DENSO
Crafting the Core

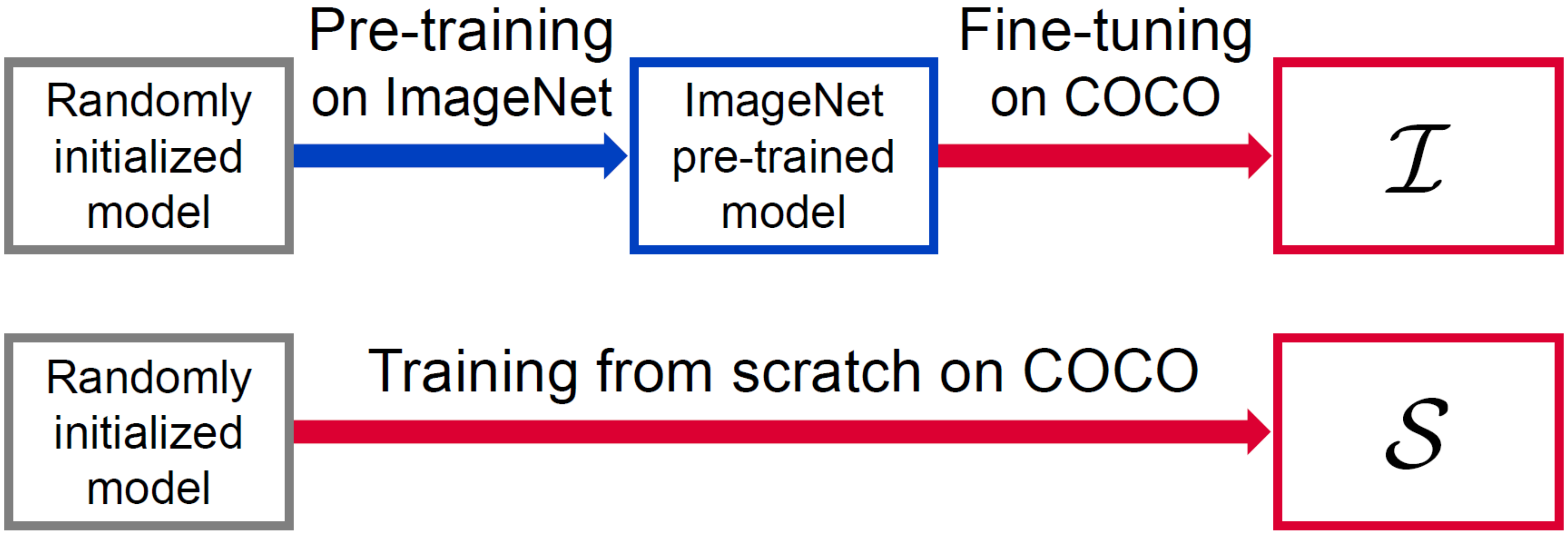
Understanding the Effects of Pre-Training for Object Detectors via Eigenspectrum

https://arxiv.org/abs/1909.04021

ICCV 2019 Neural Architects Workshop

Yosuke Shinya (DENSO CORPORATION)
Edgar Simo-Serra (Waseda University)
Taiji Suzuki (The University of Tokyo / RIKEN)

