Politecnico di Torino

Politecnico di Torino

Computer Engineering - Artificial Intelligence and Data Analytics

A.a. 2025/2026

Graduation Session March 2026

Anomaly Detection in Hyperspectral Remote Sensing Imagery

Supervisors:

Prof. Tatiana Tommasi

Prof. Lia Morra

Candidate:

Matteo Bonsignore

Abstract

Anomaly detection in Remote Sensing imagery is a challenging problem due to the scarcity of anomalous samples, strong class imbalance, and the heterogeneity of sensors and acquisition conditions. Although recent geospatial foundation models provide powerful feature representations, their effective use for anomaly detection remains underexplored.

This thesis investigates an embedding-based anomaly detection pipeline for multispectral and hyperspectral satellite imagery, combining foundation model backbones with PatchCore, a patch-level anomaly detection method originally developed for industrial inspection. Multiple foundation models and datasets are evaluated, highlighting the limitations of cross-sensor transfer and motivating a focused study on the FLOGA wildfire dataset, which mitigates sensor mismatch.

To improve anomaly separability in the embedding space, a semi-supervised fine-tuning strategy for foundation model backbones is introduced, based on center-based regularization, teacher–student distillation, and a margin-based objective that explicitly pushes anomalous samples away from the normal embedding distribution.

Extensive experiments demonstrate that the proposed fine-tuning substantially improves anomaly detection performance compared to frozen backbones, achieving substantial gains in AUROC.

The results highlight the importance of representation adaptation for satellite anomaly detection and show that combining foundation models with lightweight, semi-supervised fine-tuning and patch-based scoring is a viable and effective approach in data-scarce scenarios.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Wildfires are rare but high-impact natural hazards that affect ecosystems, economic activity, and human safety every year. Rapid detection and localization, including in remote areas, can enable timely intervention and substantially reduce damage and casualties.

Remote sensing provides a scalable tool for wildfire monitoring thanks to its wide-area coverage, frequent revisit time, and the availability of rich spectral information beyond standard RGB imagery. In particular, multispectral and hyperspectral sensors capture reflectance across multiple wavelength bands, often extending into the near- and shortwave-infrared ranges, which can provide additional information about vegetation condition, smoke, active fires, and post-event burn scars.

Unlike RGB images, which store three values per pixel, multispectral images typically provide a limited set of discrete bands (from a few to tens), while hyperspectral images provide hundreds of contiguous bands. This increased spectral resolution may improve the separability of rare events from background land cover, but it also introduces additional challenges when models must handle heterogeneous spectral configurations.

Despite the growing interest in satellite-based wildfire monitoring, operational anomaly detection still faces key limitations. First, wildfire events are rare, and publicly available datasets capturing active fires or burned areas are limited and often weakly labeled or highly imbalanced. Second, sensor heterogeneity across datasets introduces systematic distribution shifts: images acquired by different satellites present different spectral responses and spatial resolutions, making cross-dataset comparison and joint modeling non-trivial.

Recent geospatial foundation models offer a potential way to learn more transferable representations, but it is unclear whether these representations are directly suitable for distance-based anomaly detection.

1.1 Problem Statement

This thesis studies the adaptation of embedding-based anomaly detection techniques, specifically PatchCore, to wildfire-related events in multispectral and hyperspectral satellite imagery. In this work, anomalies are defined at the image level as (i) active fires and (ii) burned areas, under highly imbalanced conditions.

The key challenge is that embedding-based anomaly detection pipelines, originally developed for industrial vision, may fail under sensor shift, modality shift, and broader domain shift between datasets. Such methods rely on learning (or assuming) a standardized representation of normal samples in the embedding space, which is easier to achieve in controlled industrial environments but becomes challenging in Earth Observation, where changes unrelated to the target anomaly (e.g., atmospheric conditions or seasonal differences) can substantially alter the embedding distribution.

Therefore, the central question is whether geospatial foundation model representations can provide sufficiently structured and transferable embeddings for distance-based anomaly scoring, and whether limited supervised signal can adapt these embeddings to improve anomaly ranking in the target domain.

1.2 Research Gap

Most remote sensing anomaly detection methods treat anomalies as spectral outliers with respect to background statistics, often without leveraging semantic representations.

Meanwhile, geospatial foundation models promise transferable embeddings across Earth Observation data, but there is limited evidence that these embeddings are directly usable for PatchCore-style anomaly detection under sensor and modality shifts, and whether they remain comparable across heterogeneous spectral configurations.

Moreover, the impact of lightweight semi-supervised fine-tuning for improving anomaly ranking in this setting is still underexplored.

1.3 Research Questions

The objective of the thesis is to answer a few precise research questions:

- **RQ1: Can pre-trained foundation models detect fire-related events as anomalies using a standard unsupervised anomaly detection pipeline?**

This question evaluates whether frozen embeddings from geospatial foundation models are sufficiently structured for PatchCore-style distance-based scoring to separate normal scenes from fire-related anomalies (active fires and burned areas), without task-specific training.

- **RQ2: Are foundation model embeddings consistent and comparable across different satellite sensors?**

  This question examines whether embeddings extracted from images acquired by different sensors and spectral modalities (e.g., Sentinel-2 vs Landsat-8; hyperspectral vs multispectral) remain aligned enough to support unified memory banks and cross-sensor evaluation, or whether sensor- and modality-shift dominates distance-based anomaly scores.

- **RQ3: Can semi-supervised fine-tuning with limited labeled fire examples substantially improve embeddings for fire anomaly detection?**

  This question assesses whether lightweight backbone adaptation with limited labeled anomalies can reshape the embedding space (e.g., compacting normal clusters and increasing separation margins) and translate into improved PatchCore ranking performance, compared to frozen backbones.

1.4 Proposed Approach

The proposed pipeline applies PatchCore, an embedding-based anomaly detection method, to remote sensing imagery by using geospatial foundation model backbones (HyperFree and Copernicus-FM) as feature extractors. Given an input image, the backbone produces patch-level feature maps, which PatchCore uses to build a memory bank of normal patches and to compute anomaly scores. Image-level scores are obtained by averaging patch scores.

The pipeline is evaluated under multiple dataset configurations designed to isolate domain effects, such as sensor/modality shift. In addition, spectral invariance tests and a computational and structural comparison between the two backbones are performed. Performance is primarily compared using the Area Under the ROC Curve (AUROC).

Finally, a semi-supervised fine-tuning strategy is introduced to adapt the backbones to the target domain. The objective reshapes the embedding space by (i) compacting normal samples around a learned centroid, (ii) enforcing a margin that pushes anomalous samples away from the normal centroid, and (iii) applying a teacher–student distillation term to preserve semantic structure during adaptation.

1.5 Main Contributions

The main contributions of this thesis are:

- **PatchCore adaptation to spectral remote sensing.** PatchCore is adapted to multispectral and hyperspectral satellite imagery by using geospatial foundation model backbones (HyperFree and Copernicus-FM) as patch-level feature extractors.
- **Systematic evaluation under dataset and sensor shifts.** A set of experimental configurations is defined to isolate the effect of domain mismatch (sensor/modality changes) on memory-bank construction, anomaly scoring, and image-level ranking performance.
- **Spectral invariance testing with synthetic and native pairs.** Representation drift across spectral configurations is quantified through a large-scale SRF-based hyperspectral-to-multispectral conversion study, complemented by an analysis on co-registered PRISMA–Sentinel-2 acquisitions.
- **Semi-supervised backbone adaptation for anomaly ranking.** A fine-tuning strategy is introduced to reshape embedding geometry via centroid compactness for normal samples, margin-based separation for anomalies, and teacher–student distillation to preserve semantic structure.

1.6 Thesis Organization

The remainder of this thesis is organized as follows.

Chapter 2 reviews background and related work on anomaly detection in remote sensing, PatchCore-style methods, and geospatial foundation models. Chapter 3 describes the proposed anomaly detection pipeline, including feature extraction with foundation model backbones, PatchCore scoring, and the semi-supervised fine-tuning strategy. Chapter 4 presents the experimental protocol, datasets, and results, including baseline frozen-backbone evaluations, spectral invariance tests, fine-tuning experiments, and computational/structural comparisons between HyperFree and Copernicus-FM. Finally, Chapter 5 summarizes the main findings, discusses limitations, and outlines directions for future work.

Chapter 2

Related Works

The study can be divided into the definition of the task, the definition of the models, and the definition of the datasets.

2.1 Task

The anomaly detection task is defined in literature with three different meanings, depending on the context and the relevant scientific community. Three main tasks can be distinguished:

- Hyperspectral Anomaly Detection (HAD).
- Change Detection (CD).
- Anomaly Detection (AD).

Each one is characterized by different definitions, formalizations, approaches, and benchmarks.

2.1.1 Hyperspectral Anomaly Detection

Definition and Goals Considering hyperspectral images, the goal of the HAD task is to find pixels or areas whose spectra deviate significantly from the background, without a priori knowledge of the target. The detection itself is based on the richness of spectral information, but suffers from computational complexity and sensitivity to noise and scene changes [32] [26].

Formalization Given a hyperspectral cube and given the spectral vector associated to the pixel x, the background is modeled as a distribution, assuming that

the signal is the combination of two signals, background and anomaly, not known a priori.

A detector assigns a score $s(x)$ and declares a pixel anomalous if $s(x) > \tau$. The anomaly score is typically obtained by comparing the local spectrum with estimated background properties, for example, where d can be the Mahalanobis distance or the residual with respect to a low-dimensional subspace and are estimates of the background, obtained globally or from local windows.

Families of Methods Different families of methods exist, each one based on different structures:

- **Statistical Models-Based**, which leverages background estimates, spectral distances, RX/GLRT-type detectors, local/global approaches [26].
- **Structured Representation-Based**, Low-rank and sparse representations, adaptive graphs, manifold structures. Recent work on graphs highlights sensitivity to relational structure and the quality of graph construction [29].
- **Deep Learning-Based**, using autoencoders/GAN, masked autoencoding and transformers. These are useful methods but exposed to small benchmarks biases and overfitting. Recent works highlights the possibility to use foundational models like HyperFree in this task [17].

Benchmark Most historical comparative evaluations use small datasets, with limited number, resulting in metrics saturation on a few scenarios and poor cross-sensor/territorial generalizability. Existing studies highlight significant limitations of existing benchmarks in terms of small size, partial annotations, and lack of spectral variability and operating conditions [26] [32] [17].

2.1.2 Change Detection

Definition and Goals Change detection (CD) in remote sensing identifies variations between co-registered images acquired at different times. It covers applications such as urbanization, disasters, deforestation, and infrastructure, with both supervised and unsupervised pipelines [6] [12] [8] [18].

Formalization Change detection is typically modeled as a classification or segmentation task that takes as input a sequence of two or more hyperspectral or multispectral cubes [20]. Almost all methods formalize the problem as a sequence of two operations: a preprocessing function, which includes, for example, co-registration, and a comparison function, which depends on the type of input and the type of task.

Families of Methods Four different classes of algorithms can be identified:

- **Binary Change Detection (BCD)**: Focuses on detecting whether a pixel has changed between two time points. The output is a binary map. It includes methods such as Change Vector Analysis (CVA), image differencing, and supervised binary classifiers.

- **Multiclass Change Detection (MCD)**: Goes beyond BCD by also identifying the type of change. Methods often rely on post-classification comparison or multi-output deep networks. Difficulties can often arise due to label noise and the complexity of modeling class transitions.

- **Anomalous Change Detection (ACD)**: Aims to identify rare, unexpected, or statistically anomalous changes without prior class labels. Techniques include unsupervised approaches such as robust PCA, clustering, and autoencoders.

- **Time Series Change Detection (TSCD)**: Extends CD to more than two time points, capturing temporal dynamics. Typically uses recurrent models (LSTMs), temporal CNNs, or statistical trend analysis (e.g., BFAST).

Applications and Critical Issues The main applications for CD are building change, land cover dynamics and disaster mapping, with recent works which integrates also foundational models [19].

Talking about critical issues, the quality of the result depends on the quality of the co-registration, the availability of archives consistent over time and the management of seasonal/atmospheric variations. It's important to mention that maintaining a clean and rich history is expensive and not always feasible [12].

2.1.3 Anomaly Detection

Definition and Goals In the general literature, AD aims to identify data points that deviate from normal patterns, typically without anomalous labels or with a small number of defined anomalies. The typical setting, widely studied in industry and computer vision, involves the availability of a set of normal data during the training phase. Based on the available literature, this type of task has not been studied in earth observation, where segmentation tasks and/or applications focused on a specific type of anomaly are more common. This is because sensor heterogeneity, atmospheric variability and geographic domain shift violate the stationary assumptions of classical AD pipelines.

Families of Methods Unsupervised methods are essentially divided into two families:

- **Embedding-Based**, these methods transform complex data (text, images, logs, transactions) into vector representations (embeddings) in a latent space. Anomalies are identified as observations whose embeddings significantly deviate from the distribution of normal data, typically measured via distance-based or density-based criteria.

  In the context of image data, embedding-based methods can operate either at the global level, by associating a single embedding to each image, or at the local (patch) level, by extracting embeddings from spatial regions of the image. Patch-based embedding methods, such as PatchCore, mode the distribution of normal patch embeddings using a memory bank and compute anomaly scores by aggregating distances between test patches and normal reference patches. This local formulation has been shown to improve robustness to spatially localized anomalies and background variability.

- **Reconstruction-Based**, models such as autoencoders or GANs, that learn the "normal" and use the reconstruction error as a score, are used. Known limitations of this approach include semantic leakage, dependence on the training distribution, reconstruction artifacts that inflate false positives, and a decrease in cross-sensor performance. Finally, high-capacity models may be able to reconstruct the anomaly even if they never observed it during training.

