Semi- and weak-supervised learning for Norwegian tree species detection

Maritjn Vermeer[1], David Völgyes *[1], Tord Kriznik Sørensen[1], Heidrun Miller[2], and Daniele Fantin[1]

[1]Science and Technology AS
[2]Allskog SA

Abstract

Tree species mapping of Norwegian production forests is a time-consuming process as forest associations largely rely on manual interpretation of earth observation data. Deep learning based image segmentation techniques have the potential to improve automated tree species classification, but a major challenge is the limited quality and availability of training data. Semi-supervised techniques could alleviate the need for training label and weak supervision enables handling coarse-grained and noisy labels. In this study, we evaluated the added value of semi-supervised deep learning methods in a weakly supervised setting. Specifically, consistency training and pseudo-labeling are applied for tree species classification from aerial ortho imagery in Norway. The techniques are generic and relevant for the wider earth observation domain, especially for other land cover segmentation tasks. The results show that consistency training gives a significant performance increase. Pseudo-labeling on the other hand does not, potentially this is due to varying convergence speeds for different classes causing confirmation bias or a partial violation of the cluster assumption.

1 Introduction

For forest management the availability of complete, accurate and up-to-date forest inventories is essential. Typically, forest inventories store information about forest stands, which are roughly uniform areas within the forest that are managed as a single unit. One of the most important parameters of the forest stand is the volumetric tree species distribution. Within Norway, there are three main tree species used for production: Norway spruce, Scots pine and birch. Currently, the determination of the tree species distribution per stand is performed by a forestry expert, by visual interpretation of aerial imagery and in some cases LiDAR data. Tree species mapping is therefore expensive, error-prone and time-consuming, leading to forest inventories that are incomplete and/or outdated.

Deep Learning (DL) is getting ubiquitous in state-of-the-art land cover classification [4]. Previous approaches to tree species classification in Norway either used classic machine learning approaches [2] or are drone-based and therefore have limited scalability [6]. One of the main challenges of successfully implementing a scalable deep learning approach for tree species mapping in Norway is the limited availability and quality of labeled data. Limited quantity and quality labeled data is a common challenge within Earth Observation (EO).

In various domains semi-supervised learning techniques have proven to be successful. Two major branches within deep semi-supervised learning are consistency training and pseudo-labeling [7]. Consistency training is a method in which predictions for input samples are trained to be consistent with their noisy counterparts. [9] show that rather than adding noise, data augmentation methods can be used for consistency training. The augmentation methods that perform well in supervised learning are typically performing well in this semi-supervised setting as well. Pseudo-labeling is a method in which confident predictions are used as ground truth labels in the training process.

*Corresponding Author: volgyes@stcorp.no

https://doi.org/10.7557/18.6817

The available tree species training data are rough volumetric distributions of tree species at forest stand level. Weak supervision refers to techniques that address handling noisy and coarse-grained labels [10].

The main contribution of this study is the evaluation of the effectiveness of semi-supervised deep learning techniques for tree species detection in Norway from aerial imagery. Specifically, consistency training and pseudo-labeling will be evaluated in a weakly supervised setting. Both techniques are generic, and the results are therefore relevant for the broader EO domain.

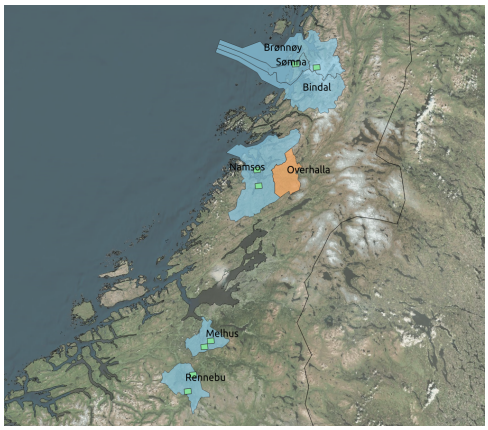

Figure 1: The study area consisting of 7 municipalities, with train, validation and test sets shown in blue, green and orange respectively.

