Home excericse - NOVA course on deep learning in remote sensing

Julius Wold

2023-06-29

Code used for experiment can be found in this repository.

Objective

Compare the effect of training a seedling detector on your own annotated dataset vs. the full dataset on the detector's performance.

Materials and methods

Annotations

Two sets of annotations were used for model tranining: **My annotations** and **All annotations**.

- **My annotations**
  - Annotations made by the author.
- **All annotations**
  - Annotations from all students merged.

My annotations consists of 48 images with 280 annotations of trees while **All annotations** consists of 387 images with 5062 annotations. 70% of data was randomly assigned to training and 30% to validation. An overview of images and annotations used in the tranining and validation split are shown in Table 1.

Table 1: Number of images and annotatations for each dataset.

Count	Split	All annotations	My annotations
Number of images	Train	271	34
Number of images	Val	116	14
Number of images	Sum	387	48
Number of trees	Train	3492	201
Number of trees	Val	1570	79
Number of trees	Sum	5062	280

Test data

Two sets of test data was used for evaluating models: **Tiled test data** and **Drone RGB orthomosaics**.

- **Tiled test data**
  - Used for ML metrics evaluation.
  - Tiled and annotated images.
- **Drone RGB orthomosaics**
  - Used for domain metrics evaluation.
  - Orthomosaics of four sites.
    - Each site contains four plots (~0.1 ha) with tree posistions measured in field.

Model training

YOLOv8 models were trained using the dataset **My annotations** and the dataset **All annotations**. A grid search were performed for model sizes *Nano*, *Medium* and *Xtra large* and image sizes *256*, *640* and *1024* (Table 2). The best model for each dataset were selected using mAP@.5.

Table 2: Grid search.

Model	Image size
yolov8n.pt	256
yolov8n.pt	640
yolov8n.pt	1024
yolov8m.pt	256
yolov8m.pt	640
yolov8m.pt	1024
yolov8x.pt	256
yolov8x.pt	640
yolov8x.pt	1024

Model evaluation

The selected models from the model training were evaluated using machine-learning metris and domain metrics on the test data.

ML metrics

ML metrics were evaluated by testing the models against the **tiled test data** using the inbuilt val mode at default settings.

Domain metrics

Residual Mean Square Error (RMSE) and Mean Deviation (MD) were calculated according to Equation 1 and Equation 2 for trees per ha. Additionally RMSE (%) and MD (%) were calculated as size of RMSE and MD relative to the observed mean.

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad (1)$$

$$MD = \frac{\Sigma_{i=1}^{n}(\hat{y}_i - y_i)}{n} \qquad (2)$$

Results & Discussion

Model training

Detailed view of trained models can be found in this Comet project.

Training performance

Models trained by the **My annotations** dataset achived better performance in training than the models trained on **All annotations**. For **My annotations** the highest performance was achived by the *nano* model with image size *640* (mAP@.5 of 0.59 at epoch 94). The best perfoming model for **All annotations** was the *meduim* model with an image size of *640* (mAP@.5 of 0.36 at epoch 42). Summary of training results for are shown in Table 3 (a) and Table 3 (b) for **My annotations** and **All annotations** respectively.

Table 3: Training performance of best models.

(a) My annotations

model	img sz	metrics/mAP50(B)	metrics/mAP50-95(B)	metrics/precision(B)	metrics/recall(B)	model/speed_PyTorch(ms)
yolov8x.pt	1024	0.545	0.228	0.551	0.544	24.577
yolov8x.pt	640	0.511	0.235	0.664	0.456	10.645
yolov8x.pt	256	0.536	0.211	0.551	0.575	6.930
yolov8m.pt	1024	0.542	0.269	0.595	0.544	9.502
yolov8m.pt	640	0.564	0.256	0.535	0.656	6.573
yolov8	25	0.559	0.246	0.599	0.671	7.859

m.pt	6					
yolov8 n.pt	1024	0.519	0.228	0.692	0.455	4.001
yolov8 n.pt	640	0.593	0.281	0.602	0.669	4.287
yolov8 n.pt	256	0.553	0.252	0.590	0.570	4.787