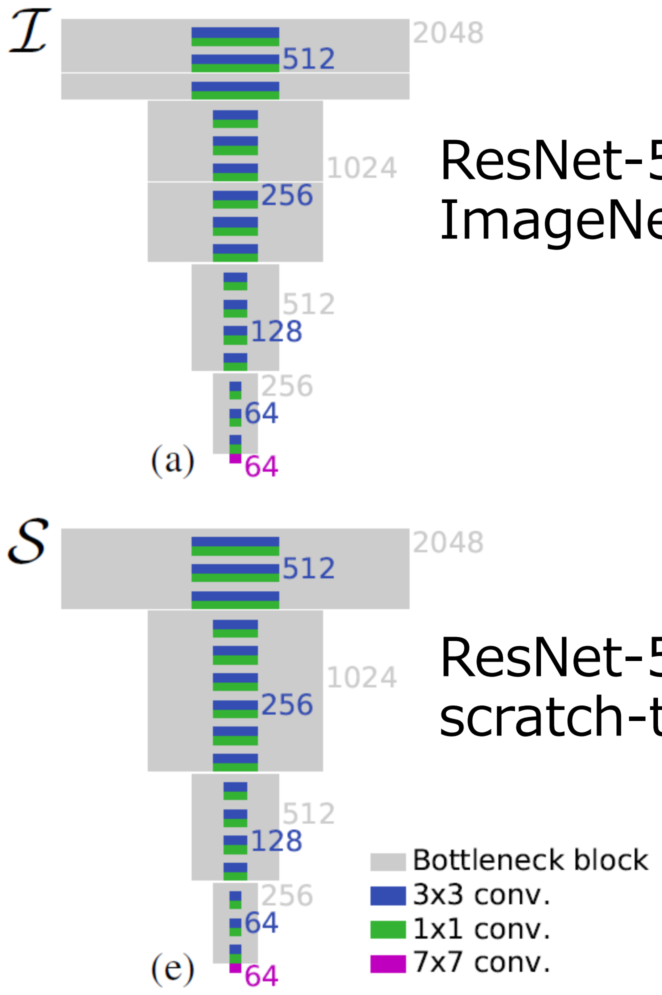
Background



- Prior works:
  - $\mathcal{I}$ can achieve high accuracy [Girshick et al., CVPR2014]
  - $\mathcal{S}$ can achieve similar accuracy to $\mathcal{I}$ [He et al., ICCV2019]
- This work:

  Do $\mathcal{I}$ and $\mathcal{S}$ converge to similar models? **No!**

(Extrinsic) Architecture



ResNet-50 backbone of
ImageNet fine-tuned detector $\mathcal{I}$

ResNet-50 backbone of
scratch-trained detector $\mathcal{S}$

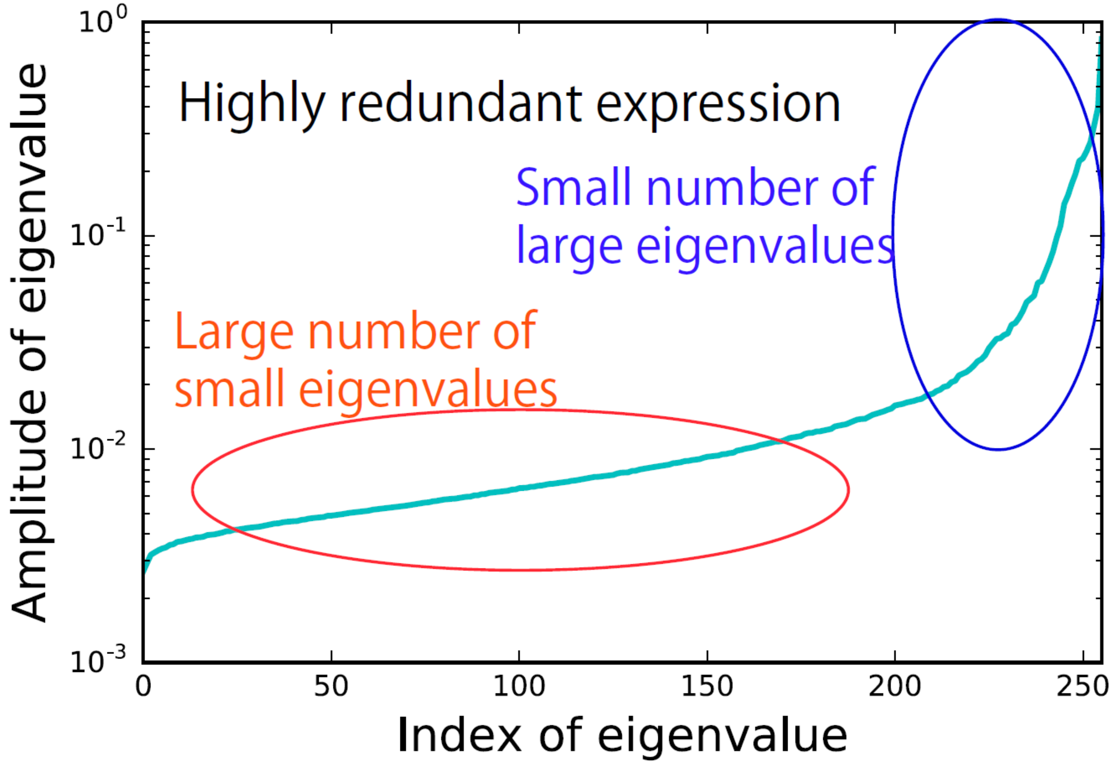
- Same (extrinsic) architectures
- Similar accuracy

How to analyze the difference?

