POLITECNICO DI TORINO

Master Degree course in Data Science and Engineering

Master Degree Thesis

Curriculum Learning for Satellite Earth Observation

Supervisors

Prof. Paolo GARZA

Prof. Daniele REGE CAMBRIN

Candidate

Fabrizio GALLO

ACADEMIC YEAR 2024-2025

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Chiara Succi, for her invaluable guidance, patience, and support throughout this process and my studies at ESCP Business School.

I am also very grateful to Professor Daniele Rege Cambrin for his helpful feedback and constant encouragement, which were so important in shaping the direction and quality of my work. As the supervisor on the Politecnico di Torino side, this research was possible thanks to his ability to connect technical knowledge with a practical business view, as well as the support of his department.

My thanks also go to the faculty and staff at Politecnico di Torino for creating a stimulating academic environment and for their dedication to education. I especially want to thank my classmates and colleagues during these two challenging and rewarding years. Their collaboration, discussions, and support enriched my experience so much.

I am deeply thankful to my sister, Elena, whose support never stopped and who always believed in me, giving me strength and motivation. I am also very grateful to my parents, Monica and Vincenzo, for their constant encouragement and for standing by me through all the ups and downs of my academic journey. They always accepted my decisions and supported me in everything I chose. A special thanks goes to my grandmother Maria Carmela, who always believed in my potential and never doubted that I could achieve great things, inspiring me to always aim higher.

I also want to thank my best friend, Pasquale, who never abandoned me despite our different paths, and with whom I continue to share passions and goals for the future.

I am equally grateful to all my friends from my studies in Turin, Paris, and Berlin, for their constant support and belief in me during this challenging journey. Their encouragement gave me the strength to persevere.

Finally, I extend my gratitude to all the professionals and industry experts who generously gave their time and knowledge, making this research more relevant and complete.

This thesis would not have been possible without the help and support of all these people.

Thank you.

Abstract

This thesis explores the application of Curriculum Learning to the classification of satellite images, a key challenge in the field of Earth Observation. Satellite imagery provides an invaluable source of information for analyzing Earth's surface, yet its complexity, volume, and diverse data formats make it difficult to process efficiently with conventional machine learning techniques. Deep Learning models, particularly Residual Networks (ResNet), have shown promise in image classification tasks, but often require extensive labeled datasets and high computational resources.

This research analyzes how Curriculum Learning, an approach that tries to follow the human learning process by introducing training samples from simple to complex, can improve the efficiency and effectiveness of training models on satellite data. The study begins by reviewing the fundamentals of supervised and unsupervised learning in image classification for Earth Observation and then highlights the difficulties in satellite imagery, including data volume, variable quality, class imbalance, and domain specific labeling needs.

The thesis then focuses on how Curriculum Learning can address these challenges by structuring the training process to start with clearer, easier-to-label images and progressively introducing more complex data. This structured approach not only improves generalization but also enhances labeling efficiency and model robustness.

The research culminates in the proposal of a structured Curriculum Learning framework, integrating modern deep learning architectures such as ResNet and Vision Transformers, tailored for satellite image classification. The thesis also outlines potential future directions, including self-supervised learning, dynamic difficulty scheduling, and multimodal learning.

Contents

Chapter 1

Introduction

In this master thesis, we explore how machine learning techniques can be applied to classify satellite images, highlighting the main challenges that need to be addressed to achieve accurate and robust results. A particular focus is given to a relatively recent paradigm known as *Curriculum Learning.* Satellite imagery represents a complex type of data, often stored in specialized formats that are not easily interpretable by either the human eye or a standard computer. Therefore, statistical models and advanced algorithms are required to extract meaningful patterns and enable the recognition of land cover and scenes on the Earth's surface.

In the early stages of satellite image recognition research, simple techniques were employed to detect specific features from these datasets. Initial studies often relied on methods such as region-based segmentation, edge detection, and similar approaches [3]. These techniques mainly used deterministic algorithms or basic learning-based procedures. In contrast, this work focuses on modern statistical models that leverage machine learning to automatically identify relevant image features. Over the years, several algorithms have been introduced for classification tasks, including *Random Forests*, *Support Vector Machines*, and different forms of *Neural Networks*, particularly *ResNets and ViTs*, which are especially effective for satellite image classification.

In this thesis, partilcular use cases of Convolutional Neural Networks are utilized under supervised and unsupervised learning settings. The central goal, however, is to emphasize the importance of applying *Curriculum Learning* to the task of satellite image classification. This approach trains models progressively, by introducing data of increasing complexity in a step-by-step manner, mimicking a structured human learning process.

Before proceeding with the main analysis, we briefly introduce the principles of supervised and unsupervised learning to provide background knowledge for the reader. Although these paradigms will be examined in greater depth in later chapters, a short overview is provided here:

- **Supervised learning:** The model is trained on labeled data (for instance, images associated with specific information such as their geographical location). Through these labeled examples, the algorithm learns to identify relevant patterns and can then generalize its predictions to new, unseen data.

- **Unsupervised learning:** The model is trained on unlabeled data and must autonomously uncover underlying structures, such as clusters or relationships between samples. In this way, it effectively generates new groupings or representations of the data for the first time.

This distinction is important in the subject of Curriculum Learning since one of the most notable blockers come from partially unlabeled or completely unlabeled data. The research first examines methods based on the supervised learning approach and subsequently explores possible unsupervised strategies to compare their respective performance. The ultimate objective is to demonstrate the effectiveness of machine learning in the context of satellite image classification. The thesis is written to be accessible even to readers without a deep technical background, ensuring that the core ideas and results remain clear throughout the discussion.

1.1 Earth Observation context

Satellite imagery plays a fundamental role in the classification of images for Earth observation purposes. The field of *remote sensing*, which focuses on gathering information about the Earth's surface without direct physical interaction, is tightly connected to advancements in *Earth Observation* technologies. These systems make it possible to capture highly detailed information about both the surface and the atmosphere directly from space. The resulting images provide valuable insights that support researchers, policymakers, and institutions in monitoring temporal changes, identifying anomalies, and making informed, data-driven decisions [11]. Examples include tracking deforestation rates, evaluating agricultural productivity, or studying the evolution of ice sheets-tasks that have become significantly more accurate and efficient thanks to satellite data analysis [2].

In recent years, the number of satellites launched by governments and private companies has increased, leading to an rapid rise in the amount of available images and datasets. This continuous data growth offers new opportunities but also introduces major challenges, particularly in the efficient processing, annotation, and interpretation of satellite data.

1.2 Challenges in labeling high-resolution satellite imagery

Labeling satellite imagery is both computationally and financially intensive. Modern deep learning models generally rely on large-scale labeled datasets and considerable computational resources. The process entails significant expenses, including hardware such as GPUs, datacenter infrastructure, and energy consumption, as well as software and labor costs related to annotation and model training. Beyond these direct costs, several major challenges are commonly encountered:

- **High data volume:** Contemporary satellite constellations can generate terabytes of imagery each day. Manually annotating even a small portion of this data requires substantial human effort and can quickly become cost-prohibitive. While automated or semi-automated labeling techniques can reduce manual workload, they often introduce inaccuracies that must later be corrected, increasing the overall processing burden.

- **Variable image quality:** The quality of satellite data can fluctuate considerably due to a range of factors, including atmospheric conditions (for example humidity or cloud cover), differences between sensors (such as resolution and spectral sensitivity), and variations in lighting. These inconsistencies introduce noise that complicates the labeling process and increase the likelihood of misclassification.

- **Class imbalance:** In many real-world applications, certain categories are underrepresented compared to others. For instance, urban areas may occupy less surface area in an image than vegetation or water bodies. Such imbalance can lead to biased models that perform poorly on minority classes unless these disparities are addressed during both labeling and training stages.

- **Need for specialized expertise:** Accurate labeling often requires domain-specific knowledge, such as understanding of land cover types, vegetation characteristics, or features observable only in specific spectral bands. Non-expert annotators may struggle to correctly label subtle or complex features, which can result in inconsistencies or errors that must later be reviewed and corrected.

1.3 The concept of Curriculum Learning and its relevance

Curriculum Learning is a way to feed the algorithm of Machine Learning where the model is exposed to data in an organized, step-by-step manner. Starting with a defined "easy" pull of examples, gradually the algorithm proposes more complex data to be fed. Inspired by the way humans learn, from simpler concepts to more complex ones, this approach guides the model build robust feature learning and analysis, reducing the risk of being stuck in complex attribute characteristics in the beginning of the learning process [21].

In the context of labeling satellite images, curriculum learning shows results for several reasons:

- Progressive complexity:

  By initially training on clearer and less ambiguous satellite images, the model learns the important features in the beginning. Once it has these core information mastered, it can tackle more complex cases step-by-step more effectively.

- Better generalization:

  Curriculum learning can help the model generalize better. In fact, using complex images the noise and complexity can confuse the learning process. Introducing data with more and more information gradually lets the model adapt and refine its learned features.

- Labeling efficiency:

Curriculum Learning can solve the human interaction problem when it comes to label the dataset. By using step-by-step learning, the human annotator can rapidly identify, when the labels become more complex, the general patterns thanks to previous well trained steps with less complex images, and bring small changes only there-after. The previous level complexity of the images guide not only the model, but also the human correction in next steps.

1.4 Technical implications of Curriculum Learning in Earth Observation

While Curriculum Learning has primarily been explored in the domain of artificial intelligence and computer vision, its implications can be extended beyond technical performance improvements. In the context of Earth Observation, the strategic and operational benefits are many. By integrating Curriculum Learning into the classical machine learning processes areas like performance, time of deployment, computational resources and labor costs will have a benefit.

1.4.1 Speed and scalability

As mentioned in the literature, one of the principal operational challenges in remote sensing and Earth observation projects is the data costs. The cost lays both in saving and processing this kind of voluminous sets. Despite great improvements in technological capabilities, handling this type of data on cloud (currently the most used way to train machine learning algorithms) is not yet efficient [20]. Since traditional algorithms rely on randomly sampling the dataset, GPU hours on inefficient training are spent. Hence costs of cloud infrastructure rise and hardware can easily depreciate. Curriculum learning, among other solutions presented in literature, can be a solution to this challenge. It offers a structured training approach that uses simpler examples first at a lower costs and higher efficiency. Moreover, Curriculum Learning is linked to bring faster convergence that cuts the costs of infrastructure even more.

In current economic situation, the demand for Earth observation is growing rapidly in many industries, such as agriculture, urban planning, environment monitoring and others that were cited in this master thesis. This growth impacts directly on the strategies needed to facilitate scalability on the company processes to provide the use of this data. In this scenario, Curriculum Learning can help machine learning model to be deployed more efficiently in various and different platforms. The need to train a model with long times of implementation and the necessity of specializing the algorithms can be not longer an issue. Thanks to Curriculum Learning, the model can be trained with a smaller amount of images that are generally needed to learn the classification, and then specialized only on the complex images suited for the sector of use. This will enable companies to scale their services to new markets with a greater confidence and scalability.

1.5 Goal of the thesis and first Curriculum Learning framework proposed for Earth Observation

As we discussed in the introduction on this thesis, the volume of satellite images is representing both an opportunity and a challenge for the scientific world and the social impact it can bring. The fast increase of satellites, that are built specifically for the job of Earth Observation, and the new levels of depth of resolution (new spectral layers are added every time to analyze different aspects of Earth surface and more), is pushing the boundaries of the amount of data that is being stored and made available to the scientific community. We saw, however, that there is an equal pressure in computational demand, especially because of the training of deep learning models that require vast amount of labeled data and extended training time. For this reason, this thesis will aim to investigate deeply the effectiveness of Curriculum Learning in a general point of view and, especially, applied to Earth Observation, as a mean to improve the efficiency and efficacy of training models in a such expensive but promising context.

Since Curriculum Learning is inspired by human learning process, this promoted an organized and structured rethinking of training the Machine Learning models. The process starts from simple examples, going towards more complex examples. However, this promising restructuring has not being studied deeply in all the potential use cases where it can be leveraged. This thesis will be a guide and review on how Curriculum Learning is built, analyzing deeply the theory behind it, to construct a process to be used for the context of high-resolution and complex Earth Observation datasets. By breaking down the methodology of Curriculum Learning and studying the literature behind Earth Observation techniques, this work will examine whether structuring the training process through curriculum strategies can possibly have an impact on resources, time and results.

A primary goal of this research is to assess how the ordering of training samples could affect the learning process of deep neural networks applied to satellite images. This thesis will propose, among other things, objective criteria for ranking image labels based on their complexity and elements contained, aiming to construct a learning process that allows the model to start with easy images and slowly growing difficulty of batches. This progression is hypothesized to help the model establish strong initial representations, making it more robust and efficient as training proceeds.

To evaluate the benefits of CL, the literature usually compares models with traditional random sampling techniques. The criteria that usually are set to quantify performance are:

- Convergence speed:

  How quickly the model reaches an acceptable level of performance during training.

- Computational Efficiency:

  The reduction in GPU time and memory usage across different training regimes. Metrics that quantify the efficiency are usually loss measures. However this results to be difficult to use. While loss can be measured at each epoch, different models use different magnitudes, loss functions and different images, especially in the case of CL, where images are clustered in different ways based on the Curriculum strategy. This results in different curves that can be compared only if the models are trained with same images and loss functions.

- Data Efficiency:

  Whether a model can achieve comparable or better accuracy using fewer labeled samples, addressing the challenge of limited or partially annotated datasets. This in particular can be easily evaluated in curriculum settings measuring the amount of patches used to reach similar performances in accuracy and generalization.

- Generalization Performance:

  The extent to which trained models maintain accuracy on unseen test data and across varying image types. This is a general metric that works in settings of CL as much as other settings. Earth Observation suffers generalization performance due to very broad contexts that are present in the surface of Earth.

To evaluate the potential of Curriculum Learning, we will analyze some tests made on more standard techniques used to classify satellite images. Furthermore we will analyze different techniques present in literature that are under the dome of Curriculum Learning techniques [21] used in other contexts. Thanks to this different approaches we can find commonalities with already studied techniques we can already predict the potential benefit of Curriculum Learning for Earth Observation.

In summary, this research will:

1. Benchmark the performance of conventional training methods across multiple metrics compared to Curriculum Learning, after a careful review of the theoretical basis.

2. Design a possible Curriculum Learning technique tailored to satellite image classification tasks.

3. Understand next steps in use cases and impacts of Curriculum Learning in Earth Observation.

Chapter 2

Background on Earth Observation and Machine Learning

Earth Observation is an essential pillar for many sectors, starting from urban planning to ecosystem management, to environmental monitoring. It is based on satellite images of Earth surface that provide multi-spectral (hence different layers of photography that most of the times go beyond the visible spectrum) and high-resolution images. Since the amount of the satellites has been increasing exponentially, the domain of usage is expanding. This brings, as we presented in the introduction, also extensive challenges on the techniques to leverage the information.

To address these challenges, Machine Learning, and in particular Deep Learning, has been increasingly adopted as a method to enhance Earth Observation learning. The techniques, mentioned in the literature [17], comprise of various usages depending on the application. They space from image classification, object detection or semantic segmentation. In this particular thesis, the image classification will be the focus of the research and analysis. In literature, Machine Learning techniques space from Neural Networks, Convolutional Neural Networks, but also simpler algorithm such as Random Tree Forest, Support Vector Machines, Maximum Likelihood for supervised techniques have been used, while K-Nearest Neighbors or ISODATA are implemented for unsupervised techniques. Beyond this basic methods, researchers often differentiate object-oriented techniques and pixel-based as other approaches. The thesis the Convolutional Neural Network will be used as main method to focus on object-oriented analysis, and will be analyzed further in the next chapters. In the specific background of this thesis, ResNets and ViT will be tested against remote sensing images, with the intent to discover way to make training fast, reliable, efficient with Curriculum Learning.

This chapter will provide a structured and deep overview of the Earth Observation science, to then dive deep on the its Machine Learning integration pipeline. The research will cover the types of satellite data used, the preprocessing required and used in the scientific community, a guide on Machine Learning techniques and the usage on satellite images. Understanding the capabilities and limitations of these methods is key to understand how it can be leveraged and how Curriculum Learning can play a pivotal role.

