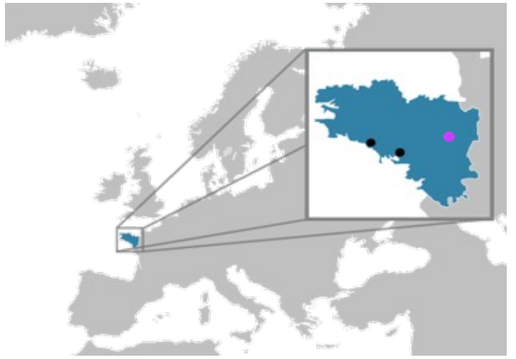
Sébastien Lefèvre
Université Bretagne Sud / IRISA
The Arctic University of Norway / ESA
sebastien.lefevre@irisa.fr

AI and Remote Sensing

Workshop on
Artificial Intelligence for Mapping Building Damages after Earthquake and Storms Based on Remote Sensing

JRC, Ispra, 10-12 November 2026

ABOUT ME



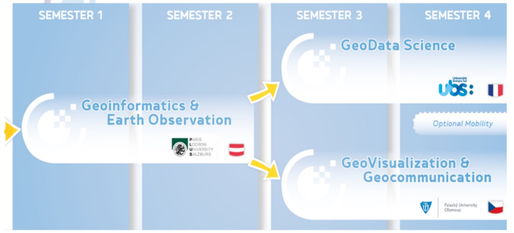
- Full Professor at University of South Brittany since 2010: www.univ-ubs.fr
- Adjunct Professor at UiT – The Arctic University of Norway since 2023
- Visiting Professor at ESA – Phi-lab since 2025
- Founder and former head of OBELIX group at IRISA:
  www.irisa.fr/obelix (25 researchers on ML, CV, and EO)
  leading the follow-up IDEFIX group on AI for EO, Environment, Astrophysics

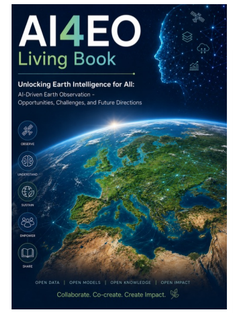


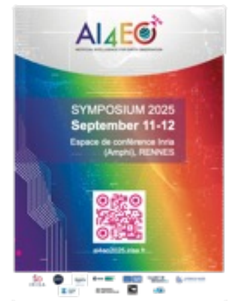
- Chair of the GeoData Science track
  EMJM Copernicus Master in Digital Earth
  www.master-cde.eu

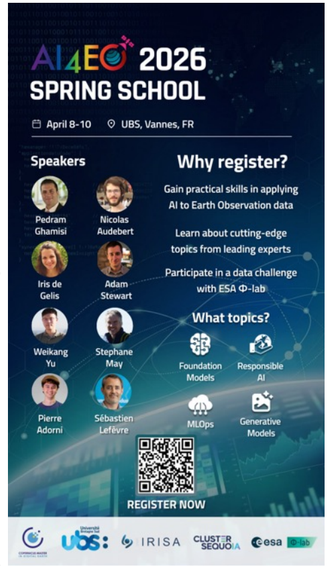
  Call open until 22 Jan 2027











- Holder of an Experienced Chair within AI Cluster SequoIA
- Involved in various AI4EO activities: collaborative book, symposium in 2025, yearly Spring School from 2026, next edition **30 Mar – 2 Apr 2027**
- 20y connection with JRC: JURSE in Paris (M.Pesaresi), Mathematical Morphology (P.Soille), **3 joint PhD started in 2026** through CDP with Unit E1 – Disaster Risk Management

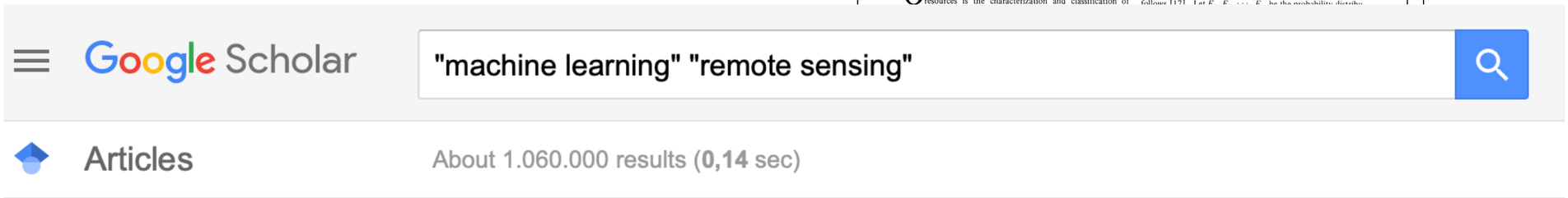
AI & REMOTE SENSING:

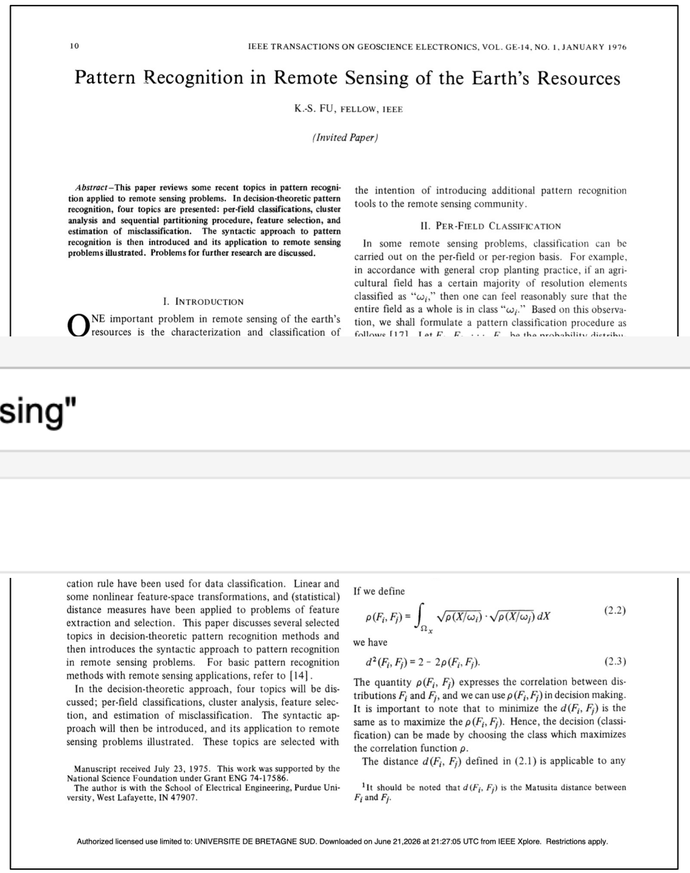
A SHORT HISTORICAL PERSPECTIVE

MACHINE LEARNING MET EO MORE THAN 50 YEARS AGO…

Seminal works in the 1970s

Widely used since then





10 IEEE TRANSACTIONS ON GEOSCIENCE ELECTRONICS, VOL. GE-14, NO. 1, JANUARY 1976

Pattern Recognition in Remote Sensing of the Earth's Resources

K.-S. FU, FELLOW, IEEE

(Invited Paper)

Abstract—This paper reviews some recent topics in pattern recognition applied to remote sensing problems. In decision-theoretic pattern recognition, four topics are presented: per-field classifications, cluster analysis and sequential partitioning procedure, feature selection, and estimation of misclassification. The syntactic approach to pattern recognition is then introduced and its application to remote sensing problems illustrated. Problems for further research are discussed.

I. INTRODUCTION

ONE important problem in remote sensing of the earth's resources is the characterization and classification of

the intention of introducing additional pattern recognition tools to the remote sensing community.

II. PER-FIELD CLASSIFICATION

In some remote sensing problems, classification can be carried out on the per-field or per-region basis. For example, in accordance with general crop planting practice, if an agricultural field has a certain majority of resolution elements classified as "ω_i," then one can feel reasonably sure that the entire field as a whole is in class "ω_i." Based on this observation, we shall formulate a pattern classification procedure as

cation rule have been used for data classification. Linear and some nonlinear feature-space transformations, and (statistical) distance measures have been applied to problems of feature extraction and selection. This paper discusses several selected topics in decision-theoretic pattern recognition methods and then introduces the syntactic approach to pattern recognition in remote sensing problems. For basic pattern recognition methods with remote sensing applications, refer to [14].

In the decision-theoretic approach, four topics will be discussed; per-field classifications, cluster analysis, feature selection, and estimation of misclassification. The syntactic approach will then be introduced, and its application to remote sensing problems illustrated. These topics are selected with

Manuscript received July 23, 1975. This work was supported by the National Science Foundation under Grant ENG 74-17586.
The author is with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

If we define

$$\rho(F_i, F_j) \equiv \int_{\Omega_X} \sqrt{p(X/\omega_i)} \cdot \sqrt{p(X/\omega_j)}\, dX \tag{2.2}$$

we have

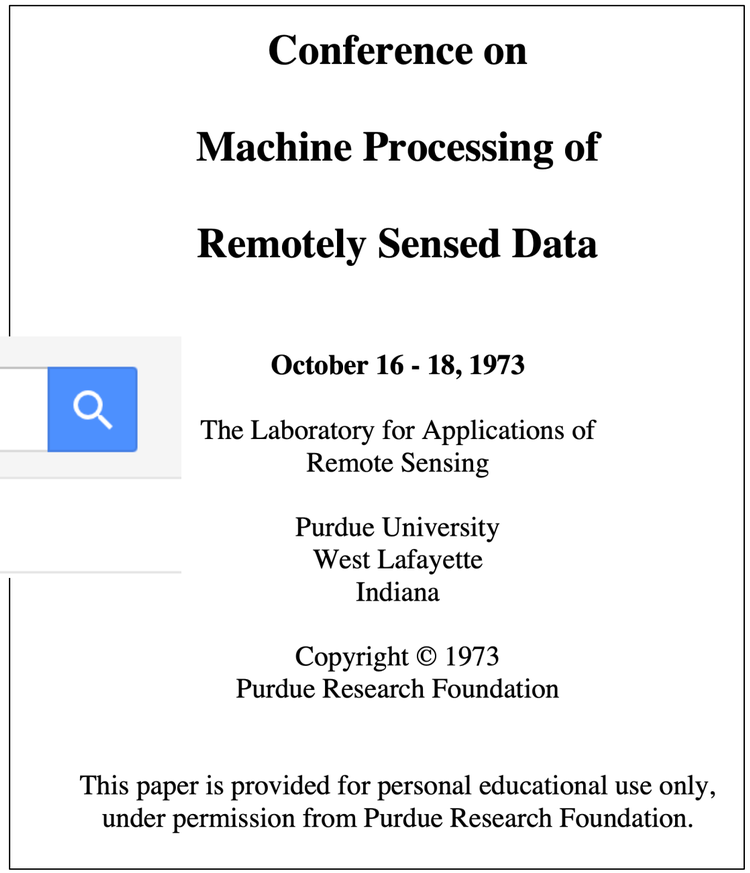
$$d^2(F_i, F_j) \equiv 2 - 2\rho(F_i, F_j). \tag{2.3}$$

The quantity $\rho(F_i, F_j)$ expresses the correlation between distributions F_i and F_j, and we can use $\rho(F_i, F_j)$ in decision making. It is important to note that to minimize the $d(F_i, F_j)$ is the same as to maximize the $\rho(F_i, F_j)$. Hence, the decision (classification) can be made by choosing the class which maximizes the correlation function ρ.

The distance $d(F_i, F_j)$ defined in (2.1) is applicable to any

[1]It should be noted that $d(F_i, F_j)$ is the Matusita distance between F_i and F_j.

Authorized licensed use limited to: UNIVERSITE DE BRETAGNE SUD. Downloaded on June 21,2026 at 21:27:05 UTC from IEEE Xplore. Restrictions apply.



Conference on

Machine Processing of

Remotely Sensed Data

October 16 - 18, 1973

The Laboratory for Applications of Remote Sensing

Purdue University
West Lafayette
Indiana

Copyright © 1973
Purdue Research Foundation

This paper is provided for personal educational use only, under permission from Purdue Research Foundation.

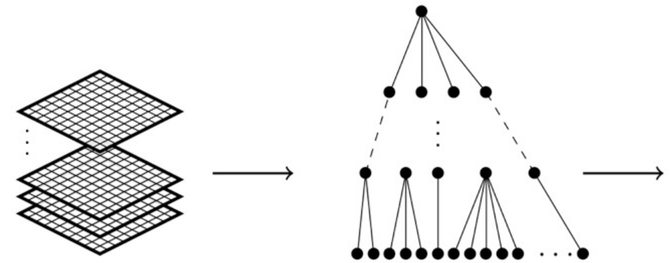
Three main components:

1. Training samples (labels)
2. Supervised classifier (KNN, SVM, RF)
3. Expert / handcrafted features

- Spectral signatures
- Geometric and spatial features
- Texture, Morphological features (DMP, etc)

MORPHOLOGICAL PROFILES AT SCALE

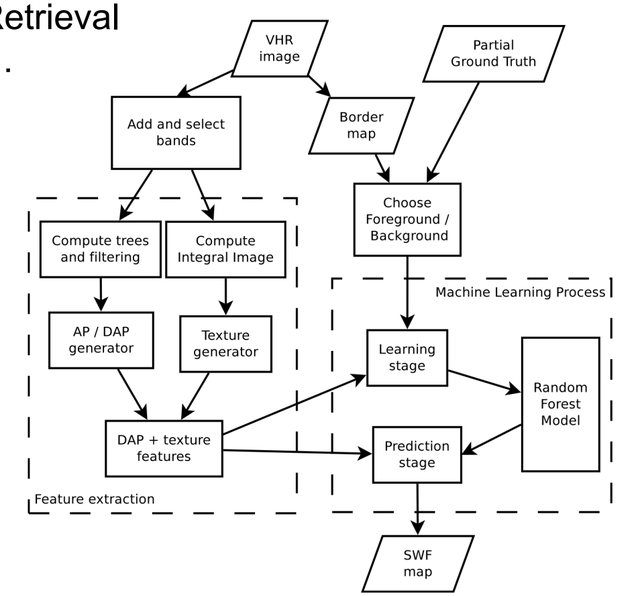
Mathematical Morphology + efficient algorithms → morphological hierarchies



Complex data

Tree construction and management

Applications
- Feature extraction
- Description
- Segmentation
- Classification
- Retrieval
- …



Embedded in an interactive RF classification pipeline

to process EU at VHR and map SWF for Copernicus



©SHUTTERSTOCK.COM/LAVIZZARA

A decade of advances

DEISE SANTANA MAIA, MINH-TAN PHAM, ERCHAN APTOULA, FLORENT GUIOTTE, AND SÉBASTIEN LEFÈVRE

Morphological attribute profiles (APs) are among the most prominent methods for spatial–spectral pixel analysis of remote sensing images. Since their introduction a decade ago to tackle land cover classification, many studies have been contributed to the state of the art, focusing not only on their application to a wider range of tasks but also on their performance improvement and extension to more complex Earth observation data.

Despite the overwhelming proliferation of deep learning-based methods in the past five years, APs are far from obsolete, due mainly to their high flexibility, low computational cost, lower training data requirement, and rigorous mathematical foundation. In this survey, an entire decade's worth of more than 100 AP-related contributions to the field of remote sensing have been compiled, providing an extensive panorama of this robust and effective tool. Moreover, a collective experimental comparison of the reviewed AP variations is provided as well, not only in terms of classification performance but, for the first time, in terms of their generalization capacity, too.

Digital Object Identifier 10.1109/MGRS.2021.3051859
Date of current version: 1 March 2021

BACKGROUND

Classification constitutes one of the paramount tasks of remote sensing image analysis for Earth observation. Its performance is critical for the success of land use and cover mapping and monitoring. The rapid development of high-resolution (HR) and very-HR (VHR) image acquisition technologies has led to increasingly complex images with higher levels of detail. Consequently, the importance of the joint exploitation of spatially contextual information along with spectral pixel characteristics became clear early on [1].

Morphological profiles (MPs) [2] were introduced almost two decades ago to address this exact issue. In essence, they produce multiscale descriptions of their input through the application of a sequence of morphological reconstruction-based filters using structuring elements of various sizes (and shapes) [2]. Their rigorous mathematical foundation and inherent ability to capture spatial–spectral information have led to the development of several variants [3]–[6].

However, as spatial resolutions and image sizes progressively increased, their relatively high computational cost has been drastically accentuated. Moreover, their initially celebrated capacity of capturing size and shape variations eventually became extremely limited when confronted with thematic classes differing in terms of alternative properties, such as contrast, homogeneity, and so on. In an effort to overcome these shortcomings, MPs have been

Authorized licensed use limited to: UNIVERSITE DE BRETAGNE SUD. Downloaded on October 18,2021 at 07:00:11 UTC from IEEE Xplore. Restrictions apply.

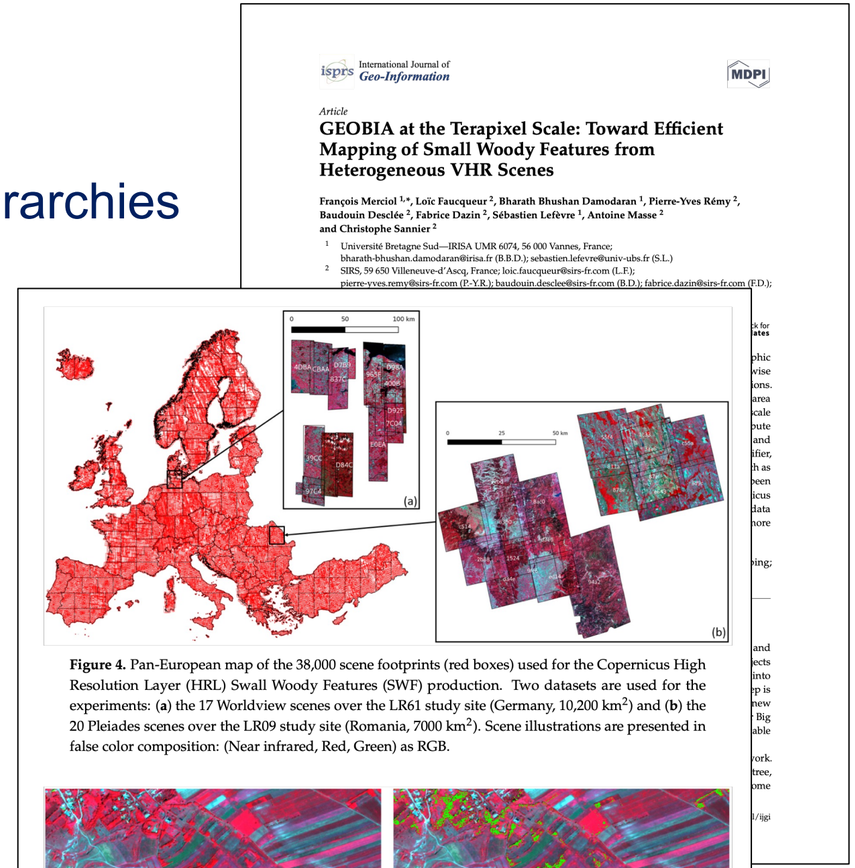
MORPHOLOGICAL PROFILES AT SCALE

Mathematical Morphology + efficient algorithms → morphological hierarchies

Embedded in an interactive
RF classification pipeline

to process EU at VHR
and map SWF for Copernicus





isprs International Journal of Geo-Information

MDPI

Article

GEOBIA at the Terapixel Scale: Toward Efficient Mapping of Small Woody Features from Heterogeneous VHR Scenes

François Mercier [1,*], Loïc Faucqueur [2], Bharath Bhushan Damodaran [1], Pierre-Yves Rémy [2], Baudouin Desclée [2], Fabrice Dazin [2], Sébastien Lefèvre [1], Antoine Masse [2] and Christophe Sannier [2]

[1] Université Bretagne Sud—IRISA UMR 6074, 56 000 Vannes, France; bharath-bhushan.damodaran@irisa.fr (B.B.D.); sebastien.lefevre@univ-ubs.fr (S.L.)

