

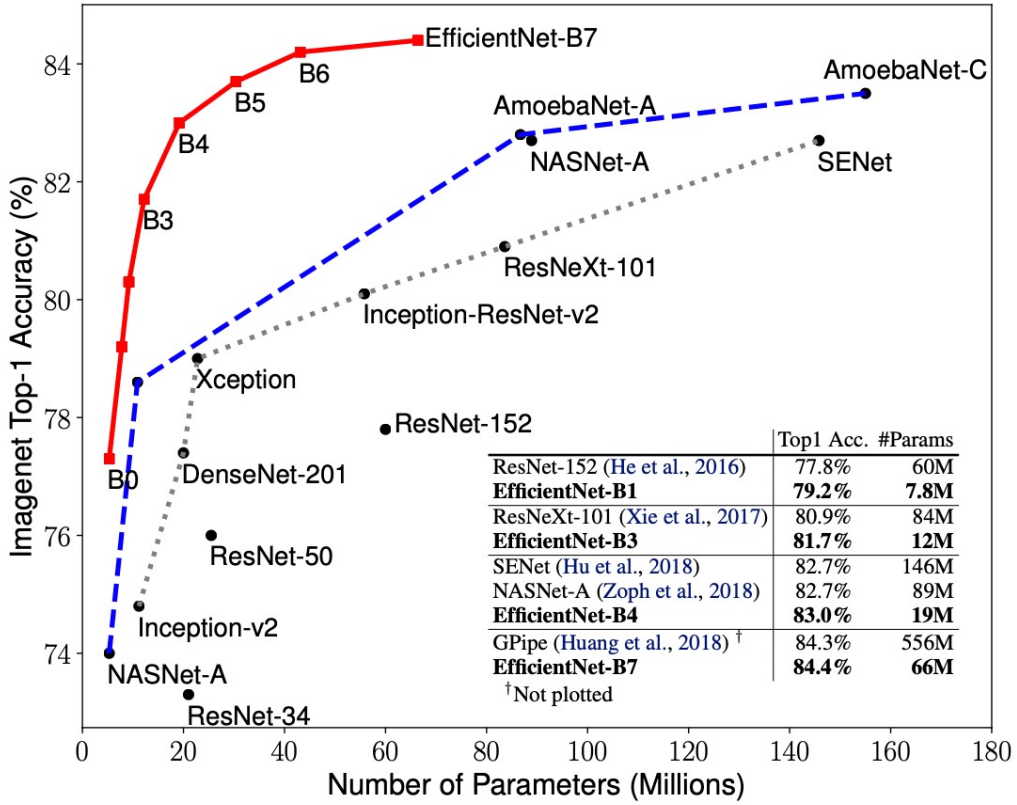
AdaSearch: *Many-to-One* Unified Neural Architecture Search via A Smooth *Curriculum*