A detailed overview of the state-of-the-art and the available models is available at the following link: Awesome Industrial Anomaly Detection Github.

2.2 Foundational Models for Remote Sensing

Geospatial foundational models (FMs) are models trained in a self-supervised manner on large EO datasets. These models are designed to be fine-tuned to multiple downstream tasks, potentially reducing labeling costs and increasing cross-domain/sensor transferability. In recent years, several models have been proposed, often with very strong performance claims. However, it is important to note that, compared to vision-language models (VLMs), geospatial foundational models are often trained on smaller datasets and tend to be based on a single sensor/modality.

Moreover, the majority of datasets and models do not consider the temporal and non-stationary nature of satellite data [20].

The models proposed in the literature differ based on their training dataset, architecture, and training method.

The use of foundational models in the satellite field potentially has numerous advantages [14]:

- **Self-supervised pre-training**: produces robust embeddings, useful for developing embedding-based anomaly detection and change detection tools under variable conditions.
- **Lightweight fine-tuning**: allows specialization on specific scenes/spectral regimes while maintaining generality, useful for AD in open-set and data-scarce scenarios.
- **Federated/edge learning**: architectures designed for distributed fine-tuning across constellations or operations on edge arrays open up continuous updates and on-orbit detection, subject to bandwidth/computation constraints [35].

Works such as HyperFree [17] indicate that in the hyperspectral field some FMs are potentially able to achieve comparable performance with specialized models in various tasks, including hyperspectral anomaly detection and hyperspectral change detection. Other benchmarks such as PANGAEA [20] reach more cautious conclusions: on multispectral data, FMs do not necessarily offer superior performance to models trained directly on specific annotated datasets. In that study, a simple segmentation neural network (U-Net) often proved to be more effective than a pre-trained FM without fine-tuning, even in conditions of relative data scarcity (Table 2.1).

Model	HLS Burns	MADOS	PASTIS	Sen1Floods11	xView2	FBP	DynEarthNet	CropMap	SN7	AI4Farms	BioMassters	#Top2
CROMA	82.42	**67.55**	32.32	90.89	53.27	51.83	38.29	49.38	59.28	25.65	36.81	2
DOFA	80.63	59.58	30.02	89.37	59.64	43.18	39.29	51.33	61.84	27.07	42.81	2
GFM-Swin	76.90	64.71	21.24	72.60	59.15	67.18	34.09	46.98	60.89	27.19	46.83	1
Prithvi	83.62	49.98	33.93	90.37	49.35	46.81	27.86	43.07	56.54	26.86	39.99	1
RemoteCLIP	76.59	60.00	18.23	74.26	57.41	**69.19**	31.78	52.05	57.76	25.12	49.79	2
SatlasNet	79.96	55.86	17.51	90.30	52.23	50.97	36.31	46.97	61.88	25.13	41.67	0
Scale-MAE	76.68	57.32	24.55	74.13	**60.72**	67.19	35.11	25.42	**62.96**	21.47	47.15	3
SpectralGPT	80.47	57.99	35.44	89.07	48.40	33.42	37.85	46.95	58.86	26.75	36.11	1
S12-MoCo	81.58	51.76	34.49	89.26	51.59	53.02	35.44	48.58	57.64	25.38	40.21	0
S12-DINO	81.72	49.37	36.18	88.61	50.56	51.15	34.81	48.66	56.47	25.62	41.23	1
S12-MAE	81.91	49.90	32.03	87.79	50.44	51.92	34.08	45.8	57.13	24.69	41.07	0
S12-Data2Vec	81.91	44.36	34.32	88.15	51.36	48.82	35.90	**54.03**	58.23	24.23	41.91	1
UNet Baseline	**84.51**	54.79	31.60	**91.42**	58.68	60.47	**39.46**	47.57	62.09	**46.34**	**35.67**	6
ViT Baseline	81.58	48.19	**38.53**	87.66	57.43	59.32	36.83	44.08	52.57	38.37	38.55	2

Table 2.1: Performance obtained by different FMs, without fine-tuning, on a set of benchmarks compared with standard supervised architectures (U-Net and ViT). Of particular interest are the HLS Burns dataset [22], related to the identification of burned areas, and Sen1Floods11 [2], related to flood recognition. Table reproduced from [20].

The results are dependent on the downstream task: the benefit of training on specific datasets is stronger for binary classification/segmentation tasks, such as

recognizing burned areas and floods, while more complex multi-class segmentation and change detection tasks tend to benefit more from pre-training. The same study shows strong variation between different FMs, with some models like CROMA performing very well, and a substantial performance degradation when there are variations in the geographic domain between training and testing data, regardless of the model and the setting chosen (U-Net or FM). In other words, models trained on a single dataset generalize poorly to geographic areas different from those on which they were trained, but using a FM as a backbone does not always lead to more robust results in the settings investigated in this study.

2.2.1 Models Overview

The following provides an overview of all the FM analyzed for the purpose, with a brief description, starting with the models that take as input just MSI:

CROMA [10] Dual ViTs architecture, taking as input both radar (SAR) and optical data. A multi-modal encoder fuses the two inputs and a lightweight decoder reconstructs the masked image patches during training.

Efficient Onboard Multitasking AI Architecture [13] Based on GhostNet V2, it is composed of two symmetric convolutional branches, inspired by BYOL, which are trained through contrastive learning.

msGFM [11] Can handle RGB, SAR, Sentinel-2 and DSM inputs. Each modality has a distinct embedding layer, then a shared encoder learns joint representations. Finally a separate decoder for each modality reconstructs the images. The training is based on masked autoencoding.

SatMAE++ [21] Based on ViT, it is able to learn scale-invariant representations from multispectral satellite images using a masked autoencoding objective.

AnySat [1] Handles different sensor inputs through modality-specific projectors, which encodes the input samples, then aggregated with a shared transformer and fused with a modality combiner. It uses a JEPA (Joint Embedding Predictive Architecture) setup with teacher-student objective which encourages cross-modal and cross-resolution consistency.

RobSense [8] Versatile model that works with uni-modal or multi-modal inputs (MS + SAR), temporal and static data and even with complete or incomplete datasets. MAE based, the architecture is based on two uni-modal encoders, followed by a multi-modal encoder.

DINOv3 [27] Big FM originally trained to work with RGB inputs, can easily be adapted to MSI. ViT based, the training is based on the teacher-student paradigm.

AlphaEarth [5] Embedding field model, designed to create dense, spatio-temporal embeddings of earth's surface, leverages an encoder-decoder architecture.

TerraMind [15] It is an any-to-any generative multi-modal model for EO, which uses tokenization before an encoder-decoder architecture to generate multi-modal latent tokens.

Copernicus-FM [31] ViT-based backbone, able to process different spectral (Sentinel-1 to 3) and non-spectral (DEM) inputs. Thanks to its spectral hypernet, it is able to handle inputs with variable number of bands. Metadata can also be encoded through Fourier encoding.

During the analysis, considering the novelty of the field, few models able to handle HSI were found:

HyperFree [17] ViT-based model, uses a channel-adaptive embedding module that dynamically builds convolutional kernels from a wavelength-indexed weight dictionary. This ensures the model to handle hyperspectral inputs with variable number of bands. The model even supports promptable training using point prompts to handle zero-shot EO tasks.

FoMo-Net [3] Also ViT-based backbone, designed to process images from multiple sensors and spectral resolutions. Trained using a Spectral-MAE strategy, masking random spectral bands to let the model reconstruct them.

The final choice was to focus on HyperFree and Copernicus-FM.

2.3 Hyperspectral and Multispectral Datasets

Regarding datasets, the analysis has been more difficult, because of the reduced amount of available data, given by the novelty of the field, but more importantly by differences across sensors, which became a problem in a task where more than one dataset was needed.

The sensor problem is even more problematic considering that datasets using images coming from the same sensor usually perform different pre-processing pipelines on them, discarding different bands and modifying the spectral signature of the images.

The first group of analyzed dataset can be found at Table 2.2, presenting foundation datasets, which are big unlabeled datasets mainly used for the training of FM.

Table 2.2: Foundation Datasets

Name	# of Samples	Shape	Sensor
Hyper-Seg [17]	42k	$512 \times 512 \times 224$	AVIRIS
HyperGlobal-450k [30]	450k	$64 \times 64 \times 242 - 330$	EO-1
SpectralEarth [4]	500k	$128 \times 128 \times 202$	EnMap
HySpecNet-11k [9]	11k	$128 \times 128 \times 224$	EnMap
Copernicus Pretrain [31]	1M	$264 \times 264 \times 13$	Sentinel-2

Moving on there are the task-specific datasets or benchmarks, which are labeled datasets with much less samples, used to test the models performances on domain specific tasks. These can be divided by task in: flood related in Table 2.3, and fire related in Table 2.4.

Table 2.3: Flood Datasets and Benchmarks

Name	# of Samples	Shape	Sensor
WorldFloodsv2 [23]	509	$256 \times 256 \times 13$	Sentinel-2
Sen1Floods11 [2]	12	$512 \times 512 \times 2 - 13$	Sentinel-1 -2
UrbanSARFloods [34]	8879	$512 \times 512 \times 8$	Sentinel-1

Table 2.4: Fire Datasets and Benchmarks

Name	# of Samples	Shape	Sensor
Sen2Fire [33]	2466	$512 \times 512 \times 12$	Sentinel-2
Active Fire [7]	150k	$256 \times 256 \times 10$	Landsat-8
Land8Fire [28]	20k	$256 \times 256 \times 10$	Landsat-8
FLOGA [25]	170k	$256 \times 256 \times 9$	Sentinel-2

After an accurate analysis of all the mentioned datasets, the decision fell on SpectralEarth [4], Copernicus Pretrain [31], Active Fire [7] and FLOGA [25] as the main datasets used for the experiments. A detailed description of the selected datasets, including preprocessing steps, splits, and their role in the proposed pipeline, is provided in Section 4.4.

Chapter 3

Method

The problem of image-level anomaly detection in multispectral and hyperspectral satellite imagery is addressed. Given a collection of satellite images representing normal conditions, and little or no explicit supervision on anomalous events, the goal is to identify samples that deviate from the normal data distribution. In this work, anomalies correspond to rare events such as burned areas or active fires, which are not explicitly modeled during training but are expected to emerge as outliers in a learned feature space.

In this chapter the proposed pipeline for anomaly detection on multispectral and hyperspectral satellite imagery is described. Starting from the overall architecture of the pipeline (3.1), then moving to the foundation model backbones used for feature extraction (3.2), the PatchCore anomaly detection module (3.3), and finally the semi-supervised fine-tuning strategy adopted (3.4).

3.1 Pipeline Overview

The idea behind the pipeline design was to reuse an anomaly detection technique originally designed for industrial visual inspection, namely PatchCore [24], and adapt it to satellite imagery inputs.

To do so, a foundation model backbone is added as the first block of the pipeline, capable of handling multispectral and hyperspectral inputs. Its role is to extract and project the features from the input images to the embedding space. After feature extraction there is the head of the pipeline, which is the actual anomaly detection module, built using the PatchCore method. Starting from the extracted features in the embedding space its role is to build the memory bank and compute the anomaly scores for each sample.

The full pipeline diagram is represented in Figure 3.1.

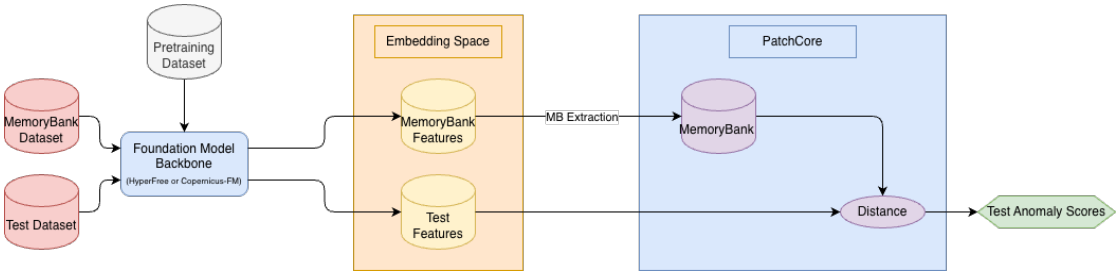


Figure 3.1: Pipeline Overview

3.2 Foundation Model Backbone

The foundation model backbone is a fundamental part of the pipeline. It takes as input the raw satellite images, analyzes them and then projects the result in the embedding space, giving the starting point for the anomaly detection itself. For these reasons, it is fundamental to have a backbone able to represent the input data correctly for this task. Foundation models are selected because these are models trained on a large amount of unlabeled data, with solid generalization ability, adaptable to any task.

After a preliminary analysis of the available FMs (Section 2.2), considering that one of the main faced constraints, due to the available data, was the need to analyze samples coming from different sensors, with different spatial and spectral signatures, with the same backbone, the final choice was to use HyperFree [17] and Copernicus-FM [31], which, among all the FMs, were the only two able to handle inputs with variable number of bands natively.

3.2.1 HyperFree

HyperFree [17] is a channel adaptive model, meaning it can process HSI with any number of channels and any wavelength using a module called channel adaptive embedding module (CAE), trained on the proprietary HyperSeg hyperspectral dataset.

It is based on the prompt engineering paradigm (P-E), which allows to design task-specific prompts to perform downstream tasks without finetuning. The model is composed of the meta-architecture of segment anything model (SAM) [16], with the addition of two specific modules, which allows it to handle different input structures and use the P-E paradigm.