(b) All annotations

model	img sz	metrics/mAP50(B)	metrics/mAP50-95(B)	metrics/precision(B)	metrics/recall(B)	model/speed_PyTorch(ms)
yolov8 x.pt	1024	0.325	0.115	0.388	0.448	25.644
yolov8 x.pt	640	0.324	0.121	0.426	0.447	10.408
yolov8 x.pt	256	0.316	0.111	0.403	0.468	2.897
yolov8 m.pt	1024	0.328	0.126	0.413	0.460	9.942
yolov8 m.pt	640	0.364	0.129	0.393	0.526	4.113
yolov8 m.pt	256	0.336	0.111	0.376	0.445	2.009
yolov8 n.pt	1024	0.326	0.122	0.418	0.411	2.320
yolov8 n.pt	640	0.342	0.124	0.408	0.466	1.595
yolov8 n.pt	256	0.325	0.114	0.384	0.462	1.216

There was little impact on training perfomance with varying image size and model size for models trained with **All annotations**. For models trained with **My annotations** a difference can be observed between *nano* models and *medium* and *xtra large* models. *Medium* and *xtra large* seems to improve faster than *nano* models (higher performance at earlier epoch) but starts to overfit earlier on the data. mAP@.5 curves for models are shown in Figure 1.

It is suprising that *xtra large* models did not perform better than the smaller model sizes, low quality of annotations or the small size of the dataset might be reasons for this.

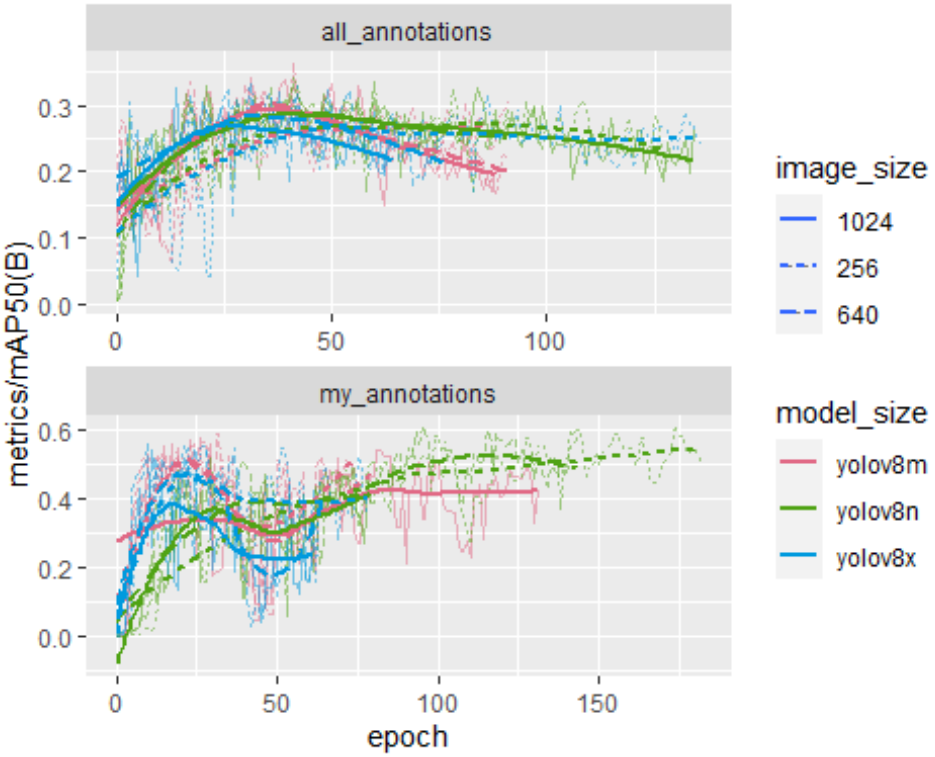
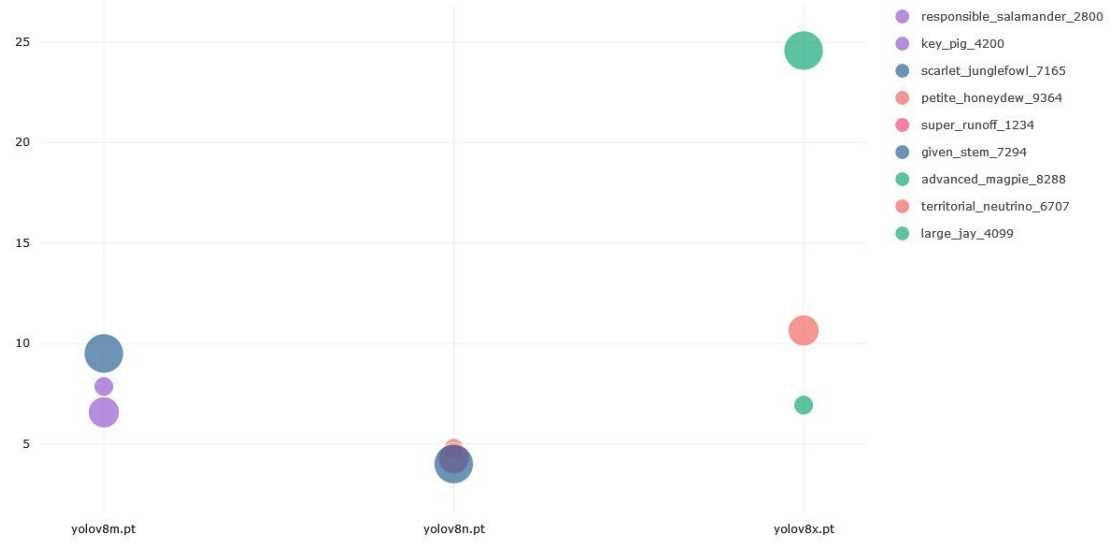


