Transfer Learning with Self-Supervised Vision Transformer for Large-Scale Plant Identification

Mingle Xu[1,2], Sook Yoon[3], Yongchae Jeong[1], Jaesu Lee[4] and Dong Sun Park[1,2]

Jeonbuk National University, South Korea
Core Research Institute of Intelligent Robots
Multimedia Lab Under Prof. Park and Prof. Yoon
Mingle Xu

Sep 8, 2022
Bologna, Italy

Road Map

- A glimpse of PlantCLEF2022 challenge
- Our analysis of PlantCLEF2022 dataset
- Our strategy and results
- Extra experiments
- Summary

A glimpse of PlantCLEF2022

1 Global-scale plant identification.

Trusted: 80k classes, 2.9M images

Web: 57k classes, 1.1M images

2 Observation-level image classification.

Testing dataset: 26,868 observations with 55,306 images.

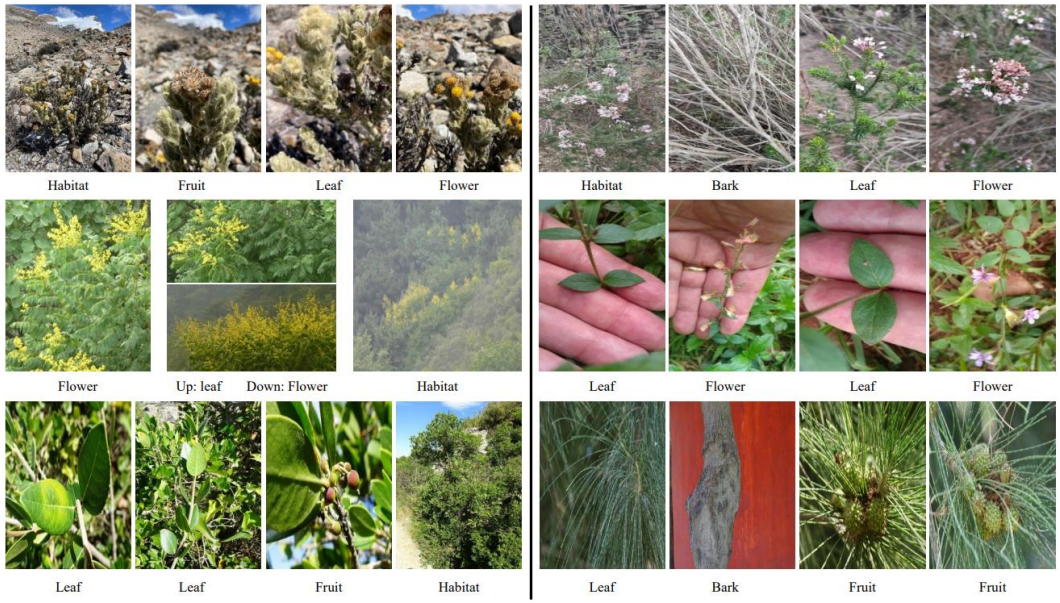


Figure 4: Six observations of testing dataset in PlantCLEF2022. One observation refers to an actual plant and we can take multiple images for single observation. The PlantCLEF2022 challenge requires

$$MA-MRR = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{O_n}\sum_{i=1}^{O_n}\frac{1}{rank_i},$$

3 Evaluation metric:

MA-MRR

Class 1 Score 1

Class 2 Score 2

…

Class 30 Score 30

Our analysis 1

A few-shot learning (FSL) task.

- PlantCLEF2022
  - Trusted: ~36 image/per class, 80k classes, 2.9M images
  - Web: ~19 image/per class, 57k classes, 1.1M images
- ImageNet-1k: ~1281 image/per class
- Flowers-102: 40-258 image/per class
- Places: >5k image/per class

Goëau, H., Bonnet, P. and Joly, A., 2022. Overview of PlantCLEF 2022: Image-based plant identification at global scale. *Working Notes of CLEF*, pp.1526-1539.

Our analysis 2

Huge image variations.

