Learning Features from Herbariums to Perform Automatic Classification of Field Images

Participation to LifeCLEF Plant Challenge 2020

Juan Villacis (jvillacis@ic-itcr.ac.cr)
Instituto Tecnológico de Costa-Rica - internship @ AMAP-lab Cirad Montpellier, France in collaboration with ZENITH team Inria Montpellier, France
Supervisors: Hervé Goëau, Pierre Bonnet, Alexis Joly & Erick Mata

18-06-2020

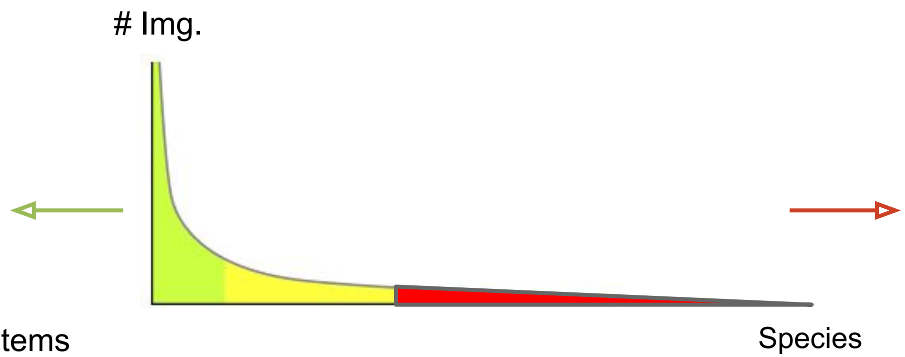
Objective

Train a machine learning model that can perform automatic classification on photos of tropical flora

nowadays automated systems perform well in **temperate regions**

- deep learning
- big data

Top1, PlantCLEF 2018 : 0,88



Plant biodiversity long tail distribution

but poorly in **tropical regions**:

- lack of data
- great visual diversity
- difficult access

Top1, PlantCLEF 2019 : 0,24

Objective

Train a machine learning model that can perform automatic classification on photos of tropical flora **with the help of herbarium collections**

-> potentially millions of underexploited digitized herbarium sheets (eReColNat, iDigBio)

-> performances? State of the art approaches based on photos in the field suitable?

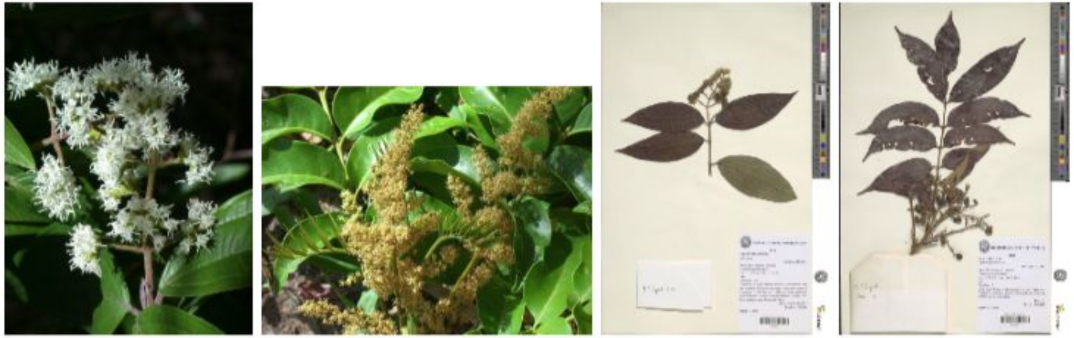


Fig. 1. Photos in the field and herbarium sheets of the same individual plant (Tapirira guianensis Aubl.). In spite of very different visual appearances between the two domains, similar structure and shapes of the flowers, fruits and leaves can be observed.

