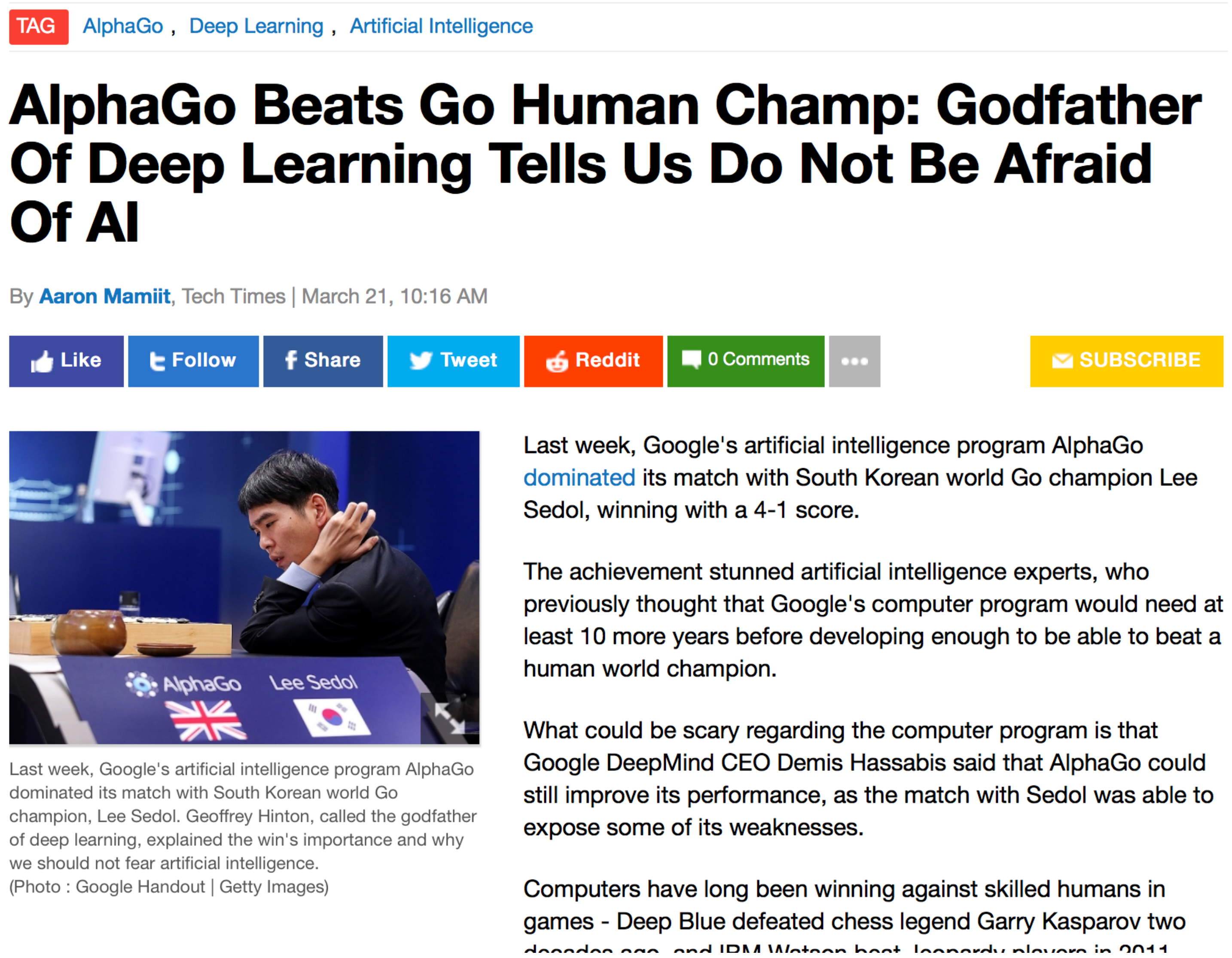
VIRTUAL JUNE 19-25

Convolutional Networks for Mobile Applications

Gao Huang

Department of Automation, Tsinghua University

Deep Learning


TAG AlphaGo , Deep Learning , Artificial Intelligence

AlphaGo Beats Go Human Champ: Godfather Of Deep Learning Tells Us Do Not Be Afraid Of AI

By **Aaron Mamiit**, Tech Times | March 21, 10:16 AM

Like | Follow | Share | Tweet | Reddit | 0 Comments | SUBSCRIBE

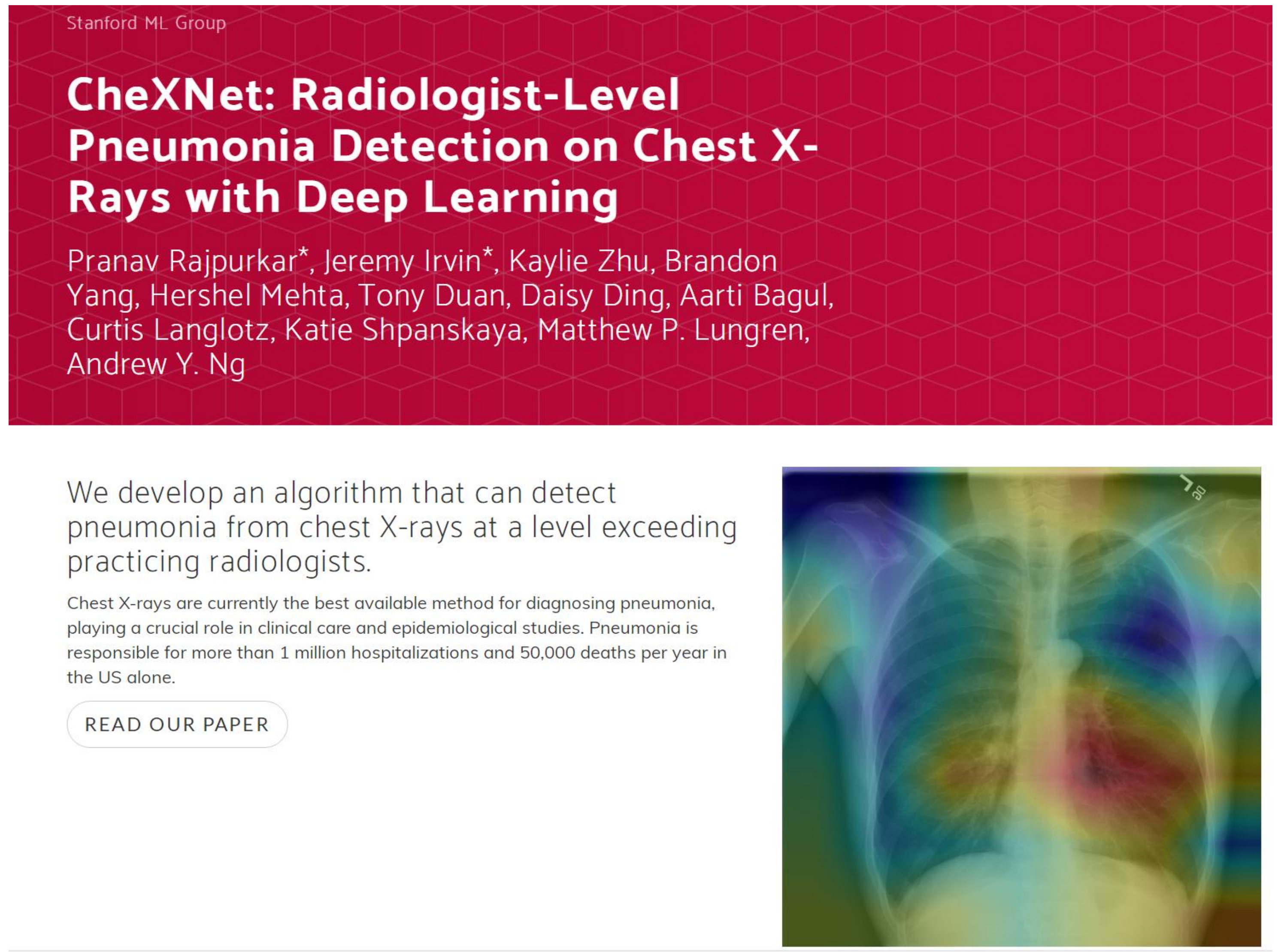
Last week, Google's artificial intelligence program AlphaGo dominated its match with South Korean world Go champion Lee Sedol, winning with a 4-1 score.

The achievement stunned artificial intelligence experts, who previously thought that Google's computer program would need at least 10 more years before developing enough to be able to beat a human world champion.

What could be scary regarding the computer program is that Google DeepMind CEO Demis Hassabis said that AlphaGo could still improve its performance, as the match with Sedol was able to expose some of its weaknesses.

Computers have long been winning against skilled humans in games - Deep Blue defeated chess legend Garry Kasparov two

Last week, Google's artificial intelligence program AlphaGo dominated its match with South Korean world Go champion, Lee Sedol. Geoffrey Hinton, called the godfather of deep learning, explained the win's importance and why we should not fear artificial intelligence.
(Photo : Google Handout | Getty Images)


Stanford ML Group

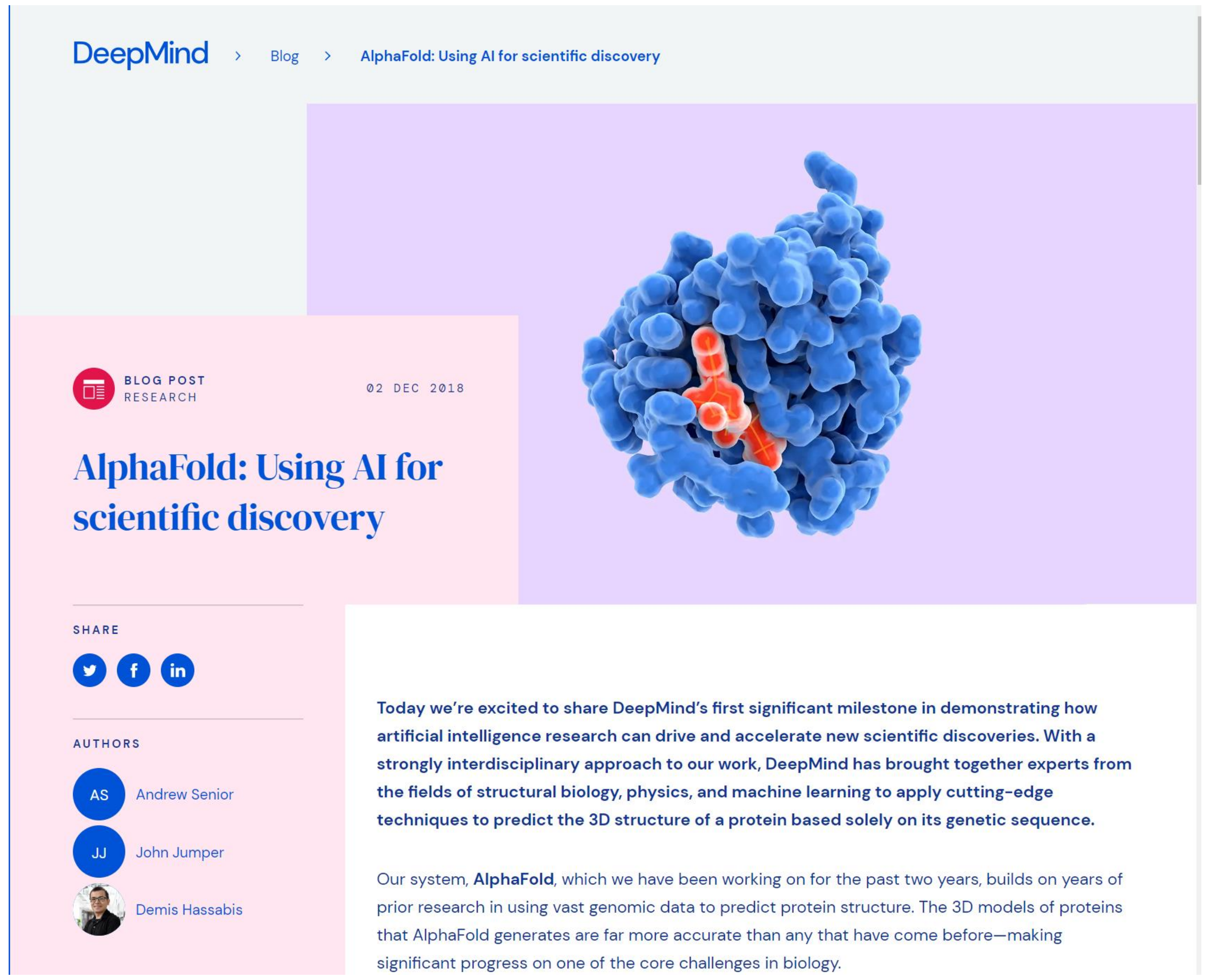
CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar*, Jeremy Irvin*, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng

We develop an algorithm that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists.

Chest X-rays are currently the best available method for diagnosing pneumonia, playing a crucial role in clinical care and epidemiological studies. Pneumonia is responsible for more than 1 million hospitalizations and 50,000 deaths per year in the US alone.

READ OUR PAPER


DeepMind › Blog › AlphaFold: Using AI for scientific discovery

BLOG POST RESEARCH 02 DEC 2018

AlphaFold: Using AI for scientific discovery

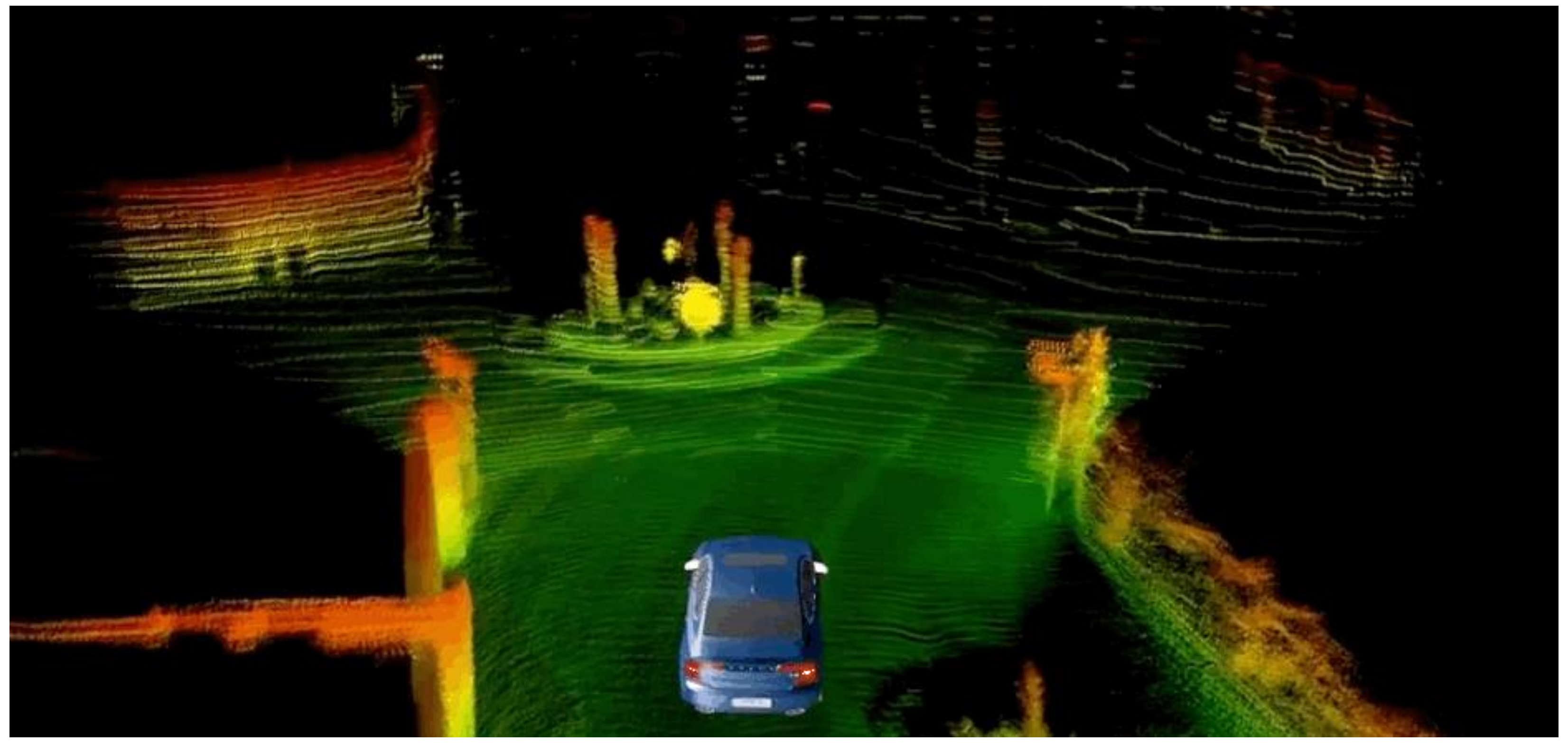
SHARE

AUTHORS

Andrew Senior
John Jumper
Demis Hassabis

Today we're excited to share DeepMind's first significant milestone in demonstrating how artificial intelligence research can drive and accelerate new scientific discoveries. With a strongly interdisciplinary approach to our work, DeepMind has brought together experts from the fields of structural biology, physics, and machine learning to apply cutting-edge techniques to predict the 3D structure of a protein based solely on its genetic sequence.

Our system, **AlphaFold**, which we have been working on for the past two years, builds on years of prior research in using vast genomic data to predict protein structure. The 3D models of proteins that AlphaFold generates are far more accurate than any that have come before—making significant progress on one of the core challenges in biology.



Outline

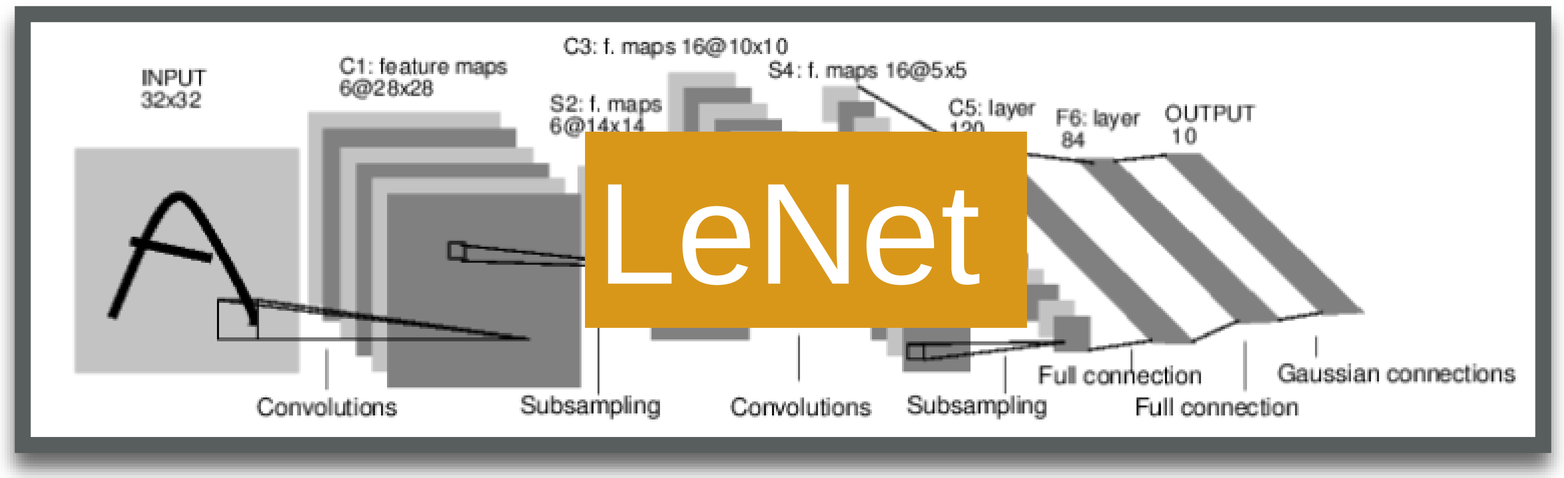
1. Overview of CNN backbones
2. Architecture design for mobile CNNs
3. Dynamic CNNs for mobile applications

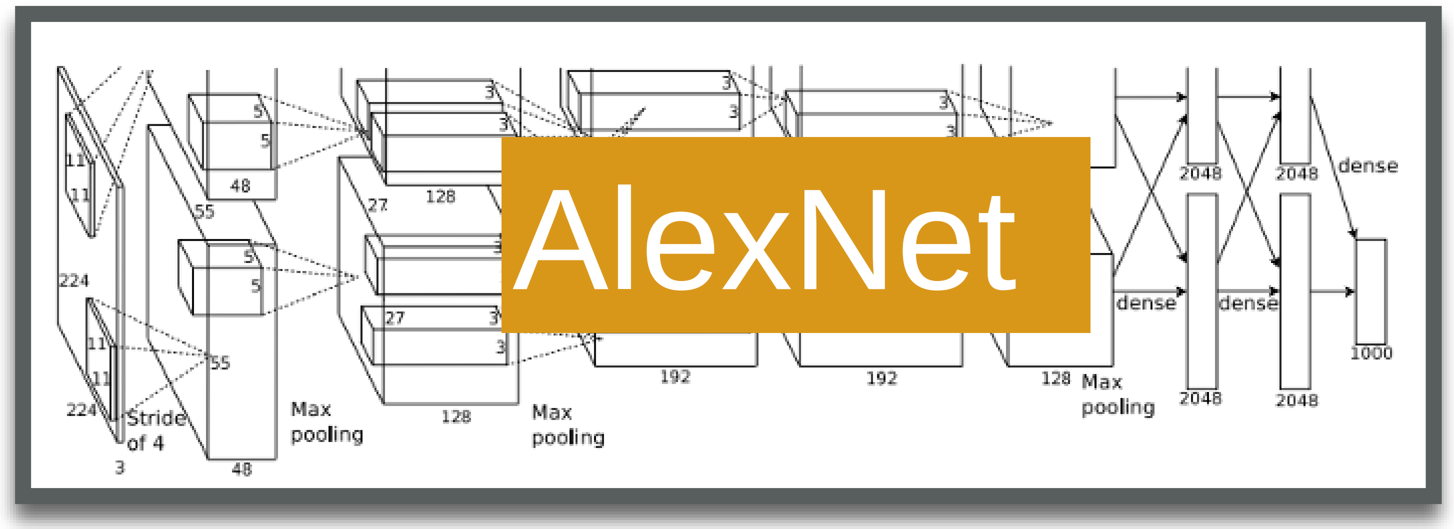
Outline

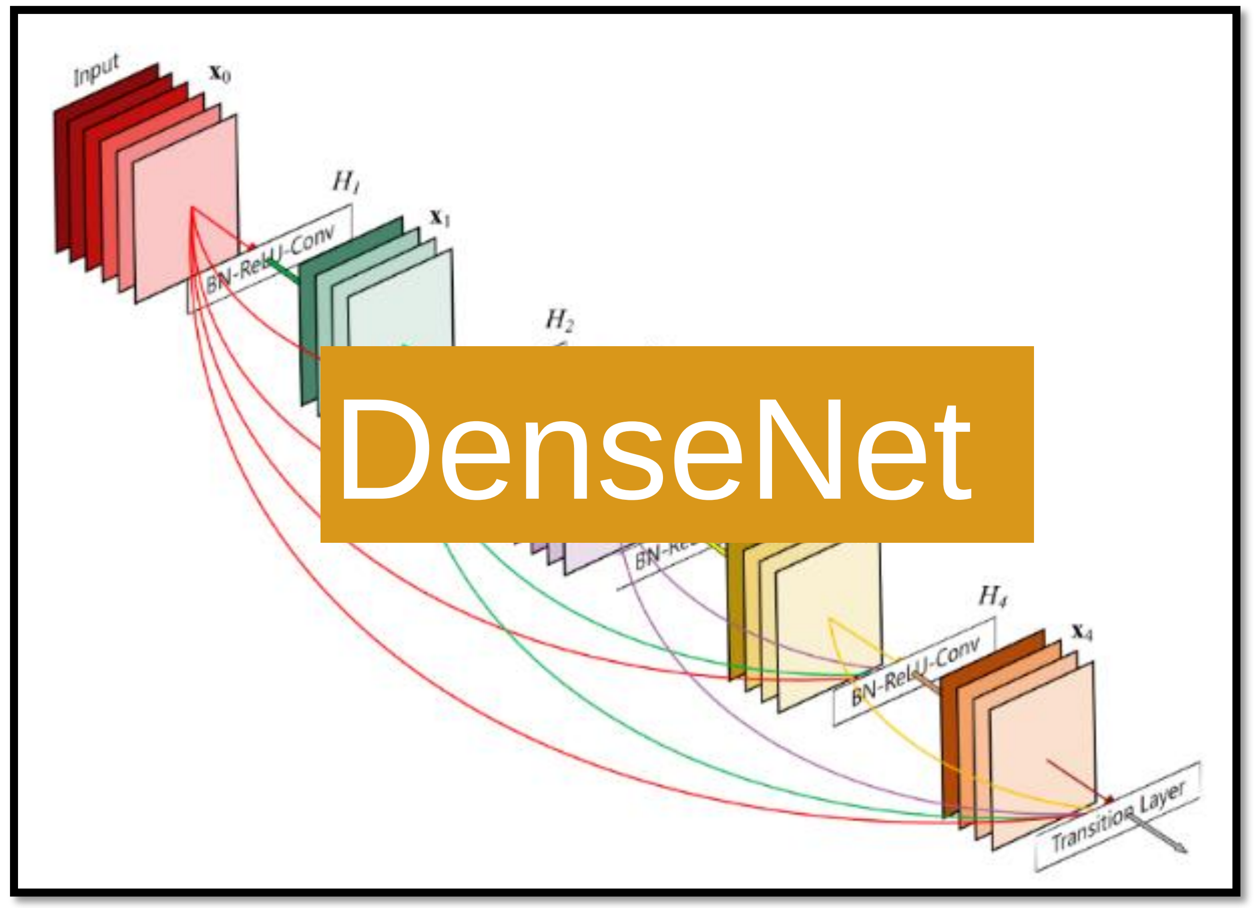
清華大學
Tsinghua University

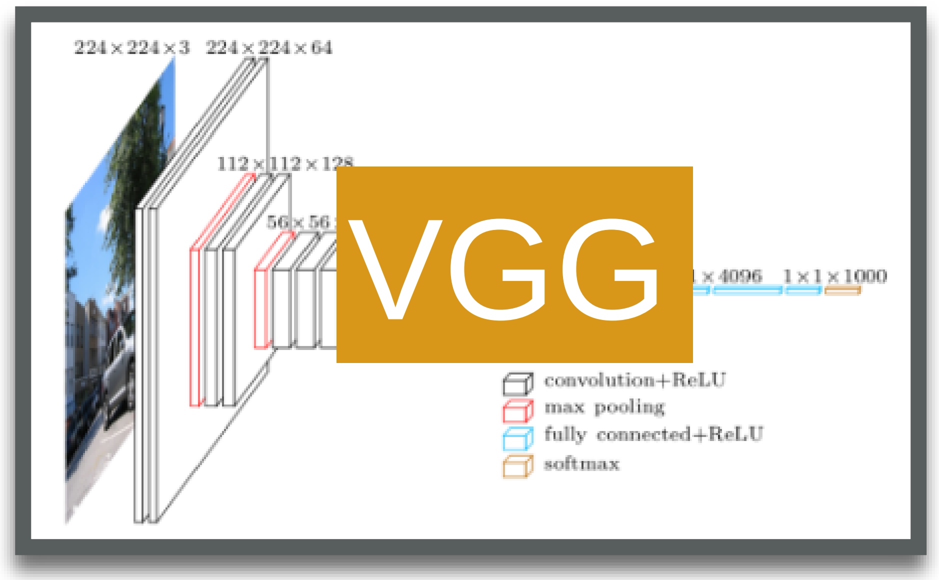
1. Overview of CNN backbones
2. Architecture design for mobile CNNs
3. Dynamic CNNs for mobile applications

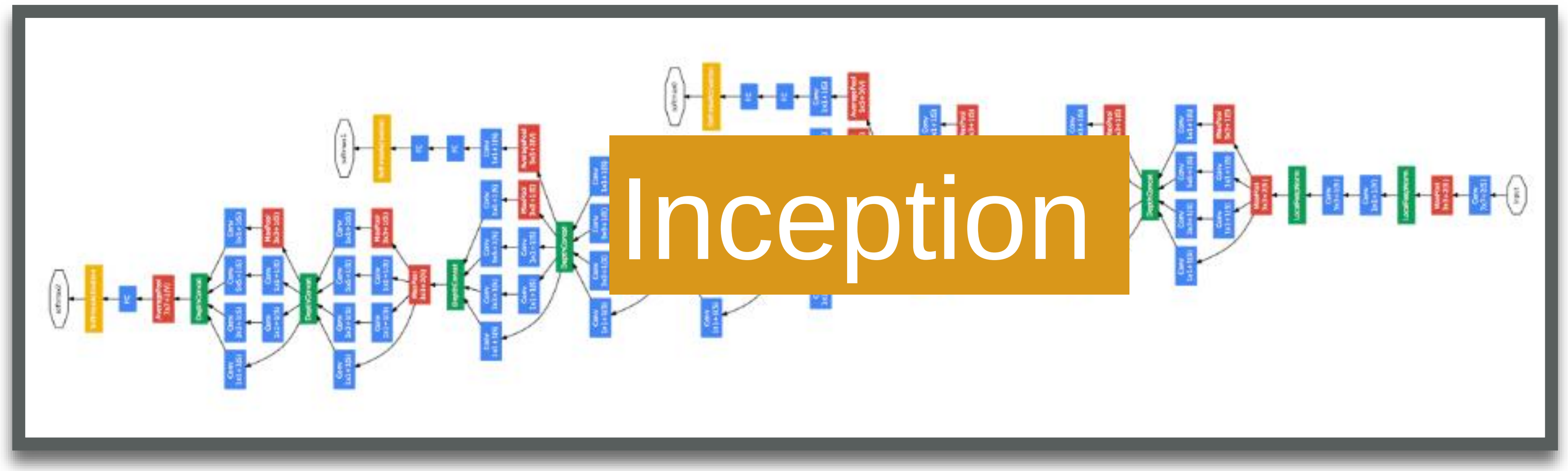
Convolutional Networks

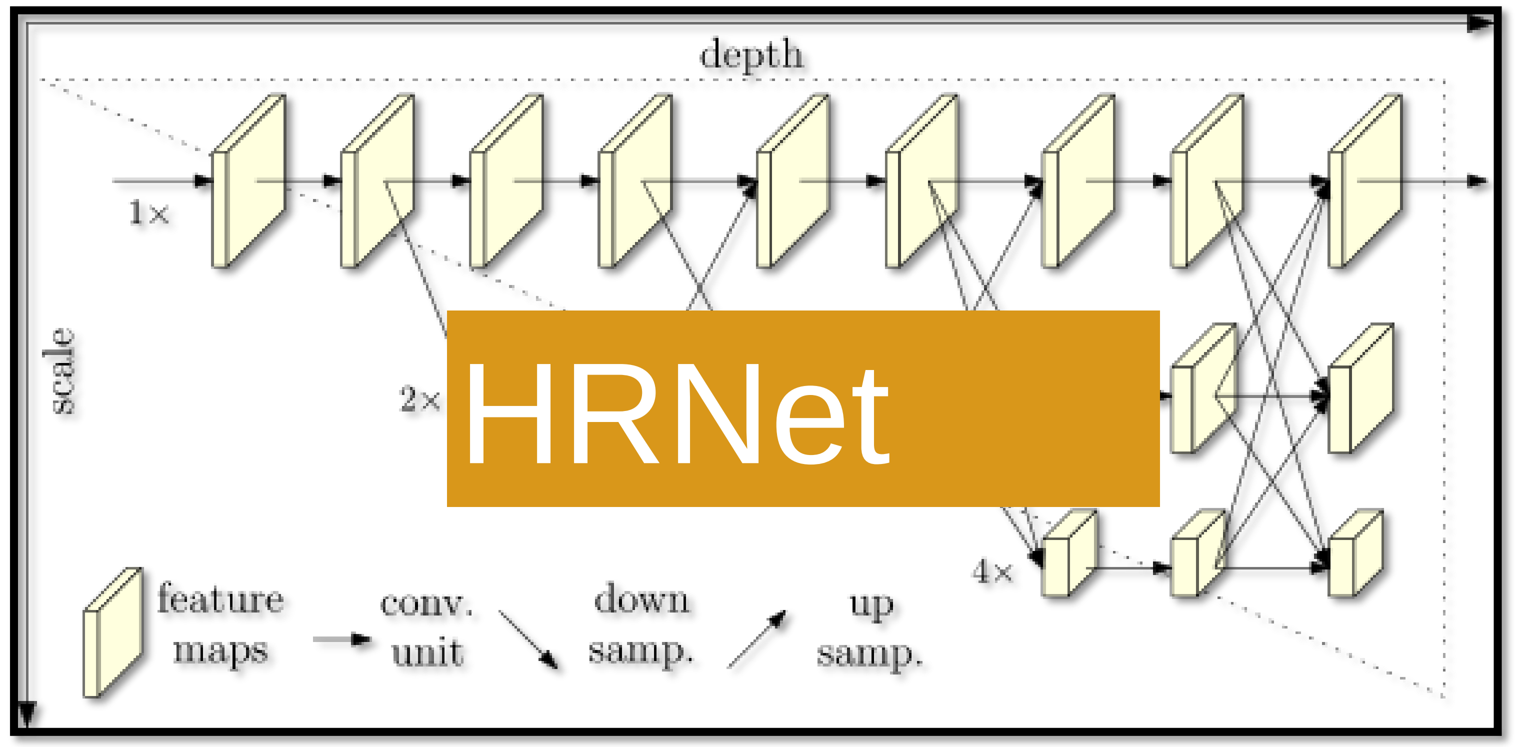


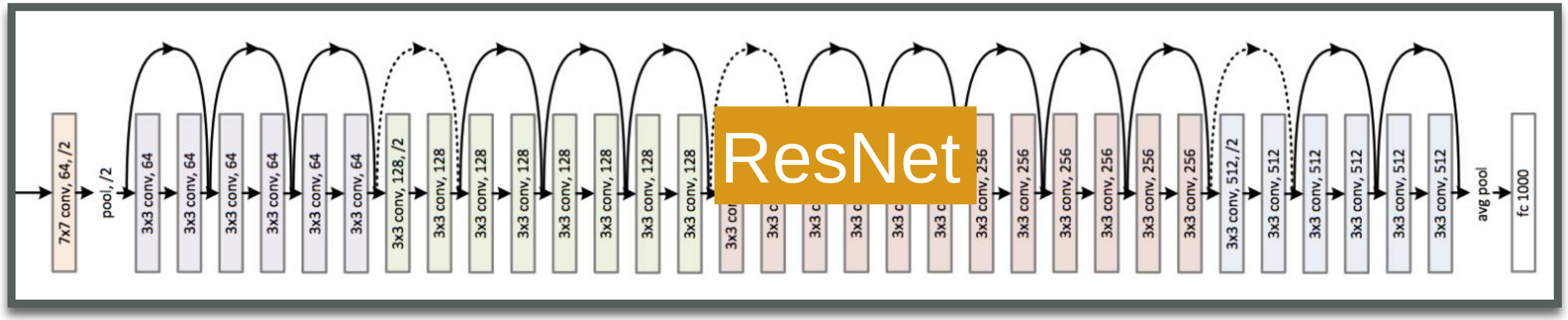












Why architecture matters?

Representation power

Optimization Characteristics

Generalization

Efficiency

Advances in CNN Architecture Design

- **AlexNet**
- ZF-Net
- DSN
- NIN
- **VGG**
- **GoogleNet**
- …

2012-2015

Fast developing stage

Aim for high accuracy

- Highway Networks
- FractalNet
- **ResNet**
- **DenseNet**
- ResNeXt
- Dual Path Network
- …

2015-2017

Mature stage

Aim for simple design principles

2017-Present

Prosperous stage

Aim for better accuracy-speed tradeoff

Light-weighted models

- **MobileNet** (V1, V2, V3)
- **CondenseNet**
- **ShuffleNet** (V1, V2)
- …

Neural Arch. Search

- **NASNet**
- DARTS
- …

Dynamic models

- **MSDNet**
- Block-Drop
- Glance and Focus
- …

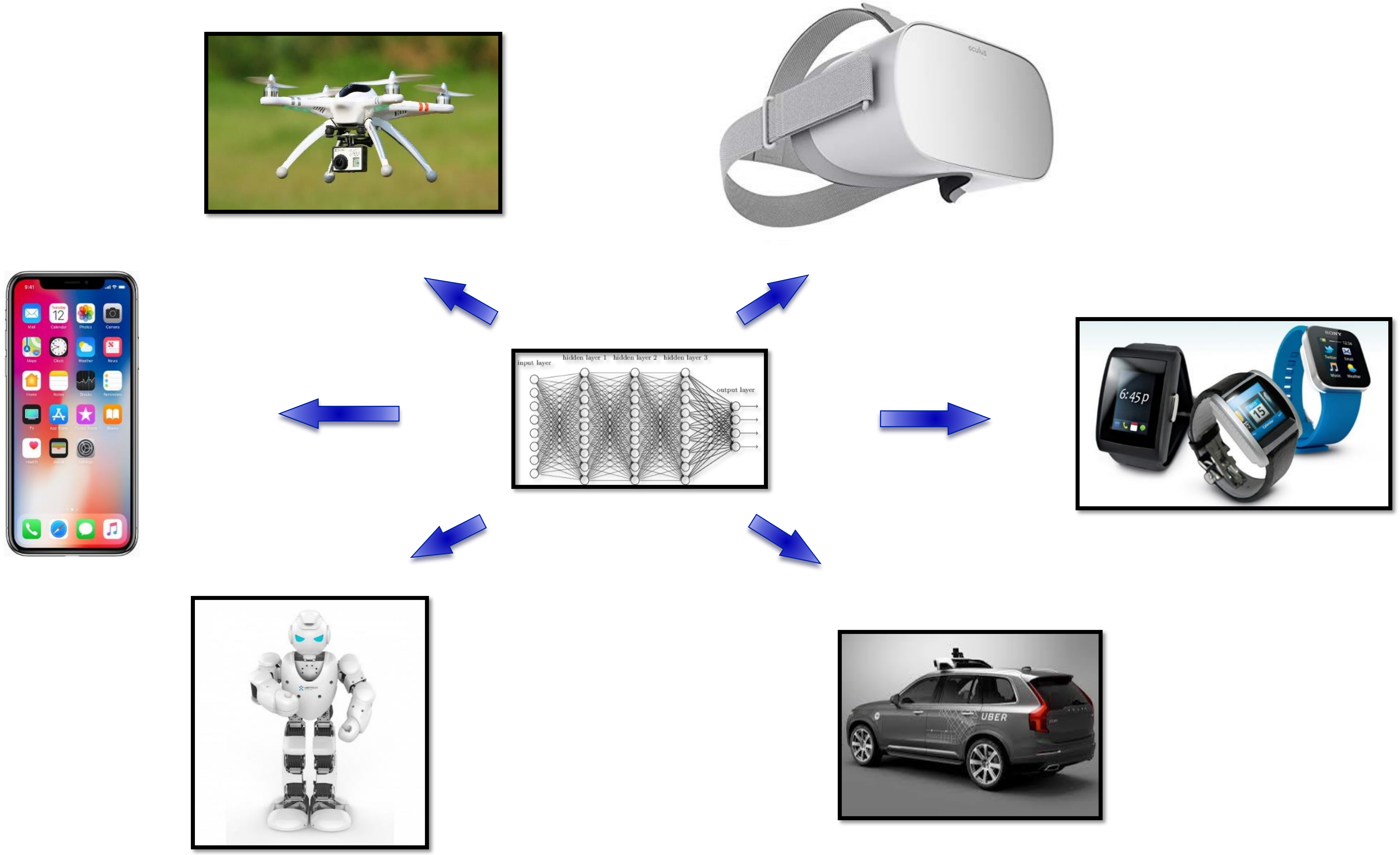
Transformers?!