https://aragakiyuiii.github.io/data/aaai22w_adasearch.pdf

Chunhui Zhang*, Yongyuan Liang*, Yifan Jiang*,

Brandeis University, Sun Yat-sen University, University of Texas at Austin

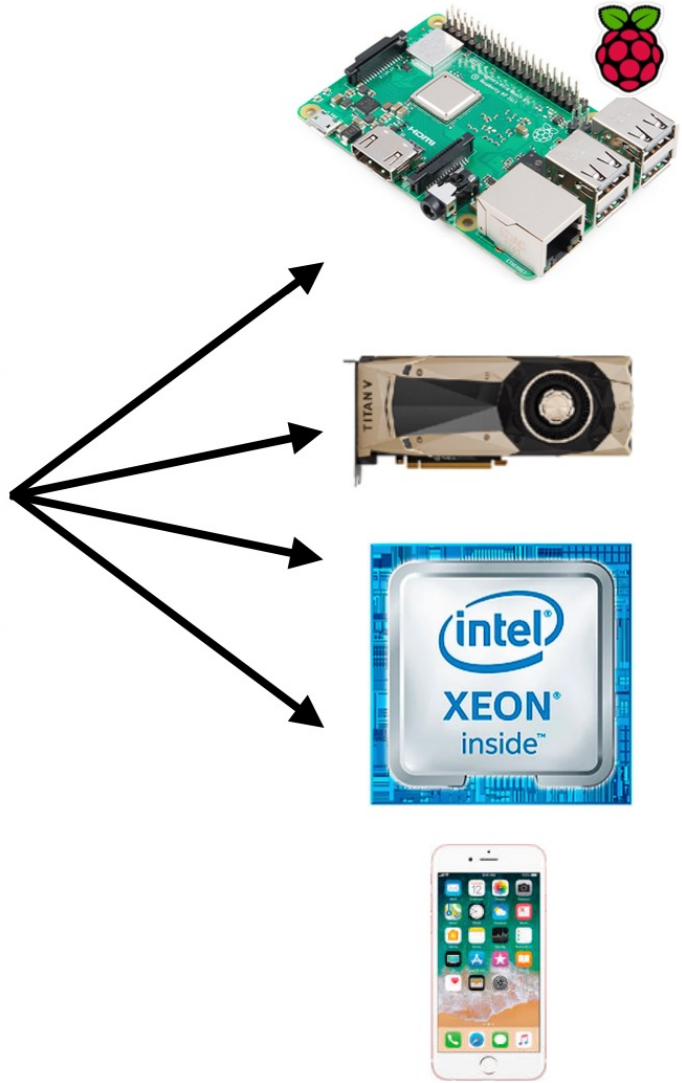
Model Size vs. Accuracy: a trade-off



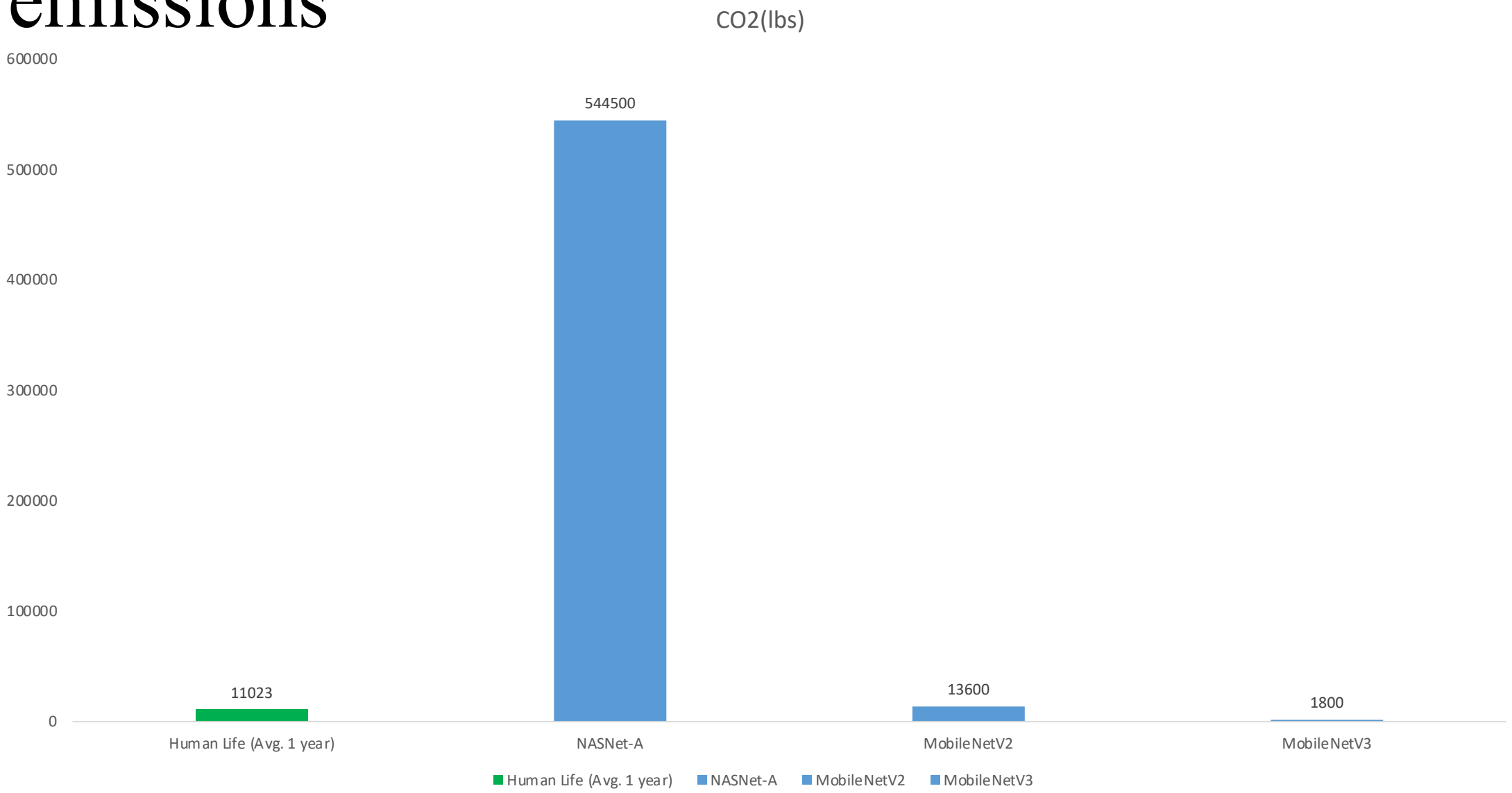
	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.2%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.7%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**83.0%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.4%**	**66M**

†Not plotted

Figure 1. **Model Size vs. ImageNet Accuracy.** All numbers are for single-crop, single-model. Our EfficientNets significantly outperform other ConvNets. In particular, EfficientNet-B7 achieves new state-of-the-art 84.4% top-1 accuracy but being 8.4x smaller and 6.1x faster than GPipe. EfficientNet-B1 is 7.6x smaller and 5.7x faster than ResNet-152. Details are in Table 2 and 4.



Cost of Neural Architecture Search: big CO2 emissions



Current Methods (e.g. PDARTS)

