THE UNIVERSITY OF SYDNEY

快速有效的NAS和基于NAS启发的模型压缩

欧阳万里

大纲

- 简介
- 快速有效的NAS
- 基于NAS启发的模型压缩
- 结论

Deep learning vs non-deep learning

- Automatically learn features from data → Achieved by deep learning

Deep learning – not fully automatic

- Automatically learn features from data → Achieved by deep learning
- Number of layers?
- Number of channels at each layer?
- What kind of operation in each layer?
- How one layer is connected to another layer?
- Data preparation?
- Objective/Loss function?
- …

Manual tuning is required

Automatically learning them is possible by AutoML

AutoML

- The problem of automatically (without human input) producing test set predictions for a new dataset within a fixed **computational budget** [a].
- Target: low error rate with low computational budget （高精度+高效率）

[a] Feurer, Matthias, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. "Efficient and robust automated machine learning." In *Advances in neural information processing systems*, pp. 2962-2970. 2015.

AutoML – Our works

- NAS:
  - Dongzhan Zhou*, Xinchi Zhou*, Wenwei Zhang, Chen Change Loy, Shuai YI, Xuesen Zhang, W. Ouyang, "EcoNAS: Finding Proxies for Economical Neural Architecture Search", CVPR, 2020.
  - Xiang Li, Chen Lin, Chuming Li, Ming Sun, Wei Wu, Junjie Yan, W. Ouyang, "Improving One-shot NAS by Suppressing the Posterior Fading", CVPR, 2020.
  - Liang F, Lin C, Guo R, Sun M, Wu W, Yan J, Ouyang W. “Computation Reallocation for Object Detection", ICLR, 2020.
- Data Augmentation:
  - Chen Lin, Minghao Guo, Chuming Li, Xin Yuan, Wei Wu, Junjie Yan, Dahua Lin, W. Ouyang. "Online Hyper-parameter Learning for Auto-Augmentation Strategy", Proc. ICCV, 2019.
- Loss:
  - Chuming Li, Xin Yuan, Chen Lin, Minghao Guo, Wei Wu, Junjie Yan, W. Ouyang. "AM-LFS: AutoML for Loss Function Search", Proc. ICCV, 2019.

AutoML – Our works

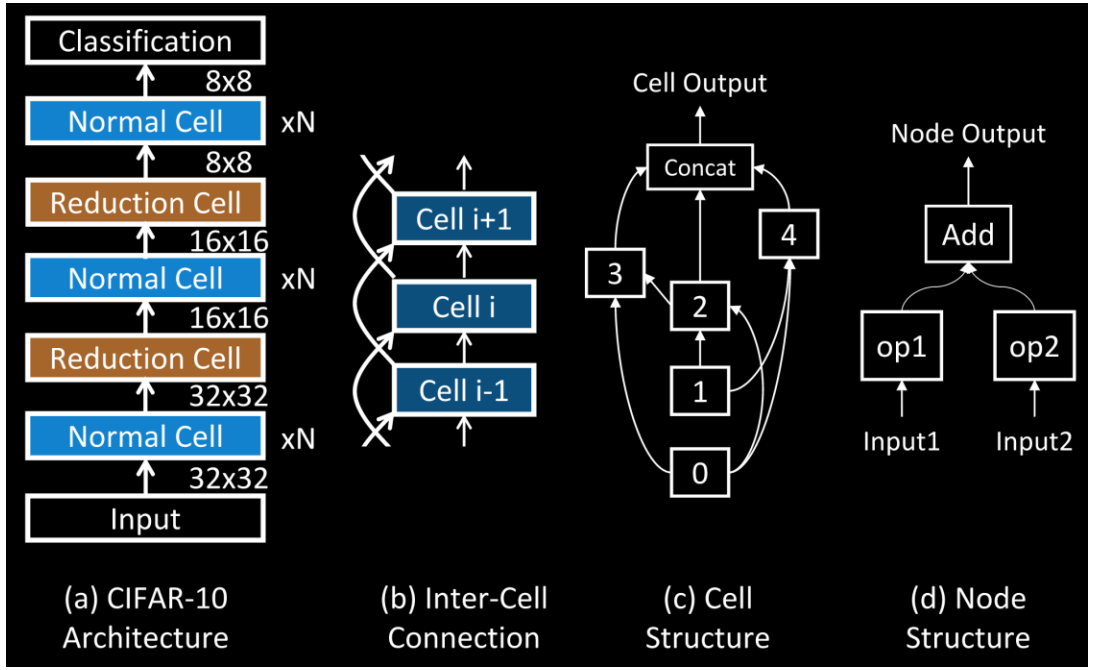
- NAS:
  - Dongzhan Zhou*, Xinchi Zhou*, Wenwei Zhang, Chen Change Loy, Shuai YI, Xuesen Zhang, W. Ouyang, "EcoNAS: Finding Proxies for Economical Neural Architecture Search", CVPR, 2020.

Network Architecture Search (NAS)

- Automatically search the suitable network architecture for specific tasks
- Time consuming

Search Space

Network Structure (from DARTS [b])



Ops:

3x3 avg pooling	**3x3 Separable Conv**
Identity	5x5 Separable Conv
3x3 max pooling	3x3 Dilated Conv
zero	5x5 Dilated Conv

[b] Liu, H., Simonyan, K., & Yang, Y. Darts: Differentiable architecture search. *ICLR 2019*.

$24^8 = 110{,}075{,}314{,}176 \sim 1 \times 10^{11}$

Search Space

- Possible choices for 24 layers with 8 operations per layer:
  - $24^8 = 110{,}075{,}314{,}176 \sim 1 \times 10^{11}$
- Suppose each choice requires 1 hour:
  - About 12,000,000 ~ 12 million years

3x3 avg pooling	**3x3 Separable Conv**
Identity	5x5 Separable Conv
3x3 max pooling	3x3 Dilated Conv
zero	5x5 Dilated Conv

Architecture	GPU Days	Method
NASNet-A [c]	1800	Reinforcement Learning
AmoebaNet-A [d]	3150	Evolution

[c] Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: CVPR (2018)
[d] Real, Esteban, et al. "Regularized evolution for image classifier architecture search." *In: AAAI*. 2019.

提纲

- 简介
- 快速有效的NAS（高效率搜索）
- 基于NAS启发的模型压缩（高效率部署）

EcoNAS: Finding Proxies for Economical Neural Architecture Search

Dongzhan Zhou, Xinchi Zhou, Wenwei Zhang, Chen Change Loy,
Shuai Yi, Xuesen Zhang, Wanli Ouyang

CVPR 2020

Motivation

- Too time consuming

Architecture	GPU Days	Method
NASNet-A [b]	1800	Reinforcement Learning
AmoebaNet-A [c]	3150	Evolution

[b] Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: CVPR (2018)
[c] Real, Esteban, et al. "Regularized evolution for image classifier architecture search." *In: AAAI*. 2019.

Proxy

3x3 avg pooling	**3x3 Separable Conv**
Identity	5x5 Separable Conv
3x3 max pooling	3x3 Dilated Conv
zero	5x5 Dilated Conv

- A proxy is a computationally reduced setting, e.g.
  - Reduced number of training epochs

Computation	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
Training Epochs (***e***)	600	300	150	75

- Compared with the original network, the proxy has the same
  - Operation
  - Number of layers
  - Relative ratio for the numbers of channels between two layers

Dongzhan Zhou, Xinchi Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, and Wanli Ouyang. "EcoNAS: Finding Proxies for Economical Neural Architecture Search." *CVPR* 2020.

Proxy

3x3 avg pooling	**3x3 Separable Conv**
Identity	5x5 Separable Conv
3x3 max pooling	3x3 Dilated Conv
zero	5x5 Dilated Conv

- A proxy is a computationally reduced setting, e.g.
  - Reduced number of training epochs
  - Reduced input resolution
  - Reduced number of channels
  - Reduced number of samples
- Compared with the original network, the proxy has the same
  - Operation
  - Number of layers
  - Relative ratio for the numbers of channels between two layers

Dongzhan Zhou, Xinchi Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, and Wanli Ouyang. "EcoNAS: Finding Proxies for Economical Neural Architecture Search." *CVPR* 2020.

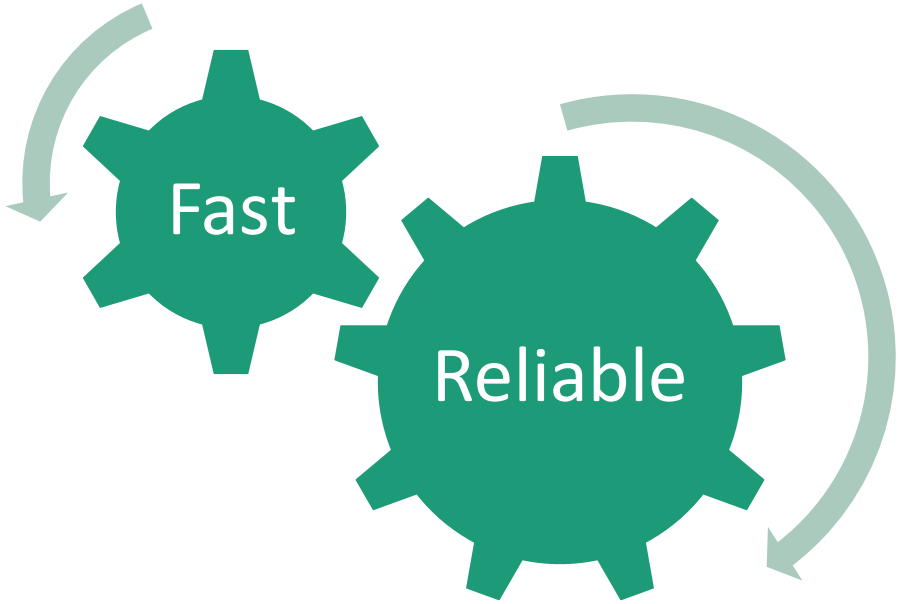
Proxy

3x3 avg pooling	**3x3 Separable Conv**
Identity	5x5 Separable Conv
3x3 max pooling	3x3 Dilated Conv
zero	5x5 Dilated Conv

- A proxy is a computationally reduced setting, e.g.
  - Reduced number of training epochs [19]
  - Reduced input resolution
  - Reduced number of channels [23]
  - Reduced number of samples [17, 19 31]
- Compared with the original network, the proxy has the same
  - Operation
  - Number of layers
  - Relative ratio for the numbers of channels between two layers

[7] Boyang Deng, Junjie Yan, and Dahua Lin. Peephole: redicting network performance before training. CoRR, abs/1712.03351, 2017.
[17] Dmytro Mishkin, Nikolay Sergievskiy, and Jiri Matasa. Systematic evaluation of cnn advances on the imagenet. CVIU, 2017.
[23] Kailas Vodrahalli, Ke Li, and Jitendra Malik. Are all training examples created equal? an empirical study. CoRR, abs/1811.12569, 2018
[19] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Regularized evolution for image classifier architecture search. In AAAI, 2019.
[31] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In CVPR, 2018.

What is a good proxy?



This paper

- A systematic and empirical study on the proxy
- Appropriate use of proxy can
  - Make NAS fast
  - Get architectures with better accuracy

Proxy – reliability

Existing proxies behave differently in maintaining rank consistency.

Example:

	Real Ranking	**Ranking in Proxy 1**	**Ranking in Proxy 2**
Network A	1	1	3
Network B	2	2	4
Network C	3	3	1
Network D	4	4	2
		Good Proxy	Bad Proxy

Finding reliable proxies is important for Neural Architecture Search.

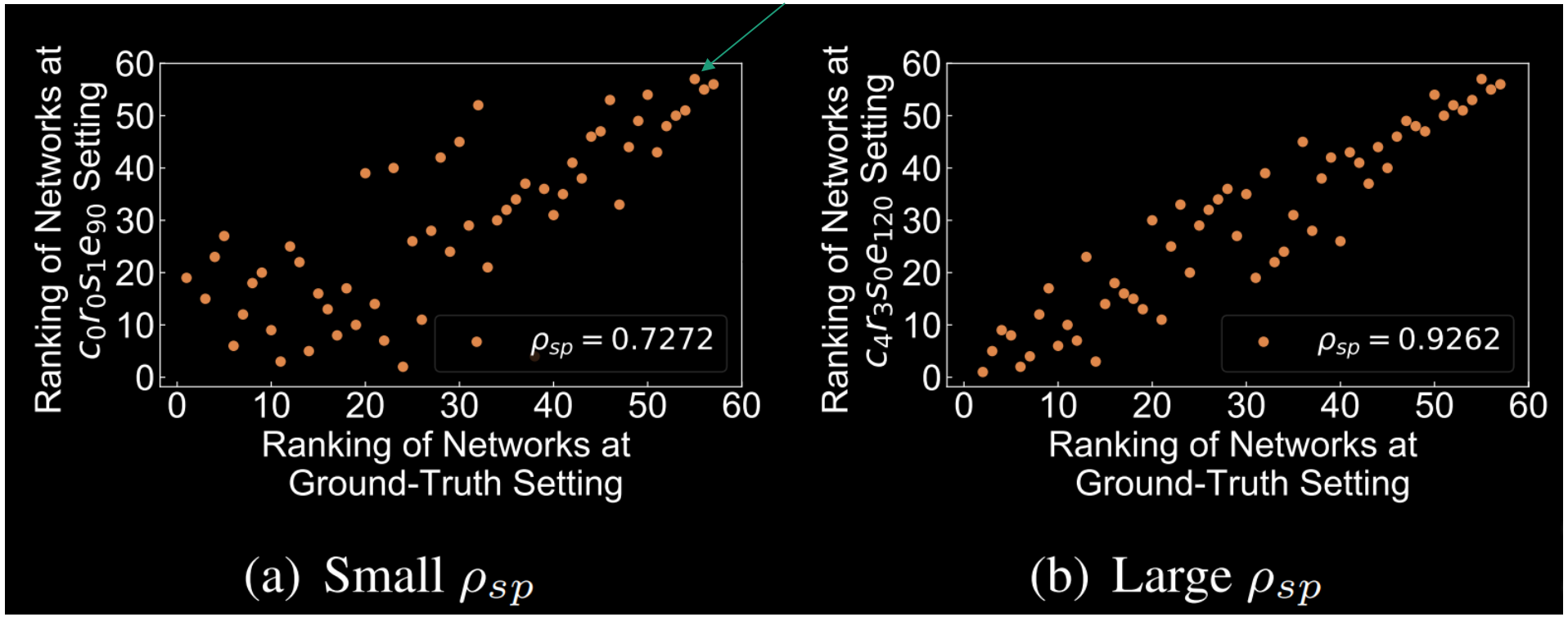
Dongzhan Zhou, Xinchi Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, and Wanli Ouyang. "EcoNAS: Finding Proxies for Economical Neural Architecture Search." *CVPR* 2020.