Figure 1: Training results

Model speed

Increasing model size (*Nano* -> *Medium* -> *Xtra large*) resulted in slower models. The effect of image size on model speed was greater with increasing model size. Image size had little effect on *nano* models but lead to much slower models for *medium* and *xtra large* models. Figure 2 shows the effect of model size and image size on model speed.



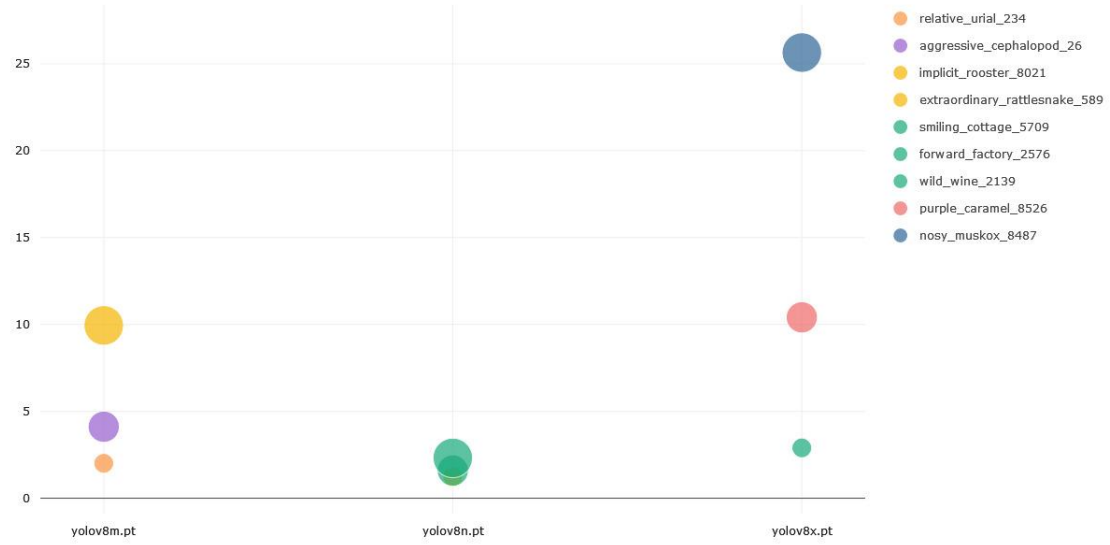


Figure 2: Model speed (size of dots represents image size).

Model evaluation

ML metrics

There was no large difference in performance between the models trained on **My annotations** or **All annotations**. **All annotations** performed slighty better with an mAP@.5 of 0.38 compared to 0.36 for **My annotations**. ML metrics of the models are shown in Table 4.