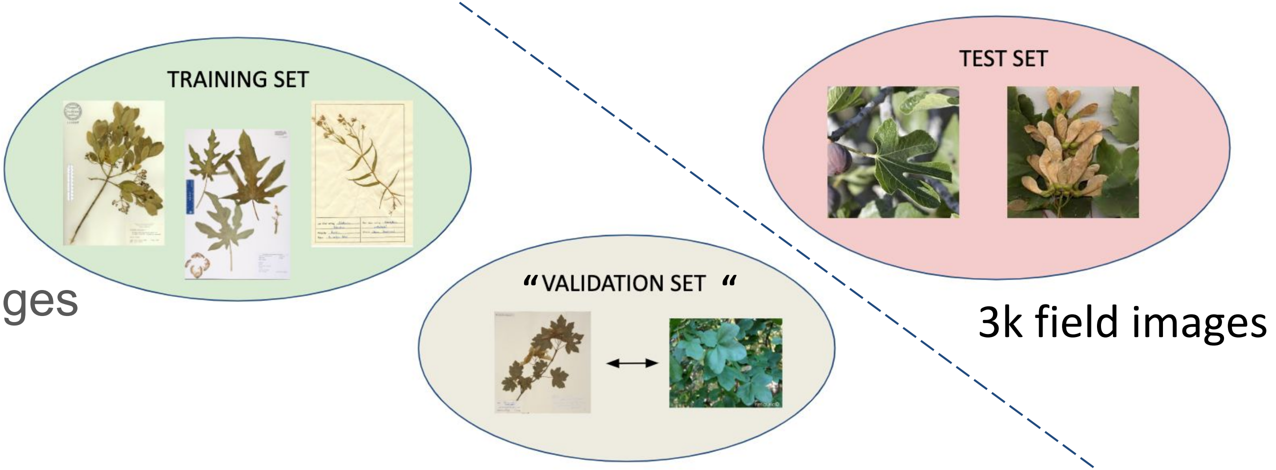
Dataset (PlantCLEF2020)

iDigBio
Integrated Digitized Biocollections

CAY L'HERBIER IRD DE GUYANE

330,752 sheets
997 species
+ 4,482 field images
from 375 sp

Cross-domain Plant Identification



3k field images

350 specimens associated with both herbarium sheet(s) and photo(s) in the field

Not enough field images to train an adequate model

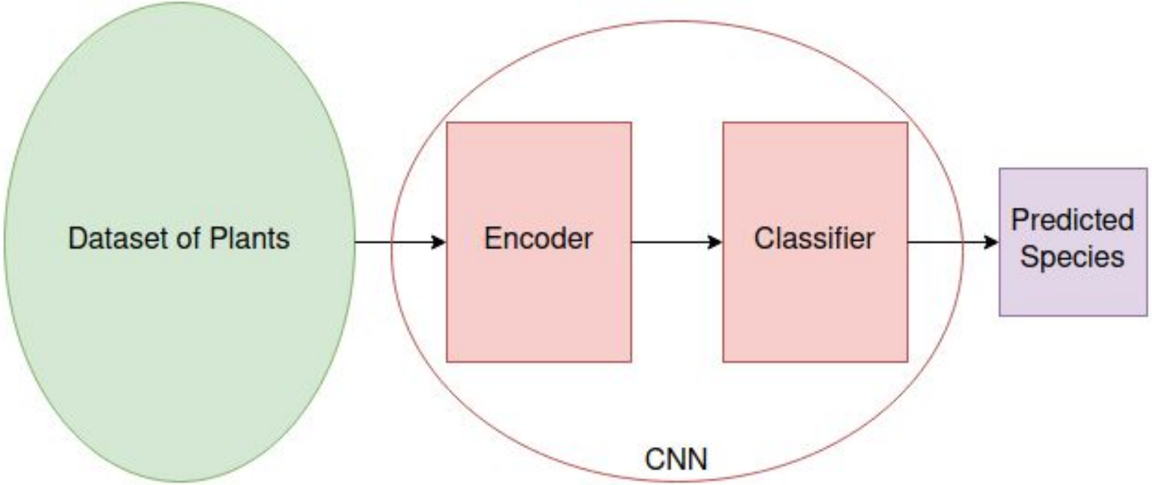
Will try to take advantage of existing herbarium images to compensate for missing data

1. Use a Convolutional Neural Network (CNN)

Use the dataset provided and train a model (ResNet50) in three stages:

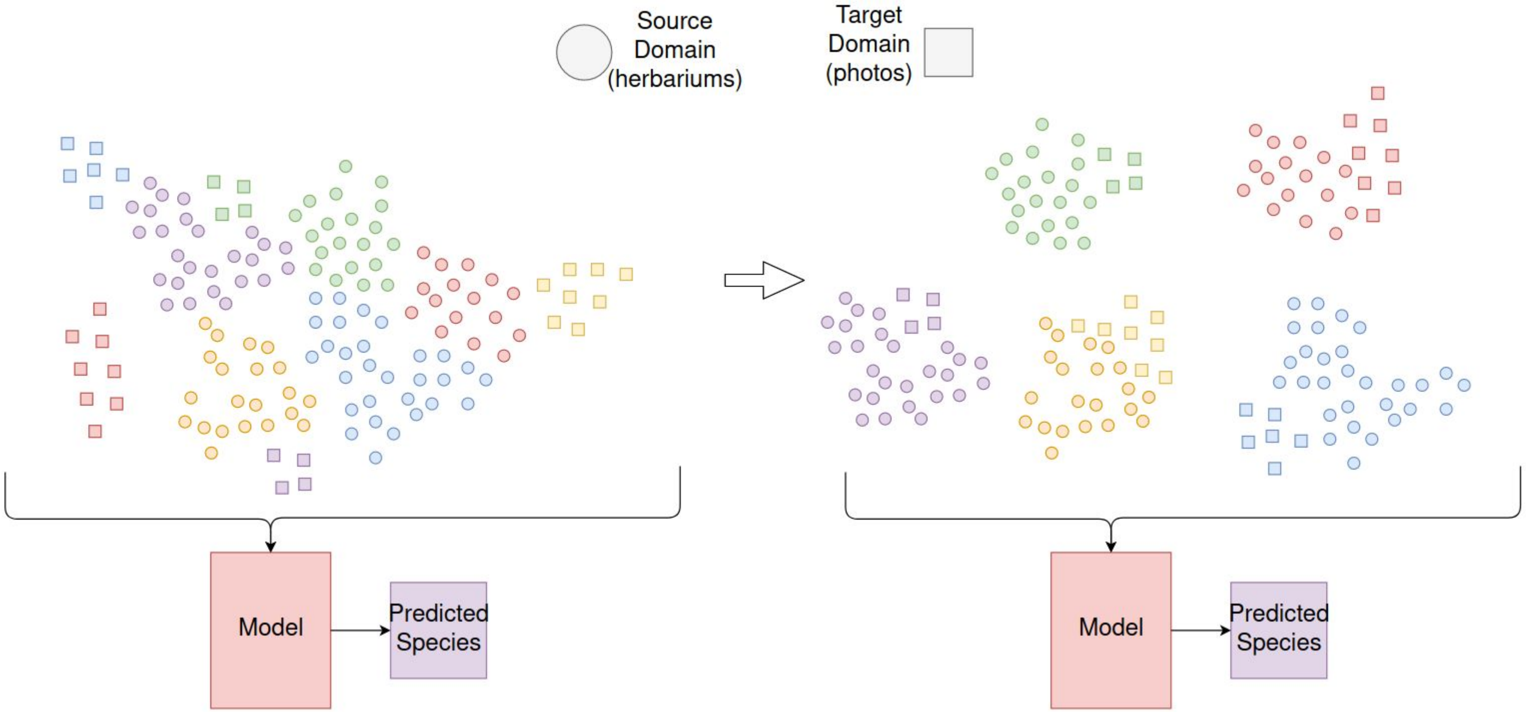
1. Imagenet
2. Finetune with Herbariums
3. Finetune with photos

Try to use herbarium features to obtain a better model: naive approach



Carranza-Rojas, J., Goeau, H., Bonnet, P., Mata-Montero, E., & Joly, A. (2017). Going deeper in the automated identification of Herbarium specimens. *BMC evolutionary biology*, *17*(1), 181