How to evaluate the reliability of Proxies?

Spearman Coefficient of original ranking (Ground-Truth Setting) and proxy ranking (reduced setting).

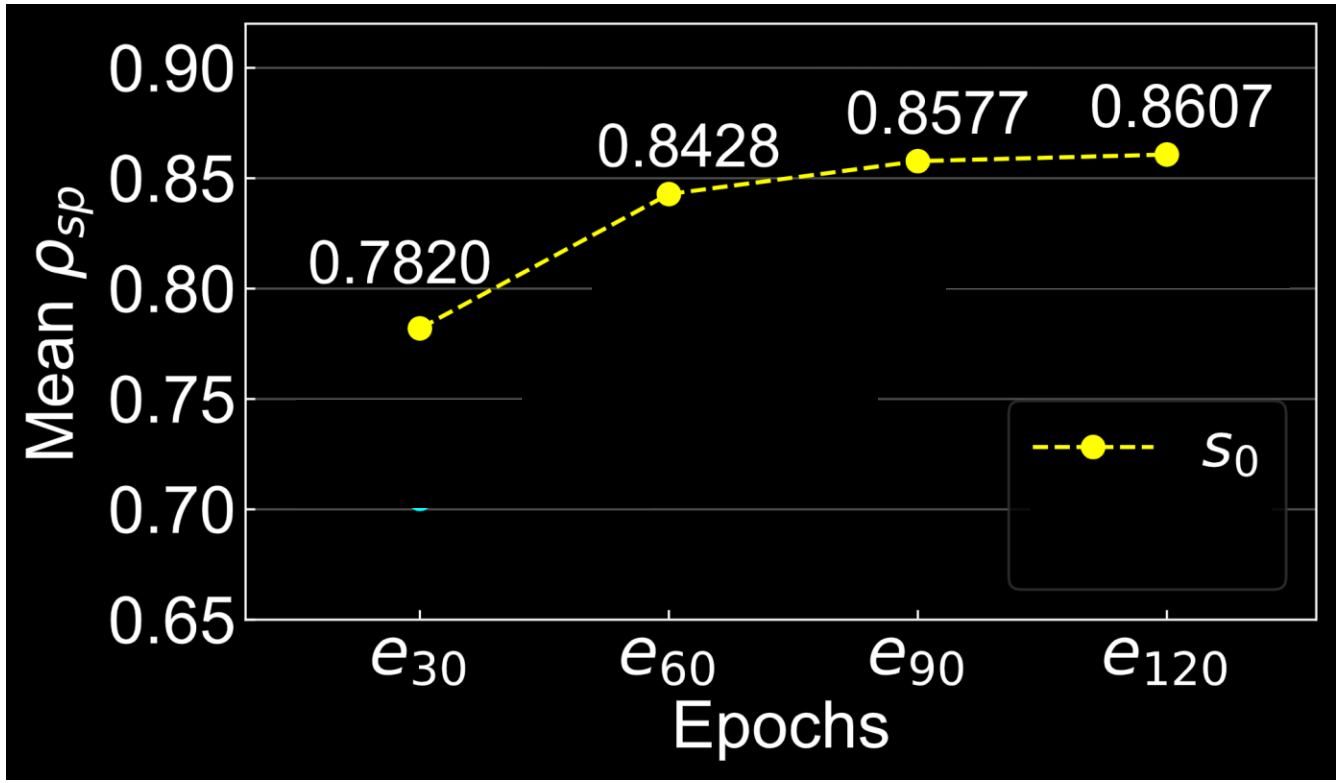
- Value range [-1, 1], higher absolute value indicates stronger correlation.
- Positive value for positive correlation, vice versa.

A model sampled from the search space



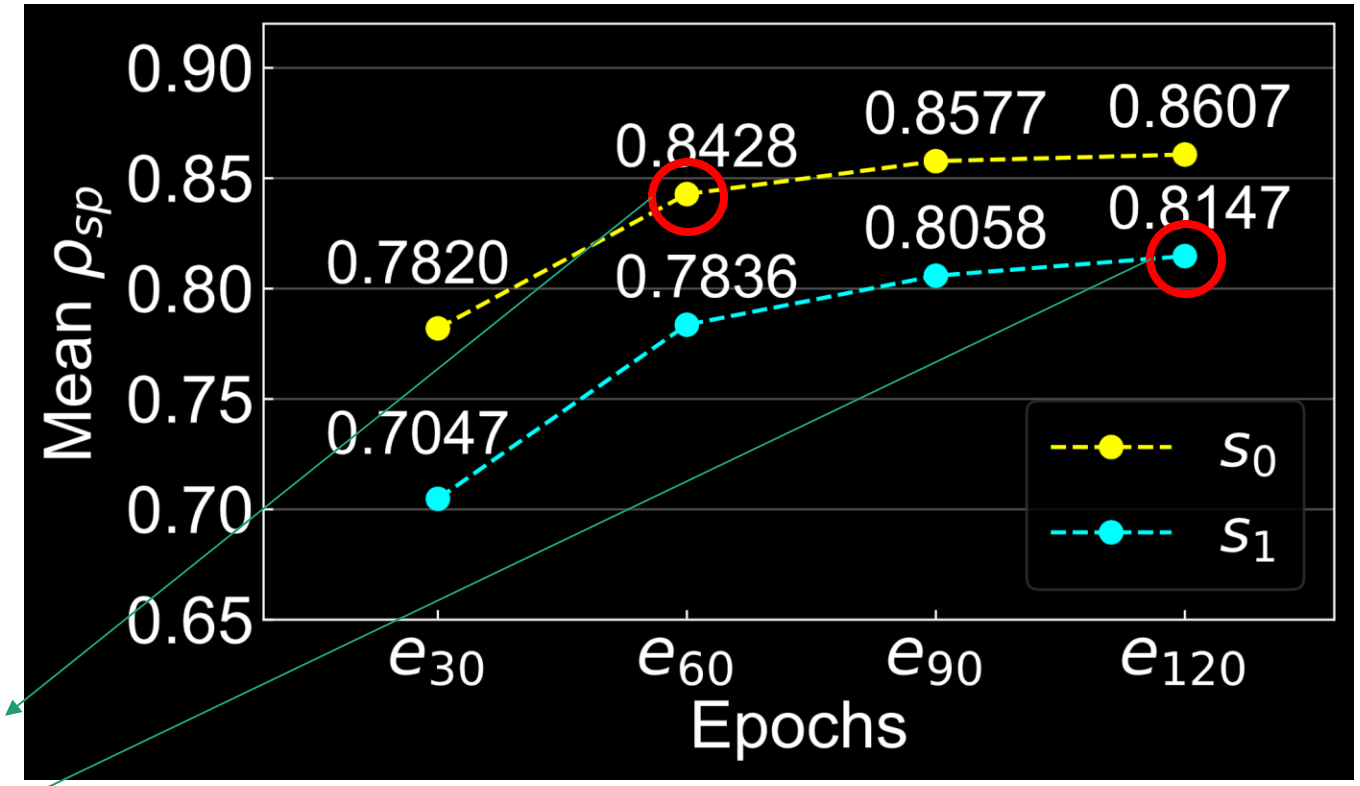
Influence of sample ratio (***s***) and epochs (***e***)

With the same iteration numbers, using more training samples with fewer training epochs could be more effective than using more training epochs and fewer training samples.



Influence of sample ratio (***s***) and epochs (***e***)

With the same iteration numbers, using more training samples with fewer training epochs could be more effective than using more training epochs and fewer training samples.



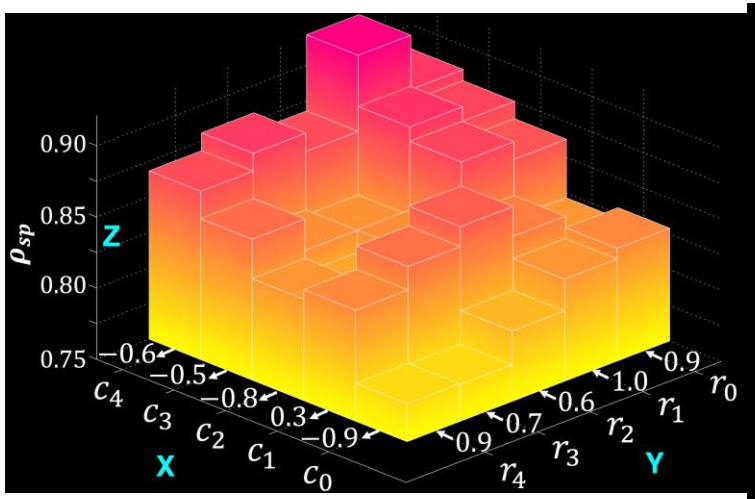
60 epochs, 100 iters per epoch

120 epochs, 50 iters per epoch

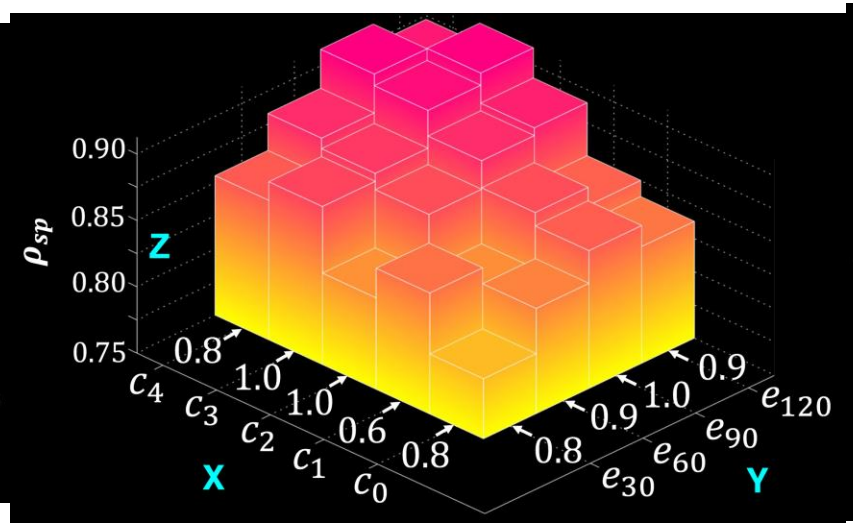
Influence of channels (***c***) and resolution (***r***)

Reducing the resolution of input images is sometimes feasible

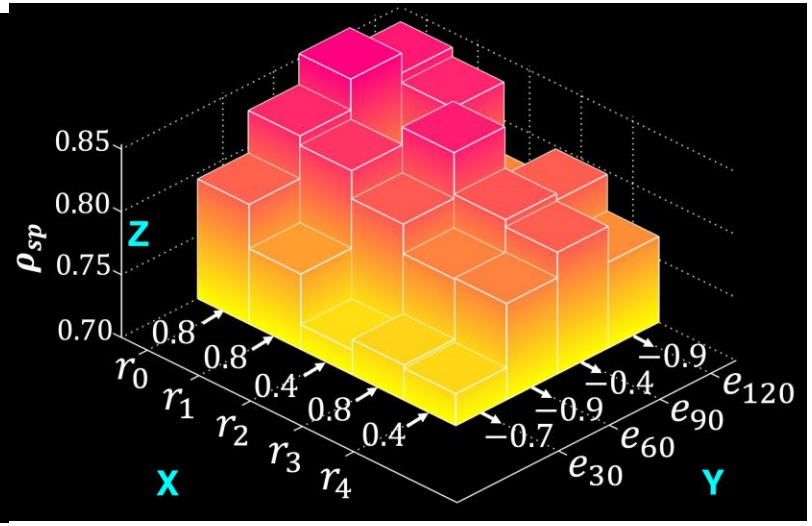
Reducing the number of channels of networks is more reliable than reducing the resolution.



