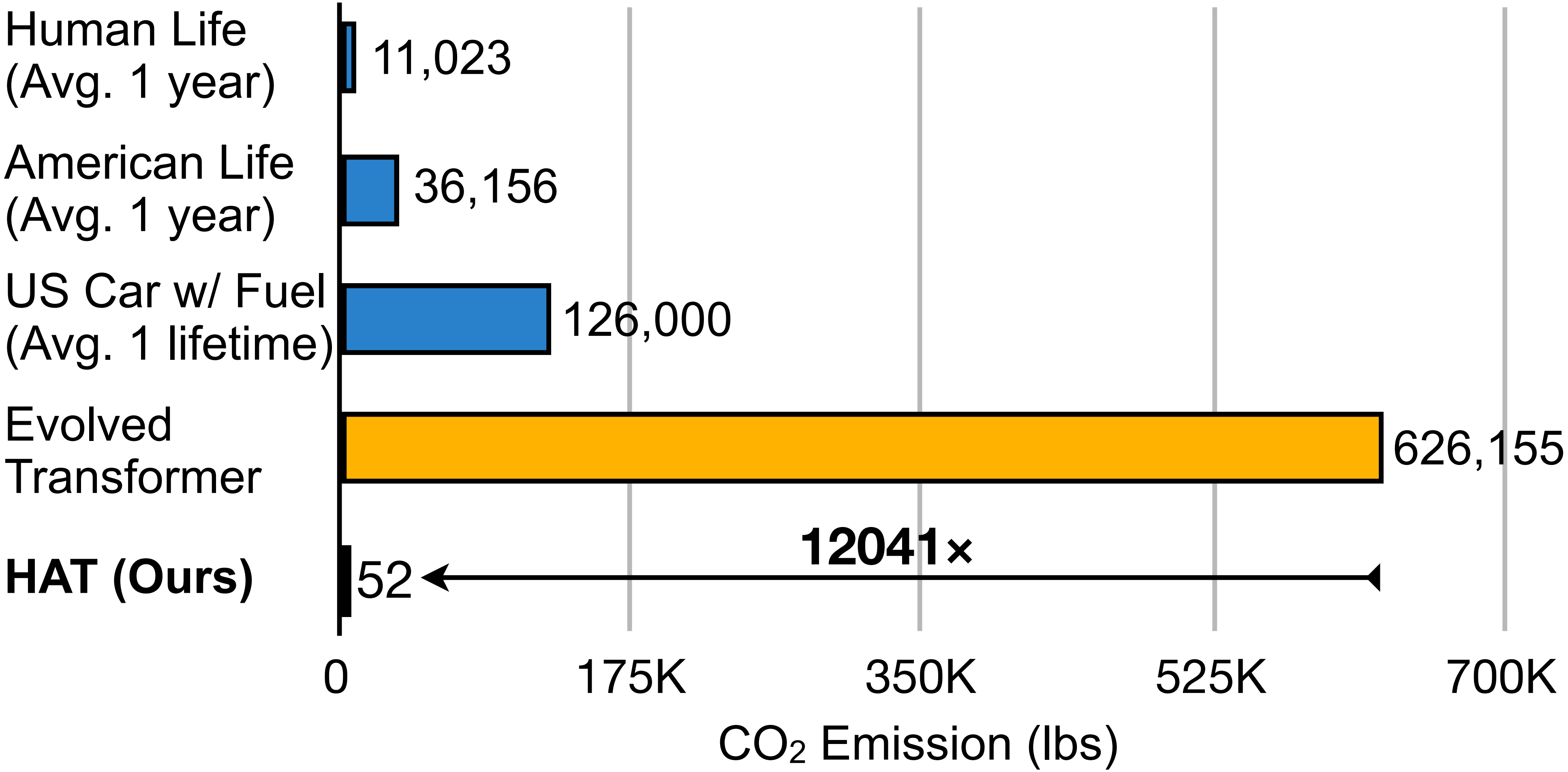


On WMT'14 En-De, same performance, **3.7x** smaller model size;
3x, 1.6x, 1.5x faster on Raspberry Pi, CPU, GPU, respectively than Transformer Baseline

HAT: Hardware-Aware Transformer

		Hardware-Aware	Hetero. Layers	Latency	#Params	BLEU	GPU Hours	CO_2e (lbs)	Cloud Computation Cost
WMT'14 En-Fr	Transformer	✗	✗	23.2s	176M	41.2	8×30	68	\$178 - \$595
	Evolved Transformer	✗	✗	20.9s	175M	41.3	$8\times274K$	626K	\$1.6M - \$5.5M
	HAT (Ours)	✓	✓	**7.8s**	**48M**	**41.4**	$8\times(13+14)$	61	\$159 - \$534
	HAT (Ours)	✓	✓	9.1s	57M	**41.8**	$8\times(13+15)$	64	\$166 - \$555

HAT is Environmental Friendly



CO_2 Emission of HAT training is only **52** pounds, while that of Evolved Transformer is **626,155** pounds

HAT is Quantization Friendly

	BLEU	Model Size	Reduction
Transformer Float32	41.2	705MB	–
HAT Float32	41.8	227MB	3×
HAT 8 bits	**41.9**	**57MB**	**12×**
HAT 4 bits	41.1	**28MB**	**25×**

NeurIPS MicroNet Challenge (NLP Track)

	Sparsity	Quantization	Test Perplexity	Score
Model 1	42.12%	9 bits	34.95	0.0482
Model 2	40.12%	9 bits	34.65	0.0485
Model 3	33.85%	8 bits	34.95	**0.0475**

Winning 1st place in the NeurIPS MicroNet Challenge

Resource is Limited.

We need Once-For-All.

Efficient Deep Learning on the Edge

- **Efficient 3D Algorithms:**
  - PVCNN for efficient point-cloud recognition [NeurIPS'19, spotlight]
  - TSM for efficient video recognition [ICCV'19]
- **Compression / NAS**
  - Deep Compression [NIPS'15, ICLR'16]
  - ProxylessNAS, AMC, HAQ [ICLR'19, ECCV'18, CVPR'19, oral]
  - Once-For-All (OFA) Network

Make AI Efficient

Any Human Resource

Any Computational Resource

MIT HAN LAB
Hardware, AI and Neural-nets

hanlab.mit.edu
github.com/mit-han-lab