Table 4: Machine learnining metrics.

	metrics/precision(B)	metrics/recall(B)	metrics/mAP50(B)	metrics/mAP50-95(B)	fitness
my_annotations_yolov8n.pt_640	0.49	0.42	0.36	0.11	0.14
all_annotations_yolov8m.pt_640	0.56	0.45	0.38	0.12	0.14

Domain metrics

RMSE and MD of predicted trees per. ha are shown in Table 5 and predicted and observed values are shown in Figure 3. The difference in domain metrics between models were also here quite low but here the model trained on **My annotations** performed better. RMSE and RMSE (%) over all sites were respectivley 623 trees/ha and 42% for **My annotations** and 632 trees/ha and 43% for **All annotations**. Both models struggled with detections on the site Braatan, with RMSE (%) of 74% and 67% respectivley for **My annotations** and *All annotations**. The best performance was found on site Hobol with RMSE (%) of 10% and 9%.

Both models consistently underestimated the number of trees on all sites. MD and MD (%) over all sites were respectivley -473 trees/ha and -32% for **My annotations** and -499 trees/ha and -34% for **All annotations**.

Table 5: Domain metrics.

Model	aoi_name	RMSE	RMSE (%)	MD	MD (%)
my_annotations_yolov8n.pt_640	braatan	1077	74	-1018	-70
my_annotations_yolov8n.pt_640	galbyveien	504	25	-475	-23
my_annotations_yolov8n.pt_640	hobol	111	10	-57	-5
my_annotations_yolov8n.pt_640	krakstad	357	27	-344	-26
my_annotations_yolov8n.pt_640	all	623	42	-473	-32
all_annotations_yolov8m.pt_640	braatan	981	67	-891	-61
all_annotations_yolov8m.pt_640	galbyveien	692	34	-662	-33
all_annotations_yolov8m.pt_640	hobol	99	9	-90	-8
all_annotations_yolov8m.pt_640	krakstad	386	29	-353	-27
all_annotations_yolov8m.pt_640	all	632	43	-499	-34

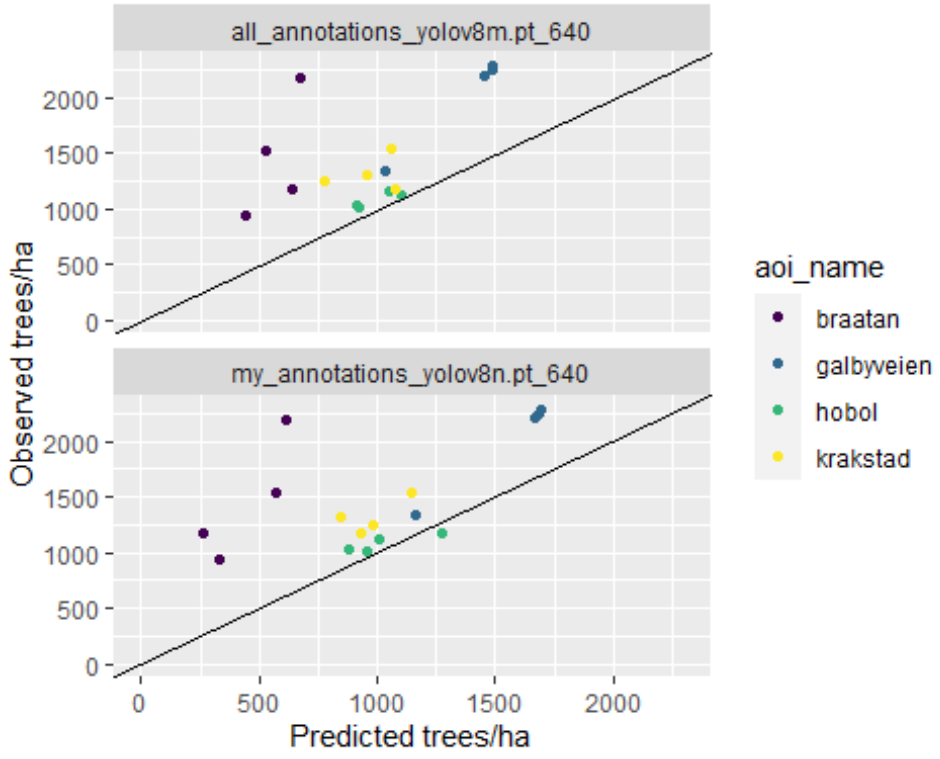


Figure 3: Predicted vs. observed trees/ha. Both models underestimate the number of trees per. ha and struggle with predictions at the sites Braatan and Galbyveien.