$c_x r_y s_0 e_{60}$



$c_x r_0 s_0 e_y$

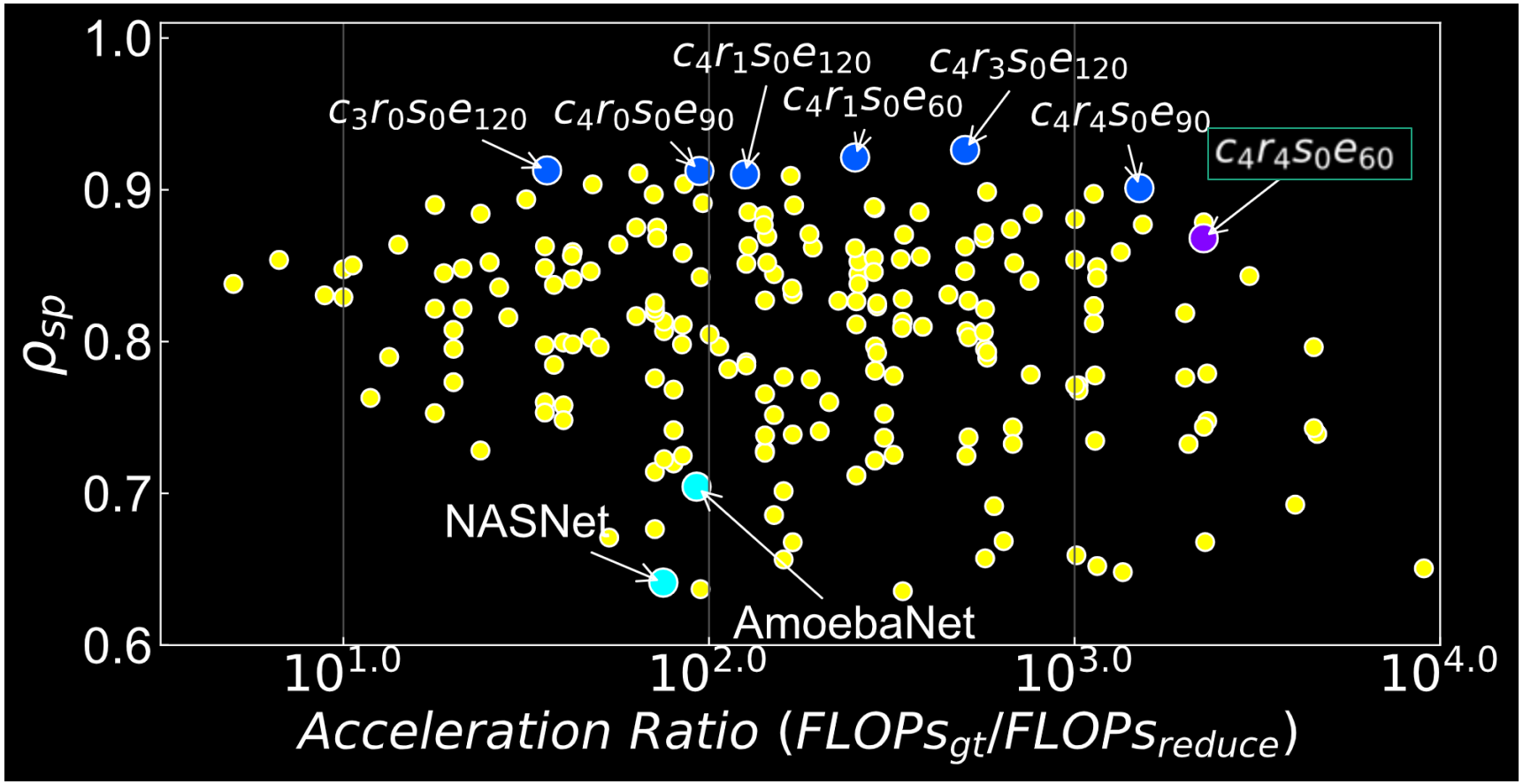


$c_0 r_x s_0 e_y$

Dongzhan Zhou, Xinchi Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, and Wanli Ouyang. "EcoNAS: Finding Proxies for Economical Neural Architecture Search." *CVPR* 2020.

Efficient proxies

An efficient proxy does not necessarily have a poor rank consistency.



Dongzhan Zhou, Xinchi Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, and Wanli Ouyang. "EcoNAS: Finding Proxies for Economical Neural Architecture Search." *CVPR* 2020.

Proxy – Example

Use reduction factors for training

$$\frac{1}{16} \times \frac{1}{16} \times \frac{1}{1} \times \frac{1}{10} = \frac{1}{2560}$$

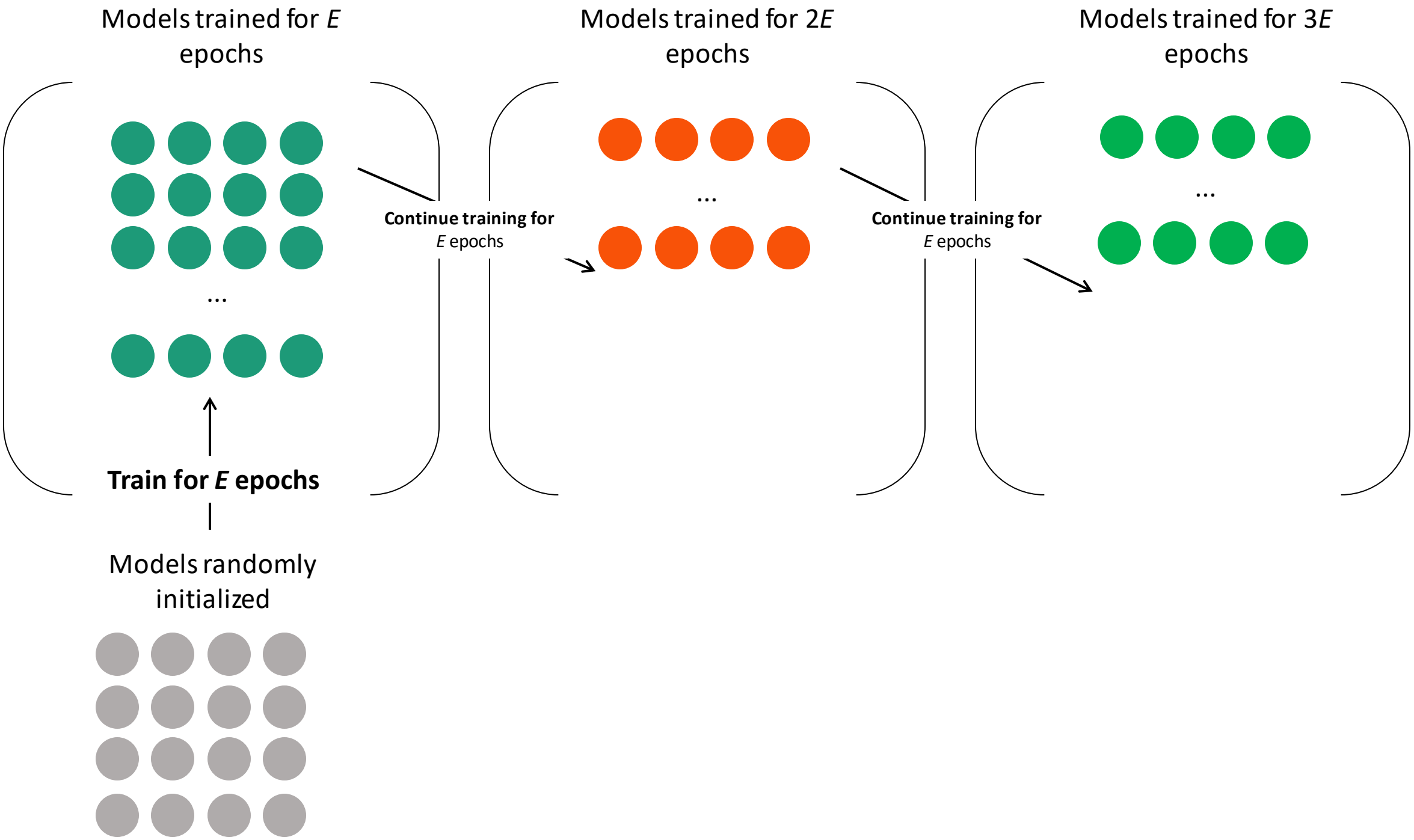
$c_4 r_4 s_0 e_{60}$

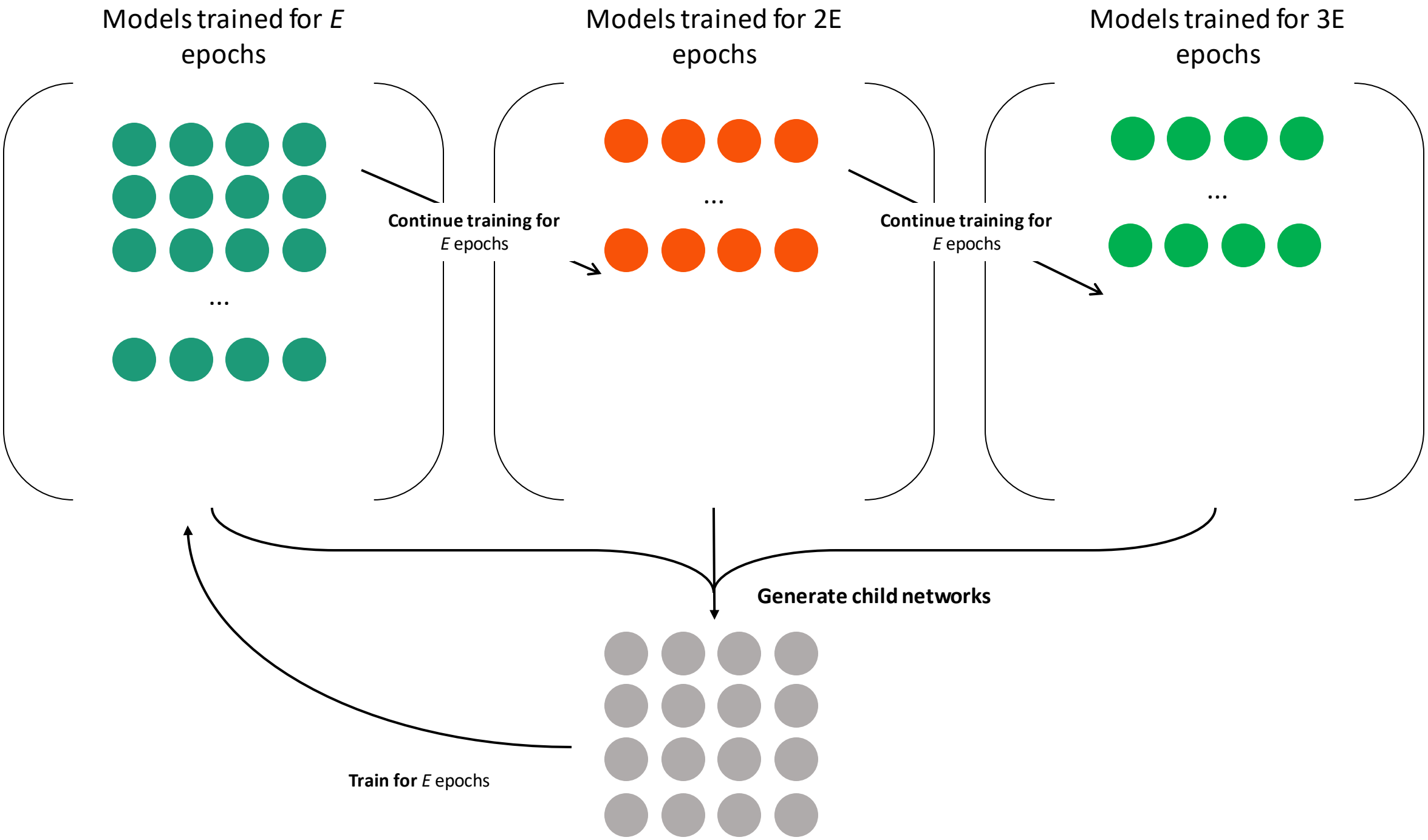
Original setting:		Proxy:
Conv(l): 36 × c channels	c_4	9 × c
Conv(l+1): 36 × c channels	c_4	9 × c
Input resolution: 32 ×32	r_4	8×8
Training data: 50000	s_0	50000
Training epochs: 600	e_{60}	60

EcoNAS: Economical evolutionary-based NAS

1. Select a fast and reliable proxy.
2. Hierarchical proxy strategy: train networks with different proxies based on their accuracy.

Dongzhan Zhou, Xinchi Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, and Wanli Ouyang. "EcoNAS: Finding Proxies for Economical Neural Architecture Search." *CVPR* 2020.


Models trained for *E* epochs
Models trained for 2*E* epochs
Models trained for 3*E* epochs
Continue training for *E* epochs
Continue training for *E* epochs
...
...
...
Train for *E* epochs
Models randomly initialized


Models trained for *E* epochs
Models trained for 2E epochs
Models trained for 3E epochs
...
...
...
Continue training for *E* epochs
Continue training for *E* epochs
Generate child networks
Train for *E* epochs

EcoNAS: Economical evolutionary-based NAS

1. Select a more efficient and consistent reduced setting as proxy.
2. Hierarchical proxy strategy: train networks with different proxies based on their accuracy.

Setting: Three population sets P_E, P_{2E}, P_{3E}, which store networks trained for E, 2E, 3E epochs, respectively.

For each cycle:

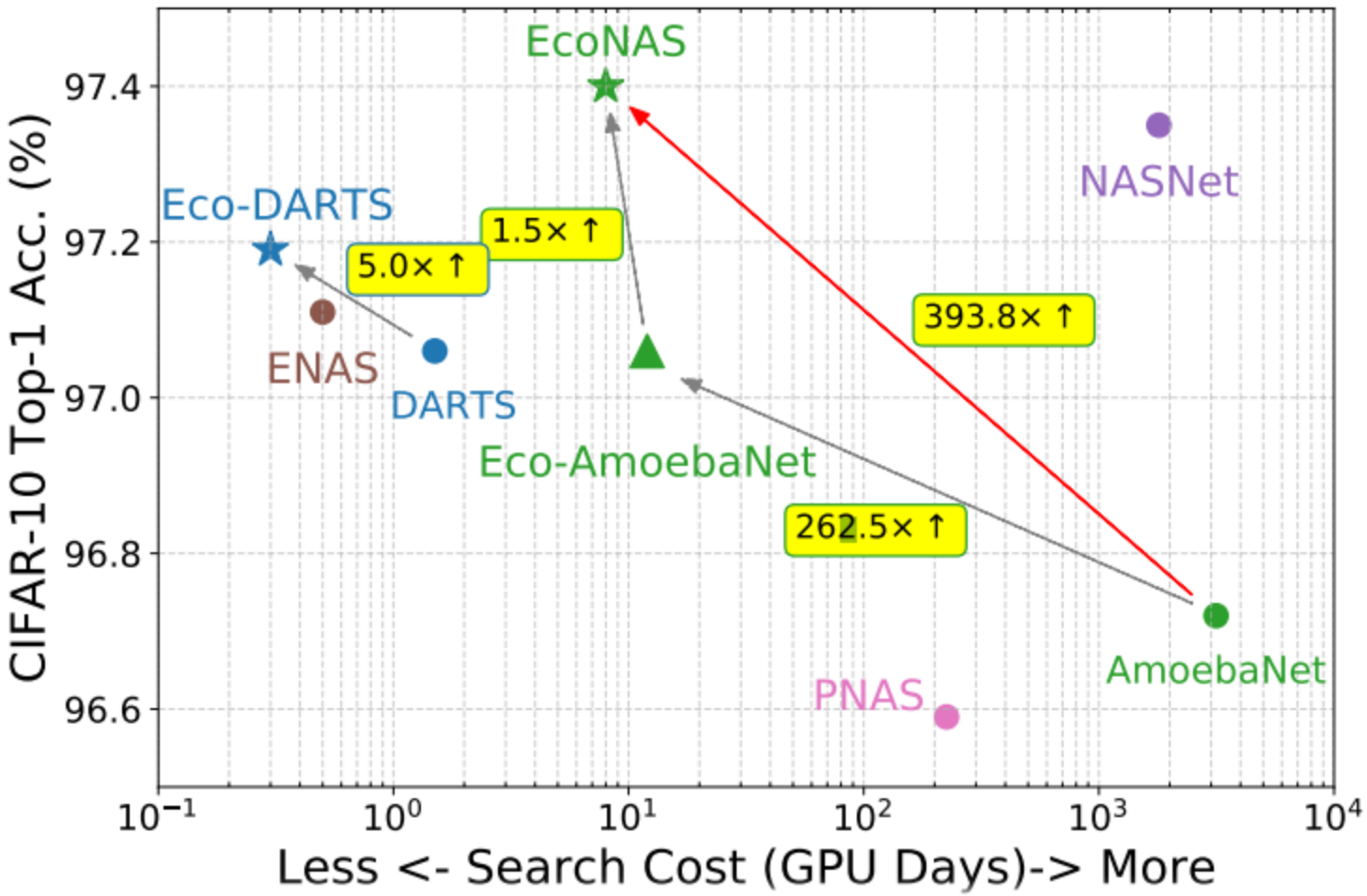
Step 1. A batch of networks are randomly sampled from P_E, P_{2E}, P_{3E} and mutated. Networks with higher accuracy are more likely to be chosen. Train the mutated networks for E epochs and add them to P_E.