[2] SIRS, 59 650 Villeneuve-d'Ascq, France; loic.faucqueur@sirs-fr.com (L.F.); pierre-yves.remy@sirs-fr.com (P.-Y.R.); baudouin.desclee@sirs-fr.com (B.D.); fabrice.dazin@sirs-fr.com (F.D.);

Figure 4. Pan-European map of the 38,000 scene footprints (red boxes) used for the Copernicus High Resolution Layer (HRL) Swall Woody Features (SWF) production. Two datasets are used for the experiments: (**a**) the 17 Worldview scenes over the LR61 study site (Germany, 10,200 km^2) and (**b**) the 20 Pleiades scenes over the LR09 study site (Romania, 7000 km^2). Scene illustrations are presented in false color composition: (Near infrared, Red, Green) as RGB.

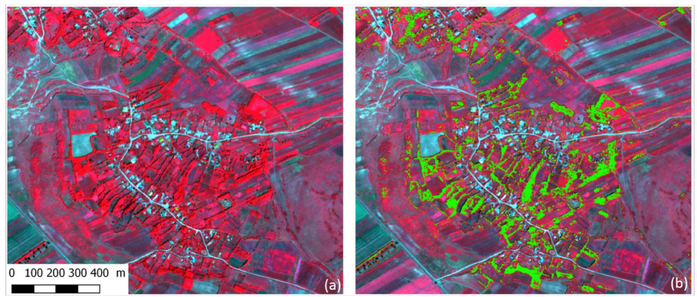


Figure 5. Illustration of classification results: (**a**) false color composition of Pleiades scene #c55a over LR09 study site in Romania and (**b**) the same false color composition with superimposition of the Small Woody Features layer classification output (green).

THE DEEP LEARNING BREAKTHROUGH (ALREADY 10-15 YEARS AGO)

No need to design the features anymore, simply learn them from the (annotated) data

Various models for various tasks

- Classification
- Semantic segmentation
- Object detection
- …

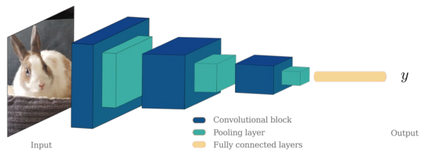
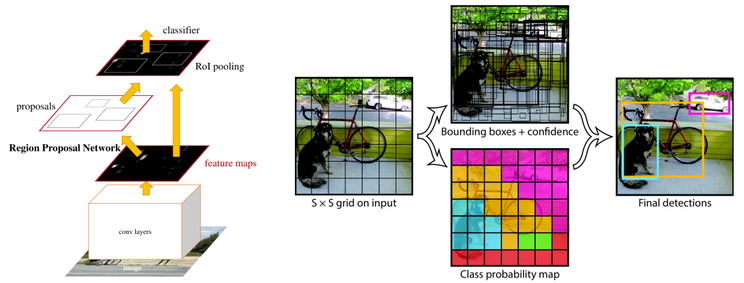
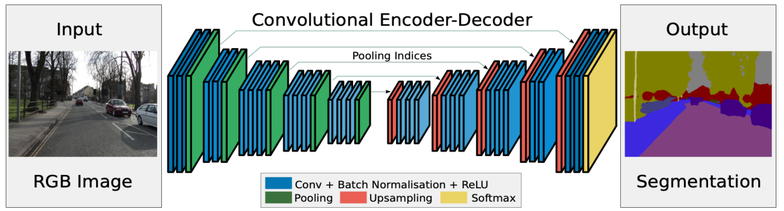


Figure 1.8 – Standard CNN architecture. A CNN is usually composed of convolutional blocks (sequence of convolutional layers, plus activation function, pooling layers and fully connected layers at the end.

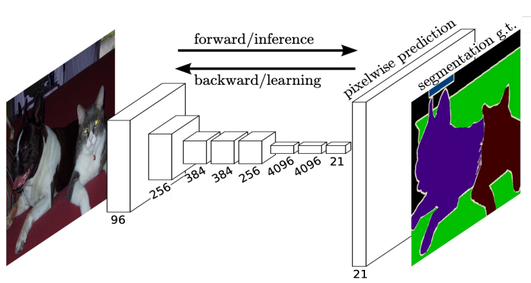


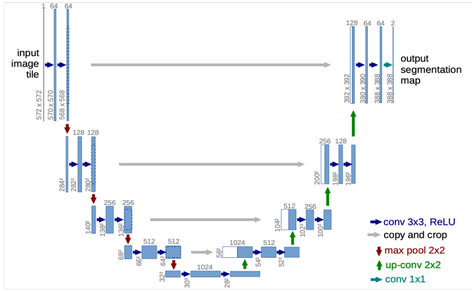


how many different deep learning architectures have been introduced in computer vision?

For the entire field of computer vision, I would use the following orders of magnitude:

Definition	Approximate scale
Major architectural families	15–20
Influential named architectures	Hundreds
Published architectural variants	Thousands
All proposed architectures, including minor modifications	Potentially tens of thousands

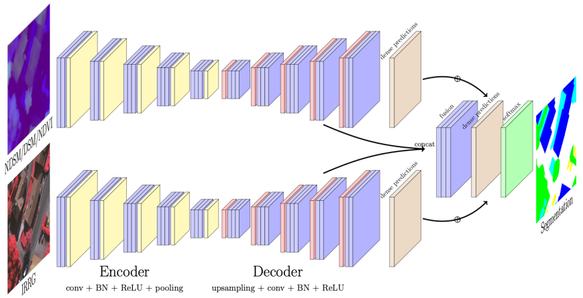


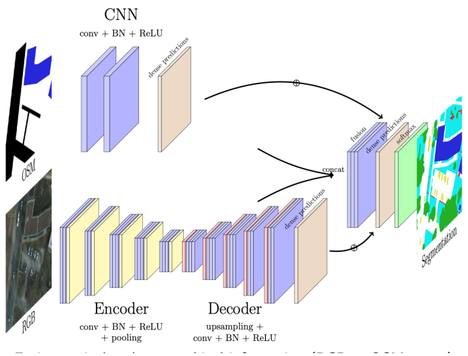


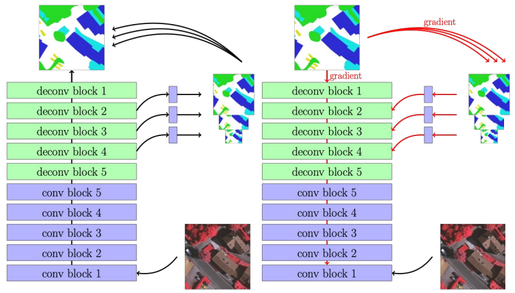
AMONG THE FIRST WORKS IN DEEP LEARNING FOR REMOTE SENSING…

Nicolas Audebert PhD with ONERA (2015-2018)



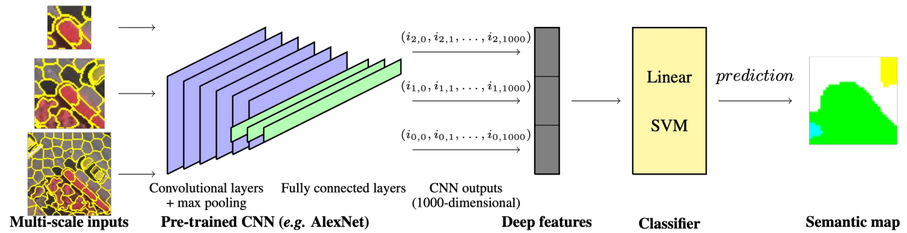


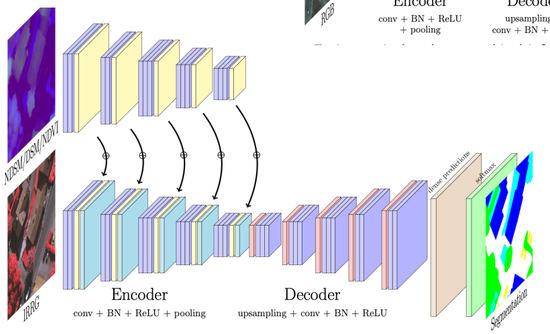


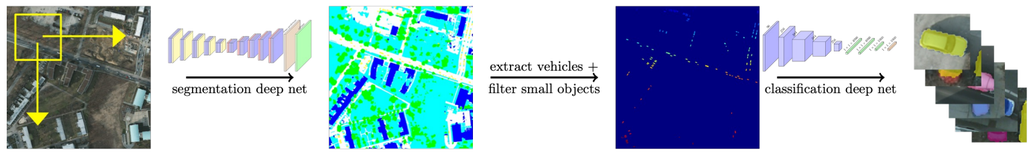


(a) Multi-scale prediction using SegNet. (b) Backpropagation at multiple scales.

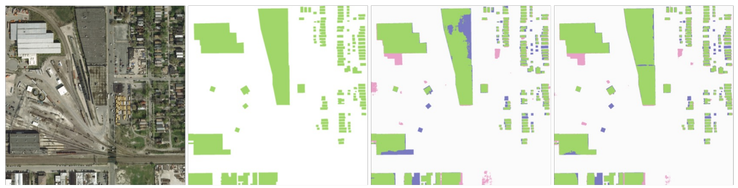
Figure 3: Multi-scale deep supervision of SegNet with 3 branches on remote sensing data.



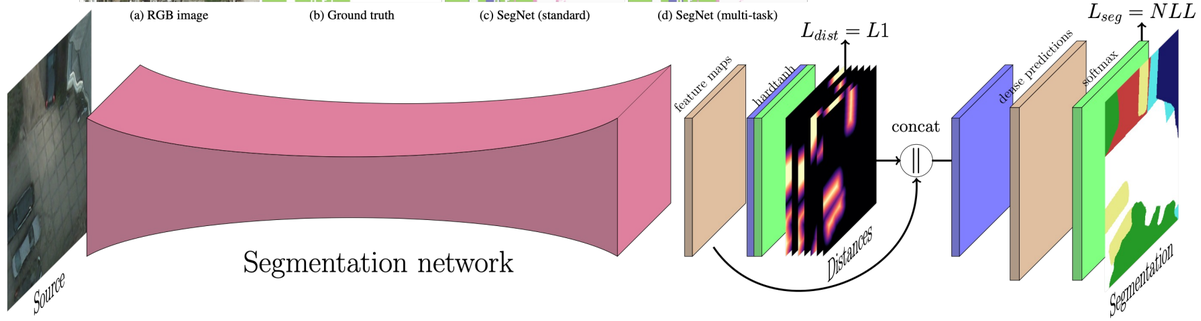


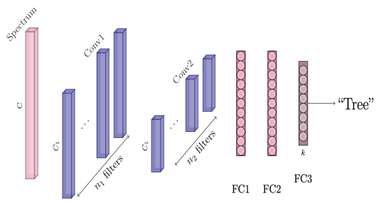


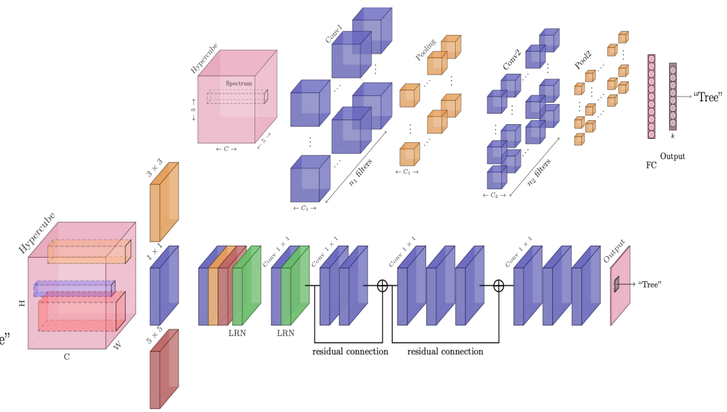
Sliding window on RGB data | Semantic map | Vehicle mask | Classified vehicles

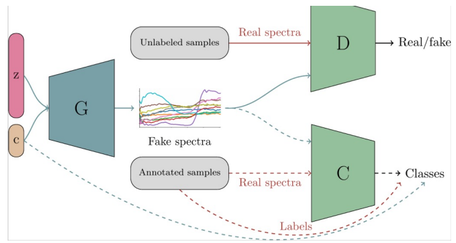


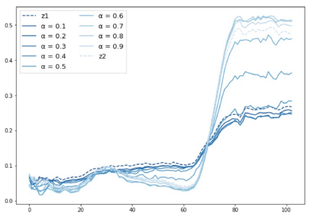
(a) RGB image (b) Ground truth (c) SegNet (standard) (d) SegNet (multi-task)

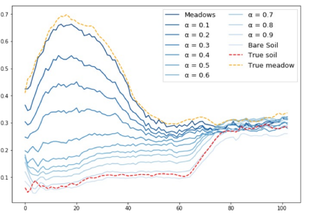












LARGE-SCALE SEMANTIC CLASSIFICATION: OUTCOME OF THE FIRST YEAR OF INRIA AERIAL IMAGE LABELING BENCHMARK

Bohao Huang[1], *Kangkang Lu*[2], *Nicolas Audebert*[3,4], *Andrew Khalel*[5], *Yuliya Tarabalka*[6], *Jordan Malof*[1]
Alexandre Boulch[3], *Bertrand Le Saux*[3], *Leslie Collins*[1], *Kyle Bradbury*[1], *Sébastien Lefèvre*[4], *Motaz El-Saban*[5]

[1] Duke University; [2] NUS; [3] ONERA; [4] Univ. Bretagne-Sud, IRISA; [5] Raisa energy; [6] UCA, Inria
Email: yuliya.tarabalka@inria.fr

Train	Tiles*	Total area	Test	Tiles*	Total area
Austin, TX	36	81 km^2	Bellingham, WA	36	81 km^2
Chicago, IL	36	81 km^2	San Francisco, CA	36	81 km^2
Kitsap County, WA	36	81 km^2	Bloomington, IN	36	81 km^2
Vienna, Austria	36	81 km^2	Innsbruck, Austria	36	81 km^2
West Tyrol, Austria	36	81 km^2	East Tyrol, Austria	36	81 km^2
Total	180	405 km^2	Total	180	405 km^2

Table 1: IAIL dataset statistics. *Tile size: 5000^2 px. (0.3 m/pixel).

released, and a contest has been launched online to classify data from five test cities.

In this paper, we discuss the outcomes of the first year of the Inria benchmark contest. We first briefly recall the compo-

2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops

Joint Learning from Earth Observation and OpenStreetMap Data to Get Faster Better Semantic Maps

Nicolas Audebert, Bertrand Le Saux
ONERA, *The French Aerospace Lab*
F-91761 Palaiseau, France
{nicolas.audebert,bertrand.le_saux}@onera.fr

Sébastien Lefèvre
Univ. Bret
F-
seba

Abstract

In this work, we investigate the use of OpenStreetMap data for semantic labeling of Earth Observation images. Deep neural networks have been used in the past for remote sensing data classification from various sensors, including multispectral, hyperspectral, SAR and LiDAR data. While OpenStreetMap has already been used as ground truth data for training such networks, this abundant data source remains rarely exploited as an input information layer. In this paper, we study different use cases and deep network architectures to leverage OpenStreetMap data for semantic labeling of aerial and satellite images. Especially, we look into fusion based architectures and coarse-to-fine segmentation to include the OpenStreetMap layer into multispectral-based deep fully convolutional networks. We illustrate how these methods can be successfully used on two public datasets: ISPRS Potsdam and DFC2017. We show that OpenStreetMap data can efficiently be integrated into the vision-based deep learning models and that it significantly improves both the accuracy performance and the convergence speed of the networks.

1. Introduction

Dense labeling of remote sensing data is a common task to generate semantic maps of large areas automatically, either to perform cartography of urban areas or to determine land use covers at a large scale. Many techniques that originated in the computer vision community for semantic segmentation can be applied on Red-Green-Blue (RGB) remote sensing data to this end and has been successfully used with state-of-the-art results, *e.g.* Convolutional Neural Networks (CNN) for land use classification [4] or Fully Convolutional Networks (FCN) for semantic labeling of urban areas [1]. Yet, remote sensing data is rarely limited to RGB channels. Often, multiple sensors are used to obtain complementary physical information about the observed area,

with the final pro
LiDAR and SAR
tigated semantic l
sion of multiple h
age complementa
is still a huge unt
geo-information s
tary crowdsourced
the very popular
valuable source o
training ground tr
roads, buildings fo
can also be seen a
vide another infor
sensors. Consequ
question: how ca
based semantic lab

2. Related wor

Semantic label
deep neural netwo
in the literature.
using CNN [24], r
works for automat
The most recent
on the use of Ful
chitecture initially
multimedia image
mote sensing dat
deep models obta
lution datasets [28
ysis [1], graphical
ary prediction [22
cessful for buildin
ages [21, 31, 34] a

Although most
e.g. RGB or mul
geneous data laye

2160-7516/17 $31.00 © 2017 IEEE
DOI 10.1109/CVPRW.2017.199

Authorized licensed use limited to: UNIVERSITE DE BRETAGNE SUD. Downloaded on November 09,2026 a

ISPRS Journal of Photogrammetry and Remote Sensing 140 (2018) 20–32

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks

Nicolas Audebert [a,b,*], Bertrand Le Saux [a], Sébastien Lefèvre [b]

[a] *ONERA, The French Aerospace Lab, F-91761 Palaiseau, France*
[b] *Univ. Bretagne-Sud, UMR 6074, IRISA, F-56000 Vannes, France*

ARTICLE INFO

Article history:
Received 28 February 2017
Received in revised form 11 November 2017
Accepted 13 November 2017
Available online 23 November 2017

Keywords:
Deep learning
Remote sensing
Semantic mapping
Data fusion

ABSTRACT

In this work, we investigate various methods to deal with semantic labeling of very high resolution multi-modal remote sensing data. Especially, we study how deep fully convolutional networks can be adapted to deal with multi-modal and multi-scale remote sensing data for semantic labeling. Our contributions are threefold: (a) we present an efficient multi-scale approach to leverage both a large spatial context and the high resolution data, (b) we investigate early and late fusion of Lidar and multispectral data, (c) we validate our methods on two public datasets with state-of-the-art results. Our results indicate that late fusion make it possible to recover errors steaming from ambiguous data, while early fusion allows for better joint-feature learning but at the cost of higher sensitivity to missing data.
© 2017 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

1. Introduction

Remote sensing has benefited a lot from deep learning in the past few years, mainly thanks to progress achieved in the computer vision community on natural RGB images. Indeed, most deep learning architectures designed for multimedia vision can be used on remote sensing optical images. This resulted in significant improvements in many remote sensing tasks such as vehicle detection (Chen et al., 2014), semantic labeling (Audebert et al., 2016; Marmanis et al., 2016; Maggiori et al., 2017; Sherrah, 2016) and land cover/use classification (Penatti et al., 2015; Nogueira et al., 2017). However, these improvements have been mostly restrained to traditional 3-channels RGB images that are the main focus of the computer vision community.

On the other hand, Earth Observation data is rarely limited to this kind of optical sensor. Additional data, either from the same sensor (e.g. multispectral data) or from another one (e.g. a Lidar point cloud) is sometimes available. However, adapting vision-based deep networks to these larger data is not trivial, as this requires to work with new data structures that do not share the same underlying physical and numerical properties. Nonetheless, all these sources provide complementary information that should be used jointly to maximize the labeling accuracy.

In this work we present how to build a comprehensive deep learning model to leverage multi-modal high-resolution remote sensing data, with the example of semantic labeling of Lidar and multispectral data over urban areas. Our contributions are the following:

- We show how to implement an efficient multi-scale deep fully convolutional neural network using SegNet (Badrinarayanan et al., 2017) and ResNet (He et al., 2016).
- We investigate early fusion of multi-modal remote sensing data based on the FuseNet principle (Hazirbas et al., 2016). We show that while early fusion significantly improves semantic segmentation by allowing the network to learn jointly stronger multi-modal features, it also induces higher sensitivity to missing or noisy data.
- We investigate late fusion of multi-modal remote sensing data based on the residual correction strategy (Audebert et al., 2016). We show that, although performing not as good as early fusion, residual correction improves semantic labeling and makes it possible to recover some critical errors on hard pixels.
- We successfully validate our methods on the ISPRS Semantic Labeling Challenge datasets of Cramer (2010), with results placing our methods amongst the best of the state-of-the-art.