2.1 Introduction to Earth Observation

Earth Observation refers to the collection, analysis, and interpretation of data about the Earth's surface and atmosphere through remote sensing technologies. Remote sensing is the technique of acquiring images from a distance of earth surface, in which the main acquisition platforms are satellites or aircraft. We will focus on satellite images for the purpose of the thesis. Thanks to advancement in technology involved in this field, both in longevity of satellites, better photography apparatus and computer processing, the relevance has seen unstopped growth. Extensive selection of uses of satellite images has been mentioned in this thesis and it includes both civil and military use cases.

The technology behind satellite image acquisition relies on Electromagnetic (EM) energy emanation [18]. Using this energy field, the satellite leverages the propriety of all surfaces to reflect the EM signal in different ways based on their proprieties. For each different reflected wave, it correspond a different view of the object, hence a different layer. For satellites, the main source of EM field is the sun since it provides natural source of visible and infrared-ultraviolet frequencies. A rappresentation of the image patches creation can be seen here 2.1.

In the images that will be retrieved by the satellite, each layer will have a different number of spectrum-layers, that brings important information but also great complexity in the dataset. The satellites relies on two or more sensors to read the received the reflected EM waves. Based on weather the EM field is created by the sun or the satellite, the sensor will accordingly defined passive 2.9 or active 2.3.

2.1.1 The need of remote sensing and Earth Observation

In this section, we will understand the usages and the reasons that brought Earth Observation to become popular in some researches of Machine Learning and statistical learning.

Most notably, the Earth observation remains as an important field for the academic landscape. Apart from the numerous papers mentioned in this thesis, many more show the growth of publications and citations that fuels both the scientific community and the economic landscape.

Despite the advances and usage, the use of Earth Observation is strongly limited by the complexity of its data. High dimensionality, extensive labeling need, and computational intensity pose as barriers for the research and use. Satellites provide great variability of sensor types, image quality and in numerous atmospheric conditions with unlimited geographic contexts, complicating the processing and usage even more. Hence, Machine Learning plays a pivotal role to make the field more accessible, and Curriculum Learning poses itself as one of the chances to make it possible.

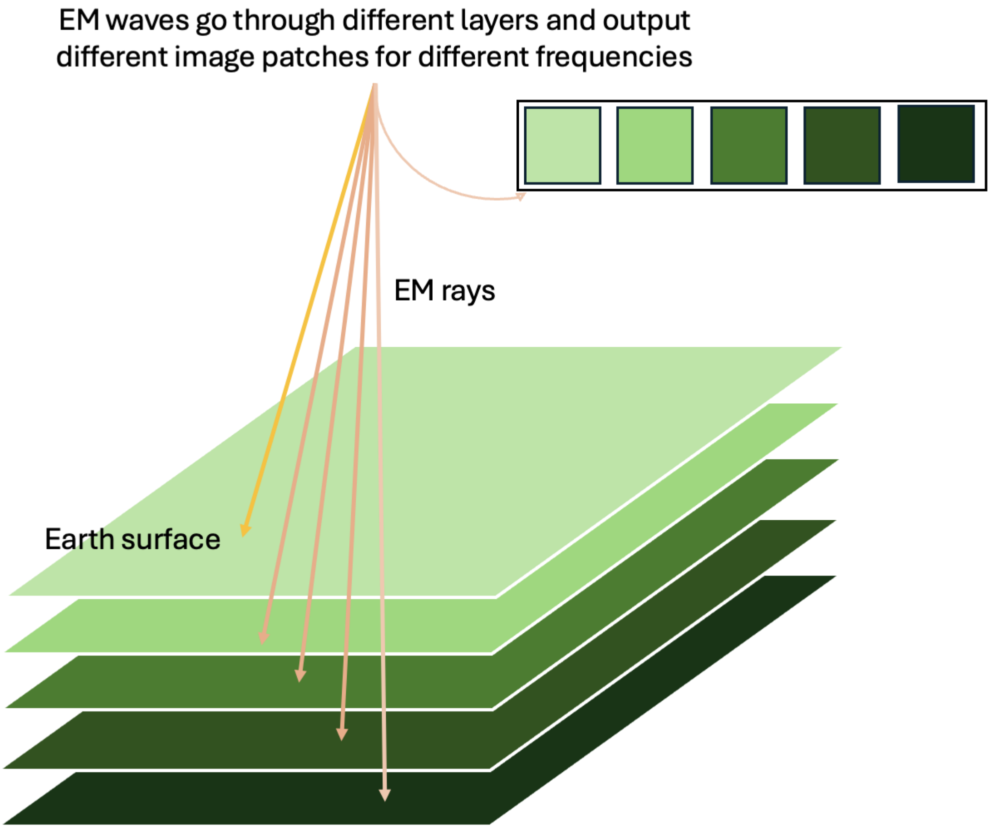
EM waves go through different layers and output different image patches for different frequencies
EM rays
Earth surface

Figure 2.1. Image patches creation via different frequencies

2.1.2 Satellite image data in Curriculum Learning: characteristics and challenges

The different datasets that are available for satellite image classification range in size, image type and resolution. For the purpose of this thesis, only some publicly available datasets will be used and will be explained to help the course of this research. The datasets will be used for the thesis discussed here are: BigEarthNet, EuroSAT, and SSL4EO. All three are derived from the Sentinel-2 satellite constellation but vary considerably in their size, spatial resolution, number of channels, and primary application. A concise overview of each dataset is provided below, followed by more detailed descriptions.

Since the mentioned datasets come from Sentinel-2, a brief table will show all the bands captured by the satellite. Some of the datasets may use all the bands, while others use only smaller spatial resolution for the purpose.

- **BigEarthNet**

  BigEarthNet is a large scale dataset containing Sentinel-2 image patches collected

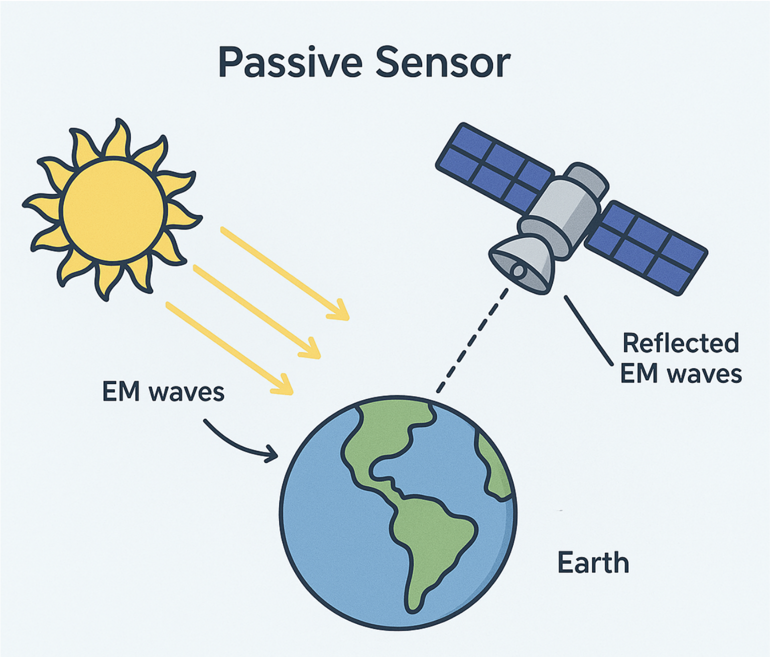

Figure 2.2. Passive sensors example

from multiple European regions. Each image measures 120×120 pixels and contains 10 distinct spectral bands (layers), with visible and infrared wavelengths. A characteristic that distinguishes BigEarthNet is its multi-label nature, since each patch may contain multiple classes, hence complexity. This setup creates challenges, since the model needs to handle class imbalance and scale the classification algorithms, considering also all the difficulties that we mentioned, making BigEarthNet an excellent test for evaluating the robustness and adaptability of different Machine Learning and Deep Learning models.

Detailed description:

- *Size:* Over 590,000 pairs of image patches
- *Resolutions:* image resolution of 120×120 pixels, spatial resolution of 20m or 10m
- *Source satellite:* Sentinel-2
- *Number of channels:* Originally 13 spectral bands but only 10 are used to guarantee quality of the dataset, excluding the 60m resolution patches.
- *Usage:* Multi-label land-cover classification across Europe which serves as a benchmark for class imbalance and large-scale image analysis

In the example taken from of BEN dataset here 2.4

- **EuroSAT**

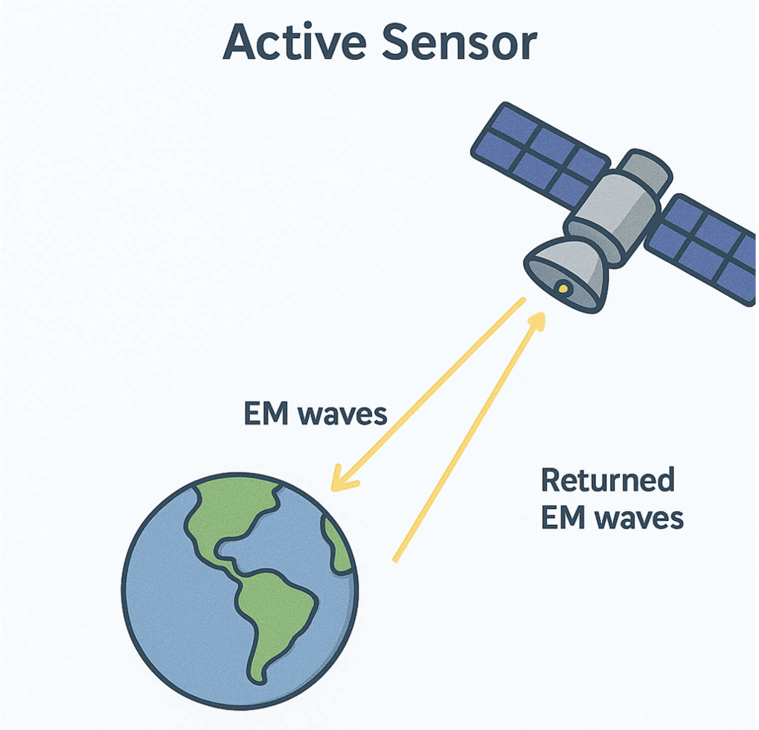

Figure 2.3. Active passive sensors example

EuroSAT is a smaller in size dataset with around 27,000 Sentinel-2 image patches, each with a dimension of 64×64 pixels and similarly describing 10 different spectral layers. Unlike BigEarthNet, EuroSAT assigns only a single label to each image, dividing the dataset into 10 and balanced classes. Due to its smaller and more balanced composition, EuroSAT is usually choice for an initial testing and prototyping, hyperparameter tuning, and quick benchmarking of new algorithms. Hence it can be used to test an initial setup of Curriculum Learning pipeline.

Detailed description:

- *Size:* Approximately 27,000 image patches
- *Resolutions:* image resolution of 64×64 pixels, spatial resolution of 10m
- *Source satellite:* Sentinel-2
- *Number of channels:* 13 spectral bands
- *Usage:* Balanced, single-label classification of 10 distinct land use and land cover categories; frequently employed for rapid benchmarking and transfer learning

Example of EuroSAT dataset here 2.5

Table 2.1. Spectral bands of Sentinel-2

Band	**Name**	**Wavelength (nm)**	**Spatial resolution**	**Typical use**
B01	Coastal aerosol	443	60m	Aerosol detection, atmospheric correction
B02	Blue	490	10m	Water body analysis, atmospheric correction
B03	Green	560	10m	Vegetation monitoring, urban area analysis
B04	Red	665	10m	Vegetation health assessment, land cover mapping
B05	Red Edge 1	705	20m	Chlorophyll content estimation, vegetation stress
B06	Red Edge 2	740	20m	Biomass estimation, crop monitoring
B07	Red Edge 3	783	20m	Additional vegetation indices
B08	Near Infrared (NIR)	842	10m	Vegetation vigor, biomass mapping
B8A	Narrow NIR	865	20m	Crop analysis, canopy structure
B09	Water vapor	945	60m	Atmospheric water vapor estimation
B10	SWIR-Cirrus	1375	60m	Cirrus cloud detection
B11	SWIR 1	1610	20m	Soil moisture, burnt area mapping
B12	SWIR 2	2190	20m	Geology, soil and snow discrimination

- **SSL4EO**

  SSL4EO (self-supervised learning for Earth Observation) is a repository aimed at fostering advancement unsupervised and semi-supervised techniques in the remote sensing domain. The dataset contains a large pair collection of Sentinel-2 image patches, with a range of size of 264×264, with labels being either incomplete or entirely absent. Its extended geographic coverage and significant variability in seasonal and atmospheric conditions are needed for stronger model generalization. Furthermore, SSL4EO's variable structure is particularly usable for exploring Curriculum Learning methods, since it allows incremental labeling and more complex addition to labels and classes to train the algorithm.

  Detailed description:

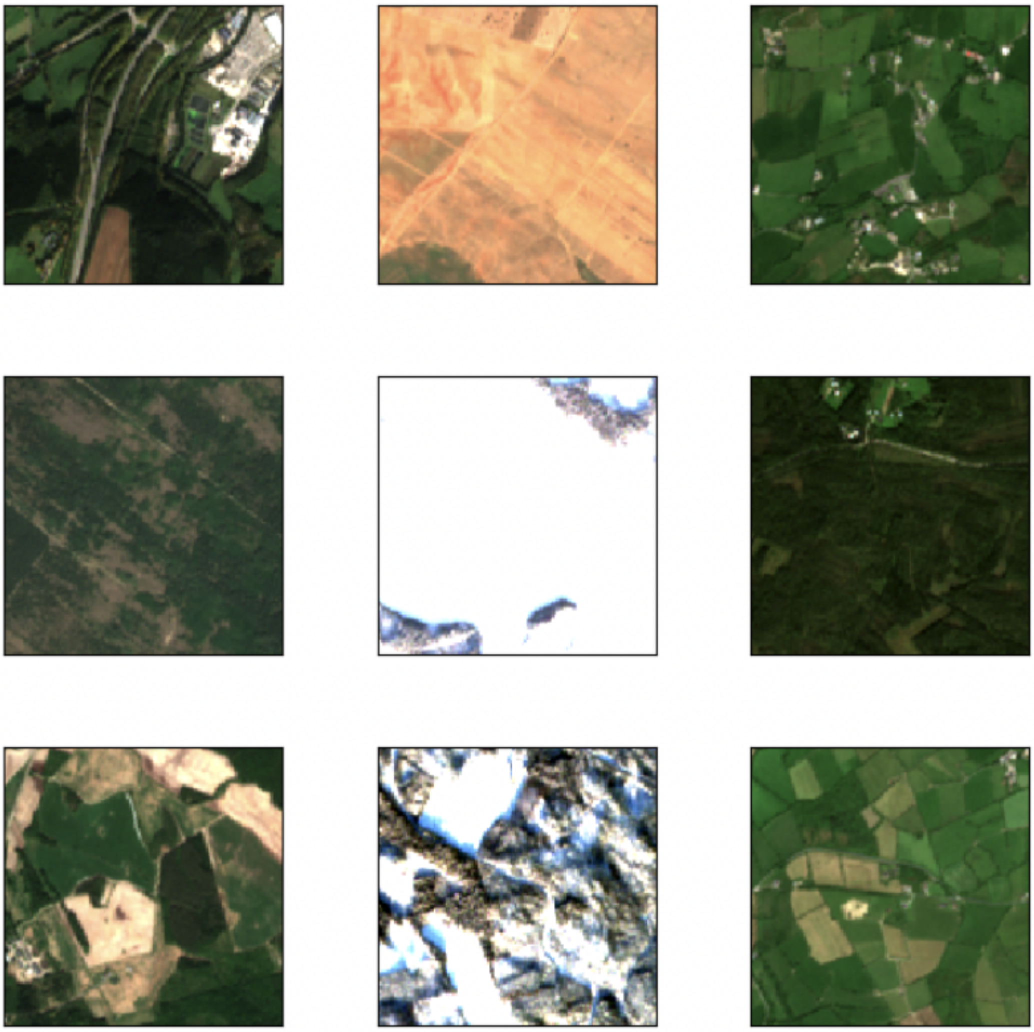

Figure 2.4. BigEarthNet example of patches.

