Scaling and Benchmarking Self-Supervised Visual Representation Learning

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra

Facebook AI Research

2019

Presenter: Kobee Raveendran
Faculty: Dr. Yogesh S. Rawat
CAP6412, Spring 2020

Topic Overview

- Introduction
- Background Information
- Scaling Self-supervised Learning
- Domain Transfer
- Benchmarking Suite
- Conclusion

Introduction

- Supervised learning:

$$\min \frac{1}{N} \Sigma\, loss(X, Y)$$

Credit: digitweek

car

Introduction

- Datasets: ImageNet
  - 14+ million images
    - 1 million with bounding boxes
  - 20,000+ classes
    - 3,000 have bounding boxes
  - Human annotated via crowdsourcing

Introduction

- Supervised methods are data-inefficient
    - Requires an abundance of high-quality, labeled training data

Introduction

- Supervised methods are data-inefficient
  - Requires an abundance of high-quality, labeled training data
- This data can be hard to obtain

Introduction

- Supervised methods are data-inefficient
  - Requires an abundance of high-quality, labeled training data
- This data can be hard to obtain
  - Scraping is susceptible to noisiness

Introduction

- Supervised methods are data-inefficient
  - Requires an abundance of high-quality, labeled training data
- This data can be hard to obtain
  - Scraping is susceptible to noisiness
  - Some data sets require extensive domain expertise for proper labeling

Introduction

- Supervised methods are data-inefficient
  - Requires an abundance of high-quality, labeled training data
- This data can be hard to obtain
  - Scraping is susceptible to noisiness
  - Some tasks require extensive domain expertise for proper labeling
  - Expensive with respect to time and money

Introduction

- Semi-supervised
  - Partially labeled, partially unlabeled

Introduction

- Semi-supervised
  - Partially labeled, partially unlabeled
- Weakly-supervised
  - Coarse-grained labels

Credit: Jisoo Jeong, Seungeu Lee, Jeesoo Kimm, and Nojun Kwak. *Consistency-based Semi-supervised Learning for Object Detection*

Introduction

- Semi-supervised
  - Partially labeled, partially unlabeled
- Weakly-supervised
  - Coarse-grained labels
- Unsupervised
  - No labels

Credit: Jisoo Jeong, Seungeu Lee, Jeesoo Kimm, and Nojun Kwak. *Consistency-based Semi-supervised Learning for Object Detection*

Introduction

- Self-supervised learning
  - Benefits from the availability of unlabeled data

Introduction

- Self-supervised learning
  - Benefits from the availability of unlabeled data
  - Pretext tasks
    - Ground truth can be derived from the attributes of the input itself

Introduction

- Self-supervised learning
  - Benefits from the availability of unlabeled data
  - Pretext tasks
    - Ground truth can be derived from the attributes of the input itself
  - Downstream tasks

Introduction

Rotation

Credit: Shin'ya Yamaguchi, Sekitoshi Kanai, Tetsuya Shioda, Shoichiro Takeda. *Multiple Pre-text Task for Self-Supervised Learning via Mixing Multiple Image Transformations*

UCF

Introduction

Inpainting

Credit: Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S. Huang. *Generative Image Inpainting with Contextual Attention*

Introduction

- Previous methods haven’t yet capitalized on the scalability of unlabeled data
    - Confined to the scale of ImageNet

Introduction

- Previous methods haven’t yet capitalized on the scalability of unlabeled data
  - Confined to the scale of ImageNet
- Scaling along multiple axes:
  - Data set size

Introduction

- Previous methods haven’t yet capitalized on the scalability of unlabeled data
  - Confined to the scale of ImageNet
- Scaling along multiple axes:
  - Data set size
  - Network capacity

Introduction

- Previous methods haven’t yet capitalized on the scalability of unlabeled data
  - Confined to the scale of ImageNet
- Scaling along multiple axes:
  - Data set size
  - Network capacity
  - Pretext problem complexity

Introduction

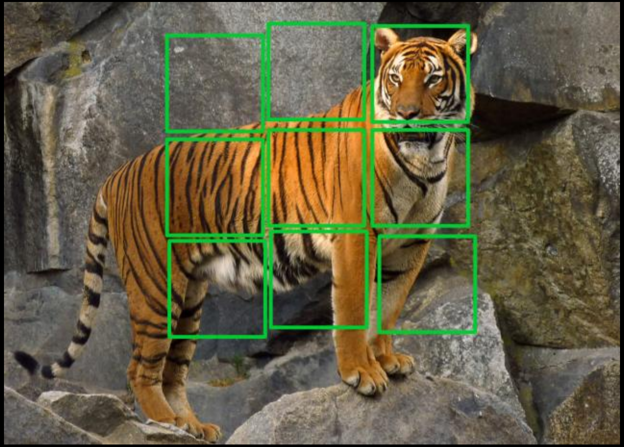
- Benchmarking suite for representation evaluation
- Good methods should:
  - Generalize to a variety of tasks
  - Require little to no supervision and fine-tuning

Introduction

- Pretext tasks
  - Multi-modal
    - i.e. autonomous vehicles sensor fusion for perception, videos with sound, etc.
  - Visual only

