DGK Ausschuss Geodäsie (DGK)
der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 976

Jonathan Simon Prexl

Sensor-Informed Self-Supervised Representation Learning for Multi-Modal Earth Observation Data

München 2026

Verlag der Bayerischen Akademie der Wissenschaften, München

ISSN 0065-5325 ISBN 978-3-7696-5388-5

DGK Ausschuss Geodäsie (DGK)
der Bayerischen Akademie der Wissenschaften

Reihe C Dissertationen Heft Nr. 976

Sensor-Informed Self-Supervised Representation Learning for Multi-Modal Earth Observation Data

Vollständiger Abdruck der von der Fakultät für Luft- und Raumfahrttechnik
der Universität der Bundeswehr München
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
angenommenen Dissertation.

M.Sc. Jonathan Simon Prexl

München 2026

Verlag der Bayerischen Akademie der Wissenschaften, München

ISSN 0065-5325 ISBN 978-3-7696-5388-5

Adresse des Ausschusses Geodäsie (DGK)
der Bayerischen Akademie der Wissenschaften:

Ausschuss Geodäsie (DGK) der Bayerischen Akademie der Wissenschaften
Alfons-Goppel-Straße 11 ● D – 80 539 München
Telefon +49 – 89 – 23 031 1113 ● Telefax +49 – 89 – 23 031 - 1283 / - 1100
e-mail post@dgk.badw.de ● http://www.dgk.badw.de

Gutachter:

1. Prof. Dr.-Ing. Michael Schmitt

2. Prof. Dr. Sébastien Lefèvre

Die Dissertation wurde am 02.06.2025 bei der Universität der Bundeswehr München
eingereicht und durch die Fakultät für Luft- und Raumfahrttechnik
am 29.09.2025 angenommen.

Die mündliche Prüfung fand am 29.09.2025 statt.

Diese Dissertation ist auf dem Server des Ausschusses Geodäsie (DGK)
der Bayerischen Akademie der Wissenschaften unter <http://dgk.badw.de/>
sowie auf dem Publikationsserver der Universität der Bundeswehr München unter
<https://athene-forschung.unibw.de/doc/154088/154088.pdf> elektronisch publiziert

ISSN 0065-5325 ISBN 978-3-7696-5388-5

Wir schaffen das!
– Angela Merkel

Abstract

English Version The growing number of satellite missions – equipped with sensors across diverse modalities – has significantly increased both the volume and complexity of available *Earth observation* (EO) data. This wealth of information holds great potential for addressing a wide range of scientific and socially relevant challenges. Modern deep learning approaches offer effective means for extracting meaningful insights from such data. However, the unique characteristics of EO data – such as variations in sensor parameters across satellites, differences in acquisition geometries and modality-specific data characteristics – are often underutilised in existing deep learning methods. This thesis explores how sensor-specific knowledge can be systematically integrated into deep learning workflows to improve model performance and enable the processing of data originating from multiple sensors in a *sensor-informed* manner.

Three approaches are proposed to address different facets of *sensor-informed* learning for EO data. First, the *SenPaMAE* framework introduces a parameter-aware masked autoencoder for multi-spectral data, incorporating relevant sensor parameters during both the pre-training and fine-tuning stages. This enables training across datasets with varying sensor characteristics and supports sensor-independent inference. Second, the *SARFormer* framework leverages geometric acquisition parameters and imaging modes specific to high-resolution *synthetic aperture radar* (SAR) data. Integrating this information leads to improved performance on downstream tasks such as segmentation and 3D reconstruction. Third, the *IaI-SimCLR* framework extends existing multi-sensor contrastive learning techniques by introducing loss functions that stem from deeper considerations of the sensor design. This approach enhances the ability of models to learn semantically meaningful and generalizable representations via multi-modal pre-training strategies.

Together, these contributions advance the development of more general and scalable EO machine learning models – a step toward broader and more accessible automated analysis of EO data for users from diverse backgrounds.

Deutsche Fassung Die wachsende Zahl an Satellitenmissionen – ausgestattet mit Sensoren unterschiedlicher Modalitäten – hat sowohl das Volumen als auch die Komplexität der verfügbaren Erdbeobachtungsdaten erheblich erhöht. Diese Fülle an Infor-

mationen bietet großes Potenzial zur Beantwortung vielfältiger wissenschaftlicher und gesellschaftlich relevanter Fragestellungen. Moderne Ansätze des maschinellen Lernens stellen dabei wirkungsvolle Werkzeuge zur Verfügung, um aus diesen Daten relevante Erkenntnisse zu gewinnen. Die besonderen Eigenschaften von Erdbeobachtungsdaten – wie Unterschiede in den Sensorparametern zwischen verschiedenen Satelliten, variierende Aufnahmewinkel sowie Partikularitäten von Daten bestimmter Modalitäten – werden von bestehenden Methoden jedoch häufig nicht ausreichend berücksichtigt. Diese Arbeit untersucht, wie sensorspezifisches Wissen systematisch in Modelle basierend auf Techniken des maschinellen Lernens integriert werden kann, um diese zu verbessern und die Verarbeitung von Erdbeobachtungsdaten – aufgenommen mit unterschiedlichen Sensoren – auf eine *sensor-informierte* Weise zu ermöglichen.

Es werden drei Ansätze vorgestellt, die verschiedene Aspekte des *sensor-informierten* maschinellen Lernens mit Erdbeobachtungsdaten adressieren. Erstens stellt das in dieser Arbeit eingeführte *SenPaMAE* Framework ein Modell für multispektrale Erdbeobachtungsdaten vor, das relevante Sensorparameter – sowohl während eines selbstüberwachten Vortrainierens als auch später, während der zweiten Phase bei der annotierte Erdbeobachtungsdaten mit einbezogen werden – berücksichtigt. Dies ermöglicht das Training über Datensätze mit unterschiedlichen Sensorcharakteristiken hinweg und unterstützt eine sensorunabhängige Inferenz. Des Weiteren ermöglicht das *SARFormer*-Framework die geometrischen Aufnahmeparameter sowie den Bildgebungsmodus hochauflösender SAR-Daten für den Lernprozess mit einzubeziehen. Die Integration dieser Informationen führt zu einer verbesserten Vorhersage bei Aufgaben wie Segmentierung und 3D-Rekonstruktion. Drittens erweitert das vorgestellte *IaI-SimCLR*-Framework bestehende kontrastive Lernansätze auf Daten verschiedener Modalitäten durch Verlustfunktionen, die auf einer tieferen Berücksichtigung des Sensordesigns basieren. Dieser Ansatz verbessert die Fähigkeit von Modellen, semantisch aussagekräftige und generalisierbare Repräsentationen im Rahmen des selbstüberwachten Vortrainierens zu erlernen.

Diese Beiträge zur bestehenden Literatur treiben die Entwicklung breit einsetzbarer und skalierbarer Modelle für Erdbeobachtungsdaten voran. Dies ist ein Schritt hin zu einer vereinfachten, breiteren und für Nutzer mit unterschiedlichen fachlichen Hintergründen leichter zugänglichen automatisierten Auswertung von Erdbeobachtungsdaten.

Contents

Acronyms

1

Introduction

Over the past decade, the field of satellite-based *Earth observation* (EO) has experienced rapid advancements, with even more progress projected for the near future. One significant milestone – particularly for the European community – was the introduction of the open data policy *Copernicus* constellation, developed and operated by the *European Space Agency* (ESA). Notably, the flagship of the constellation, *Sentinel-2*, is equipped with a multi-spectral sensor that began delivering data at a significantly improved spatial resolution of 10 m GSD, surpassing that of other freely available sources. Since this resolution approximately corresponds to the characteristic scale of many human-made structures, it has enabled a wide range of new applications. This development coincided with rapid advances in the computer vision community, particularly the shift toward solving vision tasks using *deep learning* (DL) in an end-to-end manner. Due to the symbiotic relationship between the two disciplines, research conducted around applying DL to satellite-based EO data has grown significantly over the past decade. As a result, DL has become the predominant approach for many EO-related tasks, while EO data has emerged as an increasingly prominent application area at major computer vision research conferences.

Still, both fields continue to develop rapidly. The number of sensors in space and their respective capabilities are growing quickly. This growth is partially driven by the launch of commercial satellites as well as subsequent missions by major space agencies, including ESA and *National Aeronautics and Space Administration* (NASA). To name a few examples: on the optical multi-spectral side, companies like *PlanetLabs* have introduced daily revisit products with 3 m GSD, multiple companies (e.g. *Pixxel*, *Satellogic* or *Wyvern*) are working on hyper-spectral missions and companies like *Iceye*, *Capella*, and *Umbra* are providing high-resolution SAR data. In addition to that, new governmental missions such as *Biomass*, *NISAR*, and *EnMAP* have been launched, and space

agencies are already planning the next generation – e.g., the Landsat successor or the next generation of Sentinel-2 with significantly improved temporal, spectral, and spatial resolution. Remarkably, all of the above-listed examples only cover the most well-known missions and do not even remotely encompass the full spectrum of missions measuring the Earth's surface with sensors of various modalities.

At this point – given the above-stated amount of various EO missions – it is important to note that all sensors, even those belonging to the same modality class, have different properties and measure different aspects of the Earth's surface. Rarely, two instruments from different constellations are identical. The numerous technical design decisions and trade-offs involved result in a diverse set of sensors, and consequently, a wide range of derivable information. While increased data access is fundamentally positive – since new applications are opening up – the volume and diversity of the data make manual analysis difficult, highlighting the need for automated solutions, such as deep learning-based approaches.

However, (fully supervised) DL methods require large amounts of labeled data for training, which poses a major challenge in satellite-based EO for several reasons. For example, annotating data is especially difficult for humans when observing a relatively small spatial extent of the data or when the resolution is coarser than the characteristic length scales of objects or other human-made structures. Moreover, many of the important features utilized by DL approaches – such as spectral characteristics or temporal trends – are due to the high dimensional nature difficult for humans to capture and hence hard to annotate. The alternative approach – collecting labeled data through field studies – is intrinsically limited by scale. This leads to a bottleneck for fully supervised training approaches.

One strategy emerging from the computer vision community (or more generally, the DL community) to address the issue of an insufficient amount of annotated data – common to many other disciplines as well – is self-supervised pre-training. Here, DL models are pre-trained on unlabeled data using related, synthetically created so-called *pretext* tasks. This presents a promising approach for EO data due to the intrinsic availability of vast amounts of unlabeled data collected by the growing number of sensors. While self-supervised pre-training has been successfully applied to many computer vision tasks, the concept got significantly more attention after the successful applications to the domain of natural language processing and the resulting modern large language models. Even though the fact that larger models (more tunable parameters) present – if trained correctly – the opportunity to perform better on many tasks [1, 2, 3] is a long-standing presence in the field of machine learning, the concept of increasing model size together with self-supervised pre-training got re-branded to the term *foundation model*. At the time of writing this thesis – and during the preceding years in which most of its content was developed – this term had been adopted by related fields such as EO and climate science, although the definition and generalizability of models referred to in this way had not yet been fully established. Since, on a technical level, no new strategies besides the utilization of self-supervised pre-training and the scaling of model sizes are introduced, this thesis does not make use of this terminology.

This thesis proposes novel approaches to perform self-supervised learning in the above-described scenario of a growing number of diverse EO data. If the diversity of sensor data should be leveraged, careful design decisions in the algorithms and learning pipelines are necessary. Neglecting the differences in sensors or acquisition conditions often leads to unwanted invariances in the resulting models. This thesis addresses this challenge from different angles, where the focus lies on incorporating the differences between sensors and acquisition conditions into learning pipelines based on self-supervised pre-training followed by fine-tuning for specific downstream tasks. To this end, knowledge about the design of sensors in multi-sensor setups, the specific geometric acquisition conditions, and the information content captured by different modalities is explicitly utilized. This contributes toward the long-term goal of observing our planet and its ecosystems at dense temporal scales using multiple sensors, while maintaining the distinctions between different sensors and modalities.

Structure of the Thesis

The rest of this thesis is structured as follows: After providing an overview of all publications produced during the course of this PhD, Chapter 2 presents the foundations of satellite-based Earth observation as well as basic deep learning concepts. This includes a brief introduction of general EO concepts, followed by an outline of the physical principles behind optical multi-spectral imagery and *synthetic aperture radar* (SAR). In both cases, emphasis is placed on the sensor parameters relevant to the remainder of the thesis. Subsequently, Chapter 2 also offers a brief introduction to the general field of *deep learning* (DL), covering basic architectural designs and self-supervised pre-training paradigms. Here, the two most common approaches are discussed in more detail, which form the basis for the methodological part of this thesis later on. Following the introductory sections, Chapter 3 outlines the current state of the art in self-supervised learning approaches, including more specifically sensor-informed learning strategies in the context of EO data, before putting the contributions of this thesis into perspective. This is followed by Chapter 4, which describes in detail the three methodological frameworks developed over the course of this dissertation, followed by a description of the datasets and corresponding sensors presented in Chapter 5. Chapter 6 outlines the experimental design used to evaluate the proposed frameworks and presents the corresponding results. In Chapter 7, the results are interpreted and connections between the different approaches – where applicable – are discussed. Finally, Chapter 8 places the findings in a broader context and outlines open questions for future research in the field.

List of Publications

This thesis, in its core, outlines the results and thoughts behind three main publications that deal with the topic of sensor-informed self-supervised learning for EO data. Those publications are centered around the questions of training a DL network that is capable

of processing data from multiple multi-spectral satellites in a sensor-informed manner, the incorporation of geometrical acquisition conditions in the case of high resolution SAR images, as well as the utilization of multi-modal data for pre-training. Hence, a majority of the results presented in later chapters can also be found in the three double-blind peer-reviewed publications that are published in the proceedings of reputable computer vision conferences:

- Jonathan Prexl and Michael Schmitt. SenPa-MAE: Sensor parameter aware masked autoencoder for multi-satellite self-supervised pretraining. In Proceedings of GCPR, pages 317 – 331, 2024.

- Jonathan Prexl, Michael Recla, and Michael Schmitt. SARFormer – an acquisition parameter aware vision transformer for synthetic aperture radar data. In Proceedings of CVPR Workshops, in press, 2025.

- Jonathan Prexl and Michael Schmitt. Multi-modal multi-objective contrastive learning for Sentinel-1/2 imagery. In Proceedings of CVPR Workshops, pages 2135 – 2143, 2023.

In addition, this thesis builds upon a further contribution in the form of a textbook chapter entitled *Self-supervised learning for multi-modal Earth observation* published in the textbook:

- Jonathan Prexl and Michael Schmitt. Chapter 8: Self-supervised learning for multi-modal Earth observation data. In Saha Sudipan, editor, Deep Learning for Multi-Sensor Earth Observation, pages 181–199. Elsevier, 2025.

While this thesis mainly presents the content of the three above-mentioned works, during the time of this PhD, various other publications in the broader context of DL for EO data have been published and listed here for the sake of completeness. Even though results are not directly presented within this thesis, during the creation, collaborations, and participation at conferences, many valuable experiences were gained that highly influenced the main contributions with respect to implementation and experimental design. This comprises the four peer-reviewed articles:

- Jonathan Prexl and Michael Schmitt. The potential of Sentinel-2 data for global building footprint mapping with high temporal resolution. In Proceedings of JURSE, pages 1–4, 2023.

- Ribana Roscher, Marc Russwurm, Caroline Gevaert, Michael Kampffmeyer, Jefersson A Dos Santos, Maria Vakalopoulou, Ronny Hänsch, Stine Hansen, Keiller Nogueira, Jonathan Prexl, and Devis Tuia. Better, not just more: Data-centric machine learning for Earth observation. IEEE Geoscience and Remote Sensing Magazine, 12(4):335–355, 2024.

- Jonathan Prexl, Anton Baumann, and Michael Schmitt. A comparison of uncertainty estimation methods for building footprint change detection from Sentinel-2 imagery. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 10:339–346, 2024.

- Deepika Mann, Jonathan Prexl, Michael Schmitt, and Felicitas Sommer. Multiscale assessment of land use efficiency in functional urban areas of Munich and Augsburg. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 10:231–236, 2024.

as well as two additional abstract reviewed contributions to conferences centered around self-supervised learning for EO data:

- Jonathan Prexl and Michael Schmitt. The effect of contrastive pretraining on downstream tasks in optical remote sensing. In Proceedings of IGARSS, pages 526–529, 2023.

- Jonathan Prexl and Michael Schmitt. Sensor parameter encoding for multi-sensor self-supervised learning via masked autoencoders. In Proceedings of IGARSS, pages 2837–2840, 2024.

and two further abstract reviewed contributions covering downstream task-related works:

- Jonathan Prexl, Sudipan Saha, and Michael Schmitt. High precision mapping of building changes using Sentinel-2. In Proceedings of IGARSS, pages 6744–6747, 2023.

- Deepika Mann, Jonathan Prexl, Sudipan Saha, and Michael Schmitt. Mapping high-resolution building development over Delhi NCR using Sentinel-2. In Proceedings of IGARSS, pages 3190–3193, 2024.

2
Foundations

This chapter lays the theoretical ground for the methodological innovations presented in this thesis and is divided into two sections. First, the basic principles of satellite-based Earth observation are introduced, covering physical foundations, terminology, and sensor classes as well as, in more detail, the technologies underlying multi-spectral and SAR based Earth Observation data. This is followed by an introduction to the foundations of deep learning, aspects of the architectural designs, and the basics of the self-supervised learning approaches that form the basis of this thesis. The structure of the first section is loosely inspired by commonly referenced textbooks in the field, such as [4, 5, 6], while introducing variations in the order and depth of topics according to the needs of this dissertation. The second part is partially based on a previously published book chapter, which was produced during the research period covered by this dissertation [7] as well as on commonly referred review articles [8, 9, 10].

2.1 Satellite-Based Earth Observation

The field of remote sensing covers a broad range of sensors and sensor platforms, all with the shared goal of observing the Earth's surface via its interaction with electromagnetic radiation. Within this thesis, all argumentation is restricted to the subclass of spaceborne remote sensing observing the land surface, hence leaving out classes of sensors that are dedicated to the measurement of atmospheric or oceanic properties, as well as the particularities of airborn systems. Beyond that, the term remote sensing will be used interchangeably with the term EO, which refers in all cases to spaceborne Earth observation. Below, the basic ideas of satellite-based EO are introduced to give readers from other disciplines a basic intuition.

Note: Depending on the definition, the term remote sensing also often includes various disciplines, e.g., photogrammetric near field measuring techniques. Within this thesis, it refers strictly to satellite-based Earth observation.

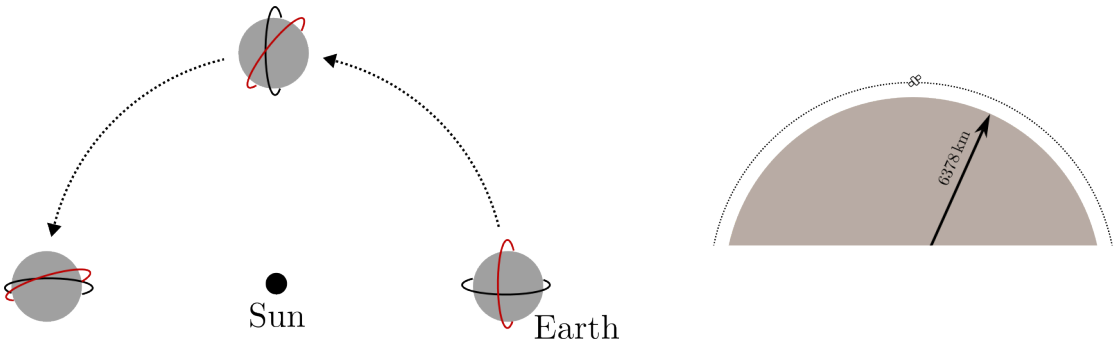


Figure 2.1: (Left) Demonstration of a sun-synchronous orbit. Over the different periods of the year, the sun-synchronous orbit (black) keeps a stable relative position with respect to the Sun. (Right) A to-scale illustration of the orbit height for a LEO with respect to the Earth's radius.
(The left part of the graphic is adapted from [11].)

The Position of the Sensors In order to observe the Earth's surface from space, the satellites carrying the sensors are placed in the *low Earth orbit* (LEO) several hundred kilometers above the Earth's surface[1]. All sensors that are used in this thesis are on (near) polar orbits that are *sun-synchronous*. This means that the season does not affect the revisit schedule or lighting conditions during measurements. Additionally, the energy supply of the satellite is independent of seasonal variations. The geometrical relation of the sensors with respect to the Sun in case of a *sun-synchronous* orbit is graphically illustrated in Fig. 2.1.

The Sensor's Field of View and Revisits Satellites in LEO typically operate in one of two sensor modes: either continuous monitoring in a fixed direction or scheduled observations that can target any land surface visible from the sensor's position. In the continuous monitoring mode, the sensor points directly downward[2] – referred to as nadir – along the satellite's ground track, providing consistent coverage of the Earth's surface. The width of the observed area perpendicular to the flight direction is referred to as the sensor's swath (compare Fig. 2.2). In scheduled observation mode, the sensor can be programmed to image at nadir or at an angle (off-nadir), allowing flexibility to capture specific areas. This mode is commonly used for higher-resolution imaging with a smaller swath, while medium-resolution missions operated by state authorities typically aim for periodic global coverage with the sensor always steered in nadir direction. In both modes, the time between two subsequent observations of the same location is referred to as the sensor's revisit time. In continuous monitoring mode, this concept is more straightforward, as the satellite regularly passes over the same areas. In scheduled observation mode, revisit time also depends on the sensor's ability to steer off-nadir and on how imaging opportunities are prioritized. For continuous monitoring missions, the swath width where the sensor acquires data is typically between several tens and several hundreds of kilometers for medium-resolution sensors. As explained later, there

[1] LEO is considered to be in the range of 180 km to 2000 km (compare [5]), whereas all sensors used within this thesis are within the $\sim$ 450 km to $\sim$ 780 km range.
[2] Except in the case of SAR sensors, which are always side-looking, as will be discussed later.

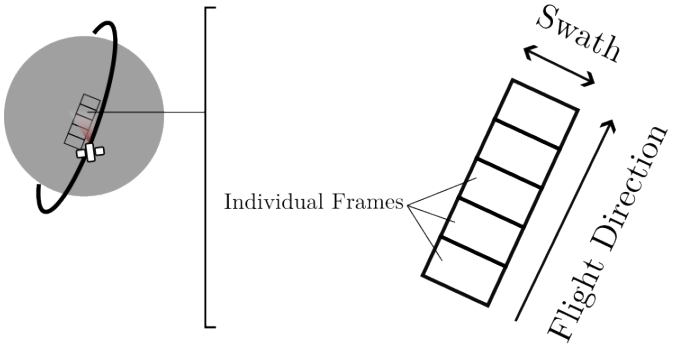


Figure 2.2: The width of the section that the sensor is capturing data along track is referred to as the sensor's swath, while data acquired by the end-user is usually subdivided into individual frames.

is a trade-off between resolution and revisit time (along with other sensor parameters). These trade-offs in sensor design result in a landscape where a complete picture of the Earth's surface is obtained through multiple sensors, each with its own specialization.

Multi-Modal Earth Observation Sensors technologies utilized in satellite-based EO are typically divided into modality classes according to their measurement principle and type of electromagnetic radiation they are observing. In a categorization of first order, sensors can be distinguished by their active or passive measurement principle, depending on wether the electromagnetic radiation for the sensing process is emitted from the sensor (active) or if *natural* sources (sun or thermal radiation) provide the electromagnetic radiation for the measurement (passive). In the following two sections, an example of each (multi-spectral and SAR) will be discussed in detail. In addition to the sensing strategy, the wavelength of the utilized electromagnetic radiation as well as the spectral resolution help to set apart different modality classes. Resulting examples would be the passive sensor of the modality classes multi-spectral, hyper-spectral, or thermal sensors. Fig. 2.3 gives a graphical overview of the spectral position of commonly used modalities in the field of remote sensing, together with the emission spectrum of the sun and the Earth (which necessarily restricts the potential wavelength regimes for passive sensors) as well as the *atmospheric transmittance spectrum* (ATS) of the Earth which further limits potential wavelength regimes when designing sensors.

Note: The term modality can also refer to audio, natural language, or other data types, but here it is used in a stricter sense.

Growing Amount of Sensors Spaceborn EO data for social and civilian (in contrast to military) purposes was first acquired by launch of the still ongoing Landsat program (started in the year 1972, compare [13]) by NASA. With the successful succession of Landsat missions – from Landsat-1 to the current Landsat-9 – and a decades-long archive of Earth observation data acquired by NASA, a major shift in accessibility and impact for both the public and the scientific community occurred in 2008 by the transition to a free and open data policy. This marked the beginning of a new era in which

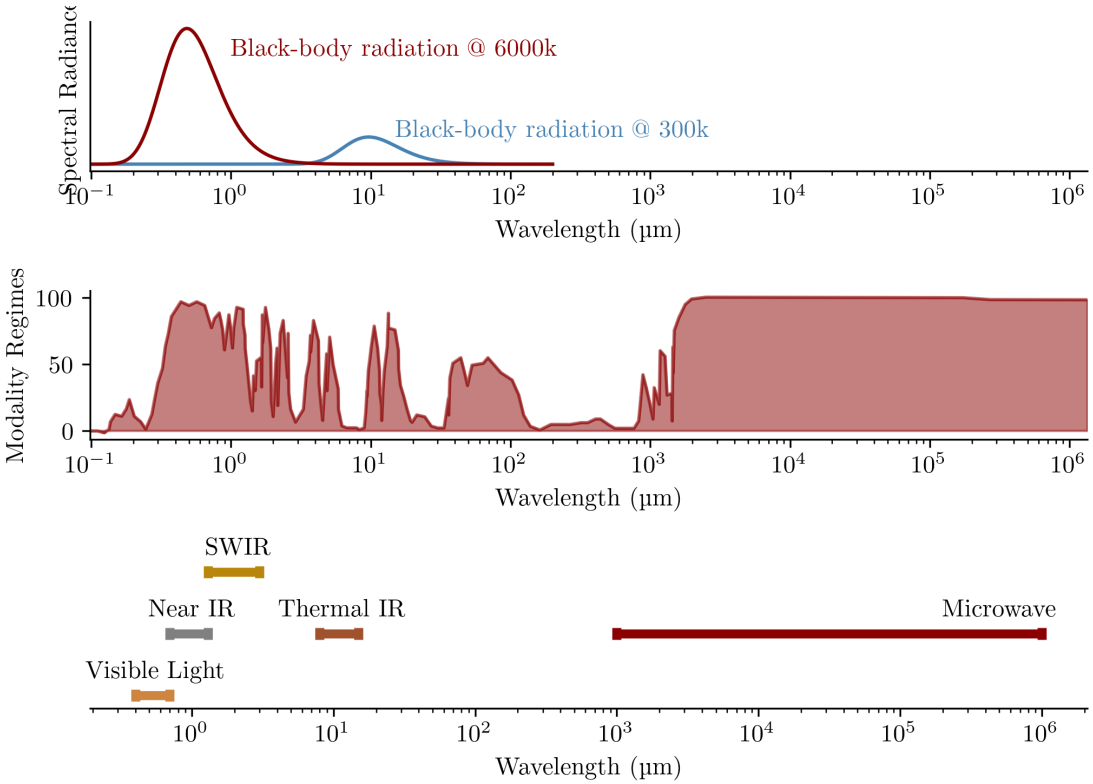


Figure 2.3: (Top) Black-body radiation for 6000 K (approx. the Sun's temperature) and 300 K (approx. the Earth's temperature) accordingly to Plank's law and under the black body assumption (compare [12]). The black body radiation of the Earth is artificially enhanced by several orders of magnitude for visibility. (Middle) The transmissivity of the Earth's atmosphere as a function of the wavelength. (Bottom) The spectral ranges commonly used in the context of Earth observation technology, used to categorize modality classes. (Graphic adapted from [4].)

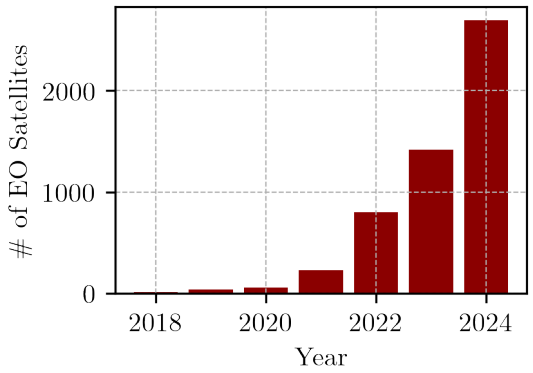


Figure 2.4: The number of EO satellites as a function of the corresponding launch year. (Data underlying this graphic is taken from adapted from [14].)

satellite data became an essential resource for monitoring changes on Earth's surface. Another major milestone was reached in 2014 with the launch of the *Sentinel* program by ESA, which introduced open data access for medium resolution SAR data as well as data from its multi-spectral sensors. Meanwhile, the two space agencies, among others, are planning successors to their respective missions, pushing the technical limitations of this freely available data, while commercial companies are revolutionizing the sector with satellite constellations delivering near-real-time, high-resolution imaging. As a result, the number of sensors is continuously increasing and is expected to keep growing (compare Fig. 2.4).

Within this thesis, methods are introduced that interoperate domain knowledge to design more efficient self-supervised pre-training strategies for the two modalities of multi-spectral and SAR spaceborne Earth observation imagery. The following two sections describe the two modality classes in more detail.

2.1.1 Optical (Multi-spectral) Earth Observation

Multi-spectral cameras on board of satellites are passive sensors that measure the reflected sunlight of the Earth's surface, divided into multiple wavelength ranges as so-called bands or channels. The measurements are commonly conducted in the visible and the near infrared parts of the spectrum and divided into tens of bands, whereas sensors with significantly more bands – sometimes ranging into the low hundreds – would be considered hyper-spectral. This section covers the relevant physical and technological concepts for multi-spectral EO data.

The Sun as an Energy Source Multi-spectral sensors, as a passive sensing technique, use the electromagnetic radiation emitted by the Sun, reflected by the Earth's surface, and detected by satellite-mounted sensors. The initial distribution of electro-

magnetic radiation intensity is determined by the Sun's emitted light, which can be approximated using Planck's law for black-body radiation [12]. Fig. 2.3 shows this approximated distribution as a function of the light's wavelength, where, approximately within the interval [200 nm, 1600 nm], the intensity of the emitted light exceeds 10 % of the distribution's maximum which lies around $\sim$ 480 nm. From this, the potential wavelength regime for all passive sensor systems (excluding thermal cameras, which use the Earth's emission spectrum as a source) is limited to the range where sufficient electromagnetic radiation is present. Further restrictions are imposed by the atmospheric interactions described in the next paragraph.

Atmospheric Effects The atmosphere and its specific chemical composition at a given point in time affect the measurement of Earth's surface reflectivity from space in various ways. Several physical processes alter the intensity as a function of the wavelength and hence the spectral composition of the observed signal. Those physical processes act along the two-fold transit through the atmosphere, once from the sun to the Earth's surface and after the reflection, back to the sensor. The dominating processes are absorption (compare ATS in Fig. 2.3) and scattering processes that weaken the intensity along the main path of directed radiation. Scattering takes place at small particles and molecules as well as on larger particles such as dust, water vapor, or water droplets. Further, molecules in the atmosphere absorb electromagnetic radiation and – due to the narrow energy modes excited in the molecules – comparably sharp absorption lines show up in the overall atmospheric transmission spectrum [12]. Those processes, among all others, are a function of the wavelength and typically lead to non-uniform changes of the relative amplitudes in various channels. Beyond the effects described above, which weaken the directional radiation, the scattering effects further introduce indirect radiation paths to influence the distribution of the electromagnetic radiation observed by the sensor. This, again, can be separated along the two-fold transit through the atmosphere. Firstly, incoming scattered light enhances the radiation density on the target through diffuse sunlight (skylight) which was – before the scattering process – not directed to the position on the ground that is observed. Secondly, scattered light can be directly deflected in upper layers of the atmosphere and directed to the sensor without interacting with the Earth's surface or any lower atmospheric layer. Both mechanisms are again a function of the wavelength of the electromagnetic radiation. A graphical illustration of the concepts and the direct and indirect radiation paths that influence the observed radiation distribution is given in Fig. 2.5.

The above-described processes lead to a change in observed intensity and spectral composition by the sensor, all depending on the exact chemical composition of the atmosphere. Satellite providers commonly correct for those effects by taking the atmospheric state into account. Corrected imagery is commonly referred to as *bottom of atmosphere* (BOA) (or *surface reflectance*), in contrast to *top of atmosphere* (TOA), where the correction for the atmospheric state is not applied. Fig. 2.5 provides a comparison of BOA and TOA data of a Sentinel-2 image captured over the Munich metropolitan area. It can be observed that for most positions on the ground, the BOA values exceed the non-corrected TOA since the absorption and scattering of the directed radiation dominate.

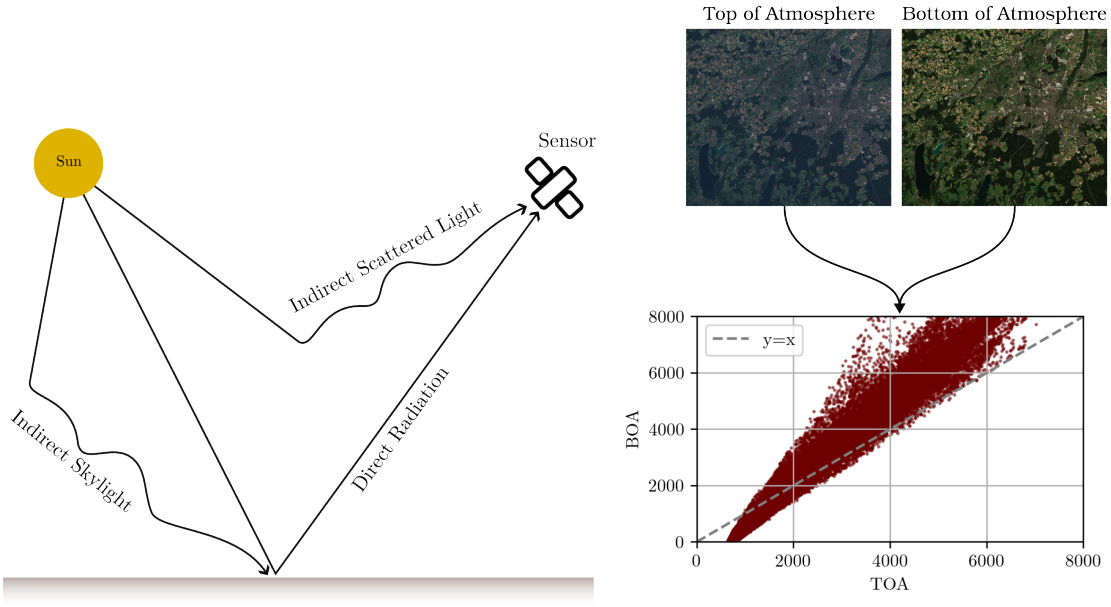


Figure 2.5: (Left) Illustration of the radiative transfer process in a typical Earth observation setting. The target gets illuminated by direct sunlight as well as diffuse skylight. The sensor observes the analog phenomena by capturing the directly reflected light as well as diffuse scattered sunlight. (Right top) Example images for *top of atmosphere* (TOA) data and *bottom of atmosphere* (BOA) data for a test area around Munich captured by Sentinel-2. (Right bottom) Correlation of the numerical values for the blue channel (B02) of Sentinel-2 for TOA data- and BOA data format. (The left part of the graphic is adapted from [4].)