The original SAM architecture is composed of three fundamental elements:

- **Image Encoder**, which is a MAE pre-trained Vision Transformer (ViT), adapted to process high resolution inputs, taking in input the patch tokens from the CAE.

- **Prompt Encoder**, represents point prompts by positional encoding summed with learned embeddings.

- **Mask Decoder**, maps the image embedding, prompt embedding and output token to a segmentation mask.

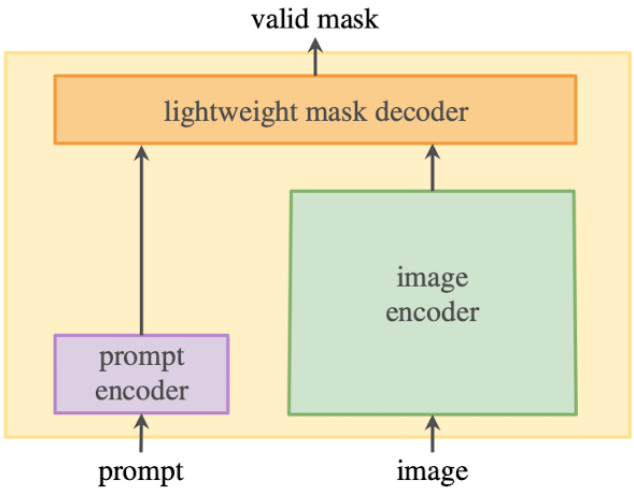


Figure 3.2: SAM Architecture

Channel-Adaptive Embedding (CAE)

This is the first innovative module presented. Its role is to map the input image with a variable number of channels into tokens with a fixed length. Placed before the SAM architecture, it works as a preprocessing step for the input images.

The main component is the weight dictionary, learned during training, that is a collection of 221 weight matrices, each one mapping a specific wavelength interval. Each interval is $10nm$, and the model covers the $400nm - 2500nm$ range of wavelengths. The weight matrices have size $221 \times p \times p \times j$, where $p \times p$ is the patch size and j is the number of output channels.

Besides the image itself, this module requires in input also the list of central wavelengths for all the input bands. This is fundamental to assembling the convolutional filter, given by picking the corresponding weight matrices. The convolutional filter is then used as the kernel for a depthwise convolution operation on the input image to obtain the output tokens.

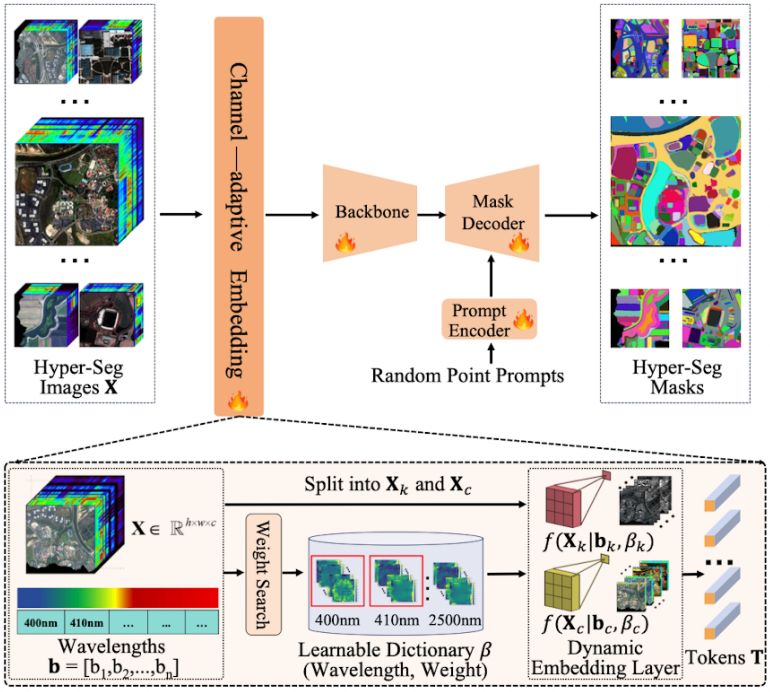


Figure 3.3: Channel-Adaptive Embedding (CAE)

Prompt-Mask-Feature (PMF) Interactive Inferring

Since many task require to segment out all the masks belonging to the same semantic category, this module connects the prompt (tells the model an object reference in task), mask (outputs accurate location information) and feature (reflects semantic information) together in feature space, using the cosine distance of features to measure the semantic similarity.

Specifically, there are two interaction modes:

- **Mode 1**, from prompt to feature. It converts a prompt (x, y) into a semantic vector $(d_{(x,y)} \in R^j)$.

  In other words, starting from the set of masks extracted by the backbone modules, it selects the one that contains the prompt. The semantic representation of the prompt is then given by the average of all the features belonging to that selected mask.

- **Mode 2**, given a reference feature $(d_{(x,y)})$, it finds all masks in M that are semantically similar. For each predicted mask it computes the average of the corresponding features, and through cosine similarity it makes a comparison with the reference feature. Only masks whose similarity is above a threshold (τ) are considered.

An usage example is the HAD task, where only mode 2 is used. Specifically, mode 2 is leveraged to perform a self-similarity filtering. It considers the assumption that anomalies are typically small and semantically different from the background, so for each predicted mask it computes the number of pixels and the features average, discarding large or common masks and highlighting small and different patches as the anomalies. The classification is done considering a threshold (τ).

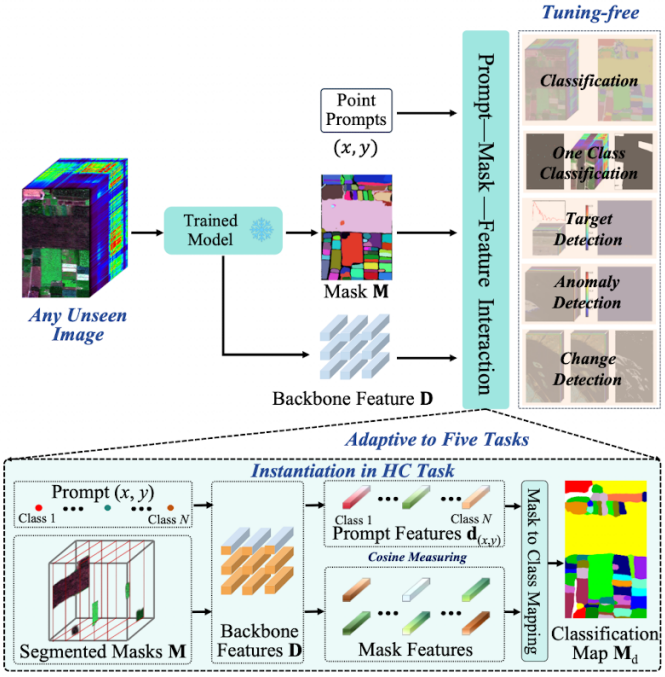


Figure 3.4: Prompt-Mask-Feature (PMF) Interactive Inferring

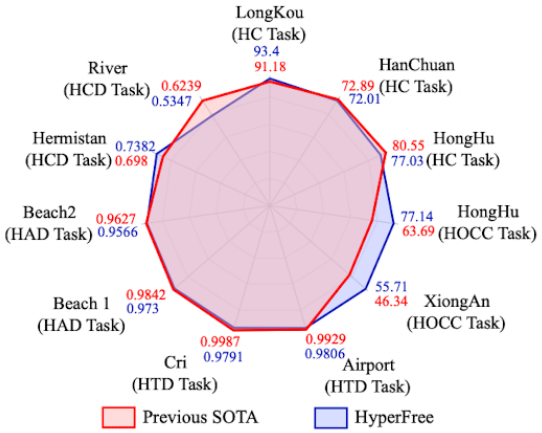


Figure 3.5: HyperFree performance on different downstream tasks, image taken from [17]

Model Extraction

Considering the differences between the working method of HyperFree and the proposed pipeline, it was not possible to reuse the whole model for the purpose. In fact, HyperFree backbone is originally thought to output segmentation masks, thanks to the mask decoder and the prompt encoder.

The objective was to have a rich and informative embedding space, in order to use it as the starting point for our AD pipeline. For these reasons, only some parts of the full model were extracted and reused:

- **Channel-Adaptive Embedding (CAE)**, which allows to analyze both MSI and HSI, without any constraint about the dimensionality of the input, in terms of number of bands. This is optimal to be able to analyze images coming from different sensors without any adaptation.

- **Image Encoder**, from the SAM architecture, that takes as input the output tokens of CAE, in order to project the images in the embedding space.

In the adopted notation (Section 3.3.1), the composition of CAE and the SAM image encoder plays the role of the backbone f_θ that maps a multispectral or hyperspectral image x to a feature map $F = f_\theta(x)$.

3.2.2 Copernicus-FM

Copernicus Foundation Model (Copernicus-FM) [31], developed by ESA, is a unified backbone based on Vision Transformer (ViT) architecture, extended to operate on spectrally heterogeneous and multi-resolution inputs. It introduces a series of spectral-aware components to handle arbitrary band combinations across sensors, and leverages also input metadata to enrich the data representation.

The model has been trained using a proprietary dataset, called Copernicus Pretrain, which is a Sentinel-2 MSI dataset. In fact, this model, even though it is able to work with variable inputs shapes, cannot natively work with HSI.

Model Architecture

The overall architecture (Figure 3.6) shows the main components used by the model.

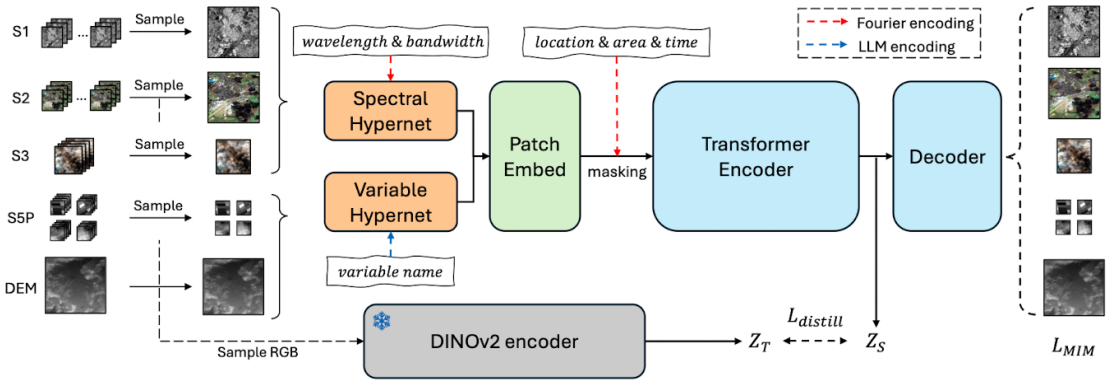


Figure 3.6: Copernicus-FM Pipeline

The key element is the spectral hypernetwork, a module that dynamically generates patch embedding weights starting from the wavelength and bandwidth of each input channel, that are embedded through Fourier encoding. The output vector then passes through an MLP, that reshapes it to obtain a band-specific convolutional kernel. Eventually, the variable hypernetwork, which works in parallel with the spectral hypernetwork, can integrate textual information about the input image, such as atmospheric information in DEM format. In this case an LLM encoder is used instead of the Fourier encoding.

Stacking up the obtained convolutional kernels, the dynamic patch embedding is able to perform the convolution operation over the input image. The resulting tokens can be enriched with other metadata, such as location, time and area, before entering the transformer encoder, which finally projects the input image in the embedding space.

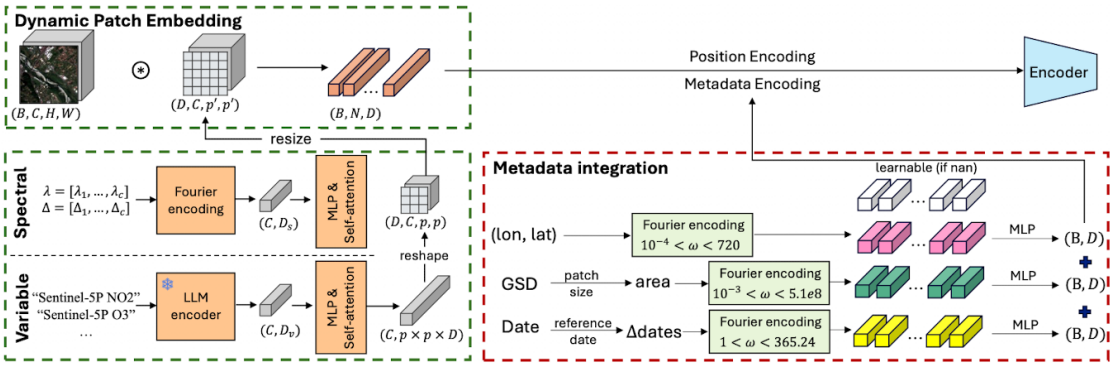


Figure 3.7: Copernicus-FM Patch Embedding

To evaluate the generalization capabilities of the model, the authors introduce Copernicus Benchmark, which is a curated suite of remote sensing tasks spanning

various sensors, spatial scales and annotation types. The tasks include scene classification, semantic segmentation, cloud detection and change detection.

Model Extraction

This pipeline allows the model to be used as a universal feature extractor for multispectral tasks, this is why, differently from HyperFree, the whole model could be reused for the pipeline. The only module that is not exploited, because of the dataset used in the experiments, was the variable hypernetwork, which needs specific atmospheric data.

3.3 PatchCore

PatchCore [24] is a state-of-the-art, training-free anomaly detection method, which relies on comparing deep feature embeddings of image patches to a memory bank, which represent the normal samples. It is composed of two main parts, which are the memory bank construction and the anomaly scoring algorithm.