Background Information

- Pretext tasks: Jigsaw puzzle

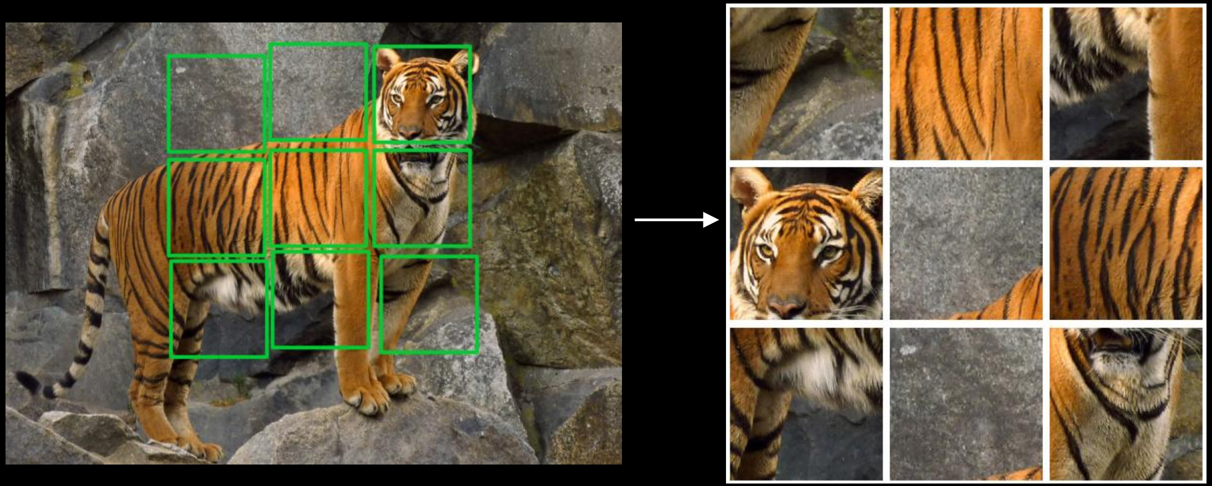


Credit: Mehdi Noroozi and Paolo Favaro. *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*

UCF

Background Information

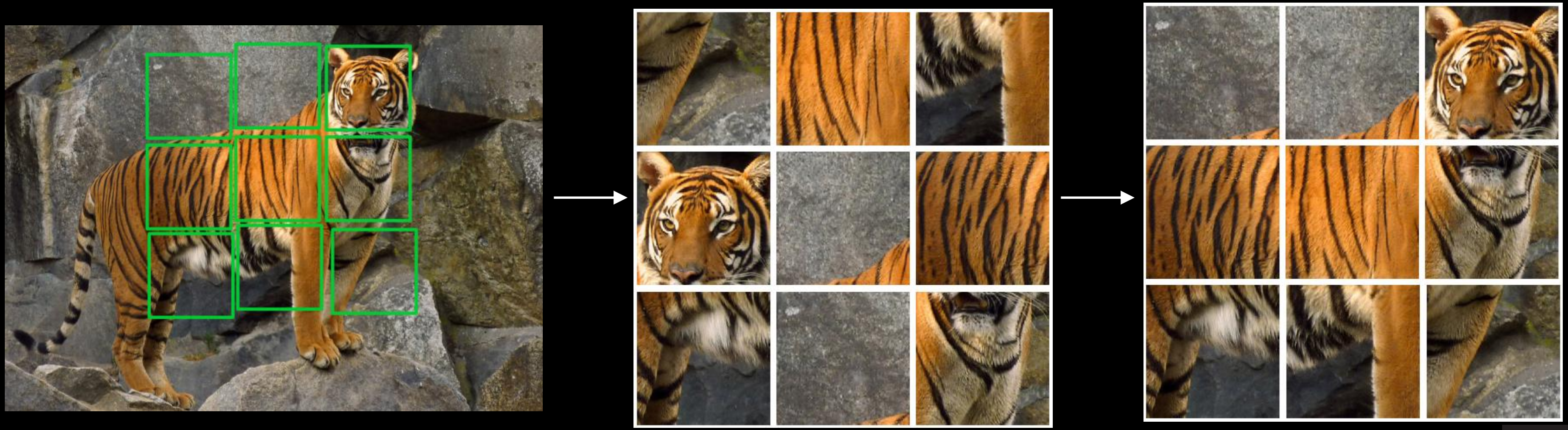
- Pretext tasks: Jigsaw puzzle



Credit: Mehdi Noroozi and Paolo Favaro. *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*

Background Information

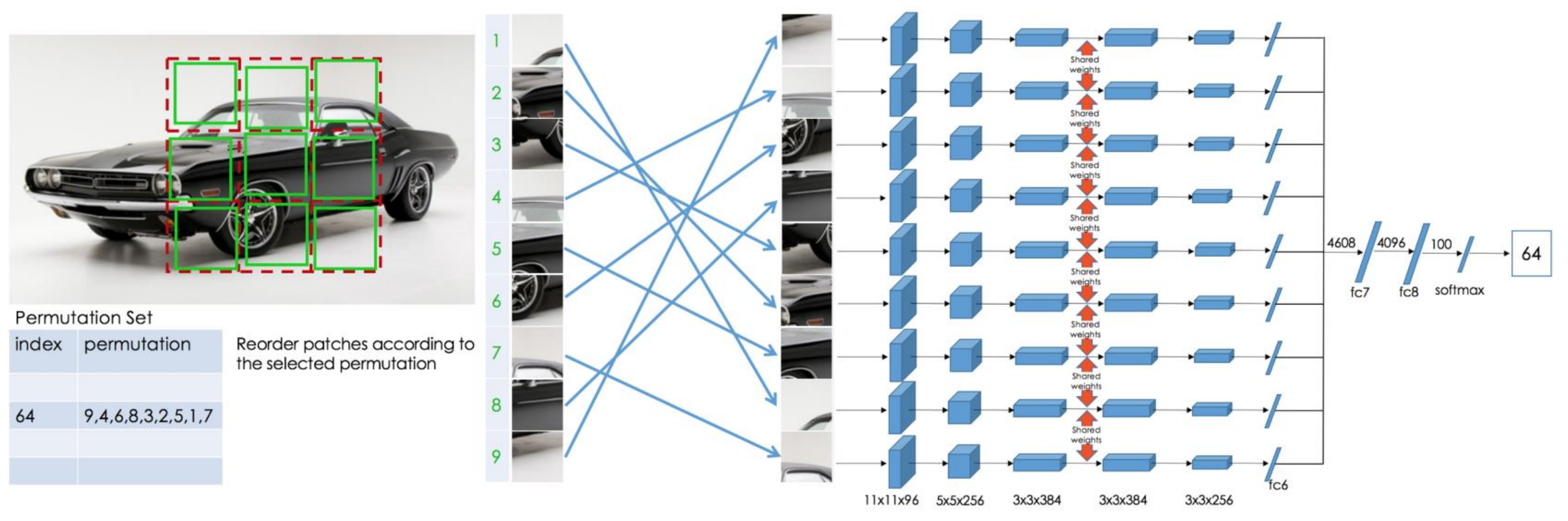
- Pretext tasks: Jigsaw puzzle



Credit: Mehdi Noroozi and Paolo Favaro. *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*