Intrinsic dimensionalities

[Suzuki, AISTATS2018]

- Effective dimension of networks is less than actual number of parameters
- Intrinsic dimensionalities can be quantified by eigenspectra of covariance matrices of feature maps



[Suzuki et al., Tokyo Deep Learning Workshop 2018]

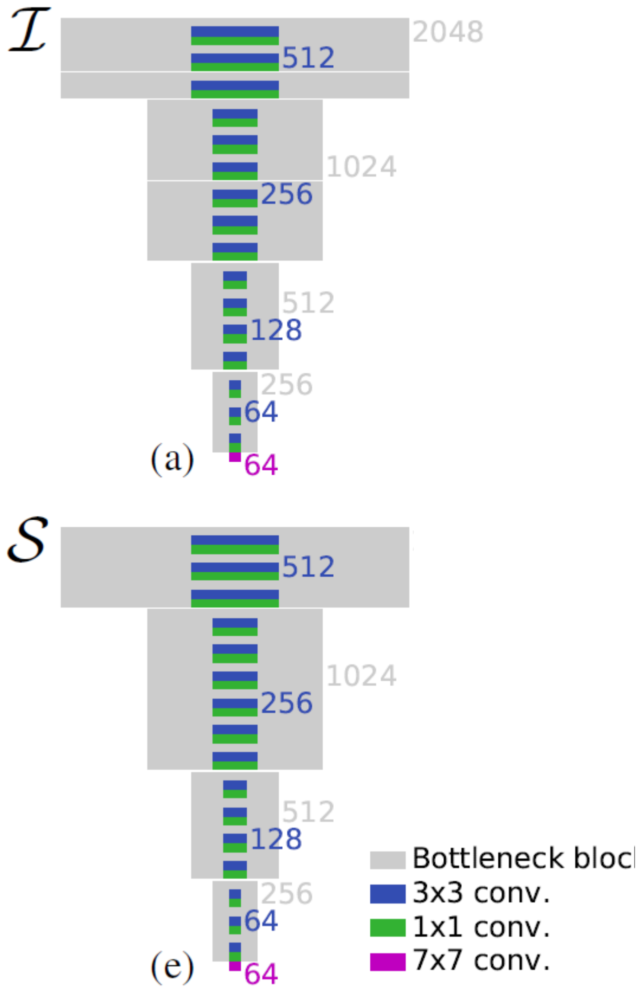
<u>Settings in our experiments</u>

- Covariance matrix:

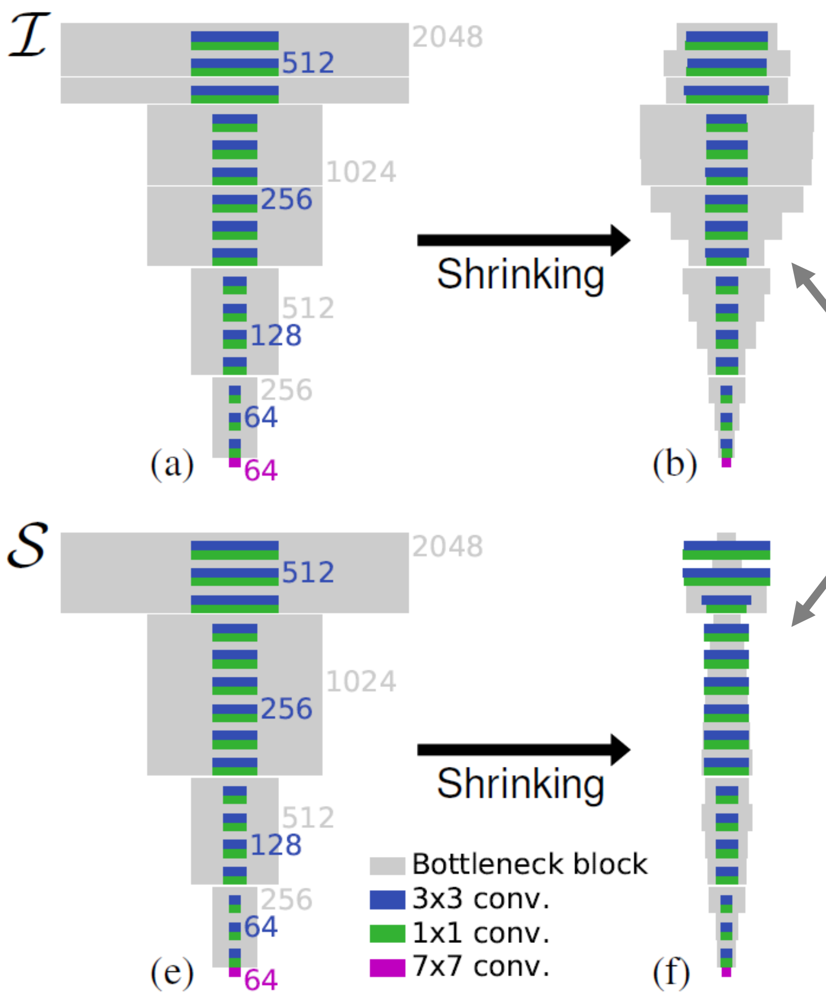
$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{W_i H_i}\sum_{x=1}^{W_i}\sum_{y=1}^{H_i} F_{i,x,y}F_{i,x,y}^{\top}$$