- **ViT and its variants**
- …

CNNs for Mobile Applications

Goal:

- Low compute
- Low latency
- Low memory cost



Outline

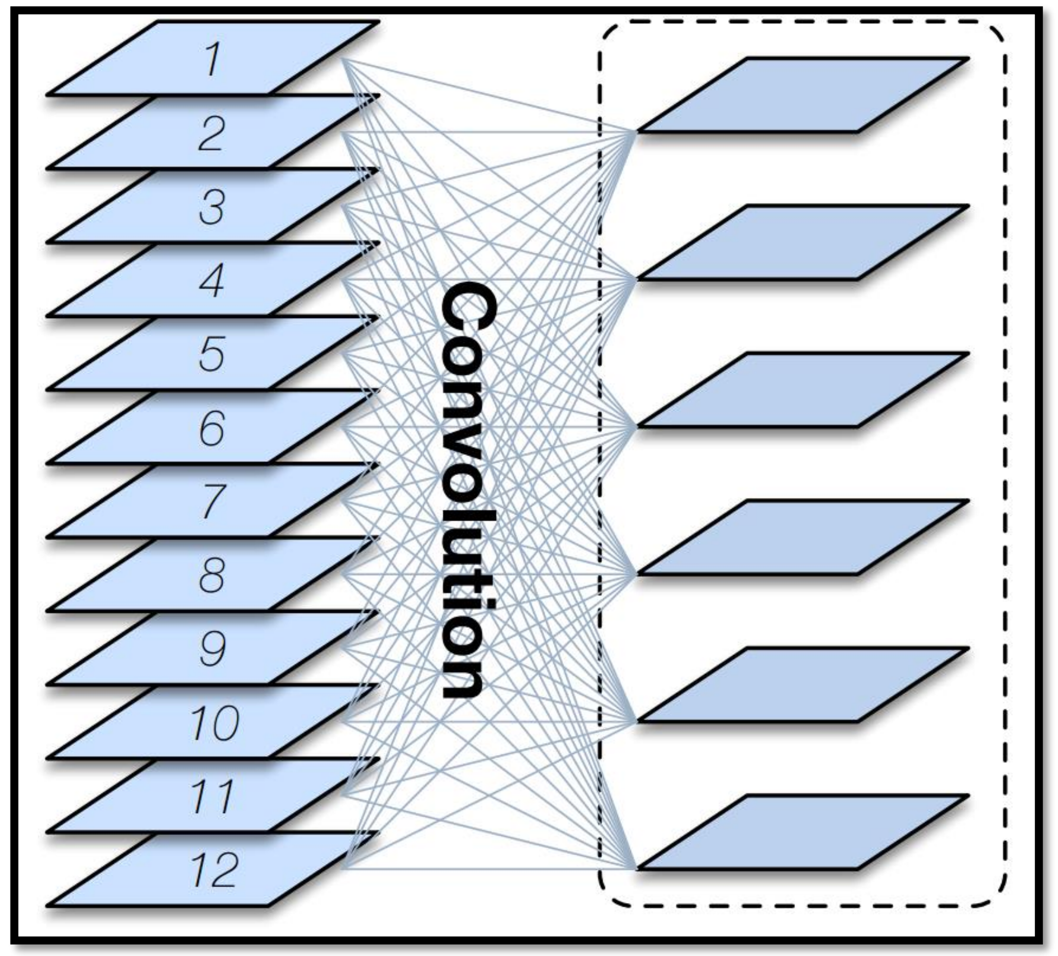
1. Overview of CNN backbones
2. Architecture design for mobile CNNs
3. Dynamic CNNs for mobile applications

Group Convolution

Main Idea:

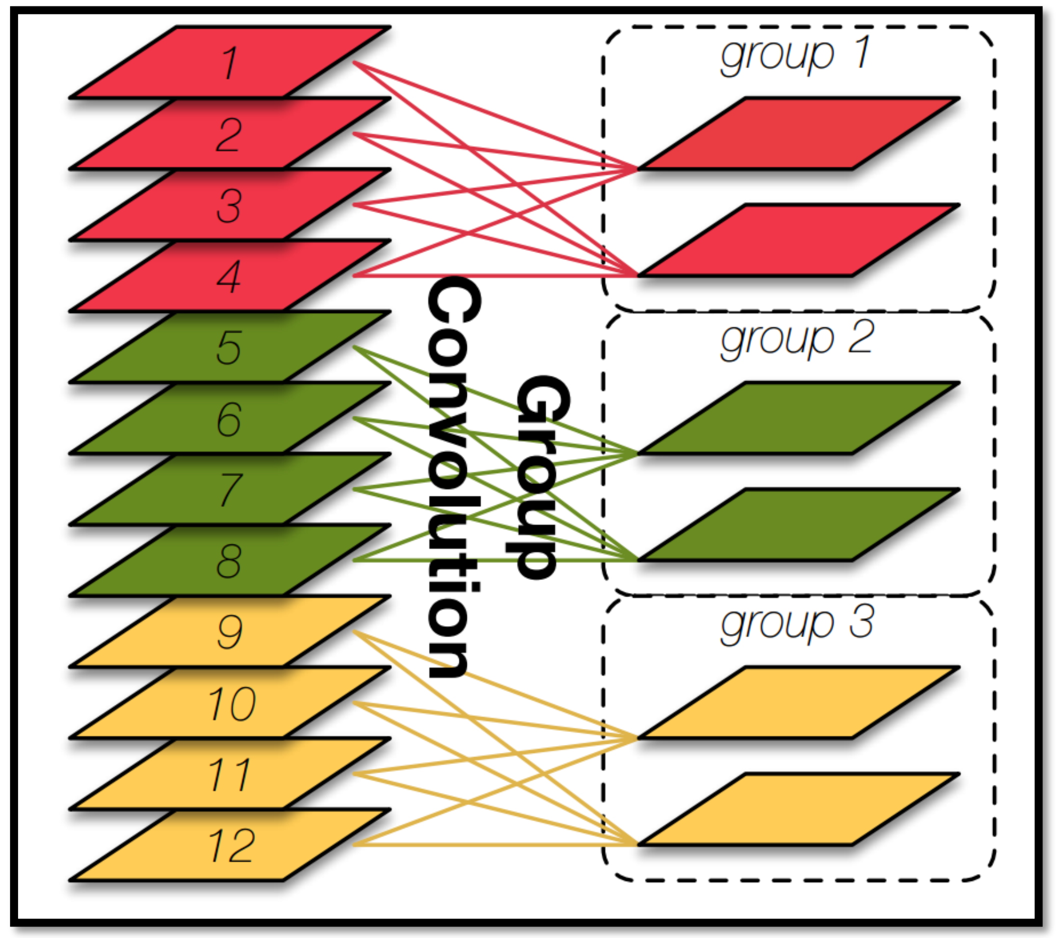
✓ Split convolution into multiple groups

Standard Convolution



$$O(C \times C)$$

Group Convolution



$$O\left(\frac{C \times C}{G}\right)$$

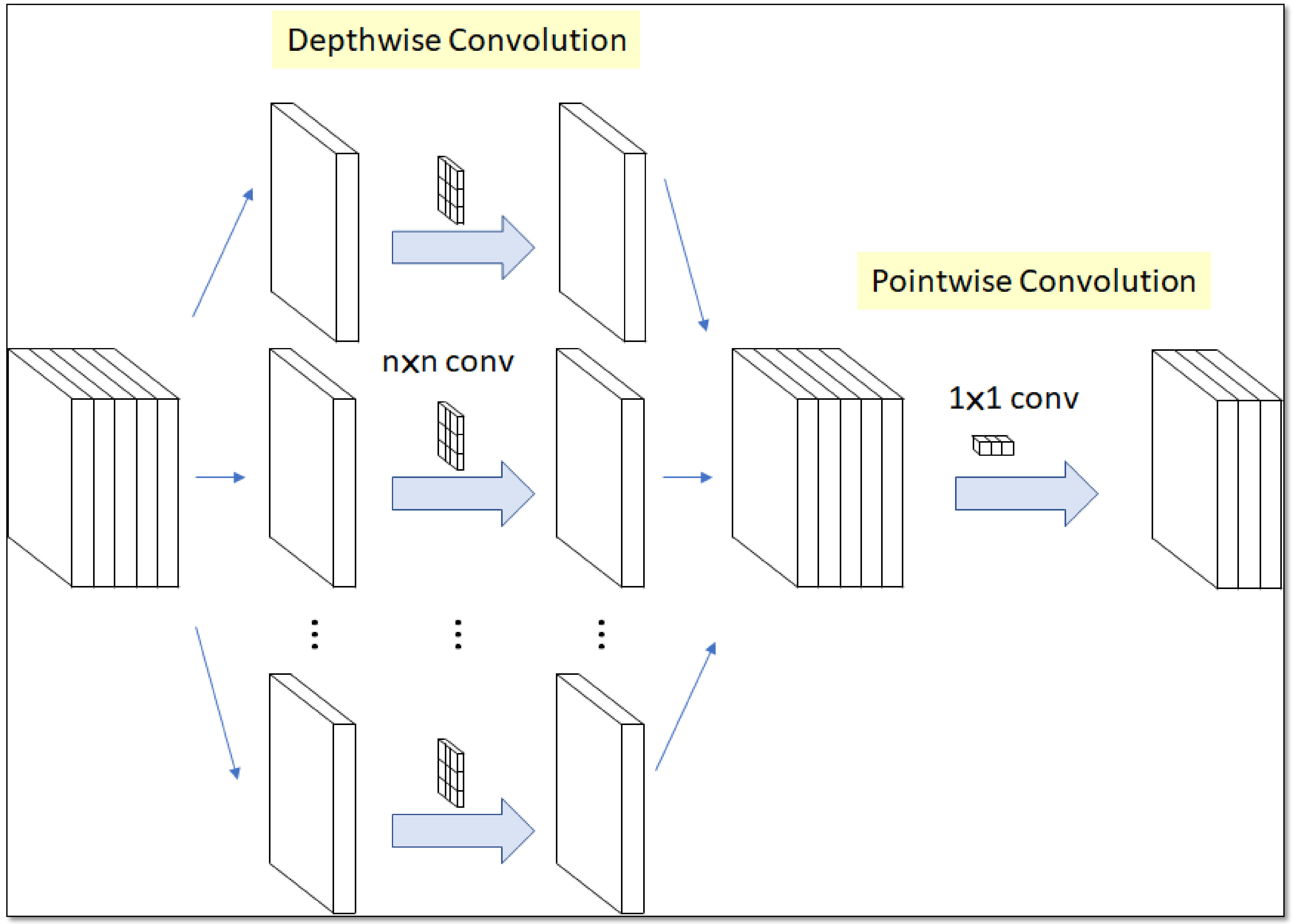
CNNs using Group Convolution:

✓ AlexNet (Krizhevsky et al, NIPS'12)
✓ ResNeXt (Xie et al, CVPR'17)
✓ CondenseNet (Huang et al, CVPR'18)
✓ ShuffleNet (Zhang et al, CVPR'18)
✓ …

Depth-wise Separable Convolution (DSC)

Main Idea:

✓ Split convolution into multiple groups, each group has one channel



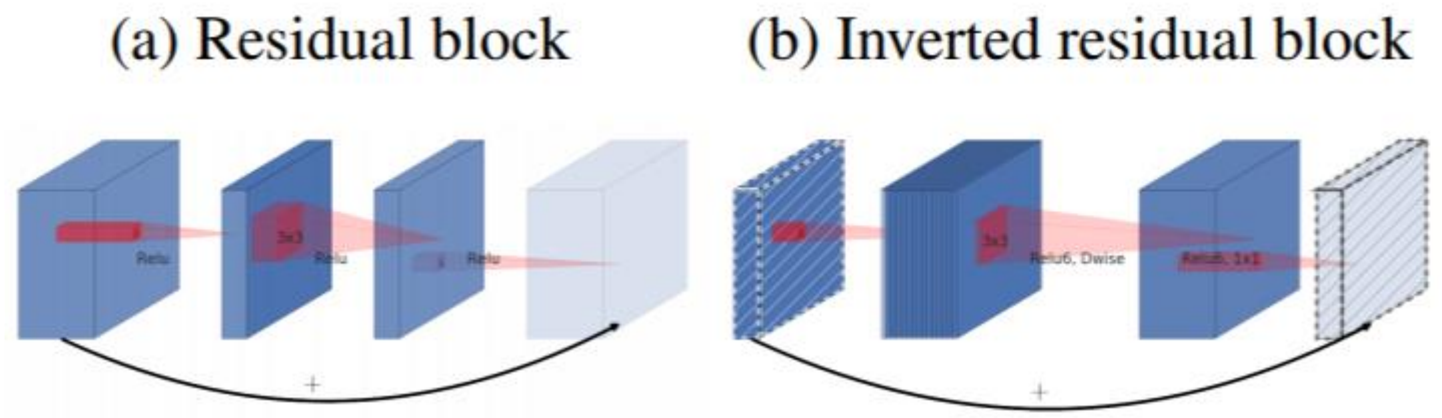
Networks using DSC:

✓ Xception (Chollet, CVPR'17)
✓ MobileNet (Howard et al, CVPR'18)
✓ MobileNet V2 (Sandler et al, 2018)
✓ ShuffleNet V2 (Ma et al, CVPR'19)
✓ NasNet (Zoph, CVPR'18)
✓ …

MoblieNets

MobileNet v1 [Howard et al, CVPR'18]

✓ Depth-wise separable convolution



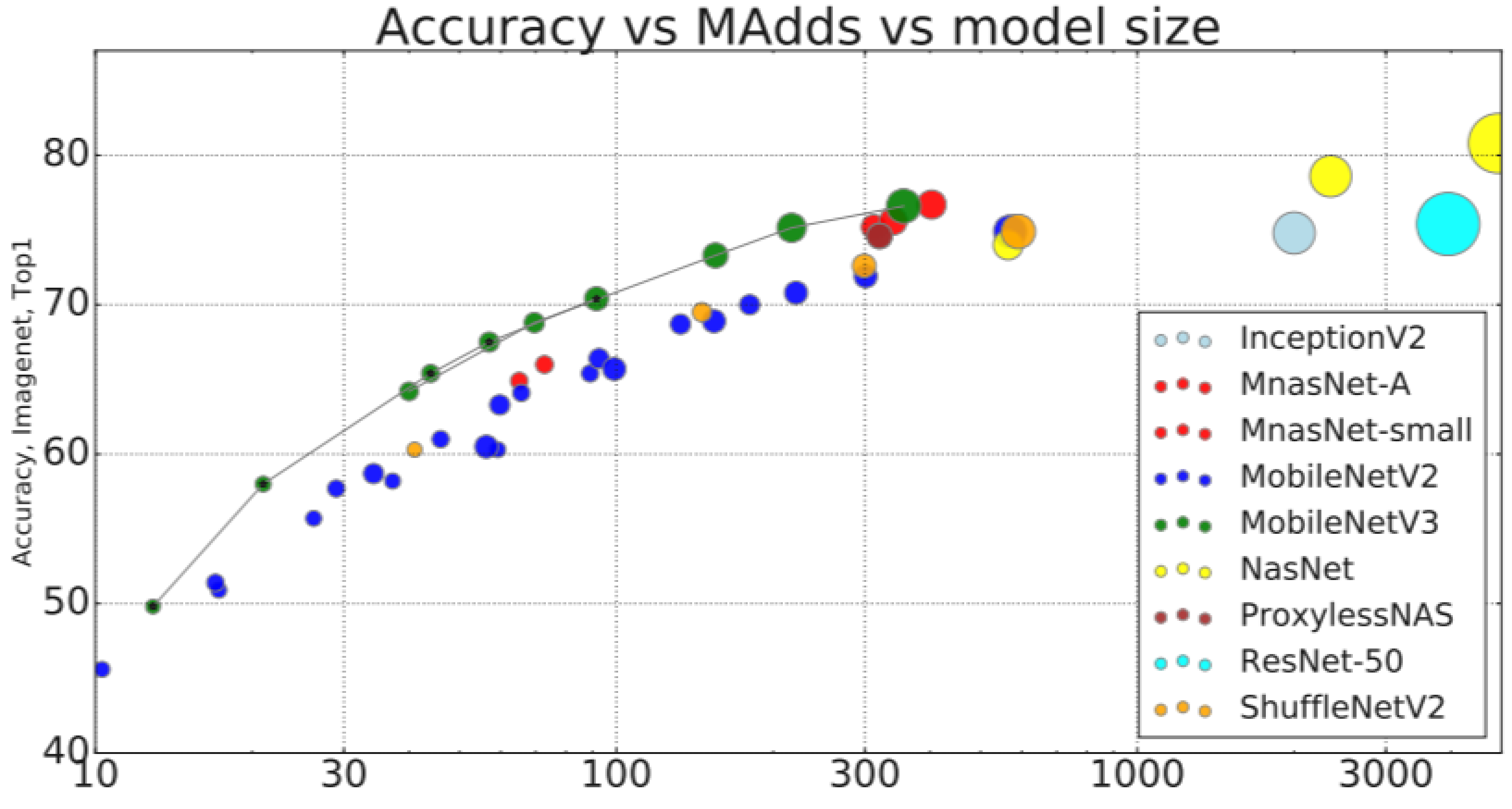
Inverted Residuals in MoblieNet v2

MobileNet v2 [Sandler et al, CVPR'19]

✓ Inverted Residuals and Linear Bottlenecks

MobileNet v3 [Howard et al, CVPR'19]

✓ Introducing neural architecture search



Comparison of MoblieNet v3 and other models

ShuffleNets

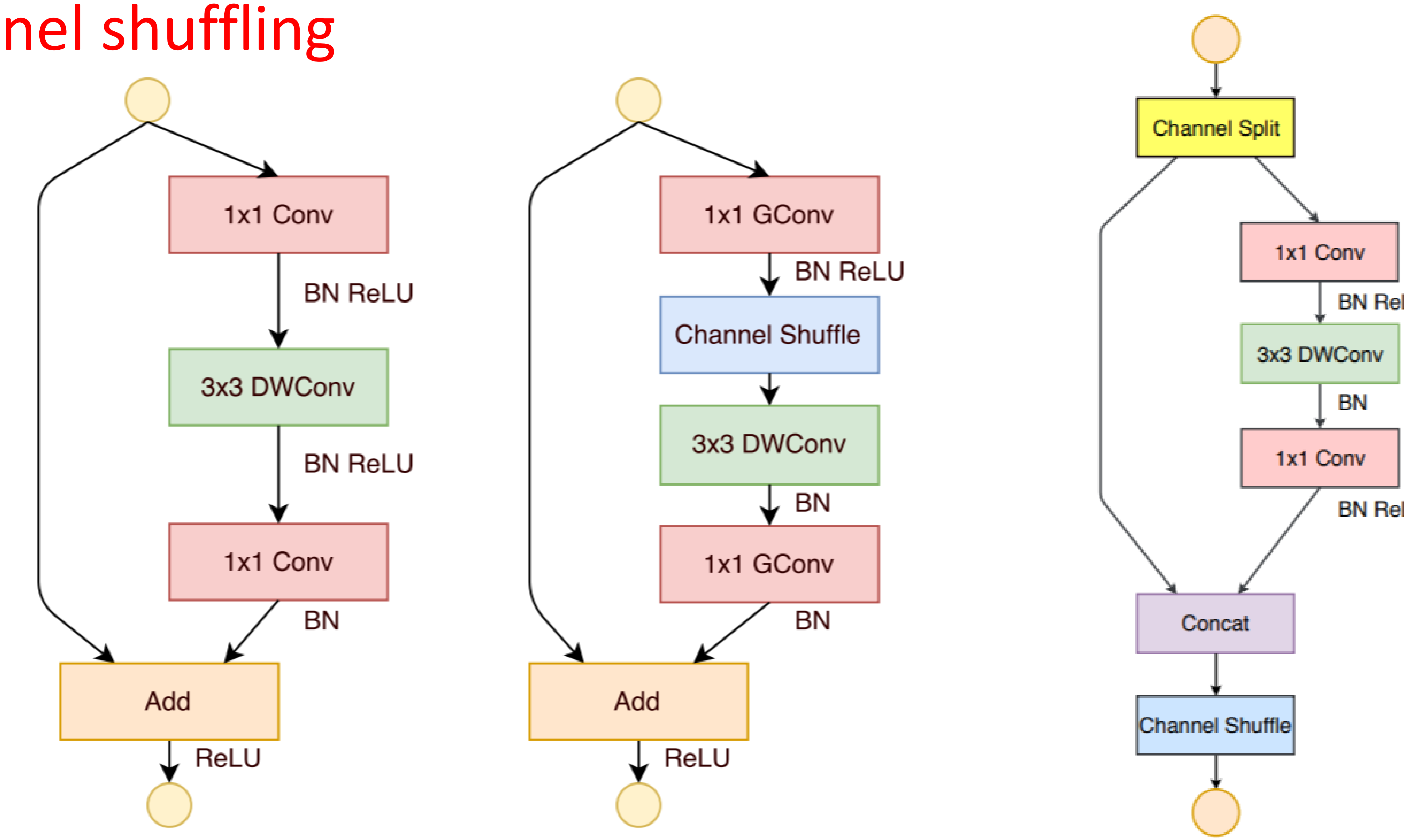
ShuffleNet v1 [Zhang et al, CVPR'18]

✓ Consecutive group convolution with channel shuffling

ShuffleNet v2 [Ma et al, ECCV'18]

✓ Feature reuse with dense connection

✓ Special design for hardware efficiency



Residual unit with DWC

ShuffleNet v1 : Group Conv with channel shuffle

ShuffleNet v2 : Channel split and dense connection

CondenseNets

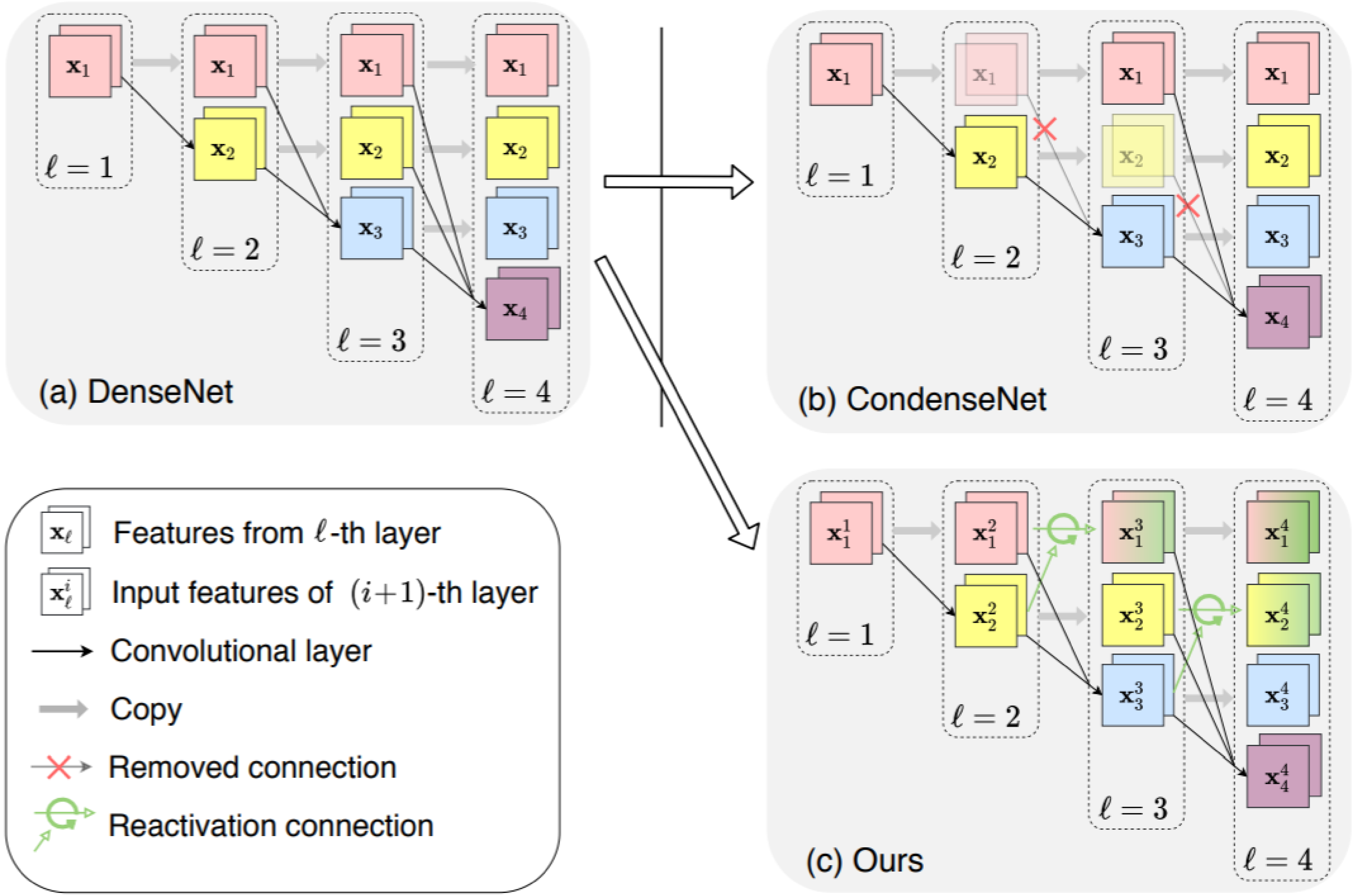
CondenseNets v1 [Huang et al, CVPR'18]

✓ Sparsified dense connections

✓ Learned group convolution

CondenseNets v2 [Yang et al, CVPR'21]

✓ Feature reactivation

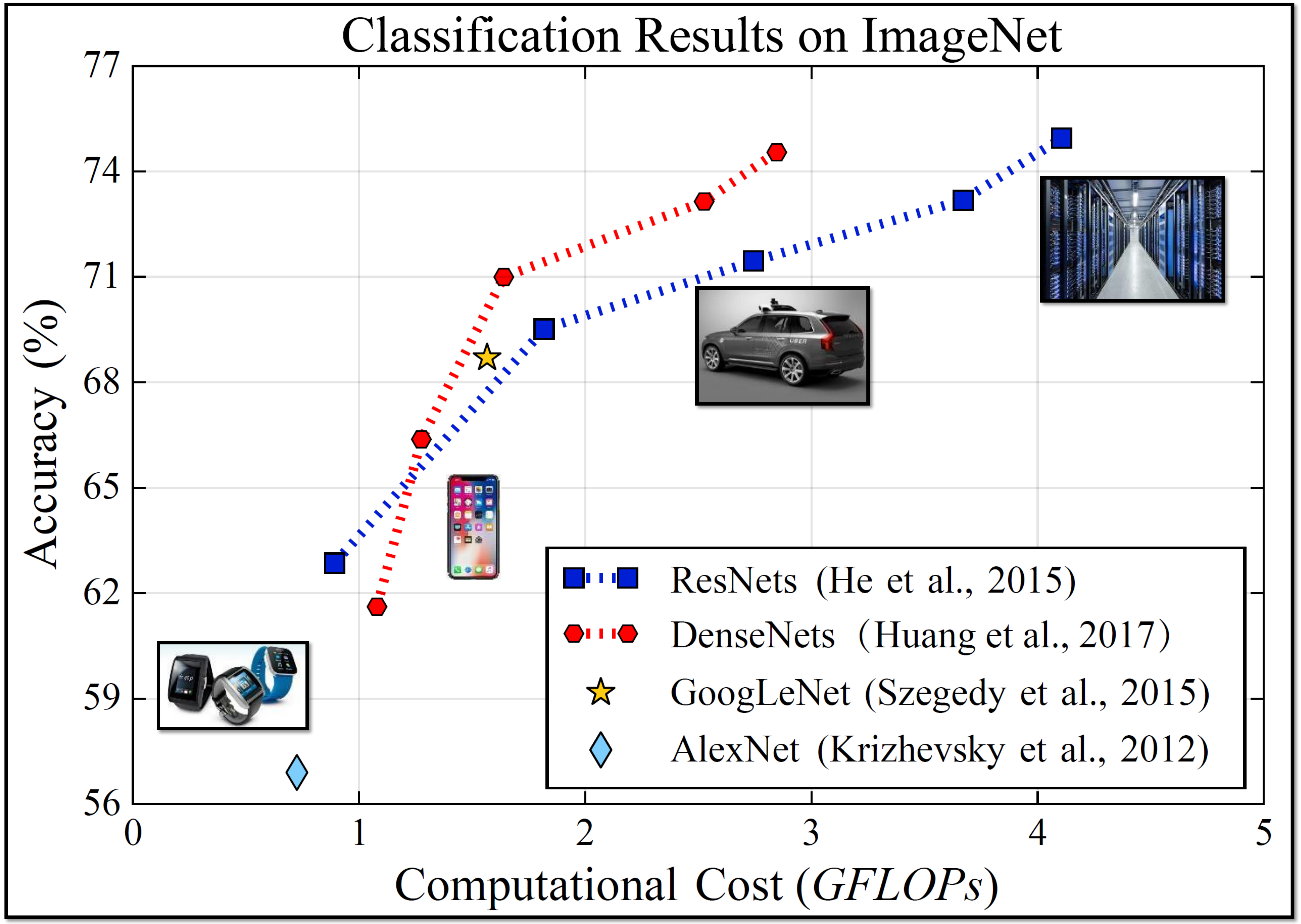


Outline

1. Overview of CNN backbones
2. Architecture design for mobile CNNs
3. Dynamic CNNs for mobile applications

Why do we need dynamic neural networks?

Accuracy-Time Tradeoff



Bigger is better

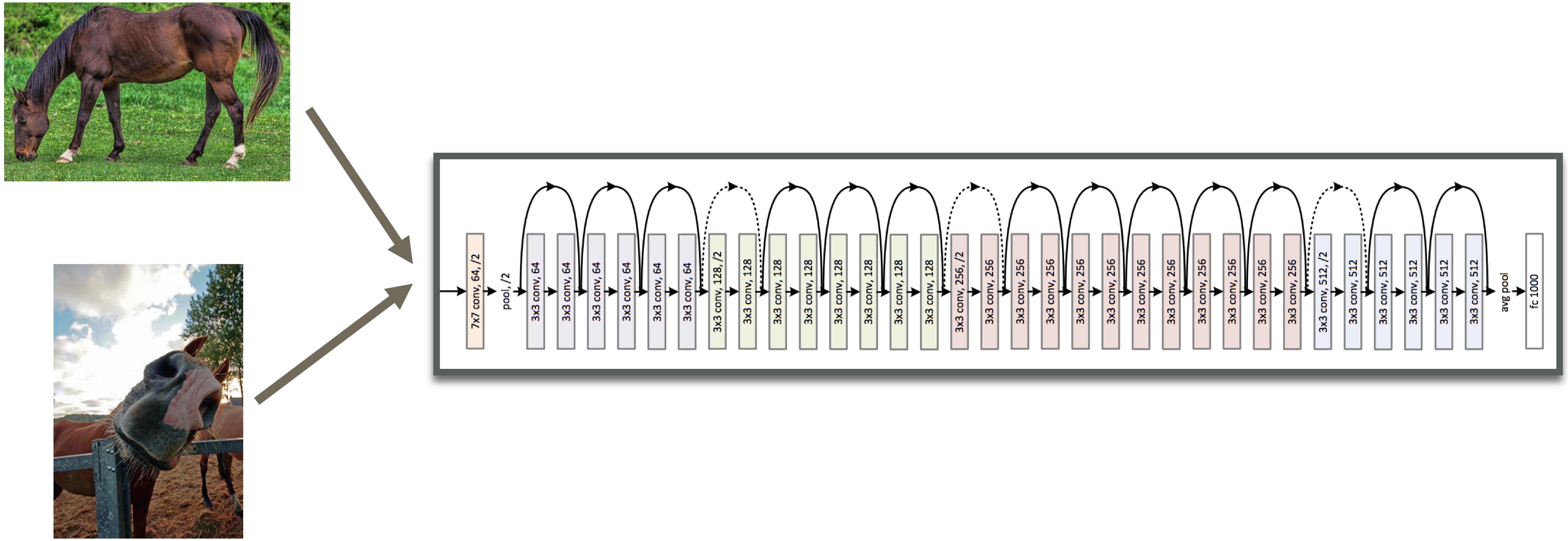
Bigger models are needed for those noncanonical images.

**Photo Courtesy of Pixel Addict (CC BY-ND 2.0)*

Bigger is better

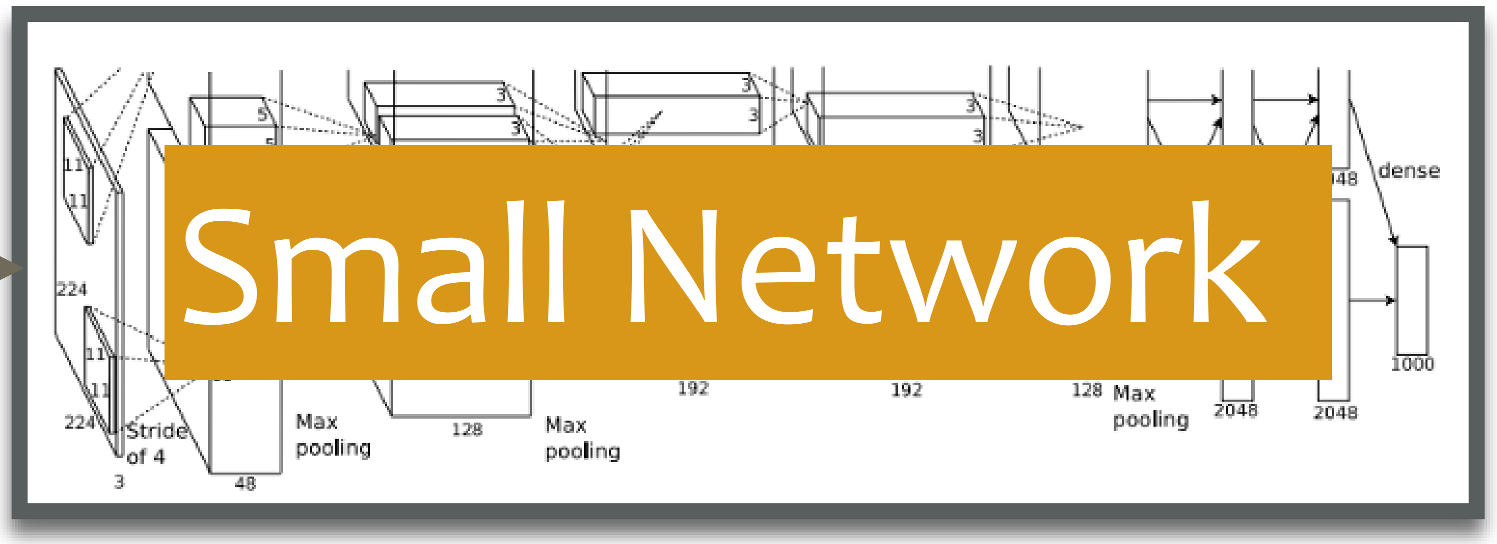
*Photo Courtesy of Willian Doyle(CC BY-ND 2.0)

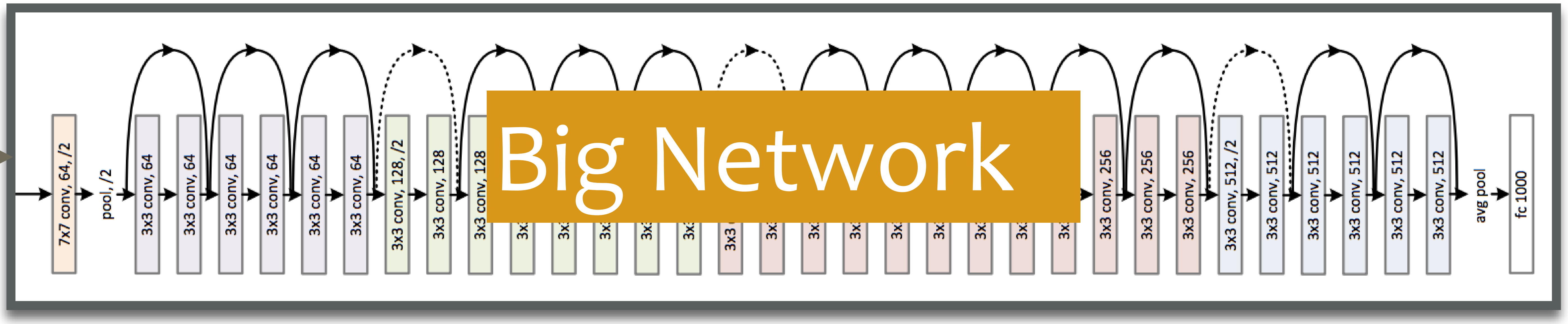
*Why do we use the **same** expensive model for all images?*



The model architecture (depth, width, etc) should be

conditioned on the input!