Background Information

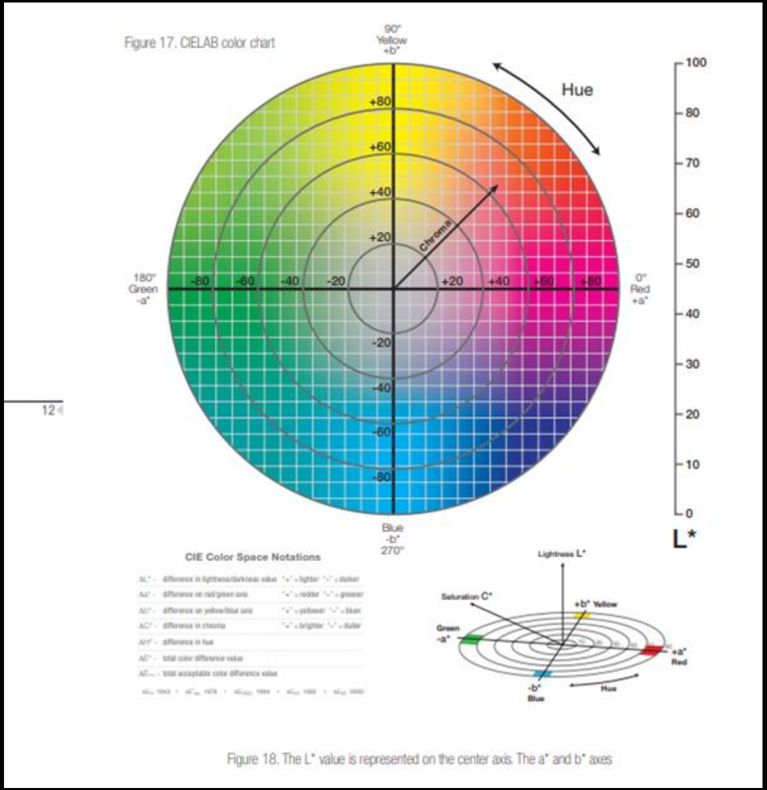
- Self-supervised learning on Jigsaw
  - *N*-way Siamese network



Credit: Mehdi Noroozi and Paolo Favaro. *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*

Background Information

- Lab color space



Credit: https://www.xrite.com/blog/lab-color-space

Background Information

- Pretext tasks: Image colorization

Credit: Richard Zhang, Philip Isola, and Alexei A. Efros. *Colorful Image Colorization*

Background Information

- Hard- vs. soft-encoding
  - Y = [0, 0, 0, 1, 0]

Background Information

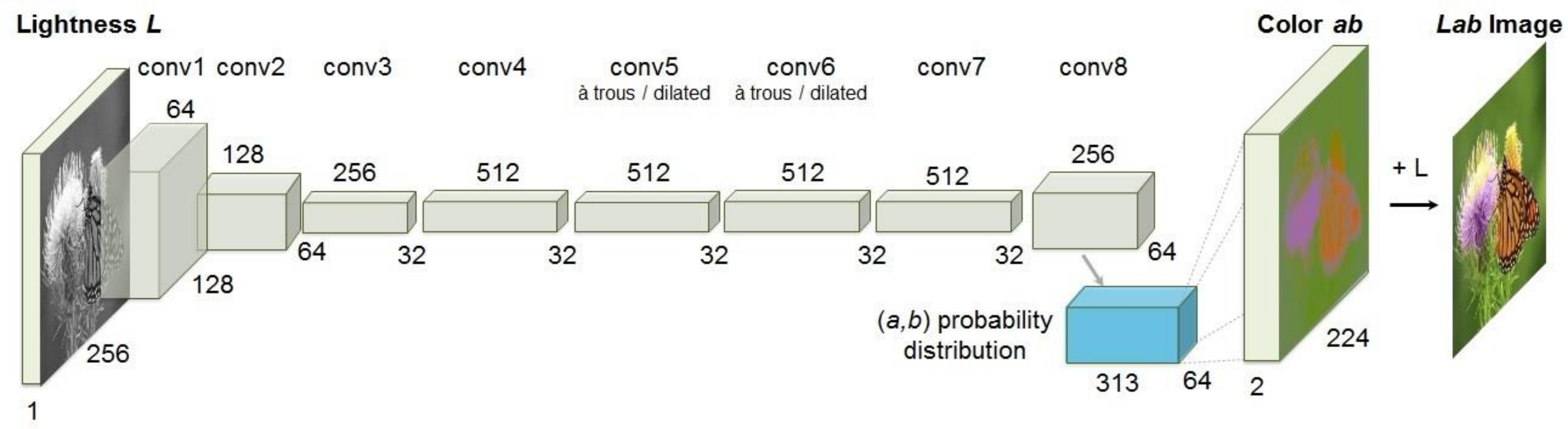
- Hard- vs. soft-encoding
    - Y = [0, 0, 0, 1, 0]
    - What if we don't need to exactly match the GT?
        - Multiple correct answers:
            - Y = [0, 1, 0, 0, 0] (turquoise)
            - Y = [0, 0, 1, 0, 0] (cyan)
            - Y = [0, 0, 0, 1, 0] (light blue)

Background Information

- Hard- vs. soft-encoding
  - Y = [0, 0, 0, 1, 0]
  - What if we don't need to exactly match the GT?
    - Multiple correct answers:
      - Y = [0, 1, 0, 0, 0] (turquoise)
      - Y = [0, 0, 1, 0, 0] (cyan)
      - Y = [0, 0, 0, 1, 0] (sky blue)
    - Instead: Y = [0, 0.33, 0.33, 0.33, 0]

