



**UNIVERSITY
OF ICELAND**

**Advancing Land-Cover and Land-Use Classification with
Semantically Consistent and Data-Efficient Deep Learning
on Multi-Source Earth Observation Data**

Surbhi Sharma

2026

Advancing Land-Cover and Land-Use Classification with Semantically Consistent and Data-Efficient Deep Learning on Multi-Source Earth Observation Data

Surbhi Sharma

Dissertation submitted in partial fulfillment of a
Philosophiae Doctor degree in Computational Engineering

Supervisor
Prof. Dr. Morris Riedel

Doctoral Committee
Prof. Dr. Morris Riedel
Associate Prof. Dr. Gabriele Cavallaro
Assistant Prof. Dr. Claudia Paris

Opponents
Prof. Dr. Vito Pascazio
Assoc. Prof. Dr. Gülşen Taşkın Kaya

Faculty of Industrial Engineering, Mechanical Engineering and
Computer Science
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, October 2026

Advancing Land-Cover and Land-Use Classification with Semantically Consistent and Data-Efficient Deep Learning on Multi-Source Earth Observation Data
(Advancing LULC Classification with Semantically Consistent and Data-Efficient DL on Multi-Source EO Data)

Dissertation submitted in partial fulfillment of a *Philosophiae Doctor* degree in Computational Engineering

Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
School of Engineering and Natural Sciences
University of Iceland
Dunhagi 5
107 Reykjavík
Iceland

Telephone: +354 525-4700

Bibliographic information:

Surbhi Sharma (2026) *Advancing Land-Cover and Land-Use Classification with Semantically Consistent and Data-Efficient Deep Learning on Multi-Source Earth Observation Data*, PhD dissertation, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, 146 pp.

ISBN 978-9935-578-13-6

ORCID: <https://orcid.org/0000-0001-6774-6542>

Copyright © 2026 Surbhi Sharma

This thesis may not be copied in any form without author permission.

Printing: Háskólaprent, Fálagata 2, 107 Reykjavík
Reykjavík, Iceland, October 2026

Abstract

Land-use and Land-cover (LULC) mapping from Earth Observation (EO) imagery plays a fundamental role in environmental monitoring, urban planning, agricultural management, and climate change assessment. Recent advances in satellite missions, Machine Learning (ML), Deep Learning (DL), and foundation models have significantly improved large-scale mapping capabilities. However, important challenges remain, including semantic inconsistencies between satellite imagery and *in-situ* observations, the lack of large-scale benchmark datasets with detailed LULC information for training and evaluating advanced DL models, heterogeneous data sources, mixed pixel effects, limited availability of annotated training data, and the need for computationally scalable processing frameworks. Addressing these challenges requires methodologies that integrate multiple data modalities while improving semantic consistency, data efficiency, and spatial representation.

This thesis presents a comprehensive framework for advancing satellite-based LULC mapping through four complementary research directions. First, it investigates the semantic relationship between satellite observations and *in-situ* field data by leveraging geo-tagged field photographs as an intermediate observational layer. The analysis demonstrates that apparent disagreements between satellite-derived and field-based information often arise not only from limitations of satellite observations but also from ambiguities in field interpretation and inconsistencies in land-cover (LC) taxonomies. Second, the thesis produces the Sen4Map, a large-scale benchmark dataset, integrating Sentinel-2 imagery with harmonized Land Use/Cover Area Frame Survey (LUCAS) field observations and ancillary information. The benchmark provides semantically rich annotations designed to support the development and evaluation of robust DL models for LULC mapping. Third, the thesis develops methods for sub-pixel LULC estimation to capture fine-scale landscape heterogeneity beyond the native spatial resolution of satellite imagery. As a proof of concept, the proposed framework estimates built-up fractions from Sentinel-2 imagery using geospatial foundation model embeddings, demonstrating the potential of pretrained spectral-spatial representations for sub-pixel characterization. And finally, the thesis explores data-efficient learning strategies for EO by investigating few-shot learning with meta-learning on publicly available Remote Sensing (RS) benchmark datasets. These approaches demonstrate improved generalization under limited annotation availability, reducing the dependence on extensive labeled datasets while maintaining competitive mapping performance across diverse environments.

Beyond methodological advances, the thesis also contributes scalable computational

solutions through a Hierarchical Data Format, version 5 (HDF5)-based data architecture optimized for High-Performance Computing (HPC) environments, enabling efficient storage, access, and processing of large multi-source EO datasets. Collectively, the proposed methodologies contribute toward more semantically consistent, data-efficient, and scalable satellite-based LULC mapping systems. As EO transitions toward foundation models, unified embedding representations, and agentic geospatial intelligence, the contributions presented in this thesis establish the datasets, semantic frameworks, and embedding-centric methodologies required to support the next generation of LULC mapping.

Útdráttur

Kortlagning landnotkunar og landþekju (LULC) úr jarðathugunargögnum (EO) gegnir grundvallarhlutverki í umhverfissvöktun, borgarskipulagi, landbúnaðarstjórnun og mati á loftslagsbreytingum. Nýlegar framfarir í gervitunglförum, vélanámi (ML), djúpnámi (DL) og grunnlíkönum hafa verulega bætt kortlagningarmöguleika á stórum svæðum. Hins vegar eru enn til staðar miklar áskoranir, þar á meðal merkingarleg ósamræmi milli gervitunglamynda og staðbundinna mælinga, skortur á stórum viðmiðunargagnasöfnum með nákvæmum upplýsingum um LULC til þjálfunar og mats á háþrúðum DL-líkönunum, fjölbreyttar gagnaveitur, blönduð pixlaáhrif, takmarkað framboð af merktum þjálfunargögnum og þörf á reikniritanlega skalanlegum úrvinnsluröðmum. Til að takast á við þessar áskoranir þarf að beita aðferðum sem samþætta margar gagnagerðir og bæta merkingarlegt samræmi, gagnanýtni og framsetningu rúmupplýsinga.

Þessi ritgerð kynnir heildstæðan ramma til að efla LULC-kortlagningu með gervitunglum með fjórum samverkandi rannsóknarstefnum. Í fyrsta lagi kannar hún merkingarlegt samband milli gervitunglamælinga og staðbundinna vettvangsgagna með því að nýta landfræðilega merktar vettvangsmyndir sem millilag athugana. Greiningin sýnir að áberandi ósamræmi milli gervitungls og vettvangsbundinna upplýsinga stafar oft ekki einungis af takmörkunum gervitunglamælinga heldur einnig af óskýrleika í túlkun á vettvangi og ósamræmi í flokkunarkerfum landþekju. Í öðru lagi býr ritgerðin til Sen4Map, stórt viðmiðunargagnasafn sem sameinar Sentinel-2-myndir við samræmdar svæðismælingar (Area Frame Survey) á landnotkun/landþekju (LUCAS) og aðrar aukaupplýsingar. Viðmiðunin inniheldur ítarlegar merkingar sem ætlaðar eru til að styðja við þróun og mat á traustum DL-líkönunum fyrir LULC-kortlagningu. Í þriðja lagi þróar ritgerðin aðferðir til að áætla LULC á undirpixlastigi (sub-pixel) til að fanga fínskalaða misleitni í landslagi sem fer út fyrir upprunalega rúmupplausn gervitunglamynda. Sem sönnun á virkni hugmyndarinnar áætlað ramminn sem lagður er fram hlutfall byggðs lands úr Sentinel-2 myndum með innfellingum (embeddings) landfræðilegra grunnlíkana, sem sýnir möguleika forþjálfaðra framsetninga, sem fanga litrófs- og rúmupplýsingar, til undirpixlaeinkenninga. Að lokum kannar ritgerðin gagnanýtnar námsaðferðir fyrir jarðathugun (EO) með því að rannsaka fárra sýna nám (few-shot) með metanámi (meta-learning) á opinberlega aðgengilegum viðmiðunargagnasöfnum í fjarkönnun (RS). Þessar nálganir sýna bættu alhæfingarhæfni við takmarkaða merkingu, draga úr þörf fyrir umfangsmikil merkt gagnasöfn og viðhalda samkeppnishæfum kortlagningarafköstum í fjölbreyttum umhverfum.

Fyrir utan aðferðafræðilegra framfara leggur ritgerðin einnig fram skalanlegar reikn-

anlegar lausnir með gagnalíkani byggðu á HDF5 sem er fínstillt fyrir háafkasta tölvu-
umhverfi (HPC), sem styður við skilvirka geymslu, aðgang og vinnslu stórra fjöl-
heimilda EO gagnasetta. Samanlagt stuðla aðferðir sem lagðar eru fram að merk-
ingarlega samræmdari, gagnanýtnari og skalanlegri gervitunglsmiðuðum kerfum til
LULC-kortlagningar. Þegar jarðathugun (EO) færist í átt að grunnlíkönum, sameinuðum
innfellingum og aðgerðamiðaðri landfræðilegri greind, leggja framlögin í þessari ritgerð
grunninn að gagnasöfnum, merkingargrindum og innfellingarmiðuðum aðferðum sem
nauðsynlegar eru til að styðja næstu kynslóð LULC-kortlagningar.

The universe does not assign labels, nor does it classify. It simply exists. It is we, as humans, who create categories in our pursuit of understanding.

Contents

Abstract	iii
Útdráttur	v
Dedication	vii
Contents	ix
List of Figures	xi
List of Tables	xv
List of Original Publications	xvii
Abbreviations	xix
Acknowledgments	xxi
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	3
1.3 Outline	6
1.3.1 Thesis Structure	6
1.3.2 Publications	7
1.4 Contributions	8
1.4.1 Research Publications	8
1.4.2 Code and Data Repositories	9
1.4.3 Additional Activities	9
1.4.4 Dissemination	10
2 Background	13
2.1 Multi-source EO Data	13
2.2 Satellite-based LULC Classification	18
2.3 Challenges in Satellite-Based LULC Mapping	18
2.3.1 The Semantic Gap between Ground and Satellite Observations	19
2.3.2 Landscape Heterogeneity and Mixed Pixels	20
2.3.3 Reference Data Limitations	22

2.3.4	Benchmark Data Limitations	23
2.4	Large-Scale Deep Learning for Earth Observation	23
2.4.1	High-Performance Computing for Large-Scale LULC Mapping	24
2.4.2	Foundation Models for Earth Observation	25
2.4.3	Efficient Data Management: HDF5 Dataset	26
3	Related Work	29
3.1	Bridging Spatial-Semantic Gaps Using Geo-Tagged Field Observations	29
3.2	Large-Scale Benchmark Datasets for Detailed LULC Mapping	32
3.3	Sub-Pixel LULC Modeling for Fine-Scale Landscape Representation	33
3.4	Generalization of DL Models under Limited Annotation Availability	36
4	Summary of Publications	39
4.1	<i>Objective 1: Develop methods that leverage geo-tagged field data to characterize and bridge the spatial and semantic gaps between ground-level observations and satellite imagery</i>	<i>39</i>
4.1.1	Paper II	39
4.1.2	Paper VI	40
4.2	<i>Objective 2: Establish large-scale, multi-source benchmark dataset with detailed LULC information to facilitate the training and evaluation of advanced DL models</i>	<i>42</i>
4.2.1	Paper III	42
4.3	<i>Objective 3: Build methods for sub-pixel LULC estimation to capture fine-scale landscape heterogeneity beyond the native spatial resolution of satellite imagery</i>	<i>46</i>
4.3.1	Paper IV	46
4.3.2	Paper V	48
4.4	<i>Objective 4: Investigate and evaluate the generalization capability of DL models that are resilient to limited annotation availability</i>	<i>50</i>
4.4.1	Paper I	50
5	Conclusions	53
5.1	Summary	53
5.2	Future Work	56
	Paper I	59
	Paper II	64
	Paper III	79
	Paper IV	97
	Paper V	102
	Paper VI	119
	References	127

List of Figures

1.1	Example of a location depicted in red (a) Sentinel-2 yearly composite [29], (b) a Very High Resolution Very High Resolution (VHR) Google Map Image [30], (c) cropped 3×3 Sentinel-2 patch (20m radius around the point), and (d) Associated street-level images with a downward-facing photo of the target point (P) in the middle, and four landscape photos facing the cardinal directions (N, E, S, W) [31].	2
1.2	BPMN flow diagram of the thesis with graphical representation of thesis objectives and temporal flow of the thesis with publications and objectives.	11
2.1	Data from multiple sources: (a) Satellite data [59], (b) <i>In-situ</i> data [31], and (c) Crowd-sourced data [13].	14
2.2	Overview of the Land Use/Cover Area frame Survey (LUCAS) sampling design from the year 2018. The first phase involves a systematic stratified sampling over a 2 km grid covering Europe, generating approximately 1.1 million points for Land-Cover (LC) interpretation using orthophoto imagery. The second phase includes field-based surveys and photo-interpretation, resulting in around 337,000 validated <i>in-situ</i> observations with detailed LC and Land-Use (LU) labels. This hierarchical sampling strategy ensures spatially consistent and statistically representative ground reference data for Earth Observation (EO) applications [31, 64].	15
2.3	Example of a large-scale global LC product: ESA WorldCover 2020 [79] generated from satellite observations at 10m spatial resolution, supporting applications in environmental monitoring, climate studies, and sustainable land management.	17
2.4	Example of LUCAS point labeled as “Artificial Land” (a) High-resolution Google Earth Satellite View, (b) LUCAS street-level Point image, (c) zoomed-in view of the residential area containing LUCAS point, (d) Sentinel-2 image located in the LUCAS point. [Extracted from Paper IV [100] © 2024 IEEE.]	20
2.5	Figures (a) and (b) show the Sentinel-2 and High-resolution Google Earth Satellite View for a LUCAS point with “LC1” as “Artificial Land”. Figures (c) shows the presence of multiple classes/endmembers via Abundance Map in heterogeneous environment, as obtained from the Sentinel-2 yearly composite. [Extracted from Paper IV [100] © 2024 IEEE.]	21

2.6	Modular Supercomputing Architecture from Dynamical Exascale Entry Platform (DEEP) Systems (Dynamical Exascale Entry Platform–Extreme Scale Technologies (DEEP-EST)) [126]	24
2.7	Schematic view of an embedding. A pretrained encoder compresses the observations of a geographic location into a dense d -dimensional latent vector that jointly encodes spectral patterns, spatial texture, temporal dynamics, structural context, and LC semantics. Such embeddings serve as reusable, task-agnostic inputs for downstream Land-Use and Land-Cover (LULC) analysis.	26
2.8	Sen4Map HDF5Dataset Architecture.	27
3.1	Examples of LUCAS <i>in-situ</i> samples associated with the “Woodland” LC class: (a)(b) LUCAS street-level images, (c)(e) Sentinel-2 satellite data, (d)(f) ESA WorldCover thematic product. The sample’s locations are highlighted in black. [Extracted from Paper II [153], © 2023 SPIE.]	30
4.1	Flowchart of the proposed procedure used to analyze the semantic gap between <i>in-situ</i> and satellite-based land-cover (LC) data. The semantic LC consistency analysis is carried out comparing the LC information: (1) collected <i>in-situ</i> , (2) estimated in the satellite-based thematic products, and (3) derived from the semantic segmentation of street-level images. [Extracted from Paper II [153], © 2023 SPIE.] . .	41
4.2	Spatial and semantic diagnostic framework for assessing <i>in-situ</i> -satellite LC discrepancies using field photos, satellite images, a concept vocabulary (defined by the field data collection protocol), and <i>in-situ</i> LC labels. [Extracted from Paper VI, ©2026 IEEE.]	42
4.3	LC3 top-1 destination frequencies computed on the set of samples where <i>in-situ</i> and satellite data disagree. [Extracted from Paper VI, ©2026 IEEE.]	43
4.4	Example of a Sentinel-2 time series from the Sen4Map benchmark dataset associated with one LUCAS point. In this qualitative example, only the images acquired from the 21st of February until the 18th of November are visualized to capture the temporal component that can characterize different LC classes. [Extracted from Paper III [168]] . .	44
4.5	True color composition of 64×64 Sentinel-2 patches belonging to the Sen4Map dataset, which shows the variability of the LC classes located in different countries. [Extracted from Paper III [168]]	45
4.6	Statistical distribution depicted as a violin plot of the number of Sentinel-2 images associated with each (a) LC class and (b) LU class. The first level of the hierarchy is reported. [Extracted from Paper III [168]] . .	45
4.7	Abundance Map obtained after Spectral Unmixing for four different LUCAS points, marked as red and yellow, labeled as ‘Artificial Land’. Figures (1.a)-(4.a) and (1.b)-(4.b) show the LUCAS points in Sentinel-2 and High-resolution Google Earth Satellite View. Figures (1.c) - (4.c) represent the Abundance Map obtained from the Sentinel-2 yearly composite of respective LUCAS points. [Extracted from Paper IV [100], ©2024 IEEE.]	47

4.8	Workflow of the proposed method based on three main steps: (1) Data Preparation and Harmonization, (2) Sub-pixel Data Preparation, and (3) Sub-Pixel Built-up Feature Detection and Mapping. [Extracted from Paper V [200]]	48
4.9	For each of the four patches: (a) High-resolution Google Earth Image, (b) Sentinel-2 yearly composite for reference. Highlighted in orange are the unidentified overlaid built-up features from OpenStreetMap (OSM) in three global LC maps, (c) Predicted LC Map based on the highest fraction obtained with the proposed model, (d) Updated LC map based on a Threshold of 0.20, (e) Highlighted in yellow are the identified built-up features. [Extracted from Paper V [200]]	49
4.10	Visual representation of the framework: (a) Meta-learning approach to FSL, and (b) baseline finetuning approach, where f_θ is the feature extraction layer parameterized by θ and g is the embedding function using the ResNet-18 architecture. [Extracted from Paper I [201], ©2022 IEEE.]	51

List of Tables

1.1 Relationship between thesis objectives and papers. Filled circles identify each paper’s principal thesis objective; open circles denote supporting or enabling contributions. Each paper has one principal objective. 9

2.1 Comparison of widely used global LULC products. 19

3.1 Comparison of Sen4Map with widely used benchmark datasets for LULC mapping. 33

5.1 Summary of the main findings corresponding to each research objective. 54

List of Original Publications

The following publications, listed in chronological order, are part of the thesis. They are included as Appendix, and their summary and explanation can be found in Chapter 4.

- Paper I:** **S. Sharma**, R. Roscher, M. Riedel, S. M. Memon, and G. Cavallaro, “Improving Generalization for Few-Shot Remote Sensing Classification with Meta-Learning”, in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Kuala Lumpur, Malaysia, 2022, pp. 5061-5064, <http://doi.org/10.1109/IGARSS46834.2022.9884699>
- Paper II:** C. Paris, L. Martinez-Sanchez, M. van der Velde, **S. Sharma**, R. Sedona, and G. Cavallaro, “Accuracy assessment of land-use-land-cover maps: the semantic gap between in-situ and satellite data”, in *Proceedings Volume 12733, Image and Signal Processing for Remote Sensing XXIX, SPIE Remote Sensing*, Amsterdam, Netherlands, 2023, <http://doi.org/10.1117/12.2679433>
- Paper III:** **S. Sharma**, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Sen4Map: Advancing Mapping with Sentinel-2 by Providing Detailed Semantic Descriptions and Customizable Land-Use and Land-Cover Data”, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, vol. 17, pp. 13893 - 13907, 2024, <http://doi.org/10.1109/JSTARS.2024.3435081>
- Paper IV:** **S. Sharma**, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Enhancing Land Cover Mapping: A Novel Automatic Approach To Improve Mixed Spectral Pixel Classification”, in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Athens, Greece, 2024, pp. 1017-1020, <http://doi.org/10.1109/IGARSS53475.2024.10641976>

- Paper V:** S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Beyond Pixels: Identifying Built-Up Features at Sub-Pixel Level for Enhanced Satellite-Based Land Cover Mapping”, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, 2026, <http://doi.org/10.1109/JSTARS.2026.3712773>
- Paper VI:** S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Diagnosing the Ground–Satellite Gap in Land Cover Using Vision Foundation Models and Field Photos”, in *Proceedings of the IEEE India Geoscience and Remote Sensing Symposium (InGARSS)*, 2026 (accepted)

Abbreviations

ADMIRE	Adaptive Multi-Tier Intelligent Data Manager for Exascale
AEDs	AlphaEarth Embeddings
AVHRR	Advanced Very High Resolution Radiometer
BPMN	Business Process Model and Notation
CAP	Common Agricultural Policy
CLC	CORINE Land Cover
CLIP	Contrastive Language–Image Pre-training
CNNs	Convolutional Neural Networks
CPU	Central Processing Unit
DEEP	Dynamical Exascale Entry Platform
DEEP-EST	Dynamical Exascale Entry Platform–Extreme Scale Technologies
DL	Deep Learning
EO	Earth Observation
ESA	European Space Agency
ETM+	Enhanced Thematic Mapper Plus
EU	European Union
EuroHPC	European High-Performance Computing Joint Undertaking
FSL	Few-Shot Learning
GFM	Geospatial Foundation Models
GPU	Graphics Processing Unit
HDF5	Hierarchical Data Format, version 5
HPC	High-Performance Computing

I/O	Input/Output
IGBP	International Geosphere–Biosphere Programme
JTI	Joint Technology Initiative
LC	Land-Cover
LSMM	Linear Spectral Mixture Model
LU	Land-Use
LUCAS	Land Use/Cover Area frame Survey
LULC	Land-Use and Land-Cover
MERIS	Medium Resolution Imaging Spectrometer
ML	Machine Learning
MODIS	Moderate Resolution Imaging Spectroradiometer
MPI	Message Passing Interface
MSI	MultiSpectral Instrument
OLI	Operational Land Imager
OSM	OpenStreetMap
RF	Random Forests
RS	Remote Sensing
SAR	Synthetic Aperture Radar
SMA	Spectral Mixture Analysis
SVR	Support Vector Regression
TM	Thematic Mapper
UMD	University of Maryland
VFM	Vision Foundation Models
VHR	Very High Resolution
ViTs	Vision Transformers
VLMs	Vision-Language Models
VViT	Video Vision Transformers
WP	Work Package

Acknowledgments

Completing this thesis has been a challenging and a rewarding journey, marked by countless opportunities for learning, growth, discovery, and endless cups of coffee. Along the way, I have been fortunate to receive the support, guidance, encouragement, and kindness of many people, without whom this work would not have been possible.

First and foremost, I would like to express my deepest gratitude to my doctoral committee for their invaluable mentorship, understanding, and constructive feedback throughout my doctoral research. My most sincere thanks to Prof. Dr. Morris Riedel whose generosity, wisdom, and guidance at the University of Iceland have made a lasting impact on my doctoral journey. I owe my deepest gratitude to Prof. Dr. Gabriele Cavallaro, whose remarkable support, trust, and constant encouragement have been a source of strength throughout my PhD. His unwavering belief in me made it easier to persevere through times of pressure, uncertainty, and self-doubt. I am immensely grateful to Assistant Prof. Dr. Claudia Paris. Her confidence in my abilities inspired me to embrace challenges, remain curious, and continually strive for scientific excellence.

I am particularly grateful to Prof. Helmut Neukirchen, whose inspiring lectures and passion for research sparked many of the ideas that have accompanied me throughout my PhD, and to Prof. Ribana Roscher for her invaluable insights and the significant contribution she has made to my PhD journey.

I am also grateful to all the current and former colleagues of the Simulation and Data Lab at the Jülich Supercomputing Centre (JSC) and the High-Productivity Data Processing group in Iceland for fostering a great and encouraging working atmosphere. I extend my special thanks to Dr. Rocco Sedona, for being an outstanding collaborator and a wonderful friend throughout my PhD journey. Your invaluable contributions to our co-authored research, insightful discussions, and thoughtful feedback on both this thesis and our papers have greatly enriched my work. Special thanks to Chadi, Amer, Stefano, Josefine, Edoardo and Kennedy. The many scientific discussions, shared experiences, and moments of laughter made this journey both enjoyable and rewarding.

I gratefully acknowledge the University of Iceland and the Faculty of Industrial Engineering, Mechanical Engineering and Computer Science for providing an outstanding academic environment to pursue my doctoral studies. I am also thankful to Forschungszentrum Jülich and JSC for the excellent research facilities, computational resources, and administrative support that greatly facilitated the completion of this work. I am

also thankful to my colleagues in the ADMIRE project from across Europe for their collaboration, expertise, and willingness to share their knowledge.

I would also like to express my sincere appreciation to Prof. Morris Riedel, Prof. Gabriele Cavallaro, Katrín, and Sigga for organizing the PhD retreats, which created an inspiring environment for scientific discussions, exchanging ideas, and building connections with fellow researchers.

To my husband, Jay, words cannot fully express my gratitude. This chapter of my life would never have unfolded without you. Thank you for believing in me when I struggled to believe in myself, for your endless patience during the long days and late nights, for every cup of chai and coffee, and for always reminding me that every challenge was temporary. Your love, strength, and unwavering faith carried me through this journey, and I will forever be grateful to have you by my side.

To my beloved family, thank you for your unconditional love, and support. Each of you has played a meaningful role in shaping the person I am today, and I carry your love and values with me every day. I would especially like to thank my father Dinesh, my mother Pratima, my in-laws Asha and Rajesh; Vani Chachi and Ajay Chacha, for their constant encouragement, patience, and belief in me. This achievement belongs not only to me but to all of you.

To Bhamini, Ayush, and Shriya, thank you for your endless love, constant inspiration, and for always celebrating my successes as if they were your own.

To my wonderful friends, both near and far, thank you for standing by me throughout this journey, helping me through the setbacks, and reminding me to find joy and balance amidst the demands of a PhD. Special thanks to Riddhi, Smit, Francesca, Rocco, Hanna, Gabi, Timo, Disha, and Bela aunty. Your presence has made this journey richer, brighter, and far more meaningful.

To the loving memory of my grandmother, Sneha Lata, who dreamed of seeing me earn a doctoral degree; and my beloved daughter, Nandini (my Matariyu), who shared a precious part of this journey with me. I dedicate this thesis to both of you!

Finally, I am grateful to The One, whose blessings, mercy, and guidance have illuminated every step of this journey.

1 Introduction

This chapter provides an introduction to the thesis, laying out the motivation in Section 1.1 and objectives in Section 1.2, as well as giving an overview of the outline in Section 1.3 and the main contributions made in Section 1.4.

1.1 Motivation

Earth observation EO encompasses the systematic collection of information about the Earth's physical, chemical, and biological systems, as well as human activities, using observations acquired from space, air, land, and sea [1]. This broad data ecosystem includes Remote Sensing (RS) data, *in-situ* observations, citizen science data, and exogenous or proxy data, each providing complementary information about the Earth's surface and its dynamics [2, 3, 4, 5]. The last decade has witnessed an unprecedented growth in the availability of openly accessible EO data, fundamentally transforming the landscape of Land-Use and Land-Cover LULC mapping [6]. LULC mapping refers to the characterization of the Earth's surface based on how land is used and the physical cover present on it [7]. It provides essential information for understanding environmental changes, monitoring human activities, and supporting sustainable land management and planning [8, 9, 10].

Global satellite missions, such as the Sentinel-2 constellation, now provide continuous, high-resolution multispectral time-series imagery with frequent revisit cycles, enabling detailed monitoring of land dynamics at regional to global scales [11]. Complementing these observations is the increasing availability of large-scale *in-situ* datasets, including systematically collected field surveys from European initiatives [12] and crowd-sourced geo-tagged street-level imagery, which provide valuable ground-level perspectives that are often unavailable from satellite observations alone [13]. Together, these data sources offer rich and complementary information about the Earth's surface, capturing both large-scale spatial patterns and fine-grained local characteristics [14]. Leveraging these complementary sources of EO data, Deep Learning (DL) has emerged as a powerful approach for automatically learning complex spatial and spectral patterns [15, 16]. Its ability to learn hierarchical representations directly from data has significantly advanced the accuracy and scalability of LULC classification [17]. However, the remarkable success of DL models for LULC classification is largely dependent on the availability of

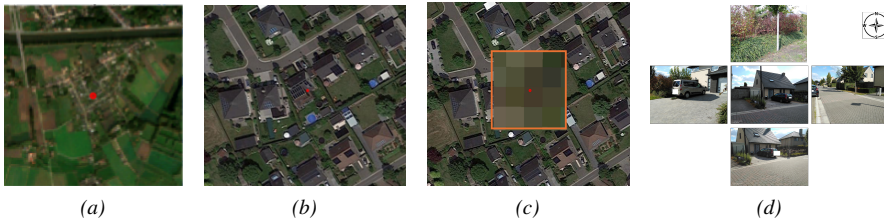


Figure 1.1. Example of a location depicted in red (a) Sentinel-2 yearly composite [29], (b) a Very High Resolution VHR Google Map Image [30], (c) cropped 3×3 Sentinel-2 patch (20m radius around the point), and (d) Associated street-level images with a downward-facing photo of the target point (P) in the middle, and four landscape photos facing the cardinal directions (N, E, S, W) [31].

vast quantities of high-quality annotations that are both informative and representative of the diversity present in real-world landscapes [18].

Effectively harnessing the potential of these diverse datasets remains a major scientific challenge, primarily due to the significant semantic gap that exists between heterogeneous data modalities [19]. Satellite imagery provides a top-down spectral representation of the Earth’s surface, whereas ground-level observations capture object-level details from entirely different viewpoints [20], as depicted in Figure 1.1. The figure illustrates the semantic gap between different sources and representations of EO data for the same geographic location, marked in red. While these observations describe the same location, they capture the landscape at substantially different spatial resolutions, viewing perspectives, and levels of detail. The Sentinel-2 imagery, Figure 1.1 (a), provides a coarse, top-down representation of the surrounding landscape, whereas the VHR image, Figure 1.1 (b), reveals finer spatial structures and individual objects. In contrast, the street-level images, Figure 1.1 (d), provide ground-based views of the target point, including a downward-facing image of the point (P) and landscape views in the four cardinal directions (N, E, S, W). This diversity highlights the semantic gap that arises when observations of the same location differ in spatial detail, viewing geometry, and visual context [21]. These differences in sensing mechanisms, spatial resolution, scale, perspective, and contextual information create inconsistent feature representations that limit the effectiveness of conventional multimodal learning approaches [22]. Furthermore, ground-level field photos are typically available only for a sparse subset of locations rather than for every pixel in EO imagery, making direct supervision and large-scale multimodal alignment inherently difficult [3]. Since collecting such data is expensive, labour-intensive, and often impractical at large spatial scales, available reference data remain sparse, geographically biased, and unevenly distributed across LC classes [23, 24]. As a result, models trained on these data often struggle to generalize to unseen regions and diverse environmental conditions [25, 26, 27, 28].

Additionally, the challenge is exacerbated by the inherent heterogeneity of real-world landscapes, where individual satellite pixels often contain mixtures of multiple LC types

[32, 33, 34]. As can be seen from Figure 1.1 (c) and (d), the target Sentinel-2 pixel is highly heterogeneous, encompassing diverse land-cover elements that are visible from both the satellite and ground-level perspectives. Such mixed pixels are particularly prevalent in transitional environments, fragmented agricultural systems, urban-rural interfaces, and sparsely vegetated regions, where the assumption of homogeneous pixel composition is frequently violated [35, 36]. Consequently, conventional pixel-based classification approaches may fail to adequately represent the underlying LC distribution, leading to information loss and reduced mapping accuracy [37, 38]. Addressing these limitations requires methods that can move beyond discrete pixel-level predictions and characterize LC composition at the sub-pixel level while effectively integrating complementary information from diverse data modalities [39, 40, 41].

Although large parts of the Earth’s surface have already been mapped using EO data, existing mapping approaches often provide limited representations of the complexity and heterogeneity of the landscape, especially in generating a semantic representation of the Earth [21]. This also highlights the need for benchmark datasets that are highly representative of real-world landscapes, capturing diverse LC and LU types, spatial and temporal configurations, and geographic contexts to enable robust training and comprehensive evaluation of LULC mapping methods. Therefore, there is a pressing need for frameworks that can effectively exploit publicly available, large-scale, multi-source EO datasets while bridging the semantic gap between heterogeneous modalities, enabling accurate and detailed LULC mapping, capturing fine-scale landscape heterogeneity, and generalizing across diverse geographic regions and environmental conditions [42]. Addressing these interconnected challenges is essential for building robust, scalable, and transferable LULC classification systems capable of delivering reliable mapping products across a wide range of spatial scales and landscapes [43].

Furthermore, advances in computing power, including the increasing availability of petascale and exascale supercomputers [44, 45, 46], have significantly reduced the barriers to processing massive datasets, enabling large-scale DL models, such as Geospatial Foundation Models (GFM), to be trained and deployed at continental and global scales [47]. The pretrained embeddings generated by these models provide rich, transferable representations of the Earth’s surface that can be leveraged across a wide range of RS tasks [48]. Together, these technological and data-driven advances present an unparalleled opportunity to develop next-generation LULC mapping systems that fully exploit large-scale, multi-source information to produce more comprehensive, timely, accurate, and detailed LULC products [17, 49].

1.2 Objectives

This thesis attempts to answer the following research questions: *How can publicly available, large-scale, multi-source datasets (including multispectral satellite imagery and geo-tagged images) be effectively exploited using advanced DL models to improve*

satellite-based LULC mapping while addressing:

- (i) *semantic and spatial gap between diverse data modalities,*
- (ii) *the need for more detailed and complex LULC classification legends,*
- (iii) *fine-scale landscape heterogeneity, and*
- (iv) *the scarcity of high-quality reference data.*

The question above inspired the definition of the following **thesis objectives (TO)**:

- **TO1:** Develop advanced DL methodologies that leverage geo-tagged field data and state-of-the-art vision foundation models to bridge the inherent spatial and semantic gaps between *in-situ* ground-level observations and satellite imagery, enabling the systematic study, characterization, and interpretation of ambiguities and taxonomic inconsistencies to improve the accuracy and reliability of LULC mapping.
- **TO2:** Establish a large-scale multi-source benchmark dataset with detailed LULC annotations by integrating multi-temporal satellite imagery with harmonized ground-reference, which advances research in DL-based LULC classification by providing a standardized resource for model training and evaluation for large-scale LULC mapping and monitoring.
- **TO3:** Build robust sub-pixel LULC mapping methods that leverage pretrained EO embeddings together with spectral, spatial, and multi-scale contextual information, to estimate fractional LC compositions within mixed pixels in order to better characterize heterogeneous landscapes and enhance the precision and reliability of LULC mapping beyond the native spatial resolution of satellite imagery.
- **TO4:** Investigate and evaluate annotation-efficient DL models for LULC classification that maintain high predictive performance with limited labeled data, with the objective of reducing annotation requirements and enabling scalable, transferable, and cost-effective RS and EO applications across diverse geographic regions and acquisition conditions.

Although modern EO classification systems have enabled comprehensive global monitoring of the Earth's surface, translating satellite measurements into meaningful descriptions of LULC remains a non-trivial task [21, 50]. A persistent challenge in EO is the semantic gap between what is visible in satellite imagery and the actual conditions on the ground [21]. Satellite images provide an indirect representation of the Earth's surface, often constrained by spatial resolution, sensor characteristics, viewing geometry, atmospheric effects, and temporal variability [50]. As a result, visually similar image patterns may correspond to different ground conditions, while distinct surface characteristics may appear spectrally alike, leading to ambiguities in interpretation [51].

These discrepancies are amplified by the inherent heterogeneities present in the scene under observation [33]. Intrinsically, they limit the ability of DL models to accurately infer real-world phenomena from remotely sensed data [15]. Therefore, there is a critical need to develop methods that can identify and bridge this semantic gap by learning representations that better capture the relationship between image observations and ground truth. Such approaches are essential for improving the reliability, interpretability, and generalization of RS DL models across diverse landscapes and environmental conditions, which are addressed in **TO1**.

The focus of **TO2** is on developing large-scale, multi-source dataset that serve as a foundation for next-generation LULC mapping. Rather than supporting only traditional semantic LC classification, the dataset is designed to enable advanced DL models to integrate heterogeneous information from multispectral satellite imagery and geo-tagged ground-level observations. By learning complementary spectral, spatial, and semantic representations across modalities, such models can move beyond conventional mapping paradigms to produce more detailed, context-aware, and robust descriptions of LULC [49]. As a result, the dataset not only facilitates the training and benchmarking of DL methods but also promotes the development of richer EO models capable of capturing the complexity and heterogeneity of real-world environments.

Advancing LULC mapping is not only limited to the pixel level but extends beyond it to capture finer spatial heterogeneity within individual pixels. In many real-world landscapes, especially those characterized by mixed or transitional LC types, a single pixel often represents a combination of multiple classes rather than a homogeneous surface [33, 36]. This intrinsic mixing introduces uncertainty in traditional classification frameworks and limits their ability to accurately represent complex environments [38]. Addressing this challenge requires the development of sub-pixel level methods that can estimate fractional class compositions by exploiting spectral–spatial relationships and multi-scale contextual information [39, 41]. Building such methods is the main focus of **TO3**, and is essential for enhancing the fidelity of LULC mapping, improving the characterization of heterogeneous landscapes, and enabling more reliable analysis for environmental monitoring and decision-making applications.

Despite the remarkable success of DL methods in RS and EO, their practical deployment remains constrained by the reliance on large volumes of high-quality annotated data, which are often expensive, time-consuming, and difficult to obtain [52]. Models that can maintain robust predictive performance when trained with sparse or weakly annotated datasets are essential for enabling scalable, cost-effective, and transferable solutions across diverse geographic regions and acquisition conditions [53]. **TO4** aims to investigate such models that are resilient to limited annotation availability.

1.3 Outline

This thesis follows a cumulative style, presenting the results of the conducted research as a collection of articles published in conferences and journals.

1.3.1 Thesis Structure

The thesis is organized as follows:

- **Chapter 1** introduces the research domain and describes the motivation behind this thesis. It defines the research objectives, presents the organization of the thesis and its constituent chapters, summarizes the associated publications, and outlines the principal scientific contributions. In addition, the chapter provides an overview of the developed datasets and code repositories, as well as the research projects and dissemination activities that supported and resulted from this work.
- **Chapter 2** presents the background and related concepts relevant to this thesis. It discusses multi-source EO and benchmark datasets, introduces the fundamentals of LULC mapping with key challenges, reviews widely used global LULC products and *in-situ* reference datasets, and summarizes recent advances in Machine Learning (ML), DL, GFM, and sub-pixel LULC estimation. The chapter also highlights the role of High-Performance Computing in enabling scalable EO workflows, thereby providing the theoretical and methodological foundation for the research presented in this thesis.
- **Chapter 3** summarizes the related work upon which this thesis builds and extends. It reviews existing research corresponding to the four research objectives, including semantic understanding of satellite and *in-situ* observations, benchmark datasets for EO, foundation models, sub-pixel LULC estimation, and data-efficient DL methods. The chapter concludes by identifying the key research gaps addressed throughout this thesis.
- **Chapter 4** presents the conference and journal publications comprising this thesis in detail. For each publication, it discusses the research motivation, methodology, experimental evaluation, and principal findings, while highlighting how the individual contributions collectively address the research objectives defined in Chapter 1. The chapter includes both published works and manuscripts currently under review at peer-reviewed journals and conferences.
- **Chapter 5** concludes the thesis by synthesizing the principal findings and contributions across all research objectives and publications. It revisits the research questions and objectives presented in Chapter 1, discussing how they have been addressed through the proposed methodologies and their associated results. The chapter further reflects on the scientific and practical implications of the work,

outlines its current limitations, and identifies promising directions for future research. Finally, it discusses the broader relevance of the thesis in the context of next-generation EO systems, highlighting how the proposed datasets, semantic frameworks, scalable data infrastructures, and embedding-centric methodologies contribute to the emerging paradigm of foundation models, unified geospatial representations, and agentic EO.

1.3.2 Publications

The following publications, listed in chronological order, are part of the thesis. They are included as Appendix, and their summary and explanation can be found in Chapter 4.

Paper I: S. Sharma, R. Roscher, M. Riedel, S. M. Memon, and G. Cavallaro, “Improving Generalization for Few-Shot Remote Sensing Classification with Meta-Learning”, in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Kuala Lumpur, Malaysia, 2022, pp. 5061-5064, <http://doi.org/10.1109/IGARSS46834.2022.9884699>

Paper II: C. Paris, L. Martinez-Sanchez, M. van der Velde, S. Sharma, R. Sedona, and G. Cavallaro, “Accuracy assessment of land-use-land-cover maps: the semantic gap between in-situ and satellite data”, in *Proceedings Volume 12733, Image and Signal Processing for Remote Sensing XXIX, SPIE Remote Sensing*, Amsterdam, Netherlands, 2023, <http://doi.org/10.1117/12.2679433>

Paper III: S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Sen4Map: Advancing Mapping with Sentinel-2 by Providing Detailed Semantic Descriptions and Customizable Land-Use and Land-Cover Data”, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, vol. 17, pp. 13893 - 13907, 2024, <http://doi.org/10.1109/JSTARS.2024.3435081>

Paper IV: S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Enhancing Land Cover Mapping: A Novel Automatic Approach To Improve Mixed Spectral Pixel Classification”, in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Athens, Greece, 2024, pp. 1017-1020, <http://doi.org/10.1109/IGARSS53475.2024.10641976>

Paper V: S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Beyond Pixels: Identifying Built-Up Features at Sub-Pixel Level for Enhanced Satellite-Based Land Cover Mapping”, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, 2026, <http://doi.org/10.1109/JSTARS.2026.3712773>

Paper VI: S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Diagnosing

the Ground–Satellite Gap in Land Cover Using Vision Foundation Models and Field Photos”, in *Proceedings of the IEEE India Geoscience and Remote Sensing Symposium (InGARSS)*, 2026 (accepted)

1.4 Contributions

1.4.1 Research Publications

All papers presented in Chapter 4 relate to the TOs, as shown in Table. 1.1. Each paper has one principal objective which is marked by filled circles while open circles denote supporting or enabling contributions.

To address the challenge of the semantic gap between field observations and satellite imagery in LULC mapping, and to address **TO1**, special attention was put on identifying and overcoming the sources of uncertainty in LULC accuracy assessment. In **Paper II**, the semantic mismatch between *in-situ* observations and satellite-derived LULC data was identified, and the street-level imagery was used as an intermediate representation to bridge this gap. A follow-up was done in **Paper VI**, where the main emphasis was put on diagnosing the inconsistencies/semantic gap between satellite and *in-situ* data, with the help of state-of-the-art Vision Foundation Models (VFM). To further improve the performance of LULC classification systems, attention was put on developing large-scale and multi-source dataset, accomplishing **TO2**. In **Paper III**, the Sen4Map¹ benchmark dataset was produced. It is a large-scale benchmark dataset that combines Sentinel-2 time-series imagery with detailed semantic annotations derived from the LUCAS *in-situ* survey. The dataset provides customizable hierarchical LULC labels providing extensive flexibility. Additionally, baseline evaluations using traditional ML and transformer-based DL models were presented, demonstrating the effectiveness of spatio-temporal architectures for LULC mapping. This dataset is further used in Paper VI, thereby also contributing to **TO1**. Since the study of heterogeneous and mixed landscapes forms an integral part of LULC classifications, special attention was put on enhancing LULC mapping for mixed pixels in **Paper IV**. This was the motivation for **TO3**, where study was further extended for LULC mapping at a sub-pixel level. In **Paper V**, both methods and datasets were built to estimate fractional class compositions by exploiting spectral–spatial relationships and multi-scale contextual information, contributing to both **TO2** and **TO3**. And finally, to accomplish **TO4**, a thorough literature review on the state-of-the-art transfer learning and meta-learning methods for Few-shot remote sensing image classification was done. This was the main motivation for **Paper I**. Special attention was put on meta-learning, self-supervised, and semi-supervised learning approaches along with the available benchmark datasets to improve generalization on the unseen samples with limited annotation availability.

¹<https://datapub.fz-juelich.de/sen4map>

Table 1.1. Relationship between thesis objectives and papers. Filled circles identify each paper’s principal thesis objective; open circles denote supporting or enabling contributions. Each paper has one principal objective.

	Paper I	Paper II	Paper III	Paper IV	Paper V	Paper VI
TO1		●	○			●
TO2			●		○	
TO3				●	●	
TO4	●					

The graphical representation of the thesis objectives and the temporal flow of the thesis with publications detailing the period from the beginning to the completion of the doctoral studies is shown in Fig. 1.2, following Business Process Model and Notation (BPMN).

1.4.2 Code and Data Repositories

The following code and data repositories have been published:

- Few-shot Remote Sensing Classification with meta-learning (**Paper I**):
<https://gitlab.jsc.fz-juelich.de/sdlrs/FSL>
- Sen4Map Data and Code (**Paper III**):
<https://datapub.fz-juelich.de/sen4map>
- Beyond Pixels: Fractional Data and Code (**Paper V**):
<https://gitlab.jsc.fz-juelich.de/sdlrs/sub-pixel-mapping>
- Spatial-Semantic Diagnostic Framework (**Paper VI**):
[https://gitlab.jsc.fz-juelich.de/sdlrs/agrosemanticingarss,tag v2.1.1](https://gitlab.jsc.fz-juelich.de/sdlrs/agrosemanticingarss,tag%20v2.1.1)

1.4.3 Additional Activities

During the PhD, the following contributions were consistently made to the project Adaptive Multi-Tier Intelligent Data Manager for Exascale (ADMIRE) which received funding from the European Union’s Horizon 2020 Joint Technology Initiative (JTI)-European High-Performance Computing Joint Undertaking (EuroHPC) research and innovation programme under the grant agreement no. 956748, as part of Work Package (WP) 7: Applications “Remote Sensing”:

- As the representative of WP7, the RS application is provided in the project to co-design the novel high-performance I/O technologies.

- Regular profiles and traces were generated to support the integration of application with WP5 and WP6 artifacts using Score-P, TAU, and Darshan.
- In order to make the RS application malleable and easy integration with WP3 objectives, Elastic Horovod was introduced for the application.
- In addition, significant efforts are being made for the identification of “the safe-reconfiguration regions/control points” in the RS application for integration with the Intelligent Controller.

1.4.4 Dissemination

A number of dissemination activities related to the presented research have also been performed, which are summarized in the following:

- Online Tutorial on Message Passing Interface (MPI) in the course “Artificial Intelligence for Earth Observation powered by High Performance Computing and Quantum Computing”, UNIVERSITÀ DEGLI STUDI DEL SANNIO, Benevento, Italy, 2021
- Webinar for “Remote Sensing” use-case for ADMIRE application (2022), <https://youtu.be/PMYek3VZNB8?si=ZZ1cWHqd6OI6XAsI>
- Lecture at CRS, Iceland (2022): “Few-shot Remote Sensing Classification with Meta-Learning: Challenges and Opportunities”, https://crs.hi.is/?page_id=295
- Presented at 12th IHPC Workshop, University of Iceland, 2023: “Enhancing LULC classification systems: Harnessing large-scale satellite and street-level geotagged images” <https://www.ihpc.is/events/12th-icelandic-hpc-community-workshop>
- Teaching assistance and lectures in the course “Machine Learning for Earth Observation powered by Supercomputers”, REI506M University of Iceland, 2023 and 2024

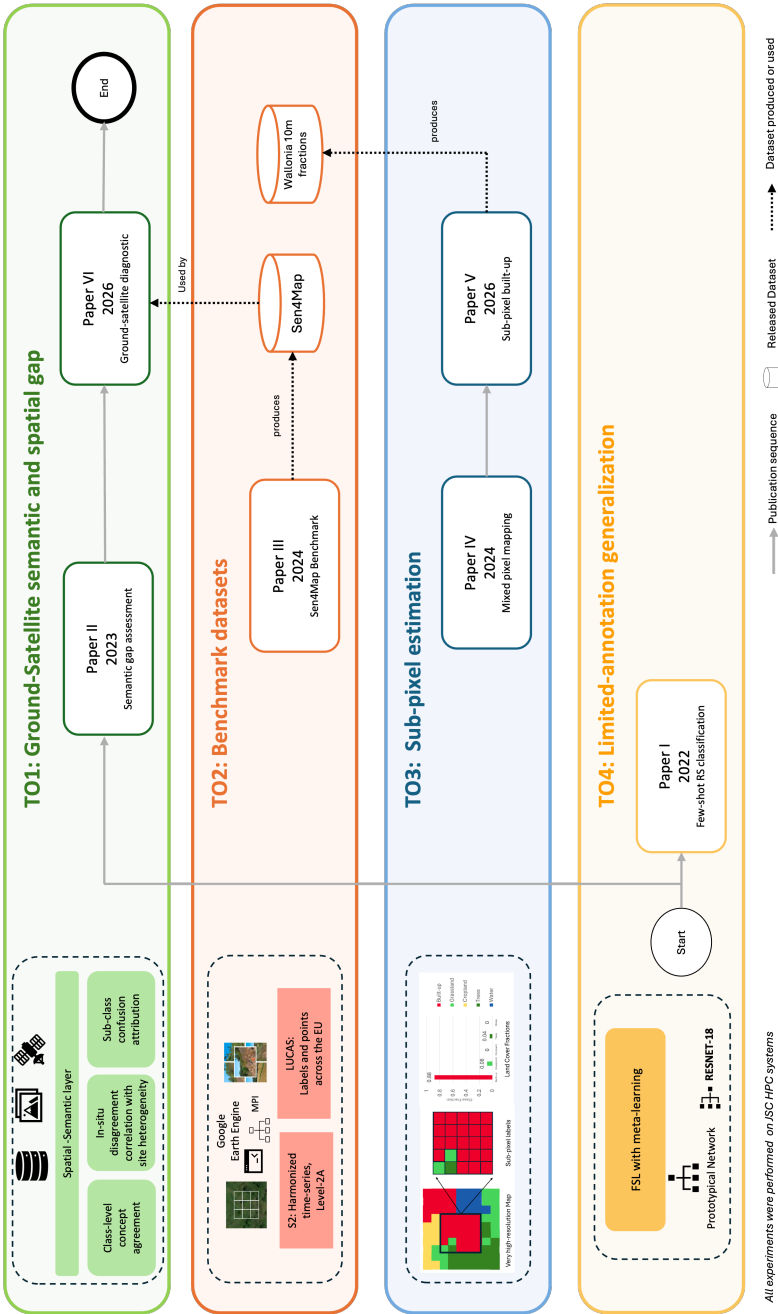


Figure 1.2. BPMN flow diagram of the thesis with graphical representation of thesis objectives and temporal flow of the thesis with publications and objectives.

2 Background

This chapter provides the necessary background to contextualize the research gaps addressed throughout this thesis. In Section 2.1, multi-source EO datasets for mapping LULC are introduced. Section 2.2 presents an overview of satellite-based LULC classification and Section 2.3 provides a brief explanation of the key challenges in satellite-based LULC mapping. Additionally, large-scale DL is discussed in 2.4.

2.1 Multi-source EO Data

EO satellite missions provide continuous, large-scale observations of the Earth’s surface, forming the foundation of modern RS applications [2]. Satellite sensors continuously acquire observations over regional to global extents, providing systematic, repetitive, and geographically consistent measurements of land, oceans, and the atmosphere. Modern satellite missions employ a variety of sensing technologies, including multispectral, hyperspectral, thermal, and Synthetic Aperture Radar (SAR) sensors, each capturing complementary physical properties of the Earth’s surface [54]. Optical satellites such as the NASA/United States Geological Survey (USGS)’s Landsat-9 [55] and the European Space Agency (ESA)’s Copernicus Sentinel-2 [29] measure reflected solar radiation across multiple spectral bands, enabling the discrimination of vegetation, water, soil, and built-up areas based on their spectral signatures [56]. The increasing availability of freely accessible satellite imagery, combined with improvements in spatial, spectral, radiometric, and temporal resolution, has established satellite acquisition as the cornerstone of various RS applications [10].

The Sentinel-2 constellation has become one of the most widely used sources of optical satellite imagery for environmental monitoring and LC analysis [50]. The mission consists of two polar-orbiting satellites equipped with the MultiSpectral Instrument (MSI), providing global observations with a revisit time of approximately five days at the equator [11]. Sentinel-2 acquires imagery in 13 spectral bands spanning the visible, near-infrared, and shortwave infrared regions of the electromagnetic spectrum, with spatial resolutions of 10, 20, and 60 m [29, 57]. This combination of relatively high spatial resolution, broad spectral coverage, frequent temporal revisit, and free and open data access has made Sentinel-2 the standard data source for numerous EO applications, including vegetation monitoring, crop mapping, forestry, disaster

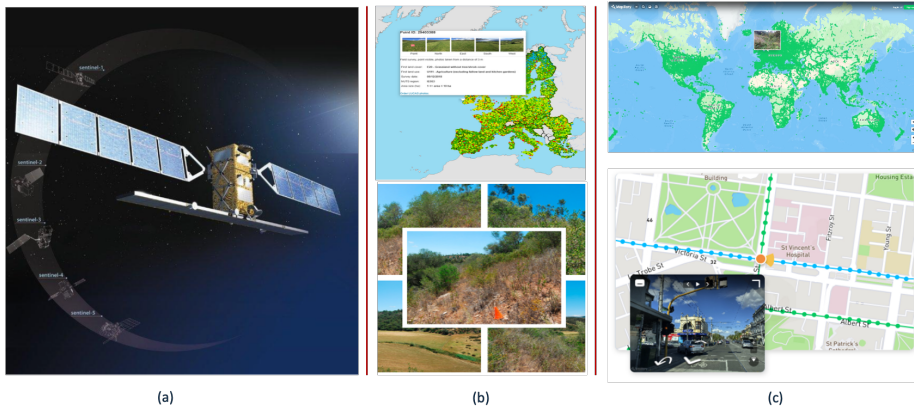


Figure 2.1. Data from multiple sources: (a) Satellite data [59], (b) In-situ data [31], and (c) Crowd-sourced data [13].

management, and LULC classification [58]. The continuous acquisition of multispectral time-series observations has created unprecedented opportunities to monitor landscape dynamics, detect environmental changes, and generate consistent, repeatable geospatial information at continental and global scales [16]. Alongside satellite imagery, large-scale *in-situ* datasets, including systematically collected field surveys and crowdsourced geo-tagged street-level imagery, provide valuable ground-level observations that cannot be directly captured from space [20], see Figure 2.1.

The *in-situ* data provide direct example of information about the physical characteristics of the Earth’s surface [3]. These data are typically collected through field surveys, environmental monitoring campaigns, geo-tagged photographs, and other ground-based measurements [54]. Unlike satellite imagery, which observes the Earth’s surface from an overhead perspective, *in-situ* observations capture objects from a human viewpoint and often contain detailed semantic information regarding vegetation species, land management practices, infrastructure, and local environmental conditions [3]. Such data play a crucial role in validating satellite-derived products, constructing reference and benchmark datasets, and supporting the development, training, and evaluation of a wide range of ML and DL approaches [60, 61]. However, field observations are generally sparse, labour-intensive to acquire, geographically uneven, and collected at different spatial and temporal scales than satellite imagery [28, 62, 63].

One of the largest publicly available sources of standardized ground reference data in Europe is the Land Use/Cover Area frame Survey LUCAS [12]. Coordinated by Eurostat, LUCAS is an *in-situ* survey conducted every three years across European Union (EU) member states to collect harmonized information on LULC [31]. Surveyors visit hundreds of thousands of systematically selected sampling locations across Europe and record standardized LULC observations following a harmonized survey protocol [65]. At each location, six geo-referenced field photographs are acquired, comprising four landscape photographs facing the cardinal directions—North (N), East (E), South

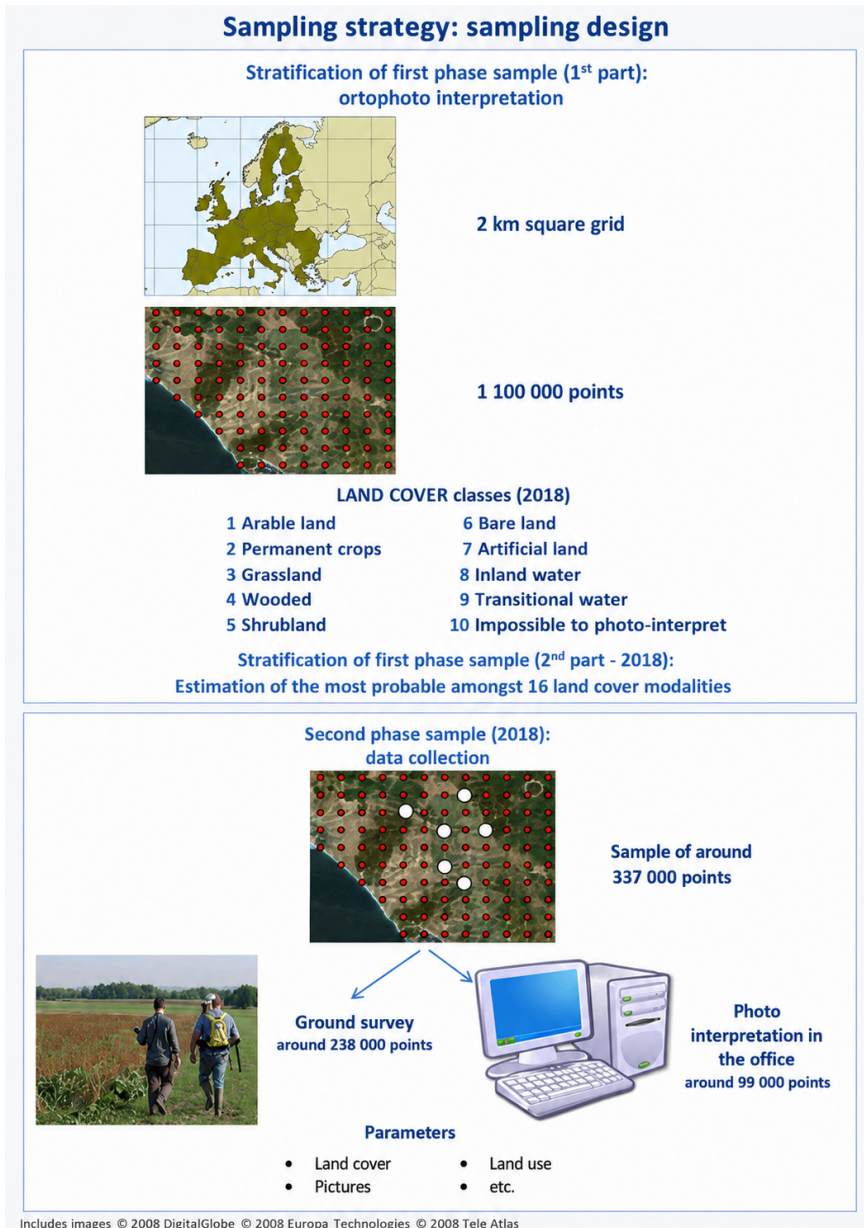


Figure 2.2. Overview of the LUCAS sampling design from the year 2018. The first phase involves a systematic stratified sampling over a 2 km grid covering Europe, generating approximately 1.1 million points for LC interpretation using orthophoto imagery. The second phase includes field-based surveys and photo-interpretation, resulting in around 337,000 validated in-situ observations with detailed LC and LU labels. This hierarchical sampling strategy ensures spatially consistent and statistically representative ground reference data for EO applications [31, 64].

