AI4Good 2022: Team 3A

ReforesTree: A Dataset for Estimating Tropical Forest Carbon Stock with Deep Learning and Aerial Imagery

Final Presentation

Klim Troyan, Silviu Nastasescu, Dominic Wong

Agenda

- Paper review
- Pre-processing
  - Outlier detection
  - Data imputation
- Manual data annotation
- Tree crown detection model
- Matching algorithm
- Regression model pipeline
- Possible future directions
- Questions

Paper review

- Paper's main purpose is to publish a dataset and present future directions to work with it

- GitHub repo and experiments are far from complete
  → Contacting the authors was needed

- Paper (with appendix) is not enough
  → Had to refer to the Master thesis (e.g., for missing AGB formula)
  → Incorrect formula in the paper, incorrect values (e.g., regarding the number of deviating banana trees, total number of trees, etc.)
  → No ground-truth on how each type of trees looks like

- Quality of the proposed data is low
  → Units of the measurements and description of (non-obvious) features are absent
  → Obvious and complicated to solve issues in the field_data file
  → Good and feasible ideas but ill-posed problem w.r.t. the considered data quality

- Make public and more transparent the relevant steps performed in the pipeline so that the results can be reproduced (e.g., for the baseline model, experiment setting, etc.)
  → How to compare? Evaluation process becomes hardly feasible

Pre-processing: Outlier detection

- Formulas require only the diameter and the species group
- The small number of samples allowed manual OD
  - Two trees with DBH < 1 cm
  - Nine trees with DBH > 50 cm
  - Only 0.2% samples in total

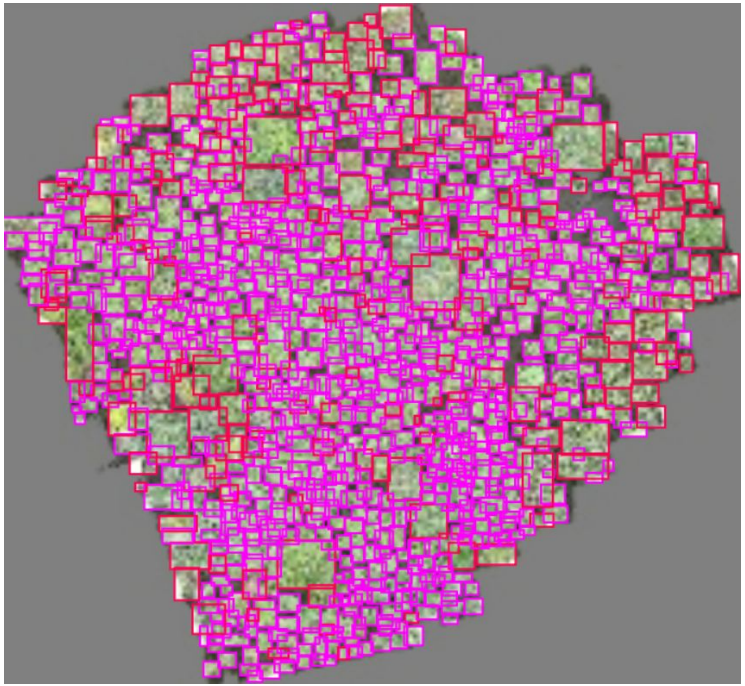
Pre-processing: Data imputation

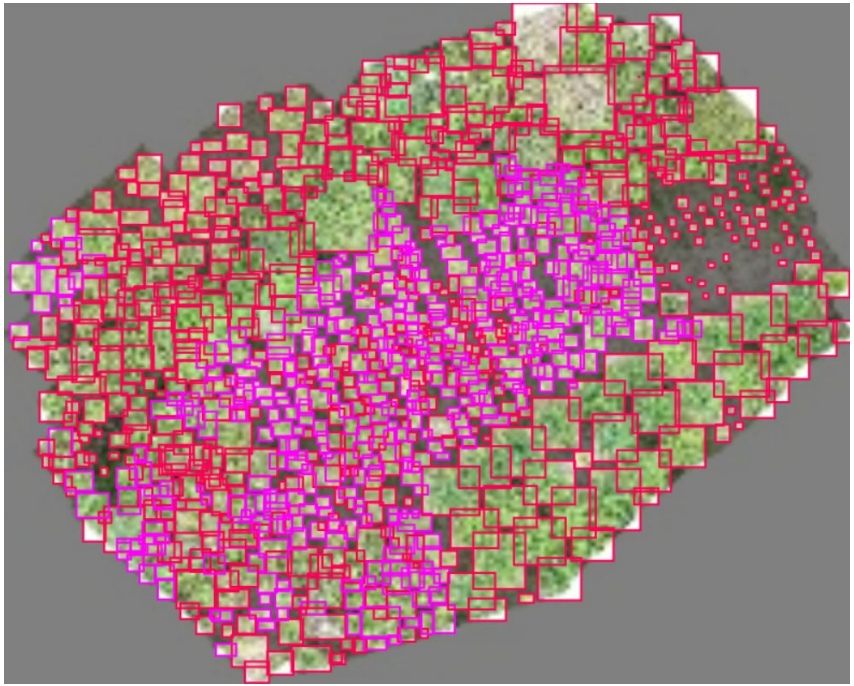
- 44% DBH values are missing
- Paper implementation: mean imputer based on year and species group
- Ours: kNN imputer based on the year, species group, species name, site
- Categorical labels were one-hot encoded