Background Information

- Self-supervised learning on Colorization



Credit: Richard Zhang, Philip Isola, and Alexei A. Efros. *Colorful Image Colorization*

Scaling Self-Supervised Learning

- Scaling along three axes:
  - Data set size
  - Model capacity
  - Pretext task complexity
- Setup
  - Train linear SVMs on output of CNN
  - YFCC-100M dataset for self-supervised pre-training
    - 99.2 million images
    - 0.8 million videos
  - Transfer learning for image classification on VOC 2007 data set

Scaling Self-Supervised Learning

- Setup
  - YFCC-100M dataset for self-supervised pre-training
    - 99.2 million images
    - 0.8 million videos
  - Train linear SVMs on output of CNNs from pretext
  - Transfer learning for image classification on VOC 2007 data set

Scaling Self-Supervised Learning

- Scaling data set size
  - Training on multiple randomly-sampled subsets of YFCC-100M
    - 1, 10, 50, and the full 100 million images
    - Problem complexity is kept constant
      - |P| = 2000
      - K = 10
    - Size variations also tested on both AlexNet and ResNet-50

Scaling Self-Supervised Learning

- Scaling model capacity
  - Trained on shallow and deep models
    - AlexNet and ResNet-50
    - Problem complexity is kept constant
      - |P| = 2000
      - K = 10
    - Tested for each subset of the full data shown on the previous slide

Scaling Self-Supervised Learning

- Scaling problem complexity
  - Jigsaw
    - Tested various configurations of the number of permutations of each puzzle
      - {100, 701, 2000, 5000, 10000}
  - Image colorization
    - Tested on various values of K
      - {2, 5, 10, 20, 40, 80, 160, 313}
  - Data set size kept constant at 1 million images
  - Evaluated on both AlexNet and ResNet-50

Scaling Self-Supervised Learning

- Scaling problem complexity
  - Jigsaw
    - Tested various configurations of the number of permutations of each puzzle
      - {100, 701, 2000, 5000, 10000}

Scaling Self-Supervised Learning

- Scaling problem complexity
  - Jigsaw
    - Tested various configurations of the number of permutations of each puzzle
      - {100, 701, 2000, 5000, 10000}
  - Image colorization
    - Tested on various values of K
      - {2, 5, 10, 20, 40, 80, 160, 313}

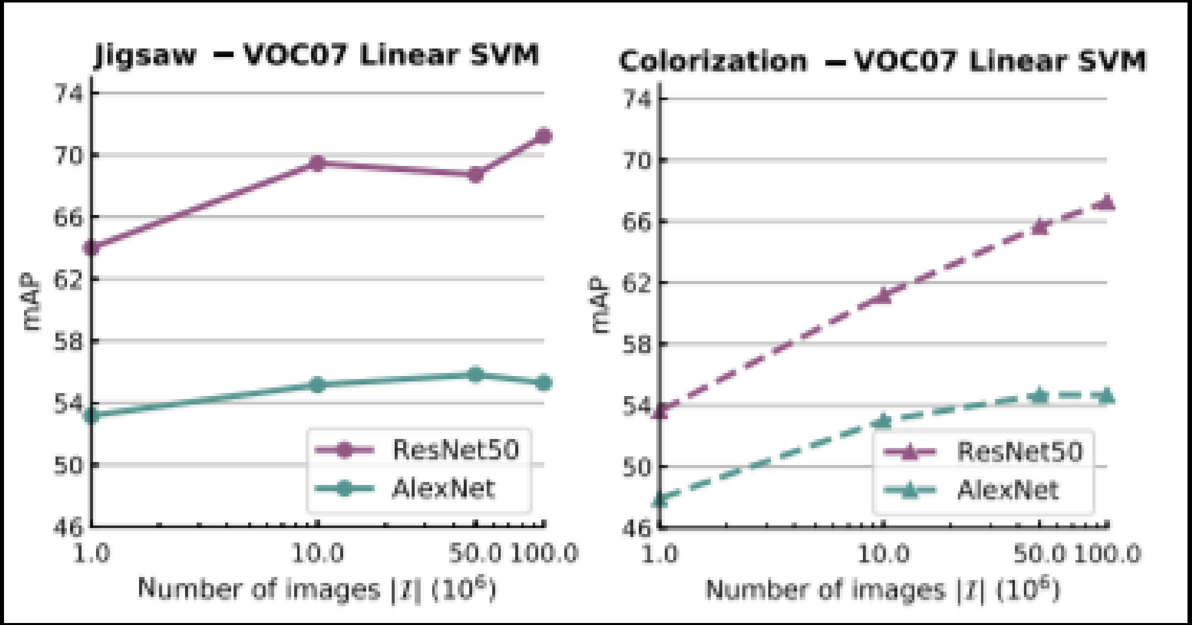
UCF

Scaling Self-Supervised Learning

- Scaling problem complexity
  - Jigsaw
    - Tested various configurations of the number of permutations of each puzzle
      - {100, 701, 2000, 5000, 10000}
  - Image colorization
    - Tested on various values of K
      - {2, 5, 10, 20, 40, 80, 160, 313}
  - Data set size kept constant at 1 million images
  - Evaluated on both AlexNet and ResNet-50