2. Related work

Semantic labeling of remote sensing data relates to the dense pixel-wise classification of images, which is called either "semantic

* Corresponding author at: ONERA, The French Aerospace Lab, F-91761 Palaiseau, France.
E-mail addresses: nicolas.audebert@onera.fr (N. Audebert), bertrand.le_saux@onera.fr (B. Le Saux), sebastien.lefevre@irisa.fr (S. Lefèvre).

https://doi.org/10.1016/j.isprsjprs.2017.11.011
0924-2716/© 2017 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

HOW USEFUL IS REGION-BASED CLASSIFICATION OF REMOTE SENSING IMAGES IN A DEEP LEARNING FRAMEWORK ?

Nicolas Audebert[1, 2], *Bertrand Le Saux* [1], *Sébastien Lefèvre* [2]

[1] ONERA, *The French Aerospace Lab*, F-91761 Palaiseau, France
[2] Univ. Bretagne-Sud, UMR 6074, IRISA, F-56000 Vannes, France
Emails: {nicolas.audebert, bertrand.le_saux}@onera.fr, sebastien.lefevre@irisa.fr

ABSTRACT

In this paper, we investigate the impact of segmentation algorithms as a preprocessing step for classification of remote sensing images in a deep learning framework. Especially, we address the issue of segmenting the image into regions to be classified using pre-trained deep neural networks as feature extractors for an SVM-based classifier. An efficient segmentation as a preprocessing step helps learning by adding a spatially-coherent structure to the data. Therefore, we compare algorithms producing superpixels with more traditional

2. CLASSIFICATION FRAMEWORK

Deep learning has been the state-of-the-art in computer vision for a few years. Neural networks operate at the pixel level by simultaneously learning which features to extract and how to classify them. Convolutional neural networks are composed of layers of neurons computing convolutions on the previous layer outputs. These layers are stacked and combined with max-pooling (i.e. sampling maximum activations) and elementwise non-linear transfer functions (e.g. $tanh$) to extract

remote sensing

MDPI

Article

Segment-before-Detect: Vehicle Detection and Classification through Semantic Segmentation of Aerial Images

Nicolas Audebert [1,2,*], Bertrand Le Saux [1] and Sébastien Lefèvre [2]

[1] ONERA, The French Aerospace Lab, F-91761 Palaiseau, France; bertrand.le_saux@onera.fr
[2] Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), University Bretagne Sud, UMR 6074, F-56000 Vannes, France; sebastien.lefevre@irisa.fr
* Correspondence: nicolas.audebert@onera.fr

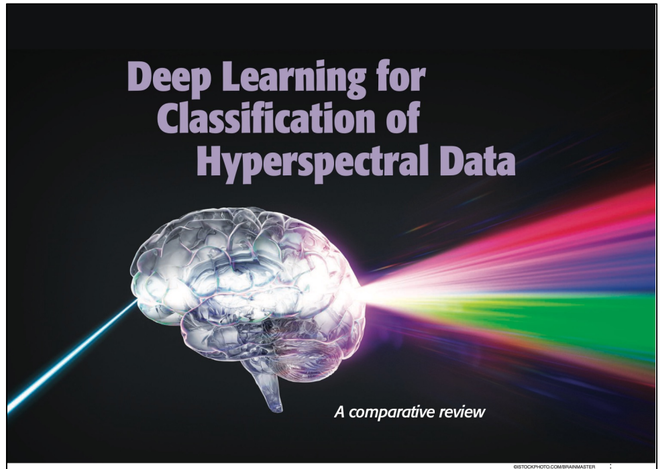
Academic Editors: Norman Kerle, Markus Gerke and Prasad S. Thenkabail
Received: 28 December 2016; Accepted: 7 April 2017; Published: 13 April 2017

Abstract: Like computer vision before, remote sensing has been radically changed by the introduction of deep learning and, more notably, Convolution Neural Networks. Land cover classification, object detection and scene understanding in aerial images rely more and more on deep networks to achieve new state-of-the-art results. Recent architectures such as Fully Convolutional Networks can even produce pixel level annotations for semantic mapping. In this work, we present a deep-learning based *segment-before-detect* method for segmentation and subsequent detection and classification of several varieties of wheeled vehicles in high resolution remote sensing images. This allows us to investigate object detection and classification on a complex dataset made up of visually similar classes, and to demonstrate the relevance of such a subclass modeling approach. Especially, we want to show that deep learning is also suitable for object-oriented analysis of Earth Observation data as effective object detection can be obtained as a byproduct of accurate semantic segmentation. First, we train a deep fully convolutional network on the ISPRS Potsdam and the NZAM/ONERA Christchurch datasets and show how the learnt semantic maps can be used to extract precise segmentation of vehicles. Then, we show that those maps are accurate enough to perform vehicle detection by simple connected component extraction. This allows us to study the repartition of vehicles in the city. Finally, we train a Convolutional Neural Network to perform vehicle classification on the VEDAI dataset, and transfer its knowledge to classify the individual vehicle instances that we detected.

Keywords: deep learning; vehicle detection; semantic segmentation; object classification

1. Introduction

Deep learning for computer vision grows more popular every year, especially thanks to Convolutional Neural Networks (CNN) that are able to learn powerful and expressive descriptors from images for a large range of tasks: classification, segmentation, detection, etc. This ubiquity of CNN in computer vision is now starting to affect remote sensing as well, as they can tackle many tasks such as land use classification or object detection in aerial images. Moreover, new architectures have appeared, derived from Fully Convolutional Networks (FCN) [1], able to output dense pixel-wise annotations and thus able to achieve fine-grained classification. Such architectures have quickly become state-of-the-art for popular datasets such as PASCAL VOC2012 [2] and Microsoft COCO [3]. In an Earth Observation context, these FCN models are now especially appealing, as dense prediction allows us performing semantic mapping without requiring any preprocessing tricks. Therefore, using FCN for Earth Observation means we can shift from superpixel segmentation and region-based classification [4–6] to fully supervised semantic segmentation [7].



GENERATIVE ADVERSARIAL NETWORKS FOR REALISTIC SYNTHESIS OF HYPERSPECTRAL SAMPLES

Nicolas Audebert [1, 2], *Bertrand Le Saux* [1], *Sébastien Lefèvre* [2]

[1] ONERA, *The French Aerospace Lab*, F-91761 Palaiseau, France
[2] Univ. Bretagne-Sud, UMR 6074, IRISA, F-56000 Vannes, France

ABSTRACT

This work addresses the scarcity of annotated hyperspectral data required to train deep neural networks. Especially, we investigate generative adversarial networks and their application to the synthesis of consistent labeled spectra. By training such networks on public datasets, we show that these models are not only able to capture the underlying distribution, but also to generate genuine-looking and physically plausible spectra. Moreover, we experimentally validate that the synthetic samples can be used as an effective data augmentation strategy. We validate our approach on several public hyperspectral datasets using a variety of deep classifiers.

***Index Terms*—** hyperspectral image classification, generative models, deep learning, data augmentation.

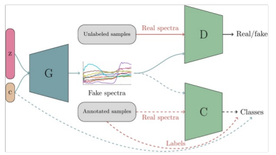


Fig. 1: The generative adversarial network used for hyperspectral sample synthesis. Red arrows denote the training of the classifier and discriminator, while the blue arrows denote the training of the generator. Dashed lines denote connections used in the supervised setting.

1. INTRODUCTION

Data augmentation consists of introducing unobserved samples into the optimization process of a statistical model [1]. Since the reintroduction of Convolutional Neural Networks (CNN) for image classification [2], this practice has been critical to avoid overfitting of deep networks. Therefore, as annotated hyperspectral data is scarce, overfitting is an even more common pitfall compared to multimedia image processing. Although recent efforts have been made to use CNN for hyperspectral image classification [3, 4, 5, 6], successes are limited to small datasets that do not leverage the generalization capacity of deep networks.

To this end, recent works have started to investigate data augmentation as a way to artificially enlarge the quantity of annotated samples. For example, [7] suggested a model of relighting to simulate the same hyperspectral pixel under different illuminations. With a more data-driven approach, [8] introduced a label propagation strategy to incorporate observed but unlabeled samples to the training set. However, these methods require either unlabeled samples or physics-related assumptions and modeling. On the other hand, [9] introduced generative models in hyperspectral image processing by using variational autoencoders to find the endmembers composition of spectral mixtures.

N. Audebert's work is funded by the Total-ONERA project NAOMI.

In this work, we introduce a way to artificially synthesize new annotated hyperspectral samples using a purely data-driven approach based on generative adversarial networks [10]. More specifically, we use a GAN to approximate the distribution of the observed hyperspectral samples and to generate new plausible samples that can be used to train deep networks. Our method can exploit both labeled and unlabeled samples and is validated on several datasets covering both aerial and satellite sensors over rural and urban areas.

2. GENERATIVE MODELS

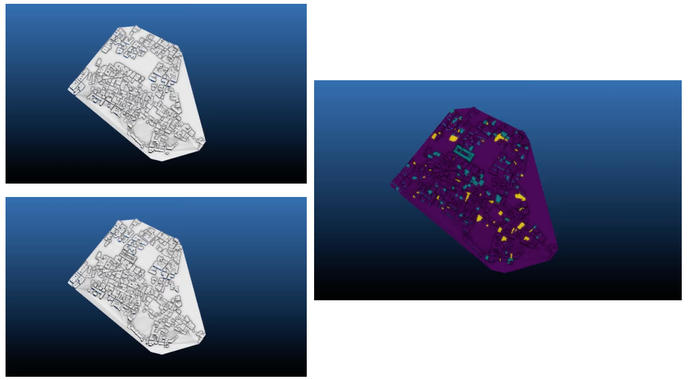
The Generative Adversarial Network framework has been introduced in [10]. It uses deep neural networks to approximate an unknown data distribution based on its observations. The idea is to generate new samples of a given distribution by training a generator to map random noise from the latent space to the distribution. However, the target distribution is observed only on some data points and we wish to use the generator to create new data points that also still belong to the underlying distribution. To this end, the generator is trained to approximate the distribution using an adversarial objective function. This function is obtained by introducing another

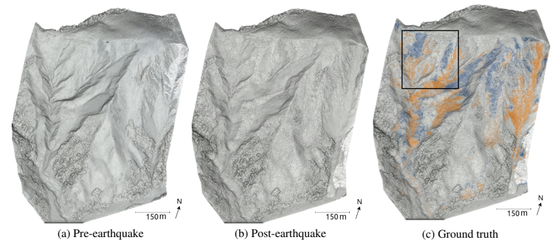
978-1-5386-7150-4/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: UNIVERSITE DE BRETAGNE SUD. Downloaded on November 09,2026 at 23:28:30 UTC from IEEE Xplore. Restrictions apply.

WHEN DEEP LEARNING MEETS POINT CLOUDS & CHANGE DETECTION

Method	mAcc (%)	mIoU (%)	Per class IoU (%) Unchanged	Source	Deposit
Encoder Fusion SiamKPConv	93.87	83.84	93.58	74.38	83.57

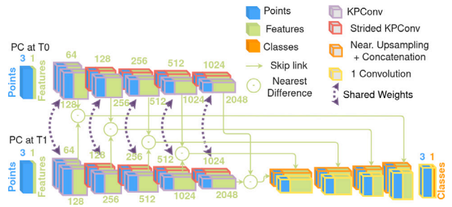


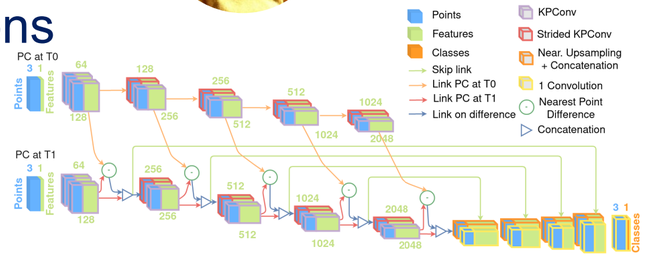


Iris de Gélis PhD with CNES/Magelium (2020-2023)

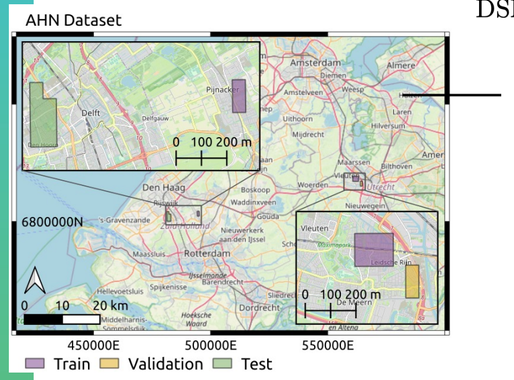
- Preparing AI for Co3D
- Various 3DCD models
- Various applications







Method	mAcc	mIoU_{ch}
Siamese KPConv (ours)	85.65 ± 1.55	**72.95** ± 2.05
Pseudo-Siamese KPConv (ours)	**87.87** ± 1.89	69.33 ± 1.99
DSM-Siamese	50.87 ± 1.15	30.96 ± 2.48
DSM-Pseudo-Siamese	70.71 ± 5.09	48.85 ± 7.03
DSM-FC-EF	71.47 ± 1.43	45.57 ± 0.98
RF	47.94 ± 0.02	29.45 ± 0.02



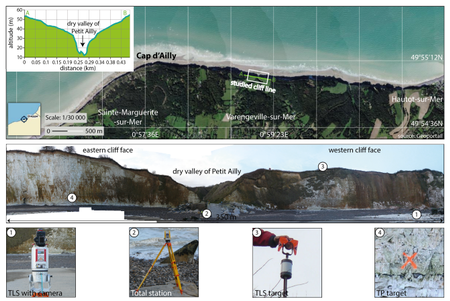


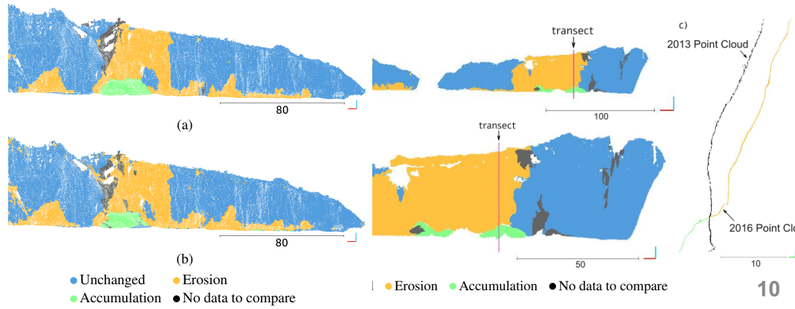
(a) PC 1: AHN3 (b) PC 2: AHN4 (c) GT

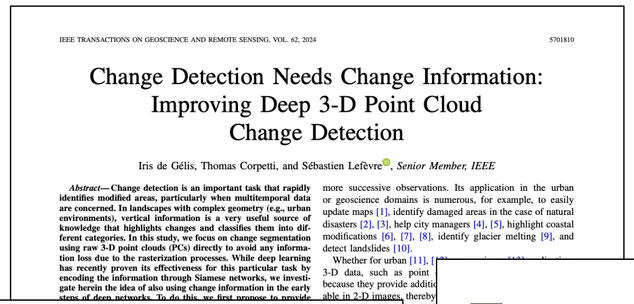
Unchanged · New Building · Demolition · New Clutter

- Public datasets
- Supervised/self-supervised/unsupervised models
- PhD awards @ AFIRF & MAGIS
- Mobility at TUM

	Per class IoU(%) Unchanged	Erosion	Acc.	mIoU_{int}
Siam. KPConv	91.94	83.86	**70.28**	**82.03**
M3C2 + threshold	**95.06**	**87.23**	48.20	68.12







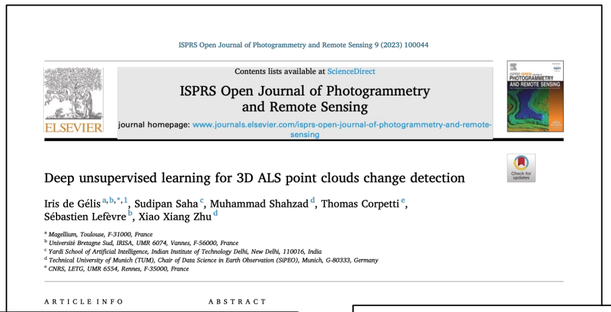
IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 62, 2024 5701810

Change Detection Needs Change Information: Improving Deep 3-D Point Cloud Change Detection

Iris de Gélis, Thomas Corpetti, and Sébastien Lefèvre, *Senior Member, IEEE*

***Abstract*— Change detection is an important task that rapidly identifies modified areas, particularly when multitemporal data are concerned. In landscapes with complex geometry (e.g., urban environments), vertical information is a very useful source of knowledge that highlights changes and classifies them into different categories. In this study, we focus on change segmentation using raw 3-D point clouds (PCs) directly to avoid any information loss due to the rasterization processes. While deep learning has recently proven its effectiveness for this particular task by encoding the information through Siamese networks, we investigate herein the idea of also using change information in the early steps of deep networks. To do this, we first propose to provide** more successive observations. Its application in the urban or geoscience domains is numerous, for example, to easily update maps [1], identify damaged areas in the case of natural disasters [2], [3], help city managers [4], [5], highlight coastal modifications [6], [7], [8], identify glacier melting [9], and detect landslides [10].

Whether for urban [11], [
3-D data, such as point
because they provide additi
able in 2-D images, thereby



ISPRS Open Journal of Photogrammetry and Remote Sensing 9 (2023) 100044

Contents lists available at ScienceDirect

ISPRS Open Journal of Photogrammetry and Remote Sensing

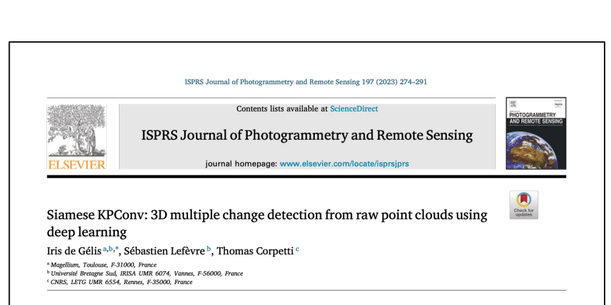
journal homepage: www.journals.elsevier.com/isprs-open-journal-of-photogrammetry-and-remote-sensing

Deep unsupervised learning for 3D ALS point clouds change detection

Iris de Gélis [a,b,*,1], Sudipan Saha [c], Muhammad Shahzad [d], Thomas Corpetti [e], Sébastien Lefèvre [b], Xiao Xiang Zhu [d]

[a] Magellium, Toulouse, F-31000, France
[b] Université Bretagne Sud, IRISA, UMR 6074, Vannes, F-56000, France
[c] Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi, 110016, India
[d] Technical University of Munich (TUM), Chair of Data Science in Earth Observation (SiPEO), Munich, G-80333, Germany
[e] CNRS, LETG, UMR 6554, Rennes, F-35000, France

ARTICLE INFO ABSTRACT



ISPRS Journal of Photogrammetry and Remote Sensing 197 (2023) 274–291

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Siamese KPConv: 3D multiple change detection from raw point clouds using deep learning

Iris de Gélis [a,b,*], Sébastien Lefèvre [b], Thomas Corpetti [c]

[a] Magellium, Toulouse, F-31000, France
[b] Université Bretagne Sud, IRISA UMR 6074, Vannes, F-56000, France
[c] CNRS, LETG UMR 6554, Rennes, F-35000, France

ARTICLE INFO

Keywords:
3D point clouds
Change detection
Deep learning
Siamese network
3D Kernel Point Convolution

ABSTRACT

This study is concerned with urban change detection and categorization in point clouds. In such situations, objects are mainly characterized by their vertical axis, and the use of native 3D data such as 3D Point Clouds (PCs) is, in general, preferred to rasterized versions because of significant loss of information implied by any rasterization process. Yet, for obvious practical reasons, most existing studies only focus on 2D images for change detection purpose. In this paper, we propose a method capable of performing change detection directly within 3D data. Despite recent deep learning developments in remote sensing, to the best of our knowledge there is no such method to tackle multi-class change segmentation that directly processes raw 3D PCs. Thereby, based on advances in deep learning for change detection in 2D images and for analysis of 3D point clouds, we propose a deep Siamese KPConv network that deals with raw 3D PCs to perform change detection and categorization in a single step. Experimental results are conducted on synthetic and real data of various kinds (LiDAR, multi-sensors). Tests performed on simulated low density LiDAR and multi-sensor datasets show that our proposed method can obtain up to 80% of mean of IoU over classes of changes, leading to an improvement ranging from 10% to 30% over the state-of-the-art. A similar range of improvements is attainable on real data. Then, we show that pre-training Siamese KPConv on simulated PCs allows us to greatly reduce (more than 3,000×) the annotations required on real data. This is a highly significant result to deal with practical scenarios. Finally, an adaptation of Siamese KPConv is realized to deal with change classification at PC scale. Our network overtakes the current state-of-the-art deep learning method by 23% and 15% of mean of IoU when assessed on synthetic and public Change3D datasets, respectively. The code is available at the following link: https://github.com/IdeGelis/torch-points3d-SiameseKPConv.



IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10641348

LANDSLIDE DETECTION IN 3D POINT CLOUDS WITH DEEP SIAMESE CONVOLUTIONAL NETWORK

Iris de Gélis[1,2], Thomas Bernard[3], Dimitri Lague[3], Thomas Corpetti[4], Sébastien Lefèvre[2]*

[1] Magellium, F-31000 Toulouse, France
[2] Université Bretagne Sud, IRISA UMR 6074, F-56000 Vannes, France
[3] Univ Rennes, CNRS, Géosciences Rennes, UMR 6118, F-35000 Rennes, France
[4] CNRS, LETG UMR 6554, F-35000 Rennes, France

ABSTRACT

Generally caused by extreme events, landslides cause severe landscape modifications and may endanger local population. It is important to be able to map them in order to better understand landscape evolution. 3D LiDAR point clouds (PCs) are a relevant choice compared to 2D imagery to directly sense ground shape modification under vegetated areas. Most of the studies propose to rely on rasterization of PCs, or multi-step semi-automatic process with tedious manual results refinement in the 3D PCs. In this study, we aim at experimenting a deep learning method to directly extract landslide sources and deposits from raw 3D PCs. To this end, we train an Encoder Fusion SiamKPConv network, designed for 3D PCs change detection, for the specific task of landslides identification in PCs acquired before and after Kaikōura earthquake (New-Zealand). The experimental results (93.87% of accuracy) show the relevance of this model.

***Index Terms*—** Landslide detection, 3D point clouds, deep learning

1. INTRODUCTION

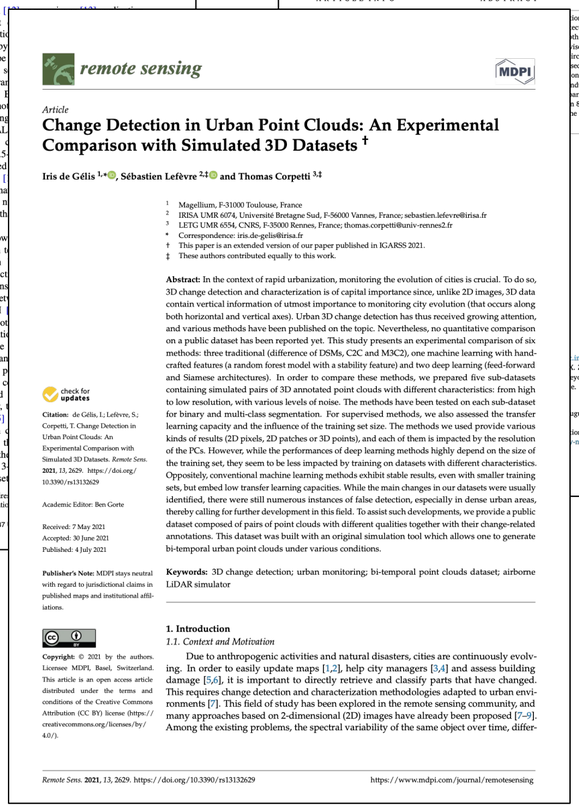
Extreme events such as earthquakes and storms often cause thousands of landslides leading to hazards for local populations [1], and cause severe hillslope and landscape erosion [2]. Landslide analysis is essential to better understand the mechanics of landslides, hillslope erosion [3], and to predict hydro-sedimentary hazards such as alteration of river dynamics due to landslide sediment deposit [4]. To this end, the creation of efficient and exhaustive landslide maps is essential. In mountainous areas, topographic changes are difficult to localize because of the access difficulty and the large size of studied areas. Thereby, landslide mapping is performed through 2D satellite or aerial images [3, 5] or 3D Point Clouds (PCs) acquired via Aerial Laser Scanning (ALS) or photogrammetry [6, 7]. In particular, elevation data ease landslide volume computation and change detection [8] while 2D images may lead to omitted topographic changes due to limited spectral variations or change below forested areas [7]. For this reason, 3D data appear again to be an interesting solution. Yet, the majority of studies rely on 2.5D rasterization of 3D data into Digital Elevation Models (DEMs) of each epoch [8], and then subsequently using a Difference of DEM (DoD) [7, 8]. However, on top of information loss due to the rasterization process (cell size and interpolation), DoD leads to strong errors in very steep surfaces [7]. To overcome these issues, it is important to deal directly with 3D PCs.

In this study, we propose to apply a deep learning method processing the raw PCs to detect landslide sources and deposits. 3D PCs change detection has been tackled by deep learning only recently, with a model called Siamese KPConv [9]. Inspired by both deep learning methods for 2D change detection as well as recent progress in the definition of convolutions dedicated to 3D PCs, Siamese KPConv performs multi-change segmentation directly from raw 3D PCs. Even more recently, Encoder Fusion SiamKPConv outperformed Siamese KPConv thanks to an encoder that fuses both monodate and change information [10]. While both networks have been first developed in urban environments, Siamese KPConv has been successfully applied to erosion detection in steep cliffs [11]. However, landslide detection differs from cliff erosion for several reasons. First, when forested areas are present, incorrect ground classification in the first LiDAR epoch can lead to false positive detections. Second, high variability in slopes steepness and orientation makes the change detection process more complex. Thereby, we propose to experiment and retrieve landslides sources and deposits using Encoder Fusion SiamKPConv deep network on two PC acquisitions made pre- and post-earthquake.

In Section 2, the study area and data are presented. Then, a description of the method is given in Section 3. Finally, results are described and discussed in Section 4.

*Contact author (iris.degelis@orbsgm.fr). I. de Gélis work is funded by Magellium and supported by a CNES grant. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011011754R2 made by GENCI.

Authorized licensed use limited to: UNIVERSITE DE BRETAGNE SUD. Downloaded on November 09,2026 at 23:33:39 UTC from IEEE Xplore. Restrictions apply.



remote sensing

MDPI

Article

Change Detection in Urban Point Clouds: An Experimental Comparison with Simulated 3D Datasets †

Iris de Gélis [1,*], Sébastien Lefèvre [2,‡] and Thomas Corpetti [3,‡]

1 Magellium, F-31000 Toulouse, France
2 IRISA UMR 6074, Université Bretagne Sud, F-56000 Vannes, France; sebastien.lefevre@irisa.fr
3 LETG UMR 6554, CNRS, F-35000 Rennes, France; thomas.corpetti@univ-rennes2.fr
* Correspondence: iris.de-gelis@irisa.fr
† This paper is an extended version of our paper published in IGARSS 2021.
‡ These authors contributed equally to this work.

Abstract: In the context of rapid urbanization, monitoring the evolution of cities is crucial. To do so, 3D change detection and characterization is of capital importance since, unlike 2D images, 3D data contain vertical information of utmost importance to monitoring city evolution (that occurs along both horizontal and vertical axes). Urban 3D change detection has thus received growing attention, and various methods have been published on the topic. Nevertheless, no quantitative comparison on a public dataset has been reported yet. This study presents an experimental comparison of six methods: three traditional (difference of DSMs, C2C and M3C2), one machine learning with handcrafted features (a random forest model with a stability feature) and two deep learning (feed-forward and Siamese architectures). In order to compare these methods, we prepared five sub-datasets containing simulated pairs of 3D annotated point clouds with different characteristics: from high to low resolution, with various levels of noise. The methods have been tested on each sub-dataset for binary and multi-class segmentation. For supervised methods, we also assessed the transfer learning capacity and the influence of the training set size. The methods we used provide various kinds of results (2D pixels, 2D patches or 3D points), and each of them is impacted by the resolution of the PCs. However, while the performances of deep learning methods highly depend on the size of the training set, they seem to be less impacted by training on datasets with different characteristics. Oppositely, conventional machine learning methods exhibit stable results, even with smaller training sets, but embed low transfer learning capacities. While the main changes in our datasets were usually identified, there were still numerous instances of false detection, especially in dense urban areas, thereby calling for further development in this field. To assist such developments, we provide a public dataset composed of pairs of point clouds with different qualities together with their change-related annotations. This dataset was built with an original simulation tool which allows one to generate bi-temporal urban point clouds under various conditions.

Keywords: 3D change detection; urban monitoring; bi-temporal point clouds dataset; airborne LiDAR simulator

1. Introduction

1.1. Context and Motivation

Due to anthropogenic activities and natural disasters, cities are continuously evolving. In order to easily update maps [1,2], help city managers [3,4] and assess building damage [5,6], it is important to directly retrieve and classify parts that have changed. This requires change detection and characterization methodologies adapted to urban environments [7]. This field of study has been explored in the remote sensing community, and many approaches based on 2-dimensional (2D) images have already been proposed [7–9]. Among the existing problems, the spectral variability of the same object over time, differ-

check for updates

Citation: de Gélis, I.; Lefèvre, S.; Corpetti, T. Change Detection in Urban Point Clouds: An Experimental Comparison with Simulated 3D Datasets. *Remote Sens.* **2021**, *13*, 2629. https://doi.org/10.3390/rs13132629

Academic Editor: Ben Gorte

Received: 7 May 2021
Accepted: 30 June 2021
Published: 4 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).



ISPRS Journal of Photogrammetry and Remote Sensing 206 (2023) 168–183

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

DC3DCD: Unsupervised learning for multiclass 3D point cloud change detection

Iris de Gélis [a,b,*], Sébastien Lefèvre [b], Thomas Corpetti [c]

[a] Magellium, Toulouse, F-31000, France
[b] Université Bretagne Sud, IRISA UMR 6074, Vannes, F-56000, France
[c] CNRS, LETG UMR 6554, Rennes, F-35000, France

ARTICLE INFO

Keywords:
3D point clouds
Change detection
Unsupervised deep learning
Deep clustering

ABSTRACT

In a constant evolving world, change detection is of prime importance to keep updated maps. To better sense areas with complex geometry (urban areas in particular), considering 3D data appears to be an interesting alternative to classical 2D images. In this context, 3D point clouds (PCs), whether obtained through LiDAR or photogrammetric techniques, provide valuable information. While recent studies showed the considerable benefit of using deep learning-based methods to detect and characterize changes into raw 3D PCs, these studies rely on large annotated training data to obtain accurate results. The collection of these annotations are tricky and time-consuming. The availability of unsupervised or weakly supervised approaches is then of prime interest. In this paper, we propose an unsupervised method, called DeepCluster 3D Change Detection (DC3DCD), to detect and categorize multiclass changes at point level. We classify our approach in the unsupervised family given the fact that we extract in a completely unsupervised way a number of clusters associated with potential changes. Let us precise that in the end of the process, the user has only to assign a label to each of these clusters to derive the final change map. Our method builds upon the DeepCluster approach, originally designed for image classification, to handle complex raw 3D PCs and perform change segmentation task. An assessment of the method on both simulated and real public dataset is provided. The proposed method allows to outperform fully-supervised traditional machine learning algorithm and to be competitive with fully-supervised deep learning networks applied on rasterization of 3D PCs with a mean of IoU over classes of change of 57.06% and 66.69% for the simulated and the real datasets, respectively. The code is available at https://github.com/idegelis/torch-points3d-dc3dcd.

1. Introduction

While urban environments are continuously and rapidly evolving, change detection between several temporal acquisitions is a way to quickly highlight modified areas in order to update maps (Demir et al., 2012) or to identify damaged objects (Dong and Shan, 2013) in case of natural disasters. With regard to the complexity of urban landscapes, sensing the area using also vertical information appears to be judicious. Indeed, such three-dimensional (3D) information allows to better characterize environment geometry and to avoid two-dimensional (2D) image problems such as the difference of viewing angles between distinct acquisitions, spectral variability of objects over time, perspective, and distortion effects (Qin et al., 2016). 3D data are acquired thanks to LiDAR sensor or photogrammetric process, both resulting in 3D Point Clouds (PCs). No matter the type of acquisition, multi-temporal 3D data are generalizing. Indeed, more and more national mapping agencies opt for full territory Aerial Laser Scanning (ALS) campaign, as in the Netherlands with Actueel Hoogtebestand Nederland (AHN) multi-temporal campaigns (Sande et al., 2010), or in France with the LiDAR high density (HD) project whose objective is to propose a complete 3D high resolution coverage of France with regular updates in the future. In the same time satellites mission for 3D sensing are multiplying, e.g., Pléiades (Bernard et al., 2012), Pléiades Néo (Jérôme, 2019), 3D Optical Constellation (CO3D) (Lebègue et al., 2020) missions. In civil engineering, Terrestrial Laser Scanning (TLS) or unmanned aerial vehicle (UAV) photogrammetry are becoming unavoidable for an accurate sense of complex objects. Thereby, this calls for methods able to analyze these multi-temporal 3D data.

Whether related to urban environment (Stilla and Xu, 2023) or geosciences (Okyay et al., 2019), many studies have been focused on handling 3D data for accurate change extraction. Recently, some

* Corresponding author.
E-mail addresses: iris.de-gelis@irisa.fr (I. de Gélis), sebastien.lefevre@irisa.fr (S. Lefèvre), thomas.corpetti@cnrs.fr (T. Corpetti).

https://doi.org/10.1016/j.isprsjprs.2023.10.022
Received 15 May 2023; Received in revised form 23 September 2023; Accepted 30 October 2023
Available online 16 November 2023
0924-2716/© 2023 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.



ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume V-3-2022
XXIV ISPRS Congress (2022 edition), 6–11 June 2022, Nice, France

CLIFF CHANGE DETECTION USING SIAMESE KPCONV DEEP NETWORK ON 3D POINT CLOUDS

I. de Gélis[1,2,*], Z. Bessin[3,4,*], P. Letourneau[4], M. Jaud[5,3], C. Delacourt[3], S. Costa[6], O. Maquaire[6], R. Davidson[6], T. Corpetti[7], S. Lefèvre[2]

[1] Magellium, F-31000 Toulouse, France
[2] IRISA UMR 6074, Université Bretagne Sud, F-56000 Vannes, France - iris.de-gelis@irisa.fr
[3] Geo-Ocean - UMR 6538, Univ Brest, CNRS, F-29280 Plouzané, France - zoe.bessin@univ-brest.fr
[4] LETG - UMR 6554, Univ Brest, CNRS, F-29280 Plouzané, France
[5] European Institute for Marine Studies (IUEM) - UMS 3113, Univ Brest, CNRS, F-29280 Plouzané, France
[6] IDEES - UMR 6266, Normandie Univ, UNICAEN, CNRS, F-14000 Caen, France
[7] CNRS, LETG UMR 6554, F-35000 Rennes, France

Commission III, WG III/7

KEY WORDS: Erosion detection, change detection, cliffs, 3D point clouds, deep learning

ABSTRACT:

Mainly depending on their lithology, coastal cliffs are prone to changes due to erosion. This erosion could increase due to climate change leading to potential threats for coastal users, assets, or infrastructure. Thus, it is important to be able to understand and characterize cliff face changes at fine scale. Usually, monitoring is conducted thanks to distance computation and manual analysis of each cliff face over 3D point clouds to be able to study 3D dynamics of cliffs. This is time consuming and inclined to each one judgment in particular when dealing with 3D point clouds data. Indeed, 3D point clouds characteristics (sparsity, impossibility of working on a classical top view representation, volume of data, ...) make their processing harder than 2D images. Last decades, an increase of performance of machine learning methods for earth observation purposes has been performed. To the best of our knowledge, deep learning has never been used for 3D change detection and categorization in coastal cliffs. Lately, Siamese KPConv brings successful results for change detection and categorization into 3D point clouds in urban area. Although the case study is different by its more random characteristics and its complex geometry, we demonstrate here that this method also allows to extract and categorize changes on coastal cliff face. Results over the study area of Petit Ailly cliffs in Varengeville-sur-Mer (France) are very promising qualitatively as well as quantitatively: erosion is retrieved with an intersection over union score of 83.86 %.

1. INTRODUCTION

The dynamics of cliff erosion is a complex phenomenon triggered by various factors whose relative contribution is still difficult to estimate. Since cliff erosion is likely to increase with sea level rise due to climate change (Slott et al., 2006; Ashton et al., 2011; Masson-Delmotte et al., 2021), understanding changes in coastal cliffs in order to better manage them would ensure the safety of communities and infrastructure threatened by erosion. In cliff erosion, we distinguish mass movements and debris falls. According to Varnes (1978), what we call debris falls refer to tiny rocks falling from the cliff face, while mass movements refer to rock falls corresponding to larger scale movements of part or all of the cliff. To quantify cliff retreat rates over decades, the simplest way is to measure the distance from different cliff top locations using aerial photographs or historical maps (Lee and Clark, 2002; Brooks and Spencer, 2010). But since changes at the cliff top may not be representative of changes over the entire cliff face (Young et al., 2009), comparisons between cliff face 3D Point Clouds (PCs) are also used. For this second approach, methods such as terrestrial laser scanning (TLS), airborne laser scanning (ALS), mobile laser scanning, unmanned aerial vehicle photogrammetry (UAVP), or terrestrial photogrammetry (TP) are used to monitor the cliff face (Young et al., 2010; James and Robson, 2012; Michoud et al., 2014; Letortu et al., 2018). Surveys are taken over time and PCs are compared using tools such as Cloud to Cloud (C2C) (Girardeau-Montaut et al., 2005) in CloudCompare software (Girardeau-Montaut, 2016). Manual analysis of the differences is then performed to determine areas of erosion on the cliff face and accumulation at the cliff foot. PCs differences provide cliff retreat rates and allow for estimation of eroded volume (Letortu et al., 2015). Different methods trying to automate cliff faces changes extraction and cliff top and toe delineation exist, however most of them do not process 3D data directly but 2.5D rasterization of PCs and finally use Digital Elevation Model (DEM) differences (Young and Ashford, 2006; Swirad and Young, 2021). For example, in Young and Ashford (2006), the elevation variability is calculated through DEM difference where negative cells represent erosion and positive cells represent accumulation, enabling to highlight significant erosion areas on cliffs with slight slope. In Swirad and Young (2021), inventories of erosion and deposition objects are retrieved with a method consisting in four steps: 1) PC processing, 2) cliff face identification, 3) change object inventories, and 4) object classification. Erosion rates are calculated using DEM 2.5D analysis combined with vertical and planimetric detection thresholds, Normalized Difference Vegetation Index (NDVI), machine learning processes and manual quality control. These methods, although relevant, are not applicable on vertical or very steep cliffs. A direct 3D approach allows to study vertical cliffs and to limit the loss of information generated by the data rasterization.

