**PROJECT: Fast Tree Species Annotation via Unsupervised and Active Deep Learning**

**Project description:**

Vegetation distribution in the Alps is directly related to geomorphic processes, water availability, plant dispersal modes (e.g. animals, wind) and indirectly to human activity from agriculture to leisure, and tourism. Nevertheless, recent warming trends have begun to affect the limits and spatial structure of vegetation colonies in the Alps thereby threatening existing ecosystems in the upper altitudes. While these changes can be monitored locally, a region-wide characterization is needed to accurately model and forecast potential change scenarios. To address this need broad scale species distributions are required, accurately linking in-situ observations with Earth observation (EO) data.

However, Automatic classification of forest cover using high resolution data is still a challenging problem due to discrepancies in images resolution and appearance. Moreover, obtaining ground truth labels via In-situ species observations and/or manual classification of large scale EO datasets of forest cover down to the tree level is infeasible on the scale that is required to sufficiently train a model for such a task.

**Task description:**

The goal of this project is to address this object labeling challenge through the use of unsupervised and self-supervised approaches. Unsupervised approaches can identify pseudo-classes in the data based on a set of prior assumptions. These pseudo classes can then be labeled by a human annotator based on the segmentation of an unsupervised method, a process comparable to a refined form of Active-learning. Finally, these labels can be used to calibrate and refine the segmentation and classification algorithm, this time using semi-supervised learning. Similar approaches have already been used for basic labeling tasks in vegetation remote sensing, but it remains to be seen if they can be scaled to large regions and through incorporating airborne laser scanning and multi-spectral image data.

The main tasks in this project are summarized as follows:

* Review and identify relevant literature on unsupervised segmentation task in particular related to 3D data
* Implement the most promising method to perform the task of unsupervised point cloud segmentation
* Integrate segmentation results into an active/semi-supervised learning framework to obtain the final species labeled objects.
* The results should be prepared for visualization and sharing via standard GIS systems.

Inputs:

* 10cm GSD airborne orthophoto / LiDAR pointcloud ~ 20 pnts/m2
* ~700 in-situ localized tree species observations
* 10cm Imaging spectroscopy/multispectral image data

**Delivarables:**

* Report summarizing findings of the investigation
* Code implementation published to lab GitLab account
* Output data prepared in a format readable by a standard GIS software

[**Self-supervised learning**](https://www.v7labs.com/blog/self-supervised-learning-guide)**:** Self-Supervised Learning (SSL) is a Machine Learning paradigm where a model, when fed with unstructured data as input, generates data labels automatically, which are further used in subsequent iterations as ground truths. Those labels are generated by disturbing the content of the data that are feed and training the model to find the original back.

<https://www.youtube.com/watch?v=iGJ1XSkCyU0>

[**Semi-supervised learning**](https://www.v7labs.com/blog/semi-supervised-learning-guide)**:** Semi-supervised learning is a broad category of machine learning that uses labeled data to ground predictions, and unlabeled data to learn the shape of the larger data distribution.

The labeled data are used to do traditional supervised learning and the unlabeled ones are used to help finetune the decision boundary. Either by adding an unsupervised loss or by doing pseudo-labelling.

[**Structured vs unstructured data:**](https://www.g2.com/articles/structured-vs-unstructured-data)

**Structured data** is most often categorized as quantitative data, and it's the type of data most of us are used to working with. Think of data that fits neatly within fixed fields and columns in [relational databases](https://www.g2.com/categories/relational-databases) and spreadsheets.

Examples of structured data include names, dates, addresses, credit card numbers, stock information, geolocation, and more.

Structured data is highly organized and easily understood by machine language. Those working within relational databases can quickly input, search, and manipulate structured data using a relational database management system (RDBMS). This is the most attractive feature of structured data.

The programming language for managing structured data is called structured query language, also known as SQL. IBM developed this language in the early 1970s, and it is particularly useful for handling relationships in databases.

**Unstructured data** is often categorized as qualitative and cannot be processed and analyzed using conventional data tools and methods. It is also known as "schema independent" or "schema on read" data.

Examples of unstructured data include text, video files, audio files, mobile activity, social media posts, satellite imagery, surveillance imagery – the list goes on and on.

Unstructured data is difficult to deconstruct because it has no predefined data model, meaning it cannot be organized in relational databases. Instead, non-relational or [NoSQL databases](https://www.g2.com/categories/nosql-databases) are the best fit for managing unstructured data.

Another way to manage unstructured data is to have it flow into a data lake or pool, allowing it to be in its raw, unstructured format.

Finding the insight buried within unstructured data isn’t an easy task. It requires advanced analytics and high technical expertise to make a difference. Data analysis can be an expensive shift for many companies.

**QUESTIONS:**

* Is the idea to match the different species observation with the different clusters of 3D points?
* The methods that I found for “unsupervised” segmentation are actually **self**-supervised methods. Wouldn’t it be more adequate for 3D-points segmentation problems?
* The unsupervised method will identify pseudo-classes based on a set of prior assumptions. What are those assumptions?
* Are the image data supposed to be the support on which the clusters are placed and then compared with in-situ localized tree species observation? Or is it that the unsupervised segmentation method is going to be tested on the LIDAR pointcloud **and** on the orthophoto?
* Is the active/semi-supervised learning framework already defined?




* (Is the forecasting aspect part of the project?)

**Unsupervised / selfsupervised segmentation task:**

* [**STEGO**](UNSUPERVISED_SEMANTIC_SEGMENTATION.pdf) **(Self-supervised Transformer with Energy-based Graph Optimization)**
* [**POINT GCC**](POINT_GCC.pdf)**: Universal Self-supervised 3D Scene Pre-training via Geometry-Color Contrast**

**Semi-supervised learning:**