Contribution ID: **2** Type: **Session 2**

# SELF-SUPERVISED BACKBONES FOR FOREST REMOTE SENSING

Self-Supervised Learning (SSL) presents an opportunity to leverage large volumes of unlabeled data to improve outcomes for tasks for which training data is limited. This is the case in forest remote sensing where high quality annotated datasets are often too small to fully take advantage of data hungry deep learning models. Recognizing this, we have pre-trained two classes of neural network backbones that can be adapted to multiple use cases: different Vision Transformers (ViTs) (Dosovitskiy, 2020) and ResNet-50 (He et al, 2016). These backbones have been pre-trained on publicly available high resolution aerial image data from North Rhine-Westphalia (NRW), using two SSL frameworks: DINO (Caron et al, 2021) and MAE (He et al, 2016). We conducted four comparisons:
1. Pre-training on either remote sensing data or the ImageNet dataset, to assess whether pre-training with remote sensing data offers an advantage.
2. Pre-training the backbones on the NRW dataset to assess whether ViTs are superior to ResNet-50.
3. Training directly in a supervised manner to assess performance improvements due to pre-training.
4. Pre-training using DINO and MAE to assess the influence of the pre-training method.

The results can be summarized as follows:
- Pre-trained SSL models outperform supervised models on downstream tasks.
- Pre-training with NRW dataset offers improved downstream task performance over pre-training with ImageNet dataset.
- ViTs perform better than ResNet-50 at the pre-training task which translates to improved performance on downstream tasks.
- In terms of performance on downstream tasks, DINO emerges as a better pre-training method than MAE.

The outcomes of testing the models on three downstream tasks are summarized below:
- Classification: On a task involving orthophotos of 7334 trees across 10 species from Gartow (Germany), the Vision Transformer achieved the highest performance.
- Binary Semantic Segmentation: On a task involving 39 orthophotos with binary segmentation masks (tree and background) from Gartow, the Vision Transformer, coupled with a ClipSeg segmentation decoder (Lüddecke et al, 2022), achieved the highest performance.
- Multiclass Semantic Segmentation: On a task involving 47 orthophotos of forest across 15 tree species from the FORTRESS dataset (Schiefer et al, 2022), the Vision Transformer, coupled with ClipSeg segmentation decoder, achieves the highest performance.

Our experiments show that in the case of small scale datasets, pre-training with remote sensing data significantly improves outcomes in comparison to initializing models with ImageNet pre-trained weights. They further show that Vision Transformers outperform ResNet-50 and require fewer epochs to reach optimal accuracy. These results point to a possibility for building a full-scale foundational model, trained on even larger areas, that could help to improve diverse forest related remote sensing tasks.

**Primary authors:** JAZIB ZAFAR, Muhammad (Forest Inventory and Remote Sensing, University of Göttingen); Mr FREUDENBERG, Maximilian (Forest Inventory and Remote Sensing, University of Göttingen); Dr LÜDDECKE, Timo (Neural Data Science, University of Göttingen); Dr NÖLKE, Nils (Forest Inventory and Remote Sensing, University of Göttingen)

**Session Classification:** Results from the community