



16TH EUROPEAN CONFERENCE ON  
**COMPUTER VISION**

[WWW.ECCV2020.EU](http://WWW.ECCV2020.EU)

# Hardware-aware Deep Neural Architecture Search

ECCV 2020 Tutorial on  
From HPO to NAS: Automated Deep Learning

Peizhao Zhang

Mobile Vision, Facebook

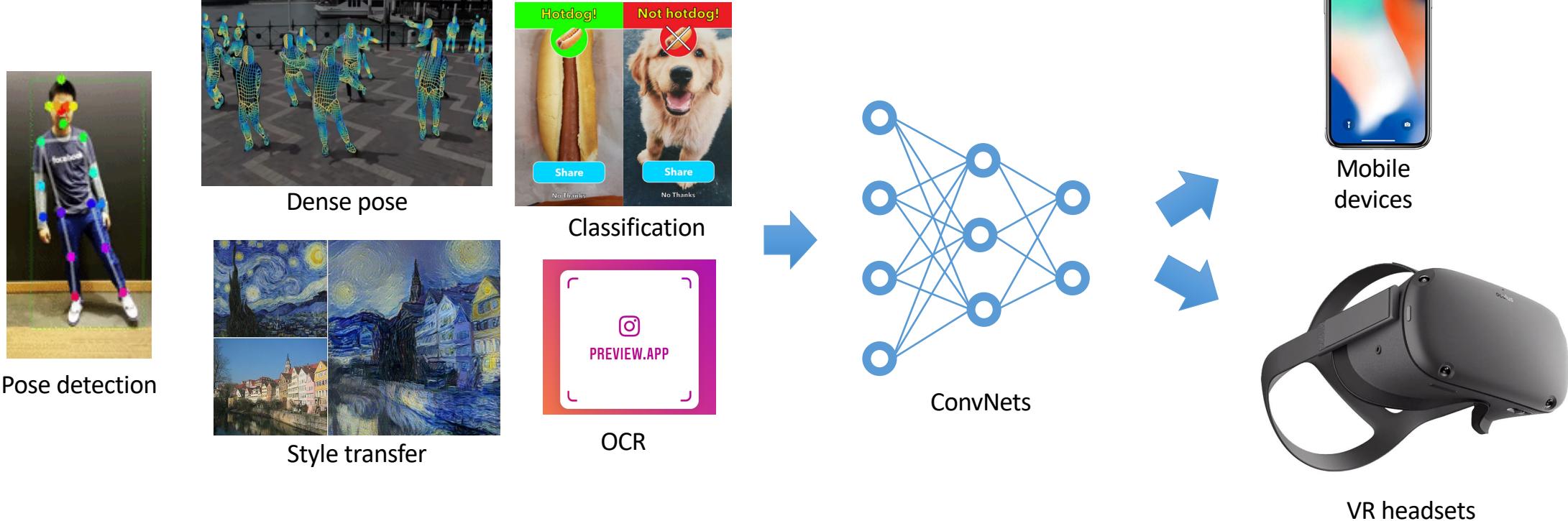
Acknowledge: Bichen Wu, Xiaoliang Dai, Alvin Wan, Peter Vajda

# Outline

- Introduction
- Hardware-aware Architecture Search
  - Differentiable Search: **FBNet** and **FBNetV2**
  - Predictor-based Search: **ChamNet** and **FBNetV3**
- Applications

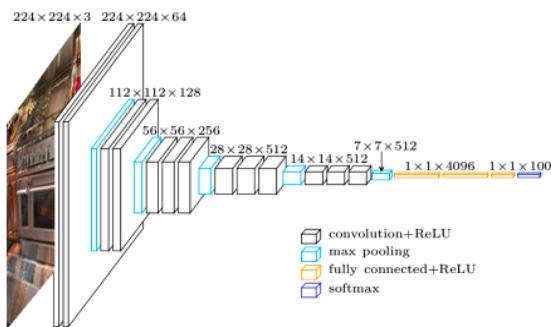
# Goal: Accuracy and Efficiency

- Mobile and embedded computer vision applications require accurate and efficient ConvNets



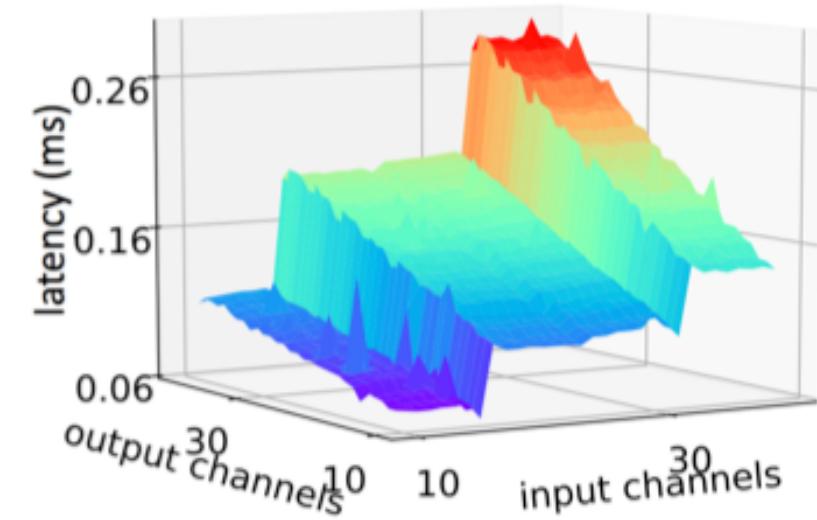
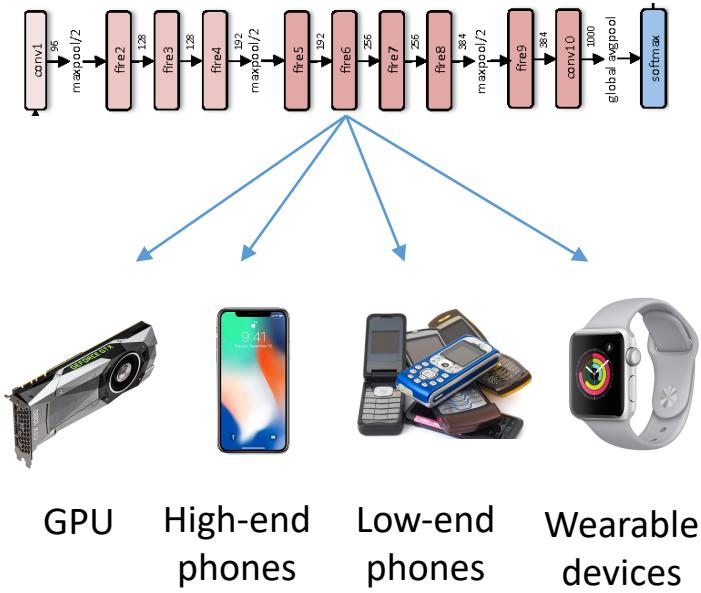
# Challenges

- Intractable design space
- Conditional optimality
- Inaccurate metrics

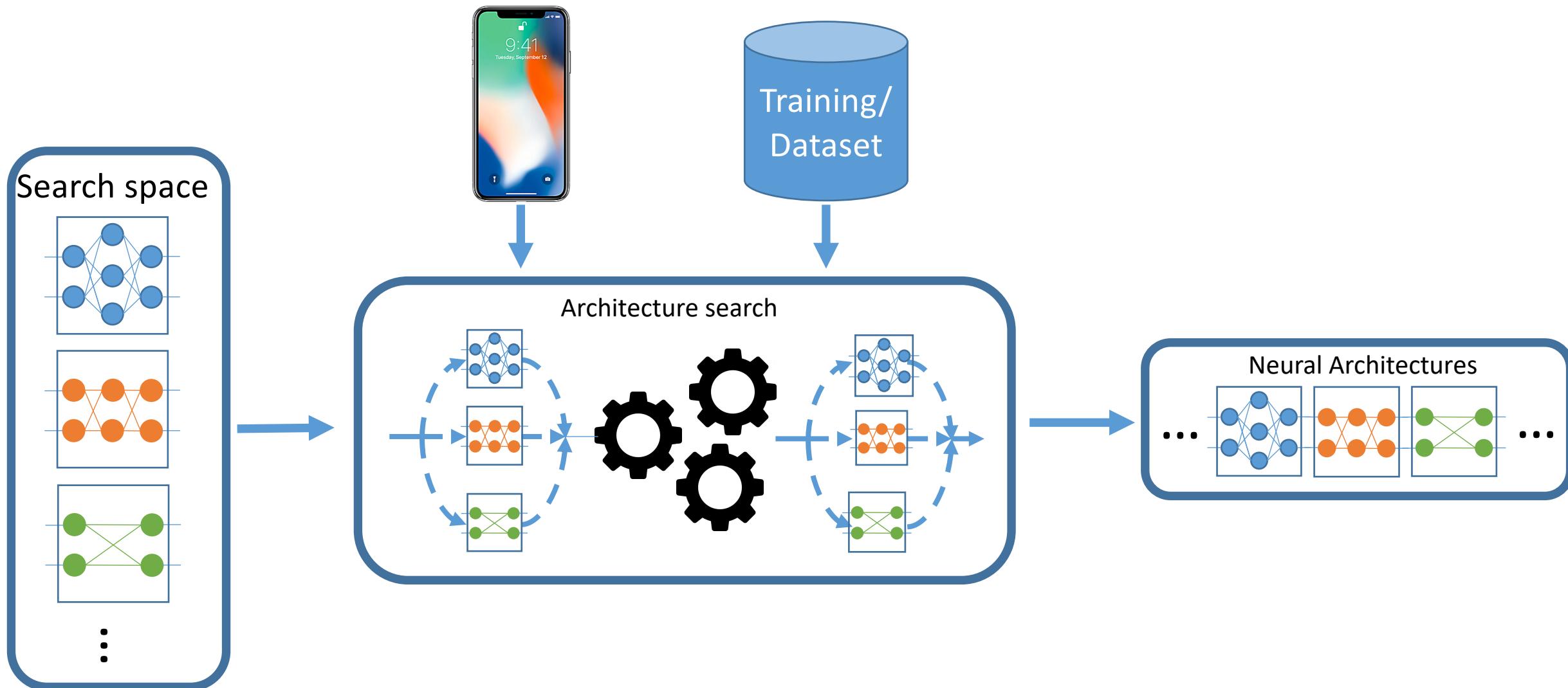


- Kernel size = {1, 3, 5}
- Channel size = {16, 32, 64, 128, 256}

$(3 \times 5)^{13} = 2e15$  possible architectures

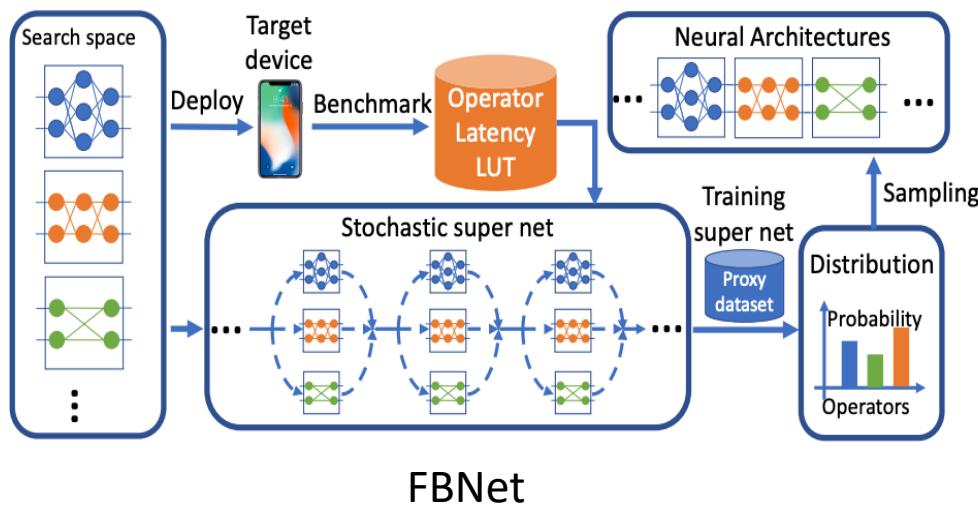


# Hardware-aware Architecture Search

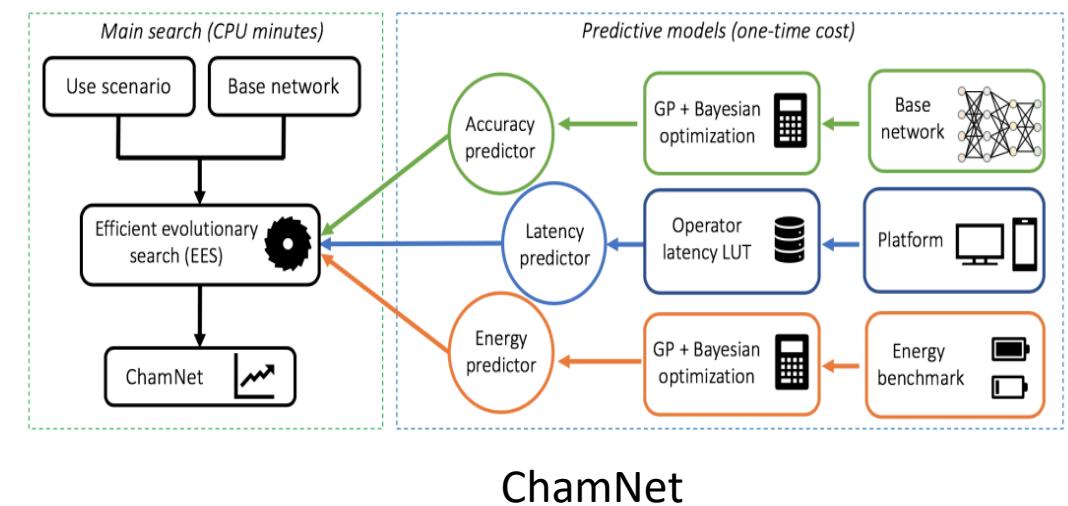


# Hardware-aware Architecture Search

- Differentiable Architecture Search: **FBNet** and **FBNetV2**
- Predictor-based Architecture Search: **ChamNet** and **FBNetV3**