Still, for generally lower reflection values, the effects of indirect skylight dominate and TOA values exceed the non-corrected BOA (compare bottom right of Fig. 2.5).

Interaction with the Earth Surface Treating the atmospheric interactions as an accountable variable, the main subject of the measurement in multi-spectral EO is the (spatially) integrated reflective properties over the corresponding position of the ground as a function of the wavelength. This spectral dependency of the reflection curve is a descriptive feature for e.g. the *land-use land-cover* (LULC) class as well as certain physical quantities like soil moisture [15], vegetation parameters [16], water quality [17], agriculture related classification [18] or biomass [19]. This adds to the geometrical features that are given by the image layout of the data. An example of the statistically averaged spectra for different land-cover classes can be found in Fig. 2.6 where the discriminative nature of the spectra of different LULC is shown.

Sensor Parameters After the interaction with the Earth's surface and the second pass through the atmosphere, the multi-spectral camera measures the intensity of the incoming radiation divided into several wavelength ranges, so-called bands or channels.

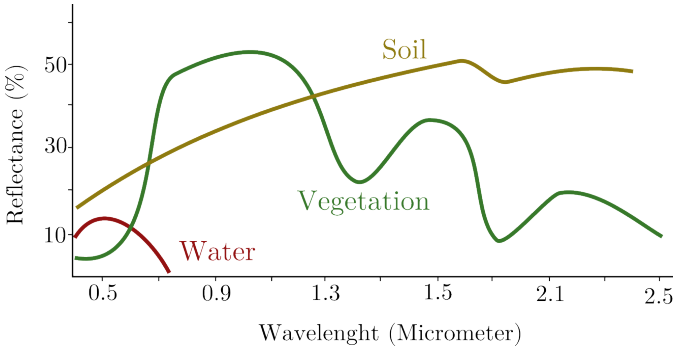


Figure 2.6: Illustration of the different reflection patterns for different LULC classes. Most notably, vegetation-related classes show a strong increase around 700 nm, whereas the water class has (next to generally lower overall reflectance) almost no reflectivity for longer wavelengths.
(The data for this graphic is extracted from `https://seos-project.eu/classification/classification-c01-p05.html`, accessed 16.05.2025.)

For this, either filter materials or optical diffractive components (e.g. diffraction gratings or prisms, which are more commonly used for hyper-spectral missions) are utilized. All optical sensors used in this thesis operate in a scanning mode that forms images by exploiting the satellite's along-track motion. Some sensors use a *push broom* scanner configuration, where a line perpendicular to the flight direction is projected onto a linear detector array, and successive lines are captured to build a two-dimensional image. Others employ a *frame-based* strip sensor layout, where narrow image strips are acquired and subsequently mosaicked along the orbital path to form a continuous image. The most relevant parameters to interpret measured multi-spectral intensities are the corresponding *spectral response function* (SRF) for each channel, which combine the optical properties of the diffractive medium as well as the absorption and reflection characteristics of all other components in the optical system. Fig. 2.7 illustrates this

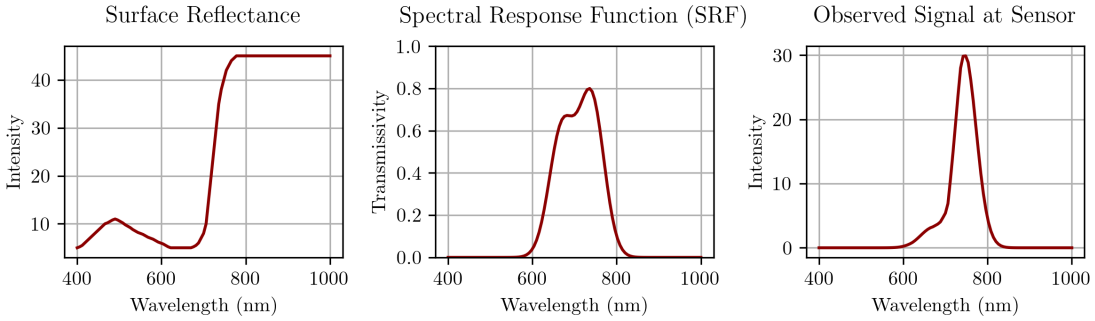


Figure 2.7: Illustration of the spectral response function and the effect on the observed signal. (Left) Incoming light with a certain spectrum is observed through an instrument with an effective spectral response function (middle), which combines all reflective and absorption characteristics of all installed optical components. This leads to an altered observed distribution of electromagnetic radiation with different wavelengths at the imaging plane (right).

concept. Independent of the magnitude of the incoming radiation (left), the measured

relative intensities (right) are a function of the SRF of the sensor (middle). This is – for every band of any sensor in orbit – a known quantity and is indispensable if measured intensities should be linked to physical or chemical quantities or when data from different sensors are subject to a comparison or merging operation. As will be seen later, the corresponding SRF of the most used multi-spectral satellite missions change significantly with regard to their central wavelength position as well as their width and shape.

In addition to the SRF and the number of spectral bands – which together define a sensor's spectral resolution – the spatial (or geometrical) resolution is a crucial parameter, as it directly influences the smallest distinguishable feature size in the final image. Spatial resolution is determined by the optical system's *point spread function* (PSF), which defines the minimum separation at which two point sources can be resolved. A graphical illustration of this concept is shown in Fig. 2.8. For two infinitesimally small point sources of electromagnetic radiation, the optical system's PSF causes their signals to spread in the imaging plane. The sensor's pixel spacing is typically chosen to match its resolution limit, commonly defined as the point where the distance between sources exceeds the *full width half maximum* (FWHM) of the intensity distribution. This spacing is referred to as the sensor's *ground sampling distance* (GSD), which may vary across different channels of the system.

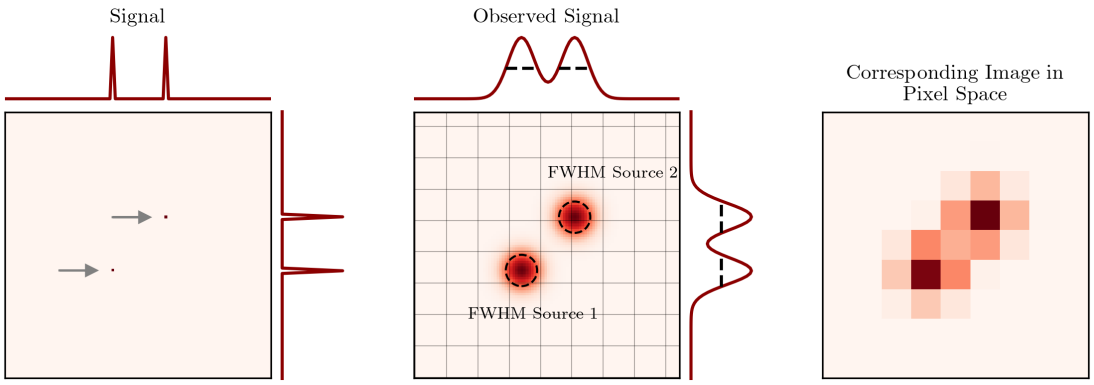


Figure 2.8: Illustration of the concept of a *point spread function* (PSF) for a given optical system. The left subplot shows the original signal (two infinitesimal point sources indicated by gray arrows). The middle subplot displays the measured signal at the sensor plane after passing the optical system, where the intensity after the imaging process is blurred out. Here, the PSF is approximated by a Gaussian bell curve. Pixel spacing for sensors is commonly set to match the resolution limit given by the FWHM value and is sketched in gray. A visualization of the resulting image in pixel space is given in the right subplot.

When designing a satellite mission, key sensor parameters such as swath width, revisit time, spectral resolution (number and width of channels), and spatial resolution (linked to GSD) involve trade-offs and must be carefully balanced against one another. Reducing the spatial area where light is collected decreases the overall signal intensity. Similarly, narrowing the spectral width of the channels also reduces the intensity of collected signal.

When both spatial and spectral reductions occur simultaneously, the *signal-to-noise ratio* (SNR) can become critically low, posing challenges for reliable data acquisition.

2.1.2 Synthetic Aperture Radar

Note: Radar is an acronym itself and stands for *Radio Detection And Ranging*

As previously mentioned, another commonly used sensing technique in the field of satellite-based Earth observation is given by *synthetic aperture radar* (SAR) imaging. One of the most valued feature of a SAR system is given by the fact that the wavelength of the utilized electromagnetic radiation has the ability to penetrate clouds and is therefore capable of capturing an image under all weather conditions and - due to the active sensing technique - also during nighttime. Those properties make SAR data a well-complementary data modality to the previously discussed multi-spectral sensors. Generally, SAR data acquisition, signal processing, and related applications encompass a wide range of fields and methodologies. Here, often the phase information of the recorded signal is utilized for interferometric and tomographic applications. In this section, the discussion is limited to focused backscatter SAR images, as the methods developed within this thesis are specifically designed to process data of this type. The following is a discussion of the underlying measurement principles as well as the relevant sensor parameters that are utilized throughout this thesis.

Distance Measuring Principle SAR sensors belong to the class of active microwave remote sensing and hence work with measuring the reflections (echos) of actively emitted electromagnetic pulses with wavelengths in the order of $1\,\mathrm{mm} - 1\,\mathrm{m}$ (compare Fig. 2.3). Due to the side-looking acquisition geometry characteristic of all SAR systems (compare Fig. 2.9), the recorded echoes exhibit a time delay that varies with the ground position where the electromagnetic radiation gets reflected. Fig. 2.9 displays the difference in travel time for reflections occurring at different positions on the ground graphically. Here, in addition to the time delay caused by varying distances, the observed intensity of the returned signal also contains information about the material properties and geometry of the reflecting surfaces. The strength of the returned signal is commonly expressed in terms of the *backscatter coefficient*, denoted by σ^0 (sigma nought). This quantity represents the radar signal power scattered back to the sensor per unit area on the ground and is influenced by variables such as surface roughness, dielectric properties, and the incidence angle. For easier interpretation and to handle the wide dynamic range of radar returns, σ^0 is commonly expressed on a logarithmic scale in *decibels* (dB). However, the observed amplitude does not only reflect the macroscopic properties of the surface. Due to the coherent nature of radar systems, SAR images are also affected by the so-called *speckle* effect. Speckle results in a granular noise-like pattern, resulting from the constructive and destructive interference of radar waves reflected by many unresolved targets within a single resolution cell [6]. While resulting speckle patterns appear random, it is a deterministic effect inherent to the coherent nature of SAR imaging and introduces a significant variability in the pixel-wise measure of σ^0. It is worth mentioning here – especially after introducing the bottom of atmosphere

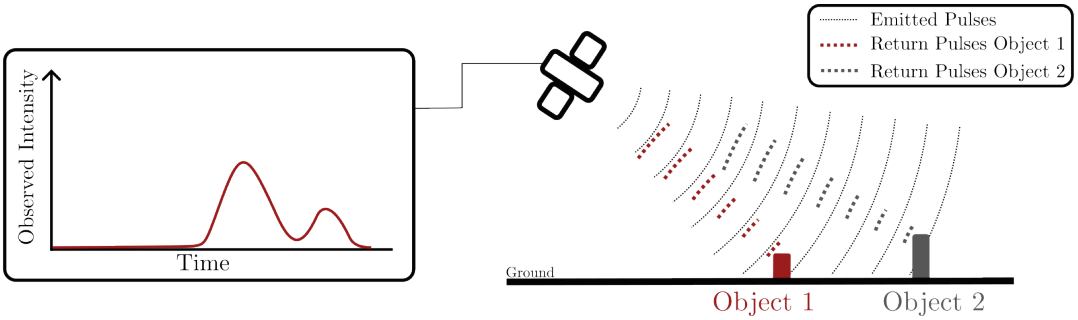


Figure 2.9: Illustration of the side-looking SAR geometry, where emitted radar pulses sent from the sensor are subject to different travel times depending on the position of the reflection on the ground (here object one and two). Next to the time delay, the overall returned intensity can vary due to geometrical and material properties, as indicated in this figure, where the returned intensity of object two is lower compared to object one.
(Graphic adapted from [4].)

reflection concept used in the previous section – that the amplitude of the backscatter signal is not meant to be interpreted as the reflectivity of the surface with respect to the radar wavelength. Quite the opposite, strongly reflecting surfaces – such as water under certain geometric conditions – can appear dark in the final image due to a lack of backscattered radiation directed toward the sensor[3].

Due to the constant travel speed along-track, subsequent measurements can be represented in an image-like data format, where one dimension corresponds to the sensor's position along the flight path and the other to the range distance (directly correlated with the above mentioned travel time of the pulses), where pixel values encode the intensities of the recorded echoes. This is referred to as the *slant range* representation since no direct correspondence to the actual position of the ground can be made. *Slant range* images can be transferred into *ground range* image by projecting the signals onto a reference surface[4]. Fig. 2.10 illustrates the concepts of slant- and ground range and their respective conversion visually.

Resolution Radar image resolution is defined separately in two directions: *range* and *azimuth*. The range direction refers to the line-of-sight distance from the sensor to the target, while the azimuth direction follows the sensor's flight path. These dimensions are subject to different physical and processing constraints.

Range resolution is primarily governed by the duration of the transmitted radar pulse, which is commonly characterized by its bandwidth.[5] Shorter pulses enable finer discrimination between echoes from closely spaced targets. However, reducing the pulse

[3] Similarly, effects such as volume scattering or double-bounce reflections (see [6]), which are beyond the scope of this thesis, also do not permit such an analogy.
[4] In the simplest case, the reference surface can be approximated by the Earth's ellipsoidal shape if no detailed elevation model is available.
[5] The bandwidth is inversely proportional to the pulse duration.

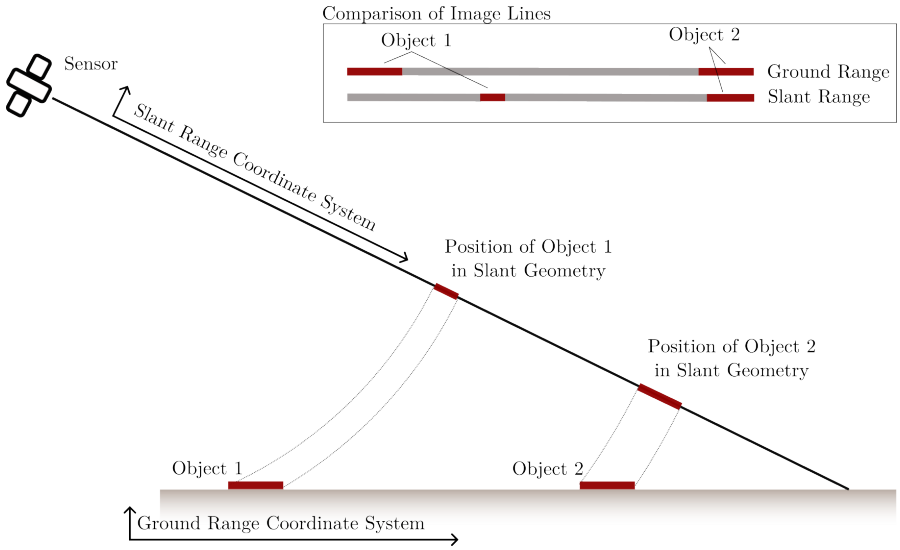


Figure 2.10: Visualization of the slant-range to ground-range projection. The projection step introduces transformations (so-called slant range distortions) in object sizes as a function of the distance to the sensor, such as for the equally sized object 1 and object 2, which – due to different positions on the ground – are not of equal size in the slant-range geometry. (Graphic adapted from [4].)

length is constrained by practical signal-to-noise limitations. This challenge is addressed through chirp modulation, where the pulse is frequency-modulated over time. By exploiting the known modulation pattern, it becomes possible to better localize the origin of reflections – even when pulses overlap – thereby effectively enhancing the resolution in the range direction. Azimuth resolution is initially limited by the antenna beamwidth, which broadens with increasing distance to the target (compare Fig. 2.11). As the platform moves forward, each ground target remains within the antenna footprint for several pulses. Utilizing the resulting Doppler frequency signature of the targets caused by the moving platform and coherently aligning and combining these echoes based on their Doppler characteristics, SAR systems synthesizes a longer virtual antenna, referred to as a *synthetic aperture*. This dramatically improves azimuth resolution beyond the limits imposed by the physical antenna. Overall, these concepts enable high-resolution radar imaging with range and azimuth resolution that are effectively independent of the distance from the sensor to the target (assuming slant-range geometry is maintained).

Nomenclature of the Acquisition Geometry Due to the intrinsic off nadir viewing geometry and the corresponding terrain-induced effects (described later), the exact geometrical position relative to the target influences the appearance of the final image in a much stronger way than in the case of medium resolution Earth observation data, described previously.

The two parameters that describe the geometrical constellation of the image acquisition are the *looking angle* θ and *azimuth angle* Az. The looking angle θ describes the off nadir component and - as will be shown later in this chapter - is highly correlated to the

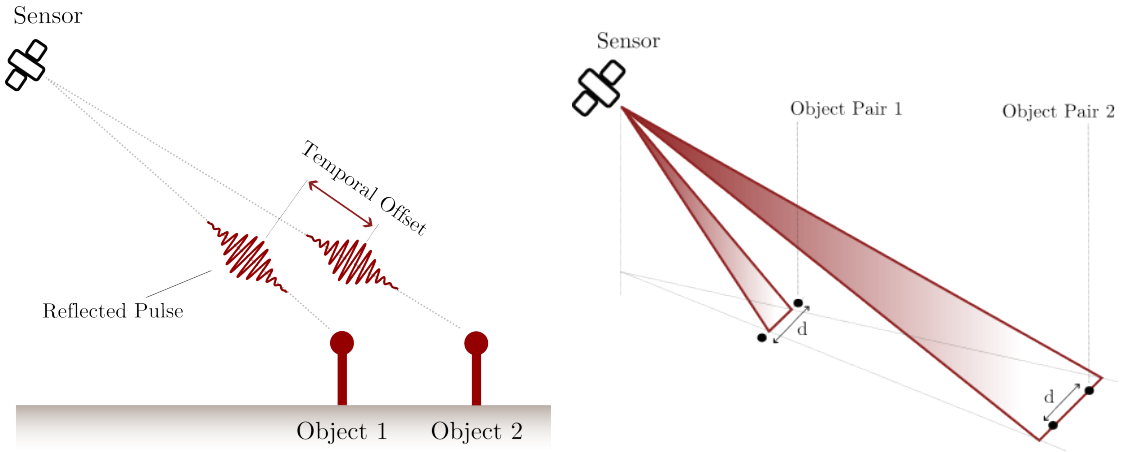


Figure 2.11: (Left) Illustration of how range resolution is determined by the duration of the radar pulse, where overlapping returns from different ground points lead to ambiguity of origin, limiting the resolution in range direction. (Right) Geometry of radar imaging in azimuth direction, highlighting how two ground points (A and B) with distance d can or cannot be resolved depending on the distance to the sensor if no advanced processing is applied. (Graphic adapted from [4].)

SAR typical terrain-induced geometric distortions. The azimuth angle Az is determined by the orbit and measured as the angle from the line of sight to the true north. Both angles are displayed graphically in Fig. 2.12.

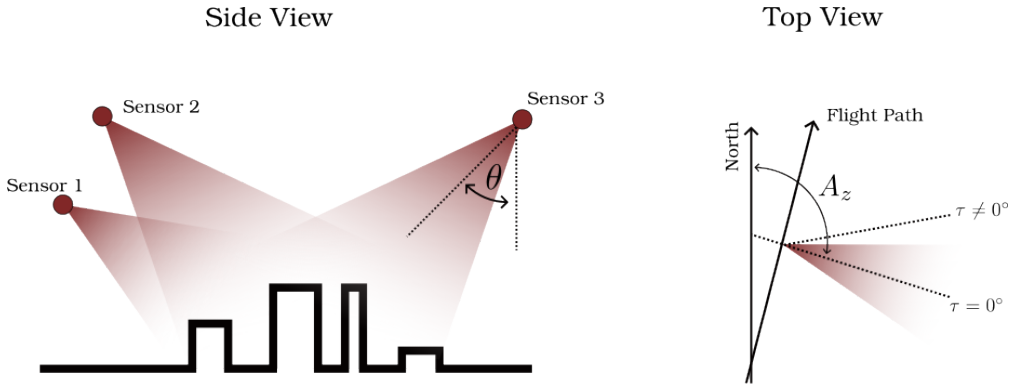


Figure 2.12: (Left) Three different configurations with different looking angles θ. (Right) The definition of the azimuth angle Az used within this thesis, as well as different squint angles τ utilized for high-resolution operation modes.

Terrain-Induced Effects SAR images are inherently unintuitive to interpret due to their measuring principle, which differs from human visual perception. Several effects can be isolated that differentiate a SAR image from its optical counterpart. Foreshortening describes the phenomenon that a sensor facing slope in the terrain appears shorter in the corresponding SAR image due to the fact that the distance from the elevated part of the slope comes closer to the sensor. This effect is a function of the slope's angle,

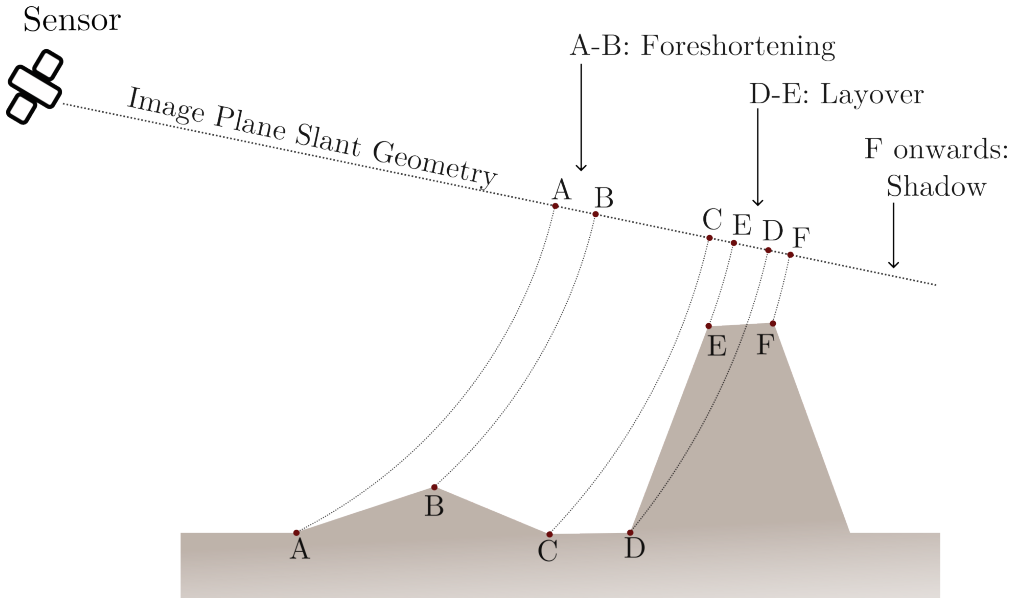


Figure 2.13: Illustration of SAR typical imaging effects. The slope on the ground surface between point A and B leads to the foreshortening effect, where the area appears smaller in the corresponding SAR image. For steep slopes (depending on the looking angle), this effect leads to a reversed order of the points in the SAR image, such as for points D and E. Further, steep elevated structures lead to parts of the ground without any illumination by the sensor (such as all positions onward from point F), which is referred to as a radar shadow.

and for large enough angles, the bottom point and top point of the sloped area change order in the SAR image. The latter represents an extreme case of foreshortening and is referred to as Layover. Beyond this step, slopes and raised structures such as buildings also generate shadowed areas without any illumination from the sensor, and hence also no corresponding echoes. All three effects are graphically displayed in Fig. 2.13 and are crucially important to take into account for interpreting SAR imaginary.

Scan Modes SAR systems can be operated with various so-called acquisition modes. The base configuration, which is outlined above and has parallels to the push broom scanner in the multi-spectral case, is the so-called *Strip-map* mode. Here, the sensor records echos without any steering of the emitted pulses towards one specific target. Beyond this, it is also possible to widen the observed area (swath) by steering the beam to different looking angles to cover multiple swaths, which is referred to as *ScanSAR*. Since the integration time for each swath is effectively smaller, *ScanSAR* exhibits a lower azimuth resolution. The rather opposite effect can be achieved by operating the sensor in *Spotlight* mode, where the target gets illuminated over a longer integration time by steering the beam along the azimuth direction. Fig. 2.14 gives a graphical overview of the three different classes of imaging modes described above.

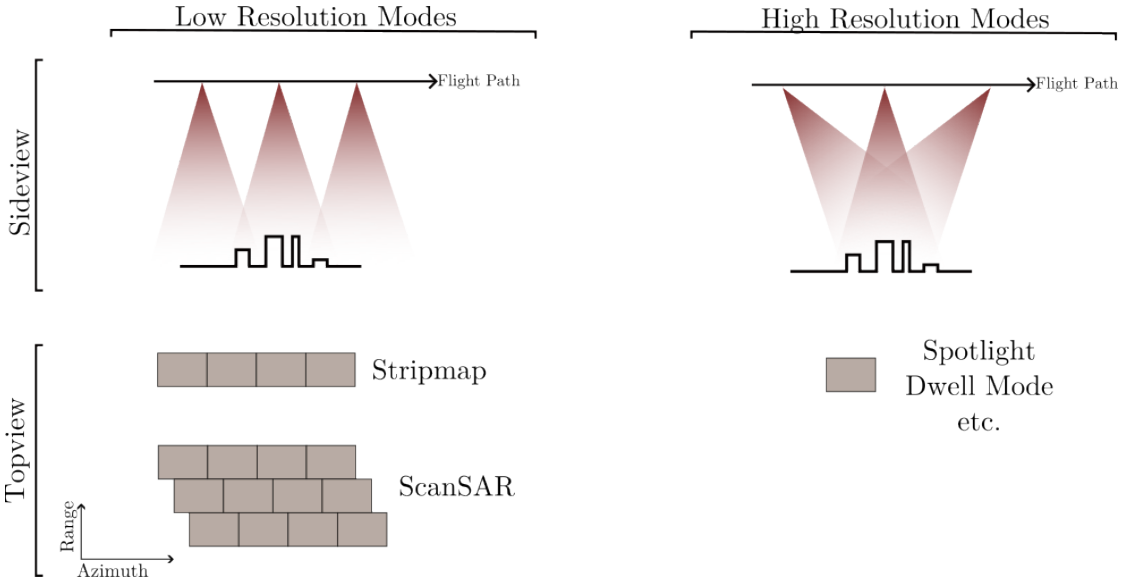


Figure 2.14: Common imaging modes for SAR operation. (Left) For the, in relation, lower resolution imaging modes, the sensor is pointed with a constant squint angle of 90° perpendicular to the flightpath. An even larger coverage can be created by additionally changing the looking angles along the flight path, which results in the so-called ScanSAR mode. Higher resolution in azimuth direction can be generated by long illumination times achieved by steering the sensor with different squint angles to a target area of choice.

SAR Bands SAR sensors operate in various frequency ranges, each offering distinct advantages and opening up diverse applications. Tab. 2.1 gives an overview of the commonly utilized frequency regimes for spaceborn EO SAR missions. While longer wavelengths enable deeper penetration into vegetation and soil, allowing the analysis of underlying structures, they come at the cost of lower azimuth resolution due to smaller Doppler shifts that are – as described above – used to enhance the resolution. Hence, sensors that operate with longer wavelengths are commonly used in environmental monitoring scenarios [20, 21], such as biomass estimation, while shorter wavelength sensors are more typically employed to monitor man-made structures [22, 23].

SAR Band	Frequency Range (GHz)	Wavelength (cm)
P-band	0.3–1.0	30–100
L-band	1.0–2.0	15–30
S-band	2.0–4.0	7.5–15
C-band	4.0–8.0	3.75–7.5
X-band	8.0–12.0	2.5–3.75

Table 2.1: Common frequency ranges for SAR sensors combined with their naming convention.

Polarizations SAR sensors are capable of transmitting and receiving signals with different polarization states, where the electromagnetic wave can be oriented either horizontally (H) or vertically (V). During both transmission and reception, the system can separate the backscattered signal into its horizontal and vertical components to analyze whether the reflection on the ground caused a change in polarization. The community-standard notation to describe these configurations uses two-letter acronyms, such as VV (vertically transmitted, vertically received), HH (horizontally transmitted, horizontally received), or cross-polarization modes like VH (vertically transmitted, horizontally received). This adds another dimension for distinguishing features in a SAR image, as certain scattering mechanisms – such as volume scattering from vegetation – are known to alter the polarization state, while man-made structures typically produce more stable, surface-dominated backscatter without such effects.

2.2 Deep Learning and Self-Supervision Concept

This section presents the second foundational block of theory relevant to this thesis: deep learning techniques, with a particular focus on self-supervised pre-training. The section is organized into two parts. First, the general concepts of deep learning and the specific architectures used throughout this work are introduced. Second, the core ideas and main approaches related to self-supervised pre-training are discussed in detail, following the book chapter [7] written during the time of this thesis.

2.2.1 Deep Learning - Concepts and Architecture

The field of DL has become the dominant approach for many image analysis applications, following breakthroughs in the early 2010s [24]. These advances leveraged newly understood scaling laws related to model size, along with the computational benefits of modern hardware, to enable the training of significantly larger models. The key premise of a DL model – which sets it apart from early *machine learning* (ML) approaches – is the use of a large number of trainable parameters $\boldsymbol{\theta}$ (also referred to as the model size) to approximate a parameterized function f that maps input data $\mathbf{x}$ to corresponding annotations $\mathbf{y}$. These parameters are learned by minimizing a task-specific loss function $\mathcal{L}$ over a dataset $\mathcal{D} = (\mathbf{x}_i, \mathbf{y}_i)$:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \mathcal{L}\left(f(\mathbf{x}_i, \boldsymbol{\theta}), \mathbf{y}_i\right) \ , \tag{2.1}$$

resulting in a set of optimized parameters (weights) $\boldsymbol{\theta}^*$. The next two sections outline the different design choices for the parameterized function f, while the corresponding parameters $\boldsymbol{\theta}^*$ are derived through stochastic optimization via backpropagation in combination with gradient descent-based optimizers [25].

Fully Connected Neural Networks One of the most common architectural designs is the *fully connected neural network* (FCNN) – also known as a multilayer perceptron – which represents the most fundamental type of artificial neural network. An FCNN consists of multiple layers of neurons, where each neuron in a layer is fully connected to every neuron in the subsequent layer. Here, the input of the next layer is given by the weighted sum of all outputs of the previous layer[6] where the weights are learnable parameters, optimized during training. The forward pass of an FCNN is defined by a series of matrix multiplications followed by nonlinear activations:

$$\mathbf{h}_l = \sigma(\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l) \tag{2.2}$$

where $\mathbf{W}_l$ and $\mathbf{b}_l$ denote the weight matrix and bias vector of layer l, $\mathbf{h}_{l-1}$ the output of the previous layer ($\mathbf{h}_0$ would then represent the input to the model), and σ is a nonlinear activation function [26, 27, 28]. FCNNs are powerful function approximators and have been widely used in various applications, and also often serve as an essential building block of more advanced architectures such as *convolutional neural network* and *vision transformer* described in the following.

Convolutional Neural Networks *Convolutional neural network*s (CNNs), originally proposed in [29], is a class of deep learning architectures where the characteristic information extraction principle is based on deriving feature maps via convolution with so-called *kernels* (or *filters*). CNNs are designed for extracting information from grid-like data with spatial context, e.g. images, and hence play a crucial role in the processing of EO data.

As is generally the case in all DL-based approaches, data is propagated layer-wise sequentially. This is commonly achieved by blocks, consisting of a combination of convolutional layers, as well as pooling layers for dimensionality reduction [29], often followed by additional layers for numerical stabilization, such as batch normalization layers [30], as well as task-specific heads. Each layer contains kernels with learnable parameters, and after activation, the resulting outputs are referred to as feature maps of the corresponding layer. Fig. 2.15 illustrates this concept where, after every block of layers, the spatial extent of the data is reduced (pooling and convolution), whereas the number of resulting feature maps and the level of abstraction increases. An example of internal representations (feature maps), illustrating the increasing level of abstraction for different layers i, is shown in Fig. 2.16.

Vision Transformer While CNNs represented the *state of the art* (SOTA) architectural choice for most deep-learning image analysis task for many years, researchers in [1] proposed a strategy to use the transformer architecture – originally proposed for the processing of natural language sequences [33] – on image like data, commonly referred to as *vision transformer* (ViT). The underlying transformer architecture operates on a set

Note: Tokens are abstract high-dimensional represenations of a datum. This term gets interchangeably used with the term embeddings or features, which is more commonly used for the CNN case.

[6] The weighted sum usually additionally undergoes a nonlinear activation function as well as the addition of a learnable bias.

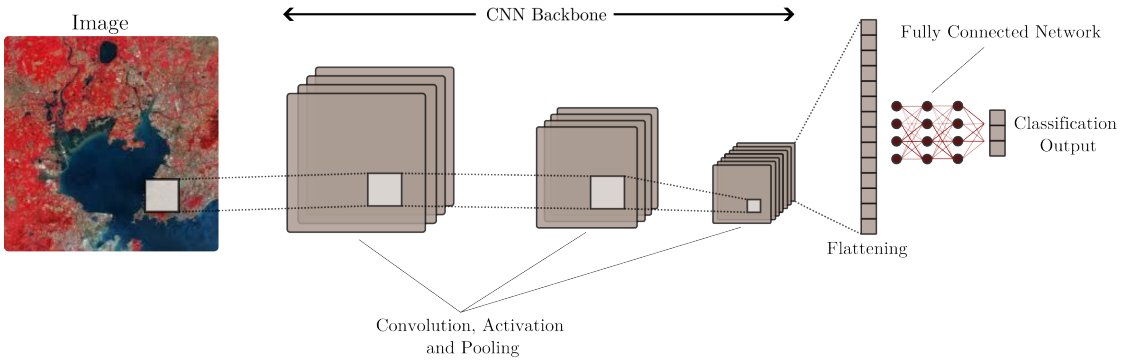


Figure 2.15: Illustration of the information flow for convolutional neural network type architectures. Starting from grid-like input data (left), hierarchical features are extracted by sequentially applying convolution and activation as well as pooling operations.

of tokens, which, in the case of natural language processing, are vector representations of words or sub-parts of words derived from a learned lookup table. In [1], the researchers proposed adapting this idea to vision tasks by dividing an image into fixed-size patches and generating one token[7] for each patch using an embedding network. This allowed images to be processed analogously to text. Notably, this straightforward adaptation – which remains a SOTA approach – achieved widespread impact, building on a broader line of research that explored integrating attention mechanisms into CNN-based architectures [34, 35, 36].