This technique requires three different splits of data for testing:

- **Memory Bank Data**, normal samples used to extract the memory bank.
- **Normal Test Set**, normal samples analyzed during test phase.
- **Anomalous Test Set**, anomalous samples analyzed during test phase.

Anomaly is not defined apriori, neither what the normality is. Depending on the test datasets, everything inside the normal splits will represent what is normal, leaving as anomalous everything else unseen in the normal set. A specific test case may be, like in this case, to consider as normal, satellite images without any wildfire or burned area, leaving images representing these events as anomalous.

The overall performance of PatchCore strongly depends on the backbone, because the core assumption, considering it is distance-based, is that normal samples will be represented close to each other in the embedding space, while anomalous samples will be represented further away from the norm.

Let's now see how the memory bank is built and how the scoring algorithm works.

3.3.1 Memory Bank Construction

Let $f_\theta : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{D\times h\times w}$ denote the foundation model backbone used in the pipeline (HyperFree or Copernicus-FM, see Section 3.2), where $f_\theta(x)$ produces a spatial feature map with D channels and spatial resolution $h \times w$.

Given an input image $x \in \mathcal{X}_N$ belonging to the set of normal images, the backbone outputs a feature tensor

$$F = f_\theta(x) \in \mathbb{R}^{D \times h \times w}.$$

Each spatial location (i, j) of F corresponds to a *patch embedding*

$$\mathbf{z}_{i,j} \in \mathbb{R}^D.$$

By flattening the spatial dimensions, each image is represented as a set of patch embeddings

$$\mathcal{Z}(x) = \left\{ \mathbf{z}_k \in \mathbb{R}^D \;\middle|\; k = 1, \ldots, h \cdot w \right\}.$$

Let $\mathcal{X}_N^{\mathrm{MB}} \subset \mathcal{X}_N$ denote the subset of normal images used for memory bank construction. The full (uncompressed) memory bank is defined as the union of all patch embeddings extracted from these images:

$$\mathcal{M}_{\mathrm{full}} = \bigcup_{x \in \mathcal{X}_N^{\mathrm{MB}}} \mathcal{Z}(x) \;\subset\; \mathbb{R}^D.$$

Since $\mathcal{M}_{\mathrm{full}}$ may contain millions of patch embeddings, PatchCore applies a *coreset subsampling* strategy to obtain a compact yet representative memory bank. In this work, the random coreset strategy is adopted, with a fixed sampling ratio $\rho = 0.1$, where a fraction ρ of patches is uniformly sampled from the full set of normal patch embeddings:

$$\mathcal{M} \subset \mathcal{M}_{\mathrm{full}}, \qquad |\mathcal{M}| = \rho \, |\mathcal{M}_{\mathrm{full}}|.$$

The resulting set $\mathcal{M} = \{\mathbf{m}_1, \ldots, \mathbf{m}_{|\mathcal{M}|}\}$ is referred to as the memory bank and represents the distribution of normal patch embeddings.

3.3.2 Anomaly Scoring

At test time, given a query image x, its patch embeddings

$$\mathcal{Z}(x) = \left\{ \mathbf{z}_k \in \mathbb{R}^D \right\}_{k=1}^{h \cdot w}$$

are extracted using the same backbone f_θ.

For each patch embedding $\mathbf{z}_k$, PatchCore computes its distance to the memory bank using a nearest-neighbor criterion:

$$s_k(x) = \min_{\mathbf{m} \in \mathcal{M}} \left\| \mathbf{z}_k - \mathbf{m} \right\|_2.$$

This defines a *patch-level anomaly score*, where larger values indicate a higher deviation from the normal feature distribution.

To obtain an *image-level anomaly score*, patch-level scores are aggregated across the image. In this work, this is done through mean aggregation, even though another possible option is to assign the highest patch-level anomaly score to the whole image:

$$S(x) = \frac{1}{h \cdot w} \sum_{k=1}^{h \cdot w} s_k(x).$$

The scalar $S(x)$ represents the final anomaly score assigned to image x.

Given a test set composed of normal and anomalous images, anomaly detection performance is evaluated by thresholding $S(x)$ and computing standard metrics such as AUROC, F1-score, precision, and recall. Importantly, PatchCore does not require any training on anomalous data: the entire notion of abnormality emerges from the distance between test-time patch embeddings and the memory bank built from normal samples only.

In the experiments, the *decision threshold* on $S(x)$ is selected as the value that maximizes the F1-score on the validation set. Alternatively, one could set the threshold based on a chosen percentile of the anomaly score distribution, trading off between false positives and false negatives depending on the application. An overview of the PatchCore pipeline can be found in Figure 3.1.

3.4 Semi-Supervised Fine-Tuning Strategy

While PatchCore operates in a training-free manner, its effectiveness critically depends on the quality and geometry of the feature representations produced by the backbone. In particular, distance-based anomaly detection methods implicitly assume that normal samples form a compact distribution in the embedding space, whereas anomalous samples lie at larger distances from this distribution.

Although foundation models provide strong general-purpose representations, they are not explicitly optimized to satisfy this geometric constraint for a specific anomaly detection task or domain. To address this mismatch, a semi-supervised fine-tuning strategy is introduced, aimed at adapting the backbone feature space to better align with the requirements of patch-based anomaly detection.

The proposed fine-tuning approach preserves the generalization capabilities of the pre-trained foundation model while reshaping the embedding space using a small amount of task-specific supervision.

Specifically, the approach leverages three fundamental loss terms, each one with its specific function:

- **Center Loss**, acts on normal samples.
- **Margin Loss**, acts on anomalous samples.

- **Distillation Loss**, acts on both normal and anomalous samples.

The goal was to reshape the embedding space such that normal samples could be more compact with respect to their center, anomalous samples would be pushed away, at least by a specified margin, while keeping the spatial informativeness of the embedding given by the baseline model.

First of all, to define the syntax, let $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N_{\text{train}}}$ denote the semi-supervised training set, with labels $y_i \in \{0,1\}$ indicating normal ($y_i = 0$) or anomalous ($y_i = 1$) samples, f_{θ_0} denotes the frozen teacher and f_θ the student encoder.
The normal center ($c \in \mathbb{R}^D$), which is the central point in the embedding space of the normal samples cluster, is computed from the teacher embeddings of normal training samples only. Defining

$$\mathcal{N} = \{i \mid y_i = 0\}, \qquad \mathcal{A} = \{i \mid y_i = 1\},$$

it is possible to estimate

$$c \;=\; \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} f_{\theta_0}(x_i).$$

3.4.1 Center Loss

Center loss acts only on normal samples ($\mathcal{N}$), it is defined as the squared distance between each normal embedding and the normal center. It has been introduced to push the normal samples toward the normal center, in order to obtain a more compact representation, as expected from PatchCore.

The formal definition of center loss is:

$$\mathcal{L}_{\text{center}}^{\text{norm}}(\theta) \;=\; \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \Big\| f_\theta(x_i) - c \Big\|_2^2.$$

3.4.2 Margin Loss

This is the loss term that is just considered by anomalous samples ($\mathcal{A}$). Instead of pulling them toward the normal center, a margin is enforced on their distance from c. Let $d_i = \|f_\theta(x_i) - c\|_2$ denote the Euclidean distance of an anomalous embedding to the center and let $m > 0$ be a fixed margin, the loss is defined as:

$$\mathcal{L}_{\text{center}}^{\text{anom}}(\theta) \;=\; \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \Big[\max(0,\, m - d_i)\Big]^2,$$

It penalizes the anomalous embeddings which lie closer than m to the normal center, becoming zero when $d_i \geq m$.

3.4.3 Distillation Loss

Keeping the two center-based loss terms without adding a mitigation could lead to the collapse of the whole embedding space. That's why distillation loss was introduced. It acts both on normal and anomalous samples, and its role is to keep the geometry of the pre-trained representation, while normal are pulled toward the center and anomalous are pushed away.

The implementation is based on the teacher-student paradigm, where the teacher (f_{θ_0}) is frozen and represented by the baseline model, and the student (f_θ) is updated during finetuning. Specifically, it computes the square distance between the teacher representation and the student representation:

$$\mathcal{L}_{\text{distill}}(\theta) \;=\; \frac{1}{|\mathcal{N}| + |\mathcal{A}|} \sum_{i \in \mathcal{N} \cup \mathcal{A}} \left\| f_\theta(x_i) - f_{\theta_0}(x_i) \right\|_2^2.$$

3.4.4 Final Objective and Validation

The two center-related terms are combined together

$$\mathcal{L}_{\text{center}}^{\text{total}}(\theta) \;=\; \mathcal{L}_{\text{center}}^{\text{norm}}(\theta) \;+\; \lambda_{\text{anom}}\, \mathcal{L}_{\text{center}}^{\text{anom}}(\theta),$$

And the final objective is

$$\mathcal{L}(\theta) \;=\; \lambda_c\, \mathcal{L}_{\text{center}}^{\text{total}}(\theta) \;+\; \lambda_d\, \mathcal{L}_{\text{distill}}(\theta),$$

Four hyperparameters are used to balance between compactness of the normal cluster, separation of anomalies from the center and faithfulness to the baseline embeddings:

- λ_c, regulates the center-related terms.
- λ_d, acts on the distillation term.
- λ_{anom}, controls the relative importance of the anomaly margin with respect to the clustering of normal samples.
- m, the anomaly margin, defines the minimum distance between anomalous samples and the normal center.

The validation is performed leveraging PatchCore at the end of each epoch on the validation set. Its normal samples are split in two half, the first one used for the creation of the memory bank and the second one for normal test, while the totality of anomalous samples are used for test. After the feature extraction each test sample receives an anomaly score, and AUROC is computed. The best checkpoint is selected on validation AUROC.

Chapter 4

Experiments

This chapter presents the experimental evaluation of the proposed anomaly detection pipeline based on geospatial foundation models and PatchCore. The analysis is structured to progressively investigate representation quality, domain robustness, and adaptability.

First, the evaluation protocol, metrics, and dataset roles are defined. Frozen-backbone performance is then analyzed under both homogeneous and heterogeneous settings to assess cross-domain generalization. A dedicated spectral invariance study further investigates whether embeddings remain stable under spectral response function (SRF) transformations.

Finally, a semi-supervised fine-tuning strategy is introduced to reshape the embedding space and mitigate domain shift effects. The impact of fine-tuning is evaluated both quantitatively and geometrically through image-level and patch-level analyses. The chapter concludes with a computational comparison of the considered backbones.

4.1 Experimental Goals

The focus in this work has been to adapt an industrial anomaly detection technique, PatchCore, to a completely different domain, remote sensing imagery, where the number of input bands for each image and the acquisition sensor change dramatically from the original domain of the technique.

Starting from this main objective, a set of sub-goals has been defined:

- Evaluate whether large-scale geospatial foundation models, HyperFree and Copernicus-FM, can serve as effective feature extractors for patch-based anomaly detection methods on multispectral and hyperspectral satellite imagery.

- Analyze the separability between normal and anomalous samples in the embedding space produced by frozen foundation model backbones, and how it changes when introducing a fine-tuning strategy.
- Quantify the anomaly detection performance of PatchCore across different datasets, sensor configurations and backbone choices, in order to evaluate robustness and generalization capabilities.
- Investigate an effective fine-tuning strategy for the backbones to improve the separability of normal and anomalous samples in specific experimental configurations.

4.2 Evaluation Protocol

In this section the evaluation protocol adopted to assess the proposed anomaly detection pipeline is described. Specifically, the protocol defines how datasets are split and used, how the foundation model backbones are used and fine-tuned, and how PatchCore is applied and evaluated.

All the experiments conducted in this thesis follow a two-stage evaluation:

- Feature extraction using a foundation model backbone, whether the baseline or finetuned version.
- Anomaly detection using PatchCore.

4.2.1 Data Splits and Usage

To evaluate the pipeline, three fundamental data splits are needed:

- **Memory Bank Data**, which is a split of normal data used to build the PatchCore memory bank, represents the distribution of normality in the embedding space.
- **Normal Test Set**, a split of normal data used for testing, so their anomaly scores with respect to the memory bank are computed, to then find a threshold for evaluation.
- **Anomalous Test Set**, the split of anomalous data, used exclusively for evaluation.

These three splits are always needed, whether the baseline pipeline or the finetuned version is evaluated. The memory bank data and normal test set may also come from the same normal dataset. In particular, in the experimental configurations that required both splits to come from the same dataset, the split protocol is well defined:

- The entire dataset is split between even and odd samples, one half is used for the memory bank construction and the other half for testing. Using this deterministic splits reproducibility and fair comparison between different configurations are ensured.

Talking specifically about the experimental configurations, they can be divided between the baseline pipeline and the fine-tuned pipeline. For the baseline pipeline configurations, no training is performed. Normal datasets are used exclusively to construct the memory bank and the normal test set, while anomalous datasets are used only at test time. Otherwise, for the fine-tuned pipeline, data are split in train, validation and test set.

Considering the finetuning is performed on FLOGA dataset only, it is possible to be more precise about the train, validation and test splits. Table 4.1 quantitatively describes the splits, while the percentage refers to the total.

Table 4.1: Train, validation, and test split statistics for the FLOGA dataset.

Type	**Train** (60%)	**Val** (8%)	**Test** (32%)
Normal	103,475	13,794	55,190
Anomalous	245	34	130

The split proportions were chosen to preserve a large training set for fine-tuning while maintaining a sufficiently representative test set for evaluation, also considering the high imbalance of the chosen dataset.