Nowadays deep learning is the leading framework in classification and even change identification in remote sensing data. Recently, Siamese KPConv network was proposed to highlight and

* Corresponding authors

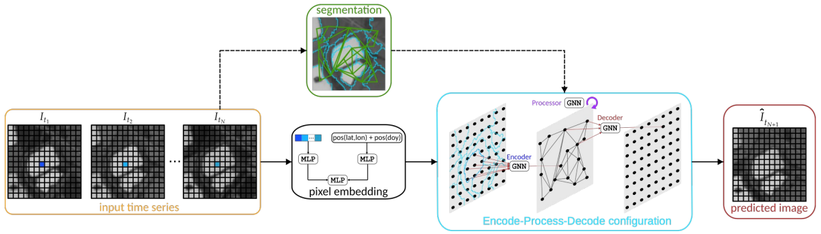
This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2022-649-2022 | © Author(s) 2022. CC BY 4.0 License. 649

GRAPHS NEURAL NETWORKS FOR SATELLITE IMAGE TIME SERIES

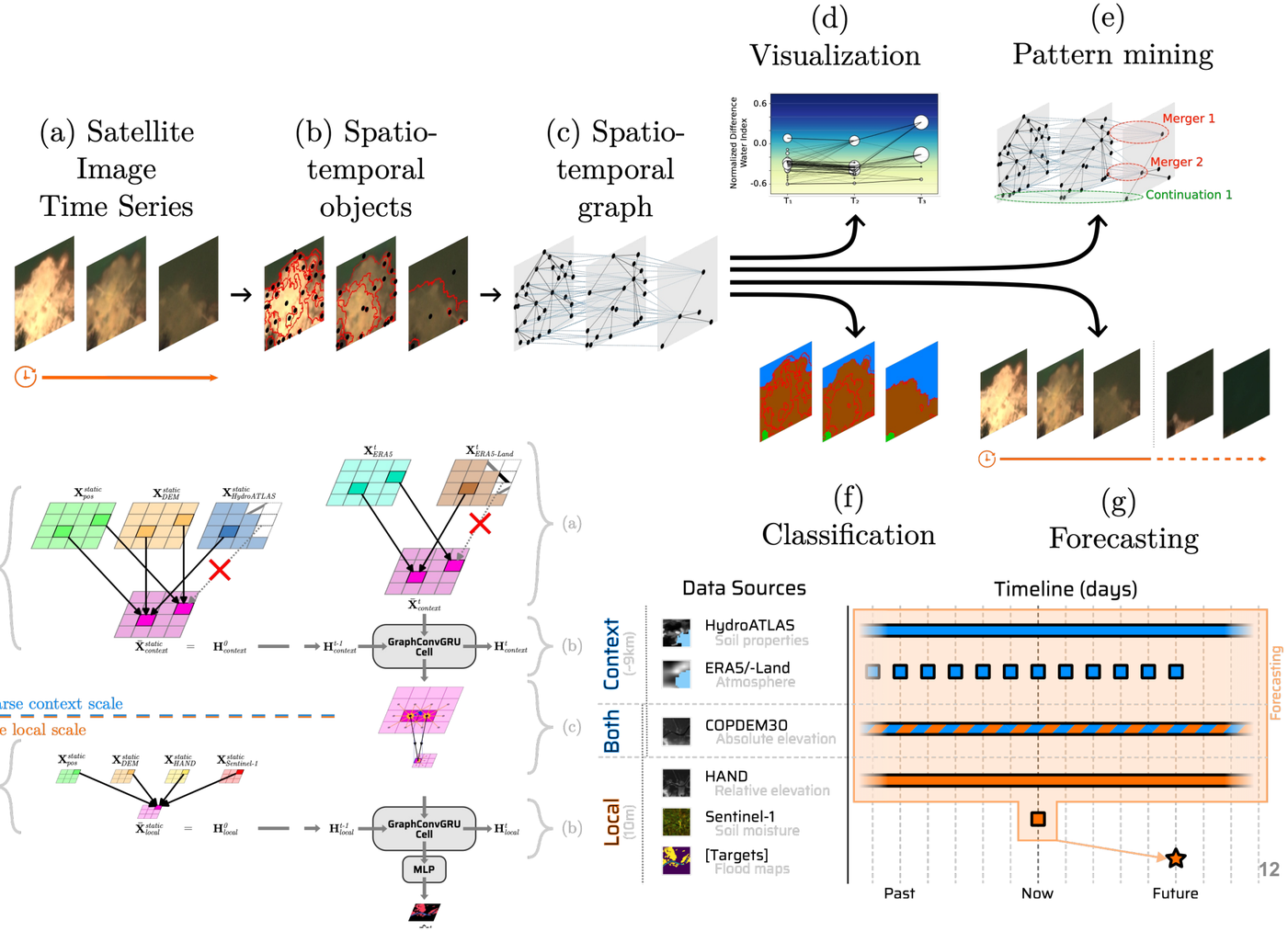


Corentin Dufourg PhD
with CNES (2022-2025) and ESA mobility

- Benchmarking GNNs
- Novel models including some adapted to multiscale/multisource data
- Various applications from mapping (land cover) to forecasting (water resource and floods)



	# Params	RMSE ↓	PSNR ↑	SSIM ↑	Runtime (min)
Input average	-	0.1550	23.32	0.7465	-
Persistence	-	0.1332	25.03	0.7897	-
LSTM	17,345	$0.1162_{\pm 0.0005}$	$25.53_{\pm 0.05}$	$\mathbf{0.8282_{\pm 0.0005}}$	22
ConvLSTM	150,721	$0.1197_{\pm 0.0029}$	$25.28_{\pm 0.19}$	$0.8113_{\pm 0.0030}$	26
TDCNN-ConvLSTM	407,681	$0.1111_{\pm 0.0008}$	$25.68_{\pm 0.08}$	$0.8083_{\pm 0.0008}$	46
Graph-based	228,673	$\mathbf{0.1097_{\pm 0.0035}}$	$\mathbf{26.42_{\pm 0.27}}$	$0.8170_{\pm 0.0070}$	41



GRAPHS NEURAL NETWORKS FOR SATELLITE IMAGE TIME SERIES



Corentin Dufourg PhD
with CNES (2022-2025) and ESA mobility

- Benchmarking GNNs
- Novel models including some adapted to multiscale/multisource data
- Various applications from mapping (land cover) to forecasting (water resource and floods)



On the Use of Graphs for Satellite Image Time Series

A comprehensive review

CORENTIN DUFOURG, CHARLOTTE PELLETIER, STÉPHANE MAY, AND SÉBASTIEN LEFÈVRE

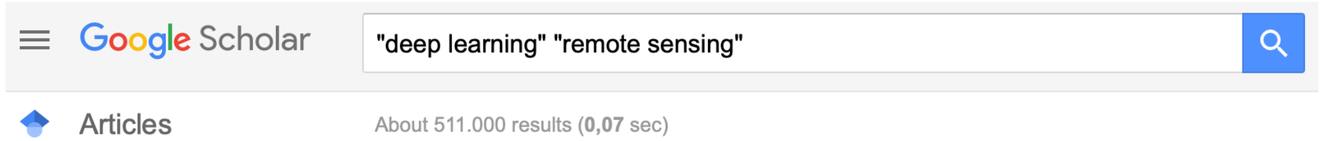
The Earth's surface is subject to complex and dynamic processes, ranging from large-scale phenomena, such as tectonic plate movements, to localized changes associated with ecosystems, agriculture, or human activity. Satellite images enable global monitoring of these processes with extensive spatial and temporal coverage, offering advantages over in situ methods. In particular, the resulting satellite image time series (SITS) datasets contain valuable information. To handle their large volume and complexity, some recent works focus on the use of graph-based techniques that abandon the regular Euclidean structure of satellite data to work at an object level. Graphs also enable the modeling of spatial and temporal interactions between identified objects, which are crucial for pattern detection, classification, and regression tasks. This article is an effort to examine the integration of graph-based methods in spatiotemporal remote sensing analysis. In particular, it aims to present a versatile graph-based pipeline to tackle SITS analysis. It focuses on the construction of spatiotemporal

Digital Object Identifier 10.1109/MGRS.2025.3622200
Date of publication 10 December 2025; date of current version 6 February 2026.

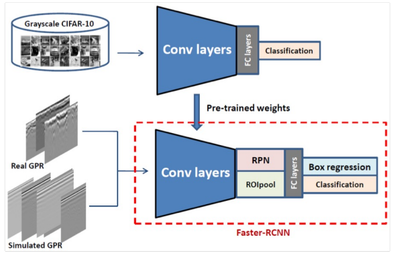
2473-2397 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.
Authorized licensed use limited to: UNIVERSITE DE BRETAGNE SUD. Downloaded on February 23,2026 at 14:04:48 UTC from IEEE Xplore. Restrictions apply.

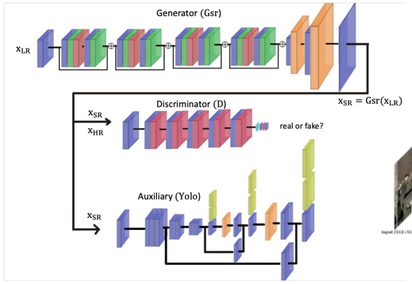
WHY DEEP LEARNING IS NOT ENOUGH?

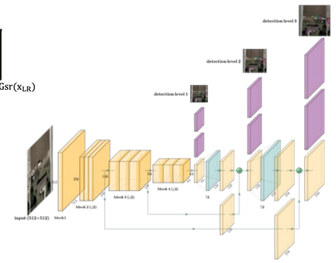
Deep learning has gained a high interest



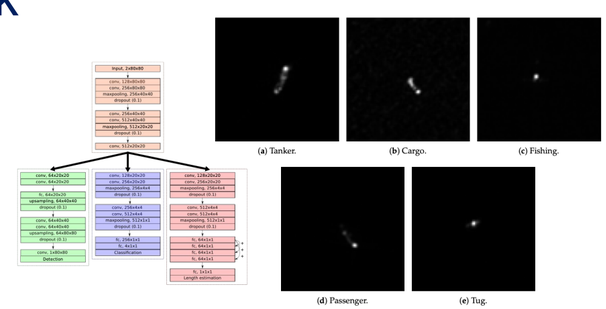
Every new task / new use case / new modality requires possibly a new network

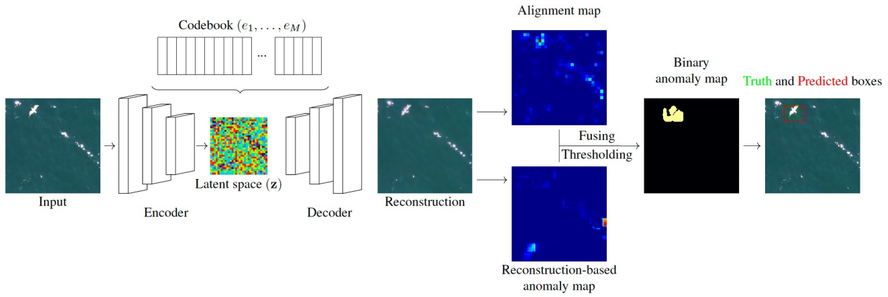


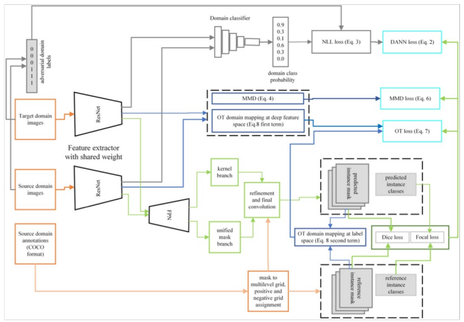


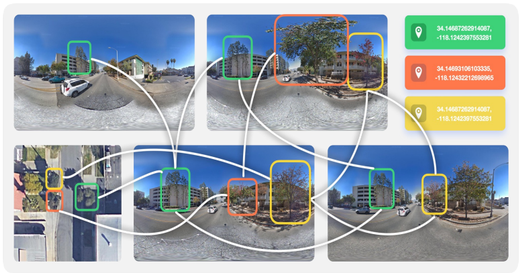














Labelling data is a mandatory, time-consuming, and complex step

GEOSPATIAL FOUNDATION MODELS

THE RECENT TREND IN AI AND REMOTE SENSING…

(BEFORE THE NEXT BREAKTHROUGHS: AGENTIC AI, WORLD MODELS, …)

FOUNDATION MODELS, THE NEXT AI GENERATION

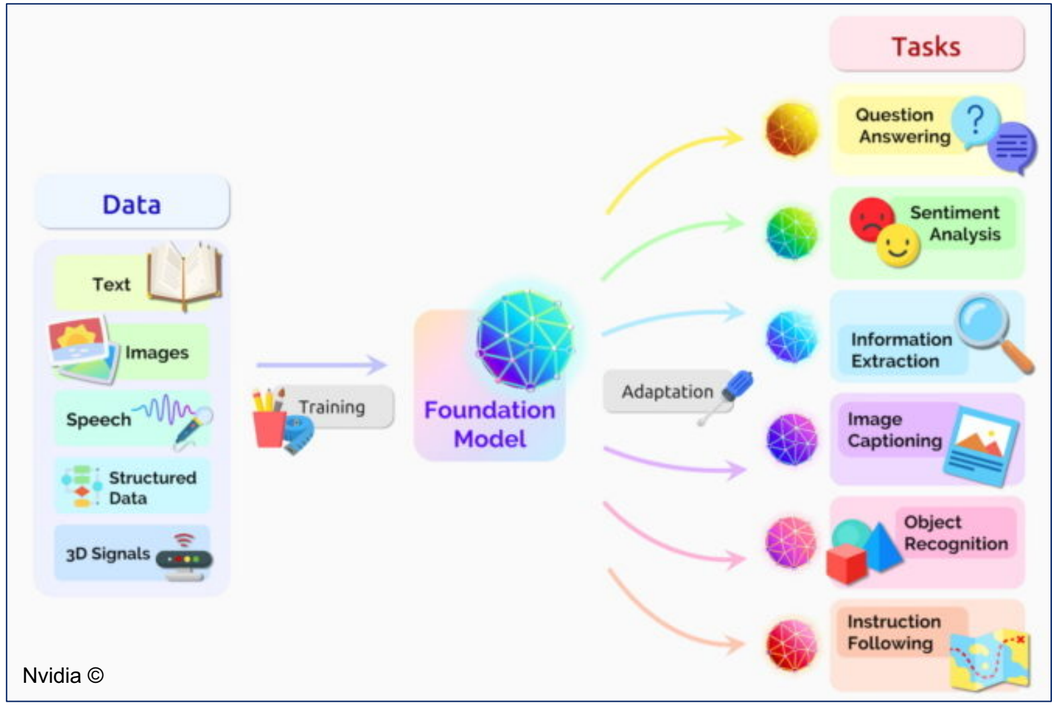
Definition from Stanford Institute for Human-Centered Artificial Intelligence's (HAI) Center for Research on Foundation Models (CRFM) in August 2021:

> any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks

Deep learning does not need handcraft features
but still requires handcrafting the network architecture

Foundation models (FM) solved the problem

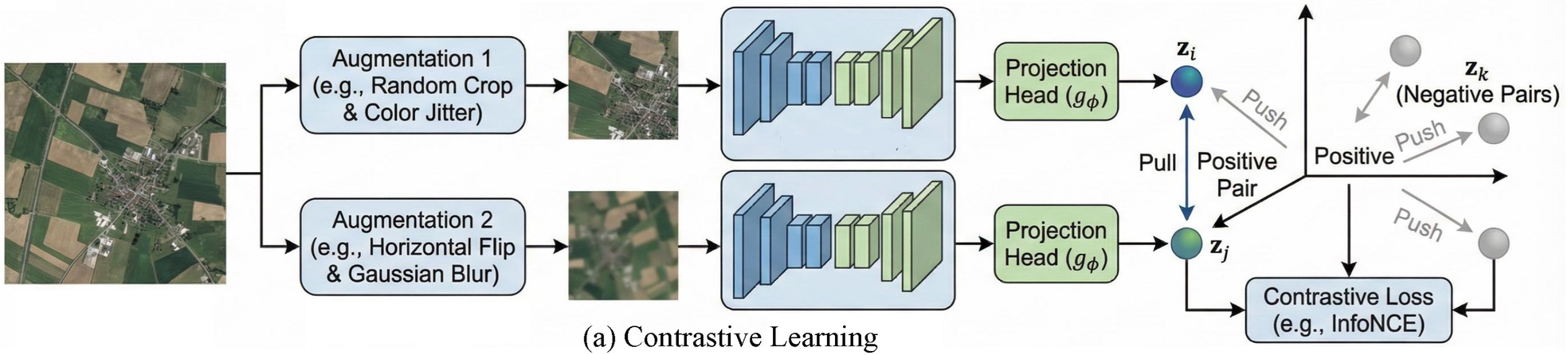
- No need to design a complex architecture
- No need to retrain a network to learn the features
- Adaptation at a minimum cost (in theory)