2 Methodology

2.1 Study area and data splits

The study area of the project consists of 7 municipalities in the Trøndelag and Nordland regions, see Figure 1. The areas were selected based on the availability of forest inventory data. Within each training municipality some areas are left out for validation. One complete municipality is left out and used for independent testing. This spatial separation between train and test/validation is important in order to avoid overfit, otherwise the model could overfit on terrain features, e.g. valley orientation, instead of learning the features of the tree species. Table 1 gives an overview of the stand data in the different sets. Figure 2 shows an example of the 20cm resolution ortho imagery that is used overlaid with tree species stand data.

The image data is available at the Norge i Bilder website[1]. Label data is not directly accessible, it is property of Allskog AS, but similar labels are publicly available through the SR16 map[3].

Table 1: Training, validation and test data splits.

	Stands			km^2
	Spruce	Pine	Birch	
Train	30757	10474	4821	531.6
Val	3264	1009	471	44.6
Test	9817	1438	981	121.7

Figure 2: Closeup of the aerial imagery overlain with stand data colored by their dominant species: spruce (green) pine (brown) and birch (light green).

2.2 Deep learning approach

A five level deep U-Net [8] was used for predicting tree species at the pixel level, the main network and training settings are given in Table 2. Tree species training data was only available as a distribution of tree species per forest stand and not at pixel level. Therefore, a custom zonal loss was used which optimizes for correct tree species distributions on stand level rather than pixel level. It does so by averaging the pixel predictions on stand level for each class. This yields a distribution at zone (stand) level (Eq. 1), which can be compared to the ground truth label. The loss per zone (L_z) was calculated by applying a Huber loss (H) on the L_1 vector norms of the zone level difference between prediction (p) and label (l) (Eq. 2). The Huber loss has a slope of 1 above δ and below quadratically flattens out as $x \to 0$ (Eq. 3). This was designed to reduce overfit as a reasonable well trained network should get diminishing returns for optimizing an already well converged stand prediction. Finally, the zone level losses are weighted by zone size (N_z) and inversely weighted with class frequencies ($\frac{1}{f_z}$). This yields the final zonal Huber loss (Eq. 4).

$$p_c = \frac{1}{N_z} \cdot \sum_i p_{c,i} \tag{1}$$

$$\mathrm{L_z}(p, l) = \mathrm{H}(||p - l||_1) \tag{2}$$

$$H(x) = \begin{cases} \frac{1}{2\delta} x^2 & \text{if } x < \delta \\ x - \frac{1}{2}\delta & \text{if } x \geq \delta \end{cases} \tag{3}$$

$$\mathrm{L_{zh}}(p, l) = \sum_z^{zones} N_z \frac{1}{f_z} \mathrm{L_z}(p, l) \tag{4}$$

The zonal Huber loss can be considered a weakly-supervised technique. The zone level label can be a mix of different species, e.g. 70% spruce, 30% pine and 0% birch. Every pixel can only belong to one species, therefore the labels on pixel level are unknown. As a result cross entropy and similar loss functions cannot be applied, but the zonal Huber loss can.

Similar to the loss, performance metrics are also reported on stand level. The main performance metric is the macro averaged F1-score for the dominant tree species. For data augmentation hue, saturation, brightness and contrast were randomly changed up to a certain factor, see Table 2. The transformation factors approximately reflect the natural variation that was observed in the aerial imagery. For example, brightness is more strongly affected by atmospheric conditions and the time of the day than hue.

Table 2: Main network and training settings.

	parameter	values
unet	blocks	5
	kernels	16, 32, 64, 128, 256
	normalization	instance
	activation	ReLU
	input channels	3 (RGB)
	output channels	4
train	learning rate	1e-4
	optimizer	ADAM
	zonal huber gamma	0.4
augments	brightness	30%
	saturation	30%
	contrast	20%
	hue	10%

2.3 Semi-supervised losses

For consistency training we used a consistency loss (L_c), which is defined as the pixel based distance (D) between the prediction by the model (m) of the original image (x) and its augmented counterpart ($\mathcal{A}(x)$) (Eq. 7). Both the $L1$ and $L2$ distance were evaluated for the distance function (Eq. 5, 6). For pseudo-labeling a pseudo loss (L_p) is defined, which utilizes the focal loss (L_f) [5] for comparing the prediction with the pseudo-label at pixel level (Eq. 8). The consistency and pseudo loss were only applied on pixels where the prediction is above a threshold value (th). The total loss is the weighted sum of the supervised and semi-supervised losses (Eq. 9).