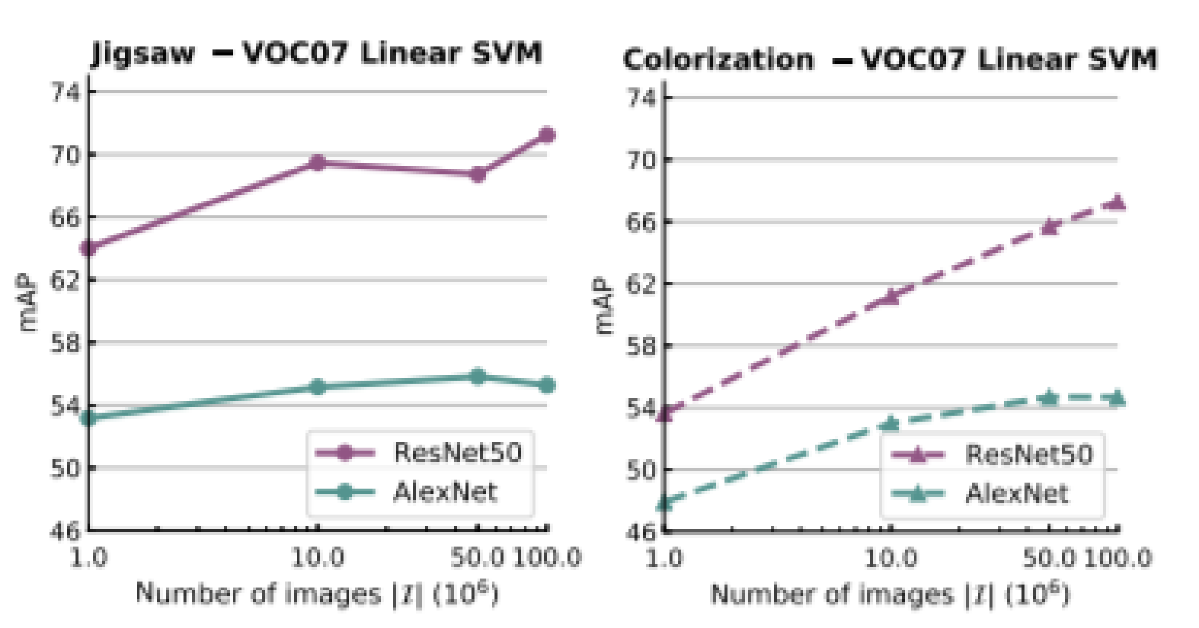
Scaling Self-Supervised Learning

- Scaling data set size: Observations



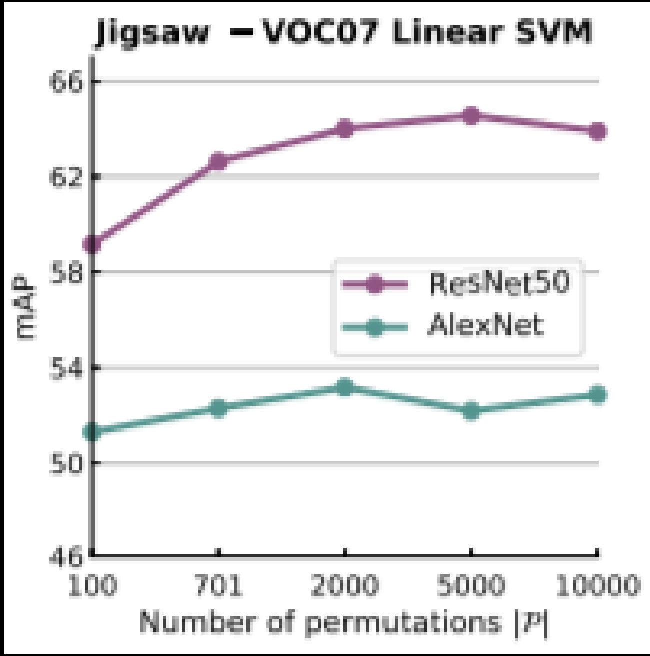
Scaling Self-Supervised Learning

- Scaling model capacity: Observations



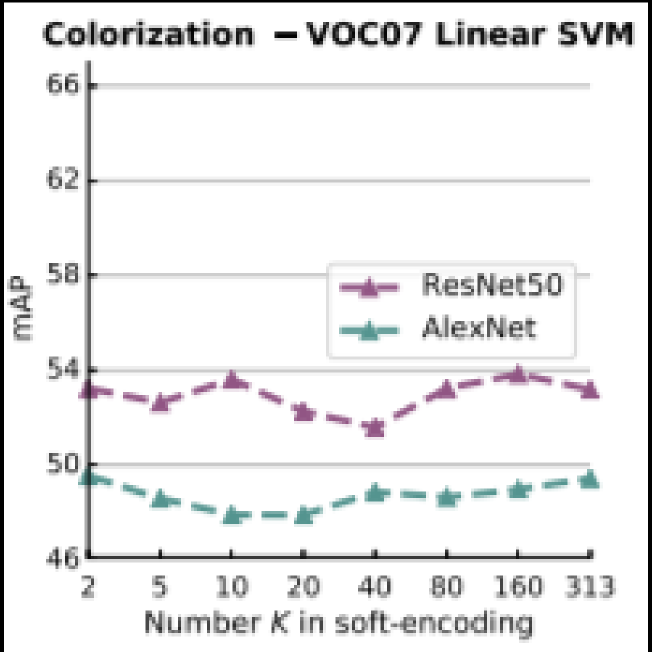
Scaling Self-Supervised Learning

- Scaling problem complexity: Observations
  - Jigsaw



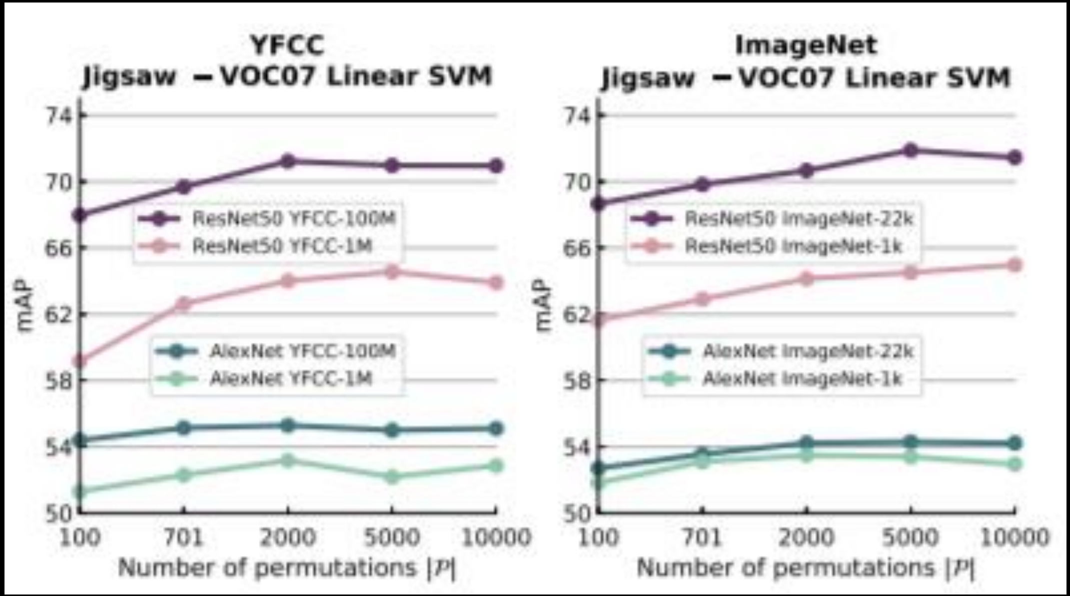
Scaling Self-Supervised Learning

- Scaling problem complexity: Observations
  - Image colorization



Scaling Self-Supervised Learning

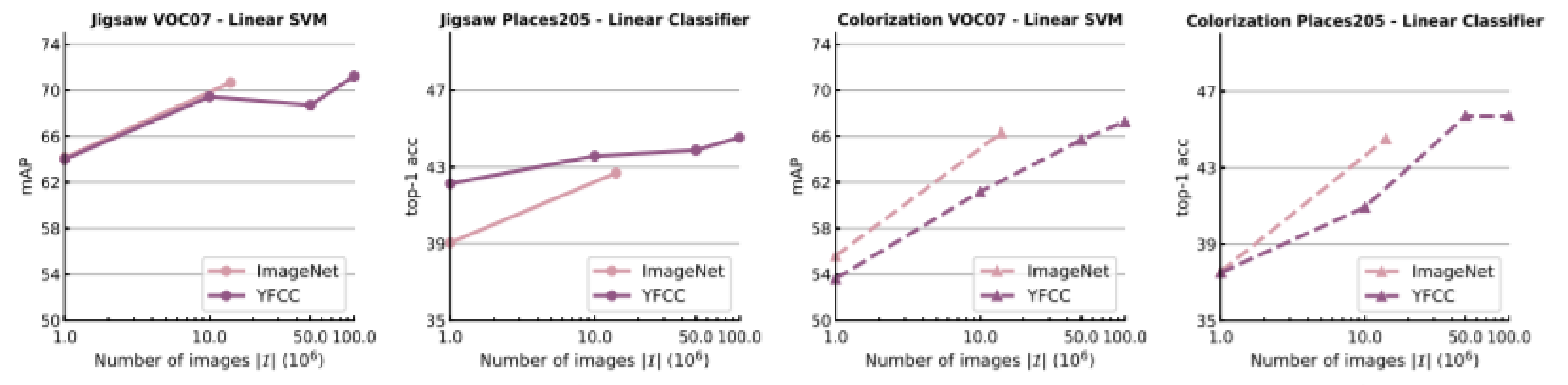
- Comprehensive Observations
  - The three scaled axes complement each other



UCF

Domain Transfer

- Effects of pre-training and transfer learning domain differences



UCF

Benchmarking Suite for Self-Supervision

- Based on the premise that good representations should:
  - Generalize to a diverse set of tasks
  - Require limited supervision and fine-tuning

Benchmarking Suite for Self-Supervision

- Setup:
  - Self-supervised pre-training
    - Jigsaw or Colorization
    - YFCC-xM, ImageNet-1k, or ImageNet-22k