The core of the transformer architecture is given by the attention mechanism – which, due to its slightly more unintuitive nature – in contrast to the information extraction via convolution – will be described in more detail within this section. To follow along Fig. 2.17 gives a graphical illustration of the inner workings. Starting from the image patches that are projected into an embedding space of dimension d_{embed}, which is denoted as a token $\mathbf{t}_i \in \mathbb{R}^{d_{embed}}$ with $i \in [1, \ldots, N_t]$ indexing the tokens and N_t represents the overall number of tokens, positional embeddings are added to the patch tokens before they are fed into the transformer layers to maintain positional awareness. This ensures that the model can distinguish between different spatial locations where the information within each token originates from. This can either be done by adding a learnable positional encoding vector or modulating them with a signal of a given frequency. In the following, it is assumed that the token $\mathbf{t}_i$ has already undergone the positional encoding step. A single transformer block then combines two processing steps: Firstly, the attention mechanism that allows each token to be updated by (a modified version of) the content of all other tokens, and secondly, a feed-forward neural network step that allows for individual updates on all tokens.

The attention step works as follows. For each token $\mathbf{t}_i$, a so-called key $\mathbf{K}_i$ and query vector $\mathbf{Q}_i$ of lower dimension d_{head} is constructed by multiplying each token with two

[7] A token is an embedded latent representation of a piece of information within a transformer network, similarly to features-maps in the case of internal representation during the processing via a CNN architecture.

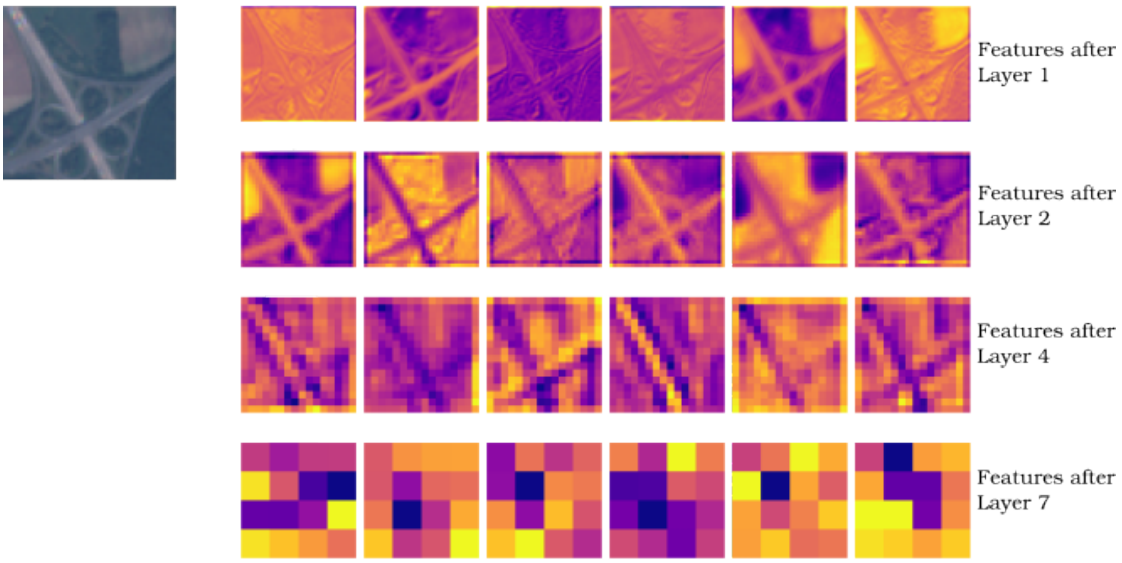


Figure 2.16: Feature maps extracted from different convolutional layers of a CNN architecture (*vgg11* from [31]), tasked to predict the patch wise land-cover [32] class for a satellite image. The leftmost image represents the original input (RGB), while the right grid displays feature maps at different layers. Each row corresponds to a specific convolutional layer, with six feature maps shown per layer. As depth increases, the extracted features become more abstract and less spatially detailed, capturing higher-level representations of the input image.

corresponding matrices with learnable parameters $\mathbf{W}_K$ and $\mathbf{W}_Q$ such as:

$$\mathbf{K}_i = \mathbf{t}_i \mathbf{W}_K \quad \mathbf{Q}_i = \mathbf{t}_i \mathbf{W}_Q, \ . \tag{2.3}$$

Subsequently, the dot product of all keys $\mathbf{K}_i$ and queries $\mathbf{Q}_j$ is calculated, which represents the attention pattern $\mathbf{A} \in \mathbb{R}^{N_t \times N_t}$. The value of $\mathbf{A}$ at position i, j can be interpreted as how much token i should *attend* to the token at position j. In order to normalize all the attention scores, the softmax operation is applied along the columns of the attention matrix $\mathbf{A}$. With the normalized attention scores $\mathbf{A}$, the update of the token goes as follows: For each token, the corresponding value vector $\mathbf{V}_i$ is introduced by again multiplying each token by a corresponding weight matrix $\mathbf{W}_V$:

$$\mathbf{V}_i = \mathbf{t}_i \mathbf{W}_V \ . \tag{2.4}$$

The normalized attention scores along a row of the matrix $\mathbf{A}$ are then used to calculate the update for a given token by summing the value matrices of all other tokens, weighted by the attention score.

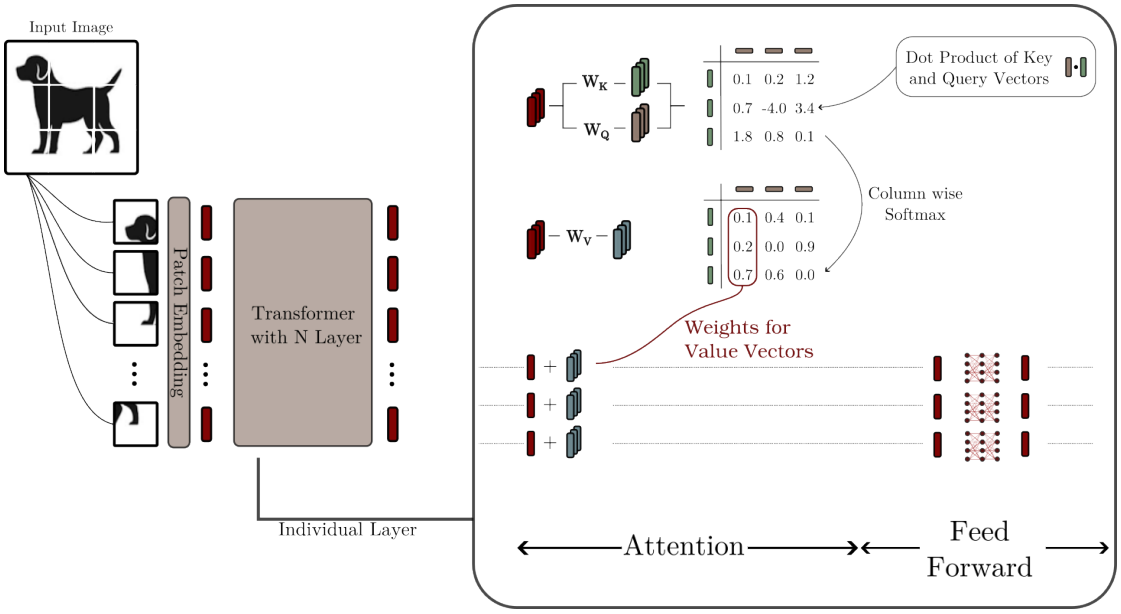


Figure 2.17: Illustration of the inner workings of an individual layer within a transformer architecture-based encoder network on image-related tasks. Here, the graphic sketches the procedure for a three token case. After the embedding of the patches of a given input image, the corresponding tokens undergo in each layer two main steps of applying the attention mechanism (information transfer between tokens) and subsequent further processing via FCNN acting on individual tokens. Within the attention step, from each token (red) corresponding key (green), query (brown) and value vectors (blue) get derived. For the update the value vectors get summed on each token weighted by the corresponding column of the attention pattern.

The above-described process is referred to as one head of attention. In practice, in every layer within the architecture, multiple heads are utilized in parallel, which is referred to as *multi-head attention*. Subsequent to the attention step, each token undergoes further processing by a densely connected feed-forward network. Notably, this step allows only for further refinement and increased abstraction of the embedding, but no information transfer from other tokens is possible, as these layers operate on individual tokens independently.

In contrast to the CNN approach – where a global receptive field is gradually built through stacked convolutional layers – the ViT architecture captures long-range dependencies inherently through the attention mechanism, which is especially effective at modeling global context.

While ViT offers powerful feature extraction, it also comes with challenges. The computational complexity of self-attention scales quadratically with the number of tokens [1], making ViTs more resource-intensive than CNNs, especially for larger images. As will be seen later, the computational complexity poses a significant problem, especially in

the context of multi-spectral imagery. While larger images are generally desirable in a remote sensing application in order to capture larger geospatial context, putting multiple spectral bands into individual tokens (which is desirable for cross-channel attention) leads to a large number of tokens and hence a high computational cost.

2.2.2 Self-Supervised Learning

The main principle of *self-supervised learning* (SSL) – which forms the methodological basis of this thesis – is to incorporate unlabeled data into a supervised machine learning training workflow, which would otherwise fully depend on annotated data label pairs as formulated in Eq. 2.1. To follow along, a graphical illustration of the complete SSL workflow can be seen in Fig. 2.18.

The first step of SSL is to construct a so-called *pretext task* (PTT), which is used to derive synthetic data-annotation pairs by some sort of manipulation on the original unlabeled data, which effectively constructs an annotated dataset with respect to the self-defined PTT. The following sections will dive into the particularities of the different approaches to designing a PTT. For now – to follow along – the reader might just imagine the simple case where the PTT for a set of images is to estimate the mean of the distribution of pixel values in the blue channel (usually the last channel in a RGB image) for a given input image $\mathbf{x}$, where the PTT label $\mathbf{y}^{PTT}$ can be derived for any image by simple calculations. Given a dataset $\mathcal{D} = \{\mathbf{x}, \mathbf{y}^{PTT}\}$ a neural network architecture composed of two functions f and g can be optimized to derive the set of weights $\boldsymbol{\theta}_f^*, \boldsymbol{\theta}_g^*$ such that:

$$(\boldsymbol{\theta}_f^*, \boldsymbol{\theta}_g^*) = \arg\min_{\boldsymbol{\theta}_f, \boldsymbol{\theta}_g} \mathcal{L}(g(f(\mathbf{x}, \boldsymbol{\theta}_f), \boldsymbol{\theta}_g), \mathbf{y}^{PTT}) \ . \tag{2.5}$$

Training a neural network architecture on the PTT therefore leads to a pre-trained architecture f with optimized weights $\boldsymbol{\theta}_f^*$. The adaptation to the downstream task potentially requires changes in g (in the artificial example above, g would output a scalar value), here denoted with a new architecture h, to match the particularities of the downstream task and hence requires some supervised fine-tuning on a labeled dataset $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$:

$$\boldsymbol{\theta}_h^* = \arg\min_{\boldsymbol{\theta}_h} \mathcal{L}(h(f(\mathbf{x}, \boldsymbol{\theta}_f), \boldsymbol{\theta}_h), \mathbf{y}) \ . \tag{2.6}$$

where $\mathbf{y}$ is the label for the downstream task and h with weights $\boldsymbol{\theta}_h$ the adapted architecture. Here, the weights $\boldsymbol{\theta}_f$ are referred to as frozen since they are not subject to the optimization routine. Generally, depending on the specific situation also the complete set of weights $(\boldsymbol{\theta}_g, \boldsymbol{\theta}_h)$ can be fine-tuned:

$$(\boldsymbol{\theta}_f^*, \boldsymbol{\theta}_h^*) = \arg\min_{\boldsymbol{\theta}_f, \boldsymbol{\theta}_h} \mathcal{L}(h(f(\mathbf{x}, \boldsymbol{\theta}_f), \boldsymbol{\theta}_h), \mathbf{y}) \ . \tag{2.7}$$

Since here, fewer tunable parameters have to be determined (at least in the case of frozen encoder f), it can be expected that this setup is more label-efficient and less prone to overfitting [37]. Still, predicting if the features derived by the pre-trained encoding network f are descriptive for the downstream application is far from trivial

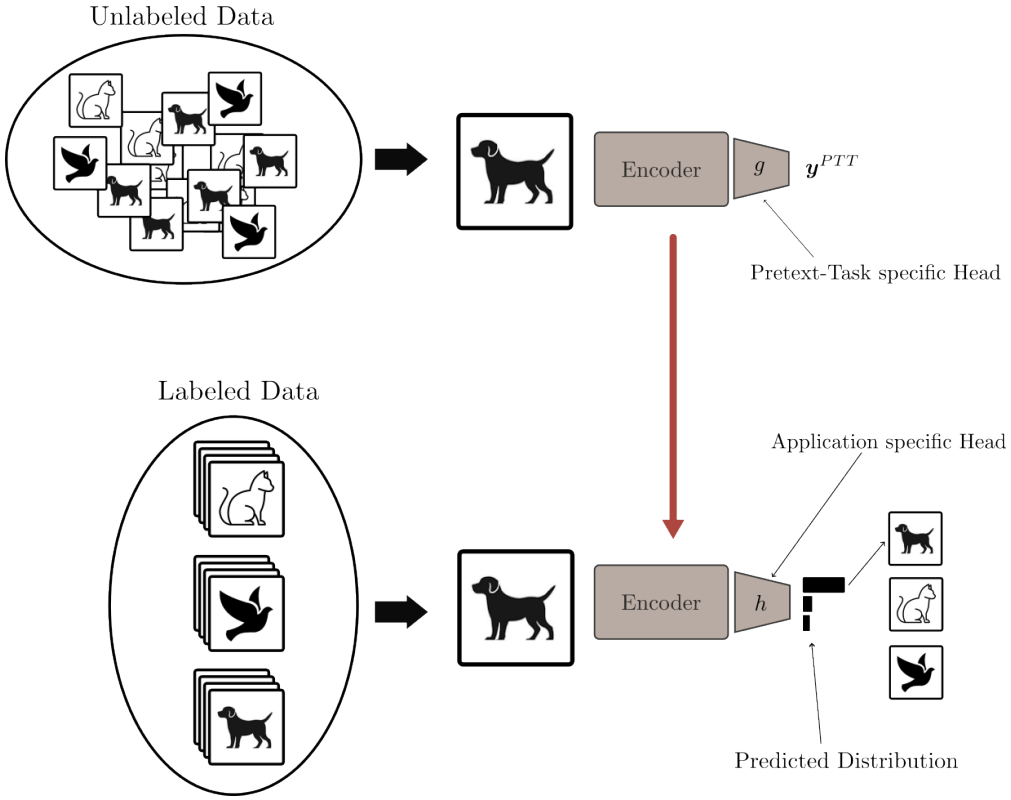


Figure 2.18: The overall self-supervised workflow. Given a set of unlabeled data, one trains a neural network architecture to solve a pretext task, with synthetically generated labels $\mathbf{y}^{PTT}$ which can be derived from the data itself. Parts of the pre-trained architecture are then used to serve as a feature extractor for a downstream task where supervised fine-tuning will be applied in order to solve the desired problem. In this example, the downstream task would be the classification of different labelled animal pictures.

and can generally only be validated empirically. A good example of this is the - at the time of writing - surprising success large language models have when pre-trained in a self-supervised fashion for the tasks of predicting the next word in a sentence. Non trivially, this leads to pre-trained architectures that demonstrate impressive capabilities of general text and context understanding, far beyond the complexity that one might expect to be learned from the pre-text task.

Even though there is no general statement that can be made about whether a pretext task will lead to predictive features, some things can be reasonably assumed. The most important thing to avoid is a so-called *mode-collapse* while pre-training. Here, the model predicts a non-descriptive set of features, which leads to poor performance on downstream tasks. One possible reason for this could be that the PTT is designed poorly, e.g. very trivial or too difficult to solve. Without experimental evidence, it could be expected that the example mentioned above (predicting the mean blue channel pixel value) would not lead to very well-described image features, and hence would not be a

good feature extractor for way more complicated tasks like scene segmentation. Given the thoughts above, it can be reasoned that it is by no means meaningful if the accuracy (or any other metric) during pre-training is particularly good. Rather, the opposite is true, where experimental evidence shows that a harder PTT leads to a better feature extractor [38] (within reasonable bounds).

In the following two sections, the two approaches of designing a PTT are outlined that are used within this thesis.

Contrastive Learning One of the most commonly utilized concepts for SSL pre-training is given by contrastive self-supervised learning. In this section, the contrastive learning concept is outlined based on the approach of the publication *SimCLR* [38]. It is worth mentioning that multiple alternative approaches, successor methods as well as subsequent publications, exist [39, 40, 41, 42, 43, 44] which still follow the general concept and only differ in their technical implementation.

The main idea of contrastive learning is to have a set of positive pairs of images, which is defined by two *views* of the same object or images with the same type of object. In the same way, negative pairs are constructed by having images that show different objects. Since large-scale datasets of the same object with different views are not available, in practice, the views are constructed by image augmentations like colorization, cropping, flipping, or converting into grayscale. Given the set of positive pairs, i.e. augmented versions of the input image $\mathbf{x}$ denoted by $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$, the objective is to learn those image features that are shared across the positive pairs as they are expected to be discriminative for the recognition of the desired object itself. For this, a feature extractor in the form of an arbitrary deep neural network architecture, e.g. a *ResNet-50* in [38], is used to obtain the features $\mathbf{h}_i = f(\tilde{\mathbf{x}}_i)$. A subsequent smaller network called projection head $\mathbf{z}_i = g(\mathbf{h}_i)$ is applied to map the features to the latent space, where the loss function gets applied. For all N images in a batch, augmented versions are being created, leading to N postive pairs and $2(N-1)$ negative counterparts. The loss function sums over all N images and forces the similarity for the obtained $\mathbf{z}_i$ for each postive pair while maximizing the distance in the latent space to all negative pairs $\mathbf{z}_j$ such that:

$$\mathcal{L}_k = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(\mathrm{sim}(z_i, z_k)/\tau)}{\sum_{j=1}^{2N} \mathbb{1}_{[j\neq i]} \exp(\mathrm{sim}(z_i, z_j)/\tau)}\right) , \tag{2.8}$$

is minimized. Here, sim denotes the *cosine similarity measure*, τ a temperature scaling factor, and $\mathbb{1}_{[j\neq i]}$ an operator that evaluates to zero if $i = j$ and one otherwise. Intuitively, the loss function encourages the features of positive pairs to be close in the latent space, while pushing those of negative pairs further apart. A graphical illustration of the process can be found Fig. 2.19. A feature embedding space that semantically encodes image content – for example, where one region of the latent space contains embeddings of images with cars, and another, distant region contains embeddings of images with animals – naturally constitutes a useful representation for image classification. In such a space, a subsequently trained linear classifier can effectively separate the classes based on the extracted features.

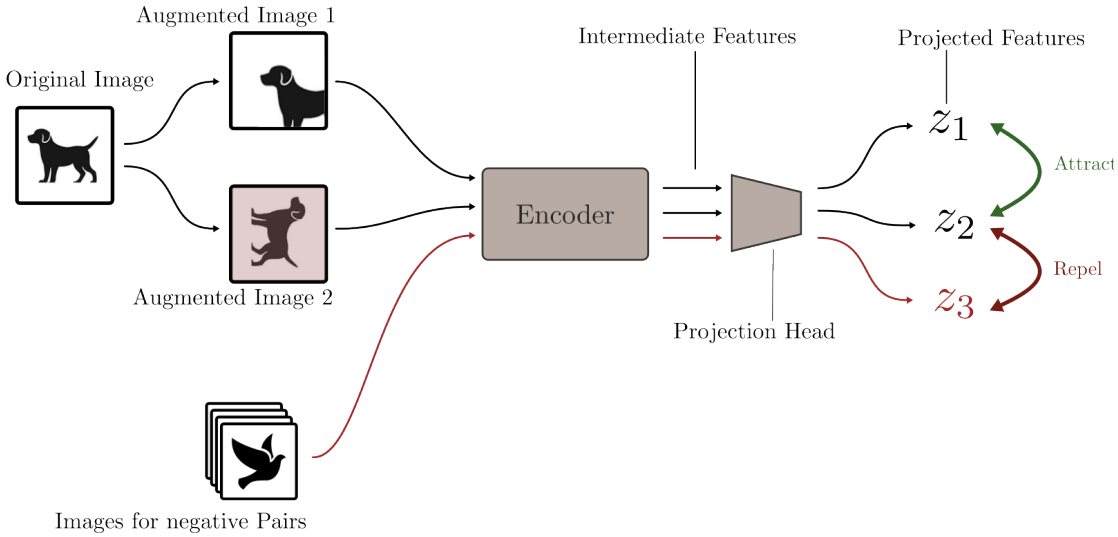


Figure 2.19: The contrastive learning paradigm. After encoding the image features, the contrastive loss gets applied, which forces the network to encode similar images (positive pairs) close in latent space while maximizing the distance to all negative encoded examples.

One of the most influential parameters for the contrastive learning setup is the set of augmentations $\mathcal{A}$ which is used to generate the two views $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ of an input image. Fig. 2.20 provides a visualization of commonly applied augmentations. In [38], the authors made an ablation study of applying the augmentations $\mathcal{A}$ separately (always two of the augmentations active) to find the influence of the design choice. The results show that the most influential augmentation is represented by the color augmentation. The authors further argued that if the color augmentation is not active, a color histogram of the two augmented views is already a descriptive feature of the contrastive PTT, which does not lead to good features for any subsequent downstream task. Fig. 2.21 displays the numerical results obtained from the study in [38] by applying sets of augmentations and plotting subsequent downstream task performance for the task of image classification accuracy.

Given the carefully engineered spectral properties of optical EO sensors described earlier, the choice of appropriate data augmentations poses a major challenge for adapting the pre-training task. As will been seen later this thesis proposed some effective techniques in order to approach this issue of choosing suitable augmentations with respect to contrastive pre-training on EO data.

Figure 2.20: Common type of augmentations used for the generation of positive pairs in the contrastive learning setup.

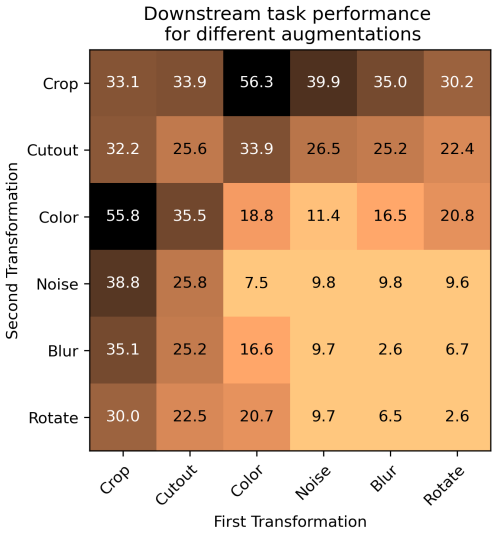


Figure 2.21: Top-1 accuracy on ImageNet by doing a linear evaluation on features obtained by the contrastive learning setup for different sets of augmentation. (Figure and numbers taken from [38].)

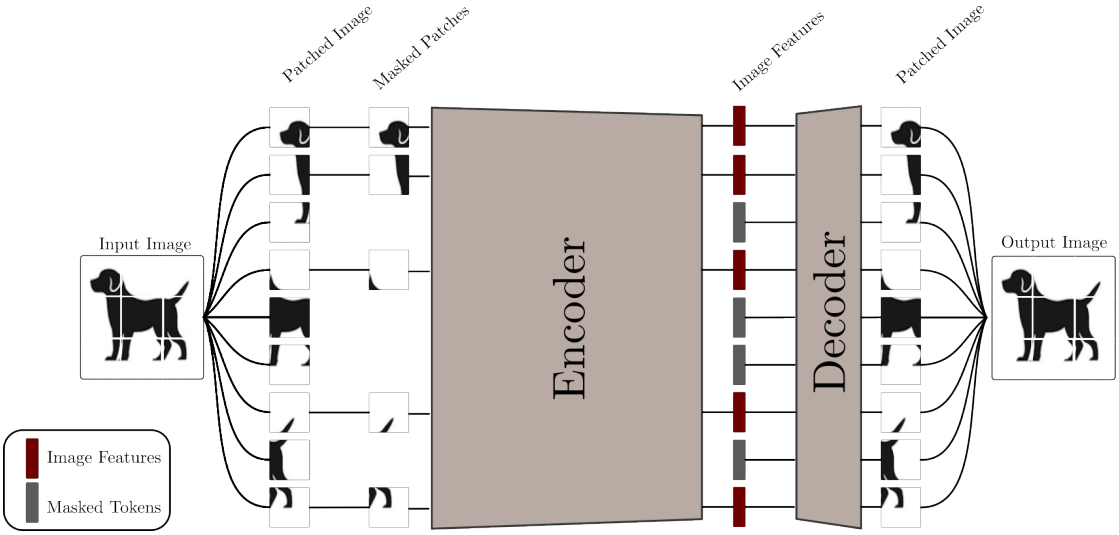


Figure 2.22: The concept of *Masked-Autoencoders*. An image divided into patches gets partially (after masking distinct image regions) processed by a transformer-based architecture. A second (smaller) decoding network is asked to restore the original input, given the encoded image patches as well as placeholder masked tokens for the missing regions.

Masked Image Modeling The second category of PTT utilized within the set of developed methods in this thesis is given by masked image modeling, outlined below.

The reconstruction of an image given a masked version represents a PTT where a detailed understanding of each object in the scene, as well as its orientation, is necessary to allow realistic inpainting of the missing part. This idea was taken up early [45], however, implemented with a CNN-based architecture, where the technical limitation is given by the masking step. In order to mask a random portion of a given input image processed by a CNN architecture, one has to set the corresponding pixel values for the masked area to some constant value c_m. Later, during the fine-tuning of the architecture, those values c_m are not present anymore, which presents a distribution shift of the input signal with respect to the data seen during pre-training.

With the rise of transformer-based architectures for image processing outlined above, the premise is changing. With the embedded patchified version of an input image, masking can be easily achieved by not presenting certain tokens (corresponding to certain image positions) to the transformer-based encoder model. Given the fact that all steps within the transformer are independent of the number of tokens, no shape mismatch is given with respect to subsequent fine-tuning with the full image.

This concept was first successfully implemented by [46], where the authors implement *masked auto-encoder* (MAE) where a significant portion of the input patches (75%) are masked before processing them via an encoder transformer network **E**. The resulting features get padded by learnable mask tokens before a decoder network **D** restores the full input image. Unlike prior approaches that rely on modest masking ratios, the authors

in [46] demonstrate that the high masking ratios can be utilized without degrading reconstruction quality. This aggressive masking encourages the model to learn rich and global scene representations, making it especially effective for downstream tasks. A further noteworthy design decision in the MAE framework lies in the seemingly simple imbalance between the sizes of the encoder and decoder networks. While large models can be used for the encoder **E**, its contribution to the overall computational cost remains relatively small, as it operates only on the unmasked portion of the input. In contrast, the decoder **D** can be kept relatively shallow, thereby encouraging the model to extract most of the descriptive features during the encoding stage. Fig. 2.22 gives a graphical illustration of the overall process.

3

Related Work

Based on the two main methodological approaches to self-supervised pre-training described at the end of the previous chapter, this chapter outlines the current state of research on single- and multi-modality *self-supervised learning* (SSL) methodologies with respect to EO data. The focus is on adaptations from classical computer vision to the EO domain, as well as on sensor-informed design choices in the corresponding frameworks. After discussing general trends in contrastive and masked image modeling approaches, the chapter highlights more recent developments centered on sensor-informed learning. Finally, the contributions of the works developed within this thesis are outlined.

3.1 Related Work in Self-Supervised Learning for Earth Observation

The methodologies presented in this thesis span multiple areas, including self-supervised learning for Earth observation (EO) data – covering both multi-spectral and SAR modalities – as well as more specialized approaches regarding the development of model pre-trained in a *sensor-informed* manner. At the time of writing the thesis, both fields, especially the latter, remain relatively new and are rapidly evolving. Fig. 3.1 exemplary displays the number of publications that overall deal with the topic of self-supervised pre-training for EO data, which highlights the dynamic and recent trend of increasing research interest.

Earth observation data differs in a few key aspects from classical object-centred imagery. This fact can be explicitly exploited for the development of domain-specific self-

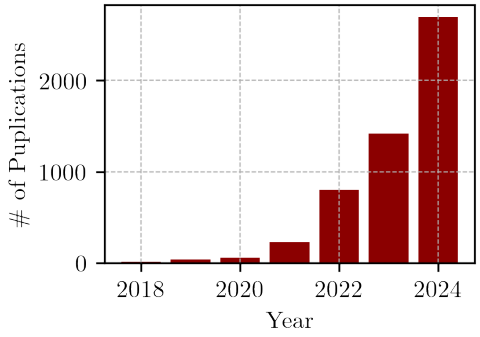


Figure 3.1: Number of publications according to `https://www.scopus.com/` matching the search query *self-supervised* AND *Earth observation* restricted to the publication type of *article* OR *conference paper* (hence ignoring *meta reviews* or other types of publications) as a function of publication year.

supervised learning algorithms. The methods presented within this thesis are focusing – as described before – on sensor-informed pre-training and making use of the unique data properties. Hence, the collection of prior works discussed in the following sections is selected to have either a thematic large overlap with the developed methods or serve as a direct predecessor (in which case it will be mentioned). Further, due to the fact that the major developments in this field happened during the time of developing the methods underlying this thesis, some of the publications came out after the specific method to which it is related. Nevertheless, even though those works are not direct predecessors, they will still be mentioned in order to present a full picture of the state of the field at the time of writing this thesis.

Before diving into the particularities of specific works of the three categories: *contrastive approaches*, *masked image modeling*, and *sensor-informed learning*, where the categories somewhat reflect the historical development of the field, a short meta-analysis of the field is presented.

Overall Developments - Meta Analysis Driven by the strong success of self-supervised pre-training in the field of classical computer vision, the research community around DL-related analysis of EO data quickly adapted to this trend. This was mainly due to the fact that unlabeled data is broadly available. Quite parallel to developments in the classical computer vision domain, early efforts were centered around contrastive learning approaches, followed by a focus on masked image modeling. The field was partially rebranded after the introduction of the term *foundation model*, mainly driven by advances in natural language processing. Nevertheless, the overall objective did not dramatically change, and clear definitions of the foundation model terminology are still pending. Starting from the development of SSL techniques for individual sensors, current trends now focus on the incorporation of multiple (or even arbitrary) sensors (to which the works in this thesis also contribute), as well as the incorporation of additional data or modalities such as natural language, where a detailed discussion of this is

omitted, as it falls beyond the scope of this thesis. While the field is further expanding quickly and approaches, as well as corresponding datasets and benchmark strategies are being released, it is still an open question to what extent and which aspects of available *foundation models* are getting adapted by end-users.

Contrastive Approaches Methods based on the contrastive learning paradigm have long been studied and applied to pretrain DL architectures for EO image analysis.

One of the first works employing contrastive self-supervised learning in the context of EO data is a work titled *Seasonal Contrast* [47], where the researchers used the intrinsic property of EO data that multiple observations at different time points can always be collected. With a set of augmented images, *seasonal augmentations* and *synthetic augmentations*, multiple embedding subspaces are defined where resulting augmentations are either invariant to all augmentations or distinctly the *seasonal augmentations* or *synthetic augmentations*, respectively. It should be mentioned that, within this framework, the resulting augmentations are built using the *MoCo* [42, 48] training paradigm, which is an alternative to the *SimCLR* approach described in the previous chapter, where a similar concept of positive (near) and negative (far) embeddings is utilized.

The authors of [49] build on the *seasonal contrast* approach by additionally introducing training examples with longer time spans between them. Samples with a larger time difference may exhibit significant changes due to LULC dynamics. In contrast to seasonal variations – where a network should ideally learn to be invariant – significant LULC changes should not be treated as such variations. Thus, the approach in [49] treats samples with a short time difference as positive pairs (encouraging invariance to seasonal changes), while for samples with a larger time difference, a check is performed to determine if the feature space has changed significantly. This serves as a change prior without relying on explicit change labels.

Both of the above-mentioned contributions illustrate how domain-specific data properties can be leveraged to adapt the contrastive learning setup. It should be noted that several other approaches have addressed similar problems – often tailored to specific applications – such as [50, 51, 52, 53, 54, 55]. However, these are not directly relevant to the current discussion and are only mentioned here for completeness. Importantly, the mentioned methods operate on a single modality. Utilizing the diversity of EO data presents an additional opportunity to further adapt the contrastive learning setup for EO applications, as will be discussed in the following.

One of the first works utilizing multi-modal views in the contrastive learning setup was presented in [56], with a subsequent extension to pixel-wise downstream tasks in [57]. In this approach, samples from two different modalities – multi-spectral and SAR – are collected over the same location. Fed through two separate encoders, the resulting

embeddings can be used in a contrastive setup without the need for further strong[1] augmentations. The core idea behind this approach is to reduce the unwanted effects of invariances introduced through augmentations, an issue already studied in the context of classical computer vision [58]. In [56], all evaluations are conducted on patch-wise classification tasks, whereas the extension in [57] firstly extends the framework to ViT-based encoders as well as pixel-wise segmentation tasks during evaluation. As will be seen later, these two works build the foundation for one of the methods developed within this thesis. The approach of using the multi-modality nature of EO data has been widely adopted, and various datasets [59, 60, 61, 62] and methods [63, 64, 65, 66] have emerged.

Masked Image Modeling While contrastive methods can also be utilized to pre-train ViT-based encoders (as done in [57]), the use of the earlier described masked auto-encoder approach is more frequently used.

One of the earliest works regarding the application of MAE to the field of EO was presented by [67] named *SatMAE*, where the classical MAE was extended to work with time-series data through the use of a temporal masking strategy. This approach, similarly to contrastive seasonal methods, leverages one of the distinct characteristics of EO data. This work is not only among the first applications of MAE to EO data, but it also introduces an important concept: the design of the masking strategy, which is crucial when adapting MAE-type methods to the EO domain. Specifically, the authors propose two types of masking functions: consistent masking and individual masking. The latter applies individual masks to each image in the time series and, according to the experimental section, represents the inferior approach.

Another relevant prior work is represented by *ScaleMAE* [68], where the authors introduced the incorporation of image scale into the model, enabling it to reliably work with images of different scales and resolutions. This marks an important prior work for the later contributions introduced in this thesis. Additionally, alongside scale encoding, the authors proposed a multi-scale reconstruction procedure, where low- and high-frequency details of the input image are predicted separately and measured with different loss functions.

Many successor publications build on the concepts described above. For example, in *S2MAE* [69], the authors introduce separate tokens per channel, effectively allowing attention-based updates in the backbone over various tokens from the same spatial position. To still differentiate between tokens from different positions, learnable spectral embeddings are introduced that encode the band index of the corresponding token.

Generally, many works center around the question of how to incorporate the multi- or even hyper-spectral nature of EO data into ViT backbones, e.g. [70, 71], or focus more on scaling on the compute side [72, 73, 74, 75, 76, 77].

[1] The authors still employ random cropping of the data but eliminate stronger augmentations such as intensity modulating augmentations or Gaussian blurring.

The next section introduces the additional advances of incorporating sensor knowledge into the learning process.

Sensor-Informed Learning The term *sensor-informed* learning is – to this day – not strictly defined (commonly also described as *sensor-aware* or *any-sensor* model). Nevertheless, this section tries to give an overview of selected self-supervised learning methods that use the domain knowledge beyond the more straightforward meta information, such as location or time.