Step 2. Choose top networks from P_E, P_{2E}, load from checkpoints and train E more epochs, then add to P_{2E}, P_{3E}.

Step 3. Remove dead networks from all populations.

Zhou, Dongzhan, Xinchi Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, and Wanli Ouyang. "EcoNAS: Finding Proxies for Economical Neural Architecture Search." *CVPR* 2020.

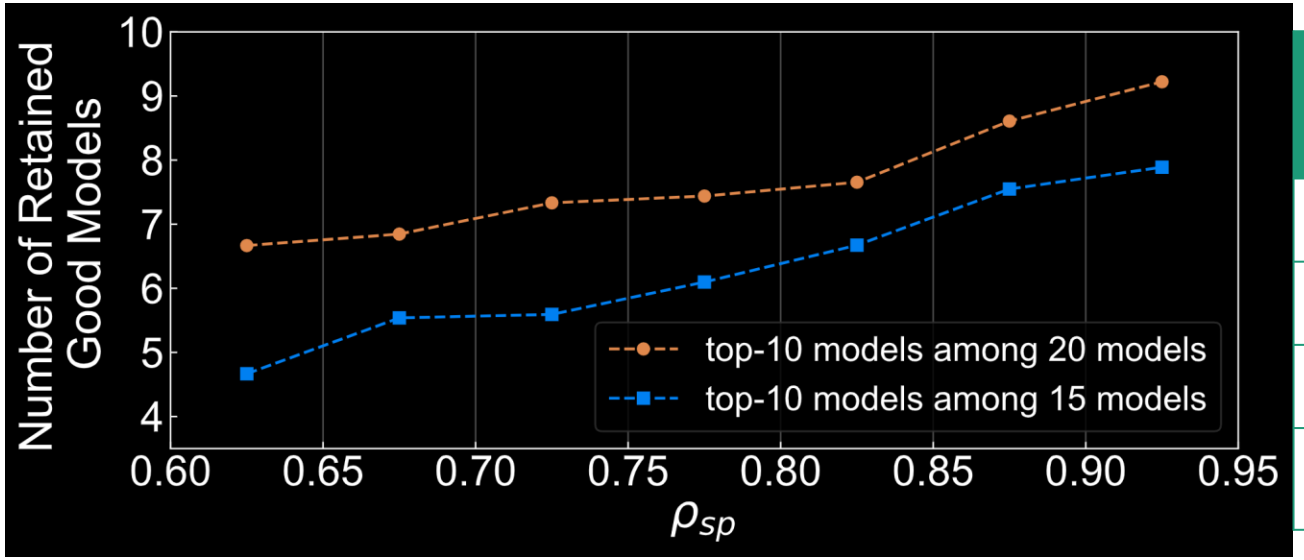
Experimental results on CIFAR-10



EcoNAS analysis

Not only save searching costs.

But also save re-training costs. Reliable proxies need not many networks to be re-trained.

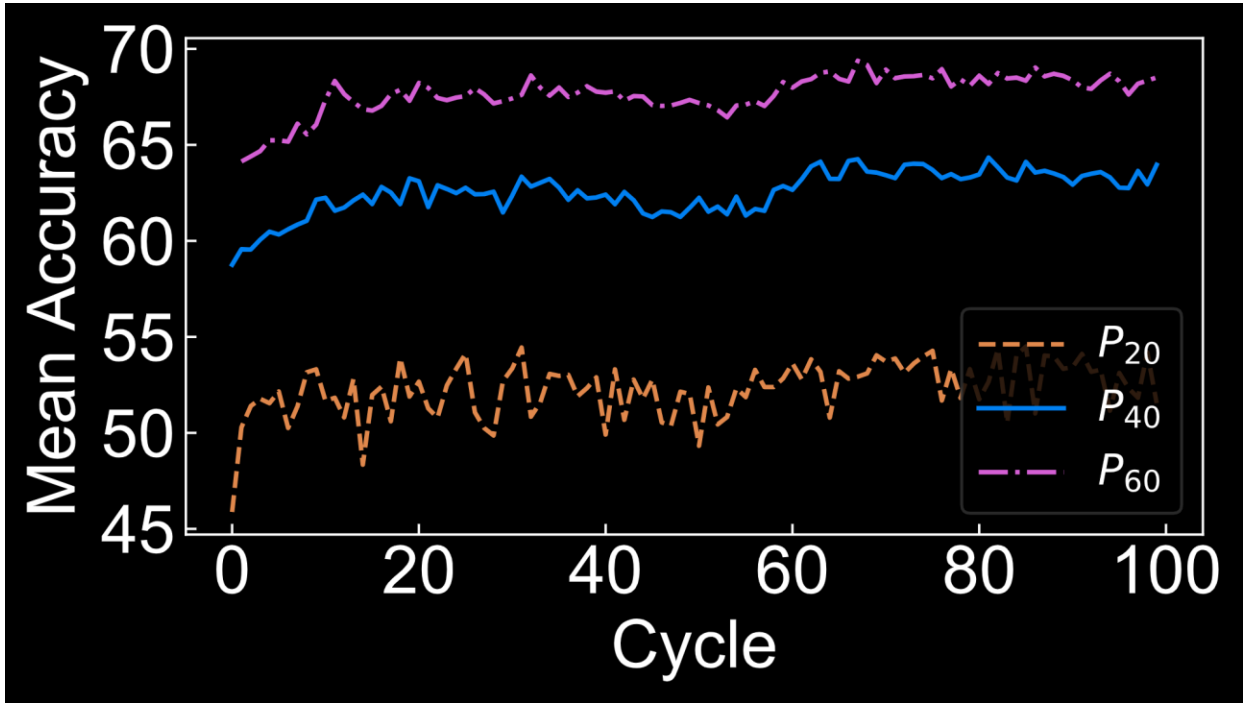


Method	Number of Re-training Networks
BlockQNN	100
NASNet	250
AmoebaNet	20
EcoNAS (ours)	5

Zhou, Dongzhan, Xinchi Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, and Wanli Ouyang. "EcoNAS: Finding Proxies for Economical Neural Architecture Search." *CVPR* 2020.

EcoNAS analysis

Provides more diverse structures, which allows searching algorithms to find accurate structures with fewer costs.



Method	Network numbers
BlockQNN	11k
NASNet	45k
AmoebaNet	20k
EcoNAS (ours)	1k

Zhou, Dongzhan, Xinchi Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, and Wanli Ouyang. "EcoNAS: Finding Proxies for Economical Neural Architecture Search." *CVPR* 2020.

EcoNAS ablation study on CIFAR-10

1. Reliable proxy and hierarchical proxy strategy will reduce both searching cost and error rate.

Reduced Setting (w/o hierarchical proxy)	Cost (GPU days)	Spearman Coefficient	Params. (M)	Error Rate (%)
AmoebaNet	3150	0.70	3.20	3.34 ± 0.06
$C_4r_4s_0e_{35}$ (ours)	12	0.74	3.18	2.94

Reduced Setting (w. hierarchical proxy)	Cost (GPU days)	Spearman Coefficient	Params. (M)	Error Rate (%)
NASNet Proxy	21	0.65	2.89	3.20
$C_3r_2s_1e_{60}$	12	0.79	2.56	2.85
$C_4r_4s_0e_{60}$ (ours)	8	0.85	3.40	2.60