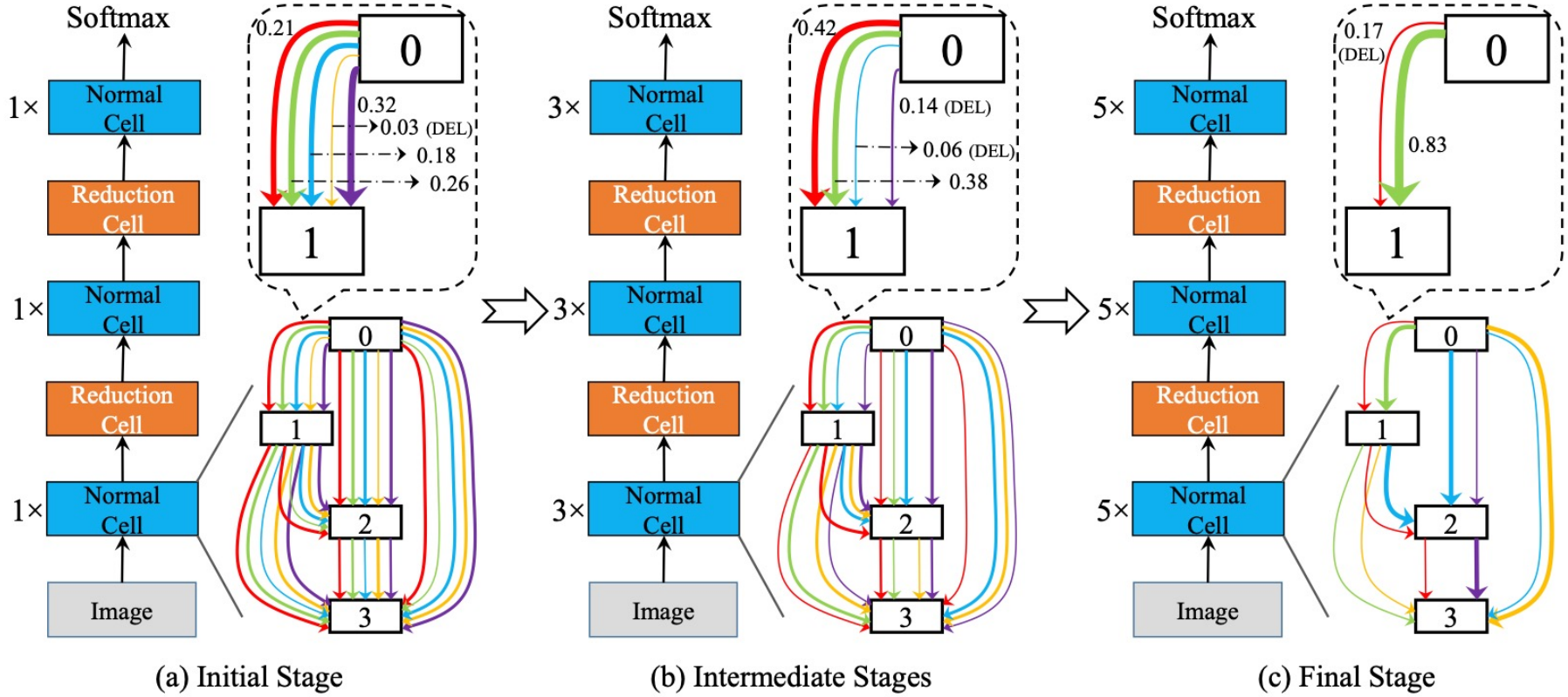


Figure 2: The overall pipeline of P-DARTS (best viewed in color). For simplicity, only one intermediate stage is shown, and only the normal cells are displayed. The depth of the search network increases from 5 at the initial stage to 11 and 17 at the intermediate and final stages, while the number of candidate operations (shown in connections with different colors) is shrunk from 5 to 3 and 2 accordingly. The lowest-scored ones at the previous stage are dropped (the scores are shown next to each connection). We obtain the final architecture by considering the final scores and possibly additional rules.

Current Methods (e.g. CNAS)

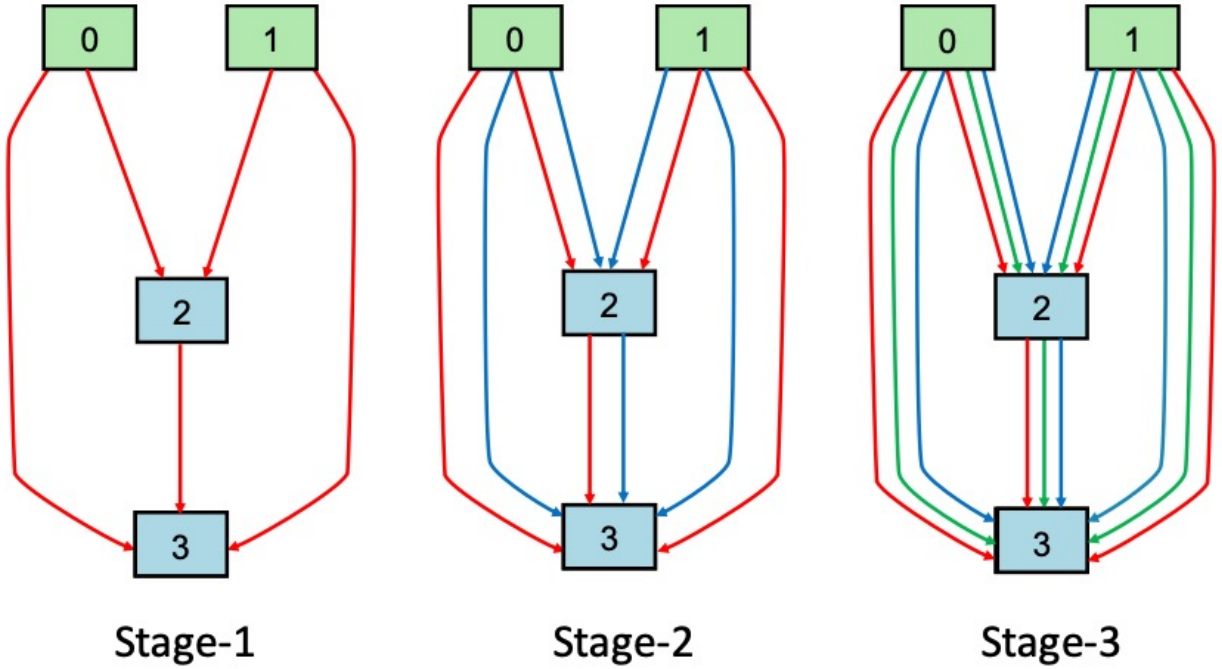


Figure 3. An overview of the search space used by CNAS. We show the candidate operations of the super network in different stages. The edges with different colors denote different operations. For simplicity, we omit the output node in this figure.

Key Question

How might we collectively search diverse networks with different size, with their ***weights sharing***, under one unified search procedure?

Answer

Expanding the curriculum learning from ***data*** level to ***model*** level for neural architecture search.