(S), and West (W)—together with a Point (P) photograph centered on the survey location and a close-up Cover (C) photograph documenting the observed LULC element [31, 66]. These images are complemented by detailed ancillary observations, including standardized LULC classifications, environmental characteristics, land management information, and other site-specific metadata describing the surrounding landscape [67]. The LUCAS survey follows a hierarchical and statistically designed sampling strategy to ensure spatially representative *in-situ* observations across Europe [64, 65]. As illustrated in Fig. 2.2, the first phase employs a systematic stratified sampling framework based on a 2 km grid, enabling large-scale coverage and initial LC interpretation using high-resolution orthophoto imagery. This step generates a dense set of candidate sampling points across diverse LC classes. In the second phase, a subset of these points is visited for field-based surveys and photo-interpretation, resulting in a refined set of validated observations with detailed LULC labels. This two-stage design ensures both spatial consistency and statistical robustness, making LUCAS a highly reliable source of ground reference data for RS-based LULC mapping.

Furthermore, the LUCAS survey adopts a hierarchical classification scheme for both LULC, enabling observations to be described at multiple levels of semantic detail while maintaining consistency across Europe [68]. The LC hierarchy consists of three nested levels, namely LC1, LC2, and LC3, where each successive level refines the classification of its parent category [67]. The first level, LC1, defines broad LC categories describing the dominant physical cover observed at the sampling location. These categories include artificial land, cropland, woodland, shrubland, grassland, bare land, water, and wetlands. LC1 therefore provides a generalized representation of the Earth’s surface that is suitable for continental-scale analyses and harmonization across countries. The second level, LC2, subdivides each LC1 category into more specific LC types. For example, the cropland class is further differentiated into cereals, root crops, industrial crops, and other agricultural categories, while woodland is divided according to forest composition, such as broadleaved, coniferous, or mixed forests. Similarly, artificial land is refined into residential, industrial, transport, and other developed LC types. The third level, LC3, provides the highest level of thematic detail by further distinguishing subclasses within each LC2 category. These fine-grained labels capture detailed vegetation types, crop species, forest characteristics, and other specific LC attributes observed during the field survey [69]. Consequently, the hierarchical structure allows users to perform analyses at varying semantic resolutions, ranging from broad continental LC classes to highly detailed local descriptions [19].

In addition to LC information, LUCAS also records hierarchical LU annotations through LU1, LU2, and LU3 levels [68]. Unlike LC, which describes the physical material present on the Earth’s surface, LU characterizes the socio-economic function or human activity associated with that land, such as agriculture, forestry, residential use, recreation, or transportation [67]. The simultaneous availability of hierarchical LULC annotations makes LUCAS a unique benchmark for studying both the physical characteristics and functional use of landscapes [31]. The dataset is extensively used for validating satellite-derived products, training ML and DL models, and investigating the relationship between ground-level observations and satellite imagery [70, 71, 72,



Figure 2.3. Example of a large-scale global LC product: ESA WorldCover 2020 [79] generated from satellite observations at 10m spatial resolution, supporting applications in environmental monitoring, climate studies, and sustainable land management.

73]. Nevertheless, LUCAS samples only a sparse set of locations and therefore cannot provide exhaustive reference information for every satellite pixel, highlighting one of the fundamental challenges in large-scale EO analysis [43].

In addition to field surveys, OpenStreetMap (OSM) provides an important source of complementary geospatial information [74, 75, 76]. As a collaboratively maintained global geographic database, OSM provides detailed vector information describing roads, buildings, railways, water bodies, LU polygons, and numerous other geographic features contributed by volunteers worldwide [74]. Although OSM exhibits varying levels of completeness and quality across different regions, its semantic richness has made it an increasingly valuable resource for generating training samples, refining classification results, and incorporating contextual geographic knowledge into LULC mapping workflows [77, 78]. Effectively exploiting these heterogeneous datasets presents a significant opportunity to develop more accurate, scalable, and transferable LULC mapping systems [42, 75, 78].

2.2 Satellite-based LULC Classification

LULC mapping is one of the most fundamental and widely investigated applications of RS [54]. While LC describes the physical materials present on the Earth's surface, such as forests, croplands, water bodies, grasslands, or built-up areas, LU refers to the manner in which humans utilize these LC types for activities such as agriculture, transportation, residential development, or industrial production [80]. Accurate and up-to-date LULC information provides the foundation for numerous scientific investigations and policy decisions, including ecosystem monitoring, carbon accounting, biodiversity assessment, food security, urban planning, hydrological modelling, and climate adaptation strategies [8, 9, 81]. As anthropogenic pressures and environmental changes continue to reshape landscapes worldwide, timely and reliable LULC maps have become increasingly important for understanding land dynamics and supporting evidence-based decision making [8, 54].

The field of satellite-based LC mapping has evolved substantially over the past three decades, driven by continuous advances in EO technologies and computational methods [50, 82]. During the 1990s, large-scale LC products primarily relied on coarse-resolution satellite sensors which provided near-daily global coverage, such as the Moderate Resolution Imaging Spectroradiometer (MODIS) at a spatial resolution of 250–1000 m; typically 500 m for LC products [83], ENVISAT's Medium Resolution Imaging Spectrometer (MERIS) at 300 m or Advanced Very High Resolution Radiometer (AVHRR) at approximately 1 km [84]. These datasets enabled the first continental and global LC inventories, but their relatively coarse spatial resolution limited the characterization of heterogeneous landscapes and fine-scale LC patterns [85]. The launch of higher-resolution optical missions, most notably the Landsat programme [86] and later the Copernicus Sentinel-2 constellation [11], marked a significant paradigm shift. These sensors provide multispectral imagery at spatial resolutions of 10–30 m, enabling the identification of considerably finer landscape structures while maintaining regular temporal revisit cycles [87, 88]. Recently, these missions enable the generation of standardized high-resolution LC maps of the Earth's surface at regional to global scales and serve as fundamental resources for environmental monitoring, climate modelling, ecosystem assessment, agricultural management, biodiversity conservation, and sustainable development [89]. Notable examples include the “Esri Land Cover” [90], “ESA WorldCover” [79], refer to Fig. 2.3, and “Dynamic World” [91], along with “Copernicus Global Land Cover” [92], “GlobeLand30” [93], and “FROM-GLC” products [52, 87, 94]. A comparison of widely available global LULC products is shown in Table 2.1

2.3 Challenges in Satellite-Based LULC Mapping

Despite significant advances in EO technologies, as mentioned in Section 2.2, satellite-based LULC mapping faces several key challenges. The following subsections examine

Table 2.1. Comparison of widely used global LULC products.

Product	Year	Sensor	Resolution	Classes
IGBP DISCover [95]	1997	AVHRR	1 km	17
UMD Land Cover [96]	1998	AVHRR	1 km	14
GLC2000 [97]	2003	SPOT-4 VEGETATION	1 km	22
MODIS MCD12Q1 [98]	2001–present	MODIS	500 m	17
GlobCover [99]	2004–2009	MERIS	300 m	22
Copernicus Global Land Cover [92]	2015–2019	PROBA-V	100 m	23
GlobeLand30 [93]	2000/2010/2020	Landsat TM/ETM+/OLI	30 m	10
FROM-GLC [87]	2010	Landsat TM/ETM+	30 m	10
FROM-GLC10 [52]	2017	Sentinel-2	10 m	10
ESA WorldCover [79]	2020/2021	Sentinel-1/2	10 m	11
Dynamic World [91]	2015–present	Sentinel-2	10 m	9
Esri Land Cover [90]	2017–present	Sentinel-2	10 m	9

these challenges in greater detail.

2.3.1 The Semantic Gap between Ground and Satellite Observations

A fundamental challenge in satellite-based LULC mapping is the “semantic gap” between satellite-derived spectral and spatial information and the detailed semantic categories used to describe real-world land surface conditions [21, 70]. Satellite sensors observe the Earth from an overhead perspective and measure reflected or emitted electromagnetic radiation across multiple spectral bands, whereas *in-situ* observations describe the landscape from a human perspective using semantic concepts such as “forest”, “grassland”, “cropland”, or “urban” [20]. The two data sources provide fundamentally different representations of the same environment: satellite imagery captures physical properties over an area, while field observations describe semantic characteristics at specific locations [21]. Consequently, translating satellite measurements into LC information requires an abstraction from a continuous, heterogeneous physical reality to a discrete semantic representation, where each image pixel is commonly assigned a single hard class label despite potentially containing multiple land-cover components [33, 36, 81].

This discrepancy extends beyond differences in sensing modality and also encompasses differences in spatial scale, perspective, and contextual information [51, 73]. A field observation typically represents a single point or object, whereas a satellite pixel may cover tens or hundreds of square metres and therefore contain multiple LC types [20]. Moreover, LC classes are inherently semantic concepts whose interpretation often depends on local context and observation scale [13]. For example, a field survey may classify a location as “Artificial land”, while the corresponding satellite pixel may be dominated by nearby cropland, grassland, or exposed soil following harvest and, therefore, exhibit spectral characteristics similar to cropland or bare ground, as can be

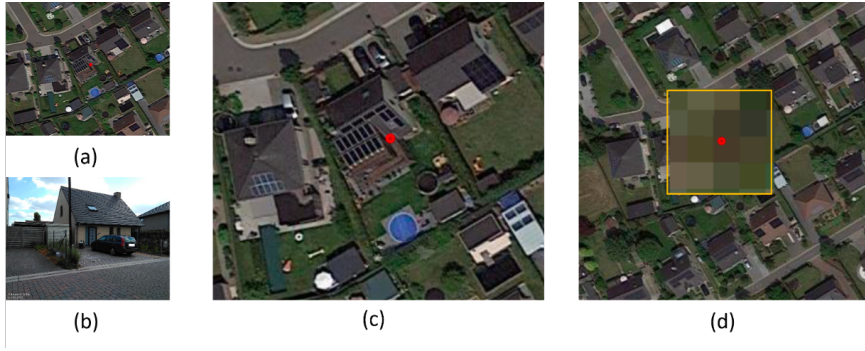


Figure 2.4. Example of LUCAS point labeled as “Artificial Land” (a) High-resolution Google Earth Satellite View, (b) LUCAS street-level Point image, (c) zoomed-in view of the residential area containing LUCAS point, (d) Sentinel-2 image located in the LUCAS point. [Extracted from Paper IV [100] © 2024 IEEE.]

seen from Figure 2.4. Conversely, two satellite patches with nearly identical spectral appearance may correspond to different LC labels because they represent distinct land uses or were acquired under different seasonal or environmental conditions [69].

These semantic and spatial discrepancies have important implications for supervised learning and model evaluation [63]. Apparent classification errors do not necessarily indicate deficiencies in the model itself; rather, they may arise from inconsistencies between ground reference labels and the information actually observable in satellite imagery [101]. Differences in annotation protocols, temporal acquisition, spatial support, or semantic definitions can all produce disagreements between predicted and reference labels, even when the satellite observations are visually indistinguishable [62, 102]. Understanding these discrepancies is therefore essential for correctly interpreting model performance, developing robust multimodal learning frameworks, and constructing reference datasets that more faithfully represent the relationship between ground observations and satellite imagery [21, 26, 82].

2.3.2 Landscape Heterogeneity and Mixed Pixels

The most inherent limitation of satellite-based EO is that the spatial resolution of a sensor rarely matches the complexity of the real-world landscape it observes [103]. Natural and human-modified environments are inherently heterogeneous, with multiple LC types often occurring within areas smaller than a single satellite pixel [34, 36]. Consequently, the recorded spectral response frequently represents a combination of several surface materials rather than a single homogeneous LC class, giving rise to the “mixed pixel problem” [104, 105]. The occurrence of mixed pixels is particularly common in fragmented urban areas, agricultural mosaics, forest edges, wetlands, and other transitional landscapes, where abrupt or gradual changes in LC occur over short

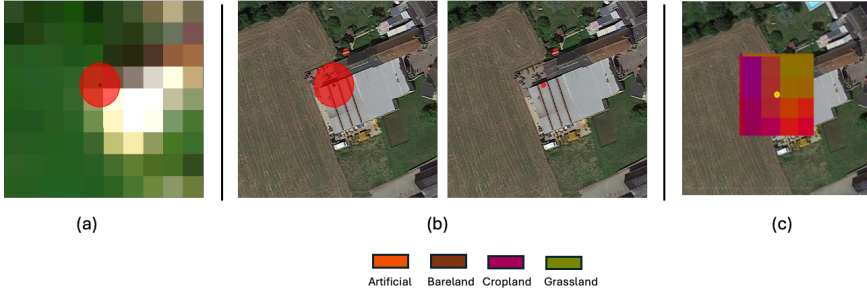


Figure 2.5. Figures (a) and (b) show the Sentinel-2 and High-resolution Google Earth Satellite View for a LUCAS point with “LC1” as “Artificial Land”. Figures (c) shows the presence of multiple classes/endmembers via Abundance Map in heterogeneous environment, as obtained from the Sentinel-2 yearly composite. [Extracted from Paper IV [100] © 2024 IEEE.]

distances [35, 36]. Even with modern sensors such as Sentinel-2, which provide imagery at spatial resolutions of 10–20 m, a single pixel may simultaneously contain multiple LC classes, making it difficult to assign a single label that accurately represents the underlying landscape [32, 39]. For example, a pixel located at the boundary between an “urban area” and a “cropland” may contain “buildings”, “trees”, “grass”, “bare land”, “crops” and “paved surfaces” in varying proportions, as depicted in Figure 2.5. Assigning each pixel solely to its dominant class discards valuable information about its sub-pixel composition and can create inconsistencies between satellite-derived labels and finer-scale reference observations used for validation. For example, a reference observation may correctly identify a “building” or “narrow river” that occupies only a small fraction of the corresponding satellite pixel; expecting a hard pixel-based classifier to assign that pixel to the feature’s class is therefore inappropriate without explicitly modeling its sub-pixel composition [39, 106].

Traditional LULC mapping approaches typically formulate the problem as a hard classification task, assigning each pixel to one discrete LC class [54]. This implicitly assumes that every pixel is homogeneous and belongs entirely to a single category [7]. While this assumption may be appropriate for large, homogeneous regions, it becomes increasingly problematic when the spatial resolution of the imagery is coarse relative to the scale of landscape heterogeneity, as individual pixels may contain multiple LC types contributing to the observed spectral signal [38, 104]. This limitation has traditionally been addressed in coarse-resolution LC products either through fractional representations of LC or through legend classes that semantically combine multiple LC components [107, 108].

The mixed pixel problem therefore reflects a fundamental mismatch between the spatial resolution of satellite observations and the spatial complexity of real-world landscapes [104, 106]. Although higher-resolution sensors can partially alleviate this limitation, they are often constrained by data availability, revisit frequency, and acquisition costs

[34]. Consequently, considerable research has focused on “sub-pixel analysis”, which seeks to estimate the proportion of different LC classes within each pixel rather than assigning a single dominant label [107]. By explicitly modeling landscape heterogeneity, sub-pixel approaches provide richer and more realistic representations of the Earth’s surface, enabling more detailed environmental analyses while fully exploiting the global coverage and frequent revisit capabilities of both medium and high-resolution satellite missions [39, 40, 106, 109].

2.3.3 Reference Data Limitations

A major bottleneck in satellite-based LULC mapping is the scarcity of high-quality reference data [53]. The need to generate timely and regularly updated LULC maps that capture ongoing landscape changes necessitates a sustained supply of representative reference data, highlighting the need for scalable and sustainable approaches to reference-data generation [43]. However, collecting *in-situ* observations requires extensive field campaigns that are labor-intensive, time-consuming, and costly, making it infeasible to obtain exhaustive annotations over large geographic regions [61]. As a result, reference datasets provide labels for only a small fraction of the Earth’s surface [12]. Beyond their limited availability, reference data are also affected by differences in spatial scale [110]. In addition, *in-situ* observations describe conditions at a specific location, whereas satellite sensors record the aggregated spectral response over an area corresponding to the sensor’s spatial resolution [62]. For example, a Sentinel-2 pixel with a spatial resolution of 10 m may encompass “vegetation”, “bare soil”, “roads”, and “small built structures” simultaneously, while the corresponding ground observation records only the dominant LC class at the sampling point. Consequently, the reference label may not fully represent the heterogeneous composition observed by the satellite sensor, particularly in fragmented landscapes or at class boundaries, or vice versa [33].

Temporal inconsistencies introduce an additional source of uncertainty [62]. Field surveys and satellite acquisitions are rarely collected simultaneously, and LC conditions may change substantially between observation dates due to seasonal vegetation dynamics, agricultural practices, urban development, or natural disturbances [111]. For example, a location labeled as “cropland” during a field campaign may correspond to a “harvested field”, “emerging vegetation”, or “exposed soil” in the available satellite image, depending on when the satellite observation was acquired. Such temporal mismatches can alter the spectral appearance of the landscape without changing its semantic label, increasing the difficulty of learning robust relationships between satellite observations and reference annotations [102].

Finally, reference datasets are themselves subject to annotation uncertainty [112]. Positional inaccuracies, differences in survey protocols, interpreter subjectivity, and inconsistencies in LC definitions can all contribute to disagreements between reference labels and satellite observations [111]. Such limitations have therefore motivated the development of label-efficient learning paradigms, including Few-Shot Learning (FSL) [113], semi-supervised learning [114], and transfer learning [115], which seek to maximize

model performance while minimizing the reliance on extensive manually annotated datasets [52].

2.3.4 Benchmark Data Limitations

Addressing the above challenges requires the development of advanced DL models capable of learning from heterogeneous observations, capturing fine-scale landscape complexity, and generalizing across diverse conditions [17]. However, developing and evaluating such models depends on the availability of high-quality benchmark datasets that adequately represent the complexity and variability of real-world landscapes [21, 26, 43]. Existing benchmark datasets often employ relatively coarse classification schemes and rarely provide detailed, harmonized representations of both LC and LU, limiting their capacity to support the development of models that can capture the full semantic complexity of real-world landscapes [116, 117]. Consequently, they offer limited support for learning rich multimodal representations, characterizing LU at a detailed semantic level, and developing foundation models capable of integrating complementary EO data sources [118, 119]. These limitations highlight the need for benchmark datasets that combine large-scale satellite observations with detailed, consistently defined, and independently collected reference information, thereby providing a more comprehensive foundation for developing and evaluating next-generation DL models for LULC mapping [56].

2.4 Large-Scale Deep Learning for Earth Observation

The expanding access to large-scale EO data has been accompanied by significant advances in DL methodologies and computational capabilities [16, 42, 44]. The emergence of DL has substantially transformed LULC mapping by enabling models to learn hierarchical representations directly from EO data, reducing reliance on handcrafted spectral and spatial features and improving the ability to capture complex spatial, spectral, and semantic patterns [17]. Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and other deep architectures have significantly improved the ability to model complex landscapes, while modern computational infrastructures have enabled their training on continental and global-scale datasets that were previously computationally infeasible [120, 121, 122]. Coupled with advances in High-Performance Computing (HPC), modern DL architectures can now be trained on massive multi-source EO datasets, facilitating large-scale, data-driven analysis with improved accuracy, scalability, and generalization [15, 82]. These developments have paved the way for increasingly sophisticated models, culminating in the emergence of foundation models that learn transferable representations from vast collections of unlabeled EO data [123, 124].

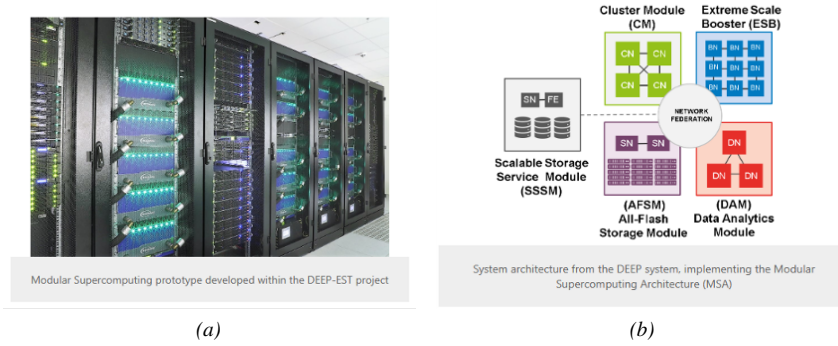


Figure 2.6. *Modular Supercomputing Architecture from DEEP Systems (DEEP-EST)* [126]

2.4.1 High-Performance Computing for Large-Scale LULC Mapping

The accelerated rise in EO data has transformed satellite-based LULC mapping into a highly data-intensive and computationally demanding task [125]. Over the past decade, open satellite missions such as Sentinel and Landsat have enabled continuous global monitoring of the Earth’s surface, generating petabytes of multispectral and multitemporal imagery [59, 86]. In parallel, the increasing availability of VHR satellite imagery, airborne observations, crowdsourced geographic information, and ancillary geospatial datasets has significantly expanded both the volume and diversity of available data [4, 14, 120]. These heterogeneous data sources provide vast opportunities for detailed characterization of the Earth’s surface, but also introduce substantial challenges related to data storage, management, processing, and analysis, particularly in integrating complementary information across different modalities, including satellite imagery, photographs, and other geospatial data.[82].

To tackle these challenges, modern satellite-based LULC mapping increasingly relies on HPC infrastructures that combine distributed computing, parallel processing, high-speed storage systems, and dedicated hardware accelerators such as graphics processing units (GPUs) [126]. These infrastructures enable scalable data preprocessing, feature extraction, model training, hyperparameter optimization, and large-scale inference, substantially reducing computational time while improving reproducibility and resource efficiency [127, 128]. More recently, modular supercomputing architectures have further advanced HPC by integrating heterogeneous compute modules, including Central Processing Unit (CPU) clusters, Graphics Processing Unit (GPU) accelerators, and data-centric modules, interconnected through high-bandwidth networks [44, 126, 129]. A modular supercomputing architecture is an HPC system composed of multiple interconnected computing modules, each optimized for different types of workloads [130]. By combining heterogeneous resources such as CPU- and GPU-based modules within a unified architecture, it enables efficient execution of diverse scientific applications

while improving scalability and resource utilization [131]. An example of modular supercomputing architecture from Dynamical Exascale Entry Platform (DEEP) systems, i.e., Dynamical Exascale Entry Platform–Extreme Scale Technologies (DEEP-EST) is shown in Figure 2.6. This modular design allows individual workflow components to execute on the most suitable hardware resources, improving scalability and performance for increasingly complex, data-intensive EO applications [132, 133]. Consequently, HPC has become an indispensable component of modern EO pipelines, supporting the development, training, and deployment of advanced ML and DL models for large-scale LULC mapping [127, 134].

2.4.2 Foundation Models for Earth Observation

Recent advances in foundation models have further shifted the focus of EO research from task-specific models toward learning general-purpose representations that can be adapted to diverse downstream applications with limited supervision [135]. Similar to VFMs developed for natural images, GFMs are large-scale pretrained models designed to learn general-purpose representations of the Earth’s surface from massive and diverse collections of satellite observations [123]. They are typically trained using self-supervised learning objectives, allowing the models to learn meaningful representations directly from the data without requiring extensive manually labelled training samples [47, 136]. Rather than relying solely on raw multispectral observations for downstream tasks, these models learn dense feature representations, commonly referred to as *embeddings*, that capture high-level semantic and contextual information about the observed landscape [137], see Figure 2.7. By exploiting spatial, spectral, temporal, and geographic relationships within the data, these models learn transferable representations that generalize across sensors, regions, and downstream tasks, reducing the dependence on large labeled datasets [118].

GFMs have developed along several complementary directions, differing in their input modalities, pretraining strategies, architectures, and learned representations [123]. Recent examples include SatMAE [138], Prithvi [139], Galileo [140], TerraMind [141], and AlphaEarth Embeddings (AEDs) [142], which have demonstrated strong performance across a wide range of EO applications. Models such as SatMAE and Prithvi primarily investigate representation learning from multispectral and multitemporal satellite imagery, using self-supervised pretraining to learn transferable representations for downstream EO tasks [138, 139]. More recent multimodal GFMs, including Galileo and TerraMind, seek to integrate a broader range of EO modalities and learn representations that capture relationships across optical imagery, SAR, temporal observations, and other geospatial variables [140, 141]. Within this broader landscape, AEDs introduces a distinct geospatial embedding-field paradigm that learns a continuous representation of the Earth’s surface from large-scale, multimodal observations [142]. These embeddings provide compact representations that can subsequently be used for downstream tasks such as LULC classification, change detection, and environmental monitoring, reducing the need to learn task-specific representations from raw EO data [18]. This formulation

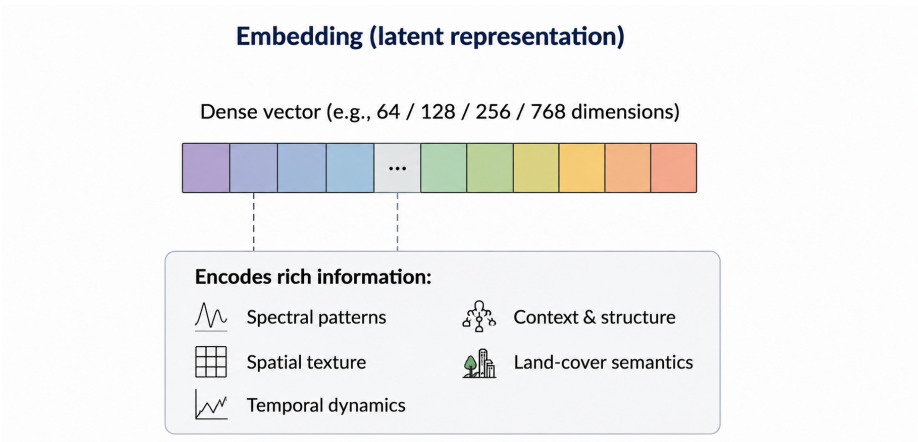


Figure 2.7. Schematic view of an embedding. A pretrained encoder compresses the observations of a geographic location into a dense d -dimensional latent vector that jointly encodes spectral patterns, spatial texture, temporal dynamics, structural context, and LC semantics. Such embeddings serve as reusable, task-agnostic inputs for downstream LULC analysis.

is particularly relevant to large-scale geospatial analysis because it provides a common representation that can be reused across regions, time periods, and downstream mapping tasks without retraining the underlying representation model [137].

2.4.3 Efficient Data Management: HDF5 Dataset

Efficient storage and data management are equally important for large-scale learning [143]. Rather than storing each data modality separately, modern benchmark datasets frequently adopt hierarchical storage formats such as HDF5[144], which support chunked storage, compression, and fast random access [145]. Such designs substantially reduce Input/Output (I/O) overhead during model training and enable efficient loading of multimodal data on HPC systems [146]. Following this principle, the benchmark dataset introduced in this thesis adopts an HDF5-based architecture to organize time series of multispectral satellite imagery with detailed LULC annotations within a unified data structure, refer Figure 2.8. This design improves scalability, reproducibility, and interoperability while facilitating efficient training of large-scale DL models [147].

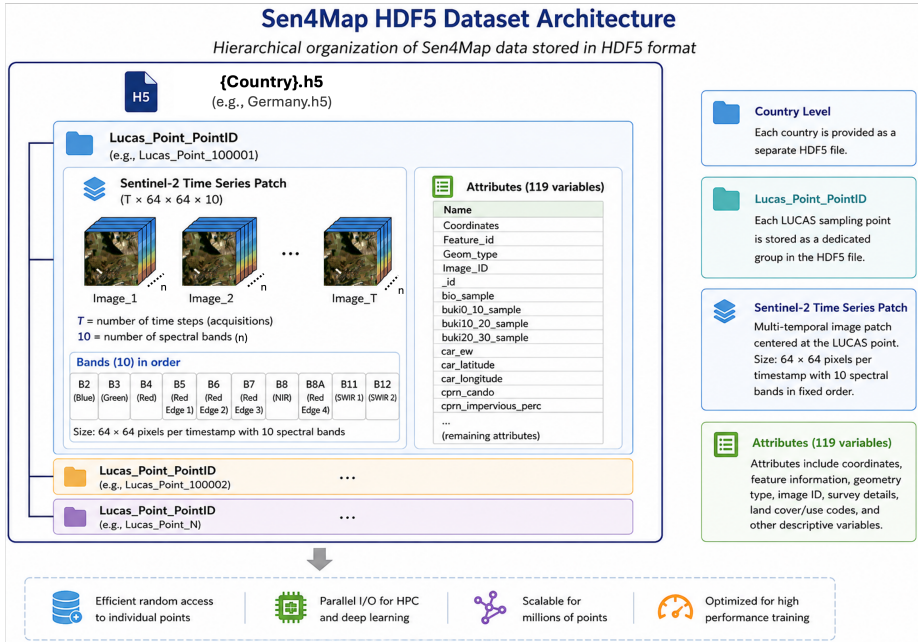


Figure 2.8. Sen4Map HDF5Dataset Architecture.

3 Related Work

This chapter summarizes the related work around the four main objectives that the thesis addresses in Section 3.1, 3.2, 3.3, and 3.4.

3.1 Bridging Spatial-Semantic Gaps Using Geo-Tagged Field Observations

The increasing availability of crowdsourced and geo-tagged ground-level imagery has stimulated growing interest in its use for generating reference data across a wide range of EO applications [148, 149]. Compared with traditional field surveys, street-level images offer a cost-effective and scalable means of collecting *in-situ* information over large geographic areas [148, 150]. For example, D’Andrimont et al. [61] proposed a framework to support the European Union’s Common Agricultural Policy (CAP) by validating farmers’ grassland declarations using Sentinel-1 and Sentinel-2 imagery. Parcels identified as potentially misclassified were verified using geo-referenced street-level photographs acquired with vehicle-mounted cameras, demonstrating a substantial increase in data collection efficiency compared with conventional field campaigns [61]. Similarly, publicly available street-level images from crowdsourcing platforms such as Mapillary and Google Street View have been exploited for crop-type monitoring by selecting the most informative images based on metadata such as camera orientation and focal length [13, 30, 60]. In addition, the Geo-Wiki platform, widely used for crowdsourced interpretation of satellite imagery for LC mapping, has been extended to support the visual interpretation of street-level photographs, further highlighting the growing role of ground-level imagery as a complementary source of reference information for EO applications [151, 152].

More recently, the role of geo-tagged field observations has expanded beyond reference data collection toward improving the understanding of the relationship between ground-level observations and satellite imagery [31, 154]. Unlike conventional *in-situ* surveys, which typically provide a single LC label at a geographic location, geo-tagged photographs preserve the visual appearance of both the sampled point and its surrounding landscape, offering substantially richer semantic context for interpreting satellite observations [155, 156], as can be seen from Figure 3.1.

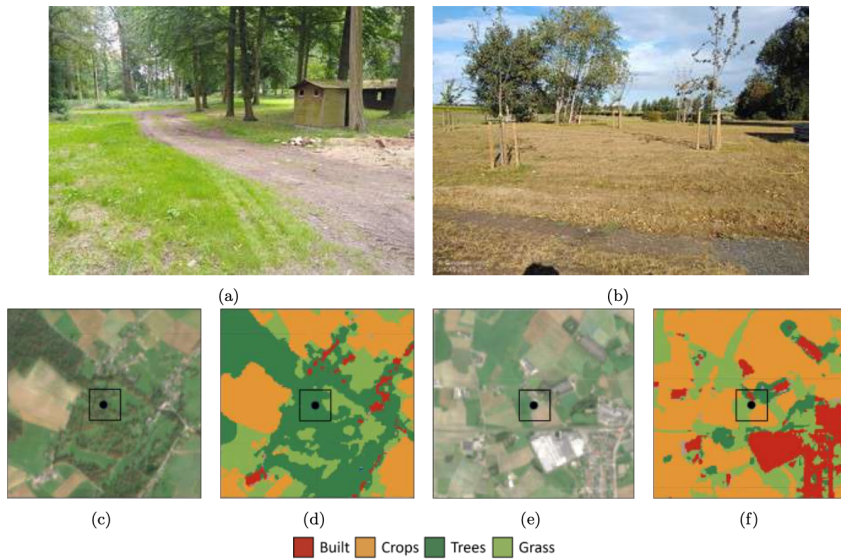


Figure 3.1. Examples of LUCAS in-situ samples associated with the “Woodland” LC class: (a)(b) LUCAS street-level images, (c)(e) Sentinel-2 satellite data, (d)(f) ESA WorldCover thematic product. The sample’s locations are highlighted in black. [Extracted from Paper II [153], © 2023 SPIE.]

Recent studies have demonstrated that incorporating this contextual information can improve the interpretation and validation of satellite-derived LC products by revealing semantic discrepancies that are difficult to identify using point-based annotations alone [69]. At the same time, advances in VFM, including both unimodal vision encoders and Vision-Language Models (VLMs), have enabled the learning of transferable semantic representations from large-scale image collections [136]. VLMs have recently emerged as a powerful paradigm for learning joint representations of visual and textual modalities [157]. These models align image and language embeddings in a shared latent space using contrastive learning, enabling zero-shot classification, cross-modal retrieval, and open-vocabulary understanding [158, 159]. In the context of RS, VLMs are particularly valuable for LULC mapping, where semantic ambiguity and limited labeled data pose significant challenges [160, 161].

Alongside vision-language models, unimodal self-supervised encoders form the second family of VFM relevant to this thesis. DINO demonstrated that ViTs trained by self-distillation on unlabeled images develop patch-level features that capture the spatial layout of a scene without any class supervision [162]. DINOv3 scales this recipe to larger models and corpora, yielding dense features that remain stable and discriminative across domains [163]. Because such representations describe image structure rather than assign categories, they provide a label-free basis for quantifying scene composition; in this thesis, they underpin the spatial-heterogeneity analysis of ground-level photographs in **Paper VI**.

Within the vision-language family, a foundational model is Contrastive Language–Image Pre-training (CLIP), which learns aligned representations by maximizing agreement between image-text pairs using a contrastive objective [164]. Given an image encoder $f_I(\cdot)$ and a text encoder $f_T(\cdot)$, CLIP optimizes similarity in a shared embedding space, enabling strong zero-shot generalization. However, its performance in RS is limited due to the domain gap between natural images and satellite imagery [165]. Still, several RS-specific VLMs have been proposed. RemoteCLIP introduces large-scale vision-language pretraining tailored to satellite imagery, leveraging curated image-caption datasets and improving semantic alignment for downstream EO tasks such as classification and retrieval [160]. Similarly, RS5M and GeoRSCLIP construct large-scale RS image-text datasets and fine-tune CLIP-based architectures, achieving significant improvements in zero-shot LULC classification and cross-modal retrieval tasks across multiple benchmarks [166].

More recently, SenCLIP [159] proposes a novel strategy for improving zero-shot LULC mapping by bridging satellite imagery with ground-level geotagged photographs. Instead of relying solely on generic prompts, SenCLIP enhances semantic alignment through ground-level prompting, enabling more expressive and context-aware representations for Sentinel-2 imagery [159]. This approach demonstrates that incorporating cross-view supervision significantly improves classification performance in geographically diverse environments. Building on temporal information, TimeSenCLIP [161] extends vision-language modeling to satellite time series by integrating spectral-temporal embeddings with ground-level semantic supervision. This enables improved modeling of phenological and seasonal variations in LULC classes, further enhancing robustness in dynamic environments [161]. Overall, these models provide a powerful framework for comparing heterogeneous visual modalities by projecting ground-level photographs and satellite imagery into a shared representation space, thereby facilitating the analysis of semantic similarities and differences across viewpoints [167].

Despite these advances, cross-view alignment has so far served almost exclusively as an end in itself, directed at improving predictive performance through training-data collection, product validation, or zero-shot classification [159, 160]. Its use as a means of interrogating the reference data, determining where and why *in-situ* annotations and satellite-derived products disagree, has received far less attention. In this diagnostic setting, the two VFM families play complementary roles: unimodal encoders characterize the internal structure of each modality, while VLMs relate them within a shared semantic space. This highlights the need for frameworks that leverage geo-tagged field observations to use shared representations for assessing reference-data reliability, ultimately improving the robustness, interpretability, and generalization of satellite-based LULC mapping. In this thesis, **Paper II** and **Paper VI** aim to characterize and bridge the semantic gap between ground-level observations and satellite imagery by exploiting geo-tagged field photographs. While **Paper II** investigates how street-level images can improve the interpretation and assessment of satellite-derived LC products by providing richer semantic context than conventional point-based reference data, **Paper VI** extends this direction by leveraging pretrained VFM, pairing a unimodal encoder (DINOv3) for spatial structure with a vision-language model (SenCLIP) for

semantic alignment across modalities. Together, these works demonstrate how geo-tagged field observations can be used not only for validation but also for understanding the relationships between ground-level and satellite perspectives, ultimately enabling more robust and semantically informed LULC mapping.

3.2 Large-Scale Benchmark Datasets for Detailed LULC Mapping

The rapid progress of DL has simultaneously increased the demand for large-scale benchmark datasets capable of supporting the training and evaluation of increasingly sophisticated models [125]. Several datasets have been developed for LC classification, including BigEarthNet [116], EuroSAT [117], Esri Land Cover [90], ESA WorldCover [89], and Dynamic World [91]. BigEarthNet and EuroSAT have become widely adopted benchmarks for image-level classification, whereas ESA WorldCover, Dynamic World, and Esri Land Cover provide pixel-level LC products derived from Sentinel imagery. ESA WorldCover was generated using Sentinel-1 and Sentinel-2 observations together with a Random Forests (RF) classifier trained on approximately 141,000 manually annotated samples, while Dynamic World and Esri Land Cover employ convolutional DL models trained using labels obtained through visual interpretation of Sentinel-2 imagery. More recently, the European CORINE Land Cover (CLC)+ product has extended high-resolution LC mapping by combining Sentinel-1 and Sentinel-2 time series with temporal DL models [92].

Consequently, there is a growing need for benchmark datasets that better reflect the complexity of real-world environments by providing fine-grained annotations, multimodal observations, and richer semantic context [119]. Such resources are critical for enabling detailed LULC understanding and for supporting the development of more accurate, data-efficient, and transferable DL models [82]. Therefore, **Paper III** introduces a large-scale, time-series benchmark dataset that combines multispectral satellite imagery with detailed LULC information obtained from ground-level observations spread across EU. By providing richer semantic descriptions, the dataset is specifically designed to facilitate the training and evaluation of next-generation DL models capable of producing more detailed, context-aware, and transferable LULC maps.

As shown in Table 3.1, several widely used benchmark datasets primarily support conventional LC classification using a relatively small number of semantic classes derived from satellite imagery or existing LC products [88, 169]. In contrast, Sen4Map combines multi-temporal Sentinel-2 imagery with detailed LUCAS annotations, providing richer semantic information and a hierarchical representation of LULC. This multi-source design facilitates the development of advanced DL models capable of learning complementary spectral, spatial, and semantic representations beyond traditional single-source classification benchmarks.

Table 3.1. Comparison of Sen4Map with widely used benchmark datasets for LULC mapping.

Dataset	Resolution	Scale	Classes	Label Source
Sen4Map (Paper III) [168]	10 m	Across EU	Hierarchical	LUCAS + geo-tagged photos
BigEarthNet [116]	10–60 m	590,326 patches	43 (CLC L3)	CORINE Land Cover
EuroSAT [117]	10 m	27,000 patches	10	Manual interpretation
ESA WorldCover [89]	10 m	Global	11	Manual VHR annotation
Dynamic World [91]	10 m	Global	9	Manual Sentinel-2 labels
Esri Land Cover [90]	10 m	Global	9	Manual Sentinel-2 labels

3.3 Sub-Pixel LULC Modeling for Fine-Scale Landscape Representation

Sub-pixel LULC mapping addresses the challenge of representing mixed LC composition within a single satellite pixel [38]. Early global and regional LC products derived from coarse-resolution sensors such as AVHRR, MODIS, and MERIS already incorporated fractional representations, as mixed pixels were an inherent consequence of their relatively large spatial footprints [40, 170, 171]. These early efforts demonstrated the feasibility of estimating LC proportions at continental and global scales, typically at spatial resolutions of around 1000 m or coarser, where sub-pixel variability dominates the observation process [172]. More recent work has extended this paradigm to medium-resolution sensors, including studies on global land characterisation using LC fractions at 100 m resolution, enabling more detailed spatial representation while still relying on relatively coarse observational scales [107].

With the advent of higher-resolution missions such as Sentinel-2, research has increasingly shifted toward estimating LC fractions at 10 m spatial resolution, enabling fine-grained characterization of heterogeneous landscapes [39]. This transition introduces new challenges, as the reduction in pixel size increases spatial detail but also demands more precise modeling of local variability and class mixtures [173, 174]. At the same time, temporal modeling approaches, such as those used in improving global LC fraction change mapping using temporal DL, have demonstrated that incorporating time-series information can significantly enhance the estimation of LC dynamics and fractional changes, particularly in environments with seasonal or phenological variability [175].

Traditional satellite-based LULC mapping relies on hard classification, where each pixel is assigned a single discrete class [54]. While effective in relatively homogeneous regions, this assumption breaks down in heterogeneous landscapes such as urban–vegetation interfaces, agricultural mosaics, wetlands, and fragmented ecosystems [33, 106]. Sub-pixel mapping addresses this limitation by estimating the fractional abundance of multiple LC classes within each pixel, providing a continuous representation that better captures spatial transitions and mixed LC structures [109, 176].

Early methodological approaches to sub-pixel LULC mapping were dominated by spectral unmixing techniques, particularly the Linear Spectral Mixture Model (LSMM), which assumes that the observed reflectance of a mixed pixel is a linear combination of a finite set of pure endmember spectra weighted by their fractional abundances [108, 130]. Formally, the linear spectral mixture model is expressed as in Equation 1,

$$\mathbf{R} = \mathbf{S}\mathbf{A} + \mathbf{E}, \quad (1)$$

where $\mathbf{R} \in \mathbb{R}^B$ denotes the observed spectral reflectance vector of a mixed pixel with B spectral bands, $\mathbf{S} \in \mathbb{R}^{B \times n}$ is the endmember matrix whose columns correspond to the spectral signatures of the n endmembers, $\mathbf{A} = [a_1, a_2, \dots, a_n]^\top$ is the abundance vector, where a_i represents the fractional contribution (or abundance) of the i -th endmember within the mixed pixel, and $\mathbf{E}$ is the residual error accounting for measurement noise, atmospheric effects, model simplifications, and nonlinear mixing processes. The objective of spectral unmixing is to estimate the abundance vector $\mathbf{A}$ from the observed spectrum while satisfying the physical constraints as represented in Equations 2 and 3, and described as follows:

$$a_i \geq 0, \quad i = 1, \dots, n, \quad (2)$$

and

$$\sum_{i=1}^n a_i = 1, \quad (3)$$

where n denotes the number of endmembers. The non-negativity constraint ensures that each abundance represents a physically meaningful fractional contribution of an endmember, while the sum-to-one constraint guarantees that the estimated abundance fractions collectively account for the entire pixel composition. Consequently, the abundance vector provides a continuous representation of the relative proportions of different LC classes within a mixed pixel, forming the basis of sub-pixel LC characterization [130].

Due to their strong physical interpretability, LSMM and Spectral Mixture Analysis (SMA) have been extensively applied to vegetation fraction estimation, soil–vegetation–shade decomposition, and urban LC characterization [177]. However, these methods rely on simplifying assumptions, including linear spectral mixing and fixed endmember signatures, which are often violated in heterogeneous landscapes where multiple materials interact in complex ways [38, 106, 177]. To overcome these limitations, fuzzy-set-based approaches were introduced to explicitly model the inherent uncertainty and gradual transitions between LC classes [178]. Rather than assigning each pixel to a single class, these methods allow pixels to exhibit partial membership across multiple

classes, providing a more flexible representation of mixed LC conditions [179]. Consequently, fuzzy approaches have proven particularly effective in situations characterized by strong spectral overlap, ambiguous class boundaries, or uncertain endmember definitions, offering a more realistic description of sub-pixel landscape heterogeneity [180, 181, 182].

More recently, ML methods have emerged as powerful data-driven alternatives for sub-pixel fraction estimation [35]. Unlike traditional spectral unmixing approaches, these methods directly learn the relationship between multispectral observations and LC fractions from reference data, enabling them to capture complex nonlinear interactions without requiring explicit endmember definitions [183, 184]. Algorithms such as RF, Support Vector Regression (SVR), and Gradient Boosting have demonstrated superior performance, particularly when trained using synthetic reference data for Sentinel-2 imagery at 10 m spatial resolution [109, 176, 183, 184]. Their ability to model nonlinear relationships makes them particularly well suited to heterogeneous environments, where spectral responses are influenced by complex interactions among “vegetation”, “soil”, “water”, and “built-up materials”, etc. [41, 109].

Among these methods, RF regression has gained widespread adoption due to its robustness, ability to model high-dimensional nonlinear relationships, and resistance to overfitting [185]. Given an input feature vector $\mathbf{x}$, the predicted LC fraction is obtained by averaging the predictions of an ensemble of T regression trees, as shown in Equation 4:

$$\hat{y}(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^T f_t(\mathbf{x}), \quad (4)$$

where $f_t(\mathbf{x})$ denotes the prediction of the t -th regression tree. By aggregating multiple trees, RF effectively reduces prediction variance while maintaining high predictive accuracy [107]. These advantages have made RF one of the most widely used ML algorithms for sub-pixel LC fraction estimation [186, 187, 188].

Despite these advances, accurate sub-pixel estimation at Sentinel-2 resolution remains a challenging problem [39]. The scarcity of high-quality fractional reference data at this spatial scale has led many methods to rely on synthetically generated training data, limiting their ability to represent the complexity and variability of real-world landscapes [184]. Furthermore, spectral similarity between LC classes and the limited ability of conventional methods to exploit contextual spatial information further constrain estimation accuracy [104]. These limitations motivate the development of more advanced learning approaches that can leverage large-scale EO data to produce accurate and transferable high-resolution fractional estimates [41]. Building on these limitations, **Paper IV** investigates ML approaches for sub-pixel LULC estimation under constrained supervision. **Paper V** further extends this direction by leveraging GFM embeddings and directly producing high-resolution fractional LC estimates at 10 m resolution, enabling a more detailed and spatially consistent representation of fine-scale

landscape heterogeneity.

3.4 Generalization of DL Models under Limited Annotation Availability

The success of DL for satellite-based LULC mapping has largely been driven by the availability of large annotated datasets [189]. However, acquiring high-quality reference labels for EO remains expensive, labour-intensive, and often requires expert interpretation or field surveys [12]. Consequently, many practical applications must operate under limited annotation availability, motivating the development of learning algorithms that generalize effectively from only a small number of labeled examples [53].

To address this challenge, several label-efficient learning paradigms have been proposed. Transfer learning adapts models pretrained on large-scale datasets to new geographic regions or downstream tasks with limited supervision, significantly reducing annotation requirements [115]. Semi-supervised learning further exploits the abundance of unlabeled satellite imagery by jointly learning from labeled and unlabeled observations [114]. More recently, self-supervised learning has emerged as a powerful alternative, enabling models to learn transferable spectral-spatial representations from massive collections of unlabeled EO data through pretext or contrastive learning objectives [190]. These approaches have substantially improved the robustness and generalization of DL models while reducing their dependence on manually annotated training data [53].

Among label-efficient learning paradigms, FSL has received considerable attention for scenarios where only a handful of labeled samples are available for each LC class [113]. Rather than learning a task-specific classifier, FSL aims to learn representations that can rapidly adapt to previously unseen classes or geographic regions using only a few labeled examples [191]. Since its introduction in the ML community [192, 193], FSL has been successfully applied to a range of RS problems, including scene classification, hyperspectral image analysis, and continual learning across multiple EO datasets [191, 194]. These approaches are particularly attractive for RS applications, where the acquisition of representative training samples is often constrained by the cost and complexity of field campaigns [53]. Meta-learning has emerged as one of the dominant paradigms for FSL by enabling models to rapidly adapt to new tasks from only a limited number of labeled examples [195]. Among the various meta-learning approaches, metric-based methods have demonstrated particularly strong performance owing to their simplicity, computational efficiency, and ability to learn discriminative embedding spaces [196, 197, 198]. A representative example is Prototypical Networks [199], which represent each class by a prototype computed as the mean embedding of its support samples. Given an embedding function $g(\cdot)$ and a support set S_k for class k , the prototype is defined as,

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{(\mathbf{x}_i, y_i) \in S_k} g(\mathbf{x}_i), \quad (5)$$

where $\mathbf{c}_k$ denotes the prototype of class k , see Equation 5. A query sample $\mathbf{x}^*$ is first mapped into the embedding space using the same encoder, and its probability of belonging to class k is computed based on the Euclidean distance to each class prototype as shown Equation 6:

$$p(y^* = k | \mathbf{x}^*, S) = \frac{\exp(-\|g(\mathbf{x}^*) - \mathbf{c}_k\|_2^2)}{\sum_{k'} \exp(-\|g(\mathbf{x}^*) - \mathbf{c}_{k'}\|_2^2)} \quad (6)$$

Rather than learning explicit decision boundaries, Prototypical Networks learn an embedding space in which samples from the same class are clustered around their corresponding prototypes while remaining well separated from other classes [199]. This formulation has proven particularly effective for RS applications characterized by high intra-class variability and strong spectral similarity between LC classes, including scene classification and hyperspectral image analysis [197, 198]. The effectiveness of this paradigm under limited annotation availability is further investigated in **Paper I**. Despite the considerable progress achieved by few-shot and meta-learning methods, generalizing across diverse geographic regions, acquisition conditions, and LC distributions remains challenging [53]. Performance often degrades under domain shift, where the target distribution differs substantially from the source domain, highlighting the need for more robust and transferable learning strategies [27, 81, 198]. Addressing this challenge, **Paper I** investigates FSL for satellite image classification by extending metric-based meta-learning approaches and systematically evaluating their generalization capabilities across multiple RS benchmark datasets.

4 Summary of Publications

This chapter summarizes the scientific contributions of the thesis in relation to the research objectives introduced in 1.2. Rather than presenting the publications in chronological order, the papers are organized according to the objective they address, providing a coherent overview of how the individual contributions collectively advance the overarching research goals. Each section briefly introduces the corresponding objective before summarizing the associated paper(s) and highlighting their principal contributions.

4.1 *Objective 1: Develop methods that leverage geo-tagged field data to characterize and bridge the spatial and semantic gaps between ground-level observations and satellite imagery*

The first research objective addresses the semantic discrepancies that arise between ground-level observations and satellite imagery due to differences in sensing perspective, spatial scale, and semantic representation. Understanding and characterizing these discrepancies is essential for improving the interpretation, validation, and integration of heterogeneous EO data. This objective is addressed through **Paper II** and **Paper VI**, which investigate complementary aspects of semantic alignment using geo-tagged field observations and modern VFM.

4.1.1 Paper II

C. Paris, L. Martinez-Sanchez, M. van der Velde, **S. Sharma**, R. Sedona, and G. Cavallaro, “Accuracy assessment of land-use-land-cover maps: the semantic gap between in-situ and satellite data”, in *Proceedings Volume 12733, Image and Signal Processing for Remote Sensing XXIX, SPIE Remote Sensing*, Amsterdam, Netherlands, 2023, <http://doi.org/10.1117/12.2679433> ²

²Copyright 2023 Society of Photo-Optical Instrumentation Engineers (SPIE). One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in

*This paper contributes to **TO1** by developing methods to identify and overcome the significant semantic gaps between diverse data modalities, and serves as a preparatory work for spatial-semantic framework developed in Paper VI.*

This paper investigates a key limitation in the accuracy assessment of LULC maps, namely the semantic gap between *in-situ* observations and satellite-derived data. The study highlights that ground-based datasets, while commonly used as reference “ground truth,” are fundamentally different from satellite observations in terms of scale, view-point, and representation, which can lead to inconsistencies when evaluating LULC classification performance. To explore this mismatch, the *in-situ* data are compared with three high-resolution satellite-based LC products and street-level imagery is introduced as an intermediate representation that can partially bridge the gap between field observations and overhead satellite views, as shown in Figure 4.1. Their analysis shows that many apparent classification errors arise not from model deficiencies but from differences in how LC is perceived and labeled across data sources. By demonstrating the role of intermediate visual context in improving interpretability and agreement between datasets, the paper argues for more nuanced accuracy assessment frameworks that explicitly account for semantic and scale discrepancies rather than treating *in-situ* labels as absolute ground truth.

The contribution to **Paper II** focused on the curation of the data, including investigating how *in-situ* observations could be related to satellite observations. This included examining the complementary information and exploring their potential to address the semantic gap between them. In particular, this work provided the basis for investigating how *in-situ* observations can serve as an intermediate representation for relating heterogeneous observations and identifying their semantic differences. These methodological and data-related developments provided the preparatory foundation for **Paper VI**, which further develops this concept into a spatial-semantic diagnostic framework.

4.1.2 Paper VI

S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Diagnosing the Ground–Satellite Gap in Land Cover Using Vision Foundation Models and Field Photos”, in *Proceedings of the IEEE India Geoscience and Remote Sensing Symposium (InGARSS)*, 2026 (accepted)³

*This paper contributes to **TO1** by developing methods to identify and overcome the significant semantic gaps between diverse data modalities. This paper naturally extends the work done in **Paper II**.*

This paper presents a zero-shot diagnostic framework for investigating the spatial and semantic discrepancies between *in-situ* LC observations and satellite-derived LC

this publication for a fee or for commercial purposes, and modification of the contents of the publication are prohibited.

³©2026 IEEE.

4.1 *Objective 1: Develop methods that leverage geo-tagged field data to characterize and bridge the spatial and semantic gaps between ground-level observations and satellite imagery*

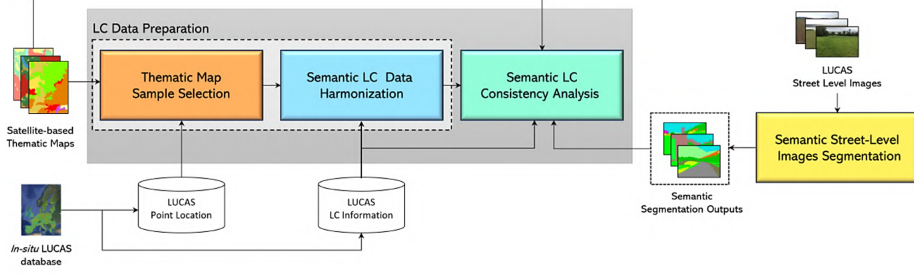


Figure 4.1. Flowchart of the proposed procedure used to analyze the semantic gap between in-situ and satellite-based land-cover (LC) data. The semantic LC consistency analysis is carried out comparing the LC information: (1) collected in-situ, (2) estimated in the satellite-based thematic products, and (3) derived from the semantic segmentation of street-level images. [Extracted from Paper II [153], © 2023 SPIE.]

products using geo-tagged field photographs as an intermediate source of evidence, as shown in Figure 4.2.

The framework builds on two frozen VFM used in complementary roles: DINOv3 patch tokens, computed over five of the six LUCAS photographs of each site (P, N, E, S, W), capture spatial structure, while CLIP/SenCLIP concept activations over the LC3 vocabulary capture semantics; neither requires task-specific training. Spatial heterogeneity (H_g) is summarized as the mean pairwise cosine distance among the pooled tokens, as in Equation 7,

$$H_g = \frac{2}{Q(Q-1)} \sum_{i < j} (1 - \cos(\mathbf{z}_i, \mathbf{z}_j)), \quad (7)$$

where $\mathbf{z}_i$, $i = 1, \dots, Q$, denote the DINOv3 patch tokens pooled across the five views, so that the score is comparable across sites. Higher H_g indicates greater within-site landscape complexity.

The framework is evaluated on the LUCAS dataset, pairing *in-situ* observations and geo-tagged field photos with Sentinel-2 imagery across 169,582 locations in Europe. Results show that spatial heterogeneity within field photos is strongly associated with satellite-*in-situ* disagreement, with agreement decreasing from 61% in homogeneous sites to 32% in highly heterogeneous landscapes, highlighting the impact of sub-pixel mixing and landscape complexity on classification reliability. Furthermore, the semantic analysis demonstrates that field photos can help determine whether disagreements originate from limitations in satellite representations or from ambiguities and inconsistencies in field-level interpretation and LC taxonomies, see Figure 4.3. By providing a richer understanding of the intrinsic properties of reference samples, the proposed framework enables the identification of more reliable labels for the construction of improved training and validation datasets.