- Intrinsic dimensionalities: numbers of eigenvalues > 10^{-3}

Our method: Intrinsic Architecture Search

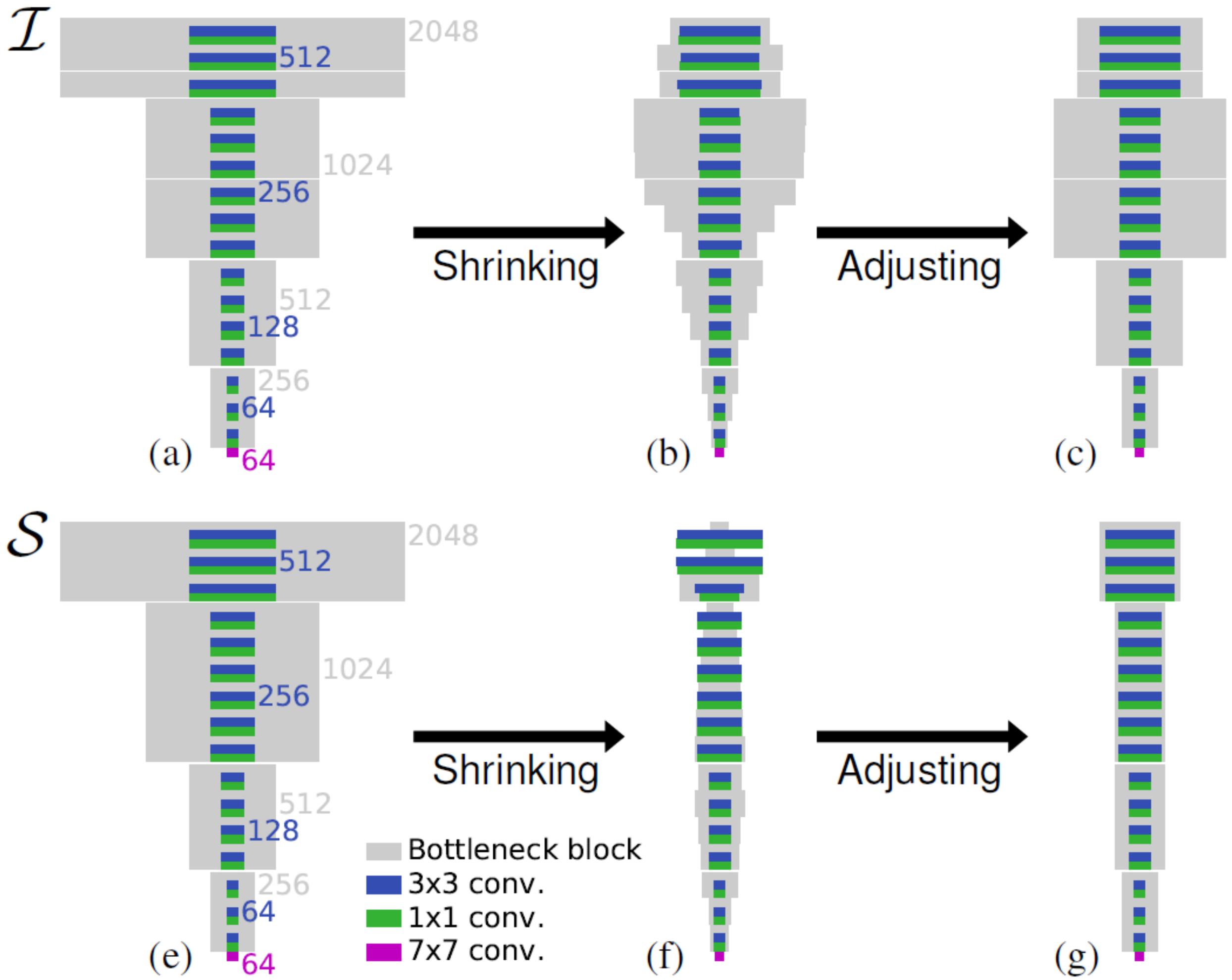