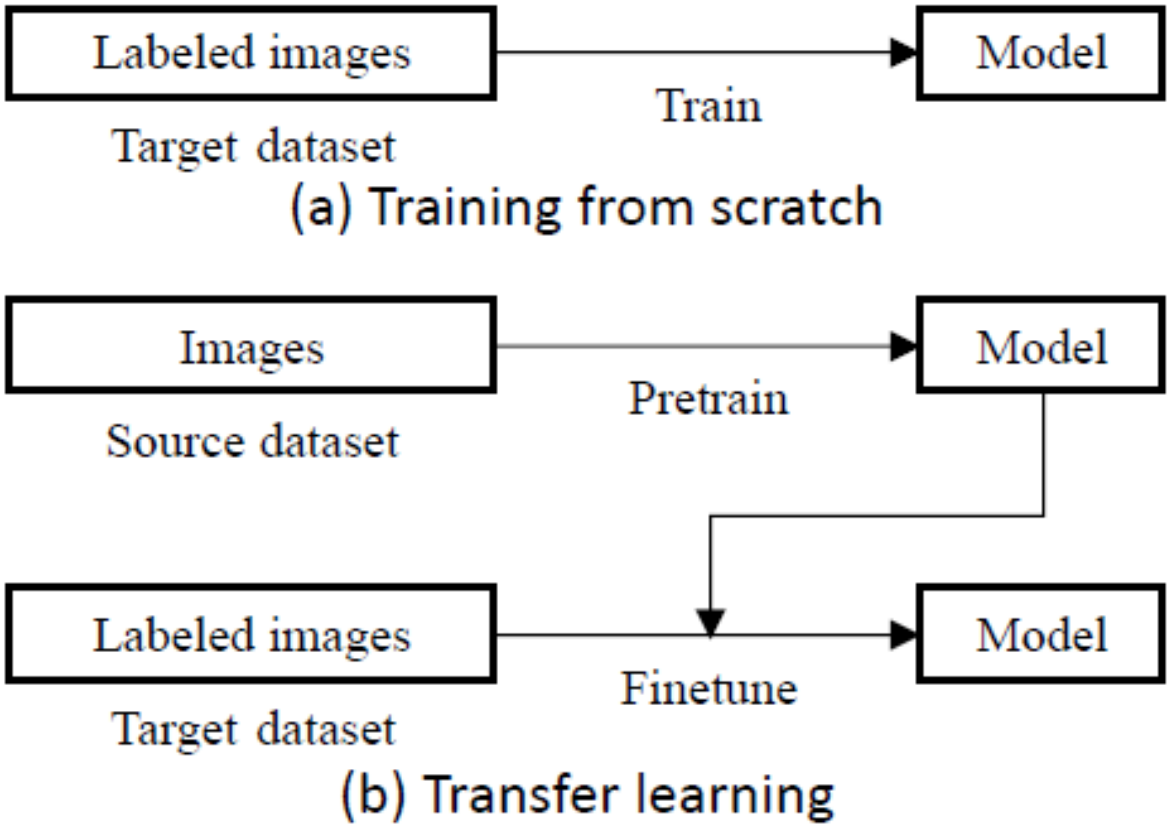
Background, plant organ, color, illumination, viewpoint, scale, …

Figure 2: Images of Cycas armstrongii Miq species from PlantCLEF2022 training dataset. The images from the same species are heterogeneous in background, viewpoint, and size.

Figure 3: Images of Aralia nudicaulis L. species from PlantCLEF2022 training dataset. The images from the same plant species are heterogeneous in background, illumination, and color.

Our Strategy: motivation

A few-shot learning (FSL) task.
Huge image variations.



Our Strategy: motivation

A few-shot learning (FSL) task.

Huge image variations.

Transfer learning

Better accuracy in ImageNet, higher transfer accuracy.

If the source dataset is far from the target dataset, supervised loss-based transfer accuracy may be low.