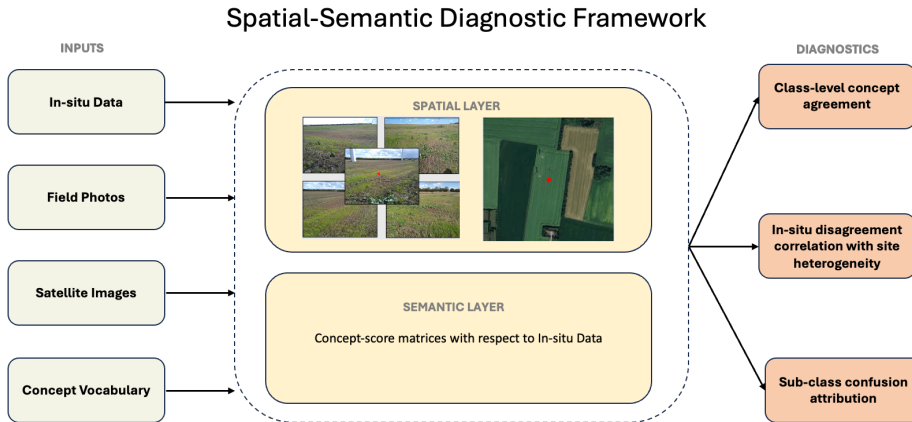


Figure 4.2. Spatial and semantic diagnostic framework for assessing in-situ-satellite LC discrepancies using field photos, satellite images, a concept vocabulary (defined by the field data collection protocol), and in-situ LC labels. [Extracted from Paper VI, ©2026 IEEE.]

4.2 Objective 2: Establish large-scale, multi-source benchmark dataset with detailed LULC information to facilitate the training and evaluation of advanced DL models

The second research objective focuses on the development of large-scale, multi-source benchmark datasets that support the training and evaluation of next-generation DL models for detailed LULC mapping. This objective is addressed through **Paper III**, which introduces the Sen4Map benchmark dataset.

4.2.1 Paper III

S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Sen4Map: Advancing Mapping with Sentinel-2 by Providing Detailed Semantic Descriptions and Customizable Land-Use and Land-Cover Data”, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, vol. 17, pp. 13893 - 13907, 2024, <http://doi.org/10.1109/JSTARS.2024.3435081>

*This paper primarily contributes to **TO2**, producing large-scale and multi-source datasets to improve the performance of LULC classification systems. The data is publicly available at <https://datapub.fz-juelich.de/sen4map>. Additionally, the Sen4Map dataset is used in **Paper VI**, thereby also contributing to **TO1**.*

4.2 Objective 2: Establish large-scale, multi-source benchmark dataset with detailed LULC information to facilitate the training and evaluation of advanced DL models

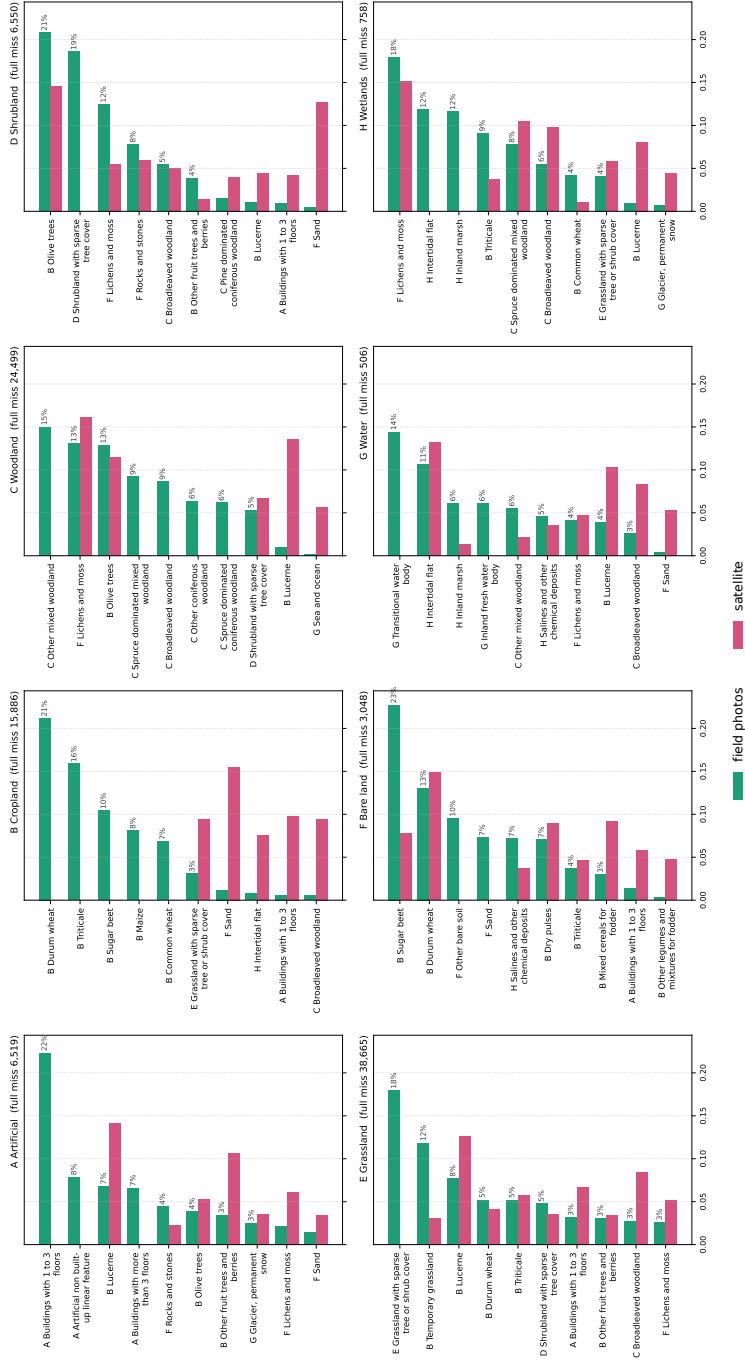


Figure 4.3. LC3 top-1 destination frequencies computed on the set of samples where in-situ and satellite data disagree. [Extracted from Paper VI, ©2026 IEEE.]

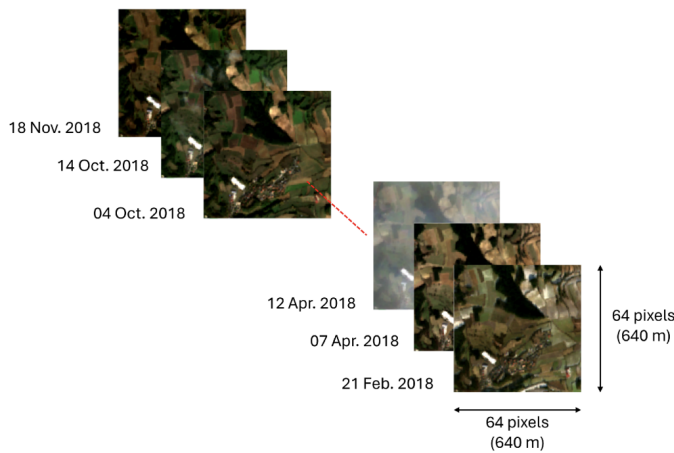


Figure 4.4. Example of a Sentinel-2 time series from the Sen4Map benchmark dataset associated with one LUCAS point. In this qualitative example, only the images acquired from the 21st of February until the 18th of November are visualized to capture the temporal component that can characterize different LC classes. [Extracted from Paper III [168]]

The paper introduces Sen4Map, a large-scale benchmark dataset designed to improve LULC mapping from Sentinel-2 imagery by combining multi-temporal satellite observations with detailed *in-situ* annotations from the 2018 LUCAS survey. Unlike existing benchmark datasets that primarily provide coarse LC labels, Sen4Map contains 335,125 georeferenced Sentinel-2 time-series patches (64×64 pixels) covering 28 European countries, enriched with hierarchical LC and LU classes, refer, Figure 4.4, 4.5, and 4.6.

The dataset preserves detailed semantic descriptions from LUCAS, making it suitable for a wide range of tasks including LC classification, LU mapping, crop-type identification, and research on semantic ambiguity in EO. To establish performance baselines, RF, pixel-based Transformers, ViTs, and Video Vision Transformers (VViT) are benchmarked, demonstrating that deep transformer-based architectures consistently outperform traditional machine learning methods, particularly when exploiting the spatial and temporal context available in Sentinel-2 image sequences. By providing rich semantic annotations, customizable class hierarchies, and publicly available benchmark splits, Sen4Map addresses the limitations of previous Sentinel-2 datasets and offers a valuable resource for developing, evaluating, and comparing ML, DL, and foundation models for large-scale LULC mapping and environmental monitoring.

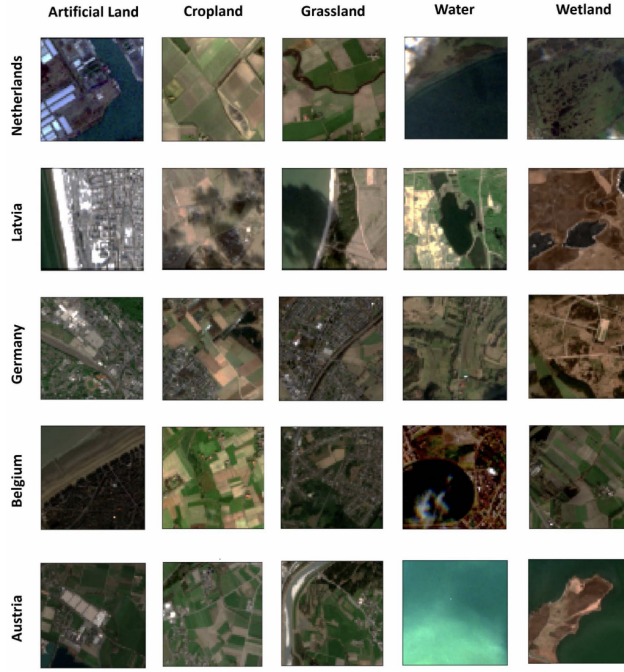


Figure 4.5. True color composition of 64×64 Sentinel-2 patches belonging to the Sen4Map dataset, which shows the variability of the LC classes located in different countries. [Extracted from Paper III [168]]

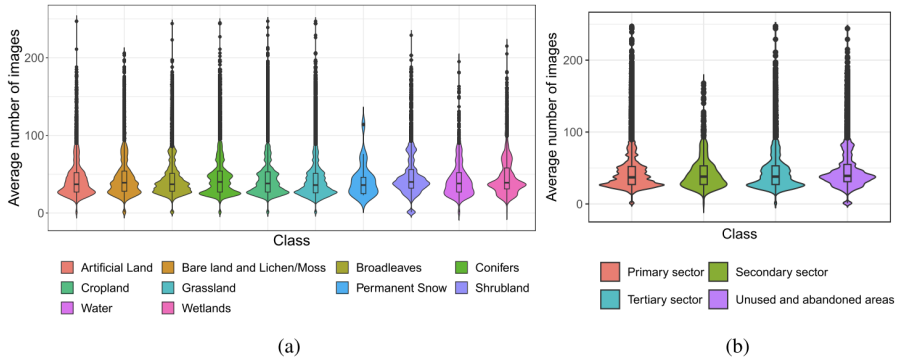


Figure 4.6. Statistical distribution depicted as a violin plot of the number of Sentinel-2 images associated with each (a) LC class and (b) LU class. The first level of the hierarchy is reported. [Extracted from Paper III [168]]

4.3 Objective 3: Build methods for sub-pixel LULC estimation to capture fine-scale landscape heterogeneity beyond the native spatial resolution of satellite imagery

The third research objective focuses on overcoming the limitations imposed by the native spatial resolution of satellite imagery by developing methods for sub-pixel LULC mapping. This objective is addressed through **Papers IV and V**, which investigate ML and GFM approaches for high-resolution fractional LULC estimation.

4.3.1 Paper IV

S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Enhancing Land Cover Mapping: A Novel Automatic Approach To Improve Mixed Spectral Pixel Classification”, in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Athens, Greece, 2024, pp. 1017-1020, <http://doi.org/10.1109/IGARSS53475.2024.10641976>⁴

*This paper contributes to **TO3** as it builds methods to enhance LULC classification at sub-pixel level.*

This paper addresses one of the major limitations of pixel-based LC mapping using Sentinel-2 satellite imagery: the presence of mixed spectral pixels, where multiple LC classes coexist within a single pixel. Such mixed pixels often lead to systematic misclassification, particularly in heterogeneous landscapes where buildings, grasslands, and croplands occur in close proximity. To mitigate this issue, a simple yet effective automatic post-processing framework is proposed based on spectral unmixing. The approach estimates the fractional composition of mixed pixels and identifies the dominant LC class, allowing initially misclassified pixels to be relabeled without retraining the classifier. Using Sentinel-2 imagery and reference data from the LUCAS survey, the study demonstrates that many pixels originally classified as cropland actually contain predominantly built-up or grassland cover, see Figure 4.7. By incorporating the concept of spectral unmixing, refer Equations 1, 2, and 3 from Section 3.3, the proposed method substantially improves the classification of these mixed pixels and increases the overall accuracy of the LC map. The results highlight the potential of lightweight unmixing techniques as an efficient and scalable solution for enhancing global LC mapping, particularly in regions where mixed pixels are common and high-resolution reference data are unavailable.

⁴©2024 IEEE. [100]

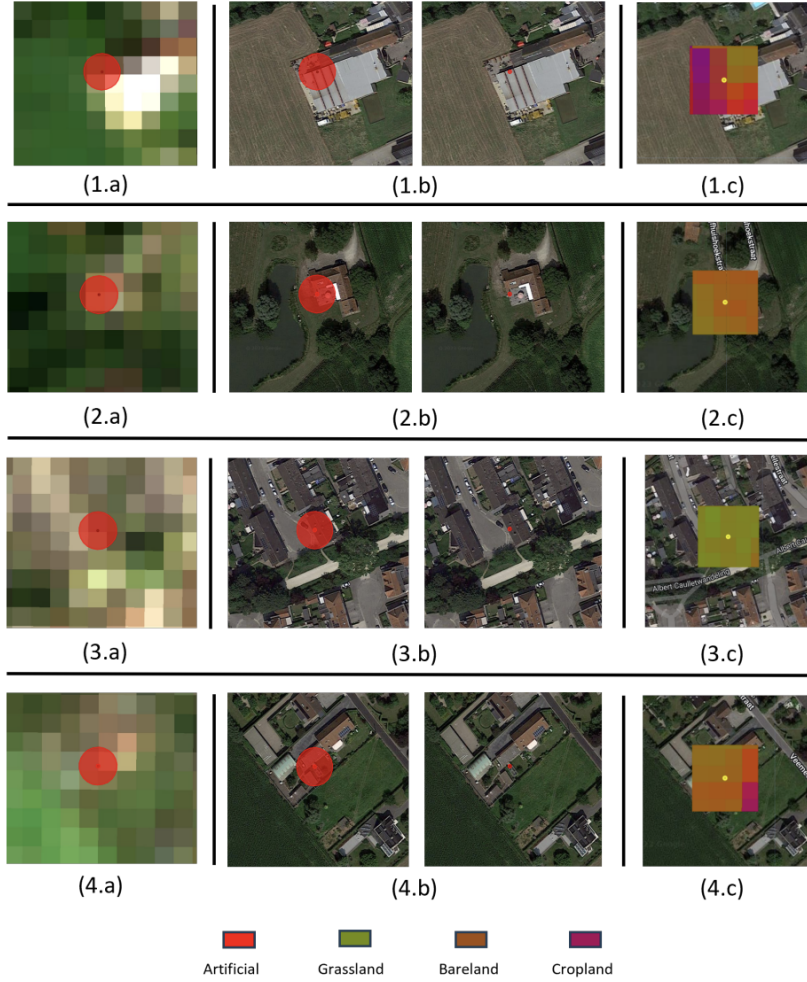


Figure 4.7. Abundance Map obtained after Spectral Unmixing for four different LUCAS points, marked as red and yellow, labeled as 'Artificial Land'. Figures (1.a)-(4.a) and (1.b)-(4.b) show the LUCAS points in Sentinel-2 and High-resolution Google Earth Satellite View. Figures (1.c) - (4.c) represent the Abundance Map obtained from the Sentinel-2 yearly composite of respective LUCAS points. [Extracted from Paper IV [100], ©2024 IEEE.]

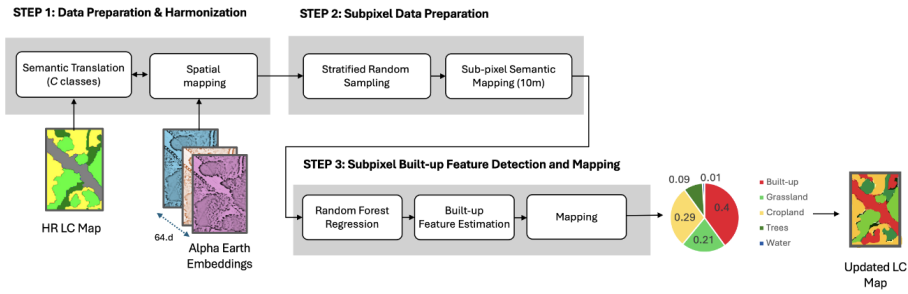


Figure 4.8. Workflow of the proposed method based on three main steps: (1) Data Preparation and Harmonization, (2) Sub-pixel Data Preparation, and (3) Sub-Pixel Built-up Feature Detection and Mapping. [Extracted from Paper V [200]]

4.3.2 Paper V

S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, “Beyond Pixels: Identifying Built-Up Features at Sub-Pixel Level for Enhanced Satellite-Based Land Cover Mapping”, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, 2026, <http://doi.org/10.1109/JSTARS.2026.3712773>

*This paper mainly contributes to **TO3** as it builds methods to enhance LULC classification at sub-pixel level. This paper also extends the work done in **Paper IV**. In addition, this paper also produces a fractional dataset, thus contributing to **TO2**.*

This paper produces a fractional dataset at 10 m spatial resolution, with the help of VHR map, and presents a regression-based framework, as discussed in Section 3.3, and Equation 4 for sub-pixel detection of urban and built-up features using thematic LC products, Sentinel-2 imagery, and ancillary data, such as OSM [74], refer, Figure 4.8. The study addresses the limitations of existing global LC maps, which often fail to capture small, fragmented, or structurally complex built-up features at 10 m spatial resolution. This study utilizes AEDs instead of raw Sentinel-2 data to produce efficient and scalable workflows. While the incorporation of AEDs yields only modest improvements in predictive performance compared to a Sentinel-2 baseline, the primary contribution lies in demonstrating that pretrained embedding representations can support scalable and efficient sub-pixel inference while maintaining comparable accuracy.

In addition, the experimental results show that the method successfully recovers 101 previously unidentified built-up features out of 400 candidates using a threshold of 0.20, including buildings as small as 20 m² and narrow road-like structures represented by buffers of approximately 1.5 m, as shown in Figure 4.9. The findings highlight the potential of combining foundation-model embeddings with sub-pixel regression to enhance the characterization of heterogeneous urban landscapes.

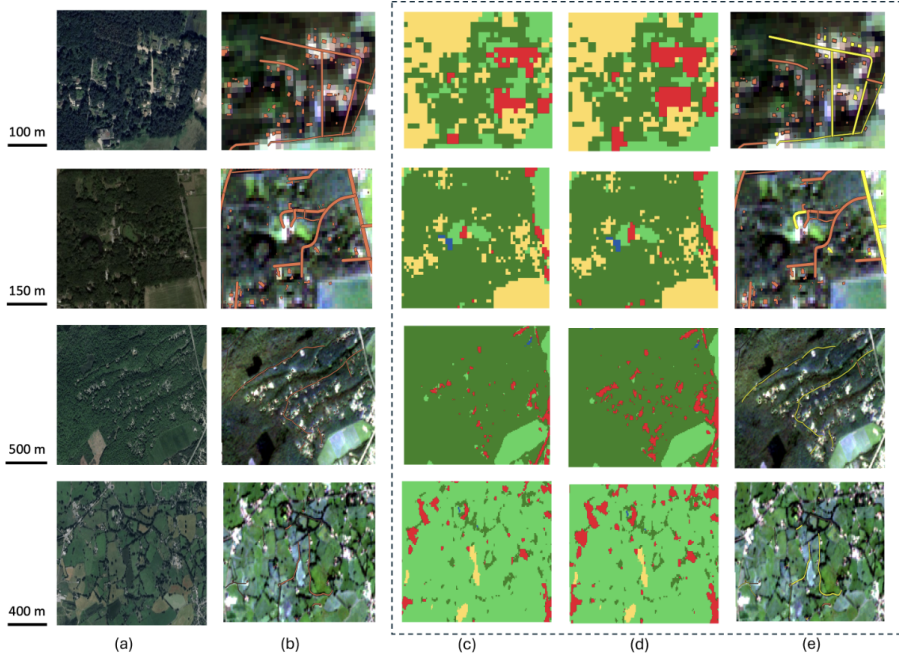


Figure 4.9. For each of the four patches: (a) High-resolution Google Earth Image, (b) Sentinel-2 yearly composite for reference. Highlighted in orange are the unidentified overlaid built-up features from OSM in three global LC maps, (c) Predicted LC Map based on the highest fraction obtained with the proposed model, (d) Updated LC map based on a Threshold of 0.20, (e) Highlighted in yellow are the identified built-up features. [Extracted from Paper V [200]]

4.4 **Objective 4: Investigate and evaluate the generalization capability of DL models that are resilient to limited annotation availability**

The fourth research objective focuses on improving the generalization capability of DL models when only limited labeled training data are available. This objective is addressed through **Paper I**, which investigates FSL and meta-learning strategies for improving model generalization under limited annotation availability.

4.4.1 Paper I

S. Sharma, R. Roscher, M. Riedel, S. M. Memon, and G. Cavallaro, “Improving Generalization for Few-Shot Remote Sensing Classification with Meta-Learning”, in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Kuala Lumpur, Malaysia, 2022, pp. 5061-5064, <http://doi.org/10.1109/IGARSS46834.2022.9884699>⁵

*This paper contributes to **TO4** as it investigated and evaluated the generalization capability of DL models that are resilient to limited annotation availability.*

This paper addresses a core challenge in RS image classification: poor generalization due to limited labeled data. In real-world RS applications, annotated samples are expensive, inconsistent across annotators, and often come from different distributions, making standard DL models prone to overfitting. The paper evaluates whether meta-learning-based few-shot models can maintain strong performance when annotated data is scarce, focusing specifically on generalization under limited supervision. Instead of training a model on a single dataset, the approach trains across multiple datasets with different feature distributions, encouraging the model to learn more transferable representations. The experimental results demonstrate that the meta-learning framework with prototypical networks, refer Section 3.4, and Equations 5 and 6, consistently improves generalization performance under few-shot and limited-annotation settings across standard remote sensing benchmarks, including EuroSAT, UCMerced, AID, WHU-RS19, and NWPU-RESISC45. A visual representation of the framework is shown in Figure 4.10. Compared to conventional supervised DL models, the meta-learned approach maintains higher accuracy in few-shot settings, even when only a small support set is available. It also shows strong cross-dataset generalization, with significantly less performance degradation when moving between datasets, and learns more transferable feature representations, making it less dependent on dataset-specific characteristics.

⁵©2022 IEEE. [201]

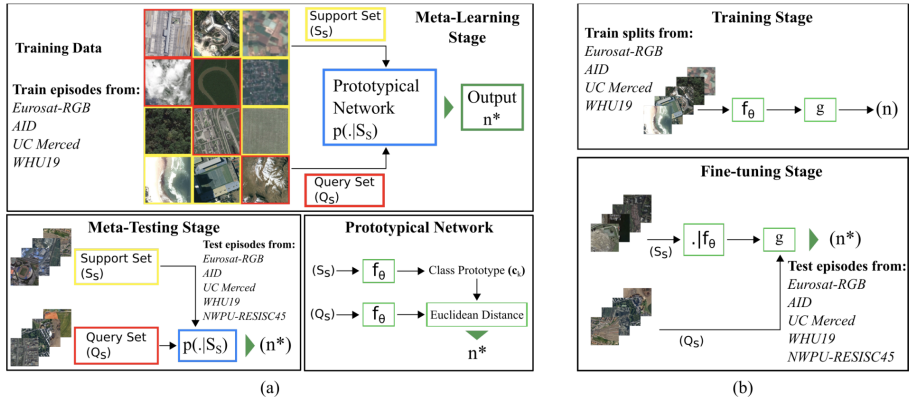


Figure 4.10. Visual representation of the framework: (a) Meta-learning approach to FSL, and (b) baseline finetuning approach, where f_θ is the feature extraction layer parameterized by θ and g is the embedding function using the ResNet-18 architecture. [Extracted from Paper I [201], ©2022 IEEE.]

5 Conclusions

In this chapter, the main findings of the thesis are summarized in Section 5.1. In addition, possible future research directions are highlighted in Section 5.2.

5.1 Summary

This thesis has attempted to advance LULC classification with scalable DL using HPC with the help of multi-source EO data. First, the thesis addresses the semantic discrepancies between heterogeneous EO data modalities that arise from differences in sensing perspective, spatial scale, and semantic representation, **TO1**. Geo-tagged *in-situ* field photographs serve as an intermediate observational layer linking ground-level surveys with satellite imagery, enabling a systematic investigation of the semantic gap between the two domains. The analysis demonstrates that apparent discrepancies between satellite-derived observations and field-based annotations stem not only from the limitations of satellite imagery, but also from ambiguities in field interpretation, inconsistencies in LC taxonomies, and differences in semantic representations across data sources.

The thesis subsequently focuses on the development of large-scale, multi-source benchmark datasets that provide the foundation for training and evaluating robust, transferable, and generalizable DL models across diverse geographic regions and environmental conditions, **TO2**. To this end, the Sen4Map benchmark dataset was developed by integrating multi-temporal Sentinel-2 imagery with detailed LULC annotations and geo-tagged field observations from the 2018 LUCAS survey and is made publicly available. By combining complementary EO modalities within a unified framework, Sen4Map enables the development of next-generation models capable of learning richer spectral, spatial, and semantic representations for detailed LULC mapping. This dataset also serves as a bridge between satellite observations and *in-situ* information, enabling systematic investigation of the semantic gap between satellite imagery and ground-level observations, thus also contributing to **TO2**.

The thesis then moves beyond conventional pixel-wise LULC by investigating sub-pixel modeling to better represent fine-scale landscape heterogeneity, **TO3**. Instead of assigning a single class to each pixel, the proposed approaches estimate the fractional

Table 5.1. Summary of the main findings corresponding to each research objective.

Objective	Paper(s)	Key Result	Main Finding
TO1	II, VI, III	Satellite- <i>in-situ</i> agreement decreases from 61% in homogeneous sites to 32% in highly heterogeneous landscapes	Paper II established the use of geo-tagged field photographs as an intermediate observational layer linking <i>in-situ</i> surveys with satellite imagery, while Paper VI demonstrated that from a semantic perspective, field photos can help disentangle whether disagreements arise from limitations in satellite representations or from inherent ambiguity already present in field-level visual interpretation and inconsistencies in LC taxonomy.
TO2	III, V	Time-series patches at 335,125 geo-tagged locations available publicly at https://datapub.fz-juelich.de/sen4map	The Sen4Map benchmark provides a large-scale, multi-source dataset linking Sentinel-2 time-series with detailed LUCAS annotations for training and evaluating next-generation DL models.
TO3	IV, V	Recovered 101 previously unidentified built-up features (out of 400 candidates) missed by global LC maps	Regression-based sub-pixel modeling enables estimation of built-up fractions at Sentinel-2 resolution, providing a more detailed representation of heterogeneous landscapes than conventional hard classification.
TO4	I	Accuracy of 81.83 ± 0.87 on NWPU-RESISC45 evaluated over 600 variable-way, variable-shot episodes.	Meta-learning significantly improves generalization under limited annotation availability, enabling robust adaptation using only a small fraction of labeled training data.

composition of LC classes, enabling a more realistic characterization of heterogeneous environments and mixed LC conditions. This richer representation enhances the ac-

curacy and usefulness of LULC products for downstream environmental monitoring and decision-making. Specifically, the thesis develops fractional dataset and regression-based methods for estimating LC fractions from Sentinel-2 imagery and further explores the use of GFM embeddings to generate high-resolution LC maps that capture urban heterogeneity beyond the native spatial resolution of the sensor, contributing to both **TO2** and **TO3**.

Finally, the thesis addresses the challenge of limited annotation availability by developing and evaluating learning strategies that generalize effectively from sparse or weakly labeled data, **TO4**. While modern DL models have achieved state-of-the-art performance in EO, their dependence on extensive, high-quality reference datasets limits their scalability and transferability across geographic regions and sensing conditions. To overcome this limitation, the thesis investigates FSL and meta-learning techniques that enable models to learn transferable representations and adapt to new tasks with only a small number of labeled samples. These approaches demonstrate improved generalization across diverse RS datasets, supporting the development of more robust and data-efficient LULC mapping systems.

In addition, the thesis also leverages VFM and HPC to enhance the scalability and effectiveness of the proposed approaches. VFM, pretrained on large and diverse datasets, provide rich and transferable feature representations that reduce the dependence on task-specific annotations and improve generalization across different geographic regions and sensing conditions. By adapting these models to RS data, the work benefits from strong pretrained priors that help bridge data scarcity and domain variability in downstream tasks. In parallel, HPC enables efficient processing of large-scale, multi-source geospatial datasets and supports the training and evaluation of computationally intensive models. This combination of VFM and HPC infrastructure is essential for handling the complexity and scale of modern LULC mapping problems, ultimately enabling more robust, data-efficient, and scalable EO workflows.

Overall the main findings corresponding to each research objective are summarized in Table 5.1.

Overall, this thesis addresses the main research question as follows:

How can publicly available, large-scale, multi-source datasets (including multispectral satellite imagery and geo-tagged images) be effectively exploited using advanced DL models to improve satellite-based LULC mapping while addressing:

- (i) *semantic and spatial gap between diverse data modalities*: This thesis uses *in-situ* field observations (street-level images) to bridge the semantic gap between what is visible from satellite and ground, as in **Paper II**. In addition, this thesis uses VFM along with field photos to diagnose the spatial-semantic gap in LC, in **Paper VI**, using Sen4Map from Paper III. The work presented is focused on quantifying within-site visual heterogeneity in field photographs to capture landscape complexity and mixed LC conditions through spatial analysis, while the semantic analysis evaluates the consistency of LC interpretations across field

images, satellite imagery, and *in-situ* reference annotations.

- (ii) *the need for more detailed and complex LULC classification legends*: This thesis presents **Sen4Map**, a large-scale, multi-source benchmark dataset specifically designed to support the development and evaluation of DL models capable of learning rich spectral, spatial, and semantic representations for detailed LULC mapping, in **Paper III**. The dataset is publicly available at <https://datapub.fz-juelich.de/sen4map>. Secondly, a fractional LC dataset from Paper V was also introduced to support sub-pixel estimation, providing fine-grained reference annotations for training and evaluating regression-based models using Sentinel-2 imagery.
- (iii) *fine-scale landscape heterogeneity*: To identify and map LC classes that are smaller than the sensor’s nominal pixel resolution or exhibit complex mixed LC signatures, this thesis uses a Random Forests classification and regression-based sub-pixel mapping approach, **Paper IV, V**. In addition, this thesis leverages pretrained AEDs, which offer rich, multi-temporal feature representations that help separate and characterize built-up information within mixed pixels, enabling the capture of fine-scale landscape heterogeneity.
- (iv) *the scarcity of high-quality reference data*: This thesis investigates DL models and methods that are resilient to limited annotation availability, such as FSL with meta-learning, as in **Paper I**. The presented work generalizes well on multiple unseen RS datasets.

5.2 Future Work

Although this thesis demonstrates promising advances in satellite-based LULC mapping, several open questions remain that motivate future research.

1. The proposed methods were primarily evaluated using European datasets, particularly the 2018 LUCAS survey and Sentinel-2 imagery. Although these datasets provide extensive geographic coverage and diverse landscape characteristics, the generalization of the proposed approaches to regions with substantially different climates, ecosystems, LU practices, and annotation protocols remains to be investigated. Future work should extend the evaluation to globally distributed benchmark datasets and investigate domain adaptation, continual learning, and data-efficient learning strategies to improve robustness and transferability across geographic regions and sensing conditions.
2. The semantic analysis relies on geo-tagged field photographs as an intermediate representation linking ground-level observations and satellite imagery. While these photographs provide substantially richer contextual information than conventional point-based annotations, they are still constrained by acquisition viewpoint, image quality, seasonal conditions, and incomplete visual coverage of the

surrounding landscape. Consequently, certain semantic discrepancies between ground observations and satellite imagery may remain unresolved, particularly in densely built or highly heterogeneous environments. Future work should investigate richer semantic vocabularies, additional satellite encoders, and representations learned across multiple spatial scales to better capture landscape semantics and improve the robustness, interpretability, and transferability of LULC mapping systems.

3. Although the Sen4Map dataset substantially enriches existing benchmark datasets by integrating multi-temporal Sentinel-2 imagery with detailed LUCAS annotations and geo-tagged field photographs, it remains limited to a single acquisition year and a specific annotation protocol. Extending the dataset to multiple years, additional sensing modalities, and more diverse geographic regions would further improve its applicability for training and evaluating foundation models and multi-modal EO systems.
4. The sub-pixel mapping framework focuses primarily on estimating built-up fractions from Sentinel-2 imagery. While this demonstrates the feasibility of exploiting foundation model embeddings for fine-scale landscape characterization, the methodology has not yet been extended to simultaneously estimate the fractional composition of multiple LC classes. Furthermore, accurately separating spectrally similar LC classes, such as built-up areas, bare ground, and sparsely vegetated surfaces, remains a challenging problem, particularly in complex urban and peri-urban environments. Future research should investigate the use of GFM embeddings such as AEDs, whose rich spectral-spatial representations offer considerable potential for disentangling these classes and improving sub-pixel LC characterization. Finally, extending and validating the framework beyond Belgium across diverse geographic regions will be essential for demonstrating its scalability and generalizability.
5. Although pretrained VFM and GFM provide highly transferable representations, their use in this thesis is limited to feature extraction rather than end-to-end adaptation or fine-tuning. As larger EO foundation models continue to emerge, jointly optimizing these representations for downstream LULC tasks may further improve performance and generalization. In addition, the computational requirements associated with training and deploying increasingly large foundation models remain an important practical consideration for operational large-scale mapping.

Beyond these specific directions, the contributions of this thesis connect to a broader convergence in EO, in which heterogeneous observations from satellites, aerial imagery, *in-situ* surveys, and other geospatial data sources are increasingly compressed into tokens and embedding fields [141, 142, 202], made addressable through natural language [159], and consumed by agentic systems that autonomously orchestrate geospatial analysis workflows [203]. As these paradigms mature, future EO systems will increasingly rely on interoperable representations that seamlessly integrate multi-modal observations, support zero- and few-shot reasoning, and enable scalable analysis across spatial and

semantic levels. In this landscape, the elements developed in this thesis gain rather than lose relevance. Semantically detailed benchmarks such as Sen4Map provide the grounding and evaluation data that token- and embedding-based models require; the diagnostic framework of Papers II and VI offers a way to verify whether learned representations remain consistent with *in-situ* reality before their outputs are acted upon; and the embedding-centric workflow of Paper V, together with the HDF5 data layer on HPC systems, demonstrates how such representations can be stored, served, and analyzed efficiently at scale. The thesis can therefore be read as contributing the data, semantics, and trust layer on which this next generation of EO analysis will rest.

Paper I

Improving Generalization for Few-Shot Remote Sensing Classification with Meta-Learning

S. Sharma, R. Roscher, M. Riedel, S. M. Memon, and G. Cavallaro

IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2022

©2022 IEEE. Reprinted, with permission, from [201]

Surbhi Sharma wrote the code and the paper, ran all experiments on the HPC machine and performed the analysis.

IMPROVING GENERALIZATION FOR FEW-SHOT REMOTE SENSING CLASSIFICATION WITH META-LEARNING

Surbhi Sharma^{1,2}, Ribana Roscher^{3,4}, Morris Riedel², Shahbaz Memon¹, and Gabriele Cavallaro¹

¹ Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany

² School of Engineering and Natural Sciences, University of Iceland, Iceland

³ Institute of Geodesy and Geoinformation, University of Bonn, Germany

⁴ Data Science in Earth Observation, Technical University of Munich, Germany

ABSTRACT

In Remote Sensing (RS) classification, generalization ability is one of the measure that characterizes the success of Machine Learning (ML) models, but is often impeded by the scarce availability of annotated training data. Annotated RS samples are expensive to obtain and can present large disparities when produced by different annotators. In this paper, we utilize Few-Shot Learning (FSL) with meta-learning to address the challenge of generalization using limited amount of training information. The data used in this paper is leveraged from different datasets that have diverse distributions, that means distinct feature spaces. We tested our approach on publicly available RS benchmark datasets to perform few-shot RS image classification using meta-learning. The results of the experiments suggest that our approach is able to generalize well on the unseen data even with limited number of training samples and reasonable training time.

Index Terms— Few-shot learning, meta-learning, deep learning, classification, remote sensing.

1. INTRODUCTION

To achieve sufficient generalization performance in different RS classification problems with Deep Learning (DL) models, it is essential to have access to large amounts of annotated training samples with diverse data distribution [1]. However, such large amounts of annotated training samples are difficult to obtain. Training only on a single dataset (collection of related data) where training and validation data follow the same distribution decreases the generalization capability of the classifier. As a result, trained classifiers achieve better results on the test data of the same distribution than on samples

of a different distribution, and thus are generally restricted to similar areas in the world.

Besides that, to rapidly discover and represent new classes with few observations and with high generalization capability, Few-Shot Learning (FSL) can be adapted with meta-learning paradigm [2]. FSL is a ML paradigm for supervised learning from a limited number of samples [3]. Together with meta-learning, they tackle meta-problems where the learning algorithms learn how to learn new classes from a limited number of training samples. The learning process becomes faster as they turn to the previous tasks for accumulated meta-knowledge. There are several categories for few-shot classification based on meta-learning such as distance-based approaches [4], optimization-based method [5], class or task-wise network modulation with amortization [6].

The main objective of this paper is to increase the generalization capability of ML and DL models for few-shot RS classification using meta-learning. The main contributions are: (1) a meta-dataset workflow ¹ adapted from an existing pipeline ² to create a more realistic classification approach for RS images (i.e., this pipeline can input different datasets that are acquired by distinct sensors) and (2), a distance-based meta-learning approach applied for few-shot RS classification [4] using the meta-dataset. The experimental results demonstrate that our approach can generalize on unseen RS data with a reasonable training time.

2. RECENT WORKS

Most works in the direction of FSL in RS include meta-learning approaches such as the utilization of a meta-learning model for scene recognition in RS images [7] for lifelong learning. To serve lifelong FSL, the knowledge is transferred from one dataset to another, with just a few samples. For few-shot scene classification in RS, approaches using meta-agnostic ML has been applied with promising results [8]. The authors of [9] introduce an inductive transfer learning

Part of this work was performed in the ADMIRE project, which received funding from the European Union's Horizon 2020 JTI-EuroHPC research and innovation programme under the grant agreement no. 956748. The authors gratefully acknowledge the computing resources from the DEEP-EST project, which received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement no. 754304. The work is partly funded by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (Grant number: 01DD20001).

¹<https://gitlab.jsc.fz-juelich.de/sdlrs/few-shot-remote-sensing-classification-with-meta-learning>

²https://github.com/google-research/meta-dataset/tree/axxiv_v1

problem where few data samples from a single region allow a model to adapt to an unseen region. They used the technique to evaluate the model agnostic meta-learning algorithm on classification and segmentation tasks using globally and regionally distributed datasets. In [10], a few-shot scene classification scheme is proposed. The authors combine multiple attention mechanisms and the attention-reference mechanism into the deepEMD network for few-shot classification on benchmark RS datasets. More recent studies considering few-shot classification mainly address classification based on meta-learning by prediction and finetuning the weights of the classifiers for new classes or methods that learn to augment data such as generative adversarial networks (GANs) [11].

3. METHODOLOGY

We consider two methods for few-shot RS classification: The first method is a meta-learning based approach and the other is a baseline approach without meta-learning (for comparison purposes).

3.1. Meta-learning approach to FSL

Meta-learning methods include two stages: **meta-learning and meta-testing stage**: In the meta-learning stage, as shown in Fig.1 (a), the meta-learner trains the learner on a training set that contains a large number of different classes. In so-called episodes, the meta-learning model is trained end-to-end with support sets (subsets of the training data) in order to make the classifier learn from the few samples. In this phase, the model learns common feature representations of all classes based on the acquired meta knowledge. In case of a new class, the model is fine-tuned with a small class-specific training data set (support set from the test episode) and evaluated using a query set.

In FSL, the model is able to generalize for a given set of episodes having C classes and few labeled samples per class ($k_l, l \in \{1, \dots, C\}$), when trained on a support set $S_s = \{(\mathbf{m}_1, n_1) \dots (\mathbf{m}_K, n_K)\}$, where $K = \sum_l k_l$, and tested on a query set $Q_s = \{(\mathbf{m}_1^*, n_1^*) \dots (\mathbf{m}_K^*, n_K^*)\}$, where $\mathbf{m}$ is the input vector ($\mathbf{m} \in \mathbb{R}$) and n is the label ($n \in \{1, \dots, C\}$). Here, $\mathbf{m}^*$ and n^* are the corresponding input vector and label for the query set. If episodes have training sets that are balanced $k_l = k, \forall l$, then they are called ‘ C way, k shot’ episodes.

In this paper, a distance based meta-learner (Prototypical Networks) [4] as described below is used. The choice for this meta-learner is based on its simplicity and robustness. The approach has been proved beneficial for few-shot classification, having a potential to outperform many sophisticated meta-learning methods [12].

Prototypical Networks: Here, a prototype of each class is constructed at the time of training. During testing, each query sample of the class is classified based on the ‘nearest’

Euclidean distance from the prototype. The probability that a query sample $\mathbf{m}^*$ belongs to class k is defined as:

$$p(n^* = k | \mathbf{m}^*, S_s) = \frac{\exp(-\|g(\mathbf{m}^*) - \mathbf{c}_k\|_2^2)}{\sum_{k' \in \{1, \dots, C\}} \exp(-\|g(\mathbf{m}^*) - \mathbf{c}_{k'}\|_2^2)} \quad (1)$$

where $\mathbf{c}_k$ is the ‘prototype’ for class k : the average of the embeddings of the training samples of class k .

3.2. Baseline approach to FSL

In this, a baseline is used for few-shot classification [11, 13]. A classifier is trained on all training classes at once, which is parameterized by the network parameters having a single output per class as shown in Fig. 1 (b). We use batch training and ResNet-18 [14] as a backbone, however also other DL model may be used. To classify new classes, the model is finetuned using a given test episode on top of g using the support set and performs classification for the query set [11, 12].

4. EXPERIMENTAL SETUP

4.1. Dataset

In this paper, we use the following datasets in RGB channels: Eurosat [15], Aerial Image Dataset (AID) [16], UCMerced Land Use [17], WHU-RS19 [18], and NWPU-RESISC45 [19]. The data is split in train (70%), validation (15%) and testing (15%) sets. Each set includes classes that are not shared with the other sets. The dataset NWPU-RESISC45 is used for testing purposes only and classes from this dataset are not included for training or validation. For evaluation, we report the mean Overall Test Accuracy (OA) and mean Cohen’s kappa coefficient (κ) for each dataset [20].

4.2. Models

4.2.1. Model A: Few-shot RS classification using meta-learning (prototypical network, as meta-learner)

The prototypical network model is trained episodically with the generated episodes using training splits of the datasets. The model is trained from scratch with the embedding function g as ResNet-18, having the input size of 256×256 , with a run time as shown in Tab. 1. For episodic sampling, this paper uses variable ways/shots. The minimum of classes in any episode is set to 2 with a maximum upperbound of 50. The size of maximum number of query samples is set to 10 and support set to 500 (with maximum support size per class to 100). The model is evaluated using 600 evaluation episodes, with checkpoints recorded after every 500 steps till model reaches 50k steps. The mean Validation Accuracy (VA) is found to be 88.2694. The mean OA and mean κ (600 episodes) for different datasets is shown in Tab. 2, Model A.

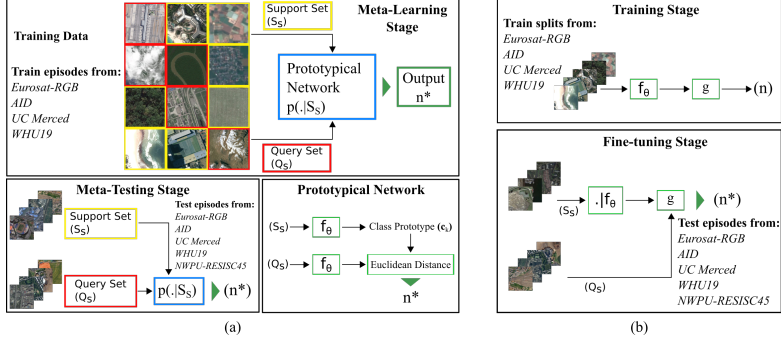


Fig. 1. (a) Meta-learning approach to FSL, and (b) baseline finetuning approach, where f_θ is the feature extraction layer parameterized by θ and g is the embedding function using the ResNet-18 architecture.

Table 1. Training Time on a Tesla V100-PCI-E-32GB GPU.

Models	Model A	Model B	Model C
Training Time	45 hours	90 hours	30 hours

4.2.2. Model B: Few-shot RS classification without meta-learning (using Baseline Finetune)

Baseline Finetune model is trained from scratch with the embedding function g as ResNet-18, having the input size of 256×256 and also the batch size of 256. Finetuning with ADAM is performed, with 75 steps and learning rate of 0.01. The mean VA is found to be 92.7264. The mean OA and mean κ (600 episodes) for different datasets is shown in Tab. 2, Model B. The evaluation procedure is similar to Model A. The run time of model for training is shown in Tab. 1.

4.2.3. Model C: Few-shot RS classification using meta-learning with initialization through Baseline

The Prototypical model is initialized using the pre-trained Baseline model. The Baseline model is trained on all the datasets and then used to initialize the weight of the meta-learner. The best pre-trained Baseline model is evaluated and their parameters are used as the initialization point for the Prototypical model with the same configuration as described in Sec. 4.2.1. The mean VA is found to be 93.09583. The mean OA and mean κ (600 episodes) for different datasets is shown in Tab. 2, Model C. The run time of model for training is shown in Tab. 1.

5. RESULTS AND DISCUSSION

Tab. 2 shows that Model C outperforms both the Model A and Model B in terms of their OA and κ for the test classes of all datasets and with the smallest run time. In addition to this, it could also be seen that Model A takes half the training time

as compared to Model B. However, Model B performs better for an unseen dataset NWPU-RESISC45 having the potential to be used as a better initializer. These results imply that the meta-learning approach can be beneficial for few-shot RS classification, if we carefully choose the model parameters or use a proper initialization technique. Using the meta-learning approach with proper initialization also reduces the overall training (compute) time of the model which becomes essential with large datasets.

In addition, adapting the meta-dataset pipeline for RS images, helps to give diversity to the training and test data which is important to study how well our model is in terms of generalization, even in presence of limited number of training samples. Moreover, the lower values of κ for the NWPU-RESISC45 and AID datasets for all the models suggests that there is a higher class imbalance in these datasets which needs to be accounted. Apart from this, the Baseline approach is also having a potential, however in future, carefully choosing the backbone and model parameters can result in a better training time, thereby giving this approach a better chance.

6. CONCLUSIONS

We performed few-shot RS classification using meta-learning (prototypical network as a meta-learner) using RS datasets (Eurosar, Aerial Image Dataset, UCMerced Land Use, WHU-RS19, and NWPU-RESISC45) as leveraged from the adapted meta-dataset pipeline. For comparison purposes, we also performed few-shot RS classification using a Baseline model without meta-learning. We could see that even in the presence of a limited amount of training samples, a comparable generalization performance was achieved using meta-learning approaches in Few Shot Learning. Also, with a proper initialization, the approach helps to generalize on the unseen data, with a significantly shorter training time.

Table 2. Few-shot RS classification mean Overall Test Accuracies OA in percentage and mean κ computed on 600 episodes. (The percentage in the Dataset column shows the reserved data split for testing purposes)

Test classes for Datasets	Model A		Model B		Model C	
	OA	κ	OA	κ	OA	κ
EuroSAT (15%)	98.4750 +/- 0.2863	0.972333	97.4750, +/- 0.3630	0.946167	99.1833, +/- 0.1833	0.988167
AID (15%)	78.3736 +/- 0.8904	0.668083	82.9958, +/- 0.7779	0.722250	85.8972, +/- 0.7071	0.781139
UC Merced (15%)	88.7305, +/- 0.7843	0.814000	95.1583, +/- 0.4619	0.920333	95.2000, +/- 0.4841	0.929583
WHU19 (15%)	99.2306, +/- 0.2274	0.987167	98.6278, +/- 0.2529	0.982000	99.7694, +/- 0.1032	0.994333
NWPU-RESISC45 (100%)	69.0711, +/- 1.0950	0.596833	81.8329, +/- 0.8740	0.747792	79.9027, +/- 0.8811	0.738003

7. REFERENCES

- [1] A. Stoian, V. Poulain, J. Inglada, V. Poughon, and D. Derksen, "Land Cover Maps Production with High Resolution Satellite Image Time Series and CNNs: Adaptations and Limits for Operational Systems," *Remote Sensing*, vol. 11, no. 17, 2019.
- [2] S. Thrun and L. Pratt, Eds., *Learning to Learn*, Kluwer Academic Publishers, 1998.
- [3] Li Fei-Fei, R. Fergus, and P. Perona, "One-Shot Learning of Object Categories," *IEEE TPAMI*, vol. 28, no. 4, pp. 594–611, 2006.
- [4] J. Snell, K. Swersky, and R. Zemel, "Prototypical Networks for Few-Shot Learning," in *Advances in Neural Information Processing Systems*, 2017.
- [5] L. Zintgraf, K. Shiarlis, V. Kurin, K. Hofmann, and S. Whiteson, "Fast Context Adaptation via Meta-Learning," in *Proceedings of the 36th ICML*, 2019.
- [6] J. Gordon, J. and Bronskill, S. Nowozin, M. Bauer, and R. E. Turner, "Meta-learning probabilistic inference for prediction," in *Proceedings of the 7th ICLR*, 2019.
- [7] M. Zhai, H. Liu, and F. Sun, "Lifelong Learning for Scene Recognition in Remote Sensing Images," *IEEE GRSL*, vol. 16, no. 9, pp. 1472–1476, 2019.
- [8] D. A. Alajaji and H. Alhichri, "Few Shot Scene Classification in Remote Sensing using Meta-Agnostic Machine," in *Proceedings of the 6th CDMA*, 2020, pp. 77–80.
- [9] M. Wang, S. and Rußwurm, M. Körner, and D. B. Lobbell, "Meta-Learning For Few-Shot Time Series Classification," in *IEEE IGARSS*, 2020, pp. 7041–7044.
- [10] Z. Yuan, W. Huang, L. Li, and X. Luo, "Few-Shot Scene Classification With Multi-Attention Deepemd Network in Remote Sensing," *IEEE Access*, vol. 9, pp. 19891–19901, 2021.
- [11] C. Wei-Yu, L. Yen-Cheng, K. Zsolt, F. Wang Yu-Chiang, and H. Jia-Bin, "A Closer Look at Few-shot Classification," 2020.
- [12] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, P. A. Swersky, K. Manzagol, and Larochelle H., "Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples," 2019.
- [13] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," in *International Conference on Learning Representations*, 2020.
- [14] K. He, R. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015.
- [15] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," in *IEEE IGARSS*, 2018, pp. 204–207.
- [16] G. S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A Benchmark Data set for Performance Evaluation of Aerial Scene Classification," *IEEE TGRS*, 2017.
- [17] Yi Yang and Shawn Newsam, "Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification," in *Proceedings of the ACM GIS*, 2010.
- [18] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Maitre, and H. Sun, "Structural High-Resolution Satellite Image Indexing," *ISPRS*, 2010.
- [19] G. Cheng, J. Han, and X. Lu, "Remote Sensing Image Scene Classification: Benchmark and State of the Art," 2017.
- [20] Margherita Grandini, Enrico Bagli, and Giorgio Visani, "Metrics for multi-class classification: an overview," 2020.

Paper II

Accuracy assessment of land-use-land-cover maps: the semantic gap between in-situ and satellite data

C. Paris, L. Martinez-Sanchez, M. van der Velde, **S. Sharma**, R. Sedona, and G. Cavallaro

Image and Signal Processing for Remote Sensing XXIX (SPIE), 2023

Copyright 2023 Society of Photo-Optical Instrumentation Engineers (SPIE). One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in this publication for a fee or for commercial purposes, and modification of the contents of the publication are prohibited.

Surbhi Sharma contributed to data curation.

Accuracy Assessment of Land-Use-Land-Cover Maps: the Semantic Gap Between In-Situ and Satellite Data

Claudia Paris¹, Laura Martinez-Sanchez², Marijn van der Velde², Surbhi Sharma^{3,4},
Rocco Sedona³, and Gabriele Cavallaro^{3,4}

¹ Department of Natural Resources, ITC, University of Twente, Enschede, The Netherlands

² European Commission, Joint Research Centre (JRC), Ispra, Italy

³ Jülich Supercomputing Centre, Forschungszentrum Jülich, 52428 Jülich, Germany

⁴ School of Engineering and Natural Sciences, University of Iceland, 107 Reykjavik, Iceland

1. ABSTRACT

The availability of high-resolution, open, and free satellite data has facilitated the production of global Land-Use-Land-Cover (LULC) maps, which are extremely important to monitor the Earth's surface constantly. However, generating these maps demands significant efforts in collecting a vast amount of data to train the classifier and to assess their accuracy. Although *in-situ* surveys are generally regarded as reliable sources of information, it is important to note that there may be inconsistencies between the *in-situ* data and the information derived from satellite data. This can be attributed to various factors (1) differences in viewpoint perspectives, i.e., aerial versus ground views, and (2) spatial resolution of the satellite images versus the extent of the Land-Cover (LC) present in the scene. The aim of this paper is to explore the feasibility of using geo-referenced street-level imagery to bridge the gap between information provided by field surveys and satellite data. Unlike conventional *in-situ* surveys that typically provide geo-tagged location-specific information on LULC, street-level images offer a richer semantic context for the sampling point under examination. This allows for (1) an improved interpretation of LC characteristics, and (2) a stronger correlation with satellite data. The experimental analysis was conducted considering the 2018 Land Use and Coverage Area Frame Survey (LUCAS) *in-situ* data, the LUCAS landscape (street-level) images and three high-resolution thematic products derived from satellite data, namely, Google's Dynamic World, ESA's World Cover, and Esri's Land Cover maps.

Keywords: Land-Cover (LC) mapping, satellite-based thematic products, classification, accuracy, validation, *in-situ* data, Land Use and Coverage Area Frame Survey (LUCAS) database, Sentinel-2.

2. INTRODUCTION

The production of accurate Land-Use-Land-Cover (LULC) maps is extremely important for continuously monitoring the Earth's surface to support environmental monitoring and planning. While Land-Cover (LC) refers to the natural and anthropogenic land attributes (e.g., forests, wetlands, built areas, etc.), Land-Use (LU) describes the activities and purposes for which the land is used (e.g., shopping centers, reservoirs, parking lots, etc.). In the last few decades, different satellite images have been used to generate LC maps. Owing to their nearly daily temporal resolutions, optical satellites such as the Moderate Resolution Imaging Spectroradiometer (MODIS), ENVISAT's Medium Resolution Imaging Spectrometer (MERIS) or Advanced Very High Resolution Radiometer (AVHRR) were used to monitor the extent and distribution of the major LC classes at a worldwide level.¹⁻³ Although these global products are widely employed, their main drawback is the coarse spatial resolution, leading to high uncertainty LC classification and update due to the mixed pixel problem.⁴ Due to their pixel size (e.g., > 250m), these thematic products are typically characterized by spatially aggregated classes, e.g., "Mixed Forest" which includes areas with a mix of conifers and broad leaves.⁴ Furthermore, they do not provide enough level of geometrical detail to support precise environmental monitoring applications. At a 30m spatial resolution, several products have recently been produced globally using Landsat satellite data.⁵⁻⁷ However, since Landsat data are captured every 16 days, they result in limited cloud-free images in regions with frequent cloud cover.

The European Union's Copernicus program significantly transformed the LC mapping perspective, thanks to the high temporal and spatial resolutions of the Sentinel satellite data. The 10m-resolution pixels can be associated with spatially distinct LC classes, allowing for creating thematic products that provide detailed

information from the geometrical viewpoint. In particular, the Sentinel-2 multispectral data have been used to produce several global thematic products. The most popular ones are the ESA WorldCover,⁸ Google's Dynamic World,⁹ and Esri's Land Cover¹⁰ products. The quality of these products has been assessed and compared¹¹ using: (1) a global validation set made up of photo-interpreted samples,⁹ and (2) the European 2018 Land Use and Coverage Area Frame Survey (LUCAS) database,¹² consisting of point-based *in-situ* survey data. The results obtained an accuracy of 72%, 75%, and 65% for Dynamic World, Esri, and WorldCover, respectively, using the global validation database of manually labeled Sentinel-2 pixels. In contrast, the accuracy obtained with the *in-situ* data is 66%, 63%, and 71%, for Dynamic World, Esri, and WorldCover, respectively. These results indicate a discrepancy between the validation performed when using labeled samples derived by visually interpreting satellite data compared to those collected *in-situ*.

Due to the rise in crowdsourced data availability, there has been a growing interest in using street-level images to create reference data across various fields.^{13–15} D'Andrimont *et al.*¹⁵ propose a method to support the European Union's Common Agricultural Policy (CAP) that aims to simplify the control of farmers' parcel declarations. Sentinel-1 and Sentinel-2 data are used to automatically detect parcels that are potentially not grasslands but are declared by farmers as such. To verify the results obtained, a continuous collection of geo-referenced images was carried out using a high-definition camera mounted on top of a car. As expected, street-level images greatly increase the efficiency of ground reference data collection, with 1,411 parcels surveyed per day compared to 217 that can be collected through classic field visits. The street-level images publicly available in the crowdsourcing platform Mapillary were used to perform crop-type monitoring.¹⁵ For each cultivated parcel, the most informative street-level images are selected based on their meta-data (e.g., orientation, focal length of the camera). Recently, the Geo-Wiki toolbox, widely used for the interpretation of satellite images by volunteers for LC applications,^{16,17} has been updated with a novel tool for the visual interpretation of street-level images.