Manual data annotation

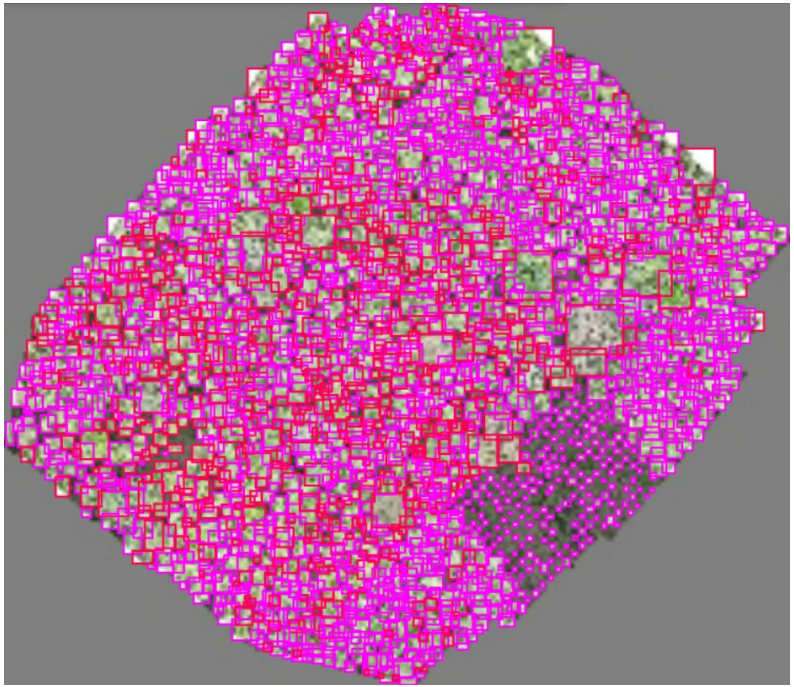
- Manual annotation of all the trees in the areas of interest, and even more!
- Annotations: bounding boxes with binary banana/non-banana labels.
  - banana: 4116
  - non-banana: 3723
- More accurate and reliable than DeepForest's predictions