- *Size:* Hundreds of thousands of unlabeled or partially labeled patches
- *Resolutions:* Image resolution of 264×264, spatial resolution of 10m for SSL4EO-S12 and 30m for SSL4EO-L
- *Source satellite:* Sentinel-1 and Sentinel-2
- *Number of channels:* 13 spectral bands
- *Usage:* Designed for self-supervised and semi-supervised experiments, encompassing wide-ranging geographic areas under diverse atmospheric and seasonal conditions

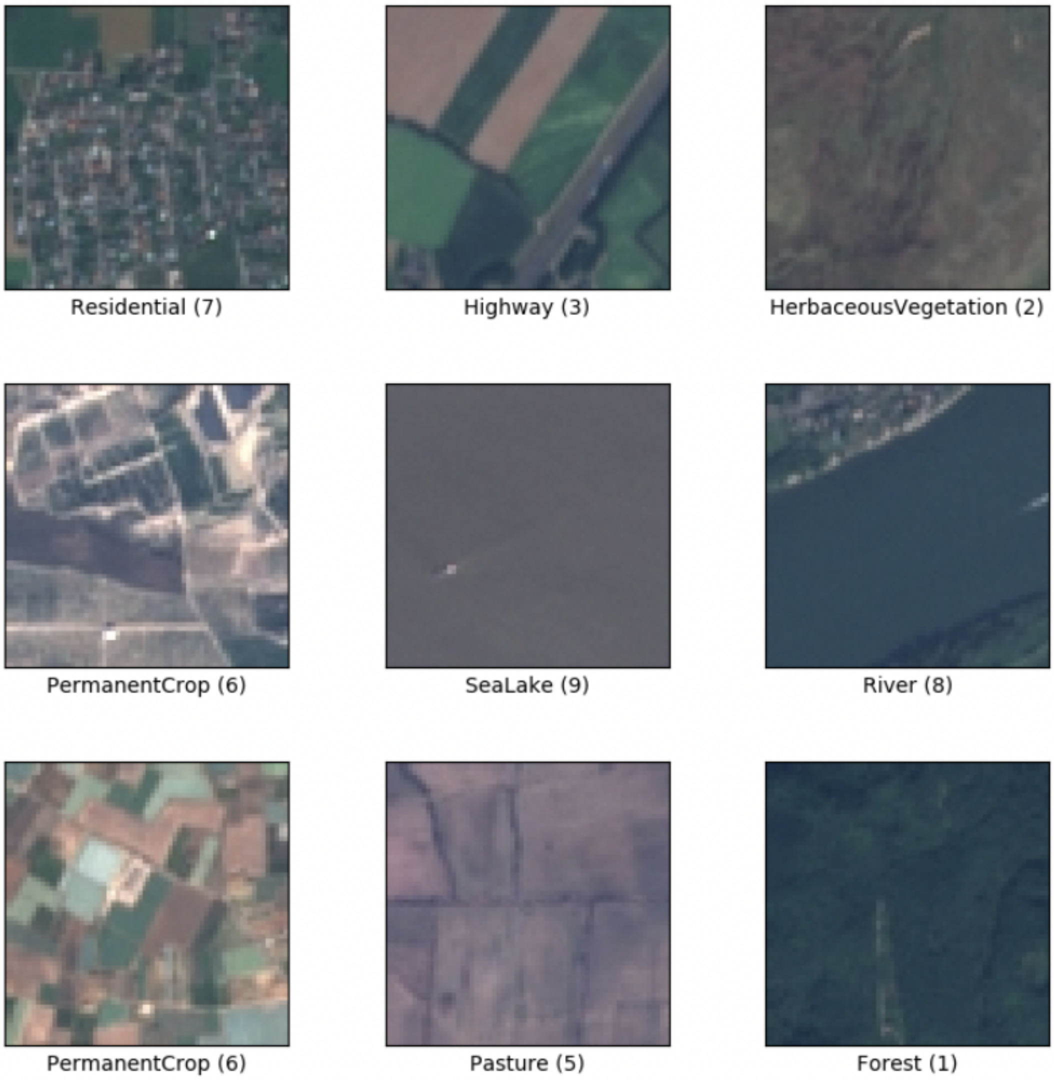

Figure 2.5. EuroSAT example of patches.

Example of SSL4EO dataset here 2.6

2.1.3 Role of Machine Learning and Deep Learning in Earth Observation

Machine Learning and Deep Learning play a pivotal role in Earth Observation as it changes the capabilities of analysis, use-cases and scalability drastically. After having described and analyzed some of the most popular datasets for statistical learning that

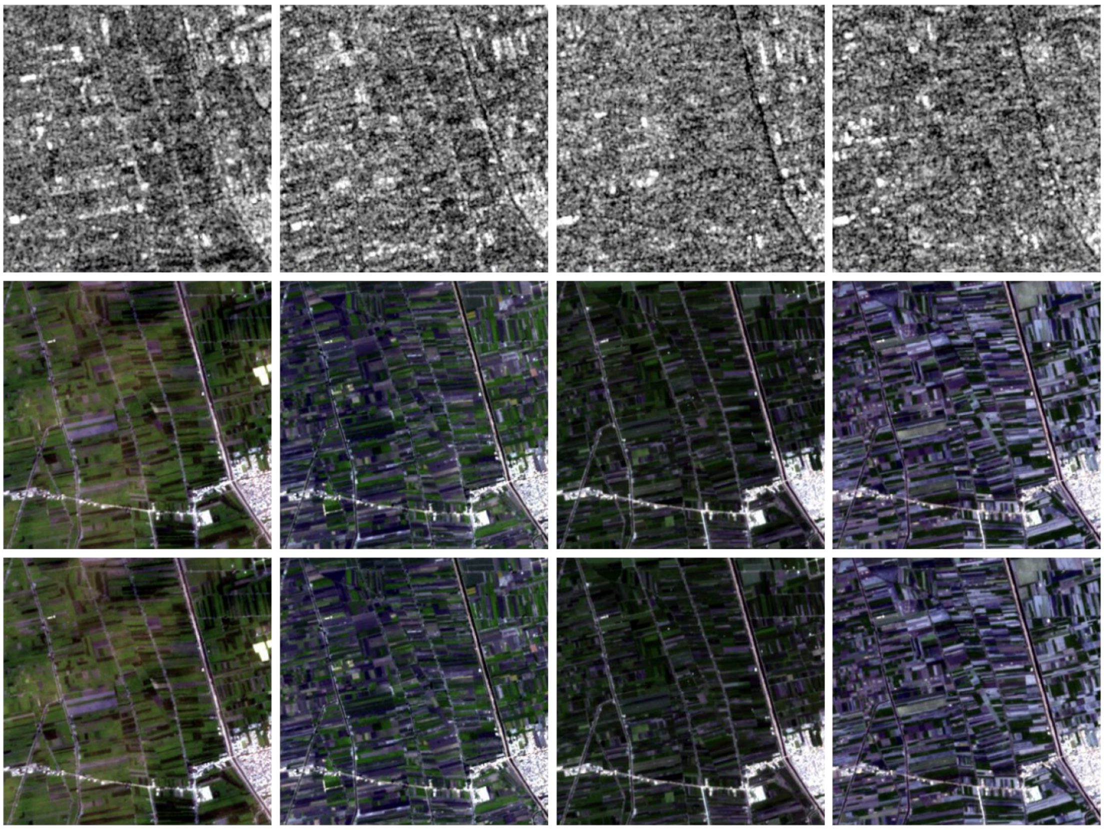

Figure 2.6. SSL4EO example of patches [24].

Sentinel-2 provides, it is clear that using more manual or deterministic approaches to label data is not viable to handle scalability, cost and efficiency. On the contrary, Machine Learning and Deep Learning are made to exactly handle these tasks: extracting statistical patterns, classify land cover types despite the variability, detect changes even in non visible contexts, and label new data thanks to trained models.

To understand deeply the role of Machine Learning and Deep Learning, this paragraph will touch some important points of the technical foundations of the technology behind it. However, a more detailed discussion will be provided in the next chapter.

Machine Learning (ML) techniques have become an essential component in the processing and analysis of Earth Observation (EO) data. In particular, their ability to extract complex spatial and spectral patterns makes them extremely suitable for remote sensing tasks. Among the wide range of ML models, approaches such as *Support Vector Machines*, *Random Forests*, and *Neural Networks* have all been successfully applied to image classification problems in this domain. However, traditional algorithms often depend heavily on data preprocessing and feature engineering, which can be time-consuming and sensitive to noise or inconsistencies in the data. These limitations directly affect a modelâs capacity to generalize effectively to new, unseen samples.

To overcome many of these challenges, *Deep Learning* methods have emerged as a more powerful and flexible alternative. Based on artificial neural network architectures, these models are capable of automatically learning hierarchical feature representations from raw data, thus minimizing the need for extensive manual preprocessing. This property is particularly advantageous in the context of Earth Observation, where image data is complex, high-dimensional, and often heterogeneous.

A key architecture within Deep Learning for image classification is the *Convolutional Neural Network* (CNN). CNNs have demonstrated outstanding performance in extracting spatially localized features directly from pixel-level inputs, making them especially effective for remote sensing applications. Their ability to process images in their near-raw form reduces preprocessing requirements and improves generalization, offering a more robust approach to understanding and classifying Earth surface features. Given its central relevance, the CNN model will be analyzed in greater depth in the following chapter.

In this thesis a particular example of subset of CNNs will be addressed and analysed for the purpose of Earth Observation. In the literature some of pre-trained and pre-structured architectures of CNNs have been proposed for the task of Earth Observation. In particular, some researches pointed out the utility of using ResNet backbones. In particular ResNet50 has showed promising results in the field of reaching high results with lower usage of data [23]. Some other researches suggested that Vision Transformers (ViTs) have state-of-the-art performance for Earth Observation [4]. Thanks to element of attention towards the key elements of the image patches this architecture demonstrates promising results if combined with the simple and powerful method of Curriculum Learning. In another chapter, we will present a deep overview of researches and results of applications of these complex but powerful structures and will be compared to the use solely of Curriculum Learning to Earth Observation. To complete the analysis, we will present a possible approach to combine the two methods.

Curriculum Learning, the central focus of this thesis, represents a strategy to leverage Machine Learning to get closer to its full potentiality and tackle complex problems that block Earth Observation to be spread in more new use cases. As it was presented before, it introduces the concept of structuring the training process by presenting the model with easier examples before gradually increasing the difficulty of the input data. In Earth Observation contexts, this could involve training on homogeneous, well labeled areas first before moving to more heterogeneous or noisy regions. Past researches showed the promising power of Curriculum Learning in context of remote sensing, and it has been explored especially on a performance point of view, where semi-supervised Curriculum Learning models surpassed even fully-supervised results. However, this thesis will deeply analyze the untapped potentiality of Curriculum Learning combined with established architectures mentioned before, especially for the case of supervised learning. To manage that, the next chapter will present in detail the theoretical basis behind Machine Learning.

2.2 Machine Learning Foundations for Earth Observation

This chapter outlines the theoretical and methodological basis of the Machine Learning techniques used throughout this work. Rather than introducing general definitions, the focus will be on the specific design choices, evaluation metrics, and learning paradigms that are most relevant to image classification in Earth Observation (EO). Understanding these principles is fundamental to grasp how *Curriculum Learning* interacts with deep architectures and why certain training and validation strategies are particularly effective in this domain.

EO data present particular challenges: high spatial dimensionality, multiple spectral channels, and large intra-class variability caused by atmospheric and illumination differences. As a result, the choice of metrics, normalization methods, and loss functions must be carefully adapted to ensure convergence, generalization, and robustness to noise.

Evaluation Metrics and Loss Functions

In the EO context, model performance cannot be measured solely by overall accuracy. Differences in class representation, spectral ambiguity, and geographical imbalance make it necessary to consider more nuanced metrics.

Metrics for Supervised Learning

- **Accuracy:**
  Represents the global proportion of correctly classified pixels or image patches. While simple, it can be misleading when the dataset is unbalanced across land-cover classes.
  $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision and Recall:**
  These metrics are crucial when certain classes, like urban areas or wetlands, are underrepresented. Precision evaluates the reliability of positive predictions, while recall quantifies the ability to detect all relevant pixels for a class.
  $$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:**
  A harmonic mean between precision and recall, often reported as a balanced indicator of per-class performance, especially for heterogeneous surfaces.
  $$\text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Metrics for Self-Supervised and Unsupervised Learning

In self-supervised learning (SSL), where labels are not available, evaluation relies on proxy metrics and downstream performance after fine-tuning. Since the SSL4EO-S12 dataset used in this thesis belongs to this class, several indicators are relevant:

- **Representation Quality:**
  The linear separability of latent representations is measured by training a lightweight classifier (for example logistic regression) on top of frozen features. The resulting accuracy reflects how well the model captures semantic structure.

- **Silhouette Coefficient:**
  Used to quantify cluster compactness and separation when representations are projected into a latent space.

  $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

  where $a(i)$ and $b(i)$ are the intra- and inter-cluster distances, respectively.

- **Davies-Bouldin Index (DBI):**
  Commonly used to validate clustering quality in self-supervised feature spaces. Lower values correspond to tighter, better-separated clusters.

  $$\text{DBI} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

- **Downstream Linear Probing (frozen encoder) used in the research of this thesis:**
  To assess the usefulness of self-supervised features for EO classification, a shallow classifier is trained on top of a *frozen* backbone and evaluated on a standard downstream dataset (for example EuroSAT; RGB or multispectral). We report top-1 accuracy (and mean±std over multiple random seeds) as the primary selection criterion across backbones and curricula. This probe isolates representation quality from end-to-end finetuning effects and is therefore complementary to SSL-internal metrics (for example contrastive loss) and clustering indices.

Loss Functions

Loss design in EO tasks plays a major role in how the model learns spatial and spectral coherence.

- **Cross-Entropy Loss:**
  Used in supervised settings for pixel- or patch-level classification. It directly compares predicted class probabilities with ground truth labels.

  $$\mathcal{L}_{\text{CE}} = -\sum y \log(\hat{y})$$

- **Contrastive and Triplet Losses:**
  In self-supervised learning, these functions encourage representations of similar images (for example same area, different time or augmentations) to be close in latent space, and dissimilar ones to be far apart.

  $$\mathcal{L}_{\text{contrastive}} = (1-y)d^2 + y\max(0, m-d)^2$$

  where d is the Euclidean distance between embeddings and m is a predefined margin.

- **Focal Loss:**
  Applied in imbalanced EO datasets to down-weight easily classified examples (such as large homogeneous regions) and focus on harder, minority classes (like roads or clouds).

  $$\mathcal{L}_{\text{Focal}} = -\alpha_t(1-\hat{y}_t)^\gamma \log(\hat{y}_t)$$

- **Contrastive Loss (NT-Xent) used in this thesis:**
  During self-supervised pre-training, the model learns to maximize agreement between augmented views of the same patch and minimize agreement between different patches. The NT-Xent (Normalized Temperature-scaled Cross-Entropy) formulation encourages discriminative yet invariant representations:

  $$\mathcal{L}_{\text{NT-Xent}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k\neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

  where $\text{sim}(\cdot)$ denotes cosine similarity and τ the temperature parameter. This objective is central to the self-supervised stages implemented for both ResNet and ViT backbones.

- **Triplet Loss:**
  Used in some comparative runs to enforce structured latent spaces. It optimizes relative distances between an anchor a, a positive p, and a negative n:

  $$\mathcal{L}_{\text{Triplet}} = \max(0, d(a,p) - d(a,n) + m)$$

  where $d(\cdot)$ is the Euclidean distance and m a fixed margin. This improves the geometric organization of representations learned from unlabelled patches.

- **Dice Loss:**
  Although more common in segmentation, Dice Loss is included for completeness as it measures the overlap between predicted and reference regions:

  $$\mathcal{L}_{\text{Dice}} = 1 - \frac{2|P \cap G|}{|P| + |G|}$$

  It highlights spatial consistency and is useful when evaluating models that predict region masks or multi-class pixel maps.