Unlike existing literature, our approach does not view street-level images as a substitute for ground reference data. Instead, this paper aims to explore the feasibility of using street-level images to bridge the semantic gap existing between *in-situ* and satellite data focusing on the specific problem of assessing the accuracy of satellite-based thematic products. Although producing accurate thematic products requires high-quality and sufficient training data, it is crucial to implement proper validation measures to ensure data quality. Unlike conventional *in-situ* surveys that provide geo-tagged location-specific information on LULC, street-level images offer a more comprehensive semantic context of the sampling point under investigation, thereby allowing for: (1) improved interpretation of the LC characteristics, and (2) better linkage to satellite data information. We conducted the experimental analysis using the 2018 LUCAS database, consisting of *in-situ* data¹⁸ and the three released satellite-based high-resolution thematic products, namely, WorldCover,⁸ Dynamic World,⁹ and Esri.¹⁰ The LUCAS survey is coordinated by the Statistical Office of the European Commission (Eurostat) and collects LULC data through field observation for a large set of geographically referenced points distributed all over Europe. Photographs are also taken on each LUCAS point, including of the LC at the point (the so-called point photo) and of the surrounding landscape in the four cardinal directions (street-level images).

The structure of the rest of the paper is as follows. Section 3 presents the problem formulation, providing a real example of the semantic gap between the LC information collected in *in-situ* and the one visible in a satellite-based thematic product. This semantic gap can be explained by checking the landscape properties visible in the street-level images. Section 4 presents the dataset and the procedure employed to carry out the experimental analysis. Section 5 presents the experimental setup, while Section 6 discusses the experimental results. Finally, Section 7 draws the conclusion of the paper.

3. PROBLEM FORMULATION

From the analysis of the literature reported in the previous section, it turned out that several LC products are now available on a global scale. Although existing thematic products currently map the extent and global distribution of the main LC categories, the new frontier of mapping aims to better understand the relationship between environmental and socio-economic indicators. These analyses are crucial for implementing and supporting complex environmental policies, such as those described in the Sustainable Development Goals (SDGs)¹,

¹<https://sdgs.un.org/goals>

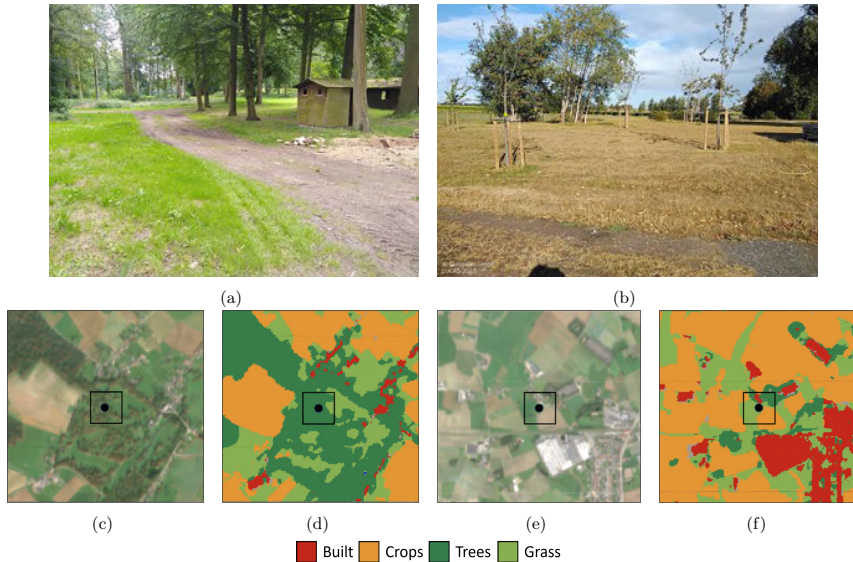


Figure 1: Examples of LUCAS *in-situ* samples associated with the “Woodland” LC class: (a)(b) LUCAS street-level images, (c)(e) Sentinel-2 satellite data, (d)(f) ESA WorldCover thematic product. The sample’s locations are highlighted in black.

which require a better investigation of the interaction between human lifestyles and the environment on a global scale. In this context, it is urgent to define solutions for continuously collecting reference data for the calibration and validation of thematic products. Furthermore, it is crucial to effectively quantify the inconsistencies between reference data and satellite-based products and estimate their uncertainties.

The collection of labeled data is typically carried out by *in-situ* surveys or by manual annotation based on the visual interpretation of satellite/aerial Very High Resolution (VHR) data. Although ground reference data are generally considered more reliable sources of information than manual photo-interpretation, it is important to note that there may be inconsistencies between *in-situ* and satellite-derived data due to: (1) differences between viewpoints, i.e., between aerial and ground views, and (2) the spatial resolution of the satellite image compared to the extent of LC present in the scene. This can lead to partial validation assessments that do not provide a real quantitative evaluation to the users and stakeholders interested in the LC information provided by the thematic product. In this context, the use of street-level images can support providing a better understanding of the LULC properties of the scene. Given the continuous growth in data gathered via crowdsourcing, an unprecedented amount of geo-tagged and time-stamped photographs information is now collected using free apps for smartphones.^{19–21} These data can aid in enhancing the comparison between ground-based and satellite-based LC information.

Fig. 1 shows a real example of the semantic gap between *in-situ* and satellite-based LC extracted from the LUCAS database and the ESA Wordcover map,⁸ respectively. Both LUCAS samples are associated with the “Woodland” LC due to the presence of trees (see Fig. 1(a) and Fig. 1(b)). However, the sample located in the dense forest (see Fig. 1(c)) is correctly classified in the satellite-based thematic product (see Fig. 1(d)), while the individual trees located in open forest area are associated in the thematic product with the “Grass” LC label (see Fig. 1(f)).

4. MATERIALS AND METHODS

To analyze the semantic gap between the satellite images and *in-situ* data, we compared the LC information of three recently released satellite-based thematic products with the *in-situ* data collected during the 2018 LUCAS survey. In this section, we first describe the properties of the thematic products and the LUCAS dataset. Then, the procedure applied to carry out the experimental analysis is reported.

4.1 Satellite-based thematic products

This study focuses on three satellite-based thematic products generated at 10m spatial resolution, i.e., WorldCover,⁸ Dynamic World⁹ and Esri.¹⁰ The Esri and Dynamic World were generated using the same reference dataset consisting of over 5 billion hand-annotated Sentinel-2 pixel patches distributed worldwide used to train deep learning models.^{9,10} In contrast, WorldCover was produced using Sentinel-1 and Sentinel-2 data and considering a random forest classifier trained on 141,000 manually annotated pixels.

Although these products are available on a global scale, the considered analysis is carried out in the European Union because of the spatial coverage of the *in-situ* LUCAS data. Since the ESA WorldCover is available for only the years 2020 and 2021, to maintain temporal consistency across all three products, we considered the thematic products corresponding to the year 2020 in the experimental analysis. While WorldCover and Esri are released on an annual basis, the Dynamic World is generated each time a Sentinel-2 image is available. The Dynamic World 2020 thematic product was generated by calculating the mode of the classes LC classes present across all Dynamic World thematic products generated during the year 2020, applied on a per-pixel basis.¹¹

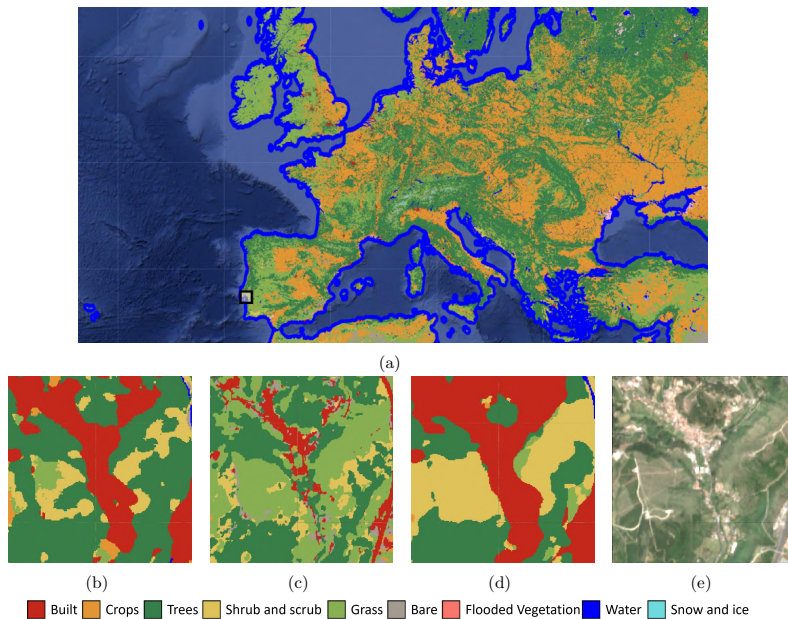


Figure 2: (a) Example of the WorldCover thematic product depicting the main LC distributed over Europe, and a zoomed-in view of the location highlighted in black that compares the: (b) 2020 Dynamic World, (c) 2020 WorldCover, (d) 2020 Esri, and (e) Sentinel-2 2020 annual composite.

While the Esri and Dynamic World depict 9 LC categories, namely, “Water”, “Trees”, “Grass”, “Flooded vegetation”, “Crops”, “Shrub and scrub”, “Built”, “Bare”, “Snow and ice”, the WorldCover product provides information on two extra classes, i.e., “Mangrove” and “Moss and lichen”. In the considered experimental analysis, we did not consider the “Mangrove”, “Moss and lichen”, “Snow and ice” classes since their extent is very limited over Europe. Fig. 2 shows a qualitative example of the 2020 WorldCover depicting the main LC distributed over Europe, together with a zoomed-in view of the location highlighted in black that compares the Dynamic World, WorldCover, Esri, and the annual composite Sentinel-2 image for the year 2020.

4.2 LUCAS database

The LUCAS survey, coordinated by Eurostat, aims to collect harmonized data LULC, agro-environmental, and soil data by field observation of geographically referenced points. This dataset has been extensively used for training and validating the quality of existing thematic products.²² Because of the hierarchical structure of its legend, the LUCAS database provides a comprehensive description of the LULC present in the scene, together with street-level pictures collected by the surveyors. In the considered experimental analysis, we focused on the 2018 LUCAS survey. To match the LC classes present in the thematic products, we focused on the LUCAS first-level LC classification, i.e., “Artificial land” (A00), “Cropland” (B00), “Bare land and lichens/moss” (F00), “Grassland” (E00), “Shrubland” (D00), “Woodland” (C00), “Wetlands” (H00) and “Water areas” (G10 to G40).

It is worth noting that the recorded LULC information lies within a circle of 1.5 m radius at each grid point, i.e., Minimum Mapping Unit (MMU). However, this is not always the case as LUCAS also employs an “extended window” of observation for 20 m for those cases where the surveyor encounters a point that falls on a boundary or a small linear feature, or where there is heterogeneity in the LC. This is relevant for instance when the percentage of tree cover determines whether an area is classified as woodland (>10%) or grassland with sparse tree coverage (5-10%)². Fig 3 shows some example of LUCAS street-level picture of “Grassland” Fig. 3(a), “Cropland” Fig. 3(b), “Artificial land” Fig. 3(c), “Wetlands” Fig. 3(d), “Water areas” Fig. 3(e), “Bare Land” Fig. 3(f), “Shrubland” Fig. 3(g), and “Woodland” Fig. 3(h).



Figure 3: Examples of LUCAS street-level pictures of (a) “Grassland”, (b) “Cropland”, (c) “Artificial land”, (d) “Wetland”, (e) “Water”, (f) “Bare Land”, (g) “Shrubland”, and (h) “Woodland”.

4.3 LC Data Preparation

Fig. 4 illustrates the flowchart of the proposed procedure utilized to analyze the semantic gap between *in-situ* and satellite-based LC data, leveraging the comprehensive semantic information provided by the street-level pictures. To achieve this goal, the first step aims to harmonize the considered data sources to facilitate their

²²Eurostat LUCAS 2018 C1 instructions

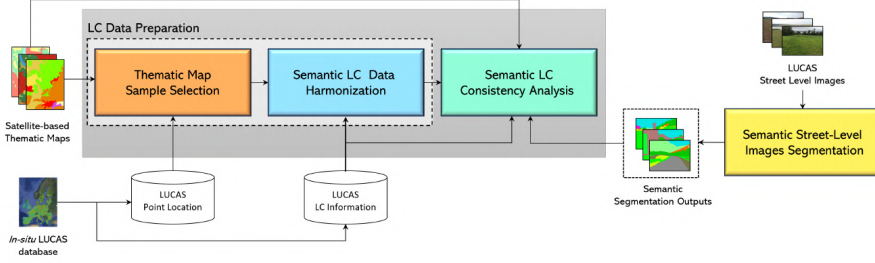


Figure 4: Flowchart of the proposed procedure used to analyze the semantic gap between *in-situ* and satellite-based LC data. The semantic LC consistency analysis is carried out comparing the LC information: (1) collected *in-situ*, (2) estimated in the satellite-based thematic products, and (3) derived from the semantic segmentation of street-level images.

comparison. First, we extract the LC information provided by the thematic products for the locations associated with the LUCAS samples. Then, the legend of the three products and the LUCAS database are harmonized according to Venter *et al.*¹¹ resulting in the following eight-class typology, namely, “Water”, “Trees”, “Grass”, “Flooded vegetation”, “Crops”, “Shrub and scrub”, “Built” and “Bare”. At the end of the LC data preparation step, the harmonized LC data can be used to analyze the consistency of the LC information: (1) collected *in-situ*, (2) estimated in the satellite-based thematic products, and (3) derived from the semantic segmentation of street-level images.

4.4 Semantic Street-Level Images Segmentation

To automatically extract the semantic context provided by the street-level images, we applied a semantic segmentation algorithm able to characterize the landscape present in street-level pictures. We employed the state-of-the-art segmentation Deeplabv3+ model^{23,24} widely used in computer vision to automatically detect and classify the objects present in the scene. The Deeplabv3+ is based on an encoder-decoder architecture. While the encoder extracts high-level multi-scale features, the decoder is implemented using a combination of low-level and high-level features to generate the final pixel-wise segmentation map. Thanks to the Atrous Spatial Pyramid Pooling (ASPP) that applies multiple dilated convolutions in the Encoder model, DeepLabV3+ can achieve accurate results on objects at various scales. In the considered analysis, we exploited the model pre-trained on the ADE20K dataset,²⁵ while the semantic segmentation inference was carried out considering the ResNet 101 layers backbone. Nevertheless, alternative semantic segmentation methodologies can be employed, provided they can effectively characterize the landscape present in street-level pictures. According to the semantic segmentation results obtained, we focused the attention on the objects that are mostly present in the scene, i.e., namely, “building”, “car”, “earth”, “fence”, “field”, “grass”, “hill”, “house”, “land”, “mountain”, “path”, “plant”, “pole”, “railing”, “road”, “rock”, “sidewalk”, “signboard”, “sky”, “tree”, “wall”, “water”.

4.5 Semantic LC Consistency Analysis

In the final step, we analyze the semantic consistency of *in-situ* and satellite-based data. First, the *in-situ* LUCAS LC labels are compared with those predicted in the satellite-based thematic products. To assess the feasibility of using street-level images for enhancing the validation of thematic products, we analyze the statistical distribution of objects identified within the street-level images’ semantic segmentation results. The analysis is carried out per LC class, comparing samples with LC agreement between *in-situ* and satellite data sources with those in which LC differs. In particular, the aim is to check whether there is a clear discrepancy in terms of the presence of identified objects within the landscape for cases of agreement and disagreement.

Fig. 5 shows a qualitative example of the procedure carried to evaluate the semantic LC consistency of a “water” sample present in the LUCAS database and described as “Inland fresh running water” (see Fig.

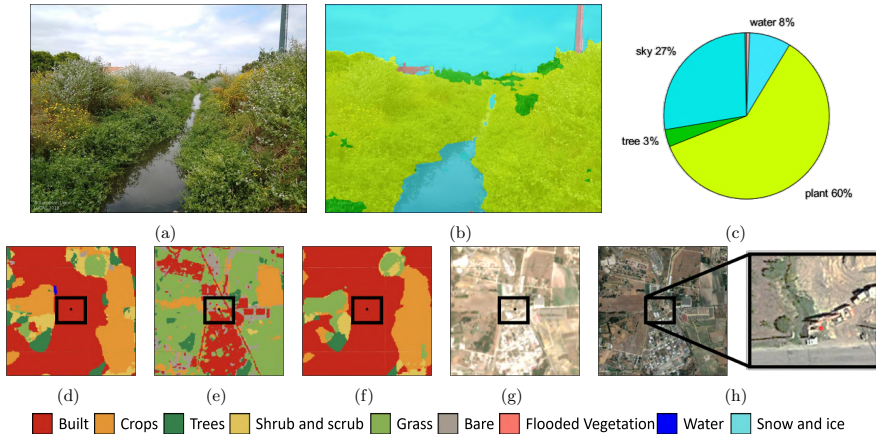


Figure 5: Qualitative example of the procedure carried out for the semantic LC consistency analysis considering a “water” sample associated with the LUCAS LC description “Inland fresh running water”: (a) LUCAS street-level picture, (b) semantic segmentation result, (c) statistical distribution of the semantic segmentation result, (d) Dynamic World, (e) WorldCover, (f) Esri, (g) Sentinel-2, and (h) VHR Google image. The corresponding LC is “Built”, “Crops”, and “Built” in the Dynamic World, WorldCover, and 2020 Esri thematic products, respectively.

5(a)). The corresponding LC information estimated in the thematic products is “Built”, “Crops”, and “Built” in the Dynamic World (see Fig. 5(d)), WorldCover (see Fig. 5(e)), and Esri thematic product (see Fig. 5(f)), respectively. This example shows disagreement between the LUCAS *in-situ* and the satellite-based LC since none of the considered thematic products identify the “water” LC category. Following the established validation approach typically used in the literature, these samples would result in misclassification, thus decreasing the accuracy of the thematic product. Although the street-level picture shows the presence of a river (see Fig. 5(a)), because of its small size it is barely visible in the VHR Google image (see Fig. 5(h)) and almost invisible in the Sentinel-2 data (see Fig. 5(g)). On one hand, it is unfeasible to assume that a photo-interpreter can manually check all misclassified samples to assess whether some of them cannot be correctly classified using satellite data. On the other hand, the result of the semantic segmentation of the street-level image (see Fig. 5(c)) provides detailed information about the composition of the landscape. In the considered example, a high number of pixels is associated with the “plant” class (60%), while a very small one is associated with “water” (8%). This analysis can be carried out automatically and can be used to characterize the uncertainty of the thematic product better.

5. EXPERIMENTAL SETUP

The experimental analysis was conducted on a subset of 10,432 LUCAS samples as a proof of concept. The samples were selected according to a stratified random sampling strategy applied all over the European Union. Fig. 6 shows the spatial distribution of LUCAS samples used to carry out the experiments. Also, the number of LUCAS samples per LC class is reported together with an example of Sentinel-2 image acquired over a LUCAS “Wetland” point. Since the objective of this paper is not to evaluate the accuracy of the considered thematic products, during the experimental analysis we concentrate on two categories of samples: those displaying a semantic discrepancy between all three LC products and the LUCAS *in-situ* LC information, i.e., *Disagreement* samples, and those having LC agreement across all the analyzed data sources, i.e., *Agreement* samples. Any LUCAS sample that agrees with only one or two thematic products is excluded from the analysis. Focusing on the agreement/disagreement of three thematic products allows us to exclude samples that may be incorrectly

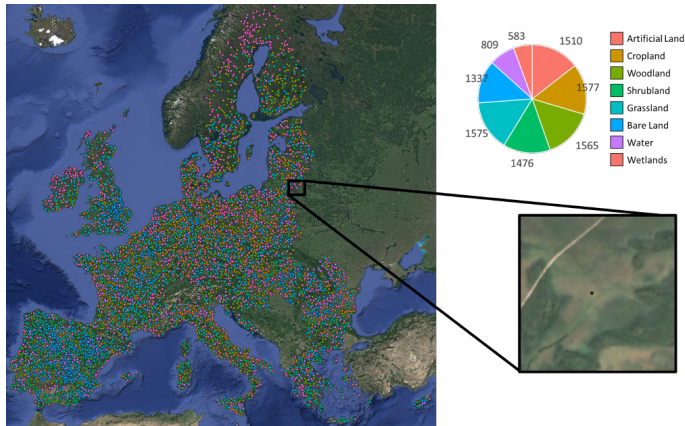


Figure 6: Spatial distribution of the subset of 10,432 LUCAS samples used to carry out the experimental analysis. The number of LUCAS samples per LC class is reported together with an example of Sentinel-2 image acquired over a LUCAS “Wetland” point.

classified due to limitations in the classification model and/or training dataset while examining the intrinsic semantic gap between ground and satellite perspectives. Tab. 1 reports the number of samples per LC class: (1) belonging to the subset of LUCAS data, (2) sharing a consistent pattern in the three thematic products, and (3) belonging to either the “Agreement” or “Disagreement” categories. Out of the 10,432 LUCAS samples, only 8,159 samples translate to either agreement (3,241 “Agreement” samples) or disagreement (4,918 “Disagreement” samples) for all three thematic products. As expected and in line with the experimental finding of Venter *et al.*,¹¹ the most critical classes having the highest disagreement between *in-situ* and satellite-based LC are “Bare”, “Shrub and scrub”, “Grass” and “Flooded Vegetation”. In the considered analysis, also several “Water” samples are misclassified since they are associated with small rivers that may be partially visible at 10m spatial resolution.

Table 1: Number of samples per LC class where there is agreement and disagreement among the labels recorded *in-situ* and those present in the three satellite-based thematic products. Out of the 10,432 LUCAS samples, 8,159 samples translate to either agreement (3,241 “Agreement” samples) or disagreement (4,918 “Disagreement” samples) for all three thematic products.

<i>in-situ</i> LUCAS data			Satellite-based Map data		LC Comparison (# samples)	
LC classes	# samples		LC classes	# samples	Agreement	Disagreement
Artificial land	1510		Built	981	575	406
Cropland	1577		Crops	1269	1174	95
Woodland	1565		Trees	1222	1086	136
Shrubland	1476		Shrub and scrub	693	81	612
Grassland	1575		Grass	602	137	465
Bareland	1337		Bare	1142	10	1132
Water	583		Water	455	86	369
Wetland	809		Flooded Vegetation	670	3	667
Total	10432		Total	8159	3241	4918

6. EXPERIMENTAL RESULTS AND DISCUSSION

Fig. 7 shows some qualitative examples of semantic segmentation output on “Agreement” and “Disagreement” samples across the LC classes, i.e., “Built”, “Crops”, “Trees”, “Shrub and scrub” and “Grass”, respectively. From qualitative analysis, one can notice that the samples misclassified in the satellite-based thematic products are characterized by complex scenes having mixed LC classes not comparable with the spatial resolution of the satellite data. In the “Built” LC, the “Agreement” sample shows that most of the pixels of the semantic segmentation output are associated with the “building”, “wall” and “road” classes, see Fig. 7(a). In contrast, the “Disagreement” sample has a strong presence of “trees”, “plants” and “grass” classes, see Fig. 7(c). Similarly, for the “Crops” LC class, in the “Disagreement” sample the semantic segmentation output represents a scene where a forest area is very close to the crop area, see Fig. 7(g). This can lead to a possible misclassification result in the thematic products because of the satellite perspective. Fig. 7(i) and Fig. 7(k) show “Agreement” and “Disagreement” samples for the “Tree” LC class. In this case, the semantic segmentation results point out that we are considering a dense and an open forest, respectively. This information can be automatically derived from the number of pixels associated with the “trees” class compared to the “earth” class. Similarly, for the “Shrub and scrub” class, the “Disagreement” sample is highly affected by the presence of bushes that can be confused with trees. Finally, Fig. 7(t) shows an example of “Grass” LC present in an urban area dominated by concrete.

These qualitative results are confirmed by the analysis of the statistical distribution of the objects present in the semantic segmentation results. Fig. 8 and Fig. 9 illustrate the average number of pixels per object present in the semantic segmentation results for both the “Agreement” and “Disagreement” samples across the various LC classes, i.e., “Built”, “Crops”, “Trees”, “Shrub and scrub” and “Grass”, “Bare”, “Water”, “Flooded vegetation”, respectively. By analyzing the results in detail, one can notice a conspicuous discrepancy between the information contained within the satellite data and the data acquired *in-situ* through surveying activities. This is highlighted by the divergence in terms of the number of objects visible in the street-level images when comparing the *Agreement* and *Disagreement* samples. For instance, for the “Built” LC class, the misclassified samples (Fig. 8(b)) have more trees in the scene than those present in the correctly classified samples (see Fig. 8(a)), which are dominated by objects belonging to artificial surfaces such as “wall”, “road”, and “building”. For the “Crops” LC class, the correctly classified samples (see Fig. 8(c)) are mainly dominated by “field” and “plant” objects, while the misclassified ones have a strong presence of “trees” in the scene (see Fig. 8(d)). For the “Water” LC class, the presence of “water” in the correctly classified samples (see Fig. 9(e)) is much higher than the one detected in street-level images associated with the misclassified samples (see Fig. 9(f)). This may happen when the recorded LC information lies within a circle of 1.5 m radius (e.g., small river), thus leading to a significantly smaller MMU compared to the spatial resolution of Sentinel-2 satellite data. Similarly, for the “Tree” LC class, the presence of “trees” in the correctly classified samples (see Fig. 8(e)) is much higher compared to the one visible in the misclassified samples (see Fig. 8(f)). In particular, the obtained semantic segmentation outputs suggest that there is a high probability that the “Agreement” and “Disagreement” samples belonging to the “Tree” class are associated with closed forest and open forest, respectively. This assumption can be derived by checking the landscape composition reported in Fig. 8(e) and Fig. 8(f). For the “Grass” class, the “Agreement” samples (see Fig. 9(a)) are dominated by the presence of “grass” in the street-level pictures, while more heterogeneous landscapes are visible in the “Disagreement” samples (see Fig. 9(b)). In general, one can notice that a sharp discrepancy in terms of object occurrences is visible in all the considered LC classes.

The obtained results demonstrate that during the validation procedure it’s essential to consider the specific properties of the satellite data in question, such as spatial resolution. Such analysis is implicitly carried out when using a validation set of samples derived through photo-interpretation. This involves utilizing the same satellite data used in the experiments, such as the global validation set made up of photo-interpreted samples based on Sentinel-2 data. While the validation dataset enables an effective comparison of LC detected by the automatic classification model and the photo-interpreted one, it does not calculate accuracy based on the real LC present in the scene. The quantitative evaluation based on LC information collected *in-situ* data combined with the semantic information provided by street-level images can lead to a comprehensive evaluation of the accuracy of current thematic products.

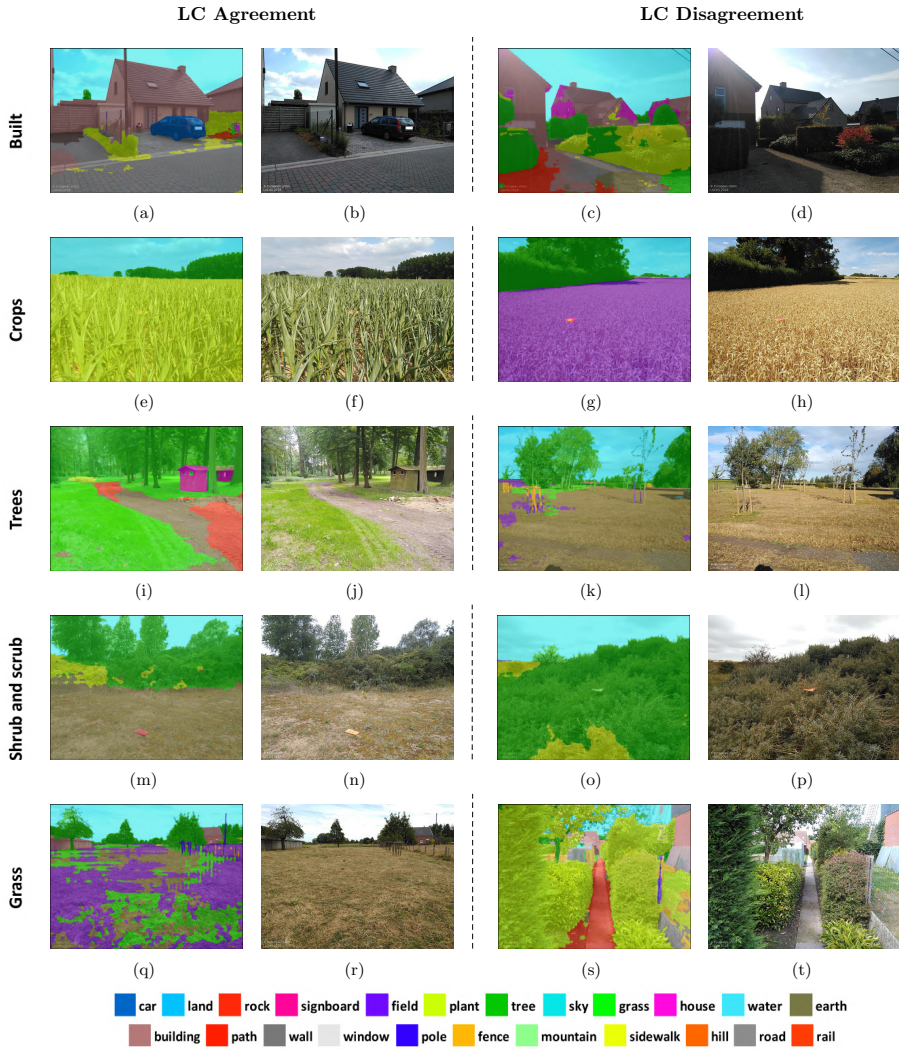


Figure 7: Qualitative examples of semantic segmentation results obtained on the street-level pictures belonging to: (a)-(d) “Built”, (e) - (h) “Crops”, (i) - (l) “Trees”, (m) - (p) “Shrub”, and (q)-(t) “Grass”. While the examples on the left belong to the *Agreement* samples, those on the right belong to the *Disagreement* samples.

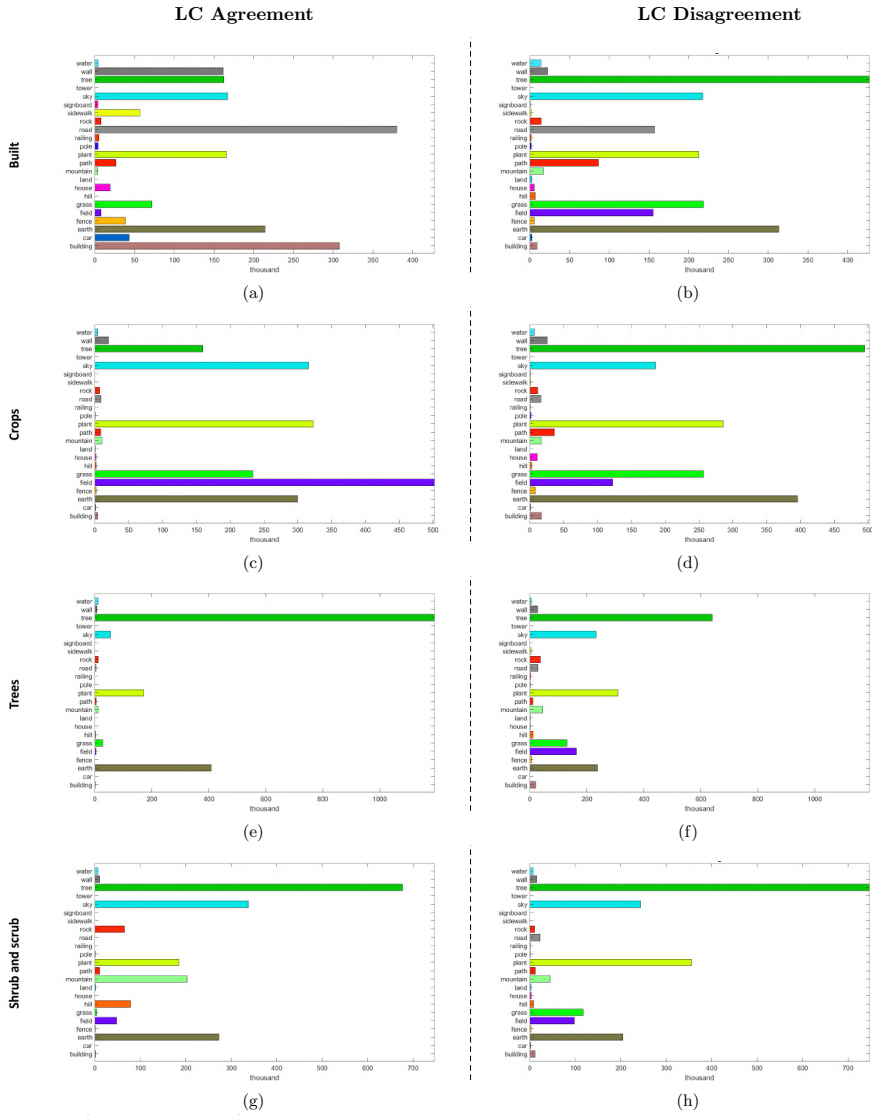


Figure 8: Average number of pixels associated with the objects present in the semantic segmentation output obtained on the “Agreement” and “Disagreement” samples, respectively, for the class: (a)(b) “Built”, (c)(d) “Crop”, (e)(f) “Tree”, and (g)(h) “Shrub and scrub”.

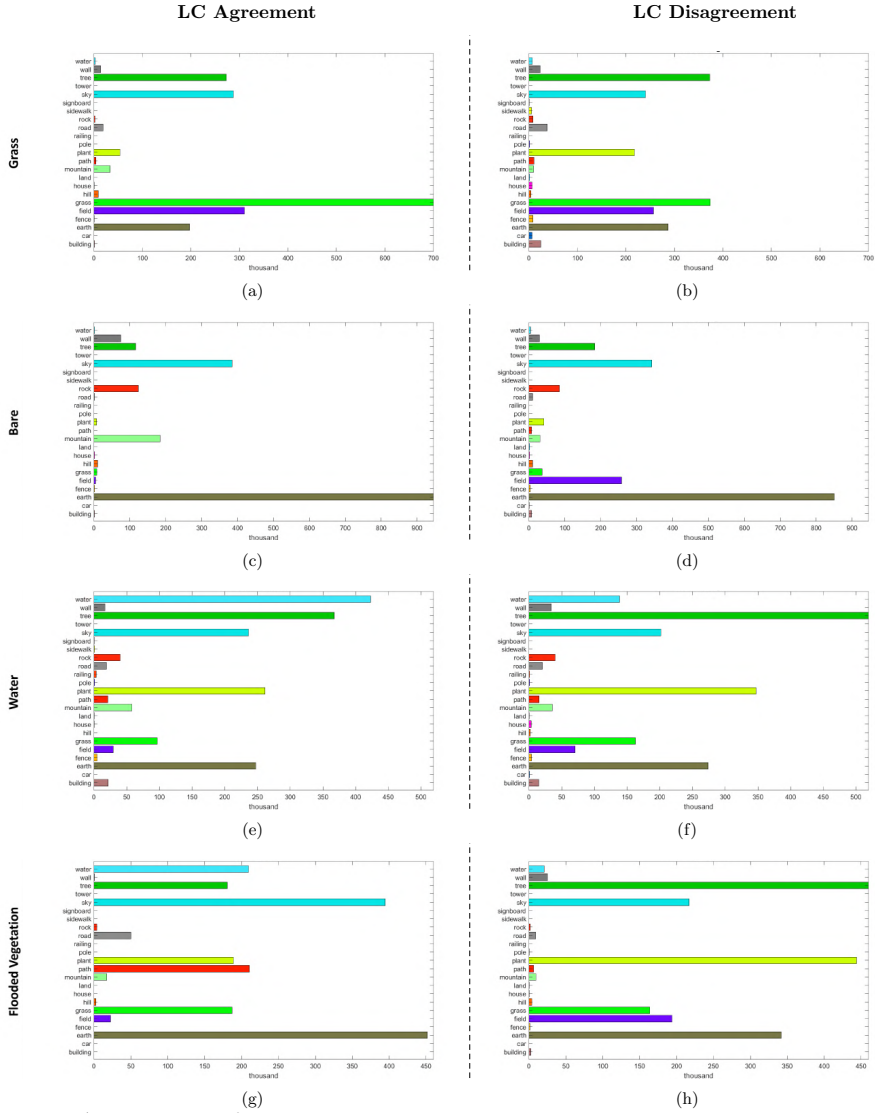


Figure 9: Average number of pixels associated with the objects present in the semantic segmentation output obtained on the “Agreement” and “Disagreement” samples, respectively, for the class: (a)(b) “Bare”, (c)(d) “Grass”, (e)(f) “Water”, and (g)(h) “Flooded vegetation”.

7. CONCLUSIONS

This paper analyzed the potential of using street-level geo-referenced images to bridge the differences between field surveys and satellite data to enhance the validation process of existing thematic products. The experimental results confirmed the expected semantic gap between *in-situ* and the satellite-based thematic products because of their different viewpoints (i.e., ground versus aerial) and different MMU of the LC present in the scene versus the spatial resolution of the satellite data. From the results obtained we can conclude that, by capitalizing on the information provided by the street-level pictures, it is possible to identify the different properties of the considered landscape automatically. Indeed, street-level imagery provides a clear overview of the objects in the scene that can be automatically retrieved with semantic segmentation algorithms. Relying also on this information in the validation of thematic satellite products can lead to more comprehensive evaluation results. We would like to remark that the considered experimental analysis has been carried out in Europe because of the availability of the LUCAS databases. However, with the continuous surge in the number of crowdsourced images, there exists an opportunity to harness street-level images for a more thorough validation of the existing thematic products on a global scale. As future development, we aim to further investigate the use of street-level images to gain more insights on both the LU and LC properties of the scene. Moreover, we would like to compare the *in-situ* LUCAS database with thematic products characterized by detailed classification schemes such as the Corine Land Cover (CLC) map having 44 classes.

8. ACKNOWLEDGMENTS

Part of this work was performed in the ADMIRE project, which received funding from the European Union's Horizon 2020 JTI-EuroHPC research and innovation programme under the grant agreement no. 956748.

REFERENCES

1. Friedl, M. A., McIver, D. K., Hodges, J. C., Zhang, X. Y., Muchoney, D., Strahler, A. H., Woodcock, C. E., Gopal, S., Schneider, A., Cooper, A., et al., "Global land cover mapping from modis: algorithms and early results," *Remote sensing of Environment* 83(1-2), 287–302 (2002).
2. Li, W., MacBean, N., Ciais, P., Defourny, P., Lamarche, C., Bontemps, S., Houghton, R. A., and Peng, S., "Gross and net land cover changes in the main plant functional types derived from the annual esa cci land cover maps (1992–2015)," *Earth System Science Data* 10(1), 219–234 (2018).
3. Loveland, T. R., Reed, B. C., Brown, J. F., Ohlen, D. O., Zhu, Z., Yang, L., and Merchant, J. W., "Development of a global land cover characteristics database and igbp discover from 1 km avhrr data," *International journal of remote sensing* 21(6-7), 1303–1330 (2000).
4. Paris, C. and Bruzzone, L., "A novel approach to the unsupervised extraction of reliable training samples from thematic products," *IEEE Transactions on Geoscience and Remote Sensing* 59(3), 1930–1948 (2021).
5. Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., et al., "Global land cover mapping at 30 m resolution: A pok-based operational approach," *ISPRS Journal of Photogrammetry and Remote Sensing* 103, 7–27 (2015).
6. Liu, H., Gong, P., Wang, J., Wang, X., Ning, G., and Xu, B., "Production of global daily seamless data cubes and quantification of global land cover change from 1985 to 2020-imap world 1.0," *Remote Sensing of Environment* 258, 112364 (2021).
7. Friedl, M. A., Woodcock, C. E., Olofsson, P., Zhu, Z., Loveland, T., Stanimirova, R., Arevalo, P., Bullock, E., Hu, K.-T., Zhang, Y., et al., "Medium spatial resolution mapping of global land cover and land cover change across multiple decades from landsat," *Frontiers in Remote Sensing* 3, 894571 (2022).
8. Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., et al., "Esa worldcover 10 m 2021 v200," (2022).
9. Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., et al., "Dynamic world, near real-time global 10 m land use land cover mapping," *Scientific Data* 9(1), 251 (2022).
10. Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., and Brumby, S. P., "Global land use/land cover with sentinel 2 and deep learning," in [2021 IEEE international geoscience and remote sensing symposium IGARSS], 4704–4707, IEEE (2021).
11. Venter, Z. S., Barton, D. N., Chakraborty, T., Simensen, T., and Singh, G., "Global 10 m land use land cover datasets: A comparison of dynamic world, world cover and esri land cover," *Remote Sensing* 14(16), 4101 (2022).
12. d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L., Haub, P., Buck, O., Haub, C., Eiselt, B., and van der Velde, M., "Lucas cover photos 2006–2018 over the eu: 874 646 spatially distributed geo-tagged close-up photos with land cover and plant species label," *Earth System Science Data* 14(10), 4463–4472 (2022).
13. Seiferling, I., Naik, N., Ratti, C., and Proulx, R., "Green streets- quantifying and mapping urban trees with street-level imagery and computer vision," *Landscape and Urban Planning* 165, 93–101 (2017).
14. Su, Y.-F., Lin, Y.-T., Jang, J.-H., and Han, J.-Y., "High-resolution flood simulation in urban areas through the application of remote sensing and crowdsourcing technologies," *Frontiers in Earth Science* 9, 756198 (2022).

15. d'Andrimont, R., Yordanov, M., Lemoine, G., Yoong, J., Nikel, K., and Van der Velde, M., "Crowdsourced street-level imagery as a potential source of in-situ data for crop monitoring," *Land* 7(4), 127 (2018).
16. Fritz, S., McCallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., Van der Velde, M., Kraxner, F., and Obersteiner, M., "Geo-wiki: An online platform for improving global land cover," *Environmental Modelling & Software* 31, 110–123 (2012).
17. See, L., Fritz, S., Perger, C., Schill, C., McCallum, I., Schepaschenko, D., Duerauer, M., Sturn, T., Karner, M., Kraxner, F., et al., "Harnessing the power of volunteers, the internet and google earth to collect and validate global spatial information using geo-wiki," *Technological Forecasting and Social Change* 98, 324–335 (2015).
18. d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L., Eiselt, B., Palmieri, A., Dominici, P., Gallego, J., Reuter, H. I., Joebges, C., Lemoine, G., et al., "Harmonised lucas in-situ land cover and use database for field surveys from 2006 to 2018 in the european union," *Scientific data* 7(1), 352 (2020).
19. Laso Bayas, J. C., See, L., Fritz, S., Sturn, T., Perger, C., Dürauer, M., Karner, M., Moorthy, I., Schepaschenko, D., Domian, D., et al., "Crowdsourcing in-situ data on land cover and land use using gamification and mobile technology," *Remote Sensing* 8(11), 905 (2016).
20. Unger, S., Rollins, M., Tietz, A., and Dumais, H., "inaturalist as an engaging tool for identifying organisms in outdoor activities," *Journal of Biological Education* 55(5), 537–547 (2021).
21. van Der Velde, M., Goëau, H., Bonnet, P., d'Andrimont, R., Yordanov, M., Affouard, A., Claverie, M., Czúcz, B., Elvekjær, N., Martinez-Sanchez, L., et al., "Pl@ntnet crops: merging citizen science observations and structured survey data to improve crop recognition for agri-food-environment applications," *Environmental Research Letters* 18(2), 025005 (2023).
22. Gao, Y., Liu, L., Zhang, X., Chen, X., Mi, J., and Xie, S., "Consistency analysis and accuracy assessment of three global 30-m land-cover products over the european union using the lucas dataset," *Remote Sensing* 12(21), 3479 (2020).
23. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L., "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence* 40(4), 834–848 (2017).
24. Guo, J., He, H., He, T., Lausen, L., Li, M., Lin, H., Shi, X., Wang, C., Xie, J., Zha, S., et al., "Gluoncv and gluonnlp: Deep learning in computer vision and natural language processing," *The Journal of Machine Learning Research* 21(1), 845–851 (2020).
25. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., and Torralba, A., "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision* 127, 302–321 (2019).

Paper III

Sen4Map: Advancing Mapping with Sentinel-2 by Providing Detailed Semantic Descriptions and Customizable Land-Use and Land-Cover Data

S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris

IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), 2024

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License

(<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

Surbhi Sharma wrote the code and the main part of the paper, ran all experiments on the HPC machine and performed the analysis.

Sen4Map: Advancing Mapping with Sentinel-2 by Providing Detailed Semantic Descriptions and Customizable Land-Use and Land-Cover Data

Surbhi Sharma, *Student Member, IEEE*, Rocco Sedona, *Member, IEEE*, Morris Riedel, *Member, IEEE*,
Gabriele Cavallaro, *Senior Member, IEEE*, and Claudia Paris, *Senior Member, IEEE*

Abstract—This paper presents Sen4Map, a large-scale benchmark dataset designed to enhance the capability of generating land-cover maps using Sentinel-2 data. Comprising non-overlapping 64×64 patches extracted from Sentinel-2 time series images, the dataset spans 335,125 geo-tagged locations across the European Union. These locations are associated with detailed land-cover and land-use information gathered by expert surveyors in 2018. Unlike most existing large datasets available in the literature, the presented database provides: (1) a detailed description of the land-cover and land-use properties of each sampled area; (2) independence of scale, as it is associated with reference data collected *in-situ* by expert surveyors; (3) the ability to test both temporal and spatial classification approaches because of the availability of time series of 64×64 patches associated with each labeled sample; and (4) samples were collected following a stratified random sample design to obtain a statistically representative spatial distribution of land-cover classes throughout the European Union. To showcase the properties and challenges offered by Sen4Map, we benchmarked the current state-of-the-art land-cover classification approaches. The dataset and code can be downloaded at: <https://datapub.fz-juelich.de/sen4map>.

Index Terms—Land Use and Coverage Area frame Survey (LUCAS), land-use and land-cover mapping, Sentinel-2, supervised classification, benchmark dataset, machine learning.

I. INTRODUCTION

The expanded deployment of Earth Observation (EO) satellites produces vast quantities of Remote Sensing (RS) data in multiple modalities, such as optical and radar, with resolutions from submetric to decametric. Extracting meaningful information from these varied datasets is crucial for effectively addressing environmental challenges such as urbanization, deforestation, climate change, and global warming [1]. This extraction provides important insights into environmental dynamics and supports sustainable decision-making processes.

Classification maps, produced through the analysis of RS data, stand as key thematic outputs. They offer crucial insights into the Earth's physical and functional characteristics by distinguishing between natural landscapes and human activities across various time periods. They differentiate land-cover, which describes physical types such as corn fields

or grasslands, from land-use, revealing how these areas are employed by humans, for instance, for agriculture. Given the vast amounts of data from satellite image time series, developing automatic methods for producing classification maps has become essential [2].

The introduction of the Copernicus Programme, providing free and open access to RS data, along with advancements in Machine Learning (ML), has revolutionized Earth's surface monitoring. This combination has significantly improved the accuracy of land-cover map generation [3]. The Sentinel-2 constellation, part of this program, acquires multispectral images having 13 spectral bands, out of which four bands at a spatial resolution of 10 m, six bands at 20 m, and three bands at 60 m. The Sentinel-2 constellation has a temporal resolution of five days [4]. These satellite image time series, covering the global surface with short revisit periods, provide the foundation for large-scale mapping such as detailed crop type mapping [5], forest monitoring [6], water body detection [7], and the production of global thematic products at a 10 m spatial resolution [8], [9].

To facilitate the development of workflows based on novel Deep Learning (DL) algorithms and the evaluation and comparison of classical ML approaches, several large-scale datasets based on Sentinel-2 data have been released. These benchmarks address different tasks, such as multi-sensor and multi-modal data fusion [10], cloud removal [11] [12], super-resolution [13], visual question answering [14], as well as large-scale datasets for unsupervised pre-training and transfer learning [15] [16] [17] [18]. The availability of web platforms such as HyperLableMe [19] also facilitates benchmarking of RS image classifiers with their harmonized large-scale dataset containing labeled and heterogeneous multispectral and hyperspectral images.

The release of EO benchmarks is an essential trend in establishing a large-scale, open, and potentially interoperable ecosystem of open-access datasets. This trend is vital for keeping pace with the escalating demands of large-scale DL models, such as foundation models, which are increasingly data-intensive [20]. DL models for EO necessitate substantial resources for collecting and compiling the datasets required for training. Providing the community with large-scale land-cover and land-use datasets not only offers an immediately beneficial resource but also acts as a model for future expansions, facilitating significant progress in mapping the Earth's surface [18]. Within the domain of land-cover classification, datasets

Surbhi Sharma, Gabriele Cavallaro and Morris Riedel are with the Jülich Supercomputing Centre (JSC), Wilhelm-Johnen-Straße 52428 Jülich, Germany, and with the University of Iceland, 107 Reykjavík, Iceland (e-mail: su.sharma@fz-juelich.de, morris@hi.is, g.cavallaro@fz-juelich.de).

Rocco Sedona is with the Jülich Supercomputing Centre (JSC), Wilhelm-Johnen-Straße 52428 Jülich, Germany (e-mail: r.sedona@fz-juelich.de.)

Claudia Paris is with the University of Twente, Enschede, the Netherlands (e-mail: c.paris@utwente.nl).

such as BigEarthNet [21], EuroSAT [22], Worldcover [23], and Dynamic World [8] are prominent. Worldcover is an European Space Agency (ESA) product at the global level, recently generated using Sentinel-1 and Sentinel-2 images by training a Random Forest (RF) classifier with a labeled dataset of 141,000 training samples generated through manual annotation of Google Maps and Microsoft Bing images. The map legend includes 11 widely spread generic classes and also has been used to generate the global Google's Dynamic World [8] and Esri's [24] land-cover products. Unlike the ESA Worldcover Map, these land-cover products were generated using convolutional DL models trained with labeled data derived from manual interpretation of Sentinel-2 images. A high spatial resolution of the CORINE land-cover (CLC) map, the CLC+, has been recently released [25] to provide the first European land-cover product in unprecedented detail, leveraging Sentinel-1 and Sentinel-2 images acquired in 2018. Based on the temporal convolutional DL model, the map aims to depict 11 land-cover categories as presented in [26]. On the other hand, using these benchmark datasets can present challenges from an operation viewpoint. They usually feature a legend with fewer than 15 land-cover classes, reducing the level of detail in Earth's surface descriptions and oversimplifying the complexity and diversity of land-cover types. When annotations are performed through photo-interpretation, labeled datasets heavily rely on the EO data used. Additionally, their specific map legends may render them inadequate for testing different classification schemes or improving the semantic detail a land-cover map can offer [27].

This paper presents a benchmark dataset, Sen4Map, which is designed to advance the state-of-the-art of land-cover mapping applied to Sentinel-2 satellite data. Differently from the existing datasets, Sen4Map is the first large-scale dataset to include 335,125 time series of 64×64 multispectral patches extracted from Sentinel-2 tiles, each associated with geolocated land-cover data collected *in-situ* by experienced surveyors [28], [29]. To this end, we leveraged the public availability of the LUCAS database, a triennial ground survey coordinated by the Statistical Office of the European Commission (Eurostat) that collects land-cover and land-use data over the whole European Union [29] since 2001. The LUCAS data have already been used to validate existing thematic products [30], [31], as well as to generate land-cover maps at country [32] or European scale [33]. However, all these approaches have made partial use of the extremely valuable information provided by the LUCAS database, generally considering only high-level land-cover categories (e.g., cropland, artificial surfaces, water bodies, etc.) and completely neglecting land-use categories. In addition, most approaches using Sentinel-2 data [32], [33] or Landsat data [34] considered only the LUCAS points associated with homogeneous land-cover, by filtering out a high amount of samples that may be difficult to classify.

In this paper, we aim to enhance the capability of Sentinel-2 for land-cover mapping by fully leveraging the challenges presented by the real *in-situ* LUCAS data collection, which provides land-cover information not specifically tailored to the properties of Sentinel-2 satellite data. Specifically, the main contributions of the proposed reference dataset are as follows:

- The dataset provides detailed, geo-tagged land-cover and land-use information associated with the central pixel of each 64×64 patch extracted from Sentinel-2 tiles, enabling an assessment of the limitations of satellite data in producing semantically detailed land-cover maps.
- The scale-independent land-cover information, collected *in-situ* by experienced surveyors, facilitates the analysis of Sentinel-2's ability to accurately map land-covers of various sizes.
- The provision of ready-to-use 64×64 Sentinel-2 time series data supports evaluating and comparing land-cover mapping algorithms that utilize both spatial and temporal dimensions.
- Sen4Map's design, based on a stratified random sampling across the European Union, assesses the generalization capabilities of land-cover mapping techniques, both within individual countries and across the EU.

The rest of this article is organized into five sections. Section II provides an overview of the proposed approach for developing the Sen4Map dataset. Section III describes the information included in the datasets, including spatial and temporal properties. Section IV discusses the experimental results obtained by state-of-the-art classification algorithms and highlights the properties and challenges presented by Sen4Map. Finally, Section V concludes this article and presents potential future developments.

II. SEN4MAP GENERATION

This section outlines the methodology for developing the Sen4Map, leveraging two data sources: the 2018 Sentinel-2 image time series and the 2018 LUCAS field data. Subsequent sections provide further details.

A. LUCAS Database

Sen4Map includes geo-tagged locations of the LUCAS points, which provide information on land-cover (the physical coverage observed on the Earth's surface) and land-use (the socio-economic function of the observed Earth's surface). Moreover, for each LUCAS point, we can link to information that is available in the dataset such as several environmental parameters, a downward-facing photo of the target point (P), and four landscape photos facing the cardinal directions (N, E, S, W), as described in [29]. These data, along with detailed metadata concerning the quality of the observations, including the acquisition date, the GPS location of the surveyor in the field, and the theoretical location of the target point (used as the central pixel of the 64×64 patch), are publicly available. Given the recent launch of Sentinel-2 in August 2015, the initial version of Sen4Map primarily utilizes the consolidated 2018 LUCAS database. This approach facilitates the creation of an annual time series of contemporary Sentinel-2 images for each LUCAS point. Nevertheless, our goal is to regularly update the Sen4Map dataset with recent Sentinel-2 satellite acquisitions following the release of the consolidated 2022 LUCAS dataset. The spatial distribution of the 2018 LUCAS samples collected all over the European Union is shown in Figure 1 by representing the main land-cover classes in different colors.

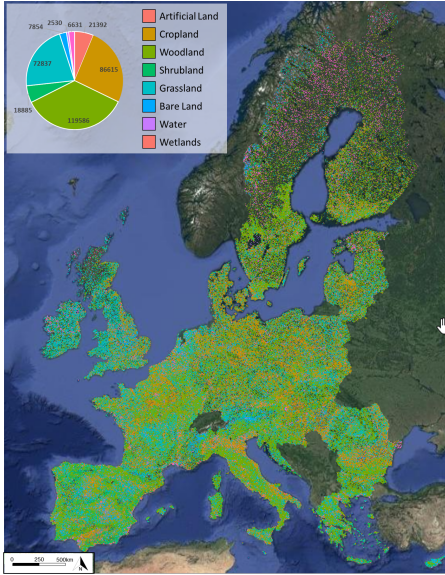


Fig. 1: Spatial distribution of the 2018 LUCAS samples collected all over the European Union. The different land-cover categories are highlighted with different colors.

B. Sentinel-2 Database

Sen4Map is based on multispectral time series of patches, each 64×64 pixels, extracted from Sentinel-2 tiles of 100×100 km, acquired in 2018 by the Sentinel-2A and Sentinel-2B satellites. Launched in June 2015 and March 2017, respectively, these land-monitoring satellites operate in a sun-synchronous orbit. They are both equipped with a Multispectral Instrument which can capture 13 spectral bands. The bands B01 (Aerosols), B09 (Water Vapour), and B10 (Cirrus) feature a spatial resolution of 60 meters and central wavelengths of 443 nm, 945 nm, and 1375 nm, respectively. Bands B05 (Red Edge 1), B06 (Red Edge 2), B07 (Red Edge 3), B08A (Red Edge 4), B11 (SWIR 1), and B12 (SWIR 2) offer a spatial resolution of 20 meters with central wavelengths of 705 nm, 740 nm, 783 nm, 865 nm, 1610 nm, and 2190 nm, respectively. Finally, bands B02 (Blue), B03 (Green), B04 (Red), and B08 (NIR) provide a spatial resolution of 10 meters with central wavelengths of 490 nm, 560 nm, 665 nm, and 842 nm, respectively. The Sen4Map dataset provides the Sentinel-2 bands only from 10 m and 20 m spatial resolution since these bands are typically used for land-cover mapping. The bands with 60 m spatial resolution such as Aerosols, Water Vapour, and Cirrus are not included in the dataset, as they are primarily intended for atmospheric corrections. Due to the short repeat cycle for the two-satellite constellation

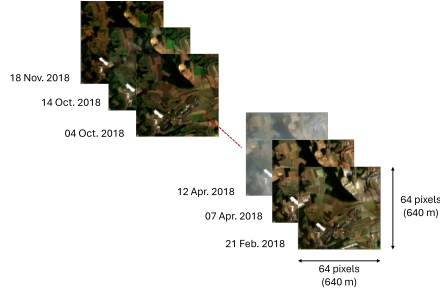


Fig. 2: Example of a Sentinel-2 time series from the Sen4Map benchmark dataset associated with one LUCAS point. In this qualitative example, only the images acquired from the 21st of February until the 18th of November are visualized to capture the temporal component that can characterize different land cover classes.