MAE

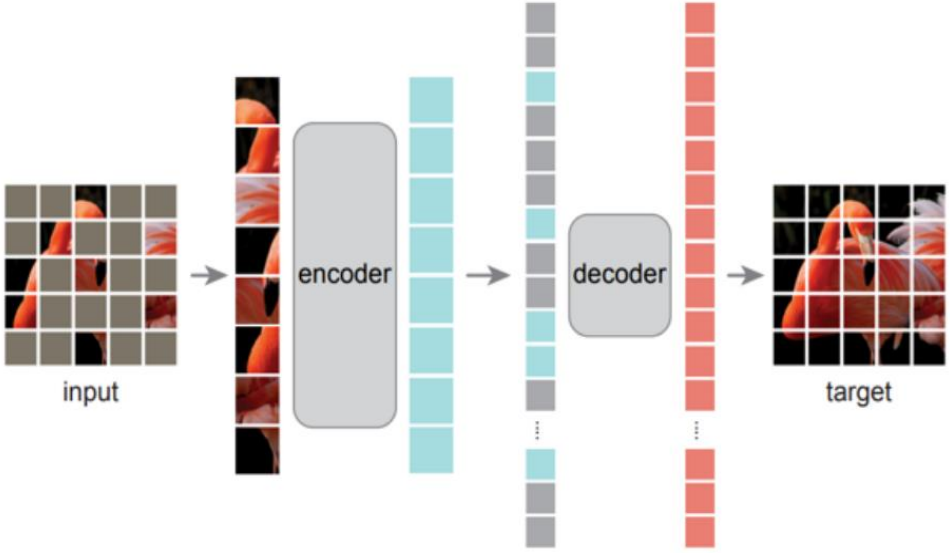
Kornblith, Simon, Jonathon Shlens, and Quoc V. Le. "Do better imagenet models transfer better?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
Kolesnikov, Alexander, et al. "Big transfer (bit): General visual representation learning." European conference on computer vision. Springer, Cham, 2020.
Wu, Zhirong, et al. "Unsupervised feature learning via non-parametric instance discrimination." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

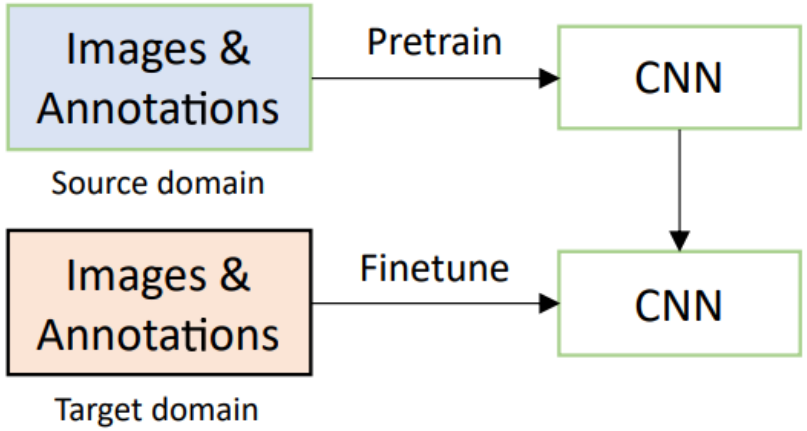
Our strategy

MAE (masked autoencoder)

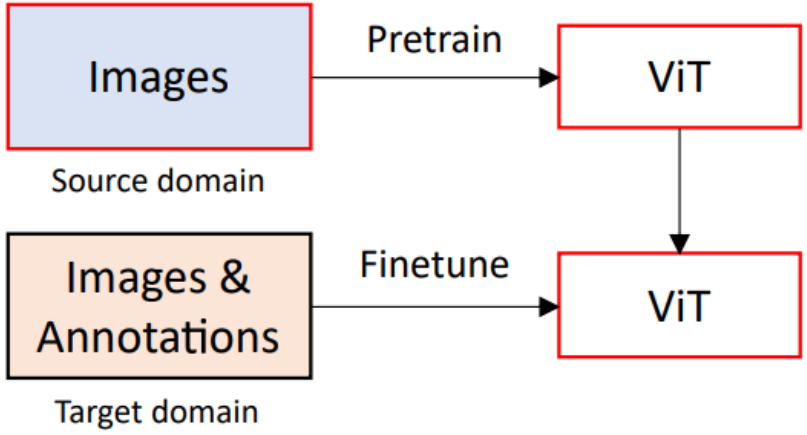
- High accuracy (ViT-based)
- Self-supervised loss