FBNet

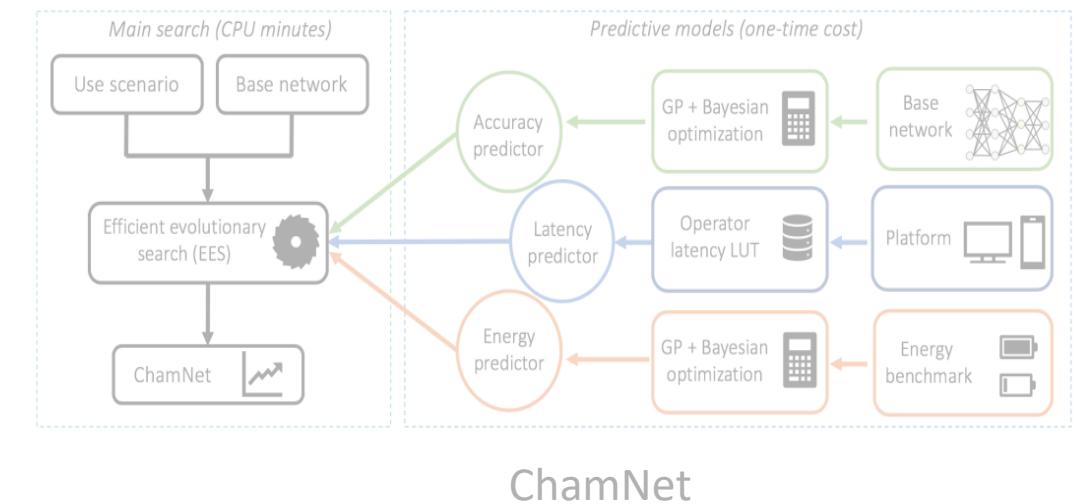
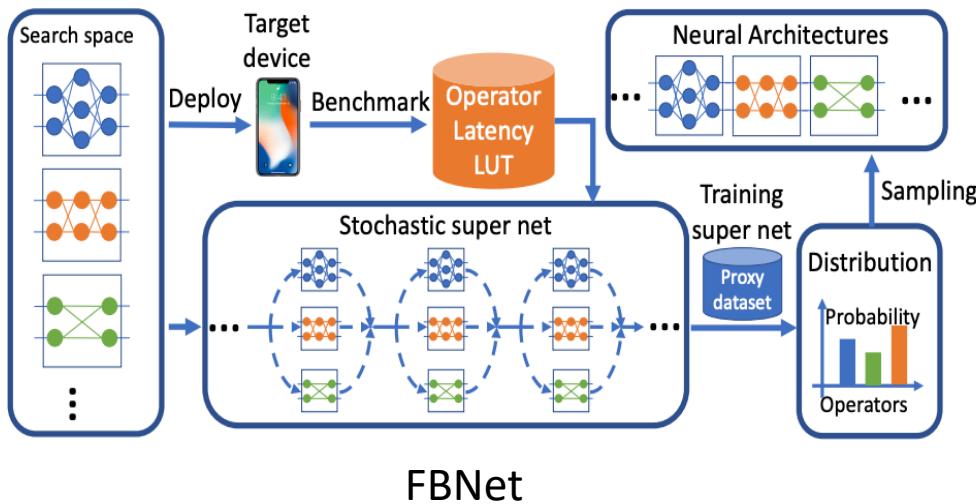


ChamNet

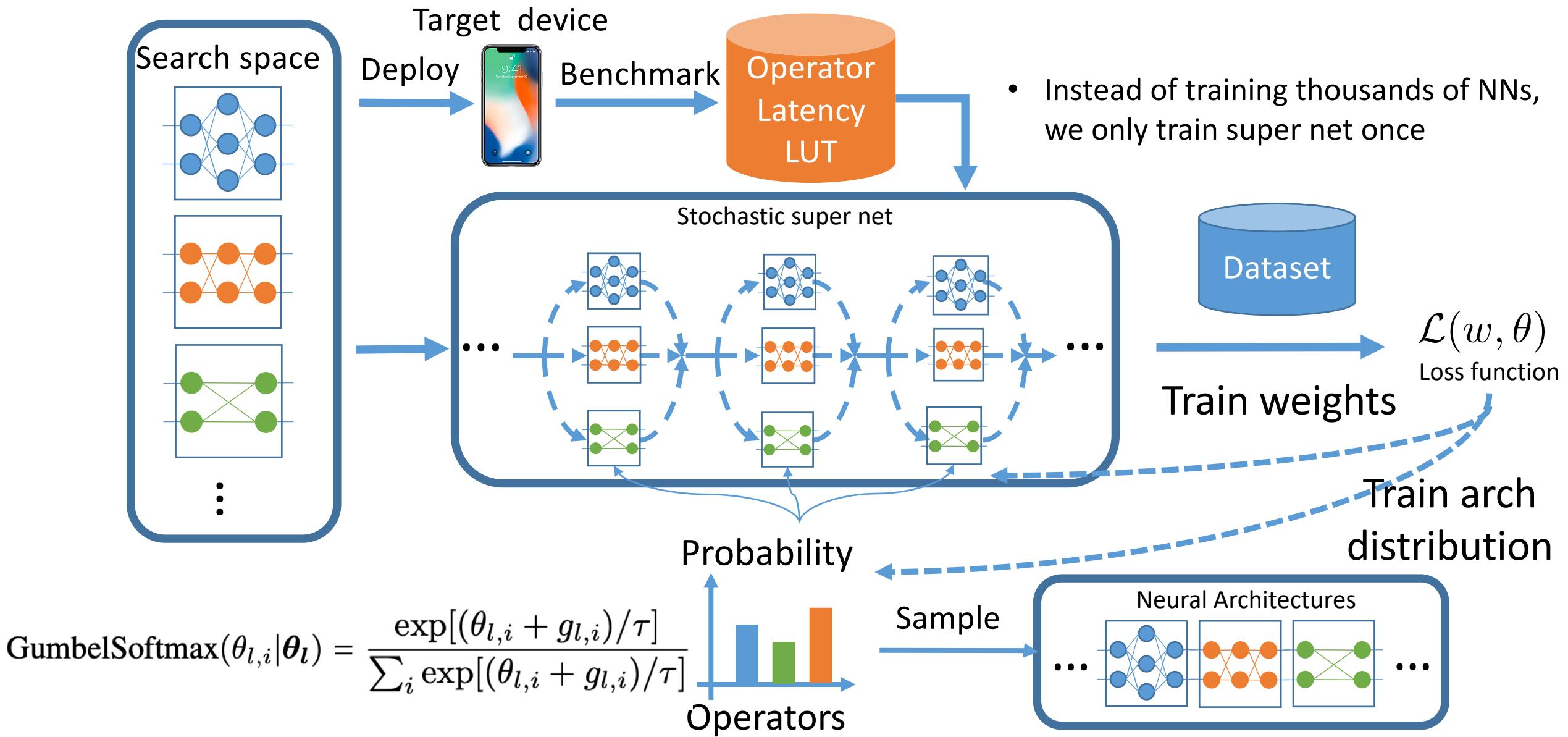
- Wu et al., FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search, CVPR 2019
- Dai et al., ChamNet: Towards Efficient Network Design through Platform-Aware Model Adaptation, CVPR 2019
- Wan et al., FBNetV2: Differentiable Neural Architecture Search for Spatial and Channel Dimensions, CVPR 2020
- Dai et al., FBNetV3: FBNetV3: Joint Architecture-Recipe Search using Neural Acquisition Function, submitted

# Hardware-aware Architecture Search

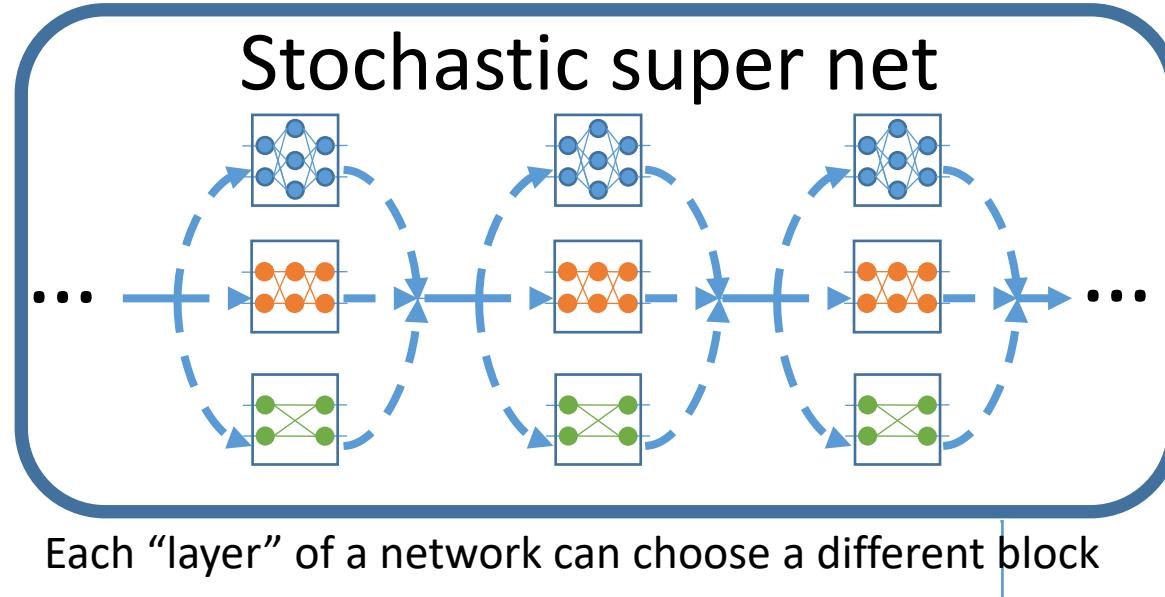
- Differentiable Architecture Search: **FBNet** and **FBNetV2**
- Predictor-based Architecture Search: ChamNet and FBNetV3