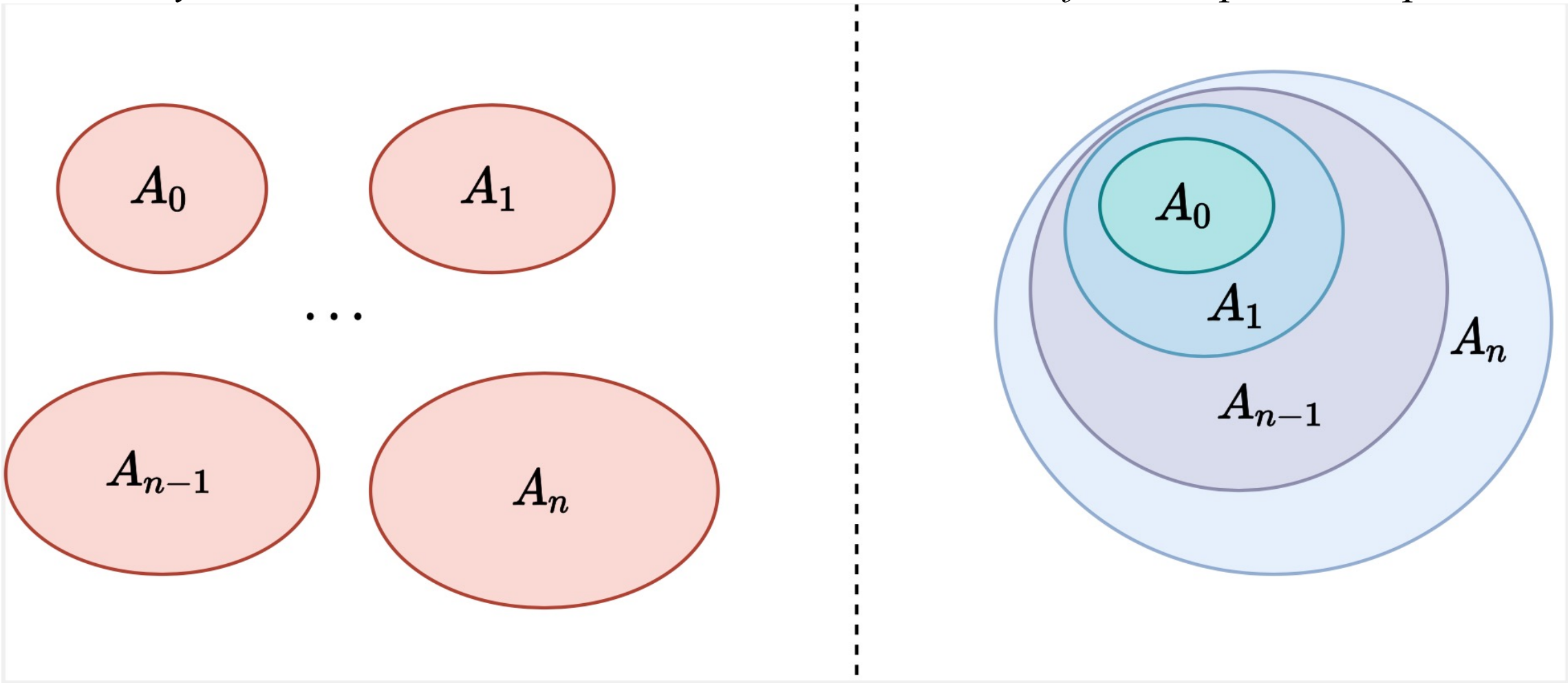
Curriculum Learning

smoothly learning representation on data level from simple to hard

similar to human learning

Curriculum Learning

smoothly search architectures on model-architecture level from simple to complex



Curriculum SuperNet Training *1. SuperNet2SuperNet*

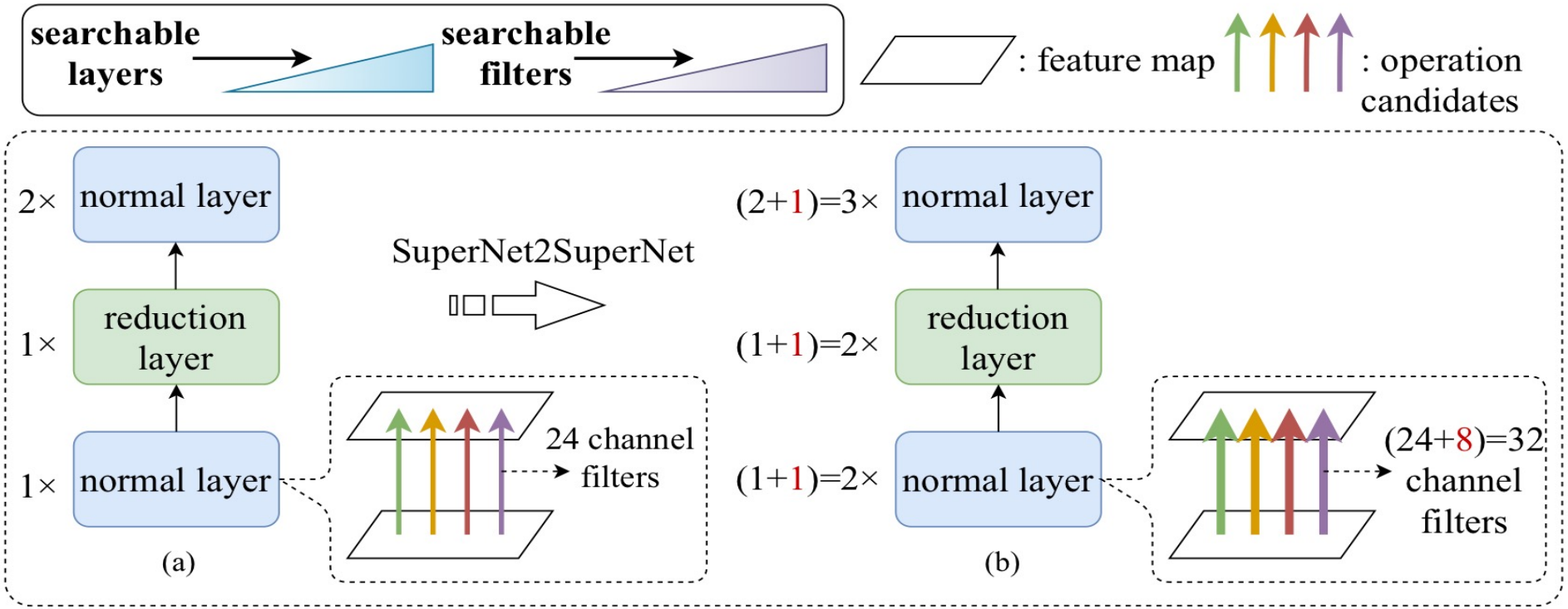


Figure 1: **SuperNet2SuperNet** (a) is a SuperNet with compact search space; (b) the search space grows up in number of layers and operation candidates' filters, which inherites parameters from (a) via SuperNet2SuperNet to smooth further search. Also, newly added layers and filters will still keep the functionality.