$$\mathcal{L} = ||input - target||$$





(a) Popular strategy with supervised CNN



(b) Our strategy with Self-supervised ViT

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R., 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16000-16009).

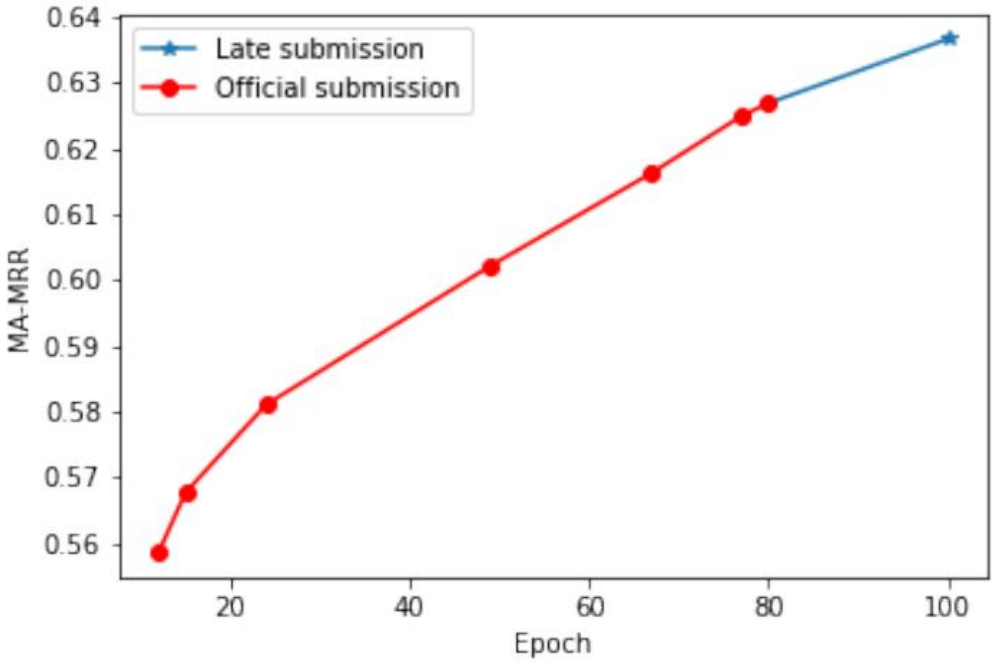
Our Result

Our strategy is simple but effective.

Huge computation. 4 RTX 3090 GPUs, ~ 20 days 100 epochs.

Core Research Institute of Intelligent Robots
Nation Research Foundation (NRF)

Team	MA-MRR
Ours	0.62692
Second place	0.60781
Third place	0.51043
Fourth place	0.46010

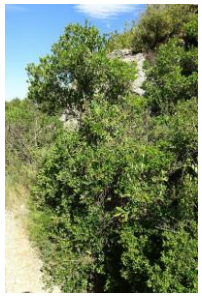


Our Strategy for Observation-Level

Observation-level recognition is one way to make deep learning-based models robust with high performance.

Observation-level classification.
Testing dataset: 26,868 observations with 55,306 images.



A 0.68
B 0.24

B 0.85
C 0.05

D 0.45
F 0.30

B 0.23
G 0.12

Observation-level strategy:
Single-random.
Single-highest.
Multi-sorted.