(Sentinel-2A and Sentinel-2B) of approximately 5 days, the time series images obtained are ideal for continuous mapping and monitoring of the Earth's surface. Figure 2 shows a qualitative example of Sentinel-2 time series reported for a 64×64 patch, where only the true color compositions are represented per acquisition.

III. SEN4MAP PROPERTIES

A. Spatio-Temporal Properties

The Sentinel-2 tiles considered for the extraction of a time series of patches were those with cloud coverage of less than 40% and with less than 5% no-pixel data. We selected the locations where LUCAS points are available (i.e., theoretical locations, 2006-2018), with the support of Google Earth Engine (GEE) [35]. LUCAS is carried out systematically using a distributed field design based on a $2 \text{ km} \times 2 \text{ km}$ grid throughout the European Union. The points of intersection from the grid are referred to as theoretical locations, “*th_long*” and “*th_lat*” [29]. This stratified spatial distribution allows for properly mapping the different land-cover conditions while reducing the risk of spatial autocorrelation, which is extremely important for land-cover mapping. Although, the field survey is carried out based on theoretical locations, the discrepancy between the actual surveyed location, “*gps_lon*” and “*gps_lat*” and the theoretical locations, could also induce spatial errors. A buffer around the point can therefore help mitigate this based on *in-situ* observations of the multi-year harmonized LUCAS survey database [29]. We created a 650 m buffer around each point to extract patches of 64×64 size. The choice of a 64×64 patch is critical for various ML algorithms. Since several Sentinel-2 tiles in the GEE image collection are rotated by differing orbits, there is a need to compensate for NO DATA at the edges, requiring the buffer size of 650 m. This results in some patches being larger than 64×64 , which are subsequently cropped to the target size. Each time series includes additional

information on scene classification maps “SCL” at a spatial resolution of 20 m, along with spatial resolutions of 10 m and 20 m.

Out of a total of 337,854 LUCAS points from the year 2018, only 2,704 points had no Sentinel-2 times series images considering less than 40 % of cloud coverage and less than 5 % of no-pixel data. Another 25 LUCAS points where no in-situ data was available (i.e., no labels) were also discarded. Therefore, the Sen4Map dataset results in a total of 335,125 time-series patches across 28 European countries: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Ireland, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, and the United Kingdom. The distribution of the LUCAS points all over the European Union is shown in Figure 3 by presenting per country the number of geotagged locations where *in-situ* data are available. The average number of Sentinel-2 images available per time-series for the respective LUCAS locations are reported in Figure 4. We considered only Sentinel-2 tiles with cloud coverage lower than 40%, leading to time series of varying lengths and temporal sampling rates across different LUCAS locations. Spatial and temporal harmonization of satellite data is a well-known challenge in large-scale land-cover mapping. Various algorithms for temporal feature sampling [36], [37] can be employed to mitigate differences in the acquisition times of the first and last images, as well as the diverse temporal sampling and sequence lengths associated with each LUCAS point time series.

B. Land-cover and Land-Use Information

For each LUCAS point, a hierarchical representation of the associated land-cover and land-use is reported. The land-cover types reported in the first level provide a high-level overview of what is present in the scene, by distinguishing “Artificial land” (A00), “Cropland” (B00), “Woodland” (C00), “Shrubland” (D00), “Grassland” (E00), “Bareland and lichens/moss” (F00), “Water areas” (G00) and “Wetlands” (H00). Similarly, the first level of land-use type aims at classifying “Primary sector” (U100), “Secondary sector” (U200), “Tertiary sector transport, utilities and residential” (U300) and “Unused and abandoned areas” (U400). The statistical distribution of the number of Sentinel-2 images associated with higher-level land-cover and land-use classes is shown in Figure 5. The detailed country-wise statistics of the Sen4Map dataset for the higher-level hierarchy of LUCAS labels are shown in the Appendix.

The land-cover labels are defined according to the land-cover Classification System (LCCS) [38], a comprehensive reference system designed for land-cover mapping. The LCCS aims to provide extensive flexibility, enabling the description of land-cover features at various scales or levels of detail. Rather than offering a single, universal thematic legend, it standardizes terminology and attributes used to define land-cover classes. The system utilizes independent diagnostic criteria, facilitating correlation with existing classifications and legends regardless of the mapping scale or methods used.

The descriptions of the land-cover **Lc1** and land-use **Lu1** types are very detailed in the second level of the hierarchy,

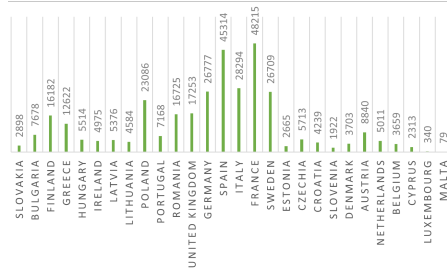


Fig. 3: Country-wise distribution of LUCAS surveyed locations with available *in-situ* data.

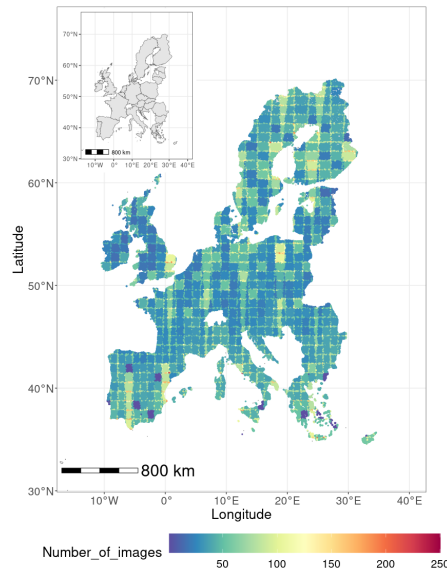


Fig. 4: Number of Sentinel-2 images having less than 40% of cloud coverage and associated with each LUCAS point.

where the specific combination of numbers and digits provides a detailed definition of what is present in the scene, e.g., the “A11” category refers to “Buildings with one to three floors.” Similarly, “U112” is associated with fallow land to describe when the land is grazed and not used for crop production. In this framework, intersecting the information provided by the land-use and land-cover descriptions allows for an extremely detailed characterization of the geo-tagged location. For instance, parking areas close to an area used for forestry purposes will have a land-cover label “A21” (Non-

built-up area features) and land-use “U120” (Areas used for forestry purposes). Moreover, for patches where more than one land-cover co-exists, a second land-cover **Lc2** and land-use **Lu2** information are reported together with their percentages of coverage.

C. Structure of the Dataset

The Sen4Map dataset is available for 28 countries in individual HDF5 files. The HDF5 format is highly parallelizable and optimized for High-performance Computing (HPC), making it scalable and faster for processing [39]. This format allows data to be stored in self-explanatory structures of the time-series images, where time-series Sentinel-2 images for each point can be processed individually with their corresponding attributes. Each HDF5 file contains datasets associated with LUCAS points from a single country in the form of a three-dimensional array of Sentinel-2 64×64 time-series patches, individually named as ‘*Lucas_Point_{Point_ID}*’ with their corresponding 119 attributes. Some examples of true color compositions of 64×64 Sentinel-2 patches belonging to different countries are shown in Figure 6.

IV. LAND-COVER MAPPING - BENCHMARKING

This section provides a comparison of the state-of-the-art classification algorithms most commonly used to generate land-cover maps, emphasizing the properties and challenges posed by the proposed Sen4Map dataset. In particular, we tested and compared (1) the pixel-based Random Forest classifier [40], widely used for its robustness in the presence of noisy data [41]; (2) the pixel-based Transformer DL model [42], which is extremely effective in handling mapping tasks requiring time series of satellite data; (3) the standard spatial ViT model [43], typically utilized for image processing through its attention mechanism; and (4) the Video Vision Transformer [44], capable of integrating both spatial and temporal information provided by time series of satellite data.

A. Experimental Setup

Thanks to the detailed description of the land-use and land-cover properties of the LUCAS samples, customized classification tasks can be defined. In the experimental setup we consider, we provide two examples: (1) the classification of the most widespread land-cover categories typically mapped at global scale [31], i.e., namely, “Water”, “Broadleaves”, “Shrubland”, “Conifers”, “Grass”, “Crops”, “Wetland”, “Artificial Land” and “Bareland”, and (2) a detailed crop classification task to map “Cereals”, “Root Crops”, “Non-permanent Industrial crop”, “Dry pulses, vegetables, flowers”, “Fodder crops”, “Bareland”, “Woodland and Shrubland”, and “Grassland” according to the map legend presented in [45]. We describe how we generate these legends from the detailed LUCAS land-cover labels in Table I and Table II, respectively. However, users have the flexibility to define alternative classification tasks based on their specific requirements.

The dataset used for training contains 70% of randomly selected points from each country. Similarly, from the exclusive

TABLE I: Detailed mapping of LUCAS labels for broad land-cover classification

Target land-cover	LUCAS land-cover
Artificial Land	A- Artificial Land (Except A22)
Bareland and Lichens/Moss	A22 – Non built-up linear features F – Bareland and Lichens/Moss
Grassland	E – Grassland B50 – Fodder Crops
Cropland	B10 – Cereals B20 – Root Crops B30 – Non-Permanent Industrial Crop B40 – Dry Pulses, Vegetables and Flowers B70 – Permanent Fruits B80 – Other Permanent Crops BX1 BX2
Broadleaves	C10 – Broadleaved Woodland
Conifers	C20 – Coniferous Woodland C30 – Mixed Woodland CXX – LUCAS SU FT
Shrubland	D- Shrubland
Water	G – Water Areas
Wetlands	H – Wetlands

remaining points per country, validation, and test sets were formed, each containing 15% of randomly selected exclusive data points from each country. In the case of the (1) land-cover classification task, we use 234,555 points for training, 50,284 points for validation, and 50,286 for testing. For the second crop classification task, we use 196,514 points for training, 48,394 points for validation, and 48,396 for testing, by considering only points having LUCAS labels as shown in Table II.

The Random Forest classifier and the pixel-based Transformer DL model were trained considering the center pixel of the 64×64 patch for both classification tasks. To harmonize the Sentinel-2 data available for different locations from the temporal point of view, a time series of 12 monthly composites was generated for each LUCAS point. To this end, we considered the widely used median composite approach described in [46], by masking out the clouds according to the information available in the “SCL” classification map associated with each Sentinel-2 image. However, we would like to remark that the proposed Sen4Map dataset provides the original time-series of Sentinel-2 images. Therefore, any approach can be used to harmonize the satellite data from the spatial and temporal viewpoint.

In contrast to the Random Forest and the pixel-based Transformer DL model, the spatial ViT model was trained using 15×15 patches, which were cropped from the original 64×64 patches, and were associated with annual composites. To highlight the properties of the Sen4Map dataset, we choose a patch size of 15×15 as a proof of concept. The selection of a 15×15 patch was a reasonable setup for some specific methods such as patch-based transformer models that later allowed us to emphasize the attention to segments of smaller 3×3 patches, which later helped with the inference and comparison with

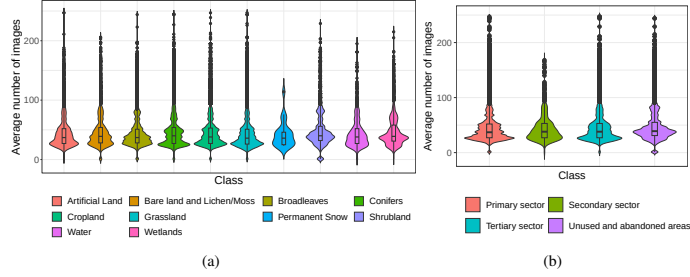


Fig. 5: Statistical distribution depicted as a violin plot of the number of Sentinel-2 images associated with each (a) land-cover class, and b) land-use class

. The first level of the hierarchy is reported.

TABLE II: Detailed mapping of LUCAS labels for crop classification

Target land-cover	LUCAS land-cover
Cereals	B11 - Common wheat B12 - Durum wheat B13 - Barley B14 - Rye B15 - Oats B16 - Maize B17 - Rice B18 - Triticale B19 - Other cereals
Root Crops	B12 - Potatoes B22 - Sugar beet B23 - Other roots crops
Non permanent industrial crops	B34 - Cotton B35 - Other fibre and oleaginous crops B36 - Tobacco B37 - Other non-permanent industrial crops
Dry pulses, vegetables and flowers	B31 - Sunflower B32 - Rape and turnip rape B33 - Soya B41 - Dry pulses B42 - Tomatoes B43 - Other fresh vegetables B44 - Floriculture and ornamental plants B45 - Strawberries
Fodder crops	B51 - Clovers B52 - Lucerne B53 - Other leguminous and mixtures for fodder B54 - Mixed cereals for fodder
Bareland	F10 - Rocks and Stone F20 - Sand F30 - Lichens and Moss F40 - Other Bare Soil
Woodland and Shrubland	B71-B77 - Fruit Trees B81 - Olive Grooves B82 - Vineyards B83 - Nurseries B84 - Permanent industrial crops C10 - Broadleaved Woodland C20 - Coniferous Woodland C30 - Mixed Woodland D10 - Shrubland with sparse tree cover D20 - Shrubland without tree cover
Grassland	B55 - Temporary Grasslands E10 - Grassland with sparse tree/shrub cover E20 - Grassland without tree/shrub cover E30 - Spontaneously re-vegetated surfaces

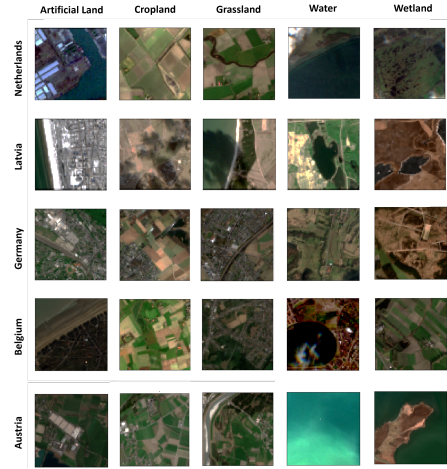


Fig. 6: True color composition of 64×64 Sentinel-2 patches belonging to the Sen4Map dataset, which shows the variability of the land-cover classes located in different countries.

other models.

Similar to the monthly composites, the annual composite was calculated by computing the median value of all the cloud-free Sentinel-2 images associated with the LUCAS location according to the cloud information provided by the “SCL” classification map. Finally, to leverage both the spatial and temporal information available in the Sen4Map dataset, the VViT model was trained on 15×15 patches (cropped from 64×64 patches) associated with the time series of 12 monthly composites. The need for harmonization of the data from a temporal viewpoint is based on the architectural limitation of the VViT model. Cross-entropy [47] is used to optimize



Fig. 7: Examples of conifer points. The 3×3 window allows the correct evaluation of the land-cover classification result with the real land-cover present in the scene.

the transformer-based models. A description of important hyperparameters used for training ML models is provided in the Appendix. In addition, to assess uncertainties in the performance of different classification algorithms, Entropy [48] is reported as shown in Table III. Higher value indicates higher uncertainty.

B. Experimental Results

Table IV and V report the experimental results obtained with the different state-of-the-art classification methods on the land-cover legend proposed in [31], and the crop classification legend proposed in [45]. Differently from the [31], the “Tree” class is separated into “Broadleaves” and “Conifers”. The obtained predictions are evaluated by comparing the land-cover information with the classification results associated with exactly the LUCAS point location and a 3×3 window centered in the LUCAS point location. In the second case, if at least one of the pixels is associated with the LUCAS land-cover label, the point is considered correctly classified. This condition allows us to mitigate possible GPS errors and handle classes for which the LUCAS observation is related to a land-cover within an observation window in an area defined by a radius of 20 m around the point. Figure 7 shows the importance of evaluating the classification results considering the 3×3 pixel window by reporting an example of two LUCAS points associated with conifer land-cover. This can be validated by comparing the overall performance of all classification methods using the center pixel, as superior results are achieved with a 3×3 window. Furthermore, all classification methods exhibited similar performance for their respective classes.

By focusing on the land-cover classification task, it is evident that the most challenging classes are Bareland and Shrubland. The classification of Shrublands poses difficulties due to the spatial disparity between the satellite’s resolution and the ground-level land-cover information. To address this issue, many approaches documented in the literature that utilized the LUCAS database to train their models often employed spatial filtering techniques to refine both the training and validation datasets, thereby improving accuracy. Despite this, transitional classes such as Bareland and Shrubland, which exhibit seasonal and spatial variability, are highly susceptible to misclassification. This phenomenon is also noticeable in classes such as Dry pulses, flowers, and Fodder crops within the crop classification task.

V. DISCUSSION

One of the most significant challenges posed by the proposed dataset Sen4Map pertains to the spatial disparity between the resolution of satellite imagery and the scale of ground-based land-cover information. Unlike existing benchmark datasets, which are generated through photo-interpretation of satellite images for land-cover mapping, Sen4Map relies on *in-situ* data. The implementation of a 3×3 window, described in Section IV-B, has been demonstrated to mitigate the disparities between satellite spatial resolution and ground-based *in-situ* observations. However, unlike existing methodologies [33], we did not apply any data cleaning procedures. This decision aligns with the dataset’s objective of advancing mapping capabilities using Sentinel-2 data, particularly on challenging land-cover samples.

In addition, the Sen4Map dataset offers the possibility to analyze classification outcomes based on their comprehensive land-use and land-cover description. For instance, Figure 8 shows a LUCAS location highlighted in red in Figure 8b, associated with the land-cover label **Lc1** (“A21” - Non-built-up area features), belonging to the class “Artificial land”, but misclassified as “Cropland”. In the LUCAS *in-situ* street-level picture shown in Figure 8a, one can observe that although the LUCAS location is correctly associated with a road (belonging to the “Artificial land” category), it is actually sealed land within an agricultural landscape. This property can also be identified by its description. The point is associated with a land-use label **Lu1** equal to “U312” (road transport) and **Lu2** equal to “U111” (agriculture excluding fallow land and kitchen areas). This land-use description is highly informative for understanding the subset of misclassified samples categorized as “Artificial land” and misclassified labeled as “Cropland”. Figure 8c displays the set of misclassified samples with land-use code ‘U111’, belonging to the “Artificial land” class. It is evident that these points are primarily misclassified as “Grassland” and “Cropland” when only considering the center pixel.

To understand the challenges associated with classifying land cover types like “Bareland”, Figure 9a presents the collective misclassification occurrences linked to the “Bareland” class, specifically concerning the land-cover label **Lc1** identified as “F40” (Other Bare Soil). A considerable portion of the misclassified samples labeled as “F40” are erroneously identified as ‘Cropland’. For instance, Figure 10 shows a location associated with LUCAS land-cover **Lc1** “F40” and land-use **Lu1** “U111” that was misclassified as “Cropland”.

To analyze further, the associated land-use from those samples that were misclassified as “Cropland” is plotted in Figure 9b. It can be seen that mainly “U111” and “U112” are responsible for these misclassifications since they clarify that this specific land-cover is associated with “Agricultural production” and “Fallow agricultural land”, respectively. Indeed, in the LUCAS classification scheme, the land-cover label “F40” links to land-use “U111” as agriculture bare land, tilled and/or prepared for seeding, and “U112” as fallow land with crop residues, becoming the main reason for such misclassifications.

TABLE III: Mean Entropy (in nat) obtained for different ML models

Land-cover	Random Forest	Transformer (pixel-based)	Vision Transformer	Video Vision Transformer
Broad land-cover [31]	1.16	0.66	0.74	0.74
Crop land-cover [45]	0.83	0.39	0.72	0.65

TABLE IV: F-score and Overall Accuracy obtained on the set of land-cover classes typically mapped [31].

Classes	Random Forest		Transformer (pixel-based)		ViT		ViViT	
	Center pixel	3x3 window	Center pixel	3x3 window	Center pixel	3x3 window	Center pixel	3x3 window
Artificial land	0.49	0.72	0.57	0.78	0.53	0.77	0.59	0.79
Bareland	0.2	0.38	0.24	0.47	0.2	0.43	0.25	0.5
Broadleaves	0.69	0.86	0.73	0.9	0.69	0.84	0.75	0.87
Conifers	0.76	0.88	0.80	0.91	0.78	0.87	0.81	0.9
Cropland	0.8	0.89	0.83	0.92	0.78	0.87	0.83	0.9
Grassland	0.69	0.84	0.73	0.89	0.68	0.84	0.73	0.87
Shrubland	0.29	0.51	0.42	0.73	0.31	0.54	0.43	0.65
Water	0.61	0.74	0.63	0.79	0.6	0.75	0.65	0.79
Wetlands	0.6	0.75	0.67	0.87	0.61	0.76	0.7	0.83
W.A. F-score	0.67	0.81	0.72	0.87	0.67	0.81	0.72	0.85
Overall Accuracy	0.68	0.83	0.73	0.88	0.68	0.82	0.73	0.85

TABLE V: F-score and Overall Accuracy obtained on the crop classification legend proposed in [45].

Classes	Random Forest		Transformer (pixel-based)		ViT		ViViT	
	Center pixel	3x3 window	Center pixel	3x3 window	Center pixel	3x3 window	Center pixel	3x3 window
Cereals	0.78	0.88	0.83	0.93	0.66	0.78	0.79	0.86
Root crops	0.56	0.69	0.69	0.88	0.24	0.44	0.64	0.75
Industrial crops	0.64	0.76	0.72	0.88	0.42	0.58	0.67	0.78
Dry pulses,flowers	0.07	0.1	0.31	0.68	0.04	0.08	0.18	0.35
Fodder crops	0.01	0.01	0.31	0.64	0.19	0.34	0.37	0.54
Bareland	0.33	0.47	0.40	0.64	0.21	0.37	0.32	0.47
Wood/Shrubland	0.88	0.96	0.90	0.97	0.85	0.93	0.88	0.95
Grassland	0.71	0.88	0.76	0.93	0.66	0.84	0.71	0.87
W.A. F-score	0.77	0.88	0.82	0.93	0.69	0.82	0.76	0.86
Overall Accuracy	0.79	0.89	0.82	0.93	0.70	0.83	0.77	0.87

In a comparable scenario involving crop classification, particularly for the challenging class “Fodder crop”, a majority of misclassifications from the “Fodder crop” were identified as “Grassland”, as depicted in Figure 11c, when considering predictions based on a center pixel. This observation is further supported by Figure 11a and 11b, where a sample with the land-cover label **Lc1** “B53” (Other legumes and mixtures for fodder) exhibits a notable semblance to temporary grasslands when scrutinized through a satellite-based viewpoint. Since these crops are highly prone to seasonal variations, the temporal component is crucial in determining accurate classification predictions. This can be validated from Table V, where it is evident that in cases of pixel-based transformers and video-vision transformers, where the attention mechanism is highly dependent on the temporal dimension, results in better predictions. Additionally, a 3×3 window further improves the classification results. Further observations within a 3×3 window indicate that the chances of correct classification increase by at least 15%.

VI. CONCLUSION

To enhance the capability of generating land-cover maps using Sentinel-2 data, this paper presents Sen4Map, a large-scale benchmark dataset. It comprises non-overlapping 64×64 patches extracted from Sentinel-2 time series images, covering over 335,125 geo-tagged locations across the European Union. These geo-tagged locations are associated with detailed land-cover and land-use information, gathered by experts as part of the LUCAS survey in 2018. The proposed dataset is highly customizable, offering flexibility across different land-cover and land-use categories. It supports both spatial and temporal classification approaches and is scale-independent. The dataset is available in country-wise, highly parallelizable HDF5 files, showcasing high modularity in terms of scale, and land-cover and land-use information. This paper demonstrates the dataset’s application through various classification approaches using state-of-the-art models for generating land-cover maps. Furthermore, it highlights the challenges posed by the schematic gap between satellite and in-situ data, offering opportunities for developing techniques and methods to

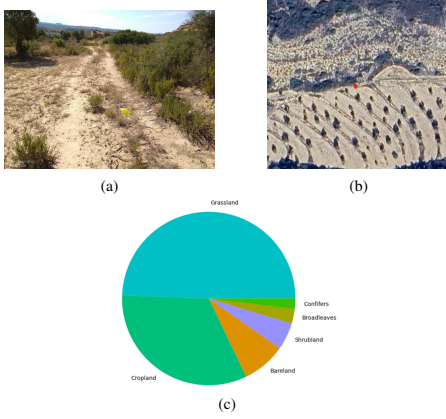


Fig. 8: Land-cover classification task: An example of misclassification from the Sen4Map dataset is depicted. (a) Shows an *in-situ* street-level image from LUCAS, (b) displays a Very High Resolution Google Map image, and (c) illustrates the number of misclassified points categorized as land-use “U111”, which belong to the class “Artificial land”.

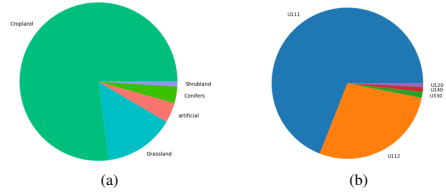


Fig. 9: Land-cover classification analysis: examining misclassified samples in the “Bareland” class; (a) Misclassifications within the “Bareland” class associated to the land-cover label Lc1 “F40”; (b) Land-use associated with misclassified samples labeled as “F40”, incorrectly identified as “Cropland”.

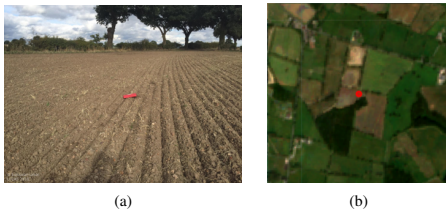


Fig. 10: Example from Sen4Map dataset from Bareland class having land-cover “F40” and land-use “U111”, misclassified as Cropland, (a) LUCAS *in-situ* street-level image, (b) yearly composite of Sentinel-2 image.

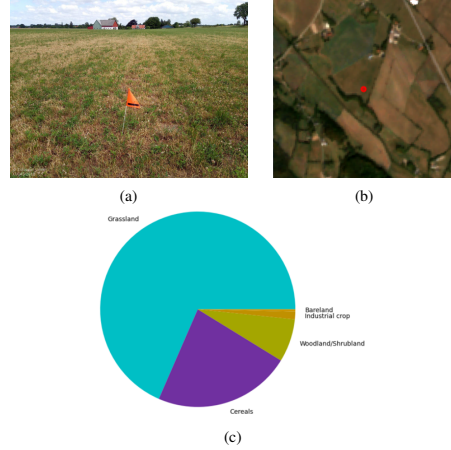


Fig. 11: Crop classification task: detailed analysis of misclassified samples from class “Fodder crops”; (a) LUCAS *in-situ* street-level image, (b) yearly composite of Sentinel-2 image, (c) All misclassifications from class “Fodder crops”.

advance current land-cover mapping efforts with Sentinel-2.

VII. ACKNOWLEDGMENT

Part of this work was performed in the ADMIRE project, which received funding from the European Union’s Horizon 2020 JTI-EuroHPC research and innovation programme under the grant agreement no. 956748. This work received complementary funding from the German Federal Ministry of Education and Research (no. 16HPC008). This work is co-financed by the EUROCC2 project funded by the European High-Performance Computing Joint Undertaking (JU) and EU/EEA states under grant agreement No 101101903. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC).

REFERENCES

- [1] C. Persello, J. D. Wegner, R. Hänsch, D. Tuia, P. Ghamisi, M. Koeva, and G. Camps-Valls, “Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 172–200, 2022.
- [2] C. Paris, L. Bruzzone, and D. Fernández-Prieto, “A novel approach to the unsupervised update of land-cover maps by classification of time series of multispectral images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4259–4277, 2019.
- [3] M. Russwurm, N. Courty, R. Emonet, S. Lefèvre, D. Tuia, and R. Tave-nard, “End-to-end learned early classification of time series for in-season crop type mapping,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 445–456, 2023.

- [4] Copernicus, "Sentinel-2 mission overview," <https://sentiwiki.copernicus.eu/web/s2-mission>, 2024, accessed: 2024-06-12.
- [5] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefèvre, and M. Körner, "Breizhrops: a tome series dataset for crop type mapping," in *XXIV ISPRS Congress, Commission II*, ser. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, N. Papadimitris, C. Mallet, F. Lafarge, F. Remondino, I. Toschi, and T. Fuse, Eds. ISPRS, Aug. 2020, pp. 1545–1551, 2020 24th ISPRS Congress - Technical Commission II ; Conference date: 31-08-2020 Through 02-09-2020.
- [6] V. Akbari, S. Solberg, and S. Puliti, "Multitemporal sentinel-1 and sentinel-2 images for characterization and discrimination of young forest stands under regeneration in norway," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5049–5063, 2021.
- [7] J. Parajuli, R. Fernandez-Beltran, J. Kang, and F. Pla, "Attentional dense convolutional neural network for water body extraction from sentinel-2 images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6804–6816, 2022.
- [8] C. F. Brown, S. P. Brumby, B. Guzder-Williams, T. Birch, S. B. Hyde, J. Mazzariello, W. Czerwinski, V. J. Pasquarella, R. Haertel, S. Ilyushchenko *et al.*, "Dynamic world, near real-time global 10 m land use land cover mapping," *Scientific Data*, vol. 9, no. 1, pp. 1–17, 2022.
- [9] R. C. Sharma, "Countrywide mapping of plant ecological communities with 101 legends including land cover types for the first time at 10 m resolution through convolutional learning of satellite images," *Applied Sciences*, vol. 12, no. 14, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/14/7125>
- [10] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7, pp. 153–160, 2019. [Online]. Available: <https://isprs-annals.copernicus.org/articles/IV-2-W7/153/2019/>
- [11] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "Sen12ms-cr-ts: A remote-sensing data set for multimodal multitemporal cloud removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
- [12] C. Aybar, L. Yshuaylas, J. Loja, K. Gonzales, F. Herrera, L. Bautista, R. Yali, A. Flores, L. Diaz, N. Cuenca, W. Espinoza, F. Prudencio, V. Lactayo, D. Montero, M. Sudmanns, D. Tiede, G. Mateo-García, and L. Gómez-Chova, "Cloudsen12, a global dataset for semantic understanding of cloud and cloud shadow in sentinel-2," *Scientific Data*, vol. 9, no. 1, p. 782, Dec 2022.
- [13] J. Cornébaris, I. Orsolic, and F. Kalaitzis, "Open high-resolution satellite imagery: The worldstrat dataset - with application to super-resolution," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [14] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery," *Advances in Neural Information Processing Systems*, vol. 35, pp. 197–211, 2022.
- [15] O. Mañas, A. Lacoste, X. Giró i Nieto, D. Vázquez, and P. Rodríguez López, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9394–9403.
- [16] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, "Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 3, pp. 98–106, 2023.
- [17] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "Satlaspretrain: A large-scale dataset for remote sensing image understanding," in *ICCV 2023*, 2022.
- [18] A. Francis and M. Czerkawski, "Major tom: Expandable datasets for earth observation," 2024.
- [19] J. Munoz-Mari, E. Izquierdo-Verdiguier, M. Campos-Taberner, A. Perez-Suay, L. Gomez-Chova, G. Mateo-Garcia, A. B. Ruescas, V. Laparra, J. A. Padron, F. Amoros-Lopez, and G. Camps-Valls, "Hyperlabelme : A web platform for benchmarking remote-sensing image classifiers," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 79–85, 2017.
- [20] J. Jakubik, S. Roy, C. E. Phillips, and *et al.*, "Foundation Models for Generalist Geospatial Artificial Intelligence," *Preprint Available on arxiv:2310.18660*, Oct. 2023.
- [21] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigeearthnet: A large-scale benchmark archive for remote sensing image understanding," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5901–5904.
- [22] P. Helber, B. Bischke, A. Dengel, and D. 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*, vol. 12, no. 7, pp. 2217–2226, 2019.
- [23] R. Van De Kerchove, D. Zana, W. Keersmaecker, N. Souverijns, J. Wevers, C. Brockmann, A. Grosu, A. Paccini, O. Cartus, M. Santoro *et al.*, "Esa worldcover: Global land cover mapping at 10 m resolution for 2020 based on sentinel-1 and 2 data," in *AGU Fall Meeting Abstracts*, vol. 2021, 2021, pp. GC451–0915.
- [24] K. Karra, C. Kontgis, Z. Statman-Weil, J. C. Mazzariello, M. Mathis, and S. P. Brumby, "Global land use/land cover with sentinel 2 and deep learning," in *2021 IEEE international geoscience and remote sensing symposium IGARSS*. IEEE, 2021, pp. 4704–4707.
- [25] M. Probeck, I. Ruiz, G. Rammingier, C. Fourie, P. Maier, M. Ickerott, C. Storch, A. Homolka, S. J. Muller, H. Tiwari *et al.*, "C1c+ backbone: Set the scene in copernicus for the coming decade," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 2076–2079.
- [26] C. Pelletier, G. I. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sensing*, vol. 11, no. 5, p. 523, 2019.
- [27] C. Paris and L. Bruzzone, "Automatic extraction of weak labeled samples from existing thematic products for training convolutional neural networks," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5722–5725.
- [28] Copernicus, "Sentinel-2a, processed by esa, msi level-2a boa reflectance product. collection 1, european space agency," <https://doi.org/10.5270/S2-znk9xsj>, 2021, accessed: 2024-06-12.
- [29] R. d'Andrimont, M. Yordanov, L. Martinez-Sanchez, B. Eiselt, A. Palmieri, P. Dominici, J. Gallego, H. I. Reuter, C. Joebges, G. Lemoine *et al.*, "Harmonised lucas in-situ land cover and use database for field surveys from 2006 to 2018 in the european union," *Scientific data*, vol. 7, no. 1, p. 352, 2020.
- [30] Y. Gao, L. Liu, X. Zhang, X. Chen, J. Mi, and S. Xie, "Consistency analysis and accuracy assessment of three global 30-m land-cover products over the european union using the lucas dataset," *Remote Sensing*, vol. 12, no. 21, p. 3479, 2020.
- [31] Z. S. Venter, D. N. Barton, T. Chakraborty, T. Simensen, and G. Singh, "Global 10 m land use land cover datasets: A comparison of dynamic world, world cover and esri land cover," *Remote Sensing*, vol. 14, no. 16, p. 4101, 2022.
- [32] M. Weigand, J. Staab, M. Wurm, and H. Taubenböck, "Spatial and semantic effects of lucas samples on fully automated land use/land cover classification in high-resolution sentinel-2 data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 88, p. 102065, 2020.
- [33] Z. S. Venter and M. A. Sydenham, "Continental-scale land cover mapping at 10 m resolution over europe (elc10)," *Remote Sensing*, vol. 13, no. 12, p. 2301, 2021.
- [34] S. M. Mirmazloumi, M. Kakooei, F. Mohseni, A. Ghorbanian, M. Amani, M. Crosetto, and O. Monserrat, "Eluc-10, a 10 m european land use and land cover map using sentinel and landsat data in google earth engine," *Remote Sensing*, vol. 14, no. 13, p. 3041, 2022.
- [35] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18–27, 2017, big Remotely Sensed Data: tools, applications and experiences. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425717302900>
- [36] C. Ji, M. Du, Y. Wei, Y. Hu, S. Liu, L. Pan, and X. Zheng, "Time series classification with random temporal features," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 9, p. 101783, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1319157823003373>
- [37] Z. Liu, X. Li, and Y. Cui, "Time series model interpretation via temporal feature sampling," in *Web and Big Data*, X. Song, R. Feng, Y. Chen, J. Li, and G. Min, Eds. Singapore: Springer Nature Singapore, 2024, pp. 346–360.
- [38] M. Herold, R. Hubald, and A. Di Gregorio, "Translating and evaluating land cover legends using the UN land cover classification system LCCS," *GFOC-GOLD Report*, vol. 43, pp. 1482–1483, 2009.

- [39] The HDF Group, "Hierarchical Data Format, version 5," 1997-NNNN, accessed on: 2024-06-12. [Online]. Available: <http://www.hdfgroup.org/HDF5/>
- [40] C. Pelletier, S. Valero, J. Inglada, N. Champion, C. Marais Sicre, and G. Dedieu, "Effect of training class label noise on classification performances for land cover mapping with satellite image time series," *Remote Sensing*, vol. 9, no. 2, p. 173, 2017.
- [41] R. C. Sharma, "An ultra-resolution features extraction suite for community-level vegetation differentiation and mapping at a sub-meter resolution," *Remote Sensing*, vol. 14, no. 13, 2022. [Online]. Available: <https://www.mdpi.com/2072-4292/14/13/3145>
- [42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000-6010.
- [43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [44] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," 2021.
- [45] R. d'Andrimont, A. Verhegghen, G. Lemoine, P. Kempeneers, M. Meroni, and M. van der Velde, "From parcel to continental scale – a first european crop type map based on sentinel-1 and lucas copernicus in-situ observations," *Remote Sensing of Environment*, vol. 266, p. 112708, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425721004284>
- [46] G. Weikmann, C. Paris, and L. Bruzzone, "Timesen2crop: A million labeled samples dataset of sentinel 2 image time series for crop-type classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4699-4708, 2021.
- [47] R. Y. Rubinstein and D. P. Kroese, *The Cross Entropy Method: A Unified Approach To Combinatorial Optimization, Monte-carlo Simulation (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2004.
- [48] A. Namdari and Z. S. Li, "A review of entropy measures for uncertainty quantification of stochastic processes," *Advances in Mechanical Engineering*, vol. 11, no. 6, p. 1687814019857350, 2019. [Online]. Available: <https://doi.org/10.1177/1687814019857350>



Surbhi Sharma (Student Member, IEEE) received the B.Tech degree in electronics and communication engineering from Amity University, India, in 2015, and the M.Sc degree in Geo-information Science and Earth Observation with a specialization in Geoinformatics jointly from the Indian Institute of Remote Sensing, Indian Space and Research Organization (ISRO), India, and the Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, the Netherlands, in 2018. She is a member of the "AI and ML for Remote Sensing"

Simulation and Data Lab at the Jülich Supercomputing Centre, Germany. She is currently pursuing the Ph.D. degree in computational engineering at the University of Iceland. Her research interest lies in scalable machine learning and deep learning methods for remote sensing applications, with a particular focus on advanced deep transfer learning methods using modern High-Performance Computing (HPC) systems.



Rocco Sedona (Member, IEEE) received his B.Sc. and M.Sc. degrees in Information Engineering from the University of Trento in 2016 and 2019, respectively, and a Ph.D. degree in Computational Engineering from the University of Iceland in 2023. He is a member of the "AI and ML for Remote Sensing" Simulation and Data Lab at the JSC in Germany. His research interests primarily lie in the field of Deep Learning and its application to Remote Sensing data. He has extensively utilized optical satellite data acquired by Landsat (NASA) and Sentinel (ESA) missions towards near real-time land-cover classification. Additionally, he specializes in Distributed Deep Learning on High-Performance Computing systems, an area of study that he has been actively engaged in since 2019.



Morris Riedel (Member, IEEE) received his PhD from the Karlsruhe Institute of Technology (KIT) and worked in data-intensive parallel and distributed systems since 2004. He is currently a Full Professor of High-Performance Computing with an emphasis on Parallel and Scalable Machine Learning at the School of Natural Sciences and Engineering of the University of Iceland. Since 2004, Prof. Dr. -Ing. Morris Riedel held various positions at the Jülich Supercomputing Centre of Forschungszentrum Jülich in Germany. In addition, he is the Head

of the joint High Productivity Data Processing research group between the Jülich Supercomputing Centre and the University of Iceland. Since 2020, he is also the EuroHPC Joint Undertaking governing board member for Iceland. His research interests include high-performance computing, remote sensing applications, medicine and health applications, pattern recognition, image processing, and data sciences, and he has authored extensively in those fields. Prof. Dr. -Ing. Morris Riedel's online YouTube and university lectures include High-Performance Computing – Advanced Scientific Computing, Cloud Computing and Big Data – Parallel and Scalable Machine and Deep Learning, as well as Statistical Data Mining. In addition, he has performed numerous hands-on training events in parallel and scalable machine and deep learning techniques on cutting-edge HPC systems.



Gabriele Cavallaro (Senior Member, IEEE) received his B.Sc. and M.Sc. degrees in Telecommunications Engineering from the University of Trento, Italy, in 2011 and 2013, respectively, and a Ph.D. degree in Electrical and Computer Engineering from the University of Iceland, Iceland, in 2016. From 2016 to 2021 he has been the deputy head of the "High Productivity Data Processing" (HPDP) research group at the Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich, Germany. Since 2022, he is the Head of the "AI and ML for Remote Sensing" Simulation and Data Lab at the JSC and an Adjunct Associate Professor with the School of Natural Sciences and Engineering, University of Iceland, Iceland. From 2020 to 2023, he held the position of Chair for the High-Performance and Disruptive Computing in Remote Sensing (HDCRS) Working Group under the IEEE GRSS Earth Science Informatics Technical Committee (ESI TC). In 2023, he took on the role of Co-chair for the ESI TC. Concurrently, he serves as Visiting Professor at the 4-lab within the European Space Agency (ESA), where he contributes to the Quantum Computing for Earth Observation (QC4EO) initiative. Additionally, he has been serving as an Associate Editor for the IEEE Transactions on Image Processing (TIP) since October 2022. He was the recipient of the IEEE GRSS Third Prize in the Student Paper Competition of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2015 (Milan - Italy). His research interests cover remote sensing data processing with parallel machine learning algorithms that scale on distributed computing systems and innovative computing technologies.

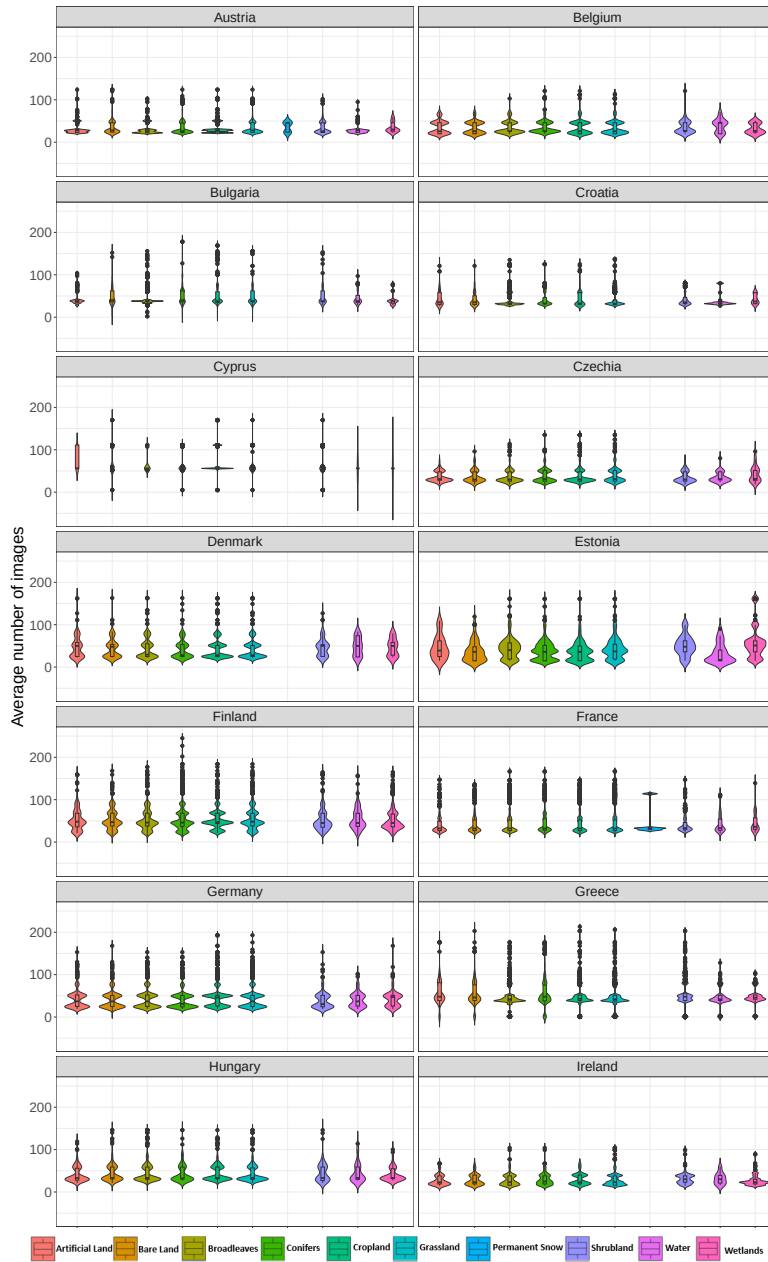


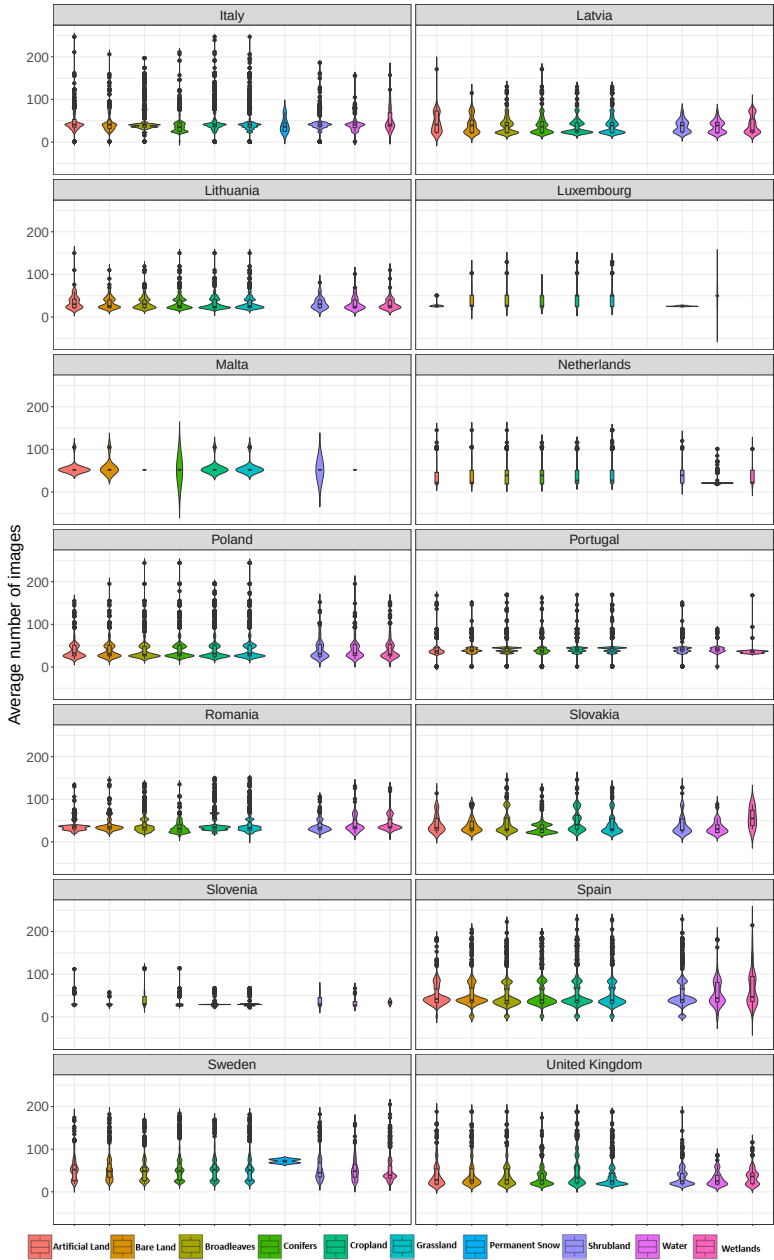
Claudia Paris (Senior Member, IEEE) received the B.S. degree and M.S. (summa cum laude) degree in Telecommunication Engineering and the Ph.D. in Information and Communication Technology from the University of Trento, Italy, in 2010, 2012, 2016, respectively. She accomplished the Honors Master Program in Research within the Master Degree in Telecommunication Engineering in 2012. Since 2021 she is an Assistant Professor in the Faculty of Geoinformation and Earth Observation Sciences (ITC) at the University of Twente. Claudia Paris'

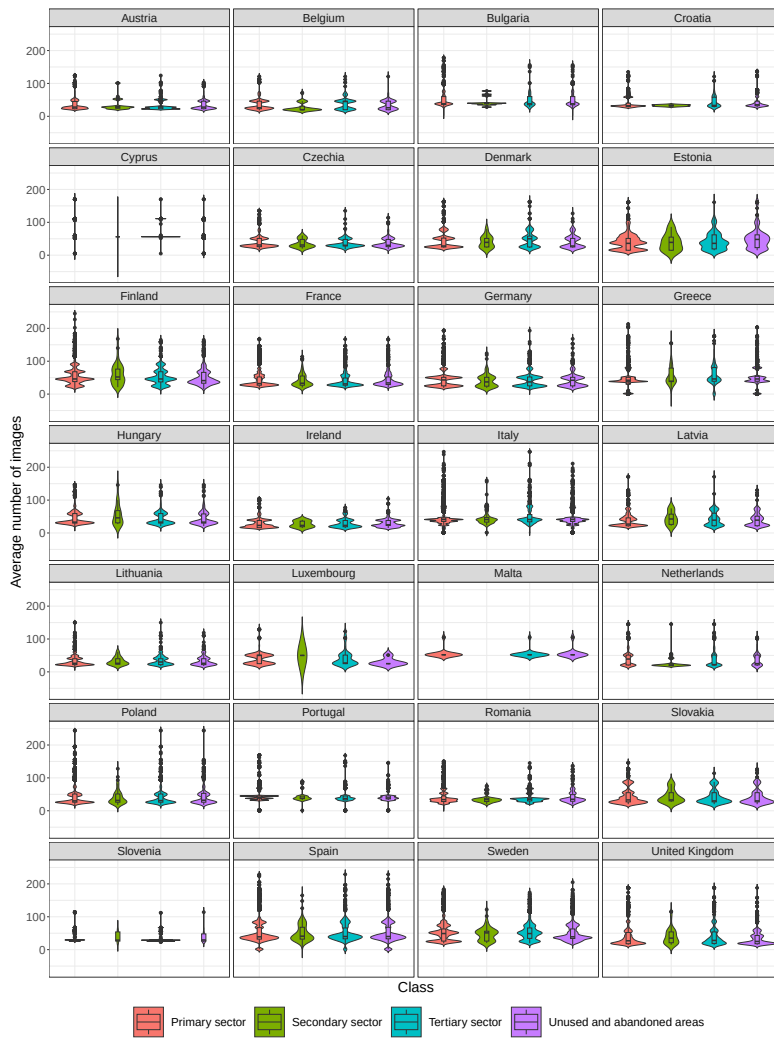
research encompasses image and signal processing, machine learning and deep learning, specifically applied to remote sensing image analysis. She focuses on designing innovative and automated workflows for the analysis and classification of large-scale Earth observation data for various applications (e.g., forest/agricultural mapping and monitoring). To this end, she works with multisource remote sensing data (e.g., LiDAR, hyperspectral, multispectral, and high-resolution optical images), multitemporal image analysis, domain adaptation methods, development of ad hoc deep learning solutions, and biophysical parameter estimation. She conducts research on these topics within the frameworks of national and international projects. Since 2023 she has been serving as an Associate Editor for the IEEE Geoscience of Remote Sensing Letters (GRSL). Dr. Paris won the very prestigious Symposium Prize Paper Award (SPPA) at the 2016 International Symposium on Geoscience and Remote Sensing (Beijing, China, 2016) and at the 2017 International Symposium on Geoscience and Remote Sensing (Fort Worth, Texas, USA, 2017). She also won the IEEE Geoscience and Remote Sensing Society 2022 Letters Prize Paper Award (exceptional paper in terms of content and impact on the GRS-Society). Since 2020, she has been a member of the program and scientific committee of several international conferences and workshops.

APPENDIX

Country-wise statistical distribution of the number of Sentinel-2 images associated with each land-cover and land-use class considering the first level of the hierarchy in Sen4Map benchmark dataset, depicted as a violin plot.







The list of some of the hyper-parameters used for training ML models is shown in tabular form.

Hyperparameters	Random Forest
Number of trees	400
criterion	gini
Max depth	Until pure
Minimum number of splits	2
Maximum number of features	sqrt

Hyperparameters	Transformer (pixel-based)
Embedding Dimension	128
Hidden Dimension	256
Number of heads	4
Number of layers	3
Number of patches	12
Learning rate	3e-4
Dropout	0.2

Hyperparameters	Vision Transformer
Embedding Dimension	256
Hidden Dimension	512
Number of heads	8
Number of layers	6
Image Patch size	3
Number of patches	25
Learning rate	3e-4
Dropout	0.2

Hyperparameters	Video Vision Transformer
Image size	15
Frames	12
Image Patch size	3
Frame Patch size	3
Embedding Dimension	256
Hidden Dimension	512
Number of heads	8
Spatial Depth	3
Temporal Depth	3
Learning rate	3e-5

Paper IV

Enhancing Land Cover Mapping: A Novel Automatic Approach To Improve Mixed Spectral Pixel Classification

S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris

IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2024

©2024 IEEE. Reprinted, with permission, from [100]

Surbhi Sharma wrote the code and the paper, ran all experiments on the HPC machine and performed the analysis.

ENHANCING LAND COVER MAPPING: A NOVEL AUTOMATIC APPROACH TO IMPROVE MIXED SPECTRAL PIXEL CLASSIFICATION

Surbhi Sharma^{1,2}, Rocco Sedona², Morris Riedel^{1,2}, Gabriele Cavallaro^{1,2}, Claudia Paris³

¹School of Engineering and Natural Sciences, University of Iceland, 107 Reykjavik, Iceland

²Jülich Supercomputing Centre, Forschungszentrum Jülich, 52428 Jülich, Germany

³University of Twente, 7514 AE Enschede, The Netherlands

ABSTRACT

The increasing availability of high-resolution, open-access satellite data facilitates the production of global Land Cover (LC) maps, an essential source of information for managing and monitoring natural and human-induced processes. However, the accuracy of the obtained LC maps can be affected by the discrepancy between the spatial resolution of the satellite images and the extent of the LC present in the scene. Indeed, several pixels may be misclassified because of their mixed spectral signatures, i.e., more than two LC classes are present in the pixel. To solve this problem, this paper proposes an approach that explores the possibility of using simple but effective unmixing approaches to enhance the classification accuracy of the mixed spectral pixels. The results showed that several pixels, including buildings and grassland LC, are typically classified as cropland. By unmixing their spectral content, it is possible to extract the most prevalent class within the area of each pixel to update the classification map, thus sharply increasing the map accuracy. These promising preliminary results indicate the potential for broader applicability and efficiency in global LC mapping.

Index Terms— Land Cover (LC) mapping, spectral unmixing, Sentinel-2, Land use/cover area frame survey (LU-CAS).

1. INTRODUCTION

With the Copernicus Programme, the availability of free and open high-resolution satellite data has completely transformed the methods used to monitor the Earth's surface. The spatial, temporal, and unique spectral properties of the Copernicus Sentinel-2 satellite data have enabled the Earth observation (EO) community to produce global thematic products

The research leading to these results was conducted within the ADMIRE project, which received funding from the European Union's Horizon 2020 JTI-EuroHPC research and innovation programme under grant agreement No. 956748. This work also received complementary funding from the German Federal Ministry of Education and Research under grant No. 16HPC008. This work is co-financed by the EUROCC2 project funded by the European High-Performance Computing Joint Undertaking (JU) and EU/EEA states under grant agreement No. 101101903.

at a 10m spatial resolution [1]. The production of accurate LC maps enables continuous monitoring of the Earth's surface, which is crucial for managing the planet's resources and observing environmental processes such as desertification, urbanization, and deforestation [2].

Multiple satellite pixels can be associated with mixed spectral signatures due to the differing spatial resolutions of satellite images compared to the extent of the LC classes present in the scene. This can lead to inaccurate classification results, thus affecting the reliability of the final thematic product [3]. In [4], the authors demonstrated that when evaluating the accuracy of existing thematic products using validation databases derived by visually interpreting satellite data, they achieved higher accuracy than when using *in-situ* data. This discrepancy typically occurs when the size of the LC class present in the scene is much smaller than the geometrical resolution of the satellite data, e.g., small rivers or tiny roads [3]. While much effort has been devoted to generating LC products at the global level [4], little has been done to mitigate the influence of mixed pixels on the LC classification result [5].

To solve this problem, the main objective of this paper is to understand and deal with the complexities of specific LC classes at a pixel level to increase the accuracy of obtained LC maps, using spectral unmixing with underlying semantic differences. First, a pre-processing step performs data harmonization of the considered time series of optical satellite data. Then, an ensemble of Random Forest (RF) classifiers is trained to produce a set of LC maps that are compared to detect discrepancies among them. This condition allows to automatically detect the samples with the highest probability of belonging to mixed spectral pixels, i.e., those with inconsistent classification results. Finally, for the inconsistent pixels, a state-of-the-art unmixing approach is applied to disentangle the different spectral components and determine the best LC label to be assigned to those pixels.

2. PROPOSED METHOD

Fig. 1 shows the block diagram of the proposed method, which is based on three main steps: (i) satellite data pre-

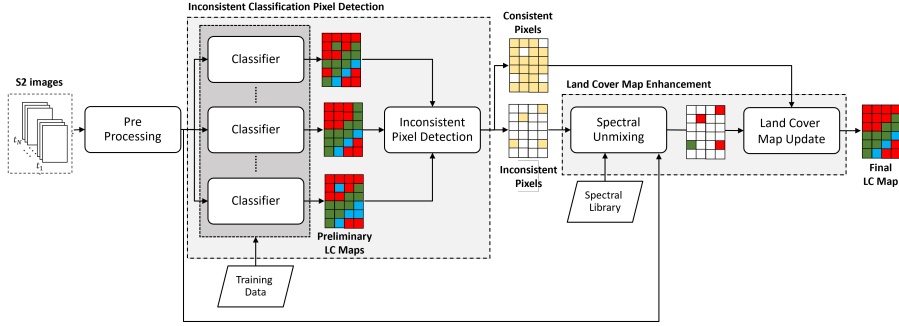


Fig. 1. Workflow of the proposed data-driven approach to improve LC mapping of the mixed spectral pixels.

processing, (ii) identification of inconsistent LC classification results, and (iii) LC map enhancement for mixed spectral pixels.