# FBNet: Differentiable Neural Architecture Search

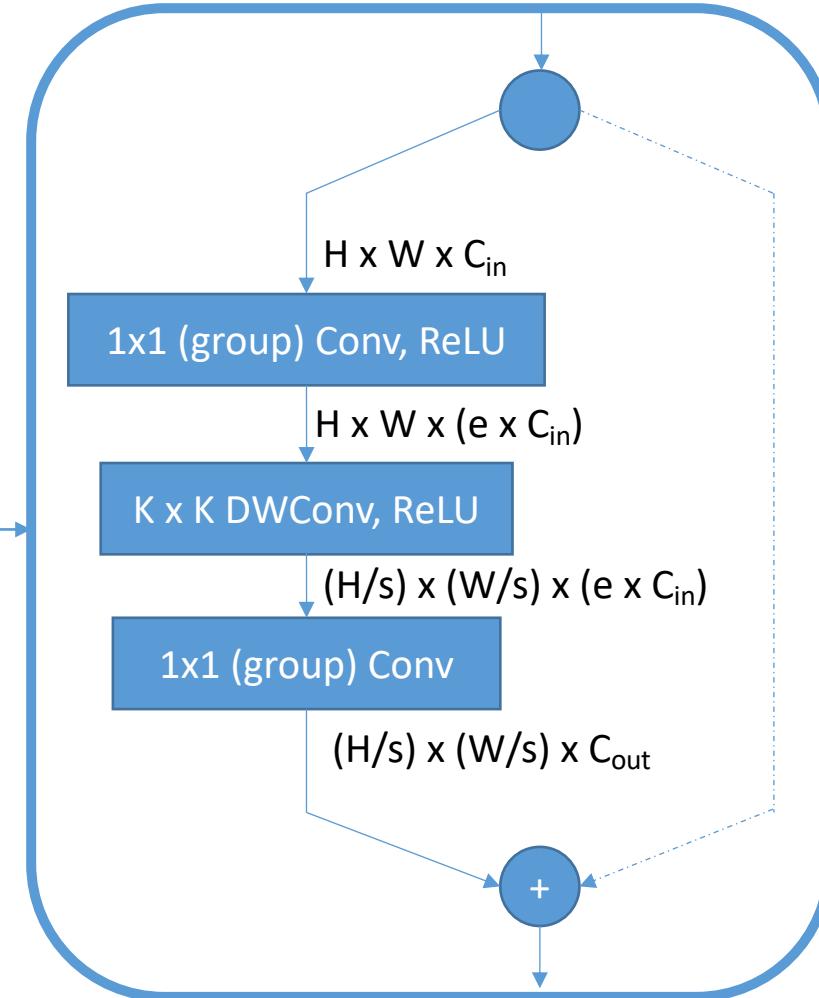


# FBNets Search Space



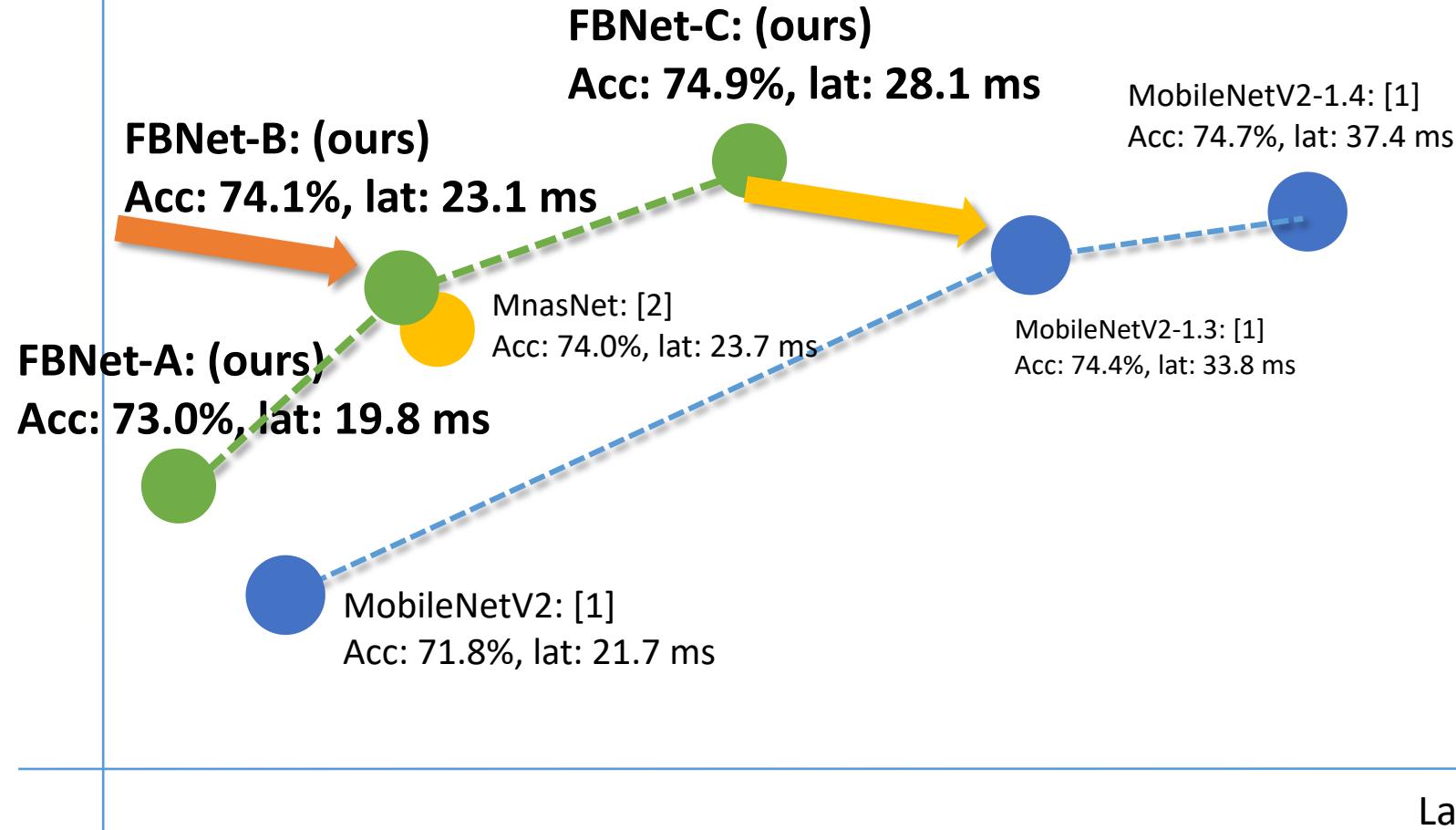
Candidate blocks with different hyper-parameters

- Kernel size: 3, 5
- Expansion rate: 1, 3, 6
- Skip: no-operation



# FBNets vs. Previous State-of-the-art

ImageNet top-1 Accuracy



- FBNet-B vs MobileNetV2[1], same accuracy, **1.5x faster, 2.4x smaller FLOPs**
- FBNet vs. MnasNet[2], search cost is 8 GPUs x 24 hours, **421x lower** than MnasNet

[1] Sandler, Mark, et al. "MobileNetV2: Inverted Residuals and Linear Bottlenecks." CVPR18

[2] Tan, Mingxing, et al. "Mnasnet: Platform-aware neural architecture search for mobile." CVPR19

# FBNets for Different Target Devices

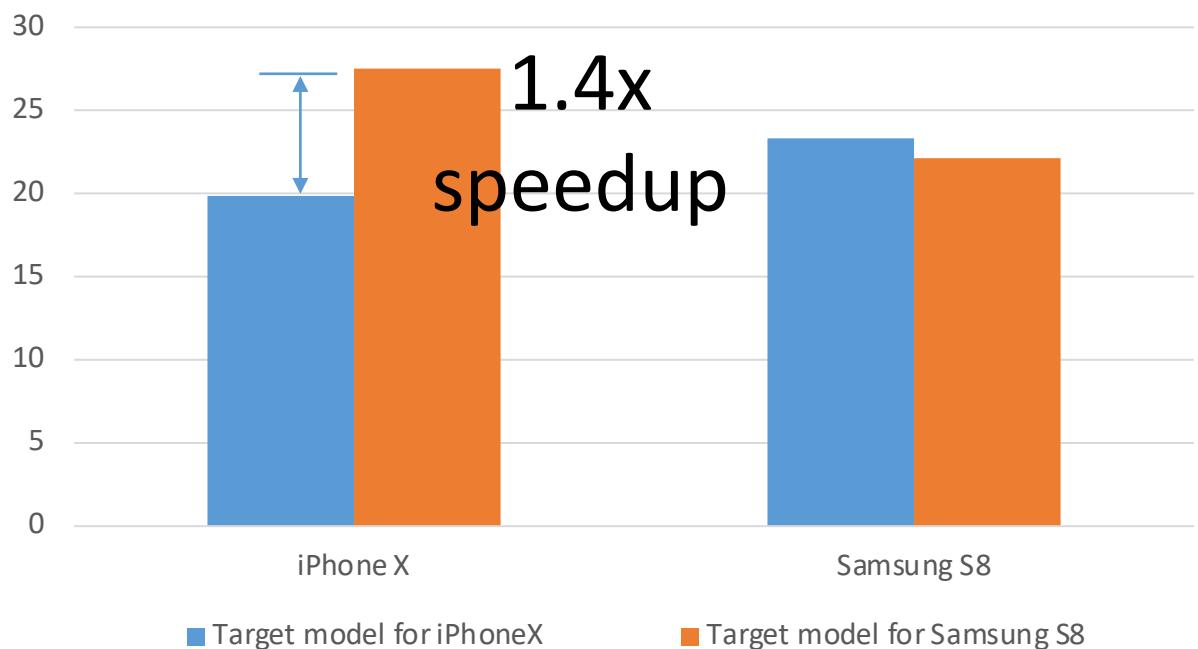


- Apple A11
- Big: 2 ARMv8 @ 2.5 GHz
- Little: 4 ARMv8 @ 1.4 GHz
- Vectorization: 4-wide 32-bit MAC
- LPDDR4x memory (30 GB/s)
- GPU + Neural Processing Engine



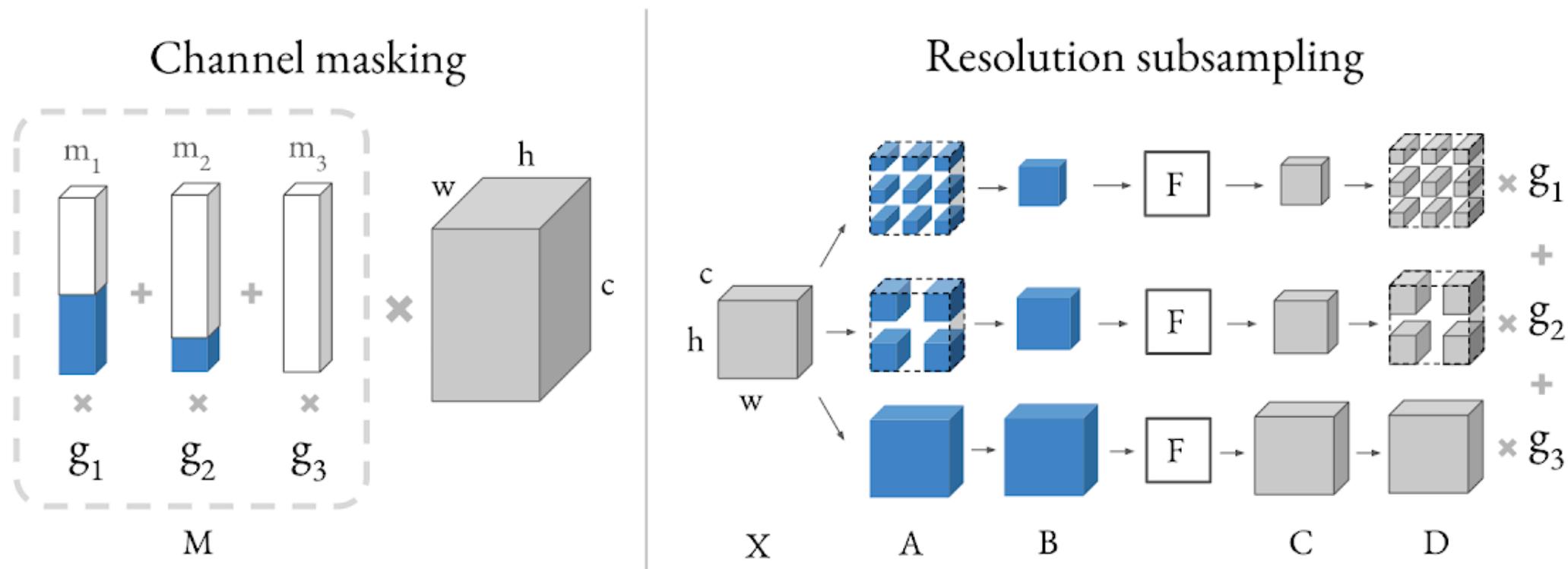
- Snapdragon 835
- Big: 4 ARMv8 @ 2.4 GHz
- Little: 4 ARMv8 @ 1.9 GHz
- Vectorization: 4-wide 32-bit MAC
- LPDDR4x memory (30 GB/s)
- Adreno 540 GPU

- Under similar accuracy constraint (73.27% vs 73.20%), FBNet optimized for iPhone-X achieves 1.4x speedup over the Samsung optimized model

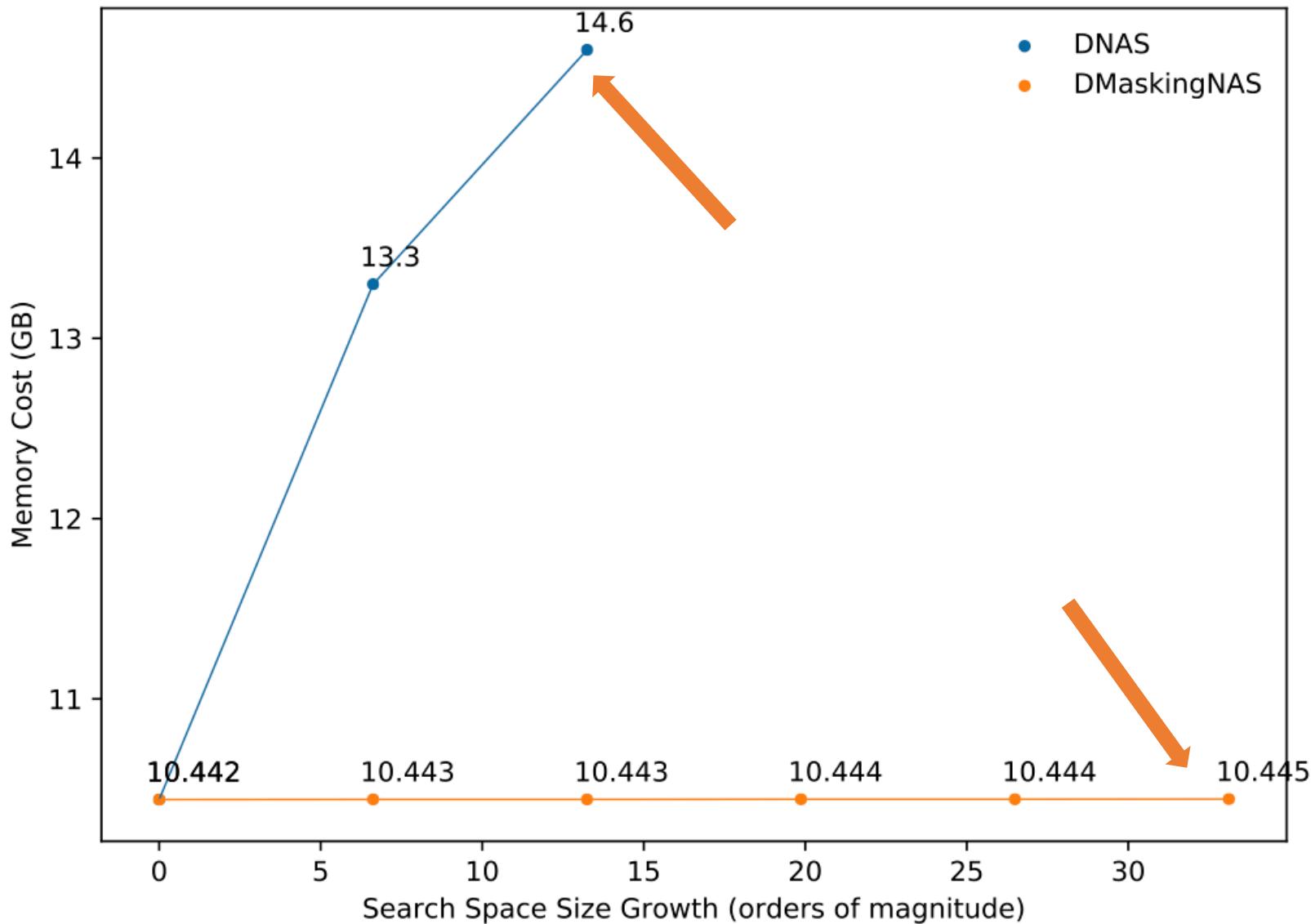


# FBNetV2: Differentiable Neural Architecture Search for Spatial and Channel Dimensions

- Searches Channels, Input Resolution, and Expansion Rate
- Expands search space by  $10^{14}$  times