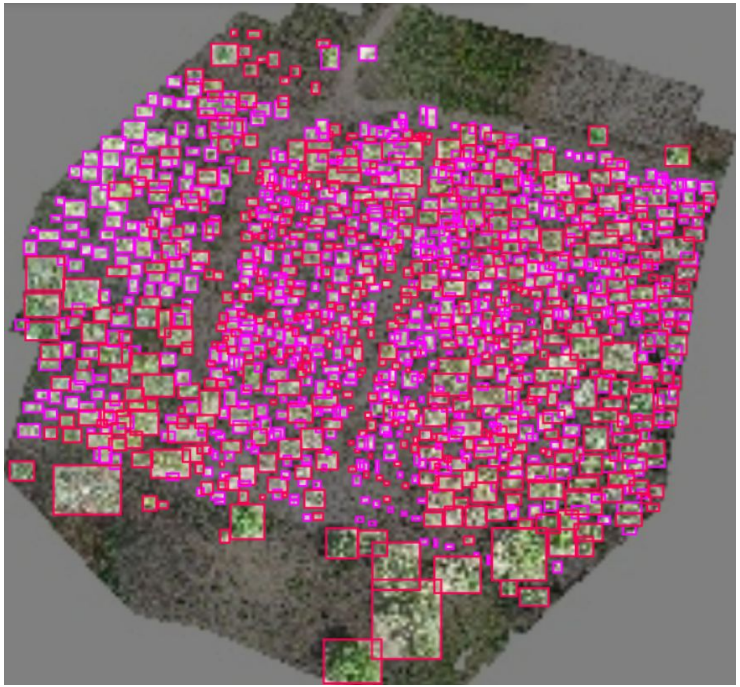
Manual data annotation



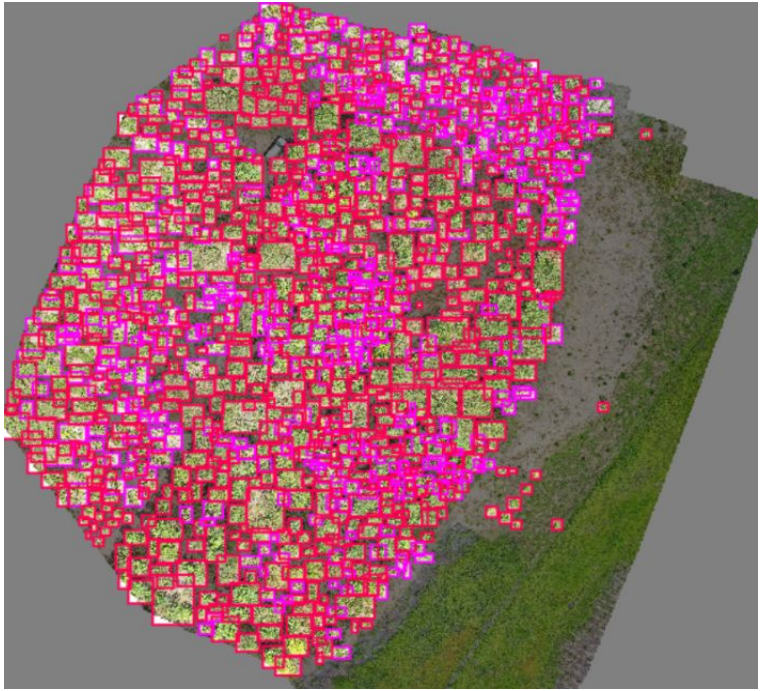


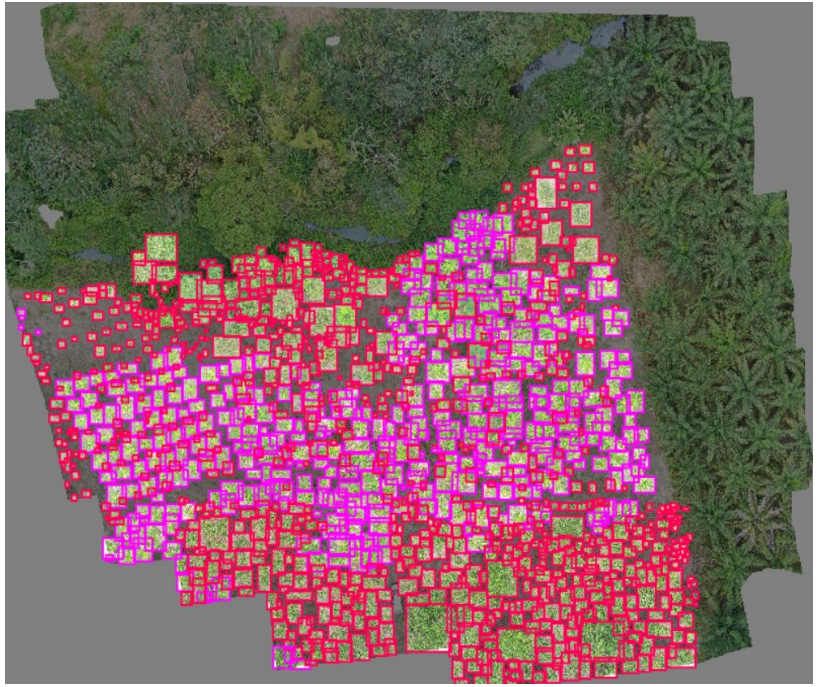
Manual data annotation





Manual data annotation



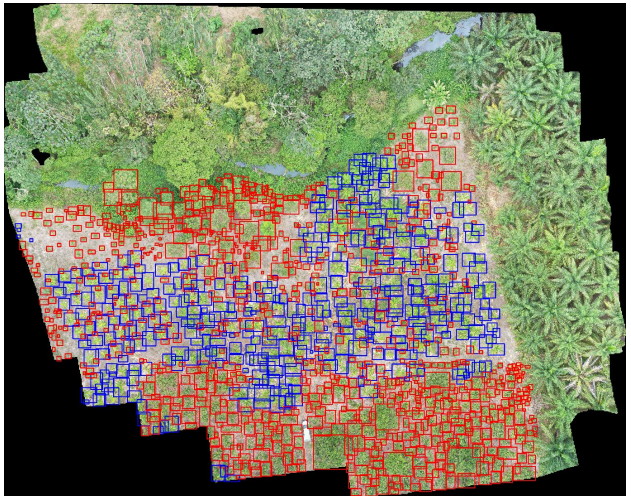


Manual data annotation: Usage

- Fine-tune DeepForest
- Train new object detection models like DeepForest
- Input for OneForest, avoiding the annotations overlap problem
- Important insight: the number of trees in the considered plantation areas is smaller than the number of trees measured