2.1. Pre-processing

In the first step of the proposed method, the time series of optical satellite images used to generate the LC map are harmonized. In the considered experiments, the Sentinel-2 satellite data, widely used for LC mapping due to their high spatial resolution (10m), high revisit time (up to 5 days), and spectral properties, are utilized [4, 6]. In particular, a time series of Sentinel-2 images acquired throughout the year is considered to generate an annual LC map. To remove cloud coverage and harmonize the data from the spatial and temporal viewpoint, the time series of images are converted into a time series of 12 monthly composites. The median approximation of the reflectance values collected over each month is computed for each pixel. Let $\mathbf{TS} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{12})$ be the time series of 12 monthly composites, with $\mathbf{x}_k \in \mathbb{R}^{1 \times 120}$ representing the k th pixel's temporal spectral vectors associated with the 120 features, i.e., 10 reflectance values $\times$ 12 months. These monthly composites are then used to generate LC maps.

2.2. Inconsistent Classification Pixel Detection

The second step of the proposed method aims to automatically identify the mixed spectral pixels with the highest probability of being misclassified. To this end, an ensemble of statistically independent Random Forest classifiers is trained using randomly chosen training samples via a bootstrap technique [7, 8]. Let Ω_c be the set of LC classes present in the scene. The initial pool of labeled data is sampled without replacement to generate of N training set, i.e., $\{T_1, T_2, \dots, T_N\}$ used to train N independent classifiers $\{C_1, C_2, \dots, C_N\}$, where $T_n = \{(\mathbf{x}_b, y_b)\}_b$ is the n th training set having $\mathbf{x}_b \in \mathbb{R}^{1 \times 120}$ and $y_b \in \Omega_c$. In this setup, it

is reasonable to assume that the obtained classification errors are uncorrelated and associated with the mixed spectral pixels, which typically lead to uncertain classification results. By comparing the ensemble of obtained LC maps, the pool of inconsistent pixels having the highest probability of being mixed from the spectral viewpoint is identified.

2.3. Land Cover Map Enhancement

Finally, the method aims to identify inconsistencies and misclassifications that provide patterns and information about mixed pixels. Here, for simplicity, we used Linear Spectral Unmixing to analyze the contribution of different endmembers, or materials within a mixed pixel [9]. However, any other unmixing techniques can be considered. The linear mixing model is often represented as follows:

$$R = S \cdot A + E, \quad (1)$$

where: R represents the observed spectrum of a mixed pixel, S is a matrix containing the spectra of the endmembers, A is a vector containing the abundance fractions of each endmember in the pixel, and E is an error term accounting for factors such as noise or atmospheric effects. The objective of spectral unmixing is to determine the abundance fractions (A) of each endmember within a mixed pixel. An abundance map illustrates the spatial distribution of these fractions throughout an image. These abundance maps are generated on the yearly composite of Sentinel-2 time-series images obtained by using the median approximation of the reflectance values collected over the whole year for each pixel. These maps are further used to improve classification accuracy by updating LC maps with the most prevalent class.

3. DATASET DESCRIPTION

To train and validate the obtained LC map, we used *in-situ* data from the Land Use and Coverage Area frame Sur-



Fig. 2. Example of LUCAS point labeled as ‘Artificial Land’ (a) High-resolution Google Earth Satellite View, (b) LUCAS street-level image, (c) zoomed-in view of the residential area containing LUCAS point, (d) Sentinel-2 image located in the LUCAS point.

vey (LUCAS) database [10], a survey coordinated by the Statistical Office of the European Commission (Eurostat), which collects harmonized data on land cover/land use, agro-environmental variables, and soil through field observation of geographically referenced points. The LC labels in the LUCAS database are defined according to the Land Cover Classification System (LCCS) [11].

The experimental setup preprocesses and downloads time series of Sentinel-2 Level-2A data for 2018. Each series consists of a 3×3 pixel patch with a 20 m spatial resolution (discarding bands with a 60 m spatial resolution) and is associated with the LC label available in LUCAS. To capture a detailed and accurate representation of ground reference data, a 3×3 patch of Sentinel-2 is used. As an example in Fig. 2, it can be seen that even though the LUCAS point clearly describes a residential area to be classified as artificial land (see Fig. 2(a), Fig. 2(b), and Fig. 2(c)), the Sentinel-2 pixel associated with the point location is mixed (see Fig. 2(d)), covering a grassland as well, potentially leading to a mixed spectral signature and misclassification.

4. EXPERIMENTS AND RESULTS

To assess the effectiveness of the proposed approach, results have been carried out at the country and continental level. In particular, we considered 3659 LUCAS points from Belgium and 337854 LUCAS points covering the European Union (EU). A pool of inconsistent pixels was identified by comparing the ensemble of LC maps obtained using the RF classifier, having the highest probability of being mixed as described in Section 2.2. The class probabilities per pixel were also calculated for 9 LC classes, namely ‘Artificial land’, ‘Bareland’, ‘Broadleaves’, ‘Conifers’, ‘Cropland’, ‘Grassland’, ‘Shrubland’, ‘Water’, and ‘Wetland’. A total of 478 pixels from Belgium and 51,806 pixels from EU with a higher class probability of being in more than one class were identified as mixed and chosen for further spectral analysis. From the obtained inconsistent samples, 220 samples from Belgium and 26,316 samples from EU are found to be misclassified.

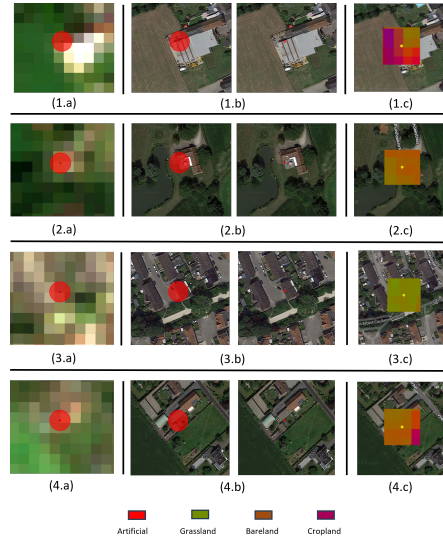


Fig. 3. Abundance Map obtained after Spectral Unmixing for four different LUCAS points, marked as red and yellow, labeled as ‘Artificial Land’. Figures (1.a)-(4.a) and (1.b)-(4.b) show the LUCAS points in Sentinel-2 and High-resolution Google Earth Satellite View. Figures (1.c) - (4.c) represent the Abundance Map obtained from the Sentinel-2 yearly composite of respective LUCAS points.

For experimentation, our study is limited to 478 pixels from Belgium and a randomly selected subset of 1000 pixels out of 51,806 inconsistent pixels from the EU to reduce the overall computational time.

Fig. 3 shows example cases corresponding to four different LUCAS points labeled as ‘Artificial land’. Since the obtained inconsistent pixels belong to some of the major LC classes, the abundance maps were created by performing spectral unmixing, focusing on finding 9 major classes, namely ‘Artificial land’, ‘Bareland’, ‘Broadleaves’, ‘Conifers’, ‘Cropland’, ‘Grassland’, ‘Shrubland’, ‘Water’, and ‘Wetland’ as shown in Fig. 3. These maps are created on a 3×3 patch for a more detailed and optimal view of what is on the ground using the annual composite, as shown in Fig. 3(1.c) - (4.c), respectively. The endmembers, or the ‘pure spectra’, were determined in a supervised way for the purpose of cross-referencing. Finally, the labels were updated based on the highest abundance fraction of the class obtained.

The overall classification accuracy and weighted average

Table 1. The weighted average F-score and Overall accuracy using the baseline and proposed method.

Classes	Belgium		Samples from EU	
	Baseline	Proposed Method	Baseline	Proposed Method
Weighted Average F1-score	0.51	0.63	0.43	0.54
Overall Accuracy	0.54	0.62	0.45	0.54

F1-score for the baseline method and our proposed method, are reported in Tab. 1. It can be seen that our proposed methodology performs well overall. The results obtained on a single country, Belgium, and at the European level suggest that assigning the pixels to the class corresponding to the highest abundance fraction can help improve the accuracies of LC classification.

5. CONCLUSION

This paper introduces a data-driven workflow for improving LC maps via spectral unmixing, specifically targeting mixed pixel issues to bolster classification accuracy. This study makes use of an automated workflow and method to generate abundance maps and update LC labels using optimized resources available on High-performance computing (HPC) systems and data parallel processing strategies. The results suggest potential for streamlining both the LC classification and enhancement processes. The study also explores the misclassification arising from semantic differences between satellite and *in-situ* data, which are attributed to their different spatial resolutions and viewpoints. The obtained results provide insights into the mixed spectral classes that are typically present at a scene and are misclassified. With the help of spectral unmixing, the proposed approach successfully updates the classification maps by extracting the most prevalent class within the area of each pixel, increasing the accuracy of the LC map.

6. FUTURE DEVELOPMENTS

The obtained abundance fractions provide valuable insights for the most prevalent class present on the scene. In cases where the highest abundance fractions do not match or justify the *in-situ* data, especially where the second abundance fraction is the actual class label, a detailed analysis of each time series component from patches and rule-based filtering can be opted for updating LC labels. In addition, advanced Deep Learning (DL) models, known for their extensive data requirements, such as Transformers, could significantly enhance the accuracy of baseline methods like RF [12]. Other unsupervised techniques for evaluating endmembers, such as Vertex Component analysis (VCA), also hold the potential for significantly expediting the workflow, which, on the other hand, relies mostly on the spectral disposition of the dataset [13, 14].

7. REFERENCES

- [1] R. Van De Kerchove, D. Zanaga, W. Keersmaecker, N. Souverijns, J. Wevers, C. Brockmann, A. Grosu, A. Paccini, O. Cartus, M. Santoro, et al., "Esa worldcover: Global land cover mapping at 10 m resolution for 2020 based on sentinel-1 and 2 data.," in *AGU Fall Meeting Abstracts*, 2021, vol. 2021, pp. GC451-0915.
- [2] A. Stoian, V. Poulain, J. Inglada, V. Poughon, and D. Derksen, "Land cover maps production with high resolution satellite image time series and convolutional neural networks: Adaptations and limits for operational systems," *Remote Sensing*, vol. 11, no. 17, pp. 1986, 2019.
- [3] C. Paris, L. Martinez-Sanchez, M. van der Velde, S. Sharma, R. Sedona, and G. Cavallaro, "Accuracy assessment of land-use-land-cover maps: the semantic gap between in situ and satellite data," in *Image and Signal Processing for Remote Sensing XXIX*. 2023, vol. 12733, SPIE.
- [4] Zander S Venter, David N Barton, Tirthankar Chakraborty, Trond Simensen, and Geethen Singh, "Global 10 m land use land cover datasets: A comparison of dynamic world, world cover and esri land cover," *Remote Sensing*, vol. 14, no. 16, pp. 4101, 2022.
- [5] Chao Yang, Guofeng Wu, Kai Ding, Tiezhu Shi, Qingquan Li, and Jinliang Wang, "Improving land use/land cover classification by integrating pixel unmixing and decision tree methods," *Remote Sensing*, vol. 9, no. 12, pp. 1222, 2017.
- [6] European Space Agency (ESA), "Copernicus Sentinel-2 MSI Level-2A BOA Reflectance Product," 2021, Collection 1.
- [7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
- [8] Gilles Louppe, "Understanding random forests: From theory to practice," 2015.
- [9] Jiaojiao Wei and Xiaofei Wang, "An overview on linear unmixing of hyperspectral data," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–12, 08 2020.
- [10] R. d'Andrimont, M. Yordanov, L. Martinez-Sanchez, B. Eisel, A. Palmieri, P. Dominici, F. J. Gallego Pinilla, H. Reuter, C. Joebges, G. Lemoine, and M. Velde, "Harmonised lucas in-situ land cover and use database for field surveys from 2006 to 2018 in the european union," *Scientific Data*, vol. 7, 10 2020.
- [11] M. Herold, R. Hubald, and A. Di Gregorio, "Translating and evaluating land cover legends using the un land cover classification system (lccs)," *GOGC-GOLD Report*, vol. 43, 2009.
- [12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.
- [13] J.M.P. Nascimento and J.M.B. Dias, "Vertex component analysis: a fast algorithm to unmix hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, 2005.
- [14] José M. Rodríguez Alves, José M. P. Nascimento, José M. Bioucas-Dias, Vítor Silva, and Antonio Plaza, "Parallel implementation of vertex component analysis for hyperspectral endmember extraction," in *2012 IEEE International Geoscience and Remote Sensing Symposium*, 2012, pp. 4078–4081.

Paper V

Beyond Pixels: Identifying Built-Up Features at Sub-Pixel Level for Enhanced Satellite-Based Land Cover Mapping

S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris

IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), 2026 (submitted)

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License
(<https://creativecommons.org/licenses/by/4.0/>).

Surbhi Sharma wrote the code and the main part of the paper, ran all experiments on the HPC machine and performed the analysis.

Beyond Pixels: Identifying Built-Up Features at Sub-Pixel Level for Enhanced Satellite-Based Land Cover Mapping

Surbhi Sharma, *Student Member, IEEE*, Rocco Sedona, *Member, IEEE*, Morris Riedel, *Member, IEEE*,
Gabriele Cavallaro, *Senior Member, IEEE*, Claudia Paris, *Senior Member, IEEE*

Abstract—Identifying buildings and urban/built-up features has become increasingly important as cities grow more complex and traditional pixel-based classification methods struggle with mixed-land-cover signatures. Despite their high spatial resolution and rich multispectral capabilities, openly available datasets such as Sentinel-2 often remain insufficient for extracting fine-scale urban information. For example, roads, buildings that are smaller than their spatial resolution, or urban areas alongside gardens, are the most challenging features to identify. To detect and map built-up structures that fall below the sensor’s nominal pixel size, or have complex mixed land-cover signatures, we propose to use regression-based sub-pixel mapping. We rely on pre-trained Alpha Earth Embeddings, which provide rich, multi-temporal feature representations that help disentangle built-up features in mixed pixels. These embeddings are combined with freely available very high spatial resolution land-cover map for training. The proposed method could successfully identify small buildings and narrow roads, having a size smaller than the spatial resolution of the considered satellite data. In particular, it was able to detect buildings with areas as small as 20 m^2 and roads represented by buffers as narrow as 1.5 m . Our results indicate that combining sub-pixel mapping with embedding representations enables improved identification of complex built-up features having a size smaller than 10 m when compared to existing land-cover maps.

Index Terms—Sub-pixel mapping, urban/built-up features, mixed land cover signatures, Sentinel-2, Random Forest (RF) regression, Open Street Map (OSM), Alpha Earth Embeddings (AEDs).

I. INTRODUCTION

THE urban or built-up feature identification is a fundamental component of land-cover mapping, as cities are dynamic systems that rapidly transform natural landscapes. Accurate detection of built-up areas is crucial to understanding urban growth patterns, infrastructure expansion, and the resulting environmental impacts [1]. Extraction of built-up areas from remote sensing imagery has traditionally relied on very high-resolution data, where individual structures can be delineated using geometric cues, shadows, and texture [2], [3].

Surbhi Sharma, Morris Riedel and Gabriele Cavallaro are with the School of Engineering and Natural Sciences (SENS), University of Iceland, 102 Reykjavik, Iceland, and the Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich, 52428 Jülich, Germany (e-mail: su.sharma@fz-juelich.de, morris@hi.is, gcavallaro@hi.is).

Rocco Sedona is with the Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich, 52428 Jülich, Germany (e-mail: r.sedona@fz-juelich.de).

Claudia Paris is with the University of Twente, Enschede, the Netherlands (e-mail: c.paris@utwente.nl).

With the advent of deep learning, encoder-decoder networks such as U-Net, DeepLabv3+, and HRNet have become the dominant tools for building footprint extraction at sub-meter or meter-scale resolution [4]–[7]. However, such architectures require high-quality, high-resolution training data and remain impractical for large-area or continental-scale applications when very high-resolution data is not consistently available.

The openly accessible Copernicus Sentinel-2 mission plays a crucial role in identifying and mapping major urban areas and overall settlement patterns, offering a valuable combination of high temporal revisit, fine spatial detail, and rich spectral information. These characteristics have enabled the development of several 10 m regional and global land-cover products [8], [9]. Qiu *et al.* [10] introduced Sen2HSE, a fully convolutional network designed for large-scale human-settlement mapping, demonstrating that Sentinel-2 imagery can effectively capture built-up areas. Despite its strong large-area performance, when working at the nominal spatial resolution of Sentinel-2, existing approaches are still unable to detect small individual buildings or sub-pixel heterogeneity inherent to Sentinel-2’s pixel. Conversely, Radoux *et al.* [11] empirically assessed the ability of Sentinel-2 to detect sub-pixel landscape features, focusing on assessing the capability of detecting objects smaller than the satellite pixel, compared to Landsat-8 and SPOT-5. The analysis demonstrated that Sentinel-2, despite its 10 m pixel size, can reliably detect many sub-pixel landscape features such as narrow roads, small linear vegetation strips, and small water bodies. This is possible when sufficient contrast exists between the target and its surrounding environment. The study further indicates that Sentinel-2 can detect roads as narrow as approximately 3 m in rural areas, and small ponds or water bodies when their diameter exceeds roughly 11 m .

Sirko *et al.* [12] demonstrated that high-resolution urban features, such as buildings and roads, can be detected from 10 m Sentinel-2 imagery using a teacher-student deep learning framework. The study trains a “student” model to replicate high-resolution building and road segmentation masks derived from very high-resolution data, demonstrating that Sentinel-2 imagery can capture fine-scale urban structure and support building counts and large-scale urban mapping. However, the approach depends on high-resolution reference data, is computationally intensive, and sensitive to temporal mismatches. Sub-pixel mapping has the potential to address this limitation by estimating the proportion and spatial distribution of mul-

tipple land-cover classes within a single pixel while reducing the computational burden [13], [14].

Regression-based sub-pixel approaches are particularly valuable for the identification of built-up features because they exhibit strong spectral mixing with vegetation, bare soil, and shadows [15]. Without sub-pixel techniques, sensors like Sentinel-2 may underestimate urbanization in dense cores or peri-urban areas [16]. Schug *et al.* [17], [18] have recently developed a fraction land-cover map of Germany and Austria at 10m spatial resolution using Sentinel-1 and Sentinel-2 spectral-temporal metrics. The map characterizes anthropogenic and landscape structures and demonstrates the potential of sub-pixel mapping for national-scale settlement mapping. To this end, a synthetic training dataset was used to simulate the mixed training data used to run the regression. Machine Learning (ML)-based regression approaches trained on synthetically mixed datasets are widely used for mapping built-up features at the sub-pixel level [19]–[21]. Yet, the reliance on such synthetic mixtures is not always advantageous, as they often fail to capture the complex non-linearities present at the ground level. These non-linearities can arise from shadowing and multiple scattering in urban canyons, material weathering, and sub-pixel vertical heterogeneity caused by interactions between buildings and vegetation, all of which are implicitly accounted for when training with traditional map-based approaches [22].

Although most of the existing land-cover products have been generated with Sentinel-2 image time series, recent advances in geospatial foundation models, such as Temporal Embeddings of Surface Spectra for Earth Representation and Analysis (TESSERA) [23] and AlphaEarth Foundations [24], which both provide embeddings at 10m resolution, offer a fundamentally different paradigm known as embeddings-as-data: precomputed, generic, task-agnostic features that can be adapted with lightweight heads by providing learned embeddings from multi-temporal, multi-source satellite remote sensing data at 10m resolution. These models leverage self-supervised learning to create reliable, well-generalizing, transferable, and analysis-ready feature representations that have been shown to achieve high accuracy with limited labeled data [25], [26], potentially overcoming the limitations of both synthetic and map-based training approaches for large-scale built-up area characterization. Despite their promise, the potential of embeddings for identifying built-up features at the sub-pixel level remains largely unexplored.

In this study, we develop a novel regression-based approach for sub-pixel mapping that leverages pre-trained AlphaEarth Embeddings (AEDs) [24] to enhance the identification of built-up features at a resolution finer than that of satellite data. AEDs are particularly well-suited for sub-pixel built-up detection due to their ability to encode rich spatial-spectral context from large-scale Earth observation data in a self-supervised manner. Unlike other geospatial or vision foundation models foundation models, AED are specifically designed for geospatial data, enabling them to capture both fine-grained structural patterns and broader contextual information relevant to land-cover characterization. This is critical for sub-pixel analysis, where built-up surfaces are often mixed with other classes and

exhibit overlapping spectral signatures. Consequently, AEDs improve the separability of built-up components from spectrally similar classes, such as bare ground, while maintaining a compact representation that can be efficiently integrated into fractional mapping frameworks.

Although AEDs are primarily trained for semantic representation, they provide rich spatial-spectral and contextual information that is highly relevant for sub-pixel regression. In our framework, AEDs are used as high-level feature representations, which are then mapped to fractional abundances through a supervised regression framework. The regression model learns the relationship between the embedding space and fractional abundances using high-resolution reference data, thereby bridging the gap between semantic features and sub-pixel estimation.

The selection of AEDs is further motivated by their availability as one of the few publicly accessible precomputed embedding products at 10m spatial resolution having global coverage. This availability enables the effective use of rich spatial-spectral and contextual representations without requiring computationally intensive embedding generation.

Therefore, unlike other approaches that rely directly on Sentinel-2 spectral bands, our method exploits the representational capacity of AEDs, which encode temporal and contextual information at 10m spatial resolution. This makes them particularly effective for identifying built-up areas, including fragmented or sub-pixel structures. In addition, the proposed method generates reliable sub-pixel fractions using regional very high spatial resolution land-cover map, and leverages complementary ancillary data like OpenStreetMap (OSM) built-up footprints (buildings, roads, and railways) and available global land-cover products at 10m for evaluation. The main contributions of this work are as follows:

- 1) A sub-pixel mapping pipeline for built-up feature identification at spatial scales finer than 10m;
- 2) A fractional dataset to support improved mapping of complex built-up structures and enable adaptation to other landscape elements.

The rest of the structure of this paper is as follows. Section II shows the problem formulation and introduces the notation of the paper. Section III describes the proposed method used in this paper. Section IV presents the considered study area and dataset, while Section V shows the experimental setup along with the obtained sub-pixel mapping results. Finally, Section VI draws the conclusion of the paper.

II. PROBLEM FORMULATION AND NOTATION

The proposed method aims to estimate sub-pixel fractions of built-up areas at 10m spatial resolution by leveraging a very high spatial resolution land-cover map as a reference in order to release the constraint of generating a synthetic dataset, which can sharply affect the accuracy of the results obtained. It is assumed that a very high spatial resolution land-cover map is available only for a subset of the study area, as such products are typically provided at regional scales, whereas the sub-pixel analysis is carried out over a much larger domain due to the global coverage of the satellite data. To this end, the approach

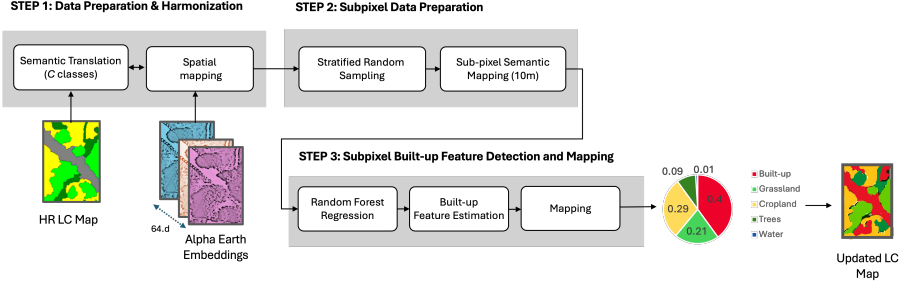


Fig. 1: Workflow of the proposed method based on three main steps: (1) Data Preparation and Harmonization, (2) Sub-pixel Data Preparation, and (3) Sub-Pixel Built-up Feature Detection and Mapping.

explores the use of AEDs as the feature representation of the input imagery, since these embeddings encode informative spectral-spatial features, enabling the regression model to infer fine-scale built-up proportions from the 10m observation grid more accurately.

A key constraint of the approach is that the semantic definitions of the reference land-cover map classes must be compatible with, and distinguishable within, the feature space encoded by the AEDs. In addition, the reference thematic product must be sufficiently reliable and accurate.

Let the spatial domain $\mathcal{S}$ be discretized into a regular grid of N pixels having height and width of 10m respectively, denoted by η_n , each associated with a unique spatial location $s_n \in \mathcal{S}$ for $n = 1, \dots, N$. A multispectral image defined over this domain is represented as a tensor $\mathbf{I} \in \mathbb{R}^{N \times B}$, where $\mathbf{I}(s_n)$ denotes the B -dimensional spectral vector of pixel η_n at location s_n . For each pixel, let $X(s_n)$ denote its AEDs, represented as a 64-dimensional embedding vector $X_{s_n} = (x_1(s_n), \dots, x_{64}(s_n)) \in \mathbb{R}^{64}$. These per-pixel embeddings together form a tensor $\mathbf{Z} \in \mathbb{R}^{N \times 64}$, which serves as the feature representation for sub-pixel mapping tasks, where each pixel η_n acts as the fundamental spatial unit.

Let $M(s)$ denote the very high spatial resolution land-cover map defined over the domain $\mathcal{S}$, discretized into P pixels denoted by ρ_p , each associated with a unique spatial location $s_p \in \mathcal{S}$ for $p = 1, \dots, P$. Each pixel $\rho_p \in \mathcal{S}$ is assigned a land-cover class from the set $\{C_1, C_2, \dots, C_c\}$, where c denotes the total number of classes. Accordingly, the map provides for every ρ_p at s_p , a label $M(s_p) \in \{C_1, C_2, \dots, C_c\}$. In the proposed framework, this map is assumed to be spatially co-registered with the AEDs, to be used to perform the sub-pixel mapping.

Let R_M and R_X denote the spatial resolutions of the high-resolution map and the AEDs, respectively, and define the resolution ratio as $\lambda = R_X/R_M$. Each satellite pixel η_n corresponds to a spatial region that contains $K = \lambda^2$ high-resolution pixels $\{\rho_1, \dots, \rho_K\}$, given that both datasets are co-registered and refer to the same ground coordinates. This establishes a direct multi-scale relationship between the K

fine-resolution elements of the $M(s)$ map and the spatial resolution of the considered satellite data.

III. METHODOLOGY

Fig.1 shows the proposed method, which is based on three steps: (1) Data Preparation and Harmonization, (2) Sub-pixel Data Preparation, and (3) Sub-Pixel Built-up Feature Detection and Mapping. While in the first step, the considered data is prepared and harmonized, the second step aims to automatically generate the sub-pixel training data that will be used to generate the final sup-pixel thematic product in step 3, focused on the identification of built-up features having a size smaller than the spatial resolution of the satellite data.

A. Data Preparation and Harmonization

In the first step of the proposed method, the data is prepared and harmonized from a spatial and semantic viewpoint. To generate the training data used to perform the sub-pixel analysis, the method relies on the semantic information provided by the very high spatial resolution land-cover map. To support the multiscale analysis for sub-pixel mapping the data must be co-registered. Since each satellite pixel η_n at location s_n corresponds to K high-resolution pixels $\{\rho_1, \dots, \rho_K\}$ at location s_k , the sub-pixel multi-label set associated with η_n at s_n is defined as

$$L(s_n) = \{M(s_k) \mid k = 1, \dots, K\},$$

where $M(s_k)$ denotes the land-cover label of the high-resolution pixel ρ_k at location s_k .

To align the map legend of the considered map $M(s)$ with a set of classes that can be discriminated using the considered satellite data $X(s)$, we leverage the Land Cover Classification System (LCCS) [27], a standardized and widely adopted land-cover terminology designed to support the translation and comparison of existing thematic legends. LCCS provides a comprehensive and globally applicable framework for this purpose. Rather than prescribing a single universal legend, LCCS offers a flexible structure that enables land-cover features to

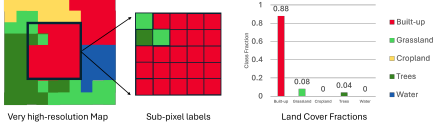


Fig. 2: Illustrative example of the computation of land-cover fractions, where all labels within a single satellite pixel are used to calculate the corresponding fractions.

be described at any scale or level of detail by standardizing the terminology and diagnostic attributes used to define classes. At the end of this step, the set of classes $\{C_1, C_2, \dots, C_c\}$ present in the considered map $M(s)$ is converted into a new set of target classes $\{C_1^{LCCS}, C_2^{LCCS}, \dots, C_c^{LCCS}\}$ based on the LCCS standard. Let us define as φ the mapping that converts each original class into one of the LCCS classes. Applying φ to the map replaces each pixel's original label with its LCCS-equivalent label. In this work, LCCS is used to guide the semantic translation of the very high spatial resolution land-cover map into a set of land-cover classes of interest for the considered downstream task, and which can be accurately classified in the feature space encoded by the AEDs.

B. Sub-pixel Data Preparation

The second step of the method generates sub-pixel data by exploiting the correspondence between each satellite pixel and the associated high-resolution land-cover labels. Per-pixel label sets are derived to represent the mixed composition of the satellite data, from which land-cover fractions are computed to provide the supervision needed for training and evaluating the sub-pixel regression model.

The samples are selected using a region-wise stratified random sampling strategy that accounts for the prior probabilities of the classes present in the study area, utilizing the semantic information provided by the very high spatial resolution land-cover map $M(s)$. Each translated $\{C_1^{LCCS}, C_2^{LCCS}, \dots, C_c^{LCCS}\}$ class serves as a stratum, ensuring that the dataset is both balanced and representative across all land-cover categories. To control for spatial autocorrelation and ensure independent evaluation, samples were drawn from geographically distinct regions, with no spatial overlap between training and validation datasets. Each sample corresponds to a satellite pixel η_n at spatial location s_n , characterized by the AEDs vector $X(s_n)$. For each η_n , a set of sub-pixel multi-labels is derived from the corresponding K high-resolution pixels, $L(s_n) = \{M(s_k) \mid k = 1, \dots, K\}$, where $M(s_k) \in \{C_i^{LCCS} \mid i = 1, \dots, c\}$ after translation of the original thematic classes into the LCCS-based class set. These sub-pixel labels for respective class are aggregated to compute land-cover fractions, which serve as ground-truth targets for training and validation. Figure 2 shows a qualitative example of land-cover fraction computation.

The training data is used to learn the mapping between the AEDs vectors X_s and the corresponding sub-pixel class

fractions, enabling the model to capture how spectral-spatial information relates to fine-scale land-cover composition. The validation data, in turn, provides an independent benchmark for monitoring generalization performance and tuning model parameters. By evaluating the prediction quality on unseen samples, the validation set guides hyperparameter selection and prevents overfitting, ultimately ensuring a more generalizable and reliable sub-pixel mapping model.

C. Sub-Pixel Built-up Feature Detection and Mapping

The final stage of the proposed method identifies built-up features at a resolution finer than that of the original satellite data. To estimate sub-pixel built-up fractions from AEDs, Random Forest (RF) regression [28] is employed because of its robustness and capacity to model non-linear relationships. Since the non-linear transformation from spectral space to semantic feature space has already been learned during pre-training, the resulting embeddings are structured such that downstream tasks can often be addressed with simpler models [29]. While a linear regressor may suffice for many applications, we employ RF regression to accommodate any residual non-linearities in the relationship between embedding features and built-up fractions, and to benefit from its robustness to outliers and feature interactions.

A RF regression model F is trained to predict the built-up fraction $f(s_n)$ from the embedding $X(s_n)$. The predicted fraction is obtained as the average output of all trees in the forest:

$$f(s_n) = \frac{1}{T} \sum_{t=1}^T f^{(t)}(s_n),$$

where T denotes the total number of trees and $f^{(t)}(s_n)$ is the prediction of the t -th tree. The model parameters θ across all trees are optimized to minimize the squared error between predicted and ground-truth fractions derived from high-resolution data:

$$\min_{\theta} \sum_{s_n} (f^{GT}(s_n) - f(s_n))^2,$$

where $f^{GT}(s_n)$ is the aggregated built-up fraction from the high-resolution map. At the end of this step, the predicted built-up fractions are transformed into discrete built-up detections using a threshold, and the resulting layer is used to update the land-cover map at 10 m spatial resolution. Newly identified buildings, including those embedded within mixed pixels, are thereby incorporated into the original map, producing a refined representation of built-up areas. This process enables the delineation of built-up structures at a finer effective resolution than that of the original satellite data, supporting more detailed mapping and analysis of built-up/urban areas.

IV. DATASET DESCRIPTION AND EXPERIMENTAL SETUP

The study area is located in Belgium, a country with diverse urban densities, making it an ideal location for mapping built-up features at the sub-pixel level. In particular, to evaluate the proposed sub-pixel mapping method, we selected four representative patches across Belgium, highlighted in yellow in Fig.

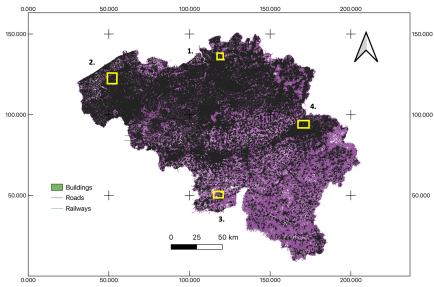


Fig. 3: (a) Built-up areas in Belgium from OSM. Highlighted in yellow are the four different areas for patch selection.

3, collectively comprising 82,380 pixels and approximately 883 built-up features. A very high spatial resolution land-cover map from 2018 is publicly available at the regional level for the Wallonia region [30]. The map was produced using a harmonized, multi-source approach that integrates high-resolution orthophotos (25 cm), 0.8 pts/m^2 LiDAR, Sentinel-1 and Sentinel-2 data, along with existing land-cover and thematic products, and achieved a high overall accuracy of 94.8% ($\pm 1.3\%$) according to an independent validation. The map has a spatial resolution of 2 m and can be used to generate a reliable reference for training and validating sub-pixel built-up estimates in this region. From a semantic viewpoint, the map present a set of 13 classes which are translated into five broader land-cover categories: “Built-up”, “Grassland”, “Cropland”, “Trees”, and “Water”, using the LCCS as described in Section III-A. The semantic translation is reported in Tab. I. This simplified legend facilitates generalization across different hierarchical class schemes while maintaining a clear focus on built-up features. It is important to note that the semantic translation from high-resolution reference data to LCCS classes represents a potential source of uncertainty in the sub-pixel framework, as misclassifications at this stage may propagate to fractional abundance estimates. Although this translation is grounded in the LCCS standard, which provides a hierarchical and structured classification scheme to support consistent and systematic class mapping, a degree of semantic ambiguity remains unavoidable.

A total of 23,711 samples are selected within the Wallonia region of Belgium, following the procedure described in Section III-B. Semantic land-cover fractions for the five target classes are aggregated at a 10 m resolution, with each 10 m unit comprising 25 high-resolution pixels from $M(s)$. Of the 23,711 samples, 70% were used to train the RF regression model, while the remaining 30% were set aside as a validation dataset. From this validation set, 30% of the samples were further reserved for hyperparameter tuning, and the remaining 70% were used to assess the model’s predictive accuracy. To minimize overfitting, hyperparameter optimization was carried out using the dedicated tuning subset, with the objective of

TABLE I: Translation of the Wallonia map legend into the target legend of interest.

Map Legend	Target Classes
Natural Material Surfaces with less than 10 % vegetation	Built-up
Artificially sealed ground surface	
Building, specific structures and facilities	
Grassland with intensive management	Grassland
Grassland and scrub of biological interest	
Inundated grassland and scrub of biological interest	
Vegetation of recently disturbed area (e.g., clear cut)	
Herbaceous in rotation during the year (e.g., crops)	Cropland
Broadleaved trees ($>=3 \text{ m}$)	Trees
Coniferous trees ($>=3 \text{ m}$)	
Small coniferous trees ($<3 \text{ m}$)	
Small broadleaved trees ($<3 \text{ m}$) and shrubs	
Water	Water

maximizing predictive performance on previously unseen data. The sensitivity analysis of the RF hyperparameter is reported in the Appendix.

Finally, to evaluate the ability of the proposed approach to detect built-up in existing land-cover products, we tested and compared three publicly available global land-cover maps at 10 m spatial resolution: ESRI 2018 Global Land Cover [31], Dynamic World 2018 [32], and ESA WorldCover 2020 [33]. The effectiveness of these maps in representing built-up areas was assessed using built-up footprint data from OpenStreetMap (OSM) [34] that includes 8,843,308 major built-up features, including 7,382,160 buildings, 1,431,364 roads, and 29,784 railways. Because roads and railways in OSM are represented as centerlines, their surface area cannot be measured directly. Following Heldens *et al.* [35], category-specific buffers are applied to roads based on OSM highway classes, and railways were buffered using assumed corridor widths representing tracks, ballast, and clearances. This approach allows consistent estimation of surface areas from linear OSM features, while acknowledging that the results are approximate. A list of OSM class/types with respective buffers used in this paper for road and railway features is shown in the Appendix.

Across the four selected patches, the three global products jointly fail to detect 400 of the approximately 883 built-up features (344 buildings, 55 roads, and one railway track). These 400 omitted features form the reference set that the proposed method seeks to recover; the full per-product comparison is reported in Section V-E.

To assess the performance of the proposed method across varying neighbourhood densities, representing increasing levels of built-up intensity, we define reference classes based on fixed density intervals: 0-0.01 (no built-up), 0.01-0.2 (sparse/fragmented), 0.2-0.4 (transitional), 0.4-0.6 (moderately dense), and 0.6-1.0 (dense) regions. Neighbourhood densities are computed within a 20 m buffer using a 2 m resolution reference built-up fraction derived from OSM data, ensuring coverage of at least a 3×3 window of Sentinel-2 10 m pixels. This formulation allows us to quantitatively assess how

well the method reproduces meaningful transitions in built-up intensity and provides a consistent, interpretable framework for evaluating model performance across all density levels.

V. EXPERIMENTAL RESULTS

The results of this study are organized into 5 key components: (1) we analyze the model's ability to detect built-up structures at the sub-pixel level, with particular emphasis on small and fragmented features; (2) we present the calibration of the proposed model along with a sensitivity analysis to examine the influence of critical parameters on its performance, (3) we assess threshold selection across built-up density classes; (4) we discuss the final steps and considerations, highlighting the practical implications and potential limitations of the approach and (5), we evaluate gaps in existing global land-cover products by systematically quantifying the omission of built-up features compared to the baseline version of the method.

A. Sub-Pixel Detection of Built-Up Structures

To obtain predictions for every pixel, we retrieved AEDs and applied the trained RF regression model as described in Section III. Figure 4 presents representative examples comparing the reference built-up fractions derived from OSM data with those predicted by the proposed approach. To facilitate interpretation of the results in relation to spatial heterogeneity, the VHR ESRI Basemap, corresponding Sentinel-2 pixels, Wallonia maps, and OSM built-up layers are also shown. For each example, the classification status is reported, indicating whether the built-up is the dominant land-cover fraction within the pixel. These representative examples were selected to illustrate a range of built-up landscape configurations, including centrally concentrated urban cores, edge-dominated urban areas, transition zones, and sparsely built rural settlements. The selected cases highlight differences in spatial arrangement and surrounding context rather than their frequency of occurrence, providing qualitative insight into sub-pixel heterogeneity. The examples demonstrate that, as expected, pixels containing centrally concentrated urban cores are more likely to be correctly classified, whereas performance tends to decline in sparse settlements and edge-dominated urban regions, where built-up features occupy a smaller proportion of the pixel and exhibit greater spatial fragmentation.

To contextualize the performance of the proposed method, validation pixels were grouped into bins based on their reference built-up fraction, and both normalized RMSE (NRMSE) and MSE were computed for each bin, as shown in Figure 5. Instead of RMSE, a stabilized NRMSE was calculated, where errors were scaled using a mean-standard deviation normalization to ensure robustness. As expected, the MSE increases with higher built-up fractions, reflecting larger absolute errors in densely built-up areas. In contrast, the NRMSE decreases with increasing fraction, indicating improved relative performance where the signal is more homogeneous. This complementary behavior highlights that while absolute errors grow with target magnitude, the model maintains higher relative accuracy in dense urban regions, whereas sparse and heterogeneous areas

remain more challenging due to sub-pixel mixing. Importantly, the elevated NRMSE in low-density and heterogeneous areas highlights the intrinsic challenge of sub-pixel estimation in mixed pixels. Despite this and considering 10 m spatial resolution, sub-pixel mapping provides meaningful fractional estimates even in sparse or fragmented urban landscapes, i.e. built-up fraction values of 0.3 to 0.5, where conventional "hard" classification would likely fail or misclassify partially built pixels. This demonstrates one of the key strengths of the approach: the ability to capture partial occupancy and fine-scale heterogeneity, offering a more nuanced and realistic representation of urban structure across all densities.

B. Selection of Minimum Built-Up Fraction Thresholds

Because the proposed method predicts fractional built-up values rather than discrete labels, converting these fractions into explicit detections requires the specification of a cut-off threshold. Indeed, relying solely on the highest fraction does not necessarily provide correct information regarding the presence of built-up areas (see Figure 4). To determine an appropriate threshold and assess its impact on performance, we conducted a sensitivity analysis aimed at balancing omission errors (missed features) and commission errors (false positives), as shown in Fig. 6 and Tab. II with respect to the reference masks of 82,380 pixels derived from OSM built-up footprints. This analysis provides insight into the trade-offs involved in translating continuous fractional predictions into actionable discrete outputs.

Using a baseline approach in which each pixel was assigned to the class with the highest predicted fraction, resulted in substantial omission errors, successfully recovering only 45 of the 400 built-up features including 12 features with an area between 50 m² and 100 m² that were missed by the global land-cover products. Lowering the threshold reduces omission errors and allows for the detection of smaller, narrow, or fragmented features that are typically unresolved by coarse-resolution products, but it also increases commission error. A visual example depicting the impact of thresholds in detecting built-up features is shown in Fig. 7. It is important to note that the reference mask derived from OSM built-up footprints (see Fig. 7(a.2)) represents all buildings, roads, and railway tracks, regardless of their size, or with sharp boundaries and does not account for spatial blurring or signal mixing introduced by the satellite sensor's PSF. As a result, commission errors are likely overestimated, as sensor-induced signal spreading is penalized as false positive detections. It can also be seen that the commission errors at lower thresholds are particularly pronounced in heterogeneous urban-rural transition zones. These false positives primarily arise from sub-pixel mixing, where non-built-up surfaces such as bare ground or sparsely vegetated areas exhibit spectral characteristics similar to built-up features. Additionally, fragmented structures and mixed land-cover compositions introduce ambiguity in the embedding space, increasing the likelihood of misclassification at lower thresholds. This highlights the inherent trade-off between sensitivity and precision, and underscores the need for careful threshold selection in complex landscapes.

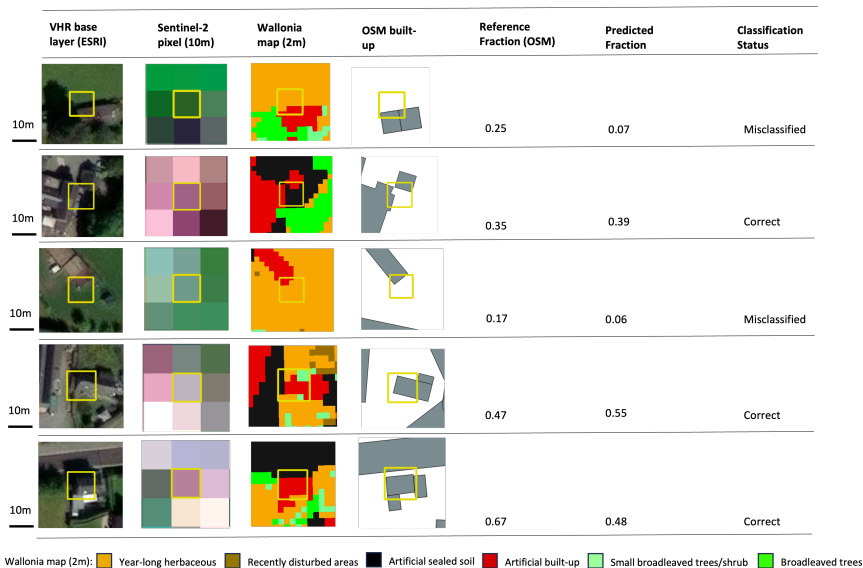


Fig. 4: Representative examples of different built-up landscape configurations, including centrally concentrated urban cores, edge-dominated urban regions, transition zones, and sparsely built rural settlements. The classification status indicated whether the built-up is the dominant land-cover fraction within the pixel.

TABLE II: Built-up Detected with different fractional Thresholds across four patches with respect to the reference mask.

Threshold	True Positives (Correctly predicted built-up)	False Positives (Overestimated built-up)	False Negatives (Missed built-up)	Omission Error	Commission Error
No threshold (highest fraction)	1284	2103	4960	0.79	0.62
Threshold 0.25	1526	2785	4718	0.75	0.64
Threshold 0.20	2043	4246	4201	0.67	0.67
Threshold 0.15	2703	7377	3541	0.57	0.73

C. Threshold Selection Across Built-Up Density Classes

The sensitivity analysis, as reported in Fig. 6, Fig. 7, and Tab. II identified a threshold of 0.20 as the optimal operating point, providing a balance between maximizing the recovery of fine-scale built-up structures and controlling overestimation. Nevertheless, this threshold is inferred from a sensitivity analysis conducted over four patches within Belgium, and its direct transferability to different geographic regions, land-cover compositions, or sensor modalities cannot be assumed. Therefore, it should be interpreted as context-dependent, with potential need for region-specific calibration in broader applications. To evaluate the robustness and empirical validity of the selected decision thresholds with respect to neighbourhood densities, three independent unsupervised classification approaches were applied. Two methods: K-means clustering ($k = 5$), and Jenks natural breaks ($k = 5$) were applied to the positive-density subset of the data, excluding zero-density values to focus on

inhabited structures. A third method, multi-Otsu thresholding ($k = 5$), was applied to the full distribution, allowing the zero-density mode to be naturally absorbed as an independent class. For each method, the threshold separating the sparse and transitional regimes were extracted and compared against the adopted reference value of $T = 0.20$. Particular emphasis was placed on the sparse and transitional regimes because these density classes are especially relevant for sub-pixel analyses, where vegetation cover is often fragmented and heterogeneous. Therefore, a data-driven assessment was conducted to evaluate the suitability of $T = 0.20$ as the boundary between sparse and transitional conditions. The extracted thresholds were subsequently compared with the reference value to determine how well the adopted threshold represents the sparse–transitional transition across methods. Across methods, the inferred boundaries show consistent but not identical estimates, reflecting sensitivity to distributional assumptions and whether the zero-

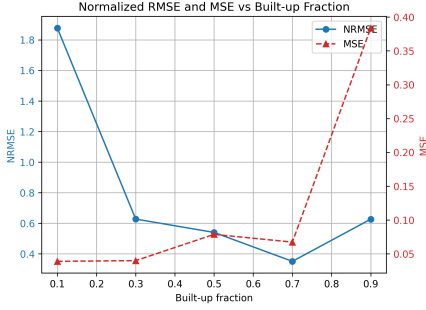


Fig. 5: Error Analysis with respect to built-up fractions on validation data

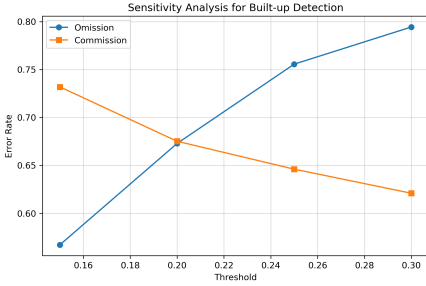


Fig. 6: Sensitivity Analysis for Built-up Detection for different thresholds.

density mode is explicitly included. The resulting values indicate a bounded range of plausible transition points, with estimates spanning approximately 0.10 to 0.30, as summarized in Table III. In particular, the adopted manuscript value of 0.20 lies within the central tendency of the tested approaches (Multi-Otsu: 0.200; Jenks: 0.274; K-means: 0.126; Multi-Otsu bootstrap mean: 0.206 ± 0.007 ; 95% [0.191, 0.219], derived from resampling 100 times with subsets of 30,000 points in patches and re-estimating the sparse — transitional break). Importantly, the analysis on spatial-subset made by splitting the patches into geographic halves and quadrants and re-estimating the empirical break on each subset, yields a narrow range (0.104–0.125), indicating limited sensitivity to local sampling variation. While the maximum local purity estimate suggests a higher possible threshold (0.30), this represents an upper-bound scenario rather than a stable partition across methods.

Overall, these results support the choice of 0.20 as a conservative and methodologically consistent compromise value that lies within the empirically observed range of plausible

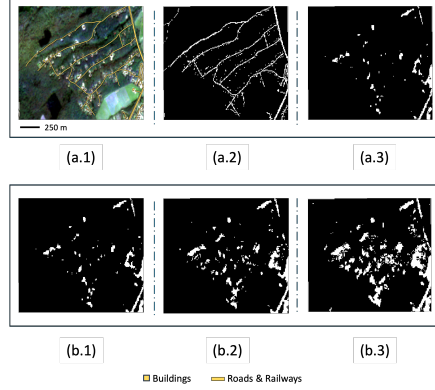


Fig. 7: An example depicting the impact of thresholds in detecting built-up features. (a.1) Sentinel-2 2018 yearly composite with overlaid built-up footprint for reference, (a.2) Reference mask for built-up features obtained using OSM built-up footprint, (a.3) Predicted built-up features using proposed method (highest fraction), (b.1)-(b.3) represent predicted built-up features with thresholds 0.25, 0.20, and 0.15, respectively.

Method	Sparse—Transitional T^*
Adopted (manuscript)	0.200
K-means (1-D, $k = 5$, positive subset)	0.126
Jenks natural breaks ($k = 5$, positive subset)	0.274
Multi-Otsu ($k = 5$, full distribution)	0.200
K-means bootstrap mean $\pm$ std	0.123 ± 0.006
K-means bootstrap 95% CI	[0.105, 0.130]
Multi-Otsu bootstrap mean $\pm$ std	0.206 ± 0.007
Multi-Otsu bootstrap 95% CI	[0.191, 0.219]
Max local-purity T^*	0.300
Spatial-subset range (min .. max)	0.104 .. 0.125

TABLE III: Threshold sensitivity analysis for the adopted ($T=0.20$), with data-driven estimates for the sparse—transitional density boundary.

thresholds, rather than an arbitrary selection. At the same time, the observed variability across methods confirms that the threshold is context-dependent, and should be interpreted as a bounded approximation rather than a universal constant.

D. Practical Implications and Limitations

Applying the optimized 0.20 threshold to the four test patches significantly improved detection performance compared to the baseline. The land-cover maps generated using the highest fraction and the 0.20 threshold detections are reported in Fig. 8(d) integrating these previously invisible sub-pixel features into a refined built-up layer. The newly identified built-up features are highlighted in Fig. 8(e), illustrating the contribution of the proposed approach beyond existing products. While this update enhances the representation of fine-scale infrastructure, it is important to note that threshold-based

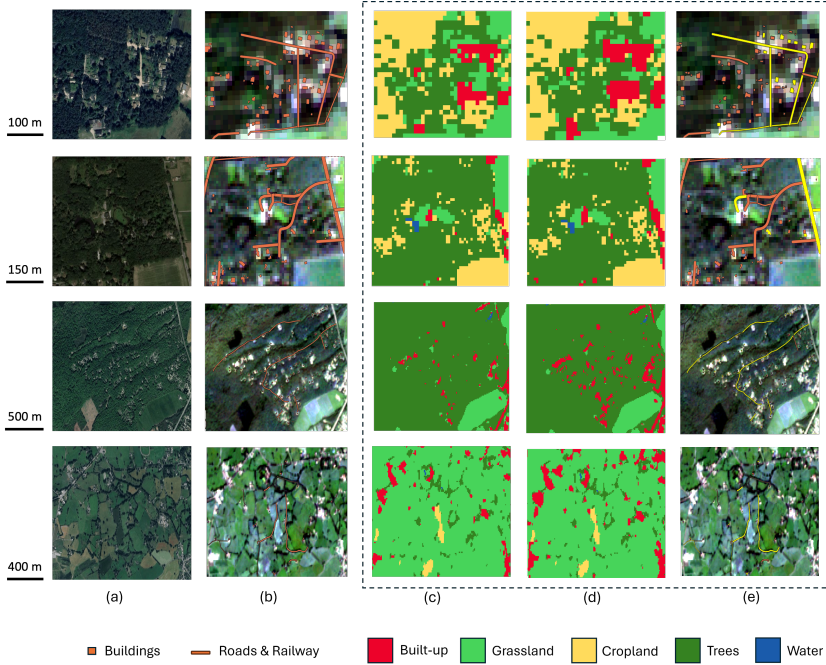


Fig. 8: For each of the four patches: (a) High-resolution Google Earth Image, (b) Sentinel-2 yearly composite for reference. Highlighted in orange are the unidentified overlaid built-up features from OSM in three global land-cover maps, (c) Predicted land-cover Map based on the highest fraction obtained with the proposed model, (d) Updated land-cover map based on a Threshold of 0.20, (e) Highlighted in yellow are the identified built-up features.

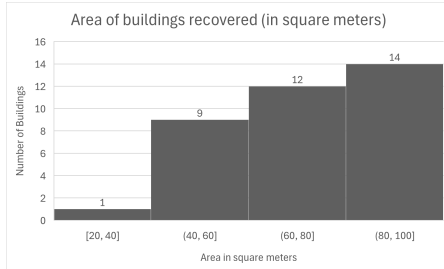


Fig. 9: Area-wise analysis of the recovered buildings across four patches.

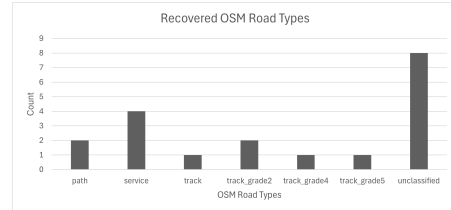


Fig. 10: OSM Road type/classes of recovered features.

integration involves a trade-off between omission and commission errors and remains constrained by sensor resolution and the effects of the PSF. These considerations should be

taken into account when interpreting the refined land-cover map and when applying the approach to other regions or sensors. Lastly, the examples across all maps indicate stronger detection of buildings than of roads or railways. This limitation arises from the sub-pixel nature of the problem, where such features occupy a smaller fraction of the pixel and are there-

fore more affected by spectral mixing and spatial ambiguity. Consequently, their representation in the embedding space is less distinct, leading to reduced detection accuracy. This highlights an inherent challenge in sub-pixel mapping and suggests the need for further methodological improvements to better capture fine-scale and fragmented built-up patterns. Also, to provide a fair assessment of the method, it is important to compute the number of buildings/roads having an area that can be detected reliably given the PSF of the sensor.

When using the 0.20 threshold, the proposed method successfully recovered 101 of the 400 built-up features that were previously undetected by global maps. Of the 344 buildings missed by the three global maps, 222 were smaller than the nominal pixel resolution (10m), or an area (100m²). The proposed method successfully recovered 81 building areas. Notably, this included 36 buildings with areas between 20m² and 100m² as shown in Fig. 9, proving the method's ability to resolve structures significantly smaller than the native satellite resolution. In addition, the method recovered 19 road features and parts of a major railway track. Analysis of OSM road types, see Fig. 10, indicates that while most detected roads were "unclassified" or "service" roads (typically ~5.5m wide), the model also identified 7 narrower features classified as "paths" and "tracks," which correspond to buffer widths of less than 3m.

E. Global Land-Cover Maps: Built-up Detection

To establish the baseline performance of existing products, we compared OSM data against three global land-cover maps at 10m (ESRI 2018, Dynamic World 2018, and ESA WorldCover 2020) across Belgium. Our analysis reveals that global products consistently fail to detect significant infrastructure. Specifically, none of the three global maps identified 289,043 buildings, 255,747 roads, and 752 railway features present in the OSM data, as shown in Fig. 11. These omissions underscore the limitations of current global 10m products and highlight the necessity for sub-pixel mapping approaches capable of resolving small or complex built-up features. It is however important to note that the comparisons with available global land-cover products are included to provide a qualitative benchmark for the sub-pixel estimates. However, differences in spatial resolution and class definitions between these products and the high-resolution reference data necessitate cautious interpretation of any apparent performance improvements. Discrepancies may arise from model performance, but can also reflect variations in mapping protocols, class semantics, and spatial detail. Therefore, these comparisons are intended to support contextual understanding rather than serve as strict quantitative validation.

Focusing on the four patches selected across Belgium to evaluate the proposed sub-pixel mapping method, the global land-cover maps failed to detect 400 built-up features (out of 883), including 344 buildings, 55 roads, and 1 railway, see Fig. 12(c–e), of which 224 features have an area of less than 100m², highlighting the limitations of existing products at this scale. The results are compared again to the proposed baseline approach, in which each pixel was assigned to the

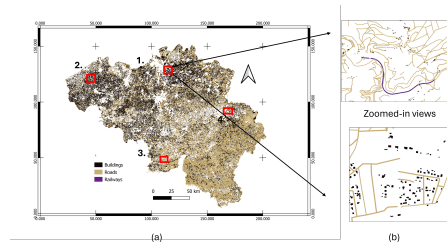


Fig. 11: (a) Built-up areas undetected by all three global land-cover maps. Highlighted in red are the four different areas for patch selection. (b) Zoomed-in views.

class with the highest predicted fraction Fig. 12(f). Given the constraints imposed by the sensor's point spread function (PSF), which describes the spatial blurring of signals across neighboring pixels, the complete detection of all built-up features is inherently impossible. Accordingly, this study focuses on quantifying how many built-up features can be effectively retrieved within these physical limitations by leveraging sub-pixel analysis. While features smaller than the pixel footprint cannot be individually resolved, sub-pixel approaches can still identify their presence, a capability that existing methods, which do not operate at this level, leave unexploited.