Single-highest.
B 0.85
C 0.05

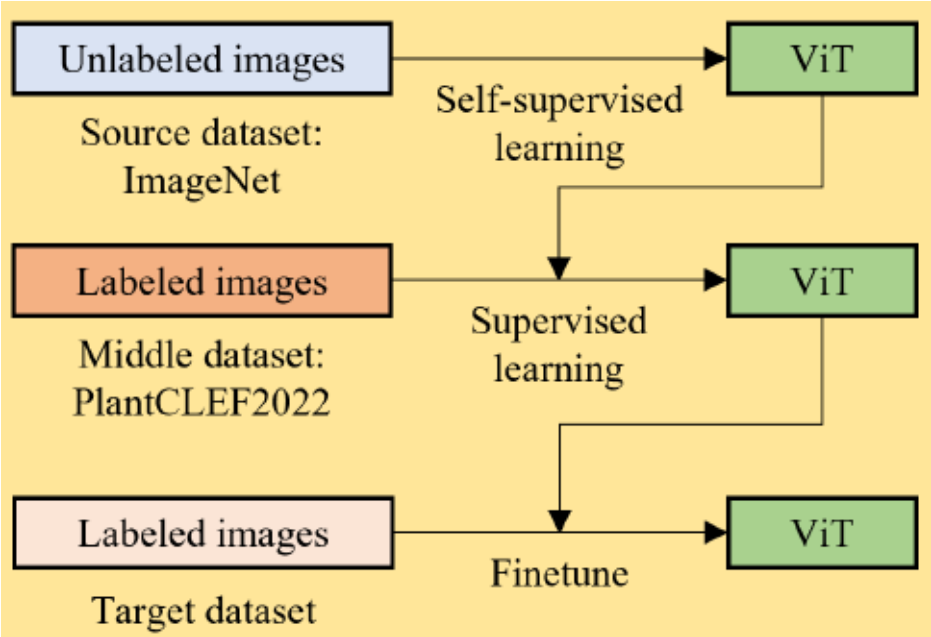
Multi-sorted.
B 0.85
A 0.68

Epoch	Single-highest	Multi-sorted
80	0.62692	No
100	0.63668	0.64079

Extra Experiments

Plant-related task.
Huge image variations.

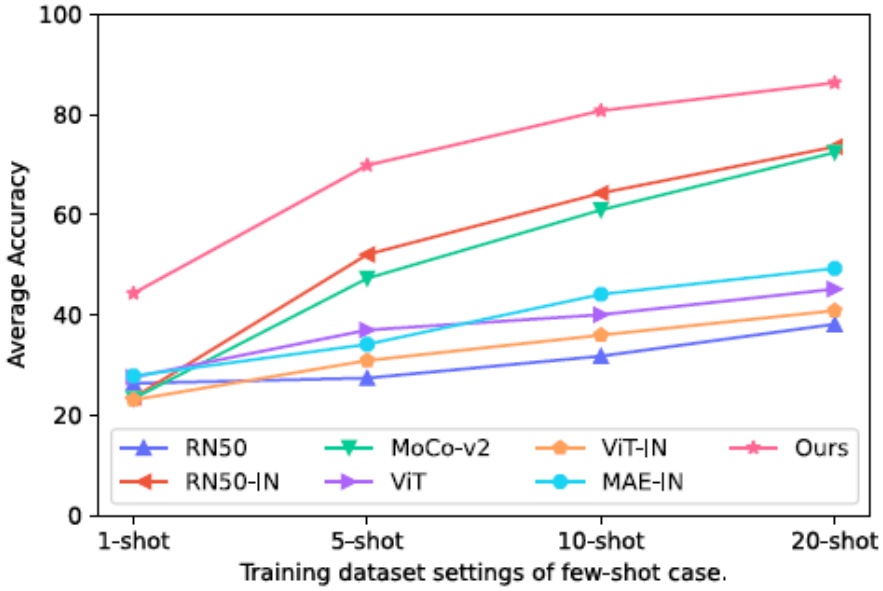
Good for plant-related task with limited data