Note: All of the methods described in this section are either published simultaneously with the methods developed within this thesis or subsequently. Nevertheless, they show the clear trend of research direction that was also partially established via the methods presented within this dissertation.

The researchers presenting *DOFA* [78] introduce the *Wavelength-conditioned dynamic patch embedding* strategy. Here, the framework is mainly based on the introduction of a hypernetwork that generates network weights based on the central wavelength of a given channel. Given input images of shape $\mathbf{X} \in \mathbb{R}^{C \times W \times H}$, the center wavelength values $\lambda_c \in \mathbb{R}$ are embedded (via sinusoidal embedding) into a higher-dimensional space d_λ. After processing with two fully connected layers, including a residual connection, they are further processed via a transformer encoder. The transformer encoder outputs weights $\mathbf{W}_\lambda$ and bias values $\mathbf{B}_\lambda$ for the patch embedding layer of the main encoder. In this way, the patch embedding layer adapts to the number of input bands for different inputs. The architecture is pre-trained on optical as well as SAR data with different numbers of channels or observed polarizations.

Another recent work addresses the problem of differently sized input information, commonly present in EO tasks. Not only are input resolution and image size often different, but in some cases, multiple modalities are available, while flexibility is still required due to missing data (e.g., because of cloud cover). In the *AnySat* framework [79], a *scale-adaptive patch encoder* is introduced, which converts a patched input into sub-patches that are subsequently merged into an overall image representation and encoded with the corresponding GSD value. In the case of multiple inputs from different modalities or timeseries data, a subsequent encoder merges the modality or temporal dimension of the input data. *AnySat* is pre-trained with two self-supervised strategies. Consistency across modalities is enforced with a contrastive loss, similarly to [56, 57]. In addition to the contrastive loss, a *JEPA*-based pre-training [80] is conducted.

In *Panopticon*, another *sensor-informed* pre-training approach is proposed, based on the *DINOv2* pre-training strategy [44]. Given a set of images, *local* and *global* crops are defined by spatially and spectrally sub-sampling the data. Then, in line with the *DINO* concept [43, 44], a student network – given the *local* views – learns to mimic the teacher network on the *global* views. The authors propose solving the variable input size problem by applying *cross-attention* over input channels, inspired by [81], leading to one token per patch regardless of the number of input channels. Further, the central wavelength, as well as orbit and polarization information for SAR images, is encoded onto the tokens similarly to *DOFA* [78]. Beyond this, the authors propose slowly increasing the diversity of sensors via a *Progressive* pre-training approach, where the model is initially trained on a smaller subset of all available sensors in a warm-up-like fashion.

3.2 Contribution of this Thesis

The contributions proposed in this thesis are threefold and all centered around the research question – raised in Chapter 1 – on sensor-informed SSL for EO data in times of growing availability and access of spaceborn observations. Three subtopics are addressed, outlined below:

SenPaMAE Firstly, given a set of multi-spectral satellites – where the premise at the time of creating this work was to train individual networks on each multi-spectral sensor individually – a central research question emerges: how can a unified model be trained across data originating from multiple, heterogeneous multi-spectral sensors? This question could potentially be solved from an engineering perspective by selecting a fixed set of sensors and informing the model – via the injection of some form of sensor ID – which sensor the input image originates from. In the *SenPaMAE* framework, developed during this PhD, a more general approach was chosen. Since the specific differences of each optical system are known, the originating sensor itself is not the meta-information of interest. Rather, it is the specific sensor parameters that describe the measurement. For this – as will be shown in the next chapter – the previously introduced main sensor parameters, namely the spectral response function and the ground sampling distance, are injected into the learning process. In contrast to a simpler approach via the injection of sensor IDs, incorporating the specific sensor parameters describes exactly all similarities and differences between multiple sensors without relying on matching observations in order to learn those. To enable the model to understand these sensor parameters, the backbones are pre-trained in a self-supervised fashion, allowing the network to learn meaningful representations from the provided sensor parameters without relying on labeled data.

SARFormer Secondly, especially for higher-resolution data that is captured in a scheduled operational mode (compare Chapter 2), the specific geometric acquisition parameters – in addition to the sensor parameters – are of interest, as they highly influence the outcome of the measurement. While this concept generally applies to both optical and SAR modalities, the latter is particularly affected due to its distinct imaging modes (see Section 2.1.2), which significantly alter the measurement. In the second contribution, the *SARFormer* framework is introduced, where these specific acquisition parameters – the exact geometric position of the sensor as well as the imaging mode – are injected into the learning process of the model in order to enhance the descriptiveness of obtained features. This – like in all contributions – is paired with self-supervised pre-training that focuses on the particularities of the pretext task in the context of high-resolution EO data.

IaI-SimCLR The third methodological contribution builds directly on the strategies described above [56, 57] to obtain representations in a self-supervised manner, utilizing

the multi-modality aspect of available Earth observation data. Here, instead of relying solely on the information shared in two views of different modalities, additional restrictions are built into the training process to reduce the loss of information not present in both views simultaneously. This is achieved via the extension of the contrastive loss function and is accompanied by further improvements with respect to previous contrastive approaches, such as enhanced batch-sampling methods, which are highly relevant for controlling the level of similarity of samples within a batch and significantly influence the learned representations. Additionally, the dependency on the set of augmentations used for constructing the contrastive learning process is studied in detail since – as described in Section 2.2.2 – those represent one of the most crucial hyperparameters for such setups.

4

Self-Supervised Sensor-Informed Learning Methods

This chapter outlines the three sensor-informed learning methods that represent the core of this thesis. First, the sensor-informed models and the corresponding pre-training strategies for both multi-sensor multi-spectral data (named *SenPaMAE*) as well as high-resolution SAR data (named *SARFormer*) are introduced. Subsequently, a sensor-informed multi-modal alignment pre-training strategy named *IaI-SimCLR* is presented. All contributions are separately published and hence, a description can also be found in the three main publications [82, 83, 84] which form the foundation of this thesis.

4.1 SenPaMAE

The objective of the *SenPaMAE* architecture is to generalize the prediction step that usually processes a multi-spectral image tensor $\mathbf{x} \in \mathbb{R}^{C \times W \times H}$ to the more general setup of processing $\mathbf{x}$ together with the corresponding sensor parameters for a complete description of the observed signal. In this context, C denotes the number of channels and W and H width and height of the image, respectively. Within this thesis, the sensor specifications are described by the set of corresponding spectral response functions $\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_C\}$ and the GSD values $\{\sigma_1, \ldots, \sigma_C\}$ for each channel. In the following, the corresponding set of sensor parameters for a given multi-spectral signal $\mathbf{x}$ is denoted as $\{\boldsymbol{\lambda}_c, \sigma_c\}$. By incorporating this generalization step, the model becomes capable of handling data from multiple sensors with differing specifications. This is facilitated by the *sensor parameter encoding* (SPE) module detailed below. As a result, the network can be pre-trained on heterogeneous sensor data, and its ability to generalize across

sensor types is later assessed through various downstream task experiments. The here introduced modifications are twofold: Firstly, a sensor parameter encoding module is introduced in order to enable the model to take the specific sensor parameters of the corresponding multi-spectral signal into account. Secondly, an information-preserving data augmentation strategy is formulated, enabling data augmentation during pre-training within the framework of sensor-informed learning. The *SenPaMAE* framework is – like all three methods introduced in this thesis – built with self-supervised pre-training in mind.

Sensor Parameter Encoding Generally, the pre-training is based upon the MAE training paradigm [46] (compare Section 2.2.2) and can be conducted in a self-supervised fashion over a dataset containing images from multiple sensors without needing to have matching observations (i.e. with the same acquisition date and location), as those are hard to collect due to the different revisit time schedules of the satellite missions.

Injecting multiple non-equal-shaped types of information (i.e. image and corresponding sensor parameter) into a neural network is a non-trivial task[1] . The *SenPaMAE* approach builds upon the ViT architecture [1] and the associated positional encoding strategies to inject the 3D multi-spectral data tensor $\mathbf{x}$ of size $H \times W$ with C channels and the set of corresponding response functions $\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_C\}$ and GSD values $\{\sigma_1, \ldots, \sigma_C\}$ into the model. To guide the reader, Fig. 4.1 gives a graphical overview of the two proposed architectural modifications to follow along with the technical description.

Note: Since the computational cost for training and inference of a transformer architecture scales quadratically with the number of tokens, this approach is – even though being elegant – computationally expensive. More details will be outlined in the discussion section.

In this framework, when the *sensor parameter encoding* (SPE) module is active, the model input takes the form $(\mathbf{x}, \{\boldsymbol{\lambda}_c, \sigma_c\})$. This approach is applied during the model pre-training and for subsequent fine-tuning on downstream tasks, as will be seen later. Unlike the conventional ViT approach, where an image is segmented into patches of size P^2, each patch having a position (p_x, p_y), and all channel information for a patch is encoded into a single token with size d_{emb}, the *SenPaMAE* framework generates a token for each patch and channel separately. This results in $N_t = C \times HW/P^2$ tokens $\mathbf{t}_i^c$ of dimension d_{emb}, where c encodes the channel of origin and i the position (p_x, p_y) within the image. This design decision is motivated by two facts: In contrast to natural images, for multi-spectral signals much of the desired information content lies in the relative reflectance statistics of individual bands. Therefore, providing separate tokens per channel enhances the flexibility of applying the attention mechanism across channels. Furthermore, this presents the opportunity to mask a significant portion of the tokens for the MAE masking step without creating extensive regions in the image devoid of geometric information. This is especially crucial since optical satellite images typically capture objects at a relatively small scale (a few pixels). In contrast to that – in natural imagery – objects often occupy the entire image, making reconstruction of

[1] Conventional deep learning architectures are typically designed for a single input format, for instance, convolutional layers expect image-like inputs. Incorporating an additional data type (such as the geographic location of an image) often requires reshaping the secondary input, e.g. by adding an auxiliary image channel containing the location information. Such approaches are often suboptimal, as the statistical distributions of the data sources may differ, and modality-specific processing weights can yield better results. Transformer architectures offer greater flexibility in this respect.

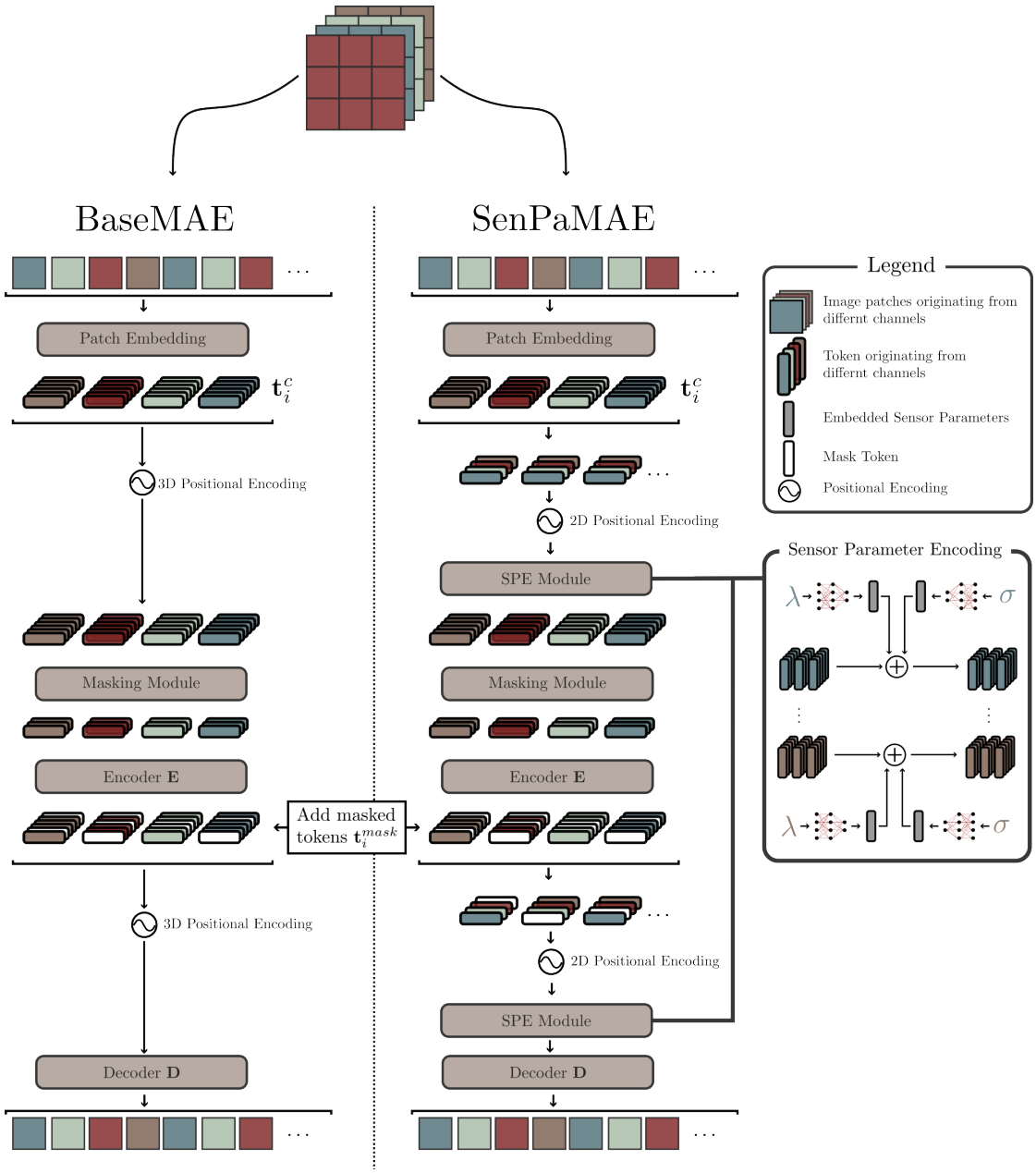


Figure 4.1: The proposed *SenPaMAE* architecture as well as the baseline model *BaseMAE*. After patch embedding, the tokens $\mathbf{t}_i^c$ undergo a three-step encoding procedure where information about patch position, spectral response function, as well as ground sampling distance, gets added. After the encoding, all tokens are processed in an MAE-like encoder-decoder setup. An analogue encoding procedure can be applied before the decoding step. The *BaseMAE* setup differs from *SenPaMAE* by taking the channel index for the positional encoding into account and neglecting the injection of the sensor parameters.

heavily masked images with extensive contiguous masked areas a viable task.

Given all tokens $\mathbf{t}_i^c$ with $i \in (1, \ldots, HW/P^2)$ and $i \in (1, \ldots, C)$ first the learnable positional encoding vector $\boldsymbol{\omega}_j^{pos} \in \mathbb{R}^{d_{emb}}$ gets applied to each token originating from one position in the multi-spectral tensor, hence j is encoding one specific position (p_x, p_y). This only encodes the position in the image since tokens originating from different channels still get added with the same positional encoding. Subsequently, to the positional encoding, the SPE module gets applied. Here, the C corresponding spectral response function $\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_C\}$ and GSD values $\{\sigma_1, \ldots, \sigma_C\}$ get used and encoded into the embedding dimension by applying two separate five-layer *multi-layer perception* (MLP) networks denoted as f_{SRF} and f_{GSD}:

$$f_{SRF} : \mathbb{R}^{2300} \to \mathbb{R}^{d_{emb}},\ \boldsymbol{\omega}_c^{SRF} = f_{SRF}(\boldsymbol{\lambda}_c) \tag{4.1}$$

$$f_{GSD} : \mathbb{R} \to \mathbb{R}^{d_{emb}},\ \boldsymbol{\omega}_c^{GSD} = f_{GSD}(\sigma_c) \tag{4.2}$$

Note: The dimension of the spectral response function of $\mathbb{R}^{2300}$ originates from the resampling step after obtaining the information from the satellite providers and corresponds to the window $[400\ nm, 2700\ nm]$ with $1\ nm$ stepwidth.

which leads to the spectral and resolution encoding vectors $\boldsymbol{\omega}_c^{SRF} \in \mathbb{R}^{D_{emb}}$ and $\boldsymbol{\omega}_c^{GSD} \in \mathbb{R}^{D_{emb}}$. Subsequently all tokens $\mathbf{t}_i^c$ originating from the corresponding channel c get summed together with the corresponding sensor parameter embeddings $\boldsymbol{\omega}_c^{SRF}$ and $\boldsymbol{\omega}_c^{GSD}$.

Like in the original MAE framework, with the set of tokens with positional, spectral and resolution encoding applied, the image features $\mathbf{f}_i^c$ get created by applying an encoder network $\mathbf{f}_i^c = \mathbf{E}(\mathbf{t}_i^c)$ to a random subset of tokens determined by the masking ratio r_{mask}. Analogue to the MAE strategy, the set of encoded tokens gets filled up by learnable mask-tokens $\mathbf{t}_i^{mask}$ to restore the original sequence length. Two different modes of encoding strategies for the decoder step are possible and will be evaluated. In both scenarios, analogue to the encoding step the features $\mathbf{f}_i^c$ get summed up with a positional encoding vector $\boldsymbol{\omega}^{pos} \in \mathbb{R}^{HW/P^2 \times D_{emb}}$ depending on the position of origin in the image (p_x, p_y). Subsequently, the analysis covers two cases: one in which the SPE module is applied again and another in which it is omitted. This follows from the intuitive idea that not encoding the sensor parameters in the decoding step forces a model to implicitly encode the information into the features $\mathbf{f}_i^c$. In contrast to that, applying a second SPE module after the decoding stage would provide the decoder $\mathbf{D}$ with the necessary information for the reconstruction task, and therefore the information is not necessarily encoded in the features $\mathbf{f}_i^c$. After the second encoding step, the features are processed by the decoder network $\mathbf{D}$, and the decoded features get transformed into the original patch dimension by a trainable linear layer.

To ensure a fair evaluation of the above-mentioned contribution, the standard MAE-like baseline – with some small modifications – will serve as a benchmark and will be denoted as *BaseMAE*. Unlike in the vanilla version, analogous to *SenPaMAE*, there will still be one token per patch position and channel. Here, the parameter encoding module is neither applied before the decoding nor the encoding network. This makes an adaptation of the positional encoding necessary since otherwise, the position of the tokens is not uniquely determined. Therefore the positional encoding is extended to a three-dimensional version $\boldsymbol{\omega}_j^{pos} \in \mathbb{R}^{d_{emb}}$ and j is encoding one specific position and the corresponding channel (C, p_x, p_y).

Spectral Superposition Augmentation Diversity and amount of data during the training of neural network architectures are known to enhance the performance significantly [85]. Still – as described above – producing models that are invariant to image manipulations like changing brightness or contrast is not an option when pursuing the objective of generating sensor parameter-aware architectures.

To tackle the problem, a data augmentation strategy, which will further be called *spectral superposition augmentation* (SSA), is proposed within the *SenPaMAE* framework. Here, synthetic multi-spectral signals $\tilde{\mathbf{x}}$ are created by the superposition of two randomly selected channels within the set of all channels from one sensor:

$$\tilde{\mathbf{x}} = \alpha_1 \mathbf{x}_m + \alpha_2 \mathbf{x}_n \tag{4.3}$$

where the corresponding spectral response function for $\tilde{\mathbf{x}}$ adds up to:

$$\tilde{\boldsymbol{\lambda}} = \alpha_1 \boldsymbol{\lambda}_m + \alpha_2 \boldsymbol{\lambda}_n \ . \tag{4.4}$$

Here, α_1 and α_2 can, in each augmentation step, be randomly chosen from the interval $0 < \alpha_1, \alpha_2 < 1$. This augmentation strategy increases the diversity of examples during training, can easily be extended to combine more than two channels, and requires a detailed understanding of $\boldsymbol{\lambda}$ to reconstruct images. This augmentation step gets applied to a random number of channels in the input image determined by p_{mix}. It is worth mentioning that SSA is purely proposed as a means to augment samples during the pre-training stage and will not be utilized in any form during the fine-tuning stage of the model.

To further enhance the diversity of data seen during training, an additional cubic down-sampling operation on $\mathbf{x}_c$ and $\tilde{\mathbf{x}}_c$ gets applied. Starting from the original GSD of the sensor σ_{org} a randomly selected target GSD from the set $\{\sigma | \sigma > \sigma_{org}, \sigma \in (5, 10, 15, 20, 30)\}$ is chosen for a random subset of images (down-sample probability p_{down} per channel) in the batch. Prior to the down-sampling operation, Gaussian blurring is applied in order to realistically mimic a sensor with lower GSD values [86].

Note: The values for the target GSD come from the available imagery in $\mathcal{D}^{PT}_{MSMS}$ dataset, which will serve for the empirical evaluation of the method in Section 6.1.

Methodological Summary The *SenPaMAE* framework allows for a more general type of input information with respect to a multi-spectral signal into a neural network. This is done by not merging information from different channels into one token as done for natural images, rather creating separate tokens, while accepting higher computational costs. From here, a large variety of meta-information could be encoded into the tokens, where within this thesis, experimental validation will be restricted to the main parameters of a multi-spectral signal, namely the spectral response function and GSD, which relates to the spatial resolution. Furthermore, an information-persevering data augmentation strategy is introduced for the self-supervised pre-training.

4.2 SARFormer

The objective of the proposed *SARFormer* architecture is to process one or more SAR acquisitions (interchangeably used with the word views in the following) of the same geographical area on the Earth's surface, while maintaining the physical integrity of the signals by taking into account the corresponding geometrical acquisition conditions as described in Section 2.1.2. The intuition behind this approach lies in the strong dependency of visual features in a SAR image on the sensor's geometric orientation relative to the target, e.g. the layover extent of elevated structures and its dependency on the looking angle. As all methodological contributions of this thesis, the *SARFormer* framework is developed with a self-supervised pre-training approach in mind in order to tackle the problem of label scarcities, present for many EO tasks. Contributions within the *SARFormer* framework are two-fold. Firstly, a module is proposed that allows the model the incorporation of the corresponding acquisition parameters into the training and inference process. Secondly, three different masking configurations – that serve as the pre-text task in the MAE setup – are proposed.

Acquisition Parameter Encoding (APE) The input data is given by a set of SAR acquisitions denoted by $\mathbf{x}_v \in \mathbb{R}^{1\times W\times H}$ where $v \in 1,\dots,N_v$ indicates the view index and W and H the width and height of the image, respectively. Without a loss of generalization, N_v will be restricted to $N_v \in (1,2,3)$ within this thesis. Each view $\mathbf{x}_v$ has a separate set of acquisition parameters (compare Section 2.1.2) namely the looking angle θ_v, the azimuth angle Az_v as well as an acquisition mode m_v which will be denoted as the set of acquisition parameters $\Phi_v = (\theta_v, Az_v, m_v)$.

This framework – similar to *SenPaMAE*– is built on the classical vision transformer [1] which gets used as an encoder network $\mathbf{E}$ to process the set of $\{\mathbf{x}_v\}$, together with the acquisition parameters $\{\Phi_v\}$. Each input view gets divided into patches of size P^2, which, after linear embedding into the latent dimension d_{emb}, leads to $\mathbf{t}_i^v$ tokens, where $i \in 1,\dots,HW/P^2$ indicates the position within the image and $v \in 1,2,3$ the corresponding originating view, leading to a overall number of tokens $N_t = N_v \times HW/P^2$. As in [1], the position of each token in the image gets encoded through the addition of learnable positional encoding vectors. For the following description of the architectural modifications Fig. 4.2 gives a graphical overview.

To incorporate the acquisition parameters Φ_v into $\mathbf{E}$ a linear layer f_{aqu} get utilized which transfers preprocessed acquisition parameters Φ_v^{pre} into the latent dimension d_{emb} of the model:

$$f_{aqu} : \mathbb{R}^4 \rightarrow \mathbb{R}^d,\ \Phi_v^{embed} = f_{aqu}(\Phi_v^{pre}) \tag{4.5}$$

where the processing of the acquisition parameters is given by:

$$\Phi_v^{pre} = (\cos(Az), \sin(Az), 1/\tan(\theta), m). \tag{4.6}$$

Here, the sine and cosine component of Az is chosen to describe the azimuth angle $Az \in [0°, 360°]$ in order to account for the circular nature of the quantity (e.g. $0°$

and 360° is effectively identical). Further, the preprocessing of the looking angle θ is chosen to match roughly the ratio between the height of a structure and its layover extent (compare [84]). For N_v input views, N_v learnable metatokens $\mathbf{t}^v_{\text{meta}} \in \mathbb{R}^{d_{emb}}$ get generated and for each view v, the corresponding embedded view parameter get added to the corresponding metatoken. The input to the transformer backbone for a two-view input is hence given by $2 \times HW/P^2$ image tokens and 2 metatokens. From here, the information from the metatokens can be transferred into the image features via the attention mechanism within the ViT encoder $\mathbf{E}$. This step is in the following referred to as *acquisition parameter encoding* (APE).

Note: Since the metatoken are built up from a learnable component as well as the embedded acquisition parameters, no positional encoding is necessary since a unique identification of each metatoken to the corresponding view is possible.

Domain Adapted Masking Similar to the *SenPaMAE* setup, the *SARFormer* architecture gets pre-trained in a self-supervised manner using the MAE framework [46]. Here, the metatokens do not fall under the masking operation and get equally processed by the decoder network.

Similar to *SenPaMAE*, large unmasked areas which lead to an ambiguous pre-text task are intuitively to be avoided. Within the experimental scope of the development of the *SARFormer* framework, three different masking strategies are tested, all for the case of two-view model ($N_v = 2$) inputs, without loss of generalizability to higher view numbers. In the single-view case ($N_v = 1$), two of the three strategies are not applicable, and only random masking will be evaluated, and the APE module is equally applied to all pre-training runs.

For pre-training with a single SAR acquisition, random masking with no further modification is applied. The same procedure for a two-view case is in the following referred to as the *random* masking strategy. Furthermore, a *preserving* masking strategy is introduced, where for each patch position, one of the two tokens $\mathbf{t}^1_i$ or $\mathbf{t}^2_i$ is left unmasked. This approach reduces the occurrence of large unmasked regions covering entire objects, which cannot be effectively reconstructed, while requiring an understanding of the relationship between different view geometries Φ_v to fill in the missing information. An extreme variant of this is also explored, in which all tokens from either view one or view two are masked, which is referred to as *blind-channel* masking. Fig. 4.3 gives a graphical overview of the three above-introduced masking strategies.

Methodological Summary The *SARFormer* framework allows for a more general type of input information with respect to a SAR signal into a neural network. This is done by the addition of metatokens (one per input view), where each metatoken carries the information about the geometrical scene acquisition parameters that can update the image tokens via the attention mechanism of the encoder (and during pre-training the decoder) network. This stands in contrast to the approach of the *SenPaMAE* framework, where the sensor parameters get directly encoded onto the image tokens and represent a more flexible approach (compare Chapter 7). Further, within the *SARFormer* framework, different masking strategies are applied in order to tackle the problem of large unmasked areas, which are – for remote sensing imagery – intuitively unwanted.

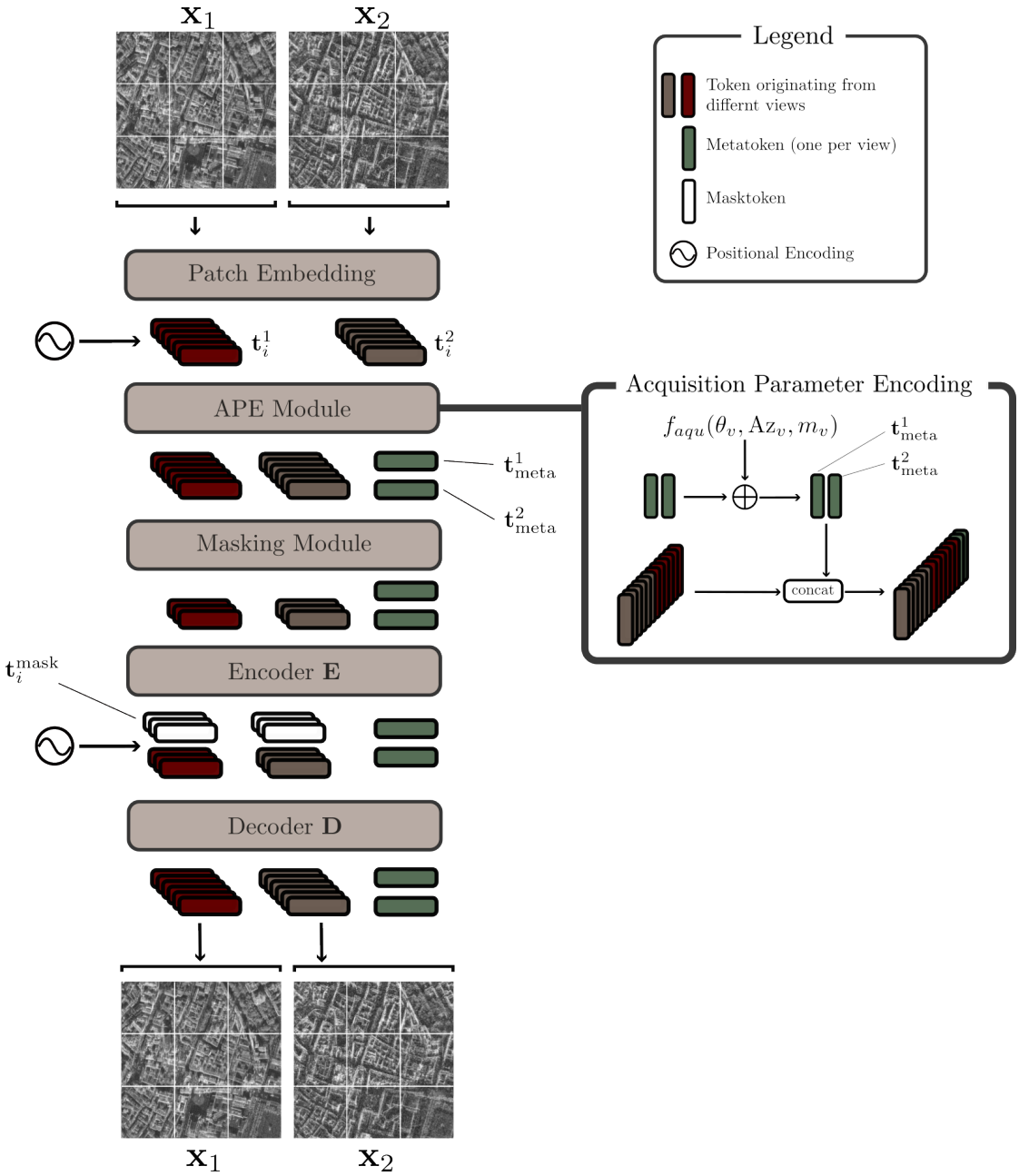


Figure 4.2: The proposed architecture modifications and the corresponding nomenclature for the pre-training scenario of the *SARFormer* framework. The scenario drawn here corresponds to a two-view case but can be generalized to one or multiple views as done in the experimental section of this work.

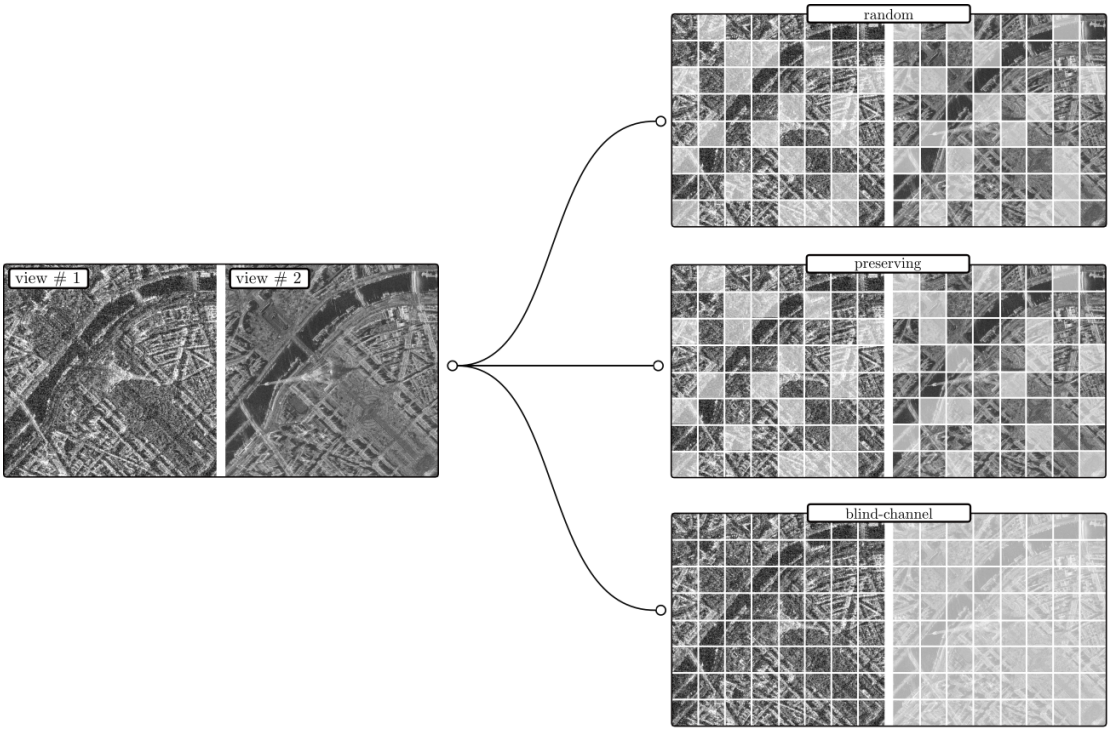


Figure 4.3: Illustration of different masking strategies for a two view SAR input to the *SARFormer*. Next to the random masking strategy, the persistent masking as well as the blind-channel masking both minimize the portion of the image where no information about the geometrical structure is available.

4.3 IaI-SimCLR

In the following section, the method *IaI-SimCLR* is introduced, where the objective lies in extending the augmentation-free pre-training approaches [56, 57] in a domain-informed manner in order to preserve the physical integrity of the signals throughout the decoding stage. As described in Chapter 3, the inherent availability of geolocation in EO data can be leveraged to design pre-training strategies that use data from multiple sensors to construct contrastive learning problems, enabling the development of descriptive encoders in a self-supervised manner. However, it can be argued that this leads to an intrinsic problem. Given two views of the same scene acquired with different sensor modalities, the encoders can only focus on information that is present in both views, since both inputs are aimed to be mapped into the identical location in latent space. In the following, this will be referred to as inter-modality information. Further, the network has to learn to be invariant to all information that only occurs in one of the two views, e.g. spectral information in multi-spectral data or SAR specific features caused by the imaging geometry, since this information can not be used to align the vectors in latent space. In the following, this will be referred to as intra-modality information.

To visually support this hypothesis, Fig. 4.4 shows an example of the same scene captured by a multi-spectral and a SAR sensor, respectively. These two modalities form the basis for the experimental validation of the hypotheses developed in this section, as will be seen later. Even though previous works [56, 57] are showing promising results while

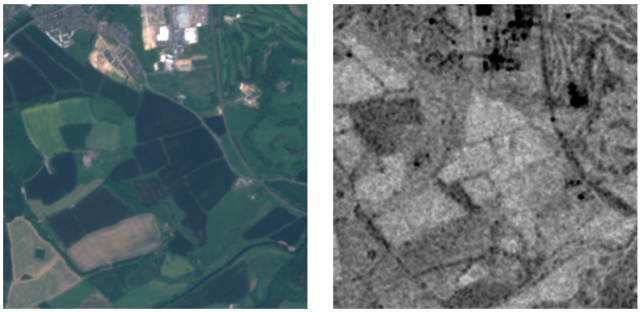


Figure 4.4: An exemplary multi-spectral (Sentinel-2) and SAR (Sentinel-1) image over the identical location. The samples are taken from the $\mathcal{D}_{MM}^{PT}$ dataset.

focusing on inter-modality views only, an increasing performance can be expected by solving the above-mentioned shortcomings.