F. Computational Performance

Furthermore, the proposed framework is computationally efficient and inherently scalable, with global applicability as a long-term objective. This is enabled by its modular architecture and the use of compact embeddings, which reduce preprocessing overhead and facilitate scalable processing. However, large-scale deployment remains beyond the scope of this study and is deferred to future work. Belgium is selected as a proof of concept, motivated in part by the availability of high-resolution thematic data from the Wallonia region. Nevertheless, the evaluation is conducted across four geographically distinct locations within Belgium, ensuring that the analysis is not confined to a single region. Moreover, land-cover classes exhibiting similar spectral characteristics are expected to generalize more effectively across regions, supporting the potential transferability of the approach.

To evaluate the added value of AEDs for sub-pixel regression, we compare performance against models using the annual composite of raw Sentinel-2 spectral bands. Using the validation subset reserved for accuracy assessment (70% of the validation dataset), the model using AEDs achieves a Mean Squared Error (MSE) of 0.04 and Root Mean Squared Error (RMSE) of 0.21, compared to an MSE of 0.06 and RMSE of 0.23 for the Sentinel-2 spectral features. Although the improvement in accuracy is moderate, the main advantage of using AEDs lies in the shift in processing requirements they enable. By replacing annual Sentinel-2 composites with AEDs, the workflow removes the most computationally demanding stage: preprocessing large volumes of raw Sentinel-2 imagery.

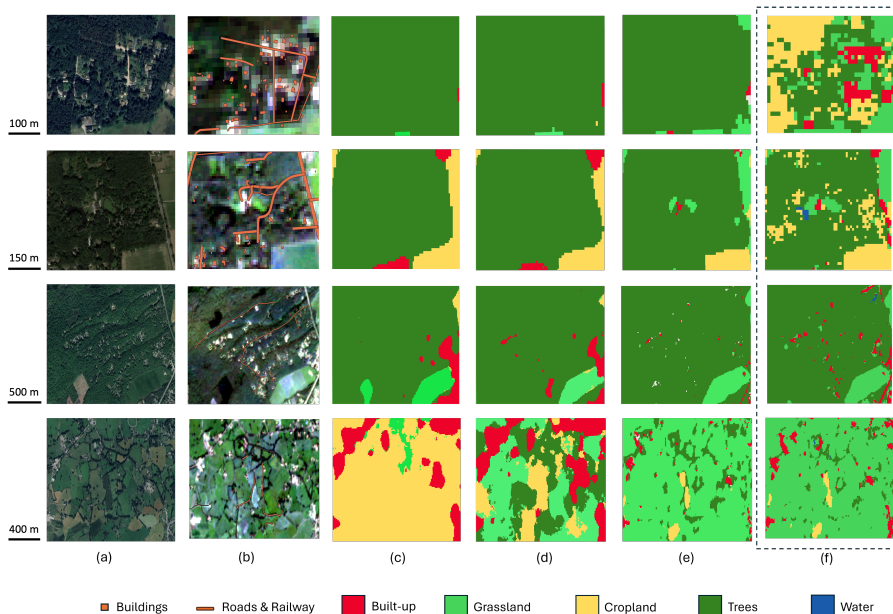


Fig. 12: For each of the four patches: (a) High-resolution Google Earth Image, (b) Sentinel-2 yearly composite for reference. Highlighted in orange are the unidentified overlaid built-up features from OSM in three global land-cover maps, (c) ESRI 2018 Global Land Cover, (d) Dynamic World 2018, (e) ESA World Cover 2020, and (f) Predicted land-cover Map based on the highest fraction obtained with the proposed model.

TABLE IV: Computational performance of the evaluated workflows. Computational cost is reported in seconds per square kilometer (s/km^2).

Data	Preprocessing (s/km^2)	Training (s/km^2)	Inference (s/km^2)	Overall (s/km^2)	Peak Memory (MiB)
Sentinel-2	967.19	26.21	0.99	993.83	3555.97
AEDs	78.25	41.22	1.08	119.94	3936.86

This substantially simplifies large-scale implementation and supports global deployment by removing the need to address region-specific cloud-cover conditions and to design appropriate cloud-mitigation strategies. Table IV summarizes the computational performance of the two workflows, reporting preprocessing, training, and inference times (s/km^2), total computational cost (s/km^2), and peak memory usage (MiB). As expected, training, inference, and memory requirements are similar for both approaches. In contrast, the overall computational cost for Sentinel-2 is significantly higher, due to the extensive preprocessing required to generate cloud-free annual composites from raw imagery. These results demonstrate that AEDs can achieve comparable or better predictive performance while significantly reducing the computational burden associated with large-scale analyses.

VI. CONCLUSION

This paper proposes a regression-based approach for the identification of urban/built-up features at a sub-pixel level, leveraging available thematic land-cover products together with ancillary data. The results indicate that a considerable portion of built-up areas missed by global land-cover products are either too small or too structurally complex to be reliably detected at 10 m resolution. Within this context, the method successfully identifies approximately 101 out of 400 previously unidentified built-up features at a 0.20 threshold, demonstrating its capacity to recover a subset of small or fragmented structures that are not captured in existing global products. This includes buildings with areas as small as approximately $20 m^2$ and narrow linear features such as roads represented by buffers down to about 1.5 m, despite being

below the native spatial resolution of the input imagery. In terms of predictive performance, the use of AEDs yields only modest improvements over the Sentinel-2-based baseline (MSE 0.04 vs. 0.06; RMSE 0.21 vs. 0.23). Rather than large gains in accuracy, the primary contribution of the approach lies in its ability to maintain comparable performance while offering improved scalability and efficient utilization of pre-trained embedding representations for sub-pixel inference. Overall, this study highlights the potential of combining pre-trained embedding representations with sub-pixel regression for improving the characterization of heterogeneous urban landscapes, with particular emphasis on scalability and robustness rather than large absolute gains in accuracy.

As future developments, the proposed method can very well be extended to leverage AEDs to disentangle bare ground components from built-up, which remain challenging to separate due to their spectral similarity. By capturing the informative spectral-spatial structure of the embedding space, AEDs can be used to explicitly model bare ground, improving fractional estimates and enhancing the robustness of the sub-pixel model in heterogeneous urban and peri-urban areas.

ACKNOWLEDGMENT

This work was partially supported by the EUROCC 2 project, funded by the European High Performance Computing Joint Undertaking (EuroHPC JU) and EU/EEA states under Grant Agreement No. 101101903, and by the European Digital Innovation Hub of Iceland (EDIH-IS), partially funded by the Digital Europe Programme.

REFERENCES

- [1] United Nations, "Sustainable development goal 11: Make cities inclusive, safe, resilient and sustainable," <https://www.un.org/sustainabledevelopment/cities/>, 2025, accessed: Dec. 4, 2025.
- [2] J. Li, X. Huang, L. Tu, T. Zhang, and L. Wang, "A review of building detection from very high resolution optical remote sensing images," *GIScience & Remote Sensing*, vol. 59, no. 1, pp. 1199–1225, 2022. [Online]. Available: <https://doi.org/10.1080/15481603.2022.2101727>
- [3] Q. Li, L. Mou, Y. Sun, Y. Hua, Y. Shi, and X. X. Zhu, "A review of building extraction from remote sensing imagery: Geometrical structures and semantic attributes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.
- [4] A. Benchabana, M. K. Kholadi, R. Bensaci, and B. Khaldi, "Building detection in high-resolution remote sensing images by enhancing super-pixel segmentation and classification using deep learning approaches," *Buildings*, vol. 13, p. 1649, 06 2023.
- [5] Y. Yuan, X. Shi, and J. Gao, "Building extraction from remote sensing images with deep learning: A survey on vision techniques," *Computer Vision and Image Understanding*, vol. 251, p. 104253, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1077314224003345>
- [6] L. Lin, P. Li, and X. Yan, "Deep learning-based building extraction from remote sensing images: A comprehensive review," *Energies*, vol. 14, p. 7982, 11 2021.
- [7] W. Sirko, S. Kashubin, M. Ritter, A. Annkah, Y. S. E. Bouchareb, Y. Dauphin, D. Keyers, M. Neumann, M. Cisse, and J. Quinn, "Continental-scale building detection from high resolution satellite imagery," 2021. [Online]. Available: <https://arxiv.org/abs/2107.12283>
- [8] Y. Akbari, S. Solberg, and S. Puliti, "Multitemporal sentinel-1 and sentinel-2 images for characterization and discrimination of young forest stands under regeneration in norway," *IEEE JSTARS*, vol. 14, pp. 5049–5063, 2021.
- [9] R. Van De Kerchove, D. Zanaga, W. Keersmaecker, N. Souverijns, J. Wevers, C. Brockmann, A. Grosu, A. Pacini, O. Cartus, M. Santoro *et al.*, "Esa worldcover: Global land cover mapping at 10 m resolution for 2020 based on sentinel-1 and 2 data," in *AGU Fall Meeting Abstracts*, vol. 2021, 2021, pp. GC451–0915.
- [10] C. Qiu, M. Schmitt, C. Geiß, T.-H. K. Chen, and X. X. Zhu, "A framework for large-scale mapping of human settlement extent from sentinel-2 images via fully convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 163, pp. 152–170, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0924271620300344>
- [11] J. Radoux, G. Chomé, D. C. Jacques, F. Waldner, N. Bellemans, N. Matton, C. Lamarche, R. D'Andrimont, and P. Defourny, "Sentinel-2's potential for sub-pixel landscape feature detection," *Remote Sensing*, vol. 8, no. 6, 2016. [Online]. Available: <https://www.mdpi.com/2072-4292/8/6/488>
- [12] W. Sirko, E. A. Brempong, J. T. C. Marcos, A. Annkah, A. Korme, M. A. Hassen, K. Sapkota, T. Shekel, A. Diack, S. Nevo, J. Hickey, and J. Quinn, "High-resolution building and road detection from sentinel-2," 2024. [Online]. Available: <https://arxiv.org/abs/2310.11622>
- [13] J. Rodríguez-Ortega, S. Tabik, Y. Benhammou, R. Khaldi, and D. Alcaraz-Segura, "Land use and land cover fraction estimation for sentinel-2 rgb images: A new lulc mapping task," *Remote Sensing Applications: Society and Environment*, vol. 39, p. 101626, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S235293852500179X>
- [14] M. Hao, G. Dou, X. Zhang, H. Lin, and W. Huo, "A subpixel mapping method for urban land use by reducing shadow effects," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 2163–2177, 2023.
- [15] J. Rosentreter, R. Hagensieker, A. Okujeni, R. Roscher, P. D. Wagner, and B. Waske, "Subpixel mapping of urban areas using enmap data and multioutput support vector regression," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 5, pp. 1938–1948, 2017.
- [16] Y. Cai, M. Zhang, and H. Lin, "Estimating the urban fractional vegetation cover using an object-based mixture analysis method and sentinel-2 msi imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 341–350, 2020.
- [17] F. Schug, D. Frantz, A. Okujeni, S. van der Linden, and P. Hostert, "Land cover fraction map of Germany at 10m spatial resolution based on Sentinel-1 and Sentinel-2 spectral temporal metrics," 2020. [Online]. Available: <https://doi.org/10.1594/PANGAEA.920894>
- [18] —, "Land cover fraction map of Austria at 10m spatial resolution based on Sentinel-1 and Sentinel-2 spectral temporal metrics," 2020. [Online]. Available: <https://doi.org/10.1594/PANGAEA.923037>
- [19] A. Okujeni, S. van der Linden, L. Tits, B. Somers, and P. Hostert, "Support vector regression and synthetically mixed training data for quantifying urban land cover," *Remote Sensing of Environment*, vol. 137, pp. 184–197, 2013. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425713002009>
- [20] A. Okujeni, S. van der Linden, S. Suess, and P. Hostert, "Ensemble learning from synthetically mixed training data for quantifying urban land cover with support vector regression," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 4, pp. 1640–1650, 2017.
- [21] F. Schug, D. Frantz, A. Okujeni, S. van der Linden, and P. Hostert, "Mapping urban-rural gradients of settlements and vegetation at national scale using sentinel-2 spectral-temporal metrics and regression-based unmixing with synthetic training data," *Remote Sensing of Environment*, vol. 246, p. 111810, 09 2020.
- [22] F. Priem, A. Okujeni, S. van der Linden, and F. Canters, "Comparing map-based and library-based training approaches for urban land-cover fraction mapping from sentinel-2 imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 78, pp. 295–305, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0303243418309942>
- [23] Z. F. *et al.*, "Tessera: Temporal embeddings of surface spectra for earth representation and analysis," 2025. [Online]. Available: <https://arxiv.org/abs/2506.20380>
- [24] C. F. Brown, M. R. Kazmierski, V. J. Pasquarella, W. J. Rucklidge, M. Samsikova, C. Zhang, E. Shelhamer, E. Lahera, O. Wiles, S. Ilyushchenko, N. Gorelick, L. L. Zhang, S. Alj, E. Schechter, S. Askay, O. Guinan, R. Moore, A. Boukouvalas, and P. Kohli, "Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data," 2025. [Online]. Available: <https://arxiv.org/abs/2507.22291>

- [25] V. Marsocci, Y. Jia, G. L. Bellier, D. Kerekes, L. Zeng, S. Hafner, S. Gerard, E. Brune, R. Yadav, A. Shibli, H. Fang, Y. Ban, M. Vergauwen, N. Audebert, and A. Nascetti, "Pangaea: A global and inclusive benchmark for geospatial foundation models," 2025. [Online]. Available: <https://arxiv.org/abs/2412.04204>
- [26] N. Dionelis, C. Fibaek, L. Camilleri, A. Luyts, J. Bosmans, and B. L. Saux, "Evaluating and benchmarking foundation models for earth observation and geospatial ai," 2024. [Online]. Available: <https://arxiv.org/abs/2406.18295>
- [27] H. M. W. CE, D. G. A. M. P. B. A. L. J. and S. C., "A joint initiative for harmonization and validation of land cover datasets," *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, vol. 44, pp. 1719–1727, 2006.
- [28] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. [Online]. Available: <https://api.semanticscholar.org/CorpusID:89141>
- [29] L. Houriez, S. Pilarski, B. Vahedi, A. Ahmadalipour, T. H. Scully, N. Affitto, D. Andre, C. Jaffe, M. Wedner, R. Mazzola, J. Jeffery, B. Messinger, S. McGinley-Smith, and S. Russell, "Scalable geospatial data generation using alphacarth foundations model," 2025. [Online]. Available: <https://arxiv.org/abs/2508.11739>
- [30] J. Radoux, A. Bourdouxhe, T. Coppée, M. De Vroey, M. Dufrière, and P. Defourny, "A consistent land cover map time series at 2 m spatial resolution—the lifewatch 2006-2015-2018-2019 dataset for wallonia," *Data*, vol. 8, no. 1, 2023. [Online]. Available: <https://www.mdpi.com/2306-5729/8/1/13>
- [31] I. Observatory, Esri, and Microsoft, "Esri 2018 global land use/land cover (10m) – sentinel-2 based," ArcGIS Living Atlas of the World, 2018, global 10m land cover map derived from ESA Sentinel-2 imagery. [Online]. Available: <https://www.esri.com/en-us/arcgis/products/arcgis-living-atlas/land-cover>
- [32] C. F. Brown, S. P. Brumby, B. Guzder-Williams, T. Birch, S. B. Hyde, J. Mazzariello, W. Czervinski, V. J. Pasquarella, R. Haertel, S. Ilyushchenko, K. Schwehr, M. Weisse, F. Stolle, C. Hanson, O. Guinan, R. Moore, and A. M. Tait, "Dynamic world, near real-time global 10 m land use land cover mapping," *Scientific Data*, vol. 9, p. 251, 2022.
- [33] D. Zanaga, R. Van De Kerchove, W. De Keersmaecker, N. Souverijns, C. Brockmann, R. Quast, J. Wevers, A. Grosu, A. Paccini, S. Vergnaud, O. Cartus, M. Santoro, S. Fritz, I. Georgieva, M. Lesiv, S. Carter, M. Herold, L. Li, N.-E. Tsendbazar, F. Raimoino, and O. Arino, "Esa worldcover 10 m 2020 v100," Dataset on Zenodo, 2021, european Space Agency (ESA) WorldCover project. [Online]. Available: <https://zenodo.org/record/5571936>
- [34] OpenStreetMap contributors, "Planet dump retrieved from <https://planet.osm.org>," <https://www.openstreetmap.org>, 2017.
- [35] W. Heldens, C. Burmeister, F. Kanani-Sühring, B. Maronga, D. Pavlik, M. Sühring, J. Zeidler, and T. Esch, "Geospatial input data for the palm model system 6.0: model requirements, data sources and processing," *Geoscientific Model Development*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:243238940>



Surbhi Sharma (Student Member, IEEE) received the B.Tech degree in electronics and communication engineering from Amity University, India, in 2015, and the M.Sc degree in Geo-information Science and Earth Observation with a specialization in Geo-informatics jointly from the Indian Institute of Remote Sensing, Indian Space and Research Organization (ISRO), India, and the Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, the Netherlands, in 2018. She is a member of the "AI and ML for Remote Sensing"

Simulation and Data Lab at the Jülich Supercomputing Centre, Germany. She is currently pursuing the Ph.D. degree in computational engineering at the University of Iceland. Her research interest lies in scalable machine learning and deep learning methods for remote sensing applications, with a particular focus on advanced deep transfer learning methods using modern High-Performance Computing (HPC) systems.



missions towards near real-time land-cover classification. Additionally, he specializes in Distributed Deep Learning on High-Performance Computing systems, an area of study that he has been actively engaged in since 2019.

Rocco Sedona (Member, IEEE) received his B.Sc. and M.Sc. degrees in Information Engineering from the University of Trento in 2016 and 2019, respectively, and a Ph.D. degree in Computational Engineering from the University of Iceland in 2023. He is a member of the "AI and ML for Remote Sensing" Simulation and Data Lab at the JSC in Germany. His research interests primarily lie in the field of Deep Learning and its application to Remote Sensing data. He has extensively utilized optical satellite data acquired by Landsat (NASA) and Sentinel (ESA)



of the joint High Productivity Data Processing research group between the Jülich Supercomputing Centre and the University of Iceland. Since 2020, he is also the EuroHPC Joint Undertaking governing board member for Iceland. His research interests include high-performance computing, remote sensing applications, medicine and health applications, pattern recognition, image processing, and data science, and he has authored extensively in those fields. Prof. Dr. – Ing. Morris Riedel's online YouTube and university lectures include High-Performance Computing – Advanced Scientific Computing, Cloud Computing and Big Data – Parallel and Scalable Machine and Deep Learning, as well as Statistical Data Mining. In addition, he has performed numerous hands-on training events in parallel and scalable machine and deep learning techniques on cutting-edge HPC systems.

Morris Riedel (Member, IEEE) received his PhD from the Karlsruhe Institute of Technology (KIT) and worked in data-intensive parallel and distributed systems since 2004. He is currently a Full Professor of High-Performance Computing with an emphasis on Parallel and Scalable Machine Learning at the School of Natural Sciences and Engineering of the University of Iceland. Since 2004, Prof. Dr. – Ing. Morris Riedel held various positions at the Jülich Supercomputing Centre of Forschungszentrum Jülich in Germany. In addition, he is the Head



Professor at the Computer Science Department of the University of Iceland. Since 2022, he has been the Head of the "AI and ML for Remote Sensing" Simulation and Data Lab at the JSC. Since August 2024, he is an Associate Professor with the Faculty of Electrical and Computer Engineering at the University of Iceland. From 2020 to 2023, he held the position of Lead for the High-Performance and Disruptive Computing in Remote Sensing (HDCRS) Working Group under the IEEE GRSS Earth Science Informatics Technical Committee (ESI TC). From 2023 to 2024, he served as Co-chair for the ESI TC. Since 2025, he has been the Chair of the IEEE GRSS Quantum Earth Science and Technology Technical Committee (QUEST TC). Concurrently, he serves as a Visiting Professor at the European Space Agency (ESA) Φ-Lab, where he contributes to initiatives in high-performance and innovative computing for Earth observation. He served as an Associate Editor for the IEEE Transactions on Image Processing (TIP) from 2022 to 2024. He was the recipient of the IEEE GRSS Third Prize in the Student Paper Competition of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2015 (Milan - Italy). His research interests cover remote sensing data processing with parallel machine learning algorithms that scale on distributed computing systems and innovative computing technologies.

Gabriele Cavallaro (Senior Member, IEEE) received his B.Sc. and M.Sc. degrees in Telecommunications Engineering from the University of Trento, Italy, in 2011 and 2013, respectively, and a Ph.D. degree in Electrical and Computer Engineering from the University of Iceland, Iceland, in 2016. From 2016 to 2021, he was the deputy head of the "High Productivity Data Processing" (HPDP) research group at the Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich, Germany. From 2022 to 2024, he served as an Adjunct Associate



Claudia Paris (Senior Member, IEEE) received the B.S. degree and M.S. (summa cum laude) degree in Telecommunication Engineering and the Ph.D. in Information and Communication Technology from the University of Trento, Italy, in 2010, 2012, 2016, respectively. She accomplished the Honors Master Program in Research within the Master Degree in Telecommunication Engineering in 2012. Since 2021 she is an Assistant Professor in the Faculty of Geoinformation and Earth Observation Sciences (ITC) at the University of Twente. Claudia Paris'

research encompasses image and signal processing, machine learning and deep learning, specifically applied to remote sensing image analysis. She focuses on designing innovative and automated workflows for the analysis and classification of large-scale Earth observation data for various applications (e.g., forest/agricultural mapping and monitoring). To this end, she works with multisource remote sensing data (e.g., LiDAR, hyperspectral, multispectral, and high-resolution optical images), multitemporal image analysis, domain adaptation methods, development of ad hoc deep learning solutions, and biophysical parameter estimation. She conducts research on these topics within the frameworks of national and international projects. Since 2023 she has been serving as an Associate Editor for the IEEE Geoscience of Remote Sensing Letters (GRSL). Dr. Paris won the very prestigious Symposium Prize Paper Award (SPPA) at the 2016 International Symposium on Geoscience and Remote Sensing (Beijing, China, 2016) and at the 2017 International Symposium on Geoscience and Remote Sensing (Fort Worth, Texas, USA, 2017). She also won the IEEE Geoscience and Remote Sensing Society 2022 Letters Prize Paper Award (exceptional paper in terms of content and impact on the GRS-Society). Since 2020, she has been a member of the program and scientific committee of several international conferences and workshops.

APPENDIX A SUPPLEMENTARY MATERIAL

A. RF Hyperparameters Sensitivity Analysis

Table V and Figure 13 present the sensitivity analysis of all RF hyperparameters. Together with the RMSE range, in Table V we reported the RMSE variation in % computed as follows:

$$\text{Variation (\%)} = \frac{\text{RMSE}_{\max} - \text{RMSE}_{\min}}{\text{RMSE}_{\min}} \times 100, \quad (1)$$

where, $\text{RMSE}_{\max}$ and $\text{RMSE}_{\min}$ are the maximum and minimum validation RMSE values obtained across the tested values of a given hyperparameter. The largest variation was observed for the number of trees, while maximum tree depth, minimum samples required for splitting, and the feature selection strategy had progressively smaller effects on model performance. Overall, these results indicate that the RF model exhibits stable predictive performance across a broad range of hyperparameter values and that the selected configuration lies within a near-optimal region rather than representing an isolated optimum.

TABLE V: List of key hyperparameters and sensitivity analysis

Hyperparameters	Search space	Selected for RF	RMSE Range	Variation (in %)
Number of trees	[50, 100, 200, 300, 400]	400	0.21171–0.21291	0.57
criterion	['squared error']	squared error	-	-
Max depth	[20, 50, 75, 100, 125, 150]	75	0.21170–0.19928	0.19
Minimum number of splits	[2, 4, 8, 10, 12, 14, 16, 18, 20]	8	0.21155–0.21191	0.17
Maximum number of features	['sqrt', 'log2']	sqrt	0.21171–0.21181	0.04

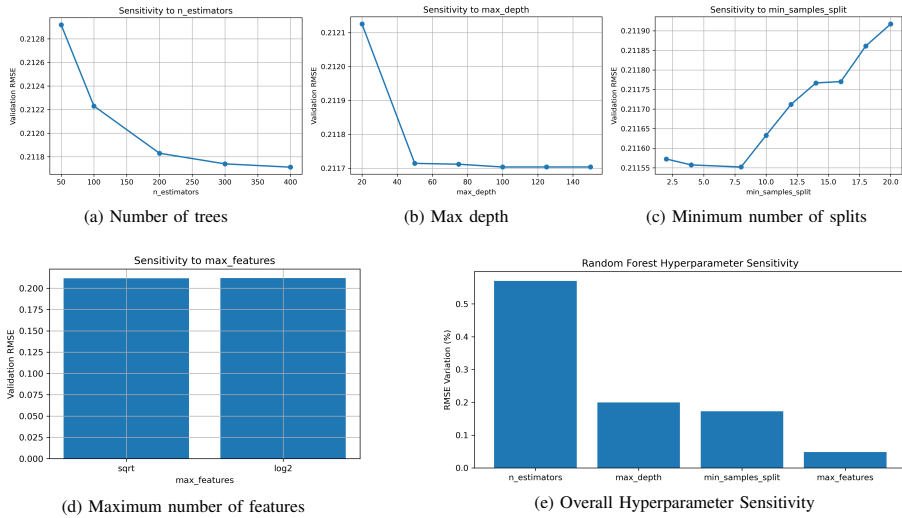


Fig. 13: RF Hyperparameter Sensitivity Analysis. Validation RMSE obtained on the dedicated tuning subset as a function of (a) the number of trees, (b) maximum tree depth, (c) minimum samples required for splitting, and (d) feature selection strategy, (e) The corresponding range-based variability was quantified as the percentage increase from the minimum to the maximum validation RMSE for each hyperparameter configuration. The small variations observed indicate that model performance was relatively insensitive to the investigated hyperparameter values.

TABLE VI: Buffer widths assigned to OSM road and railway features based on their respective types.

OSM road type (fclass)	Buffer (m)
service	5.5
residential	5.5
footway	2.5
path	1.5
unclassified	5.5
track	2.6
tertiary	5.5
cycleway	2.5
secondary	6.5
primary	7.5
track_grade3	2.5
track_grade4	2.5
track_grade2	2.5
track_grade5	2.5
steps	1.5
track_grade1	2.5
pedestrian	1.5
living_street	5.5
trunk	8
motorway	12
motorway_link	8
trunk_link	8
primary_link	7.5
bridleway	2
secondary_link	6.5
tertiary_link	5.5
busway	4
unknown	3

OSM railway type (fclass)	Buffer (m)
rail	12
tram	7
subway	9
light_rail	7
narrow_gauge	5
miniature_railway	3
monorail	4
funicular	4

B. Buffer widths for OSM features

A list of OSM class/types with respective buffers for road and railway features is shown in Table VI.

Paper VI

Diagnosing the Ground–Satellite Gap in Land Cover Using Vision Foundation Models and Field Photos

S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris

IEEE India Geoscience and Remote Sensing Symposium (InGARSS), 2026 (accepted)

©2026 IEEE.

Surbhi Sharma contributed to the conceptualization of the study, experimental design and implementation, data analysis, interpretation of results, and manuscript drafting and revision.

Diagnosing the Ground–Satellite Gap in Land Cover Using Vision Foundation Models and Field Photos

Surbhi Sharma¹, Rocco Sedona², Morris Riedel^{1,2}, Gabriele Cavallaro^{1,2}, Claudia Paris³

¹University of Iceland, Reykjavík, Iceland

²Forschungszentrum Jülich, Jülich, Germany

³University of Twente, ITC, Enschede, The Netherlands

sus21@hi.is, r.sedona@fz-juelich.de, morris@hi.is, gcavallaro@hi.is, c.paris@utwente.nl

Abstract—High-resolution satellite data enable detailed land cover mapping, yet systematic mismatches persist with *in-situ* observations. To investigate this gap, we conduct a combined spatial and semantic analysis of land cover as represented in *in-situ* annotations and satellite images, using geo-tagged field photos to bridge the two domains. This analysis is performed in a zero-shot setting using pretrained vision foundation models, relying solely on the intrinsic visual content of the data without task-specific training. While the spatial analysis aims to quantify within-site visual heterogeneity in the field photos to assess landscape complexity and mixed land cover conditions, the semantic analysis evaluates the consistency of land cover interpretation across field photos, satellite imagery, and *in-situ* reference annotations. The framework is assessed on the publicly available Land Use/Cover Area frame Survey (LUCAS) dataset, which provides *in-situ* observations and geo-tagged field photos that were paired with Sentinel-2 data. Results computed over 169,582 locations across European Union show that when satellite and *in-situ* labels differ: (1) field-photo spatial heterogeneity is a strong predictor of disagreement, with agreement decreasing from 61% in homogeneous sites to 32% in highly heterogeneous ones, and (2) a semantic consistency field-photo-*in-situ* analysis helps disentangle semantic ambiguity in landscape interpretation from potential errors in satellite-based land cover assignment.

Index Terms—Vision-language models, foundation models, embeddings, LUCAS, Sentinel-2, field photos.

I. INTRODUCTION

Satellite-based land cover mapping plays a fundamental role in environmental monitoring, policy development, and decision making [2]. Early global land cover products were commonly generated at coarse spatial resolutions, often around 1 km, limiting their ability to capture fine-scale landscape heterogeneity and small land cover features. The increasing availability of satellite missions such as Sentinel and Landsat has allowed the production of global and regional land cover maps at 10–30 m [3]. Despite the high spatial resolution of modern satellite sensors, land cover remains a highly semantic and scale-dependent concept [4]. In traditional field surveys, enumerators assign land cover labels based on direct

visual inspection and contextual reasoning, often supported by standardized classification guidelines [5]. In contrast, satellite-based approaches infer classes indirectly from spatial and temporal patterns observed in the image time series, constrained by sensor resolution and observation scale [2]. As a result, discrepancies frequently emerge between *in-situ* observation and satellite-derived interpretations, particularly for heterogeneous landscapes, transitional classes, or categories that combine land cover and land use properties [6].

At the same time, modern field data collection protocols increasingly incorporate geo-tagged photos as standard practice [7], providing an intermediate observational layer that connects the *in-situ* annotations based on visual interpretation of field enumerators with the synoptic, top-down perspective of satellite remote sensing [8]. Recent progress in vision foundation models, both unimodal and vision-language, supports this direction. In particular, DINOv3 is a self-supervised vision transformer trained by self-distillation on a large unlabelled image corpus, producing dense patch-level features that preserve local spatial structure without class supervision [9]. Another example is Contrastive Language-Image Pre-training (CLIP) [10], which jointly embeds images and text into a shared space, enabling zero-shot scoring of any image against a fixed text vocabulary. Recently, SenCLIP [11] adapts CLIP to Sentinel-2 by fine-tuning only the image branch on co-located LUCAS photo-satellite pairs, keeping the text encoder frozen and aligning satellite features with CLIP’s text space.

In this context, this paper proposes a diagnostic framework¹ for automatically identifying spatial and semantic cross-view discrepancies between *in-situ* observations and satellite representations, leveraging geo-tagged field photos as a bridge between the two domains. Understanding where and why representations diverge is a prerequisite for more informed training, validation, targeted fusion, domain adaptation, or label refinement, rather than an end in itself. The framework builds on frozen vision foundation models, including DINOv3 for capturing spatial structure and CLIP/SenCLIP for extracting semantic concept activations over a land cover vocabulary (defined from the land cover taxonomy used in the field reference data collection) to systematically charac-

This work is supported by the AgroVision project, “An Innovative Foundation Model for the Fusion of Satellite Image Time Series and Field Photos into Comprehensive Agricultural Reference Data”, funded by the European Space Agency (ESA) Φ-Lab under contract No: 4000147606. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer JUWELS [1] at Jülich Supercomputing Centre (JSC).

¹Code, configuration and all result tables:

<https://gitlab.jsc.fz-juelich.de/sdlrs/agrosemantics/garss, tag v2.1.1.1.>

terize mismatches with *in-situ* data. This reliance on existing foundation models and measures is intentional: the objective is not to introduce another representation-learning model or learning strategy, but to establish a model-agnostic spatial-semantic diagnostic framework for exposing and characterizing discrepancies between satellite representations and *in-situ* observations. Accordingly, the framework is intended as a diagnostic tool for identifying such discrepancies, rather than for producing a corrected or harmonized land cover map. Our contributions are:

- (i) A diagnostic formulation of the ground–satellite gap that jointly models spatial heterogeneity in geo-tagged field photos and semantic agreement across field photos, satellite imagery, and *in-situ* annotations, enabling cross-view comparison at the site level.
- (ii) A zero-shot framework based on frozen vision foundation models, requiring no task-specific training, combining DINOv3 for spatial structure representation and CLIP/SenCLIP for semantic alignment with a land cover vocabulary across modalities.
- (iii) An empirical evaluation on 169,582 LUCAS sites across European Union, showing that field-photo heterogeneity is strongly associated with satellite–*in-situ* disagreement, and that semantic consistency analysis further helps disentangle landscape interpretation ambiguity from potential satellite misclassification.

II. DATASET DESCRIPTION

The proposed diagnostic framework leverages publicly available geo-tagged field observations and photographs from the 2018 release of the LUCAS [5], which provides a quality-controlled dataset in which photographs have undergone prior review for completeness and image quality. For the experiments, these observations were temporally aligned with Sentinel-2 time series derived from the Sen4Map benchmark dataset [12]. LUCAS provides, for each sampled location, five field views: one point-centred view (P) and four cardinal landscape views (N, E, S, W). Each location is annotated with land cover and land use information following a hierarchical land cover taxonomy structured into three levels: Level-1 land cover class (LC1), comprising eight broad categories (artificial surfaces, cropland, woodland, shrubland, grassland, bare land, water, and wetlands); Level-2 land cover class (LC2), providing intermediate thematic refinement; and Level-3 land cover class (LC3), defining the most detailed land cover classes. For each LUCAS point, the co-located satellite observations from Sen4Map provide a time series of Sentinel-2 data for 2018, considering images with less than 40% of cloud coverage and less than 5% of no-pixel data, together with per-month cloud and band-validity flags and per-band normalisation statistics. These observations are aggregated into 12 monthly composites, using the median reflectance value for each month to represent the available observations while applying the Scene Classification Layer to mitigate cloud contamination according to [13]. This results in a harmonized

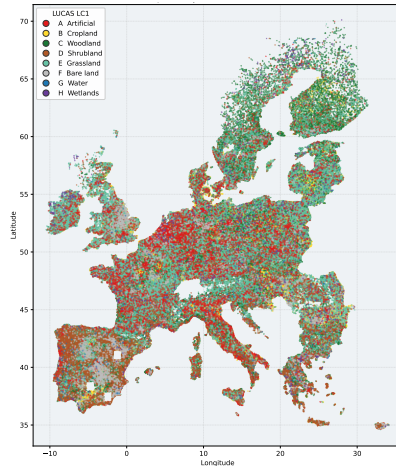


Fig. 1: Spatial distribution of LUCAS samples used in the experiments. The samples are associated with different colors according to their *in-situ* land cover LC1 labels.

monthly time series across the spatial and temporal dimensions, yielding $N=169,582$ paired sites. Fig. 1 shows the spatial distribution of the considered LUCAS points associated with different colors according to their *in-situ* land cover LC1 labels.

III. SPATIAL-SEMANTIC DIAGNOSTIC FRAMEWORK

Fig. 2 illustrates the proposed framework, designed to quantify discrepancies between *in-situ* and satellite-derived land cover information from both spatial and semantic perspectives. The framework relies on four key components: (i) surveyed points associated with *in-situ* land cover annotations collected during the field campaign, (ii) a shared land cover concept vocabulary defined by the land cover taxonomy adopted in the field data collection protocol, (iii) field photos capturing local visual context, and (iv) co-located Sentinel-2 satellite imagery.

From a spatial perspective, field photos are used to characterize within-site landscape heterogeneity, while from a semantic perspective, land cover concepts activated by field-photo and satellite representations are compared against the *in-situ* reference labels to determine ambiguity in the semantic interpretation of the observed landscape. To this end, we leverage recent advances in self-supervised and contrastive representation learning to tailor feature extraction to each analysis objective. DINOv3’s self-supervised learning objective produces strong patch-level spatial features, making it well suited for measuring within-image structural variation. Its multi-layer token representations capture both fine-grained

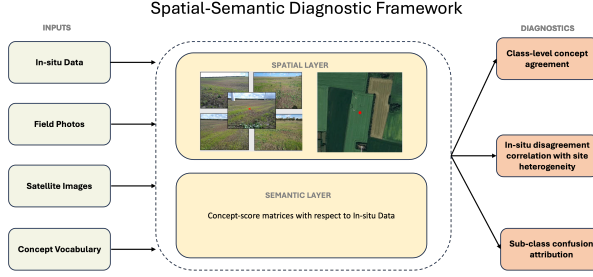


Fig. 2: Spatial and semantic diagnostic framework for assessing *in-situ*-satellite land cover discrepancies using field photos, satellite images, a concept vocabulary (defined by the field data collection protocol), and *in-situ* land cover labels.

textures and broader, macro-scale spatial patterns, allowing landscape heterogeneity to be quantified continuously rather than reduced to a small number of discrete classes. In contrast, CLIP/SenCLIP’s contrastive vision-language pretraining yields embeddings that align naturally with text concepts and so support concept-profile comparison. Overall, the proposed framework produces a set of diagnostic outputs that improve the understanding of when and why discrepancies arise between satellite-derived land cover estimates and *in-situ* observations.

A. Spatial Analysis: Landscape Complexity

Land cover discrepancies partly arise because satellite and *in-situ* observations describe the landscape under different spatial supports and viewpoints. Satellite pixels naturally aggregate reflected signals over the area determined by the sensor spatial resolution and may therefore contain mixtures of multiple land cover types within a single observation unit. In contrast, *in-situ* annotations are assigned at surveyed locations based on ground-level visual inspection and local contextual information.

The first step aims to assess the landscape heterogeneity associated with each satellite pixel using the 5 geo-tagged field photos. Let $\mathcal{P} = p_{i=1}^N$ denote the set of surveyed locations, where each location p_i is associated with five field views $\mathcal{D}_g = \{P, N, E, S, W\}$ provided by the LUCAS dataset. For each field photo, the frozen DINOv3 Vision Transformer (ViT)-L/16 encoder produces a grid of patch tokens. With an input size of 224×224 and a patch size of 16×16 , this gives $Q_0 = 196$ tokens arranged on a 14×14 grid,

$$\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_{Q_0}]^T \in \mathbb{R}^{Q_0 \times d}, \quad d=1024. \quad (1)$$

For a generic token set $\mathbf{Z}$ of size Q , we define heterogeneity as the mean pairwise cosine distance among its tokens,

$$H(\mathbf{Z}) = \frac{2}{Q(Q-1)} \sum_{1 \leq j < k \leq Q} \left(1 - \frac{\langle \mathbf{z}_j, \mathbf{z}_k \rangle}{\|\mathbf{z}_j\| \|\mathbf{z}_k\|} \right). \quad (2)$$

Normalization by the number of token pairs makes the score comparable across token sets of different sizes. For each site p_i , DINOv3 is applied independently to the five ground-level views $d \in \mathcal{D} = \{P, N, E, S, W\}$, yielding per-view token grids $\mathbf{Z}_{i,d}^g \in \mathbb{R}^{Q_0 \times d}$. The site-level ground token pool is obtained by concatenating the tokens from all views,

$$\mathbf{Z}_i^g = \text{concat}(\{\mathbf{Z}_{i,d}^g\}_{d \in \mathcal{D}}), \quad H_g^{(i)} = H(\mathbf{Z}_i^g). \quad (3)$$

The resulting score $H_g^{(i)}$ captures both local variation within individual photos and directional variation across the five views, providing a site-level proxy of landscape heterogeneity.

B. Semantic Analysis: Land Cover Consistency

The second step of the proposed approach aims to identify *in-situ* samples that may exhibit semantic ambiguity in landscape interpretation. This is achieved by focusing on cases where a disagreement arises between satellite-derived predictions and the corresponding LUCAS *in-situ* labels. In such cases, field photos are used as an intermediate source of evidence to assess and compare the land cover concepts activated in field photos and satellite data.

To perform the semantic assessment, we define a fixed land cover concept vocabulary of 74 captions, each corresponding to a class in the official LUCAS Level-3 (LC3) taxonomy as published in the Eurostat data dictionary [5]. Owing to the hierarchical structure of the LUCAS classification scheme, each LC3 class inherits its corresponding LC2 and LC1 parent categories, enabling evaluation of predictions against reference labels at multiple levels of semantic granularity (LC1, LC2, and LC3). Let $\mathcal{V} = \{v_k\}_{k=1}^K$ be the land cover concept vocabulary from LC3, where $K = 74$. Field photos are encoded with frozen CLIP ViT-B/32 [10], Sentinel-2 monthly composites with SenCLIP [11], and the LC3 concept names $\{c_k\}_{k=1}^K$ with the frozen CLIP text encoder. Both image modalities are scored against the same set of LC3 text embeddings using cosine similarity.

For each site, the field photo representation is the normalised mean of the five view embeddings. The satellite representation

is taken from the Sentinel-2 monthly composite closest in time to the LUCAS survey date, with a peak-season fallback for sites whose survey date is unavailable. Each Sentinel-2 monthly composite (B4, B3, B2 at 10 m resolution, 64×64 pixels) is normalized, upsampled to 224×224 pixels by bilinear interpolation before being passed to the frozen SenCLIP ViT-B/32 satellite encoder.

Let $v \in \{g, s\}$ index the two modalities (field photos, satellite), and let $\mathcal{V}_v \subseteq \{1, \dots, N\}$ denote the subset of sites with a valid representation for modality v , with $N_v = |\mathcal{V}_v|$. For site $i \in \mathcal{V}_v$ and concept $k \in \{1, \dots, K\}$, let

$$a_{i,k}^v = \frac{\langle e_i^v, t_k \rangle}{\|e_i^v\| \|t_k\|} \quad (4)$$

be the cosine similarity between the site embedding e_i^v and the frozen CLIP text embedding t_k of concept $c_k \in \mathcal{V}$. Scores are then centred per concept across the valid sites of each modality,

$$\tilde{a}_{i,k}^v = a_{i,k}^v - \frac{1}{N_v} \sum_{j \in \mathcal{V}_v} a_{j,k}^v, \quad (5)$$

which removes global vocabulary-level score offsets so that the top-1 decision depends on site-specific relative evidence rather than on concepts systematically favoured by the embedding geometry.

Let $\hat{c}_i^v$ be the top-scoring LC3 concept for site i and modality v after score centring,

$$\hat{c}_i^v = c_{\arg \max_k \tilde{a}_{i,k}^v}. \quad (6)$$

The predicted LC3 concept can be mapped to its LC1 parent through the codelist projection π . Semantic agreement with the LUCAS LC1 label y_i is then defined as

$$\alpha_i^v = \begin{cases} 1, & \text{if } \pi(\hat{c}_i^v) = y_i, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

At the *class level*, we report the per-class mean agreement

$$\bar{\alpha}_c^v = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \alpha_i^v, \mathcal{I}_c = \{i : y_i = c\}, \quad (8)$$

which can be evaluated against the LUCAS reference at LC1, LC2, and LC3, thereby providing a multi-level assessment of semantic consistency and taxonomy-aligned agreement. In the considered implementation, the satellite prediction is defined as the argmax over the 74 LC3 concepts, hierarchically mapped to LC2 and LC1 through the LUCAS land cover taxonomy, without additional re-optimisation or aggregation across levels.

IV. DIAGNOSTIC OUTPUTS

Fig. 3 illustrates two examples of *in-situ*-satellite land cover disagreement associated with (a) water and (d) woodland land cover. The LUCAS field photo (P) together with the corresponding Sentinel-2 true-color images (10m), and VHR Google Baselayer (30cm) are reported. Fig. 3a shows an example in which the presence of a river is correctly reported by the field enumerator. However, the limited spatial extent

of the water body, together with surrounding rock and grass, may lead to misclassification at the satellite pixel level. Fig. 3d shows an example where the presence of trees is located within a shrubland area. Although trees are present in the scene, the open woodland structure may also be interpreted as shrubland by different enumerators, and may be inconsistently represented at the satellite pixel level as well. To automatically identify cases potentially associated with incorrect land cover assignment from satellite data, the proposed framework goes beyond traditional accuracy assessment and a set of diagnostic outputs that provide interpretable indicators of cross-modal consistency across semantic, spatial, and annotation dimensions.

A. Spatial Analysis: Landscape Complexity

This step aims to quantify how mixed land cover conditions at satellite spatial resolution affect classification consistency. To this end, we exploit the landscape heterogeneity metric H_g computed on the 5 field photos. Fig. 4 illustrates the relationship between field-photo-derived landscape heterogeneity H_g and satellite-*in-situ* agreement across LC1 classes. The x-axis shows five quantiles of increasing heterogeneity, computed per LUCAS samples. The blue bars indicate agreement between satellite-derived labels and *in-situ* annotations, while the orange bars represent disagreement. Comparing single-view (P) and pooled ($P+N+E+S+W$) heterogeneity, the results show the benefit of pooling the five views. Both the single point view (a) and the five-view pool (b) reproduce a monotonic decline of agreement with increasing heterogeneity, but pooling gives a steeper trend with a wider dynamic range: agreement falls from 60.9% in the most homogeneous sites (Q1) to 31.5% in the most heterogeneous ones (Q5), a 29 pp spread against only 22 pp for P alone. Since both converge to 31.5% at Q5, this difference is concentrated at the homogeneous end, where pooling raises Q1 agreement from 53.8% to 60.9%: the cardinal landscape views isolate genuinely homogeneous, high-agreement sites that the point-centred view alone does not, making H_g a sharper indicator of satellite-*in-situ* disagreement. This additional information supports a more informed and transparent evaluation of the *in-situ* reference land cover information. Therefore, these outcomes can be used to guide reference sample selection by prioritising *in-situ* annotations that exhibit high spatial and semantic consistency, as inferred from the associated field photos. Such samples are more likely to provide reliable information for satellite-based land cover mapping, as their reduced ambiguity helps narrow the gap between *in-situ* observations and satellite-derived representations. Notably, no training data are required to obtain these results, thereby avoiding any potential influence on the analysis.

Although expected, this result confirms the feasibility of automatically assigning a landscape complexity score to LUCAS samples by leveraging the visual information contained in the field photo. This additional information enables a more informed selection of training data, allowing users to decide whether to use for training only pixels with a higher likelihood

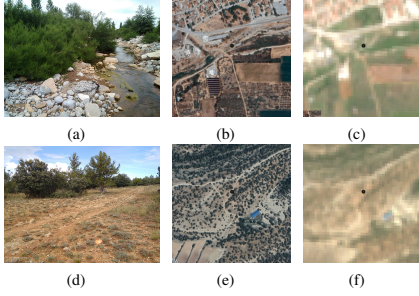


Fig. 3: Example of LUCAS field photos associated with land cover: (a) water, and (d) woodland. The corresponding locations are reported in (b)(e) VHR Google Baselayer, and (c)(f) Sentinel-2.

of representing pure land cover or, alternatively, to incorporate mixed spectral signatures. Similarly, when evaluating the accuracy of satellite-based land cover maps, this diagnostic information on landscape complexity and heterogeneity can support a more informed assessment of the map quality.

B. Semantic Analysis: Land Cover Consistency

To automatically identify samples prone to disagreement arising from subjective interpretation of land cover by field enumerators, we focus on cases where satellite-derived and *in-situ* land cover labels do not match. For all such disagreement cases, we analyse the 74 LC3 concepts activated by the corresponding field photos. This allows us to evaluate whether the observed discrepancy is also reflected in the ground-level visual evidence, thereby helping to distinguish cases where disagreement originates from satellite representation limitations from those where it is already present in the field-based interpretation of the scene (see Fig. 3d). Table I reports the per-class sample support together with the zero-shot LC1 agreement of the field photos (CLIP) and the satellite (SenCLIP) with the *in-situ* labels, obtained by mapping each top-1 LC3 concept to its LC1 parent. As expected, the field photos agree more strongly with the *in-situ* labels than the satellite across all classes (58.6% vs. 43.1% overall), with both dropping sharply on the heterogeneous and under-represented classes (Shrubland, Grassland, Water, Wetlands). The substantial class imbalance, particularly for rare classes such as Water and Wetlands therefore warrants a cautious interpretation of results.

Fig. 5 illustrates, for each LC1 class, the distribution of LC3 top-1 destination frequencies derived from field photos and satellite images. Also in this case, these results are particularly interesting as they are obtained in a zero-shot setting, without any potential bias introduced by training data. A first notable finding is the agreement at LC3 land cover concepts between satellite imagery and field photos, even in cases where

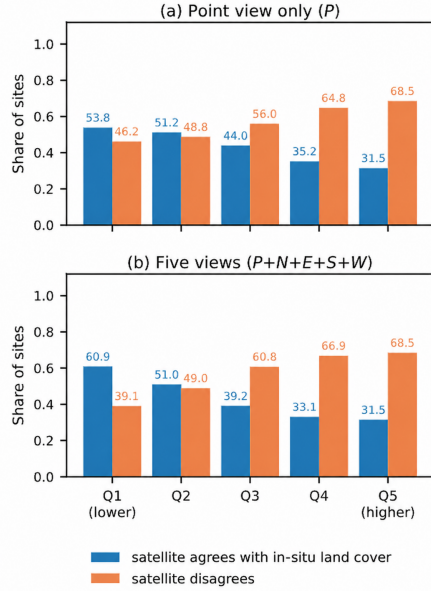


Fig. 4: Relationship between field-photo-derived landscape heterogeneity (H_g) and satellite-*in situ* agreement at LC1 level, grouped into five heterogeneity quantiles (Q1–Q5). (a) H_g from the single point-centred view (P); (b) H_g pooled over the five LUCAS views (P+N+E+S+W).

TABLE I: Per-class LUCAS sample support and zero-shot LC1 agreement (top-1 LC3 concept projected to LC1) of the field photos (CLIP) and the satellite (SenCLIP) against the *in-situ* LC1 label.

LC1 class	Samples	Field-photos agr. (%)	Satellite agr. (%)
A Artificial	10,775	53.9	39.5
B Cropland	58,669	90.9	72.9
C Woodland	44,129	62.9	44.5
D Shrubland	7,545	21.9	13.2
E Grassland	43,021	22.0	10.1
F Bare land	3,956	22.9	23.0
G Water	576	31.8	12.2
H Wetlands	911	32.8	16.8
Total	169,582	58.6	43.1

both disagree with the *in-situ* labels. For example, within the Artificial Surfaces class, several samples are consistently identified as rocks, stones, sand, or permanent vegetation (e.g., fruit trees) across both modalities. Similarly, some cropland areas appear to have been surveyed at phenological stages where crops have just been planted or harvested, and are

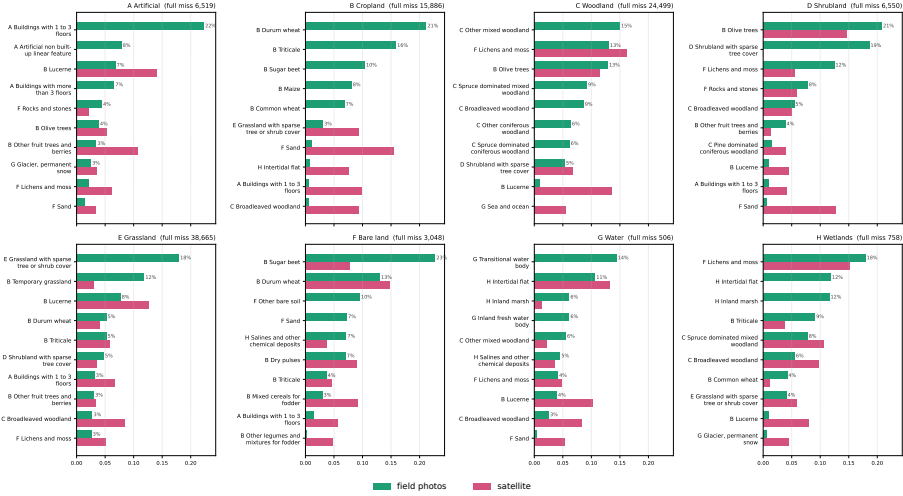


Fig. 5: LC3 destination frequencies for satellite LC1 full-miss errors. For each *in-situ* LUCAS LC1 class, we consider only samples whose satellite LC1 top-1 prediction is incorrect; the corresponding sample count is shown in the panel title. Bars report the fraction of these samples assigned to each LC3 top-1 destination by the field-photo encoder, in green, and by the satellite encoder, in purple. The plot diagnoses the semantic destinations of satellite errors and compares them with the field-photo decisions on the same samples. The ten most frequent LC3 destinations per LC1 class are shown.

therefore associated with grassland or bare soil-like categories. However, many samples still show agreement between *in-situ* information and the concepts derived from field photos, suggesting that discrepancies are often driven more by differences in viewpoint and spatial representation than by fundamentally ambiguous semantic interpretation. In the Water class, many samples are associated with transitional water body, thus providing an indication of the potential transitional nature of the class reported by the enumerator. Indeed, this automatic analysis also enables targeted visual inspection of the data, focusing on cases where the field photos indicate a clear discrepancy or ambiguity. Given the large volume of samples, exhaustive manual inspection would be impractical; instead, the proposed framework enables a selective review of the most critical cases, similar to an active learning setting. As a zero-shot diagnostic, these results provide an indirect measure of consistency between satellite representations, *in-situ* labels and LC3 concepts, and should not be interpreted as definitive evidence of label correctness.

C. Computational Efficiency

The proposed spatial-semantic diagnostic framework operates directly on embeddings rather than imagery after a one-off encoding step which takes about 35 GPU-minutes for 169,582 sites. The zero-shot scoring, class-controlled statistics, and

sensitivity analyses can be performed on cached embeddings in about two minutes on a CPU. Similarly, changes to the taxonomy or prompts can be re-evaluated within seconds. As foundation-model embeddings are increasingly distributed as data products, such as AlphaEarth [14] and TESSERA [15], label-free auditing of reference data could therefore be reduced to a CPU-side analysis of published embeddings, provided that the product is language-aligned for the semantic analysis and exposes either per-image patch-token statistics or a corresponding scalar measure of heterogeneity.

V. CONCLUSION

This paper has presented a diagnostic approach to better understand spatial and semantic discrepancies between *in-situ* and satellite-derived land cover data, leveraging geo-tagged field photos as an intermediate source of evidence. To this end, we leverage frozen vision models together with self-supervised and contrastive representation learning to extract complementary spatial and semantic information without requiring additional supervision. From a spatial perspective, results obtained at the European Union scale on $\sim 170K$ samples demonstrate that sub-pixel heterogeneity has a strong impact on satellite-*in-situ* agreement, highlighting the limitations of the satellite spatial resolution in cases of mixed land cover conditions. From a semantic perspective,

these preliminary results provide promising evidence that field photos can help disentangle whether disagreements arise from limitations in satellite representations or from inherent ambiguity already present in field-level visual interpretation and inconsistencies in land cover taxonomy. However, the resulting discrepancies should be interpreted as diagnostic indicators of potential inconsistencies rather than definitive evidence of satellite misclassification, as they may also reflect ambiguity in field observations, reference-label uncertainty, temporal land-cover change, differences between classification schemes, or limitations of zero-shot inference. Nevertheless, the proposed analysis can be used to identify more reliable labelled samples for constructing better-informed training and validation sets, by leveraging a richer understanding of the intrinsic properties of reference land cover samples. Future work will explore richer concept vocabularies, additional satellite encoders, and analysis across multiple spatial scales. Further investigation of the sensitivity of semantic agreement to the choice of LC3 concept vocabulary, prompt formulation, text descriptions, and embedding model could provide additional insights into the robustness of the semantic predictions and help identify effective configurations for hierarchical land-cover assessment.

USE OF GENERATIVE AI TOOLS

During manuscript preparation, we used Writefull's language model for English research writing to improve grammar, phrasing, and academic style. The tool was used only for language editing and did not contribute to the scientific content, analysis, results, or conclusions.

