Adapting Vision Foundation Models for Plant Phenotyping

Feng Chen[1], Mario Valerio Giuffrida[2], Sotirios A. Tsaftaris[1]
[1] The University of Edinburgh, [2] The University of Nottingham
Feng.Chen@ed.ac.uk, Valerio.Giuffrida@nottingham.ac.uk, S.Tsaftaris@ed.ac.uk

ICCV23 PARIS
CVPPA Workshop

The Work at a Glance

Motivation

- Plant image analysis tools are often **used once and then forgotten**, because they are bespoke for specific scenarios.
- Is there a model that can be **reused for various tasks**?
  - **Foundation model**

Challenges

- **Domain shift** between the pre-trained source and plant data
- Fine-tuning foundation models requires a lot of **computational power**
  - We use **adapters** to solve this

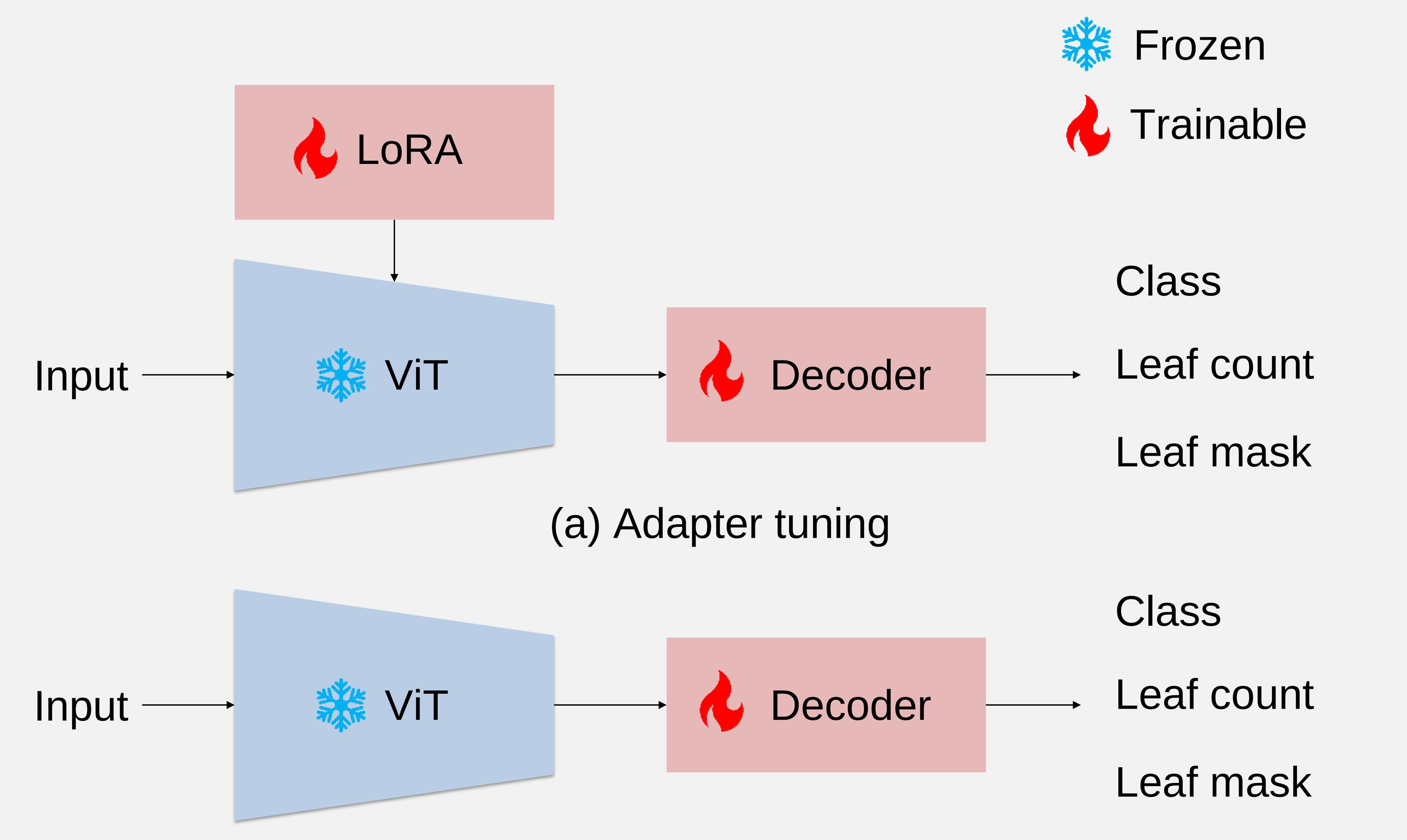
Contributions

- Benchmark the adaptation of 3 foundation models (**MAE** [1], **DINO** [2], **DINOv2** [3]) using 2 fine-tuning methods (**LoRA** [4], **decoder tuning**) on 3 plant tasks (**leaf counting, segmentation, disease classification**)

Conclusion

- Adapting a foundation model with LoRA to solve multiple plant tasks is promising (*e.g.* **MAE-LoRA** and **DINOv2-LoRA**)
- LoRA outperforms DT in most cases, except segmentation
- LoRA improves the performance in **low data regimes** and **class imbalance** (see original paper)
- The evaluated models may miss **small leaves/stems** in segmentation

Experimental Setup



(a) Adapter tuning

(b) Decoder tuning

Model

- **ViT-base** pre-trained using **MAE, DINO, DINOv2**

Adaptation

- Adapter tuning using **LoRA,** Decoder tuning (DT)

Tasks

- Leaf counting/segmentation (CVPPP: Arabidopsis/Tobacco)
- Leaf disease classification (Kaggle Cassava dataset)

Background

What is plant phenotyping?