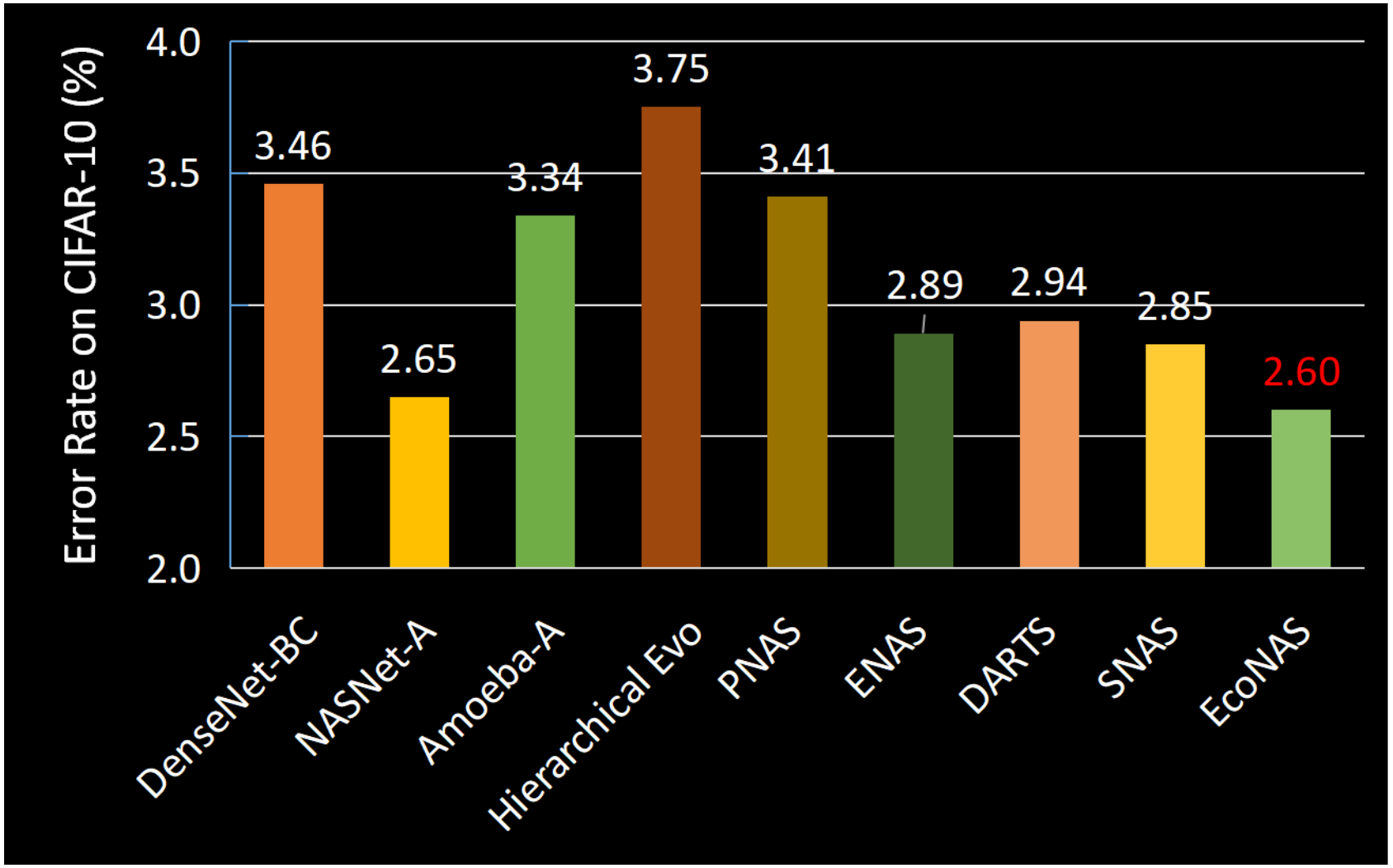
EcoNAS ablation study

2. Reliable proxy settings can be adopted in other NAS methods.

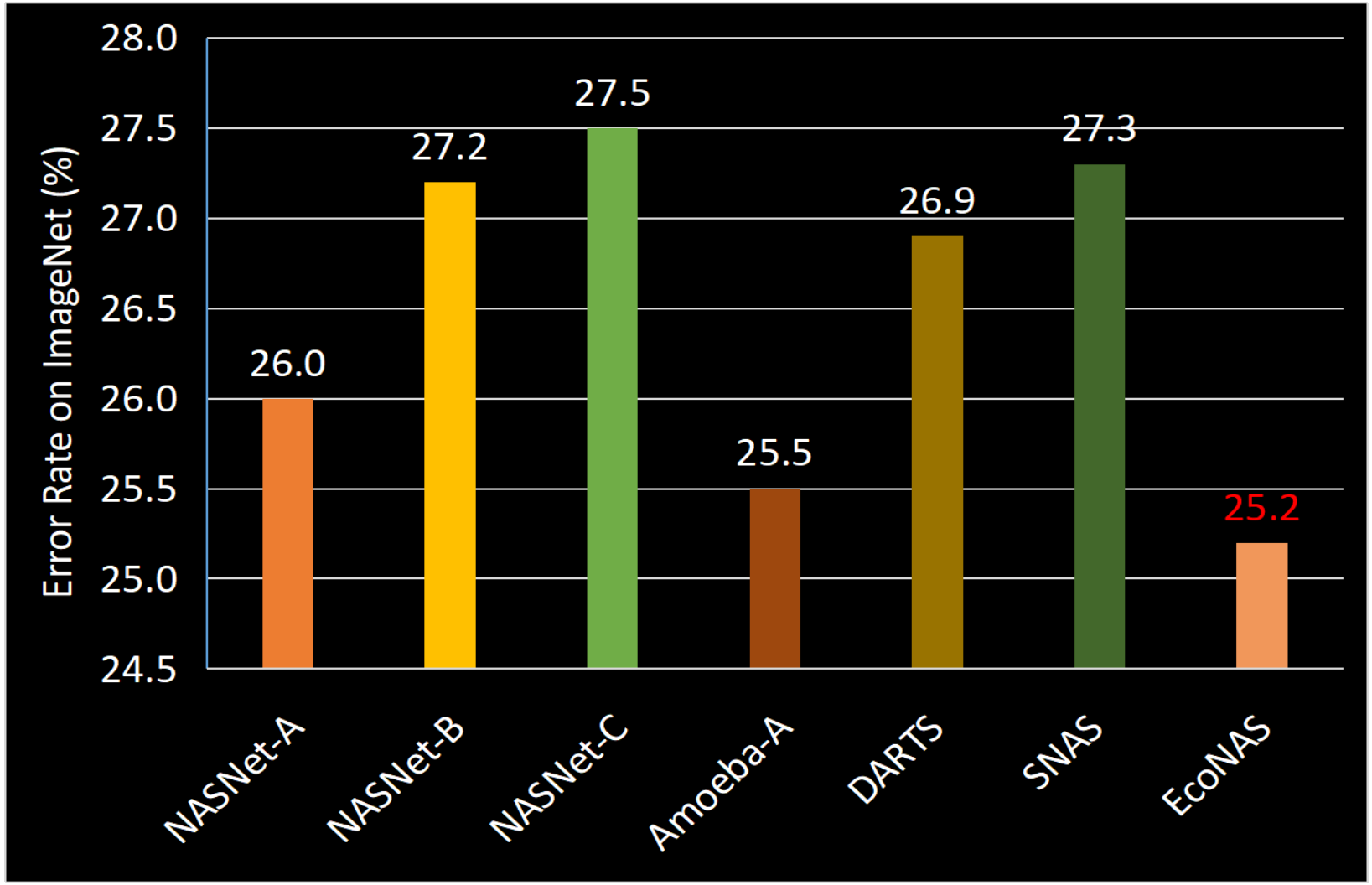
Method	Setup	Cost (GPU days)	Params. (M)	Error Rate (%)
DARTS (on CIFAR-10)	$c_2r_0s_0$	1.5	3.2	3.0
	$c_4r_2s_0$ (ours)	0.3	4.5	2.8
ProxylessNAS (on ImageNet)	$c_0r_0s_0$-S	8	4.1	25.4
	$c_0r_0s_0$-L	8	6.9	23.3
	$c_2r_2s_0$ (ours)	4	5.3	23.2

Zhou, Dongzhan, Xinchi Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, and Wanli Ouyang. "EcoNAS: Finding Proxies for Economical Neural Architecture Search." *CVPR* 2020.

EcoNAS results on CIFAR-10



EcoNAS results on ImageNet



提纲

- 快速有效的NAS （高效率搜索）
- 基于NAS启发的模型压缩

提纲

- 基于NAS启发的模型压缩（高效率部署）

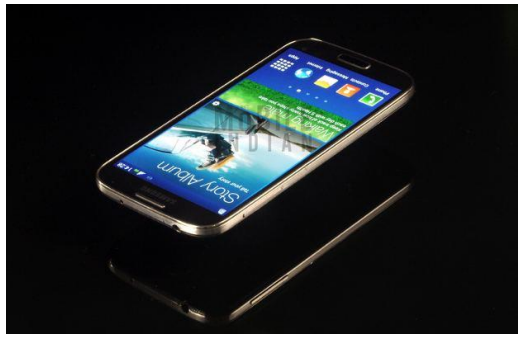
Multi-Dimensional Pruning: A Unified Framework for Model Compression

Jinyang Guo, Wanli Ouyang, Dong Xu

CVPR2020 Oral

Model Compression

Memory: 16GB / 32GB
Computation: TFLOPs/s



Memory: 8GB
Computation: GFLOPs/s

Memory: 100KB – 1MB
Computation: MFLOPs/s

Model Compression

Model Compression

Quantization

XNOR-Net, Rastegari et al., ECCV'16

HAQ, Want et al., CVPR'19

Tensor Factorization

Accelerating…, Zhang et al., T-PAMI

Compact Network Design

MobileNet, Howard et al., arXiv

ShuffleNet, Zhang et al., CVPR'18

Channel Pruning

Learning…, Liu et al., CVPR'17

Channel Pruning…, He et al., ICCV'17

Motivation

- Two redundancies are not explored:
  - Temporal-wise redundancy
  - Spatial-wise redundancy

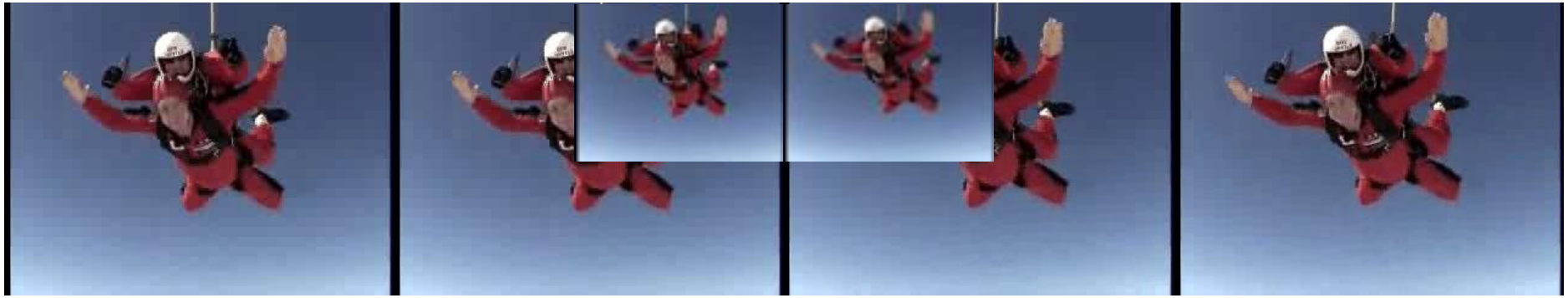
Motivation

- Two redundancies are not explored:
  - Temporal-wise redundancy

Motivation

- Two redundancies are not explored:
  - Temporal-wise redundancy
  - Spatial-wise redundancy

Downsampling



Contributions

- We proposed Multi-Dimensional Pruning (MDP):
  - Simultaneously reduce spatial/spatial-temporal and channel redundancies
  - A unified framework that can prune both 2D CNNs and 3D CNNs

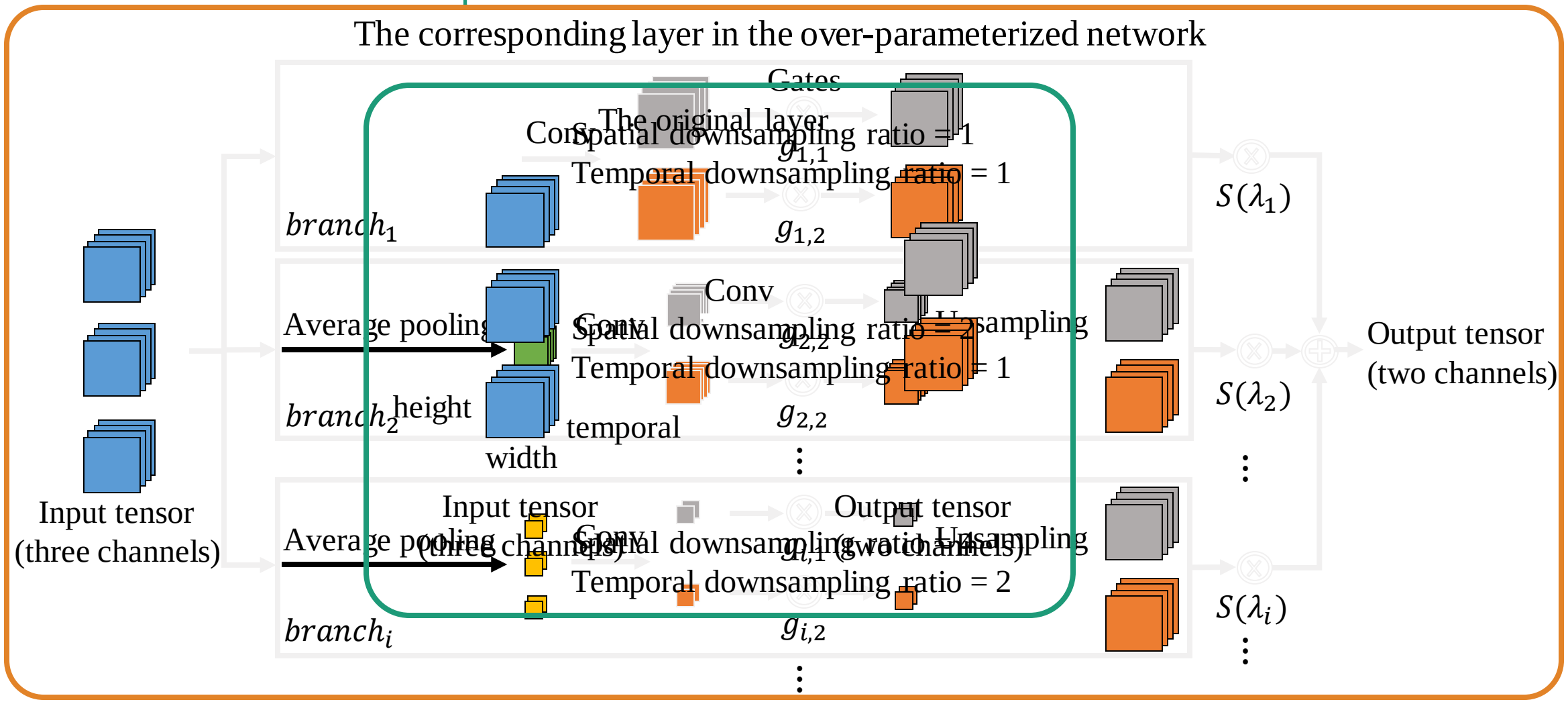
Multi-Dimensional Pruning (MDP)

The Searching Stage → The Pruning Stage → The Fine-tuning Stage

Guo, J., Ouyang, W. and Xu, D., Multi-Dimensional Pruning: A Unified Framework for Model Compression. *CVPR* 2020.

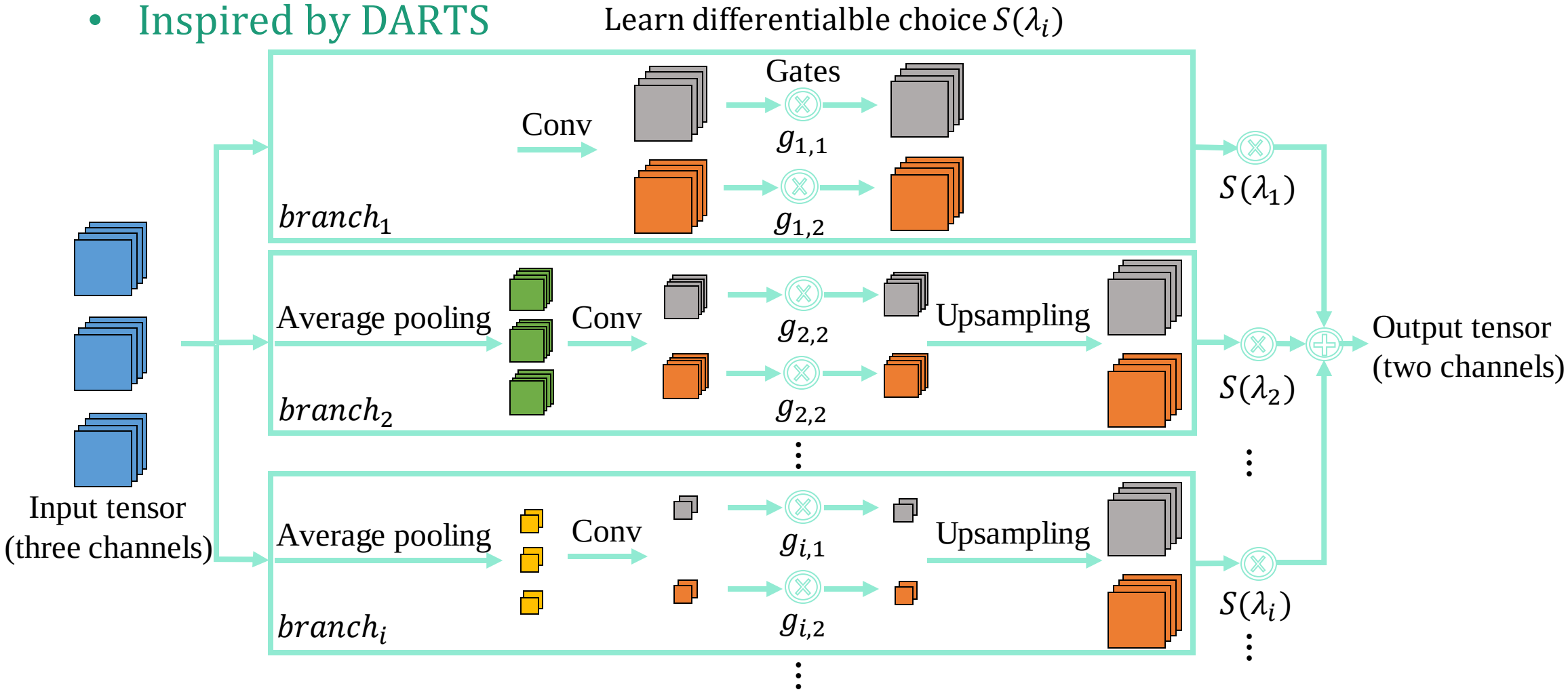
Multi-Dimensional Pruning (MDP)