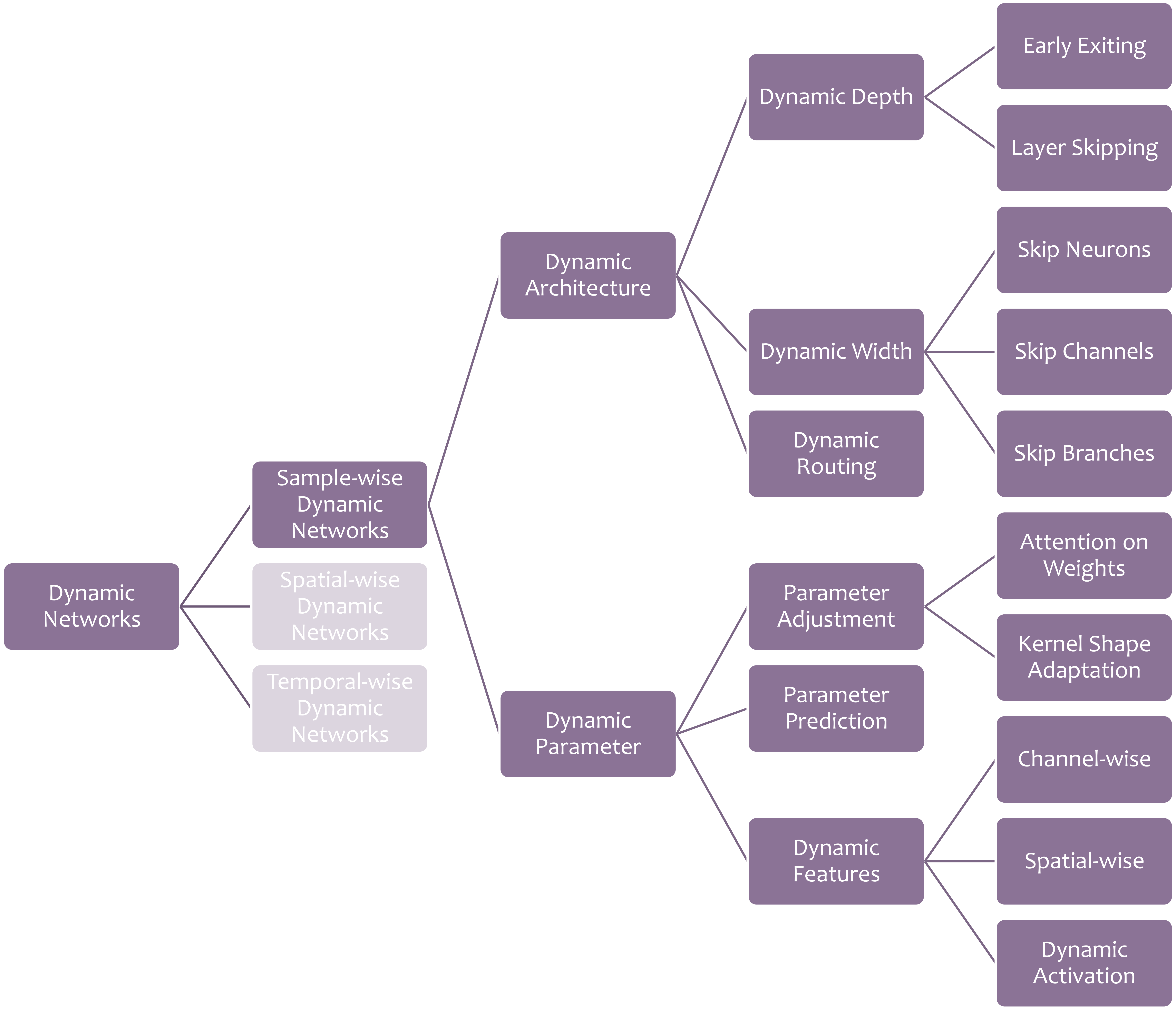




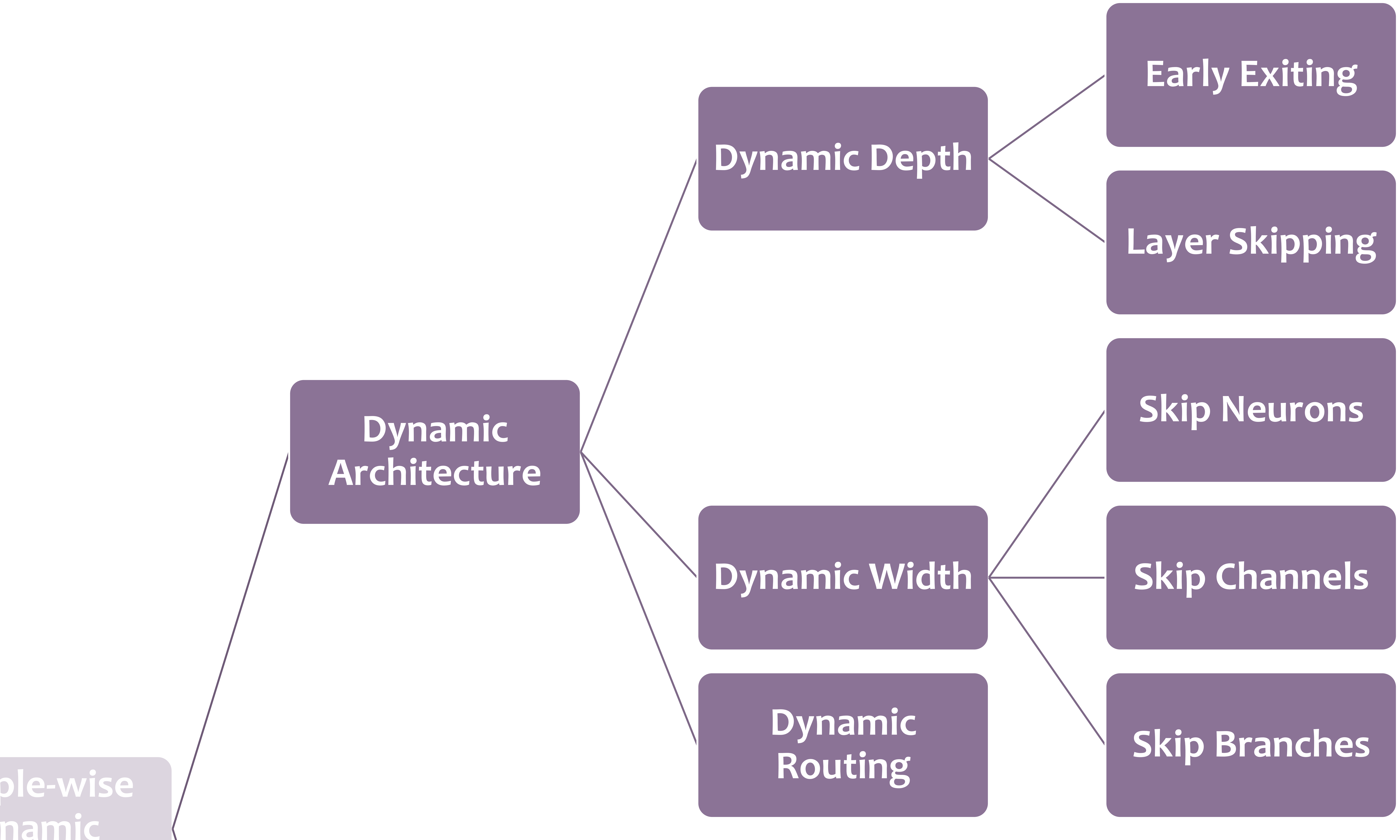
Outline

1. Overview of CNN backbones
2. Architecture design for mobile CNNs
3. Dynamic CNNs for mobile applications
   A. Sample-wise Dynamic Networks
   B. Spatial-wise Dynamic Networks
   C. Temporal-wise Dynamic Networks

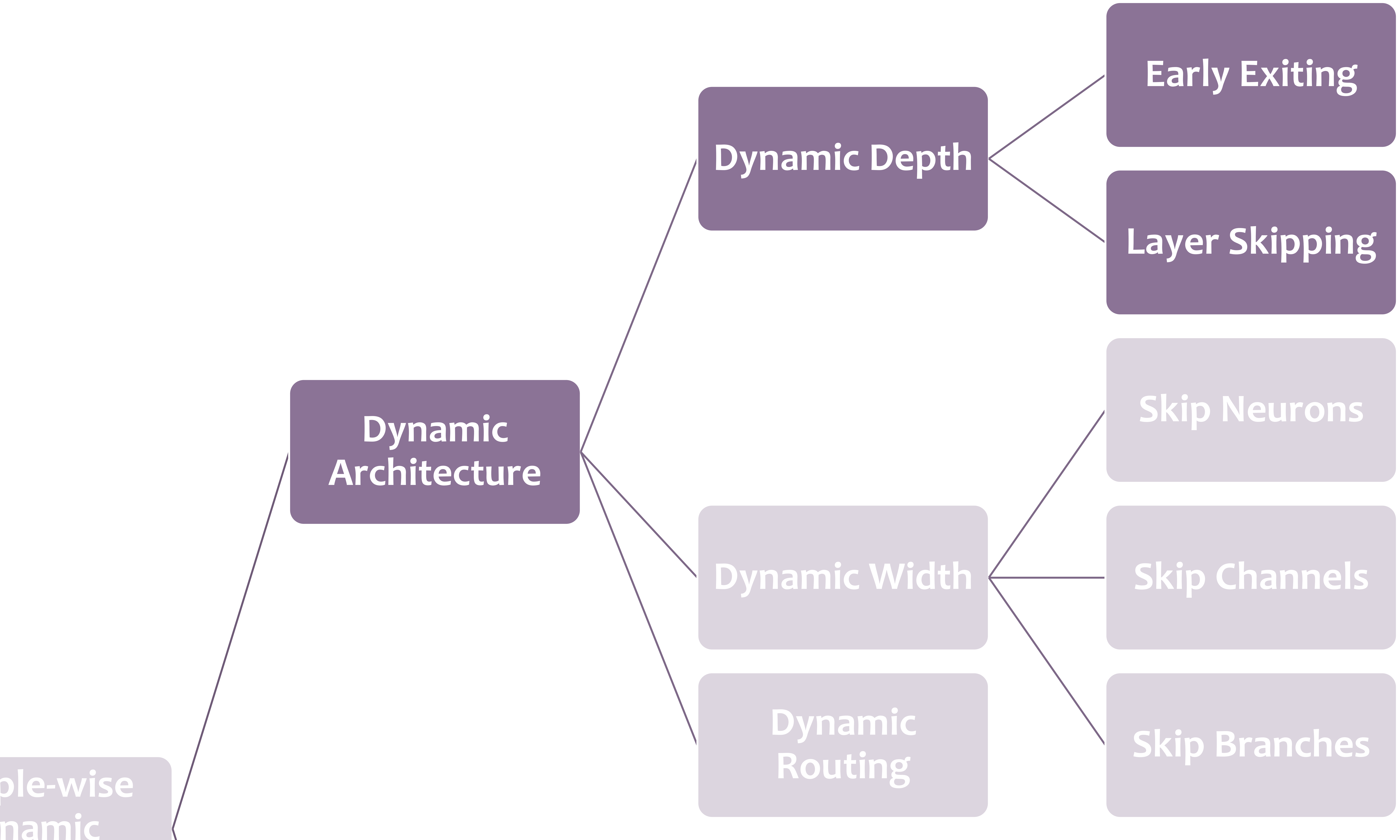
Sample-wise Dynamic Neural Networks



Sample-wise Dynamic Neural Networks

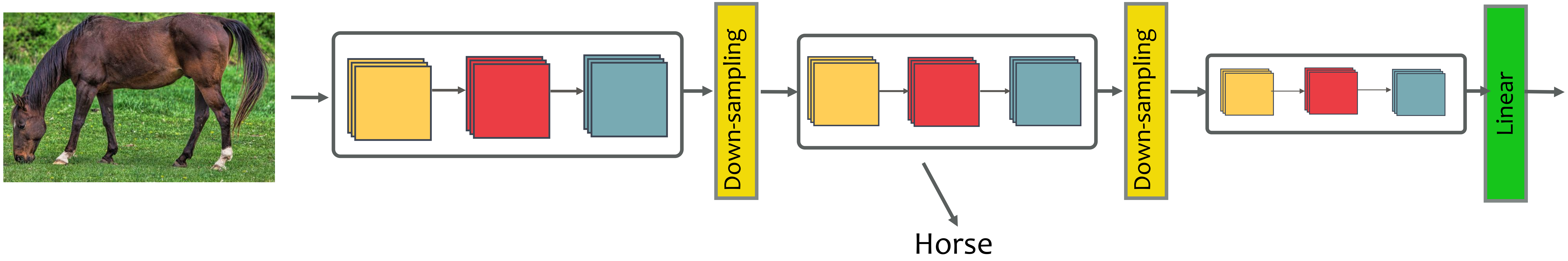


Sample-wise Dynamic Neural Networks



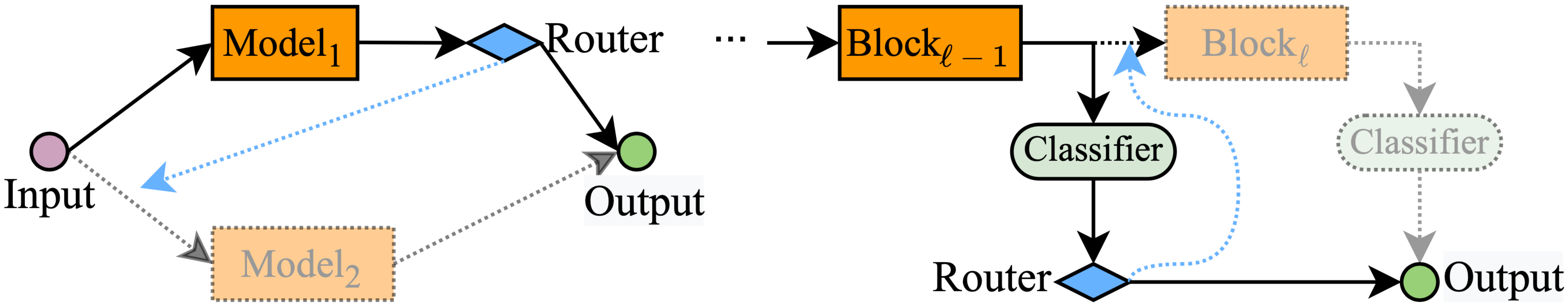
Dynamic Depth: Early Exiting

Dynamic Depth: Early Exiting



Early Exiting: Two Implementations

清華大學 Tsinghua University



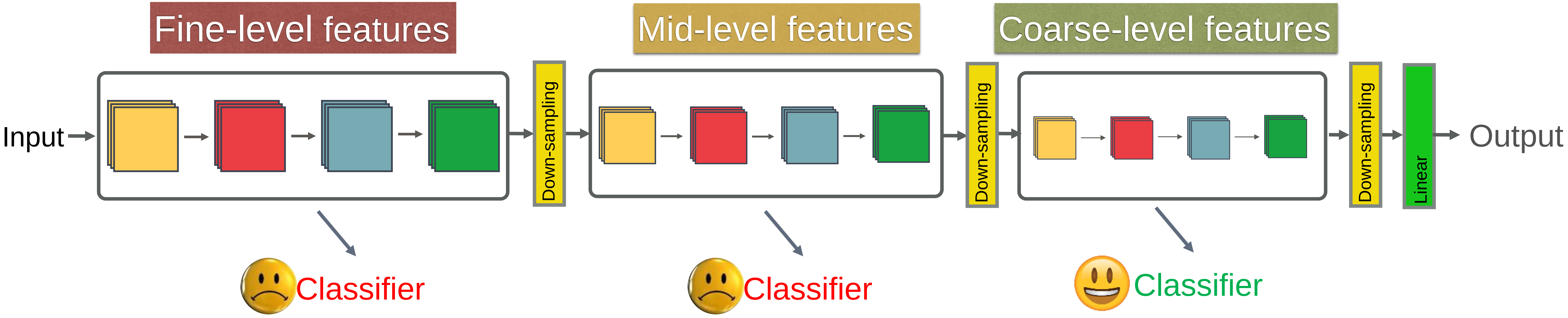
(a) Cascading of models. (b) Network with intermediate classifiers.

- Park, E., Kim, D., Kim, S., Kim, Y. D., Kim, G., Yoon, S., & Yoo, S. (2015, October). Big/little deep neural network for ultra low power inference. In *2015 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ ISSS)* (pp. 124-132). IEEE.
- Teerapittayanon, S., McDanel, B., & Kung, H. T. (2016, December). Branchynet: Fast inference via early exiting from deep neural networks. In 2016 23rd International Conference on Pattern Recognition (ICPR) (pp. 2464-2469). IEEE.
- Bolukbasi, T., Wang, J., Dekel, O., & Saligrama, V. (2017, July). Adaptive neural networks for efficient inference. In International Conference on Machine Learning (pp. 527-536). PMLR.

Dynamic Depth: Early Exiting

Tsinghua University

A challenge: Intermediate classifiers may interfere with each other

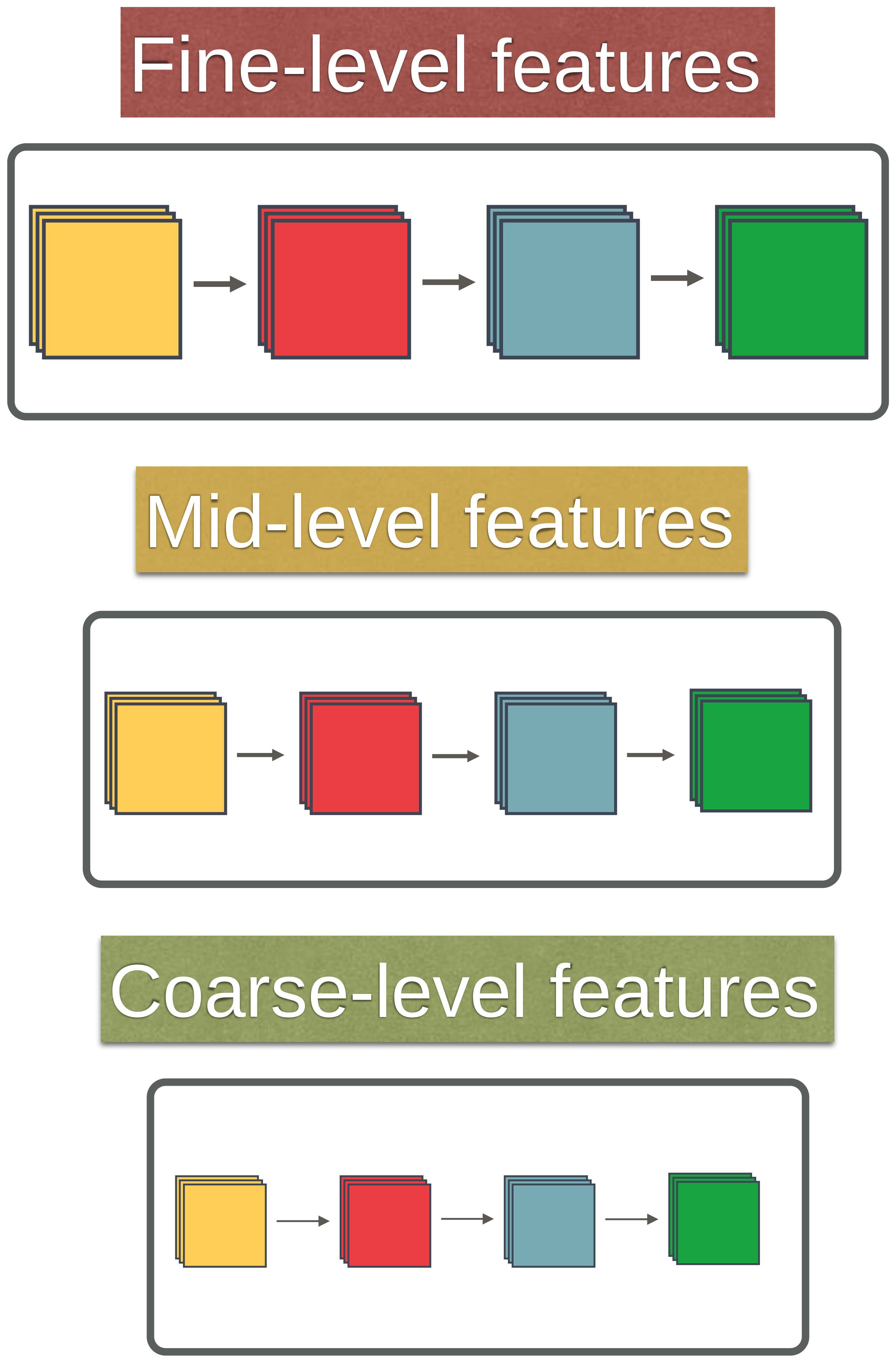


Classifiers **only work well** on **coarse-scale** feature maps

Nearly **all computation** has been done **before** getting a **coarse** feature

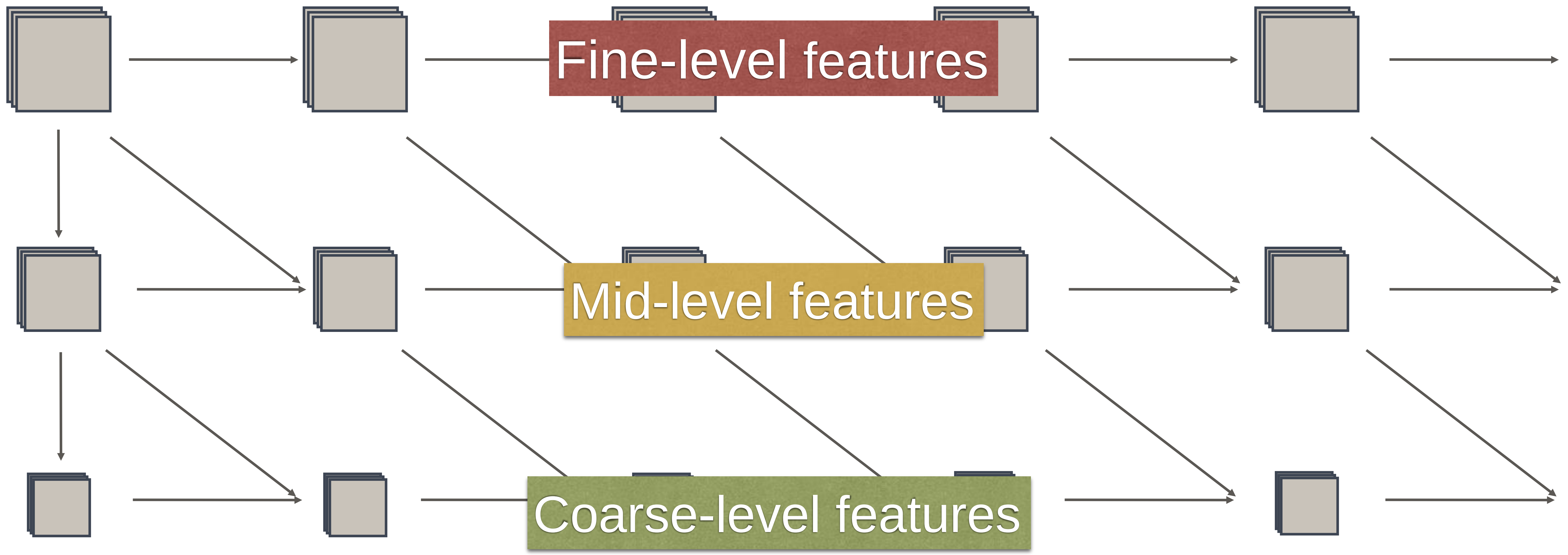
Dynamic Depth: Early Exiting

Solution: Multi-scale Architecture



Dynamic Depth: Early Exiting

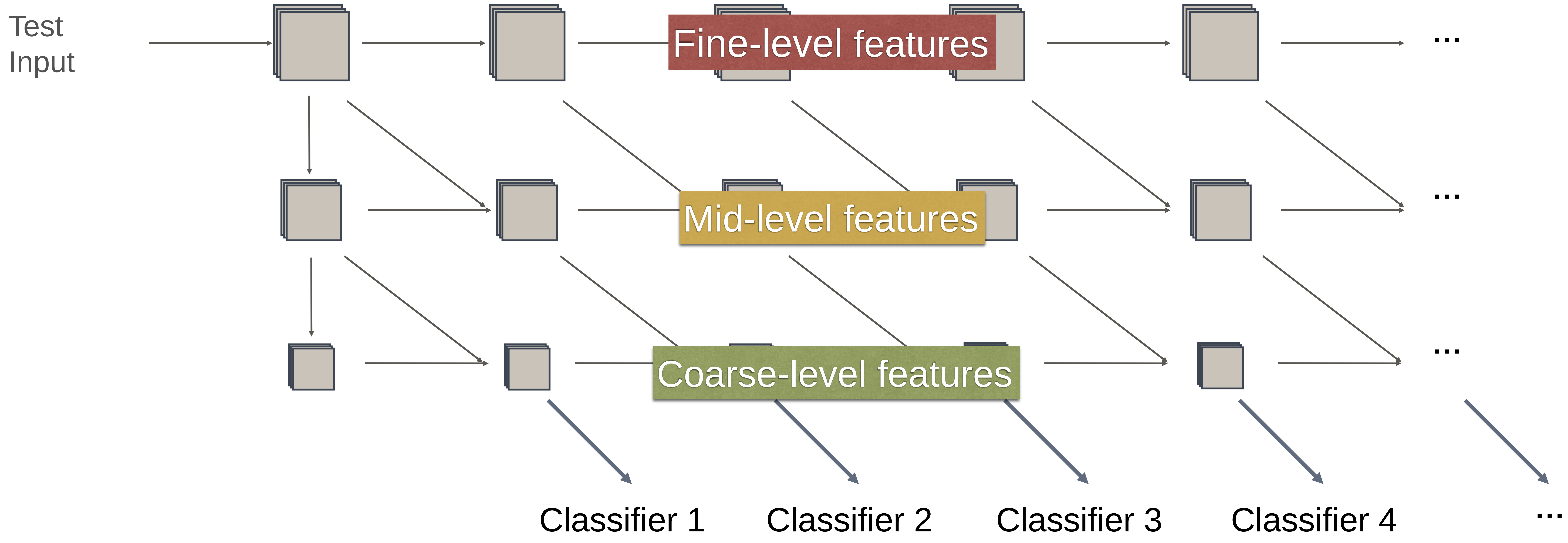
Solution: Multi-scale Architecture



- Huang, G., Chen, D., Li, T., Wu, F., van der Maaten, L., & Weinberger, K. (2018, February). Multi-Scale Dense Networks for Resource Efficient Image Classification. In International Conference on Learning Representations.

Dynamic Depth: Early Exiting

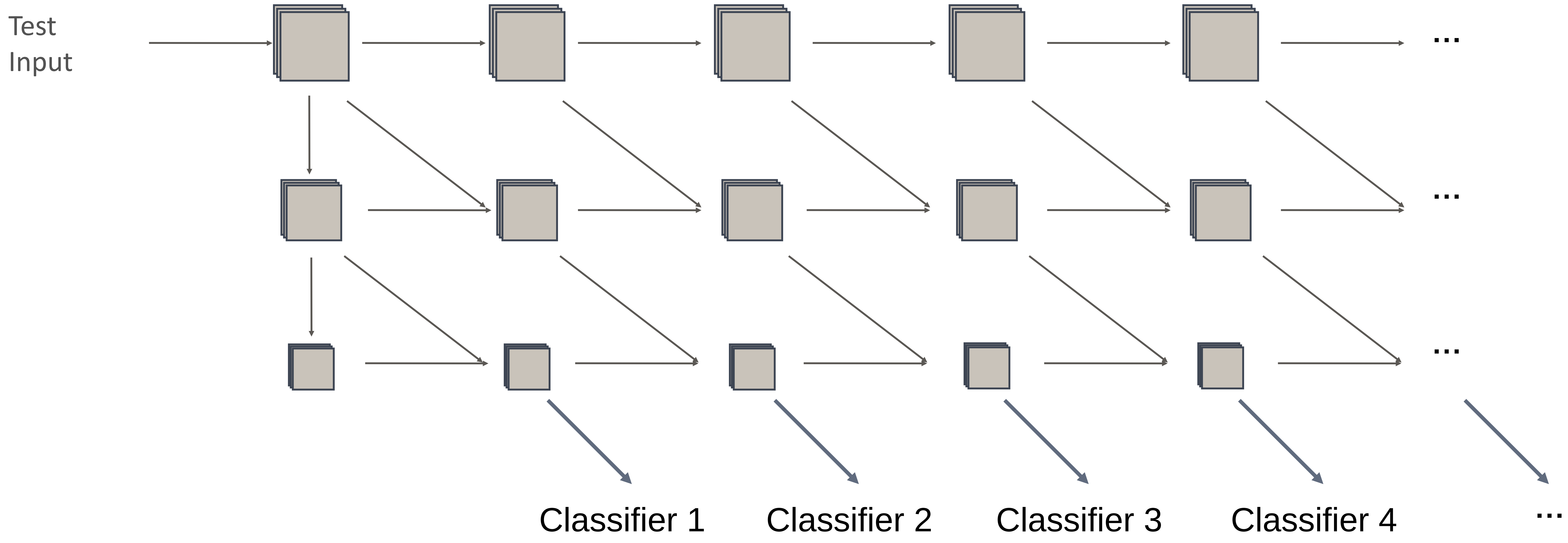
Solution: Multi-scale Architecture



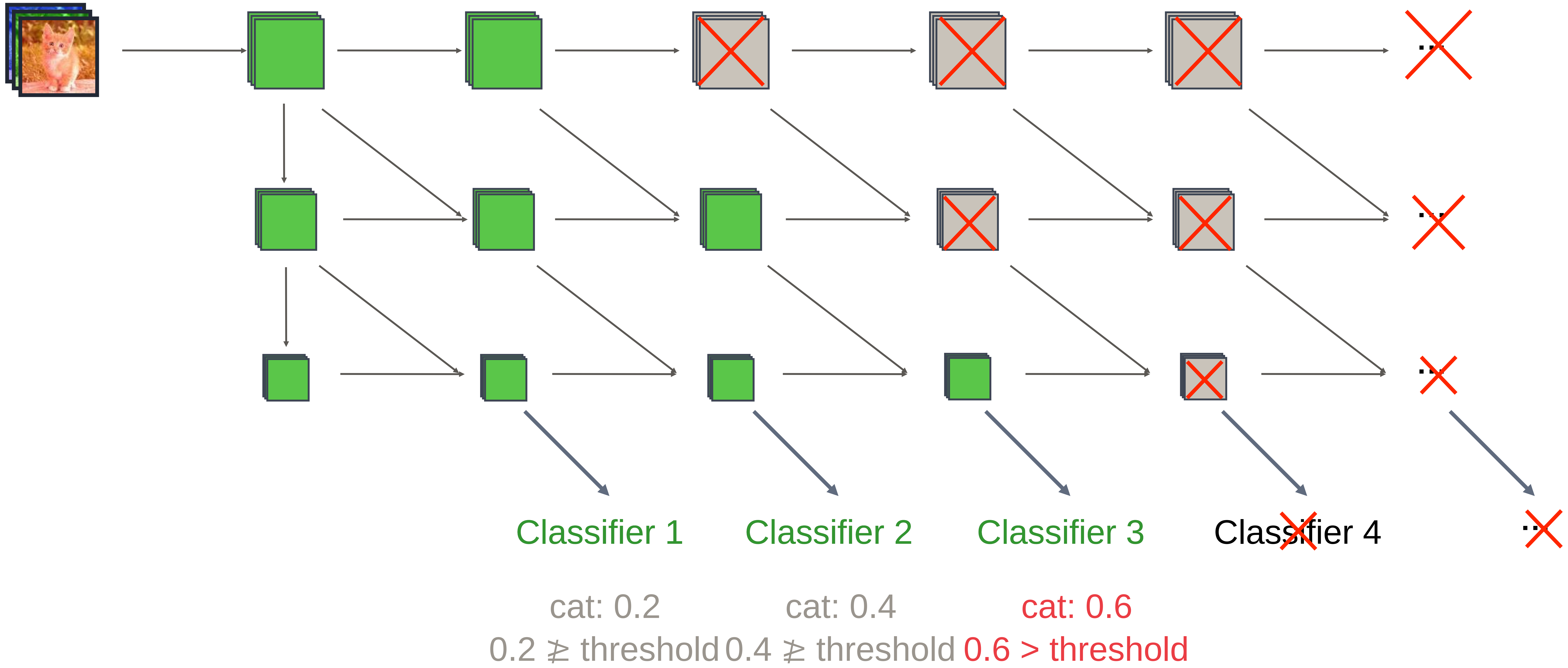
Classifiers only operate on high level features!

- Huang, G., Chen, D., Li, T., Wu, F., van der Maaten, L., & Weinberger, K. (2018, February). Multi-Scale Dense Networks for Resource Efficient Image Classification. In International Conference on Learning Representations.

Multi-scale densenet

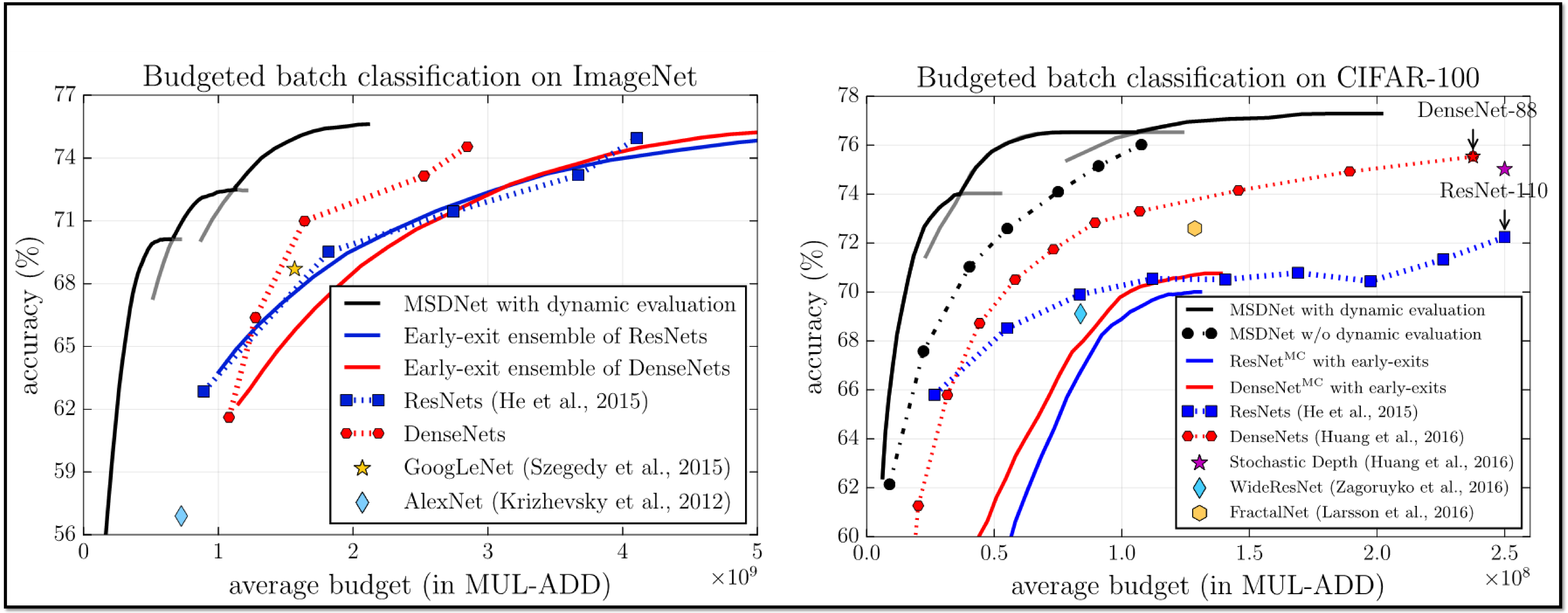


Multi-scale densenet



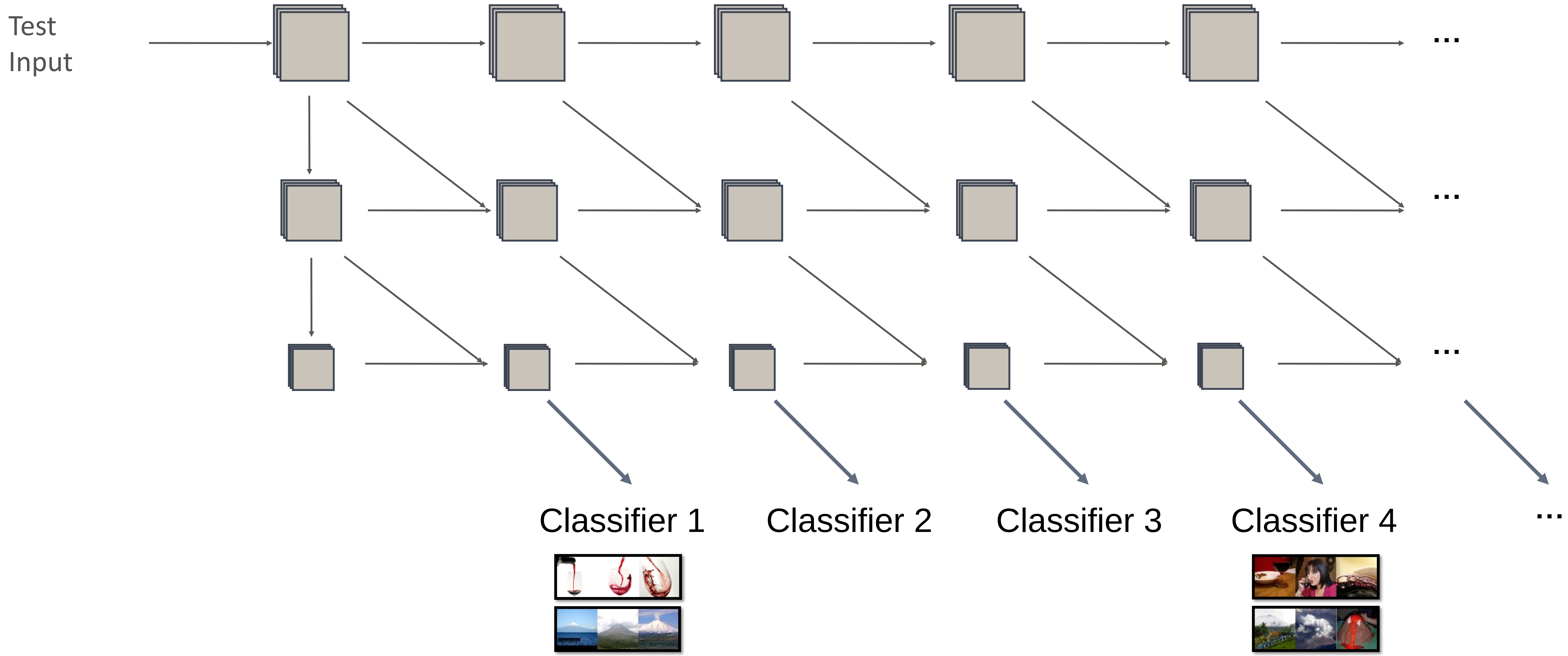
Multi-Scale DenseNet

Results



2x-5x speedup over DenseNet

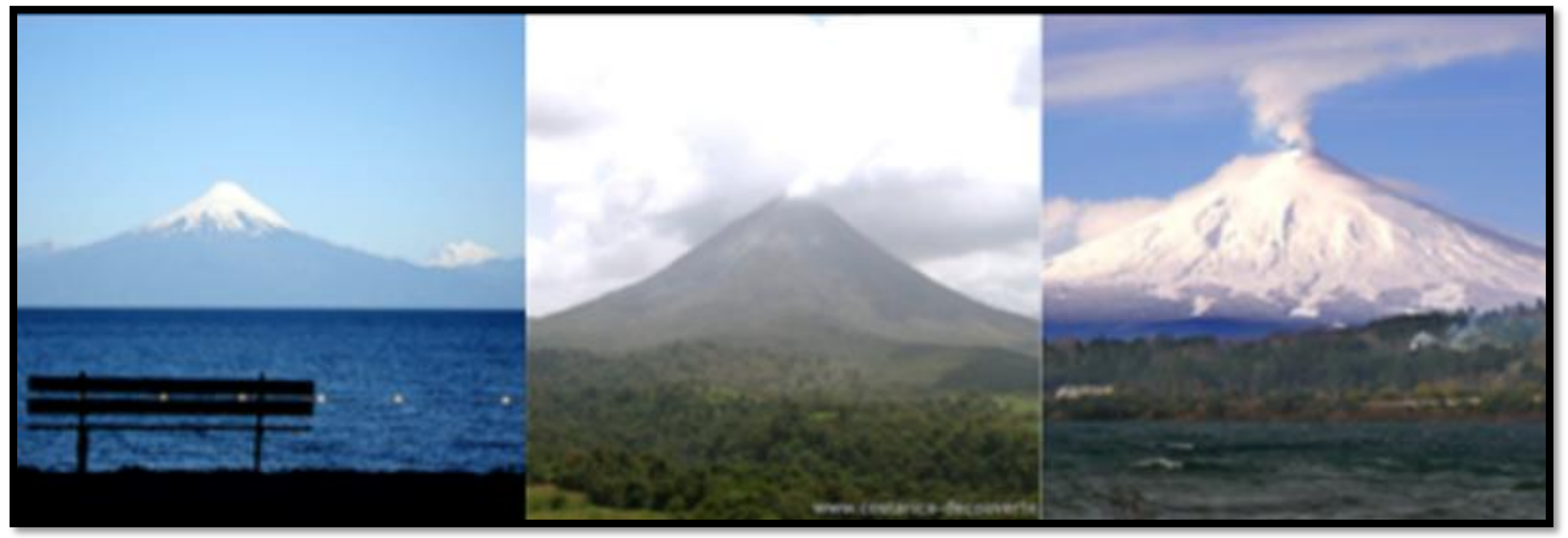
Visualization



Visualization

Class:
red wine

Class:
volcano



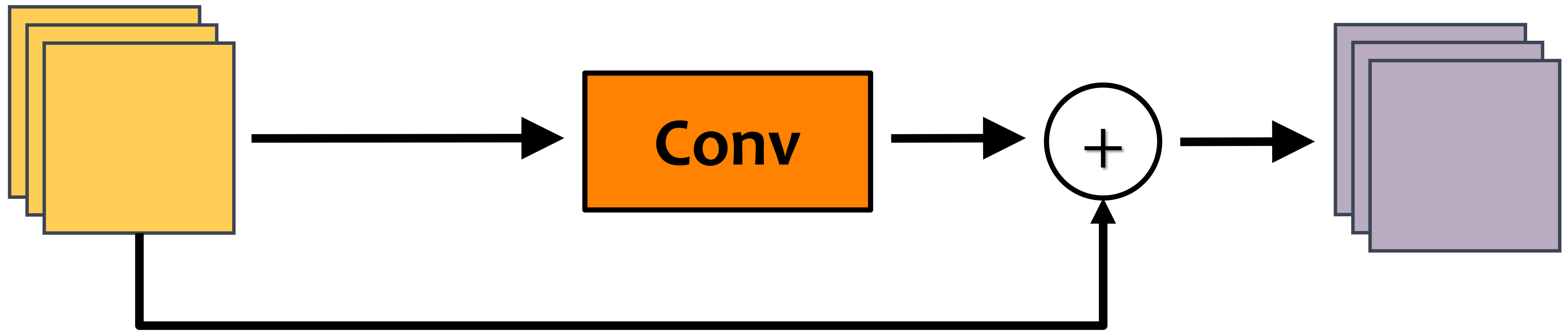
"easy"
(exit at **first** classifier)

"hard"
(exit at **last** classifier)