REFERENCES

- [1] Jülich Supercomputing Centre, "JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Jülich Supercomputing Centre," *Journal of large-scale research facilities*, vol. 7, no. A183, 2021.
- [2] Cristina Gómez, Joanne C White, and Michael A Wulder, "Optical remotely sensed time series data for land cover classification: A review," *ISPRS Journal of photogrammetry and Remote Sensing*, vol. 116, pp. 55–72, 2016.
- [3] Zander S Venter, David N Barton, Tirthankar Chakraborty, Trond Simensen, and Geethen Singh, "Global 10 m land use land cover datasets: A comparison of dynamic world, world cover and esri land cover," *Remote Sensing*, vol. 14, no. 16, pp. 4101, 2022.
- [4] Claudia Paris, Laura Martinez-Sanchez, Marijn Van Der Velde, Surbhi Sharma, Rocco Sedona, and Gabriele Cavallaro, "Accuracy assessment of land-use-land-cover maps: the semantic gap between in situ and satellite data," in *Image and Signal Processing for Remote Sensing XXXI*. SPIE, 2023, vol. 12733, pp. 187–200.
- [5] Raphaël d'Andrimont, Momchil Yordanov, Laura Martinez-Sanchez, Beatrice Eiselt, Alessandra Palmieri, Paolo Dominici, Javier Gallego, Hannes Isaak Reuter, Christian Joebges, Guido Lemoine, and Marijn van der Velde, "Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union," *Scientific Data*, vol. 7, no. 1, pp. 352, 2020.
- [6] Liangyun Liu, Xiao Zhang, Yuan Gao, Xidong Chen, Xie Shuai, and Jun Mi, "Finer-resolution mapping of global land cover: Recent developments, consistency analysis, and prospects," *Journal of Remote Sensing*, vol. 2021, 2021.
- [7] Vyrion Antoniou, Cidália Costa Fonte, Linda See, Jacinto Estima, Jamal Jokar Arsanjani, Flavio Lupia, Marco Minghini, Giles Foody, and Steffen Fritz, "Investigating the feasibility of geo-tagged photographs as sources of land cover input data," *ISPRS International Journal of Geo-Information*, vol. 5, no. 5, pp. 64, 2016.
- [8] Laura Martínez-Sánchez, Linda See, Momchil Yordanov, Astrid Verhegghen, Neijia Elvekjær, Davide Muraro, Raphaël d'Andrimont, and Marijn van der Velde, "Automatic classification of land cover from LUCAS in-situ landscape photos using semantic segmentation and a random forest model," *Environmental Modelling & Software*, vol. 172, pp. 105931, 2024.
- [9] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski, "DINOv3," 2025.
- [10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmlR, 2021, pp. 8748–8763.
- [11] Pallavi Jain, Dino Ienco, Roberto Interdonato, Tristan Berchoux, and Diego Marcos, "SenCLIP: Enhancing zero-shot land-use mapping for Sentinel-2 with ground-level prompting," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2025, pp. 5656–5665.
- [12] Surbhi Sharma, Rocco Sedona, Morris Riedel, Gabriele Cavallaro, and Claudia Paris, "Sen4Map: Advancing mapping with Sentinel-2 by providing detailed semantic descriptions and customizable land-use and land-cover data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 13893–13907, 2024.
- [13] Giulio Weikmann, Claudia Paris, and Lorenzo Bruzzone, "Time-sen2crop: A million labeled samples dataset of sentinel 2 image time series for crop-type classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4699–4708, 2021.
- [14] Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Sheldhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli, "Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data," 2025.
- [15] Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C. Lisaus, Markus Immitzer, Toby Jackson, James Ball, David A. Coomes, Anil Madhavapeddy, Andrew Blake, and Srinivasan Keshav, "Tesser: Temporal embeddings of surface spectra for earth representation and analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026, pp. 34818–34831.

References

- [1] Herbert J. Kramer. *Earth Observation Remote Sensing: Survey of Missions and Sensors*. 1st ed. Berlin, Heidelberg: Springer, 1992. ISBN: 978-3-662-00520-0. DOI: 10.1007/978-3-662-00520-0.
- [2] James B Campbell and Randolph H Wynne. *Introduction to Remote Sensing*. 5th ed. Guilford Press, 2011.
- [3] Mona Morsy, Erik Borg, and Peter Dietrich. “Editorial for Special Issue: “In Situ Data in the Interplay of Remote Sensing”.” In: *Remote Sensing* 15.8 (2023). ISSN: 2072-4292. DOI: 10.3390/rs15082056. URL: <https://www.mdpi.com/2072-4292/15/8/2056>.
- [4] Linda See et al. “Supporting Earth-Observation Calibration and Validation: A new generation of tools for crowdsourcing and citizen science.” In: *IEEE Geoscience and Remote Sensing Magazine* 4.3 (2016), pp. 38–50. DOI: 10.1109/MGRS.2015.2498840.
- [5] Doreen S. Boyd et al. “Citizen science for Earth Observation (Citizens4EO): understanding current use in the UK.” In: *International Journal of Remote Sensing* 43.8 (2022), pp. 2965–2985. DOI: 10.1080/01431161.2022.2076574. eprint: <https://doi.org/10.1080/01431161.2022.2076574>. URL: <https://doi.org/10.1080/01431161.2022.2076574>.
- [6] Claudia Paris, Lorenzo Bruzzone, and Diego Fernández-Prieto. “A novel approach to the unsupervised update of land-cover maps by classification of time series of multispectral images.” In: *IEEE Transactions on Geoscience and Remote Sensing* 57.7 (2019), pp. 4259–4277.
- [7] Eman A. Alshari and Bharti W. Gawali. “Development of classification system for LULC using remote sensing and GIS.” In: *Global Transitions Proceedings* 2.1 (2021). 1st International Conference on Advances in Information, Computing and Trends in Data Engineering (AICDE - 2020), pp. 8–17. ISSN: 2666-285X. DOI: <https://doi.org/10.1016/j.gltp.2021.01.002>. URL: <https://www.sciencedirect.com/science/article/pii/S2666285X21000029>.
- [8] Srishti Gaur and Rajendra Singh. “A Comprehensive Review on Land Use/Land Cover (LULC) Change Modeling for Urban Development: Current Status and Future Prospects.” In: *Sustainability* 15.2 (2023). ISSN: 2071-1050. DOI: 10.3

- 390/su15020903. URL: <https://www.mdpi.com/2071-1050/15/2/903>.
- [9] Yanzhao Wang et al. “A review of regional and Global scale Land Use/Land Cover (LULC) mapping products generated from satellite remote sensing.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 206 (2023), pp. 311–334. ISSN: 0924-2716. DOI: <https://doi.org/10.1016/j.isprsjprs.2023.11.014>. URL: <https://www.sciencedirect.com/science/article/pii/S0924271623003209>.
- [10] United Nations. *Sustainable Development Goal 11: Make Cities Inclusive, Safe, Resilient and Sustainable*. <https://www.un.org/sustainabledevelopment/cities/>. Accessed: Dec. 4, 2025. 2025.
- [11] European Space Agency (ESA). *Copernicus Sentinel-2 MSI Level-2A BOA Reflectance Product, Collection 1*. https://doi.org/10.5270/S2_znk9xsj. 2021. DOI: 10.5270/S2_znk9xsj.
- [12] Raphaël d’Andrimont et al. “Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union.” In: *Scientific data* 7.1 (2020), p. 352.
- [13] Gerhard Neuhold et al. “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes.” In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 5000–5009. DOI: 10.1109/ICCV.2017.534.
- [14] Jorge Rocha et al. “Land use/cover maps by RS and ancillary data integration in a GIS environment.” In: Jan. 2006, pp. 487–494. ISBN: ISBN 90-5966-042-0.
- [15] Jiayi Li, Xin Huang, and Jianya Gong. “Deep neural network for remote-sensing image interpretation: status and perspectives.” In: *National Science Review* 6 (Nov. 2019), pp. 1082–1086. DOI: 10.1093/nsr/nwz058.
- [16] Claudio Persello et al. “Deep Learning and Earth Observation to Support the Sustainable Development Goals: Current approaches, open challenges, and future opportunities.” In: *IEEE Geoscience and Remote Sensing Magazine* 10.2 (2022), pp. 172–200. DOI: 10.1109/MGRS.2021.3136100.
- [17] Shengyu Zhao et al. “Land Use and Land Cover Classification Meets Deep Learning: A Review.” In: *Sensors* 23 (Nov. 2023), p. 8966. DOI: 10.3390/s23218966.
- [18] Johannes Jakubik et al. “Foundation Models for Generalist Geospatial Artificial Intelligence.” In: *Preprint Available on arxiv:2310.18660* (Oct. 2023).
- [19] Matthias Weigand et al. “Spatial and semantic effects of LUCAS samples on fully automated land use/land cover classification in high-resolution Sentinel-2 data.” In: *Int. J. Appl. Earth Obs. Geoinformation* 88 (2020), p. 102065. URL: <https://api.semanticscholar.org/CorpusID:211830479>.
- [20] Gianpaolo Balsamo et al. “Satellite and In Situ Observations for Advancing Global Earth Surface Modelling: A Review.” In: *Remote Sensing* 10.12 (2018). ISSN: 2072-4292. DOI: 10.3390/rs10122038. URL: <https://www.mdpi.com/2072-4292/10/12/2038>.

- [21] Corina Vaduva, Florin A. Georgescu, and Mihai Datcu. “Understanding Heterogeneous EO Datasets: A Framework for Semantic Representations.” In: *IEEE Access* 6 (2018), pp. 11184–11202. DOI: 10.1109/ACCESS.2018.2801032.
- [22] Mauro Dalla Mura et al. “Challenges and Opportunities of Multimodality and Data Fusion in Remote Sensing.” In: *Proceedings of the IEEE* 103 (Sept. 2014), pp. 1585–1601. DOI: 10.1109/JPROC.2015.2462751.
- [23] Wei Xia et al. “Tackling Data Quality Challenges in Remote Sensing: Solutions for Reliable Urban Heat Island Analysis.” In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18 (2025), pp. 20117–20128. DOI: 10.1109/JSTARS.2025.3593925.
- [24] Abdallah Yussuf Ali Abdelmajeed and Radosław Juszczak. “Challenges and Limitations of Remote Sensing Applications in Northern Peatlands: Present and Future Prospects.” In: *Remote Sensing* 16.3 (2024). ISSN: 2072-4292. DOI: 10.3390/rs16030591. URL: <https://www.mdpi.com/2072-4292/16/3/591>.
- [25] Andrew Mellor et al. “Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 105 (2015), pp. 155–168. ISSN: 0924-2716. DOI: <https://doi.org/10.1016/j.isprsjprs.2015.03.014>. URL: <https://www.sciencedirect.com/science/article/pii/S0924271615000945>.
- [26] Daniel Moraes, Manuel L. Campagnolo, and Mário Caetano. “Training data in satellite image classification for land cover mapping: a review.” In: *European Journal of Remote Sensing* 57.1 (2024), p. 2341414. DOI: 10.1080/22797254.2024.2341414. eprint: <https://doi.org/10.1080/22797254.2024.2341414>. URL: <https://doi.org/10.1080/22797254.2024.2341414>.
- [27] Iwona Podsiadlo, Claudia Paris, and Lorenzo Bruzzone. “An Approach Based on Low Resolution Land-Cover-Maps and Domain Adaptation to Define Representative Training Sets at Large Scale.” In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. 2021, pp. 313–316. DOI: 10.1109/IGARSS47720.2021.9553498.
- [28] Andy Stock. “Spatiotemporal distribution of labeled data can bias the validation and selection of supervised learning algorithms: A marine remote sensing example.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 187 (2022), pp. 46–60. ISSN: 0924-2716. DOI: <https://doi.org/10.1016/j.isprsjprs.2022.02.023>. URL: <https://www.sciencedirect.com/science/article/pii/S0924271622000648>.
- [29] Copernicus. *Sentinel-2*. <https://sentinels.copernicus.eu/web/sentinel/copernicus/sentinel-2>. Accessed: 2026-06-28. 2026.
- [30] Carole Dulong et al. “Google Street View: Capturing the World at Street Level.” In: *Computer* 43 (June 2010), pp. 32–38. DOI: 10.1109/MC.2010.170.

- [31] Eurostat. *LUCAS: Land Use/Cover Area Frame Survey*. European statistical survey for land use and land cover reference data. 2018. URL: <https://ec.europa.eu/eurostat/web/lucas>.
- [32] Antonia Ludwig, Hannes Feilhauer, and Daniel Doktor. “Exploring Sentinel-2-Based Spectral Variability for Enhancing Grassland Diversity Assessments Across Germany.” In: *Applied Vegetation Science* 28.3 (2025). e70030 AVS-RES-03499.R1, e70030. DOI: <https://doi.org/10.1111/avsc.70030>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/avsc.70030>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/avsc.70030>.
- [33] J. M. Barrios et al. “Landscape heterogeneity around flux measurement stations investigated through Sentinel-2 and PROBA-V satellite imagery.” In: *Remote Sensing for Agriculture, Ecosystems, and Hydrology XX*. Ed. by Christopher M. U. Neale and Antonino Maltese. Vol. 10783. International Society for Optics and Photonics. SPIE, 2018, p. 107830D. DOI: 10.1117/12.2325620. URL: <https://doi.org/10.1117/12.2325620>.
- [34] Sabrina P. L. P. Correa et al. “Rethinking reference data quality: the role of mixed pixels in remote sensing classification.” In: *International Journal of Remote Sensing* 47.12 (2026), pp. 5194–5218. DOI: 10.1080/01431161.2026.2664863. eprint: <https://doi.org/10.1080/01431161.2026.2664863>. URL: <https://doi.org/10.1080/01431161.2026.2664863>.
- [35] Gayathri M Pillai and SK. Katiyar. “Addressing mixed pixel challenges in crop mapping: An integrated remote sensing and machine learning framework.” In: *Franklin Open* 14 (2026), p. 100516. ISSN: 2773-1863. DOI: <https://doi.org/10.1016/j.fraope.2026.100516>. URL: <https://www.sciencedirect.com/science/article/pii/S2773186326000320>.
- [36] Xiang Chen et al. “The mixed pixel effect in land surface phenology: A simulation study.” In: *Remote Sensing of Environment* 211 (2018), pp. 338–344. ISSN: 0034-4257. DOI: <https://doi.org/10.1016/j.rse.2018.04.030>. URL: <https://www.sciencedirect.com/science/article/pii/S0034425718301871>.
- [37] Jianyu Gu and Russell Congalton. “Assessing the Impact of Mixed Pixel Proportion Training Data on SVM-Based Remote Sensing Classification: A Simulated Study.” In: *Remote Sensing* 17 (Apr. 2025), p. 1274. DOI: 10.3390/rs17071274.
- [38] M PETROU. “MIXED PIXEL CLASSIFICATION: AN OVERVIEW.” In: *Information Processing for Remote Sensing*, pp. 69–83. DOI: 10.1142/9789812815705_0004. eprint: https://www.worldscientific.com/doi/pdf/10.1142/9789812815705_0004. URL: https://www.worldscientific.com/doi/abs/10.1142/9789812815705_0004.

-
- [39] Julien Radoux et al. “Sentinel-2’s Potential for Sub-Pixel Landscape Feature Detection.” In: *Remote Sensing* 8.6 (2016). ISSN: 2072-4292. DOI: 10.3390/rs8060488. URL: <https://www.mdpi.com/2072-4292/8/6/488>.
 - [40] R. Milla, M.E. Schaepman, and J.G.P.W. Clevers. “Possibilities of MERIS for sub-pixel regional land cover mapping.” In: *Proceedings of SPIE - The International Society for Optical Engineering* (Oct. 2005). DOI: 10.1117/12.627670.
 - [41] Nikolaos Bountos et al. “Sub-pixel classification of anthropogenic features using deep- learning on Sentinel-2 data.” In: May 2020. DOI: 10.5194/egusphere-egu2020-1242.
 - [42] Jiaxin Li Bat Aug et al. “Deep learning in multimodal remote sensing data fusion: A comprehensive review.” In: *International Journal of Applied Earth Observation and Geoinformation* 112 (July 2022), p. 102926. DOI: 10.1016/j.jag.2022.102926.
 - [43] Frane Gilić, Mateo Gašparović, and Martina Baučić. “Current state and challenges in producing large-scale land cover maps: review based on recent land cover products.” In: *Geocarto International* 38.1 (2023), p. 2242693. DOI: 10.1080/10106049.2023.2242693. eprint: <https://doi.org/10.1080/10106049.2023.2242693>. URL: <https://doi.org/10.1080/10106049.2023.2242693>.
 - [44] Jülich Supercomputing Centre. “JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Jülich Supercomputing Centre.” In: *Journal of large-scale research facilities JLSRF* 7.A183 (2021). DOI: 10.17815/jlsrf-7-183.
 - [45] Forschungszentrum Jülich. *Exascale Computing*. 2025. URL: <https://www.fz-juelich.de/en/research/research-fields/information/supercomputing/exascale> (visited on 08/01/2026).
 - [46] T. Breuer. *JUPITER and an overview of its support infrastructure*. Invited Talk, 2025 High-Performance Computing Knowledge Portal (HPCKP). Barcelona, Spain, June 2025. DOI: 10.34734/FZJ-2026-00067.
 - [47] Matías Mendieta et al. “Towards Geospatial Foundation Models via Continual Pretraining.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 16806–16816.
 - [48] Krzysztof Janowicz et al. “GeoFM: how will geo-foundation models reshape spatial data science and GeoAI?” In: *International Journal of Geographical Information Science* 39.9 (2025), pp. 1849–1865. DOI: 10.1080/13658816.2025.2543038. eprint: <https://doi.org/10.1080/13658816.2025.2543038>. URL: <https://doi.org/10.1080/13658816.2025.2543038>.
 - [49] Yawogan Jean Eudes Gbodjo et al. *Fine grained classification for multi-source land cover mapping*. 2020. arXiv: 2004.01963 [cs.CV]. URL: <https://arxiv.org/abs/2004.01963>.

- [50] Darius Phiri et al. "Sentinel-2 Data for Land Cover/Use Mapping: A Review." In: *Remote Sensing* 12 (July 2020), p. 2291. DOI: 10.3390/rs12142291.
- [51] Runxiang Li et al. "Scale Effect of Land Cover Classification from Multi-Resolution Satellite Remote Sensing Data." In: *Sensors* 23.13 (2023), p. 6136. DOI: 10.3390/s23136136.
- [52] Peng Gong et al. "Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017." In: *Science Bulletin* 64.6 (2019), pp. 370–373. ISSN: 2095-9273. DOI: <https://doi.org/10.1016/j.scib.2019.03.002>. URL: <https://www.sciencedirect.com/science/article/pii/S2095927319301380>.
- [53] Rahul Ghosh, Xiaowei Jia, and Vipin Kumar. *Land Cover Mapping in Limited Labels Scenario: A Survey*. 2021. arXiv: 2103.02429 [cs.CV].
- [54] Prasad S. Thenkabail, ed. *Remote Sensing Handbook, Volume IV: Forests, Biodiversity, Ecology, LULC, and Carbon*. 2nd. CRC Press, 2024. ISBN: 9781032891040. DOI: 10.1201/9781003541172. URL: <https://www.taylorfrancis.com/books/edit/10.1201/9781003541172/remote-sensing-handbook-volume-iv-prasad-thenkabail>.
- [55] Jeffrey G. Masek et al. "Landsat 9: Empowering open science and applications through continuity." In: *Remote Sensing of Environment* 248 (2020), p. 111968. ISSN: 0034-4257. DOI: <https://doi.org/10.1016/j.rse.2020.111968>. URL: <https://www.sciencedirect.com/science/article/pii/S0034425720303382>.
- [56] Volker Radeloff et al. "Need and vision for global medium-resolution Landsat and Sentinel-2 data products." In: *Remote Sensing of Environment* 300 (Jan. 2024), p. 113918. DOI: 10.1016/j.rse.2023.113918.
- [57] Dávid D.Kovács et al. "Copernicus Data Space Ecosystem establishes public cloud processing for earth observation data." In: *Scientific Data* 13 (Feb. 2026). DOI: 10.1038/s41597-026-06765-8.
- [58] Annu Kumari and S. Karthikeyan. "Sentinel-2 Data for Land Use/Land Cover Mapping: A Meta-analysis and Review." In: *SN Computer Science* 4 (Oct. 2023). DOI: 10.1007/s42979-023-02214-0.
- [59] European Space Agency (ESA). *The Sentinel Satellites*. <https://www.copernicus.eu/en/media/images/sentinel-satellites>. Copernicus Programme promotional image. Accessed: 2026-07-15. 2013.
- [60] Fernando Orduna-Cabrera et al. "Investigating the Use of Street-Level Imagery and Deep Learning to Produce In-Situ Crop Type Information." In: *Geographies* 3.3 (2023), pp. 563–573. ISSN: 2673-7086. DOI: 10.3390/geographies3030029. URL: <https://www.mdpi.com/2673-7086/3/3/29>.
- [61] Raphael d'Andrimont et al. "Crowdsourced Street-Level Imagery as a Potential Source of In-Situ Data for Crop Monitoring." In: *Land* 7 (Oct. 2018), p. 127. DOI: 10.3390/land7040127.

-
- [62] Clovis Thouvenin-Masson et al. “Satellite and In Situ Sampling Mismatches: Consequences for the Estimation of Satellite Sea Surface Salinity Uncertainties.” In: *Remote Sensing* 14.8 (2022). ISSN: 2072-4292. DOI: 10.3390/rs14081878. URL: <https://www.mdpi.com/2072-4292/14/8/1878>.
 - [63] Charlotte Pelletier et al. “Effect of training class label noise on classification performances for land cover mapping with satellite image time series.” In: *Remote Sensing* 9.2 (2017), p. 173.
 - [64] Eurostat. *LUCAS 2018: Technical reference document C-1: Instructions for surveyors*. Tech. rep. European Commission, 2018. URL: <https://ec.europa.eu/eurostat/documents/205002/8072634/LUCAS-2018-C1-Instructions.pdf>.
 - [65] Eurostat. *Redesign sample for Land Use/Cover Area frame Survey (LUCAS) 2018*. Statistical Working Paper KS-TC-18-006. European Commission, 2018. DOI: 10.2785/132365.
 - [66] Raphael d’Andrimont et al. “LUCAS cover photos 2006-2018 over the EU: 874 646 spatially distributed geo-tagged close-up photos with land cover and plant species label.” In: *Earth System Science Data* 14 (Sept. 2022), pp. 4463–4472.
 - [67] Eurostat. *Land cover and land use, landscape (LUCAS) (lan)*. 2024. URL: https://ec.europa.eu/eurostat/cache/metadata/en/lan_esms.htm (visited on 08/01/2026).
 - [68] Eurostat. *LUCAS 2018 Technical Reference Document C3: Classification (Land Cover & Land Use)*. Tech. rep. European Commission, 2018. URL: <https://ec.europa.eu/eurostat/documents/205002/8072634/LUCAS2018-C3-Classification.pdf>.
 - [69] Laura Martinez-Sanchez et al. “Automatic classification of land cover from LUCAS in-situ landscape photos using semantic segmentation and a Random Forest model.” In: *Environmental Modelling & Software* 172 (Dec. 2023), p. 105931. DOI: 10.1016/j.envsoft.2023.105931.
 - [70] Matthias Weigand et al. “Spatial and semantic effects of LUCAS samples on fully automated land use/land cover classification in high-resolution Sentinel-2 data.” In: *International Journal of Applied Earth Observation and Geoinformation* 88 (2020), p. 102065. ISSN: 1569-8432. DOI: <https://doi.org/10.1016/j.jag.2020.102065>. URL: <https://www.sciencedirect.com/science/article/pii/S0303243419307317>.
 - [71] Christos G. Karydas et al. “Use of LUCAS LC Point Database for Validating Country-Scale Land Cover Maps.” In: *Remote Sensing* 7.5 (2015), pp. 5012–5041. ISSN: 2072-4292. DOI: 10.3390/rs70505012. URL: <https://www.mdpi.com/2072-4292/7/5/5012>.
 - [72] Yuan Gao et al. “Consistency Analysis and Accuracy Assessment of Three Global 30-m Land-Cover Products over the European Union using the LUCAS Dataset.” In: *Remote Sensing* 12.21 (2020). ISSN: 2072-4292. DOI: 10.3390/rs12213479. URL: <https://www.mdpi.com/2072-4292/12/21/3479>.

- [73] Beata Hejmanowska and Piotr Kramarczyk. “Assessing Land Cover Changes Using the LUCAS Database and Sentinel Imagery: A Comparative Analysis of Accuracy Metrics.” In: *Applied Sciences* (2024). URL: <https://api.semanticscholar.org/CorpusID:275198815>.
- [74] OpenStreetMap contributors. *Planet dump retrieved from https://planet.osm.org*. <https://www.openstreetmap.org>. 2025.
- [75] Hao Li et al. “Leveraging OpenStreetMap and Multimodal Remote Sensing Data with Joint Deep Learning for Wastewater Treatment Plants Detection.” In: *International Journal of Applied Earth Observation and Geoinformation* 110 (May 2022). DOI: 10.1016/j.jag.2022.102804.
- [76] John E. Vargas-Munoz et al. “OpenStreetMap: Challenges and Opportunities in Machine Learning and Remote Sensing.” In: *IEEE Geoscience and Remote Sensing Magazine* 9.1 (2021), pp. 184–199. DOI: 10.1109/MGRS.2020.2994107.
- [77] J. Patriarca et al. “Automatic conversion of OSM data into LULC maps: comparing FOSS4G based approaches towards an enhanced performance.” In: *Open Geospatial Data, Software and Standards* 4 (Oct. 2019). DOI: 10.1186/s40965-019-0070-2.
- [78] Joaquim Patriarca, Cidália Fonte, and Alberto Cardoso. “Land Use Mapping with OpenStreetMap: A Machine Learning approach.” In: *2025 7th Experiment@ International Conference (exp.at'25)*. 2025, pp. 109–114. DOI: 10.1109/exp.at2565440.2025.11348326.
- [79] D. Zanaga et al. *ESA WorldCover 10 m 2020 v100*. Dataset on Zenodo. European Space Agency (ESA) WorldCover project. 2021. DOI: 10.5281/zenodo.5571936. URL: <https://zenodo.org/record/5571936>.
- [80] Nicolas Dalezios. *Remote Sensing Applications in Environmental and Earth System Sciences*. Mar. 2021. ISBN: 9781351680660. DOI: 10.1201/9781315166667.
- [81] Ryan Nedd et al. “A Synthesis of Land Use/Land Cover Studies: Definitions, Classification Systems, Meta-Studies, Challenges and Knowledge Gaps on a Global Landscape.” In: *Land* 10.9 (2021). ISSN: 2073-445X. DOI: 10.3390/land10090994. URL: <https://www.mdpi.com/2073-445X/10/9/994>.
- [82] Chuanrong Zhang and Xinba Li. “Land Use and Land Cover Mapping in the Era of Big Data.” In: *Land* 11 (Sept. 2022), p. 1692. DOI: 10.3390/land11101692.
- [83] M.A. Friedl et al. “Global land cover mapping from MODIS: Algorithms and early results.” In: *Remote Sensing of Environment* 83 (Nov. 2002), p. 287. DOI: 10.1016/S0034-4257(02)00078-0.

-
- [84] S. K. Maxwell, R. M. Hoffer, and P. L. Chapman. "AVHRR channel selection for land cover classification." In: *International Journal of Remote Sensing* 23.23 (2002), pp. 5061–5073. DOI: 10.1080/01431160210145588. eprint: <https://doi.org/10.1080/01431160210145588>. URL: <https://doi.org/10.1080/01431160210145588>.
 - [85] Peng Gong et al. "A new research paradigm for global land cover mapping." In: *Annals of GIS* 22.2 (2016), pp. 87–102. DOI: 10.1080/19475683.2016.1164247. eprint: <https://doi.org/10.1080/19475683.2016.1164247>. URL: <https://doi.org/10.1080/19475683.2016.1164247>.
 - [86] Michael A. Wulder et al. "The global Landsat archive: Status, consolidation, and direction." In: *Remote Sensing of Environment* 185 (2016). Landsat 8 Science Results, pp. 271–283. ISSN: 0034-4257. DOI: <https://doi.org/10.1016/j.rse.2015.11.032>. URL: <https://www.sciencedirect.com/science/article/pii/S0034425715302194>.
 - [87] Peng Gong et al. "Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data." In: *International Journal of Remote Sensing* 34.7 (2013), pp. 2607–2654. DOI: 10.1080/01431161.2012.748992. eprint: <https://doi.org/10.1080/01431161.2012.748992>. URL: <https://doi.org/10.1080/01431161.2012.748992>.
 - [88] Zander S. Venter et al. "Global 10 m Land Use Land Cover Datasets: A Comparison of Dynamic World, World Cover and Esri Land Cover." In: *Remote Sensing* 14.16 (2022). ISSN: 2072-4292. DOI: 10.3390/rs14164101. URL: <https://www.mdpi.com/2072-4292/14/16/4101>.
 - [89] R. Van De Kerchove et al. "ESA WorldCover: Global land cover mapping at 10 m resolution for 2020 based on Sentinel-1 and 2 data." In: *AGU Fall Meeting Abstracts*. Vol. 2021. 2021, GC45I–0915.
 - [90] Impact Observatory, Esri, and Microsoft. *ESRI 2018 Global Land Use/Land Cover (10m) – Sentinel-2 based*. ArcGIS Living Atlas of the World. Global 10m land cover map derived from ESA Sentinel-2 imagery. 2018. URL: <https://www.esri.com/en-us/arcgis/products/arcgis-living-atlas/land-cover>.
 - [91] Christopher F. Brown et al. "Dynamic World, Near real-time global 10 m land use land cover mapping." In: *Scientific Data* 9 (2022), p. 251. DOI: 10.1038/s41597-022-01307-4.
 - [92] Marcel Buchhorn et al. "Copernicus Global Land Cover Layers—Collection 2." In: *Remote Sensing* 12.6 (2020). ISSN: 2072-4292. DOI: 10.3390/rs12061044. URL: <https://www.mdpi.com/2072-4292/12/6/1044>.
 - [93] Jun Chen et al. "Collaborative validation of GlobeLand30: Methodology and practices." In: *Geo-spatial Information Science* 24.1 (2021), pp. 134–144. DOI: 10.1080/10095020.2021.1894906. eprint: <https://doi.org/10.1080/10095020.2021.1894906>. URL: <https://doi.org/10.1080/10095020.2021.1894906>.

- [94] Le Yu et al. "FROM-GLC Plus: toward near real-time and multi-resolution land cover mapping." In: *GIScience & Remote Sensing* 59.1 (2022), pp. 1026–1047. DOI: 10.1080/15481603.2022.2096184. eprint: <https://doi.org/10.1080/15481603.2022.2096184>. URL: <https://doi.org/10.1080/15481603.2022.2096184>.
- [95] T. R. Loveland and A. S. Belward. "The IGBP-DIS global 1km land cover data set, DISCover: First results." In: *International Journal of Remote Sensing* 18.15 (1997), pp. 3289–3295. DOI: 10.1080/014311697217099. eprint: <https://doi.org/10.1080/014311697217099>. URL: <https://doi.org/10.1080/014311697217099>.
- [96] M. C. Hansen and B. Reed. "A comparison of the IGBP DISCover and University of Maryland 1 km global land cover products." In: *International Journal of Remote Sensing* 21.6-7 (2000), pp. 1365–1373. DOI: 10.1080/014311600210218. eprint: <https://doi.org/10.1080/014311600210218>. URL: <https://doi.org/10.1080/014311600210218>.
- [97] E. Bartholomé and A. S. Belward. "GLC2000: a new approach to global land cover mapping from Earth observation data." In: *International Journal of Remote Sensing* 26.9 (2005), pp. 1959–1977. DOI: 10.1080/01431160412331291297. eprint: <https://doi.org/10.1080/01431160412331291297>. URL: <https://doi.org/10.1080/01431160412331291297>.
- [98] M. Friedl and D. Sulla-Menashe. *MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006*. [Dataset]. 2019. DOI: 10.5067/MODIS/MCD12Q1.006. URL: <https://doi.org/10.5067/MODIS/MCD12Q1.006>.
- [99] Olivier Arino et al. "GlobCover: ESA service for global land cover from MERIS." In: *2007 IEEE International Geoscience and Remote Sensing Symposium*. 2007, pp. 2412–2415. DOI: 10.1109/IGARSS.2007.4423328.
- [100] Surbhi Sharma et al. "Enhancing Land Cover Mapping: A Novel Automatic Approach To Improve Mixed Spectral Pixel Classification." In: *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*. 2024, pp. 1017–1020. DOI: 10.1109/IGARSS53475.2024.10641976.
- [101] Jonathan Smith et al. "Effect of Landscape Characteristics on Land-Cover Class Accuracy." In: *Remote Sensing of Environment* 84 (Mar. 2003), pp. 342–349. DOI: 10.1016/S0034-4257(02)00126-8.
- [102] Ruben Urraca, Christian Lanconelli, and Nadine Gobron. "Impact of the Spatio-Temporal Mismatch Between Satellite and In Situ Measurements on Validations of Surface Solar Radiation." In: *Journal of Geophysical Research: Atmospheres* 129.10 (2024). e2024JD041007. DOI: <https://doi.org/10.1029/2024JD041007>. eprint: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2024JD041007>. URL: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024JD041007>.

-
- [103] Fei Xu, Stien Heremans, and Ben Somers. “Urban land cover mapping with Sentinel-2: a spectro-spatio-temporal analysis.” In: *Urban Informatics 1* (Oct. 2022). DOI: 10.1007/s44212-022-00008-y.
 - [104] A. L. Choodarathnakara, Ashok Kumar, and Gopal Patil. “Mixed Pixels : A Challenge in Remote Sensing Data Classification for Improving Performance.” In: 2012. URL: <https://api.semanticscholar.org/CorpusID:212444889>.
 - [105] Raúl Zurita-Milla et al. “Multitemporal Unmixing of Medium-Spatial-Resolution Satellite Images: A Case Study Using MERIS Images for Land-Cover Mapping.” In: *IEEE Transactions on Geoscience and Remote Sensing* 49.11 (2011), pp. 4308–4317. DOI: 10.1109/TGRS.2011.2158320.
 - [106] Xiaodong Li et al. “Using a sub-pixel mapping model to improve the accuracy of landscape pattern indices.” In: *Ecological Indicators* 11.5 (2011), pp. 1160–1170. ISSN: 1470-160X. DOI: <https://doi.org/10.1016/j.ecolin.2010.12.016>. URL: <https://www.sciencedirect.com/science/article/pii/S1470160X10002219>.
 - [107] Dainius Masiliūnas et al. “Global land characterisation using land cover fractions at 100 m resolution.” In: *Remote Sensing of Environment* 259 (June 2021), p. 112409. DOI: 10.1016/j.rse.2021.112409.
 - [108] John B. Adams et al. “Classification of multispectral images based on fractions of endmembers: Application to land-cover change in the Brazilian Amazon.” In: *Remote Sensing of Environment* 52.2 (1995), pp. 137–154. ISSN: 0034-4257. DOI: [https://doi.org/10.1016/0034-4257\(94\)00098-8](https://doi.org/10.1016/0034-4257(94)00098-8). URL: <https://www.sciencedirect.com/science/article/pii/0034425794000988>.
 - [109] José Rodríguez-Ortega et al. “Land use and land cover fraction estimation for Sentinel-2 RGB images: A new LULC mapping task.” In: *Remote Sensing Applications: Society and Environment* 39 (2025), p. 101626. ISSN: 2352-9385. DOI: <https://doi.org/10.1016/j.rsase.2025.101626>. URL: <https://www.sciencedirect.com/science/article/pii/S235293852500179X>.
 - [110] James Wickham et al. “Thematic accuracy of the NLCD 2001 land cover for the conterminous United States.” In: *Remote Sensing of Environment* 114 (June 2010), pp. 1286–1296. DOI: 10.1016/j.rse.2010.01.018.
 - [111] Arthur Elmes et al. “Accounting for Training Data Error in Machine Learning Applied to Earth Observations.” In: *Remote Sensing* 12.6 (2020). ISSN: 2072-4292. DOI: 10.3390/rs12061034. URL: <https://www.mdpi.com/2072-4292/12/6/1034>.
 - [112] Ronald Mcroberts et al. “The effects of imperfect reference data on remote sensing-assisted estimators of land cover class proportions.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 142 (Aug. 2018). DOI: 10.1016/j.isprsjprs.2018.06.002.

- [113] Archit Parnami and Minwoo Lee. *Learning from Few Examples: A Summary of Approaches to Few-Shot Learning*. 2022. arXiv: 2203.04291 [cs.LG]. URL: <https://arxiv.org/abs/2203.04291>.
- [114] Padmanabha Y C a, Viswanath Pulabaigari, and Eswara B. “Semi-supervised learning: a brief review.” In: *International Journal of Engineering Technology* 7 (Feb. 2018), p. 81. DOI: 10.14419/ijet.v7i11.8.9977.
- [115] Fuzhen Zhuang et al. “A Comprehensive Survey on Transfer Learning.” In: *Proceedings of the IEEE* 109.1 (2021), pp. 43–76. DOI: 10.1109/JPROC.2020.3004555.
- [116] Gencer Sumbul et al. “Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding.” In: July 2019, pp. 5901–5904. DOI: 10.1109/IGARSS.2019.8900532.
- [117] Patrick Helber et al. “EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification.” In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (Aug. 2017). DOI: 10.1109/JSTARS.2019.2918242.
- [118] Philipe Dias et al. “An Agenda for Multimodal Foundation Models for Earth Observation.” In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. 2023, pp. 1237–1240. DOI: 10.1109/IGARSS52108.2023.10282966.
- [119] Danfeng Hong et al. “Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 178 (Aug. 2021), pp. 68–80. DOI: 10.1016/j.isprsjprs.2021.05.011.
- [120] Andrei Stoian et al. “Land Cover Maps Production with High Resolution Satellite Image Time Series and Convolutional Neural Networks: Adaptations and Limits for Operational Systems.” In: *Remote Sensing* 11 (Aug. 2019), p. 1986. DOI: 10.3390/rs11171986.
- [121] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” In: *International Conference on Learning Representations (ICLR)*. 2021.
- [122] Fakhri Alam Khan. “Transforming Earth Observation: An Extensive Evaluation of Vision Transformers for Satellite Images-Based Land Cover Classification.” In: *Expert Systems* 42.7 (2025). e70082 EXSY-Mar-25-1268.R2, e70082. DOI: <https://doi.org/10.1111/exsy.70082>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.70082>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.70082>.

- [123] Yuhan Ji et al. “Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations.” In: *International Journal of Geographical Information Science* 39.9 (2025), pp. 1866–1903. ISSN: 1362-3087. DOI: 10.1080/13658816.2025.2511227. URL: <http://dx.doi.org/10.1080/13658816.2025.2511227>.
- [124] Siqi Lu et al. “Vision Foundation Models in Remote Sensing: A Survey.” In: *IEEE Geoscience and Remote Sensing Magazine* 13.3 (2025), pp. 190–215. DOI: 10.1109/MGRS.2025.3541952.
- [125] Xin Zhang, Ya’nan Zhou, and Jiancheng Luo. “Deep learning for processing and analysis of remote sensing big data: a technical review.” In: *Big Earth Data* 6.4 (2022), pp. 527–560. DOI: 10.1080/20964471.2021.1964879. eprint: <https://doi.org/10.1080/20964471.2021.1964879>. URL: <https://doi.org/10.1080/20964471.2021.1964879>.
- [126] DEEP Projects. *DEEP Projects: Developing a Novel HPC Architecture that Paves the Way for European Exascale Systems*. Accessed: 2026-07-15. 2026. URL: <https://deep-projects.eu/>.
- [127] Rocco Sedona et al. “Remote Sensing Big Data Classification with High Performance Distributed Deep Learning.” In: *Remote Sensing* 11 (Dec. 2019), p. 3056. DOI: 10.3390/rs11243056.
- [128] Nathan C. Frey et al. “Benchmarking Resource Usage for Efficient Distributed Deep Learning.” In: *2022 IEEE High Performance Extreme Computing Conference (HPEC)*. 2022, pp. 1–8. DOI: 10.1109/HPEC55821.2022.9926375.
- [129] Philipp Thörnig. “JURECA: Data Centric and Booster Modules implementing the Modular Supercomputing Architecture at Jülich Supercomputing Centre.” In: *Journal of large-scale research facilities JLSRF* 7 (Oct. 2021), A182. DOI: 10.17815/jlsrf-7-182.
- [130] Yosio Shimabukuro and Flávio Ponzoni. “The Linear Spectral Mixture Model: Linear Model and Applications.” In: Jan. 2019, pp. 23–41. ISBN: 978-3-030-02016-3. DOI: 10.1007/978-3-030-02017-0_4.
- [131] Dorian Krause and Philipp Thörnig. “JURECA: Modular supercomputer at Jülich Supercomputing Centre.” In: *Journal of large-scale research facilities JLSRF* 4 (July 2018), A132. DOI: 10.17815/jlsrf-4-121-1.
- [132] E. Erlingsson et al. “Modular supercomputing design supporting machine learning applications.” In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2018, pp. 0159–0163. DOI: 10.23919/MIPRO.2018.8400031.
- [133] Liang Tian et al. “End-to-End Process Orchestration of Earth Observation Data Workflows with Apache Airflow on High Performance Computing.” In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. 2023, pp. 711–714. DOI: 10.1109/IGARSS52108.2023.10283416.

- [134] Yan Ma et al. “Remote sensing big data computing: Challenges and opportunities.” In: *Future Generation Computer Systems* 51 (2015). Special Section: A Note on New Trends in Data-Aware Scheduling and Resource Provisioning in Modern HPC Systems, pp. 47–60. ISSN: 0167-739X. DOI: <https://doi.org/10.1016/j.future.2014.10.029>. URL: <https://www.sciencedirect.com/science/article/pii/S0167739X14002234>.
- [135] Linus Scheibenreif, Michael Mommert, and Damian Borth. “Parameter Efficient Self-Supervised Geospatial Domain Adaptation.” In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 27841–27851. DOI: 10.1109/CVPR52733.2024.02630.
- [136] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision.” In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763.
- [137] Dilip Dalei, Sangeeta Shrivastava, and Narayan Panigrahi. “Geographic Places to Semantic Spaces: Analysis of Geospatial Embeddings.” In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences X-5/W2-2025* (Dec. 2025), pp. 131–135. DOI: 10.5194/isprs-annals-X-5-W2-2025-131-2025.
- [138] Yezhen Cong et al. “SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery.” In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022.
- [139] Daniela Szwarzman et al. *Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications*. 2026. arXiv: 2412.02732 [cs.CV]. URL: <https://arxiv.org/abs/2412.02732>.
- [140] Gabriel Tseng et al. “Galileo: Learning Global & Local Features of Many Remote Sensing Modalities.” In: *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. Vol. 267. Proceedings of Machine Learning Research. PMLR, 2025, pp. 60280–60300.
- [141] Johannes Jakubik et al. “TerraMind: Large-Scale Generative Multimodality for Earth Observation.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025, pp. 7383–7394.
- [142] Christopher F. Brown et al. *AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data*. 2025. arXiv: 2507.22291 [cs.CV]. URL: <https://arxiv.org/abs/2507.22291>.
- [143] Lizhe Wang and Jining Yan. “Stewardship and analysis of big Earth observation data.” In: *Big Earth Data* 4.4 (2020), pp. 349–352. DOI: 10.1080/20964471.2020.1857055. eprint: <https://doi.org/10.1080/20964471.2020.1857055>. URL: <https://doi.org/10.1080/20964471.2020.1857055>.

-
- [144] The HDF Group. *Hierarchical Data Format, version 5*. <https://www.hdfgroup.org/HDF5/>. Source code available at <https://github.com/HDFGroup/hdf5>. 2025. DOI: 10.5281/zenodo.17808614. URL: <https://www.hdfgroup.org/HDF5/>.
 - [145] Mike Folk et al. “An overview of the HDF5 technology suite and its applications.” In: Mar. 2011, pp. 36–47. DOI: 10.1145/1966895.1966900.
 - [146] Huihuo Zheng et al. “HDF5 Cache VOL: Efficient and Scalable Parallel I/O through Caching Data on Node-local Storage.” In: *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. 2022, pp. 61–70. DOI: 10.1109/CCGrid54584.2022.00015.
 - [147] Sriniket Ambatipudi and Suren Byna. *A Comparison of HDF5, Zarr, and netCDF4 in Performing Common I/O Operations*. 2023. arXiv: 2207.09503 [cs.DC]. URL: <https://arxiv.org/abs/2207.09503>.
 - [148] Ian Seiferling et al. “Green streets Quantifying and mapping urban trees with street-level imagery and computer vision.” In: *Landscape and Urban Planning* 165 (Sept. 2017), pp. 93–101. DOI: 10.1016/j.landurbplan.2017.05.010.
 - [149] Yuan-Fong Su et al. “High-Resolution Flood Simulation in Urban Areas Through the Application of Remote Sensing and Crowdsourcing Technologies.” In: *Frontiers in Earth Science* 9 (Jan. 2022), pp. 1–11. DOI: 10.3389/feart.2021.756198.
 - [150] Fan Zhang et al. “Street-level imagery analytics and applications.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 199 (May 2023), pp. 195–196. DOI: 10.1016/j.isprsjprs.2023.04.005.
 - [151] Steffen Fritz et al. “Geo-Wiki: An online platform for improving global land cover.” In: *Environmental Modelling & Software* (May 2012). DOI: 10.1016/j.envsoft.2011.11.015.
 - [152] McCallum I et al. “Geo-Wiki.org: Harnessing the power of volunteers, the Internet and Google Earth to collect and validate global spatial information.” In: *Earthzine* (Feb. 2010).
 - [153] Claudia Paris et al. “Accuracy assessment of land-use-land-cover maps: the semantic gap between in situ and satellite data.” In: *Image and Signal Processing for Remote Sensing XXIX*. Ed. by Lorenzo Bruzzone and Francesca Bovolo. Vol. 12733. International Society for Optics and Photonics. SPIE, 2023, p. 127330M. DOI: 10.1117/12.2679433. URL: <https://doi.org/10.1117/12.2679433>.
 - [154] Guang Xu et al. “Automatic land cover classification of geo-tagged field photos by deep learning.” In: *Environmental Modelling & Software* 91 (2017), pp. 127–134. ISSN: 1364-8152. DOI: <https://doi.org/10.1016/j.envsoft.2017.02.004>. URL: <https://www.sciencedirect.com/science/article/pii/S1364815216303152>.

- [155] Hanfa Xing et al. “Exploring geo-tagged photos for land cover validation with deep learning.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 141 (July 2018), pp. 237–251. DOI: 10.1016/j.isprsjprs.2018.04.025.
- [156] Vyron Antoniou et al. “Investigating the Feasibility of Geo-Tagged Photographs as Sources of Land Cover Input Data.” In: *International Journal of Geo-Information (IJGI)* 5 (May 2016), p. 64. DOI: 10.3390/ijgi5050064.
- [157] Florian Bordes et al. *An Introduction to Vision-Language Modeling*. 2024. arXiv: 2405.17247 [cs.LG]. URL: <https://arxiv.org/abs/2405.17247>.
- [158] Ting Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations*. 2020. arXiv: 2002.05709 [cs.LG]. URL: <https://arxiv.org/abs/2002.05709>.
- [159] Pallavi Jain et al. “SenCLIP: Enhancing Zero-Shot Land-Use Mapping for Sentinel-2 with Ground-Level Prompting.” In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Feb. 2025, pp. 5656–5665. DOI: 10.1109/wacv61041.2025.00552. URL: <http://dx.doi.org/10.1109/WACV61041.2025.00552>.
- [160] Fan Liu et al. “RemoteCLIP: A Vision Language Foundation Model for Remote Sensing.” In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024).
- [161] Pallavi Jain et al. “TimeSenCLIP: A time series vision–language model for remote sensing.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 236 (2026), pp. 99–119. ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2026.03.043. URL: <http://dx.doi.org/10.1016/j.isprsjprs.2026.03.043>.
- [162] Mathilde Caron et al. “Emerging Properties in Self-Supervised Vision Transformers.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 9650–9660.
- [163] Oriane Siméoni et al. *DINOv3*. 2025. arXiv: 2508.10104 [cs.CV]. URL: <https://arxiv.org/abs/2508.10104>.
- [164] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision.” In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763.
- [165] Angelos Zavras et al. “Mind the modality gap: Towards a remote sensing vision-language model via cross-modal alignment.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 228 (2025), pp. 270–287. ISSN: 0924-2716. DOI: <https://doi.org/10.1016/j.isprsjprs.2025.06.019>. URL: <https://www.sciencedirect.com/science/article/pii/S092427162500245X>.

- [166] Zilun Zhang et al. “RS5M and GeoRSCLIP: A Large-Scale Vision- Language Dataset and a Large Vision-Language Model for Remote Sensing.” In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), pp. 1–23. ISSN: 1558-0644. DOI: 10.1109/tgrs.2024.3449154. URL: <http://dx.doi.org/10.1109/TGRS.2024.3449154>.
- [167] Linlin Liang et al. “Transferring CLIP for visual grounding in remote sensing images.” In: *International Journal of Digital Earth* 18.1 (2025), p. 2512059. DOI: 10.1080/17538947.2025.2512059. eprint: <https://doi.org/10.1080/17538947.2025.2512059>. URL: <https://doi.org/10.1080/17538947.2025.2512059>.
- [168] Surbhi Sharma et al. “Sen4Map: Advancing Mapping With Sentinel-2 by Providing Detailed Semantic Descriptions and Customizable Land-Use and Land-Cover Data.” In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17 (Jan. 2024), pp. 13893–13907. DOI: 10.1109/JSTARS.2024.3435081.
- [169] David García Álvarez and Sabina Nanu. “Land Use Cover Datasets: A Review.” In: June 2022, pp. 47–66. ISBN: 978-3-030-90997-0. DOI: 10.1007/978-3-030-90998-7_4.
- [170] Patrick Broxton et al. “A MODIS-Based Global 1-km Maximum Green Vegetation Fraction Dataset.” In: *Journal of Applied Meteorology and Climatology* 53 (Aug. 2014), pp. 1996–2004. DOI: 10.1175/JAMC-D-13-0356.1.
- [171] Xubin Zeng et al. “Derivation and Evaluation of Global 1-km Fractional Vegetation Cover Data for Land Modeling.” In: *Journal of Applied Meteorology* 39.6 (2000), pp. 826–839. DOI: 10.1175/1520-0450(2000)039<0826:DAEOGK>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/apme/39/6/1520-0450_2000_039_0826_daeogk_2.0.co_2.xml.
- [172] W. He et al. “Global urban fractional changes at a 1 km resolution throughout 2100 under eight scenarios of Shared Socioeconomic Pathways (SSPs) and Representative Concentration Pathways (RCPs).” In: *Earth System Science Data* 15.8 (2023), pp. 3623–3639. DOI: 10.5194/essd-15-3623-2023. URL: <https://essd.copernicus.org/articles/15/3623/2023/>.
- [173] Franz Schug et al. *Land cover fraction map of Austria at 10m spatial resolution based on Sentinel-1 and Sentinel-2 spectral temporal metrics*. dataset. 2020. DOI: 10.1594/PANGAEA.923037. URL: <https://doi.org/10.1594/PANGAEA.923037>.
- [174] Franz Schug et al. *Land cover fraction map of Germany at 10m spatial resolution based on Sentinel-1 and Sentinel-2 spectral temporal metrics*. dataset. 2020. DOI: 10.1594/PANGAEA.920894. URL: <https://doi.org/10.1594/PANGAEA.920894>.
- [175] August Slomp, Dainius Masiliūnas, and Nandin-Erdene Tsendbazar. “Improving global land cover fraction change mapping using temporal deep learning.” In: *International Journal of Applied Earth Observation and Geoinformation* 144 (Nov. 2025), p. 104927. DOI: 10.1016/j.jag.2025.104927.

- [176] Ming Hao et al. "A Subpixel Mapping Method for Urban Land Use by Reducing Shadow Effects." In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), pp. 2163–2177. DOI: 10.1109/JSTARS.2023.3243895.
- [177] Ramandeep Kaur M. Malhi, Prashant K. Srivastava, and G. Sandhya Kiran. "6 - Identification of functionally distinct plants using linear spectral mixture analysis." In: *Hyperspectral Remote Sensing*. Ed. by Prem Chandra Pandey et al. Earth Observation. Elsevier, 2020, pp. 95–106. ISBN: 978-0-08-102894-0. DOI: <https://doi.org/10.1016/B978-0-08-102894-0.00008-5>. URL: <https://www.sciencedirect.com/science/article/pii/B9780081028940000085>.
- [178] G. M. FOODY and D. P. COX. "Sub-pixel land cover composition estimation using a linear mixture model and fuzzy membership functions." In: *International Journal of Remote Sensing* 15.3 (1994), pp. 619–631. DOI: 10.1080/01431169408954100. eprint: <https://doi.org/10.1080/01431169408954100>. URL: <https://doi.org/10.1080/01431169408954100>.
- [179] G. M. FOODY. "Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data." In: *International Journal of Remote Sensing* 17.7 (1996), pp. 1317–1340. DOI: 10.1080/01431169608948706. eprint: <https://doi.org/10.1080/01431169608948706>. URL: <https://doi.org/10.1080/01431169608948706>.
- [180] Ivan Lizarazo and Joana Barros. "Fuzzy Image Segmentation for Urban Land-Cover Classification." In: *Photogrammetric Engineering and Remote Sensing* 76 (Feb. 2010). DOI: 10.14358/PERS.76.2.151.
- [181] Hang Do et al. "Multi-scale object-based fuzzy classification for LULC mapping from optical satellite images." In: *Spatial Information Research* 27 (Feb. 2019), pp. 247–257. DOI: 10.1007/s41324-019-00240-w.
- [182] Anil Kumar, Sanjay Ghosh, and Vinay Dadhwal. "Full fuzzy land cover mapping using remote sensing data based on fuzzyc-means and density estimation." In: *Canadian Journal of Remote Sensing* 33 (June 2014), pp. 81–87. DOI: 10.5589/m07-011.
- [183] Johannes Rosentreter et al. "Subpixel Mapping of Urban Areas Using EnMAP Data and Multioutput Support Vector Regression." In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10.5 (2017), pp. 1938–1948. DOI: 10.1109/JSTARS.2017.2652726.
- [184] Franz Schug et al. "Mapping urban-rural gradients of settlements and vegetation at national scale using Sentinel-2 spectral-temporal metrics and regression-based unmixing with synthetic training data." In: *Remote Sensing of Environment* 246 (Sept. 2020), p. 111810. DOI: 10.1016/j.rse.2020.111810.
- [185] L. Breiman. "Random Forests." In: *Machine Learning* 45 (2001), pp. 5–32. URL: <https://api.semanticscholar.org/CorpusID:89141>.

-
- [186] Jeffrey Walton. “Subpixel Urban Land Cover Estimation.” In: *Photogrammetric Engineering Remote Sensing* 74 (Oct. 2008), pp. 1213–1222. DOI: 10.14358/PERS.74.10.1213.
 - [187] Lien Loosvelt et al. “Random Forests as a tool for estimating uncertainty at pixel-level in SAR image classification.” In: *International Journal of Applied Earth Observation and Geoinformation* 19 (2012), pp. 173–184. ISSN: 1569-8432. DOI: <https://doi.org/10.1016/j.jag.2012.05.011>. URL: <https://www.sciencedirect.com/science/article/pii/S0303243412001195>.
 - [188] Xihao Li et al. “Forest Fire Smoke Detection Research Based on the Random Forest Algorithm and Sub-Pixel Mapping Method.” In: *Forests* 14.3 (2023). ISSN: 1999-4907. DOI: 10.3390/f14030485. URL: <https://www.mdpi.com/1999-4907/14/3/485>.
 - [189] Panpan Xu et al. “Comparative validation of recent 10m-resolution global land cover maps.” In: *Remote Sensing of Environment* 311 (2024), p. 114316. ISSN: 0034-4257. DOI: <https://doi.org/10.1016/j.rse.2024.114316>. URL: <https://www.sciencedirect.com/science/article/pii/S0034425724003341>.
 - [190] Linus Ericsson et al. “Self-Supervised Representation Learning: Introduction, advances, and challenges.” In: *IEEE Signal Processing Magazine* 39.3 (May 2022), pp. 42–62. ISSN: 1558-0792. DOI: 10.1109/msp.2021.3134634. URL: <http://dx.doi.org/10.1109/MSP.2021.3134634>.
 - [191] D. A. Alajaji and H. Alhichri. “Few Shot Scene Classification in Remote Sensing using Meta-Agnostic Machine.” In: *Proceedings of the 6th CDMA*. 2020, pp. 77–80. DOI: 10.1109/CDMA47397.2020.00019.
 - [192] Erik G. Miller, Nicholas E. Matsakis, and Paul A. Viola. “Learning from one example through shared densities on transforms.” In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2000). ISSN: 10636919. DOI: 10.1109/CVPR.2000.855856.
 - [193] Xiang Sean Zhou and Thomas S. Huang. “Small sample learning during multimedia retrieval using BiasMap.” In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2001. DOI: 10.1109/cvpr.2001.990450.
 - [194] M. Zhai, H. Liu, and F. Sun. “Lifelong Learning for Scene Recognition in Remote Sensing Images.” In: *IEEE GRSL* 16.9 (2019), pp. 1472–1476.
 - [195] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. “Meta-Learning for Multi-Label Few-Shot Classification.” In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022, pp. 346–355. DOI: 10.1109/WACV51458.2022.00042.
 - [196] L. Zintgraf et al. “Fast Context Adaptation via Meta-Learning.” In: *Proceedings of the 36th ICML*. 2019. ISBN: 9781510886988. arXiv: 1810.03642.

- [197] Manal Hamzaoui et al. “A Hierarchical Prototypical Network for Few-Shot Remote Sensing Scene Classification.” In: *Pattern Recognition and Artificial Intelligence*. Ed. by Mounîm El Yacoubi et al. Cham: Springer International Publishing, 2022, pp. 208–220. ISBN: 978-3-031-09282-4.
- [198] Hassan Gharoun et al. “Meta-learning Approaches for Few-Shot Learning: A Survey of Recent Advances.” In: *ACM Computing Surveys* 56.12 (2024). DOI: 10.1145/3659943.
- [199] J. Snell, K. Swersky, and R. Zemel. “Prototypical Networks for Few-Shot Learning.” In: *Advances in Neural Information Processing Systems*. 2017. arXiv: 1703.05175.
- [200] Surbhi Sharma et al. “Beyond Pixels: Identifying Built-Up Features at Subpixel Level for Enhanced Satellite-Based Land Cover Mapping.” In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 19 (Jan. 2026), pp. 24115–24129. DOI: 10.1109/JSTARS.2026.3712773.
- [201] Surbhi Sharma et al. “Improving Generalization for Few-Shot Remote Sensing Classification with Meta-Learning.” In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. 2022, pp. 5061–5064. DOI: 10.1109/IGARSS46834.2022.9884699.
- [202] Zhengpeng Feng et al. *TESSERA: Temporal Embeddings of Surface Spectra for Earth Representation and Analysis*. 2025. arXiv: 2506.20380 [cs.LG]. URL: <https://arxiv.org/abs/2506.20380>.
- [203] Niloufar Alipour Talemi, Julia Boone, and Fatemeh Afghah. “Agentic AI in Remote Sensing: Foundations, Taxonomy, and Emerging Systems.” In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2026, pp. 786–799.