Curriculum SuperNet Training *1. SuperNet2SuperNet*

$$f_l(j) = \begin{cases} j & 1 \le j \le o^l \\ \text{random sample from } \{1, \cdots, o^l\} & o^l < j \le \hat{o}^l \end{cases}. \tag{1}$$

$$\hat{\boldsymbol{W}}_l[x, y, i, j] = \boldsymbol{W}_l[x, y, i, f_l(j)]. \tag{2}$$

$$\hat{\boldsymbol{W}}_{l+1}[x, y, j, k] = \frac{\boldsymbol{W}_{l+1}[x, y, f_l(j), k]}{\left|\{z | f_l(z) = f_l(j)\}\right|}. \tag{3}$$

Curriculum SuperNet Training *2. Progressive SuperNet Distillation*

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{hard} + \lambda\mathcal{L}_{soft}$$

Overview of AdaSearch

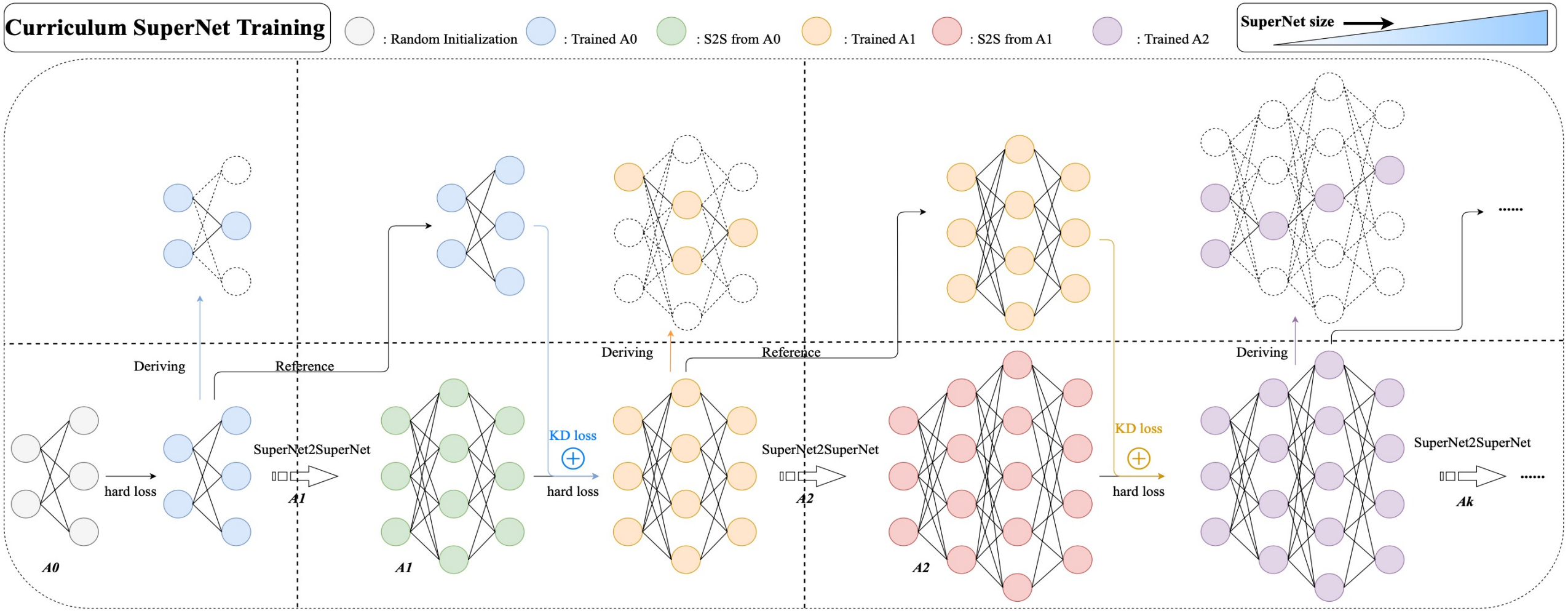
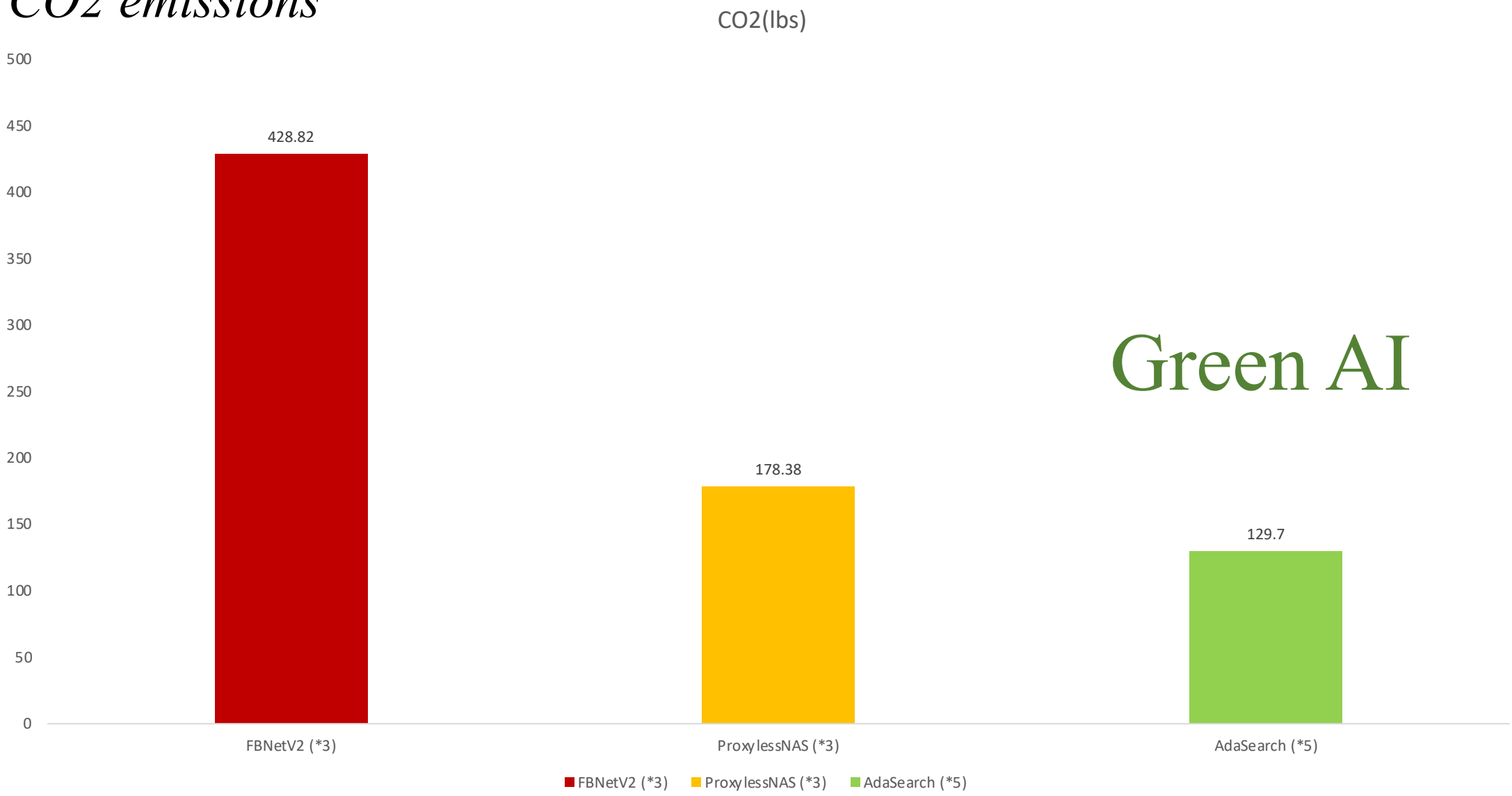