- The Searching Stage
  - Construct an over-parameterized network



Multi-Dimensional Pruning (MDP)

- The Searching Stage
  - How to find the best setting of spatial-temporal resolution?
  - Inspired by DARTS



Multi-Dimensional Pruning (MDP)

- The Searching Stage
  - Training objective function

$$\arg\min_{\boldsymbol{\theta},\boldsymbol{\lambda},\mathbf{G}} L = L_c + \alpha L_{st} + \eta L_{gate}$$

L_c: Cross-entropy loss for classification task

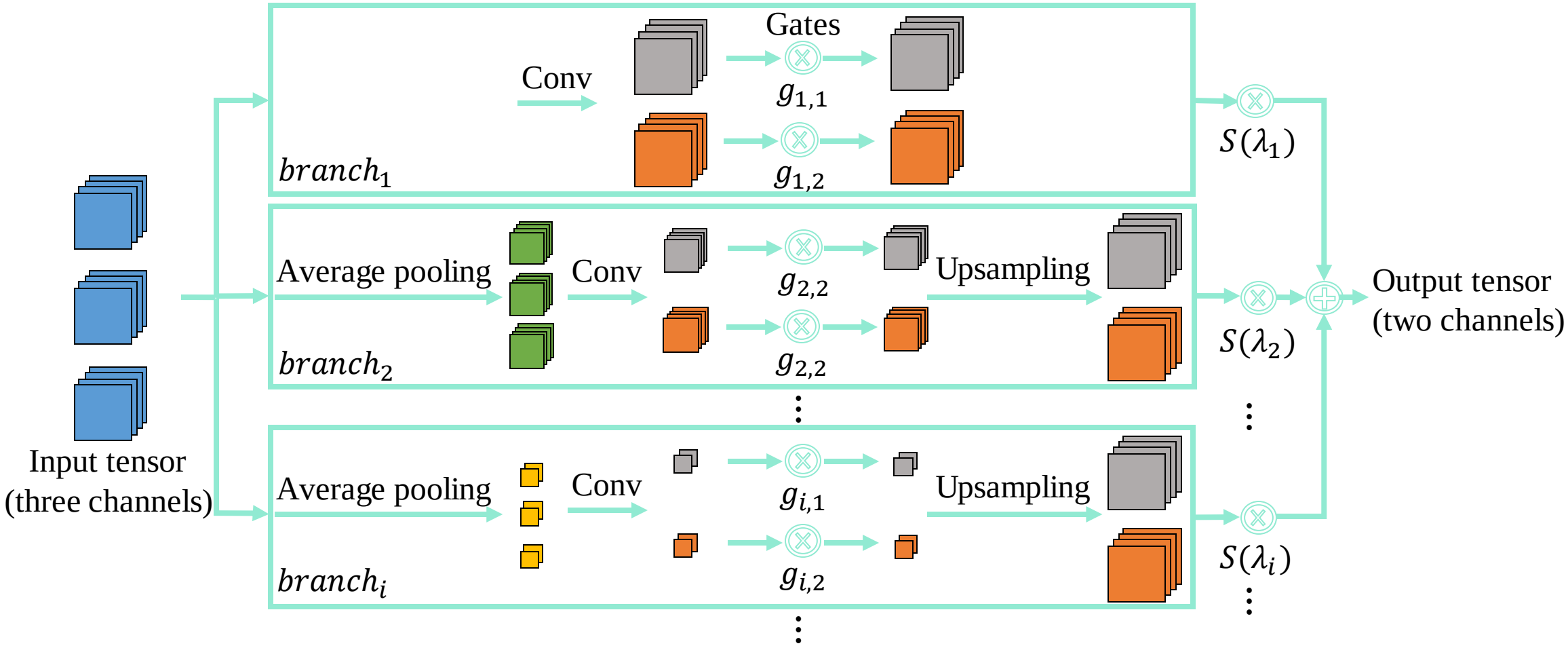
L_{st}: Penalty to introduce sparsity on branch importance for resolution selection

L_{gate}: Penalty to introduce sparsity on gates for channel pruning

Guo, J., Ouyang, W. and Xu, D., Multi-Dimensional Pruning: A Unified Framework for Model Compression. *CVPR* 2020.

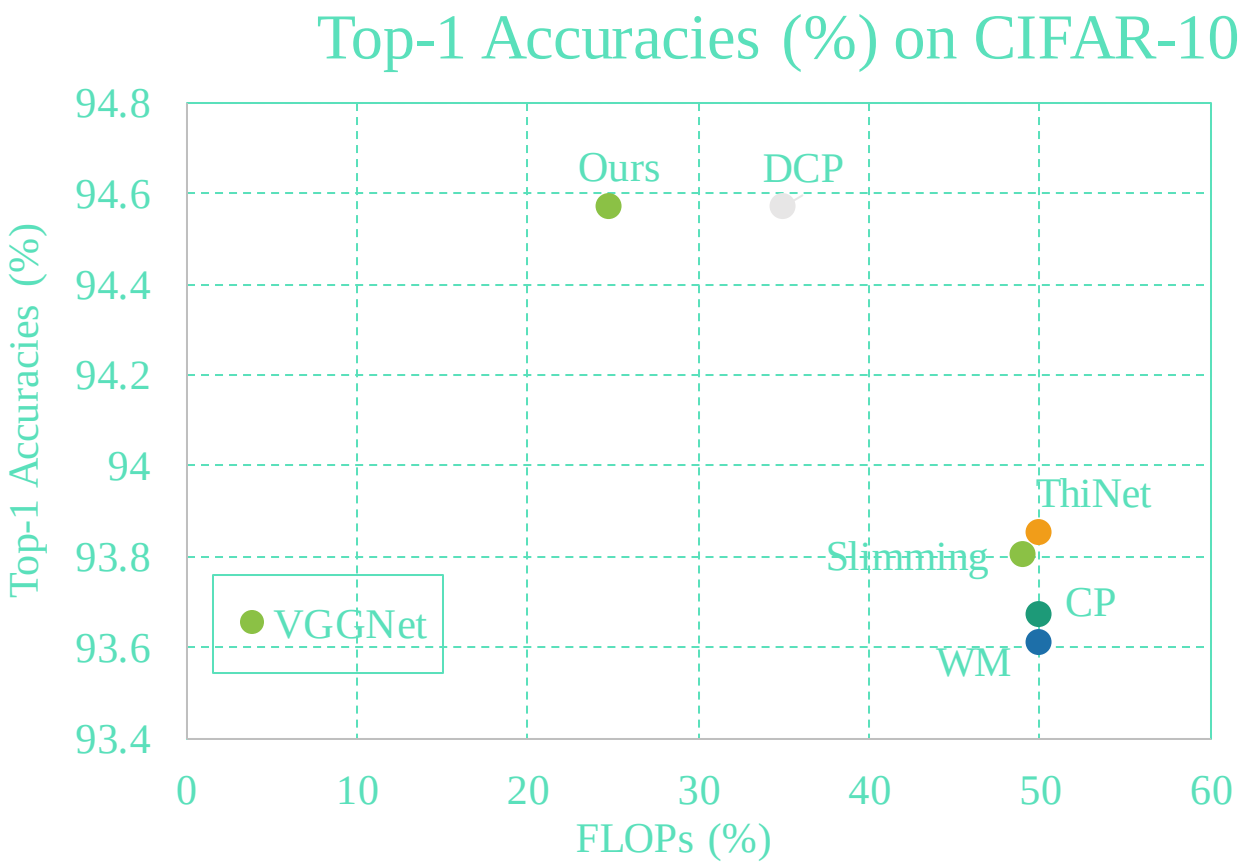
Multi-Dimensional Pruning (MDP)

- The Pruning Stage
  - Select branch with the largest branch importance score
  - Prune channels with small gate values



Experiments

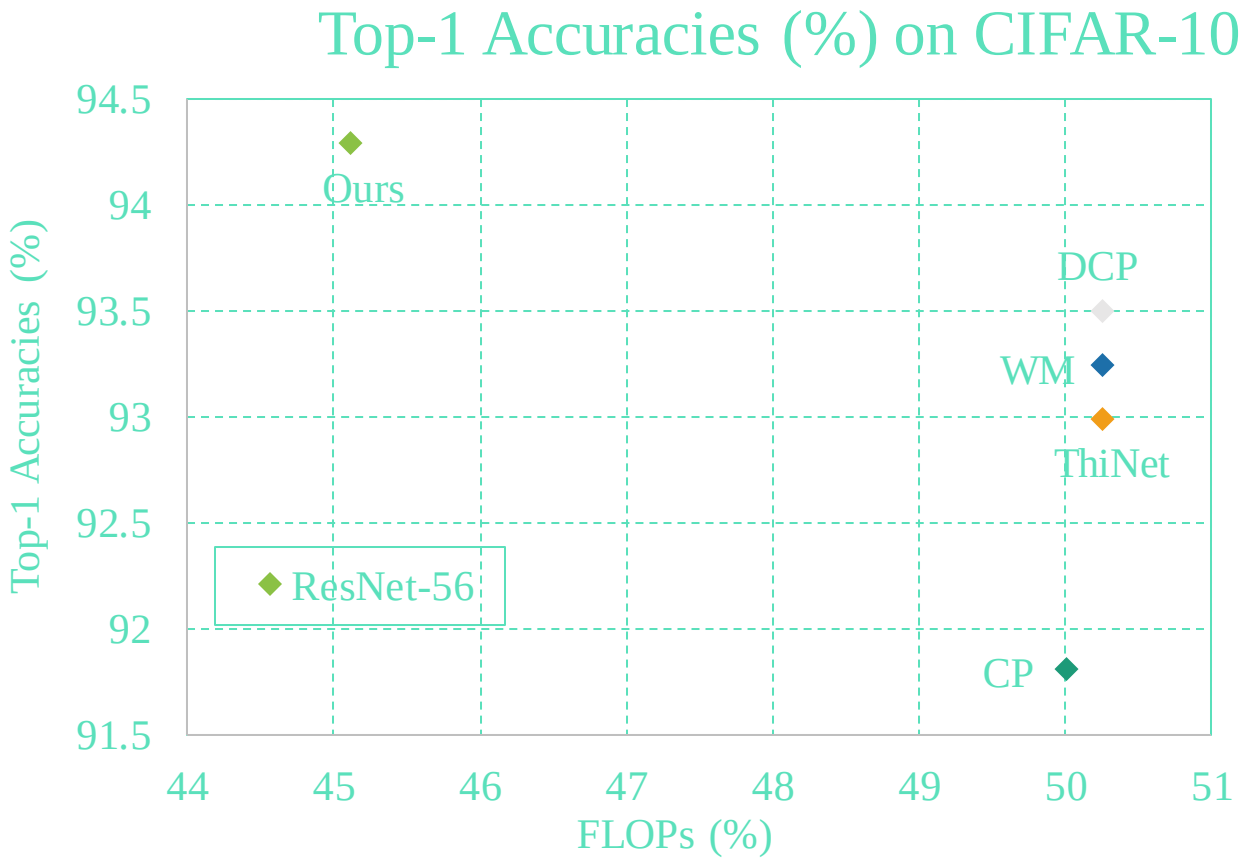
- Image Classification (2D CNNs)

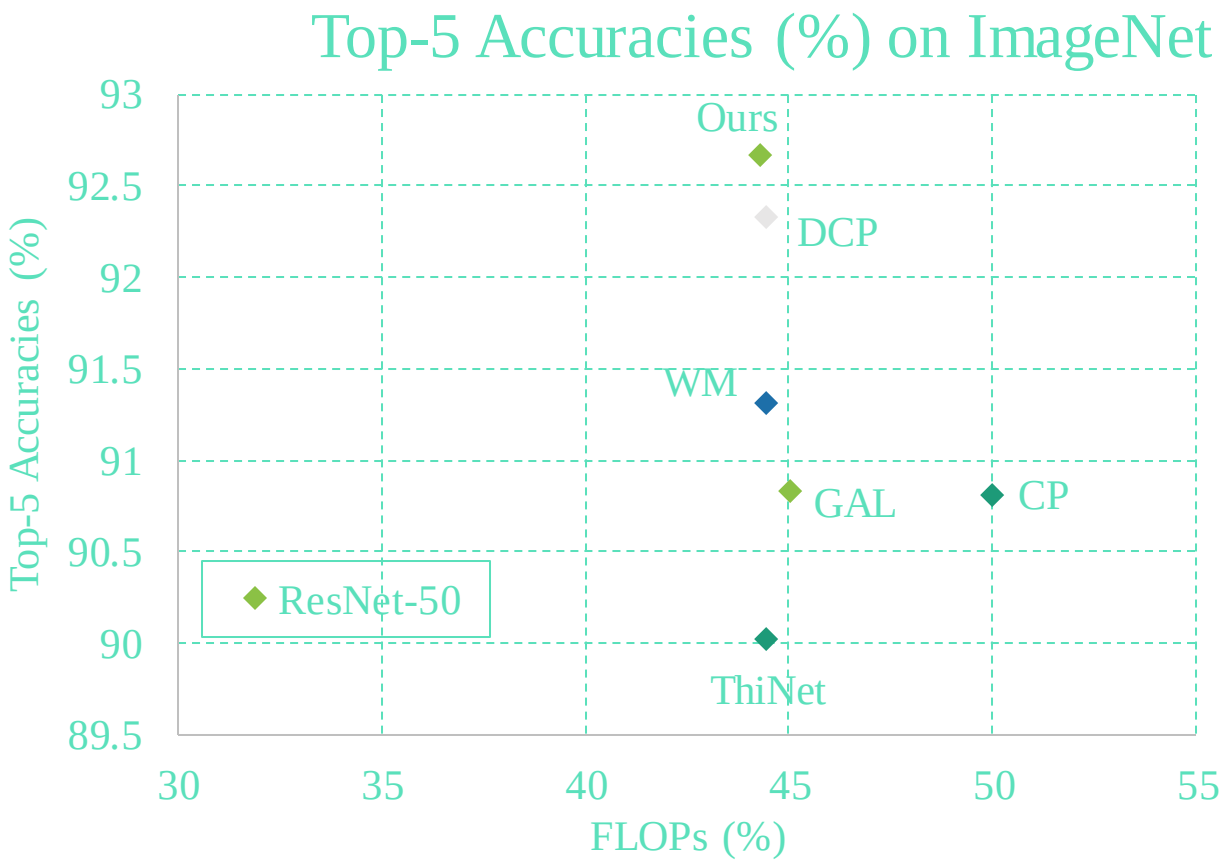


Guo, J., Ouyang, W. and Xu, D., Multi-Dimensional Pruning: A Unified Framework for Model Compression. *CVPR* 2020.