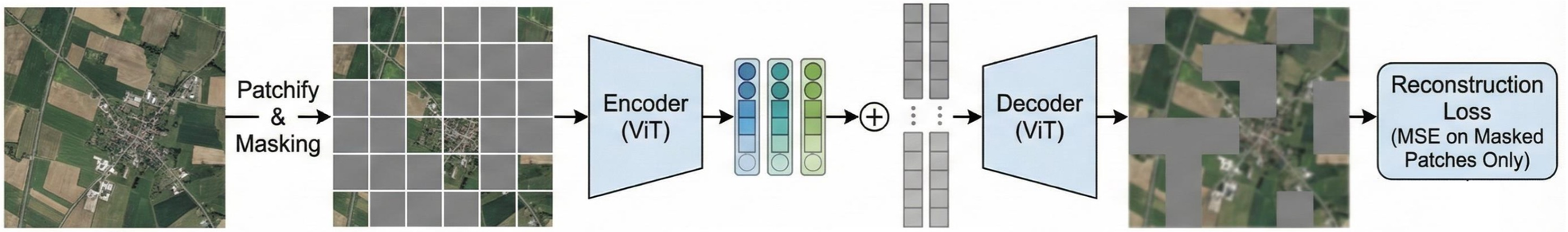
Nvidia ©

SELF-SUPERVISED LEARNING…

Various strategies to training only with unlabeled data



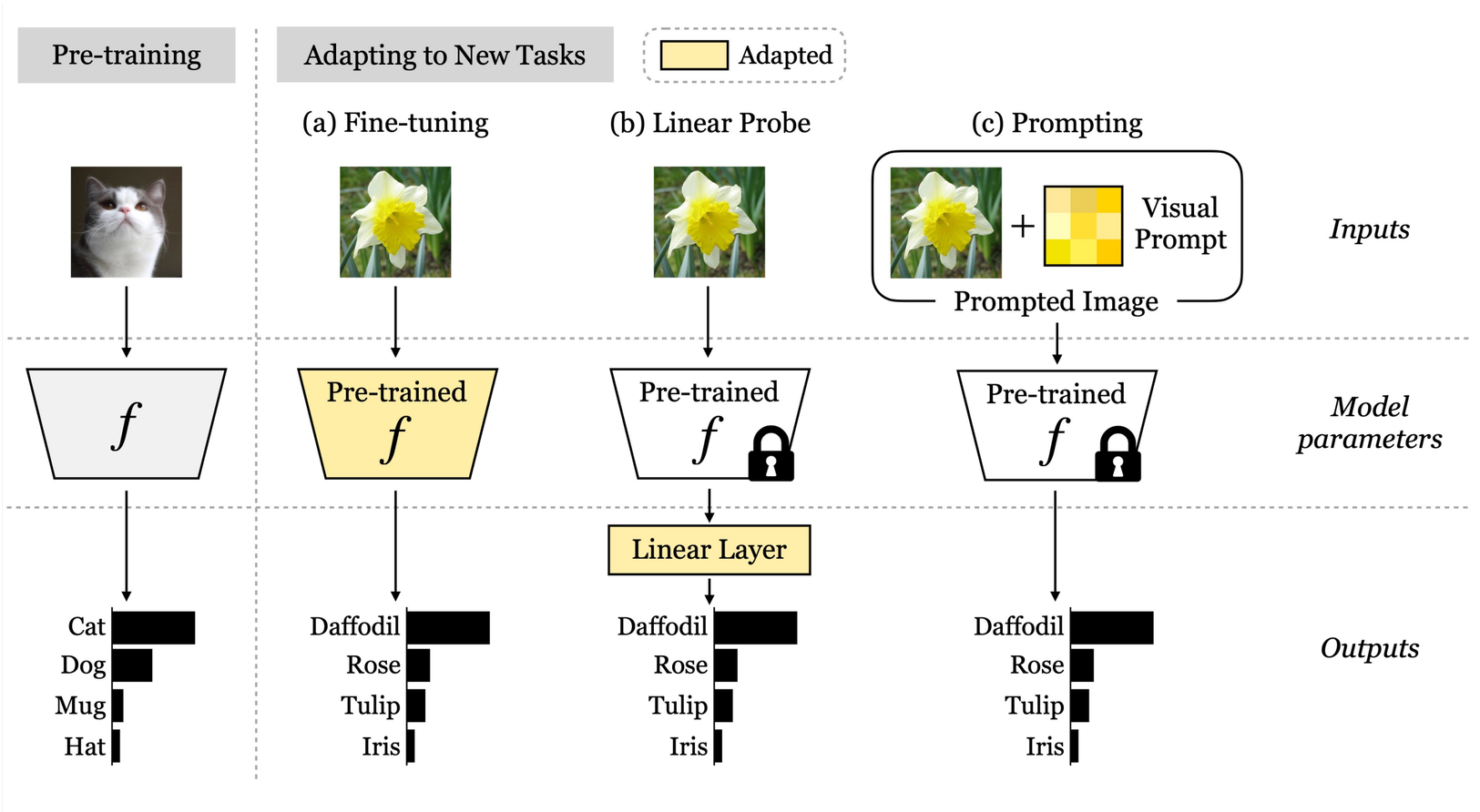
(a) Contrastive Learning



(b) Masked Image Modeling

Sang et al, 2026

… MOST OFTEN FOLLOWED TO ADAPTION TO DOWNSTREAM TASKS



Bang et al, 2022

FOUNDATION MODELS HAVE BEEN ADOPTED IN EO

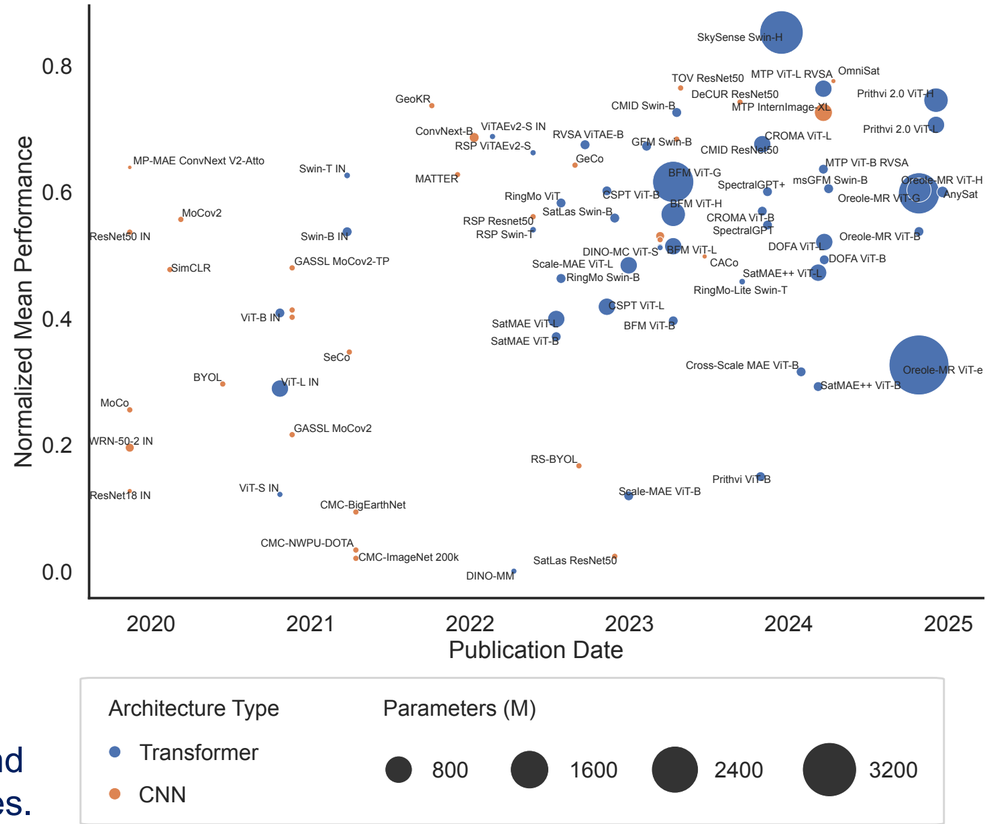
The new breakthrough in EO, as with ML on handcraft features, and then with DL

The promise: “A single generic model fine-tuned on specific tasks”

An increasing field of research…
with more than 200 models in 5 years!



Nicolas Longepe, Hamed Alemohammad, Anca Anghelea, et al. *Earth Action in Transition: Highlights from the 2025 ESA-NASA International Workshop on AI Foundation Models for EO. accepted*, IEEE Magazine of Geosciences and Remote Sensing. 2025. DOI: DOI: 10.1109/MGRS.2025.3592035



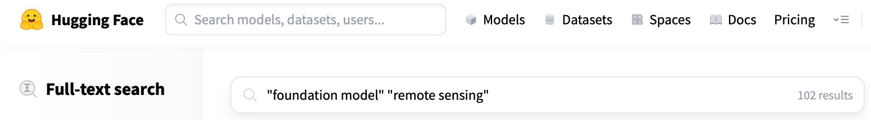
Over 850 people joined the first International Workshop on AI Foundation Model (FM) for Earth Observation (EO) co-organized by ESA and NASA 5-7 May 2025. Hosted at ESRIN (ESA’s Earth Observation Center, Italy), the event welcomed around 300 people on site, and an additional 550 online, with the promise that FMs can revolutionize EO and Earth sciences.

NUMEROUS PUBLICATIONS, CODES, MODELS, AND SERVICES ARE AVAILABLE

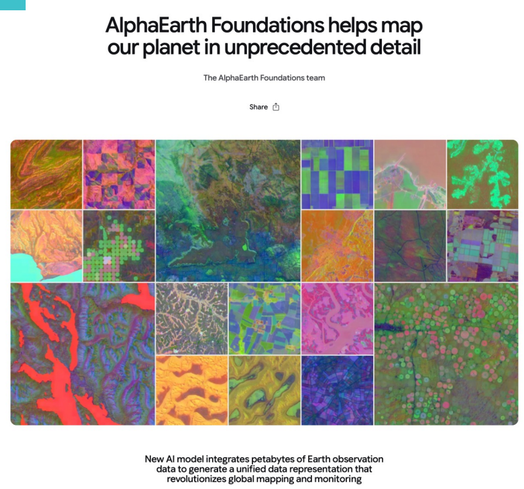
Not reaching 1M papers but it will reach it!

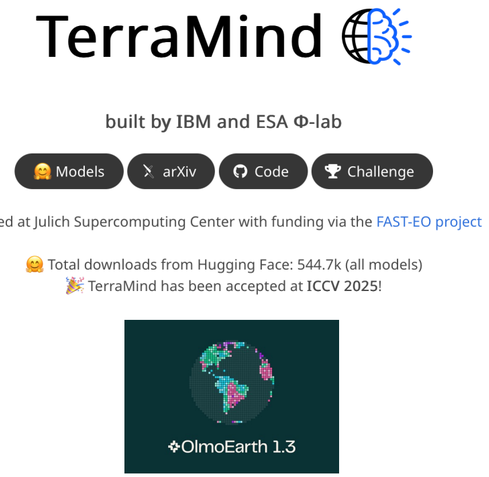
Increasing interest in general EO venues

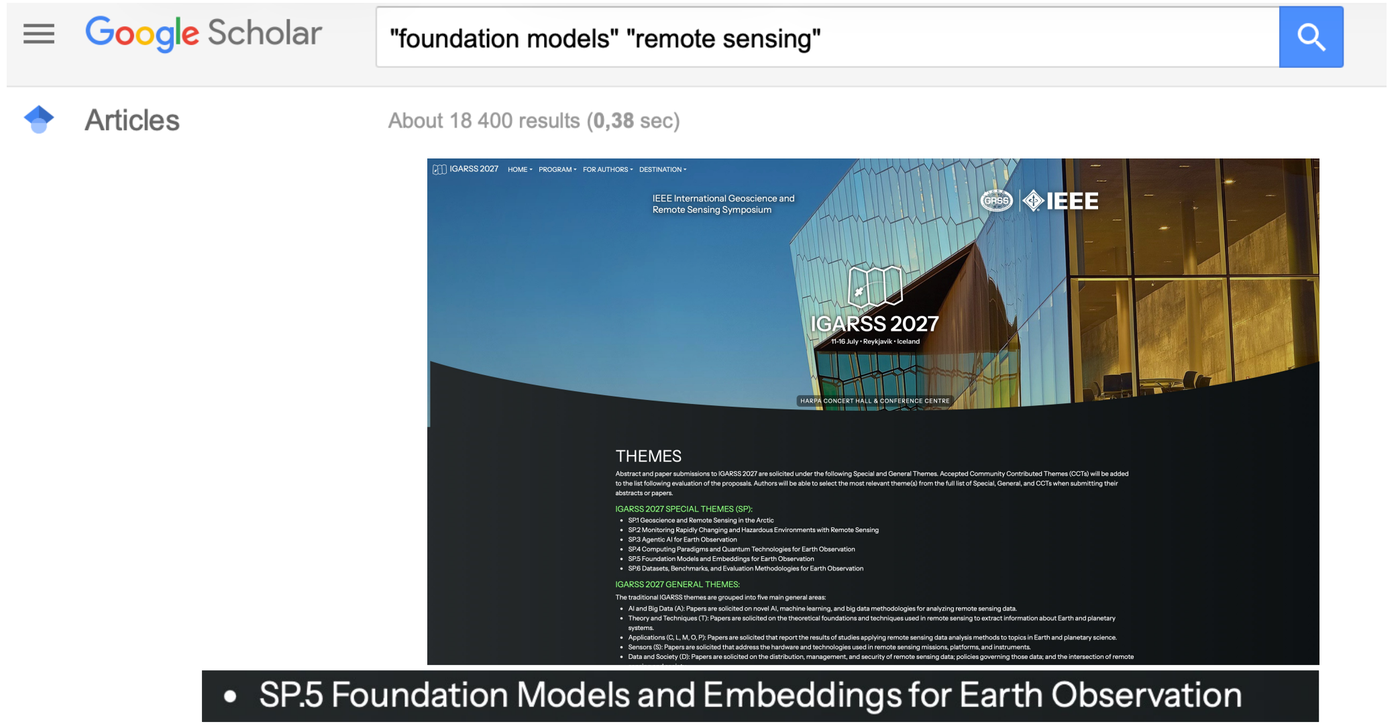
More and more models publicly available



Far beyond pure academic research







- SP.5 Foundation Models and Embeddings for Earth Observation

Looks like the field is mature enough…

So what's wrong?

NUMEROUS PUBLICATIONS, CODES, MODELS, AND SERVICES ARE AVAILABLE

GFMs are not addressing EO data and needs (yet)

GFMs are not sustainable (yet)

Looks like the field is mature enough…

So what's wrong?

GFM AND SUSTAINABILITY?

The question of sustainability occurs in various steps of the GFM life cycle

- When benchmarking to determine what model to use
- When adapting a GFM to a specific task
- When pretraining a new GFM



Pierre Adorni joint PhD with CNES (French space agency)



Giovanni Castiglioni joint PhD with U. Chile

SCALING LAWS

Joint work with ESA Phi-lab
(to be presented at GeoCV@WACV 2027)

Model performance usually related to model/dataset size?

Dataset is as important as model, calling for more data-centric machine learning…

Scaling Laws for Geospatial Foundation Models: A case study on PhilEO Bench

Nikolaos Dionelis[1], Riccardo Musto[2], Jente Bosmans[1,3], Simone Sarti[2], Giancarlo Paoletti[2], Peter Naylor[1], Valerio Marsocci[1], Sébastien Lefèvre[4], Bertrand Le Saux[1,5], Nicolas Longépé[1]*

1. European Space Agency (ESA), Φ-lab.
2. Leonardo Labs, Italy.
3. VITO.
4. IRISA, Université Bretagne Sud.
5. European Commission (EC)

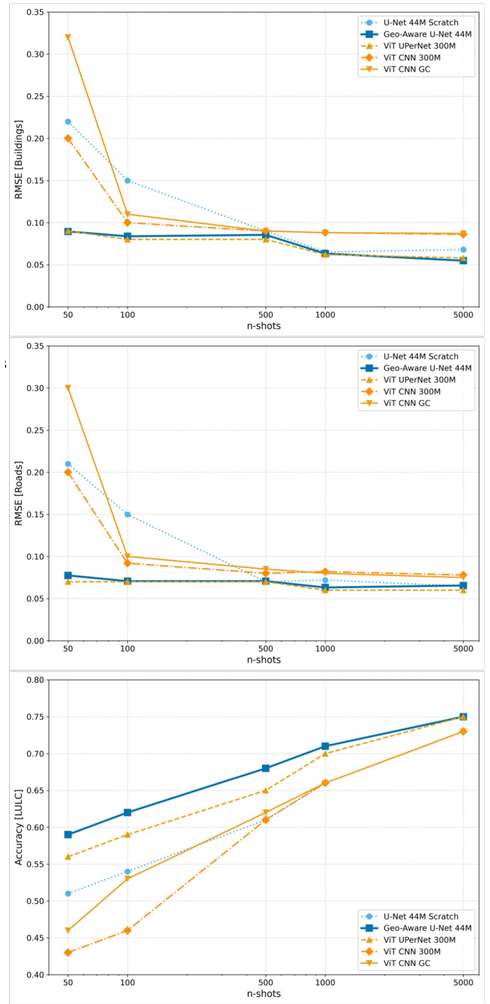


Figure 1. Performance on PhilEO benchmark tasks for models pretrained on GlobeEO 0.5TB: Building density regression (top), Road density estimation (middle), and Land cover mapping (bottom), across different n-shot settings, where the key takeaways are that pretraining helps, i.e. Geo-Aware U-Net 44M versus U-Net 44M (Scratch), and the UPerNet decoder helps, i.e. ViT UPerNet compared to ViT CNN, where the latter means that a CNN decoder is used rather than multi-scale features in the UPerNet decoder.

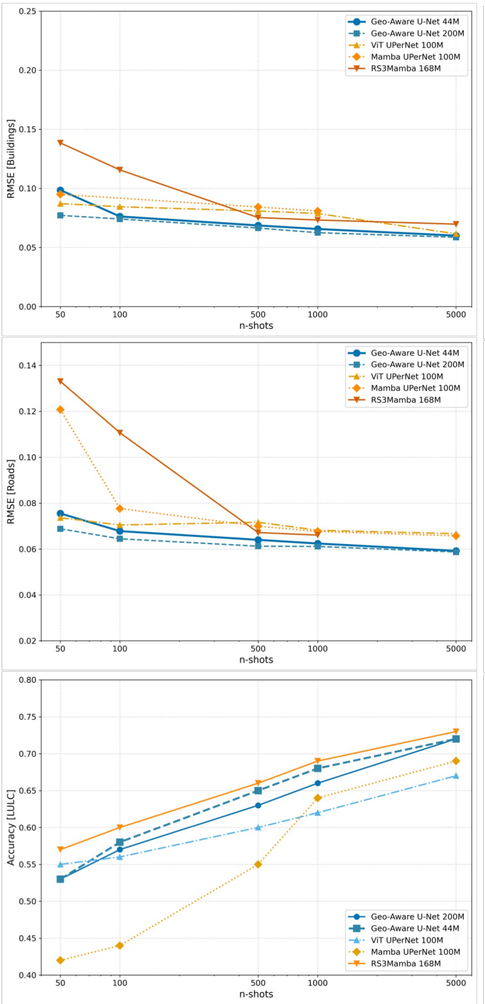


Figure 2. Evaluation over PhilEO Bench downstream tasks for various n-shots and for different model pretraining strategies on FastTOM 2TB: Building density regression (top), Road density estimation (middle), and Land cover mapping (bottom), where the key takeaways are that for pixel-wise regression downsntream tasks, U-Net 200M for most n-shots outperforms the other models, and for semantic segmentation land cover mapping, the RS3Mamba is effective.

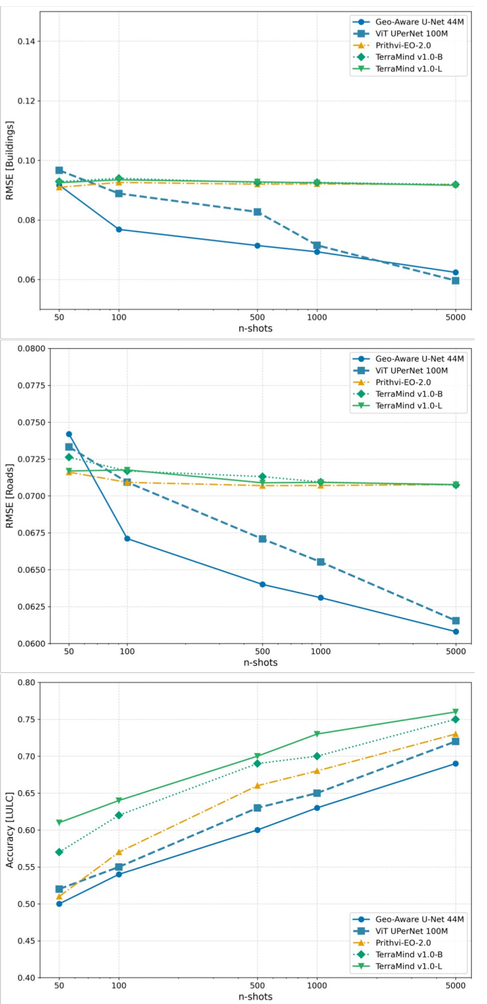


Figure 3. PhilEO benchmark results for models pretrained on MajorTOM 23TB: Building density regression (top), Road density estimation (middle), and Land cover mapping (bottom), across n-shot settings, where the key takeaways are that for pixel-wise regression tasks, U-Net 44M for most n-shots has good performance, and for large n-shots, ViT has competitive performance. We also note that for semantic segmentation downstream tasks, ViT UPerNet outperforms U-Net.

SCALING LAWS VS SUSTAINABILITY

Scaling laws have been established for foundation models…but it challenges sustainability!

- SatMAE (2022) **768** V100-hours
- Terramind (2025) **4,608** A100-hours
- Skysense (2024) **24,600** A100-hours
- DINO-v3 (2025) **61,440** H100-hours

Can we achieve same performances with a reasonable/affordable compute?

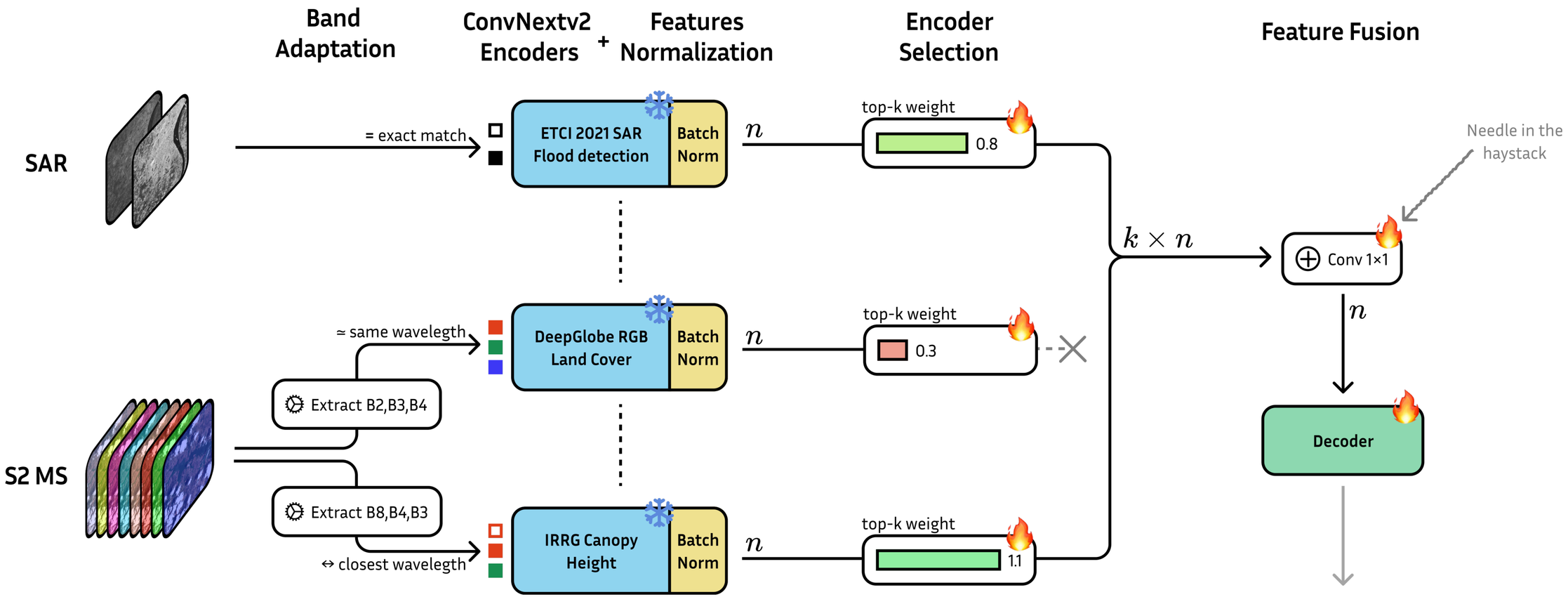
EoS-FM (presented at ECCV in September 2026) does it in 79 H100-hours!