Our method: Intrinsic Architecture Search



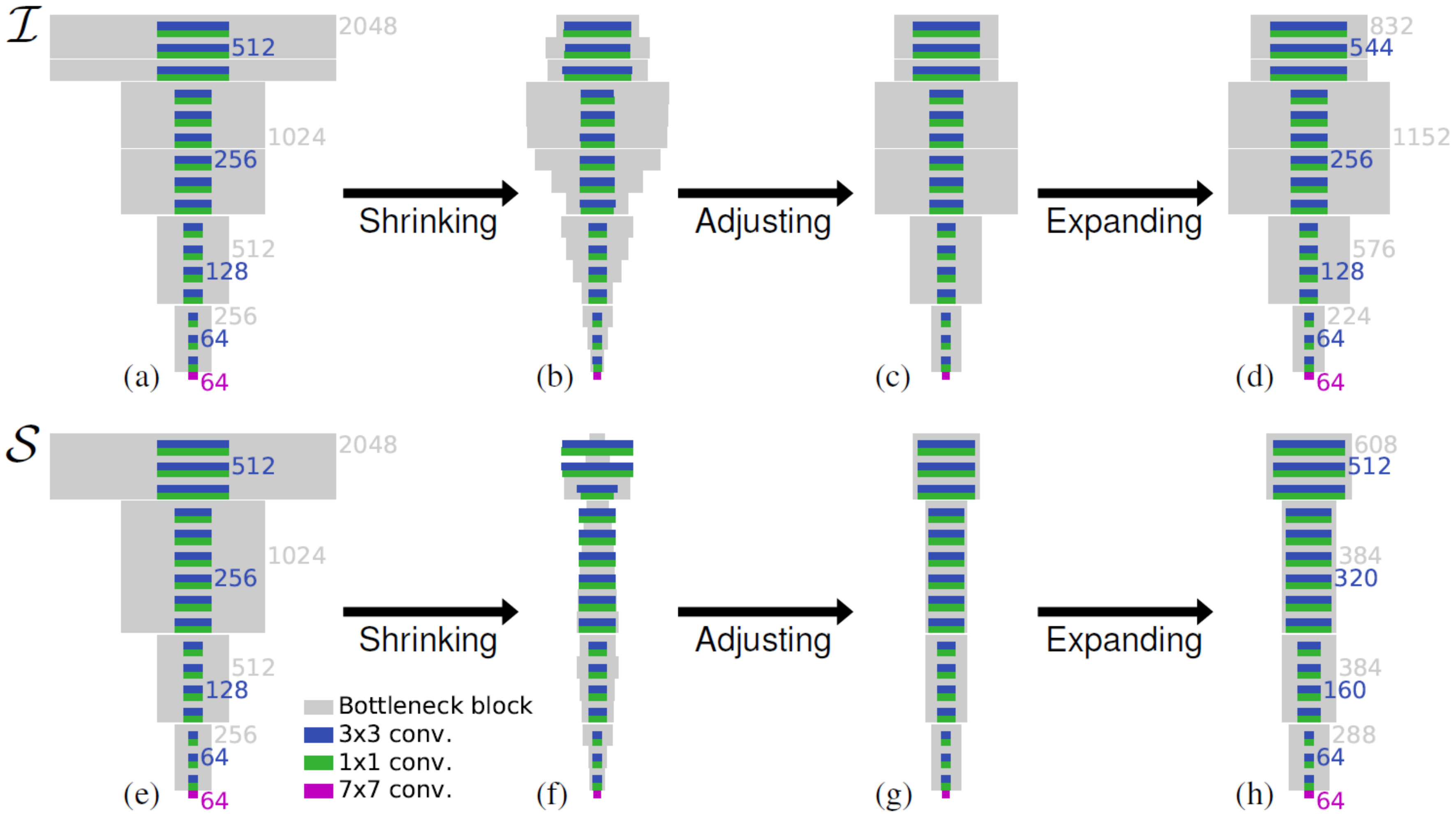
Clearly different intrinsic architectures