Further, the intrinsic access to the geolocation of the data can be used to manipulate the *degree of similarity* of samples used in the contrastive pre-training setup. This allows for an additional parameter – besides the commonly used augmentations, whose effect on EO data is intuitively problematic as discussed in Section 2.2.2 – that influences the resulting quality of representations.

In the following, two methodological advancements to previous contrastive learning se-

tups, adjusted to EO data, are introduced. Without loss of generalizability the innovations are implemented as an extension of the classical SimCLR (compare Section 2.2.2) setup and subsequent extensions to EO data [56, 57].

Intra- and Intermodality Loss-Function In order to approach the issue of vanishing inter-modality information, the ***I**ntra- **a**nd **I**nter-modality SimCLR* (*IaI-SimCLR*) is introduced. Here, in a multi-objective setup where representations are forced to fulfill both an inter-modality similarity as well as an intra-modality similarity, information that is exclusively present in one of the two images is preserved. Starting with two augmentation-free views collected from sensors with different modalities $\mathbf{x}_{S1} \in \mathbb{R}^{P\times W\times H}$ and $\mathbf{x}_{S2} \in \mathbb{R}^{C\times W\times H}$, where two separate encoders $\mathbf{h}_{S1} = \mathbf{E}^{S1}(\mathbf{x}_{S1})$ and $\mathbf{h}_{S2} = \mathbf{E}^{S2}(\mathbf{x}_{S2})$ are applied and resulting features get further projected into a lower-dimensional latent space by two separate projection heads $\mathbf{z}_{S1} = f_{S1\,inter}(\mathbf{h}_{S1})$ and $\mathbf{z}_{S2} = f_{S2\,inter}(\mathbf{h}_{S2})$. As before W and H denote the width and height of the input image, and C as well as P the number of channels of the multi-spectral signal and the number of different polarizations acquired by the SAR sensor, respectively. In the following, individual samples that form the positive pairs (same location, different modality) of the batches $\mathbf{z}_{S1}$ and $\mathbf{z}_{S2}$ get denoted as $\mathbf{z}_i$ and $\mathbf{z}_j$. From here, the contrastive loss (Normalized Temperature-scaled Cross Entropy [38]) loss can be calculated along all positive pairs of a mini-batch as:

$$\mathcal{L}_{i,j}^{inter} = -\log\left(\frac{\exp\left(\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_j)\right)/\tau}{\sum_{k=1}^{2N}\mathbb{1}_{k\neq i}\exp\left(\mathrm{sim}\left(\mathbf{z}_i,\mathbf{z}_k\right)\right)}\right) \tag{4.7}$$

where N denotes the batch size, sim defines the cosine similarity, τ the temperature parameter and $\mathbf{z}_k$ negative samples from either modality. Here the operator $\mathbb{1}_{k\neq i}$ valuates to 1 if $k \neq i$ and 0 otherwise. The intuitive impact of $\mathcal{L}_{i,j}^{inter}$ is to extract features within each input that occur in both modalities. The loss gets extended by two additional loss functions that subsequently force the intra-modality similarity with respect to applied transformations. For that, two augmented versions (per modality) of the original view are introduced, which will be further denoted by $\mathbf{x}'_{S1}$, $\mathbf{x}''_{S1}$ and $\mathbf{x}'_{S2}$, $\mathbf{x}''_{S2}$, respectively. Introducing two additional projection heads for the intra-modality similarity $f_{S1\,intra}$ and $f_{S2\,intra}$, the representations can be calculated as:

$$\begin{aligned}
\mathbf{z}'_{S1} &= f_{S1\,intra}(\mathbf{E}_{S1}(\mathbf{x}'_{S1})) \\
\mathbf{z}''_{S1} &= f_{S1\,intra}(\mathbf{E}_{S1}(\mathbf{x}''_{S1})) \\
\mathbf{z}'_{S2} &= f_{S2\,intra}(\mathbf{E}_{S2}(\mathbf{x}'_{S2})) \\
\mathbf{z}''_{S2} &= f_{S2\,intra}(\mathbf{E}_{S2}(\mathbf{x}''_{S2}))\ ,
\end{aligned} \tag{4.8}$$

analogue to the inter-modality case. With, $\mathbf{z}_i$ and $\mathbf{z}_j$ representing positive pairs from the batches from either $\mathbf{z}'_{S1}$ and $\mathbf{z}''_{S1}$ or $\mathbf{z}'_{S2}$ and $\mathbf{z}''_{S2}$ the intra-modality loss can be calculated analogue to Eq. 4.7 and will be denoted as $\mathcal{L}_{i,j}^{S1\,intra}$ and $\mathcal{L}_{i,j}^{S1\,intra}$, respectively. With equal weight of all losses, the overall loss function is given by:

$$\mathcal{L}^{total} = \mathcal{L}^{inter} + \mathcal{L}^{S1\,intra} + \mathcal{L}^{S2\,intra}\ . \tag{4.9}$$

This setup allows for three meaningful setups, namely the *DualSimCLR* setup of [56] with pure inter-modality component of the loss-function, the *Intra-SimCLR* setup, close to the

native SimCLR approach, as well as the within this thesis developed *IaI-SimCLR* where intra- as well as inter-modality contributions are incorporated for determining the degree of similarity. A graphical illustration of the proposed method and all corresponding combinations of the loss formulation studied can be seen in Fig. 4.5.

Batch Sampling Strategies The fundamental objective of any contrastive learning approach is to effectively separate positive (similar) examples from negative (dissimilar) ones. Consequently, the degree of variation in the presented imagery plays a crucial role in determining the complexity of the learning problem. Design choices that introduce shortcut opportunities during self-supervised pre-training can significantly influence the model's ability to generalize, thereby impacting its performance on downstream tasks, as discussed in Section 2.2.2. As stated above, in the context of EO data, a unique advantage lies in the inherent geolocation information, which enables a natural way to define similarity by selecting spatially proximate examples. This characteristic provides an opportunity to structure the training data in ways that directly influence the learning process. To investigate this, the impact of two distinct sampling strategies is examined within this thesis: firstly *local batch sampling* (LBS), where images are selected based on spatial proximity, and secondly *random batch sampling* (RBS), where images are drawn randomly from all available samples. A visual comparison between these two approaches is presented in Fig. 4.6. Throughout the experimental section, the results are systematically categorized and compared on whether they were trained with LBS or RBS, allowing for an assessment of the impact of spatial sampling strategies on contrastive learning performance.

Methodological Summary Pre-training through contrastive learning with multi-modal imagery in the context of remote sensing is – due to the intrinsic multi-modal availability of data within the remote sensing domain – a promising approach. The *IaI-SimCLR* framework, though the proposed loss function that operates on the intra- and inter-modality similarity level provides more flexibility to adjust resulting representations with respect to their descriptiveness. Furthermore, within the context of the *IaI-SimCLR* a new batch-sampling strategy based on spatially proximate samples for the construction of the negative pairs is proposed.

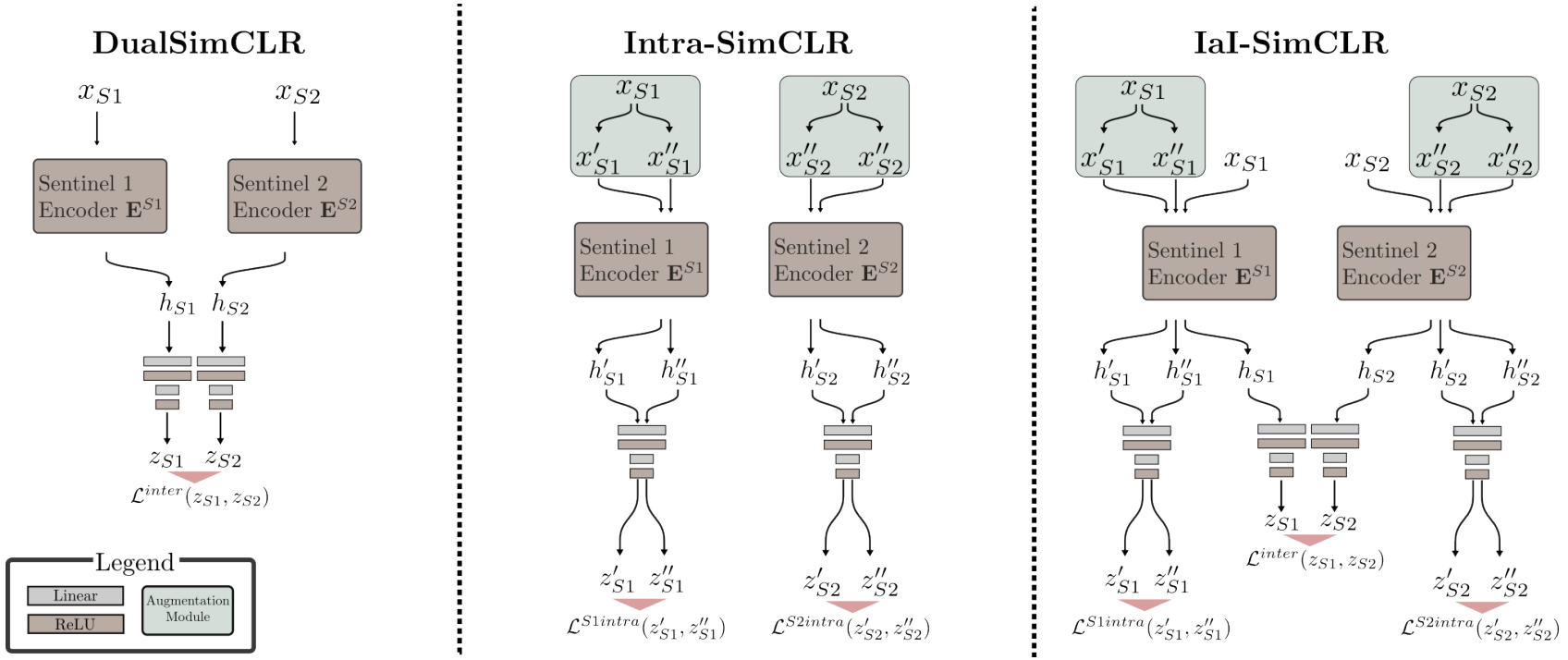


Figure 4.5: A graphical overview of the proposed experiment setup, comparing the three models *Dual-SimCLR* [56], *Intra-SimCLR*, and *IaI-SimCLR*. *Intra-SimCLR* is the original *SimCLR* [38] approach, slightly adapted to the multi-modality setup.

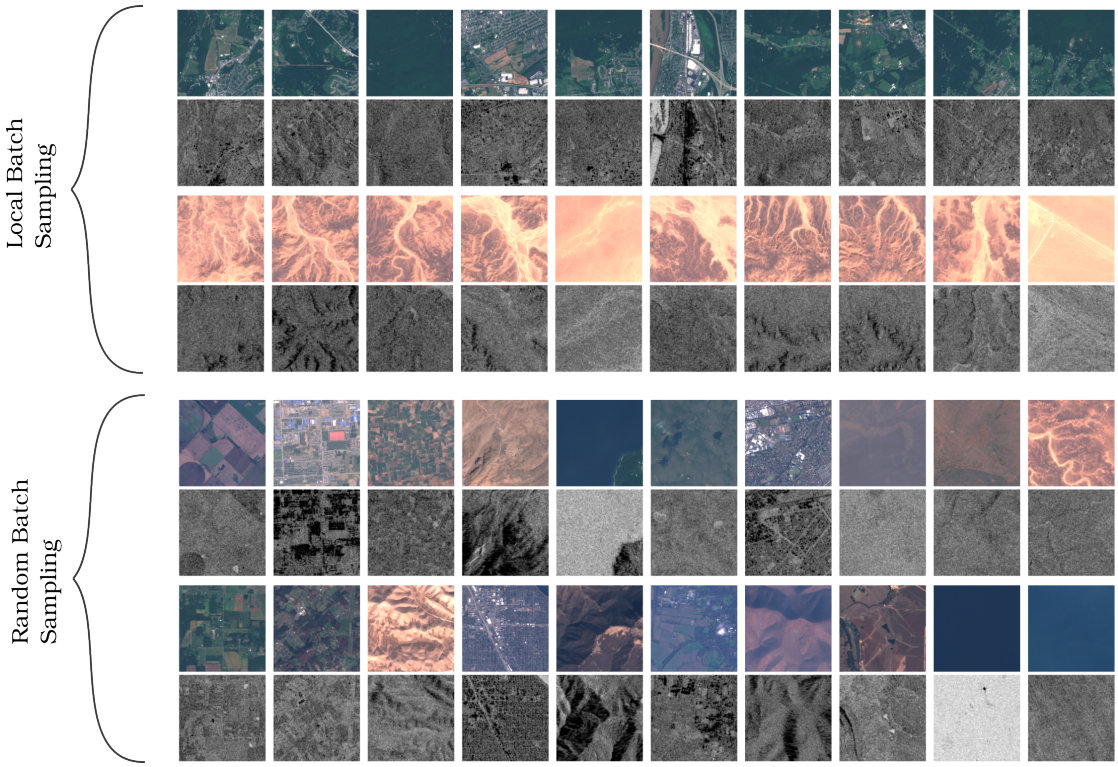


Figure 4.6: Example images for the two different batch sampling methods. The top four rows of images were drawn via LBS (twenty Sentinel-2 and corresponding Sentinel-1 samples), bottom four rows of images were drawn by RBS.

5

Sensors and Datasets

This chapter provides an overview of the technical specifications of the sensors utilized in this study, as well as the characteristics of the resulting datasets that were generated and processed for both pre-training and fine-tuning, respectively. All preprocessing steps, the geographic distribution, and any relevant considerations that impact their use in the subsequent analysis are outlined.

5.1 Used Sensors

In the following, the satellite missions and their respective sensor specifications of all sensors used throughout this thesis are introduced:

Sentinel-2 The flagship of the Sentinel constellation, operated by ESA is Sentinel-2, two satellites (indexed alphabetically) each with a multi-spectral instrument [87] capturing incoming light divided into thirteen spectral bands ranging from 442 nm to 2202 nm in center wavelength. The channels are grouped according to their respective GSD into three classes: 10 m, 20 m, and 60 m. During the experimental part of this thesis, only the 10 m and 20 m GSD bands are utilized since the 60 m bands are mainly designed around observing atmospheric properties and can reasonably be neglected for land surface monitoring tasks. The spectral position of the remaining ten channels can be found in Fig. 5.1. Each sensor has a 290 km swath, and the overall constellation has a five-day revisit at the equator. The data archive includes data ranging back to 2015 and 2017, respectively, for the two sensors. Sentinel-2 images are provided by ESA in the TOA and BOA (compare Section 2.1.1) processing levels. In the following, all references to

Note: As of the day of writing, the constellation includes three satellites as *Sentinel-2C* is operational, whereas *Sentinel-2A* is close to decommissioning.

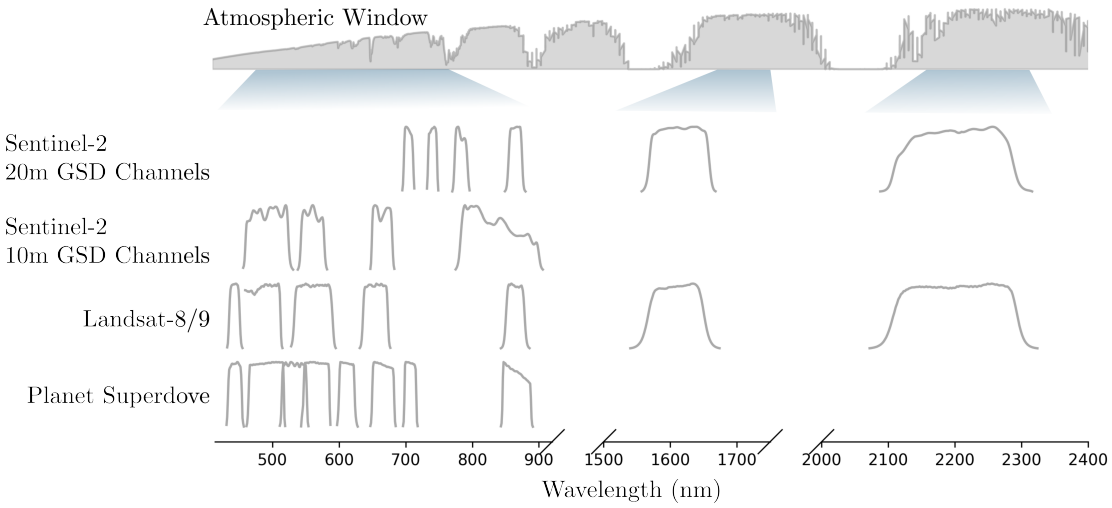


Figure 5.1: The position of the spectral response functions for the sensors Sentinel-2, Landsat and Superdove in the electromagnetic spectrum. Each response function has values ranging in $[0, 1]$, and shifts along the y axis are purely to distinguish different sensors and different sets of GSD within the set of channels. The *atmospheric transmittance spectrum* (ATS) of the Earth's atmosphere (compare Fig. 2.3), crucial when designing channels, is provided as a reference. Shaded blue boxes below the ATS indicate the position of the channels relative to the ATS (since the x-axis of the plot is split up into three regions).

Sentinel-2 data refer to the BOA level, where the atmospheric distortions are already accounted for.

Landsat The Landsat program, managed by NASA, is a long-running Earth observation mission comprising a series of successor satellites. Since the fourth generation of the mission (launched 1982), Landsat satellites carry both a multi-spectral camera named *operational land imager* (OLI) and a *thermal infrared sensor* (TIRS). Within this thesis, data from OLI of the two satellites *Landsat-8* and *Landsat-9* is used, and since the multi-spectral instrument is directly comparable for both [88], they will be referred to as Landsat without any further distinction. The sensors record data with a 30 m GSD resolution divided into seven channels. The corresponding spectral response functions are partially similar to Sentinel-2 and can be seen in Fig. 5.1. The constellation has an 8-day revisit and 185 km swath [13]. Analogue to Sentinel-2 within this thesis, all Landsat data will be used in the BOA processing level.

Superdove The Superdove constellation, operated by Planet Labs, consists of a fleet of cubesats designed for low revisit time Earth observation. Due to constant launches and decommissioning phases, the actual number of sensors in space varies but is in the range of tens to hundreds, which allows for a daily revisit even with the comparably

small swath width of ≈ 32 km [89]. Within this thesis, data denoted as Superdove will refer to the currently newest version of the sensor PSB.SD which captures data within eight spectral bands at 3.7 m to 4.7 m GSD [89]. The spectral position of the channels partially matches and partially adds to the channels provided by Sentinel-2 and Landsat (compare Fig. 5.1). As with the previous two sensors, data from Superdove will be used in the BOA processing level (referred to as surface reflectance in the context of Superdove).

Sentinel-1 Sentinel-1 is a constellation of C-band SAR satellites operated by ESA. The satellites cover a 250 km swath with a range resolution of 5 m and an azimuth resolution of 20 m while operating in the main *interferometric wide swath* (IW) mode [90]. Within this operational mode, the constellation has a revisit time of 6 days and measures the reflected echoes in two polarization modes VV and VH, allowing for surface feature discrimination across various land and maritime environments. The Sentinel-1 data used in this thesis corresponds to the *ground range detected* (GRD) format with backscatter values in logarithmic scale with already applied data preprocessing steps, including radiometric calibration and terrain correction [91].

Note: The *interferometric wide swath* is a SAR operation mode similar to the ealier disussed *ScanSAR* mode.

TerraSAR-X TerraSAR-X is a high-resolution SAR satellite mission operated by Airbus Defence and Space in partnership with the *German Aerospace Center* (DLR). The constellation is composed of two twin satellites flying in a controlled formation with a relatively small distance designed for interferometric applications [92]. Within this thesis, only the backscatter information of either of the two sensors is utilized and referred to as TerraSAR-X, even though precise nomenclature would distinguish between TerraSAR-X and TanDEM-X, which is the name for the second satellite as well as the name for the overall constellation. The satellites operate in the X-band in various acquisition modes, namely *stripmap* (SM), *high-resolution spotlight* (HS), *spotlight* (SL), and *staring spotlight* (ST) with a revisit time of 11 days. The TerraSAR-X platform can measure co- as well as cross-polarization in single or dual channel mode [93], where within this thesis, only single channel co-polarization data is utilized. Tab. 5.1 gives an overview of the resulting product specification for all four modes.

Abbreviation	Scene Extent Azimuth × Ground Range	Azimuth Resolution	Ground Range Resolution
SM	(up to) 1650 km × 30 km	3.3 m	1.7 m – 3.5 m
SL	10 km × 10 km	1.7 m	1.48 m – 3.49 m
HS	5 km × 10 km	1.1 m	1.48 m – 3.49 m
ST	∼ 2.65 km× ∼ 6 km	0.24 m	0.85 m – 1.77 m

Table 5.1: The resolution in azimuth and ground range as well as the corresponding scene extent for the different acquisition modes of TerraSAR-X mission (for a single channel observation) taken from [93, 94]. The range of the resolution in ground range direction is a consequence of different local looking angles as described in Section 2.1.2. It can be observed that resolution is inversely proportional to the extent covered by the sensor.

5.2 Description of Utilized Datasets

Throughout this work, various datasets are utilized, both pre-existing and custom-built for model pre-training and subsequent downstream task evaluation. In the following two sections, the details of all datasets are outlined. An overview of all datasets, the corresponding abbreviations used throughout this thesis, as well as their main characteristics, is given in Tab. 5.2 and Tab. 5.3 for pre-training and fine-tuning, respectively. Furthermore, Figs. 5.2 and 5.3 illustrate the geographical distribution of the utilized datasets.

5.2.1 Pre-training Datasets

In the following, the three datasets used for self-supervised pre-training of models are outlined, spanning a multi-sensor dataset for multi-spectral sensors, a high-resolution SAR dataset, as well as a multi-modal dataset incorporating spatially paired multi-spectral as well as SAR data.

Multi-Sensor Multi-Spectral Pre-training Dataset With the objective to incorporate sensor information with respect to the main sensor parameters for multi-spectral optical sensors in the *SenPaMAE* framework, the pre-training dataset $\mathcal{D}^{PT}_{MSMS}$ (**PT** for pre-training and **M**ulti-**S**ensor **M**ulti-**S**pectral) combines spatially paired data from the three satellites Sentinel-2, Landsat, and Superdove, globally sampled from 27 locations. Data from Sentinel-2 and Superdove is collected from the same date, while Landsat was sampled from within a three-month window. While all eight channels from the Superdove are included, only the 10 m and 20 m GSD channels of Sentinel-2 are considered, and Landsat is limited to its seven optical channels with 30 m GSD, neglecting data from the thermal sensors as well as the panchromatic band. The three selected

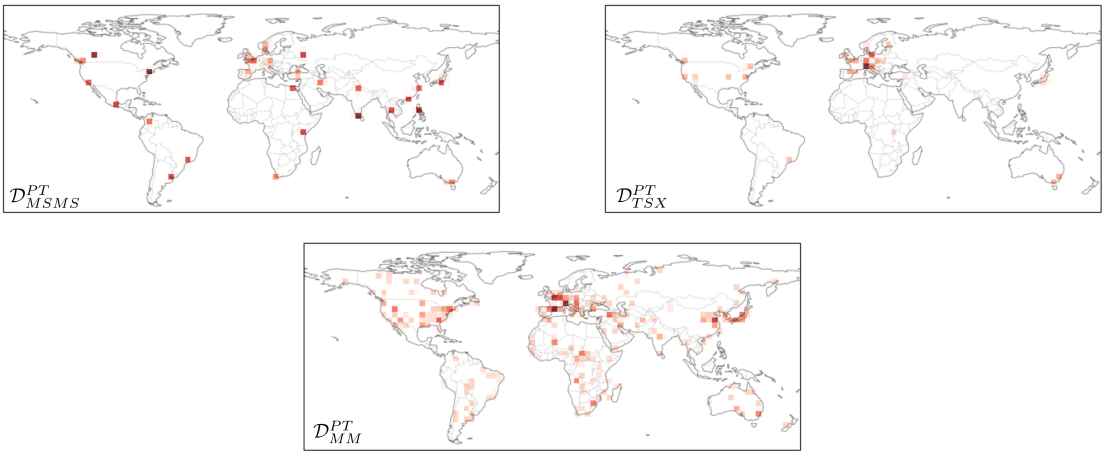


Figure 5.2: Illustration of the geographical distribution of the three pre-training datasets utilized in this thesis. Darker colors correspond to higher sample density, normalized separately for each dataset.

sensors represent some of the most frequently used data sources for optical Earth observation, offering a diverse set that varies in the number of channels, response functions, and resolution. The spectral response functions of all used channels are displayed in Fig. 5.1. From the 27 locations overall 50k patches of size 256 px × 256 px are sampled for the Sentinel-2 and Superdove images, as well as 25k further patches from Landsat and after a geographical validation split ending up with ≈ 116k images from three sensors used for training. The distribution of the training samples can be seen in Fig. 5.2. The GSD values for the above-described imagery span the values ≈ 3.5 m for Superdove to 10 m and 20 m for Sentinel-2 and 30 m for Landsat. As will be seen later, encoding the sensor resolution into the model - independent of the covered area - resampling (cubic) of all images to a 5 m pixel-spacing grid is carried out in order to yield a harmonized input signal and keeping the ground extent of the input imagery constant. All data, independent of the sensor, is preprocessed according to the respective sensor data formats to represent surface reflectance values between 0 % and 100 %, encoded in the $[0, 1]$ interval, which enables direct comparison across sensors.

Note: While access to S2 and LS imagery is free, the *Superdove* was acquired through *PlanetLabs* free research plan.

High Resolution SAR Pre-training Dataset For the incorporation of scene acquisition parameters of SAR data into the learning process via the *SARFormer* framework, the dataset $\mathcal{D}^{PT}_{TSX}$ comprises over 200 TerraSAR-X satellite acquisitions spanning multiple imaging modes, and different geometrical acquisition parameters (orbit directions, looking angle and azimuth angle) as described in Section 2.1.2. All acquisitions are predominantly centered around urban environments with a geographical bias toward European and North American cities, due to the availability of high-resolution labels for downstream tasks (compare the following section). Similar to $\mathcal{D}^{PT}_{MSMS}$ a pixel grid spacing of 1 m is chosen to generate a constant field of view on the corresponding Earth's surface independent of the imaging mode. After patchifying all acquisitions, the overall number of samples – including ones over the identical location with different viewing parameters

Abbreviation	Sensors Included	Overall Number of Available Samples	Core Concept
$\mathcal{D}_{MSMS}^{PT}$	Sentinel-2 Landsat Superdove	125,000	Temporally and spatially paired multi-sensor data, globally distributed around 29 locations
$\mathcal{D}_{TSX}^{PT}$	TerraSAR-X	661,760	Diverse acquisition geometries, spatially paired, globally distributed around urban areas
$\mathcal{D}_{MM}^{PT}$	Sentinel-1 Sentinel-2	180,662	Spatially paired multi-modal acquisitions, globally distributed and seasonally balanced

Table 5.2: Overview of the pre-training datasets used within this thesis, including the sensors, the total number of available samples, and the main characteristics of each dataset.

– counts approximately $660,000$ of size $726\,\text{px} \times 726\,\text{px}$ pixels. Due to the significantly enlarged field of view of the SM acquisition mode with respect to higher spatial resolution ones, the resulting sample distribution has bias towards stripmap acquisitions with $\approx 92\%$ in comparison to $\approx 3\%$ HS, $\approx 3\%$ SL and $\approx 1\%$ ST. This bias (along with other parameters) is accounted for during pre-training and fine-tuning through batch-sampling strategies, as described in Section 6.2. The backscatter intensities are clipped to a range of -30 to $+10\,\text{dB}$, and subsequently mapped to the $[0, 1]$ interval for neural network input.

Multi-Modal Pre-training Dataset For experiments investigating pre-training strategies through multi-modal alignment through the *IaI-SimCLR* framework, imagery from the *SEN12MS* dataset [95] is utilized during the pre-training stage and further denoted with $\mathcal{D}_{MM}^{PT}$. *SEN12MS* combines data from the two missions Sentinel-1 and Sentinel-2 with a uniform global distribution as well as a seasonally balanced time of recording. Similar to the previous approaches the Sentinel-2 data will be restricted to $10\,m$ and $20\,m$ GSD bands of *Sentinel-2*. For data of the Sentinel-1 sensor, both polarization VV and VH are utilized (compare Section 2.1.2). In total the dataset comprises $180,662$ samples of size $256\,\text{px} \times 256\,\text{px}$, where all bands are resampled to 10 m pixel spacing. Analogue to the $\mathcal{D}_{MSMS}^{PT}$ the Sentinel-2 data get preprocessed to represent the surface reflectance values between $0\,\%$ and $100\,\%$ encoded in the $[0, 1]$ interval. Sentinel-1 data on the other hand, is analogous to $\mathcal{D}_{TSX}^{PT}$, clipped to -35 to $0\,\text{dB}$, and subsequently mapped to the $[0, 1]$ interval.

5.2.2 Downstream Datasets

In the following six downstream datasets used for testing the developed methods are outlined. This encompasses a multi-sensor multi-spectral dataset paired with LULC, a high resolution SAR dataset paired with *light detection and ranging* (Lidar) derived heights and binary building footprints labels as well as four patch-wise downstream tasks on (partially) multi-modal data ranging from LULC, crop-type mapping as well as biomass regression. It should be noted that during the experiments, not necessarily the complete dataset is utilized, due to balancing or other experimental restrictions. The actual experimental setting will be described together with the experimental results in the corresponding Sections 6.1 to 6.3.

Abbreviation	Sensors Included	# Samples Train/Val	Task Description
$\mathcal{D}_{MSMS}^{EWC}$	Sentinel-2 Superdove	42k / 14k	LULC (pixel-wise)
$\mathcal{D}_{TSX}^{Buildings}$	TerraSAR-X	600k / 68k	Height Regression Building Segmentation (pixel-wise)
$\mathcal{D}_{MM}^{DFC}$	Sentinel-1 Sentinel-2	2.8k / 1.2k	Main LULC (patch-wise)
$\mathcal{D}_{MM}^{EWC}$	Sentinel-1 Sentinel-2	32k / 14k	Main LULC (patch-wise)
$\mathcal{D}_{MM}^{Crop}$	Sentinel-2	24k / 8k	Main Crop-Type (patch-wise)
$\mathcal{D}_{MM}^{Biomass}$	Sentinel-1 Sentinel-2	32k / 14k	Biomass Regression (patch-wise)

Table 5.3: Overview of the downstream datasets used, including the sensors, the total number of samples for training and testing, as well as the corresponding task type.

Multi-Sensor Multi-Spectral Downstream Dataset In order to test models that are pre-trained in a sensor-informed manner with respect to the main sensor parameter of a multi-spectral signal, the evaluation dataset $\mathcal{D}_{MSMS}^{EWC}$ is built around imagery from the two missions Sentinel-2 and Superdove acquired on the same day in analogy to $\mathcal{D}_{MSMS}^{PT}$ and paired with LULC information from *ESA World Cover* (EWC) [96]. The constant pixel spacing grid, as well as the included bands, are all designed in line with $\mathcal{D}_{MSMS}^{PT}$. The images were acquired within one day over nine locations centered around urban areas within the United States of America. After a regional split, six locations are used for training with $\approx$ 42k patches of size 256 px $\times$ 256 px, per sensor. Two locations spanning $\approx$ 14k patches are used for evaluation. After removing non-occurring classes

in the EWC labels, an eight-class classification scheme is chosen covering all existing classes over the study area, namely *Tree Cover, Shrubland, Grassland, Cropland, Built-Up, Sparse Vegetation, Water, Wetland*. The distribution of the LULC frequency is highly unbalanced and will be addressed with accordingly designed batch sampling as described in Section 6.1. In the following, the dataset will be denoted as $\mathcal{D}_{MSMS}^{EWC}$.

High Resolution SAR Downstream Dataset In order to test the validity of the approaches of incorporating the acquisition parameter injection for high resolution SAR imagery, the data underlying $\mathcal{D}_{TSX}^{PT}$ datasets get paired with two types of annotations. First, building footprint data from two sources *Open Street Map* (OSM) [97] and *Microsoft building footprints* (MBF) [98] were merged. Since OSM information yields more accurate building outlines, OSM polygons are used preferably if available. If no OSM building footprint information is available, polygons from the MBF dataset are utilized. Furthermore, height annotations are derived from publicly available Lidar data provided by regional surveying authorities. Since height labels are not available for all cities in the dataset, some locations lack the corresponding height ground truth information, which will be addressed accordingly in the fine-tuning stage as later described in Section 6.2. Absolute elevations are converted to heights above ground by subtracting a terrain model. Overall, after holding back all images of Berlin and Vancouver for validation purposes, this results in $\approx$ 90k patches with corresponding height information, whereas building footprint information is available for all samples of $\mathcal{D}_{TSX}^{PT}$. This dataset is an extension of the work presented in [99, 100, 101, 102] but due to the licensing restrictions is not published separately. In the following, the dataset will be denoted as $\mathcal{D}_{TSX}^{Buildings}$.

Multi-Modal Downstream Dataset For models pre-trained on a multi-modal alignment task, four downstream tasks were generated to test the descriptiveness of the derived representations, all designed as a patch-wise classification or regression task.

First, the *SEN12MS-DFC* dataset [103], which combines imagery of the Sentinel-1 and Sentinel-2 sensors with semi-manually created high-resolution LULC information gets utilized. The label for each sample is derived by taking the main LULC class, and the sample is only included if the main class exceeds 30 %. All patches with *Water* as the main class get discarded, which reduces the problem to *Forest*, *Shrubland*, *Grassland*, *Wetland*, *Cropland*, *Urban*, and *Barren*. After a random split, this results in overall 2808 training and 1204 validation samples of size 256 px $\times$ 256 px, respectively. The dataset will further be denoted as $\mathcal{D}_{MM}^{DFC}$.

A second LULC downstream task is based on the SEN12MS-TP [104] dataset, where Sentinel-1 and Sentinel-2 data get paired with LULC information from EWC [96], such as in $\mathcal{D}_{MSMS}^{EWC}$. Similar to $\mathcal{D}_{MM}^{DFC}$, the main class (if over 30 %) represents the patchwise label. After a random split, this results in overall 31975 training and 13790 validation samples of size 256 px $\times$ 256 px, respectively. The dataset will further be denoted as $\mathcal{D}_{MM}^{EWC}$.