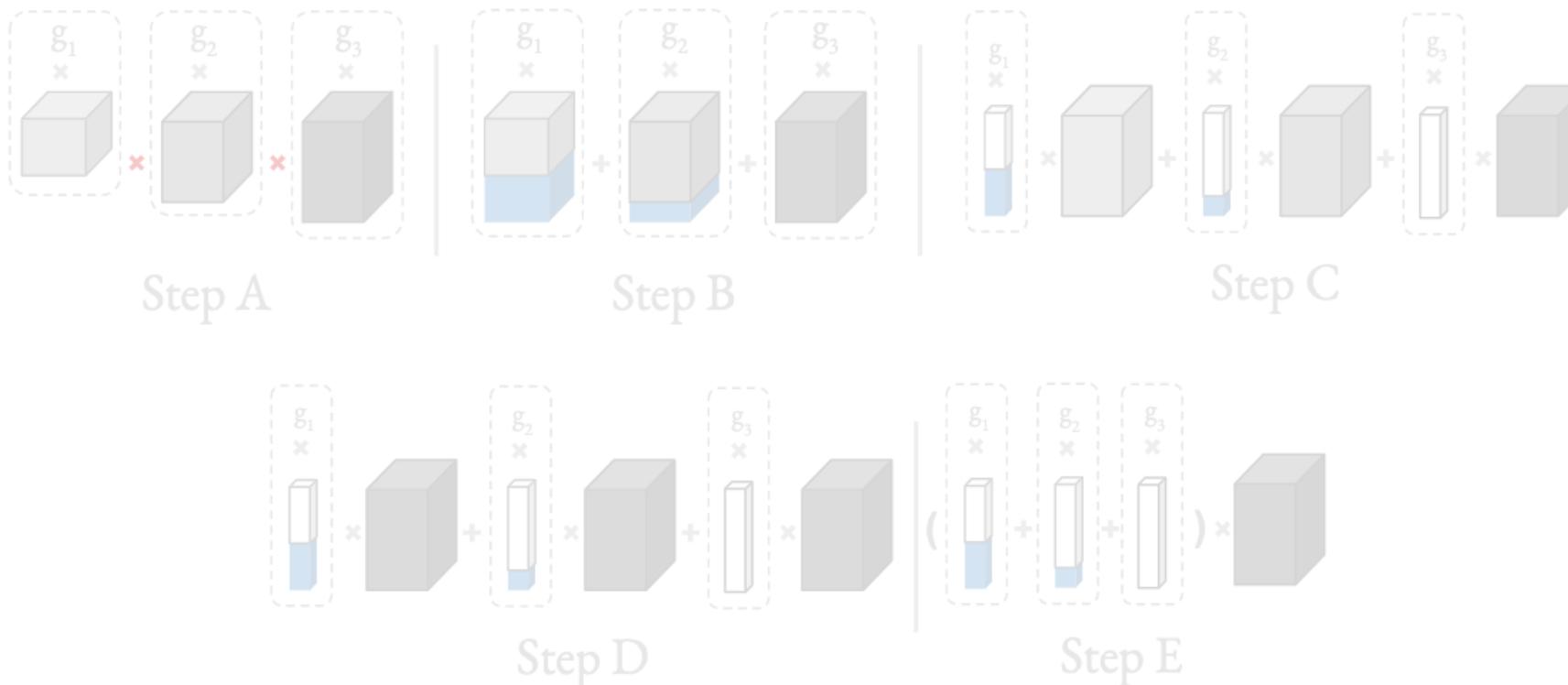
# SuperNet Memory Cost



# Channel Masking

- Expanding search space using weight-sharing approximations

**How Channel Search Works:** convolutional outputs (gray), and their architecture weights  $g_i$



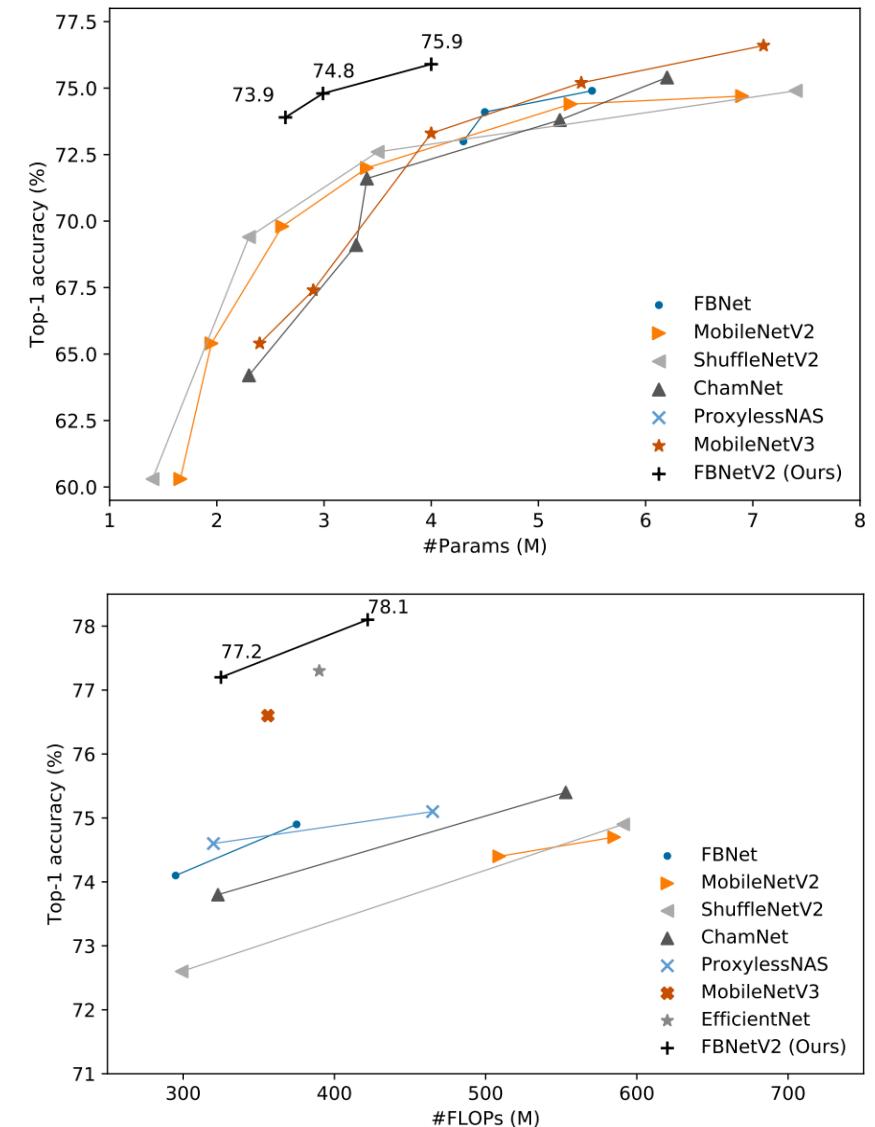
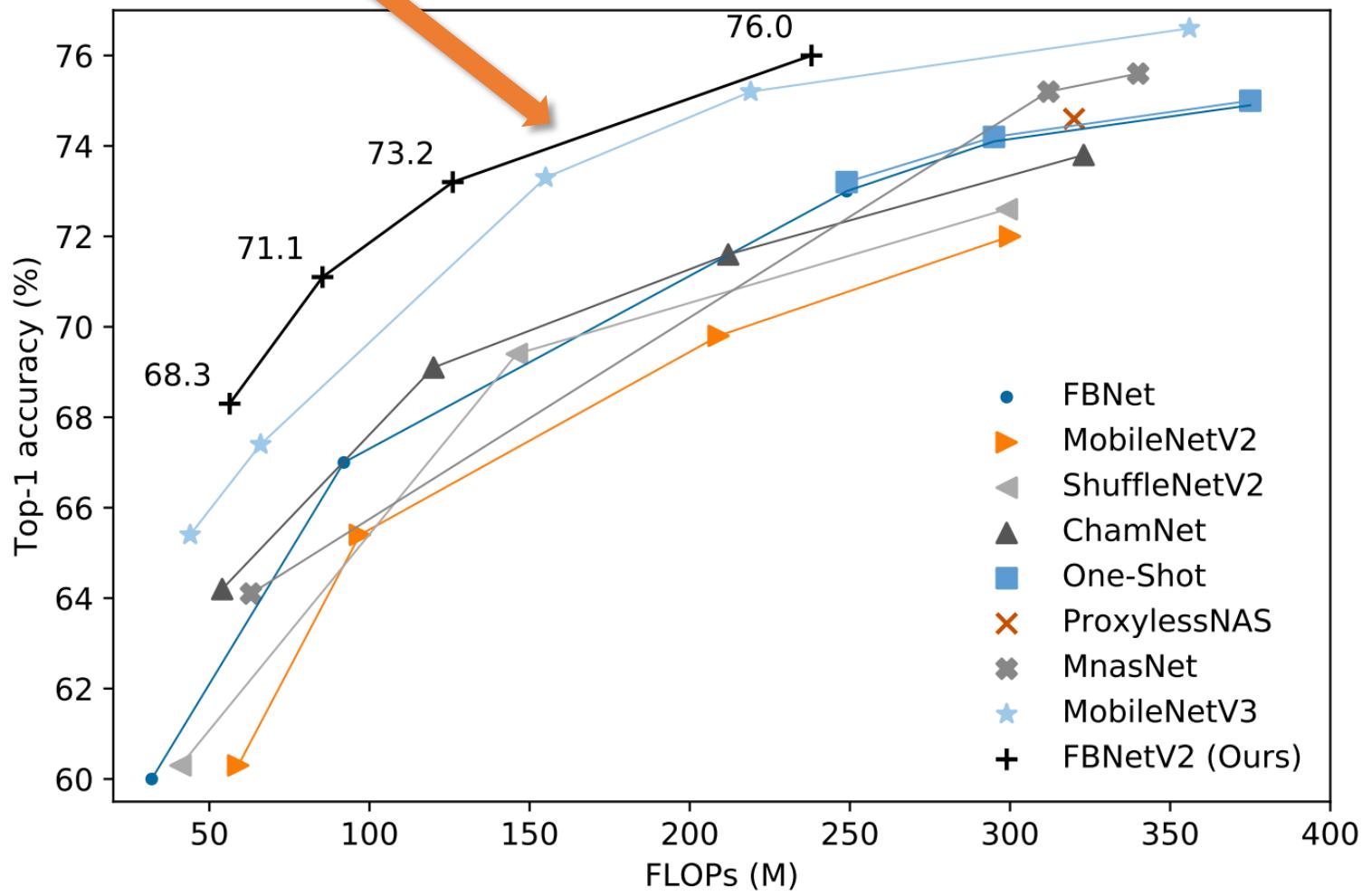
- A. Cannot sum feature maps
- B. Pad with zeros, but expensive
- C. Equivalent to masking
- D. Approximate by replacing all 3 convolutions with just 1 conv
- E. Equivalent to summing masks

# FBNetV2 Search Space

NAS algorithm	C	K	L	B	R	E
MnasNet [28]	✓		✓	✓		✓
ProxylessNAS [2]		✓	✓	✓		✓
Single-Path NAS [26]		✓	✓			✓
ChamNet [3]	✓		✓		✓	✓
FBNet [32]		✓	✓	✓		✓
DMaskingNAS	✓	✓	✓	✓	✓	✓