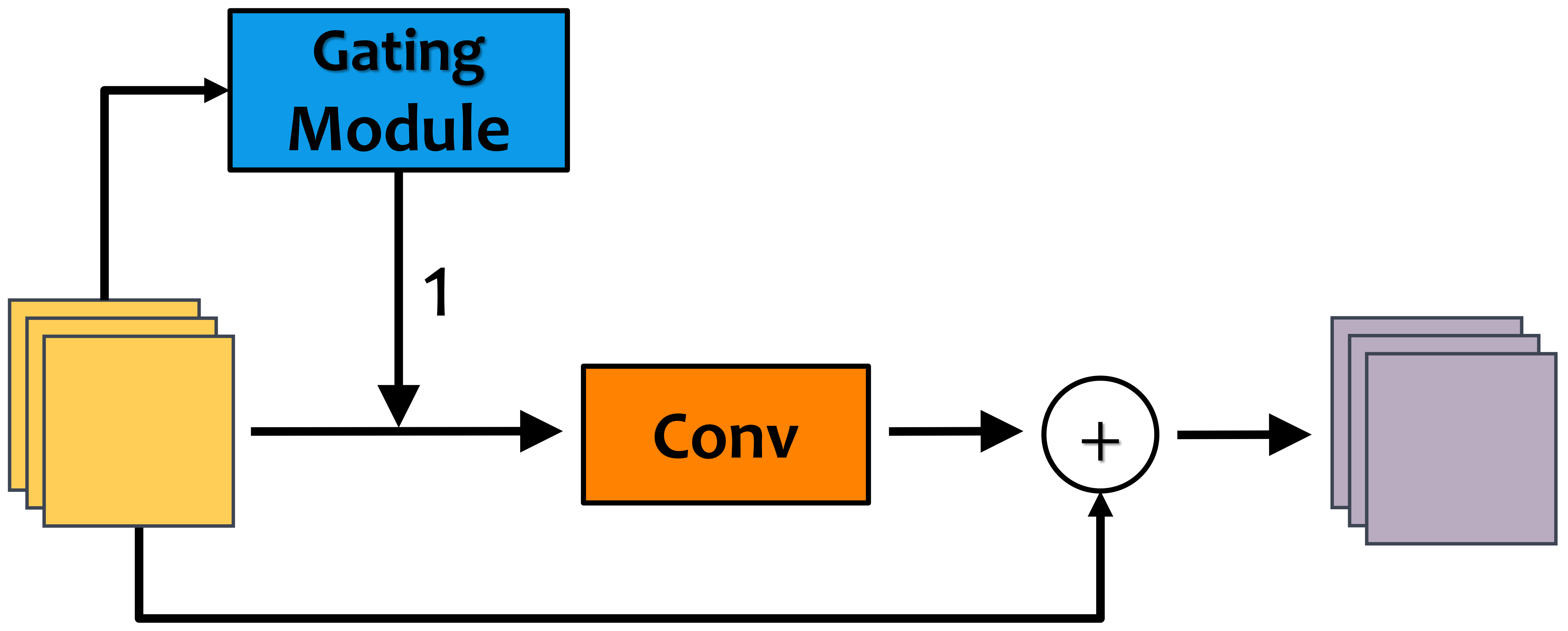
Dynamic Depth: Layer Skipping

清華大學 Tsinghua University

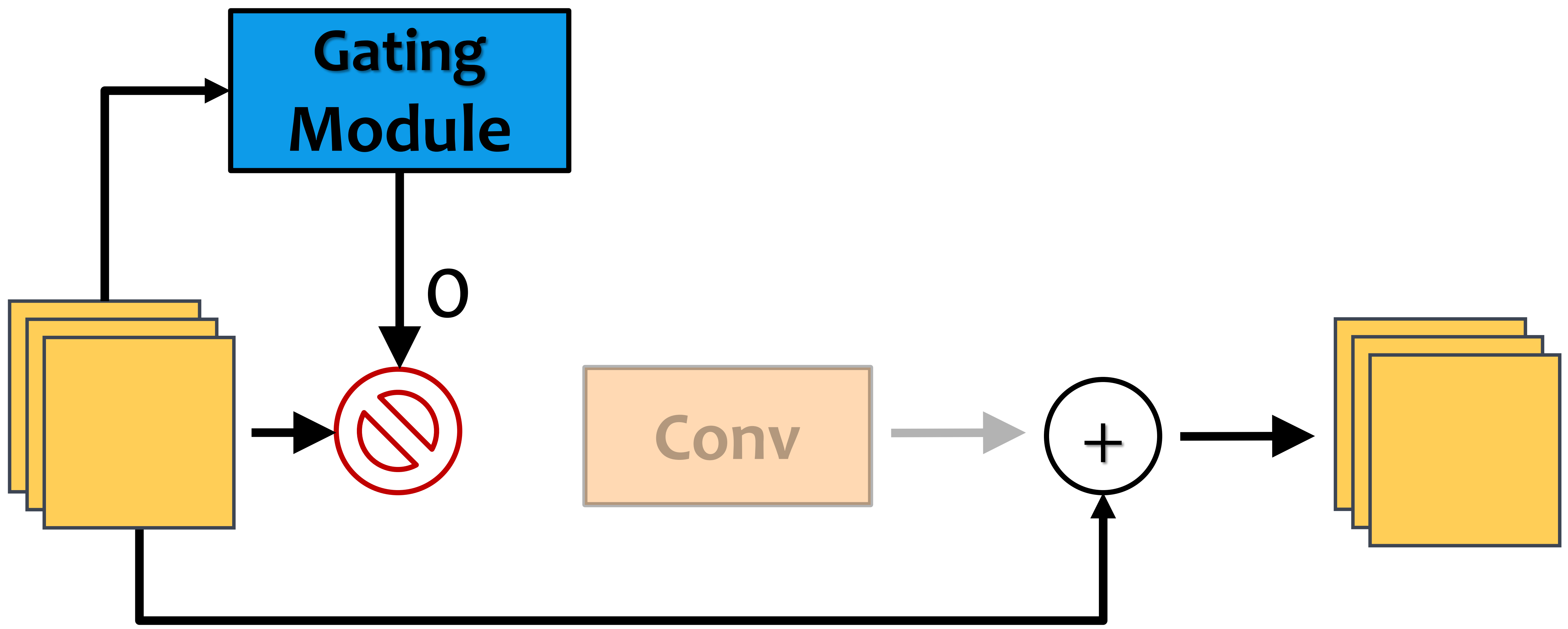
A regular residual block



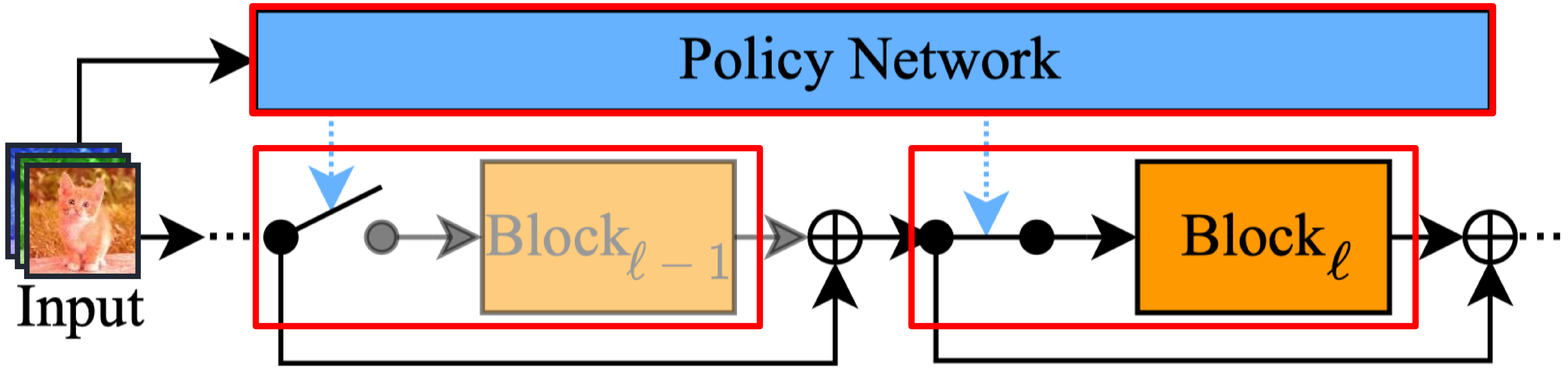
Dynamic Depth: Layer Skipping



Dynamic Depth: Layer Skipping



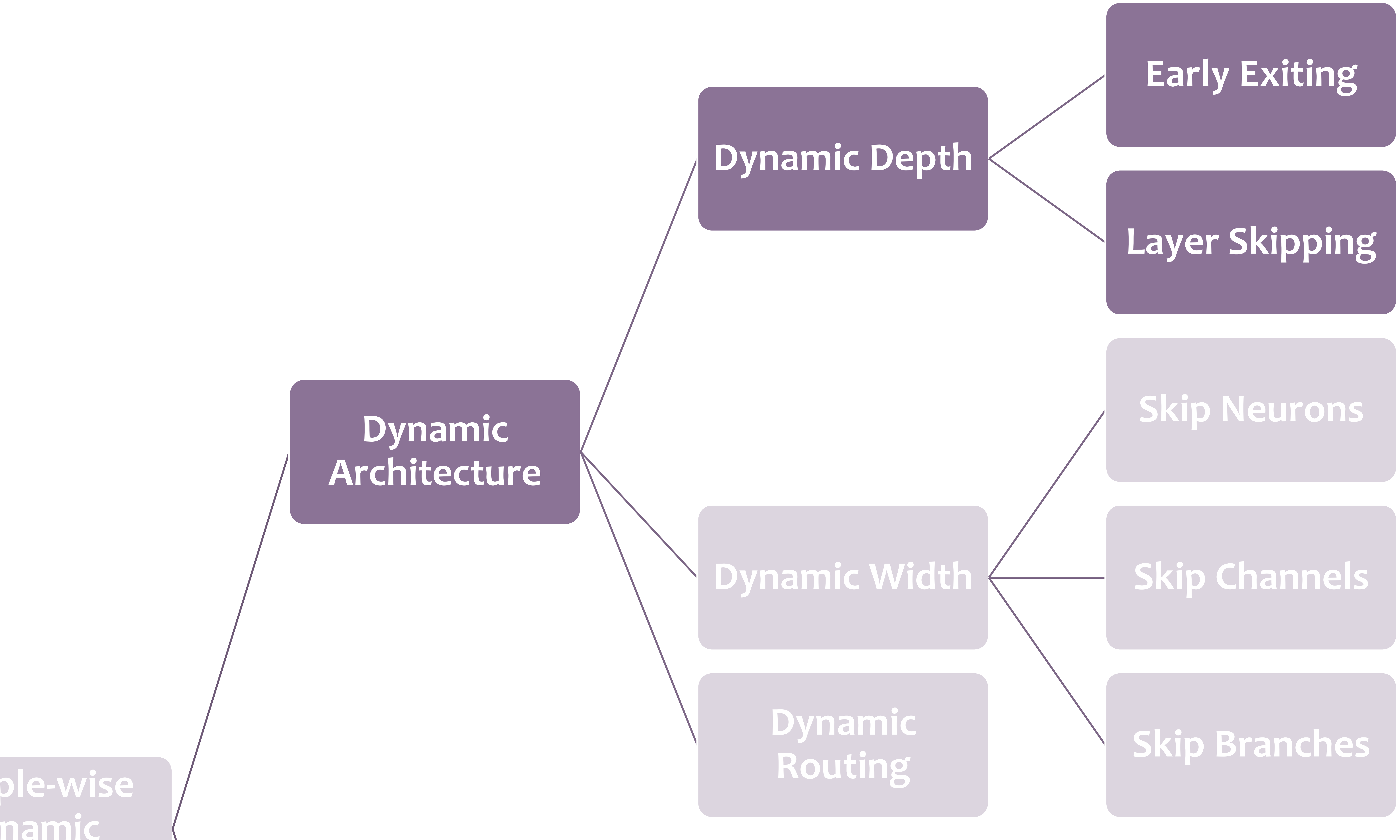
Layer Skipping Based on Policy Networks



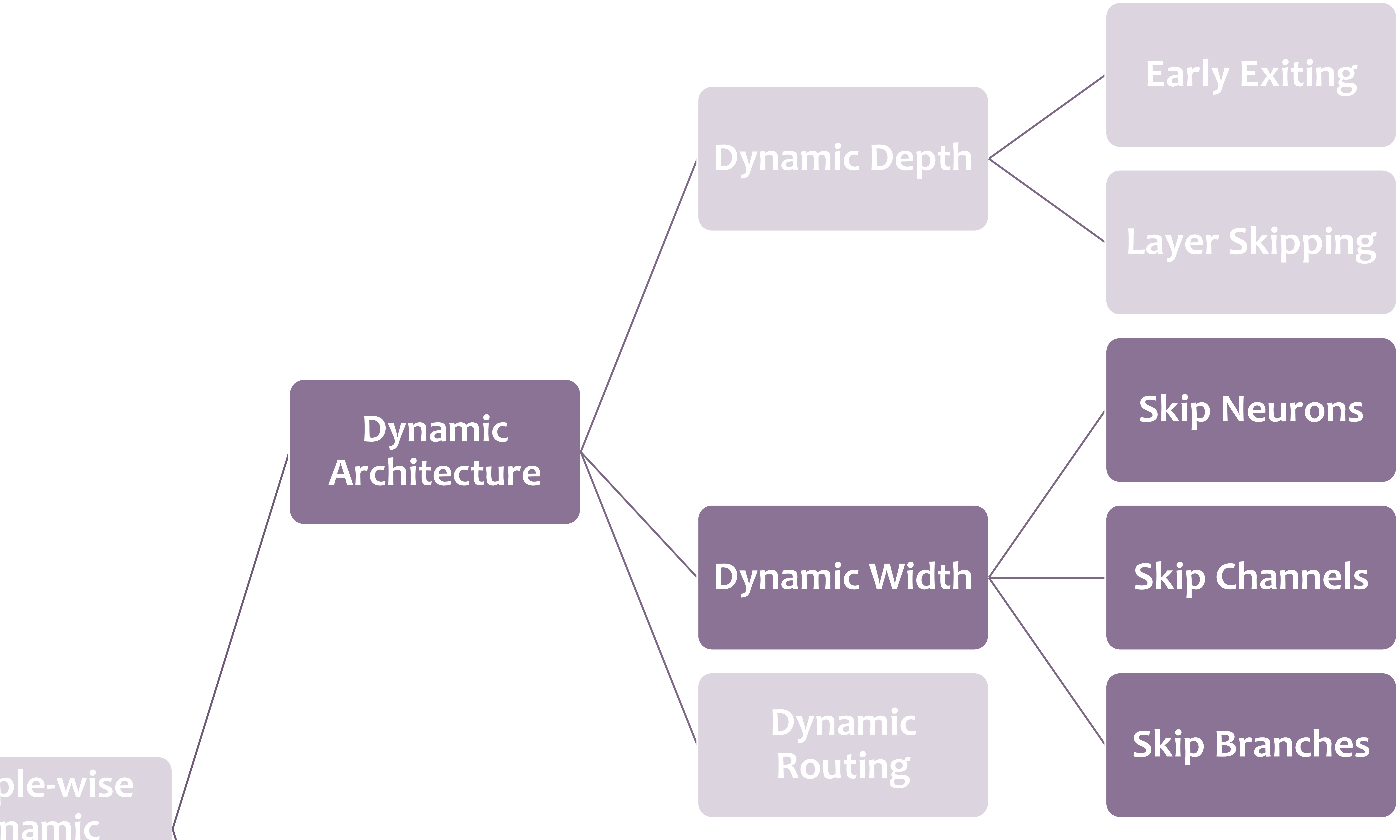
- Wu, Z., Nagarajan, T., Kumar, A., Rennie, S., Davis, L. S., Grauman, K., & Feris, R. (2018). Blockdrop: Dynamic inference paths in residual networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8817-8826).

Sample-wise Dynamic Neural Networks

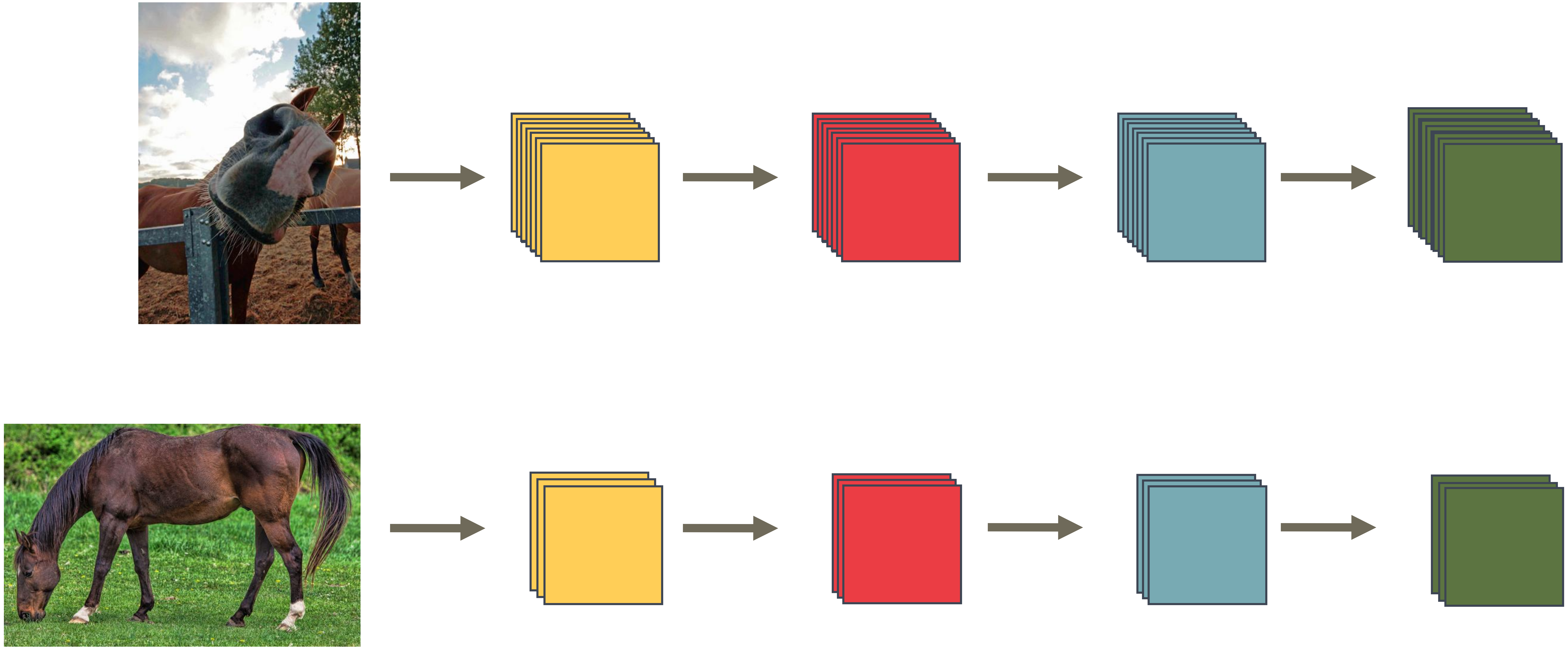
Tsinghua University



Sample-wise Dynamic Neural Networks

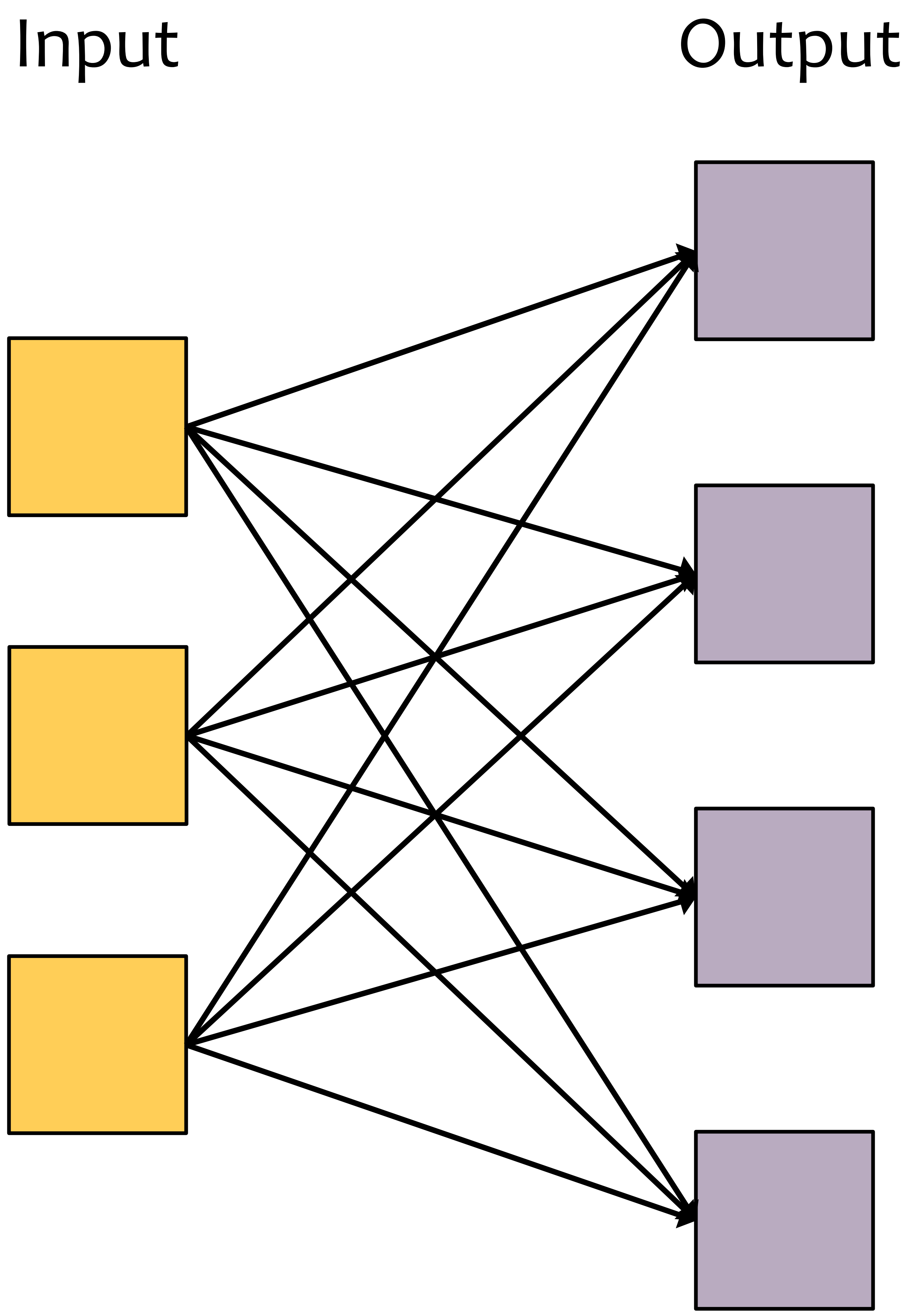


Dynamic Width

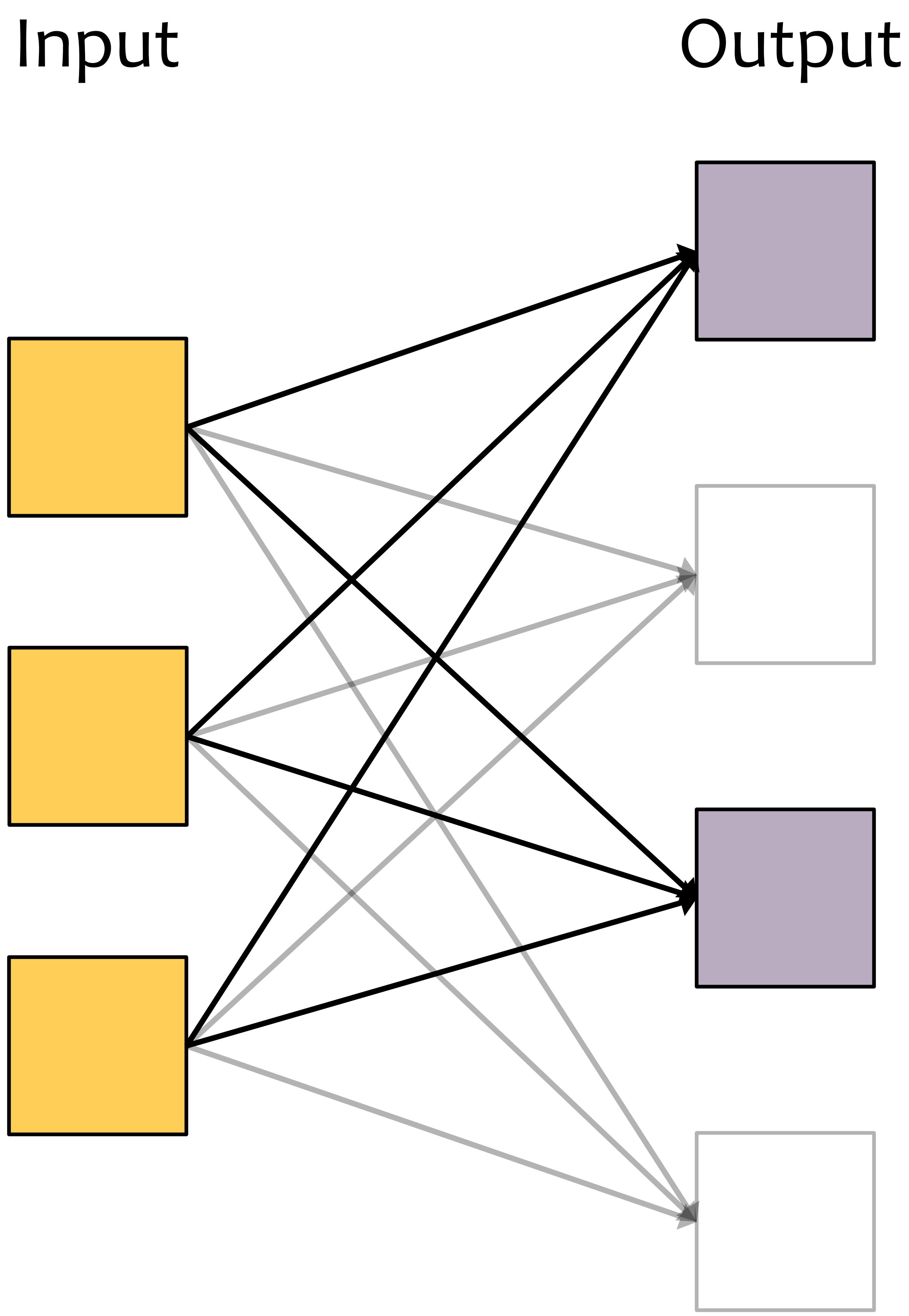


Skip Channels

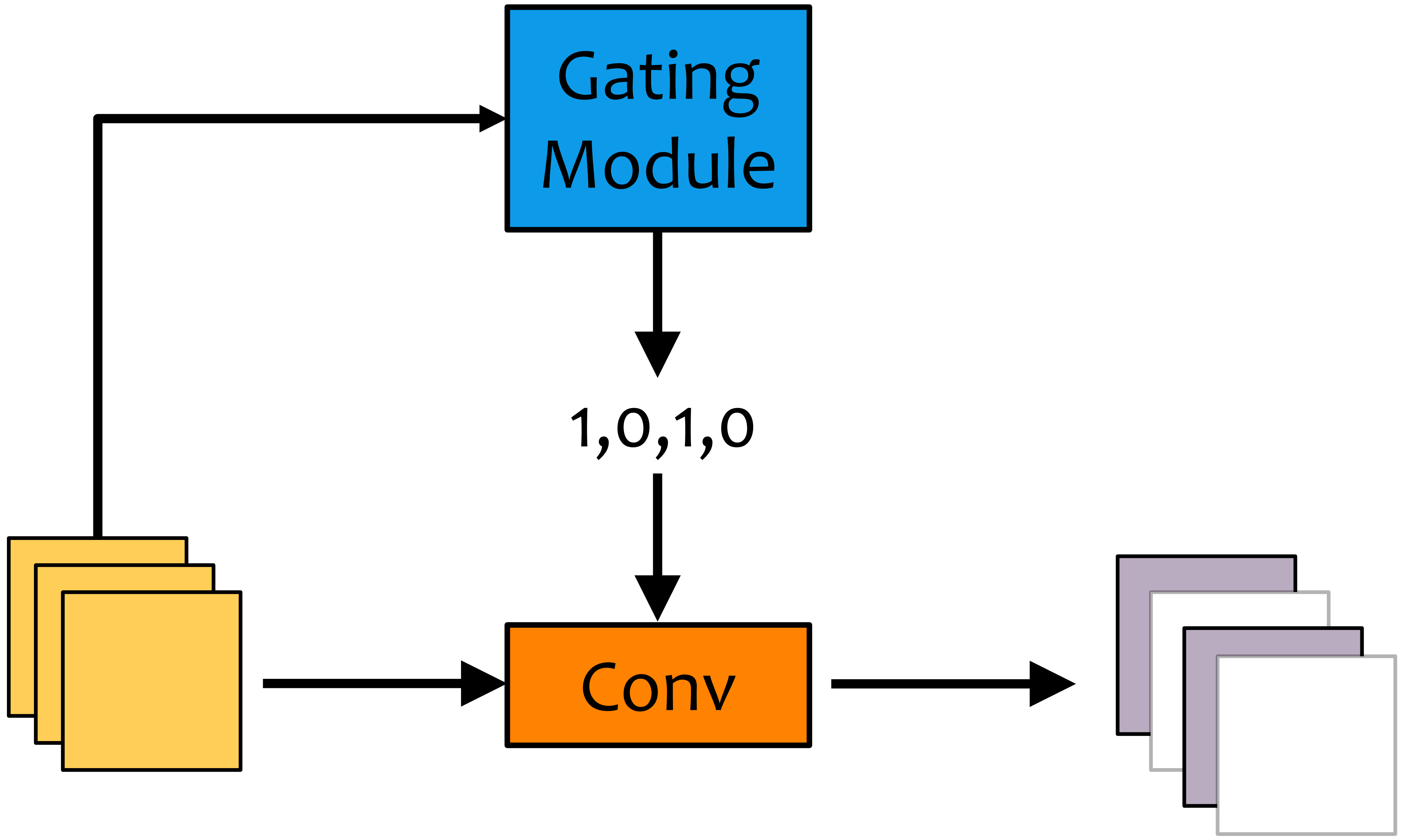
Tsinghua University



Skip Channels



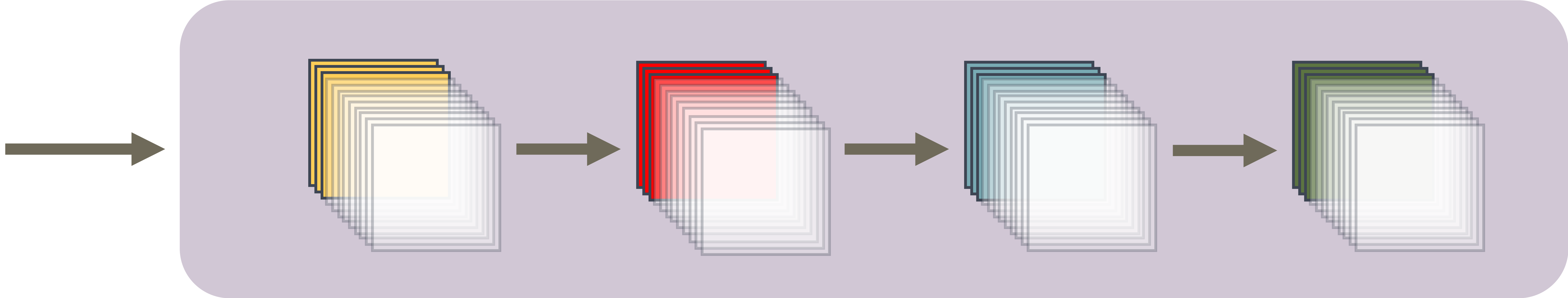
Skip Channels based on Gating Function



- Lin, J., Rao, Y., Lu, J., & Zhou, J. (2017, December). Runtime neural pruning. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 2178-2188).
- Herrmann, C., Strong Bowen, R., & Zabih, R. (2018). An end-to-end approach for speeding up neural network inference. arXiv e-prints, arXiv-1812.
- Bejnordi, B. E., Blankevoort, T., & Welling, M. (2019, September). Batch-shaping for learning conditional channel gated networks. In International Conference on Learning Representations.

Multi-stage Structure

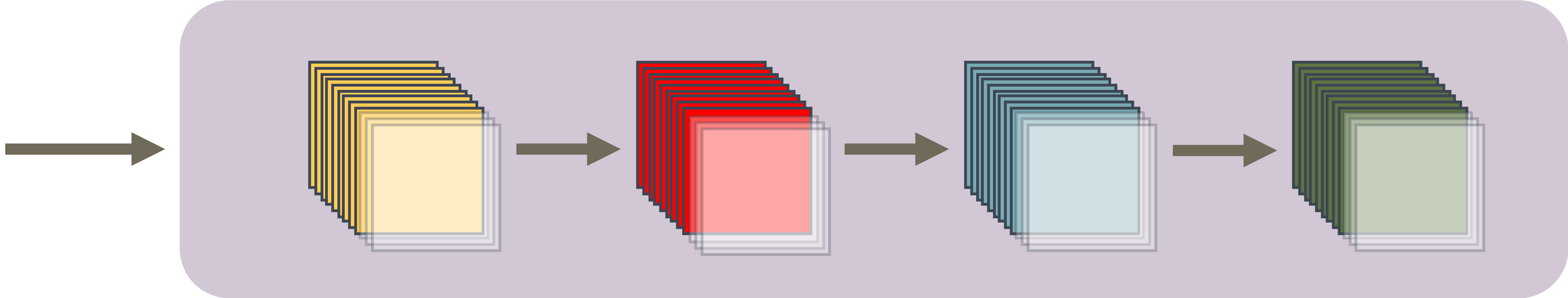
Tsinghua University



Horse: 0.2

- Yuan, Z., Wu, B., Sun, G., Liang, Z., Zhao, S., & Bi, W. (2020, August). S2dnas: Transforming static cnn model for dynamic inference via neural architecture search. In European Conference on Computer Vision (pp. 175-192). Springer, Cham.

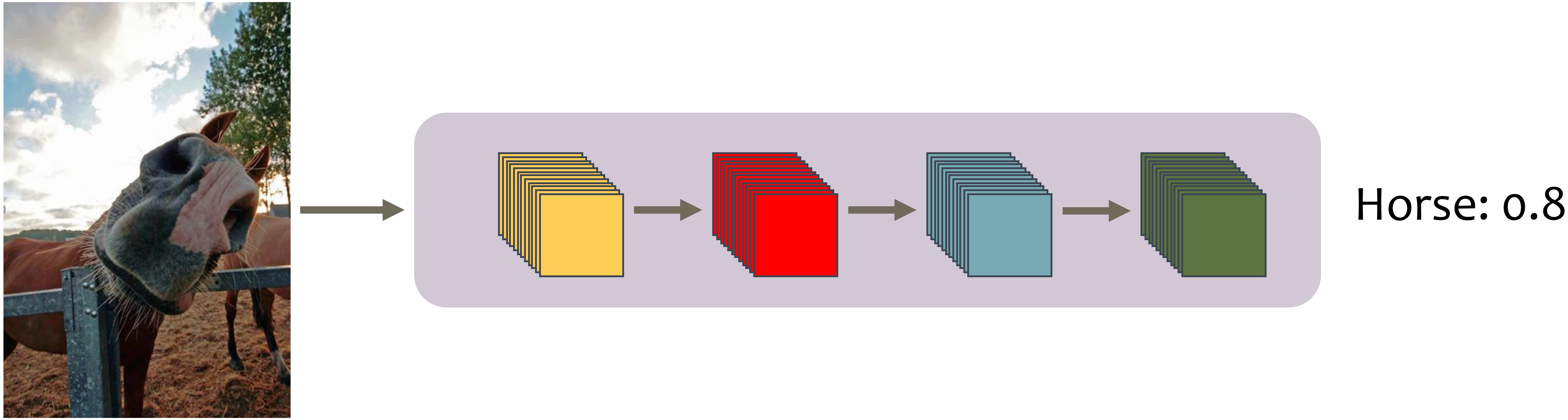
Multi-stage Structure



Horse: 0.6

- Yuan, Z., Wu, B., Sun, G., Liang, Z., Zhao, S., & Bi, W. (2020, August). S2dnas: Transforming static cnn model for dynamic inference via neural architecture search. In European Conference on Computer Vision (pp. 175-192). Springer, Cham.

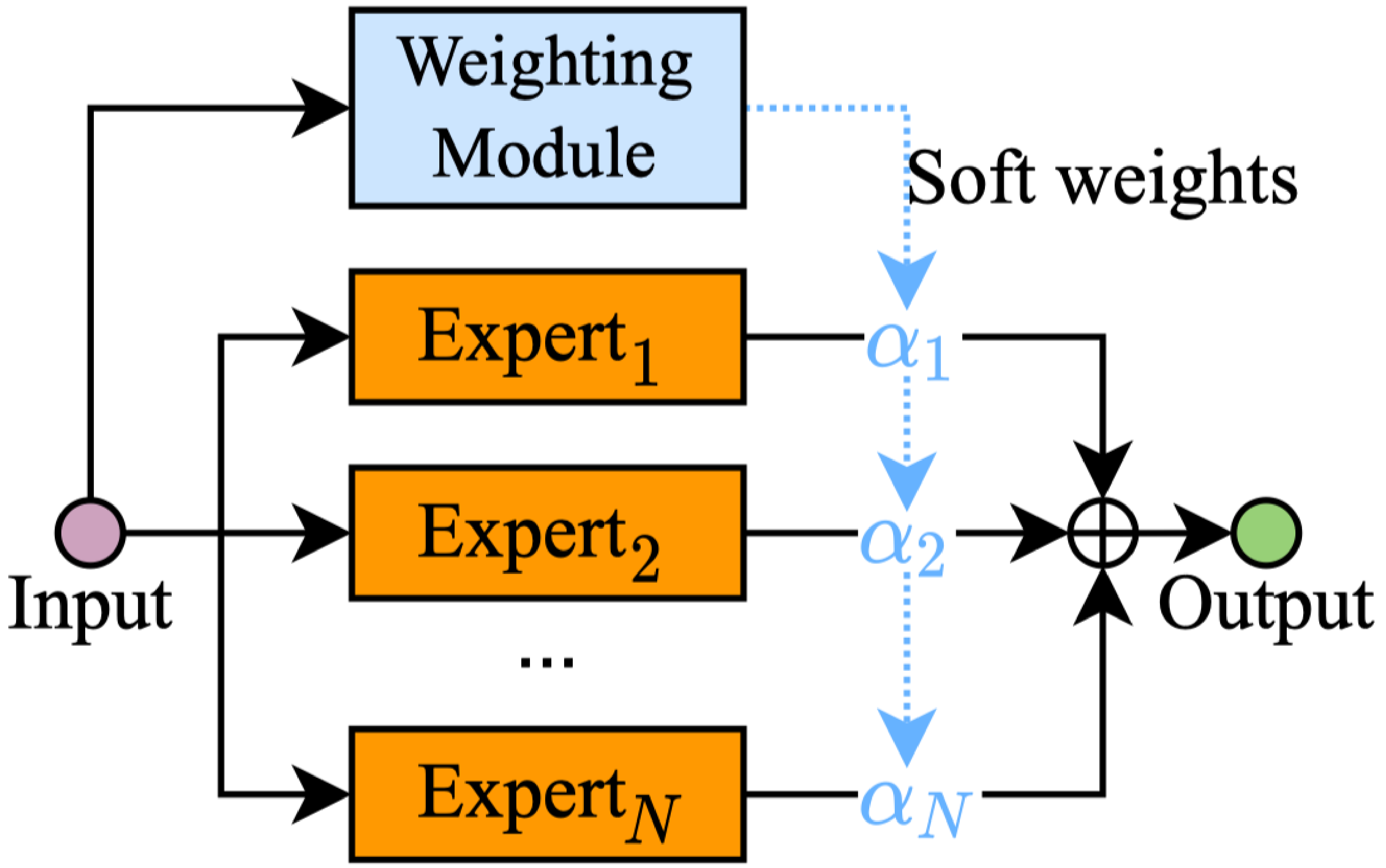
Multi-stage Structure



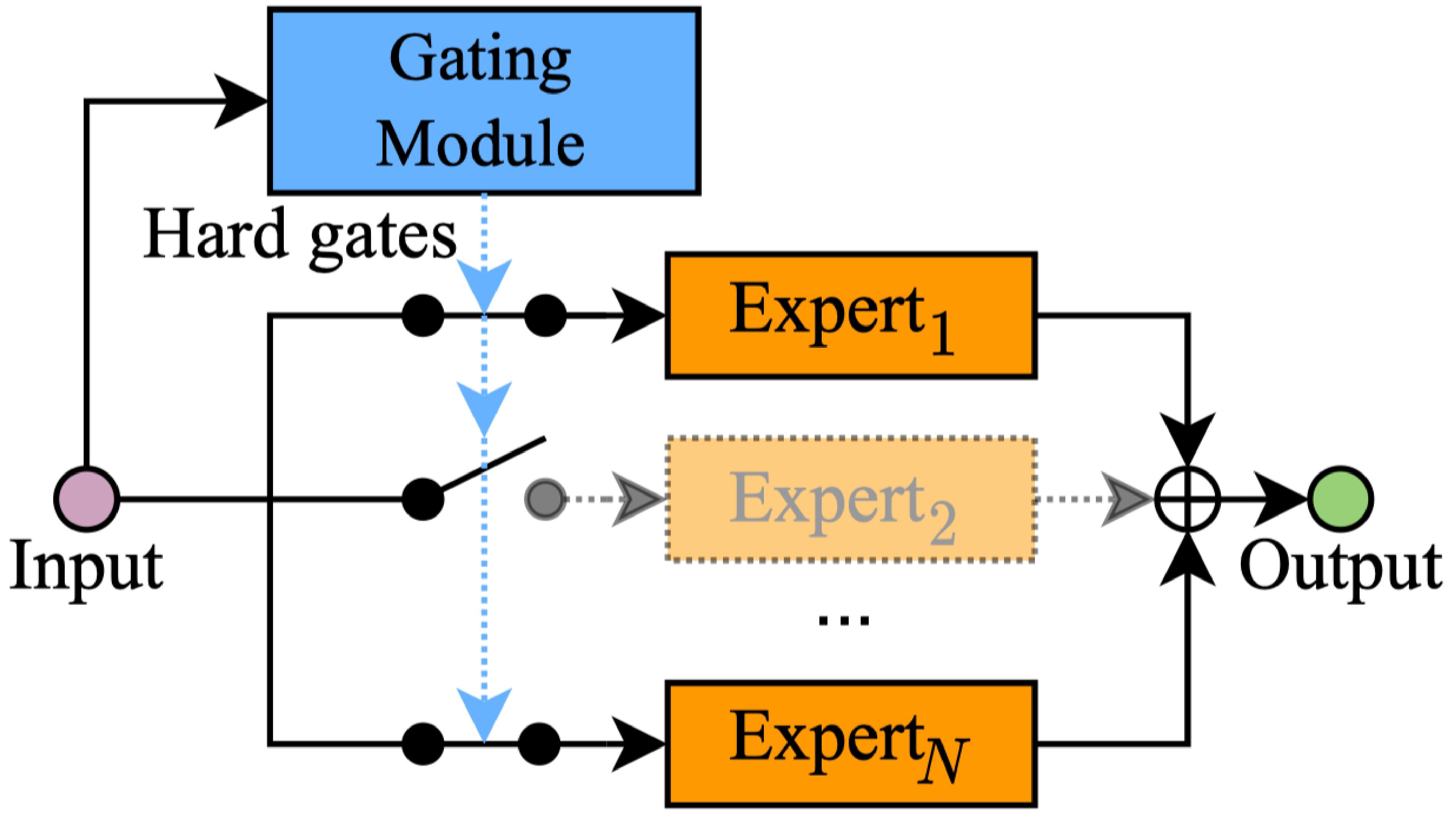
▪ Yuan, Z., Wu, B., Sun, G., Liang, Z., Zhao, S., & Bi, W. (2020, August). S2dnas: Transforming static cnn model for dynamic inference via neural architecture search. In European Conference on Computer Vision (pp. 175-192). Springer, Cham.