- **Curriculum-Aware Weighting:**
  Within Curriculum Learning, losses are monitored across difficulty stages (easy â hard). The same loss functions are retained, but their contribution is implicitly modulated by sample complexity, ensuring that optimization remains stable when introducing progressively harder samples.

- **Probing Loss (Linear Evaluation) used in this thesis:**
  In the downstream probing stage, the frozen encoderâs representations are evaluated by training a shallow linear classifier using the same cross-entropy objective. The probing loss therefore quantifies how effectively the self-supervised pre-training shaped separable and semantically meaningful latent features.

Training Strategies and Data Paradigms

Supervised Learning

In supervised EO classification, each image or pixel is labeled according to land-cover classes. The model learns a direct mapping between spectral-spatial input and target output. This approach is highly effective when labeled data are available, but scalability is limited by annotation cost and spatial inconsistency of labels. In our scenario, the examples of these.

Self-Supervised Learning

In self-supervised settings, the model learns without explicit labels by solving pretext tasks such as predicting masked regions, aligning different spectral bands, or distinguishing augmentations of the same patch. Once pretrained, the learned features are fine-tuned on a small labeled subset. This strategy is central to the SSL4EO-S12 dataset used in this thesis and aligns naturally with Curriculum Learning, since task complexity can be progressively increased-starting from simple spatial prediction to more intricate multi-band alignment.

Normalization and Preprocessing for EO

Unlike natural images, remote sensing data vary across sensors, bands, and atmospheric conditions. Proper normalization ensures that learned representations focus on meaningful surface information rather than acquisition artifacts.

- **Per-Band Normalization:**
  Each spectral channel is normalized independently using min-max scaling to preserve relative energy distribution among bands.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

- **Standardization across tiles:**
  When training over large geographical mosaics, pixel distributions differ across tiles.

Applying Z-score normalization per-tile stabilizes training and improves convergence.

$$x' = \frac{x - \mu_{\text{tile}}}{\sigma_{\text{tile}}}$$

- **Spectral consistency normalization:**
  During the transition from self-supervised pretraining on SSL4EO-S12 to supervised probing on EuroSAT, spectral and radiometric differences are harmonized by rescaling overlapping bands (for example the 10 shared Sentinel-2 bands). This step guarantees that the features learned during pretraining remain compatible with the downstream supervised classifier and prevents domain shift across datasets.

Architectural Considerations for EO Data

Even if the general mathematical foundations of neural networks remain the same, architectural choices for EO are guided by data structure and the desired spatial scale of analysis.

- **Convolutional Neural Networks (CNNs):**
  CNNs are preferred for their ability to capture local spatial dependencies and spectral correlations. For EO, kernel sizes and receptive fields are tuned to reflect the spatial resolution of the satellite sensor, balancing between small local patterns and large-scale textures.
- **Residual Networks (ResNets):**
  Deep architectures such as ResNet introduce skip connections that facilitate the backpropagation flow and enable training of very deep models without degradation. In EO, ResNets can show strong generalization on heterogeneous landscapes, particularly when combined with curriculum strategies that expose the model progressively to more complex regions.
- **Vision Transformers (ViTs):**
  ViTs replace convolutions with attention mechanisms, learning global spatial dependencies, that can strongly bring advantages to EO context. In this thesis, ViTs are also compared under curriculum-based training to assess how their global receptive field interacts with progressive data complexity.

Training and Validation in the Curriculum Framework

Curriculum Learning modifies the traditional supervised or self-supervised training pipeline by controlling the order of sample exposure. For EO, this may correspond to ordering images by cloud coverage, spectral entropy, or spatial complexity. The goal is to stabilize convergence and improve generalization to unseen scenes.

Validation and testing follow the same principles as in standard training, but performance is monitored not only for accuracy but also for stability and representation quality across difficulty stages. For SSL models, downstream probing accuracy (for example on EuroSAT) is used as the main validation criterion.

In summary, this chapter establishes the theoretical foundation for the techniques adopted in this work. Rather than focusing on abstract ML theory, it contextualizes how metrics, losses, and architectures are optimized for satellite imagery. In the following sections, we will extend these concepts to the deep architectures and Curriculum Learning strategies employed in the experiments.

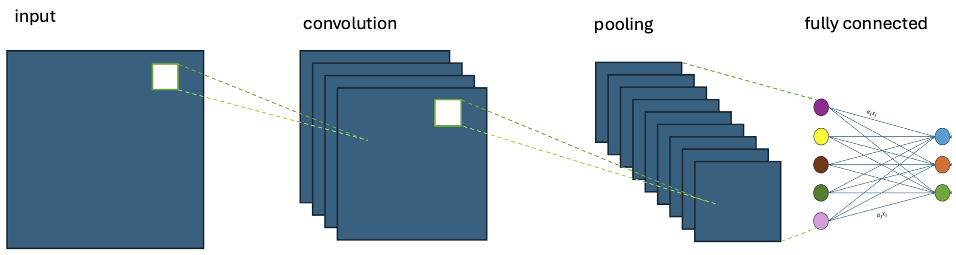

Figure 2.7. Convolutional Neural Network

- ResNet:

  One particular case on CNN design, especially tailored for complex images is the ResNet (Residual Network). The structure of this architecture is based on the idea of skipping connections during the training session. The result of this formatting is that some connections of one layer to another are cut and bypass the information [8]. The power of ResNet resides on the fact that, during the backpropagation process, the information is not lost due to the derivative of the optimization process on the loss function. When calculating the loss gradient on the back-propagation, these become very small through many layers, smoothing learning in earlier layers. Hence, skipping some of this connections helps maintaining generalization, which becomes fundamental in training on various and complex images, like the case of Earth Observation. One of the uses of ResNets is to train it on a big amount of images to create what are called "pre-trained" models. In this thesis there will be presented a method that, on the contrary, will be based on randomly initialized of the trained weights ResNet, keeping only the backbone. It represents an enhanced

CNN model but with a structure that is already known to perform on images like the context of remote sensing. The structure of a general ResNet can be seen in the figure (Figure 2.8).

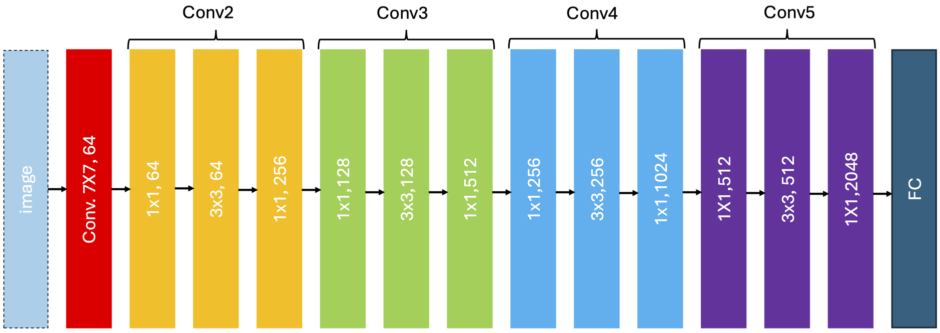

Figure 2.8. ResNet basic structure

- **ResNet-18 and ResNet-50:**
  The *Residual Network* (ResNet) family has become a reference standard for deep learning models applied to visual recognition tasks. The number in the modelâs name (for example 18 or 50) indicates the total number of layers within the architecture. In the context of Earth Observation (EO), the choice between these architectures depends on the trade-off between computational efficiency and representational power.

  – **ResNet-18:**
    A relatively shallow network with 18 layers, designed to be computationally efficient and suitable for training under limited GPU memory or time constraints (in this thesis it has been used to test and benchmark the methodoloty). It provides a good performance baseline for EO tasks where spatial complexity is moderate or where the focus is on testing curriculum strategies rather than maximizing accuracy.

  – **ResNet-50:**
    A deeper model with 50 layers and bottleneck residual blocks, offering greater

capacity to model fine-grained spatial and spectral relationships in satellite images. Due to its higher expressiveness, ResNet-50 is particularly relevant for experiments involving *Curriculum Learning*, where progressively increasing data complexity allows the model to exploit its hierarchical representation power without overfitting early stages. This backbone will therefore be emphasized in subsequent sections, especially in the self-supervised pretraining and downstream evaluation phases.

- **Vision Transformers (ViTs):**
  A more recent and conceptually different class of models is the *Vision Transformer (ViT)* architecture, which departs from the convolutional paradigm by processing images as sequences of patches. Instead of relying on convolutional kernels to extract localized spatial features, ViTs divide the image into fixed-size patches, flatten them, and embed these patches into a latent space processed by a transformer encoder originally designed for sequence modeling in natural language processing.

  The ViT architecture allows the model to capture long-range dependencies and global spatial relationships between different image regions. This property is especially advantageous in EO, where spectral and spatial correlations can span large areas-such as cloud formations, agricultural patterns, or coastline features-that may not be well captured by local convolutional operations. Recent studies [16] have shown that ViTs achieve high performance in remote sensing, particularly on heterogeneous datasets combining multiple spectral bands.

  In this thesis, ViTs are analyzed alongside CNN-based architectures to compare how their representational behavior differs under Curriculum Learning strategies. Given that ViTs are more data-hungry and sensitive to training order, this architecture provides a valuable contrast to ResNets when studying the effects of progressive learning, sample complexity, and entropy-based data staging. The models will be pretrained using self-supervised methods on the SSL4EO-S12 dataset and later evaluated through linear probing on EuroSAT, providing a direct measure of how both architectures respond to curriculum-driven pretraining.

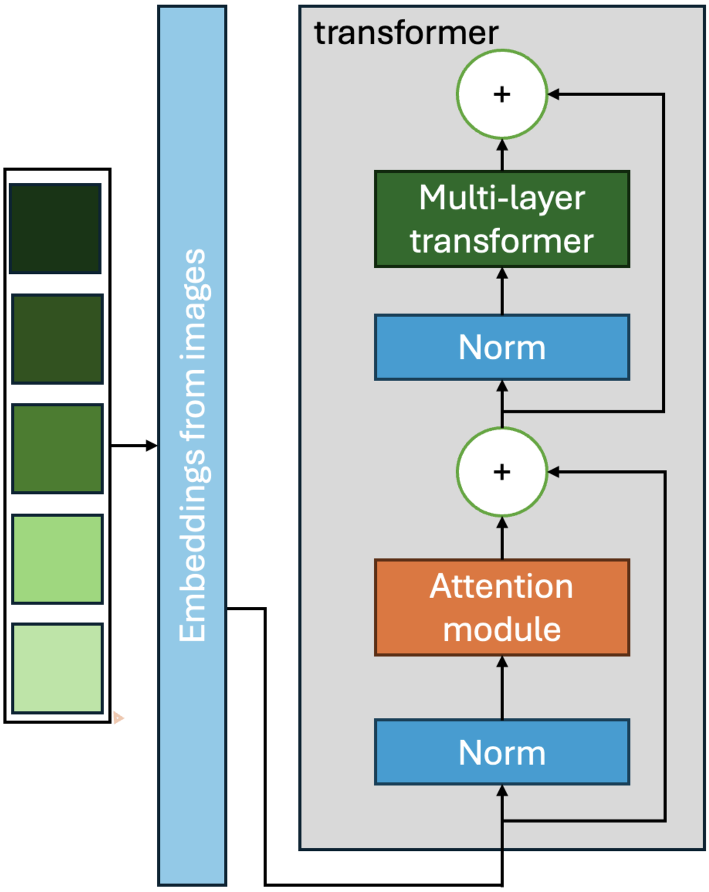

Figure 2.9. Simplified representation of the Vision Transformer architecture.

Using ResNet and ViT variants has become standard practice in computer vision tasks, from medical imaging to satellite data analysis and Curriculum Learning can strengthen their core advantages and leverage the advantages of CNNs and Transformers to discover efficiency and efficacy.

2.3 Curriculum Learning

This thesis focuses on the possible impact of Curriculum Learning on Earth Observation. Understanding how it works will help us understand how it can be merged with powerful CNNs models. Curriculum Learning relies on establishing a criterion to differentiate between "easy" and "hard" samples that are fed to the algorithm. In a general machine learning context, easy samples are usually characterized by high confidence on their class or on the classification given. For example, an easy image, can be classified with a label that the algorithm recalls to be highly certain of. On the contrary, difficult samples often include those that a baseline simple model misclassifies or where the model shows low confidence on. However, this criteria of choosing simple and complex samples has no defined way, and that is why different techniques have been explored in the literature.

2.3.1 Definition and origins

The concept of Curriculum Learning was first formally introduced by Bengio et al. [5], where the authors showed that feeding training data in a meaningful order, from easy to hard, could improve generalization of the model and get faster convergence in NNs. The intuition comes from human and animal learning processes, where basic knowledge is built with a progressive difficult and steps, with a possibility of forming a structure of knowledge and mental processes for more complex reasoning. The same idea can be abstracted to work for Artificial Neural Networks.

Going towards a more formal mathematical definition, Curriculum Learning modifies the data distribution over the epochs of the training process. Instead of taking random samples from the training set $\mathcal{D}$, the model is fed with a subset $\mathcal{D}_t \subseteq \mathcal{D}$ at each epoch t, where $\mathcal{D}_t$ contains examples up to a certain defined difficulty threshold λ_t. This threshold is gradually increased over the epochs, so that:

$$\mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \ldots \subseteq \mathcal{D}_T = \mathcal{D}$$

This strategy effectively helps how the training evolves with epochs and stabilizes, helping the gradient on the loss function stabilizing, reducing the risk of divergence, poor local minima or even vanishing gradient. It reinforces the algorithm on the learned simple patterns to prepare to slightly modify in next epochs with more complex data.

As said before, the original paradigm of Curriculum Learning was based on defining deterministically the ranking of data. Consequent studies proposed new more scalable and statistical-based ways. The way the dataset is divided is defined segmentation technique, and in the following section we will analyze some important examples.

2.3.2 Techniques of Segmentation

Heuristic-based difficulty scoring

These techniques are based on defining the sample difficulty through domain-knowledge. Hence, a human input with knowledge on what represents a simple data-point or complex will define the segmentation. In the case of images classification, the difficulty can come from visual entropy, size, number of elements or length of label name. In Earth Observation applications, heuristic scores can consider spatial homogeneity, cloud coverage, density of edges in the image, or vegetation variability in the same image, number of different types of land.

These methods are relatively easy to implement and require no additional models, but they are also limited because they rely on domain assumptions that can be subject to human bias or even error.

Self-paced curriculum learning

Self-paced curriculum learning (abbreviated SPCL), discussed by Jiang et al. [10], was proposed as a series of technique segmentation that decides on itself how to learn and which samples to learn from, at each epoch. Instead of defining before the training step, the model starts feeding the sample. When calculating the loss, the ones that bring lower loss values are defined as simple. Hence, only the simple ones will have high weight on updating the weights of the networks in the first epochs. Slowly, with time, the weight on the more loss-impacting data points will be changed to actually make the model train on those instead.

Mathematically, self-paced learning can be expressed as an optimization problem with a latent weight variable (the one that is applied to potential impact of the data point) $v_i \in \{0,1\}$ for each sample:

$$\min_{\theta,v} \sum_{i=1}^{N} v_i \cdot \mathcal{L}(f_\theta(x_i), y_i) + \lambda \sum_{i=1}^{N} v_i$$

where $\mathcal{L}$ is the loss function, f_θ is the model, and λ is a pacing parameter controlling the inclusion of harder samples with later epochs. This formulation helps the model to focus on easy samples (with low loss) in early epochs and including harder samples as learning goes on.

SPL is adaptive and often results in smoother training, but it requires properly calibrated loss function and back-propagation modeling, as well as tuned latent variable.

Teacher-student models

In this approach proposed by Matiisen et al. [15], the models rely on two different submodels: a "teacher" and a "student". The teacher, which is usually made of a pretrained model, a reinforcement learning agent or generally a model that works on segmentation of labels or images, suggests what type of ranking or segmentation to apply to a particular sample of the dataset, passing this suggestion to the student model. The student model will proceed with the actual learning from the said subset and "reports" the result to the teacher.