4.2.2 Backbone Usage and Fine-Tuning

In baseline configurations, the foundation model backbone is used as a frozen feature extractor, to assess the feasibility of using pretrained foundation models for the anomaly detection task. In fine-tuned configurations, the backbone is updated using the training split, while PatchCore is kept unchanged. In all cases, the backbone outputs patch-level embeddings that are subsequently processed by PatchCore.

4.2.3 PatchCore Application and Evaluation

PatchCore method is applied consistently across all configurations. The memory bank is always constructed using normal samples only, while anomaly scores are computed for both normal and anomalous test samples. During fine-tuning, PatchCore performance on the validation set is used for model selection using AUROC, while final results are reported on the test set. No anomalous samples are used to update the backbone parameters.

4.3 Metrics

This section describes which metrics are used to evaluate anomaly detection, dividing them into a primary, threshold-independent, metric, and a set of threshold-dependent metrics. All the computed metrics are going to be image-level metrics, even though the pipeline provides patch-level anomaly scores, they are aggregated at image-level using arithmetic mean, in order to have one single image-level anomaly score per sample.

4.3.1 Primary Metric - AUROC

The separability between normal and anomalous samples in the embedding space is a key factor in anomaly detection performance. For this reason, AUROC is selected as the primary evaluation metric. Area Under the ROC Curve (AUROC), is a performance metric that measures the model's ability to distinguish between positive and negative classes, in our case normal and anomalous, across all possible thresholds.

In other words, it can be interpreted as a separability index, representing the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Thanks to its threshold-independent behavior, it allows to objectively measure and compare the pipeline performances across multiple and different experimental configurations, considering that threshold is a configuration-specific value.

The AUROC value ranges from 0.0 to 1.0. A score of 1.0 represents a perfect model, that completely separates the two classes, while 0.5 represents a random classifier. Figure 4.1 provides a visual representation of the metric.

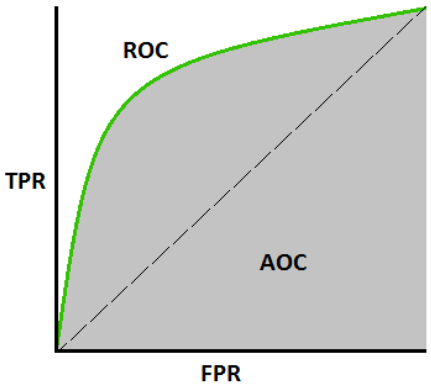


Figure 4.1: AUROC Representation

To formally define AUROC, the ROC Curve must be introduced. It is a curve constructed by finding TPR and FPR at different thresholds (τ).

True Positive Rate (TPR) measures the fraction of anomalous samples correctly detected

$$\mathrm{TPR}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FN}(\tau)}.$$

False Positive Rate (FPR) measures the fraction of normal samples incorrectly classified as anomalous

$$\mathrm{FPR}(\tau) = \frac{\mathrm{FP}(\tau)}{\mathrm{FP}(\tau) + \mathrm{TN}(\tau)}.$$

The ROC curve is defined as

$$\mathrm{ROC} = \left\{ \Big(\mathrm{FPR}(\tau), \mathrm{TPR}(\tau) \Big) \;\Big|\; \tau \in \mathbb{R} \right\}.$$

And finally, AUROC is defined as

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d(\mathrm{FPR}).$$

4.3.2 Threshold-Dependent Metrics

Besides the primary metric, a set of secondary, threshold-dependent metrics to evaluate the thresholding results of the anomaly scores have been computed. The decision threshold is chosen by picking the one that maximizes the F1-Score, ensuring a balance trade-off between precision and recall. At this point it is possible to assign to each test sample a label (normal or anomalous), and depending on that, these metrics can be computed:

- **Accuracy**, measures the fraction of correctly classified samples over the total number of test samples:

  $$\mathrm{Accuracy}(\tau) = \frac{\mathrm{TP}(\tau) + \mathrm{TN}(\tau)}{\mathrm{TP}(\tau) + \mathrm{TN}(\tau) + \mathrm{FP}(\tau) + \mathrm{FN}(\tau)}.$$

  However, accuracy can be misleading in highly imbalanced anomaly detection settings.

- **Precision**, measures reliability of anomaly predictions. High precision means samples classified as anomalous are likely to be truly anomalous:

  $$\mathrm{Precision}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FP}(\tau)}.$$

- **Recall**, measures the ability of the model to correctly identify anomalous samples:
$$\text{Recall}(\tau) = \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FN}(\tau)}.$$
- **F1-Score**, represents the harmonic mean of precision and recall, it is a balanced index:
$$\text{F1}(\tau) = 2 \cdot \frac{\text{Precision}(\tau) \cdot \text{Recall}(\tau)}{\text{Precision}(\tau) + \text{Recall}(\tau)}.$$

This set of metrics is considered secondary because threshold changes for each experimental setup, making cross-configuration comparison less straightforward. They still result useful in the anomaly detection evaluation, providing a complementary perspective on the classification behavior.

4.4 Datasets

This section describes the datasets selected for the experiments after the literature analysis (Section 2.3). As previously mentioned, remote sensing datasets vary substantially from each other. Differences may concern acquisition sensor, spectral modality, spatial resolution and annotation protocol.

When using models that are not invariant to such factors, these differences may significantly affect performance, leading to the need of mitigation solutions, for instance, fine-tuning.

Datasets can be categorized according to their role in the experimental pipeline:

- **Normal Datasets**, composed of generic satellite images. Foundation datasets, without annotations and assumed to approximate the background data distribution in the anomaly detection setting.
- **Anomalous Datasets**, satellite images capturing an anomalous event, in this specific scope active wildfire or burned area. Smaller than normal datasets, they come with anomaly annotations.

And by modality in:

- **Multispectral Datasets (MSI)**, composed of images with a limited number of discrete spectral bands, typically ranging from four to a few tens, each one covering a broad wavelength interval.
- **Hyperspectral Datasets (HSI)**, composed of images with a large number of contiguous and narrow spectral bands, ranging from tens to several hundreds of bands.

4.4.1 Copernicus Pretrain

Copernicus Pretrain [31] is a multispectral foundation dataset, built for Copernicus-FM training. It is an unannotated dataset and, in this thesis, covers the role of normal dataset. It provides images from different sensors in the multispectral range, specifically, it is composed of images acquired by Sentinel-1, Sentinel-2 and Sentinel-3, also with atmospheric information from Sentinel-5P and DEM data.

For the thesis purpose, only images from Sentinel-2 were used, with GSD $10m$ and image size $264 \times 264 \times 13$, to keep a consistent spectral response. Its $997k$ samples provide global coverage, even though for this experiments, only one third of the total samples were used, due to computation efficiency. The sub-split is created by selecting one sample every three in a deterministic manner, in order to keep the reproducibility. As a normal dataset, it has been used for memory bank construction and normal test set.

Figure 4.2: Copernicus Pretrain Sample

4.4.2 SpectralEarth

SpectralEarth [4] is a large-scale, multi-temporal dataset, designed to pre-train hyperspectral foundation models. It consists of $540k$ samples, with image size $128 \times 128 \times 202$ and GSD $30m$.

Composed of EnMAP hyperspectral acquisitions, it is used as a normal dataset under the assumption that it predominantly contains non-anomalous background scenes, as no anomaly annotations are provided. To ensure computational efficiency, a sub-split of $77k$ samples, representative for the whole dataset, has been extracted, picking a sample each fixed step, also to preserve the global coverage.

Figure 4.3: SpectralEarth Sample

SpectralEarth 7 Bands

Moved by the need of deleting the sensor-shift bias between datasets acquired with different sensors, to make them comparable, a multispectral Landsat-8-like version of SpectralEarth has been created, called SpectralEarth 7 Bands.

SpectralEarth was originally provided as hyperspectral EnMAP imagery with C narrow bands (here $C = 202$), each sample being a cube $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H and W denote the spatial dimensions and the third axis indexes the spectral channels. To obtain a multispectral version compatible with Landsat–8, 7 broad bands (B1–B7) were synthetized by convolving the EnMAP spectra with the corresponding Landsat–8 spectral response functions (SRFs).

Let $\{\lambda_k\}_{k=1}^{C}$ denote the EnMAP wavelength grid (in nm), with associated spectral steps $\Delta\lambda_k$ (in general non-uniform), and let $R(x, y, \lambda_k)$ be the reflectance value at pixel (x, y) and wavelength λ_k. For each Landsat–8 band $b \in \{1, \ldots, 7\}$, starting from the tabulated SRF

$$S_b^{\text{raw}}(\lambda) \quad \text{defined on its native wavelength grid,}$$

and interpolating it onto the EnMAP wavelength grid, obtaining a discrete SRF vector

$$S_b(\lambda_k), \qquad k = 1, \ldots, C,$$

where values outside the support of the original SRF are set to zero.

The synthesized reflectance for band b at pixel (x, y) is then defined as an

SRF-weighted spectral average of the original EnMAP spectrum:

$$I_b(x,y) = \frac{\int R(x,y,\lambda)\, S_b(\lambda)\, d\lambda}{\int S_b(\lambda)\, d\lambda} \approx \frac{\sum_{k=1}^{C} R(x,y,\lambda_k)\, S_b(\lambda_k)\, \Delta\lambda_k}{\sum_{k=1}^{C} S_b(\lambda_k)\, \Delta\lambda_k}.$$

In practice, for each hyperspectral image this discrete integral is computed, for all pixels (x, y) and for each band b, obtaining a multispectral cube

$$\mathbf{Y} \in \mathbb{R}^{H \times W \times 7},$$

which approximates Landsat–8-like bands while preserving the spatial resolution of the original SpectralEarth samples.

4.4.3 Active Fire

Active Fire [7] is a multispectral dataset labeled for active fire detection. Images come from the Landsat-8 sensor, with an image size of $256 \times 256 \times 10$ and GSD $30m$. The dataset is geographically divided in 11 splits, counting a total of 146212 images. Starting from the original 10 bands, the last two had to be excluded due to spectral incompatibility with the chosen backbones, specifically the two thermal bands, with central wavelengths of $11\mu m$ and $12\mu m$, as the backbones are able to handle bands at most at $2.5\mu m$. Therefore, only 8 out of the original 10 bands are used in the experiments, having an image size of $256 \times 256 \times 8$.

About annotations, each patch is associated with at least one binary mask, considering 0 non-fire and 1 fire. Different handcrafted algorithms, relying on statistics from the channels, for active fire detection were used to create the annotations:

- **Schroeder**, proposes a set of conditions that uses 7 Landsat-8 channels to identify active fire pixels.
- **Murphy**, relies only on three channels.
- **Kumar-Roy**, most recent, is based on 6 channels.

Besides this three, also the intersection and the voting between these three are considered for annotations. Not all the patches have all 5 possible annotations. Each image that has at least one fire pixel is considered anomalous.

The main dataset came also with an auxiliary dataset, containing 9044 samples, manually annotated. Active Fire is used in this thesis only as anomalous dataset.

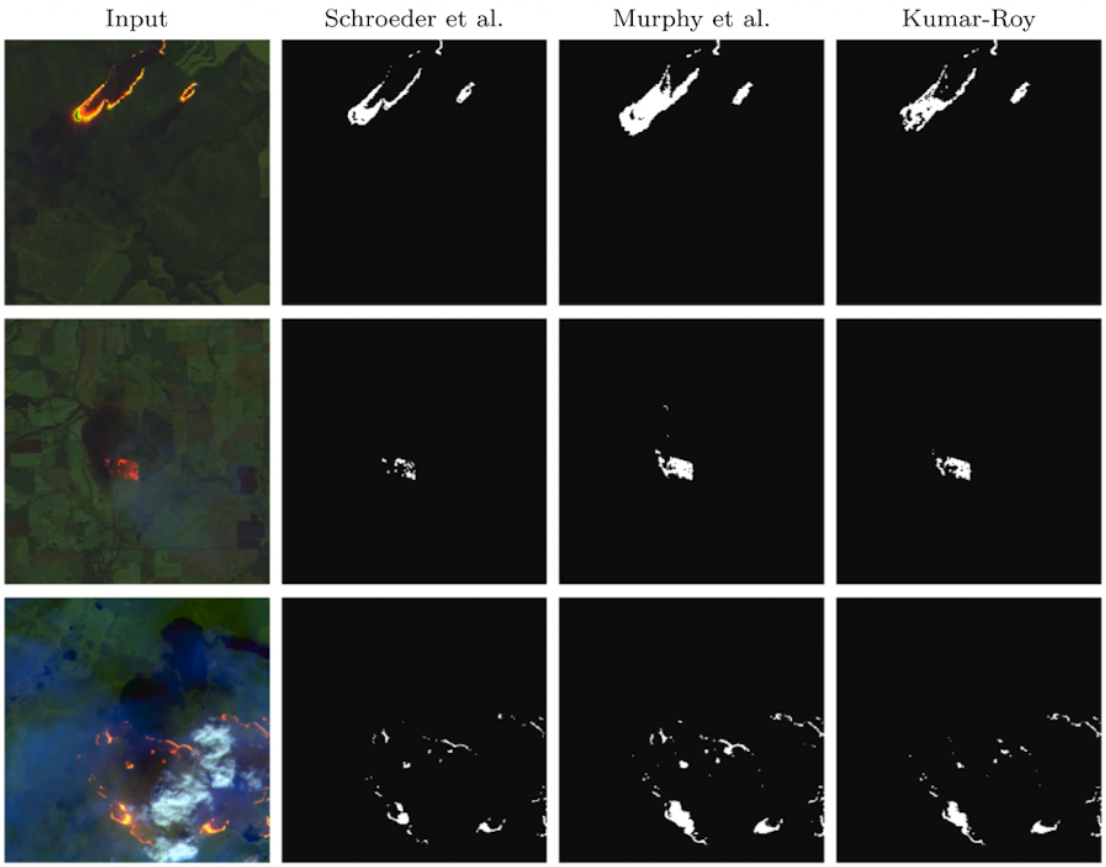


Figure 4.4: Active Fire Sample

4.4.4 FLOGA

FLOGA [25] is a multispectral dataset, originally used as a change detection dataset. Composed of Sentinel-2 images, it represents 326 wildfire events that occurred in Greece between 2017 and 2021. It comes with $86.5k$ pairs of pre- and post-event images, a total of $173k$ single images, having an image size of $256 \times 256 \times 9$ and a GSD of $20m$.