Tsinghua University

Mixture of Experts (MoE)



Soft MoE



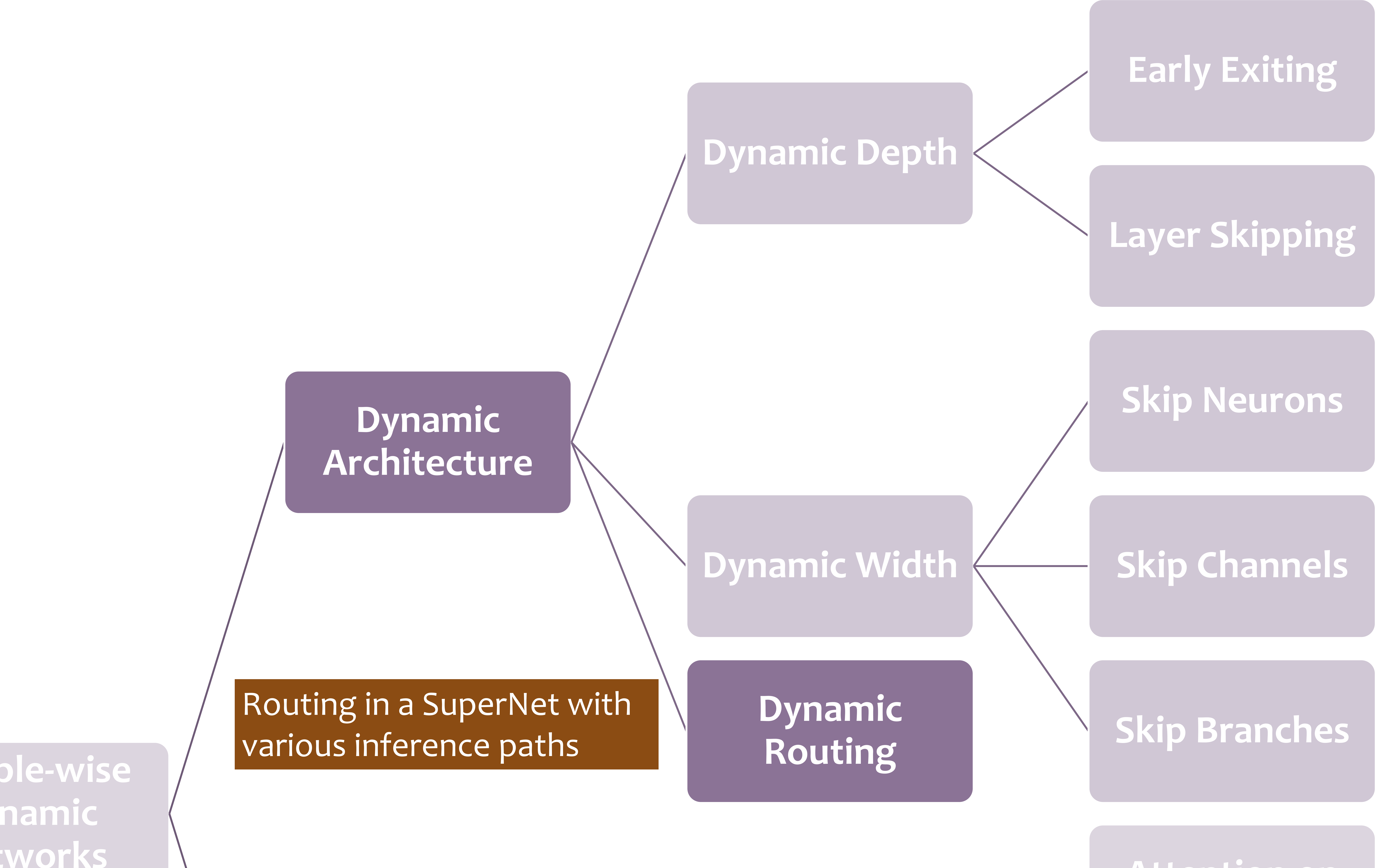
Hard MoE

- Mullapudi, R. T., Mark, W. R., Shazeer, N., & Fatahalian, K. (2018). Hydranets: Specialized dynamic architectures for efficient inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8080-8089).
- Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538.
- Fedus, W., Zoph, B., & Shazeer, N. (2021). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv preprint arXiv:2101.03961.

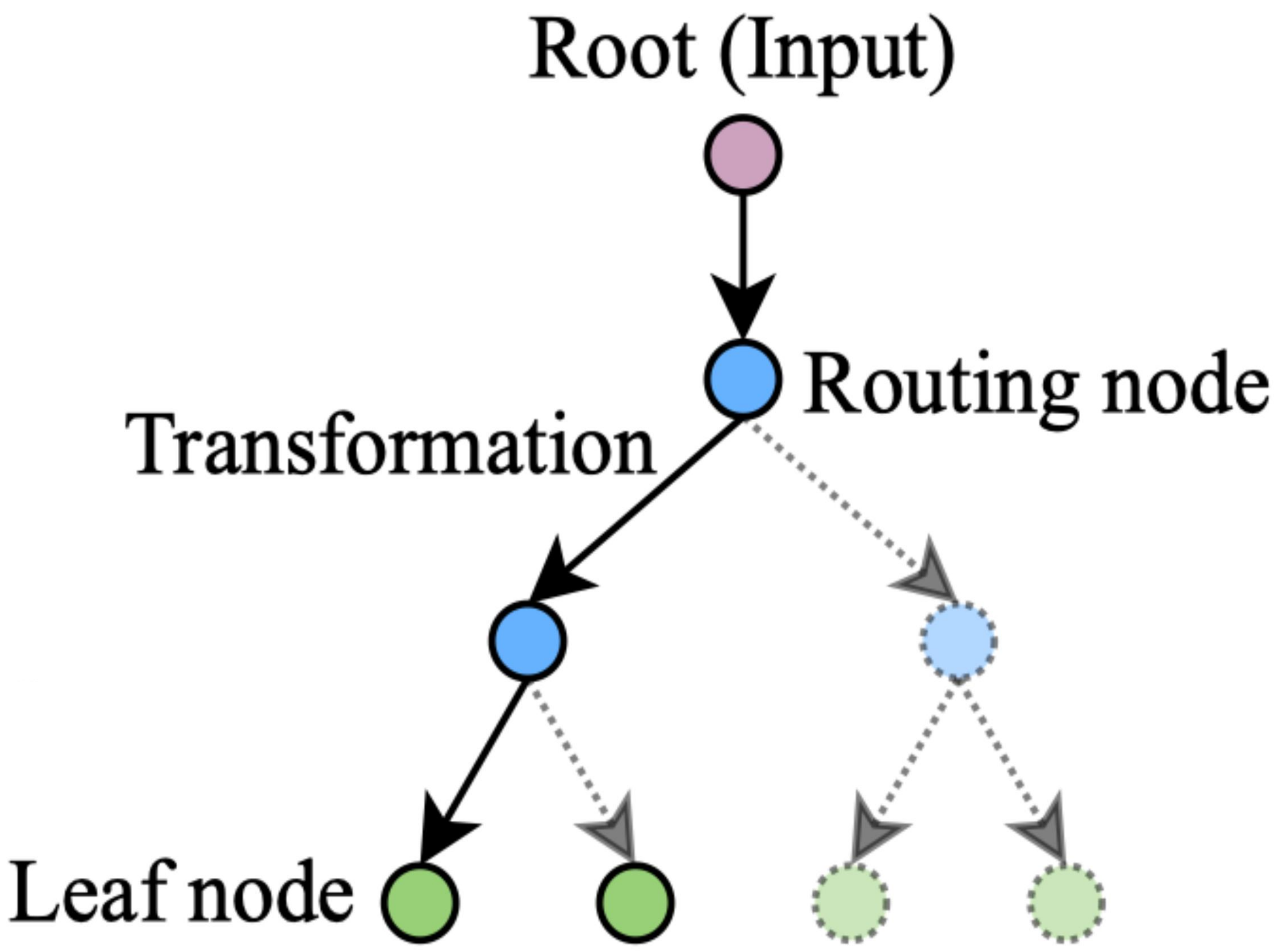
Sample-wise Dynamic Neural Networks

清華大學
Tsinghua University

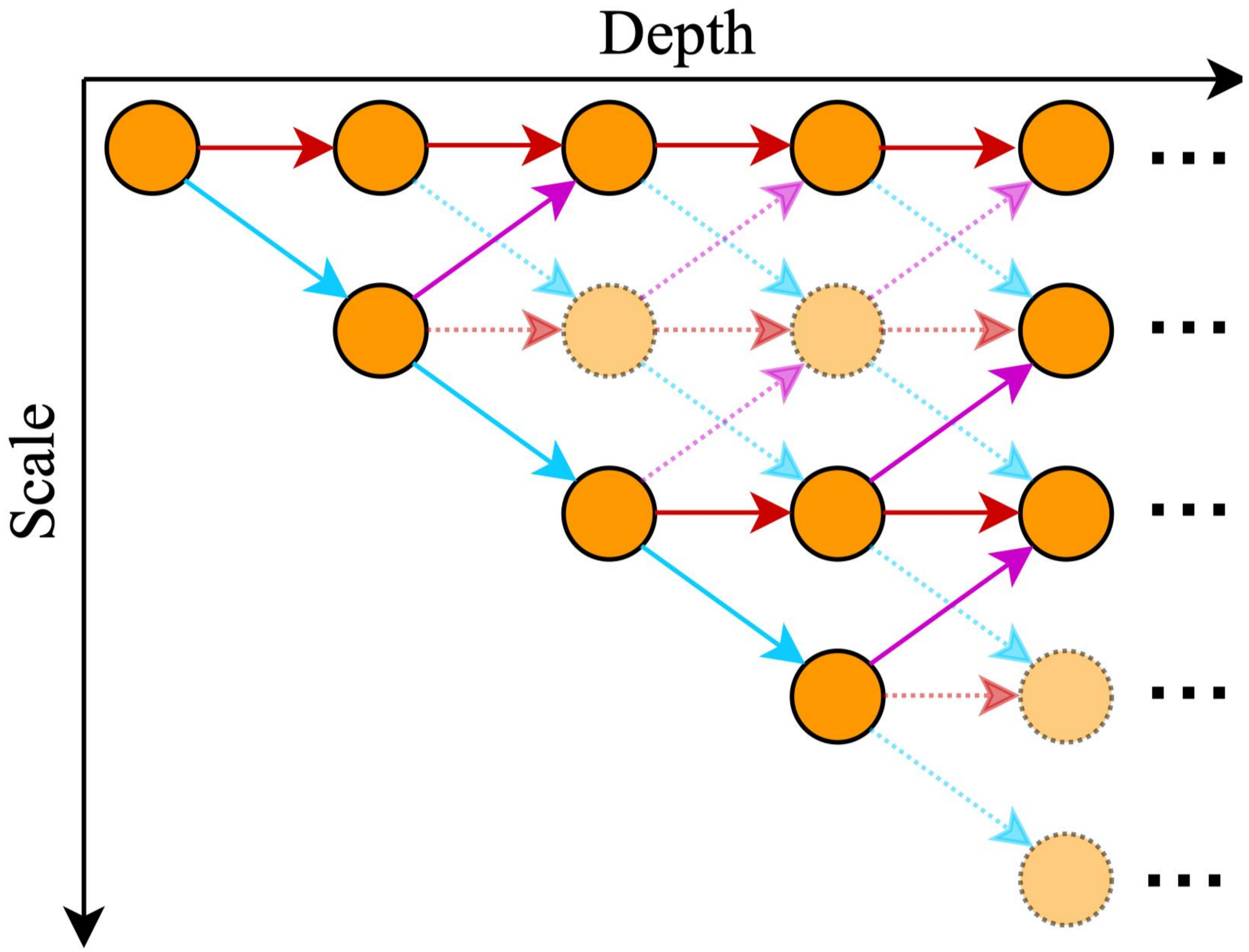
Sample-wise Dynamic Neural Networks



Dynamic Routing in SuperNets



Tree structure

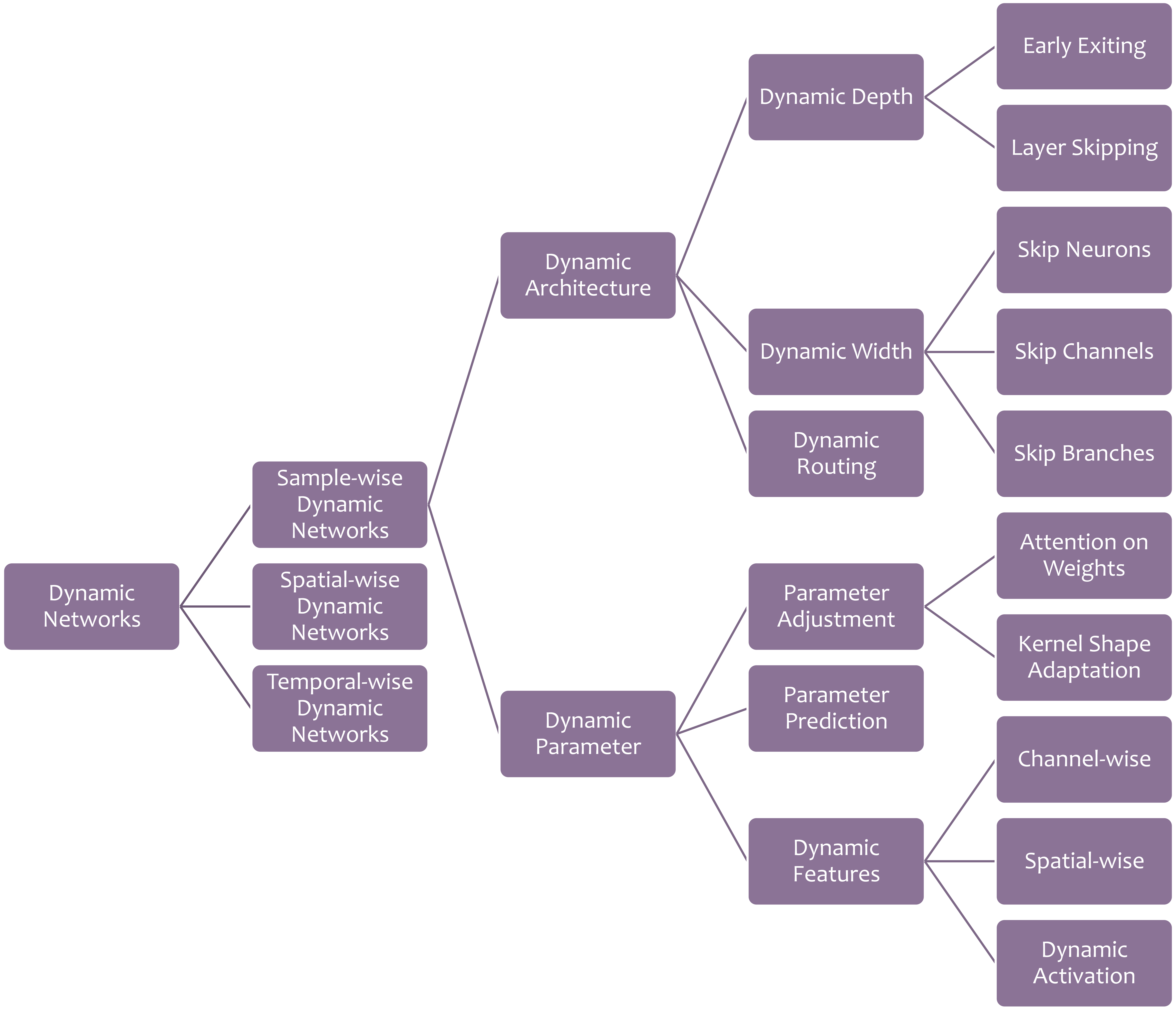


Multi-scale structure

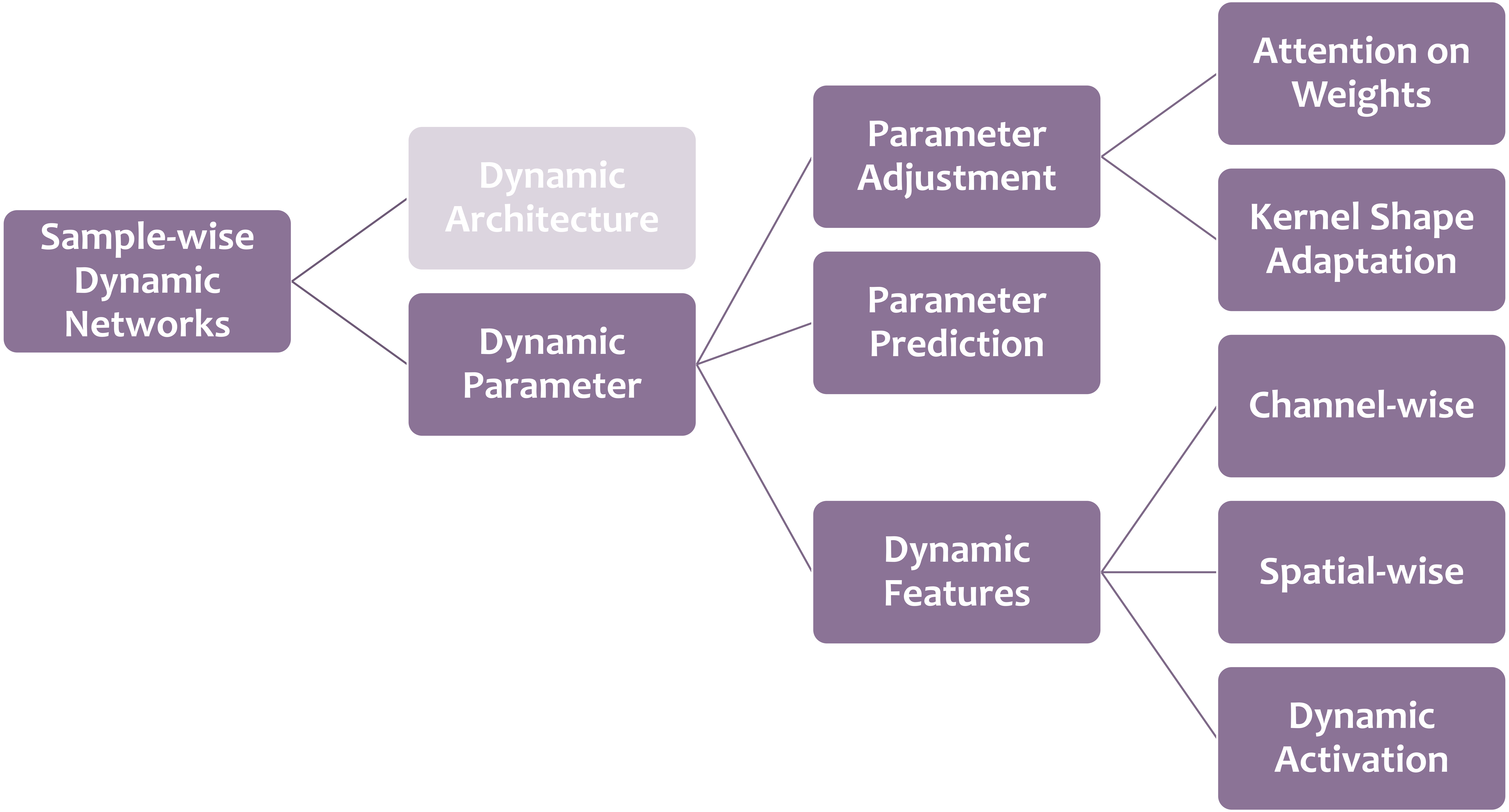
- Tanno, R., Arulkumaran, K., Alexander, D., Criminisi, A., & Nori, A. (2019, May). Adaptive neural trees. In International Conference on Machine Learning (pp. 6166-6175). PMLR.
- Li, Y., Song, L., Chen, Y., Li, Z., Zhang, X., Wang, X., & Sun, J. (2020). Learning dynamic routing for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8553-8562).

Sample-wise Dynamic Neural Networks

Tsinghua University

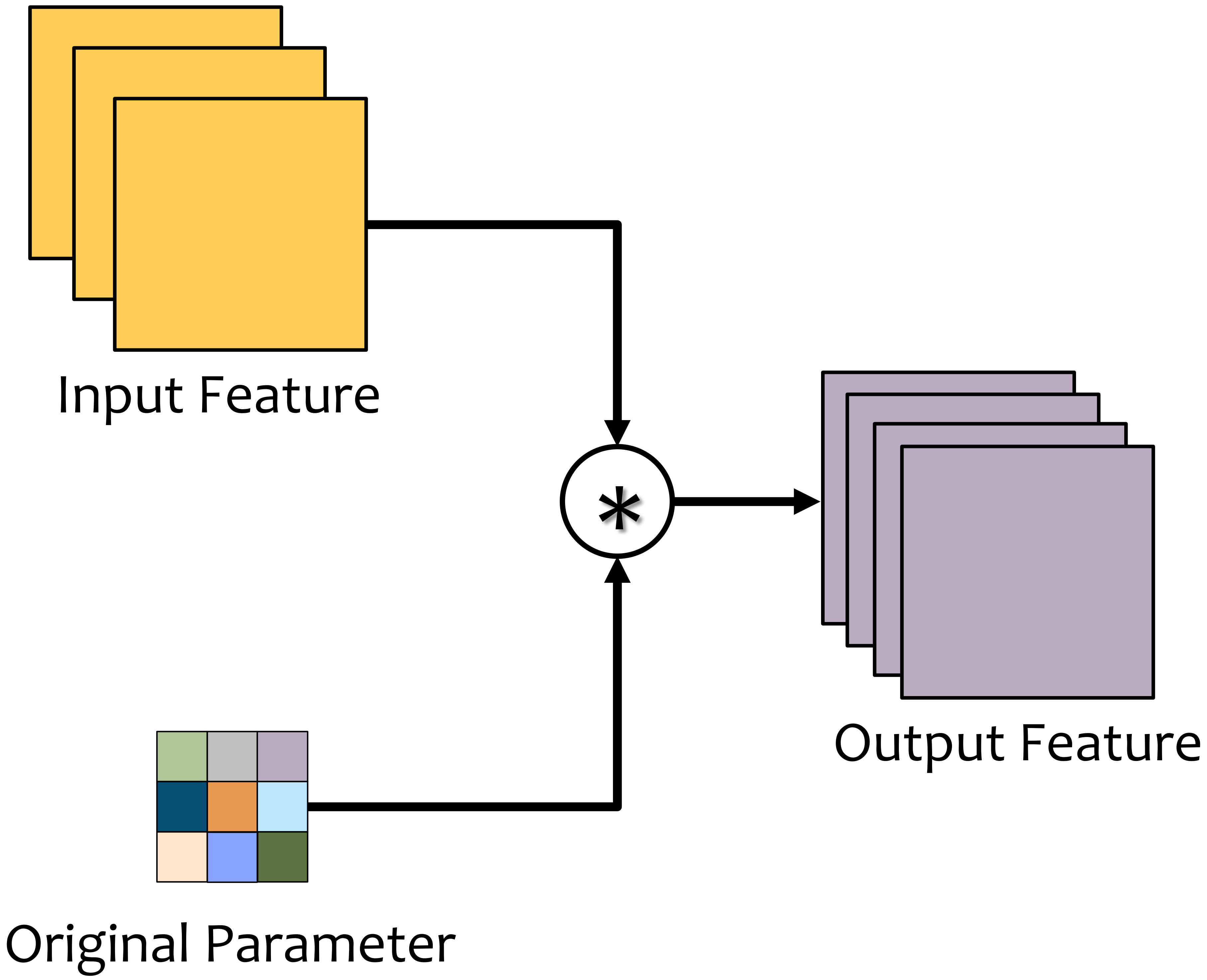


Sample-wise Dynamic Neural Networks

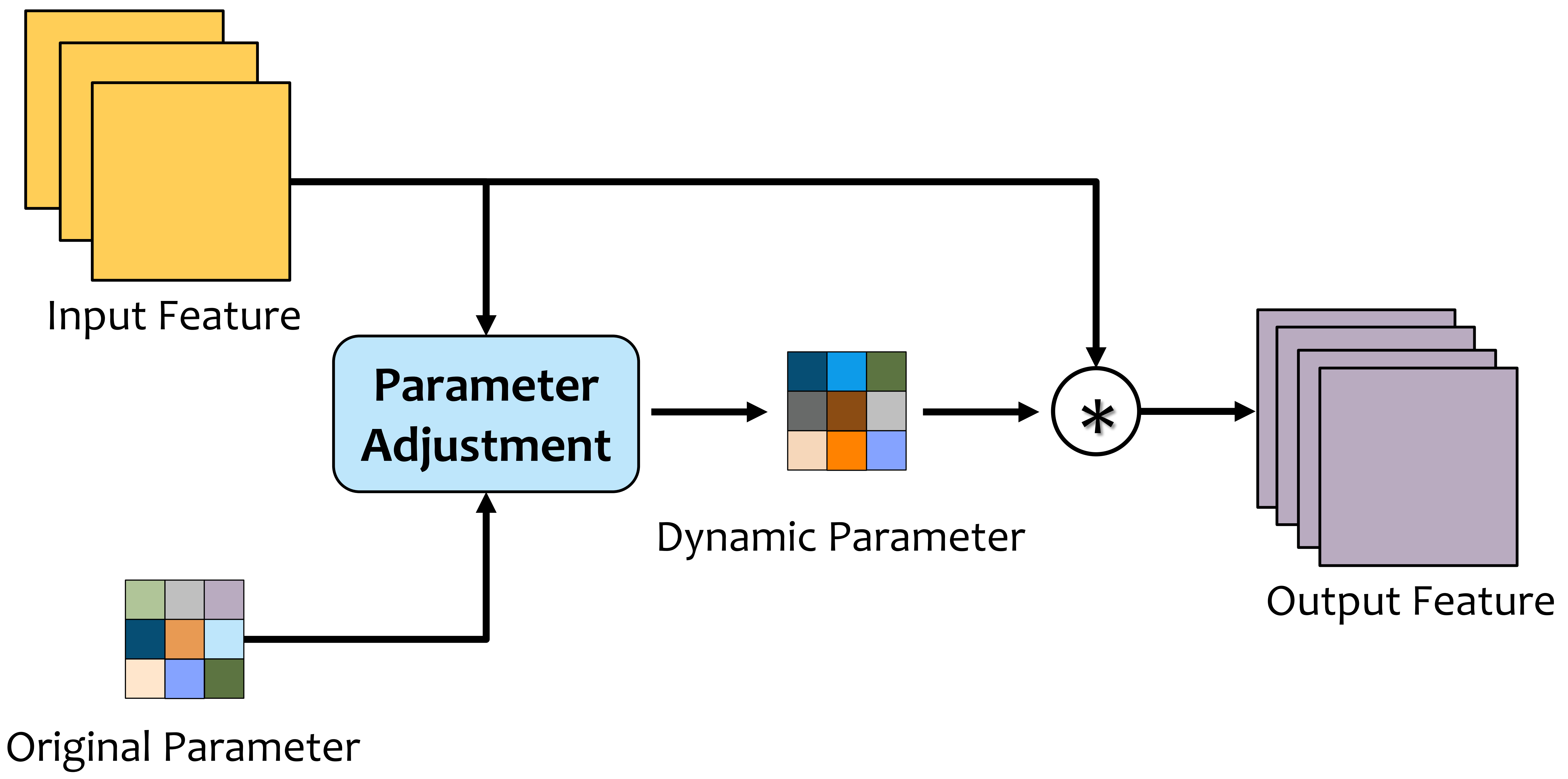


Dynamic Parameter: Weight Adjustment

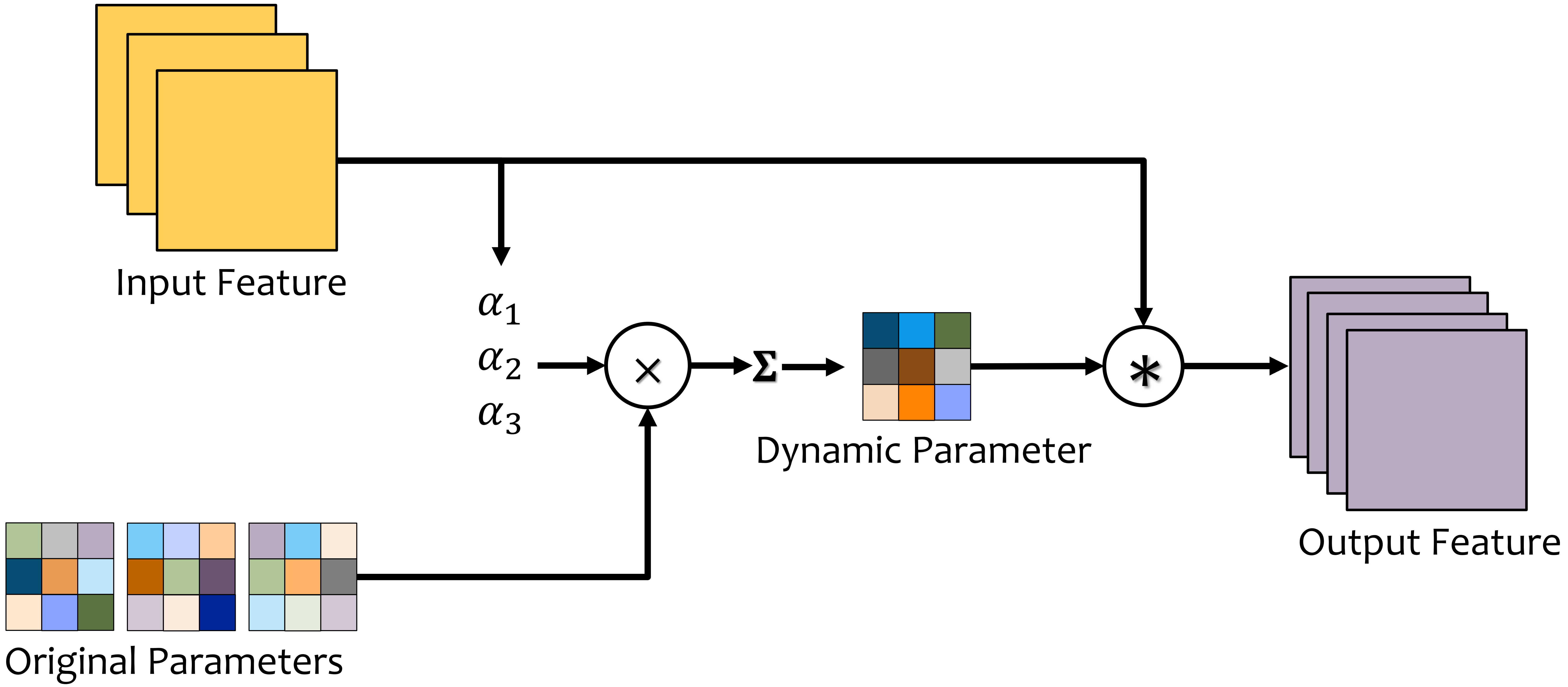
Regular convolution



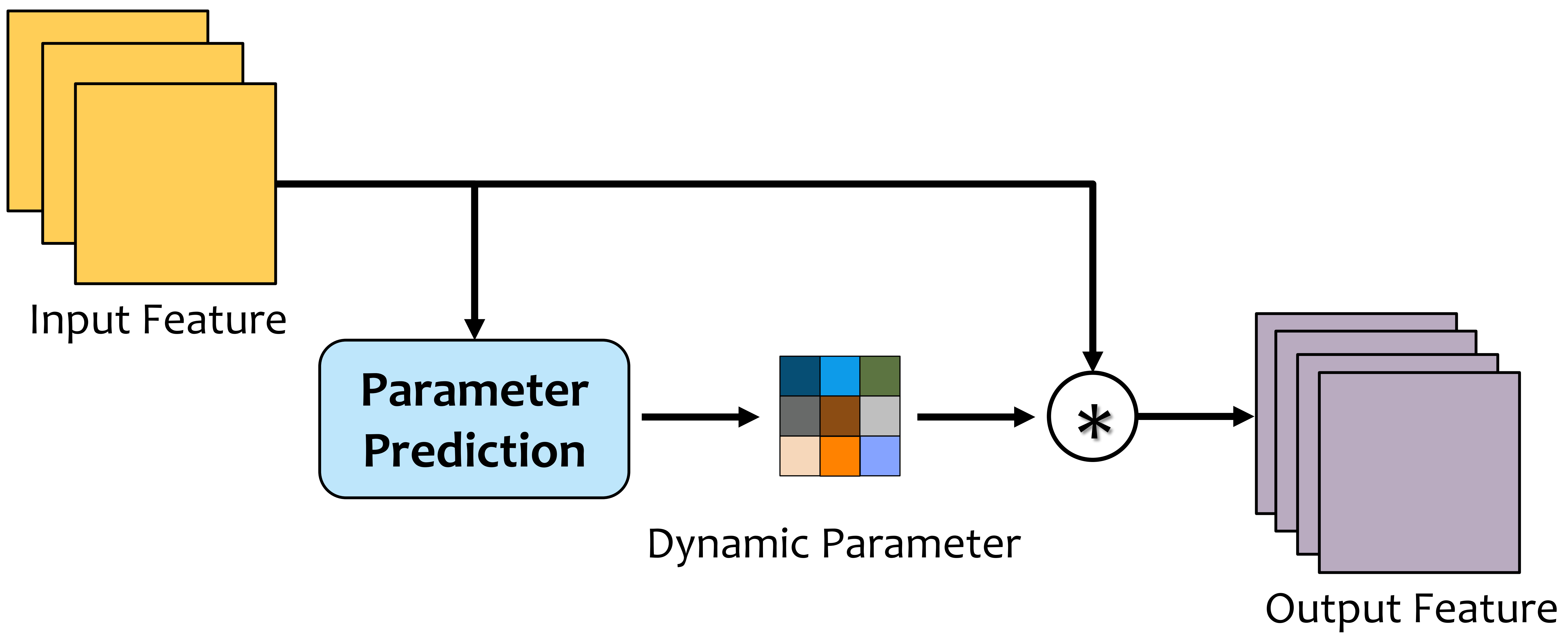
Dynamic Parameter: Weight Adjustment



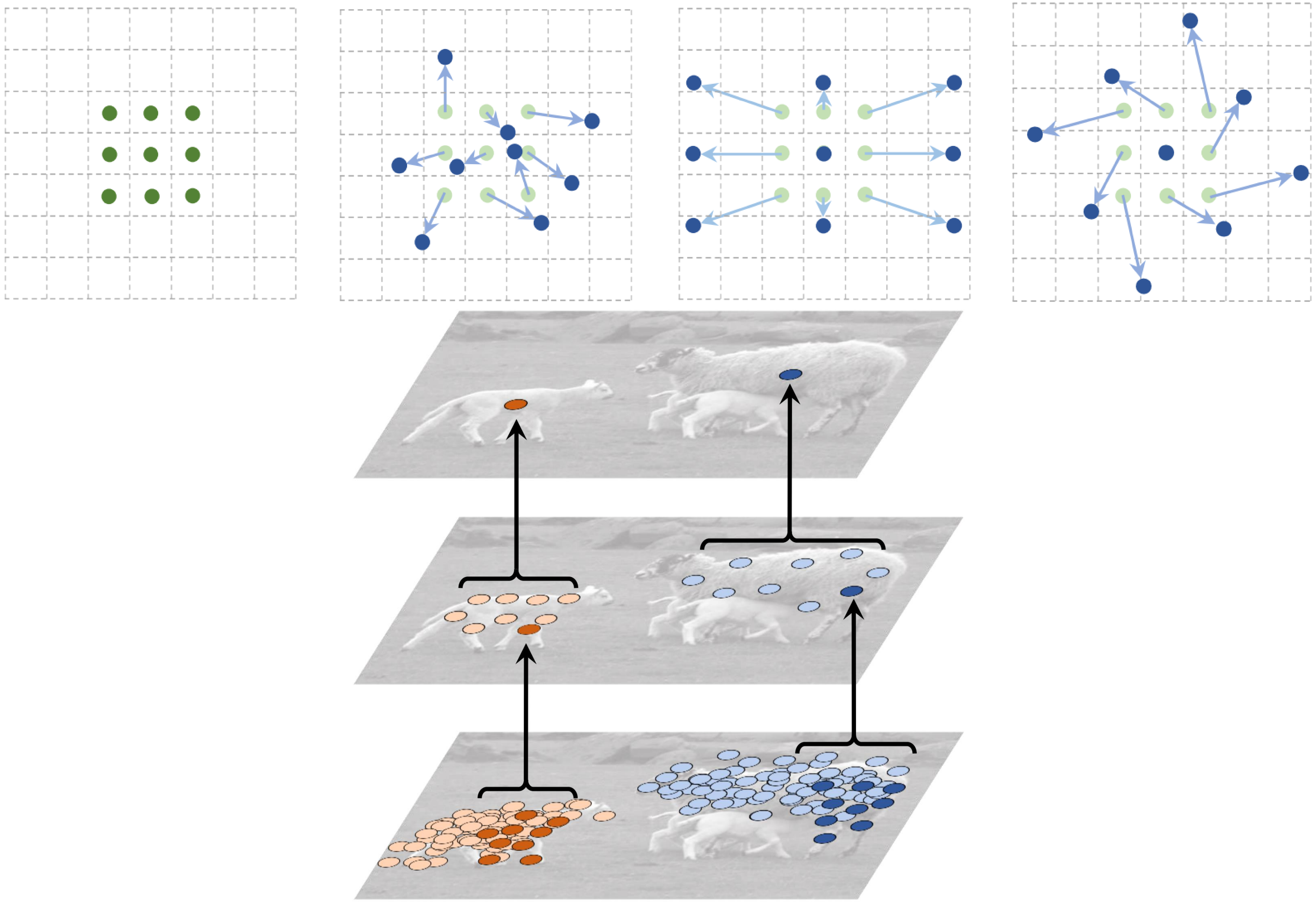
Dynamic Parameter: Weight Ensemble



Dynamic Parameter: Weight Prediction

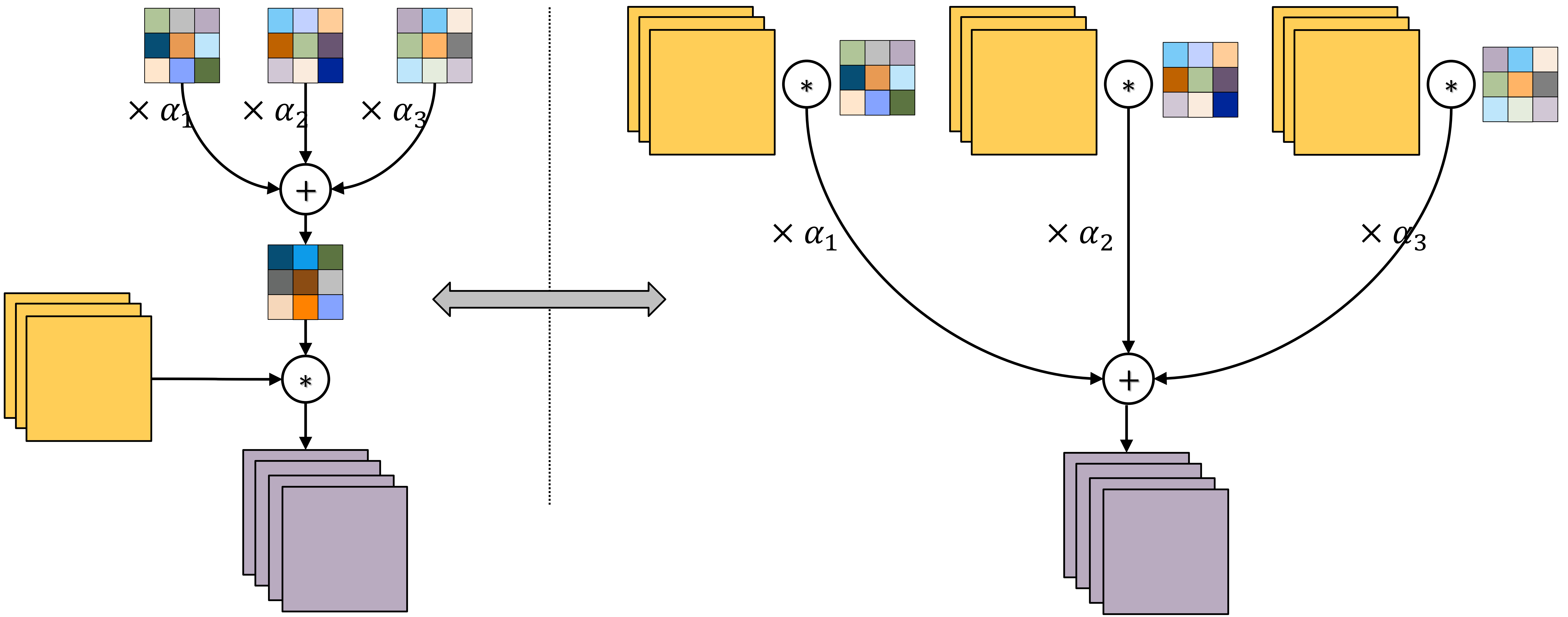


Kernel Shape Adaptation



- Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 764-773).
- Zhu, X., Hu, H., Lin, S., & Dai, J. (2019). Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9308-9316).