Experiments

- Image Classification (2D CNNs)

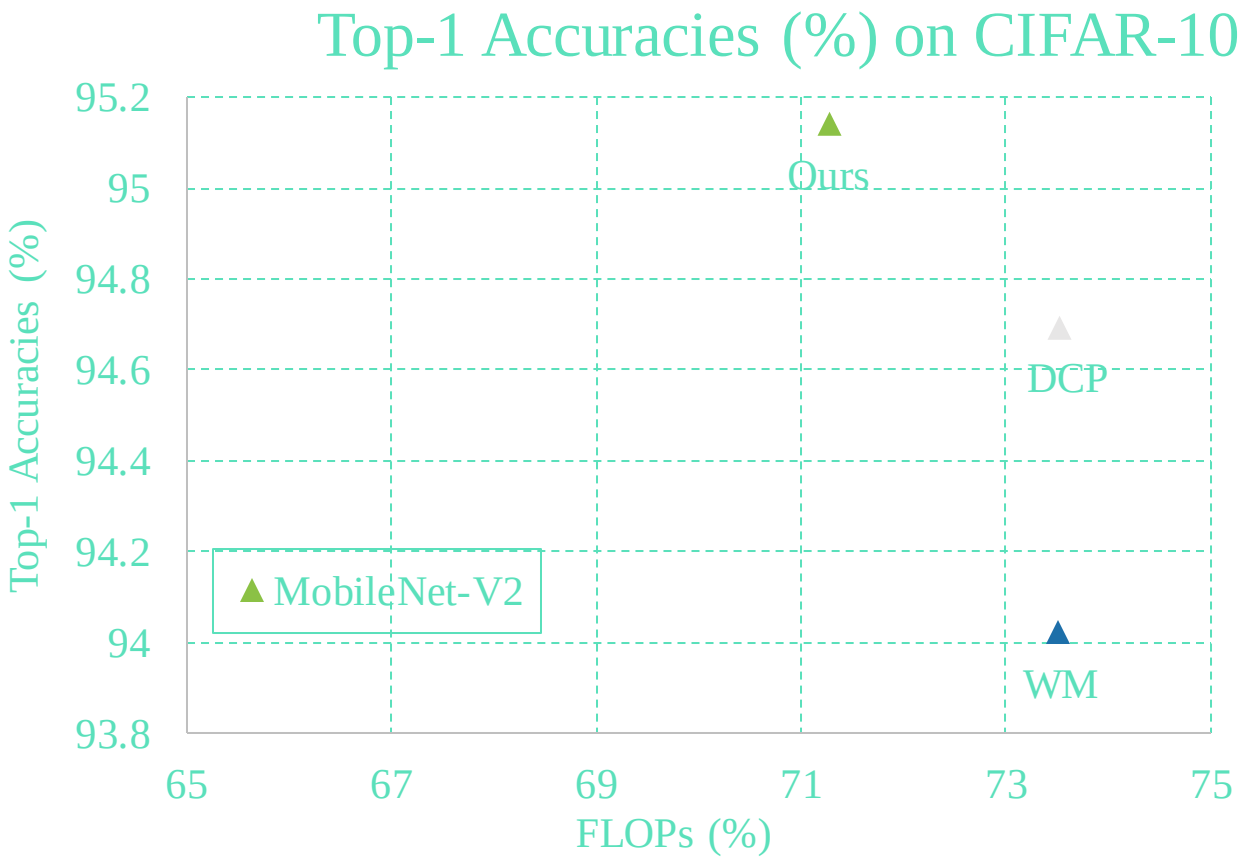


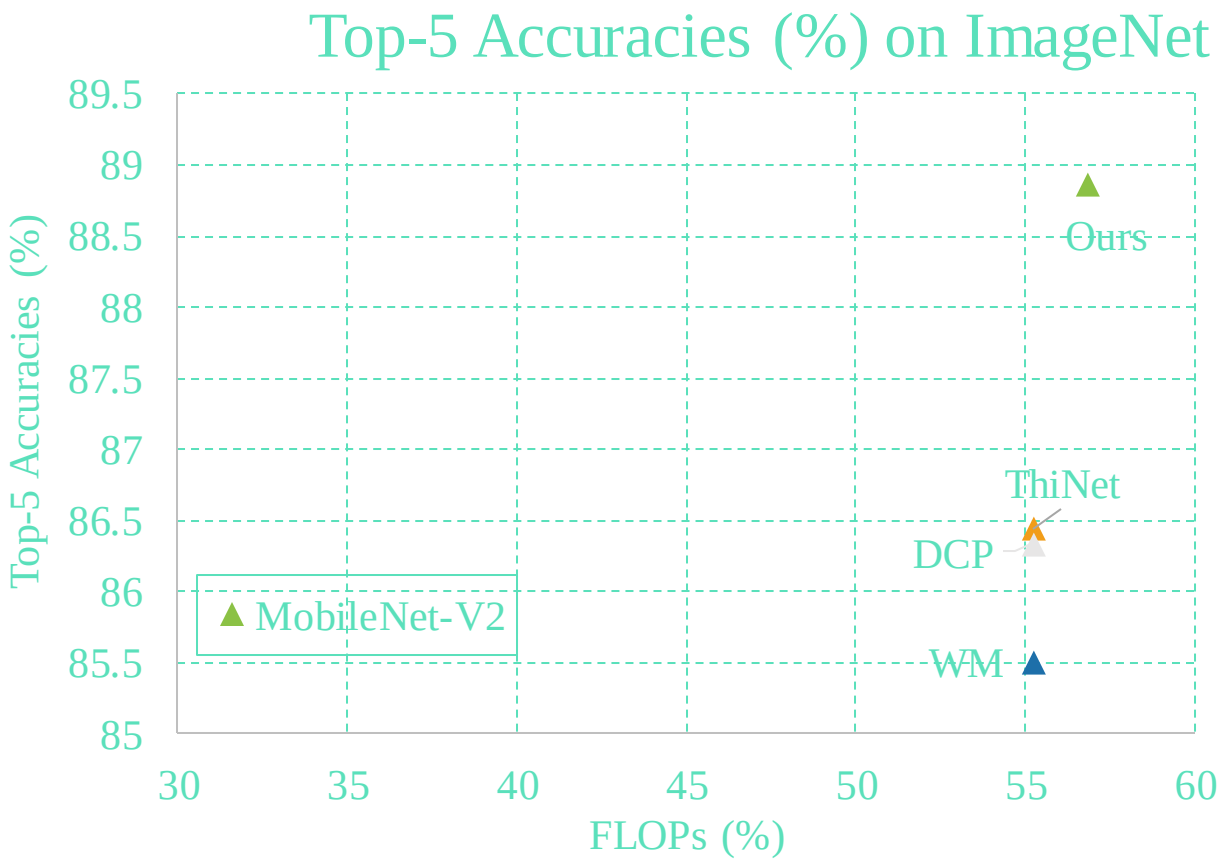


Guo, J., Ouyang, W. and Xu, D., Multi-Dimensional Pruning: A Unified Framework for Model Compression. *CVPR* 2020.

Experiments

- Image Classification (2D CNNs)



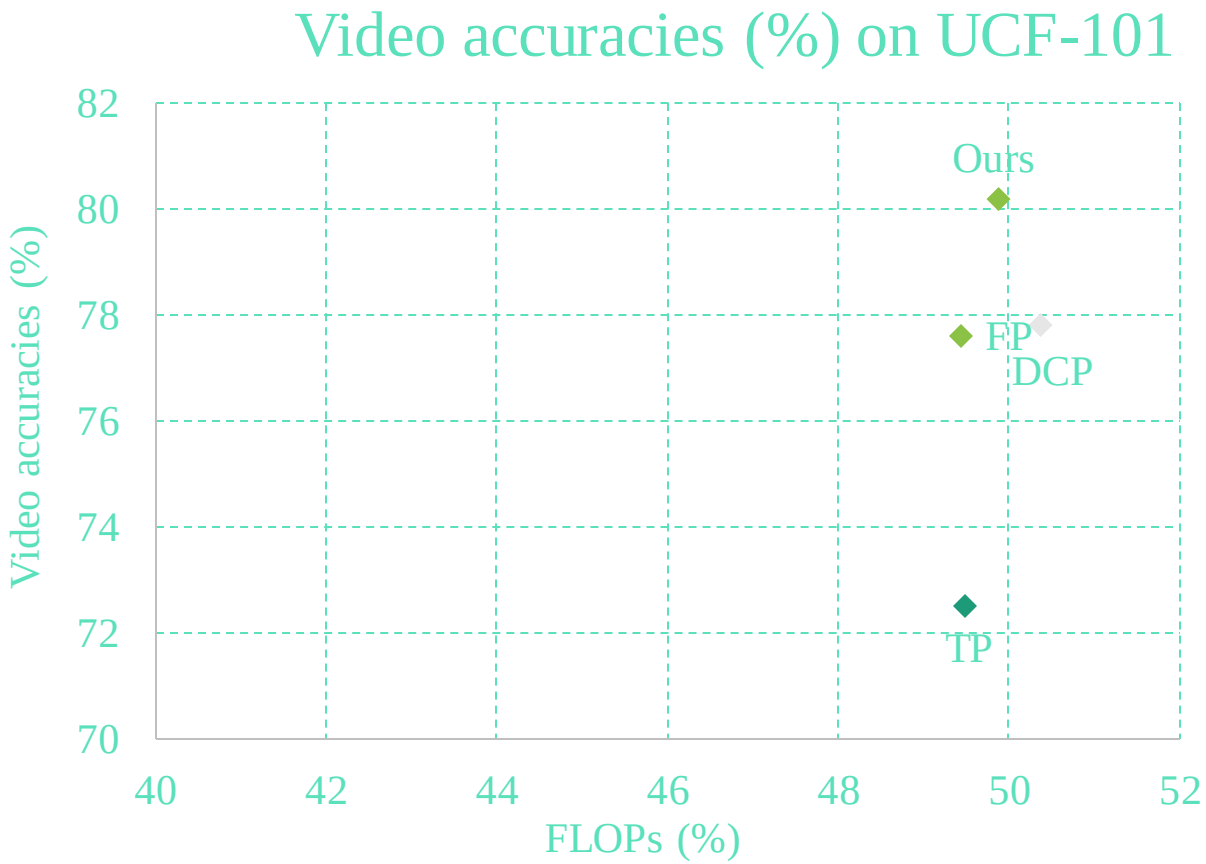


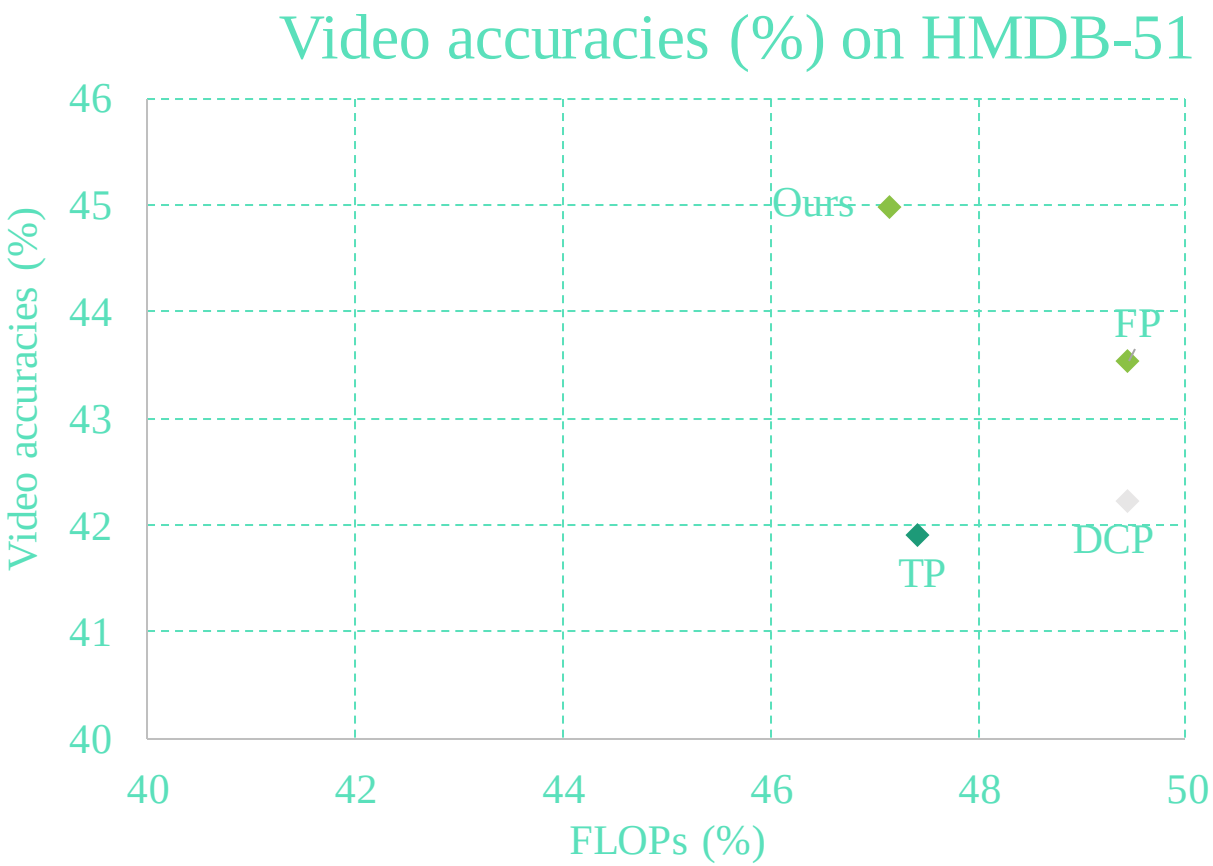
Guo, J., Ouyang, W. and Xu, D., Multi-Dimensional Pruning: A Unified Framework for Model Compression. *CVPR* 2020.