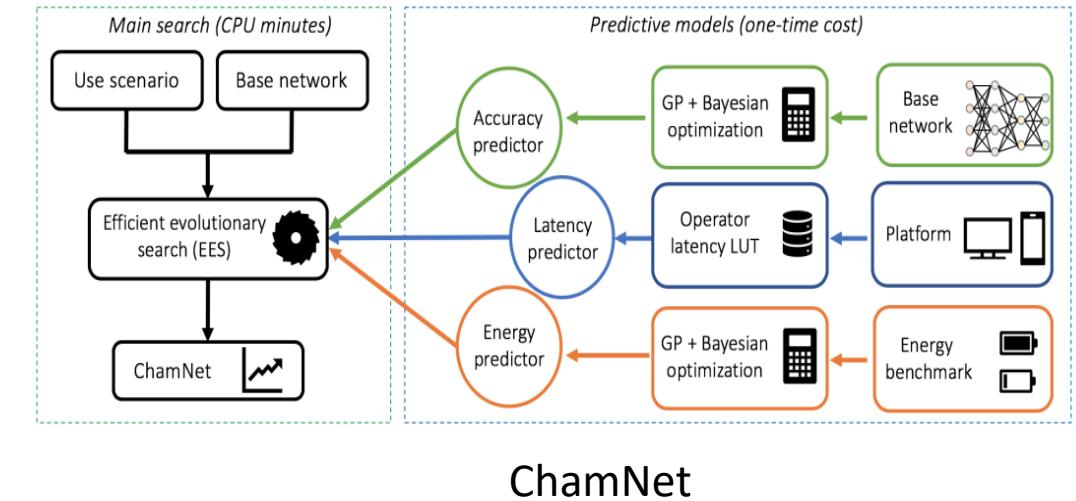
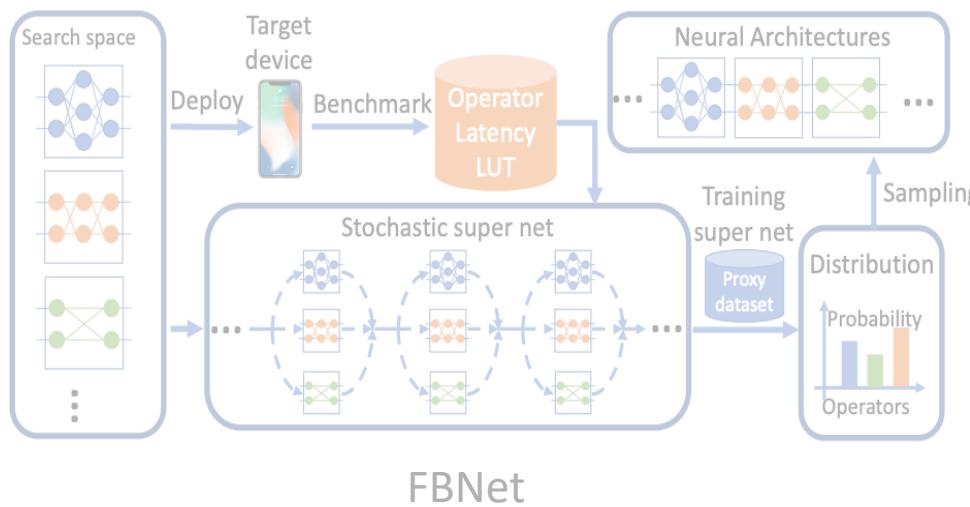
**FBNetV2’s search space** (bottom row) **eclipses other NAS’s**: Number of channels C, kernel size K, number of layers L, bottleneck type B, input resolution R, expansion rate E.

# FBNetV2 vs. previous state-of-the-art



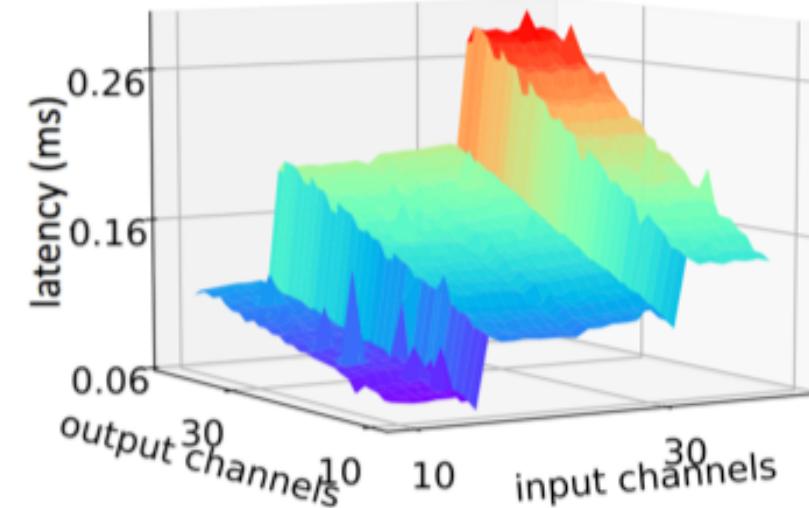
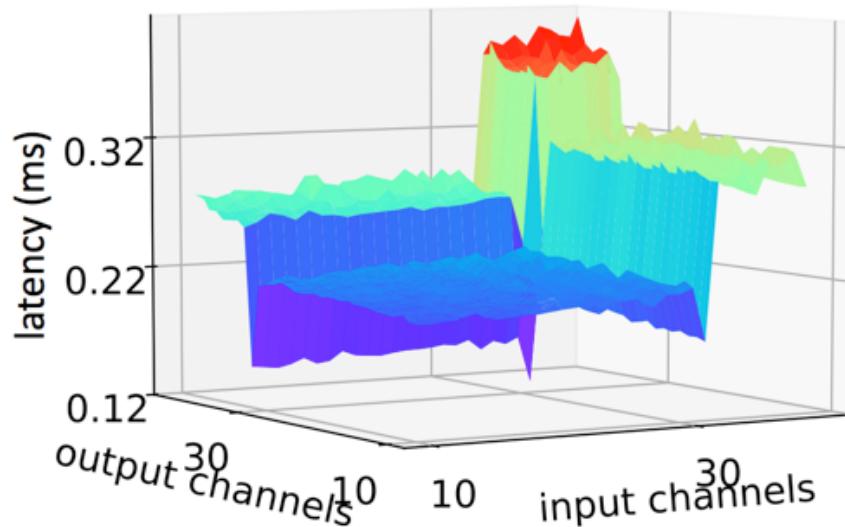
# Hardware-aware Architecture Search

- Differentiable Architecture Search: FBNet and FBNetV2
- Predictor-based Architecture Search: ChamNet and FBNetV3



# Challenges

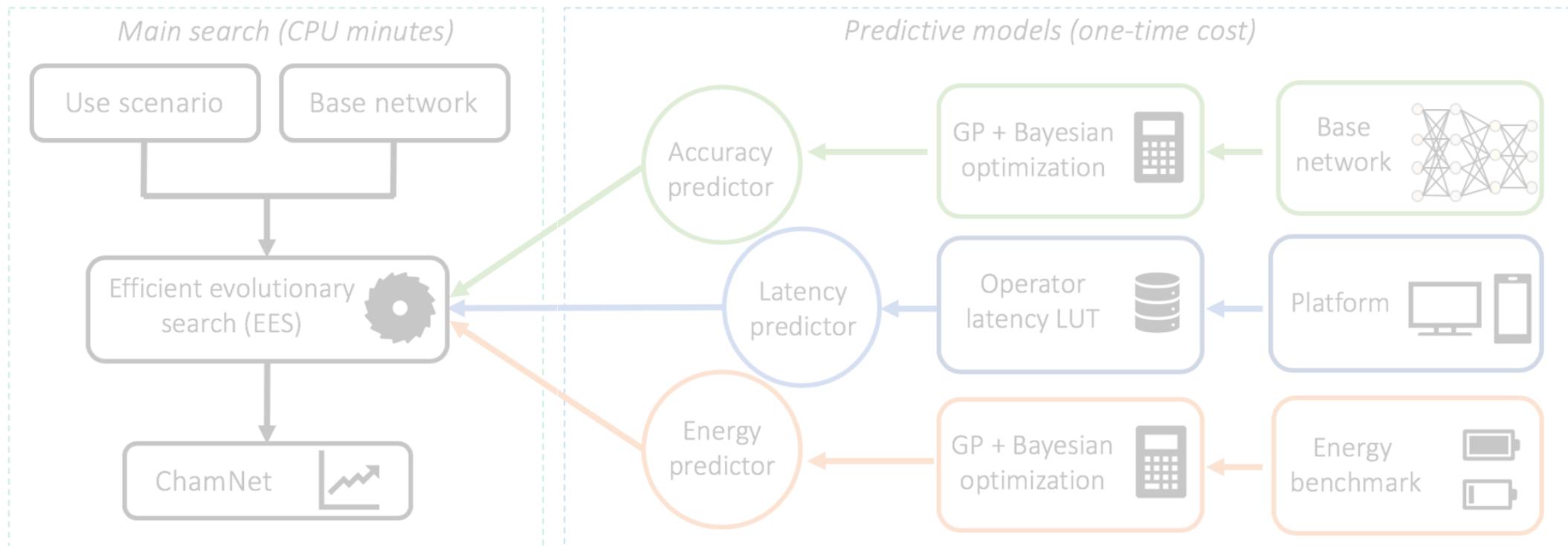
- One model does not fit all devices
- Too expensive to search per task per device



*Latency vs. #Channels for a 1x1 convolution on an input image size of 56x56 and stride 1 on (left) Snapdragon 835 GPU and (right) Hexagon v62 DSP.*

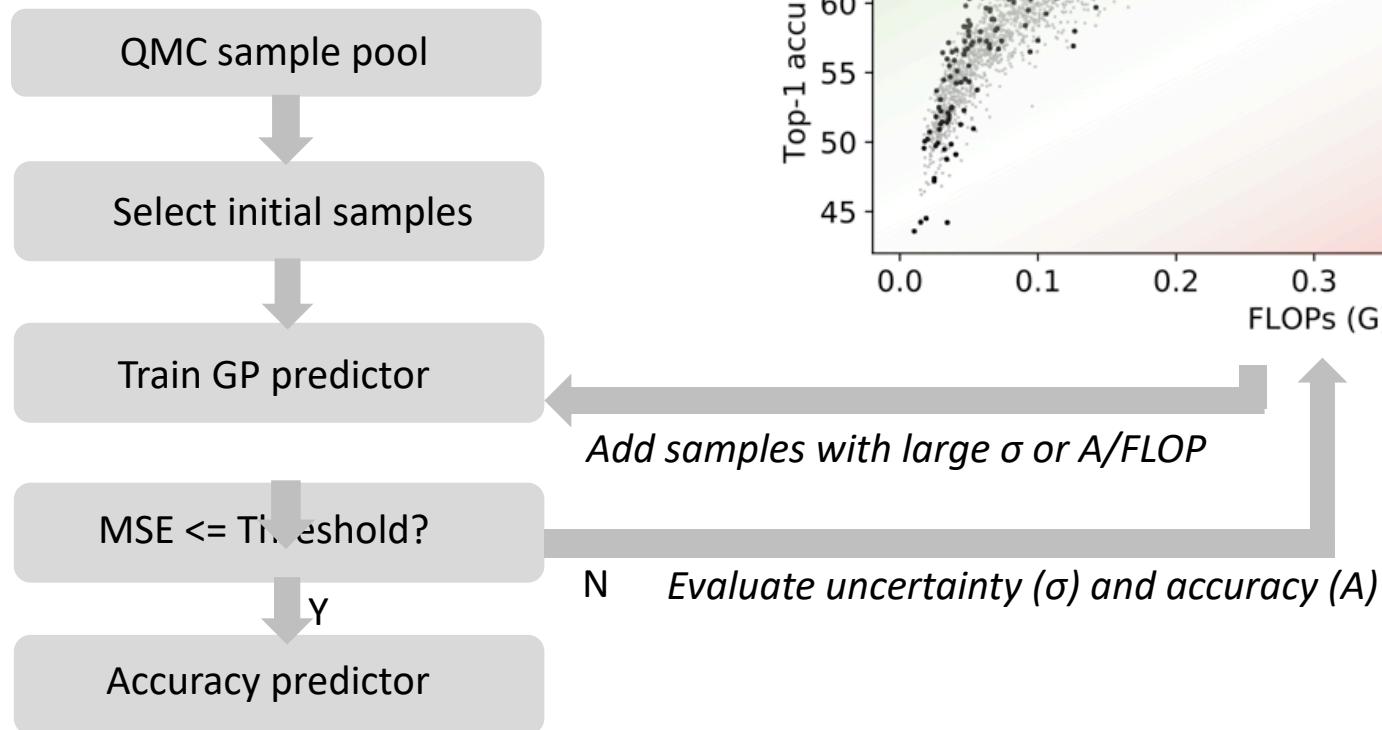
# ChamNet: Towards Efficient Network Design through Platform-Aware Model Adaptation

- Build predictive based models for **accuracy**, **latency**, and **energy**
- Search model for target devices in **minutes**