$$D_1 = ||p_c - p_c^{augmented}||_1) \tag{5}$$

$$D_2 = ||p_c - p_c^{augmented}||_2) \tag{6}$$

$$L_c = D(m(x), m(\mathcal{A}(x)) \mid m(x) > \text{th}) \quad (7)$$

$$L_p = L_f(m(x), \text{argmax}(m(x)) \mid m(x) > \text{th}) \quad (8)$$

$$L_{total} = L_{zh} + w_c \cdot L_c + w_p \cdot L_p \quad (9)$$

3 Experiments

The number of experiments is limited, due to resource/time limitations. Training a single model takes approximately 5 days on a GPU (nVidia RTX 3090, 24GB GPU RAM, 10496 CUDA cores). Hence, the following two stages of experiments were designed. Firstly, a hyperparameter search was performed to get sensible values for the semi-supervised loss parameters. Secondly, a series of experiments with the hyperparameters found in the first stage and a varying percentage of labeled data for supervision were performed. Both labeled and unlabeled data was available for the unsupervised learning.

For the first stage 10% of the labels was made available for supervised training. The reason for this relatively low share was to clearly see the effect of the semi-supervised methods. A supervised baseline model was trained using the 10% labeled data, not applying any of the semi-supervised losses. In addition, several semi-supervised models were trained, exploiting the unlabeled data as well by applying either the consistency- or pseudo loss. The hyperparameters are all related to the semi-supervised losses. For the consistency loss, these are: the weight of the loss, strength of the augmentations, the semi-supervision threshold and whether to use L1 or L2 loss for comparing distributions at pixel level. The augmentations we evaluated are brightness, contrast, saturation, hue and Gaussian noise. For the pseudo loss, the hyperparameters are only the weight of the loss and the semi-supervised threshold.

In the second stage 100%, 30%, 10%, 3%, 1% of the data is used for supervision respectively. For each of the data regimes a supervised baseline model is trained and a semi-supervised model using the parameters found in the first stage. The data regime selection is roughly logarithmic, evaluating the effect of iteratively dropping roughly 2/3 of the labeled data.

4 Results

4.1 Hyperparameter search

Table 3 shows the results of the hyperparameter search for the consistency and the pseudo loss. The default values for the pseudo are a loss weight (w) of 1 and a semi-supervised threshold (th) of 0.95. Alterations to the default parameters are shown in the table. Lowering the threshold and increasing the weight reduce the performance, vice versa increased threshold and lowered weight improve the performance. Compared to the baseline the improvement is not significant, especially since the models with the least contribution from the pseudo loss perform best. Hence, the pseudo loss is left out from the second stage of experiments. For the consistency loss the additional default parameters are the L1 loss, and for the augmentations 30% variation for brightness and saturation, 20% for contrast and 10% for hue. In the table we see that especially lowering the threshold and usage of L2 loss improve the results and to a lesser degree strengthening the augmentations and increasing the weight. The improvement compared to the baseline is significant. For the second stage of experiments the consistency training is further investigated. The eventual parameters selected for consistency training are L2 loss and a reduced threshold of 0.9. The weight and augmentations strength are not increased as performance could drop by applying all parameters that strengthen the consistency at once.

4.2 Consistency training data regimes

Figure 3 shows the results of the consistency experiments compared to the baseline models for the different data regimes. Overall we see an improvement in the order of a couple of percent in F1-score across the different data regimes. Note that even with 100% of the labeled data consistency training is beneficial, this is due to the sparse and zonal labels. The labels are sparse as they do not cover each window entirely, meaning some parts of the window are used exclusively for consistency train-

Table 3: Hyperparameter search results for pseudo-labeling and consistency training using 10 percent labeled data.

	settings	F1
Baseline		59.7
consistency HP search	default	60.2
	$th = 0.8$	64.0
	$D = L2$	63.7
	$\mathcal{A} = \mathcal{A} \cdot 2$	61.8
	$w_c = 10$	61.8
pseudo HP search	default	59.6
	$w_p = 10$	50.1
	$w_p = 0.1$	60.1
	$th = 0.8$	53.9
	$th = 0.98$	59.4

ing. Furthermore, the consistency loss acts on pixel level instead of zone level, giving more direct feedback to the network. Figure 4 shows the average offset between the ground truth and consistency model trained on the full dataset. For example if the ground truth label is 80% spruce, 10% pine and 10% birch, whilst the prediction is 100% spruce, the average offset is 13.33%.