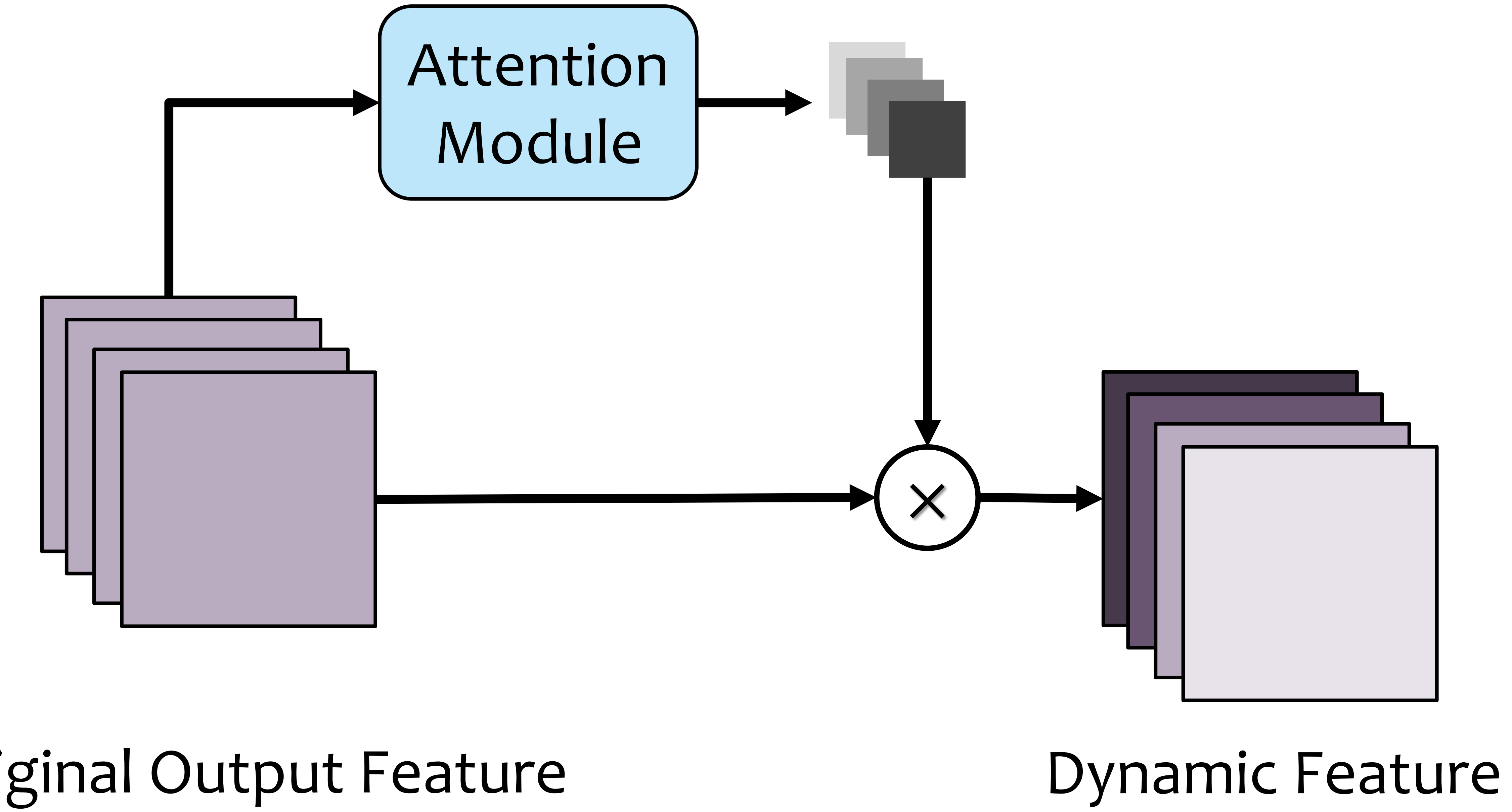
Dynamic Parameter: Weight Adjustment



$$\left(\sum_n \alpha_n \mathbf{W}_n\right) * \mathbf{x} = \sum_n \alpha_n \left(\mathbf{W}_n * \mathbf{x}\right)$$

- Yang, B., Bender, G., Le, Q. V., & Ngiam, J. (2019). Condconv: Conditionally parameterized convolutions for efficient inference. arXiv preprint arXiv:1904.04971.

Channel-wise Attention



$$(\mathbf{x} * \mathbf{W}) \otimes \boldsymbol{\alpha} = \mathbf{x} * (\mathbf{W} \otimes \boldsymbol{\alpha})$$

Dynamic Features   Dynamic Weights

- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

Outline

清華大学
Tsinghua University

1. Overview of CNN backbones
2. Architecture design for mobile CNNs
3. Dynamic CNNs for mobile applications
   A. Sample-wise Dynamic Networks
   B. Spatial-wise Dynamic Networks
   C. Temporal-wise Dynamic Networks

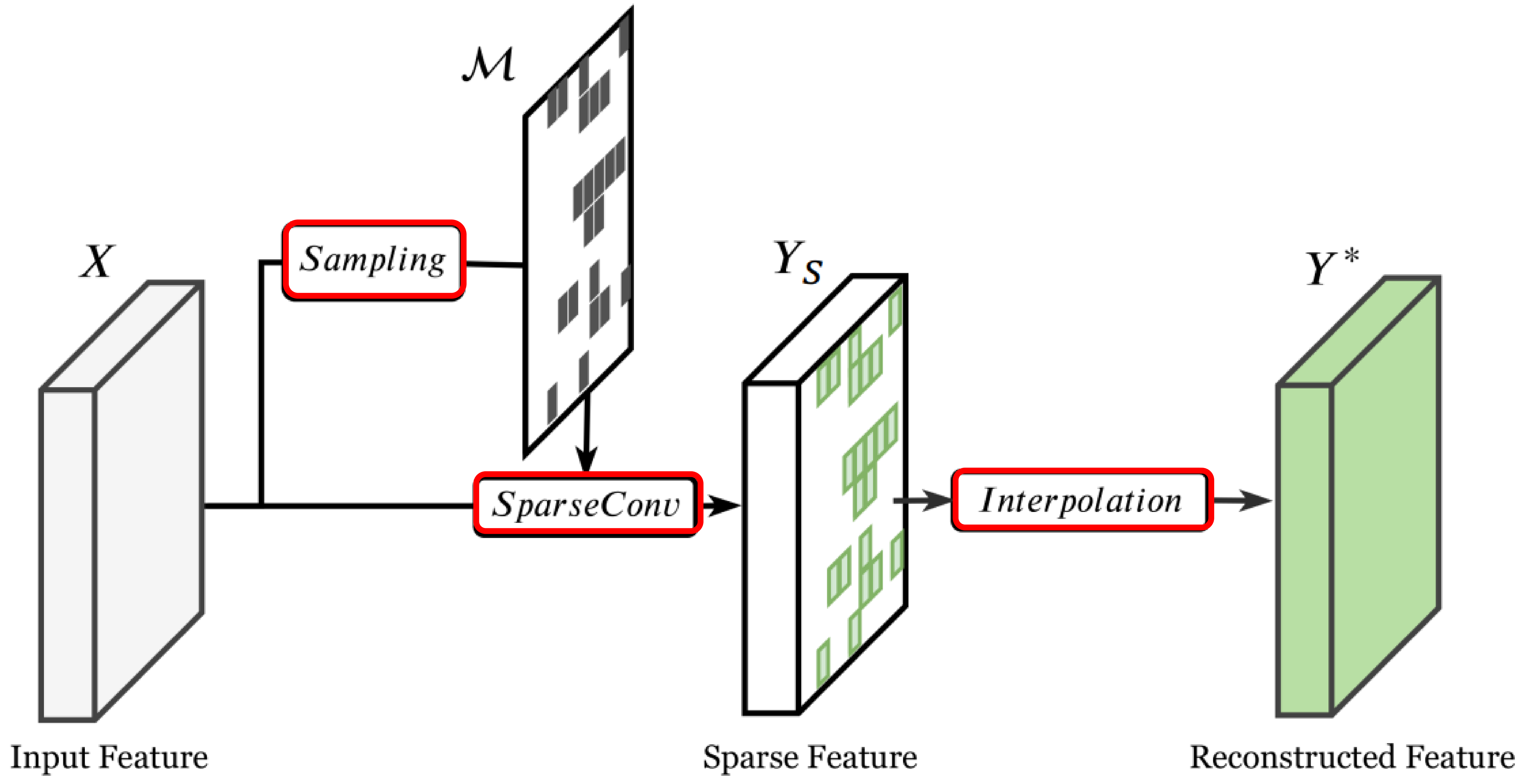
*From **Sample Adaptive** to **Spatial Adaptive***


COLUMBUS CAFÉ
COLUMBUS CAFÉ
Coca-Cola
SOUVENIRS DE PARIS
21


COLUMBUS CAFE

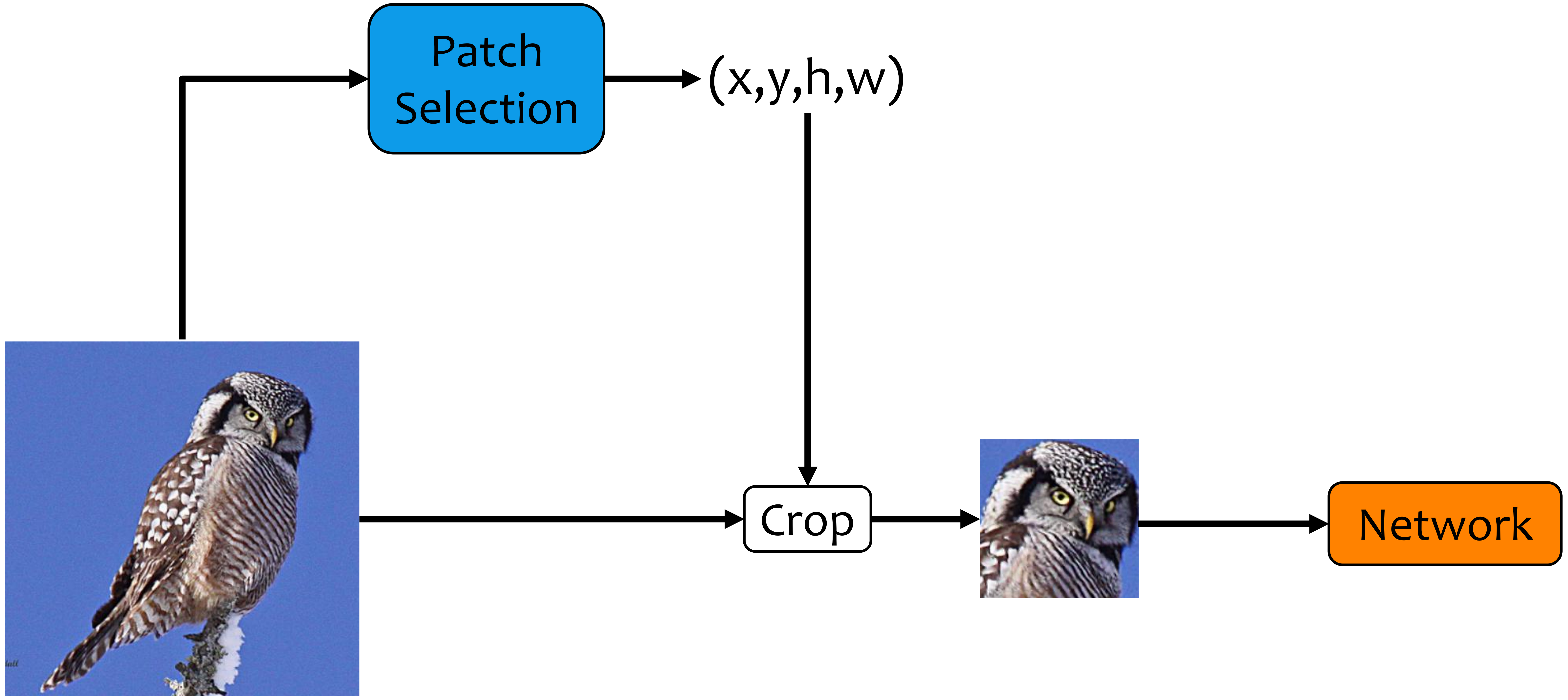
Pixel-level Dynamic Network

- **Sampling module:** generate a sampling indicator mask
- **Sparse Convolution:** compute features at sampled points
- **Interpolation module:** reconstruct entire feature map



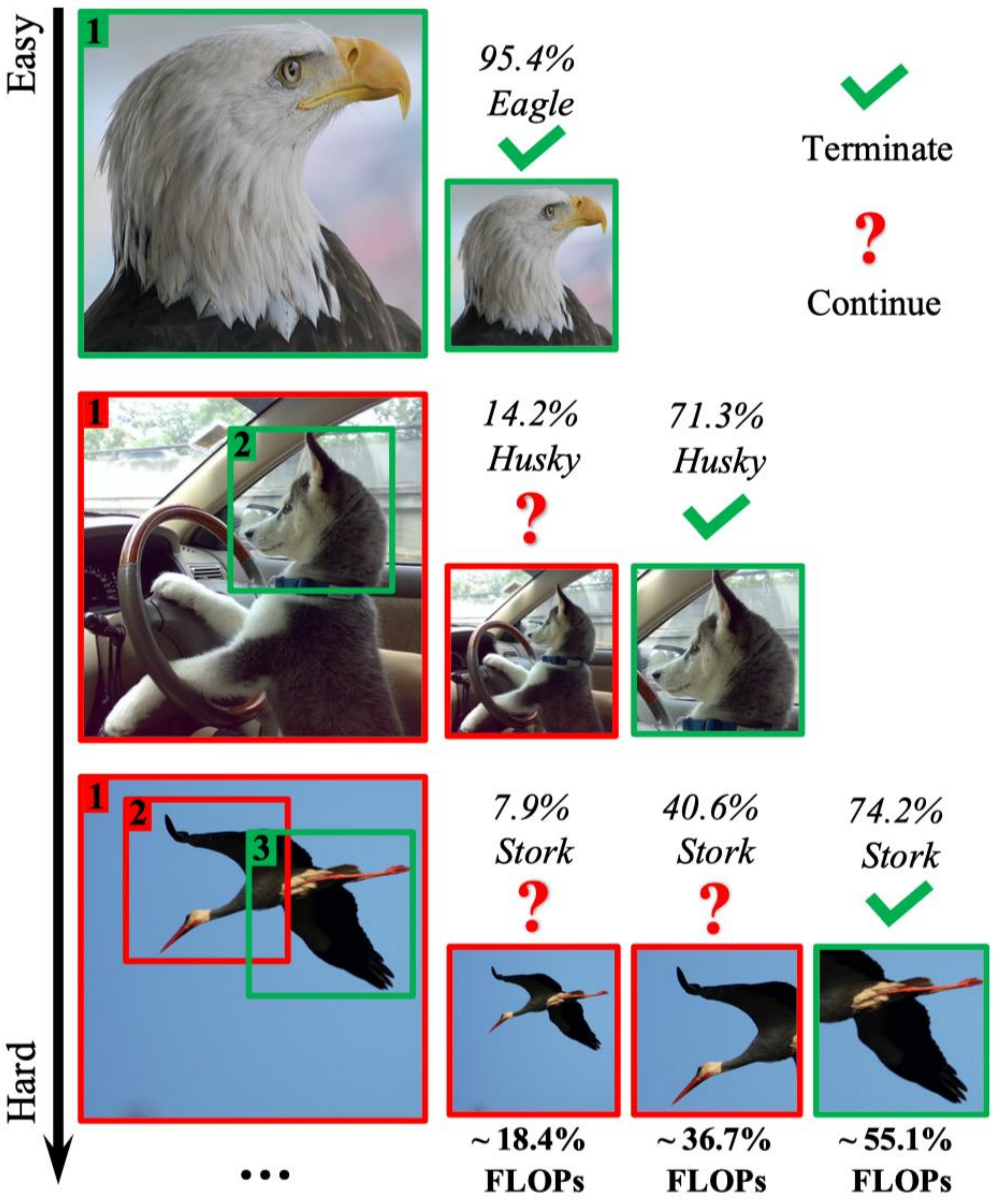
- Verelst, T., & Tuytelaars, T. (2020). Dynamic convolutions: Exploiting spatial sparsity for faster inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2320-2329).
- Xie, Z., Zhang, Z., Zhu, X., Huang, G., & Lin, S. (2020, August). Spatially adaptive inference with stochastic feature sampling and interpolation. In European Conference on Computer Vision (pp. 531-548). Springer, Cham.

Region-level Dynamic Network



Region-level Dynamic Network

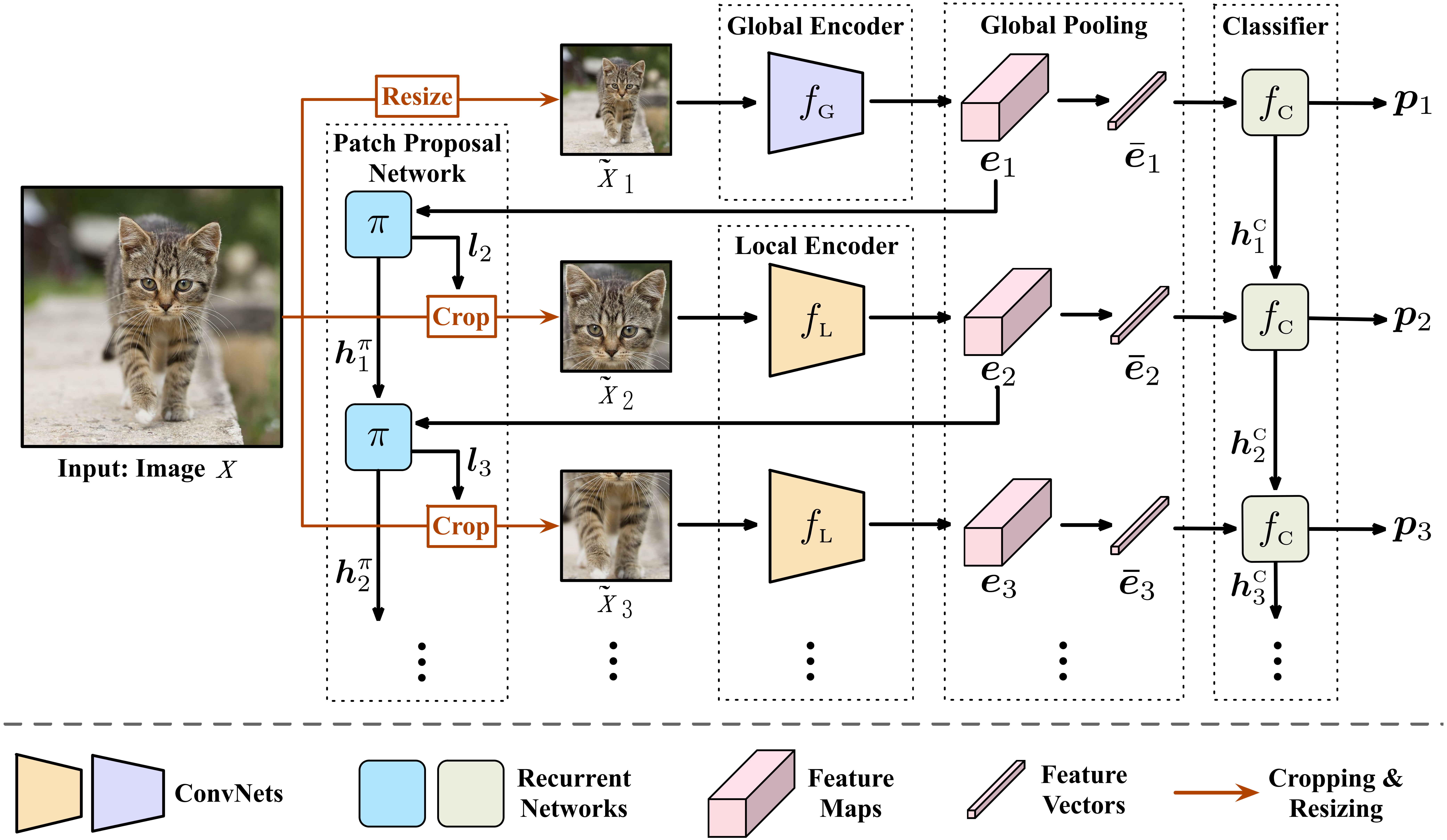
Human visual system processes information progressively.



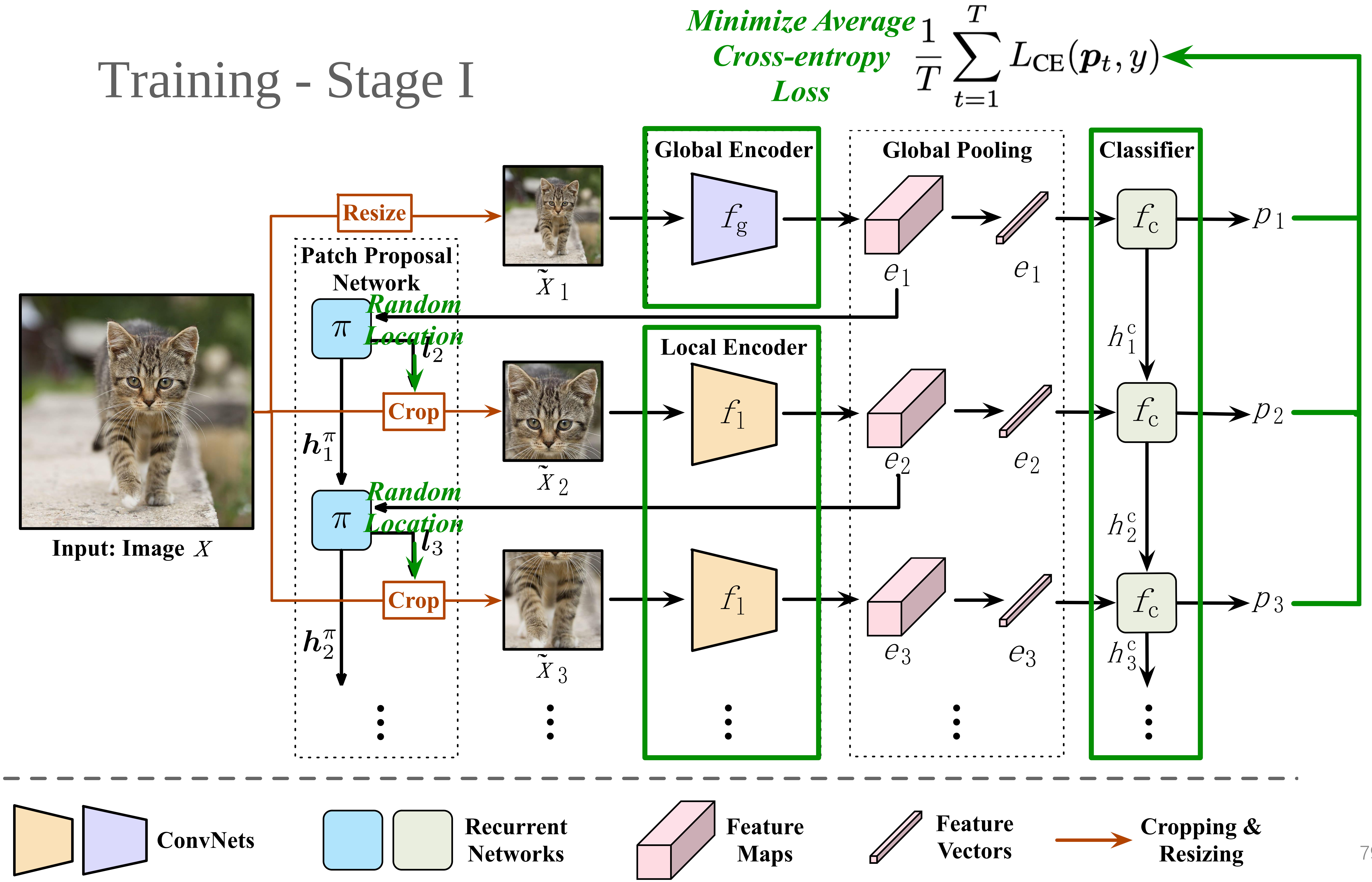
- Wang, Y., Lv, K., Huang, R., Song, S., Yang, L., & Huang, G. (2020). Glance and focus: a dynamic approach to reducing spatial redundancy in image classification. arXiv preprint arXiv:2010.05300.

Glance and Focus Network (GFNet)

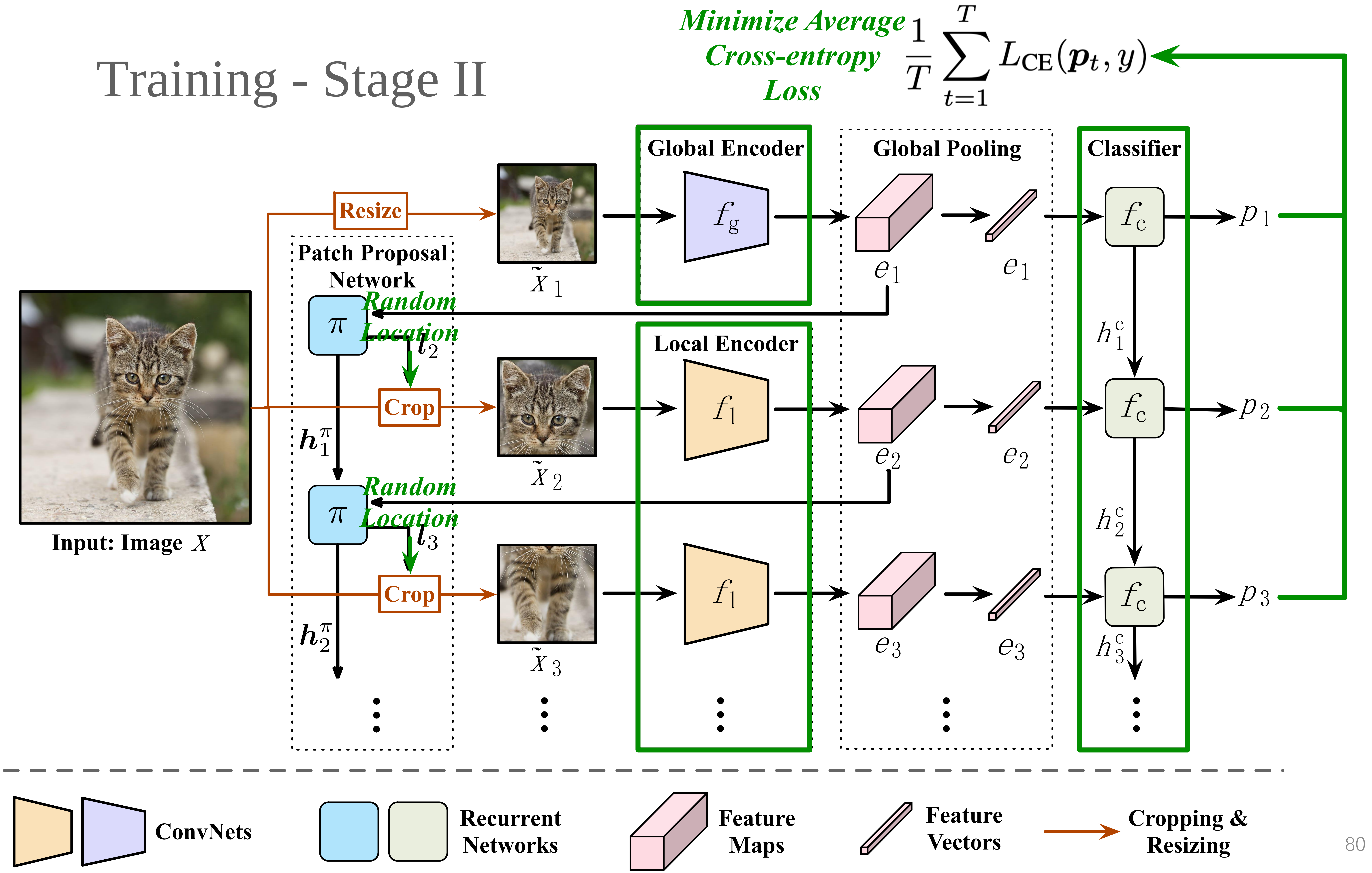
Network Architecture



Training - Stage I



Training - Stage II

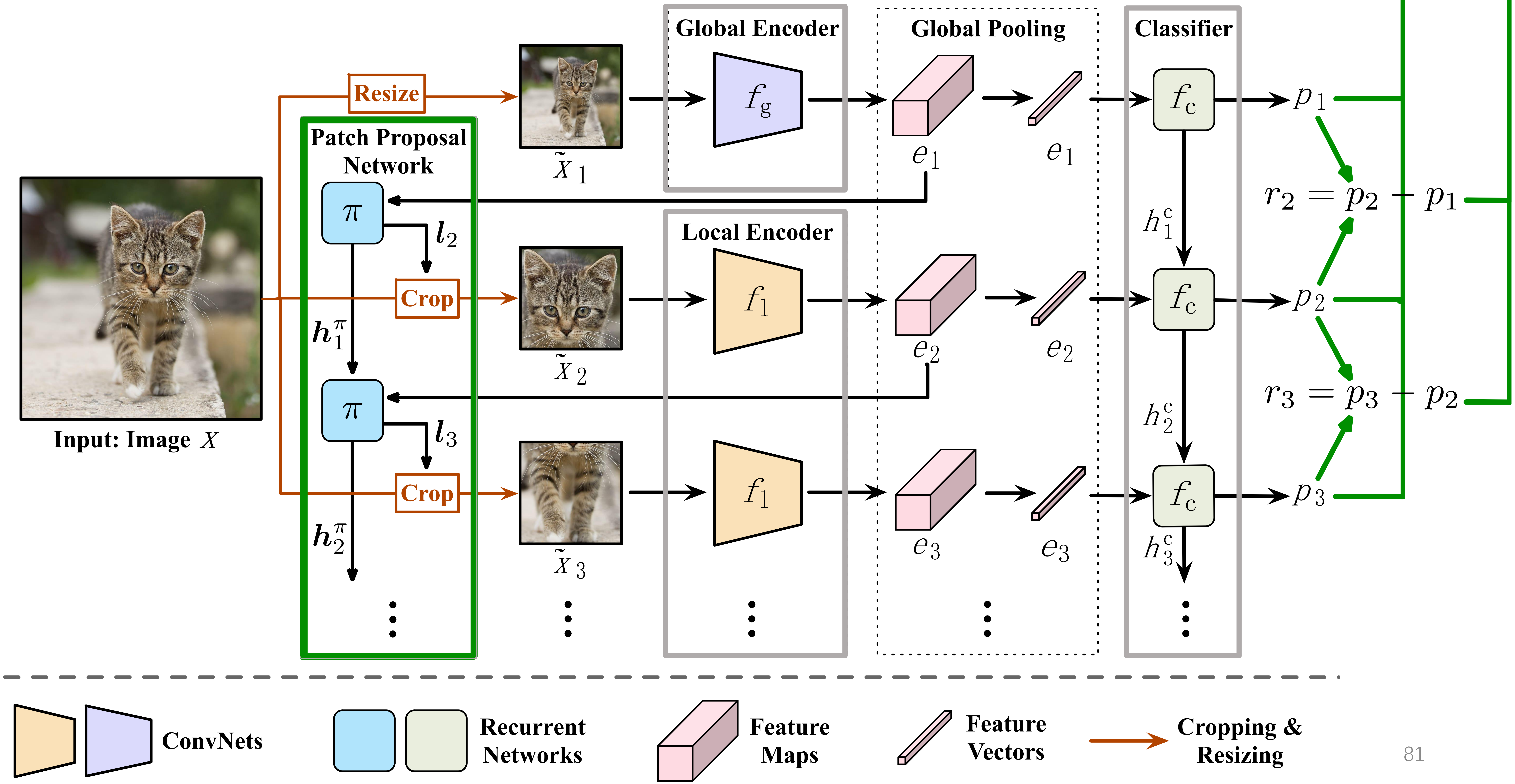


Training - Stage II

Minimize Average Cross-entropy Loss

Maximize the Discounted Rewards

$$\max_{\pi} \mathbb{E}\left[\sum_{t=2}^{T} \gamma^{t-2} r_t\right]$$

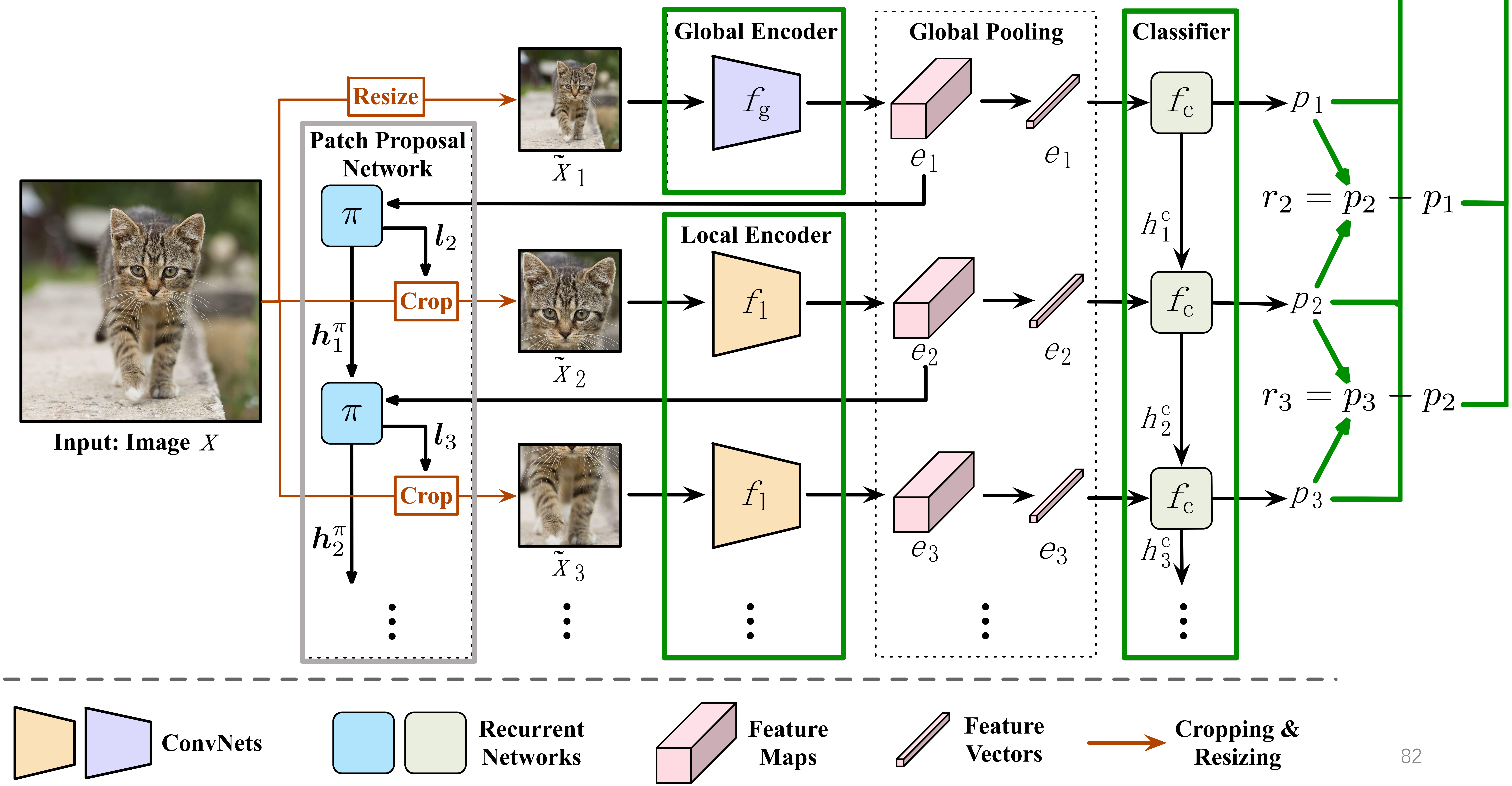


Training - Stage III

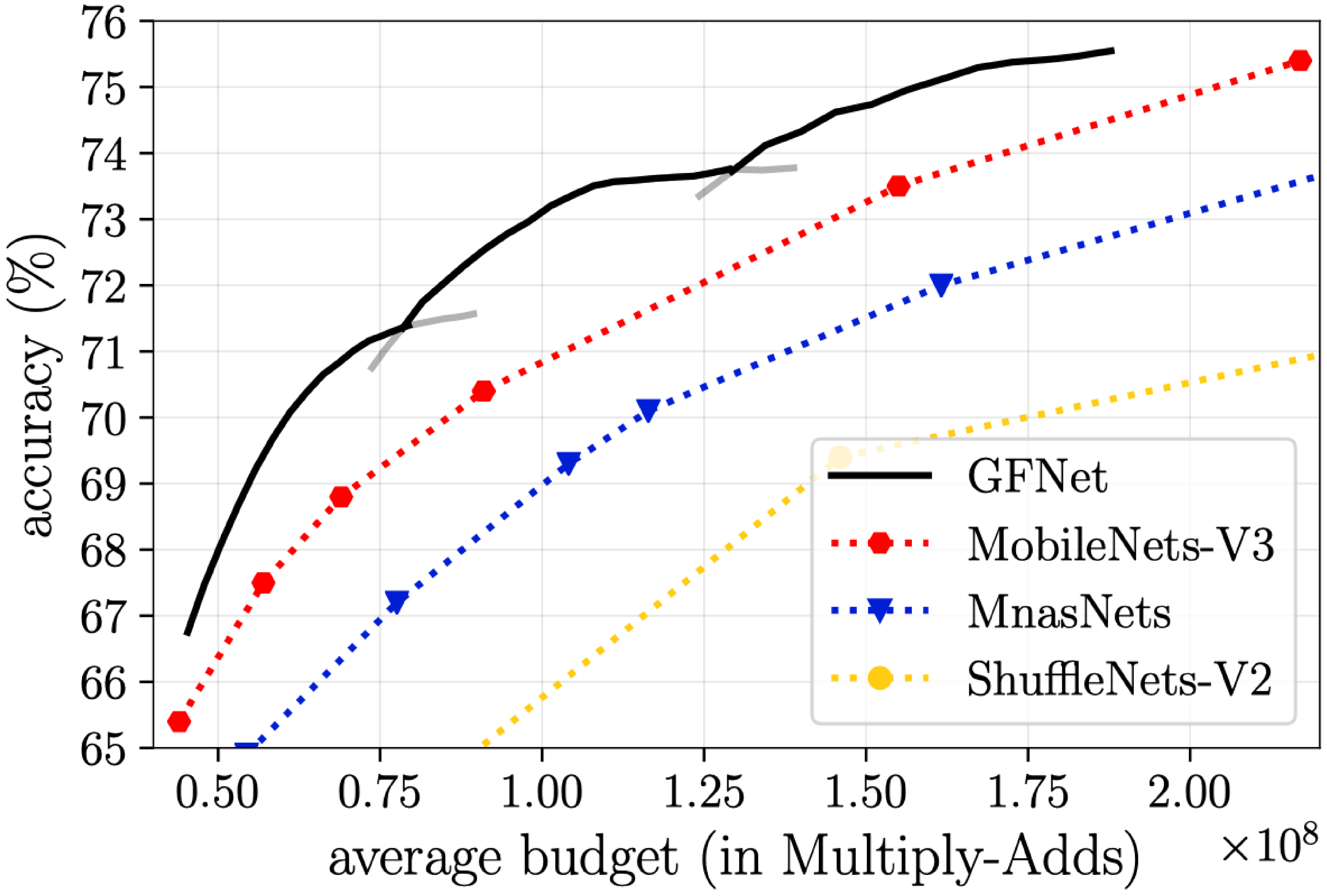
Minimize Average Cross-entropy Loss

Maximize the Discounted Rewards

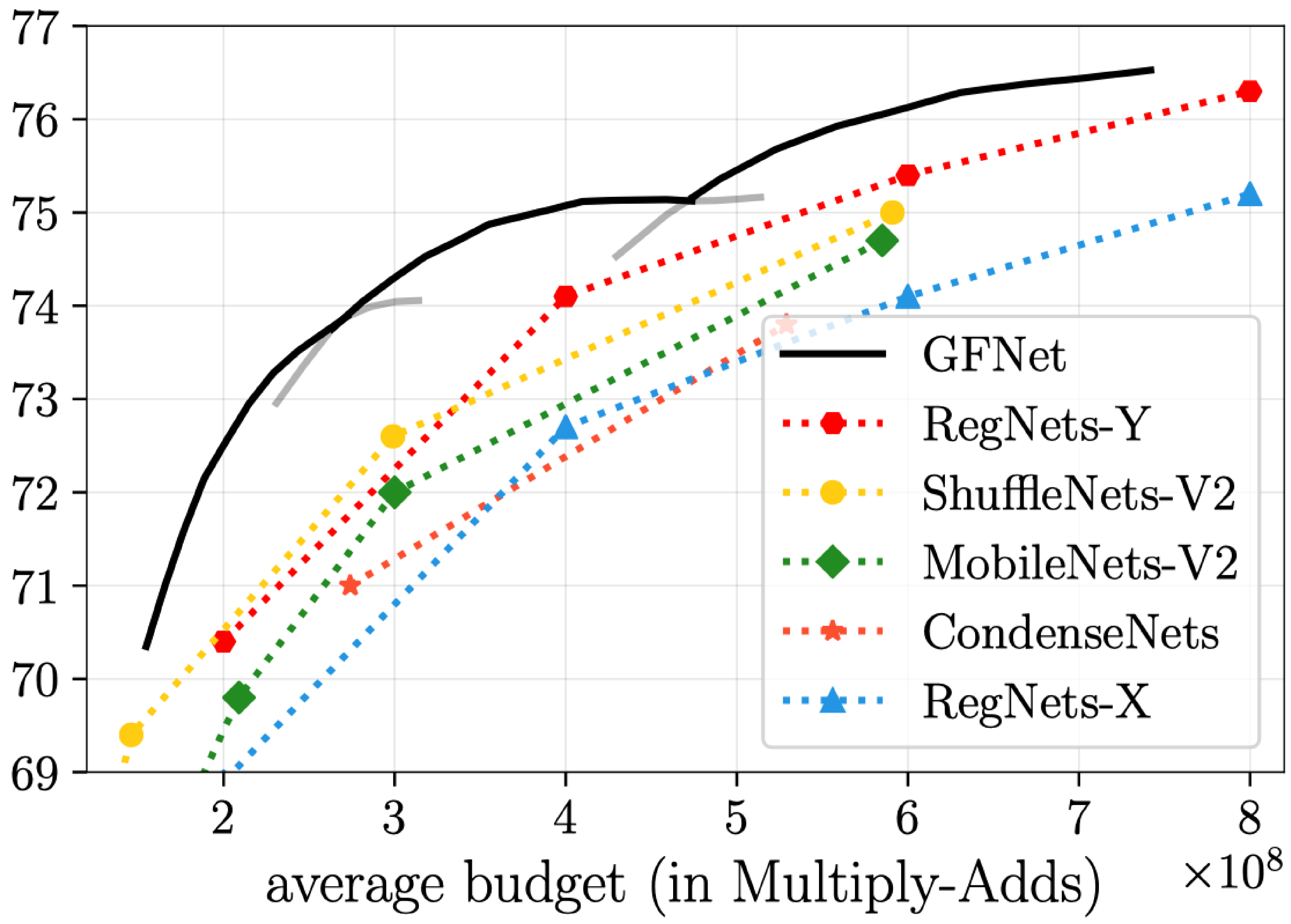
$$\max_{\pi} \mathbb{E}\left[\sum_{t=2}^{T} \gamma^{t-2} r_t\right]$$