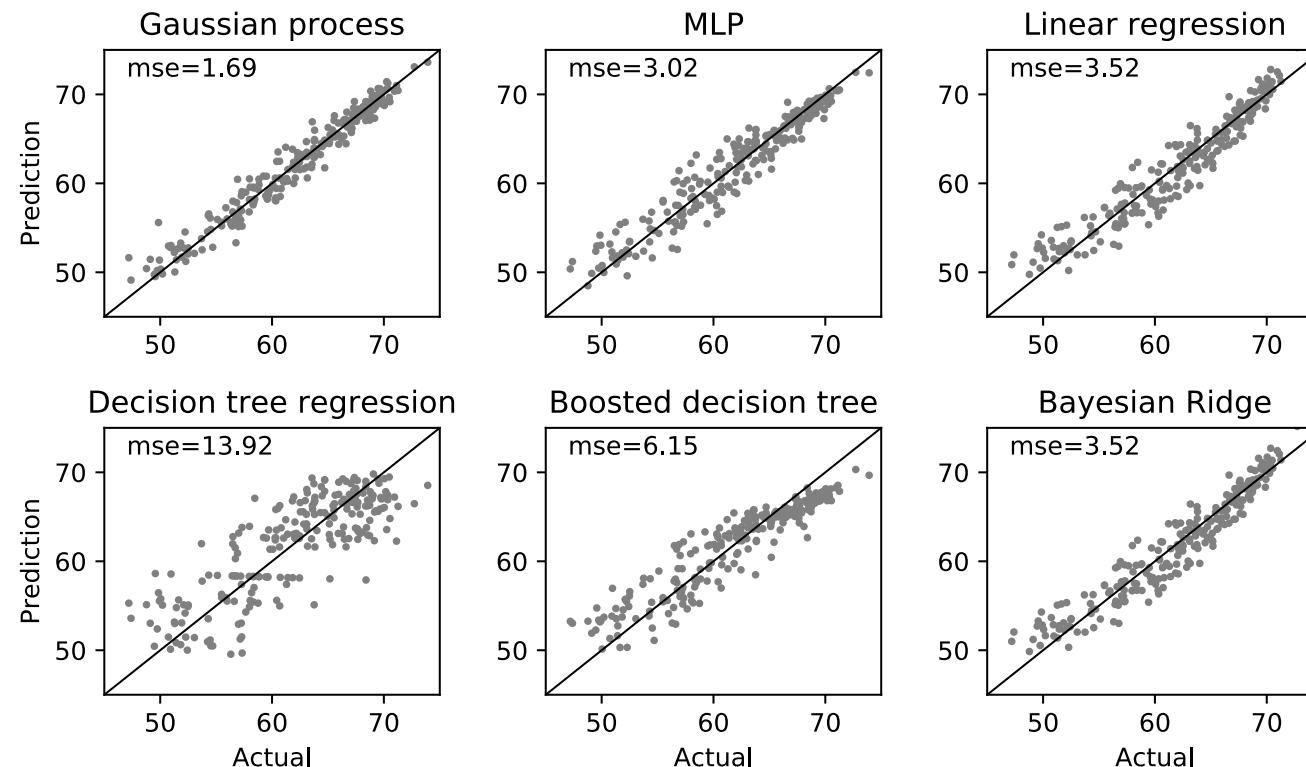
# Accuracy Predictor

- Gaussian process (GP) accuracy predictor
- Bayesian optimization for sample selection



# Accuracy Predictor

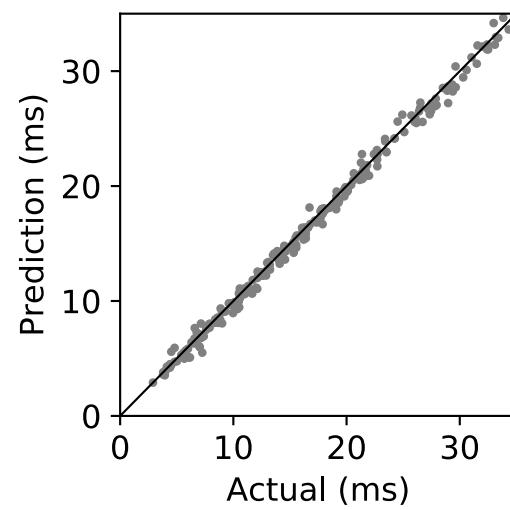
- Very efficient model evaluation
- No training during search



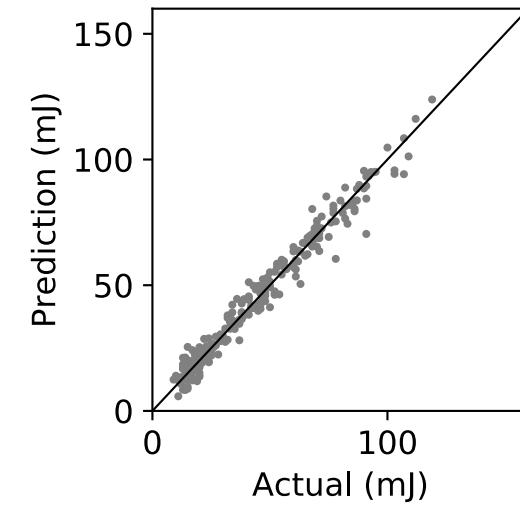
*Performance comparison of different accuracy prediction models.*

# Latency/Energy Predictor

- CNN latency look-up table (LUT)
  - Fast and reliable latency estimation
- Energy predictor
  - Based on real measurements on hardware



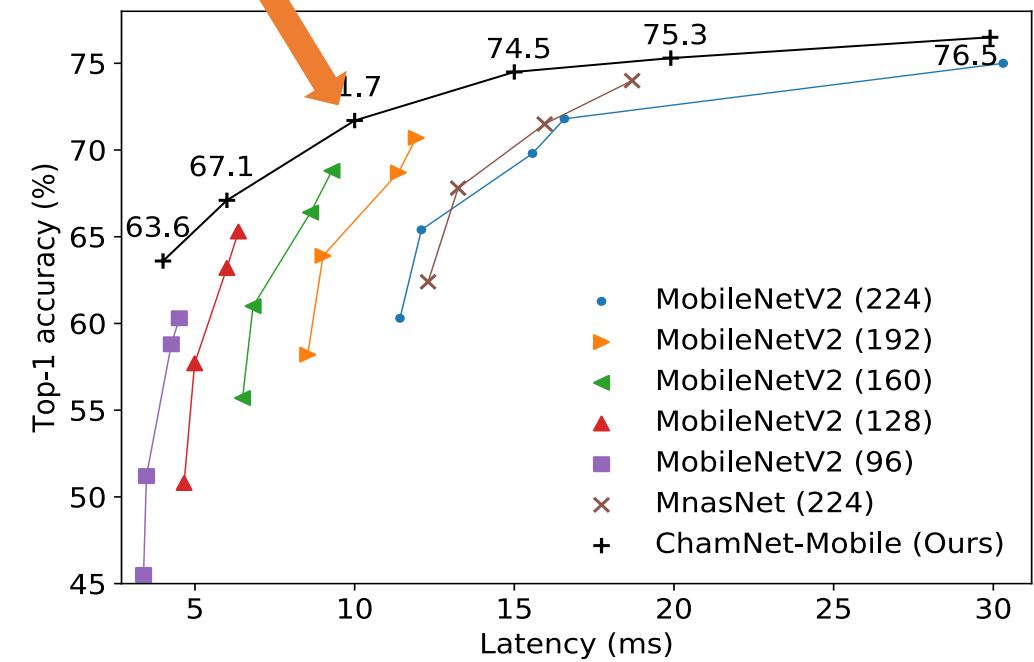
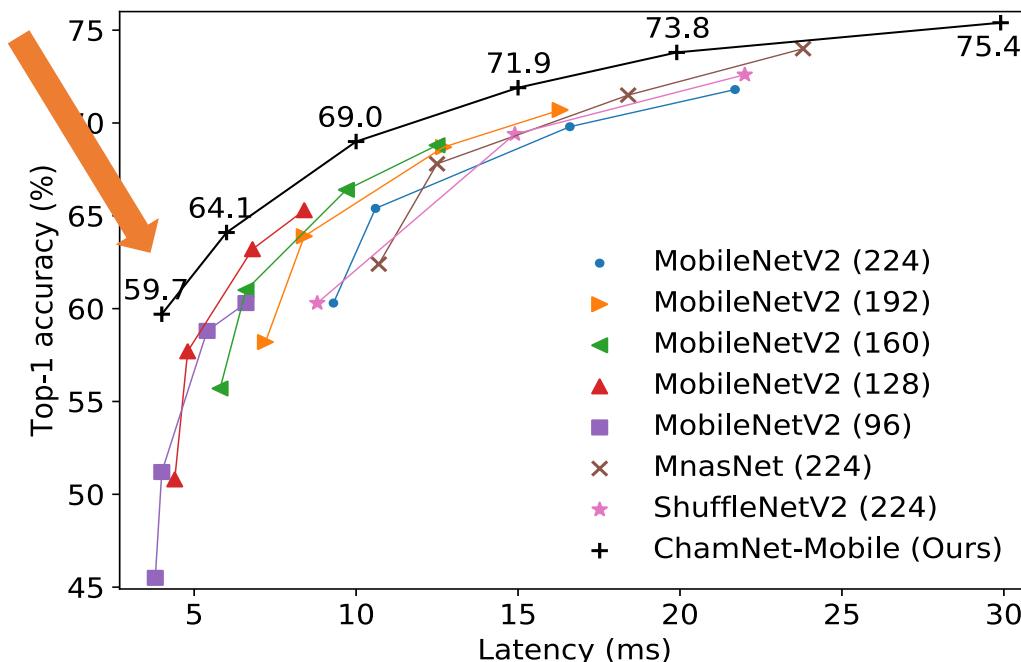
*Latency predictor*



*Energy predictor*

# ChamNet on Mobile CPU and DSP

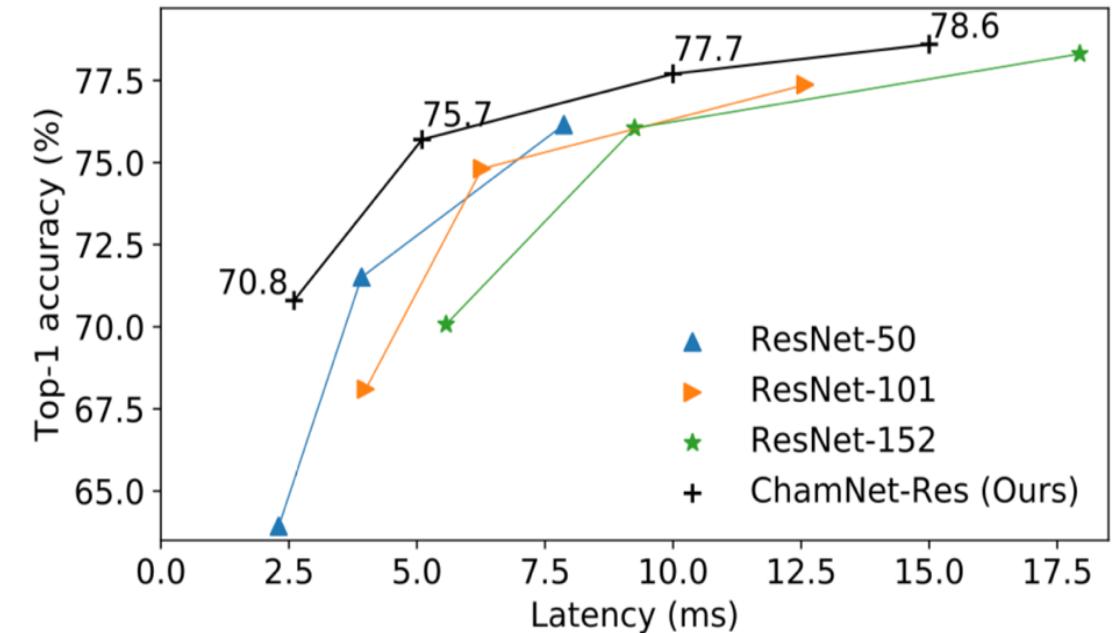
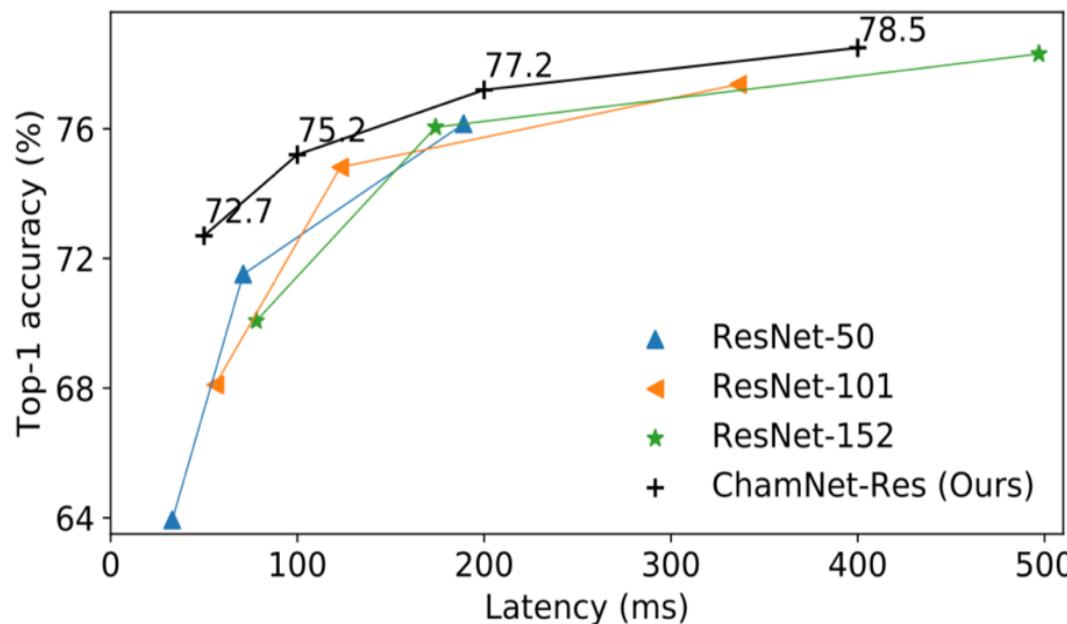
- Inverted residual block based search space
- 8.5% absolute accuracy gain at 4ms compared to MobileNetV2
- 6.6% absolute accuracy gain at 10ms compared to MnasNet



Performance of ChamNet-Mobile on (left) Snapdragon 835 CPU and (right) Hexagon v62 DSP. Numbers in parentheses indicate input image resolution.