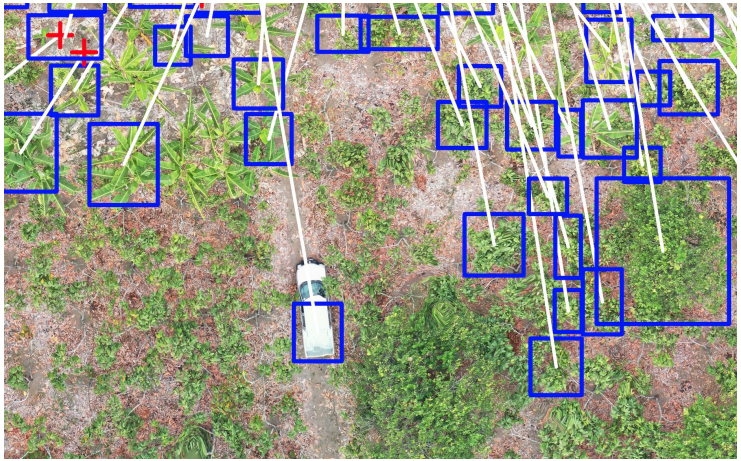
Tree crown detection model

- The current deepforest model only predicts bounding boxes of trees
- We have hand annotated bounding boxes with banana/non-banana labels
- Therefore we can finetune/train a detection model that distinguishes between bananas and non-bananas



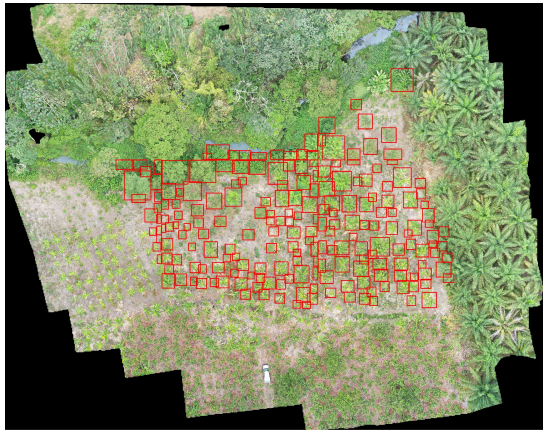
Tree crown detection model

- Dense trees
- Multiple trees in single bounding box
- Single tree split into multiple bounding boxes
- False detections

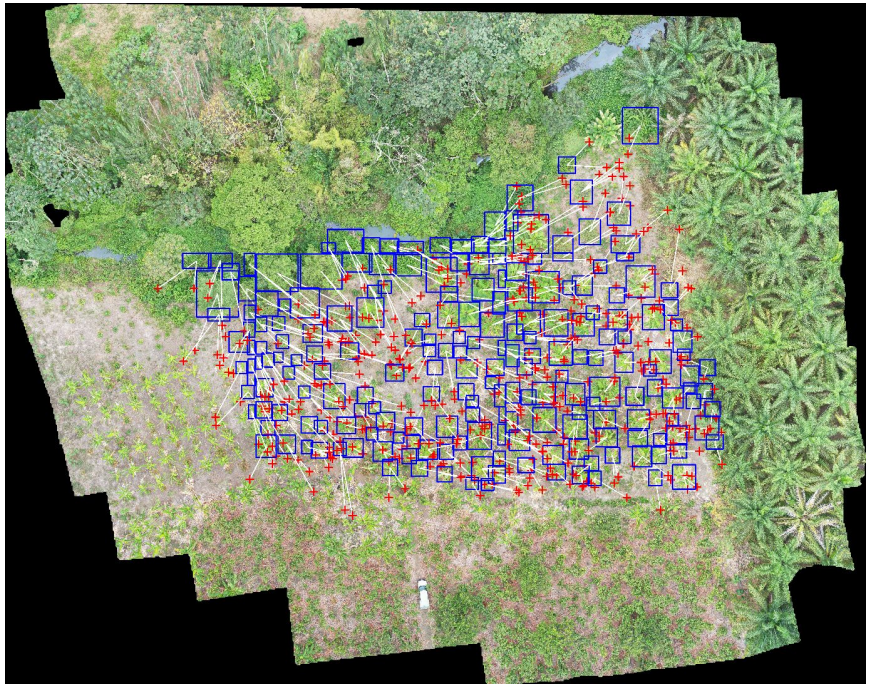


Matching algorithm

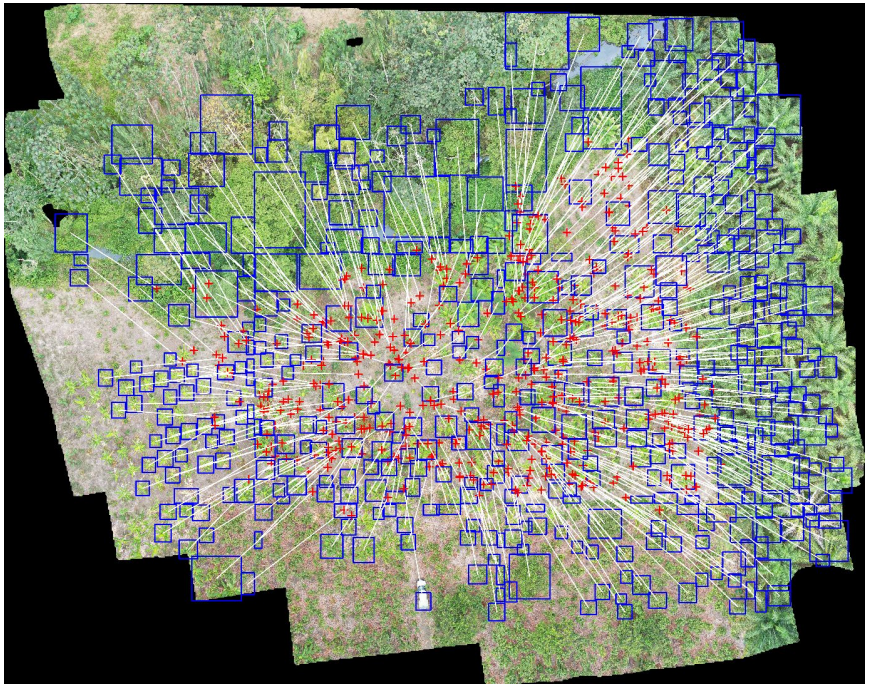
- We saw that a lot of field measurements are matched to trees that evidently do not belong to the site
- On the other hand, the number of trees inside the site polygon is much less than the number of field measurements
- Idea: dilate the given site polygon until the number of trees in the dilated polygon is equal to or more than the number of field measurements