Our method: Intrinsic Architecture Search



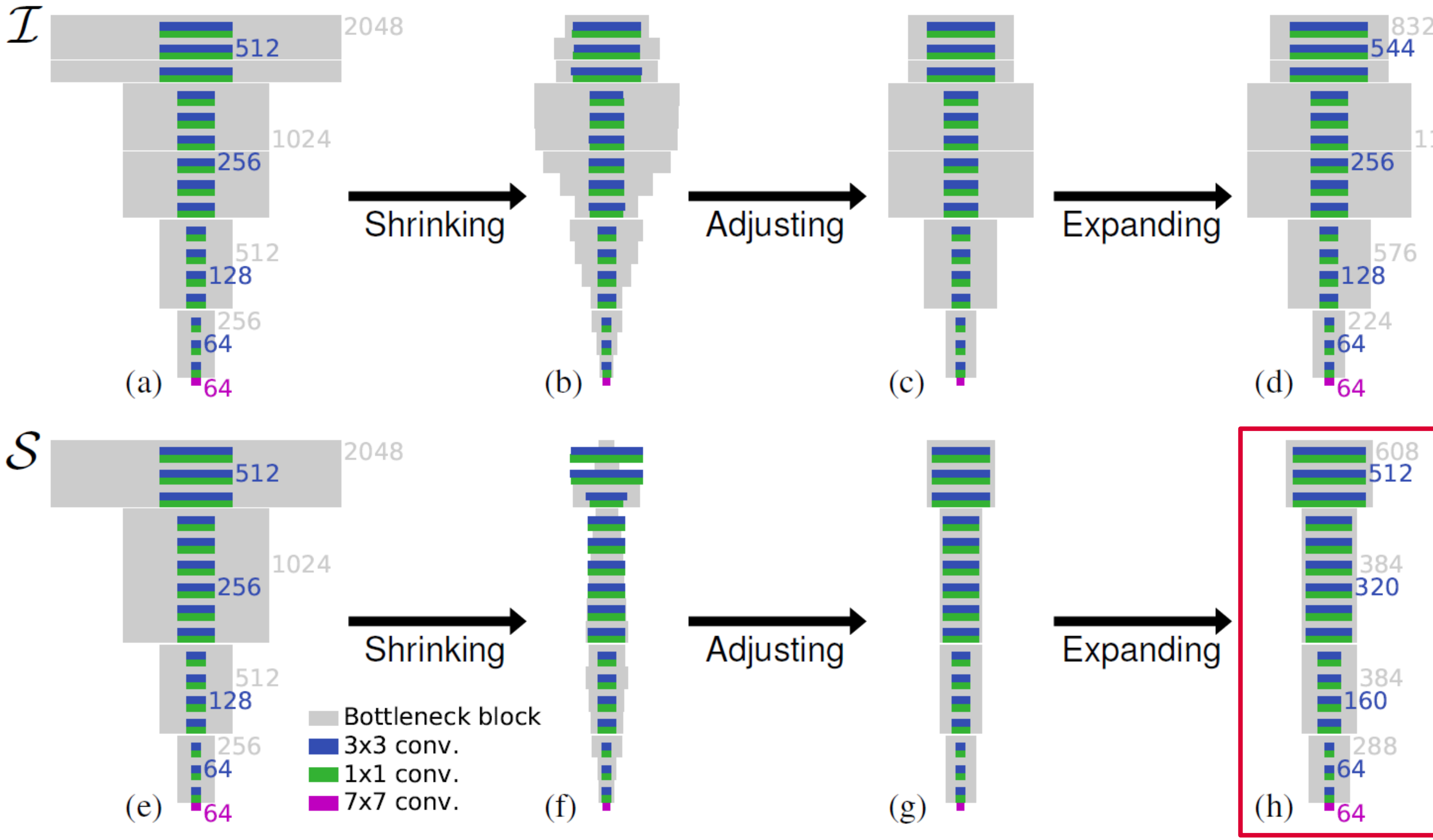
Yosuke Shinya / Applied AI R&I Dept.

Our method: Intrinsic Architecture Search



Yosuke Shinya / Applied AI R&I Dept.