Benchmarking Suite for Self-Supervision

- Setup:
  - Self-supervised pre-training
    - Jigsaw or Colorization
    - YFCC-xM, ImageNet-1k, or ImageNet-22k
  - Feature extraction from multiple layers of the CNN
    - AlexNet: following every convolution layer
    - ResNet-50: final layer of every residual block

Benchmarking Suite for Self-Supervision

- Setup:
  - Self-supervised pre-training
    - Jigsaw or Colorization
    - YFCC-xM, ImageNet-1k, or ImageNet-22k
  - Feature extraction from multiple layers of the CNN
    - AlexNet: following every convolution layer
    - ResNet-50: final layer of every residual block
  - Transfer learning on fully-supervised data sets and tasks

Benchmarking Suite for Self-Supervision

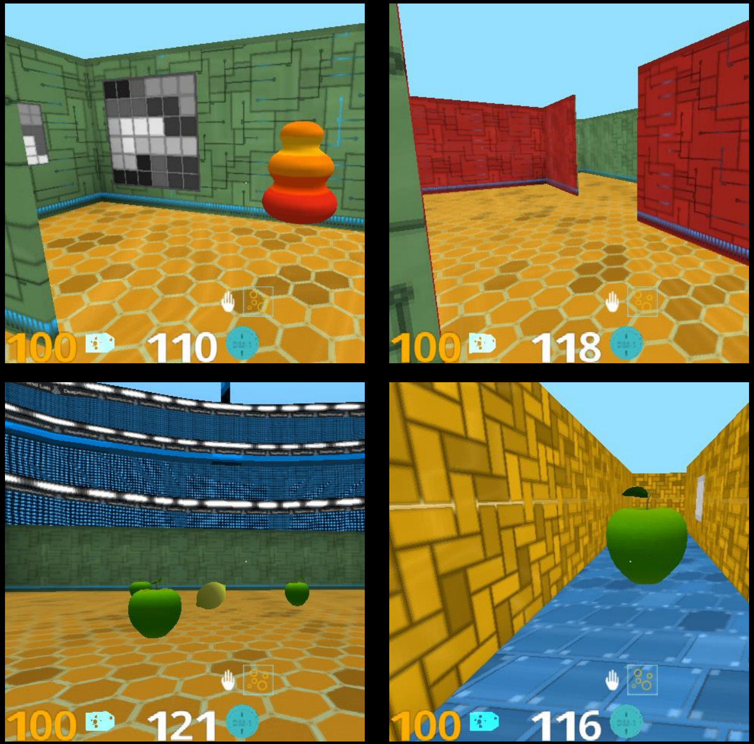
- The benchmarking suite evaluates on multiple downstream tasks:
  - Image classification
  - Low-shot image classification
  - Visual navigation
  - Object detection
  - Surface normal estimation

Benchmarking Suite for Self-Supervision

- Low-shot image classification
    - Something we’re already somewhat familiar with!