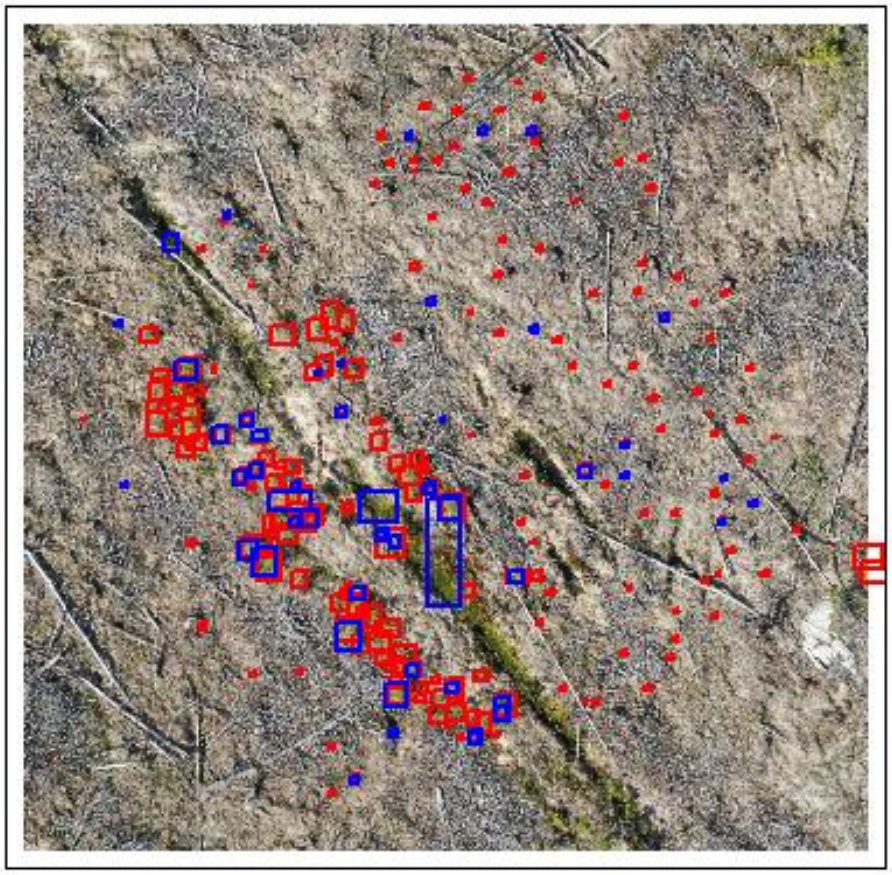
Examples of bad performance.

Both detectors performed poorly at the sites Braatan and Galbyveien. The first example is shown in Figure 4, both detectors struggle with small saplings and trees clumped together. Examples shown in Figure 5 and Figure 6 both show problems with detection of small saplings.

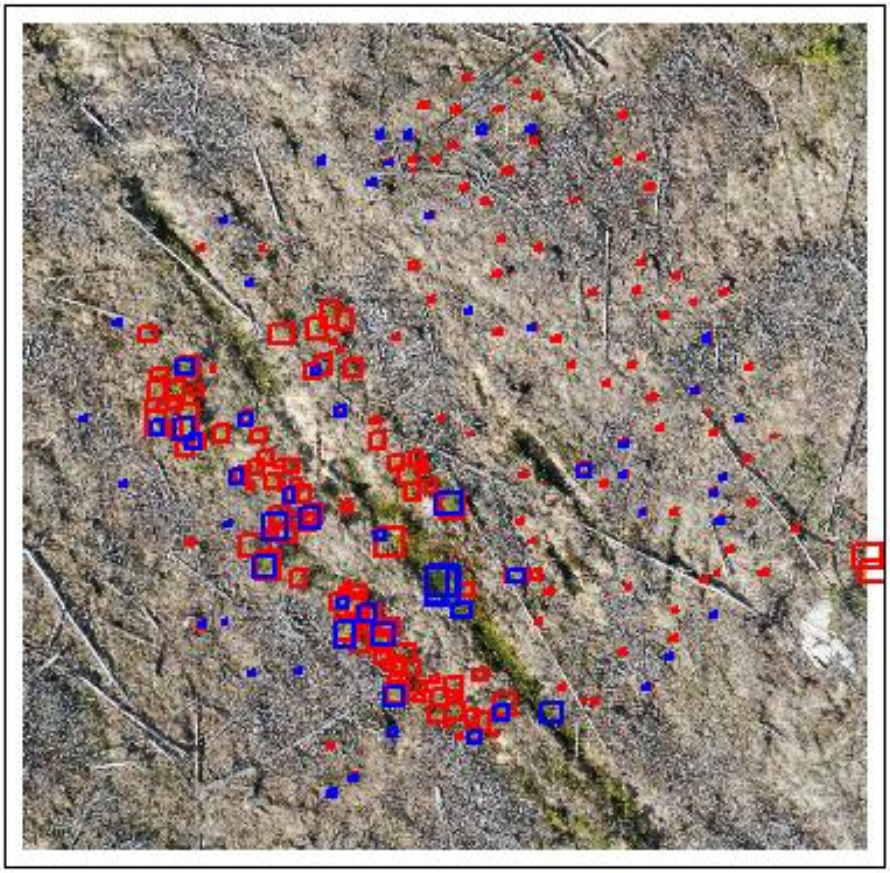
Improving the annotation quality would likley be the best strategy for improving the performance of models, especially for the small saplings. Many of the small saplings are likley to be unnannotated since they are difficult to detect for unexperienced people, which will negativly effect the models performance on small saplings. Saplings of pine are probably affected more by this, as they are more difficult to detect in annotation.