Figure 2: The overview of Curriculum SuperNet Training. Before the next search period's training, the parameters of student SuperNet are inherited from the past compact SuperNet by *SuperNet2SuperNet*. In the next search period, the soft output labels are adopted for the supervision of the student's exploring in more complex search space.

Cost of Neural Architecture Search: *ours have less CO2 emissions*



Green AI

Model	Search Cost			FLOPs
	GPU Hours(h)	Energy (kWh)	CO_2e (lbs)	
ProxylessNAS-R [Cai *et al.*, 2018]	200.0	59.46	59.46	320M
ProxylessNAS-G [Cai *et al.*, 2018]	200.0	59.46	59.46	465M
ProxylessNAS-G [Cai *et al.*, 2018]	200.0	59.46	59.46	487M
Total-cost	600.0	178.38	178.38	-
FBNetV2-F4 [Wan *et al.*, 2020]	216.0	64.21	61.26	238M
FBNetV2-L1 [Wan *et al.*, 2020]	648.0	192.63	183.78	325M
FBNetV2-L2 [Wan *et al.*, 2020]	648.0	192.63	183.78	422M
Total-cost	1512.0	449.47	428.82	-
AdaSearch-A0(ours)	**89.6**	26.75	25.52	107M
AdaSearch-A1(ours)	89.6+**74.4**=164.0	26.75+**22.19**=48.94	25.52+**21.17**=46.69	251M
AdaSearch-A2(ours)	164.0+**87.2**=251.2	48.94+**25.97**=74.91	46.69+**24.77**=71.46	369M
AdaSearch-A3(ours)	251.2+**96.8**=348.0	74.91+**28.94**=103.85	71.46+**27.61**=99.07	435M
AdaSearch-A4(ours)	348.0+**108.0**=456.0	103.85+**32.11**=135.96	99.07+**30.63**=129.70	512M
Total-cost (ours)	**456.0**	**135.96**	**129.70**	-

Table 3: Searching cost and energy consumption on ImageNet. AdaSearch obtains all architecture ($A0$-$A4$) in one procedure thus saves energy and searching time compared to others.