Teacher-student models are particularly useful when sample difficulty is hard to define heuristically and when using loss-dependent algorithm can be computationally expensive or where the risk of vanishing gradient is high. For this reason, they could be useful in Earth Observation scenarios, where one of the mentioned problems is actual heuristic labeling or computational resources. These methods are also quite flexible since the teacher model can be changed without further modifications on the student model and vice-versa. However, if the teacher model is not designed based on the needs of the context, it can become computationally more expensive as it requires two models.

Entropy-based segmentation (implemented method)

Entropy-based segmentation is one of the primary curriculum strategies employed in this thesis. Here, difficulty is derived directly from the information content of each image using the Shannon entropy measure. Given a normalized image x_i represented as a vector of pixel intensities, its entropy $H(x_i)$ is computed as:

$$H(x_i) = -\sum_{p=1}^{P} p_p \log(p_p)$$

where p_p is the empirical probability of each pixel intensity value. Low-entropy images (for example homogeneous regions such as water bodies or clouds) are considered easy, while high-entropy images (for example urban or forest areas with high spatial variation) are classified as hard. The dataset is sorted according to $H(x_i)$ and divided into three partitions:

$$\mathcal{D}_{\text{easy}},\ \mathcal{D}_{\text{medium}},\ \mathcal{D}_{\text{hard}},$$

such that:

$$H(x_i) \in \begin{cases} [H_{\min}, H_{1/3}] & \text{if } x_i \in \mathcal{D}_{\text{easy}}, \\ (H_{1/3}, H_{2/3}] & \text{if } x_i \in \mathcal{D}_{\text{medium}}, \\ (H_{2/3}, H_{\max}] & \text{if } x_i \in \mathcal{D}_{\text{hard}}. \end{cases}$$

This segmentation is entirely data-driven and computationally efficient. During training, samples are progressively introduced according to these entropy stages. The model first trains on visually simple samples and later adapts to more complex patterns, reflecting a human-like learning progression.

Masking-based segmentation (implemented method)

Another segmentation and curriculum strategy explored in this work introduces difficulty through *progressive spatial masking*. Unlike the entropy-based and Geo-AFM methods, which rank samples before training, this approach modifies the image content dynamically by randomly occluding parts of the input during the self-supervised pretraining phase.

Formally, for each image $x_i \in R^{C\times H\times W}$, a binary mask $M_i \in \{0,1\}^{C\times H\times W}$ is generated according to a Bernoulli distribution with masking probability p_m:

$$M_i \sim \text{Bernoulli}(1 - p_m)$$

The masked image is then defined as:

$$\tilde{x}_i = x_i \odot M_i$$

where $\odot$ denotes element-wise multiplication. At lower curriculum stages, p_m is small (for example $p_m = 0.05$), preserving most of the input information. As the model progresses, the masking probability increases (for example $p_m = 0.1â0.2$), forcing the encoder to infer missing spatial and spectral content from context.

This strategy effectively simulates an increasing task difficulty: the model first learns from fully visible, simpler examples and later adapts to partially occluded or corrupted inputs. Mathematically, the masking-based curriculum can be formalized as a data transformation pipeline $\mathcal{T}_m^{(k)}(x)$ parameterized by the current stage k, with stage-dependent masking probability $p_m^{(k)}$:

$$\mathcal{T}_m^{(k)}(x) = x \odot M^{(k)}, \quad M^{(k)} \sim \text{Bernoulli}(1 - p_m^{(k)}), \quad p_m^{(1)} < p_m^{(2)} < p_m^{(3)}$$

By progressively increasing $p_m^{(k)}$, the network learns to rely on more abstract, global representations that are resilient to partial information loss-a desirable property for Earth Observation models where occlusions (for example clouds or shadows) are common.

Geographical Adaptive Feature Mixing (Geo-AFM) - proposed method

The second segmentation strategy introduced in this work, *Geo-AFM*, extends the approach based on entropy by incorporating spatial and semantic diversity across geographical regions. This method is motivated by realizing that Earth Observation imagery does not vary only in texture complexity but also in regional spectral composition and environmental context.

The Geo-AFM method partitions the dataset by grouping patches into geographical cells based on latitude and longitude coordinates, then calculates intra-cell heterogeneity to define sample difficulty. Let each patch x_i be associated with a bounding box $b_i = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ projected into geographic coordinates (ϕ_i, λ_i). The dataset is divided into cells C_k representing approximately homogeneous geographic zones. For each cell, a heterogeneity score η_k is computed as the Shannon entropy of aggregated land cover distributions within that cell:

$$\eta_k = -\sum_{c=1}^{M} p_{k,c} \log(p_{k,c})$$

where $p_{k,c}$ is the normalized frequency of land-cover class c across patches in cell C_k.

The global dataset is then sorted by η_k and split into three curriculum stages:

$$\mathcal{D}_{\text{geo-easy}}, \ \mathcal{D}_{\text{geo-mid}}, \ \mathcal{D}_{\text{geo-hard}},$$

where:

$$\eta_k \in \begin{cases} [\eta_{\min}, \eta_{1/3}] & \text{for } \mathcal{D}_{\text{geo-easy}}, \\ (\eta_{1/3}, \eta_{2/3}] & \text{for } \mathcal{D}_{\text{geo-mid}}, \\ (\eta_{2/3}, \eta_{\max}] & \text{for } \mathcal{D}_{\text{geo-hard}}. \end{cases}$$

By combining geographical diversity and feature entropy, Geo-AFM captures both spatial and spectral variability, allowing the model to gradually learn from locally consistent regions before tackling heterogeneous or mixed terrains. This segmentation strategy better aligns the curriculum with the physical distribution of land types, improving representation generalization across unseen geographic regions.

Baseline segmentation (control setup)

For comparison purposes, a baseline segmentation strategy without curriculum ordering is used. In this configuration, all samples are drawn uniformly at random at each epoch:

$$x_i \sim \mathcal{U}(\mathcal{D})$$

This setup establishes a reference to assess the actual contribution of curriculum segmentation methods such as Entropy-based and Geo-AFM.

The presented segmentation techniques are just a part of the diverse selection that the literature can offer. Some of those presented here will be inspiration for the proposed method in this thesis.

2.3.3 Progressive training flow and implementation details

Once a difficulty measure has been set thanks to the segmentation method, the training set can be split into multiple difficulty levels. The simplest approach is to assign data and feeding to the model in increasing order of difficulty (level 1, level 2, level 3...). At each new stage, the model encounters more challenging examples, updating the weights with the information learned in previous stages. A typical curriculum training loop may be summarized as present in the algorithm 17.

Algorithm 1 Curriculum Learning framework for satellite image classification

```
Require: : Dataset $\mathcal{D}$ with labeled satellite images
Require: : Difficulty scoring function $Score(\cdot)$
Require: : Model architecture $M$
Require: : Number of difficulty levels $L$
Require: : Evaluation metric $Eval(\cdot)$
 1: Step 1: Data ranking and sorting
 2: for all sample $x_i$ in $\mathcal{D}$ do
 3:     Compute difficulty score $s_i \leftarrow Score(x_i)$
 4: end for
 5: Sort $\mathcal{D}$ based on $s_i$ in ascending order
 6: Step 2: split the dataset
 7: Partition $\mathcal{D}$ into $L$ subsets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_L$ by difficulty level
 8: Initialize model parameters: $M \leftarrow$ random weights
 9: Step 3: Curriculum training loop
10: for $l = 1$ to $L$ do
11:     Train model $M$ on $\mathcal{D}_l$
12:     Compute intermediate evaluation $e_l \leftarrow Eval(M, \mathcal{D}_{val})$
13:     Optionally fine-tune learning rate or optimizer settings
14: end for
15: Step 4: final evaluation
16: Evaluate final model performance on test set $\mathcal{D}_{test}$: $e_{final} \leftarrow Eval(M, \mathcal{D}_{test})$
17: return Trained model $M$, performance metrics $e_1, \ldots, e_L, e_{final}$
```

While these principles apply broadly across machine learning tasks, they can be particularly beneficial for satellite image classification. However, literature showcases different modalities in which the training steps are defined, especially in the case of Self-supervised Learning. The approach presented in this work will go deeper in the explanation of probing techniques, presented in the work (cita paper da cui hai fatto questo)

Chapter 3

Curriculum Learning in Earth Observation: Related works

3.1 Review of classification of Curriculum Learning strategies

Curriculum Learning has shown to be a powerful strategy to make training stronger, faster and in need of less computational resources. In the past chapters, the structure of general has been carefully reviewed and also existing studies have been shown. However, Curriculum Learning contains a multitude of possibilities and strategies on how to use it, especially in the Earth Observation context. Many of the existing cases have been already discussed but this chapter will go deeper and classify systematically all the Curriculum Learning strategies applicable to Earth Observation, helping to guide future research.

3.1.1 Comparison for Earth Observation classification: literature results and proposals

To understand the contributions of Curriculum Learning in the bigger context of Earth Observation, it is important to compare both conventional Earth Observation classification approaches and the relatively few studies that incorporate CL. The table 3.1 below is divided into two parts. The first presented key representative studies in Earth Observation classification without Curriculum Learning, focusing on standard Deep Learning techniques. The second part includes the more recent work that explicitly integrates CL strategies.

This extended table emphasizes how Curriculum Learning strategies are beginning to close the performance and generalization gap that exists in Earth Observation classification. While standard CNN and transformer approaches have achieved high accuracy, Curriculum Learning augmented models demonstrate improvements in convergence speed, data efficiency, and robustness, especially on noisy or complex images. This comparison further supports the motivation for the hybrid approach proposed in this thesis. To have a closer look at the effectiveness of Curriculum Learning, we tried to run some algorithms with EuroSAT dataset using ResNet18, ResNet50 and ViT-small with emptied weights 4.1. It resulted also in better accuracy, that reached respectively 81.36% and 70.99% on the test with ResNet18 trained on EuroSAT, giving us an important first result, to be furtherly analyzed, on the power of Curriculum Learning, but also a suggestion that ResNets (and potentially ViTs) are amplifiers of the capacities. The proposed method will be based on this results.

3.1.2 Taxonomy dimensions

The proposed taxonomy categorizes Curriculum Learning strategies along four primary dimensions that are:

- Supervision level: Determines the extent of labeled data required. In the beginning of the thesis we mentioned supervised and unsupervised models, and this dimension will be further explained.
- Difficulty metric: Defines how the difficulty of training samples is assessed, it will simply reinstate the main techniques proposed in the literature.
- Curriculum scheduling: Describes the strategy for ordering training samples and creating batches to feed the model.
- Model integration: Indicates how Curriculum Learning is incorporated into the training pipeline, which kind of architecture is leveraged with the curriculum setup.

Following, the table 3.2 that explains in schematic way the structure of existing Curriculum Learning for satellite image classification. curriculum-based inputs.

3.1.3 Open problems highlighted by prior work

Although Curriculum Learning (CL) has shown promise in computer vision and, to a lesser extent, in remote sensing, several challenges remain open in the EO context:

1. **Label scarcity and cost:** EO datasets are large, heterogeneous, and expensive to annotate at scale. Prior work reports that supervised models often overfit to well-represented land-cover types while underperforming on rare classes, especially when labels are noisy or region-specific.
2. **Clouds, shadows, and acquisition artifacts:** Typical curricula strategies do not model specific for Earth Observation occlusions (cloud cover, haze), being applied to standard image recognition, resulting in brittle representations when faced with partial observability or missing bands.
3. **Ambiguities in difficulty definitions:** Difficulty is often represented by task-agnostic heuristics or by loss values from a single model pass. In Earth Observation, however, *both* visual texture complexity and geographic mixing of land types matter; curricula that ignore either axis may mis-rank samples.

4. **Evaluation gaps for self-supervision:** Many works report SSL pretext losses or end-to-end finetuning scores, but fewer adopt standardized *probing* protocols across datasets (for example EuroSAT vs. BigEarthNet), making it hard to isolate representation quality from task-specific heads.

5. **Compute constraints and reproducibility:** Practical EO pipelines must run under limited GPU memory and session time (common in teaching/industry settings). Several methods assume long, uninterrupted runs or large batch sizes that are not always feasible and this thesis proposed Curriculum Learning to address this problem.

3.1.4 How this thesis solves the gaps

To respond to the limitations listed, this thesis contributes with a pipeline and analysis made exactly for EO data, combining self-supervised pretraining with curricula that explicitly target *spectral complexity*, *occlusion robustness*, and *geographic diversity*. The main elements are:

Self-supervised pretraining on multispectral EO (SSL4EO-S12): We pretrain ResNet/ViT backbones from scratch on SSL4EO-S12 using a SimCLR-style objective (NT-Xent), thus avoiding domain mismatch from natural-image pretraining and directly capturing EO spectral/spatial regularities.

Three complementary curricula aligned to EO phenomena:

- **Entropy-based curriculum** (implemented): orders samples by Shannon entropy to progress from homogeneous textures (for example water, bare soil) to highly structured scenes (for example urban, mixed vegetation).

- **Masking-based curriculum** (implemented): increases difficulty via controlled random occlusions, encouraging robustness to clouds/shadows and missing information.

- **Geo-AFM (proposed)**: partitions data into geographic cells and ranks them by intra-cell heterogeneity, aligning the curriculum with real geodiversity and land-cover mixing.

Together, these curricula cover (i) appearance complexity, (ii) observability, and (iii) geographic mixing-three axes underrepresented jointly in prior EO-CL literature.

EO-specific normalization and cross-dataset alignment:

We adopt per-band normalization and per-tile standardization; when transferring to EuroSAT/BEN, we harmonize overlapping Sentinel-2 bands to mitigate radiometric shifts. This reduces confounding due to sensor/region differences ([17]).

Standardized probing across datasets:

Representation quality is assessed with *frozen-encoder linear probes* on **EuroSAT** and a **subset of BigEarthNet**, reporting top-1 accuracy and mean±std over seeds. This isolates the effect of the SSL + curriculum choices from head capacity and finetuning heuristics.

Resource-aware experimental design: All methods are implemented in Google Colab (T4 GPUs), with curricula encoded in the data pipeline to keep memory overhead low and to tolerate session limits. This makes the setup reproducible in constrained environments without sacrificing methodological rigor.

Paper	Dataset used	Architecture	Approach	Key results
Helber et al. (2019) [9]	EuroSAT	CNN (ResNet)	Supervised classification of Sentinel-2 images	Achieved 98,75% overall accuracy on balanced dataset
Sumbul et al. (2019) [22]	BigEarthNet	Shallow CNN	Multi-label classification on Sentinel-2 patches	Introduced a new benchmark for multi-label learning in Earth Observation
P. Berg et al. (2021) [6]	SSL4EO-S	SimCLR, ViT	Self-supervised learning on EO images	Achieved strong performance on multiple downstream tasks
I. Papoutsis et al. (2022) [19]	SEN12MS	ResNet-50 + attention	Multisource fusion for classification	Improved generalization across multi-sensor inputs
E. Maggiori et al. (2017) [14]	Inria Aerial	U-Net	**Not officially curriculum strategy but it aligns with modern definition**	Reduced convergence time, better segmentation accuracy
Banerjee et al. (2021) [7]	NWPU-RESISC45, PatternNet, Indian Pines	Curriculum-driven incremental learning	**Similarity-based sample ordering Curriculum Learning**	Enhanced convergence and generalization in class-incremental learning
Li et al. (2020) [12]	Custom SAR	Custom CNN	**Heuristic Curriculum Learning (entropy/-cloud)**	Improved noise robustness in SAR imagery classification with less resources
This thesis (2025)	Sentinel-2	ResNet or ViT	**Hybrid curriculum with composite difficulty**	Conceptual model proposed for generalization and efficiency, initial tests show efficiency and fast convergence

Table 3.1. Comparative Overview of EO Classification Studies With and Without Curriculum Learning

Dimension	**Type**	**Description / Earth Observation relevance**
Supervision type	Supervised Curriculum Learning	Uses labeled data to define difficulty. Common in early Earth Observation studies with segmentation masks.
	Self-supervised Curriculum Learning	Model selects easy samples (low loss) first. Helps Earth Observation pipelines to adapt to noisy labels and class imbalance.
	Unsupervised Curriculum Learning	Difficulty based on clustering, entropy or similar, without labels. Useful for partially labeled or unlabeled image data.
Difficulty estimation	Heuristic-based	Based on specific rules applied to satellite: edge density, variability, cloud cover. Easy to compute but can be sensitive to bias.
	Loss-based	Uses loss value from the model itself to define sample difficulty. Adaptive and flexible but can be computationally slower.
	Teacher-Student	A secondary model ranks difficulty and updates student model with the assessment. Effective for complex satellite datasets.
Curriculum schedule	Static	Difficulty levels fixed before training begins. Simpler, but less flexible as the training evolves.
	Dynamic	Curriculum adapts based on training feedback. Useful for varied images but more complex to define.
Model integration	Internal	Curriculum logic built into model pipeline. Easier to manage but harder to adjust.
	External scheduler	Curriculum handled outside model. More schematic and easy to fix for Earth Observation pipelines.