# ChamNet on Server CPU and GPU

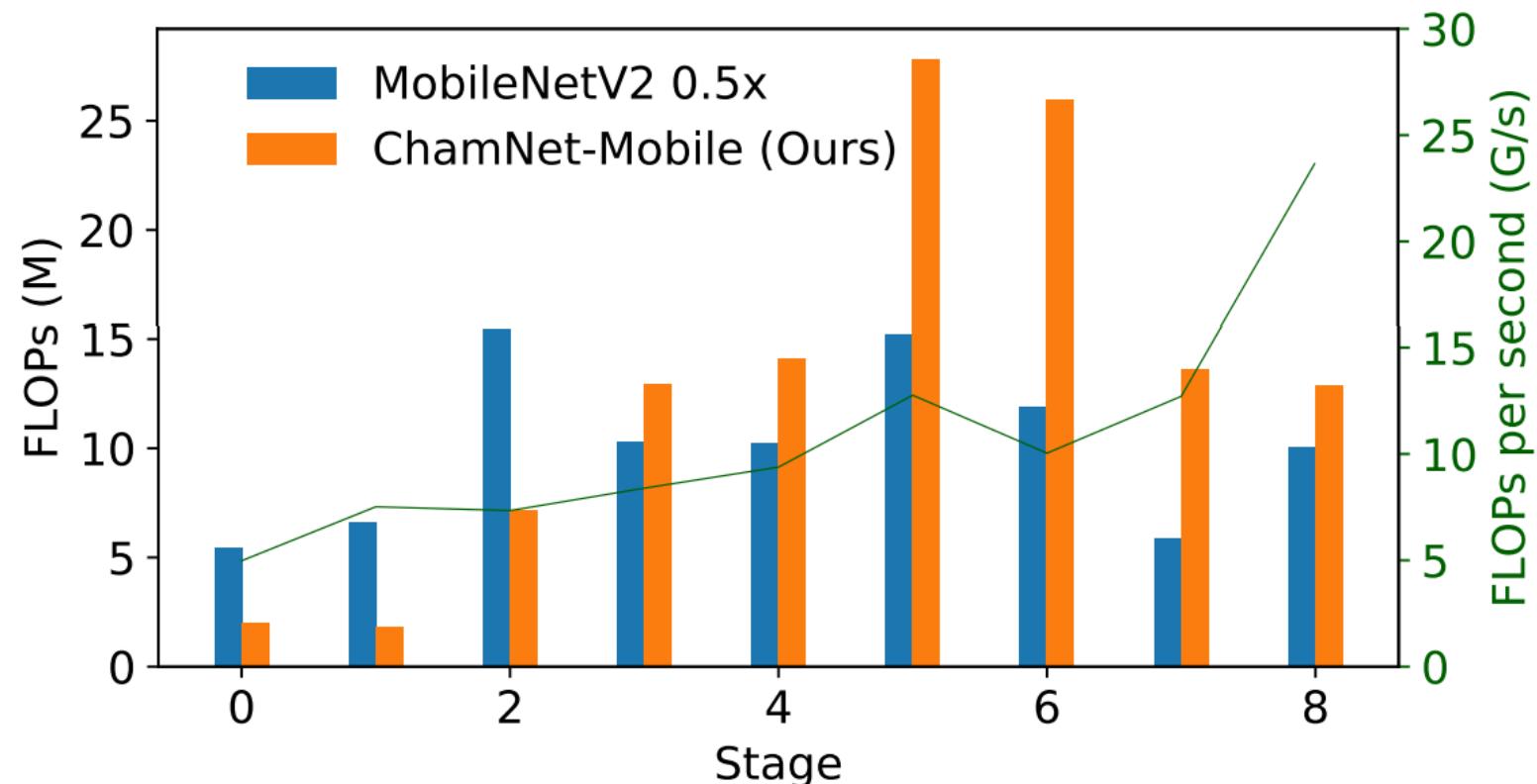
- Residual block based search space
- Significant improvement against ResNet-101 and ResNet-152



*Performance of ChamNet on (left) Intel CPU and (right) Nvidia GPU*

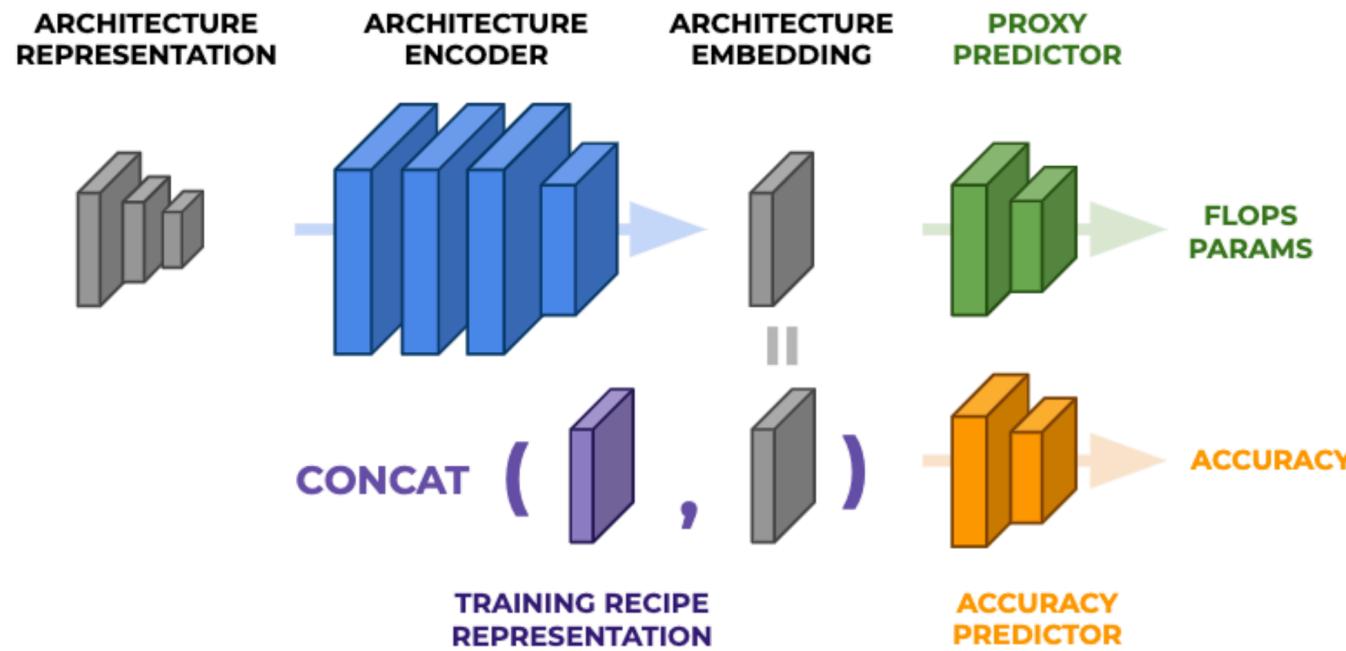
# Model FLOPS distribution

- The searched layer distribution is **heavy tailed**
- The operators in early stages with large input image size have lower efficiency (FLOPs per second)



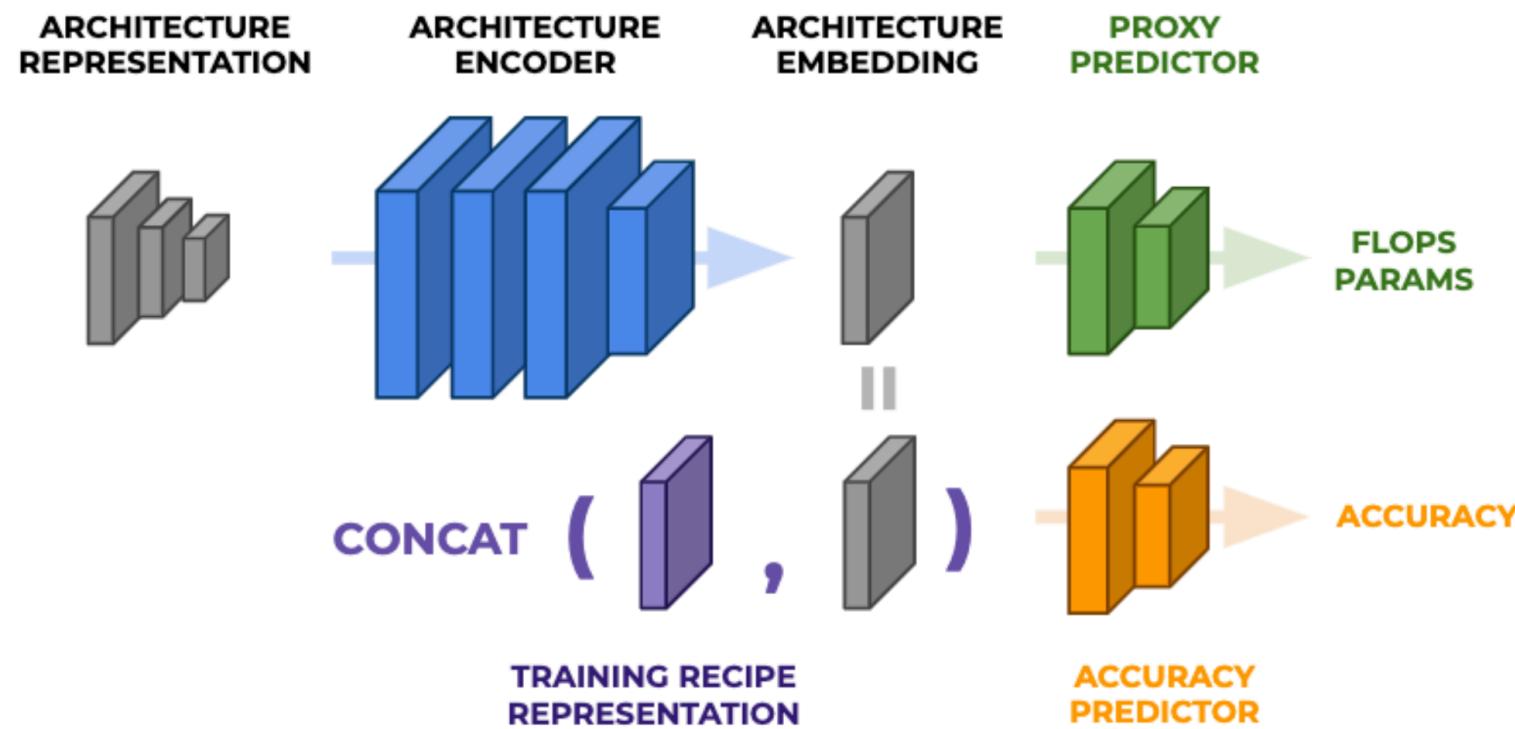
# FBNetV3: Joint Architecture-Recipe Search using Neural Acquisition Function

- Joint search of efficient architecture and training hyper-parameters
- DNN-based accuracy predictor



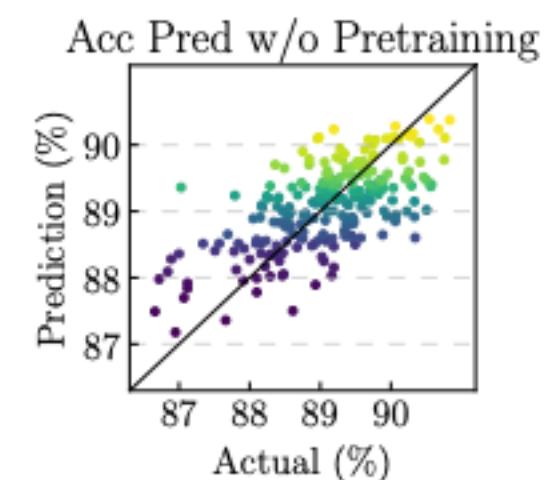
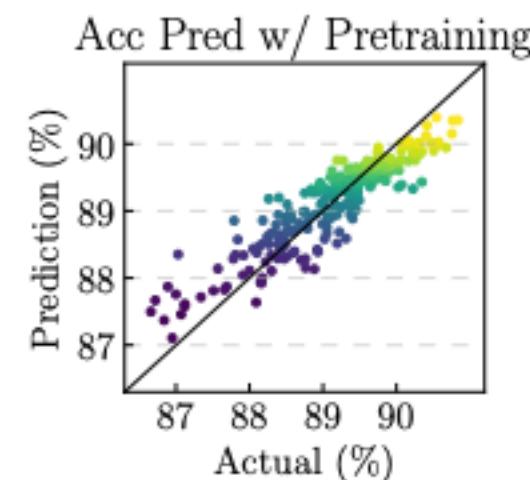
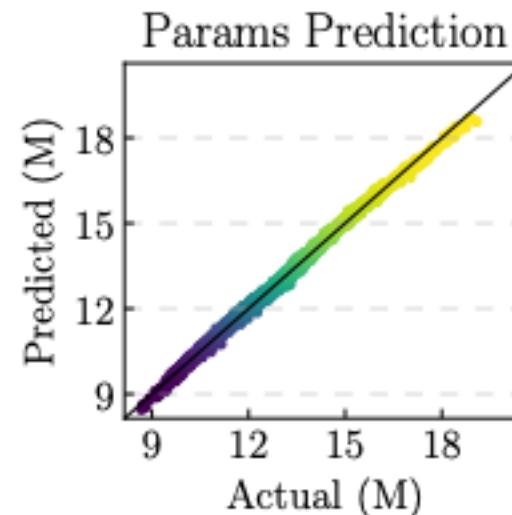
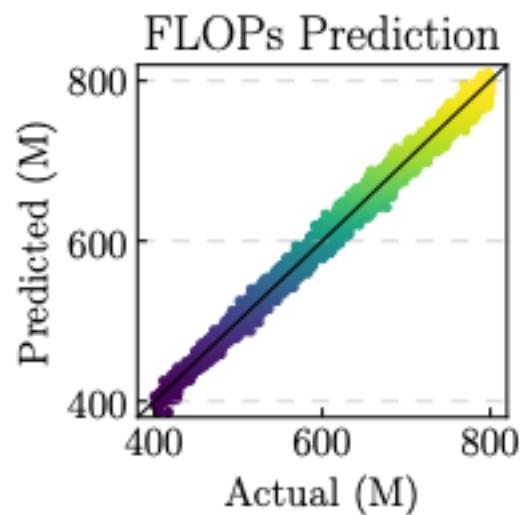
# DNN-based Accuracy Predictor