5 Discussion

5.1 Limitations

As stated in 3 training times are long, hence only a limited number of experiments can be run. The reason for this is likely the weakly supervised zonal labels. The feedback the model receives from the zonal loss is only on stand level, whilst the classification itself is on pixel level. If pixel level labels would be available the feedback would be more direct and convergence speed would increase. Additionally, the tree species stand labels at hand are relatively noisy, further slowing down convergence speed. It should be taken into account that the reported performance is based on single runs from a single seed only. Some random chance can therefore be expected when looking at the results for specific settings. The overall trend is however that consistency training outperforms the baseline model in all 10 conducted experiments. The trained models are regional, such that they can only be reliably

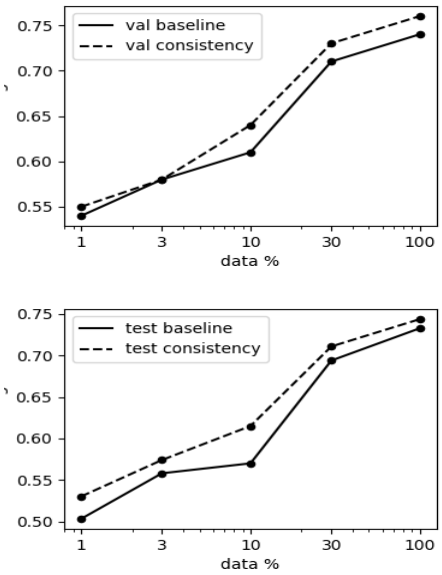

Figure 3: Macro averaged F1 score for dominant species for 1%, 3%, 10%, 30% and 100% labeled stand data. Baseline training is compared to consistency training. Top figure depicts the performance of the fully trained model on the validation set, and bottom figure shows the performance on the independent test set.

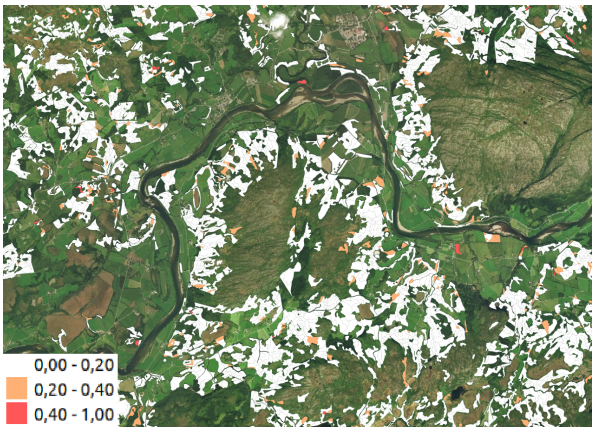

Figure 4: Tree species distribution average offset between ground truth and consistency model trained on the full dataset.

applied to Trøndelag and Nordland regions. Although the main tree species in Norwegian production forests are the same throughout the country, there are large variations in the geographic setting.

5.2 Pseudo-labeling

One of the main underlying assumptions vital for pseudo-labeling is the cluster assumption [7]. The input feature space should consist of clusters, if points are in the same cluster they are likely to be of the same class. Furthermore, the decision boundary should be in low density regions in order to clearly separate the clusters. For tree species classification from EO data it might be that such low density regions are not really present making it difficult to separate between the clusters. The forest can be an arbitrary mix of species, with the canopy showing patterns of different species simultaneously. Furthermore, it is observed during training that some classes converge before others. As pseudo-labeling happens dynamically during training an extreme imbalance occurs in the generated pseudo-labels, which is currently poorly addressed with initially set static class weights.