Table 3.2. Taxonomy of Curriculum Learning Strategies in Earth Observation

Chapter 4

Methodology: Curriculum Learning framework for satellite data and literature review

The application of Curriculum Learning in Earth Observation has a similar structure to usual application of the technique to general image classification, but it requires a particular attention. The images coming from satellites, as it has already been discussed, contain peculiar diversity, high dimensionality and spectral variability. Hence Curriculum Learning will need a particular setup that will be discussed in this chapter. For this purpose, the thesis will describe briefly a general structure to apply Curriculum Learning that summarizes common practices in research for remote sensing. Following this, there will be a literature review on general Machine Learning techniques that will serve to introduce briefly a proposed approach, unique to this thesis.

4.1 Architecture and pipeline for curriculum learning in earth observation

The implementation of Curriculum Learning in Earth Observation requires well structured to be implemented. Many researches in literature mention satellite image processing and classification steps and here a general overview will be given:

1. Data preprocessing:

   Satellite images must first go through a cleaning process. Deying et al. [13] propose a way to clean from clouds and shadows using different bands of the images that are critically useful to go beyond the obstacle or hidden details. Furthermore, spectral normalization can be applied to reduce variability caused by weather and to align the domains.

2. Difficulty scoring and complexity estimation:

   In this stage, to our knowledge, there is no universally adopted path to score satellite images for the purpose of applying Curriculum Learning. For this reason, this chapter will mention the usual scoring discussed earlier in the thesis that have a potential good result on satellite images. The first step would be to test an heuristic approach. Labels, especially on the datasets like BEN or EuroSAT (which provide their own labeling), are very descriptive and, the longer the label, the more complex the image. However this approach could be missing some important peculiarities of the image that the label does not describe. That's why further techniques could be based on the complexity of the image (teach-student approach) or go deeper into the loss-based complexity ranking.

3. Curriculum scheduling and batch construction:

   As Abid et al. [1] discusses, once the dataset is split into levels, batches need to be built to feed the model. This step is particularly needed when the task in unsupervised:

   - Static: difficulty thresholds are fixed before training.
   - Dynamic: difficulty levels evolve based on training loss or prediction confidence.

4. Model design and curriculum integration:

   Curriculum Learning can be embedded into the training pipeline in this manner:

   - Implement it in classic CNN architectures.
   - Use of hybrid backbones (ResNet or ViT).

   Literature review suggested that the most common architectures leveraged with Curriculum Learning are usually simpler than ResNets or ViT, and sometimes don't make use of Neural Networks.

5. **Validation, monitoring and adjustment**

   To verify generalization and avoid curriculum overfitting, a classic evaluation and testing has been done:

   - Keep a randomly sampled validation set across all complexity levels, but it can be forcefully tested on high difficulty batches ranked in the curriculum decision step.
   - Monitor loss curves, accuracy, and entropy reduction during training.

The structure described summarizes common practices in Machine Learning, especially in image classification, but focuses on a key complex part. The way the batches are built changes drastically the results but that is where the model can bring foundational results.

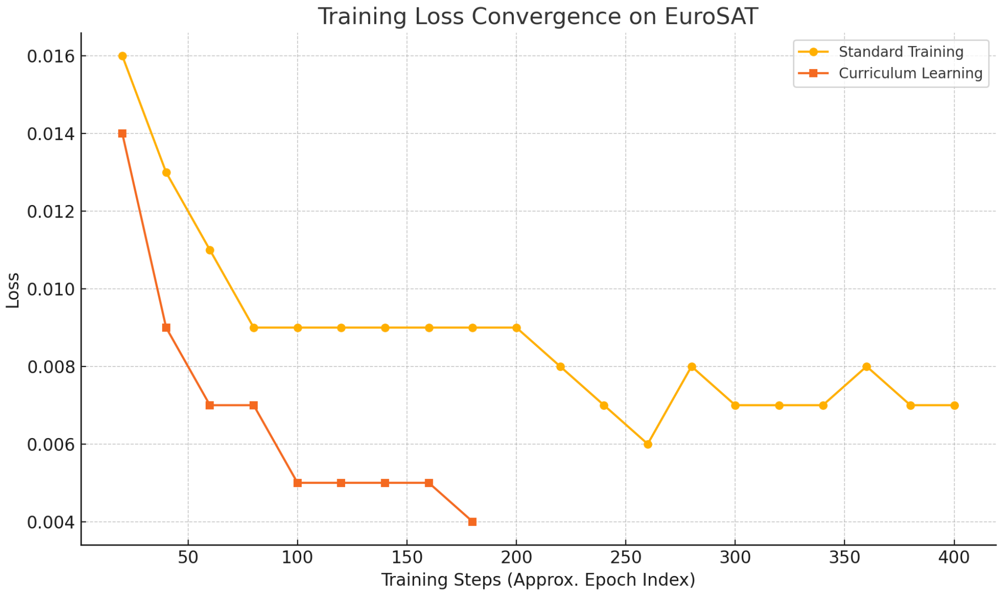

Figure 4.1. Results on EuroSAT dataset without CL and with CL

This example proves on one side the importance of Curriculum Learning in EO, but it is important to remember that the loss curves carry a bias of different dataset manipulation and ordering. Hence, from now on, the loss curves won't be presented for this case as it shows only a bias of how the dataset is being ingested by the model.

4.2 Proposed pipeline

This section presents the complete experimental pipeline developed in this thesis to integrate Curriculum Learning with state-of-the-art neural architectures for Earth Observation (EO) image classification. The work is structured around three main components:

1. Supervised baselines on EuroSAT with ResNet18 and ViT-small to assess benchmark;
2. Self-supervised contrastive pretraining on SSL4EO-S12;
3. Curriculum-based training strategies (entropy, Geo-AFM, and masking) on EuroSAT and BEN datasets.

All experiments were implemented in `PyTorch` and `TorchGeo`, executed on GPU environments, and validated through linear probing on both EuroSAT and BEN to ensure transferability of learned representations.

4.2.1 1. Baseline supervised training on EuroSAT

The first experimental stage establishes a supervised benchmark using the ResNet18 and ViT-small (ViT-B/16) architecture trained from scratch on labeled EO datasets.

The benchmark dataset used is **EuroSAT** (RGB and 13-band versions). This stage provides a baseline reference against which self-supervised and curriculum-based models are compared.

Model setup

Backbone and heads for ResNet18 We instantiate a `ResNet18` and adapt the first convolution to multispectral inputs (13 bands) and the final head to 10 EuroSAT classes:

$$\texttt{conv1}: \; R^{13\times H\times W} \to R^{64\times H'\times W'}, \qquad \texttt{fc}: \; R^{512} \to R^{10}.$$

Weights are randomly initialized (`pretrained=False`) unless otherwise stated.

Objective and optimizer We use cross-entropy loss

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{10} y_c \log \hat{y}_c,$$

and optimize all parameters with Adam (lr $= 10^{-3}$). Training is performed on GPU when available.

Data and loaders `train_loader` and `val_loader` yield dictionaries with keys `'image'` (tensor $13 \times H \times W$) and `'label'` (class index). Inputs are resized/cropped to 224×224, batched (typically 64), and normalized per band.

Curricula used

- *Non-curriculum baseline* (`baseline`): a single loader over the whole training set with random sampling per epoch.
- *Entropy / semantic-length curriculum*: the training set is stratified by a difficulty proxy (in this code variant: label-string length). We build three disjoint subsets (`short`, `medium`, `long`) via stratified splits and iterate phases in that order.
- *Coloring / spectral-variance curriculum*: an alternative curriculum that orders subsets by colorfulness or per-image variance (low→medium→high), to gradually increase spectral diversity.

Training loop (one or more epochs) For each phase in {`baseline` | `short`, `medium`, `long`}, we run a standard forwardâbackwardâupdate loop with $\mathcal{L}_{\text{CE}}$, log mini-batch loss every 20 steps, and report per-phase accuracy:

$$\text{acc} = \frac{1}{N}\sum_{i=1}^{N} 1\big[\arg\max_{c} \hat{y}_c^{(i)} = y^{(i)}\big].$$

At the end of each epoch we evaluate on the validation set with `model.eval()` and `torch.no_grad()`.

To assess efficiency on a different model, we benchmarked on EuroSAT also with a Vision Transfomer.

A ViT-B/16 model was initialized without pretrained weights. The input projection layer was modified to handle 13 spectral bands:

$$\text{Conv}_{\text{proj}} : R^{13\times H\times W} \to R^{768}$$

and the final classification head was adapted for 10 land-cover classes:

$$\text{Head} : R^{768} \to R^{10}.$$

Optimization was performed using the Adam optimizer (lr $= 10^{-4}$) with batch size 64 and cross-entropy loss.

Training protocol

Each dataset was divided into training and validation sets, and all images were resized to 224 × 224 pixels. The training objective minimized:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i),$$

where y_i are one-hot encoded labels and $\hat{y}_i$ the modelâs softmax outputs.

Three supervised curriculum configurations were trained for comparison:

- **Length-based curriculum:** datasets divided by semantic complexity of label names (short, medium, long).
- **Color-variance curriculum:** samples grouped by color variance (low, medium, high) to simulate spectral diversity.
- **Baseline:** random sample ordering without curriculum segmentation.

Algorithm 2 Supervised ViT training pipeline (EuroSAT baseline)

```
Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}$, ViT model $M_{\text{vit}}$, loss $\mathcal{L}_{\text{CE}}$, optimizer $\mathcal{O}$, epochs $T$
Split $\mathcal{D}$ into training and validation subsets
for $t = 1$ to $T$ do
    for each batch $(x, y) \in \mathcal{D}_{\text{train}}$ do
        Compute predictions $\hat{y} = M_{\text{vit}}(x)$
        Compute loss $\mathcal{L}_{\text{CE}} = -\sum y \log(\hat{y})$
        Backpropagate and update model weights using $\mathcal{O}$
    end for
    Evaluate validation accuracy on $\mathcal{D}_{\text{val}}$
end for
return Trained model $M_{\text{vit}}$ and performance metrics
```

The supervised results form the baseline for subsequent experiments and demonstrate the modelâs capability to classify EO images without self-supervised pretraining.

4.2.2 2. Self-supervised pretraining on SSL4EO-S12

SimCLR framework and NT-Xent objective

ResNet-18 backbone

A ResNet-18 encoder with 6 input bands was trained to learn representations without labels. Each image was augmented twice to generate two correlated views (x_1, x_2). The model encodes these views into embeddings (z_1, z_2) through a projection head, and the loss encourages similar views to be close while dissimilar ones remain apart:

$$\mathcal{L}_{\text{NT-Xent}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)},$$

where $\text{sim}(a, b) = \frac{a \cdot b}{\|a\|\|b\|}$ and τ is the temperature parameter (set to 0.5).

ResNet-50 backbone Alongside the ResNet-18, a deeper `ResNet-50` model was employed for self-supervised pretraining to evaluate the scalability of the curriculum strategies on a higher-capacity architecture. As with ResNet-18, the first convolutional layer was adapted to handle 6-channel inputs from the SSL4EO-S12 dataset:

$$\texttt{conv1}: \ R^{6\times H\times W} \to R^{64\times H'\times W'},$$

and the fully connected head was replaced with a projection multilayer perceptron:

$$\texttt{fc}: \ R^{2048} \to R^{1024} \to R^{128},$$

used for contrastive representation learning. Weights were initialized from scratch (`pretrained=False`) to ensure that the encoder learned spectralâspatial features directly from multispectral data instead of RGB pretraining. The optimizer, learning rate, and loss settings were identical to those of ResNet-18. This configuration allows the comparison of curriculum strategies (baseline, entropy, masking, Geo-AFM) across network depths while preserving architectural consistency.

Data augmentations and preprocessing

Each sample undergoes the following augmentations to ensure invariance to spatial and spectral transformations:

- Random resized crop (scale 0.6-1.0)
- Random horizontal flip
- Color jitter and random grayscale
- Gaussian blur
- Optional random masking (in the masking curriculum variant)

Images are normalized per spectral band to preserve energy consistency and ensure stability during contrastive learning.

Algorithm 3 Self-supervised SimCLR pretraining on SSL4EO-S12

```
Require: Dataset $\mathcal{D}$, encoder $E$, projector $P$, optimizer $\mathcal{O}$, epochs $T$
for $t = 1$ to $T$ do
    for each batch $(x_1, x_2)$ in $\mathcal{D}$ do
        Encode features: $z_1 = P(E(x_1))$, $z_2 = P(E(x_2))$
        Compute $\mathcal{L}_{\text{NT-Xent}}(z_1, z_2)$
        Backpropagate and update parameters via $\mathcal{O}$
    end for
end for
return pretrained encoder $E$
```

4.2.3 3. Curriculum Learning strategies for SSL

To evaluate the effect of sample difficulty organization, three curriculum segmentation strategies were applied during the SSL pretraining phase. Each strategy defines a structured training schedule of progressive complexity.

Entropy-based curriculum

Images are ranked by Shannon entropy:

$$H(x_i) = -\sum_{p=1}^{P} p_p \log(p_p),$$

where p_p represents the pixel intensity distribution. Low-entropy images (homogeneous areas such as water or bare soil) form the easy subset, while high-entropy (complex textures like forests or cities) form the hard subset. Training follows the ordered progression:

$$\mathcal{D}_{\text{easy}} \to \mathcal{D}_{\text{medium}} \to \mathcal{D}_{\text{hard}}.$$

Masking-based curriculum

The masking-based curriculum introduces difficulty dynamically during training by randomly obscuring parts of the input image. A binary mask $M_i \in \{0,1\}^{C\times H\times W}$ is generated as:

$$M_i \sim \text{Bernoulli}(1 - p_m), \quad \tilde{x}_i = x_i \odot M_i,$$

where p_m is the masking probability, increasing over the curriculum stages:

$$p_m^{(1)} < p_m^{(2)} < p_m^{(3)}.$$

This forces the model to learn contextual information from partially visible data, reflecting realistic EO scenarios such as cloud occlusion.

Geo-AFM (Geographical Adaptive Feature Mixing) - proposed method

The **Geo-AFM** (Geographical Adaptive Feature Mixing) method, introduced in this thesis, extends the entropy based approach by incorporating both *spatial diversity* and *geographical context* into the curriculum segmentation process. The motivation comes from a key limitation of conventional Curriculum Learning: data difficulty is often defined solely by local texture or visual entropy, ignoring spatial dependencies and geographic heterogeneity that heavily influence satellite image semantics.

Motivation: Satellite datasets like SSL4EO-S12 and BigEarthNet are multi domain by nature: samples originate from geographically distinct areas (for example Northern Europe vs. Mediterranean), each with their own spectral distributions. When models are trained without considering these regional differences, they have a tendency to overfit to specific areas, creating a *domain disalignment* problem, where features learned in one geography fail to generalize to another. Geo-AFM solves this by building a curriculum that goes from geographically homogeneous regions (where intra domain consistency is high and hence easy to train) to heterogeneous or mixed regions (where inter-domain dissimilarities increase). This progressive training exposition aligns the modelâs learning trajectory with domain diversity, effectively turning Curriculum Learning into a tool for **domain alignment**.