About annotations, FLOGA provides binary masks on burned area, having pixel value 0 representing no-burn, and pixel value 1 representing a burnt pixel. Further investigating the data, it was clear, against what could be expected, that not all post-event images contain burned pixels, but a very small percentage, 0.5% of the total.

Considering the anomaly detection pipeline requires images to be treated singularly, without any temporal link among them, the pre-event and post-event splits were redesigned in:

- **FLOGA Normal**, containing all the pre-event and post-event images without any burnt pixel, composed of a total of $172.5k$ images.

- **FLOGA Anomalous**, containing the post-event images with at least one burnt pixel, composed of a total of 409 images.

After this split, FLOGA Normal is used in this work as normal dataset, both for memory bank construction and normal test set, while FLOGA Anomalous is just used as anomalous dataset.

Figure 4.5: FLOGA Samples - Left image from FLOGA Normal - Center image from FLOGA Anomalous - Right image Ground-Truth

4.4.5 Dataset Summary and Experimental Roles

The selected datasets span three different sensors (EnMAP, Sentinel-2 and Landsat-8), multiple spatial resolutions (10m, 20m and 30m GSD), and both hyperspectral and multispectral modalities. This heterogeneity introduces spectral and spatial domain shifts that directly affect feature extraction and anomaly scoring, especially in frozen-backbone settings.

In these experiments, datasets are assigned specific operational roles:

- Copernicus Pretrain and SpectralEarth are used as large-scale normal reference distributions in the frozen-backbone experiments, providing samples for PatchCore memory bank construction and normal test evaluation.

- Active Fire is used exclusively as anomalous dataset in cross-sensor heterogeneous experiments.

- FLOGA is used both as normal and anomalous datasets after reformulation, enabling a homogeneous Sentinel-2 setting and a controlled semi-supervised fine-tuning scenario.

Regarding FLOGA, the reformulation from paired change detection to single-image anomaly detection allows the task to be aligned with PatchCore image-level anomaly scoring paradigm. The resulting class distribution is highly imbalanced, with anomalous samples representing less than 1% of the total images. This extreme imbalance reflects realistic rare-event conditions and motivates the semi-supervised fine-tuning strategy introduced in Section 3.

Overall, the dataset configurations enables evaluation under both heterogeneous (cross-sensor) and homogeneous (same-sensor) conditions, allowing to isolate the impact of domain shift and to assess whether fine-tuning mitigates performance degradation. An overview of the datasets is provided in Table 4.2.

Dataset	Sensor	Modality	Bands	GSD	# Samples Used
Copernicus Pretrain (S2)	Sentinel-2	MSI	13	10m	∼332k (1/3 of 997k)
SpectralEarth	EnMAP	HSI	202	30m	77k (subset of 540k)
SpectralEarth 7 Bands	EnMAP → L8-like	MSI (synthetic)	7	30m	77k
Active Fire	Landsat-8	MSI	8 (used)	30m	146k
FLOGA Normal	Sentinel-2	MSI	9	20m	172.5k
FLOGA Anomalous	Sentinel-2	MSI	9	20m	409

Table 4.2: Overview of datasets used in the experimental section

4.5 Frozen Backbone Baseline Experiments

This section presents a comprehensive empirical evaluation of the proposed anomaly detection pipeline, in its baseline version, using the backbone's weights released by the authors. Specifically, the following pretrained checkpoints were used

- `CopernicusFM_ViT_base_varlang_e100.pth` for Copernicus-FM.
- `HyperFree-b.pth` for HyperFree.

The analysis is organized in three main parts:

- **Homogeneous Analysis**, where a single normal dataset is used both to build the memory bank and for normal testing. More information on data splitting is provided in Section 4.2.1.
- **Heterogeneous Analysis**, where memory bank and normal test images are drawn from different normal datasets to expose cross-domain effects.
- **FLOGA Analysis**, where the anomalous class is provided by FLOGA Anomalous, shifting the focus from active fires to burned areas. Two further settings are considered:

– **Cross-Domain**, using a Copernicus Pretrain memory bank and FLOGA for testing.

– **In-Domain**, using FLOGA Normal also for the memory bank construction.

In the first two analysis, the anomalous class is given by Active Fire, defining active wildfires as the desired anomalies, while the normal datasets are Copernicus Pretrain and SpectralEarth 7 Bands. In the last part instead, the test domain is given by the burned pixels in the FLOGA Anomalous dataset, using FLOGA Normal as normal test set.

4.5.1 Homogeneous Analysis

In the homogeneous setting, the same normal dataset is used to build the PatchCore memory bank and the normal test set. The experimental settings and data splits for this analysis are described in Table 4.3, clarifying the difference between image and patch:

- **Image**, the sample entering the backbone.

- **Patch**, after passing through the backbone, an image is divided into 16 patches, arranged on a 4×4 grid.

The number of patches in the memory bank set refers to the number of patches stored in the memory bank after sub-sampling (sampling ratio 0.1).

Memory Bank Set	Normal Test Set	Anomalous Test Set
Copernicus Pretrain (249504 patches)	Copernicus Pretrain (155940 images)	Active Fire (146214 images)
SpectralEarth 7 Bands (62286 patches)	SpectralEarth 7 Bands (38929 images)	Active Fire (146214 images)

Table 4.3: Homogeneous experimental settings and data splits

In these settings anomalies are defined as active wildfires from Active Fire dataset. Table 4.4 summarizes the results for both HyperFree and Copernicus-FM backbones.

Backbone	Memory Bank	Normal Test	Anom. Test	Thr.	Acc.	Prec.	Rec.	F1	AUROC
HyperFree	Copernicus Pretrain	Copernicus Pretrain	Active Fire	0.5857	0.8997	0.8970	0.8956	0.8963	0.9620
HyperFree	SpectralEarth 7 Bands	SpectralEarth 7 Bands	Active Fire	0.4385	0.9265	0.9518	0.9553	0.9536	0.9708
Copernicus	Copernicus Pretrain	Copernicus Pretrain	Active Fire	34.497	0.7153	0.6507	0.8884	0.7512	0.8236
Copernicus	SpectralEarth 7 Bands	SpectralEarth 7 Bands	Active Fire	14.650	0.9144	0.9165	0.9811	0.9477	0.8380

Table 4.4: Homogeneous analysis results.

Image-Level Comparative Discussion. The homogeneous setting evaluates the intrinsic discriminative power of the feature representations when the memory bank and the normal test samples originate from the same dataset and sensor. Under this configuration, both backbones achieve non-trivial anomaly detection performance against Active Fire anomalies.

For HyperFree, AUROC values reach 0.9620 when using Copernicus Pretrain and 0.9708 with SpectralEarth 7 Bands. The second configuration outperforms the first one across all metrics, suggesting that HyperFree features yield a more separable normal/anomalous score distribution under the SpectralEarth 7 Bands setting.

For Copernicus-FM, performance is notably lower in terms of AUROC (0.8236 and 0.8380 respectively), despite achieving high recall values, especially with SpectralEarth 7 Bands. This behavior indicates that while anomalous samples tend to receive high anomaly scores, the ranking quality between normal and anomalous samples is weaker compared to HyperFree.

The difference of approximately 0.13 AUROC between the two backbones suggests that HyperFree produces a more separable and structured embedding representation for anomaly ranking, even before considering domain shift effects.

However, it is important to highlight a methodological aspect of this setting. Although memory bank and normal test samples originate from the same dataset (minimizing intra-class domain variability), the anomalous class (Active Fire) comes from a different dataset acquired by a different sensor (Landsat-8). Consequently, part of the separability observed in this homogeneous configuration may still be influenced by cross-sensor discrepancies between normal and anomalous domains.

This limitation motivates the heterogeneous analysis presented in the next subsection, where domain shift is explicitly introduced between memory bank and normal test data to disentangle intrinsic representational quality from cross-domain effects.

Image- and Patch-Level Analysis

To further confirm cross-sensor discrepancies and better characterize the internal behavior of the anomaly detection pipeline, anomaly scores at both patch and image level are analyzed. Beyond aggregate metrics such as AUROC, this analysis provides insight into score distributions, cluster compactness, and ranking separability.

For each configuration, the following are reported:

- **Patch-Level Histogram**, to visualize overlap between normal and anomalous score distributions.
- **Patch-Level Boxplot**, to assess dispersion and interquartile range differences.
- **Image-Level ROC Curve**, to evaluate ranking separability across thresholds.

Results are grouped by the chosen normal dataset (Copernicus Pretrain vs. SpectralEarth 7 Bands), while keeping the anomalous domain fixed (Active Fire).

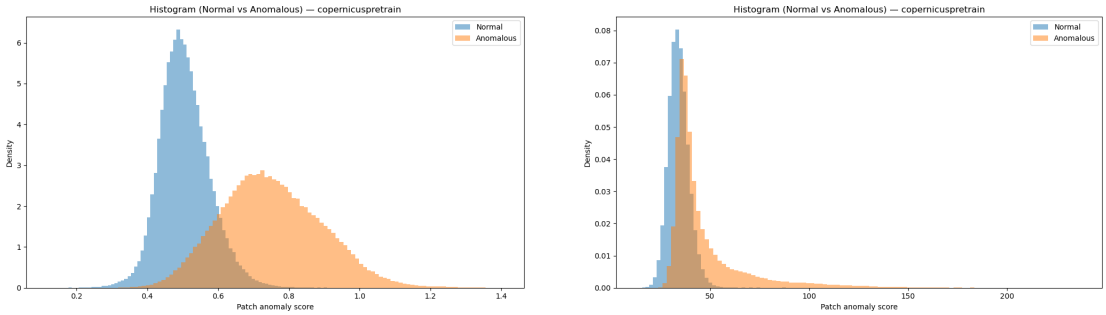
Copernicus Pretrain

Table 4.5 summarizes the quantitative results.

Backbone	Memory Bank	Normal Test	Anom. Test	Thr.	Acc.	Prec.	Rec.	F1	AUROC
HyperFree	Copernicus Pretrain	Copernicus Pretrain	Active Fire	0.5857	0.8997	0.8970	0.8956	0.8963	0.9620
Copernicus	Copernicus Pretrain	Copernicus Pretrain	Active Fire	34.497	0.7153	0.6507	0.8884	0.7512	0.8236

Table 4.5: Homogeneous analysis results - Copernicus Pretrain.

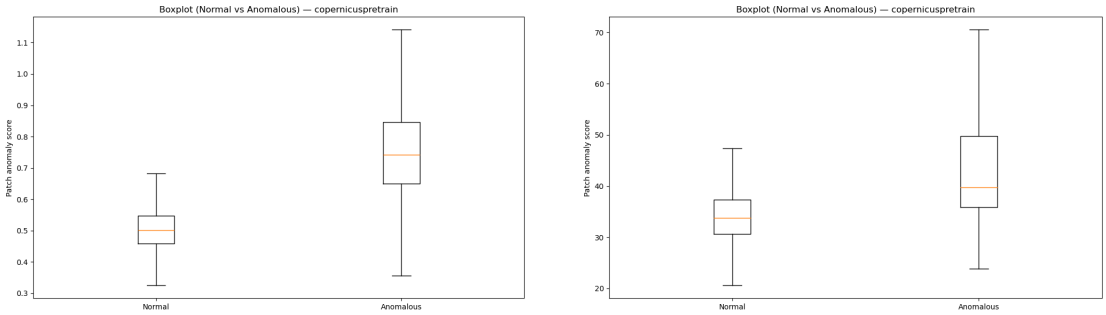
Patch-Level Histogram. As shown in Figure 4.6, HyperFree exhibits clearer separation between normal and anomalous distributions, with limited overlap between peaks. In contrast, Copernicus-FM presents stronger overlap, indicating weaker discriminative structure in the embedding space. This behavior is consistent with the AUROC gap (0.9620 vs 0.8236).



(a) HyperFree **(b)** Copernicus-FM

Figure 4.6: Patch-level anomaly score histogram (Homogeneous - Copernicus Pretrain).