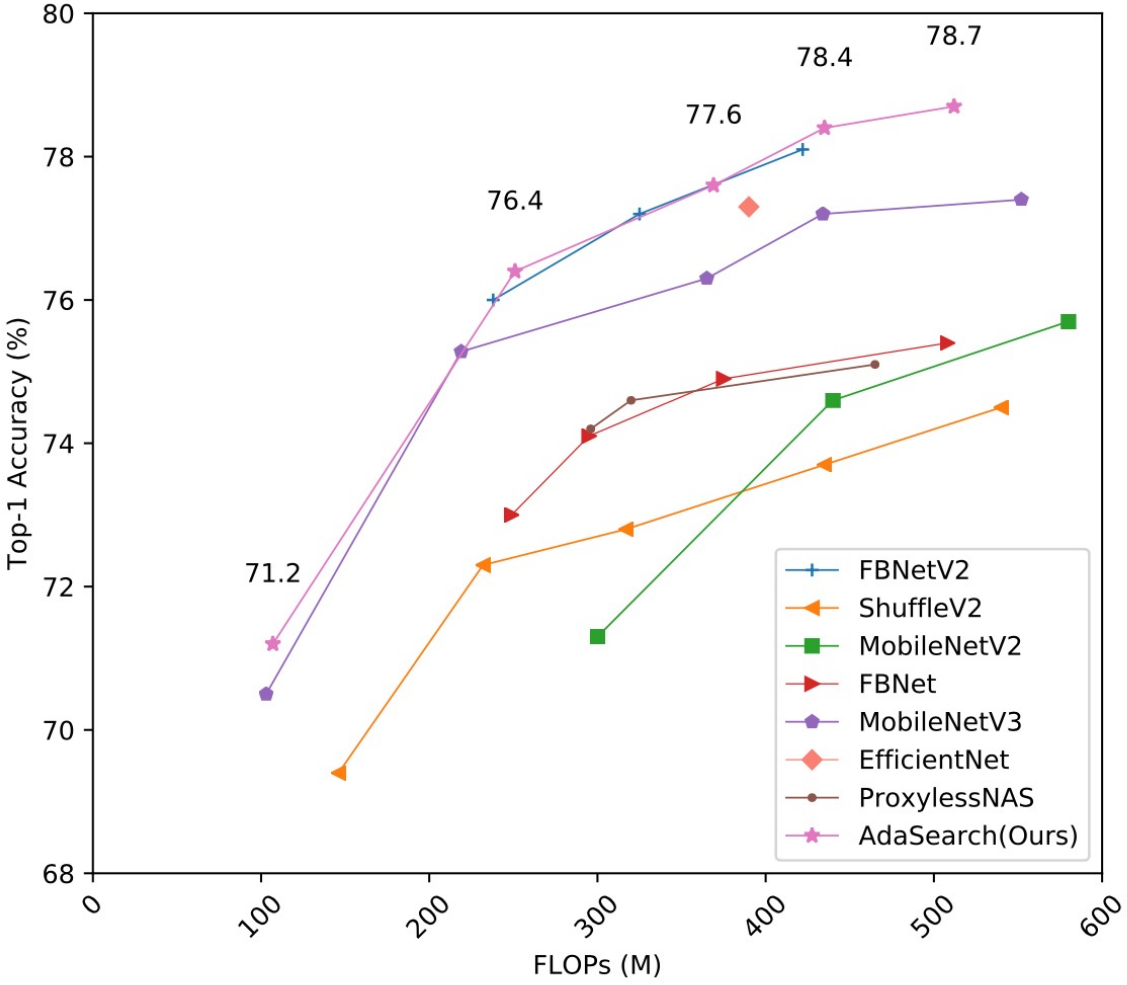


Figure 3: Accuracy vs. FLOPs

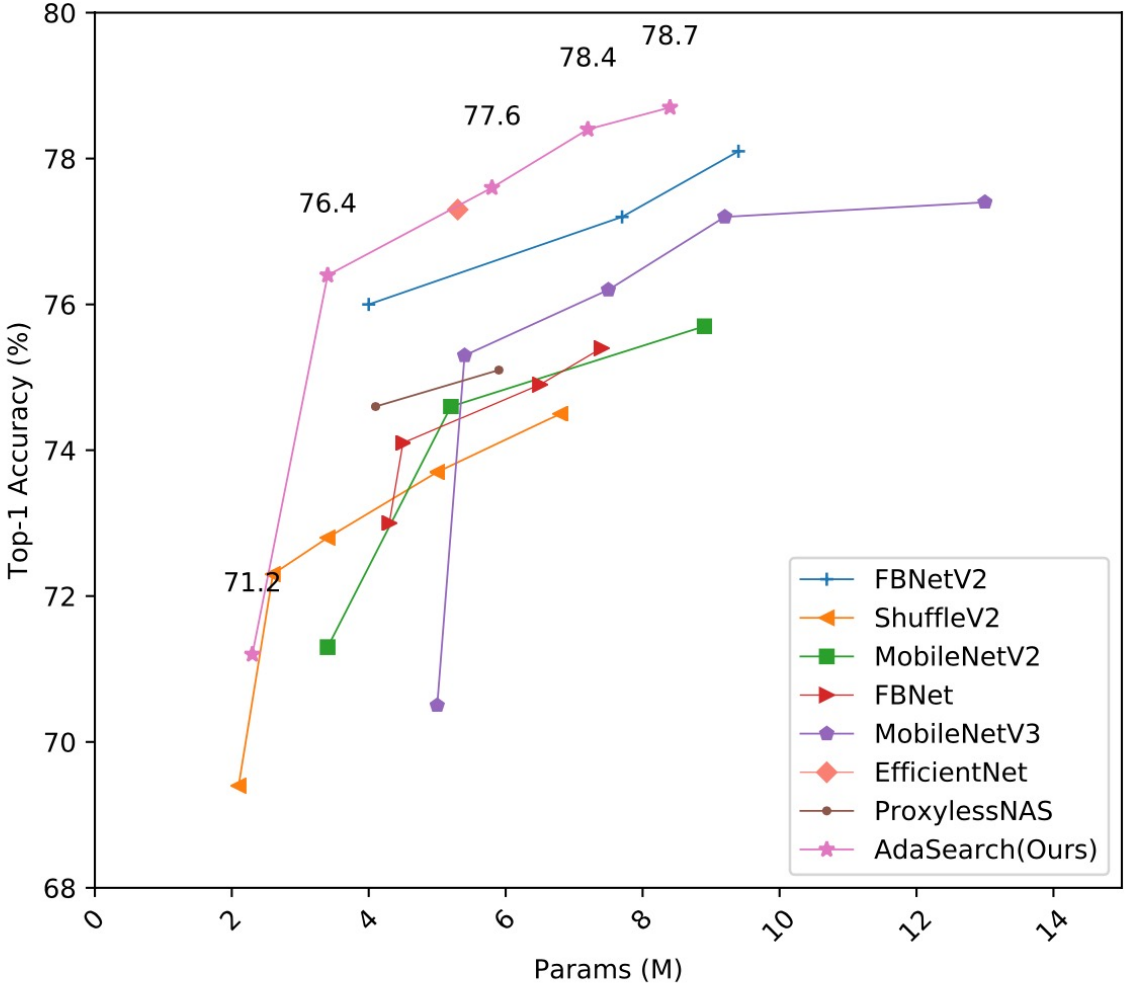
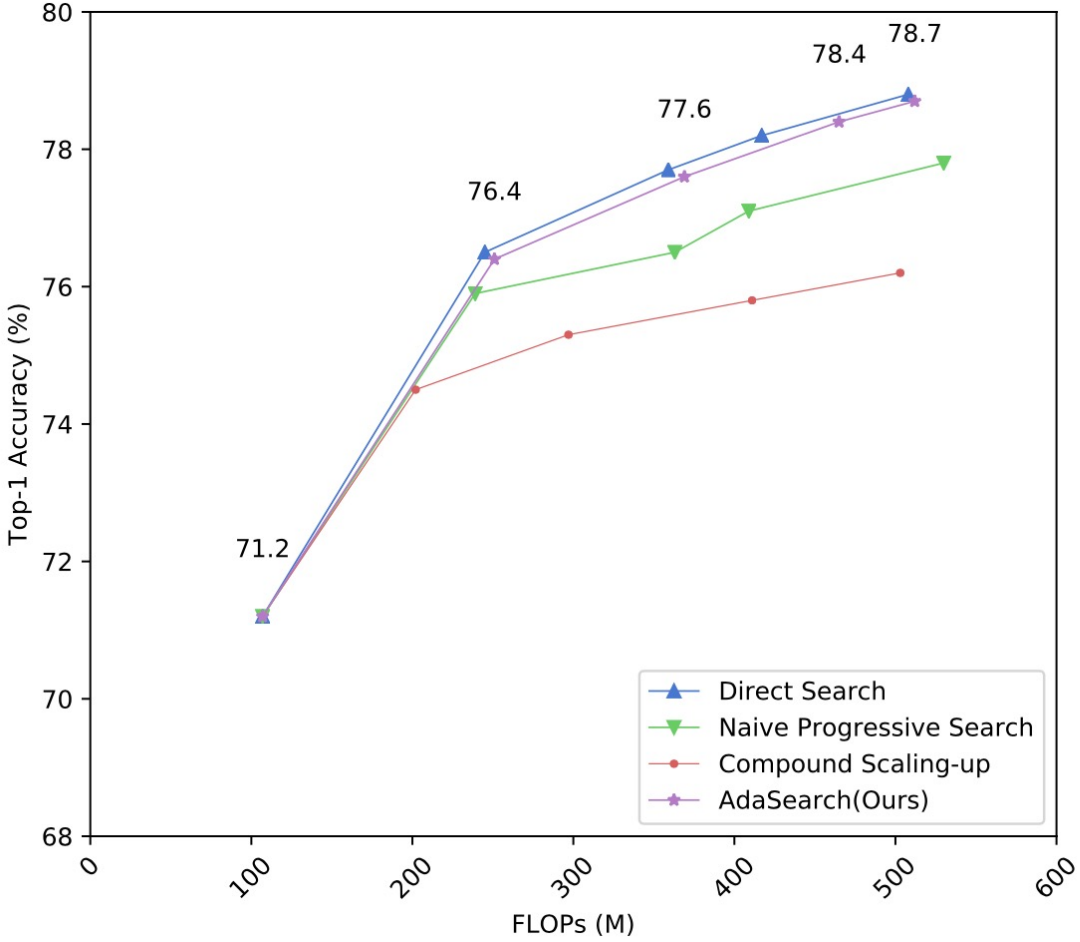
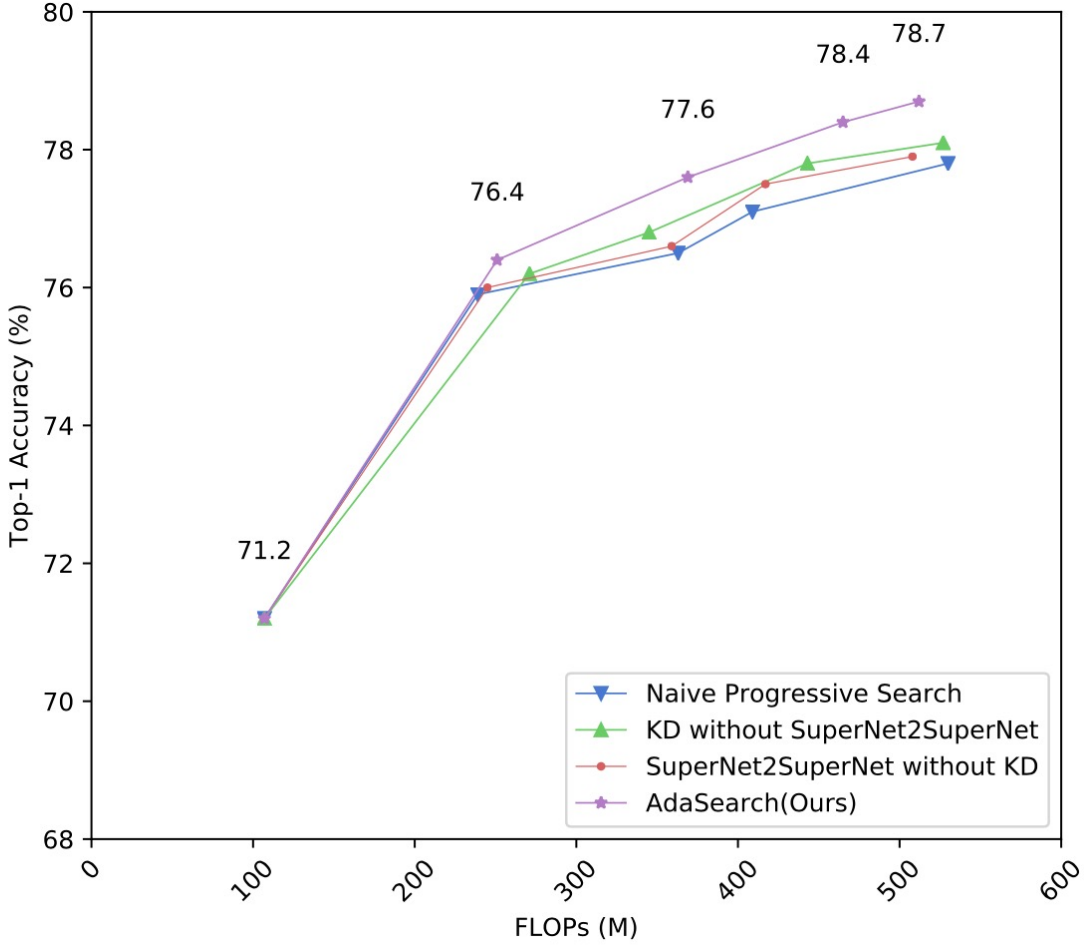


Figure 4: Accuracy vs. Params


80
78
76
74
72
70
68
Top-1 Accuracy (%)
0
100
200
300
400
500
600
FLOPs (M)
71.2
76.4
77.6
78.4
78.7
Direct Search
Naive Progressive Search
Compound Scaling-up
AdaSearch(Ours)


80
78
76
74
72
70
68
Top-1 Accuracy (%)
0
100
200
300
400
500
600
FLOPs (M)
71.2
76.4
77.6
78.4
78.7
Naive Progressive Search
KD without SuperNet2SuperNet
SuperNet2SuperNet without KD
AdaSearch(Ours)

Future Works

- Self-supervised Learning: Contrastive Learning, Predictive Learning
- etc...

Thanks for listening!