Mathematical formulation: Let each sample x_i correspond to a bounding box $b_i = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ projected into geographic coordinates (ϕ_i, λ_i). We partition the dataset into geographic cells C_k and compute for each cell a heterogeneity score η_k based on its land-cover class distribution:

$$\eta_k = -\sum_{c=1}^{M} p_{k,c} \log(p_{k,c}),$$

where $p_{k,c}$ is the normalized frequency of class c within region C_k. Each image inherits the score of its cell, and the dataset is divided into three ordered curriculum stages:

$$\mathcal{D}_{\text{geo-easy}},\ \mathcal{D}_{\text{geo-mid}},\ \mathcal{D}_{\text{geo-hard}},$$

such that:

$$\eta_k \in \begin{cases} [\eta_{\min}, \eta_{1/3}] & \text{if } x_i \in \mathcal{D}_{\text{geo-easy}}, \\ (\eta_{1/3}, \eta_{2/3}] & \text{if } x_i \in \mathcal{D}_{\text{geo-mid}}, \\ (\eta_{2/3}, \eta_{\max}] & \text{if } x_i \in \mathcal{D}_{\text{geo-hard}}. \end{cases}$$

Training interpretation: The model first learns in geographically stable regions-developing robust spectral encoders under low intra-class variance-and then gradually adapts to broader environmental conditions and atmospheric effects. This aligns the learned representation with global variability while maintaining spectral consistency across regions. Empirically, this progressive adaptation mitigates domain shifts and leads to better cross-region generalization, as confirmed by the statistically significant gains observed on both EuroSAT and BigEarthNet.

Algorithm 4 Curriculum-based self-supervised training (Entropy / Masking / Geo-AFM)

```
Require: Dataset $\mathcal{D}$, backbone $B$, curriculum strategy $\mathcal{S}$, temperature $\tau$
Partition $\mathcal{D}$ into ordered subsets $(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3)$ using $\mathcal{S}$
for stage $k = 1,2,3$ do
    for each batch $(x_1, x_2) \in \mathcal{D}_k$ do
        Compute embeddings: $z_1, z_2 = B(x_1), B(x_2)$
        Evaluate contrastive loss $\mathcal{L}_{\text{NT-Xent}}(z_1, z_2, \tau)$
        Update backbone parameters
    end for
end for
return pretrained model $B$
```

Significance: Geo-AFM thus transforms Curriculum Learning from a purely difficulty-based paradigm into a **domain-aware learning strategy**. By aligning the order of data exposure with natural geographical complexity, it bridges the gap between model-level optimization and the physical structure of EO data-a key innovation for large-scale, multi-domain remote sensing tasks.

4.2.4 Linear probing and downstream evaluation

After pretraining, the encoderâs weights are frozen and evaluated through linear probing on EuroSAT and BigEarthNet. A single-layer linear classifier is trained on top of the frozen features using cross-entropy loss:

$$y = Wz + b,$$

where z are the encoded representations and W, b are trainable parameters. The probing accuracy measures how well the representations learned during self-supervised training capture semantic separability of EO classes.

Evaluation is performed across all curriculum strategies (baseline, entropy, Geo-AFM, masking), and the results are averaged across multiple random seeds to assess stability and reproducibility. Metrics include top-1 accuracy, mean accuracy per class, and convergence rate.

Algorithm 5 Linear probing evaluation on EuroSAT and BigEarthNet

```
Require: Frozen encoder $E$, downstream dataset $\mathcal{D}_{\text{probe}} = \{(x_i, y_i)\}$, classifier $C$
Extract features $z_i = E(x_i)$ for all $x_i$
Train $C$ using cross-entropy loss on $(z_i, y_i)$
Compute accuracy on test set $\mathcal{D}_{\text{test}}$
return Representation accuracy and per-class metrics
```

4.2.5 5. Implementation summary

The complete training workflow can be summarized as follows:

1. Load the SSL4EO-S12 dataset and compute sample difficulty scores using entropy, geographic heterogeneity, or masking probability.

2. Train ResNet-50 and ViT-Small models under four configurations: *baseline*, *entropy*, *Geo-AFM*, and *masking*.

3. Apply the NT-Xent loss for self-supervised pretraining and record convergence curves for each curriculum stage.

4. Freeze encoders and train linear probes on EuroSAT and BEN to evaluate the quality of learned representations.

5. Compare probing accuracies and convergence dynamics across curriculum strategies.

This proposed pipeline provides a unified and reproducible framework for studying Curriculum Learning in EO, combining supervised, self-supervised, and curriculum-based paradigms within a consistent experimental environment. It demonstrates how progressive data organization can enhance learning efficiency and generalization across multiple spectral and spatial domains.

Chapter 5

Test environment, results and take-aways

5.1 Test environment, IDE and dataset availability

All experiments were implemented and executed in the Google Colab environment, which provides a cloud-based, GPU-accelerated interface for Python and deep learning workflows. Colab was selected for its compatibility with `PyTorch`, `TorchGeo`, and other required libraries such as `scikit-learn` and `matplotlib`, as well as its ease of integration with Google Drive for dataset management.

Hardware configuration and limitations

The environment was configured to use NVIDIA Tesla T4 GPUs (16 GB VRAM) and 12 GB of shared RAM. This configuration was sufficient for moderate-scale training of self-supervised and supervised models but introduced several practical limitations:

- **Session duration:** Colab imposes a maximum runtime of 12 hours per session, with disconnections occurring after inactivity. Long SSL pretraining runs had to be resumed manually using intermediate checkpoints.

- **Storage constraints:** The available disk space (100 GB) limited the number of datasets and model weights that could be stored simultaneously. Large datasets like BigEarthNet required selective sampling.

- **GPU memory limits:** The 16 GB VRAM capacity restricted batch size and model depth, particularly when using Vision Transformers. Batch sizes above 64 often caused out-of-memory errors, especially during curriculum stages with augmented or masked inputs.

Despite these constraints, Colab proved adequate for prototyping and validating the proposed Curriculum Learning strategies. All runs were reproducible using fixed random seeds for model initialization and dataset splitting.

Self-supervised learning (SSL4EO-S12) implementation

The self-supervised framework was built upon the `TorchGeo` implementation of the `SSL4EO-S12` dataset. The dataset was automatically downloaded and preprocessed within the Colab environment using:

```
SSL4EOLBenchmark(sensor="etm_sr", product="cdl", split="train", download=True)
```

This automatically structured the dataset into 12-band Sentinel-2 patches, which were loaded as tensors and normalized to $[0, 1]$. To control runtime and GPU load, the dataset was accessed through custom `Dataset` classes that applied spatial cropping, random augmentations, and masking as part of the curriculum.

The SSL training loop was implemented using the SimCLR paradigm with the NT-Xent loss. Augmentations and curriculum segmentation (entropy-based, masking, and Geo-AFM) were directly integrated into the `DataLoader` pipeline, ensuring progressive difficulty in the samples presented to the network. Each self-supervised run produced a pretrained encoder whose parameters were saved for later probing.

EuroSAT supervised and probing implementation

For the supervised baseline and the probing evaluation, the `EuroSAT` dataset was imported from the `torchvision.datasets` module. The training and validation subsets were preprocessed using:

```
Resize(256), CenterCrop(224), ToTensor()
```

and normalized per channel. A Vision Transformer (ViT-B/16) was trained from scratch in the supervised baseline and later used for linear probing of SSL-pretrained encoders. For probing, the frozen encoders (ResNet-50 and ViT-Small) trained on SSL4EO-S12 were connected to a shallow linear classifier and evaluated on EuroSAT to measure representation quality.

BigEarthNet subset preparation

Due to computational and storage limits in Colab, the full **BigEarthNet (BEN)** dataset could not be processed. Instead, a representative subset was manually downloaded from local drive and split into training and test sets. This subset maintained spectral and class diversity across geographic regions but significantly reduced storage usage and training time. The subset was sufficient for comparative probing and validation of model generalization under different curriculum strategies. All BEN experiments used the same preprocessing pipeline as EuroSAT, ensuring comparability between datasets.

5.2 Results and plots

At this stage, the results primarily serve to validate the correct functioning of the training pipeline, curriculum segmentation, and probing setup. Both Training and Test results are presented.

Training results and Test plots

Below the table shows the results of the test values reached after 100 epochs on Resnet18 and ViT-Small in supervised mode, while 500 epochs have been tested for the Resnet50 and ViT-Small in self-supervised mode.

Table 5.1. Summary of model accuracies (%) across training configurations, curricula, and datasets.

Model	Setting / Task	Dataset	Baseline	Entropy	Coloring / Masking	Geo-AFM
ResNet18	Supervised training	EuroSAT	65,5 %	75,9 %	86,3 %	NA
VitSmall	Supervised training	EuroSAT	62,6 %	65,1 %	67,6 %	NA
ResNet50	Linear probing	EuroSAT	68,5 %	68,8 %	68,9 %	70,8 %
	Linear probing	BEN	71,2 %	71,3 %	75,1 %	78,4 %
ViT-Small	Linear probing	EuroSAT	68,6 %	68,8 %	69,4 %	72,3 %
	Linear probing	BEN	72,1 %	71,9 %	75,1 %	78,4 %

To visualize the trend of learning, the training plots show the cases of self-supervised learning in Resnet50 and Vit-Small trained in EuroSAT and BEN datasets. Refer to image 5.1, 5.2, 5.3 and 5.4. For reference in the legend, the baseline will be BL, Entropy curriculum E, Masking curriculum M and Geo-AFM will be T.

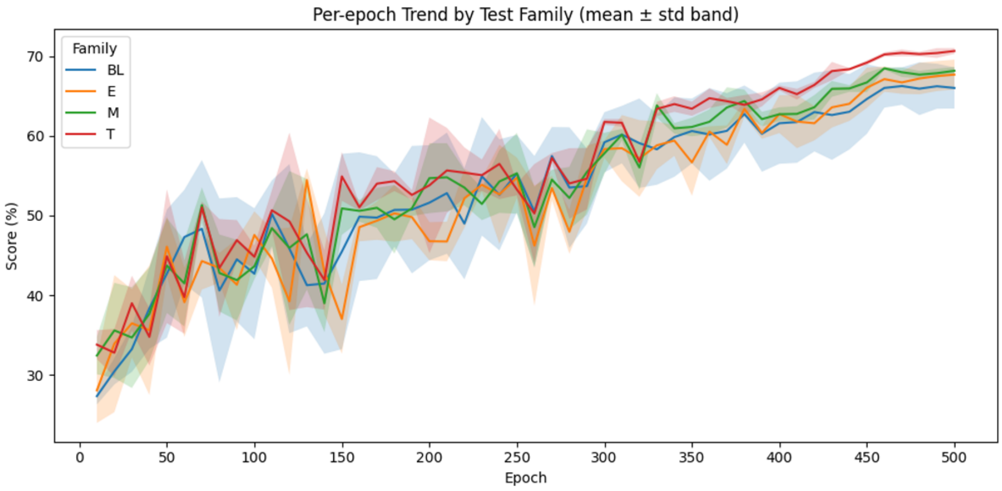

Figure 5.1. resnet50 training results for the 4 methods, probed on EuroSAT

5.2.1 Results: qualitative and quantitative notes on Training trends

Qualitative trends from the plots.

- **Monotonic improvement over epochs.** All configurations (BL = baseline, E = entropy, M = masking, TG = geo-based) show steady increases in probe accuracy with training progress (Figures 5.1-5.4). The learning curves for TG and M generally

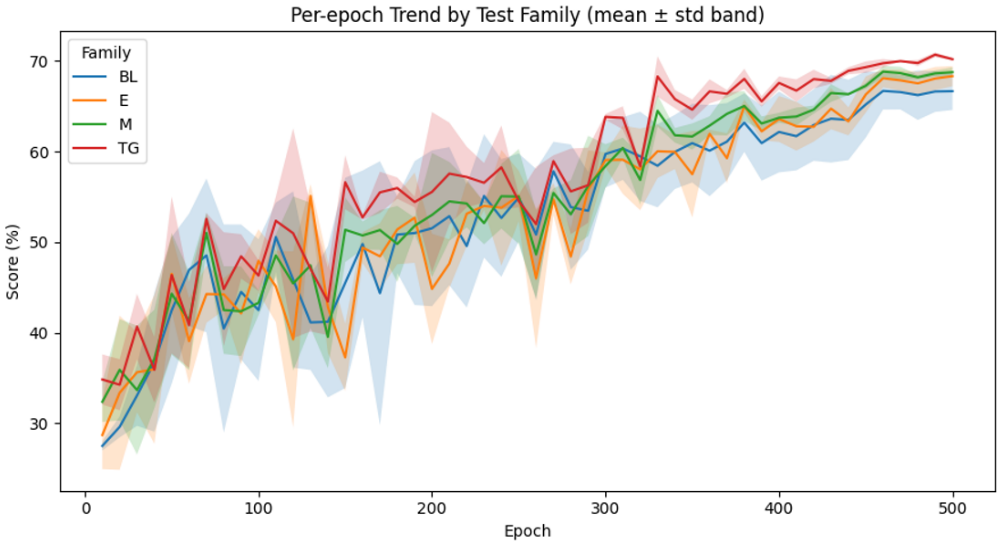

Figure 5.2. Vit-small training results for the 4 methods, probed on EuroSAT

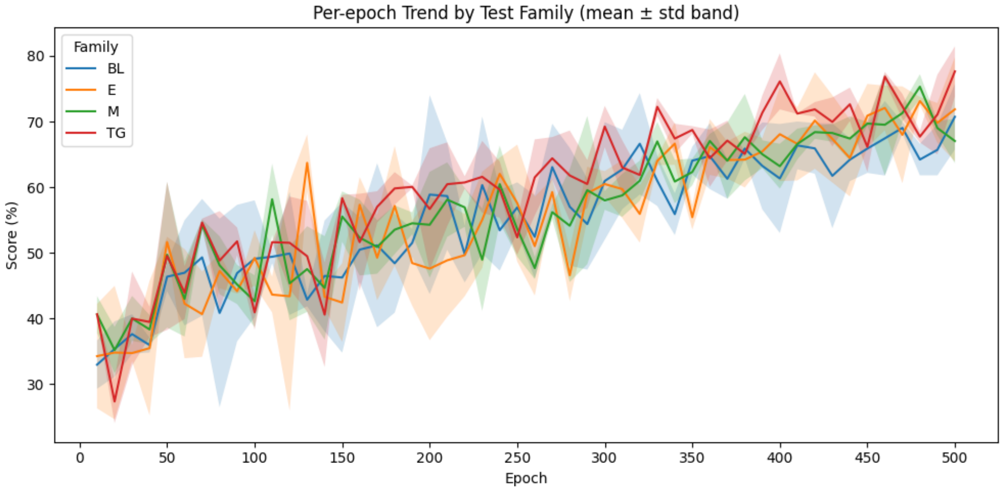

Figure 5.3. ResNet50 training results for the 4 methods, probed on BEN

track above BL after ≈150-200 epochs, with widening separation in later stages (350-500 epochs). It is clear how the results on EuroSAT result more stable than those on BEN. The latter is comprised with a vast type of views that can make training and test stability more difficult. However, the clear advantage on the TG model is evident, also signaling more stability in last 50 epochs.