5.3 Consistency training

The consistency training enforces invariance between samples and their augmented counterparts. By doing so, it reduces the risk of over-fitting on the wrong patterns. The difference in performance between the validation and test set for the low data regimes could be explained by the fact that, the validation set is more similar to the training set than the test set. The validation set is composed of tiles left out in the training municipalities, whilst the test set is composed of a completely different municipality 1. Therefore, it is as expected that especially the 1% and 3% baseline models that are heavily over-fitted on the training set give better results on the validation than on the test set. A very promising result is that the 3% consistency model outperforms the 10% baseline model, looking at the test set. This shows the added value consistency training could have when little ground truth labels are available, whilst imagery is available.

6 Future work

The future aim is to develop a country-wide model, one of the main challenges in doing so is that training data is hard to get and often has restricted usage policies. Therefore, we believe that semi-supervised techniques are important to reach this goal. As discussed convergence speed is low, transfer learning could potentially speed up training cycles and in addition improve the performance. Out of the box pre-trained models are typically trained on standard (non-EO) imagery, which potentially does not translate well to EO imagery. Therefore, it would likely be beneficial to use specific EO and in particular land cover related pre-training tasks. The pseudo-labeling could be further investigated using dynamic class weighting for the pseudo loss, to avoid self-reinforcing class imbalance. Instead of dynamic weighting, iterative student-teacher setup with adjusted class balances is also a potential strategy. The consistency training could be improved by more extensive hyper-parameter tuning and introducing domain specific augmentations. Such augmentations could for example simulate atmospheric conditions or seasonal variation in the forest.

7 Conclusion

Consistency training has shown to be effective for tree species classification from aerial imagery, significantly improving performance and by such alleviating the need for a large amount of training labels. Improvement is on the order of 1-5% for the macro averaged F1-score, corresponding to a reduction in required training labels of up to a factor of 3. Pseudo-labeling on the other hand did not yield significantly enhanced performance. The suspected reasons for this is the class imbalance in the pseudo-label loss and nature of the problem with gradual transitions between tree species classes due to mixed forest. In general results are expected to be similar for other land cover classification tasks, which are often characterized by large natural variation within classes and gradual changes between them.

Acknowledgement

This work was funded by the European Space Agency (ESA) as part of the SENTREE project (ESA Contract No 4000136015/21/I-DT-lr) in the ,,ESA AO/1-10468/20/I-FvO FUTURE EO-1 EO SCIENCE FOR SOCIETY PERMANENTLY OPEN CALL" program.

References

[1] Norge i bilder. URL `https://norgeibilder.no/`.

[2] J. Breidenbach, L. T. Waser, M. Debella-Gilo, J. Schumacher, J. Rahlf, M. Hauglin, S. Puliti, and R. Astrup. National mapping and estimation of forest area by dominant tree species using sentinel-2 data. *Canadian Journal of Forest Research*, 51(3):365–379, 2021. doi: 10.1139/cjfr-2020-0170.

[3] M. Hauglin, J. Rahlf, J. Schumacher, R. Astrup, and J. Breidenbach. Large scale mapping of forest attributes using heterogeneous sets of airborne laser scanning and national forest inventory data. *Forest Ecosystems*, 8:1–15, 2021.

[4] T. Hoeser and C. Kuenzer. Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends. *Remote Sensing*, 12 (10):1667, 2020. doi: 10.3390/rs12101667.

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. doi: 10.1109/TPAMI.2018.2858826.

[6] S. Nezami, E. Khoramshahi, O. Nevalainen, I. Pölönen, and E. Honkavaara. Tree species classification of drone hyperspectral and rgb imagery with deep learning convolutional neural networks. *Remote Sensing*, 12(7):1070, 2020. doi: 10.20944/preprints202002.0334.v1.

[7] Y. Ouali, C. Hudelot, and M. Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020. URL `https://arxiv.org/pdf/2006.05278.pdf`.

[8] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

[9] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020. URL `https://arxiv.org/abs/1904.12848`.

[10] Z.-H. Zhou. A brief introduction to weakly supervised learning. *National science review*, 5 (1):44–53, 2018. doi: 10.1093/NSR/NWX106.