2. Use domain adaptation

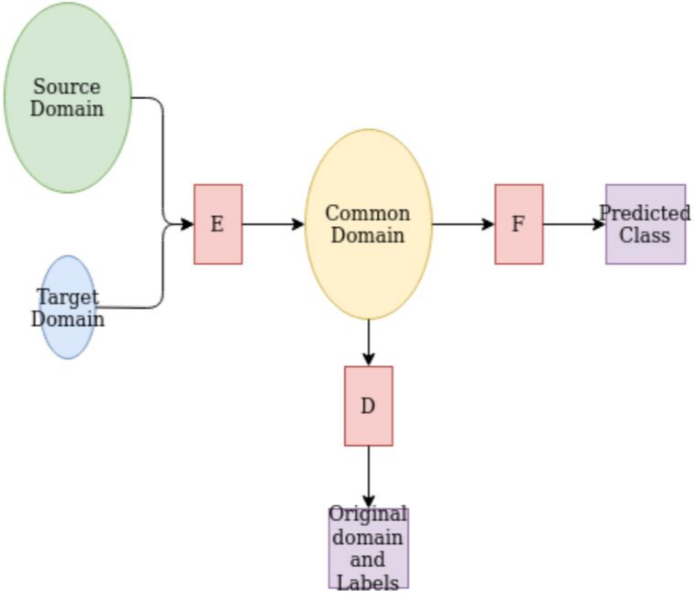


Same labels but different distributions, hard to train a model

Try to map to a similar distribution where training is easier

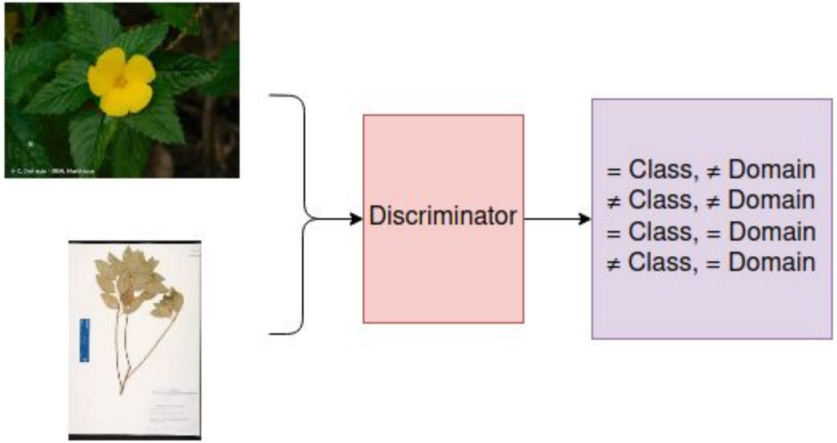
2. Use domain adaptation (FSDA)

Technique used: **F**ew-**S**hot adversarial **D**omain **A**daptation



Discriminator: trained to determine if inputs are from same class and domain

Objective: fool it



Motiian, S., Jones, Q., Iranmanesh, S., & Doretto, G. (2017). Few-shot adversarial domain adaptation. In Advances in Neural Information Processing Systems (pp. 6670-6680).

3. Extra data

Use data from sources like search engines and online repositories to compensate for the lack of images

The data was taken from PlantCLEF19 and last years winner of the PlantCLEF challenge[2]

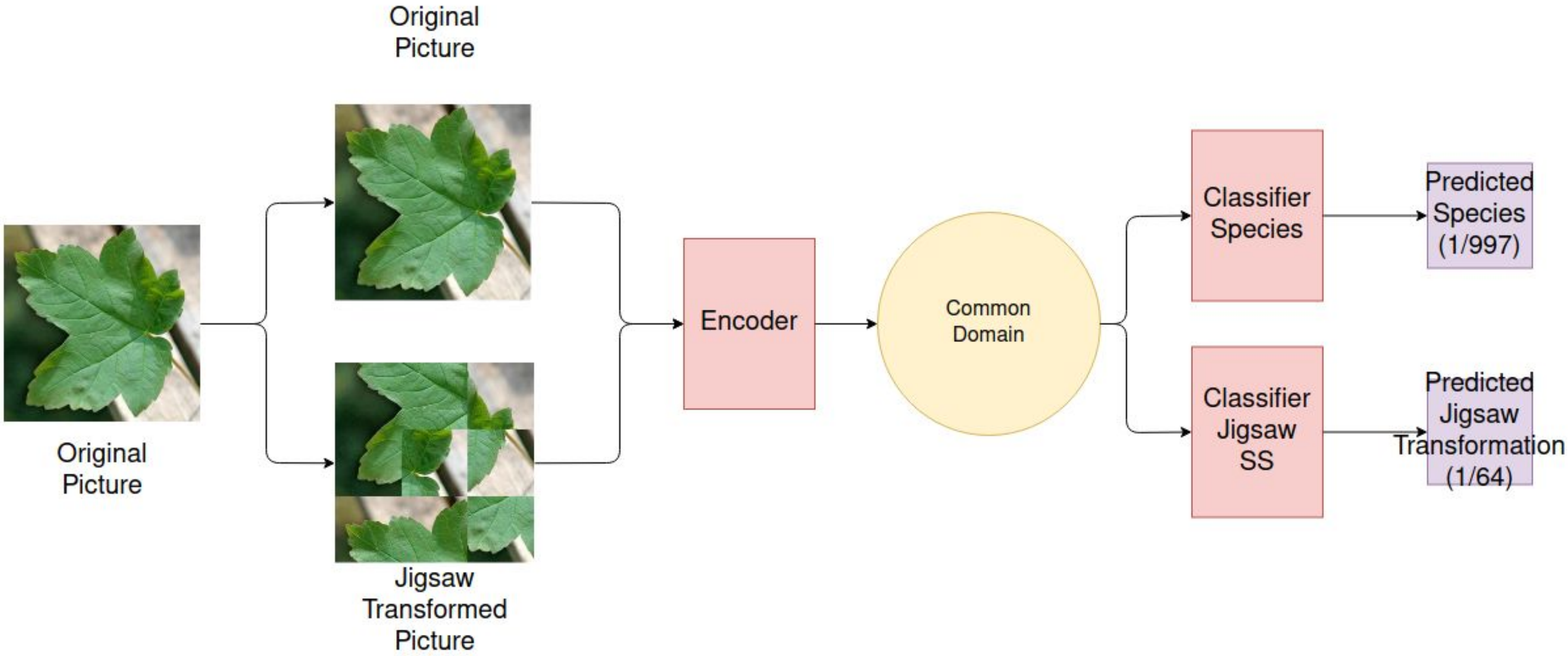
Data can be noisy and not completely related to the task