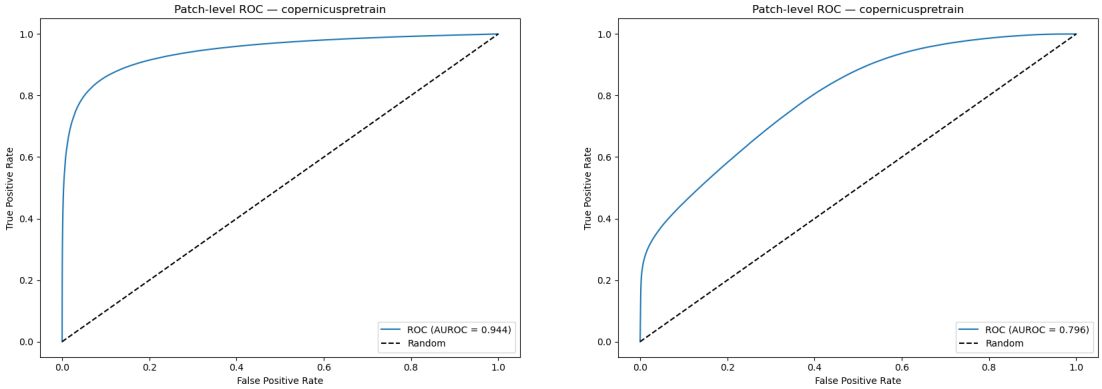
Patch-Level Boxplot. The boxplots in Figure 4.7 further highlight distribution spread. HyperFree shows tighter normal cluster compactness and clearer separation between medians. Copernicus-FM instead exhibits larger variance and overlapping interquartile ranges. Importantly, absolute score magnitude is not directly comparable across backbones, as PatchCore distances depend on embedding scaling.



(a) HyperFree **(b)** Copernicus-FM

Figure 4.7: Patch-level anomaly score boxplot (Homogeneous - Copernicus Pretrain).

Image-Level ROC Curve. Figure 4.8 confirms ranking behavior: HyperFree's curve bends sharply toward the top-left corner, while Copernicus-FM remains closer to the diagonal. This directly reflects the AUROC difference and indicates stronger ranking consistency for HyperFree.



(a) HyperFree **(b)** Copernicus-FM

Figure 4.8: Image-level ROC curves (Homogeneous - Copernicus Pretrain).

Overall, HyperFree demonstrates tighter normal cluster compactness and strong ranking separability, whereas Copernicus-FM exhibits higher intra-class variability leading to greater distribution overlap. It should be noted that although this setting is homogeneous for normal samples, anomalies originate from Landsat-8, so cross-sensor effects remain partially present.

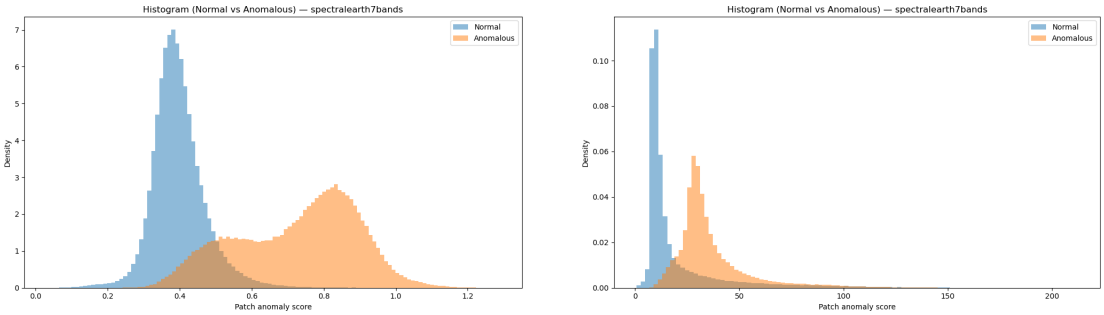
SpectralEarth 7 Bands

Table 4.6 summarizes quantitative results.

Backbone	Memory Bank	Normal Test	Anom. Test	Thr.	Acc.	Prec.	Rec.	F1	AUROC
HyperFree	Spectralearth 7 Bands	Spectralearth 7 Bands	Active Fire	0.4385	0.9265	0.9518	0.9553	0.9536	0.9708
Copernicus	Spectralearth 7 Bands	Spectralearth 7 Bands	Active Fire	14.650	0.9144	0.9165	0.9811	0.9477	0.8380

Table 4.6: Homogeneous analysis results - SpectralEarth 7 Bands.

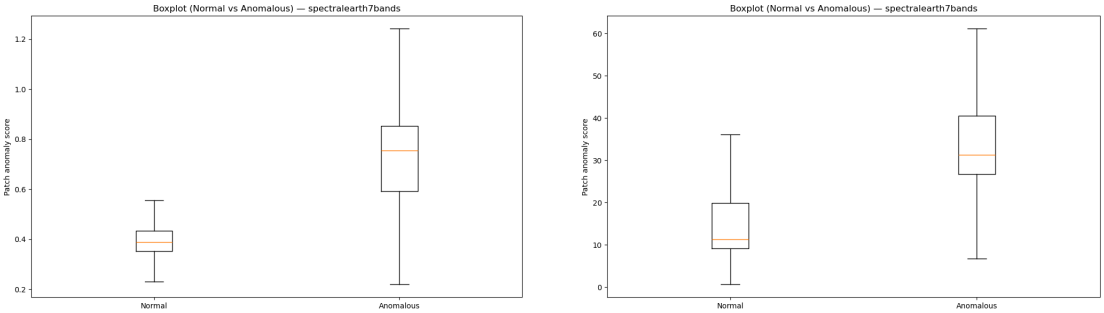
Patch-Level Histogram. Figure 4.9 shows improved separation compared to Copernicus Pretrain for both backbones. HyperFree maintains clear bimodal structure, while Copernicus-FM exhibits reduced but still noticeable overlap. AUROC improves for both models (0.9708 and 0.8380).



(a) HyperFree **(b)** Copernicus-FM

Figure 4.9: Patch-level anomaly score histogram (SpectralEarth 7 Bands memory bank).

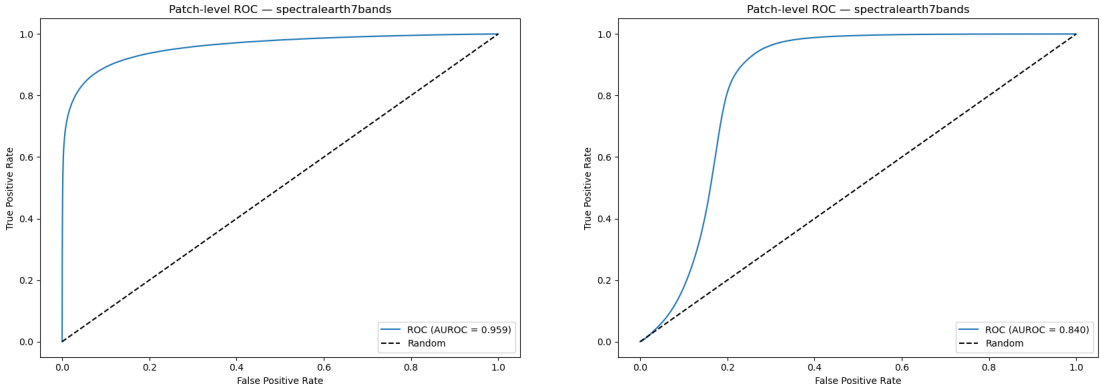
Patch-Level Boxplot. As shown in Figure 4.10, normal distributions appear more compact than in the Copernicus Pretrain case. This reduction in dispersion suggests improved cluster structure when spectral compatibility with Landsat-8 is increased.



(a) HyperFree **(b)** Copernicus-FM

Figure 4.10: Patch-level anomaly score boxplot (SpectralEarth 7 Bands memory bank).

Image-Level ROC Curve. Figure 4.11 confirms improved ranking reliability under reduced spectral mismatch. The spectral alignment between SpectralEarth 7 Bands and Landsat-8 likely contributes to this improvement.



(a) HyperFree **(b)** Copernicus-FM

Figure 4.11: Image-level ROC curves (SpectralEarth 7 Bands memory bank).

Overall, the SpectralEarth 7 Bands configuration reinforces the observation that spectral compatibility between memory bank and anomalous domain positively affects ranking reliability, while preserving the structural advantage of HyperFree over Copernicus-FM.

4.5.2 Heterogeneous Analysis

In the heterogeneous setting the memory bank is built from a normal dataset while the normal test set is drawn from a different dataset, keeping unchanged the anomalous test set (Active Fire). This configuration explicitly exposes cross-domain effects, allowing to understand if anomaly scores are driven by sensor and preprocessing differences rather than the presence of fire-related pixels.

Specifically, the analysis focuses on the following cross-domain pairs:

- **SpectralEarth 7 Bands (bank) → Copernicus Pretrain (normal)**.
- **Copernicus Pretrain (bank) → SpectralEarth 7 Bands (normal)**.

Experimental settings and data splits specified in Table 4.7.

Memory Bank Set	Normal Test Set	Anomalous Test Set
Copernicus Pretrain (62286 patches)	SpectralEarth 7 Bands (38929 images)	Active Fire (146214 images)
SpectralEarth 7 Bands (62286 patches)	Copernicus Pretrain (38929 images)	Active Fire (146214 images)

Table 4.7: Heterogeneous experimental settings and data splits

Unlike the homogeneous setting, where Copernicus Pretrain and SpectralEarth 7 Bands were not balanced, to make a fair heterogeneous comparison, especially in

terms of memory bank dimension, the two normal datasets have been balanced. Specifically, Copernicus Pretrain has been deterministically reduced to match SpectralEarth 7 Bands dimensions. Table 4.8 summarizes the results for all the four heterogeneous configurations.

Backbone	Memory Bank	Normal Test	Anom. Test	Thr.	Acc.	Prec.	Rec.	F1	AUROC
HyperFree	Copernicus Pretrain	SpectralEarth 7 Bands	Active Fire	0.3952	0.7897	0.7897	1.0000	0.8825	0.3183
HyperFree	SpectralEarth 7 Bands	Copernicus Pretrain	Active Fire	0.2706	0.7897	0.7897	1.0000	0.8825	0.0703
Copernicus	Copernicus Pretrain	SpectralEarth 7 Bands	Active Fire	33.151	0.8610	0.8700	0.9687	0.9167	0.7604
Copernicus	SpectralEarth 7 Bands	Copernicus Pretrain	Active Fire	7.9453	0.7897	0.7897	1.0000	0.8825	0.1733

Table 4.8: Heterogeneous analysis results.

Image-Level Comparative Discussion. The heterogeneous setting introduces a controlled domain shift between the memory bank and the normal test set, while keeping the anomalous domain (Active Fire) unchanged. This configuration allows disentangling semantic anomaly detection capability from cross-sensor and preprocessing biases.

The most striking observation concerns HyperFree. When the memory bank and the normal test set originate from different datasets, AUROC collapses to extremely low values (0.3183 and 0.0703 respectively). Such values indicate near-random or even inverted ranking behavior. Although recall remains artificially high (1.0000), this is a consequence of threshold selection rather than genuine separability: the model assigns systematically higher anomaly scores to the cross-domain normal samples, effectively treating them as anomalous.

This behavior suggests that, for HyperFree, anomaly scores are highly sensitive to distributional shifts between the memory bank and the test normal domain. When the embedding distribution of the normal test set differs from the bank distribution, PatchCore distances increase regardless of semantic content, leading to misclassification driven by domain mismatch rather than fire presence.

In contrast, Copernicus-FM exhibits more stable behavior under heterogeneous conditions, particularly in the configuration *Copernicus Pretrain (bank)* $\rightarrow$ *SpectralEarth 7 Bands (normal)*, where AUROC remains relatively high (0.7604). Although performance degrades compared to the homogeneous case, the ranking structure is partially preserved, indicating greater robustness to cross-domain shifts in this direction.

However, the reverse configuration *SpectralEarth 7 Bands (bank)* $\rightarrow$ *Copernicus Pretrain (normal)* also leads to a severe AUROC drop (0.1733), revealing that Copernicus-FM is not immune to domain mismatch. The asymmetry between

the two directions suggests that spectral compatibility between memory bank and anomalous domain (Landsat-8) plays a significant role: when the bank is spectrally closer to the anomalous sensor, ranking remains more stable.

Overall, the heterogeneous results demonstrate that a substantial portion of the separability observed in the homogeneous analysis is influenced by domain alignment rather than purely semantic anomaly detection. HyperFree appears particularly sensitive to memory bank distribution shifts, while Copernicus-FM shows comparatively higher robustness in one cross-domain direction. These findings confirm that sensor-induced embedding shifts can dominate PatchCore distance-based scoring, emphasizing the importance of spectral compatibility and domain consistency in memory bank construction.

Image- and Patch-Level Analysis

As in the homogeneous analysis in Section 4.5.1, to assess the separation capabilities of the heterogeneous configurations, as well as the behavior of the anomaly score distributions, an image- and patch-level analysis is performed. For each configuration, the patch-level histogram, patch-level boxplot and image level ROC curve are inspected, grouping the results by cross-domain pair.

Copernicus Pretrain (bank) → SpectralEarth 7 Bands (normal)

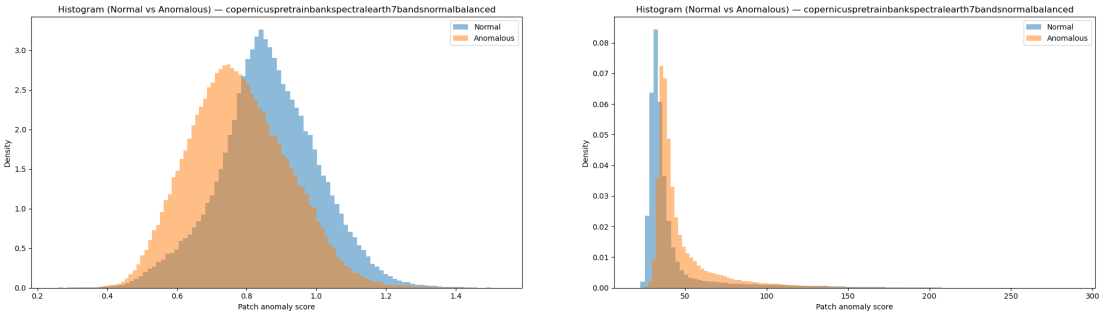
Table 4.9 summarizes the quantitative results.