Matching algorithm: dilation



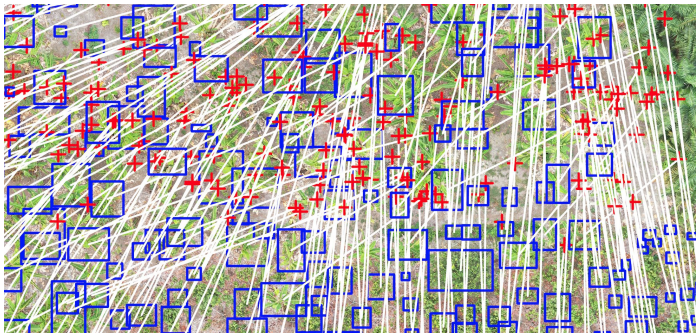
Without dilation

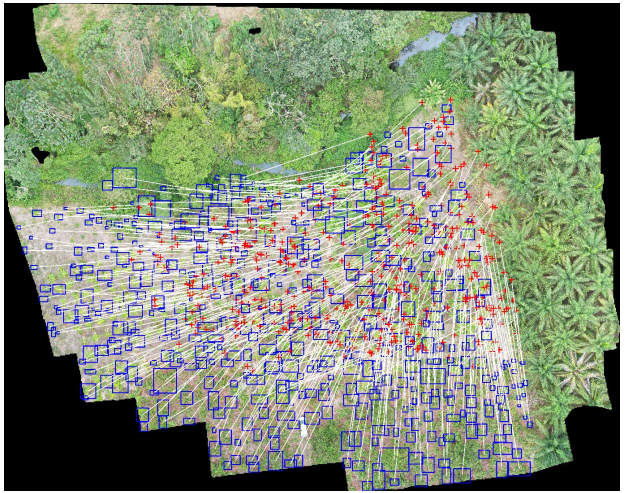


With dilation

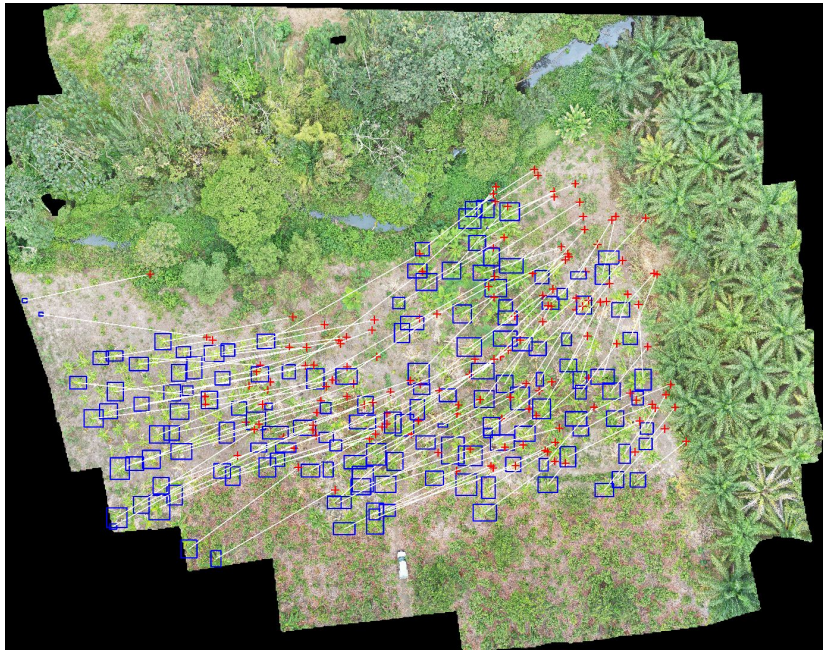
Matching algorithm with tree labels

- Optimal transport is flexible: we can constraint the matching algorithm to include priors about tree groups
- This is complementary to our proposed tree-detection model, as optimal transport uses confidence scores and minimises the cross-entropy of matches