EoS-FM: Can an Ensemble of Specialist Models act as a Generalist Feature Extractor?

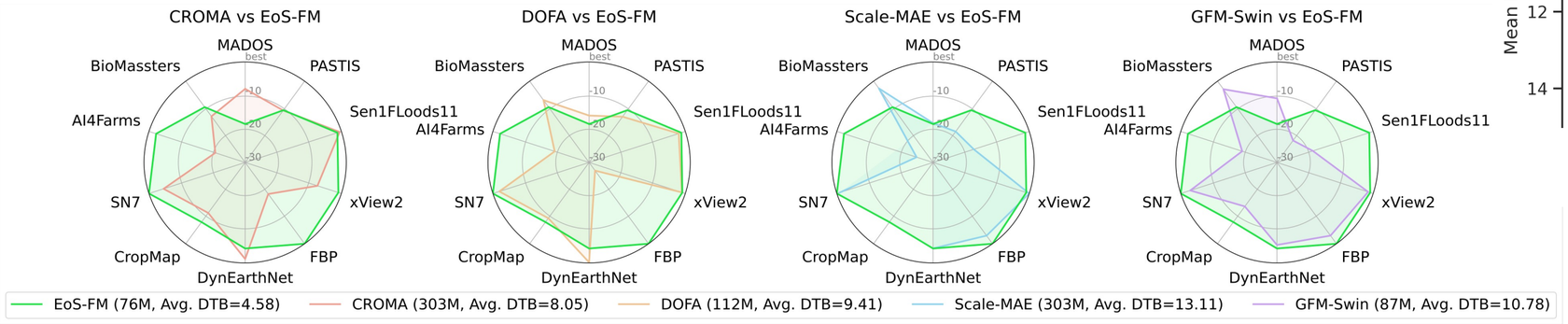
Pierre Adorni[1]✉, Minh-Tan Pham[1], Stéphane May[2], Sébastien Lefèvre[1,3]

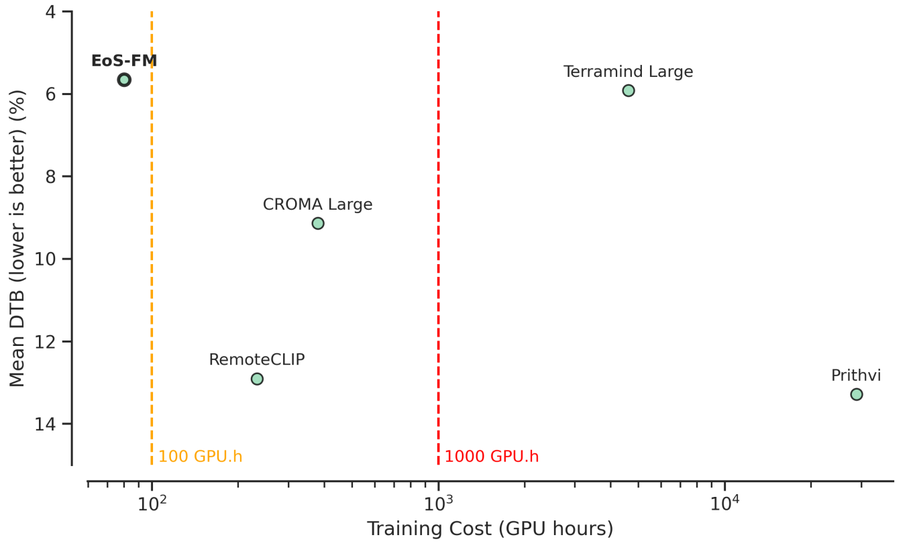
SUSTAINABLE GFM TRAINING = USE MANY SMALL EXPERT MODELS

Instead of pre-training a huge model on a huge dataset, why not relying on small specialized models?



RESULTS





Model	HLS	MAD	PAS	S1F	xV2	FBP	DEN	CM	SN7	AI4	BM ↓	Avg DTB ↓
Scale-MAE (303M) [32] ❄	76.68	57.32	24.55	74.13	**60.72**	67.19	35.11	25.42	62.96	21.47	47.15	13.11
Prithvi (87M) [2] ❄	83.62	49.98	33.93	90.37	49.35	46.81	27.86	43.07	56.54	26.86	39.99	12.20
RemoteCLIP (87M) [25] ❄	76.59	60.00	18.23	74.26	57.41	69.19	31.78	52.05	57.76	25.12	49.79	11.82
SatlasNet (87M) [3] ❄	79.96	55.86	17.51	90.30	52.23	50.97	36.31	46.97	61.88	25.13	41.67	11.64
S12-MAE (22M) [41] ❄	81.91	49.90	32.03	87.79	50.44	51.92	34.08	45.80	57.13	24.69	41.07	11.56
SpectralGPT (105M) [21] ❄	80.47	57.99	35.44	89.07	48.40	33.42	37.85	46.95	58.86	26.75	**36.11**	11.23
S12-Data2Vec (22M) [41] ❄	81.91	44.36	34.32	88.15	51.36	48.82	35.90	54.03	58.23	24.23	41.91	11.20
GFM-Swin (87M) [29] ❄	76.90	64.71	21.24	72.60	59.15	67.18	34.09	46.98	60.89	27.19	46.83	10.97
S12-DINO (22M) [41] ❄	81.72	49.37	36.18	88.61	50.56	51.15	34.81	48.66	56.47	25.62	41.23	10.78
S12-MoCo (22M) [41] ❄	81.58	51.76	34.49	89.26	51.59	53.02	35.44	48.58	57.64	25.38	40.21	10.37
DOFA (112M) [53] ❄	80.63	59.58	30.02	89.37	59.64	43.18	39.29	51.33	61.84	27.07	42.81	9.41
Thor Large (303M) [14]❄	79.47	53.73	39.88	89.55		47.62	37.29	**60.75**	59.71	26.75		9.25
CROMA (303M) [16] ❄	82.42	67.55	32.32	90.89	53.27	51.83	38.29	49.38	59.28	25.65	36.81	8.05
TerraMind Large (303M) [23] ❄	82.93	**75.57**	**43.13**	90.78		63.38	37.89	55.04	59.98	27.47		4.65
EoS-FM (76 M) ❄	87.01	61.96	30.30	89.32	59.79	**70.18**	35.15	52.76	**63.84**	44.1	40.22	**4.58**
EoS-FM Small (22M) ❄	**87.34**	62.93	19.35	90.29	57.56	65.94	32.95	49.14	63.17	38.46	40.20	5.13
UNet (~8M) [33]🔥	84.51	54.79	31.60	**91.42**	58.68	60.47	**39.46**	47.57	62.09	**46.34**	40.39	6.01
ViT-B (87M) [12] 🔥	81.58	48.19	38.53	87.66	57.43	59.32	36.83	44.08	52.57	38.37	38.55	8.78

Table 1: Results on the 11 downstream tasks of the Pangaea benchmark [28]. Our method has the best or second best performance on 5 different datasets, and exhibit the best average performance, on par with TerraMind Large. (DTB = Distance To Best). ↓ Means lower is better.

Model	HLS	MAD	PAS	S1F	xV2	FBP	DEN	CM	SN7	AI4	BM ↓	Avg DTB ↓
RemoteCLIP (87M) ❄	69.4	20.57	17.19	62.22	53.75	56.23	34.43	19.86	43.11	23.85	53.32	15.95
Scale-MAE (303M) ❄	75.47	21.47	22.86	64.74	**56.06**	48.75	35.27	13.44	49.68	26.66	54.16	14.77
GFM-Swin (87M) ❄	67.23	28.19	21.47	62.57	53.45	55.58	28.16	27.21	39.48	32.88	49.3	14.17
SatlasNet (87M) ❄	74.79	29.87	16.76	83.92	44.07	37.86	34.64	29.08	49.78	13.91	44.38	13.86
S12-Data2Vec (22M) ❄	74.38	17.86	33.09	81.91	41.6	37.27	33.63	34.11	40.66	22.85	46.52	13.81
S12-MAE (22M) ❄	76.6	18.44	31.06	84.81	39.84	35.56	30.59	35.29	40.51	23.6	43.76	13.65
Prithvi (87M) ❄	77.73	21.24	33.56	86.28	35.08	29.98	32.28	27.71	36.78	35.04	41.19	13.48
SpectralGPT (105M) ❄	71.82	20.29	34.53	83.12	35.81	39.51	35.33	31.06	36.31	37.35	**39.44**	12.46
DOFA (112M) ❄	71.98	23.77	27.68	82.84	55.6	27.82	**39.15**	29.91	46.1	27.74	46.03	12.38
S12-MoCo (22M) ❄	73.11	19.47	32.51	79.58	41.15	35.57	32.24	36.54	49.46	37.97	44.83	11.82
S12-DINO (22M) ❄	75.93	23.47	36.62	84.95	41.02	34.63	32.78	38.44	41.15	37.91	42.74	10.78
CROMA (303M) ❄	76.44	32.44	32.8	87.22	46.54	37.39	36.08	36.77	42.15	38.48	40.25	8.79
Terramind Base (86M) ❄	77.39	44.06	**39.96**	84.43		54.00	37.35	35.65	43.21	38.59		5.72
Thor Base (86M) ❄	76.90	40.67	38.93	86.29		42.80	35.21	**42.23**	55.94	38.90		5.36
EoS-FM (76 M) ❄	81.69	44.13	31.94	82.24	53.92	**66.73**	**39.79**	23.61	**61.32**	38.03	41.05	**3.67**
EoS-FM Small (22M) ❄	**82.97**	**45.38**	15.18	86.45	51.42	62.59	34.84	20.16	52.38	37.46	51.67	7.78
UNet (~8M) 🔥	79.46	24.3	29.53	**88.55**	46.77	52.58	35.59	13.88	46.08	34.84	40.39	10.14
ViT-B (87M) 🔥	75.92	10.18	38.44	81.85	44.85	56.53	35.39	27.76	36.01	**39.20**	44.89	11.05

Table 3: Results on the 11 downstream tasks of the Pangaea benchmark [28], using only **10% of the training data**. Our method has the best performance on five different datasets, and exhibit the best average performance. ↓ means lower is better.

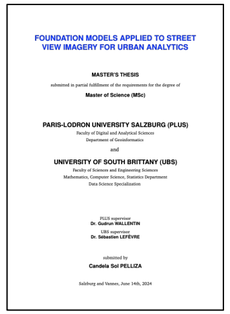
DTB = Distance To Best,

part of our sustainable benchmarking strategy

Towards Efficient Benchmarking of Foundation Models in Remote Sensing: A Capabilities Encoding Approach

Pierre Adorni[1], Minh-Tan Pham[1], Stéphane May[2], Sébastien Lefèvre[1,3]

CLOSING THE GAP BETWEEN GFMS & EO DATA AND NEEDS THROUGH EU-FUNDED MSC


FOUNDATION MODELS APPLIED TO STREET VIEW IMAGERY FOR URBAN ANALYTICS

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Master of Science (MSc)

PARIS-LODRON UNIVERSITY SALZBURG (PLUS)

Faculty of Digital and Analytical Sciences

Department of Geoinformatics

and

UNIVERSITY OF SOUTH BRITTANY (UBS)

Faculty of Sciences and Engineering Sciences

Mathematics, Computer Science, Statistics Department

Data Science Specialization

PLUS supervisor

Dr. Gudrun WALLENTIN

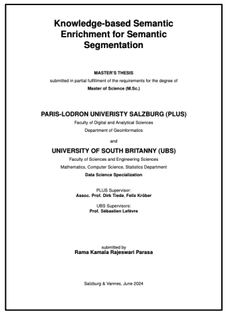
UBS supervisor

Dr. Sébastien LEFÈVRE

submitted by

Candela Sol PELLIZA

Salzburg and Vannes, June 14th, 2024


Knowledge-based Semantic Enrichment for Semantic Segmentation

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Master of Science (M.Sc.)

PARIS-LODRON UNIVERSITY SALZBURG (PLUS)

Faculty of Digital and Analytical Sciences

Department of Geoinformatics

and

UNIVERSITY OF SOUTH BRITANNY (UBS)

Faculty of Sciences and Engineering Sciences

Mathematics, Computer Science, Statistics Department

Data Science Specialization

PLUS Supervisor:

Assoc. Prof. Dirk Tiede, Felix Kröber

UBS Supervisors

Prof. Sébastien Lefèvre

submitted by

Rama Kamala Rajeswari Parasa

Salzburg & Vannes, June 2024


Benchmarking and Adapting Geospatial Foundation Models for Very High Resolution Earth Observation

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Master of Science (MSc)

UNIVERSITY OF SALZBURG

Faculty of Digital and Analytical Sciences

Department of Geoinformatics

and

UNIVERSITY OF SOUTH BRITTANY

Faculty of Sciences and Engineering Sciences

Mathematics, Computer Science, Statistics Department

Data Science Specialization

University of South Brittany supervisors: Pierre Adorni, Minh-Tan Pham, Sébastien Lefèvre

University of Salzburg supervisor: Dirk Tiede

submitted by

Kshitij Raj Sharma

Salzburg & Vannes, September 2026


Assessing the Robustness of TerraMind to Sensor Domain Gap for On-board Processing on the Φ-sat-2 Mission

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Master of Science (MSc)

UNIVERSITY OF SALZBURG

Faculty of Digital and Analytical Sciences

Department of Geoinformatics

and

UNIVERSITY OF SOUTH BRITTANY

Faculty of Sciences and Engineering Sciences

Mathematics, Computer Science, Statistics Department

Data Science Specialization

University of South Brittany supervisor: Prof. Sébastien Lefèvre

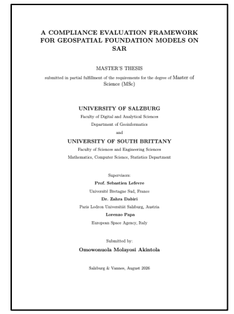
University of Salzburg supervisor: Assoc.- Prof. Hervé Nelissen

Other supervisor: Nicolas Longépé, ESA Supervisor

submitted by

Elouann Lucas

Salzburg & Vannes, August 2025


A COMPLIANCE EVALUATION FRAMEWORK FOR GEOSPATIAL FOUNDATION MODELS ON SAR

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of Master of Science (MSc)

UNIVERSITY OF SALZBURG

Faculty of Digital and Analytical Sciences

Department of Geoinformatics

and

UNIVERSITY OF SOUTH BRITTANY

Faculty of Sciences and Engineering Sciences

Mathematics, Computer Science, Statistics Department

Supervisors:

Prof. Sébastien Lefèvre

Université Bretagne Sud, France

Dr. Zahra Dabiri

Paris Lodron Universität Salzburg, Austria

Lorenzo Papa

European Space Agency, Italy

Submitted by:

Omowonuola Mofoyeni Akintola

Salzburg & Vannes, August 2026

Various MSc theses from the Copernicus Master in Digital Earth, including those I (co-)supervised

- GFMs and street views for urban analytics (Candela Sol Pelliza, 2024)
- Pretraining GFMs with semantic mapping (Rajeswari Parasa, 2024)
- Exploring SSL for individual-tree PCs (David Hansen, 2025)
- GFMs and VHR data (Kshitij Raj Sharma, 2026)
- On the SARness of GFMs (Omowonuola Akintola, 2026)
- Robustness of GFMs to on-boarding (Elouann Lucas, 2026)













General objective: how can we build impactful geospatial foundation models?

When GeoFMs Meet VHR: Benchmarking Foundation Models on Very High Resolution Earth Observation Imagery

Kshitij Raj Sharma*[1], Pierre Adorni*[1], Minh-Tan Pham[1], Dirk Tiede[2], Sébastien Lefèvre[1,3]

arXiv:2611.08003v1 [cs.CV] 5 Nov 2026

***Abstract*—Geospatial foundation models are a widely adopted starting point for remote sensing tasks. Most are pretrained on Sentinel-1/2 and Landsat imagery at a ground sampling distance (GSD) of 10m or coarser. Many applications instead rely on very high resolution (VHR) imagery, where individual buildings, roads, and amenities are resolved. We study which pretraining regimes transfer well to this scale. We introduce VHR-Bench, a benchmark suite of ten datasets at 1.5m GSD or finer spanning five task families, eight application domains, and both satellite and aerial platforms. We evaluate twelve encoders in three pretraining tiers with frozen backbones at two training-data fractions. We find that both pretraining on large natural-image collections and VHR-specific pretraining transfer effectively to VHR tasks, with complementary strengths. SkySense leads when labels are scarce; at the full label fraction the web-pretrained DINOv3 ViT-L edges ahead, with the two close together there. Encoders pretrained only on Sentinel-scale imagery transfer least well under RGB input. Guided by these complementary results, we continue the self-supervised pretraining of DINOv3 ViT-L on 1.02 million held-out VHR tiles and evaluate the resulting encoder without changing the downstream protocol. The continued model gives a clear improvement over the released backbone when labels are scarce and a modest one at the full fraction, at no increase in inference parameters or compute. These findings support a generalize-then-specialize strategy: already strong, broad natural-image pretraining can be further improved by a short VHR continuation stage while preserving its general visual features.**

***Index Terms*—Very high resolution, geospatial foundation models, benchmarking.**

I. Introduction

Self-supervised pretraining on large satellite archives has made Geospatial Foundation Models (GeoFMs), also referred to as Earth Observation Foundation Models (EO-FMs), a common backbone for remote sensing tasks. Most rely on Sentinel-1 and Sentinel-2 imagery, which is openly available on a global scale with high temporal revisit, at a Ground Sampling Distance (GSD) of 10m or coarser. This large amount of medium to high-resolution data has driven substantial progress in broad-area scene analysis, but leaves an important resolution regime largely unaddressed: Very High Resolution (VHR).

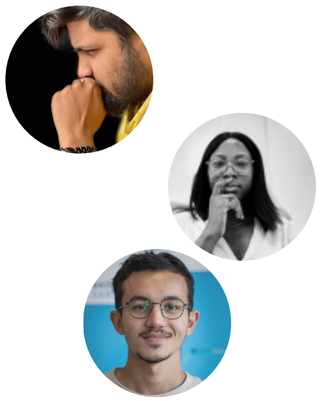
We focus on the practical demand that lies in this regime. Applications from disaster response to infrastructure mapping depend on VHR imagery, acquired from satellite platforms such as WorldView, Pléiades, and SPOT, or from aerial and drone surveys. At such resolutions, fine-grained features such as buildings, roads, amenities, and tree crowns are discriminated at the pixel level, enabling spatially precise tasks including building footprint extraction, road mapping, post-disaster damage assessment, canopy height regression, photovoltaic panel detection, tree species classification, and many more. Many disaster tasks are time-critical and demand object-level accurate spatial outputs.

Whether existing GeoFMs transfer well to this resolution regime has not been systematically evaluated. Our first hypothesis is that pretraining on medium to high-resolution Sentinel archives may be insufficient at VHR, given the substantial scale mismatch between pretraining and inference conditions. Conversely, we hypothesize that computer-vision models pretrained on large RGB natural image corpora may transfer well to optical VHR imagery: beyond sharing the same three-channel representation, VHR imagery is closer to natural images than Sentinel-2 imagery in terms of spatial scale, texture, visual diversity, and fine-grained detail. We therefore investigate whether GeoFMs without VHR pretraining underperform on VHR tasks and whether generic RGB models provide a strong alternative to VHR-specific pretraining.