Benchmarking Suite for Self-Supervision

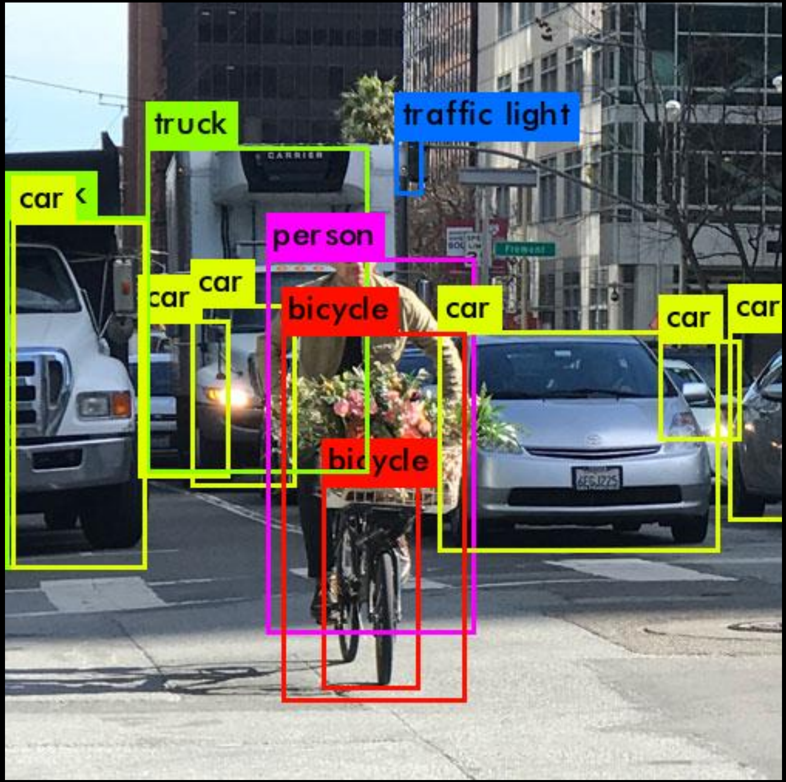
- Visual navigation



Credit: Jonas Kulhanek, Erik Derner, Tim de Bruin , and Robert Babuska. *Vision-based Navigation using Deep Reinforcement Learning*

Benchmarking Suite for Self-Supervision

- Object detection



Credit: https://towardsdatascience.com/r-cnn-3a9beddfd55a

UCF

Benchmarking Suite for Self-Supervision

- Surface normal estimation

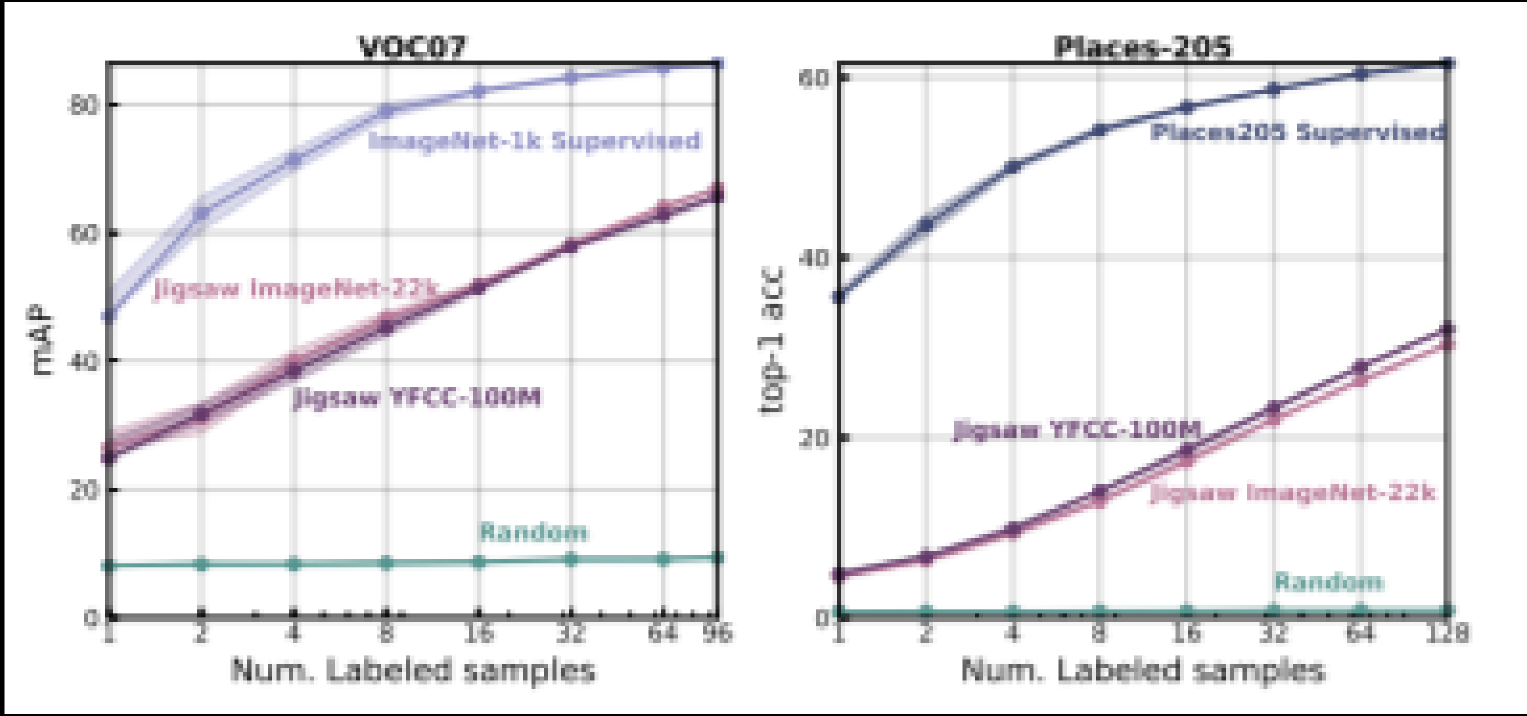
Credit: Xiaolong Wang, David F. Fouhey, and Abhinav Gupta. *Designing Deep Networks for Surface Normal Estimation*