Our method: Intrinsic Architecture Search



Discovered backbone

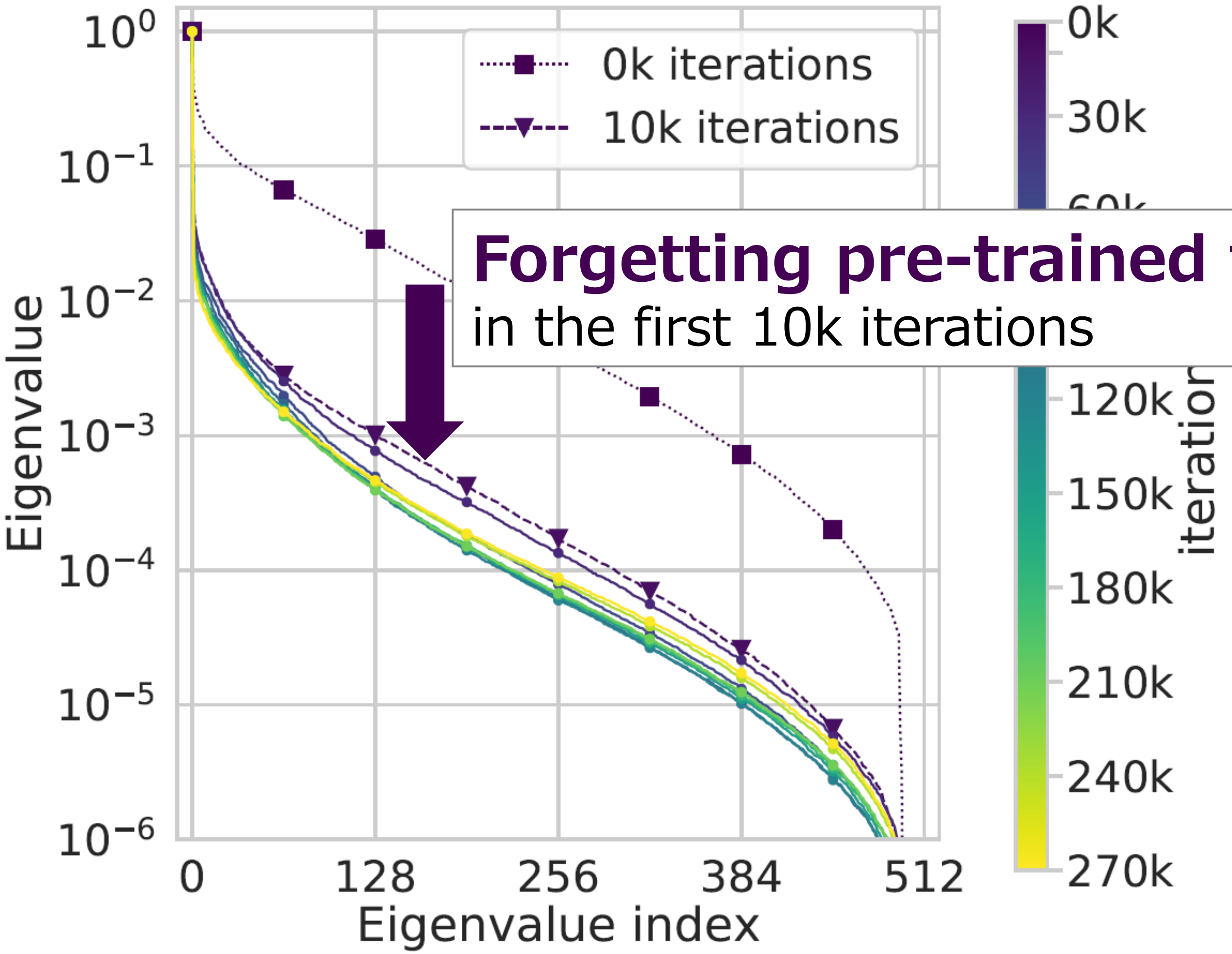
- Similar COCO AP
- Similar MACs
- 27% fewer parameters
- Lower accuracy in ImageNet classification

(Compared with ResNet-50)

Yosuke Shinya / Applied AI R&I Dept.

Eigenspectrum dynamics during fine-tuning (Dropping)