- Multi-layer perceptron with auxiliary head
  - One-hot categorical variables
  - Integral range variables



# Accuracy Predictor Pre-training

- Pretrain architecture encoder using FLOPs/number of parameters

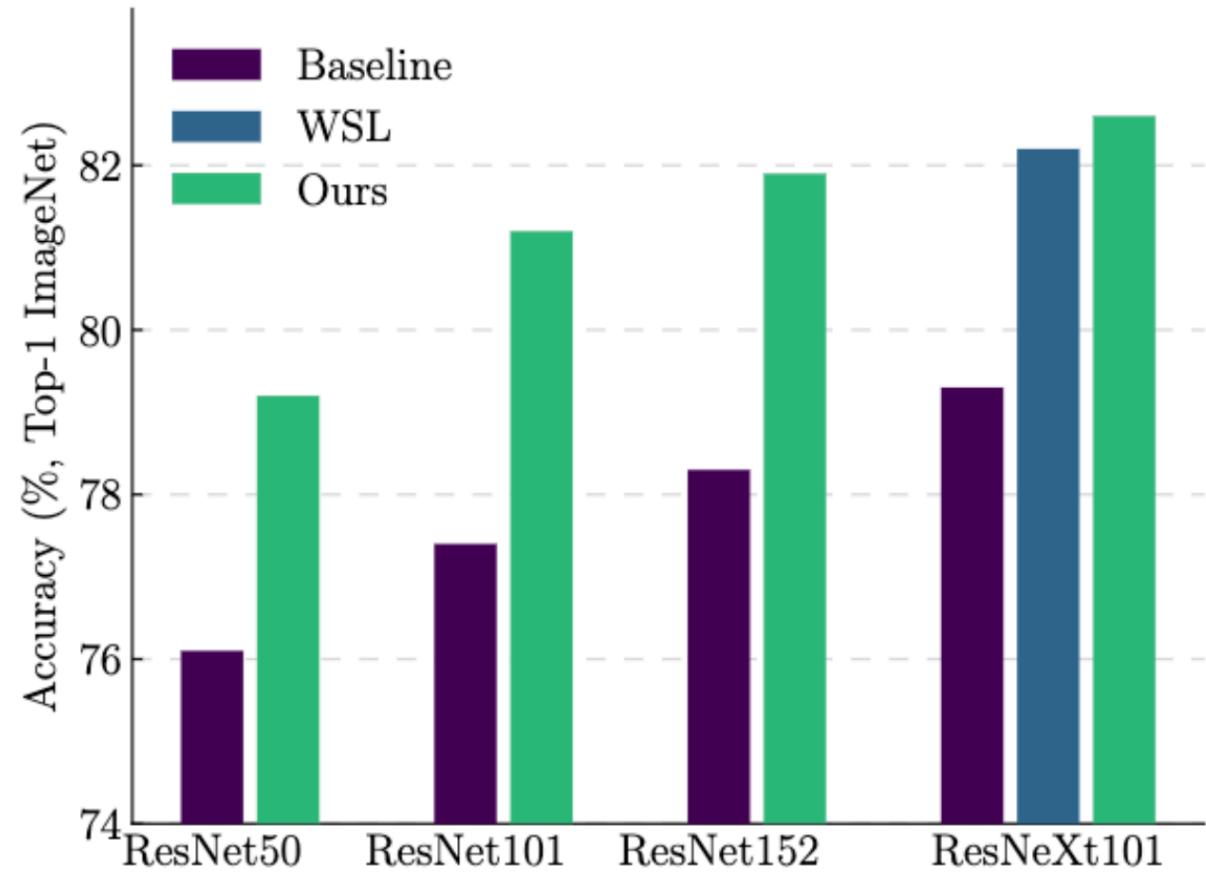


# FBNetV3 Search Space

- Architecture
  - Inverted residual block (kernel, channel, expansion, stride)
  - Number of layers
  - Squeeze-and-Excite
  - Activation function
  - Resolution
- Training hyper-parameters
  - Optimizer (type, learning rate, weight decay)
  - Regularization (dropout ratio, stochastic depth drop ratio, mix-up ratio)
  - EMA (exponential moving average)

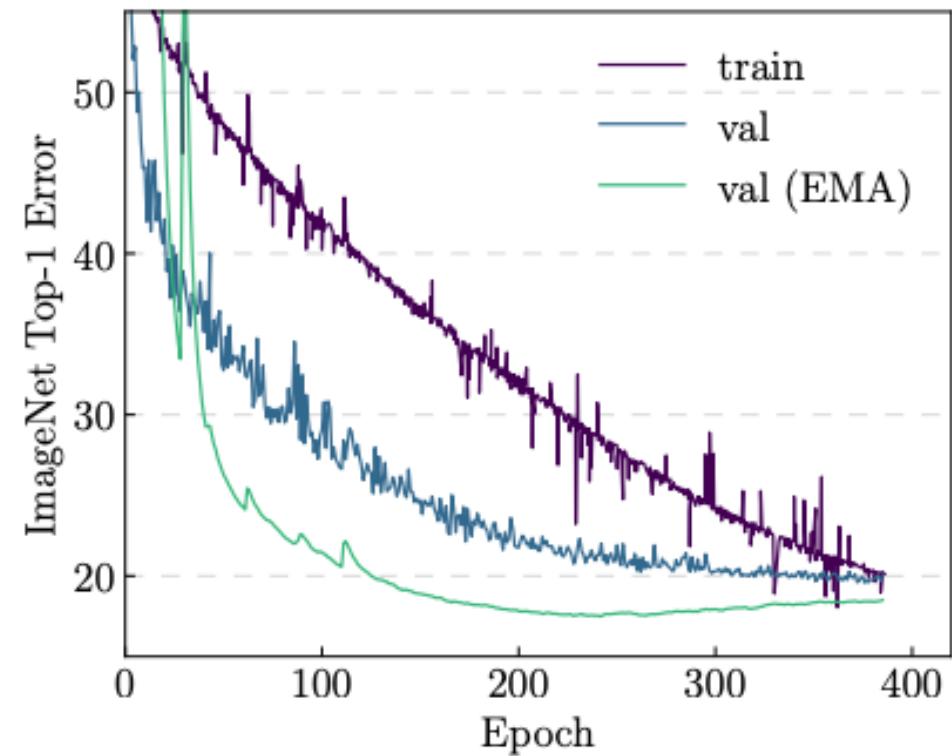
# Importance of Hyper-parameter Search

- 3~4% improvement
- Searched ResNet-50 shows 0.9% better accuracy than ResNet-152
- Searched ResNext101 surpasses the **weakly supervised learning** (WSL) model without extra 1B training data



# Importance of Hyper-parameter Search

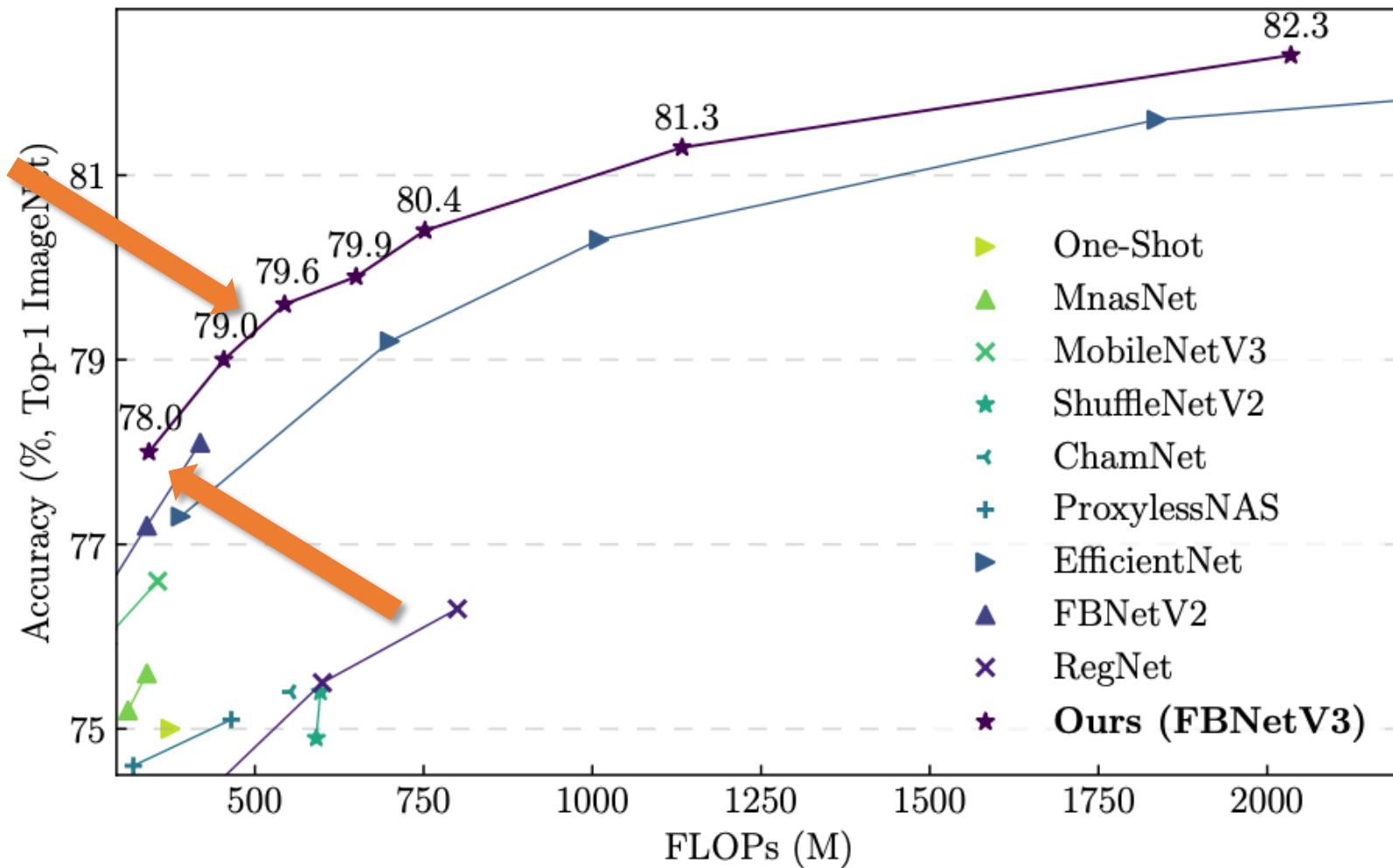
- Training with EMA shows significant improvements – but only with proper hyper-parameters settings
  - Learning rate, weight decay, learning rate decay
  - Synchronized BN



Training curve of the search recipe on ResNeXt101-32x8d

# FBNetV3 vs. previous state-of-the-art

- 28% less FLOPs with same accuracy as EfficientNet



# Architecture Search Applications for AR/VR



Single depth estimation for one shot 3D photography [1]



Real-time hand tracking on device for VR [2]

[1] Kopf et al, One Shot 3D Photography, SIGGRAPH 2020

[2] Han et al., MEgATrack: Monochrome Egocentric Articulated Hand-Tracking for Virtual Reality, SIGGRAPH 2020

# Summary

- Hardware-aware Architecture Search
  - Direct metric search (latency, energy)
  - State-of-the-art performance (same accuracy, 28% less compute)
- Differentiable Architecture Search (FBNet, FBNetV2)
  - Extremely fast: 8 GPUs for 24 hours, 421x faster search
  - $10^{14}$ x larger search space with near-constant memory cost
- Predictor-based Architecture Search (ChamNet, FBNetV3)
  - Fast adaption for different devices
  - Training hyper-parameter search

# Mobile Vision @ Facebook

## Contacts

Peizhao Zhang

<https://research.fb.com/people/zhang-peizhao/>

stzpz@fb.com

Peter Vajda

<https://research.fb.com/people/vajda-peter/>

vajdap@fb.com

Apply for **Research Scientist/Intern**

@ **MPK**, US and **Zurich**, Switzerland

mv-apply@fb.com

Thank you!

## References

- Wu et al., *FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search*
- Wan et al., *FBNetV2: Differentiable neural architecture search for spatial and channel dimensions*
- Dai et al., *ChamNet: Towards Efficient Network Design through Platform-Aware Model Adaptation*
- Dai et al., *FBNetV3: Joint Architecture-Recipe Search using Neural Acquisition Function*



Models available:

<https://github.com/facebookresearch/mobile-vision>