Results

Method	layer1	layer2	layer3	layer4	layer5
ResNet-50 ImageNet-1k Supervised	24.5	47.8	60.5	80.4	88.0
ResNet-50 Places205 Supervised	28.2	46.9	59.1	77.3	80.8
ResNet-50 Random	9.6	8.3	8.1	8.0	7.7
ResNet-50 Jigsaw ImageNet-1k	**27.1**	45.7	56.6	64.5	57.2
ResNet-50 Jigsaw ImageNet-22k	20.2	**47.7**	57.7	**71.9**	**64.8**
ResNet-50 Jigsaw YFCC-100M	20.4	47.1	**58.4**	71.0	62.5
ResNet-50 Coloriz. ImageNet-1k	24.3	40.7	48.1	55.6	52.3
ResNet-50 Coloriz. ImageNet-22k	25.8	43.1	53.6	66.1	62.7
ResNet-50 Coloriz. YFCC-100M	26.1	42.3	53.8	67.2	61.4

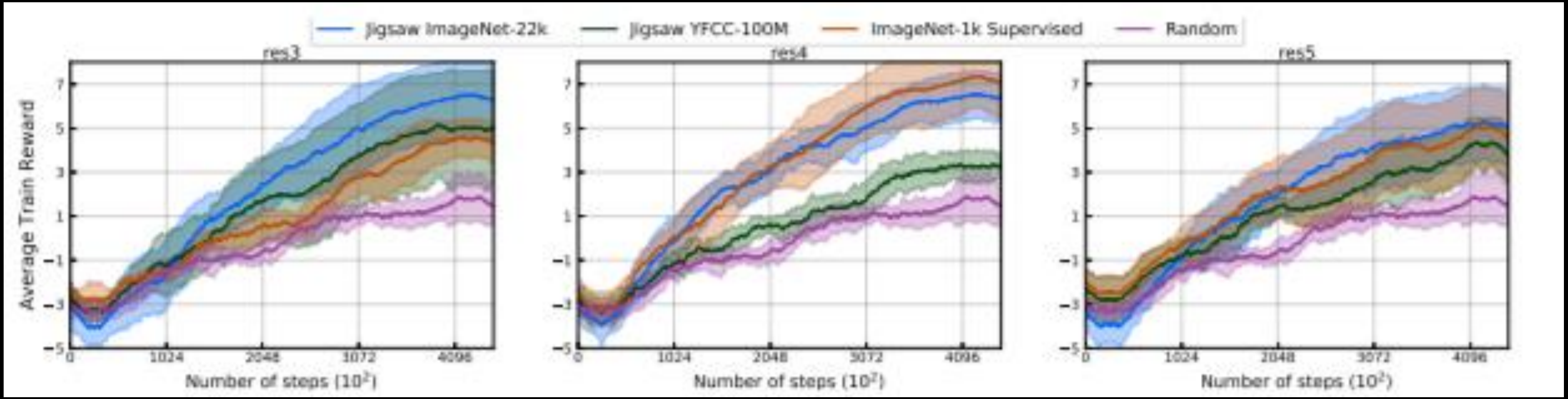
Image classification

Results



Low-shot Image classification

Results



Visual Navigation

Results

Surface Normal Estimation

Initialization	Angle Distance Mean	Angle Distance Median	Within t° 11.25	Within t° 22.5	Within t° 30
	(Lower is better)		(Higher is better)		
ResNet-50 ImageNet-1k supervised	26.4	17.1	36.1	59.2	68.5
ResNet-50 Places205 supervised	23.3	14.2	41.8	65.2	73.6
ResNet-50 Scratch	26.3	16.1	37.9	60.6	69.0
ResNet-50 Jigsaw ImageNet-1k	24.2	14.5	41.2	64.2	72.5
ResNet-50 Jigsaw ImageNet-22k	22.6	13.4	43.7	66.8	74.7
ResNet-50 Jigsaw YFCC-100M	**22.4**	**13.1**	**44.6**	**67.4**	**75.1**

Object Detection

Method	VOC07	VOC07+12
ResNet-50 ImageNet-1k Supervised*	66.7 ± 0.2	71.4 ± 0.1
ResNet-50 ImageNet-1k Supervised	$\mathbf{68.5} \pm 0.3$	$\mathbf{75.8} \pm 0.2$
ResNet-50 Places205 Supervised	65.3 ± 0.3	73.1 ± 0.3
ResNet-50 Jigsaw ImageNet-1k	56.6 ± 0.5	64.7 ± 0.2
ResNet-50 Jigsaw ImageNet-22k	$\mathbf{67.1} \pm 0.3$	$\mathbf{73.0} \pm 0.2$
ResNet-50 Jigsaw YFCC-100M	62.3 ± 0.2	69.7 ± 0.1

UCF

Conclusion

- Scaling self-supervised methods along three axes (data size, model capacity, and problem complexity) noticeably improves transfer learning performance
- Scaling along each axis complements the others

Conclusion

- Self-supervised representation learning can meet or exceed state-of-the-art *supervised* performance on some tasks
  - Surface normal estimation, visual navigation, object detection
- Falls short of supervised methods on other tasks
  - Image classification, low-shot image classification