Results (FLOPs)

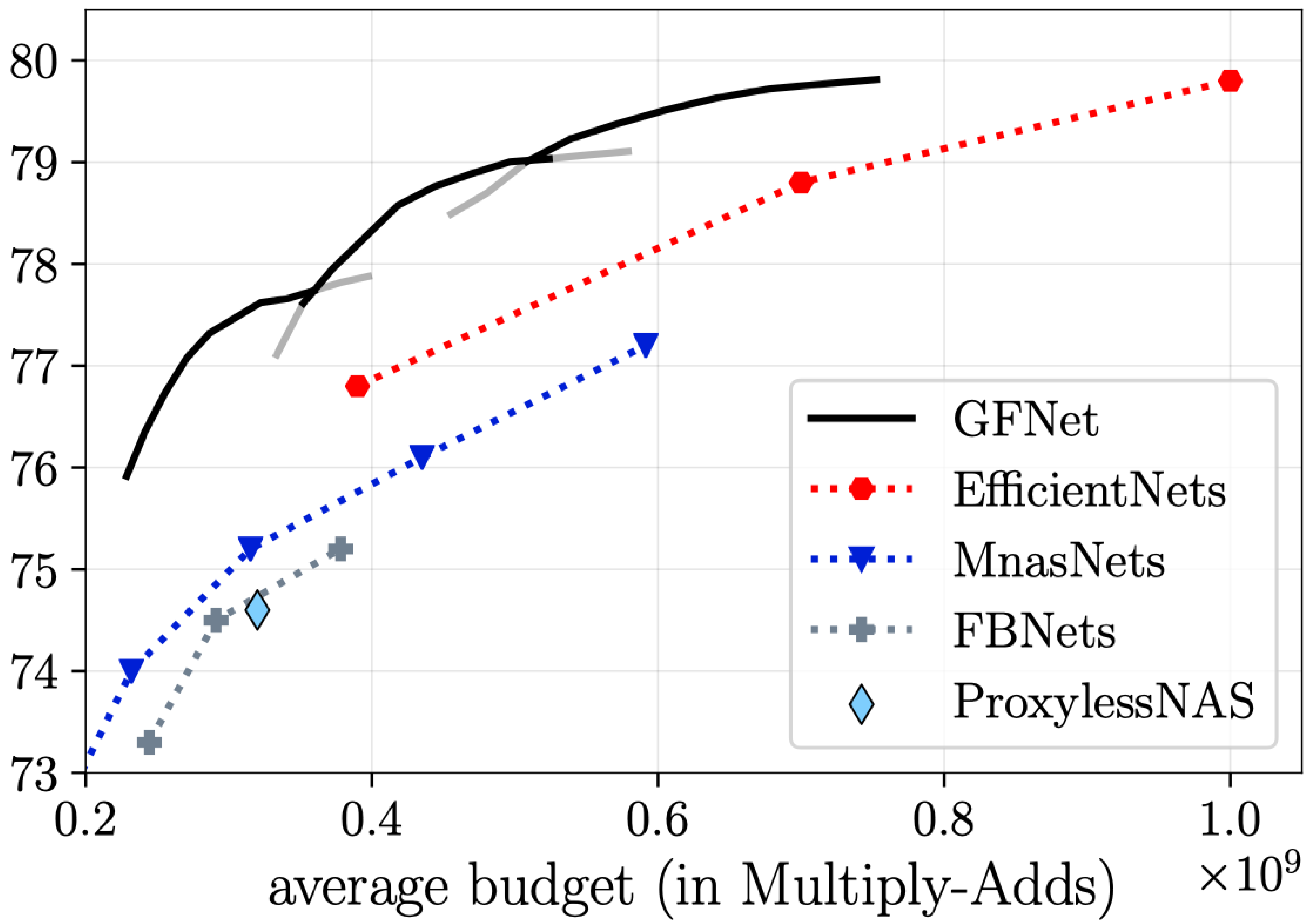


(a) MobileNet-V3



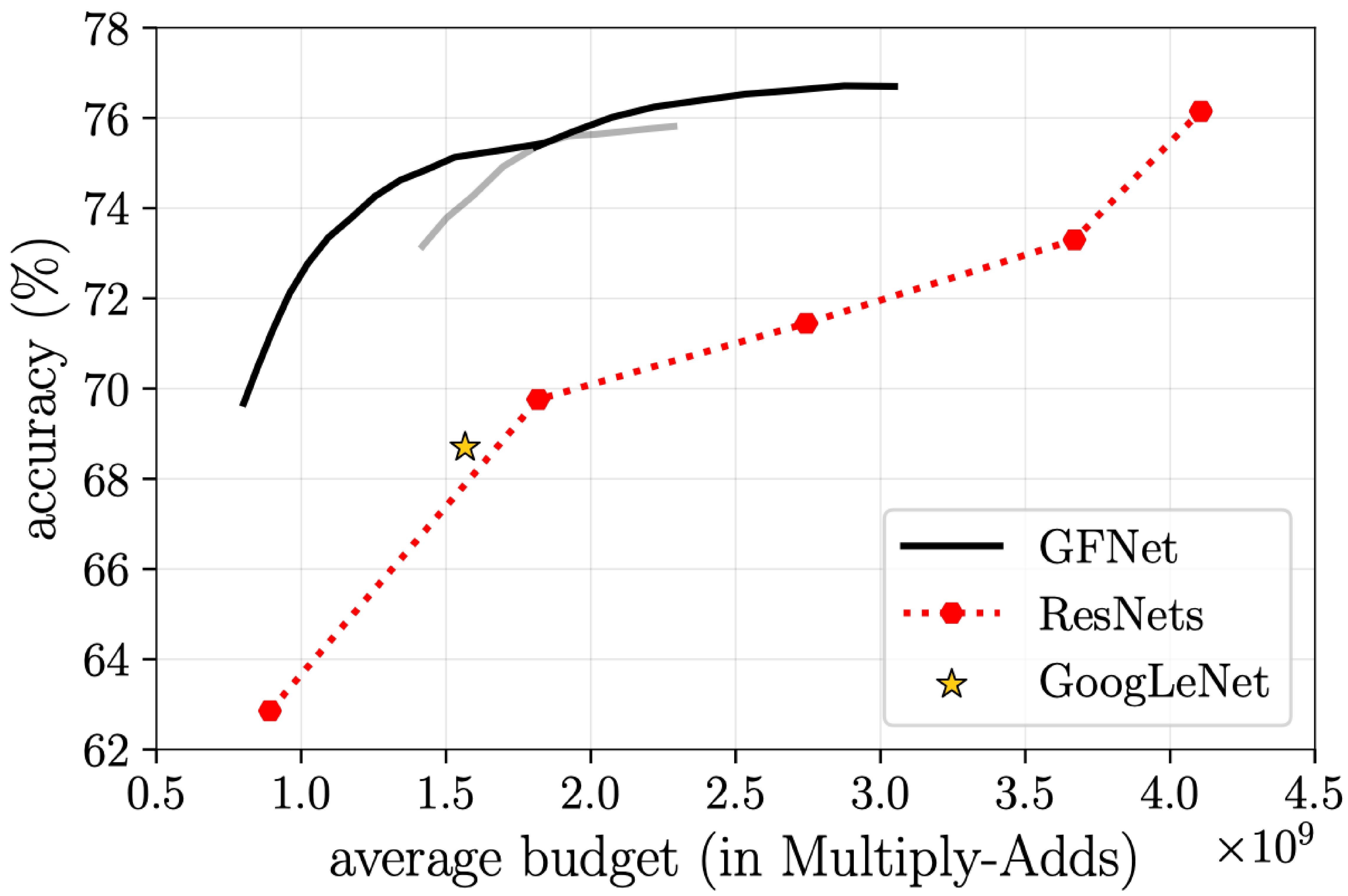
(b) RegNet-Y

Results (FLOPs)

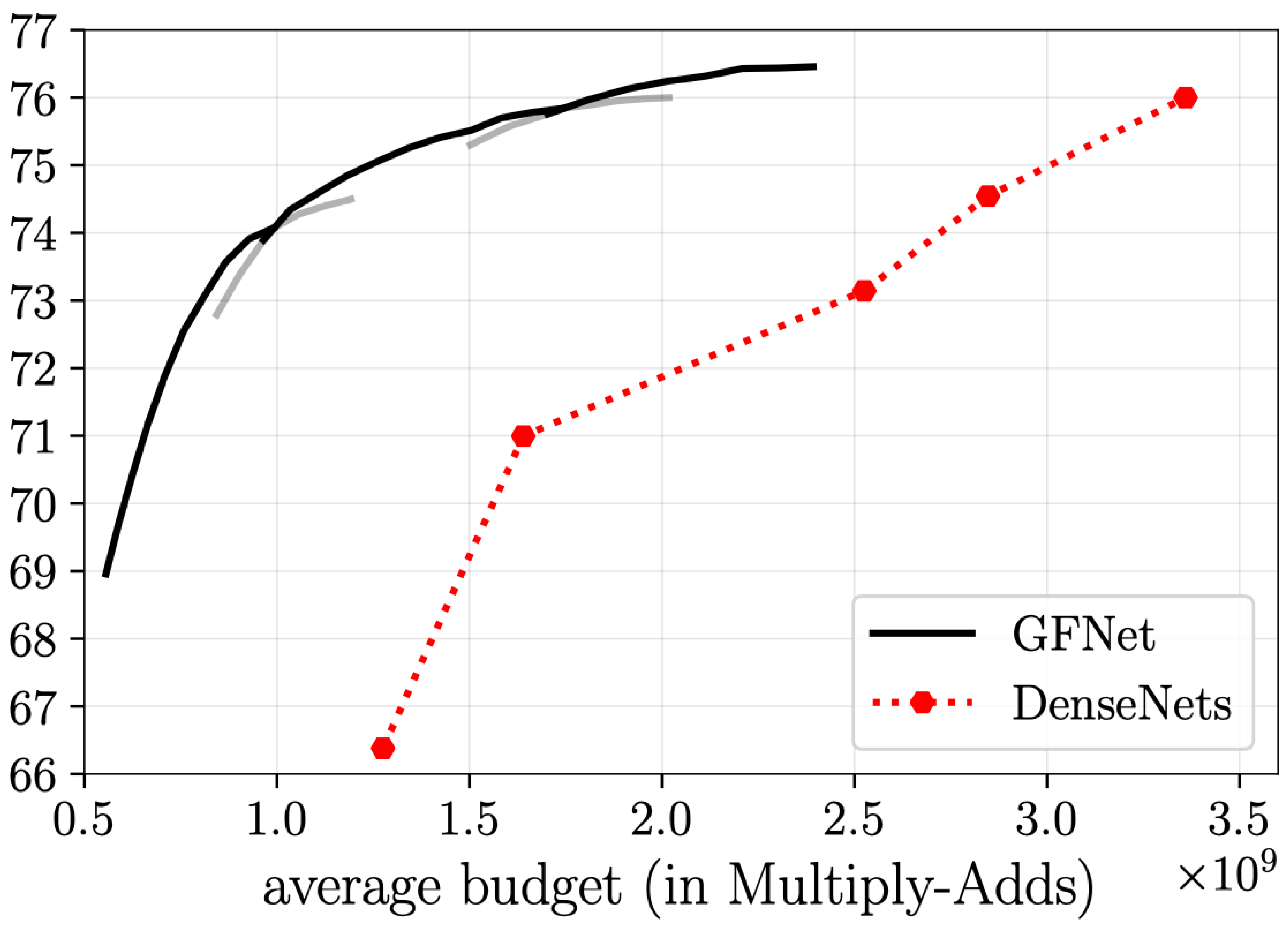


(c) EfficientNet

Results (FLOPs)

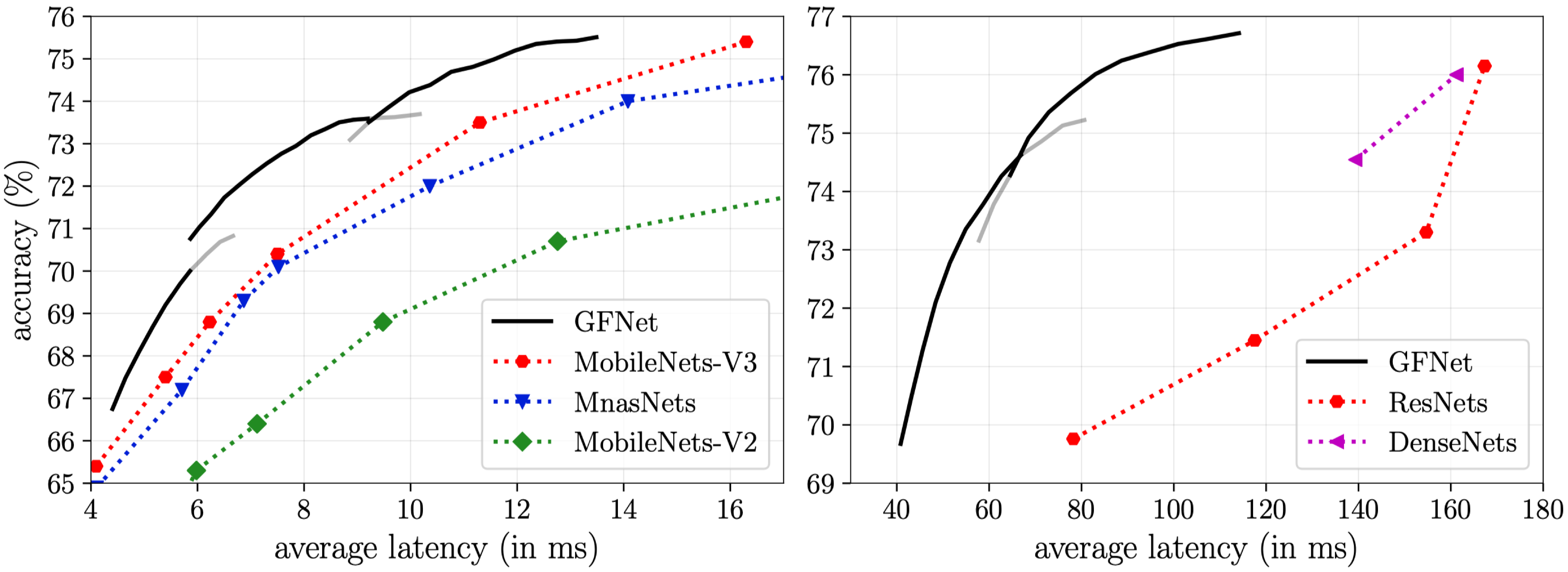


(d) ResNet



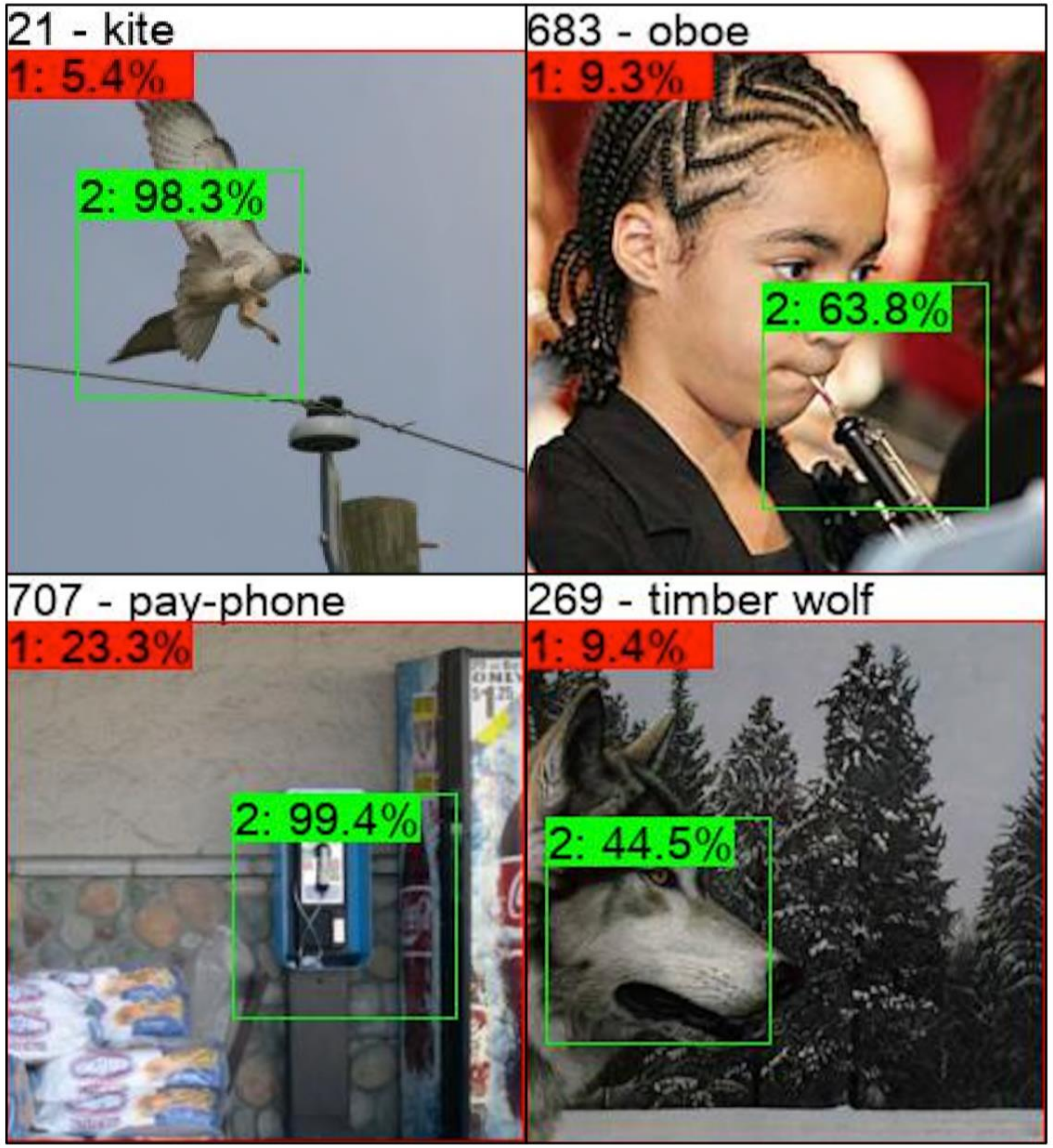
(e) DenseNet

Results (iPhone XS Max)

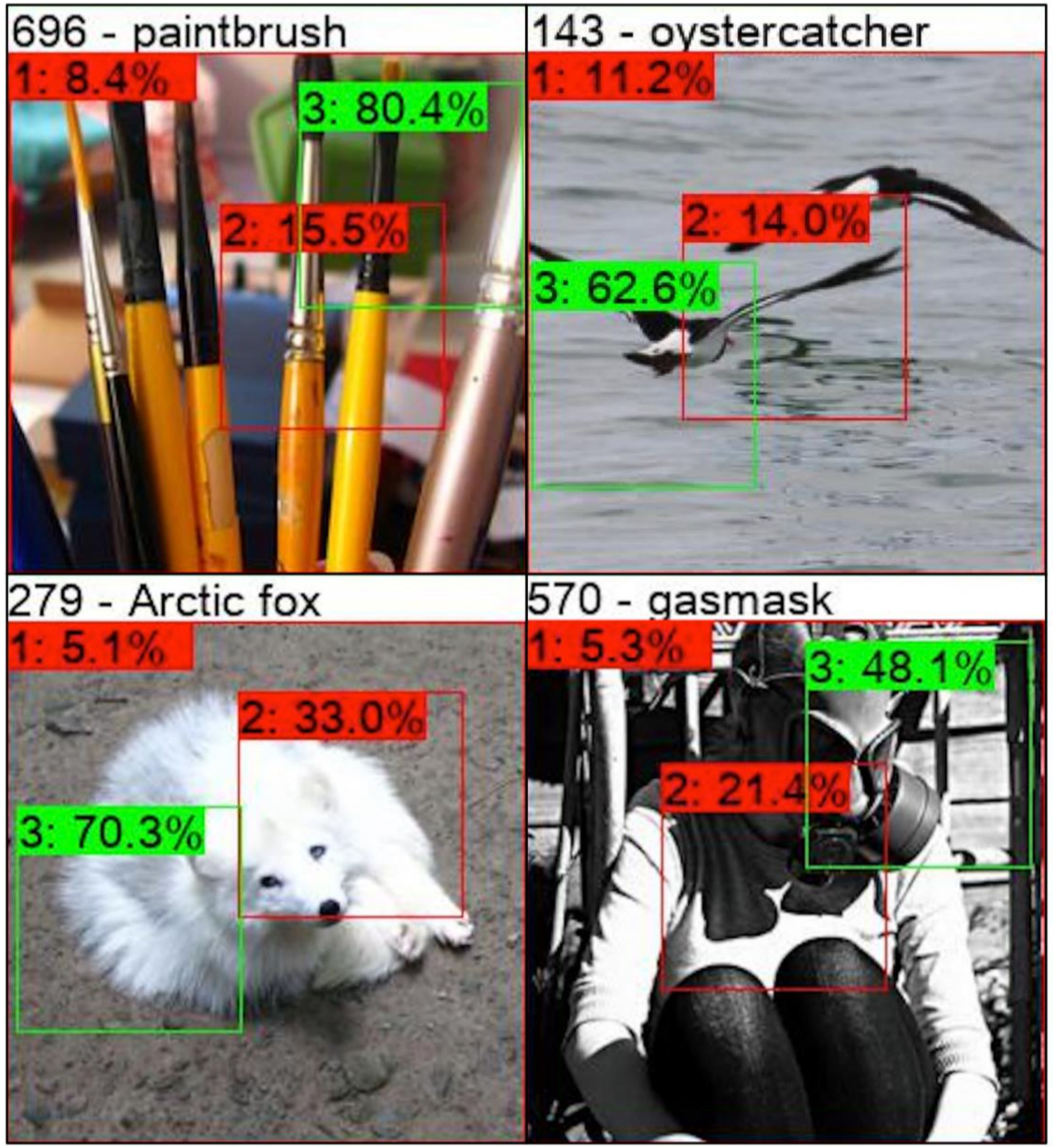


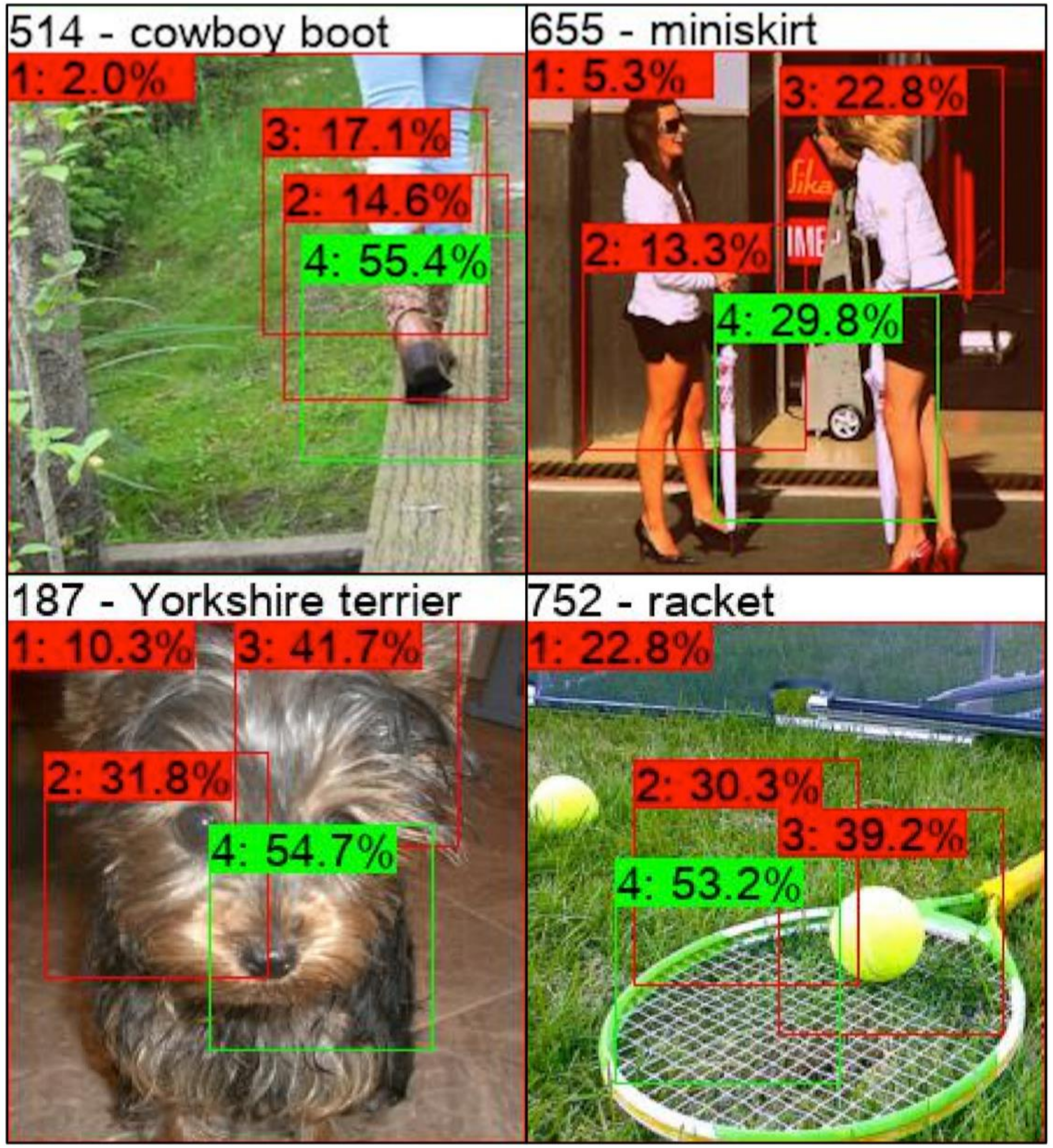
(a) MobileNet-V3 (b) ResNet

Results (Visualization)



Results (Visualization)

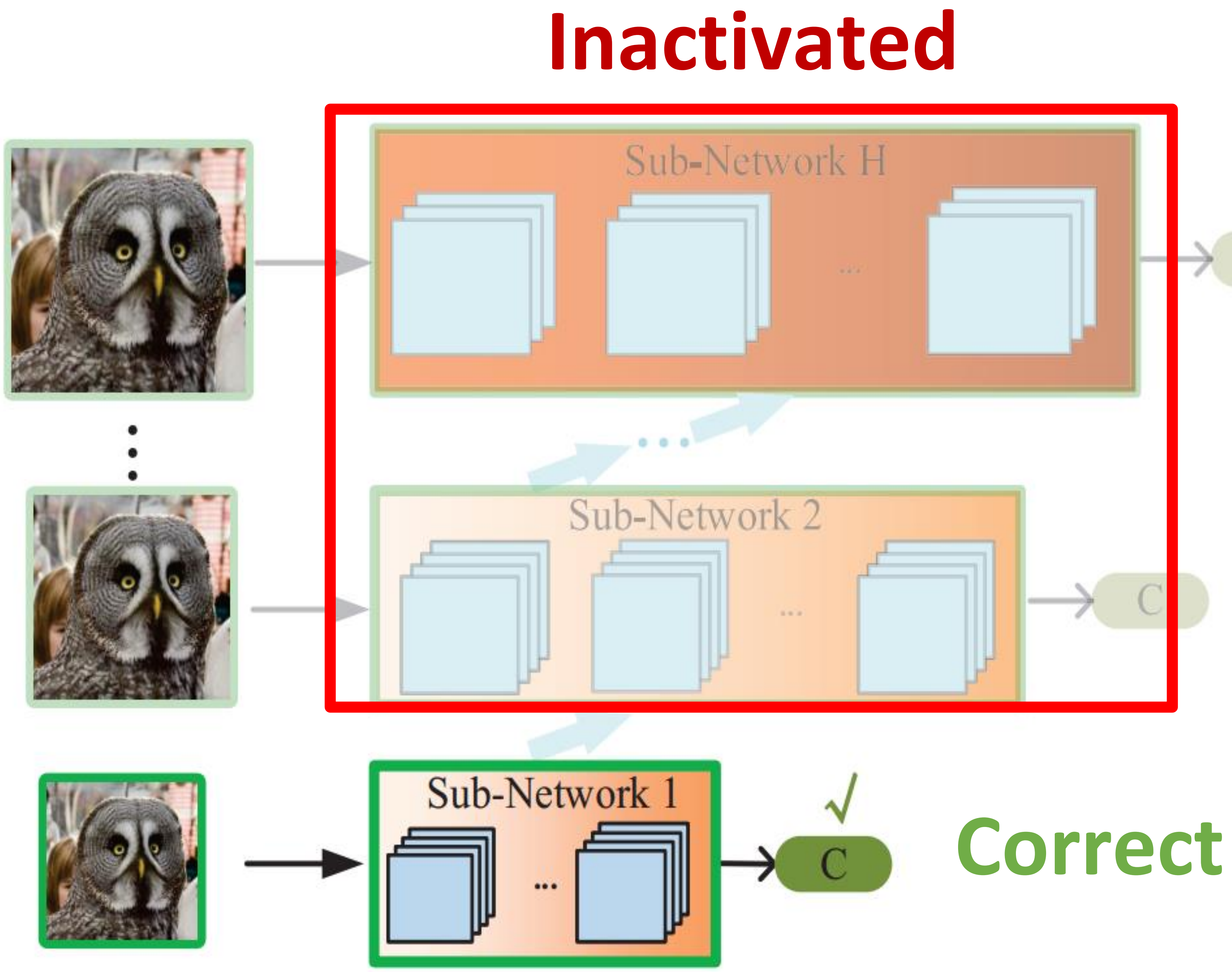




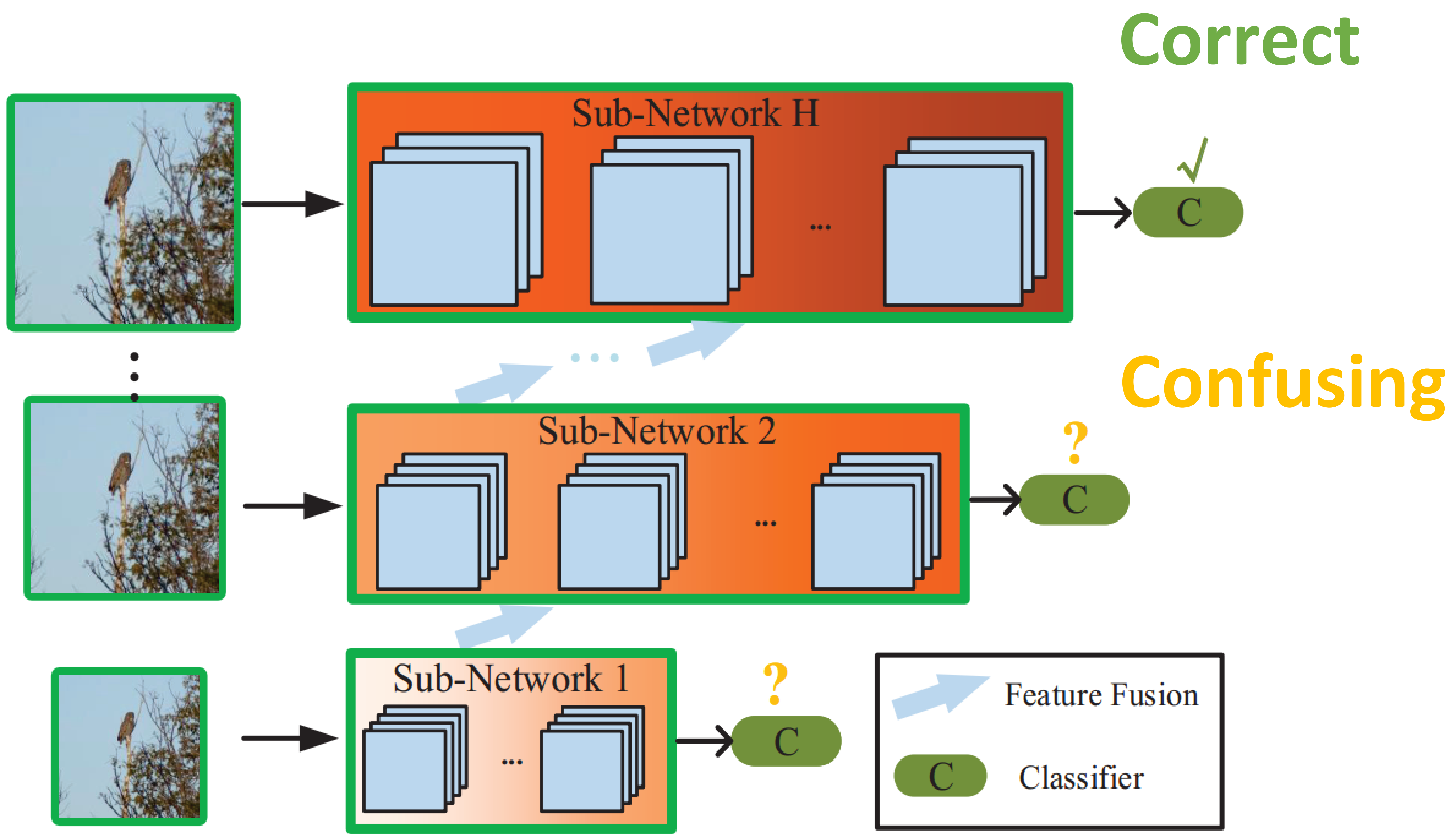
Low-resolution *representations are sufficient to recognize* ***"easy"*** *samples.*

Resolution Adaptation

- **Easy samples** (e.g. images containing large objects):

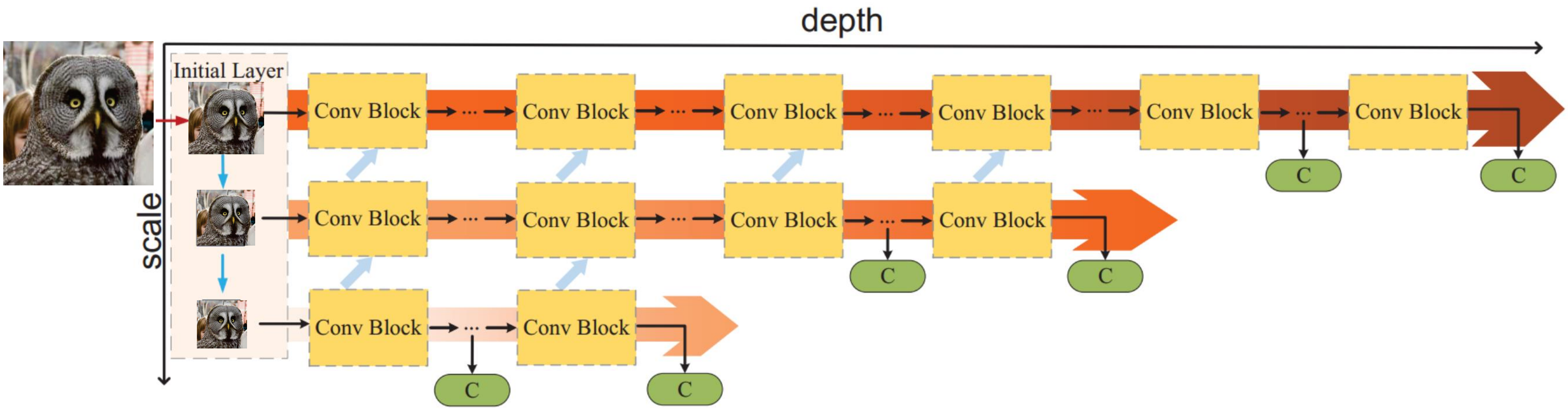


- **Hard samples** (e.g. images containing tiny objects) :



- Yang, L., Han, Y., Chen, X., Song, S., Dai, J., & Huang, G. (2020). Resolution adaptive networks for efficient inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2369-2378).

Resolution Adaptive Network



(Prediction Confidence < ε_1)

(Prediction Confidence > ε_2)

Confusing

Exit

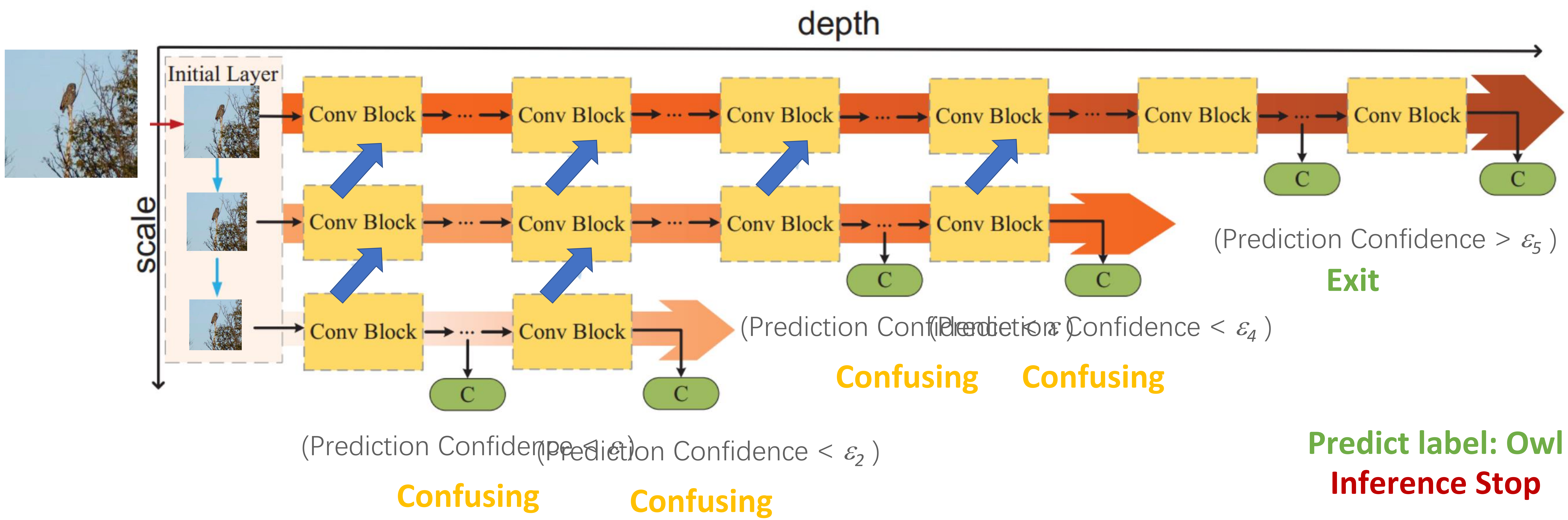
Predict label: Owl

Inference Stop

- Threshold (ε_i) controlling exiting of a sample.