Besides the land cover classification problem, another task is represented by the *Canadian-Crop* dataset [105]. In contrast to the previous two downstream tasks, this dataset only includes data of the optical modality given by Sentinel-2 data, and therefore only tests the capability of the part of the model that processes the multi-spectral data. Since this dataset is heavily biased towards certain crop classes it will be restricted to *Pasture*, *Orchard*, *Potato*, *Soybean*, *Mixedwood*, *Barley*, *Oat* and *Corn*, where – with accordingly designed batch-sampling strategies (compare Section 6.3) – a well designed learning task can be provided. After reduction the dataset contains overall 24000 training and 8000 validation samples of size 65 px × 65 px, respectively. The dataset gets denoted as $\mathcal{D}_{MM}^{Crop}$ in the following.

A further downstream for testing the multi-modal models is given by the estimation of the total biomass within a given image patch. For that, the above-mentioned *SEN12MS-TP* dataset [104] (Sentinel-1 and Sentinel-2 data) got extended and paired together with data of the spaceborn Lidar GEDI-L4A [106] mission. This results in a regression task where the corresponding overall biomass for a given patch has to be estimated accordingly to the *gridded aboveground biomass density* product [107]. The datasets get denoted as $\mathcal{D}_{MM}^{Biomass}$ in the following.

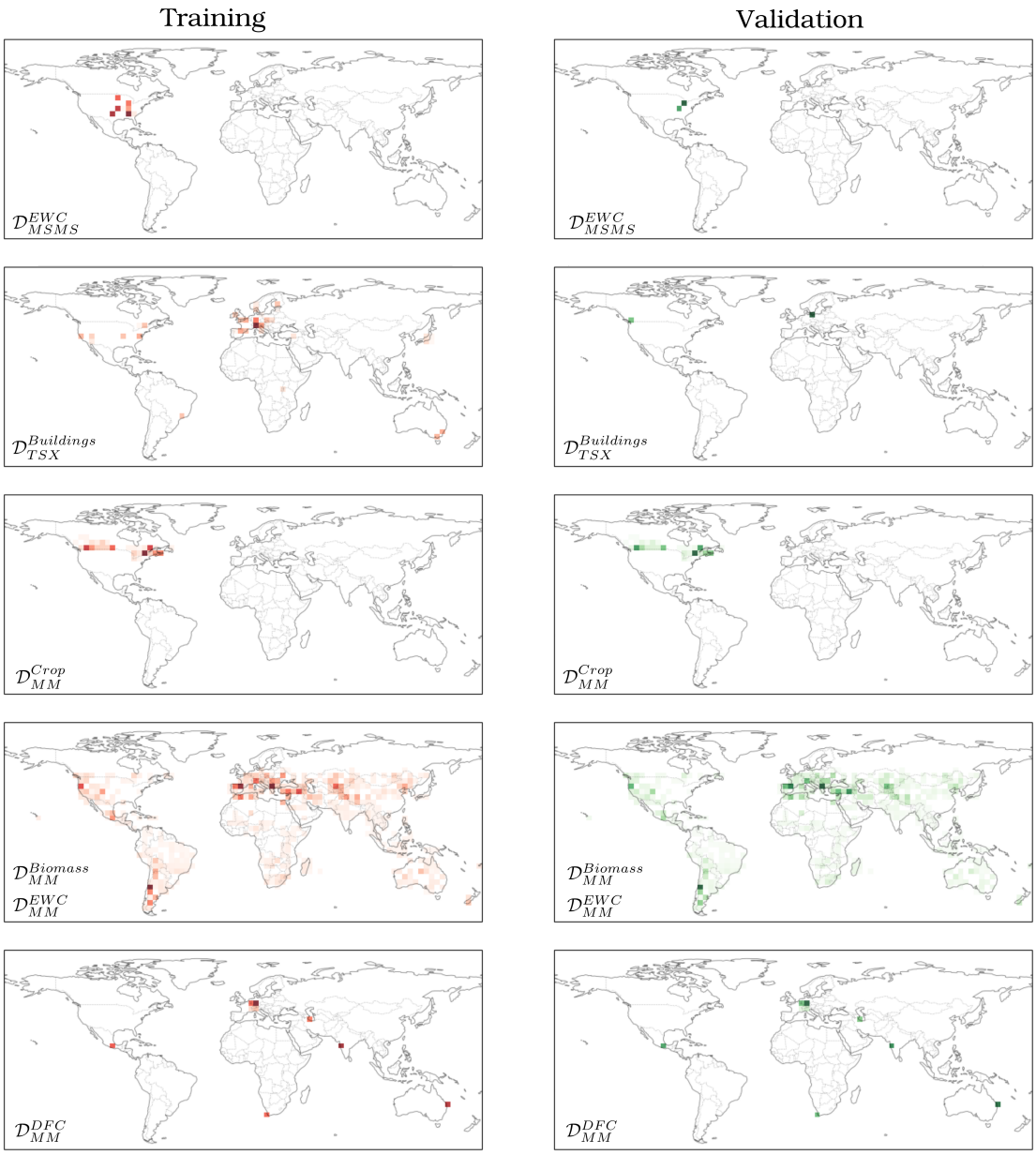


Figure 5.3: Illustration of the geographical distribution of the six fine-tuning datasets utilized in this thesis. Darker colors correspond to higher sample density, where the maximal density is scaled to one for each dataset individually.

6

Empirical Evaluation

In this chapter, the empirical evaluation of the three introduced methods – *SenPaMAE*, *SARFormer*, and *IaI-SimCLR*– is presented. For each approach, the respective pre-training procedure is described, followed by the adaptation to their corresponding downstream tasks.

6.1 Evaluation of the SenPaMAE Framework

In the following, the experimental validation of the *SenPaMAE* method, described in Section 4.1, is outlined, including a description of the pre-training procedure as well as the subsequent downstream experiments with respect to multi-sensor LULC tasks.

Pre-Training Settings *SenPaMAE* as well as the vanilla MAE-ViT version *BaseMAE* are pre-trained on the $\mathcal{D}^{PT}_{MSMS}$ dataset, described in Section 5.2.1 to reconstruct the input images with respect to the mean absolute error between the reconstruction – after the decoder – and the original input image. The ViT backbones for all models in this section have a fixed size of 12 layers with 12 heads for the encoder and 3 layers with 12 heads for the decoder, which corresponds to the ViT-Base setup, suggested in [1]. The response function information for all sensors is gathered from the satellite providers [108, 109, 110]. All $\boldsymbol{\lambda}_c$ get resampled to 300 nm - 2600 nm wavelength interval, with 1 nm step width. Hence, each channel $\mathbf{x}_c$ has a corresponding scalar GSD value $\sigma_c \in \mathbb{R}$ and as well as a response function vector $\boldsymbol{\lambda}_c \in \mathbb{R}^{2300}$.

The data, collected from the three sensors Sentinel-2, Superdove and Landsat in $\mathcal{D}^{PT}_{MSMS}$

have a varying number of channels. As described in Section 4.1, each band of the corresponding multi-spectral signal gets divided into patches – as in all ViT based approaches – and subsequently encoded into tokens via a linear transformation, resulting in $C \times H/P \times W/P$ tokens. In principle, the ViT architecture can handle data with various input lengths, still this requires severe engineering adaptations e.g. interpolation of the positional encoding of all tokens. In order to eliminate this degree of freedom, the experiments are done with a fixed size of input imagery, where all inputs are restricted to four channels by taking random subsets of all available bands. Since all data is utilized, this simulates a scenario of a high number of diverse sensors, all with a constant number of four channels[1]. It should be noted that for each input $\mathbf{x}$ the model always additionally gets the corresponding sensor parameters $\{\boldsymbol{\lambda}_c, \sigma_c\}$, hence is *informed* which random subset of the imagery is presented.

From the data included in $\mathcal{D}^{PT}_{MSMS}$, random patches of size 144 px $\times$ 144 px are sampled, and with a patch size of 16, this results in a 9×9 patch pattern per channel. All patches are transformed into the latent token space with size $d_{emb} = 768$. The masking ratio r_{mask} is set constant to 66% for all runs. If the data augmentation module SSA is active, the probabilities for channel mixing p_{mix} and downsampling p_{down} are set to 25%, with randomly two or three channels for the spectral superposition (equal probability). Training takes place over 400 epochs with 3 warm-up epochs, a cosine decay learning rate scheduler (initial learning rate of 1×10^{-4}), an *Adam optimizer*, a weight decay of 0.05, and a batch size of 128.

Resulting Reconstructions Fig. 6.1 gives a visual example of the pre-training task and the corresponding reconstruction quality for a pre-training example with four randomly chosen channels. These examples display the masking strategy, which applies individual masks per channel in order to avoid large unmasked areas, which lead to an ill-posed problem, the effect of the SSA as well as the random down-sampling operation performed on individual channels. Here, it can be seen that the model successfully reconstructs the input while successfully adjusting to the different sensor parameters. It should be noted that at this step, it is not the objective to choose the settings in a way to obtain as good a reconstruction as possible. Rather, a careful consideration of the experimental setting should be done in order to obtain a reasonably difficult problem (hence strong features must be learned in order to solve the problem) while keeping it in reasonable bounds.

[1] The design decision to work with synthetic sensors consisting of four channels each can be motivated as follows: Utilizing the full 12-, 10-, and 8-band input data would result in an experimental setup with only three distinct sensors. By generating synthetic sensors with four channels each, the diversity of sensors increases, which is precisely the aspect this method aims to capture. Working with only three sensors in a multisensor setup would drastically limit flexibility in downstream applications. For example, zero-shot inference would no longer be possible if at least two of the sensors (Sentinel-2 and SuperDove) were used during fine-tuning (using Landsat as the unseen sensor - due to its significantly lower resolution - would not serve as a fair comparison here).

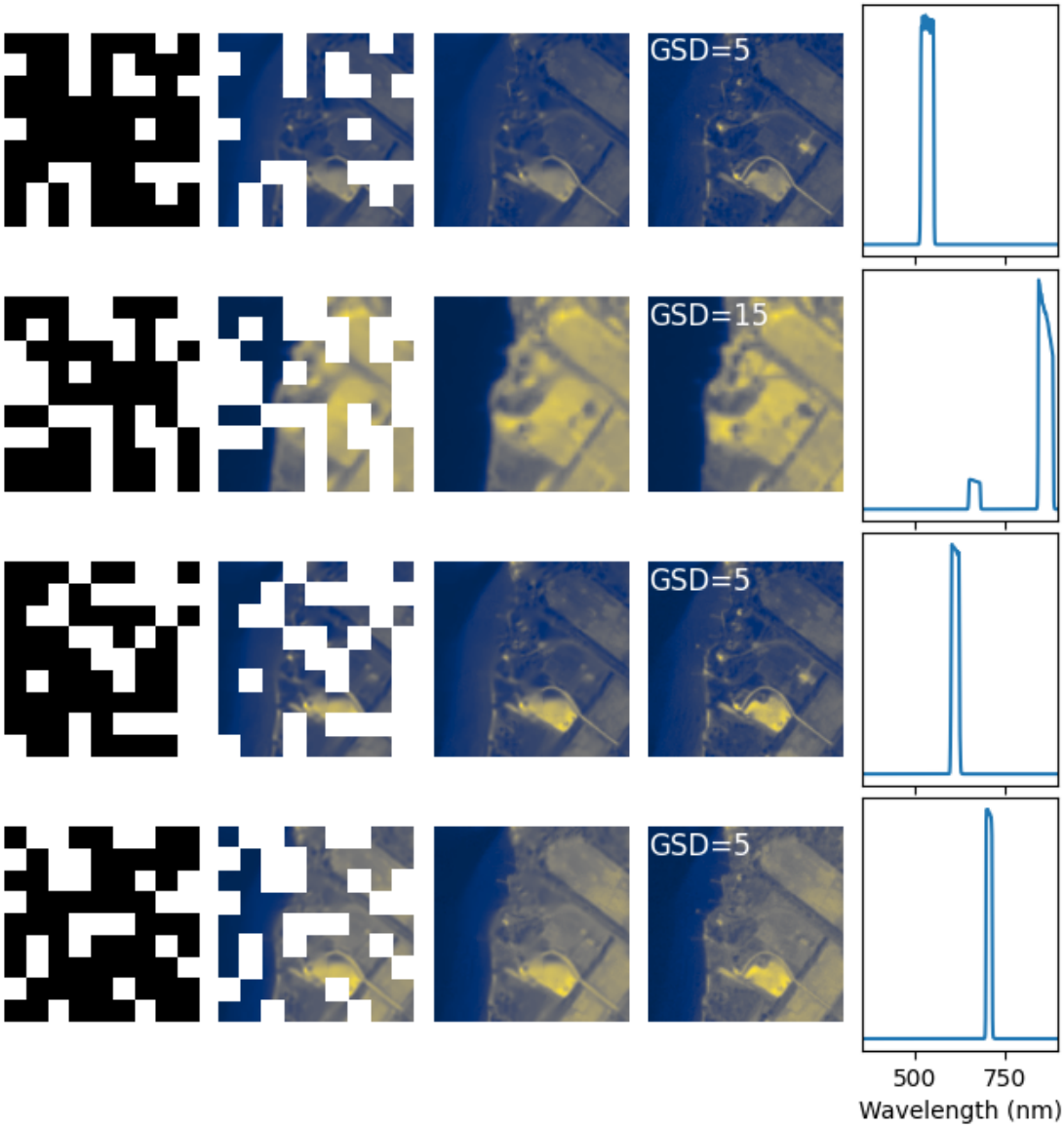


Figure 6.1: Reconstruction results for a masked four-channel input (one channel per row). From left to right, each panel shows: the channel-specific mask (black indicates masked regions, white indicates unmasked areas), the reconstructed signal within the masked regions, the reconstruction combined with the ground truth in the unmasked regions, the ground truth image annotated with the channel's GSD, and the corresponding spectral response function.

Objective of the Evaluation In this section, the capability of the pre-trained models to solve an LULC segmentation task in a multi-sensor arrangement is tested. It is worth mentioning that the overall objective of the *SenPaMAE* architecture is not necessarily restricted to achieving as high an evaluation score as possible when trained on each sensor individually. Rather, the correct injection of the sensor parameters into the model should be tested. This is in the following done by carrying out all experiments in a multi-sensor setup. For this, the data from the Sentinel-2 and Superdove sensors included in $\mathcal{D}_{MSMS}^{EWC}$ is utilized to create synthetic sensor arrangements (which is referred to as different sensors) by splitting up the available data into different channel groups. All sensors mentioned below will refer to the synthetic sensors. In the following, specific arrangements of a four-channel input for the sensors Sentinel-2 (S2) and Superdove (SD) are referred to with the four indices via subscripts. Therefore, e.g. the 10m GSD bands of Sentinel-2 get denoted as S2_{1237} and the spectrally matching counterpart of the Superdove sensors as SD_{2468} since the coastal blue channel is not present in the set of bands for the Sentinel-2 satellite. In order to follow along with the notation of the channel indices, the reader is referred to Fig. 5.1 where the bands for each of the three sensors are indexed according to ascending central wavelength. It is still important to note that all pre-trained models were exposed to data from all the available bands for each of the satellites in $\mathcal{D}_{MSMS}^{PT}$ during the pre-training step.

Two experiments are carried out: First, the model is fine-tuned on one fixed input distribution (exclusively data from one sensor), and the validation is conducted in a zero-shot manner for data from a different sensor. The second test is carried out with fine-tuning the model on data from multiple sensors simultaneously, both described in the following.

Fine-tuning Settings Fine-tuning takes place via the downstream dataset $\mathcal{D}_{MSMS}^{EWC}$ where, first, the multi-spectral image tensor $\mathbf{x}$ as well as the sensor parameters $\{\boldsymbol{\lambda}_c, \sigma_c\}$ (if the *sensor parameter encoding* (SPE) module is active) get processed by the pre-trained encoder $\mathbf{E}$. It should be noted that the decoder network of the pre-training step, as well as (if active) the second SPE module, does not get applied. The obtained processed tokens $\mathbf{t}_i^c$ with $i \in (1, \ldots, HW/P^2)$ and $c \in (1, \ldots, C)$ (compare Section 4.1) after layer j get denoted as $\mathbf{t}_{i,j}^c = \mathbf{E}_{(j)}(\mathbf{t}_{i,j-1}^c)$ where $\mathbf{E}_{(j)}$ represents the j layer of the ViT backbone. Given the set of resulting processed tokens after layer $3, 6, 9$ and 12 denoted as $\{\mathbf{t}_{i,3}^c, \mathbf{t}_{i,6}^c, \mathbf{t}_{i,9}^c, \mathbf{t}_{i,12}^c\}$ a linear projection layer f_{merge} merges all tokens originating from one image position (e.g. $\{\mathbf{t}_{i,j}^1, \mathbf{t}_{i,j}^2, \mathbf{t}_{i,j}^3, \mathbf{t}_{i,j}^4\}$ for any $j \in (3, 6, 9, 12)$) into the latent dimension d_{emb}. This step is referred to as the token-merger and is a necessary consequence of the approach of picking individual tokens for information from different channels within the same image region. This procedure results in a set of features $\{\mathbf{f}_i^3, \mathbf{f}_i^6, \mathbf{f}_i^9, \mathbf{f}_i^{12}\}$ with $i \in (1, HW/P^2)$ which carry the merged processed information from all channels. From here, the same approach as in [111] gets applied, and refereed to as *dense prediction transformer* (DPT) architecture, where intermediate features $\{\mathbf{f}_i^3, \mathbf{f}_i^6, \mathbf{f}_i^9, \mathbf{f}_i^{12}\}$ are reshaped into spatial feature maps and fused in a convolutional decoder. For reduced computational effort, all weights for the parameter encoding module, the positional encoding vectors, the patch embedding layers, as well as the first 8 transformer layers, will

be frozen for the fine-tuning step (for all runs with pre-trained encoder weights).

Fine-tuning takes place over 150 epochs with 5 warm-up epochs, a cosine decay learning rate scheduler (initial learning rate of 1×10^{-3}), an *AdamW* optimizer [112], a weight decay of 0.05 and processed with a batch size of 128. Since the class distribution of the remaining eight classes of the EWC included in $\mathcal{D}_{MSMS}^{EWC}$ over the validation area is highly imbalanced, batch sampling is applied to approach an equal distribution. For that, each sample in the batch is assigned a weight based on the inverse mean frequency of the pixel-wise class distribution within that sample. The class frequencies are computed per sample, averaged across all present classes, and the inverse of this mean is used as the corresponding sampling weight.

Zero-Shot Experiments For the zero-shot experiment, the models are trained to predict the LULC class given data of the sensor $S2_{1237}$. Subsequently, the experiments are conducted in a zero-shot manner with respect to a shift in data distribution by testing on data from the sensors with a deviating infrared channel $S2_{1234}$, a change in spatial resolution SD_{2468} (matching spectral position of the channels), as well as for the spectrally not aligned bands of the *SuperDove* sensor SD_{1357} (compare Fig. 5.1). During inference, the data as well as the corresponding sensor parameters (if parameter encoding is active) are provided to the model. Tab. 6.1 shows the results for the different encoders **E**, by varying the type of sensor parameter encoding strategy and the usage of the *spectral superposition augmentation* (SSA) augmentation module. To put the results into a broader perspective, Tab. 6.1 also contains two baselines, a vanilla *UNet* [113] as well as a version without pre-training of the *BaseMAE* architecture. Validation on the same sensor as during the fine-tuning step gives *intersection over union* (IoU) score of ≈ 0.70 for each of the models. However, conducting the validation in a zero-shot manner on the three different sensor arrangements ($S2_{1234}$, SD_{2468} and SD_{1357}) the IoU drops significantly (up to 0.21) for the non-pre-trained baseline as well as for the *BaseMAE* which does not encode sensor parameters during pre-training. Models that are pre-trained and tested with the parameter encoding module, lead to more robust embeddings, specifically when applying the module twice before the encoding and decoding step. Moreover, in both instances (one or two active parameter encoding modules), implementing the SSA yields an additional improvement in performance, highlighting the effectiveness of the proposed data augmentation strategy.

Fig. 6.3 visualizes the drop in segmentation performance. Since in Fig. 6.3a the models are validated on data from the sensor witch matches the fine-tuning distribution ($S2_{1237}$) no visually appearing significant differences can be found, Fig. 6.3b visualizes the validation of data of the synthetic sensor $S2_{1234}$. Here, a significant benefit of the pre-training procedure of the *SenPaMAE* framework can be observed as models without appropriate multi-sensor pre-training show unstable performance after the distribution shift.

Multi-Sensor Fine-tuning The second class of downstream experiments is given by fine-tuning the pre-trained models on multiple (synthetic) sensors. Two experiments

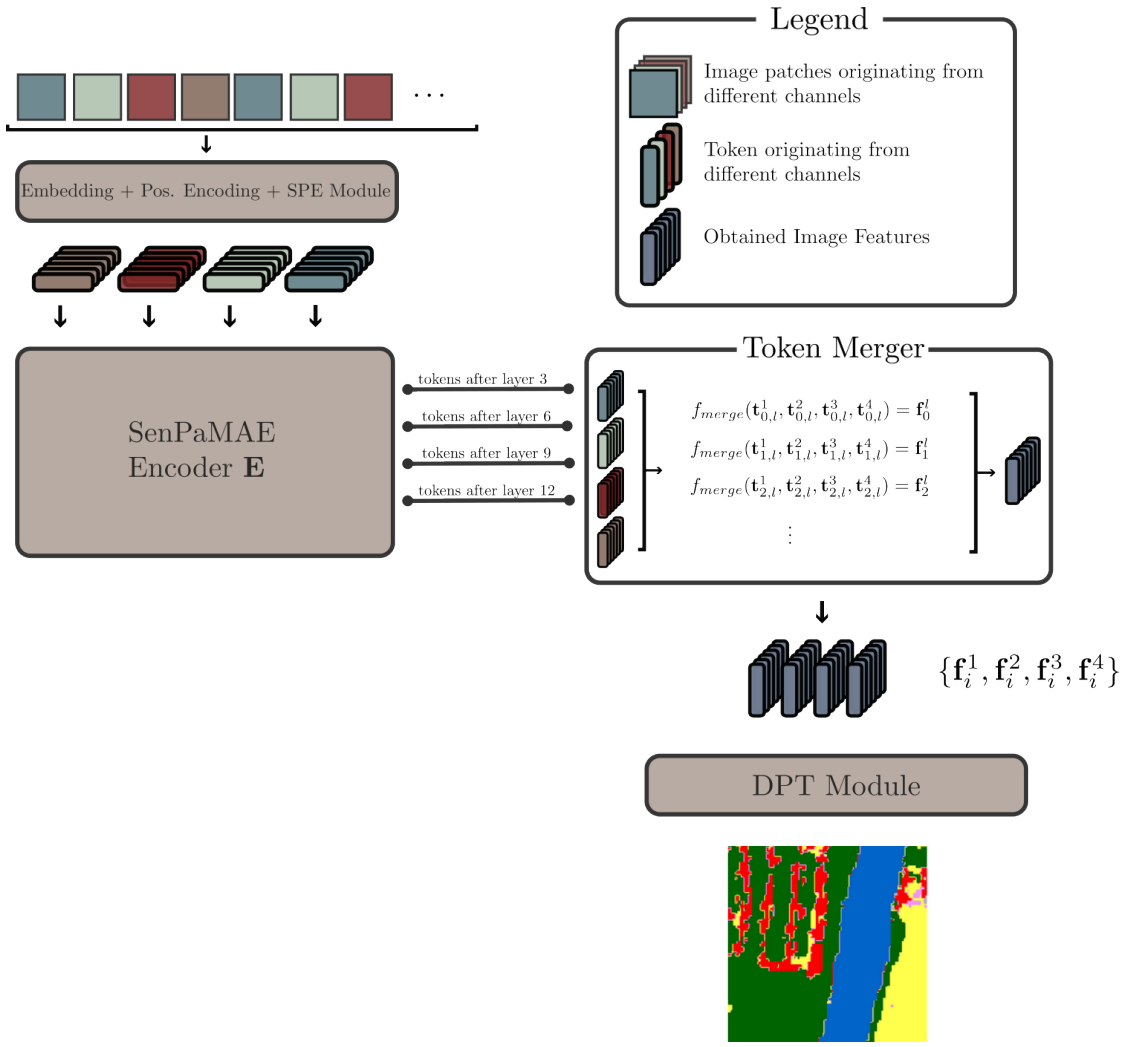


Figure 6.2: Illustration of the architectural adaptation for the backbones pre-trained within the *SenPaMAE* framework towards the downstream task evaluation. Tokens from multiple layers of the ViT backbone get processed by the token-merger module to obtain image features $\mathbf{f}^l_i$ (combining the merged processed information of all channels) which serve for the pixel wise classification task via a DPT [111] head.

Pre-Training				Micro IoU			
				Val. Set	Zero Shot		
Backbone	Pretrained	Sensor Parameter Encoding	SSA Module	$S2_{1237}$	$S2_{1234}$	SD_{2468}	SD_{1357}
UNet				0.69	0.39	0.47	0.21
BaseMAE				0.69	0.40	0.41	0.17
BaseMAE	✓			0.71	0.36	0.60	0.35
SenPa-MAE	✓	✓		0.68	0.67	0.58	0.41
SenPa-MAE	✓	✓	✓	0.70	0.68	0.64	0.46
SenPa-MAE	✓	✓✓		0.71	**0.69**	0.63	**0.52**
SenPa-MAE	✓	✓✓	✓	0.71	**0.69**	**0.65**	**0.52**

Table 6.1: Comparison of various pre-trained encoders for zero-shot evaluation. Following fine-tuning on $S2_{1237}$ to predict the LULC labels, the models are assessed in a zero-shot manner with respect to different sensors. The checkmark indicates if the sensor parameter encoding and augmentation module was active (two checkmarks for using the sensor parameter encoding also before the decoding step).

are conducted. First, the backbones and baseline models are fine-tuned on two alternating sensors, which serve as input. Secondly, synthetic input data is constructed by randomly choosing channels from either S2 or SD (equal probability), abbreviated with $S_{\text{random},}$ which serves as the input distribution for fine-tuning.

Tab. 6.2 shows the results for fine-tuning on alternating sensors $S2_{1237}$ and SD_{1357}. Here, validation scores are calculated individually on each of the two sensors. Even though the task represents the easiest possible scenario with respect to multi-sensor training – due to only two relatively similar sensors – an active SPE module leads to higher overall performance. In contrast to the results of Tab. 6.1 the SSA does only contribute in a positive manner for the case of applying the SPE once and the positive impact of this data augmentation strategy cannot be found for the case of applying the SPE twice.

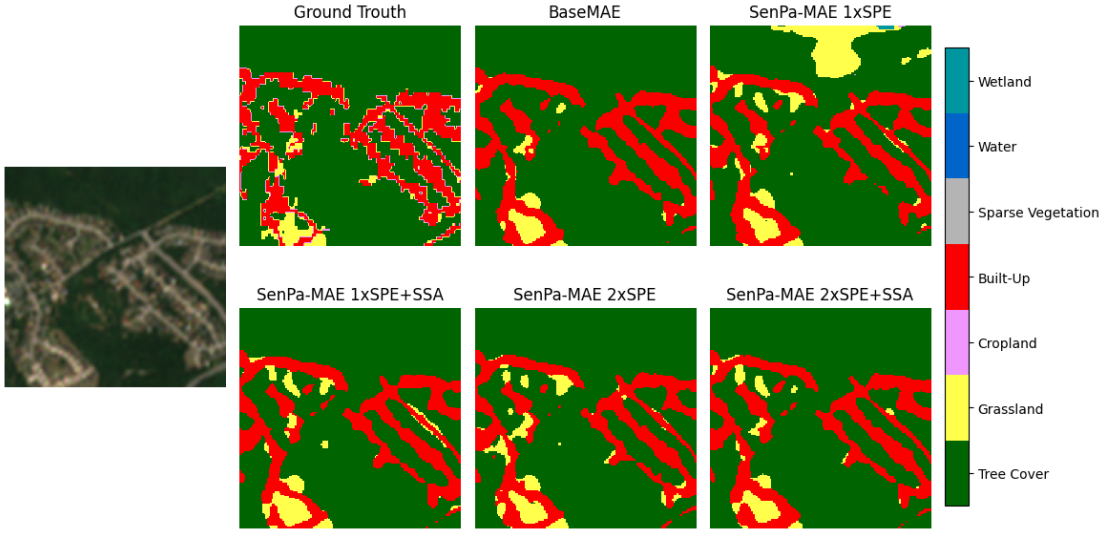
Tab. 6.3 shows the results for the more challenging task of fine-tuning on S_{random}. Here, evaluation scores are provided on a validation set of the form S_{random} as well as on three further deterministic synthetic sensor arrangements ($S2_{1234}$, SD_{2468}, and SD_{1357}). Similarly to the results in Tab. 6.2, the best performing setup is given by the *SenPaMAE* framework with the SPE applied twice. Additionally the same observation with respect to the influence of SSA as in Tab. 6.2 can be made, where data augmentation during the pre-training stage influences the fine-tuning positively for the case of applying SPE once but not for the case of applying SPE twice.

Pre-Training				Micro IoU	
Backbone	Pretrained	Sensor Parameter Encoding	SSA Module	$S2_{1237}$	SD_{1357}
UNet				0.71	0.70
BaseMAE	✓			0.72	**0.71**
SenPa-MAE	✓	✓		0.71	0.69
SenPa-MAE	✓	✓	✓	0.71	0.70
SenPa-MAE	✓	✓✓		**0.73**	**0.71**
SenPa-MAE	✓	✓✓	✓	0.72	0.70

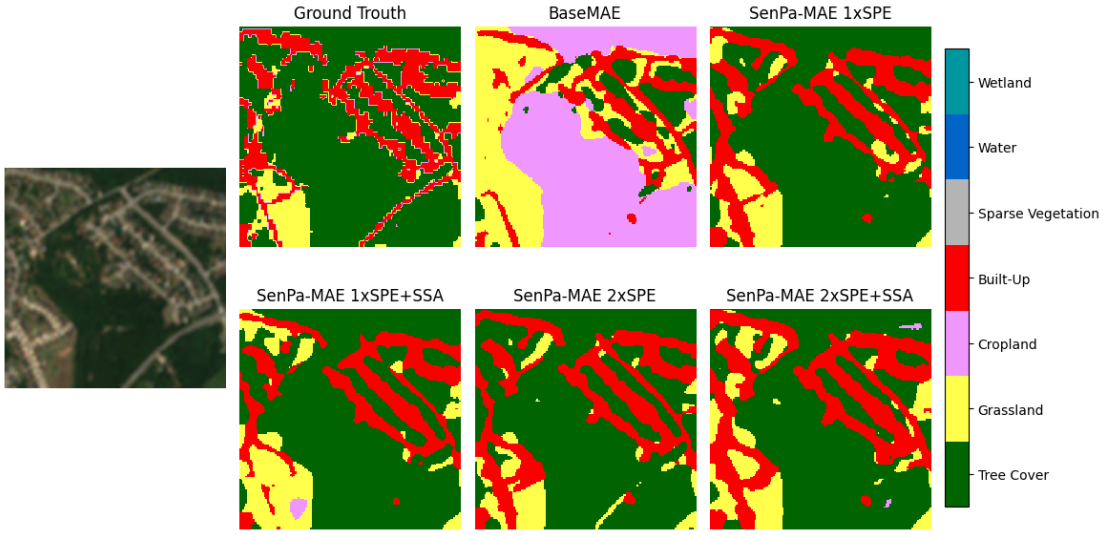
Table 6.2: Comparison of various pre-trained encoders for training on inputs from varying sensors. Input data for fine-tuning is given by alternating inputs of the form $S2_{1237}$ and SD_{1357}. Corresponding evaluation IoU-scores on the two sensors are given. The checkmark indicates if the sensor parameter encoding and augmentation module was active (two checkmarks for using the sensor parameter encoding also before the decoding step).

Pre-Training				Micro IoU			
Backbone	Pretrained	Sensor Parameter Encoding	SSA Module	S_{random}	$S2_{1237}$	SD_{2468}	SD_{1357}
UNet				0.69	0.70	0.72	0.70
BaseMAE	✓			0.71	0.71	0.73	0.72
SenPa-MAE	✓	✓		0.70	0.70	0.73	0.70
SenPa-MAE	✓	✓	✓	0.71	0.70	0.73	0.72
SenPa-MAE	✓	✓✓		**0.72**	**0.72**	**0.74**	**0.73**
SenPa-MAE	✓	✓✓	✓	0.71	0.71	0.73	0.72

Table 6.3: Comparison of various pre-trained encoders for training on inputs from varying sensors. Input data for fine-tuning is given by random channels from either S2 or SD sensor, denoted as S_{random}. Corresponding evaluation IoU-scores on S_{random}, $S2_{1237}$, SD_{2468} and SD_{1357} is given. The checkmark indicates if the sensor parameter encoding and augmentation module was active (two checkmarks for using the sensor parameter encoding also before the decoding step).



(a) Evaluated on data of the form $S2_{1237}$



(b) Evaluated on data of the form $S2_{1234}$

Figure 6.3: Visual comparison of the zero-shot evaluation results of Tab. 6.1. (Top) Output of all models evaluated on data of the form $S2_{1237}$ without any shift with respect to the fine-tuning data distribution. (Bottom) Results for evaluation with data of the form $S2_{1234}$ with a shift with respect to the fine-tuning data distribution. It can be observed that models without sensor parameter encoding undergo a significant drop in segmentation performance.

6.2 Evaluation of the SARFormer Framework

The following section presents the experimental validation of the *SARFormer* architecture, where the main objective lies in the integration of geometrical scene acquisition parameters and various pre-training strategies within the MAE setup. In contrast to the *SenPaMAE* experimental setup, where the objective is to integrate sensor parameters in a multi-satellite setting – meaning that single-sensor experiments are not necessarily expected to show additional performance gains – the *SARFormer* setup differs. Due to the incorporation of geometrical acquisition parameters, which are expected to positively impact geometrical scene understanding, performance improvements can be anticipated. In line with the previous section, pre-training and fine-tuning experimental settings are described, followed by the results of the corresponding experiments.

Pre-Training Pre-training of the *SARFormer* architecture is done with the MAE framework on the $\mathcal{D}_{TSX}^{PT}$ dataset, where the acquisition mode is encoded by the index $m \in (0, 1, 2, 3)$ for the four acquisition modes *SM*, *SL*, *HS*, and *ST* of the TerraSAR-X mission.

Two classes of models are pre-trained: Firstly, a backbone trained for processing one SAR view with corresponding geometrical scene acquisition parameters Φ_v (compare Section 4.2). Secondly, a setup processing two different views of the same geographical area, together with incorporating the two corresponding Φ_v. In both cases, the ViT-Large architecture [1], characterized by 24 transformer layers and 16 attention heads, serves as the encoder $\mathbf{E}$, as well as a shallower model with three layers for the decoder, such as proposed in [46]. In the two-view scenario, the three previously described masking strategies were employed to assess their impact on the network's learning capacity. Pre-training runs were done with 1,500 epochs with 100,000 balanced samples per epoch using the masked $\mathcal{L}_1$ loss function. Balancing is conducted with respect to discretized looking and azimuth angles as well as imaging modes, where samples are balanced with their inverse frequency. Notably, the single-view case benefits from a larger set of unique training samples (hence more diversity), as it is not constrained by the requirement of having a second view of the same area within $\mathcal{D}_{TSX}^{PT}$. Still, the overall number of back-propagation iterations is constant due to the employed batch sampling. Similarly to the pre-training of *SenPaMAE*, the *AdamW* optimizer with 3 warm-up epochs, a cosine decay learning rate scheduler (initial learning rate of 1×10^{-4}) and a weight decay of 0.05 was utilized while processing 128 samples per batch.