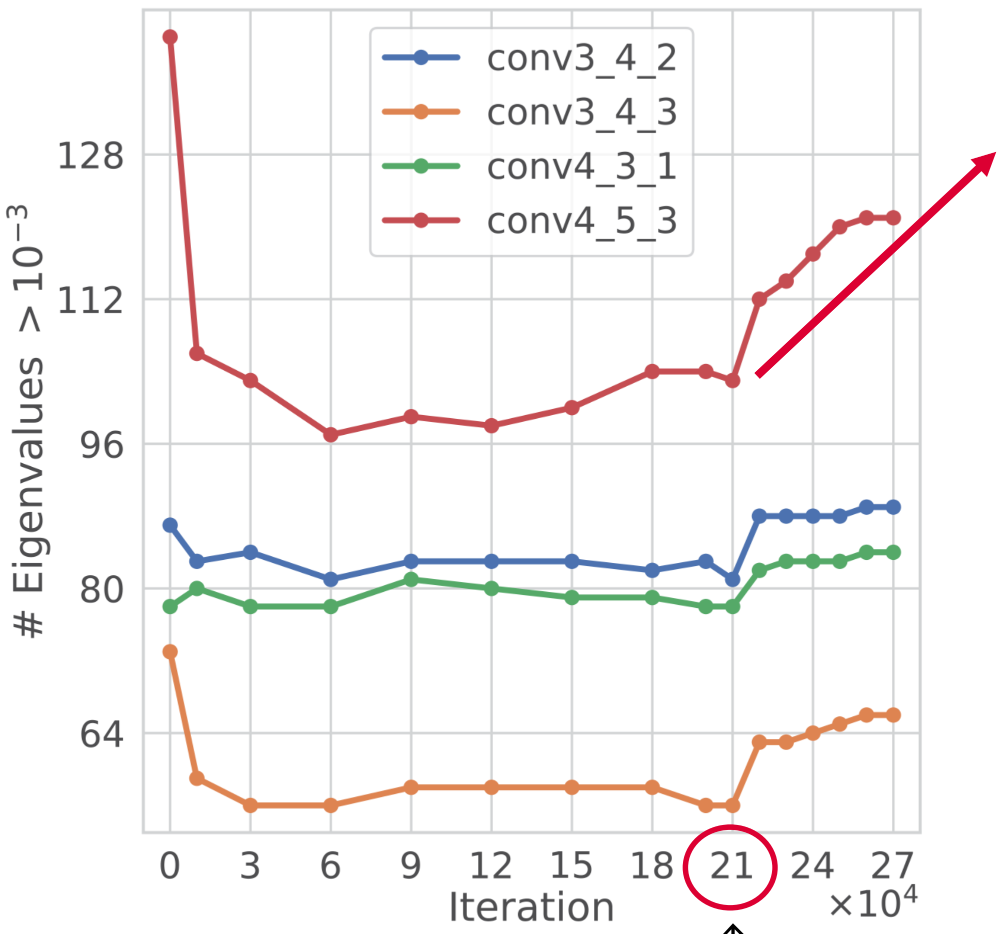
A feature map in ResNet stage 5



Forgetting pre-trained features
in the first 10k iterations

Yosuke Shinya / Applied AI R&I Dept.

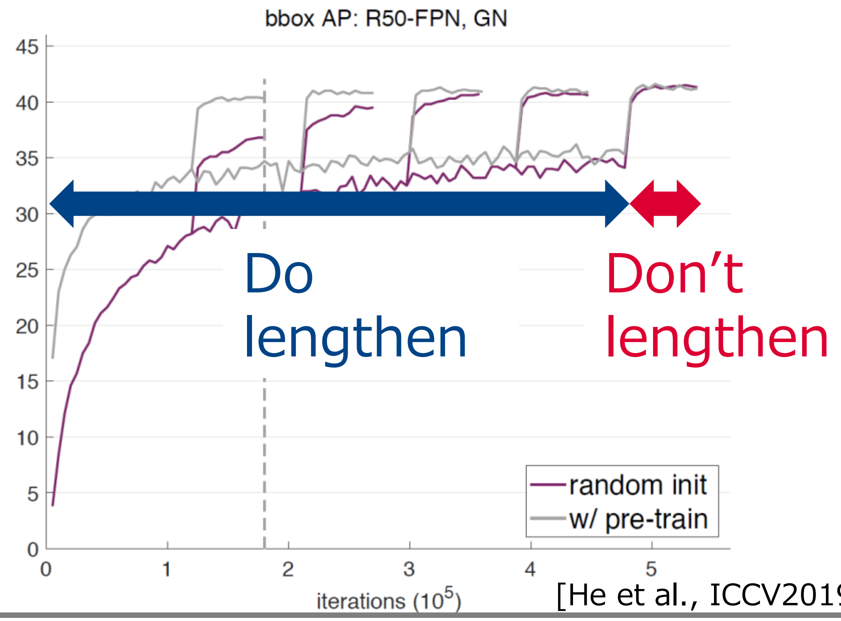
Eigenspectrum dynamics during fine-tuning (Rebounding)



↑
First learning rate decay

Training on small learning rate
→ captures detailed information about training data
→ increases eigenvalues
→ increases variance [Suzuki, AISTATS2018]
→ causes overfitting

Learning rate schedule of
"Rethinking ImageNet Pre-Training"



[He et al., ICCV2019]

Conclusions: Effects of pre-training

- ImageNet pre-training increases intrinsic dimensionalities in higher layers
- Increase of parameters caused by them
  - does not improve COCO AP
  - improve classification ability

- Current standard architectures and fine-tuning methods of object detectors are insufficient for utilizing the classification ability due to the forgetting

More appropriate architectures and knowledge-transfer methods are needed

- Hand-crafted architectures for sharing parameters → Automatic sharing
- Parameter transfer (fine-tuning) → Feature transfer or others

Where are we going?

(1) **Understanding multi-task training**

・How to deal with compression of task-irrelevant information?

(2) **Task-specific architectures and NAS methods**

・How to design architectures considering task differences?

・How to overcome long training time for object detection NAS?

- Reusing ImageNet pre-trained weights are effective but insufficient

(3) **Simultaneous optimization of architectures and parameters**

・How to optimize simultaneously?

- TWEANNs, Differentiable NAS, Neural Rejuvenation [Qiao et al., CVPR2019], ··· ?
- Eigenspectrum is related to both architectures and parameters

DENSO

Crafting the Core