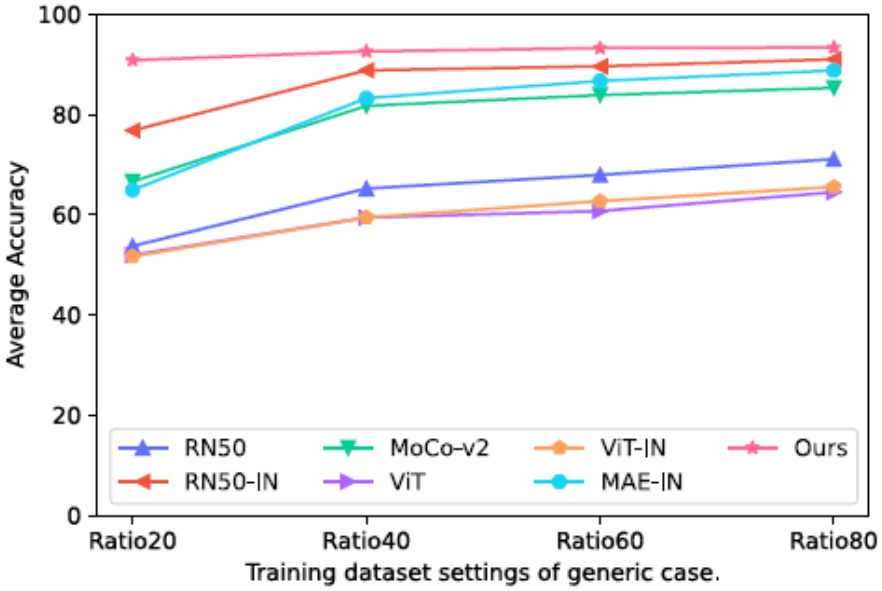


- Disease recognition: 12 datasets
- Growth stage recognition: 1 dataset
- Weed species recognition: 1 dataset

Extra Experiments: Plant disease

	1-shot	5-shot	10-shot	20-shot	Ratio20	Ratio40	Ratio60	Ratio80
RN50	26.33	27.38	31.75	38.13	53.71	65.19	67.91	71.07
RN50-IN	23.46	52.03	64.28	73.53	76.77	88.78	89.58	90.97
MoCo-v2	23.28	47.27	60.93	72.38	66.58	81.68	83.84	85.28
ViT	27.56	36.96	40.01	45.14	51.93	59.40	60.71	64.46
ViT-IN	23.02	30.87	35.94	40.83	51.64	59.42	62.67	65.53
MAE	27.81	34.11	44.08	49.26	64.90	83.23	86.65	88.76
Ours	**44.28**	**69.83**	**80.73**	**86.29**	**90.79**	**92.55**	**93.23**	**93.34**





Extra Experiments: beyond plant disease

Our pretrained model is helpful for other plant-related tasks.

	1-shot	5-shot	10-shot	20-shot	Ratio20	Ratio40	Ratio60	Ratio80
RN50	20.50	21.75	26.45	35.95	39.90	68.90	66.90	78.25
RN50-IN	45.55	75.95	87.90	87.15	60.85	98.00	98.35	98.55
MoCo-v2	45.65	70.25	84.65	86.05	66.90	96.45	96.20	97.50
ViT	32.70	39.90	44.30	51.45	56.25	65.65	75.40	80.90
ViT-IN	27.20	33.35	43.10	45.25	55.05	68.30	75.50	82.35
MAE	17.45	41.45	59.50	59.20	85.20	97.80	98.35	98.75
Ours	**73.90**	**97.60**	**97.55**	**97.85**	**99.80**	**99.35**	**98.80**	**99.70**

Summary: simple yet effective

PlantCLEF2022

A few-shot learning (FSL) task.
Huge image variations.

Transfer learning

Better accuracy in ImageNet, higher transfer accuracy.

The difference between ImageNet and PlantCLEF.

PlantCLEF is plant-related

MAE

- ViT-based
- SSL

Plant-related tasks

Thank You

Email: xml@jbnu.ac.kr

Public pretrained model and code: GitHub

Mingle Xu