- Measures the **observable features of plants**
- Indicates the growth and health of crops
- Helps develop better crops for extreme weather and food crisis

Minervini, Massimo, Hanno Scharr, and Sotirios A. Tsaftaris. "Image analysis: the new bottleneck in plant phenotyping [applications corner]." IEEE signal processing magazine 32, no. 4 (2015): 126-131.

What are foundation models?

- **Large** models (millions/billions of parameters)
- Pre-trained on **a huge amount of data**
- Able to adapt to **various new tasks**

What are adapters?

- **Light-weight trainable blocks added** to pre-trained models
- Not modify the pre-trained weights
- Low Rank Adaptation (LoRA): add trainable rank-decomposition weight matrics to each layer of Transformer

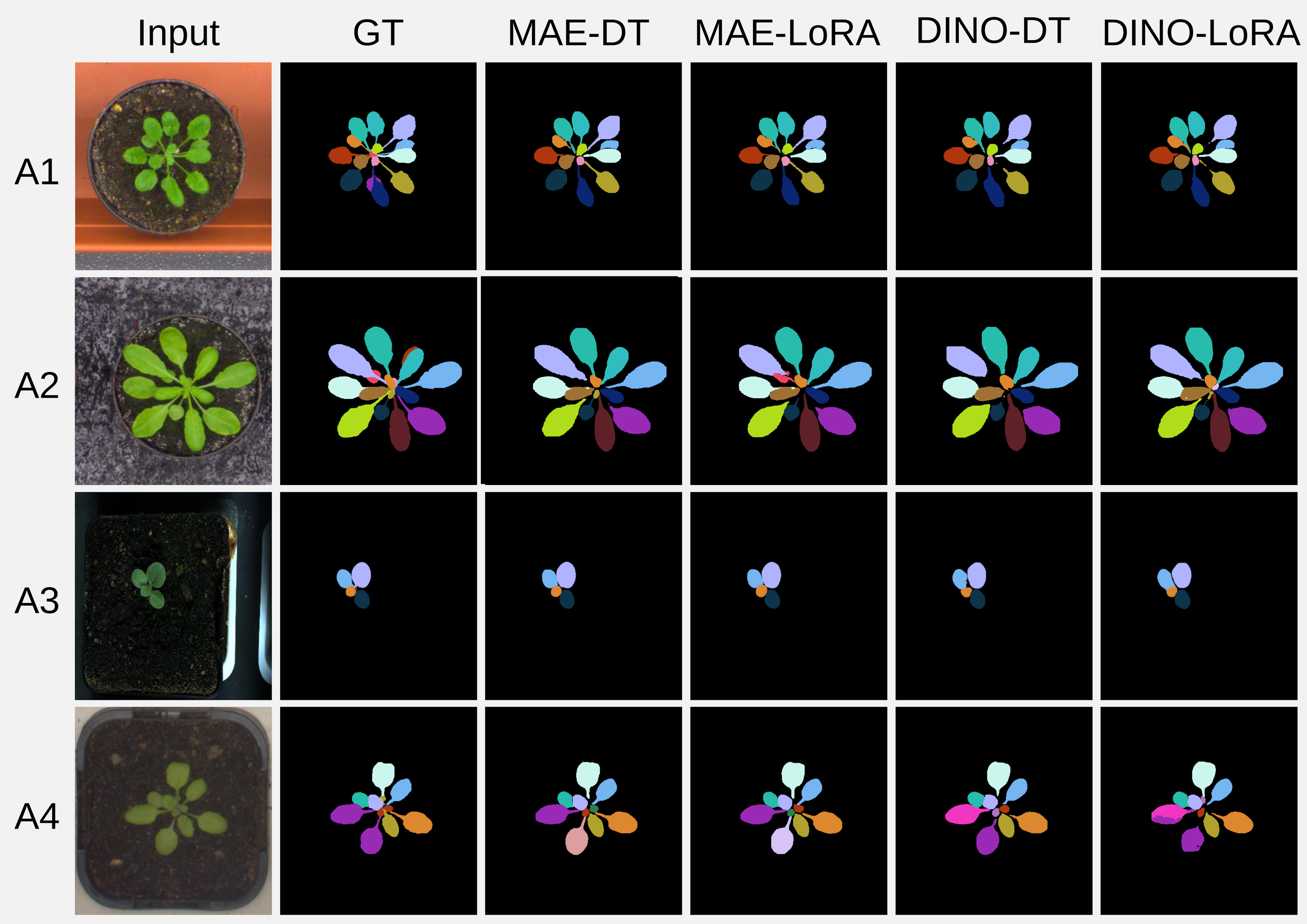
Results

- We compare the results of adapting different foundation models via LoRA and DT, **with the SoTA bespoke model** in each task.

Test results on three plant tasks

	Counting (MSE↓)	Segmentation (BestDice↑)	Classification (Acc [%] ↑)
SoTA	1.56	0.9	91.3
MAE-LoRA	1.79	0.87	88.8
DINOv2-LoRA	**1.63**	\	**89.7**
DINO-LoRA	1.88	0.82	89.0
MAE-DT	3.6	**0.88**	77.2
DINOv2-DT	1.92	\	86.1
DINO-DT	2.73	0.82	83.9

Segmentation results



Acknowledgement: This project was funded by the BBSRC grant BB/Y512333/1 “PhenomUKRI: The UK Plant and Crop Phenotyping Infrastructure”.

Key references:
[1] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." CVPR. 2022.
[2] Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." ICCV. 2021.
[3] Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." arXiv preprint arXiv:2304.07193 (2023).
[4] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).

Visit our lab (VIOS) @