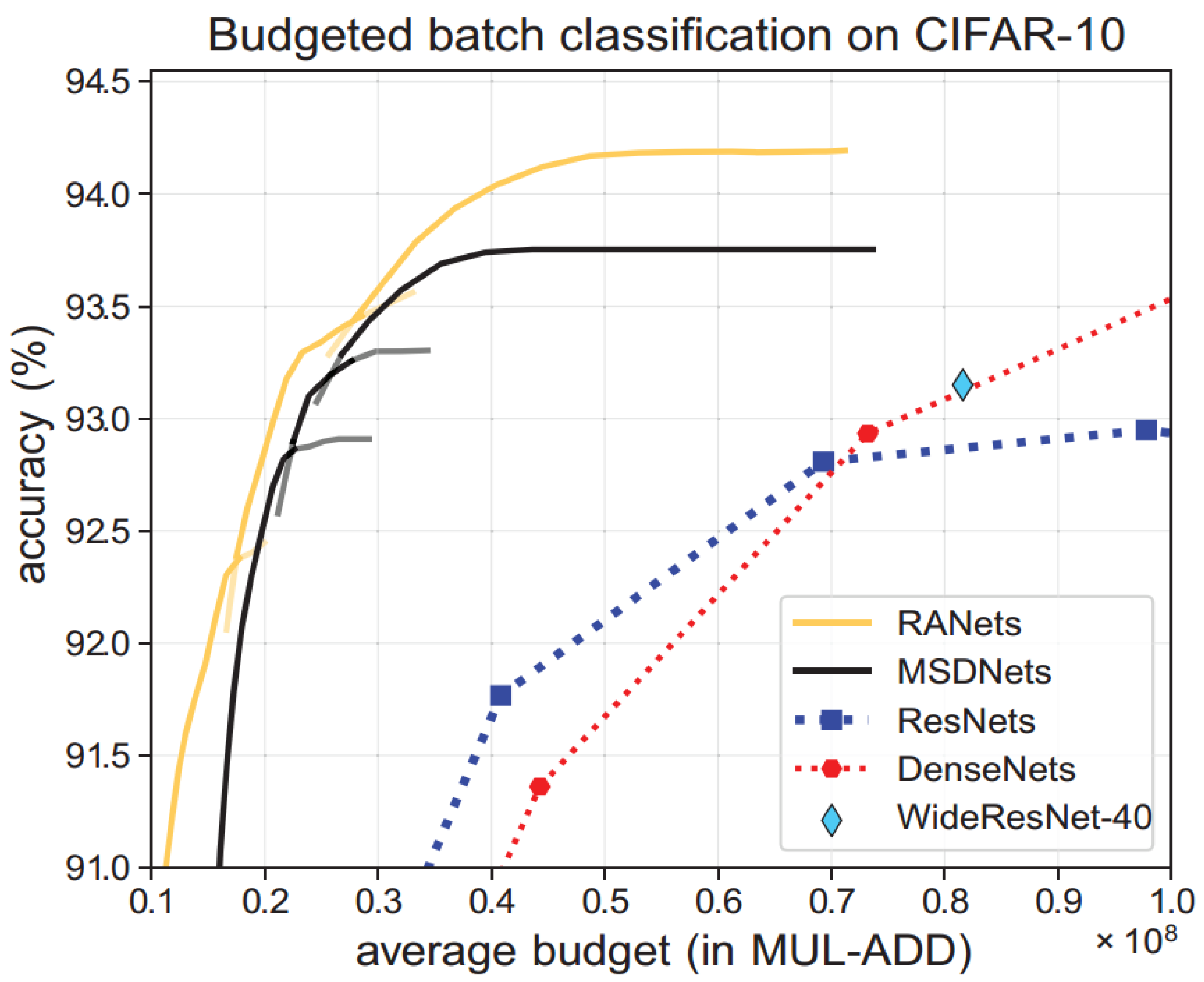
- Yang, L., Han, Y., Chen, X., Song, S., Dai, J., & Huang, G. (2020). Resolution adaptive networks for efficient inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2369-2378).

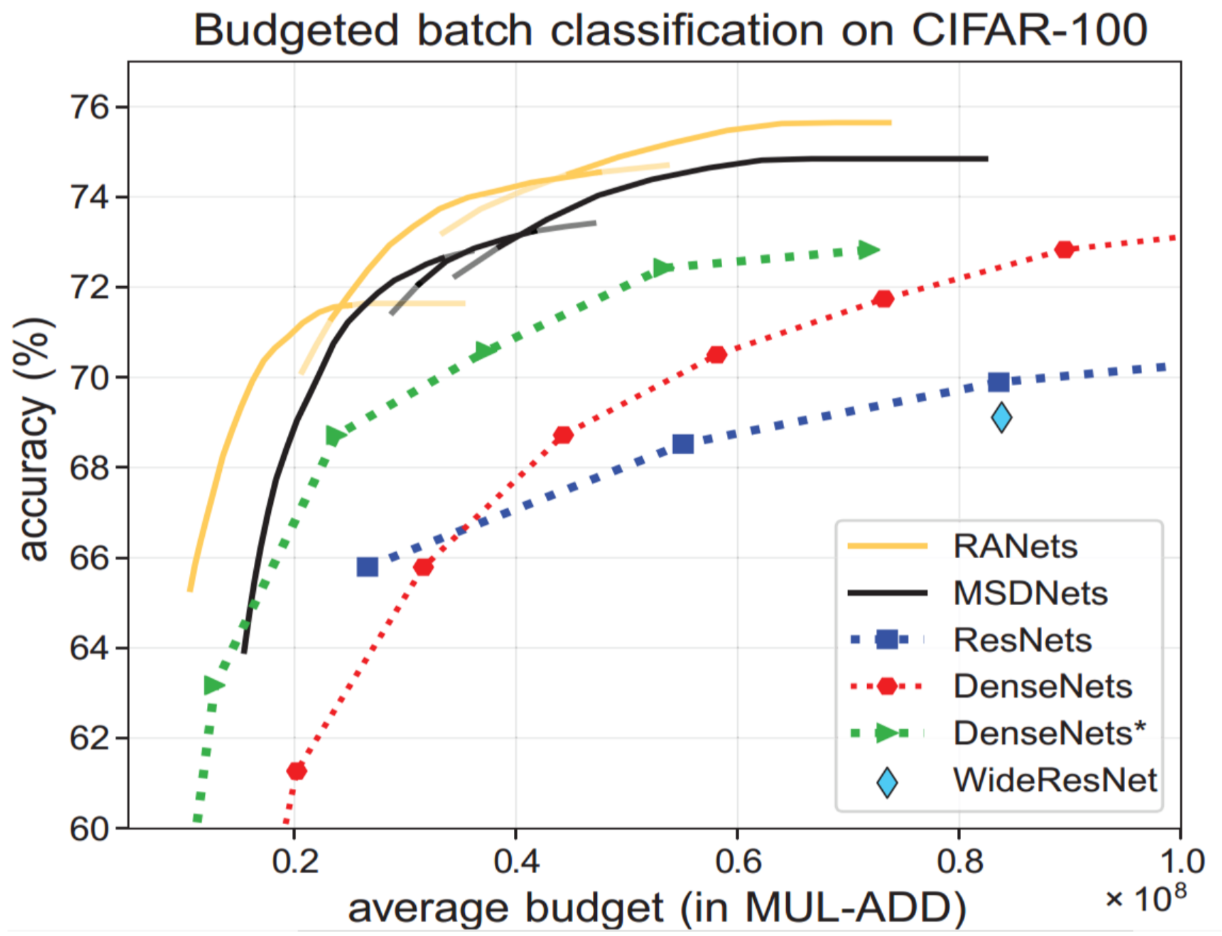
Resolution Adaptive Network

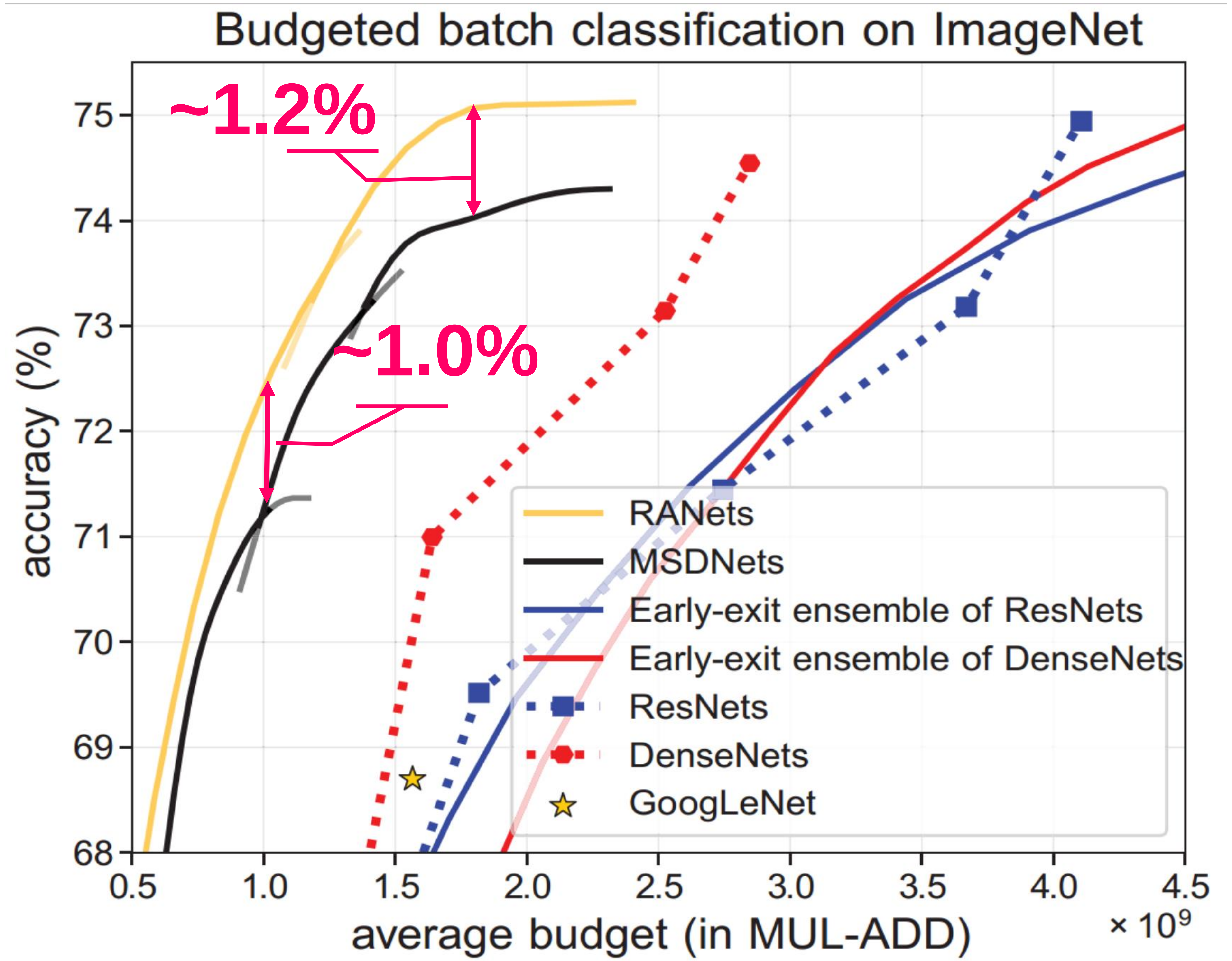


▪ Yang, L., Han, Y., Chen, X., Song, S., Dai, J., & Huang, G. (2020). Resolution adaptive networks for efficient inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2369-2378).

Results: Budgeted Batch Classification

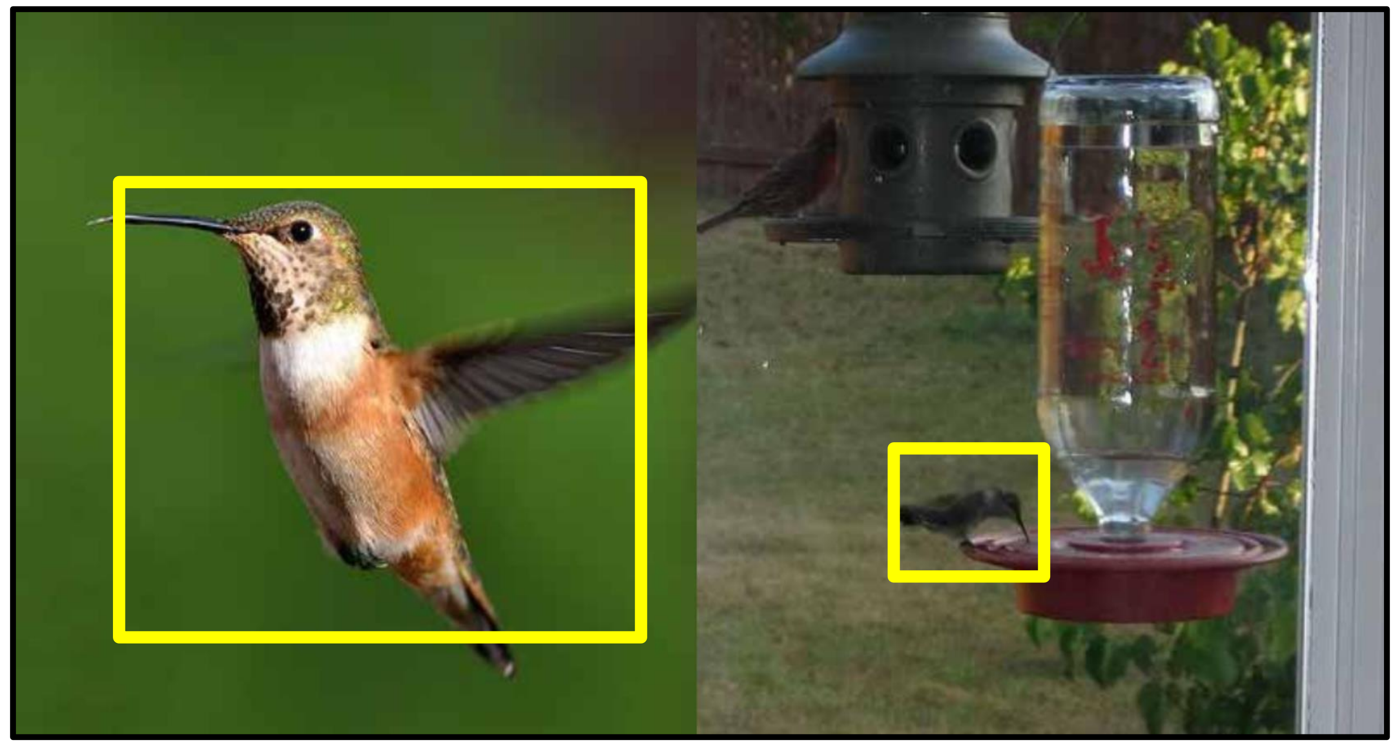






Visualization

Easy hard

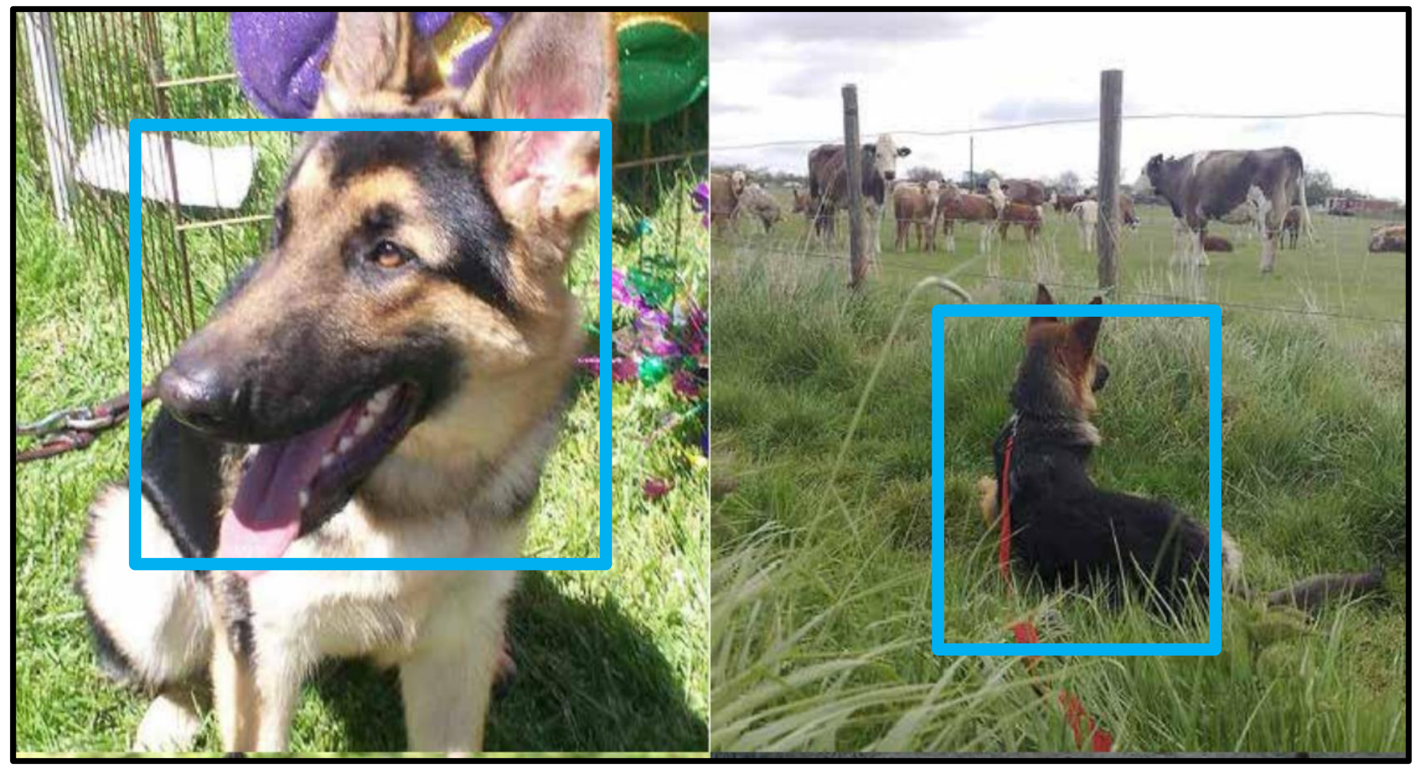


■ Images with tiny objects can be hard samples.

Easy Hard

■ Images with multiple objects can be hard samples.

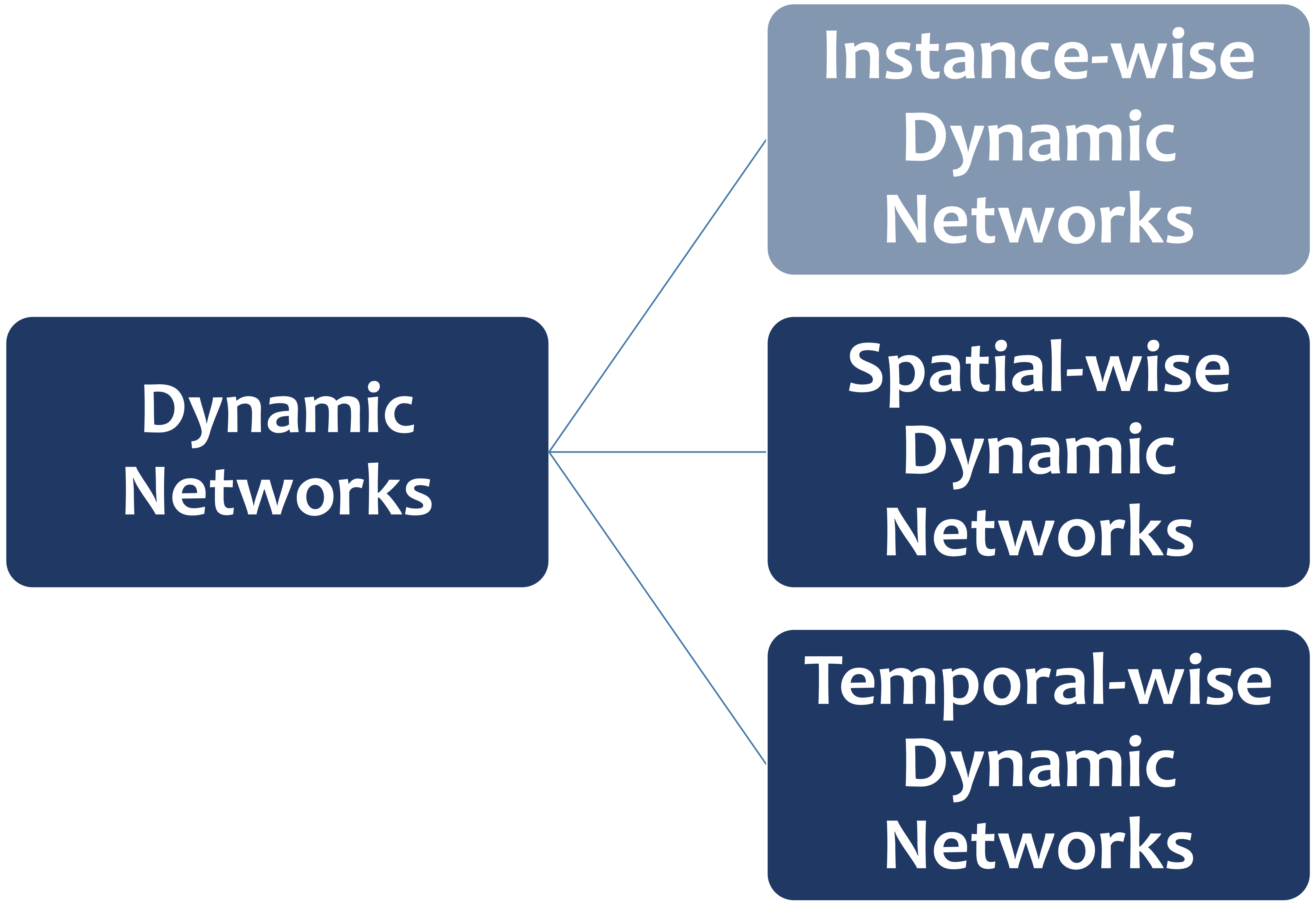
Easy Hard



■ Images with objects w/o representative characteristics can be hard samples.

1. Overview of CNN backbones
2. Architecture design for mobile CNNs
3. Dynamic CNNs for mobile applications
   A. Sample-wise Dynamic Networks
   B. Spatial-wise Dynamic Networks
   C. Temporal-wise Dynamic Networks

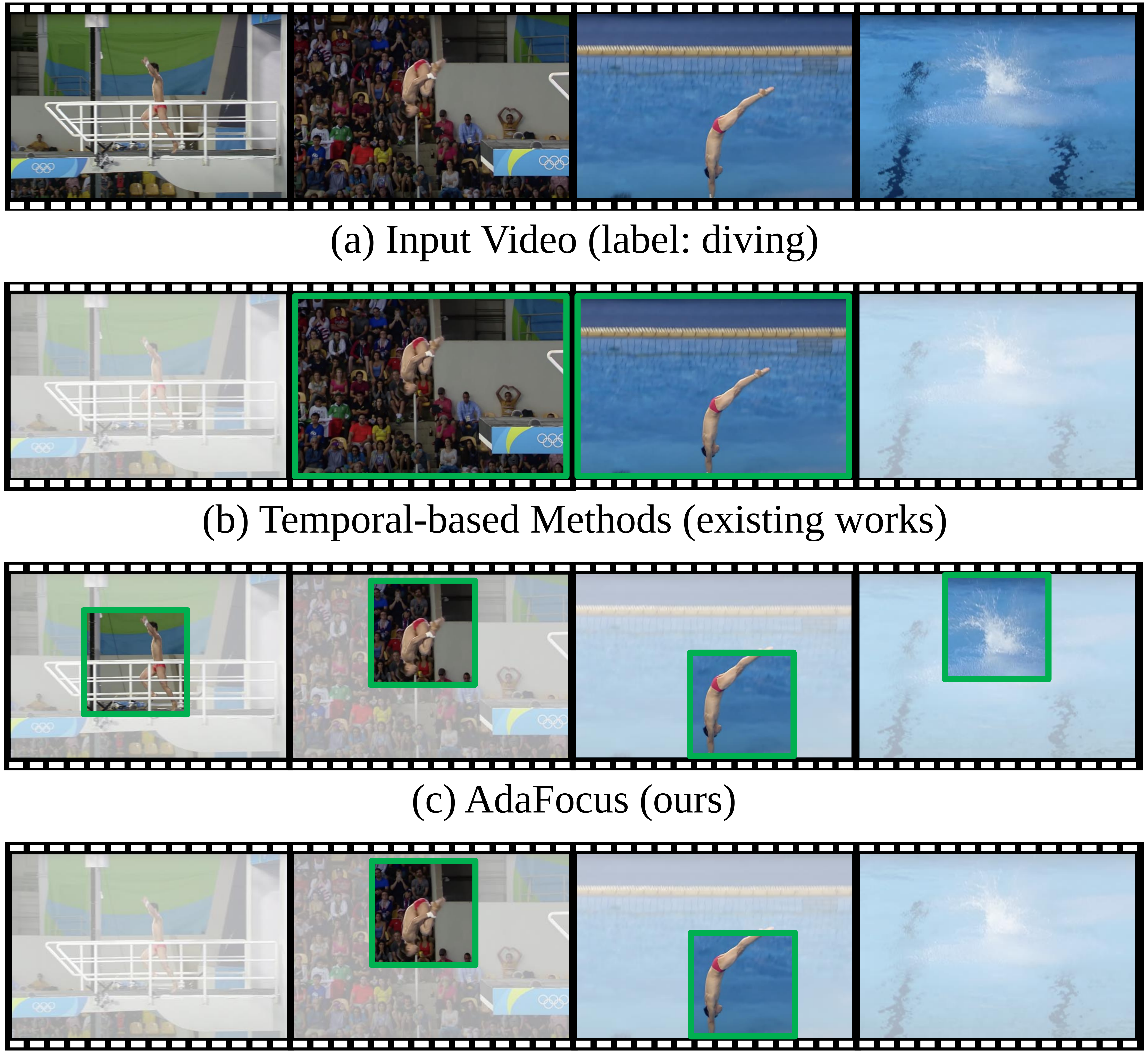
Spatially & **Temporally** Adaptive Inference for Videos



Sequential Data: Video/Text

A small portion of frames have
sufficient task-relevant information!

Adaptive Focus for Efficient Video Recognition

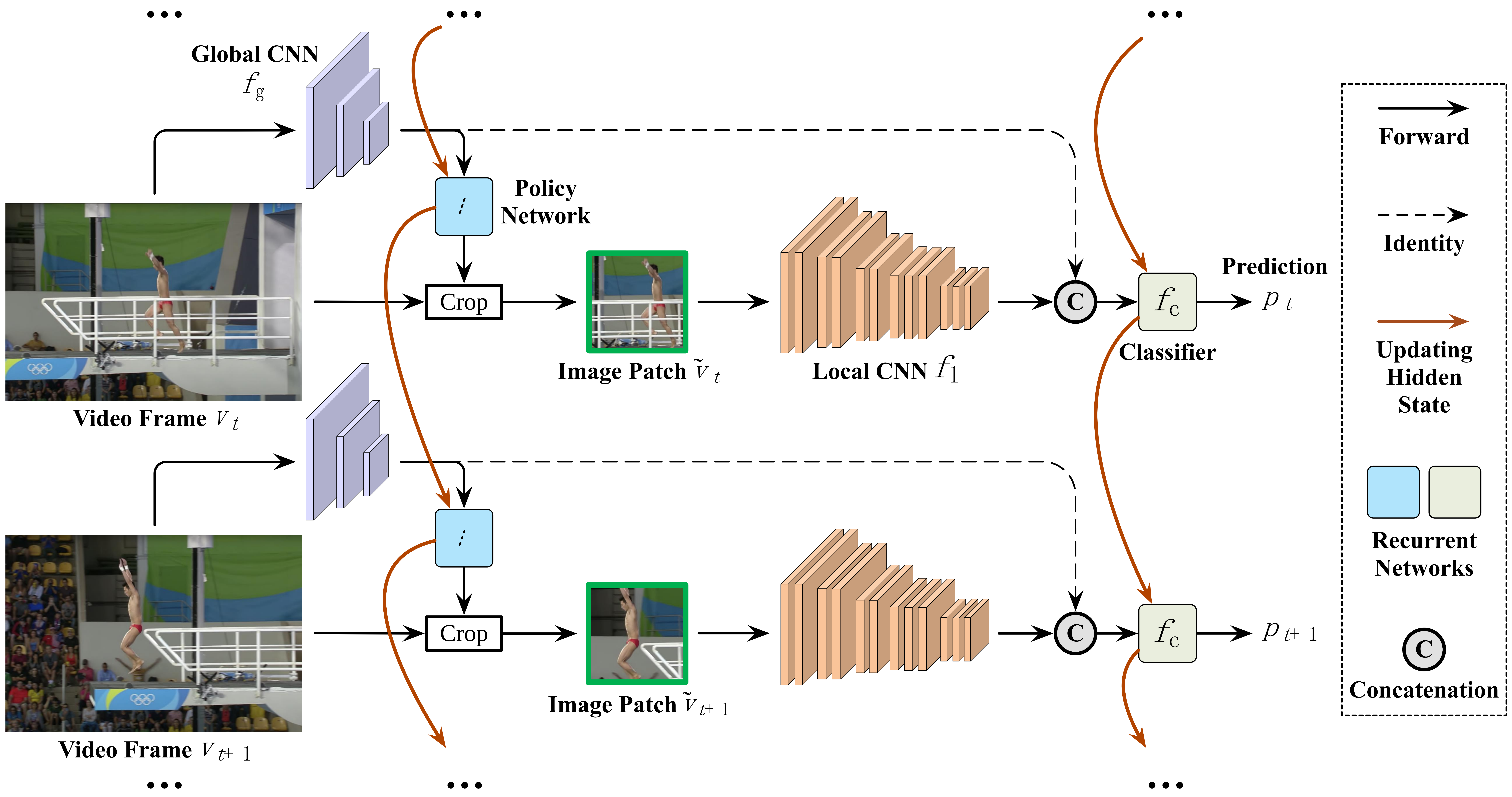


(a) Input Video (label: diving)

(b) Temporal-based Methods (existing works)

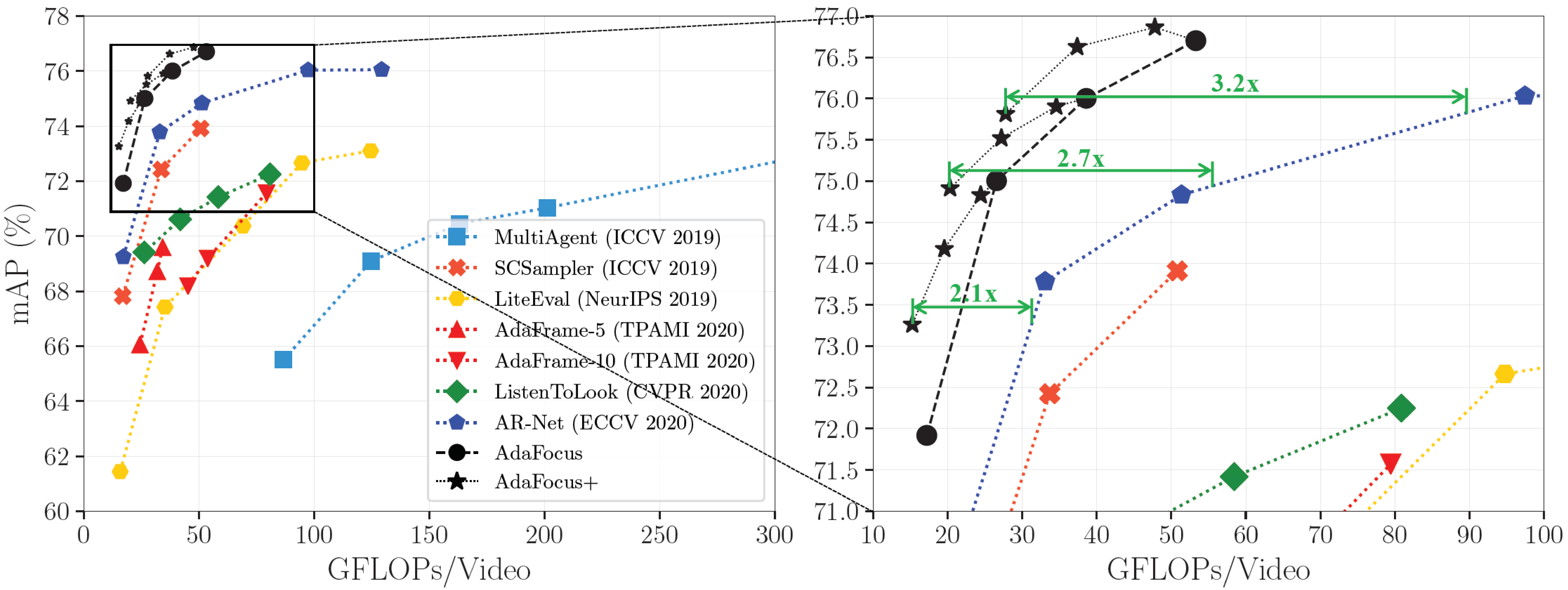
(c) AdaFocus (ours)

(d) AdaFocus+

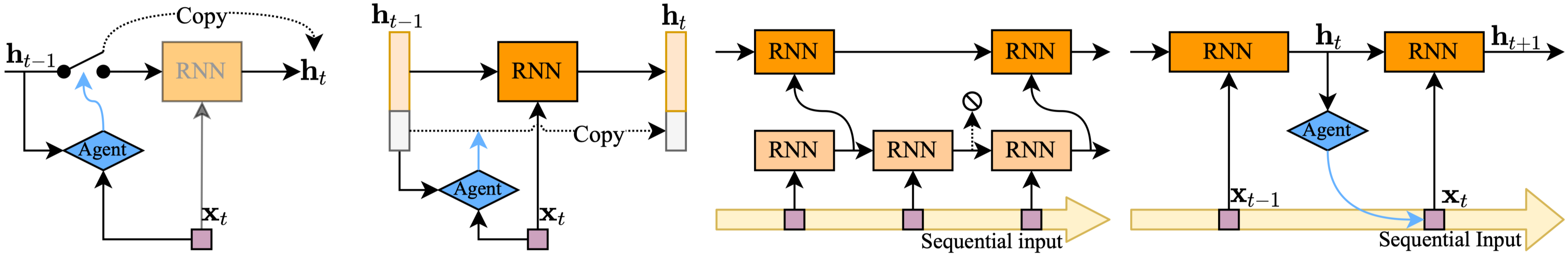


Global CNN f_g
Policy Network
Crop
Image Patch $\tilde{v}_t$
Local CNN f_l
Video Frame v_t
C
f_c
Prediction p_t
Classifier
Crop
Image Patch $\tilde{v}_{t+1}$
Video Frame v_{t+1}
f_c
p_{t+1}
Forward
Identity
Updating Hidden State
Recurrent Networks
Concatenation

Offline Video Recognition on ActivityNet



RNN-based Approaches



(a) Skip update of hidden state. (b) Partial update of hidden state. (c) Hierarchical RNN architecture. (d) Temporal dynamic jumping.

- Campos, V., Jou, B., Giró-i-Nieto, X., Torres, J., & Chang, S. F. (2017). Skip rnn: Learning to skip state updates in recurrent neural networks. arXiv preprint arXiv:1708.06834.
- Seo, M., Min, S., Farhadi, A., & Hajishirzi, H. (2017). Neural speed reading via skim-rnn. arXiv preprint arXiv:1711.02085.
- Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. In ICLR, 2017.
- Adams Wei Yu, Hongrae Lee, and Quoc Le. Learning to Skim Text. In ACL, 2017.

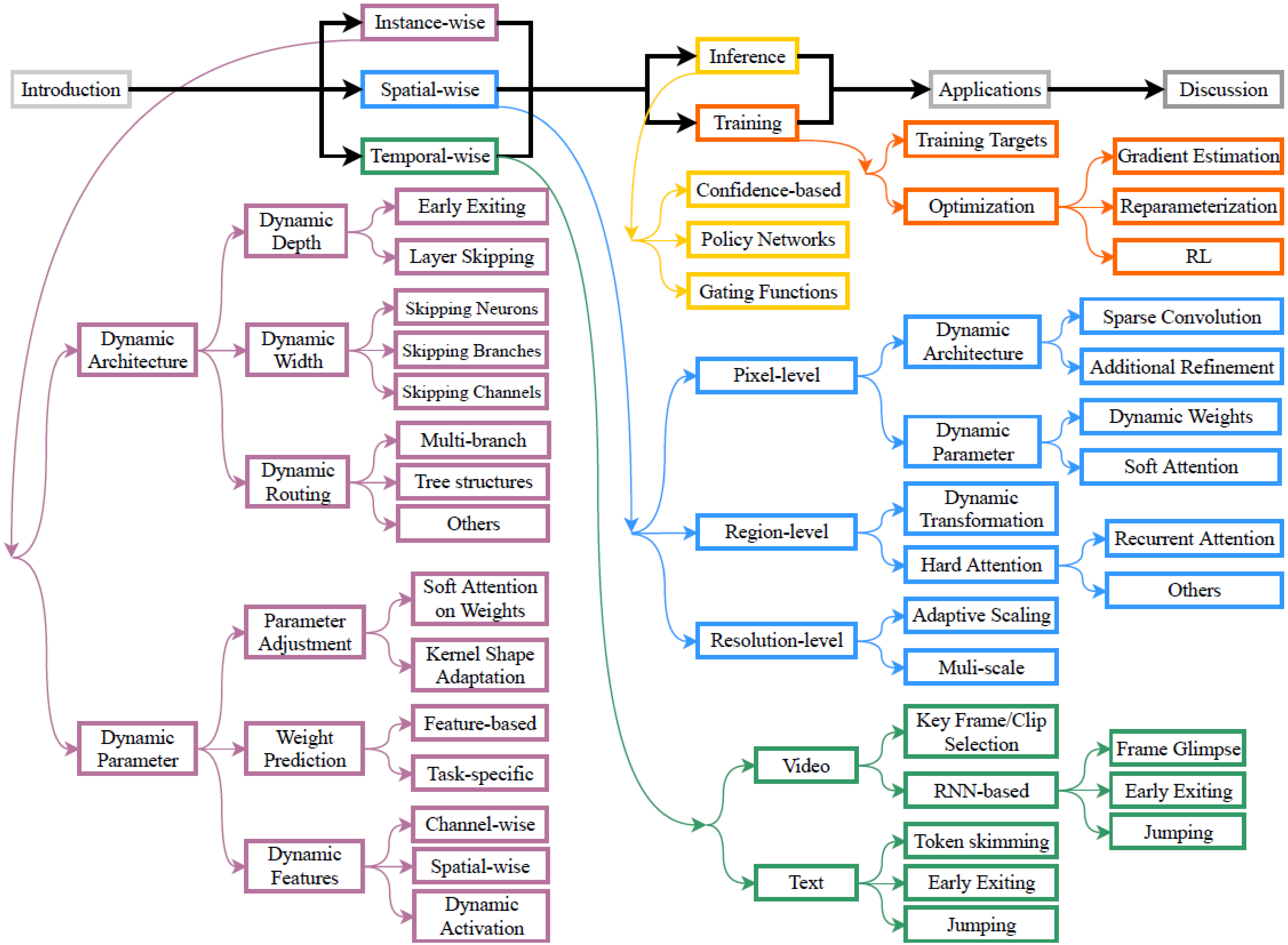
Advantages of Dynamic Neural Networks

Efficiency	Representation Power	Adaptiveness
Compatibility	Generality	Interpretability

Challenges in Dynamic Neural Networks

Theories	Architecture Design	Applicability on more diverse tasks
Gap between theoretical & practical efficiency	Robustness	Interpretability

Hierarchy of dynamic networks



arXiv Paper

Han & Huang, et al, Dynamic neural networks: A survey, arXiv:2102.04906

Thank you!

清華大學
Tsinghua University