We make three contributions. First, we introduce VHR-Bench, a benchmark of ten VHR datasets spanning five task families and covering eight application domains on both satellite and aerial optical sensors, under one common evaluation framework. Second, we evaluate twelve encoders organized into three pretraining tiers under a fixed protocol at two training-data fractions (10 and 100%). Third, as a design direction derived from the benchmark's results, and to test whether label-free adaptation can improve an already strong natural image encoder on the VHR domain, we assemble 1.02 million VHR tiles from seven open sources and use them to continue DINOv3

*Equal contribution. [1]Université Bretagne Sud, IRISA, France. [2]University of Salzburg, Department of Geoinformatics, Austria. [3]UiT The Arctic University of Norway, Tromsø, Norway. Corresponding author: skshitizraj@gmail.com

generic FMs outperform GFMs trained on Sentinel…
but continual fine-tuning with VHR helps



GFMs have not met SAR yet

TerraMind does not transfer to Phi-Sat-2

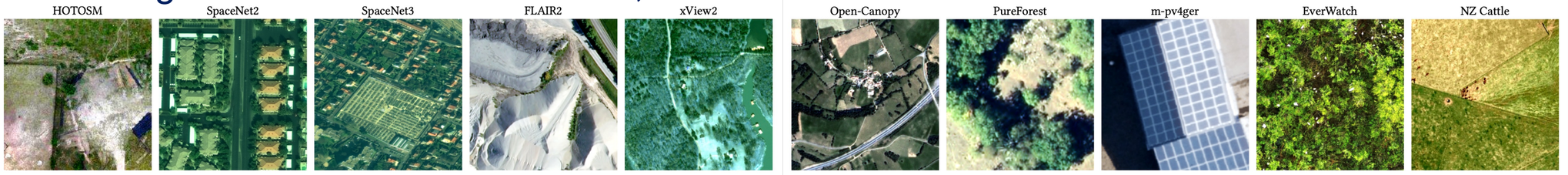
WHEN GFMS MEET VHR

VHR = very high resolution, 1.5m/pixel (GSD)

Comparison between generic foundation models and GFMs (either VHR-aware or not)

Design of a dedicated benchmark, VHR-Bench

#	Tier	Encoder	Params (M)	GFLOPs	Input (px)	Patch	Tokens	Pretraining data
1	HR-FM	Clay v1.5 ViT-L	311	723.0	256	8	1024	Sentinel-1/2, NAIP, LINZ
2		DINOv3 ViT-L (SAT)	303	164.7	256	16	256	SAT-493M (Maxar, 0.6m)
3		DOFA ViT-B	111	35.4	224	16	196	Multi-sensor (sat + aerial)
4		EoS-FM	78	206.9	256	–	–	Sentinel-1/2 and HR optical
5		SkySense SwinH	655	301.6	256	4	–	WorldView 3/4, Sentinel-1/2
6	CV-FM	ConvNeXt-XL	348	159.1	256	–	–	ImageNet-22k to 1k
7		DINOv3 ConvNeXt-L	196	89.8	256	–	–	LVD-1689M (web)
8		DINOv3 ViT-L (LVD)	303	164.7	256	16	256	LVD-1689M (web)
9		ResNet-50	24	8.2	224	–	–	ImageNet-1k
10	MidRes-FM	Prithvi-EO 2.0 600M	631	335.0	224	14	256	HLS (Landsat + S2), 30m
11		THOR v1 Base	94	44.2	224	16	196	Sentinel-1/2/3
12		TerraMind v1 L	306	253.3	224	16	196	Sentinel-1/2, DEM, LULC



#	Dataset	Platform	Sensor	GSD	Region	Task	Train / Val / Test	10% train
1	SpaceNet 2 [32]	Satellite	WorldView-3	0.3m	4 cities	Segmentation (buildings)	5186 / 1461 / 2961	517
2	SpaceNet 3 [32]	Satellite	WorldView-3	0.3m	4 cities	Segmentation (roads)	1782 / 510 / 257	178
3	xView2 (xBD) [17]	Satellite	Mixed (Maxar)	<0.8m	Global	Change detection (damage)	13384 / 917 / 933	1336
4	FLAIR 2 [12]	Aerial	Aerial RGBN	0.2m	France	Segmentation (land cover)	4049 / 1022 / 3022	1792
5	Open-Canopy [9]	Satellite	SPOT 6/7	1.5m	France	Regression (canopy height)	6634 / 737 / 13675	661
6	EverWatch [13]	Aerial	Aerial RGB	0.1m	United States	Object detection (birds)	4429 / 500 / 196	440
7	NZ Cattle [1]	Aerial	Aerial RGB (LINZ)	0.1m	New Zealand	Object detection (cattle)	524 / 66 / 65	52
8	m-pv4ger [26]	Aerial	Aerial RGB (Germany)	0.1m	Germany	Classification (PV)	11814 / 999 / 999	1181
9	PureForest [15]	Aerial	Aerial RGBN (IGN)	0.2m	France	Classification (tree species)	69111 / 13523 / 52935	6904
10	HOTOSM Buildings	Mixed	Mixed (OpenAerialMap)	0.3 to 0.5m	15 countries	Segmentation (buildings)	57890 / 7237 / 7236	5788

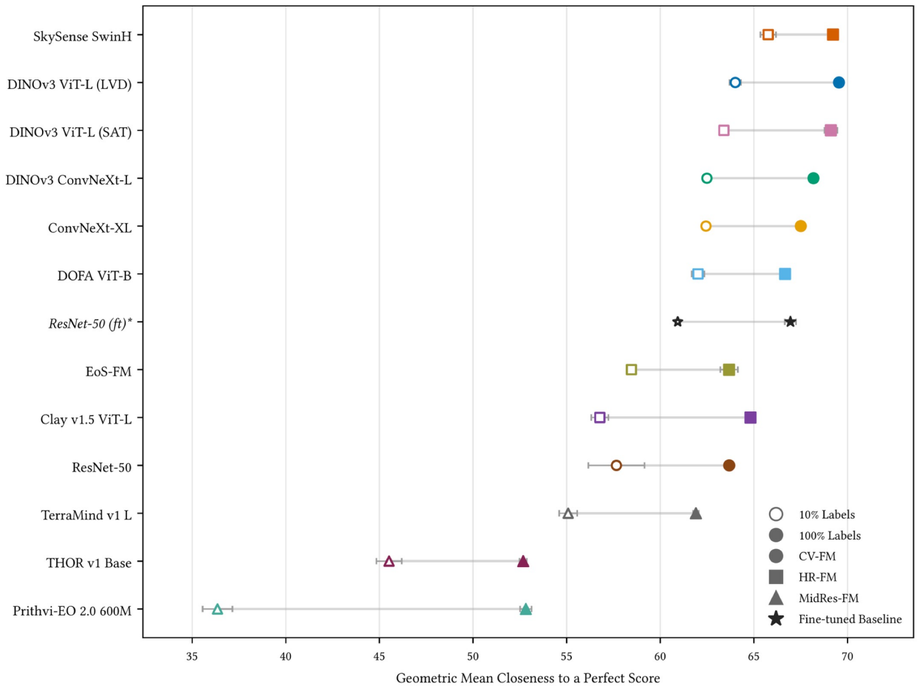
Task family	Count	Domain breakdown
Semantic segmentation	4	2 buildings (SpaceNet 2, HOTOSM Buildings), 1 roads (SpaceNet 3), 1 land cover (FLAIR 2, 13 classes)
Change detection	1	1 building damage (xView2)
Regression	1	1 forestry / canopy height (Open-Canopy)
Object detection	2	1 ecology / wading birds (EverWatch), 1 agriculture / cattle (NZ Cattle)
Classification	2	1 energy / PV panels (m-pv4ger, binary), 1 forestry / tree species (PureForest, 13 classes)

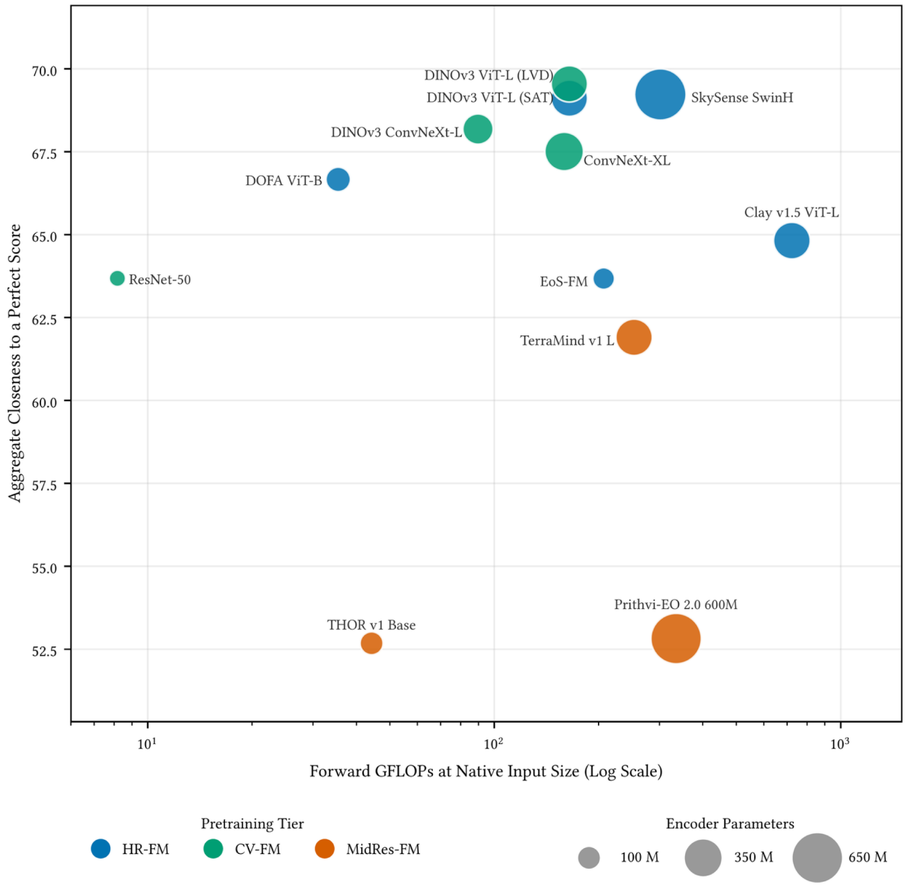
FIRST RESULTS

Closeness-to-perfect score $c_d = \begin{cases} m_d/100, & \text{bounded metrics,} \\ \max\left(1 - (\mathrm{RMSE}_d/r)^2,\ \epsilon\right), & \text{Open-Canopy,} \end{cases}$ and mean $g = \left(\prod_{d=1}^{10} c_d\right)^{1/10}$:

HR-GFMs perform weak

DINOv3 VIT-L LVD is competitive against SOTA GFM SkySense but 2x cheaper!





CONTINUAL LEARNING

Continued SSL + VHR dataset = SOTA results

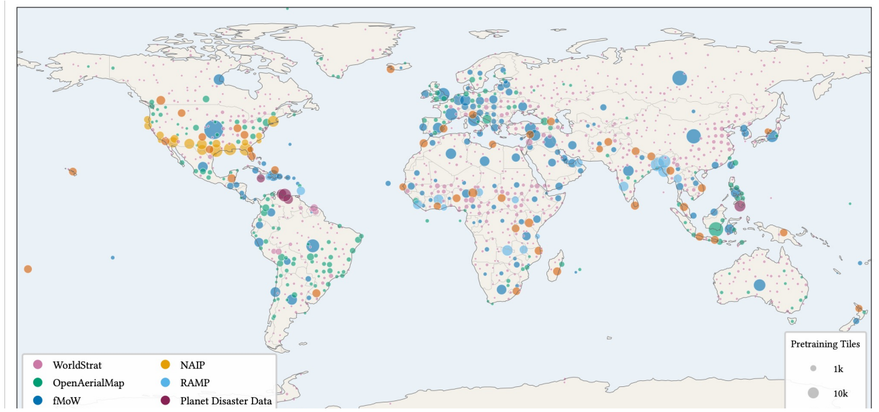


TABLE IX

PRIMARY METRIC PER DATASET AT 10 PERCENT TRAINING DATA, FROZEN BACKBONE EXCEPT INDICATED. SEGMENTATION AND CHANGE DETECTION REPORT MIOU, CLASSIFICATION MACRO F1, DETECTION MAP@50, OPEN-CANOPY-SMALL RMSE IN DECIMETRES (LOWER IS BETTER). MEAN OVER THREE DECODER SEEDS, SUBSCRIPT IS THE STANDARD DEVIATION; BEST FROZEN-ENCODER MEAN IN BOLD. THE FULLY FINE-TUNED BASELINE IS NOT COUNTED WHEN BOLDING. DATASETS: HOT: HOTOSM BUILDINGS; SN2, SN3: SPACENET 2, 3; FLR: FLAIR2; XV2: XVIEW2; OPC: OPEN-CANOPY-SMALL; PF: PUREFOREST; PV: M-PV4GER; EVW: EVERWATCH; NZC: NZ CATTLE.

Tier	Encoder	HOT	SN2	SN3	FLR	xV2	OpC	PF	PV	EvW	NZC
CV→VHR	DINOv3 ViT-L (LVD→HR)	**66.5**$_{\pm 0.4}$	**87.6**$_{\pm 0.0}$	69.7$_{\pm 0.4}$	**53.9**$_{\pm 0.6}$	**57.4**$_{\pm 0.6}$	39.8$_{\pm 0.1}$	66.2$_{\pm 0.4}$	93.4$_{\pm 0.2}$	33.9$_{\pm 0.8}$	67.0$_{\pm 2.8}$
HR-FM	Clay v1.5 ViT-L	61.2$_{\pm 0.8}$	79.9$_{\pm 0.3}$	58.8$_{\pm 0.5}$	43.9$_{\pm 0.1}$	51.8$_{\pm 0.6}$	42.8$_{\pm 0.1}$	45.7$_{\pm 0.3}$	82.2$_{\pm 0.8}$	28.3$_{\pm 1.5}$	66.8$_{\pm 1.1}$
	DINOv3 ViT-L (SAT)	65.4$_{\pm 0.7}$	85.3$_{\pm 0.3}$	68.6$_{\pm 0.4}$	51.2$_{\pm 0.3}$	55.9$_{\pm 0.3}$	**38.2**$_{\pm 0.1}$	63.0$_{\pm 1.5}$	90.2$_{\pm 0.2}$	31.9$_{\pm 0.7}$	66.0$_{\pm 1.5}$
	DOFA ViT-B	63.1$_{\pm 1.0}$	83.8$_{\pm 0.2}$	65.5$_{\pm 0.5}$	48.9$_{\pm 0.3}$	53.7$_{\pm 0.3}$	42.7$_{\pm 0.1}$	61.4$_{\pm 1.6}$	92.3$_{\pm 0.9}$	32.2$_{\pm 0.5}$	67.6$_{\pm 0.8}$
	EoS-FM	61.9$_{\pm 0.5}$	82.9$_{\pm 0.1}$	66.8$_{\pm 0.7}$	41.6$_{\pm 0.8}$	53.2$_{\pm 0.3}$	43.3$_{\pm 0.1}$	49.1$_{\pm 1.2}$	89.4$_{\pm 0.7}$	27.3$_{\pm 0.6}$	68.8$_{\pm 1.2}$
	SkySense SwinH	65.9$_{\pm 0.3}$	86.5$_{\pm 0.1}$	67.5$_{\pm 0.5}$	52.0$_{\pm 0.2}$	57.1$_{\pm 0.6}$	40.5$_{\pm 0.1}$	**67.1**$_{\pm 2.2}$	93.6$_{\pm 0.7}$	**35.9**$_{\pm 0.9}$	**75.6**$_{\pm 0.7}$
CV-FM	ConvNeXt-XL	63.8$_{\pm 0.3}$	84.3$_{\pm 0.3}$	65.8$_{\pm 0.4}$	48.3$_{\pm 0.2}$	53.7$_{\pm 0.5}$	42.7$_{\pm 0.1}$	62.1$_{\pm 2.6}$	**93.8**$_{\pm 1.0}$	34.6$_{\pm 0.4}$	64.7$_{\pm 2.6}$
	DINOv3 ConvNeXt-L	65.8$_{\pm 0.3}$	86.3$_{\pm 0.1}$	66.7$_{\pm 0.1}$	51.9$_{\pm 0.1}$	54.6$_{\pm 0.5}$	42.9$_{\pm 0.1}$	65.4$_{\pm 0.5}$	93.7$_{\pm 0.3}$	30.8$_{\pm 0.6}$	59.8$_{\pm 0.3}$
	DINOv3 ViT-L (LVD)	**66.5**$_{\pm 0.6}$	**87.6**$_{\pm 0.1}$	**69.8**$_{\pm 0.3}$	53.7$_{\pm 0.1}$	56.9$_{\pm 0.3}$	40.7$_{\pm 0.1}$	65.9$_{\pm 0.9}$	92.6$_{\pm 0.6}$	33.3$_{\pm 1.1}$	59.1$_{\pm 2.0}$
	ResNet-50	58.8$_{\pm 0.4}$	80.7$_{\pm 0.5}$	64.5$_{\pm 0.1}$	44.6$_{\pm 0.4}$	49.2$_{\pm 0.5}$	53.4$_{\pm 6.2}$	54.7$_{\pm 0.7}$	91.0$_{\pm 0.6}$	28.4$_{\pm 1.1}$	71.2$_{\pm 1.0}$
MidRes-FM	Prithvi-EO 2.0 600M	57.6$_{\pm 1.1}$	74.9$_{\pm 0.5}$	57.4$_{\pm 1.3}$	36.1$_{\pm 0.3}$	36.6$_{\pm 0.2}$	51.9$_{\pm 0.3}$	40.9$_{\pm 1.4}$	88.6$_{\pm 0.5}$	8.8$_{\pm 0.3}$	6.2$_{\pm 1.3}$
	THOR v1 Base	54.0$_{\pm 1.3}$	72.9$_{\pm 0.3}$	57.3$_{\pm 1.1}$	29.9$_{\pm 1.2}$	40.6$_{\pm 0.3}$	52.4$_{\pm 0.1}$	37.5$_{\pm 1.1}$	86.8$_{\pm 0.6}$	19.6$_{\pm 1.6}$	34.9$_{\pm 0.7}$
	TerraMind v1 L	56.8$_{\pm 2.0}$	79.9$_{\pm 0.1}$	61.6$_{\pm 0.1}$	40.6$_{\pm 0.1}$	48.4$_{\pm 0.2}$	45.6$_{\pm 0.0}$	48.4$_{\pm 2.6}$	87.9$_{\pm 1.9}$	26.3$_{\pm 0.9}$	58.3$_{\pm 1.0}$
Baseline	ResNet-50 (ft)*	61.9$_{\pm 0.6}$	81.7$_{\pm 0.8}$	65.4$_{\pm 0.1}$	44.9$_{\pm 0.5}$	52.1$_{\pm 0.6}$	47.2$_{\pm 2.9}$	56.4$_{\pm 1.4}$	94.9$_{\pm 0.9}$	33.8$_{\pm 1.2}$	72.4$_{\pm 0.7}$

BACK TO JRC… BUT STILL WITH GFMS

Global multi-task learning for mapping and characterizing human settlements from EO data (Kömürcü Kürşat)

- Perform multi-task learning directly on multiple embeddings generated by GFMs



Backcasting anthropogenic infrastructures over a century of historical EO data and maps (Clara Vydra)

- Model the data as graph and use graph neural networks (possibly Graph Foundation Models)



Explainable multimodal AI using geospatial data for rapid estimation of displacement and people in need in crises (Ivan Decostanzi)

- Use of Multimodal LLM (MLLM) and VLM to extract knowledge from multiple sources of information



CONCLUSION

Foundation models represent a new AI breakthrough, not only in NLP, CV, but also EO, pinning high hopes

However, some issues remain:

- From a user perspective, how to choose the right model?
  - No standard benchmark
  - Datasets far from real needs
  - Measuring accuracy is not enough
  - Models are not always available
- From a provider perspective, how to bring a significant contribution?
  - Scaling laws but moderate gain ignoring the negative impacts
  - Providers are limited to a very few
  - Other contributors can focus on how to use these models or develop alternatives at a lower scale
  - Validation is still very costly





Open questions

- What is the role of the scientific community beyond the very few able to enter the race of very large-scale experiments?
- How to limit the carbon footprint related to the design, evaluation, understanding, and use of GFMs?