Note: In the single view SAR case none of the other masking stategies beside *random* masking can be employed.

Adaptation to Downstream Tasks The evaluation of the *SARFormer* approach is built around solving three tasks that can be derived from $\mathcal{D}_{TSX}^{Buildings}$ in a multi-task fashion, namely predicting the pixel-wise building footprint as a classification problem, as well as predicting height in slant as well as in map geometry. The primary motivation behind a multi-task approach is often to enhance individual performances due to a more robust representation learned by solving multiple tasks simultaneously [114]. Still,

within this work, the primary motivation is to test the derived representation – when applying the proposed modifications introduced in Section 4.2 – across multiple tasks in a computationally efficient manner, since the number of corresponding downstream experiments is reduced by a factor of three.

In order to predict pixel-wise quantities, similarly to the *SenPaMAE* evaluation, the DPT [111] framework, which collects and merges information from multiple layers of $\mathbf{E}$ gets utilized. Tokens from the encoder get extracted after layers $5, 12, 18$ and 25 ($N_v \cdot N_p^2 + N_v$ tokens per layer). In the following, the extracted tokens after layer l get denoted as $\mathbf{t}_{i,l}^v$ and the *metatoken* after layer l as $\mathbf{t}_{meta,l}^v$. All tokens originating from the same spatial position i as well as the *metatokens* are merged with a linear fully connected layer followed by a GELU activation – denoted as f_{merge} – to obtain one feature $\mathbf{f}_i^l$ corresponding to the patch position i for a given layer l:

$$\begin{aligned} &f_{merge} : \mathbb{R}^{2 \cdot N_v \cdot d_{emb}} \rightarrow \mathbb{R}^{d_{emb}}, \\ &\mathbf{f}_i^l = f_{merge}(\mathbf{t}_{i,l}^1, \mathbf{t}_{i,l}^2, \ldots, \mathbf{t}_{meta,l}^1, \mathbf{t}_{meta,l}^2, \ldots) \end{aligned} \tag{6.1}$$

which carries the information merged from all available views. Hence, $N_p \cdot N_p$ features of dimension $\mathbb{R}^{d_{emb}}$ are obtained for each extraction layer l. From here, the DPT approach is utilized and extended into a multi-task output by the extension of three separate convolutional blocks composed of five convolutional layers (and corresponding activation) each. The three separate final convolutional blocks predict the height labels – in map geometry and slant geometry – as a regression task, as well as the building footprint mask for the data in $\mathcal{D}_{TSX}^{Buildings}$ as a classification problem. Similar to the pre-training, batch sampling with respect to the imaging parameters as well as the building density for the fine-tuning task is applied. Additionally, each epoch includes 25% of samples without corresponding height data, allowing the model to encounter all locations, even those lacking height labels. For these specific patches, the height estimation is excluded from the height loss calculation.

The training protocol consisted of 250 epochs using the *AdamW* optimizer, with a half-cycle cosine decay learning rate scheduler preceded by a three-epoch warm-up period. Each training epoch incorporated 100,000 randomly chosen balanced samples. The loss function for the regression task (height estimation) is given by a composite metric, combining asymmetric L1 loss, gradient loss, and normal loss [115]. For the segmentation task, binary cross-entropy loss is used. The total loss function is formulated as a weighted sum of the individual task-specific losses.

Fig. 6.4 gives a graphical overview of the architectural modifications for the fine-tuning stage.

Evaluation Procedure In contrast to previously presented evaluation procedures, the evaluation set of $\mathcal{D}_{TSX}^{Buildings}$ contains a significant number of variables that influence the overall results. For example, in a dual-view scenario, it matters greatly whether the two views are collected from different or identical orbit directions (ascending/descending).

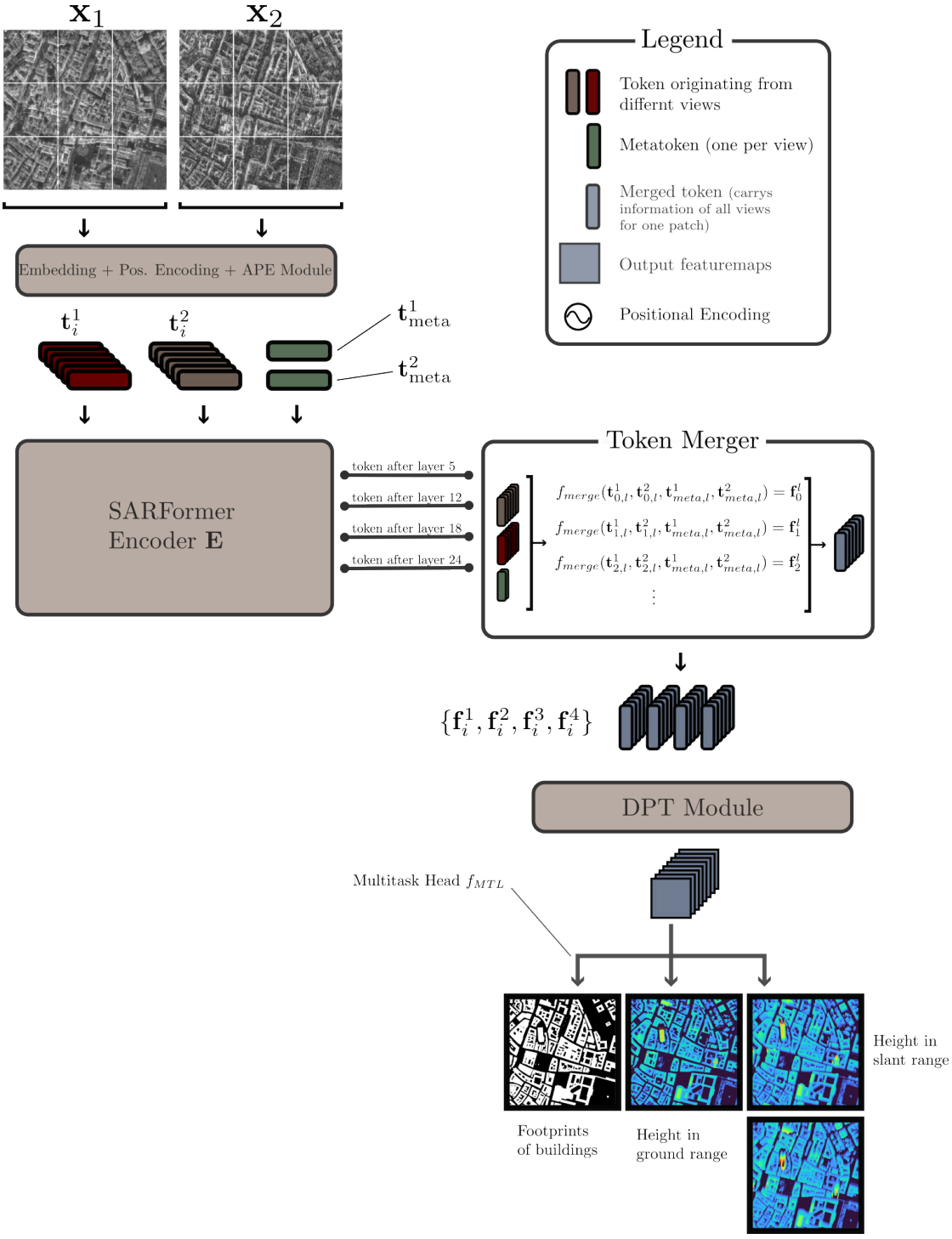


Figure 6.4: Illustration of the architectural adaptation for the backbones pre-trained within the *SARFormer* framework towards the downstream task evaluation. Tokens from multiple layers of the ViT backbone as well as the metatoken get processed by the token-merger module to obtain image features $\mathbf{f}_{i,l}$ (combining the merged processed information of all views) which serve for the pixel wise classification and regression tasks via a DPT [111] head and an adapted multitask head.

Furthermore, the combination of imaging modes also has a major influence on the overall reconstruction outcome. To account for these differences, a representative and diverse subset of images from the validation set of $\mathcal{D}_{TSX}^{Buildings}$ is chosen to enable comparisons across views. The images are selected to avoid unfair comparisons, such as comparing single-view stripmap results with dual-view results from higher-resolution imagery, even though this cannot be entirely avoided due to the intrinsic advantage of multi-view models. Overall results are obtained by averaging across six different image compositions over the two evaluation areas.

Effect of APE To evaluate the impact of the proposed *acquisition parameter encoding* (APE) module, initially, experiments are conducted with random weight initialization. In Tab. 6.4, a comparison of *SARFormer* models with the APE module both enabled and disabled, alongside a baseline model (classical UNet [113] with the same multi-task extension as described above) as a function of the number of input views N_v is presented. For the classification task the results are measured with the IoU metric as well as the *overall accuracy* (OA) (both with macro averaging) and for the regression task performance is measured with the *mean absolute error* (MAEr), *root mean squared error* (RSME) as well as *structural similarity index* (SSIM). Examining the effect of the APE module, it can be observed that it consistently improves (or in rare cases holds) performance across all metrics. Even though significant performance differences can't be found on the binary classification task of segmenting building footprints, the regression task of height reconstruction benefits significantly for metrics evaluated in slant as well as in map geometry.

			Classification Footprints		**Regression Height**			**Regression Height (Slant)**		
Number of Views	Model	APE	IoU	OA	MAEr	RMSE	SSIM	MAEr	RMSE	SSIM
1	UNet MTL		**0.68**	**0.90**	5.67	8.55	**0.82**	5.39	7.80	**0.84**
	SARFormer		0.67	0.89	5.45	7.89	**0.82**	5.35	7.70	**0.84**
	SARFormer	✓	0.67	0.89	**5.24**	**7.61**	**0.82**	**5.29**	**7.60**	**0.84**
2	UNet MTL		**0.74**	**0.92**	4.77	7.35	0.84	**4.64**	**6.89**	**0.88**
	SARFormer		0.73	**0.92**	4.61	6.87	0.85	4.77	6.99	**0.88**
	SARFormer	✓	**0.74**	**0.92**	**4.39**	**6.60**	**0.85**	4.69	6.90	**0.88**
3	UNet MTL		**0.74**	**0.92**	4.95	7.57	0.85	4.85	7.11	0.88
	SARFormer		**0.74**	**0.92**	4.63	6.90	0.85	4.88	7.15	0.88
	SARFormer	✓	**0.74**	**0.92**	**4.30**	**6.39**	**0.86**	**4.43**	**6.52**	**0.89**

Table 6.4: Numerical results of the fully-supervised experiments (no pre-training). The best values are evaluated within a specific number of views and printed in bold.

As described above, the metrics reported in Tab. 6.4 (as well as in the following tables) are mean performance over various acquisition settings to minimize effects that occur only in certain imaging modes or at certain looking- or azimuth angles. Fig. 6.5 shows the twelve separate contributions of the RSME value for the case of dual-view *SAR-*

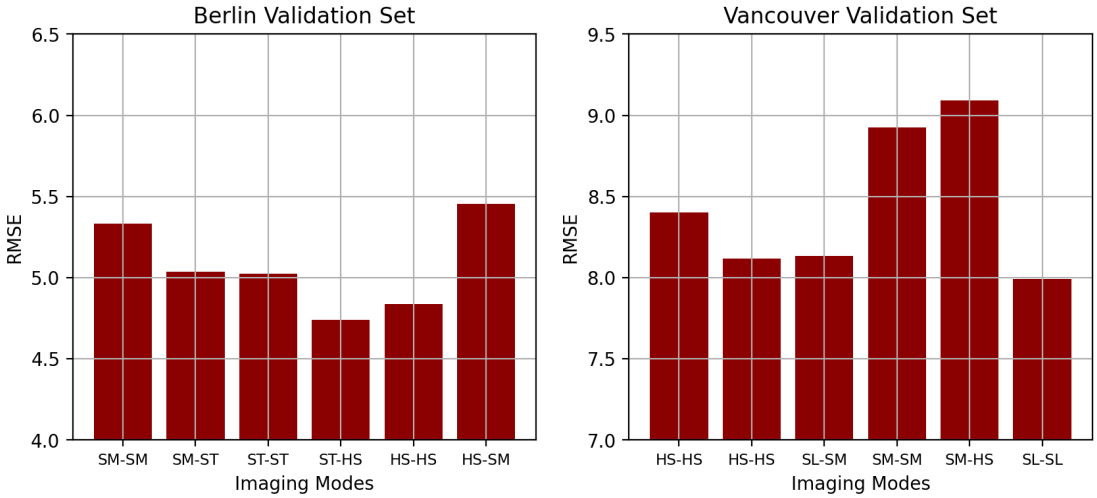


Figure 6.5: Comparison of the performance on different imaging mode pairs – *stripmap* (SM), *high-resolution spotlight* (HS), *spotlight* (SL) and *staring spotlight* (ST) – of the *SARFormer* model with active APE separated in the two test regions (Berlin and Vancouver) of the $\mathcal{D}_{TSX}^{Buildings}$.

Former with active APE as a function of the corresponding imaging modes of the two acquisitions. Here, it can be observed that the performance varies significantly across different pairs of imaging modes. To some extent, the performance correlates with the resolution obtained in the corresponding mode. e.g. pairs of two *stripmap* (SM) images (or pairs with one of the two images being a SM image), perform worse than the counterparts with images of higher resolution. Nevertheless, this trend is not fully consistent since, e.g. the best performing pair over Vancouver is given by two images with *spotlight* (SL) mode, which does not correspond to the pair of images with the highest resolution in the evaluation set. This indicates that the other quantities, such as the looking and azimuth angle, as well as the date of the acquisition (seasonal effects of vegetation), additionally influence the overall result.

Effect of Pre-training Strategies Tab. 6.5 presents the results for models trained with the three proposed MAE-based pre-training strategies. Experiments were limited to configurations with one or two SAR views, while the previously presented experiments done with random weight initialization also included the three-view case. Two scenarios for each view number are tested. Firstly, fine-tuning all model weights (the ViT encoder as well as the CNN based DPT extension) and secondly, only fine-tuning the last 1/3 of the ViT layers together with the DPT extension while keeping the remainder of the ViT weights frozen. The results indicate that the proposed *preserving* masking strategy achieves superior performance across most metrics, suggesting its suitability for non-object-centered remote sensing imagery. In contrast, the *blind-channel* strategy, representing an extreme case of *preserving*, underperformed with respect to the *preserving* masking strategy. All pre-trained and fine-tuned models in Tab. 6.5 utilized the APE module, enabling comparison with training from scratch in Tab. 6.4. In the

# Views	Frozen Weights	Masking // Type	Classification Footprints		Regression Height			Regression Height (Slant)		
			mIoU	OA	MAEr	RMSE	SSIM	MAEr	RMSE	SSIM
1	❄	Random	0.64	0.88	5.84	8.37	0.81	5.96	8.46	0.83
		Random	0.69	0.90	5.07	7.40	0.83	4.97	7.20	0.84
2	❄	Random	0.71	**0.91**	**4.49**	**6.59**	**0.85**	**4.36**	**6.41**	**0.88**
		Blind Channel	0.71	**0.91**	4.61	6.89	**0.85**	4.74	6.99	0.87
		Preserving	**0.72**	**0.91**	4.65	6.85	**0.85**	4.75	6.93	**0.88**
		Random	**0.75**	**0.92**	4.19	6.29	0.85	4.07	6.12	**0.89**
		Blind Channel	0.73	**0.92**	4.38	6.56	**0.86**	4.53	6.72	0.88
		Preserving	0.74	**0.92**	**4.12**	**6.13**	**0.86**	**3.96**	**5.90**	**0.89**

Table 6.5: Achieved metrics through pre-training and subsequent fine-tuning. Best values are evaluated according to the used pre-training method (masking strategy). The single-view case allows only for the random masking strategy. For the frozen setting noted with ❄, 2/3 of the backbone's layers were not updated during fine-tuning.

optimal configuration (*preserving* pre-training with full fine-tuning), a 2-point gain in IoU and a 4-6% improvement in MAE can be observed, compared to the from-scratch baselines.

These improvements are even more pronounced when training data is limited (as this represents one of the strongest points of self-supervised pre-training), as label scarcity makes it partially unfeasible to train larger model backbones without overfitting. To show the effects of using a pre-trained backbone, the *SARFormer* is fine-tuned using only two labeled SAR images from Paris, introducing significant domain shifts with respect to the test area of $\mathcal{D}_{TSX}^{Buildings}$. Tab. 6.6 shows the performance gains of the pre-training approach with respect to training either the *SARFormer*, or the UNet type baseline from scratch. Here, the significance of the pre-training step – which is even more pronounced that the results shown in Tab. 6.5 – which give up to 8-points improvement in IoU and a reduction of 1.56 m for MAEr can be observed.

Model	**pre-trained**	**mIoU**	**MAEr** (map)	**MAEr** (slant)
UNet MTL		0.51	7.61	7.84
SARFormer		0.54	7.22	8.51
SARFormer	✓	**0.62**	**5.66**	**6.15**

Table 6.6: Metrics achieved on the test set using a very limited labeled dataset for training given by only stripmap acquisitions over Paris. The pre-training procedure significantly benefits those cases with label scarcity.

Qualitative Comparison Visually, the differences in performance across the models are more pronounced than the numerical error metrics alone suggest. Fig. 6.6 pro-

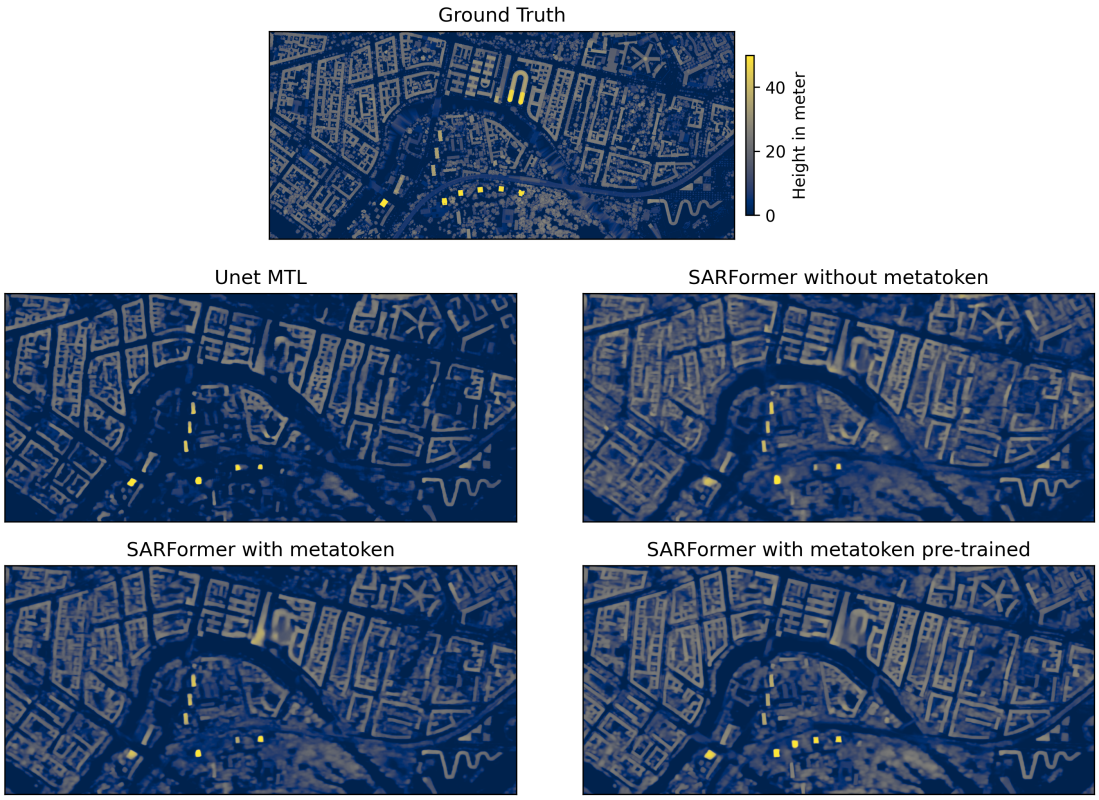


Figure 6.6: Visual examples for the results for predicting the height in map geometry (still obtained in a multi-task fashion). From top to bottom: Ground truth, UNet baseline, *SARFormer* without metatoken, *SARFormer* with metatoken as well as a pre-trained variant of the latter.

vides a side-by-side comparison of the outputs of four different configurations namely, the UNet type baseline, the *SARFormer* trained from scratch with and without the APE module active, as well as a pre-trained version of the *SARFormer* (where APE module is always active), alongside the ground truth information. In this example, especially the correct identification and reconstruction of the high-rise buildings is found to benefit significantly from both the APE module as well as preceding pre-training.

6.3 Evaluation of the IaI-SimCLR Framework

The following presents the experimental validation of the enhancement of quality of derived representation achieved with the *IaI-SimCLR* architecture with respect to previously published multi-modal contrastive methods (compare Chapter 3). In line with the previous sections, pre-training and fine-tuning experimental settings are described, followed by the results of the corresponding experiments.

Type	Probability	Applied to
Crop	100%	S1 and S2
Flip	50%	S1 and S2
Grey-scale	10%	S2
Color Augmentation	80%	None or S2
Gaussian Blur	30%	S1 and S2

Table 6.7: The augmentations comprising the set $\mathcal{A}$, along with their respective probabilities of being applied.

Pre-training Pre-training of the encoders $\mathbf{E}^{S1}$ and $\mathbf{E}^{S2}$ is performed on multi-modal imagery on the $\mathcal{D}_{MM}^{PT}$ dataset described in Section 5.2.1. All operations are performed on patches of the size $W = H = 256$ using a *ResNet18* [116] architecture for both encoders $\mathbf{E}^{S1}$ and $\mathbf{E}^{S2}$ as well as two linear layers with respective activation for the four projection heads $\mathbf{f}_{S1\,intra}$, $\mathbf{f}_{S2\,intra}$, $\mathbf{f}_{S1\,inter}$ and $\mathbf{f}_{S2\,inter}$. The feature size for the latent space after the encoding step is $\mathbf{h}_{S1}$, $\mathbf{h}_{S2} \in \mathbb{R}^{512}$ (equal dimension for augmented views), before the loss will be calculated on the reduced latent space $\mathbf{z}_{S1}$, $\mathbf{z}_{S2} \in \mathbb{R}^{128}$. For evaluation purposes, mainly the linear separability of features on the concatenated latent space $\mathbf{h}_{S1+S2} = \mathbf{h}_{S1} \oplus \mathbf{h}_{S2}$ is tested where $\oplus$ represents concatenation operation, even though each of the encoders can be applied separately to the corresponding modalities when testing on downstream task where only one of the two modalities is available (such as $\mathcal{D}_{MM}^{Crop}$). Pre-training is conducted with all meaningful combinations of the loss functions as described in Section 4.3 in order to study the effect of forcing similarity of the gained representations on multiple levels. Beyond this, models with an intra-modality loss component are trained in the two modes, with and without the color augmentation (in the following marked by nCA for **n**o **C**olor **A**ugmentation), all done with the two strategies of *local batch sampling* (LBS) and *random batch sampling* (RBS). This overall leads to $(2 \times 2 + 1) \times 2 = 10$ different models evaluated in this section. In the case of active augmentation modules, Tab. 6.7 provides a list of all applied augmentations along with their corresponding target modality.

All training has been carried out with the *Adam optimizer* (initial learning rate of 10^{-4}), a batch size of 128 images over 100 epochs, since no significant benefit could be found for longer pre-training.

Fine-tuning All ten encoder pairs are fine-tuned and evaluated based on the quality of the representations $\mathbf{h}_{S1}$ and $\mathbf{h}_{S2}$ with respect to the three downstream datasets: $\mathcal{D}_{MM}^{EWC}$, $\mathcal{D}_{MM}^{DFC}$, and $\mathcal{D}_{MM}^{Crop}$. Due to the large number of experimental settings, the number of samples in $\mathcal{D}_{MM}^{EWC}$ and $\mathcal{D}_{MM}^{Crop}$ is limited to 12,000 for training and 4,000 for validation to ensure computational feasibility. In contrast to the other two downstream tasks, the dataset $\mathcal{D}_{MM}^{Crop}$ covers only images from the optical modality, and therefore only tests the capability of $\mathbf{E}^{S2}$. To address class imbalance during fine-tuning, random weighted sampling based on the inverse frequency of class occurrences is performed. Specifically, each training sample is assigned a sampling weight inversely proportional to the frequency of

its associated class label. All three above-mentioned downstream tasks use *Cross Entropy Loss*, optimized with the Adam optimizer with an initial learning rate of 10^{-3}, a batch size of 32 samples, and training over 200 epochs. For the regression downstream task on the $\mathcal{D}_{MM}^{Biomass}$ dataset, linear regression is used to solve the task in closed form, with the motivation of compute efficient evaluation.

Comparison of Loss Function and Sampling Strategy To better understand the impact of the two proposed adaptations on performance across the four downstream datasets, all metrics are presented in two formats in Fig. 6.7 and Fig. 6.8. Fig. 6.7 shows the results divided into LBS and RBS and the different downstream tasks as a function of the pre-training epoch. This plot reveals the dynamics of the pre-training. Fig. 6.8 in contrast displays the best overall performance along all checkpoints as well as the mean performance calculated over each loss function setting.

Firstly, Fig. 6.7 (left column) shows that the representation obtained with *random batch sampling* (RBS) tends to lose descriptive quality with prolonged training, a pattern observed consistently across all models. Most pronounced is the case for *DualSimCLR* where in the case of RBS the model drastically drops in performance after ≈ 40 epochs along all four downstream tasks. This dynamics is overall not present for the case of representation obtained via the *local batch sampling* (LBS). Comparing the overall performance – independent of the learning dynamics – Fig. 6.8 shows the best evaluation score achieved across all checkpoints for the four downstream tasks. Round markers represent results using RBS, while triangular markers indicate results obtained with LBS. For the two approaches that include an intra-similarity loss component and have the augmentation module active, results are further divided into those with color augmentation (lighter shades) and those without (darker shades). Across the tasks, several consistent trends can be observed. In the *DualSimCLR* setup, RBS performs better than LBS across all four tasks. For the model using only intra-similarity loss (*Intra-SimCLR*), RBS generally performs better when color augmentation is active, a trend that does not appear when color augmentation is disabled. A similar pattern is evident in most experiments using the third loss setup, which combines intra- and inter-modality similarity (*IaI-SimCLR*).

Additional Representation Quality Metrics In order to broaden the understanding with respect to declining performance for longer pre-training, Fig. 6.9 shows additional metrics on the derived representations of *DualSimCLR* (the most drastic case of potential mode collapse) as a function of different pre-training epochs. For a given batch of 256 images (taken from the $\mathcal{D}_{MM}^{DFC}$ downstream dataset) two metrics are calculated on the corresponding embeddings $\mathbf{h}_{S1}, \mathbf{h}_{S2} \in \mathbb{R}^{512}$. First, the mean pair-wise cosine similarity is obtained by taking all pairs of images of the given batch into account and averaging the corresponding cosine similarity. Fig. 6.9 shows (for both, the embeddings of Sentinel-1 and Sentinel-2) that embeddings obtained with RBS are collapsing to an almost identical representation (the value one indicates similar embeddings), whereas the mean pair-wise cosine similarity for the LBS case is decreasing over the training.

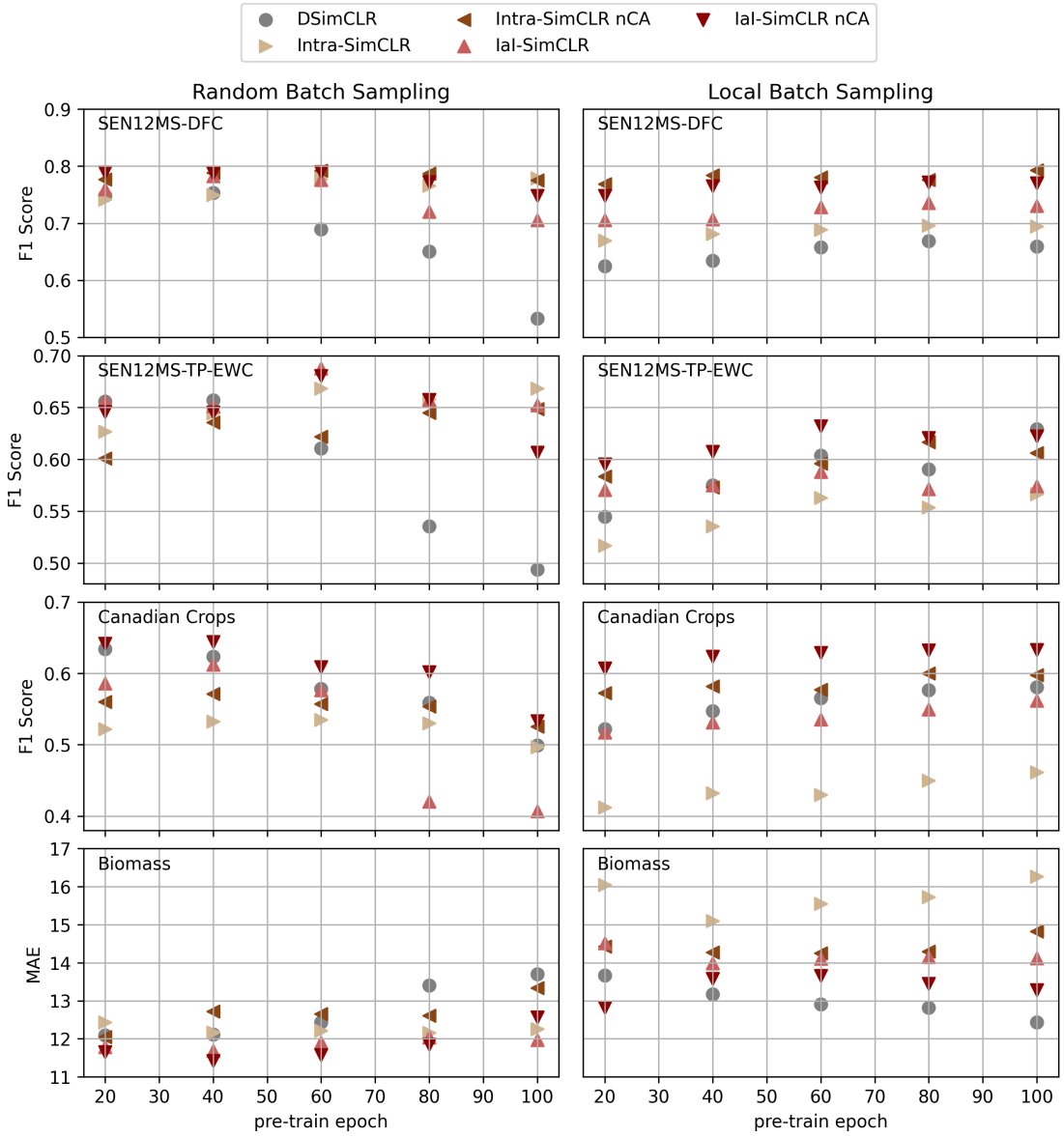


Figure 6.7: The performance of the models *IaI-SimCLR* with *DualSimCLR* and *Intra-SimCLR* with active- and non-active (nCA) color augmentation a_{color} over the four downstream tasks (rows) as a function of the pre-training epoch. Models are subdivided further into trained with random batch sampling (left column) and local batch sampling (right column).

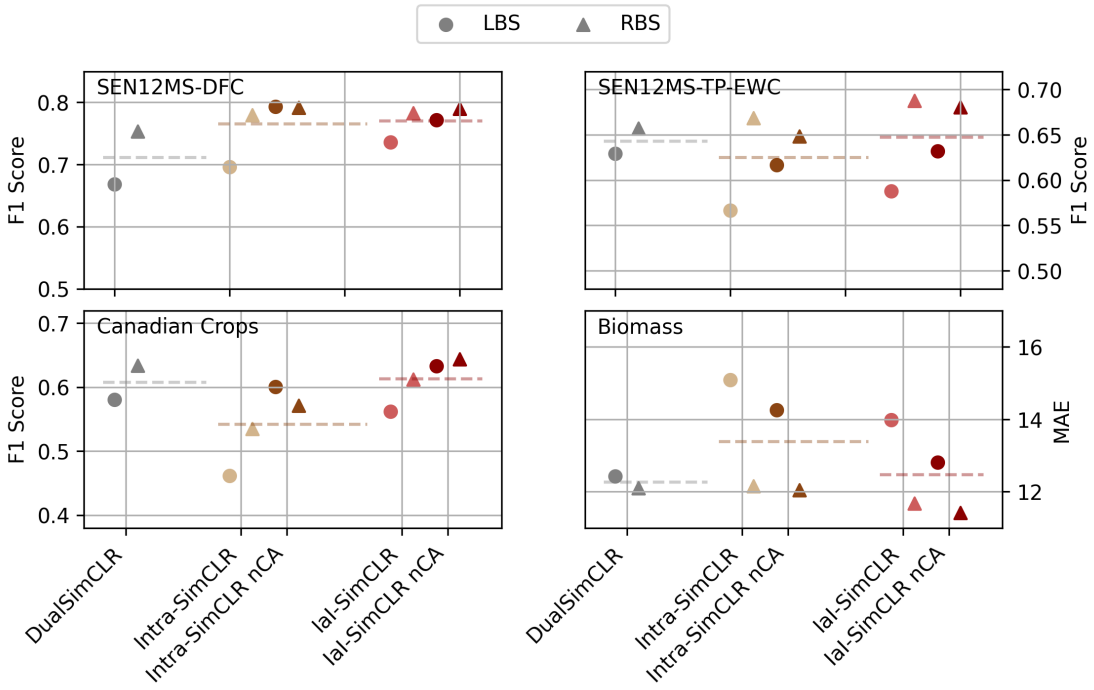


Figure 6.8: The performance of the models *IaI-SimCLR* with *DualSimCLR* and *Intra-SimCLR* with active- and non-active (nCA) color augmentation a_{color} over four different downstream tasks (rows), subdivided into trained with random batch sampling (triangles) and local batch sampling (circles).

Secondly, in a similar manner, the effective rank of the representations is investigated (once again as a function of the pre-training epoch) and displayed in the two bottom plots of Fig. 6.9. The effective rank[2] describes the magnitude of how much the representations are spread out along each dimension of the vector space. A high spread indicates a lot of capacity to encode information, whereas a low effective rank indicates that the model only encodes information along distinct directions. Similar to the mean pair-wise cosine similarly one can observe decreasing quality for RBS and increasing representation quality for the case of LBS.

[2] The effective rank is calculated by taking the batch of representations as a matrix and calculating the singular value decomposition. From there, one calculates the entropy of the normalized singular values, whereas the final effective rank is given by the exponential of the entropy.