Additional 134,457 images from 997 species

Useful but time consuming

[2] Lukas Picek, Milan Sulc, and Jiri Matas. Recognition of the amazonian flora by inception networks with test-time class prior estimation. In Working Notes of CLEF 2019

4. Self supervision in FSDA



4. Upper taxons (genus + family)

Passifloraceae

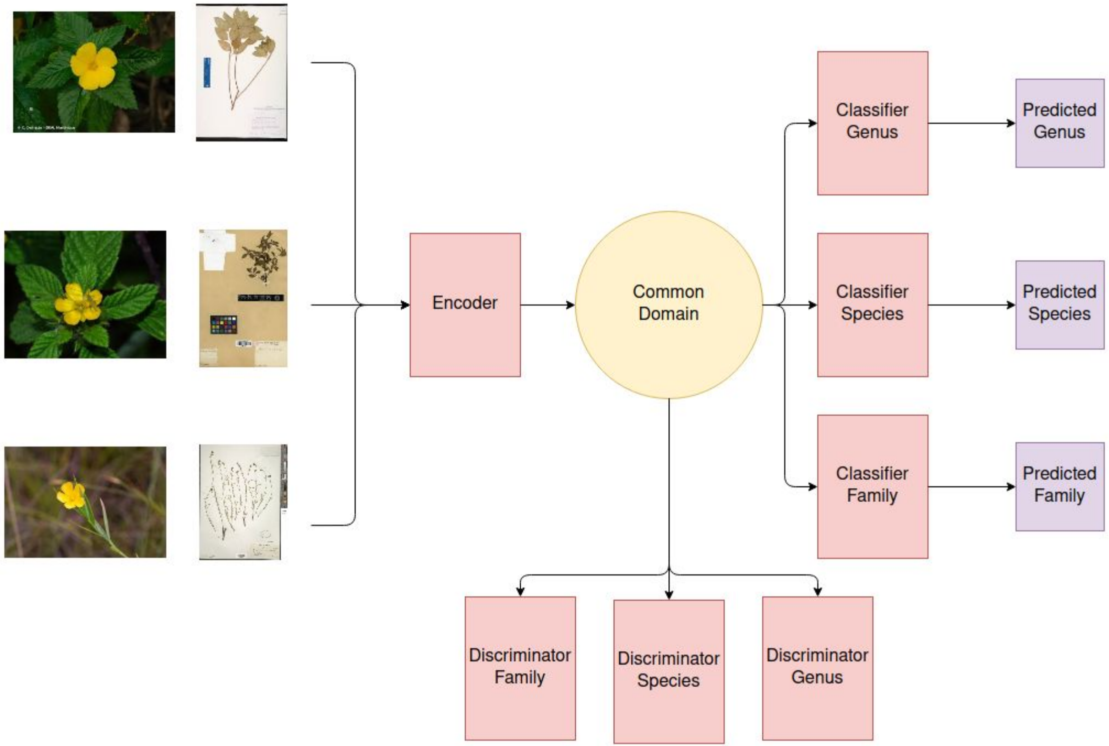
Turnera ulmifolia

Passifloraceae

Turnera odorata

Passifloraceae

Turnera guianensis



Encoder

Common Domain

Classifier Genus

Classifier Species

Classifier Family

Predicted Genus

Predicted Species

Predicted Family

Discriminator Family

Discriminator Species

Discriminator Genus

5. Combine everything

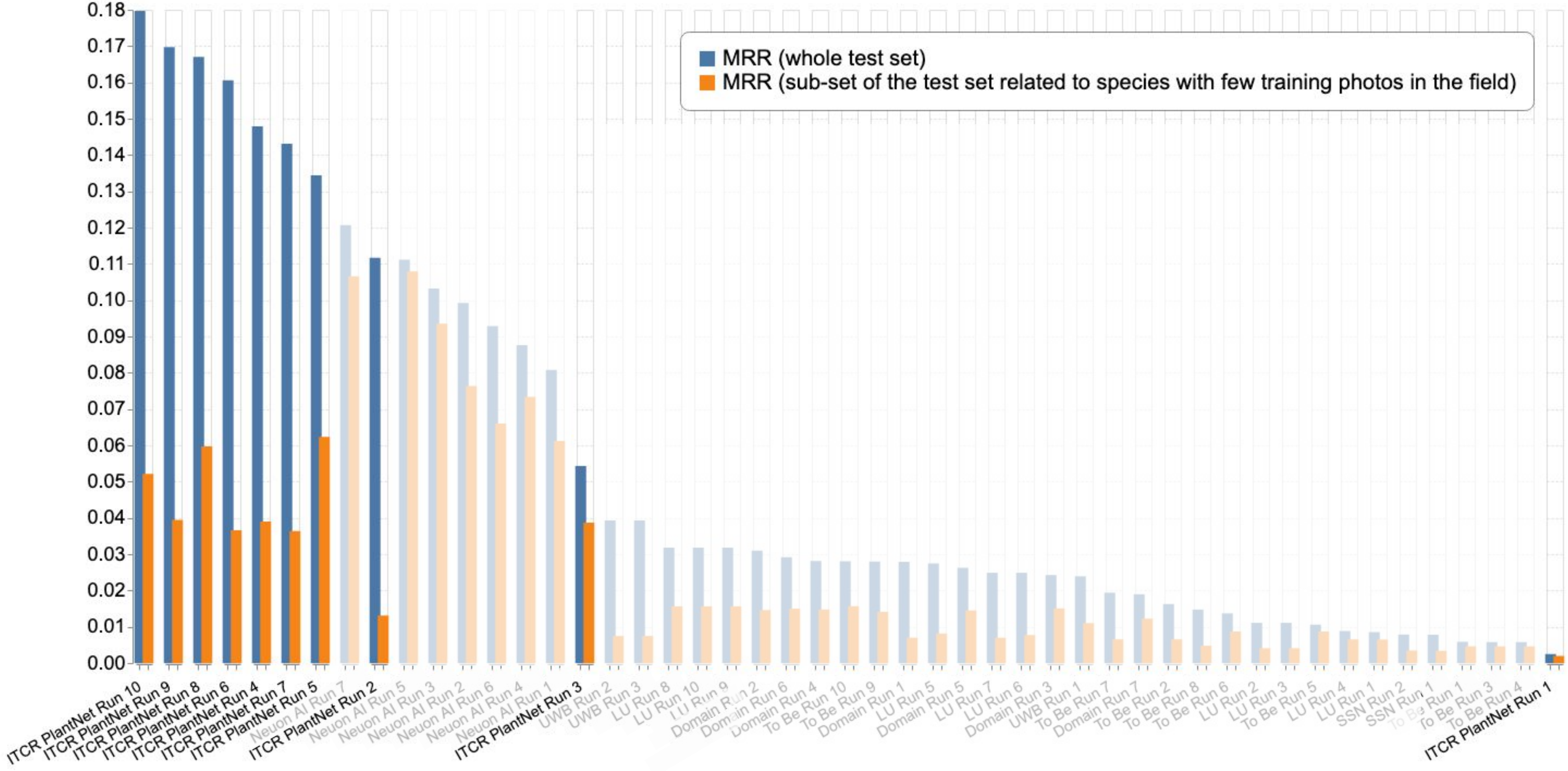
Mix the best techniques to try to obtain the best results possible

Combinations tried:

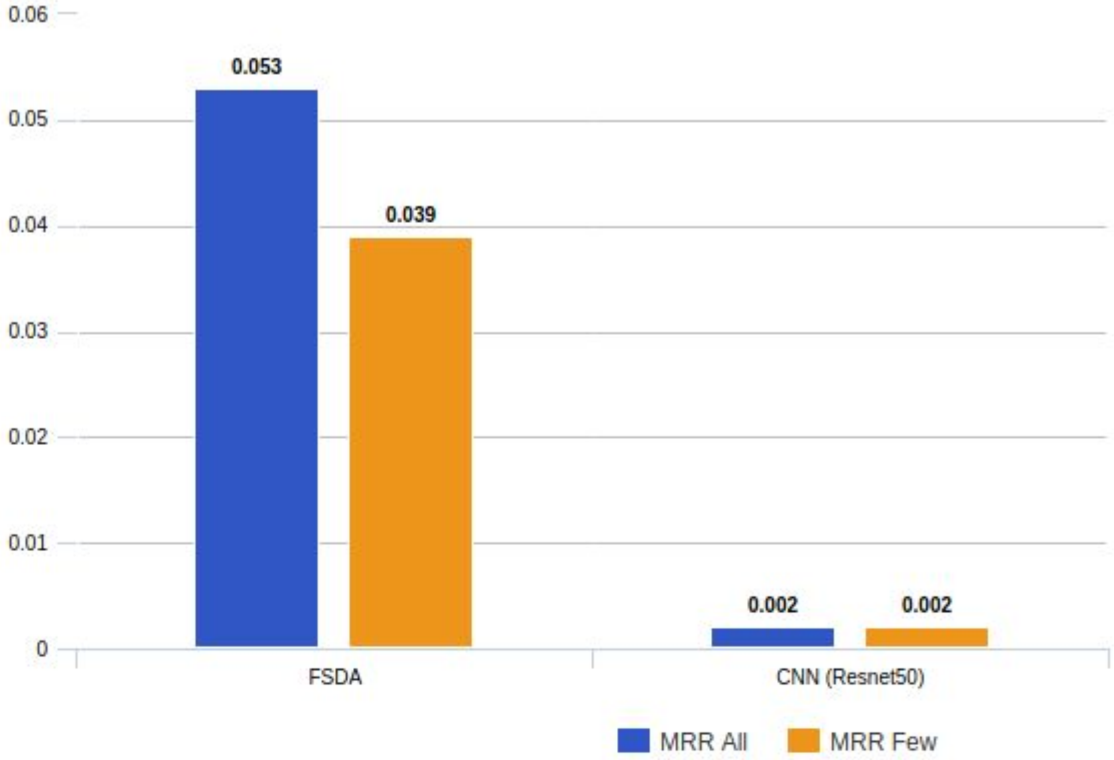
1. FSDA + extra data + self supervision
2. FSDA + extra data + upper taxons (genus & family)
3. FSDA + extra data + upper taxons (genus & family) + self supervision
4. Ensemble