- **Variance behaviour.** The shaded bands (mean ± std across seeds/epochs) are initially wide and shrink moderately as training stabilizes. TG exhibits slightly

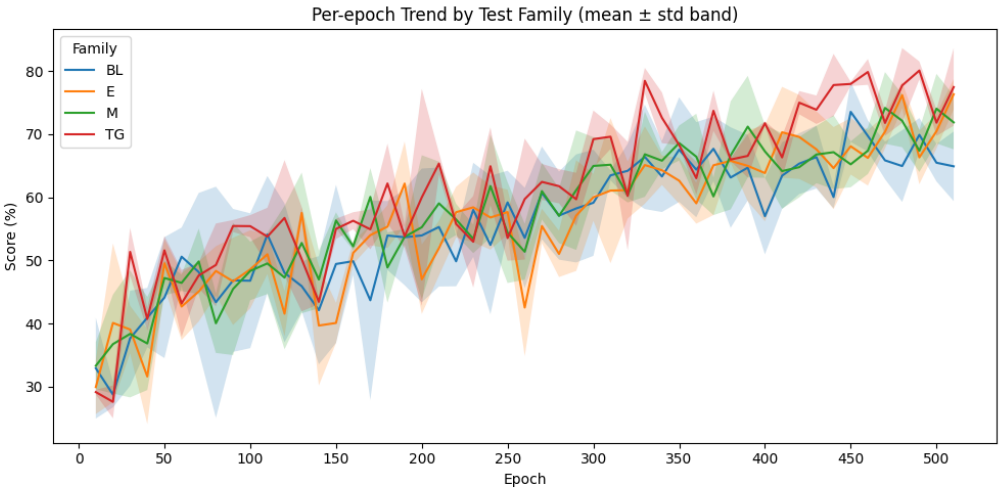

Figure 5.4. Vit-small training results for the 4 methods, probed on BEN

larger variance early but converges to a higher mean in later epochs, consistent with curricula that expose harder, more diverse samples over time.

- **Backbone sensitivity.** The qualitative gap TG > M > E≈BL is visible for *both* backbones. ResNet50 displays delayed but larger late-epoch gains for TG, whereas Vit-Small benefits earlier from M and TG. E shows better results at the end of the training but statistically is not as strong in the whole training process as the other methods.

Aggregate improvements over epochs (average %).

Setting	BL	E	M	TG
ViT-small → BEN	53.06	52.58 (−0.48)	54.22 (+1.15)	56.02 (+2.96)
ResNet-50 → EuroSAT	53.23	53.17 (+0.06)	54.63 (+1.40)	57.33 (+4.10)
ResNet-50 → BEN	55.40	55.73 (+0.33)	56.72 (+1.32)	59.65 (+4.25)
ViT-small → EuroSAT	55.83	56.28 (+0.45)	57.62 (+1.79)	61.26 (+5.43)

Note. Values are epoch-wise means; deltas vs. BL in parentheses.

Statistical relevance of the results

The figures (5.1 - 5.4) show the per-epoch linear probing accuracy for the four curriculum strategies—Baseline (BL), Entropy (E), Masking (M), and Geo-AFM (TG)—tested on both EuroSAT and BigEarthNet datasets using ResNet-50 and ViT-Small backbones. Across all experiments, a consistent upward trend in accuracy is observed as training progresses, confirming stable convergence for all methods. The shaded areas indicate the standard deviation over multiple seeds and epochs, highlighting stable learning behaviour, especially in later training phases.

Observed trends

- On both datasets, the **Geo-AFM (TG)** curriculum achieves the highest final accuracy, followed by **Masking (M)**, while **Entropy (E)** performs similarly to the baseline.
- For ResNet-50, TG reaches a mean accuracy of 57.33% on EuroSAT and 59.65% on BEN, outperforming BL by +4.10% and +4.25%, respectively.
- ViT-Small shows similar behaviour, with TG reaching 61.26% on EuroSAT and 56.02% on BEN, corresponding to improvements of +5.43% and +2.96% over BL.
- Masking (M) consistently yields moderate but statistically reliable improvements (+1-2%) across all tests, while Entropy (E) does not significantly differ from BL.

Paired one-tailed *t*-tests For each configuration, paired *t*-tests were computed across epoch-aligned results, testing the null hypothesis:

$$H_0 : \mu_{\text{tech}} \leq \mu_{\text{BL}} \quad \text{vs.} \quad H_1 : \mu_{\text{tech}} > \mu_{\text{BL}}.$$

Table 5.2 summarizes the key statistical outcomes.

Model / Dataset	**Technique**	***t*-stat**	***p*(one)**	**Significance**
ResNet50 - EuroSAT	E	-0.112	0.5445	<90%
	M	+3.780	0.0002	95%
	TG	+9.999	0.0000	95%
ResNet50 - BEN	E	+0.370	0.3564	95%
	M	+1.952	0.0283	95%
	TG	+5.765	0.0000	95%
ViT-small - EuroSAT	E	+0.543	0.2948	90%
	M	+2.527	0.0074	95%
	TG	+6.825	0.0000	95%
ViT-small - BEN	E	-0.897	0.8128	90%
	M	+3.125	0.0015	95%
	TG	+6.903	0.0000	95%

Table 5.2. Paired one-tailed *t*-tests comparing curriculum techniques vs. baseline across all settings.

Interpretation of statistical significance

- Geo-AFM (TG) shows statistically significant gains ($p < 0.001$) in all settings, confirming its consistent superiority over the baseline.

- Masking (M) demonstrates statistically meaningful improvements ($p < 0.05$) in most experiments, confirming that progressive occlusion acts as a valid difficulty signal.
- Entropy (E) does not reject H_0 in all tests ($p > 0.2$), indicating measurable advantage compared to random sample ordering, but fails the H_0 in the ResNet50 probing test with the EuroSAT dataset.
- The ranking of curriculum effectiveness is therefore: TG > M > E > BL even if it's important to remember the Entropy curriculum defects.

Summary Across all datasets and backbones, the Geo-AFM curriculum consistently produces the most reliable and statistically significant improvements in self-supervised representation quality. The masking strategy also contributes positively, especially in complex datasets like BEN where partial occlusions are common. Entropy-based ordering, while conceptually intuitive, appears insufficient alone to capture the multifactor complexity of multispectral Earth Observation imagery.

5.2.2 Take-aways

What the experiments consistently show

- **Curriculum ranking is stable:** TG > M > E ≈ BL around backbones (ResNet-50, ViT-Small) and datasets (EuroSAT, BEN). Geo-AFM (TG) guarantees the largest absolute gains (+3-5% top-1 on average), masking (M) offers smaller but reliable improvements (+1-2%), while entropy (E) rarely differs from baseline.
- **Statistical support:** One-tailed paired *t*-tests reject H_0 for TG in *all* settings ($p \ll 0.01$) and for M in most settings ($p < 0.05$), validating that the observed gains are not due to chance.
- **Backbone-specific behavior:** ViT-Small benefits most from TG in the later training stages (larger late-epoch margins), whereas ResNet-50 shows earlier, steady improvements with both M and TG. This suggests curricula help transformers âunlockâ global-context representations, while CNNs capitalize sooner on staged diversity.
- **Dataset effects:** EuroSAT curves are smoother and stabilize earlier; BEN exhibits higher variance (greater geographic and spectral heterogeneity). Despite this, TG remains the most effective strategy on BEN, indicating better cross-region generalization.
- **Variance and convergence:** Variance bands narrow as training progresses in the EuroSAT training, but less evidently in the BEN dataset. TG starts with slightly higher variance but converges to the highest mean. This supports the idea that exposing harder, diverse samples later is beneficial if preceded by easier stages where domain is considered. This alludes to a potential effect of domain generalization needed before training to more specific and complex figures.

Background needed for the proposed EO training

- **Geo-AFM can be used when geo-metadata are available:** It provides the best trade-off between complexity and gain, and scales to heterogeneous territories.
- **Default to masking if metadata are limited:** Progressive occlusion is easy to implement in the dataloader, is sensor-agnostic, and consistently improves probing scores.
- **Entropy-only ordering is rarely worth it:** On multispectral EO imagery, Entropy faces difficulty to outperforms statistically evidently the Baseline. It is preferred to use TG or M to capture spatial/spectral diversity.
- **Probe early, then fine-tune:** Linear probing is an efficient gatekeeper for representation quality, in the context of SSL. Hence, it is convenient to fine-tune following the results of the probing.

Chapter 6

Next steps in Curriculum Learning research for Earth Observation

6.1 Future Directions

The integration of Curriculum Learning into Earth Observation presents numerous possibilities for new implementations and use cases. Key future directions that are mentioned in literature or that we regard important are:

1. Multi-modal Curriculum Learning:

Earth Observation data often comprehends various modalities, such as optical images, synthetic aperture Radar (also called SAR), and LiDAR. Developing Curriculum Learning strategies that can effectively work on this cases and leverage these data sources is really important. However, it remains as an open challenge. For instance, designing different curricula that go from simpler optical data to more complex ones that are sourced in SAR could help increase the robustness of the model.

2. Dynamic Curriculum scheduling:

Since curricula are decided, in some context of Earth Observation, with deterministic approaches, a flexible optimization on the learning process can be key to uncover efficiency. As mentioned other times in this thesis, dynamic scheduling is part of the possible structuring practices when feeding a model. While this approach can be more difficult to implement, it involves future possible researches, making the model extremely efficient on focusing on impacting samples.

3. Curriculum Learning for temporal Earth Observation data:

In this thesis, a deep description on existing and promising techniques involving Curriculum Learning have been presented. All the datasets and scenarios presented only described static images with the goal of static classification. However, Earth Observation datasets often have temporal dimensions too, capturing changes over time. Developing Curriculum Learning methods that consider temporal presence can improve models' ability to classify complex scenarios based on what, not only the image suggests, but also the change of states and conditions.

Chapter 7

Conclusion

In this thesis, we investigated the role and potential of **Curriculum Learning (CL)** within the field of **Earth Observation (EO)**, a domain that is becoming increasingly data-intensive and computationally demanding. The research addressed the dual challenge of handling the growing volume and heterogeneity of satellite imagery while improving the efficiency and robustness of machine learning models used for their analysis.

We began by revisiting the theoretical foundations of Machine Learning and, in particular, Curriculum Learning, emphasizing how training paradigms inspired by human cognition can be applied to artificial learning systems. This foundation served to connect CL with the specific needs of remote sensing, where models must learn from highly variable, multi-spectral data with limited or noisy labels. We reviewed both classical and state-of-the-art architectures-Convolutional Neural Networks, ResNets, and Vision Transformers (ViTs)-and discussed their respective strengths for EO tasks.

The literature review revealed a clear research gap: while Curriculum Learning has been applied to EO in limited experimental settings, its integration with modern deep architectures and large-scale remote sensing datasets remains largely unexplored. This observation motivated the structured experimental framework proposed in this thesis, combining CL with established deep backbones (ResNet and ViT) trained from scratch or in self-supervised regimes. We implemented three EO, specific curriculum strategies-entropy-based, masking-based, and the novel **Geo-AFM (Geographical Adaptive Feature Mixing)**, and evaluated them through linear probing on the EuroSAT and BigEarthNet datasets.

The experimental findings demonstrated that CL can consistently enhance the quality of learned representations and improve downstream classification accuracy. Across all configurations, **Geo-AFM** achieved the strongest and most statistically significant gains (+3-5% top-1 accuracy over the baseline), followed by **Masking**, which provided smaller but reliable improvements (+1-2%). **Entropy-based curricula**, while intuitive, yielded results close to baseline in the overall training process and proved not completely sufficient to capture the complexity of multispectral EO data. These outcomes were validated through one-tailed paired t-tests, which confirmed the significance of Geo-AFM and Masking improvements at both 90% and 95% confidence levels. Importantly, the observed gains were not only statistically meaningful but also *practically relevant*, given the computational constraints of the real world training environments.

From a methodological standpoint, this thesis demonstrates that CL strategies can be effectively implemented even under resource-limited settings such as Google Colab, even though further tests must be done with full potentiality of training capabilities to let the model converge further more, especially in the case of the BEN dataset. Despite reduced runtime, smaller batch sizes, and partial datasets, the results showed faster convergence, smoother training, and better generalization when curriculum mechanisms were applied. This confirms the hypothesis that structured sample exposure-when aligned with spectral and geographical complexity-can guide models toward more stable and efficient learning trajectories.

Beyond the empirical analysis, we proposed a taxonomy for Curriculum Learning strategies in Earth Observation, clarifying the dimensions of supervision, difficulty estimation, scheduling, and model integration. This taxonomy serves as a framework for future research and for adapting CL to different satellite sensors, temporal scales, and downstream applications.

In summary, this thesis positions Curriculum Learning as a **viable and scalable framework** for improving the efficiency and generalization of EO classification pipelines. By connecting algorithmic learning principles with the intrinsic structure of geospatial data, CL offers a possibility for more interpretable, cost effective, and sustainable model training in the remote sensing community. Future directions include exploring adaptive pacing functions, self-supervised curricula on multimodal datasets (for example Sentinel-1/2 fusion), and geographically aware validation protocols to measure spatial generalization more rigorously.

Ultimately, Curriculum Learning shows to be a **promising paradigm for the next generation of EO systems**, capable of linking theoretical learning principles to operational impact-reducing computational cost while enhancing model reliability and transferability in a rapidly evolving global data landscape.

Bibliography

[1] Nadia Abid. *Unsupervised Curriculum Learning Case Study: Earth Observation UCL4EO*. Phd dissertation, LuleÃ¥ University of Technology, 2024.

[2] Katherine Anderson, Barbara Ryan, William Sonntag, Argyro Kavvada, and Lawrence Friedl and. Earth observation in service of the 2030 agenda for sustainable development. *Geo-spatial Information Science*, 20(2):77–96, 2017.

[3] Jatin Babbar and Neeru Rathee. Satellite image analysis: A review. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–6, 2019.

[4] Yakoub Bazi, Laila Bashmal, Mohamad M. Al Rahhal, Reham Al Dayil, and Naif Al Ajlan. Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3), 2021.

[5] Y. Bengio, JÃ©rÃ´me Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. volume 60, page 6, 06 2009.

[6] Paul Berg, Minh-Tan Pham, and Nicolas Courty. Self-supervised learning for scene classification in remote sensing: Current state of the art and perspectives. *Remote Sensing*, 14, 08 2022.

[7] S Bhat, Biplab Banerjee, Subhasis Chaudhuri, and Avik Bhattacharya. Cilea-net: A curriculum-driven incremental learning network for remote sensing image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:1–1, 05 2021.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[9] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 08 2017.

[10] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander G. Hauptmann. Self-paced learning with diversity. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[11] Pratistha Kansakar and Faisal Hossain. A review of applications of satellite earth observation data for global societal benefit and stewardship of planet earth. *Space Policy*, 36:46–54, 2016.

[12] Rui Li, Xiaodan Wang, Jian Wang, Yafei Song, and Lei Lei. Sar target recognition based on efficient fully convolutional attention block cnn. *IEEE Geoscience and*

Remote Sensing Letters, PP:1–5, 11 2020.

[13] Deying Ma, Renzhe Wu, Dongsheng Xiao, and Baikai Sui. Cloud removal from satellite images using a deep learning model with the cloud-matting method. *Remote Sensing*, 15(4), 2023.

[14] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229. IEEE, 2017.

[15] Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher-student curriculum learning, 2017.

[16] JosÃ© MaurÃcio, InÃªs Domingues, and Jorge Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9), 2023.

[17] Pablo Miralles, Antonio Scannapieco, Prerna Baranwal, Bhavin Faldu, Ruchita Abhang, Sahil Bhatia, Sebastien Bonnart, Ishita Bhatnagar, Beenish Batul, Pallavi Prasad, HÃ©ctor Ortega-GonzÃ¡lez, Harrish Joseph, Harshal More, Sondes Morchedi, Aman Panda, Marco Di Fraia, Daniel Wischert, and Daria Stepanova. Machine learning in earth observation operations: A review. 10 2021.

[18] Hafsa Ouchra and Abdessamad Belangour. Satellite image classification methods and techniques: A survey. In *2021 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–6, 2021.

[19] Ioannis Papoutsis, Nikolaos Ioannis Bountos, Angelos Zavras, Dimitrios Michail, and Christos Tryfonopoulos. Benchmarking and scaling of deep learning models for land cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195:250–268, 2023.

[20] Yogesh Rathod. A survey of machine learning techniques for artificial intelligence. *International Journal of Computer Techniques*, 11, 07 2024.

[21] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

[22] Gencer Sumbul, Marcela Charfuelan, BegÃ¼m Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904, 2019.

[23] Liyuan Wang, Yulong Chen, Xiaoye Wang, Ruixing Wang, Hao Chen, and Yinhai Zhu. *Research on Remote Sensing Image Classification Based on Transfer Learning and Data Augmentation*, pages 99–111. 08 2023.

[24] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation, 2023.