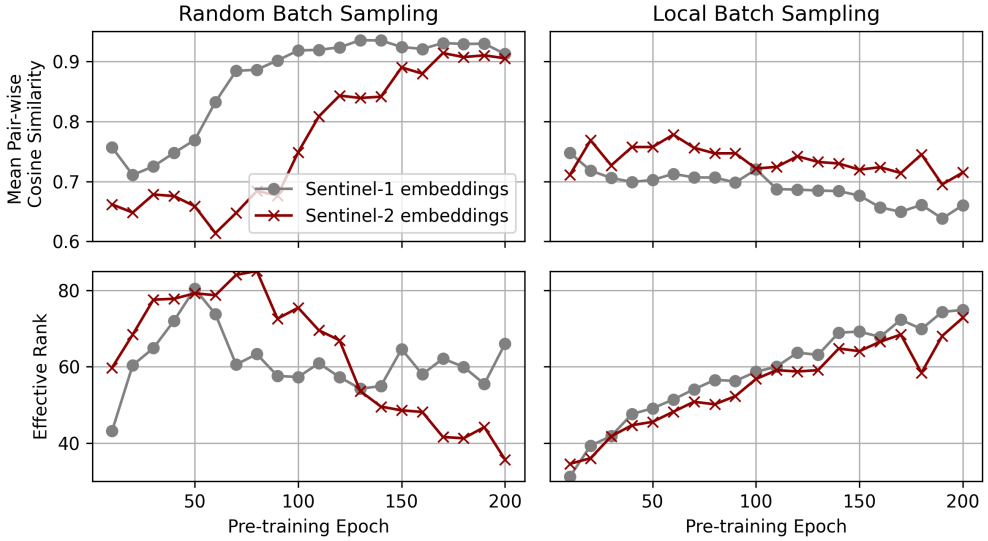


Figure 6.9: The mean pair-wise cosine similarity as well as the effective rank for the representations obtained by *DualSimCLR* as a function of the batch sampling strategy as well as the pre-training epoch.

7 Discussion

In this chapter, the results of the three Sections 6.1 to 6.3 are discussed, and relations between the three methods are established. Additionally, where applicable, insights gained from the experimental results regarding potential extensions of the methods are provided.

SenPaMAE The first of the three methods proposed in this thesis is given by the *SenPaMAE* framework, which allows for the incorporation of sensor parameters of multi-spectral satellites into the machine learning workflow and hence presents a step in the direction of multi-sensor training and inference. The experimental validation of the *SenPaMAE* framework (compare Section 6.1) is divided into two classes of experiments namely, zero-shot validation with respect to performing inference on a different sensor – compared to the (single) sensor used during fine-tuning– and secondly, the advancement of the proposed framework when training is conduced in a multi-sensor manner.

As previously shown, Tab. 6.1 presents the results of the zero-shot experiments, highlighting the gain in performance of the proposed *sensor parameter encoding* (SPE) module when performing inference on an unseen sensor (unseen with respect to sensors used during fine-tuning, not necessarily during the pre-training stage). Even a model that also got exposed to all sensors during pre-training (like the pre-trained *BaseMAE* in Tab. 6.1) shows a significant drop in performance during the zero-shot experiments compared to one trained with incorporated sensor parameters (active SPE). This suggests that multi-sensor pre-training conducted without incorporating the specific sensor parameters leads to an ill-posed learning setup with respect to the conducted downstream experiments. Further, the results for multi-sensor fine-tuning presented in Tabs. 6.2 and 6.3 indicate that the benefits of the proposed SPE module and the corresponding data augmentation

strategy SSA, while present, are somewhat less pronounced in the zero-shot evaluation setting. One observation that can be made here is that the corresponding downstream task of LULC mapping is a problem which does not necessarily (strongly) depend on the spectral characteristics of the signal, especially on the relatively high spatial resolution of 5 m to 10 m. It would be suggested for further experimental setups to additionally include vegetation-oriented tasks, such as crop- or tree-type mapping or soil and water parameter regression tasks.

It should also be mentioned that the framework could significantly benefit from the incorporation of hyper-spectral data, which would provide more spectral diversity. Even though within the current state of the framework, where each channel gets assigned to a separate token, directly inputting hyper-spectral data would be highly (computationally) inefficient. This – as described earlier – stems from the fact that the computational cost of each layer scales quadratically with the number of tokens, and generating one token per patch position and band (in the case of hyper-spectral data, hundreds of bands) would lead to a highly inefficient network. Nevertheless, with the use of the *spectral superposition augmentation* (SSA) module, arbitrary types of response functions could be generated through superposition of hyper-spectral bands and hence, highly diversify the spectral aspect of the pre-training dataset. This would greatly add to the current setup, where the spectral diversity is limited to the set of all superpositions of the channels available within the three sensors Sentinel-2, Superdove, and Landsat. In this way, hyper-spectral data could be incorporated while the model would still never be exposed to an image with the – as for hyper-spectral data typically – high number of bands individually.

While the choice of assigning separate tokens to each channel represents a hurdle for some applications, it has been chosen for the evaluation of the first version of the *SenPaMAE* framework due to the possibility of simply encoding tokens with the sensor parameters and adapting the masking function for the MAE pre-training. For future works, an alternative setup, closer to the *SARFormer* framework, is conceivable where all spectral information gets assigned to one token and an additional overall metatoken which carries the embedded information of all corresponding spectral response functions. This – provided it would lead to so similarly convincing results – would cancel the computational downside of the current approach. While this might be mistaken for a simple engineering detail, it would have larger consequences since tasks like processing time-series or doing inference with multiple input-sensors (simultaneously) would become viable. Additionally, with suitable encoding of the response functions, a metatoken incorporation of hyper-spectral data in its raw form would become feasible and would present an alternative to the above-discussed utilization of generating more diverse synthetic signals via the SSA module.

SARFormer The second main contribution, the *SARFormer* framework, allows for the incorporation of geometrical scene acquisition parameters as well as the corresponding imaging mode for high resolution SAR measurements, which highly influence the resulting measurement and hence are crucial for the interpretation of the corresponding

images.

During the evaluation of the *SARFormer* framework (compare Section 6.2) it could be shown that the proposed modification, such as the *acquisition parameter encoding* (APE) module and the pre-training strategy, significantly enhances the scene understanding capabilities of the models. In Tab. 6.4, the results for the evaluation of height reconstruction as well as segmentation of building footprints are presented for models trained from scratch. Independent of the number of input views – even though better results can be gained with a higher number of input views – the proposed architectural modifications underlying the *SARFormer* framework lead to an increase of performance in the majority of evaluation metrics. Further improvements to all models trained from scratch were made by preceding pre-training of the models, as the results in Tab. 6.5 and – even more pronounced – in Tab. 6.6 suggest. Here, where pre-training was conducted with the *masked auto-encoder* (MAE) strategy it could be shown that the exact masking configuration has significant impact and it is desirable to avoid masked areas of large extend, potentially due to the reason that reconstruction becomes infeasible. Even though the three tested masking configurations already provide insights on how to design the masking function for the task of pre-training on high resolution SAR imagery, many more configurations – especially for the multi-view case – and hyper-parameters such as masking ratio could be tested for obtaining optimal results.

Even though the model was evaluated on three distinct tasks, the evaluation would also benefit from further downstream applications. Without any empirical proof that the following argument is applicable to the results presented in Section 6.2, for any downstream task (especially in the context of moderate resolution EO data) it is always possible that the aleatoric uncertainty [117] prevents further improvements though advances in the architectural design of the model. This is intuitively hard to judge since the performance – especially when analyzing SAR data – is far beyond human capabilities. Without knowing the underlying reason, a saturation effect seems to occur in the context of the building footprint classification task, even though it can't be ruled out that further modifications still enhance the accuracy. Hence, further studies with additional downstream tasks such as object detection or semantic segmentation tasks (e.g. LULC or flood extent mapping) would allow for a deeper understanding of the effect of the *acquisition parameter awareness* introduced by the APE module.

Beyond this, future iterations of the *SARFormer* approach could potentially benefit from larger pre-training datasets. Due to data access limitations, $\mathcal{D}_{TSX}^{PT}$ and $\mathcal{D}_{TSX}^{Buildings}$ include incidental imagery. As a result, the diversity of data to which the model is exposed is not increased, as is typically the case in a standard pre-training procedure. Still, as results in Tab. 6.5 suggest, even the pre-training on the data underlying the labeled dataset yields performance increases.

Further learnings can be derived from the evaluation procedure of the *SARFormer* in Section 6.2. Within this thesis, neither *SenPaMAE* nor *SARFormer* is adapted to the processing of variable sequence length, even though the overall architectural choice would allow for such. Incorporating the engineering challenges would allow for a more

general model (instead of three distinct pre-trained backbones for processing single, dual, or triplet view inputs) and significantly reduce the amount of training and hence simplifying the evaluation process. This could free up resources that can be dedicated to investigating well-chosen hyperparameters for the masking step, as discussed above, since only the configurations presented in this thesis were tested due to computational constraints.

IaI-SimCLR For making use of the multi-modal character of available EO data for pre-training, this thesis proposes an adaptation to previously existing approaches, named *IaI-SimCLR* (compare Section 4.3). Within the experimental section regarding the results of the *IaI-SimCLR* framework (compare Section 6.3) several contributions, namely the intra- and inter-modality component of the contrastive loss, the effect of certain critical augmentations, as well as the effect of the batch-sampling strategies proposed in this thesis, have been investigated. It should be noted that the three contributions interact with one another, making it difficult to derive general statements about a clearly best-performing setup. Nevertheless, some important learning could be derived:

As discussed in Section 6.3, using samples obtained with the *random batch sampling* (RBS) strategy shows an overall declining performance with longer pre-training. One possible explanation for this behavior is referred to as mode collapse (as described in [118]), where (very low level) features get derived which are suitable for solving the contrastive problem (identifying positive pairs within a batch) but are not descriptive with respect to the downstream task evaluation. The most pronounced example here is given by the experiments with the combination of *DualSimCLR* and RBS (compare Fig. 6.7). Assumed this dynamics is responsible for the declining effects, the intuitive idea behind the LBS strategy could also serve as an explanation for the fact that the drop in performance is largely not observable in the corresponding experiments with *local batch sampling* (LBS), since low level features are less easy to find in samples collected with spatial proximity. This hypothesis is also strongly supported by the results presented in Fig. 6.9.

Nevertheless, as described in Section 6.3, the overall best performance is still represented by an early checkpoint of a model trained via RBS, hence the results within the experimental section indicate a trade-off between overall performance and a stable training dynamics. Future research could therefore be conducted around mixing the two strategies accordingly, or investigate further sampling strategies based on image similarity metrics rather than utilizing the location of the samples.

Regarding the main contribution of *IaI-SimCLR*, namely the combination of intra- and inter-modality loss functions, the evaluation in Section 6.3 shows that – when averaged over all downstream experiment settings – the proposed loss configuration represents the most effective strategy since, compared to the other two loss function setups, *IaI-SimCLR* outperformed in seven out of eight cases as can be seen in Fig. 6.8. Still, further experiments could benefit from centering the evaluation around pixel-wise tasks, as demonstrated in the cases of *SenPaMAE* and *SARFormer*. This could be achieved

either by adapting the *ResNet* backbone within a *UNet*-style decoder architecture or by switching to ViT backbones for the contrastive learning task as done in [57].

8

Conclusion

The field of *Earth observation* (EO) has evolved rapidly over the past decade, driven by the increasing availability of satellite missions equipped with sensors of diverse modalities. These developments have greatly expanded the spatial, temporal, and spectral coverage of measurements observing the Earth's surface, beneficial for a wide range of applications and scientific investigations. At the same time, the extraction of meaningful information from these heterogeneous data sources has come to rely heavily on advances in modern deep learning approaches, which are capable of learning complex patterns directly from raw observations. However, with the growing number of sensors, the complexity of the data with respect to differences of sensor parameters has also increased, since each system is carefully engineered in terms of resolutions (spatial, spectral, and temporal) for specific use-cases.

Key Contributions In light of these developments, this thesis investigates the incorporation of sensor knowledge into the deep learning workflow from multiple angles. Firstly, within a single modality, specific sensor parameters – describing the precise spectral and geometrical characteristics of measurements taken by various multi-spectral sensors – are integrated into the modeling process in the *SenPaMAE* framework [83]. This enables the development of models that can operate across data from various sensors while remaining aware of their respective differences in sensor design, which has been experimentally validated in sensor zero-shot prediction, as well as multi-sensor fine-tuning experiments. Such an approach facilitates pre-training on large datasets that include samples from multiple sensors and also opens the possibility for sensor-independent inference, an important aspect for end-user convenience. For higher-resolution data – which typically exhibits considerable variability of resulting measurement depending to the geometrical configuration at the specific moment in time where the image is acquired

(referred to as acquisition parameters) – this domain knowledge is likewise incorporated into the learning process for the case of high resolution SAR imagery in the *SARFormer* framework [84]. In the experimental section of this thesis, it has been shown that the incorporation of the acquisition parameters together with domain-adapted self-supervised learning methods significantly supports the tasks of segmentation and 3D reconstruction from high-resolution SAR images. Moreover, with the increasing number of satellites operating across various modalities, this thesis explores an approach to model pre-training via multi-modality contrastive learning in order to leverage the full spectrum of available data. Here, an extension to several successor works [56, 57] has been proposed by the *IaI-SimCLR* framework [82], which extends previous works by incorporating modality-specific loss functions in order to avoid resulting representations that do not fully reflect all information exclusive to any of the modalities.

Future Work Even though the contributions presented in this thesis aim to pave the way toward more generally applicable DL models for EO data, many scientific and engineering challenges remain.

Firstly, all approaches would benefit from larger and more diverse pre-training datasets, ideally including matched observations from various sensors and across different modalities. In this context, hybrid methods that combine contrastive and masked image modeling approaches (as well as other pretext tasks) – both investigated in this thesis – could help mitigate the individual limitations of each method. Investigations into architectural design and pre-training strategies should also be complemented by studies on scaling laws with respect to the number of trainable parameters, as such principles have proven effective in other domains but are not yet fully explored in the context of EO data.

Another crucial area of development that must go hand in hand with model adaptation and scaling of model sizes is the advancement of benchmarking methodologies. While recent research [119, 120] provides valuable first steps, many aspects underrepresented, such as diverse SAR data (most benchmarks only cover Sentinel-1 data) and the ability to cross-validate models across different geographic regions, accounting for regional biases.

On a more technical note, while research activity around low- and medium-resolution optical (multi-spectral) data is already extensive, other modalities such as SAR (especially at high resolution) and hyper-spectral imagery still hold untapped potential. In particular, open questions remain about how to fully incorporate lower-level SAR data – including both phase and amplitude – into deep learning models.

Broader Implications While this thesis investigates technical details of the academic and engineering challenges posed by multi-sensor and multi-modal pre-training, its greatest advantage lies in improving end-user accessibility. As discussed at the beginning of this thesis, the emerging terminology of *foundation models* was largely not

mentioned in this work or the corresponding publications, primarily due to the lack of a homogeneous definition. One possible way to define such models would be based solely on the adaptation of end-users. If all technical challenges are addressed and end-users can easily and reliably make use of pre-trained models for various modalities and resolutions – either on already implemented tasks for a given arbitrary (within reasonable bounds) sensors or with minimal effort of fine-tuning to new tasks – this would significantly increasing its positive impact of EO data.

In summary, this thesis contributes to the advancement of sensor-aware deep learning methods for *Earth observation* (EO) data, addressing key challenges in multi-sensor and multi-modal pre-training as well as the incorporation of sensor and acquisition parameters into the training and inference step. By contributing to this rapidly evolving field, this work aims to support the development of more robust, adaptable, and user-friendly models helping to apply EO data to tackle real world problems.

List of Figures

List of Tables

Bibliography

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of CVPR*, pages 10012–10022, 2021.

[3] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of CVPR*, pages 12104–12113, 2022.

[4] Thomas Lillesand, Ralph W Kiefer, and Jonathan Chipman. *Remote sensing and image interpretation*. John Wiley & Sons, 2015.

[5] Gareth Rees. *Physical principles of remote sensing*. Cambridge university press, 2013.

[6] Iain H Woodhouse. *Introduction to microwave remote sensing*. CRC press, 2017.

[7] Jonathan Prexl and Michael Schmitt. Chapter 8: Self-supervised learning for multi-modal Earth observation data. In Saha Sudipan, editor, *Deep Learning for Multi-Sensor Earth Observation*, pages 181–199. Elsevier, 2025.

[8] Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning. *arXiv:2304.12210*, 2023.

[9] Veenu Rani, Syed Tufael Nabi, Munish Kumar, Ajay Mittal, and Krishan Kumar. Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering*, 30(4):2761–2775, 2023.

[10] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022.

[11] Natural Resources Canada. Fundamentals of remote sensing. https://natural-resources.canada.ca/sites/nrcan/files/earthsciences/pdf/resource/tutor/fundam/pdf/fundamentals_e.pdf. Accessed: April 10, 2025.

[12] Wolfgang Demtröder. *Atoms, molecules and photons*, volume 3. Springer, 2010.

[13] Jeffrey G. Masek, hael A. Wulder, Brian Markham, Joel McCorkel, Christopher J. Crawford, James Storey, and Del T. Jenstrom. Landsat 9: Empowering open science and applications through continuity. *Remote Sensing of Environment*, 248: 111968, 2020.

[14] R Wilkinson, MM Mleczko, RJW Brewin, KJ Gaston, M Mueller, JD Shutler, X Yan, and K Anderson. Environmental impacts of earth observation data in the constellation and cloud computing era. *Science of The Total Environment*, 909: 168584, 2024.

[15] Ying Liu, Jiaxin Qian, and Hui Yue. Comprehensive evaluation of Sentinel-2 red edge and shortwave-infrared bands to estimate soil moisture. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:7448–7465, 2021.

[16] William James Frampton, Jadunandan Dash, Gary Watmough, and Edward James Milton. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 82:83–92, 2013.

[17] F. M. C. Pizani, P. Maillard, A. F. F. Ferreira, and C. C. de Amorim. Estimation of water quality in a reservoir from Sentinel-2 MSI and Landsat-8 OLI sensors. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-3-2020:401–408, 2020.

[18] Marc Rußwurm, Sébastien Lefevre, and M Körner. Breizhcrops: A satellite time series dataset for crop type identification. In *Proceedings of ICML Workshops*, volume 3, 2019.

[19] S. Puliti, J. Breidenbach, J. Schumacher, M. Hauglin, T.F. Klingenberg, and R. Astrup. Above-ground biomass change estimation using national forest inventory data with Sentinel-2 and Landsat. *Remote Sensing of Environment*, 265:112644, 2021.

[20] Peter Xaypraseuth, R. Satish, and Alok Chatterjee. NISAR spacecraft concept overview: Design challenges for a proposed flagship dual-frequency SAR mission. In *IEEE Aerospace Conference*, 2015.

[21] M. Arcioni, P. Bensi, M. W. J. Davidson, M. Drinkwater, F. Fois, C-C. Lin, R. Meynart, K. Scipal, and P. Silvestrin. ESA's biomass mission candidate system and payload overview. In *Proceedings of IGARSS*, pages 5530–5533, 2012.

[22] Rolf Werninghaus and Stefan Buckreuss. The TerraSAR-X mission and system design. *IEEE Transactions on Geoscience and Remote Sensing*, 48(2):606–614, 2010.

[23] F.Covello, F. Battazza, A. Coletta, E. Lopinto, C. Fiorentino, L. Pietranera, G. Valentini, and S. Zoffoli. COSMO-SkyMed an existing opportunity for observing the Earth. *Journal of Geodynamics*, 49(3):171–180, 2010.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Proceedings of NeurIPS*, 25, 2012.

[25] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[26] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of PMLR*, pages 315–323, 2011.

[27] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings ICML*, volume 30, page 3, 2013.

[28] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUS). *arXiv:1606.08415*, 2016.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[30] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of ICML*, pages 448–456, 2015.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[32] 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*, 12(7):2217–2226, 2019.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proceedings of NeurIPS*, 30, 2017.

[34] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of ICCV*, pages 3286–3295, 2019.

[35] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of ECCV*, pages 213–229, 2020.

[36] Chunwei Tian, Yong Xu, Zuoyong Li, Wangmeng Zuo, Lunke Fei, and Hong Liu. Attention-guided CNN for image denoising. *Neural Networks*, 124:117–129, 2020.

[37] Peng Ye, Yongqi Huang, Chongjun Tu, Minglei Li, Tao Chen, Tong He, and Wanli Ouyang. Partial fine-tuning: A successor to full fine-tuning for vision transformers. *arXiv:2312.15681*, 2023.

[38] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, pages 1597–1607, 2020.

[39] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Proceedings of NeurIPS*, 2019.

[40] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of ICML*, pages 4182–4192, 2020.

[41] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Proceedings of ECCV*, pages 776–794, 2020.

[42] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of CVPR*, pages 9729–9738, 2020.

[43] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of ICCV*, pages 9650–9660, 2021.

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023.

[45] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of CVPR*, pages 2536–2544, 2016.

[46] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of CVPR*, pages 16000–16009, 2022.

[47] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of ICCV*, pages 9414–9423, 2021.

[48] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv:2003.04297*, 2020.

[49] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of CVPR*, pages 5261–5270, 2023.

[50] Omid Ghorbanzadeh, Hejar Shahabi, Sepideh Tavakkoli Piralilou, Alessandro Crivellari, Laura EC La Rosa, Clement Atzberger, Jonathan Li, and Pedram Ghamisi. Contrastive self-supervised learning for globally distributed landslide detection. *IEEE Access*, 12:118453–118466, 2024.

[51] Mingliang Xue, Xinyuan Huo, Yao Lu, Pengyuan Niu, Xuan Liang, Hailong Shang, and Shucai Jia. Multi-scale contrastive learning for building change detection in remote sensing images. In *Proceedings of Chinese Conference on Pattern Recognition and Computer Vision*, pages 318–329. Springer, 2023.

[52] Hoàng-Ân Lê, Heng Zhang, Minh-Tan Pham, and Sébastien Lefèvre. Mutual guidance meets supervised contrastive learning: Vehicle detection in remote sensing images. *Remote Sensing*, 14(15):3689, 2022.

[53] Ya'nan Zhou, Yan Wang, Na'na Yan, Li Feng, Yuehong Chen, Tianjun Wu, Jianwei Gao, Xiwang Zhang, and Weiwei Zhu. Contrastive-learning-based time-series feature representation for parcel-based crop mapping using incomplete Sentinel-2 image sequences. *Remote Sensing*, 15(20):5009, 2023.

[54] Ankit Patnala, Scarlet Stadtler, Martin G Schultz, and Juergen Gall. Bi-modal contrastive learning for crop classification using Sentinel-2 and Planetscope. *Frontiers in Remote Sensing*, 5:1480101, 2024.

[55] Michail Tarasiou, Riza Alp Güler, and Stefanos Zafeiriou. Context-self contrastive pretraining for crop type semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.

[56] Linus Scheibenreif, Michael Mommert, and Damian Borth. Contrastive self-supervised data fusion for satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:705–711, 2022.

[57] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *Proceedings of CVPR Workshops*, pages 1422–1431, 2022.

[58] Junbo Zhang and Kaisheng Ma. Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In *Proceedings of CVPR*, pages 16650–16659, 2022.

[59] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. OmniSat: Self-supervised modality fusion for Earth observation. In *Proceedings of ECCV*, pages 409–427. Springer, 2024.

[60] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023.

[61] Gencer Sumbul, Arne De Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl. BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, 2021.

[62] Matthew Allen, Francisco Dorr, Joseph Alejandro Gallego Mejia, Laura Martínez-Ferrer, Anna Jungbluth, Freddie Kalaitzis, and Raúl Ramos-Pollán. M3LEO: A multi-modal, multi-label earth observation dataset integrating interferometric sar and multispectral data. In *Proceedings of NeurIPS*, volume 37, pages 104694–104723, 2024.

[63] Umangi Jain, Alex Wilson, and Varun Gulshan. Multimodal contrastive learning for remote sensing tasks. *arXiv:2209.02329*, 2022.

[64] Yuxing Chen and Lorenzo Bruzzone. Unsupervised CD in satellite image time series by contrastive learning and feature tracking. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

[65] Sudipan Saha, Muhammad Shahzad, Lichao Mou, Qian Song, and Xiao Xiang Zhu. Unsupervised single-scene semantic segmentation for Earth observation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

[66] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of CVPR*, pages 27672–27683, 2024.

[67] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Proceedings of NeurIPS*, volume 35, pages 197–211, 2022.

[68] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of CVPR*, pages 4088–4099, 2023.

[69] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2MAE: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *Proceedings of CVPR*, pages 24088–24097, 2024.

[70] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. SpectralGPT: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:5227–5244, 2024.

[71] Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked vision transformers for hyperspectral image classification. In *Proceedings of CVPR Workshops*, pages 2166–2176, 2023.

[72] Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv:2310.18660*, 2023.

[73] Michael J Smith, Luke Fleming, and James E Geach. EarthPT: a time series foundation model for Earth observation. *arXiv:2309.07207*, 2023.

[74] Philipe Dias, Aristeidis Tsaris, Jordan Bowman, Abhishek Potnis, Jacob Arndt, H Lexie Yang, and Dalton Lunga. OReole-FM: successes and challenges toward billion-parameter foundation models for high-resolution satellite imagery. In *Proceedings of International Conference on Advances in Geographic Information Systems*, pages 597–600, 2024.

[75] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Thorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, João Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Ankur Kumar, Myscon Truong, Denys Godwin, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Trevor Keenan, Paulo Arévolo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Christopher Hain, Robert Kennedy, Bianca Zadrozny, Gabriele Cavallaro, Campbell Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabe Moreno. Prithvi-EO-2.0: A versatile multitemporal foundation model for Earth observation applications. *arXiv:2412.02732*, 2024.

[76] Lixian Zhang, Yi Zhao, Runmin Dong, Jinxiao Zhang, Shuai Yuan, Shilei Cao, Mengxuan Chen, Juepeng Zheng, Weijia Li, Wei Liu, et al. A^2-MAE: A spatial-temporal-spectral unified remote sensing pre-training method based on anchor-aware masked autoencoder. *arXiv:2406.08079*, 2024.

[77] Nassim Ait Ali Braham, Conrad M Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. SpectralEarth: Training hyperspectral foundation models at scale. *arXiv:2408.08447*, 2024.

[78] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for Earth observation. *arXiv:2403.15356*, 2024.

[79] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: An earth observation model for any resolutions, scales, and modalities. *arXiv:2412.14123*, 2024.

[80] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of CVPR*, pages 15619–15629, 2023.

[81] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. *arXiv:2301.10343*, 2023.

[82] Jonathan Prexl and Michael Schmitt. Multi-modal multi-objective contrastive learning for Sentinel-1/2 imagery. In *Proceedings of CVPR Workshops*, pages 2135–2143, 2023.

[83] Jonathan Prexl and Michael Schmitt. SenPa-MAE: Sensor parameter aware masked autoencoder for multi-satellite self-supervised pretraining. In *Proceedings of GCPR*, pages 317–331, 2024.

[84] Jonathan Prexl, Michael Recla, and Michael Schmitt. SARFormer – an acquisition parameter aware vision transformer for synthetic aperture radar data. In *Proceedings of CVPR Workshops*, pages 2225–2234, 2025.

[85] Diane Wagner, Fabio Ferreira, Danny Stoll, Robin Tibor Schirrmeister, Samuel Müller, and Frank Hutter. On the importance of hyperparameters and data augmentation for self-supervised learning. *arXiv:2207.07875*, 2022.

[86] Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305–319, 2018.

[87] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120:25–36, 2012.

[88] Brian Markham, Del Jenstrom, Brian Sauer, Steven Pszcolka, Vicki Dulski, Jason Hair, Joel McCorkel, Geir Kvaran, Kurtis Thome, Matthew Montanaro, et al. Landsat 9 mission update and status. In *Proceedings of SPIE*, volume 11501, pages 96–102, 2020.

[89] Planet PSScene imagery product specification. `https://developers.planet.com/docs/apis/data/sensors/#the-psbsd-instrument`. Accessed: 2025-03-24.

[90] Evert Attema, Malcolm Davidson, Paul Snoeij, Björn Rommen, and Nicolas Floury. Sentinel-1 mission overview. In *Proceedings of IGARSS*, pages 36–39. IEEE, 2009.

[91] Google Earth Engine Sentinel-1 user guide. `https://developers.google.com/earth-engine/guides/sentinel1`. Accessed: 2025-03-11.

[92] Irena Hajnsek, Alberto Moreira, Manfred Zink, Stefan Buckreuss, Thomas Kraus, Markus Bachmann, and Thomas Busche. Tandem-X: Mission status and science activities. In *Proceedings of EUSAR*, pages 674–677, 2022.

[93] TerraSAR-X ground segment basic product specification document. `https://sss.terrasar-x.dlr.de/docs/TX-GS-DD-3302.pdf`, . Accessed: 2025-03-11.

[94] TerraSAR-X mage product guide. https://airbusus.com/wp-content/uploads/2020/06/r459_9_20171004_tsxx-airbusds-ma-0009_tsx-productguide_i2.01.pdf, . Accessed: 2025-05-20.

[95] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. SEN12MS – a curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Annals*, IV-2/W7:153–160, 2019.

[96] ESA WorldCover 10m 2021 v200. https://doi.org/10.5281/zenodo.7254221. Accessed: 2025-03-11.

[97] Openstreetmap building layer. https://wiki.openstreetmap.org/wiki/Key:building. Accessed: 2025-03-23.

[98] Microsoft Global Building Footprints. https://github.com/microsoft/GlobalMLBuildingFootprints. Accessed: 2025-03-23.

[99] Michael Recla and Michael Schmitt. Deep-learning-based single-image height reconstruction from very-high-resolution SAR intensity data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:496–509, 2022.

[100] Michael Recla and Michael Schmitt. Deep learning-based building footprint mapping using high-resolution SAR data. In *Proceedings of IGARSS*, pages 9983–9986, 2024.

[101] Michael Recla and Michael Schmitt. Deep learning-based DSM generation from dual-aspect SAR data. *ISPRS Annals*, X-2-2024:193–200, 2024.

[102] Michael Recla and Michael Schmitt. The SAR2Height framework for urban height map reconstruction from single SAR intensity images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211:104–120, 2024.

[103] Caleb Robinson, Kolya Malkin, Nebojsa Jojic, Huijun Chen, Rongjun Qin, Changlin Xiao, Michael Schmitt, Pedram Ghamisi, Ronny Hänsch, and Naoto Yokoya. Global land-cover mapping with weak supervision: Outcome of the 2020 IEEE GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3185–3199, 2021.

[104] Thomas Rossberg and Michael Schmitt. Towards a global model for NDVI estimation from Sentinel-1 SAR backscatter. In *Proceedings of EUSAR*, pages 245–248, 2022.

[105] Amanda A Boatswain Jacques, Abdoulaye Baniré Diallo, and Etienne Lord. Towards the creation of a canadian land-use dataset for agricultural land classification. In *Proceedings of 42nd Canadian Symposium on Remote Sensing*, 2021.

[106] Bente Eegholm, Shane Wake, Zachary Denny, Pete Dogoda, Demetrios Poulios, Barry Coyle, Pete Mulé, John Hagopian, Patrick Thompson, Luis Ramos-Izquierdo, et al. Global ecosystem dynamics investigation (GEDI) instrument alignment and test. In *Optical Modeling and System Alignment*, volume 11103, pages 53–70. SPIE, 2019.

[107] GEDI L4B gridded aboveground biomass density, version 2. https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=2017. Accessed: 2025-03-11.

[108] Sentinel-2 spectral response functions. https://sentinels.copernicus.eu/web/sentinel/search. Accessed: 2025-03-24.

[109] OLI-2 spectral response functions. https://landsat.gsfc.nasa.gov/satellites/landsat-9/landsat-9-instruments/oli-2-design/oli-2-relative-spectral-response/. Accessed: 2025-03-24.

[110] Planet developer center - the PSB.SD instrument. https://developers.planet.com/docs/apis/data/sensors/#the-psbsd-instrument. Accessed: 2025-03-24.

[111] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of CVPR*, pages 12179–12188, 2021.

[112] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017.

[113] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[114] Lukas Liebel and Marco Körner. Auxiliary tasks in multi-task learning. *arXiv:1805.06334*, 2018.

[115] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. *arXiv:1803.08673*, 2018.

[116] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR*, pages 770–778, 2016.

[117] Jonathan Prexl, Anton Baumann, and Michael Schmitt. A comparison of uncertainty estimation methods for building footprint change detection from Sentinel-2 imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10:339–346, 2024.

[118] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv:2110.09348*, 2021.

[119] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, et al. PANGAEA: A global and inclusive benchmark for geospatial foundation models. *arXiv:2412.04204*, 2024.

[120] Pierre Adorni, Minh-Tan Pham, Stéphane May, and Sébastien Lefèvre. Towards efficient benchmarking of foundation models in remote sensing: A capabilities encoding approach. In *Proceedings of CVPR Workshops*, pages 3096–3106, 2025.

[121] Jonathan Prexl and Michael Schmitt. Sensor parameter encoding for multi-sensor self-supervised learning via masked autoencoders. In *Proceedings of IGARSS*, pages 2837–2840, 2024.

[122] Jonathan Prexl and Michael Schmitt. The effect of contrastive pretraining on downstream tasks in optical remote sensing. In *Proceedings of IGARSS*, pages 526–529, 2023.

[123] Jonathan Prexl, Sudipan Saha, and Michael Schmitt. High precision mapping of building changes using Sentinel-2. In *Proceedings of IGARSS*, pages 6744–6747, 2023.

[124] Jonathan Prexl and Michael Schmitt. The potential of Sentinel-2 data for global building footprint mapping with high temporal resolution. In *Proceedings of JURSE*, pages 1–4, 2023.

[125] Ribana Roscher, Marc Russwurm, Caroline Gevaert, Michael Kampffmeyer, Jefersson A. Dos Santos, Maria Vakalopoulou, Ronny Hänsch, Stine Hansen, Keiller Nogueira, Jonathan Prexl, and Devis Tuia. Better, not just more: Data-centric machine learning for Earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 12(4):335–355, 2024.

[126] Deepika Mann, Jonathan Prexl, Sudipan Saha, and Michael Schmitt. Mapping high-resolution building development over delhi ncr using Sentinel-2. In *Proceedings of IGARSS*, pages 3190–3193, 2024.

[127] Deepika Mann, Jonathan Prexl, Michael Schmitt, and Felicitas Sommer. Multiscale assessment of land use efficiency in functional urban areas of munich and augsburg. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10:231–236, 2024.

Note of Thanks

Well, where to start... It has been a rather long journey, there were many people involved, and **all** are important!

Let's start with the professional side of things: First of all, I want to thank Michael Schmitt for giving me my second chance and providing me with all the freedom to evolve while pushing me in the correct directions whenever it was necessary! I have a hard time thinking back three years – how I felt back then – and I am super grateful that this was not the end of my academic journey! And of course, there is my previous supervisor – Bruno Eckhardt – who taught me so much and represents to this day one of my academic idols.

The time at UniBW was special, and that is for a large part because of the amazing team: So a big thank you (I will go by room number now) Max, Thomas, Anton, Deepika, Mojgan, Sanjay, Michael, Burak, Stefan, and Paolo!!! Could not have found a better group!

Then, of course, all the people – luckily now many that I can call my close friends – who helped and taught me during my time at TUM and back at my time in Marburg! Among the many noteworthy names, Lukas (soon to be Dr.), Martin (Dr.), Luise (very soon to be Dr.), and Julian (ofc. Dr.) deserve special mention! I can't overstate how important all of you were for all of this!

Last – but not even remotely less important – I want to thank my parents and my girlfriend Jenn for the unlimited support throughout this whole time!