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}$$

Results



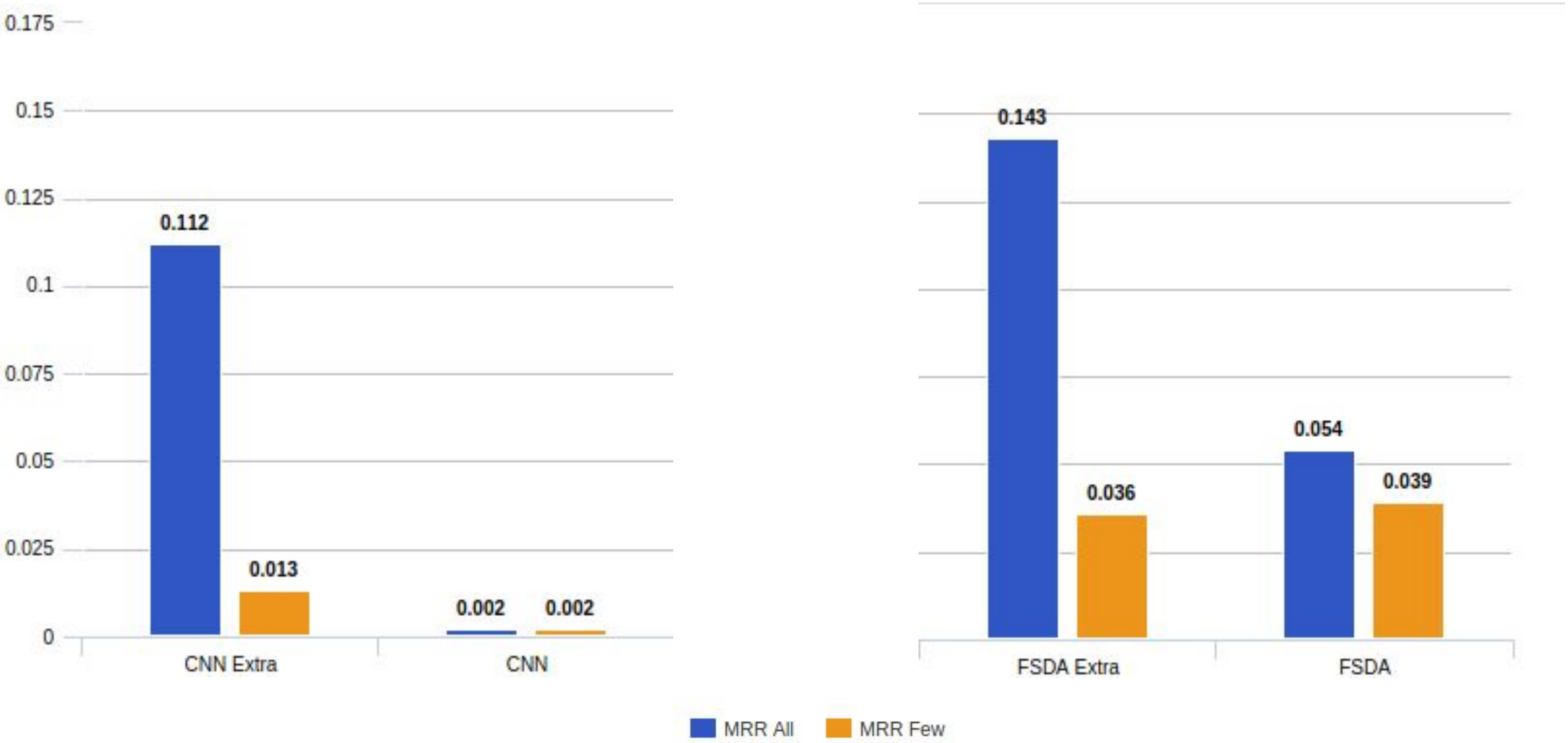
Results: Impact of Domain Adaptation



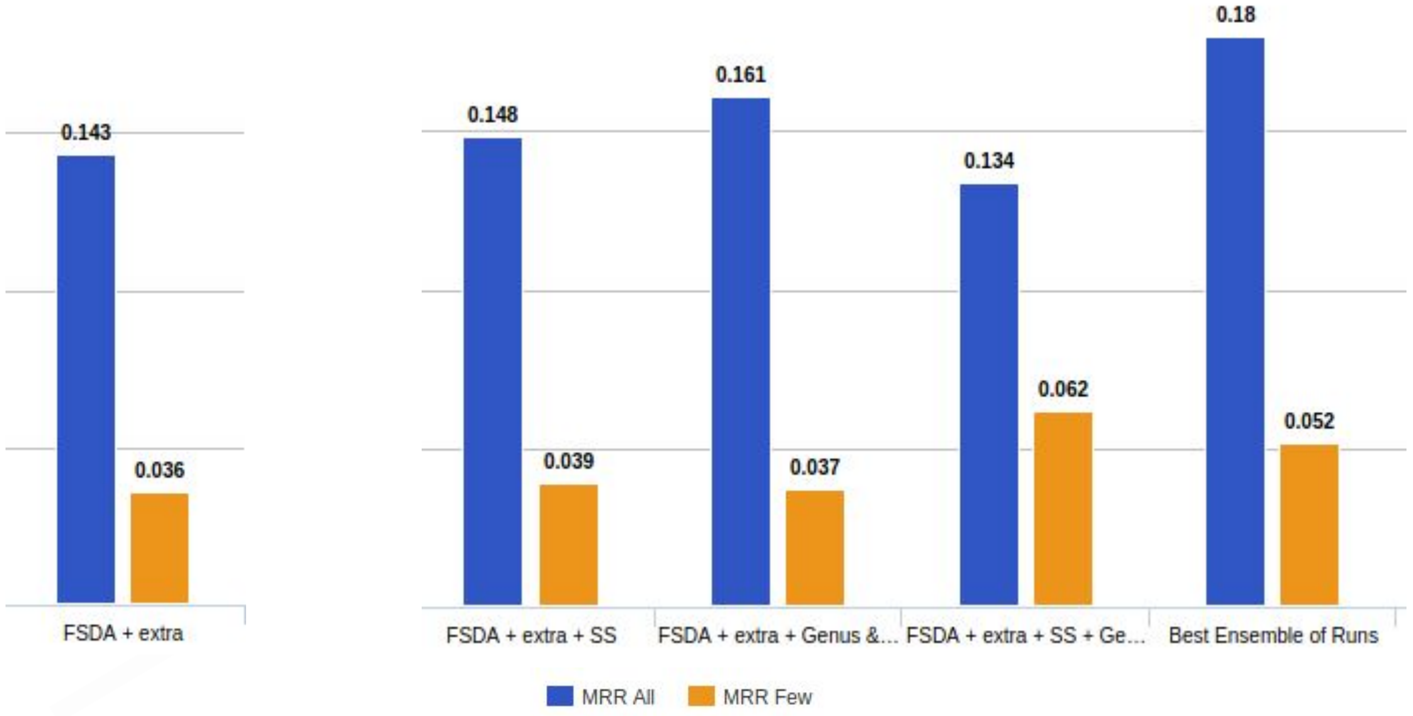
Results: Impact of extra training data

in CNN

in FSDA



Results: Impact of Performance Improving Techniques



Conclusions and Future Work

Domain adaptation increases significantly the generalization power of the models

Self supervision and information from upper taxons help the models learn

Highest score on MRR All on PlantCLEF20 but many room for improvement, particularly in difficult species

Improvements: incorporate other botanical knowledge into the process (morphology classes & metadata)