Increasing the value of the hyperparameter *imgsz* could also be a good strategy of improving performance on small saplings, as the size of the saplings in the image is quite low. Increasing image size from the default had little effect in this experiment, poor quality of annotations on small saplings can be a cause for this.

Improving the performance of trees clustered togheter is probably more difficult since the models need to differentiate between clustered trees and windfallen trees. Improving annotation quality would likley be the best option.

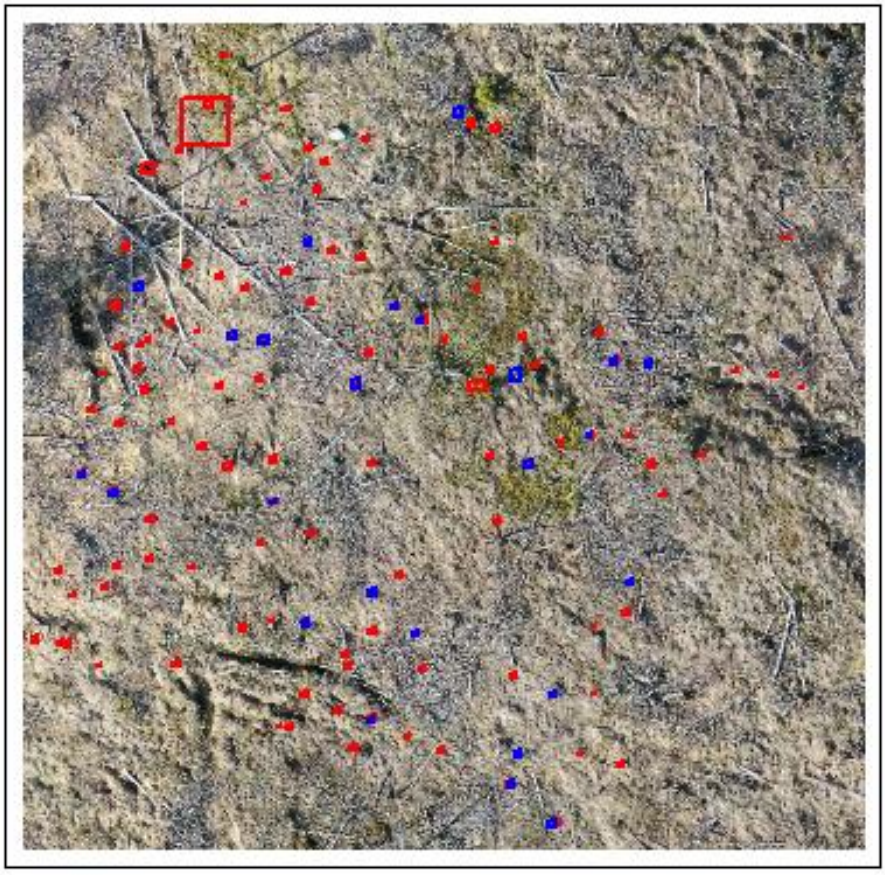


(a) My annotations

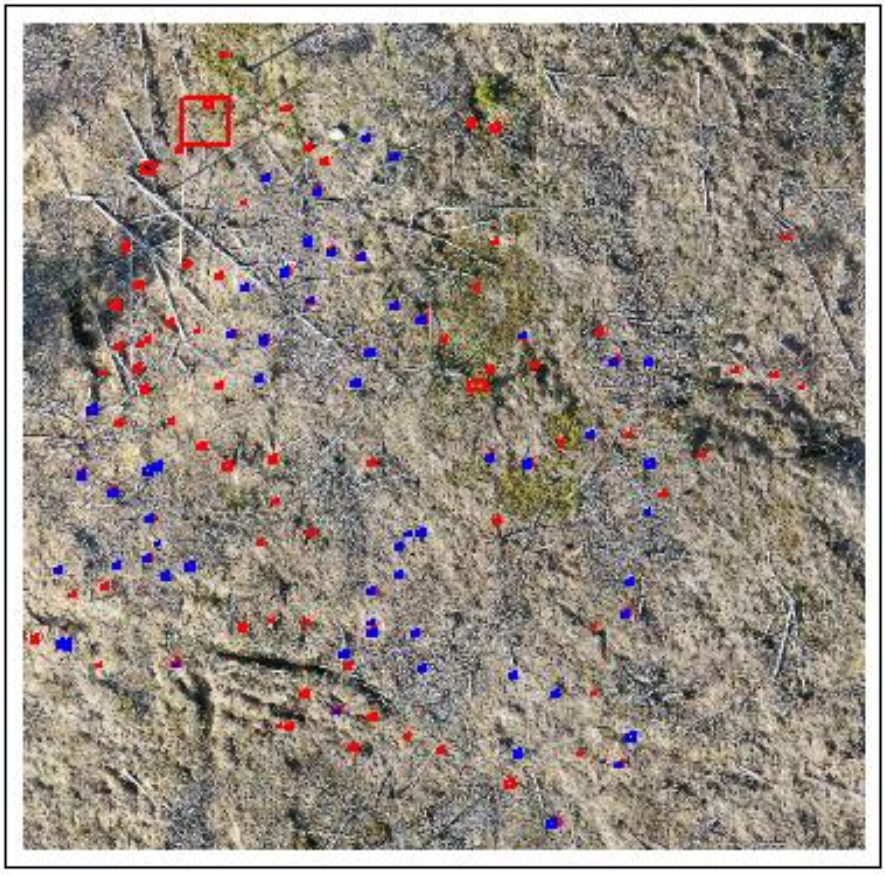


(b) All annotations

Figure 4: Poor detections at Braatan. Many larger trees close together missed by both detectors.