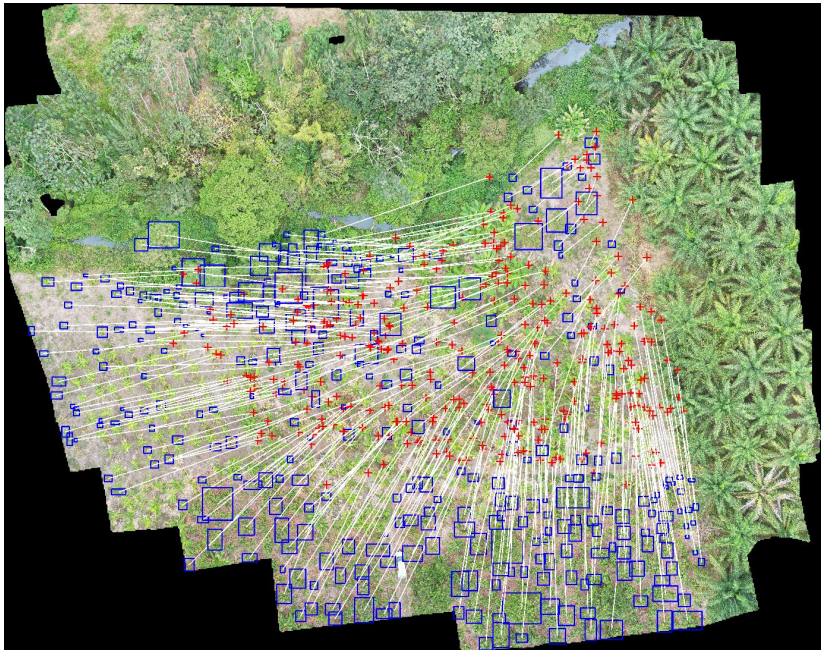




Matching algorithm with tree labels



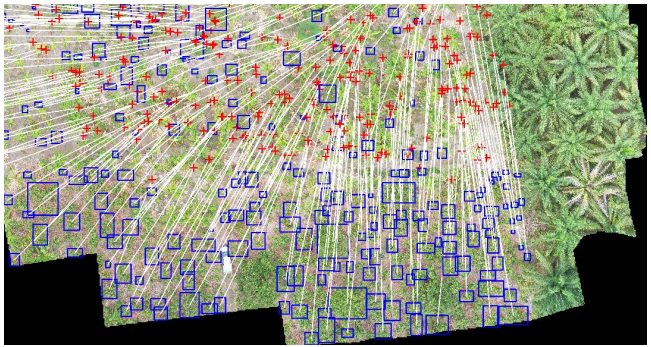
Banana

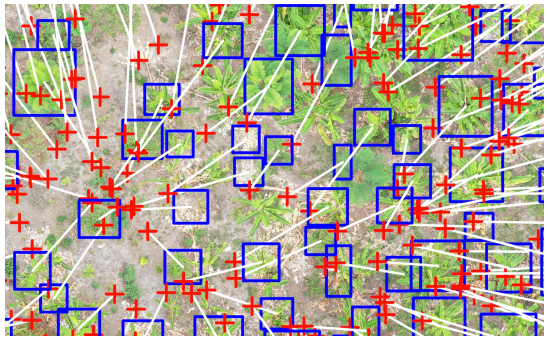


Non banana

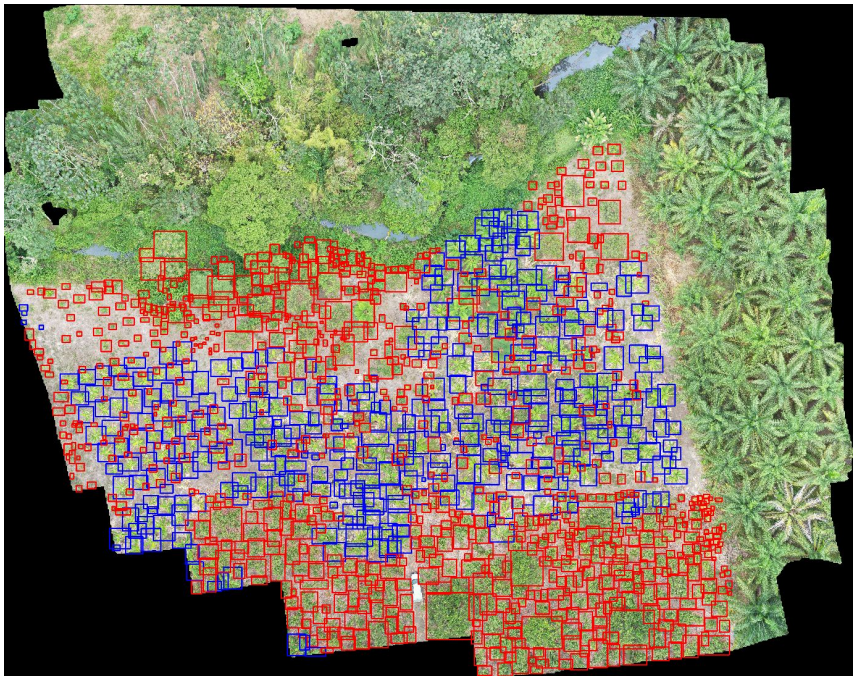
Matching algorithm: open question

- Optimal transport fits the the target distribution as “full” as possible
- For our use case, it is the spatial distribution of the matched distribution should be similar to the GPS measurements
- How can we configure/constraint optimal transport to reflect this?
- Are there other, more suitable approaches?