Experiments

- Video Classification (3D CNNs)

- C3D



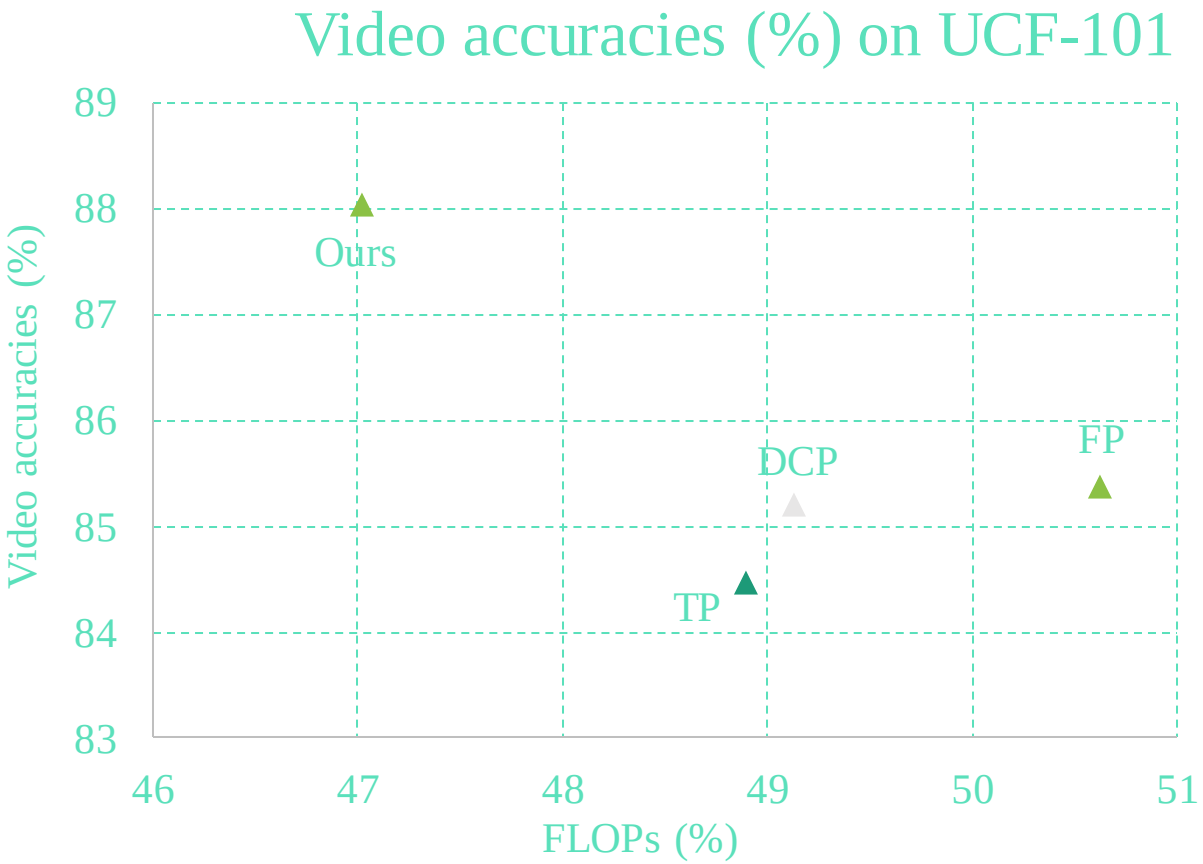


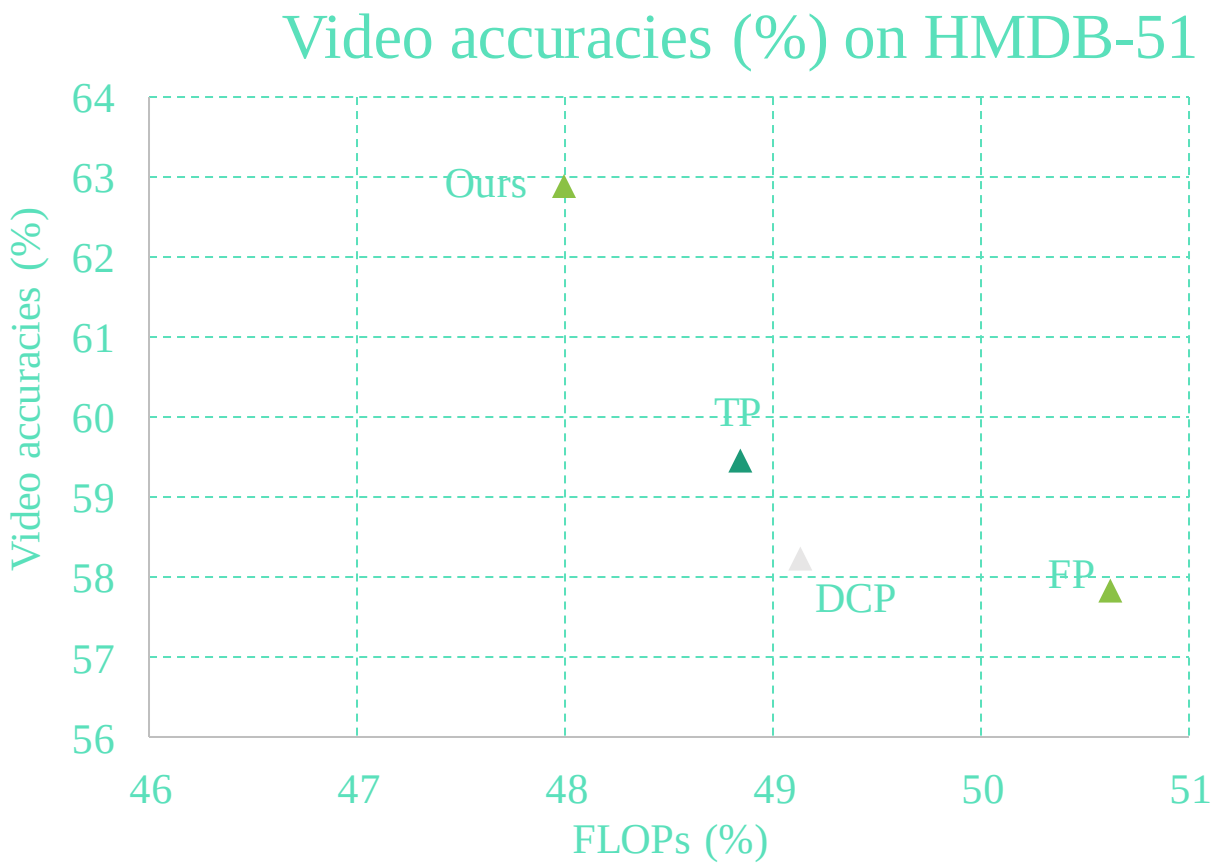
Guo, J., Ouyang, W. and Xu, D., Multi-Dimensional Pruning: A Unified Framework for Model Compression. *CVPR* 2020.

Experiments

- Video Classification (3D CNNs)

- I3D

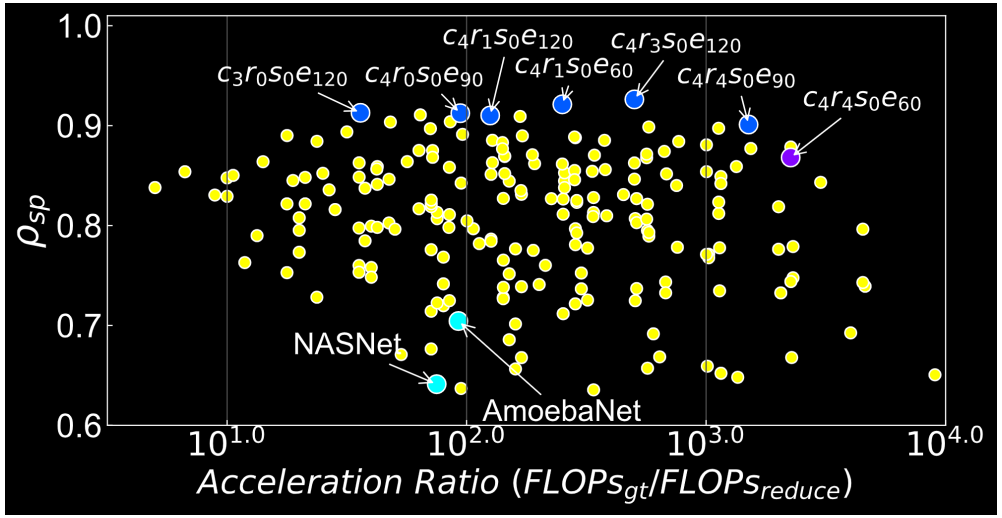


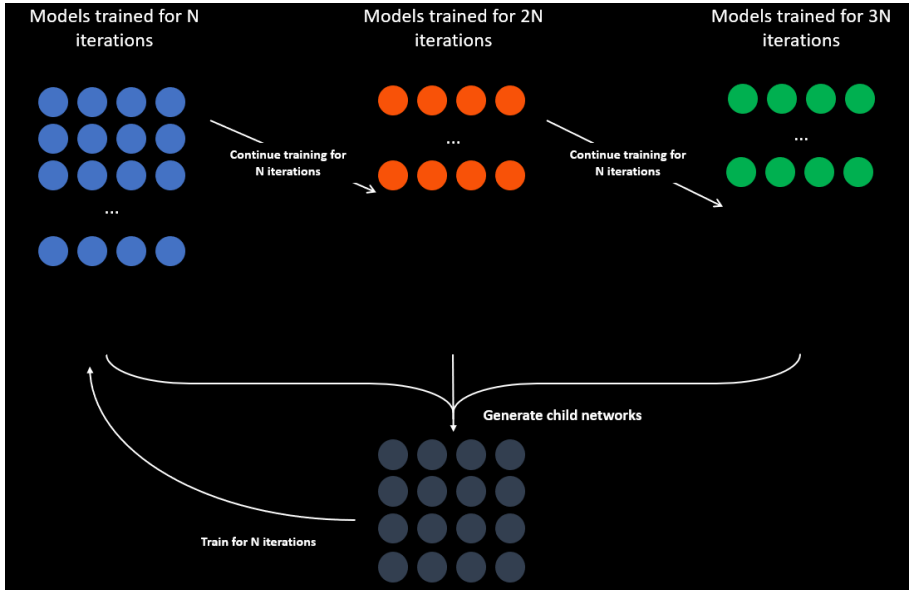


Guo, J., Ouyang, W. and Xu, D., Multi-Dimensional Pruning: A Unified Framework for Model Compression. *CVPR* 2020.

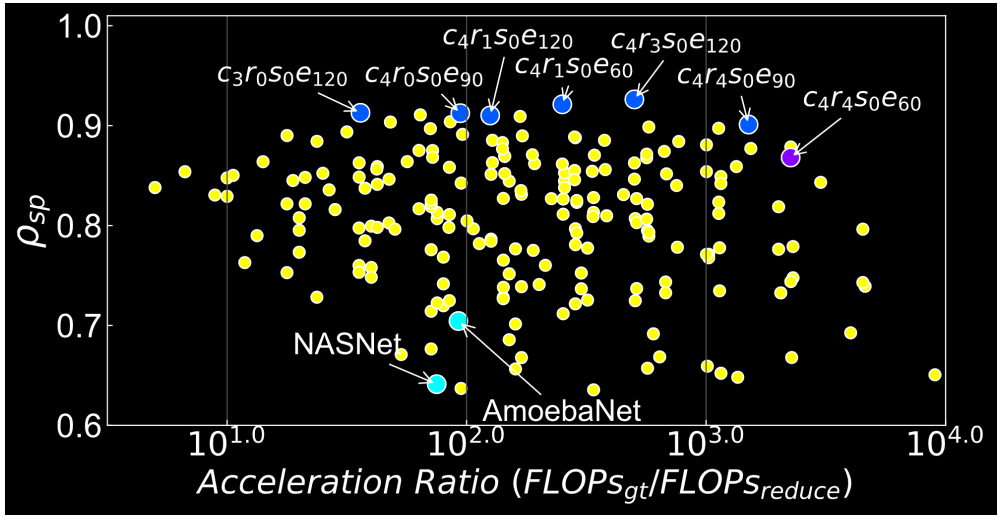
总结

- ECO-NAS
  - 可靠的搜索环境(Proxy)至关重要
  - 存在快且可靠的搜索环境(Proxy)
  - 利用环境的切换可以设计新的网络结构搜索算法

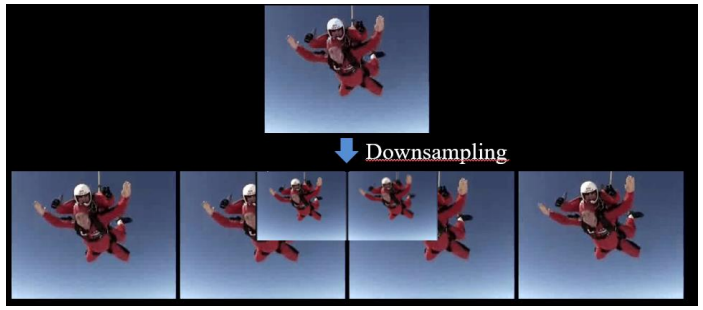




总结



- ECO-NAS
  - 可靠的搜索环境(Proxy)至关重要
  - 存在快且可靠的搜索环境(Proxy)
  - 利用环境的切换可以设计新的网络结构搜索算法
- Multi-dimensional pruning
  - 网络模型存在空间和时间的冗余
  - MDP：自动学习去掉这些冗余
  - 模型压缩可以从网络结构搜索方法得到启发



Thank you!