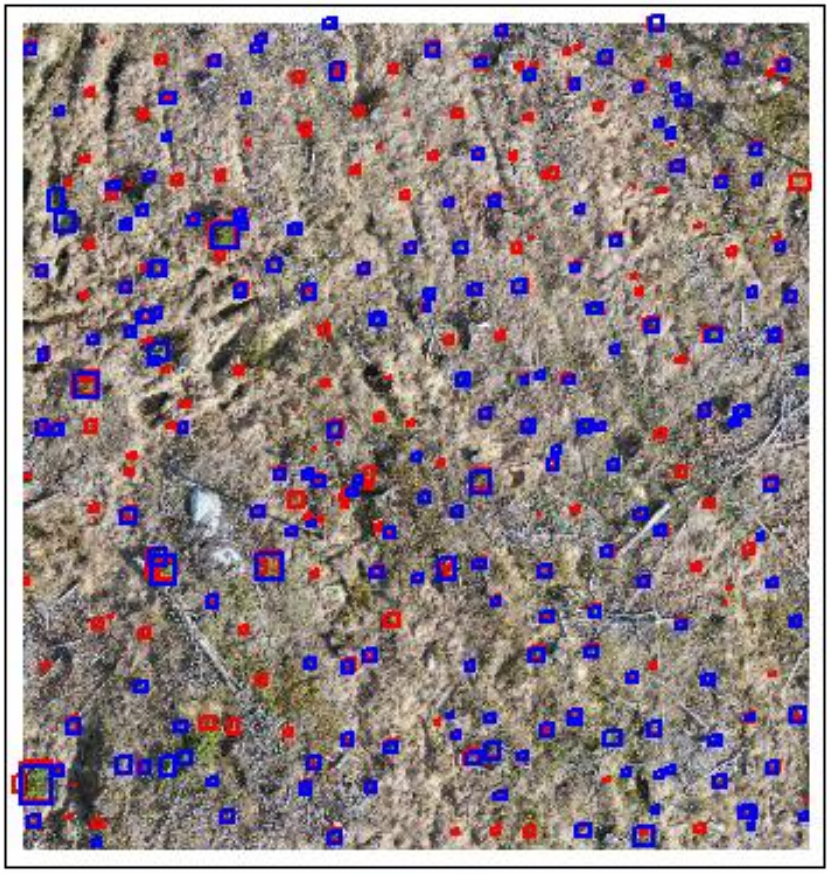


(a) My annotations

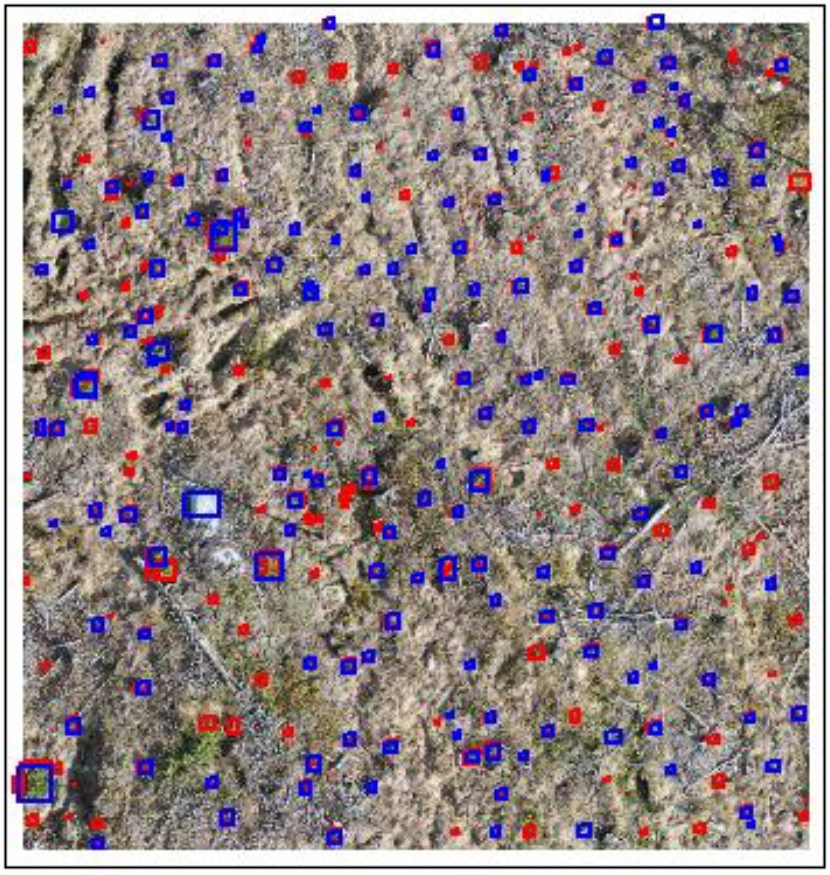


(b) All annotations

Figure 5: Poor detections at Braatan. Small saplings missed by both detectors (pine?)



(a) My annotations



(b) All annotations

Figure 6: Poor detection at site Galbyveien. Undetected smaller saplings.

Comparison of models

There was little difference in performance between the two sets of annotations. **My annotations** consisting of 48 images with 280 annotaitons performed similar to **All annotations** consisting of 387 images with 5062 annotations. Higher quality of annotations in **My annotations** may compensate for the decrease in dataset size compared to **All annotations**. The larger dataset size of **All annotations** might also be the reason for the *medium* size model performing best, compared to the *nano* model selected for **My annotations**.

Conclusion

- Quality of annotations is more important than quantity of annotations.
  - My set of 48 annotated images performed similar to the dataset of 387 images with annotations.
- Model size has the largest impact on model speed.

  - The effect of image size increases with model size.
- Training performance is not representative of domain performance.
  - **All annotations** performed worse than **My annotations** in training while the difference assessed by ML or domain metrics were minimal.