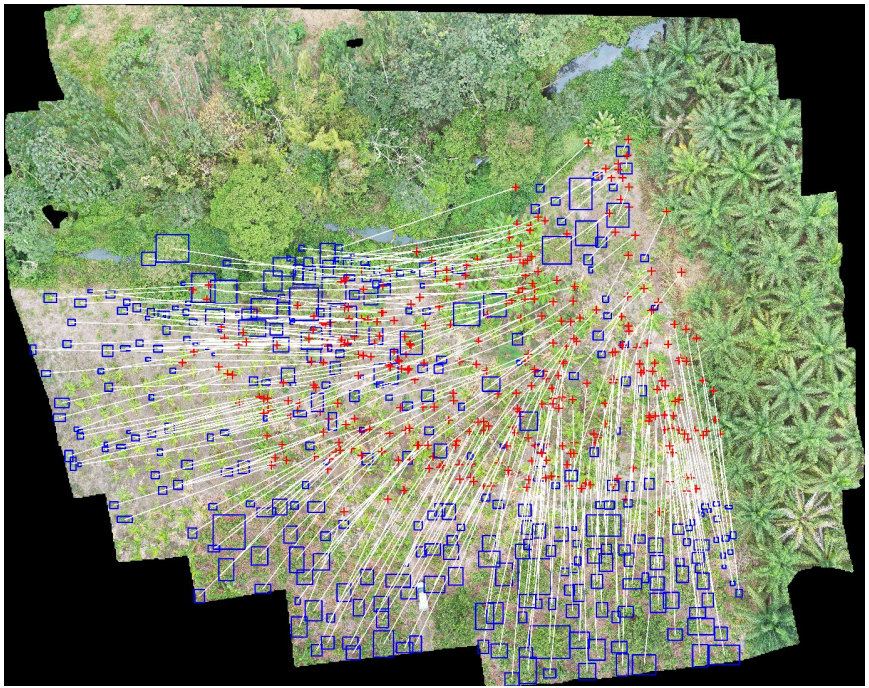




Matching algorithm: open question



Hand annotated bounding boxes



“Non-banana” matches

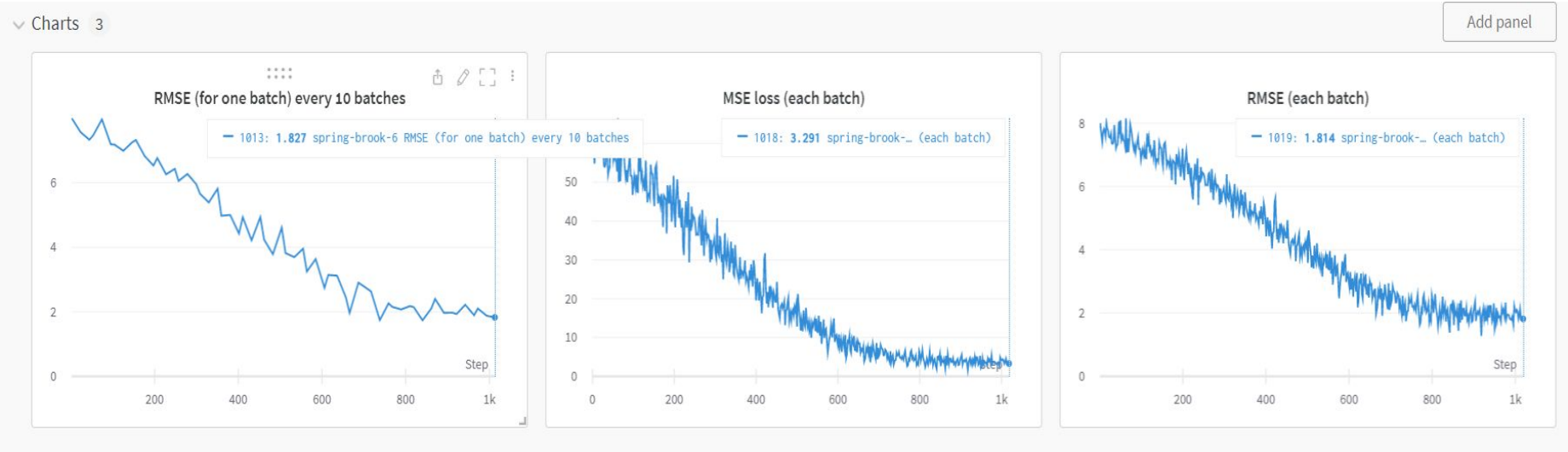
Regression model pipeline

- Implemented the full model pipeline using their final data (torchgeo and shared folder too)
  → Difficult to reproduce their baseline result... (see next slides)
- Proposed a model pipeline that incorporates some (potential) improvements
  → More robust data imputation
  → Data augmentation using flips and rotations
  → Multi-inputs model (essentially tree image + embedded categorical features such as tree group)
  → Newer ResNet backbone + freezing of the encoder or not

- Hyper-parameter search for both models

- Set and documented the evaluation procedures to obtain relatively good RMSE and R2 scores with our own model (on the initial data as well as “ours”) so that future works could compare meaningfully

Regression model pipeline

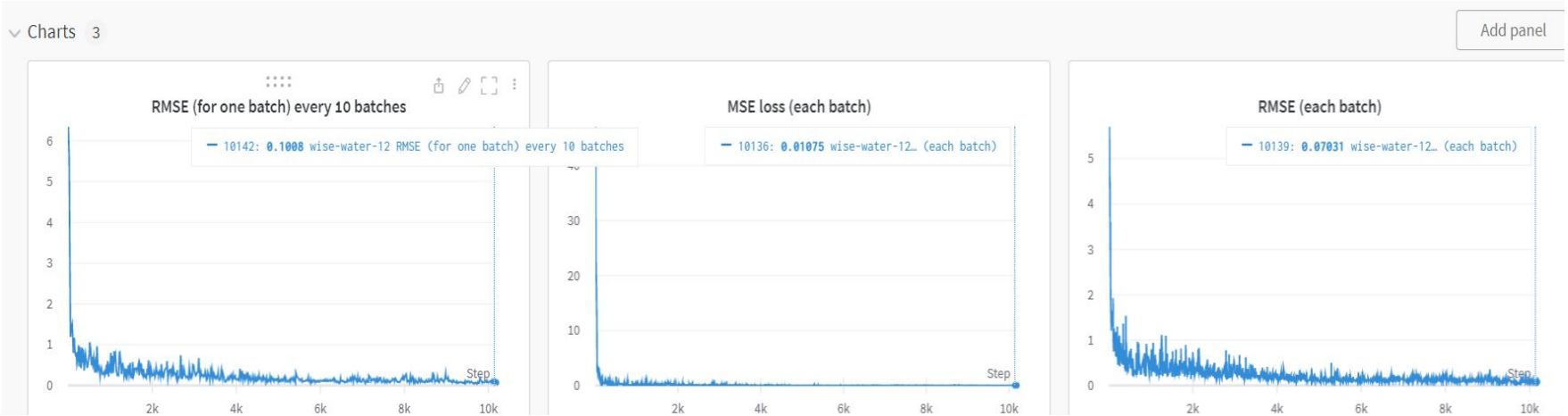


→ Not quite the results we would like… but this can be improved!

→ Use additional data pre-processing (e.g., for unmatched outlier values)

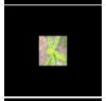
→ Training can be improved with better HPs

Regression model pipeline



→ Decent result… but this is only training! Validation and testing errors are both around 1.4

Regression model pipeline

	image	group label	target_gt	target_prediction
1		1	5.734	8.884
2		1	8.7	8.198
3		1	9.33	3.73
4		1	6.069	7.003
5		0	3.42	7.986
6		0	3.42	3.402

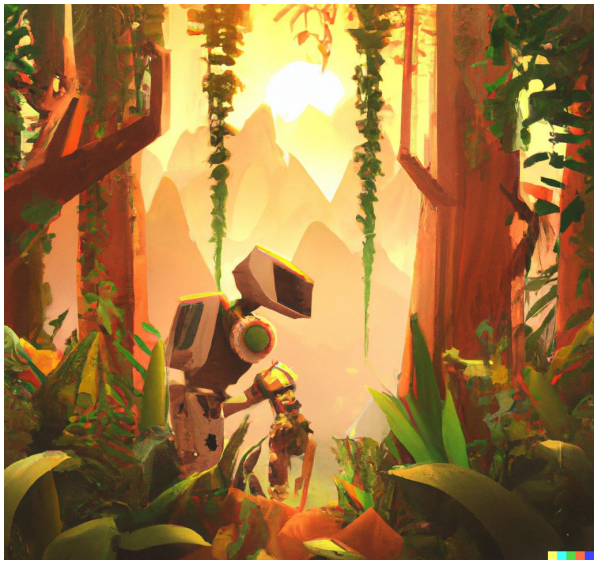
→ Labels seem good here...

→ Poor data + Imputation method + Matching algorithm are surely the cause of incoherences

Possible future directions

- Leveraging the manually annotated dataset for related tasks (cf. previous discussions)
- Use optimal transport (OT) algorithms or other approaches that preserve spatial distribution of trees
- . . .

Questions ?



Contributions

- All group members contributed equally :)