Backbone	Memory Bank	Normal Test	Anom. Test	Thr.	Acc.	Prec.	Rec.	F1	AUROC
HyperFree	Copernicus Pretrain	SpectralEarth 7 Bands	Active Fire	0.3952	0.7897	0.7897	1.0000	0.8825	0.3183
Copernicus	Copernicus Pretrain	SpectralEarth 7 Bands	Active Fire	33.151	0.8610	0.8700	0.9687	0.9167	0.7604

Table 4.9: Heterogeneous analysis results - Copernicus Pretrain (bank) → SpectralEarth 7 Bands (normal)

Patch-Level Histogram. Figure 4.12 reflects the large gap between HyperFree and Copernicus-FM AUROC values (0.3183 and 0.7604). Specifically, looking at HyperFree distribution, the anomaly score distribution for the anomalous class is shifted toward lower values than the normal ones, meaning anomalous samples are represented closer to the memory bank in the embedding space, rather than the normal samples, inverting the expected PatchCore trend.

On the other end, Copernicus-FM, with this specific cross-domain pair, tends to keep anomaly score values for anomalous samples higher than normal values, still recording a performance drop compared to the homogeneous setting (0.7604 and 0.8236).

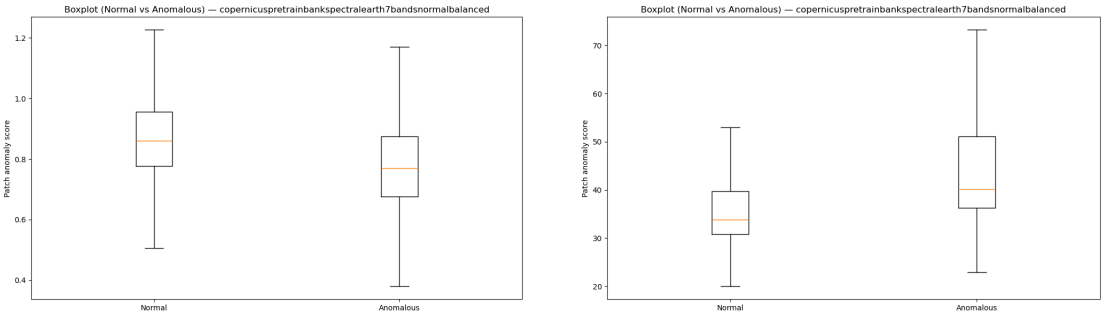


(a) HyperFree **(b)** Copernicus-FM

Figure 4.12: Patch-level anomaly score histogram (Heterogeneous - Copernicus Pretrain (bank) → SpectralEarth 7 Bands (normal)).

Patch-Level Boxplot. Boxplots in Figure 4.13 better highlight the cluster distributions, showing how, compared to the heterogeneous setting, the difference in terms of spread and standard deviation between normal and anomalous clusters is reduced. In particular, the normal cluster becomes wider for both HyperFree and Copernicus-FM, indicating increased intra-class variability.

Looking at Copernicus-FM, the anomalous cluster remains more spread out than the normal one, and, consistently with the histograms, anomalous patches tend to occupy the higher-score region.



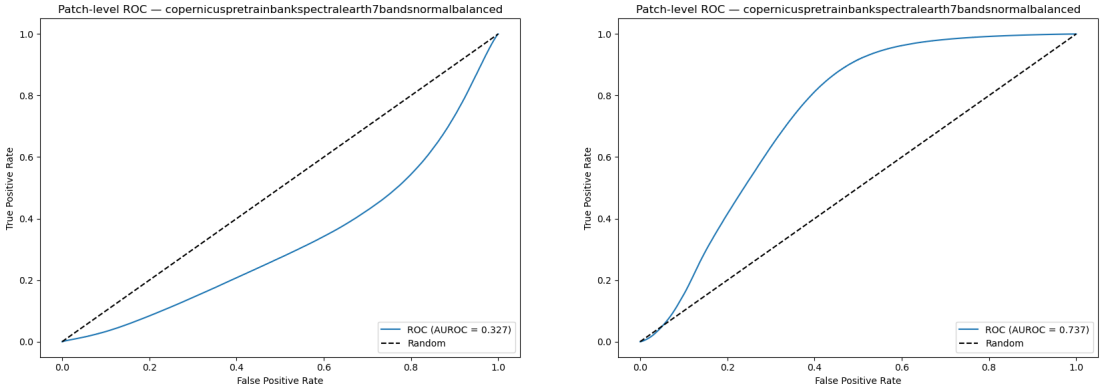
(a) HyperFree **(b)** Copernicus-FM

Figure 4.13: Patch-level anomaly score boxplot (Heterogeneous - Copernicus Pretrain (bank) → SpectralEarth 7 Bands (normal)).

Image-Level ROC Curve. Figure 4.14 confirms that HyperFree behaves in an inverted way: the ROC curve bends towards the bottom-right corner, in line with the low AUROC value (0.3183), indicating that high scores are predominantly assigned to cross-domain normal images.

Copernicus-FM instead preserves a positively oriented ROC curve, still resulting

in a curve less tending to the top-left corner of the plot, as it did in the homogeneous setting.



(a) HyperFree **(b)** Copernicus-FM

Figure 4.14: Image-level ROC curves (Heterogeneous - Copernicus Pretrain (bank) → SpectralEarth 7 Bands (normal)).

Overall, this cross-domain configuration shows that HyperFree completely loses the expected anomaly ordering when the normal test distribution shifts away from the memory bank, while Copernicus-FM retains a coherent score hierarchy and a usable ROC curve, albeit with degraded AUROC. This highlights a stronger sensitivity of HyperFree to memory-bank–test mismatch, whereas Copernicus-FM exhibits comparatively higher robustness when the bank is built on Copernicus Pretrain.

SpectralEarth 7 Bands (bank) → Copernicus Pretrain (normal)

Table 4.10 reports the quantitative results for this cross-domain pair.

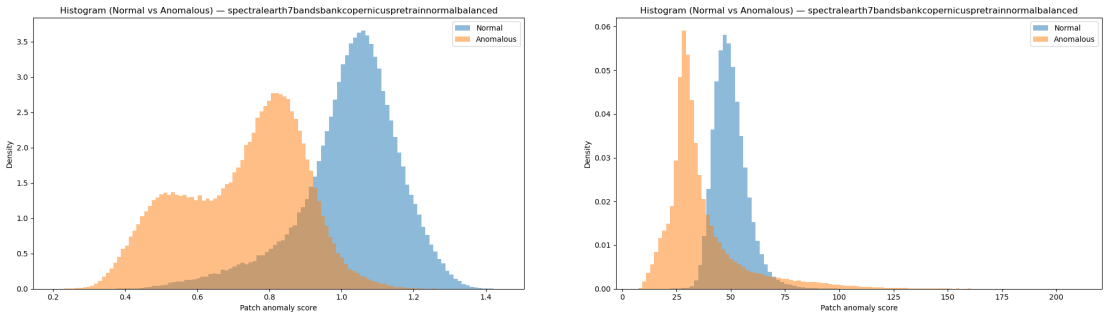
Backbone	Memory Bank	Normal Test	Anom. Test	Thr.	Acc.	Prec.	Rec.	F1	AUROC
HyperFree	SpectralEarth 7 Bands	Copernicus Pretrain	Active Fire	0.2706	0.7897	0.7897	1.0000	0.8825	0.0703
Copernicus	SpectralEarth 7 Bands	Copernicus Pretrain	Active Fire	7.9453	0.7897	0.7897	1.0000	0.8825	0.1733

Table 4.10: Heterogeneous analysis results.

Patch-Level Histogram. In this configuration, HyperFree exhibits a dramatic collapse in ranking quality, with an AUROC of 0.0703. The histogram shows that the anomaly score distribution of the normal Copernicus Pretrain samples is heavily shifted towards higher values compared to the anomalous Active Fire samples.

As in the previous cross-domain case, this indicates a complete inversion of the expected PatchCore behaviour: cross-domain normal samples are systematically assigned larger distances from the memory bank than the actual anomalies.

For Copernicus-FM, the degradation is also substantial (AUROC 0.1733), and the histogram reveals a severe overlap between normal and anomalous clusters. In this case, although the inversion effect is less visually extreme than for HyperFree, the separation between the two distributions becomes minimal, leading to near-random ranking performance.

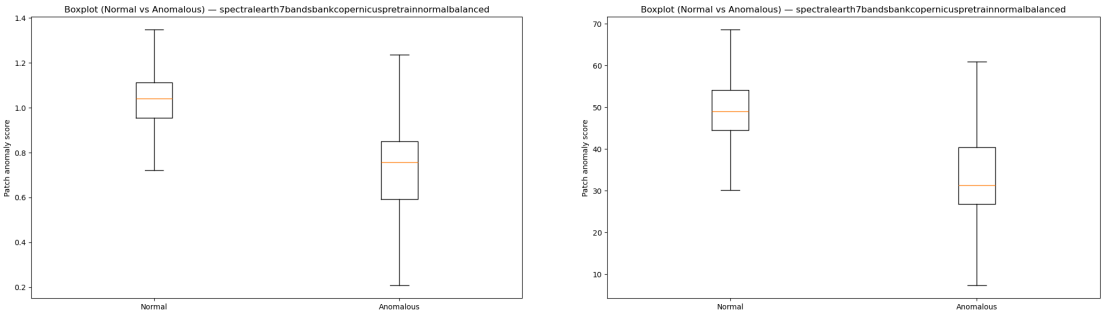


(a) HyperFree **(b)** Copernicus-FM

Figure 4.15: Patch-level anomaly score histogram (Heterogeneous - SpectralEarth 7 Bands (bank) → Copernicus Pretrain (normal)).

Patch-Level Boxplot. The boxplots further clarify the effect of the domain shift. For both backbones, the normal Copernicus Pretrain samples exhibit an elevated median anomaly scores relative to the anomalous class. This suggests that the feature representations extracted from Copernicus Pretrain are poorly aligned with the SpectralEarth-based memory bank.

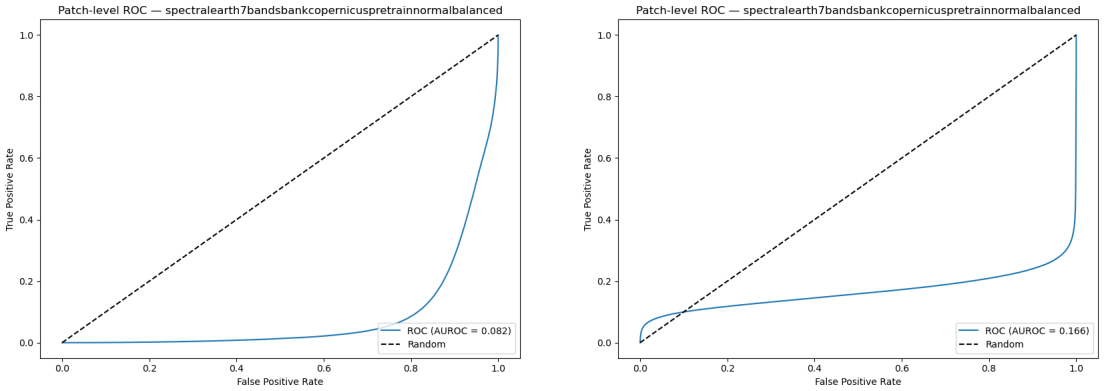
Instead, looking at the anomalous class, the median value for HyperFree is more shifted towards the normal distribution rather than Copernicus-FM.



(a) HyperFree **(b)** Copernicus-FM

Figure 4.16: Patch-level anomaly score boxplot (Heterogeneous - SpectralEarth 7 Bands (bank) → Copernicus Pretrain (normal)).

Image-Level ROC Curve. The ROC curves confirm the histogram observations. HyperFree's curve is almost entirely concentrated near the bottom-right region of the plot, consistent with the extremely low AUROC (0.0703), indicating systematic misranking. Copernicus-FM, while slightly positive at the beginning, also exhibits a curve close to the bottom-right corner, reflecting the collapse of discriminative capability in this configuration (AUROC 0.1733).



(a) HyperFree **(b)** Copernicus-FM

Figure 4.17: Image-level ROC curves (Heterogeneous - SpectralEarth 7 Bands (bank) → Copernicus Pretrain (normal)).

Overall, this cross-domain configuration demonstrates that when the memory bank is constructed from SpectralEarth 7 Bands and evaluated on Copernicus Pretrain normals, both backbones fail to maintain a consistent anomaly ordering. The severe degradation suggests that the embedding spaces learned under SpectralEarth-based statistics do not generalize to Copernicus Pretrain distribution,

highlighting the dominant role of domain alignment in PatchCore-based anomaly detection.

4.5.3 FLOGA Analysis

After analyzing the capabilities of the anomaly detection pipeline on wildfire detection, using Active Fire for the anomalous test set, the domain is shifted towards burned areas. Thanks to its structure, FLOGA can be exploited for both normal and anomalous samples, reducing the sensor-shift given by the difference of the acquisition sensor between different datasets.

Specifically, two settings have been tested:

- **Cross-Domain**, where Copernicus Pretrain is used for memory bank construction and FLOGA for testing.
- **In-Domain**, using just FLOGA for all three splits, to remove explicit sensor mismatch between memory bank and test data.

Details about experimental settings and data splits can be found in Table 4.11.

Memory Bank Set	Normal Test Set	Anomalous Test Set
Copernicus Pretrain (249504 patches)	FLOGA Normal (86229 images)	FLOGA Anomalous (409 images)
FLOGA Normal (137967 patches)	FLOGA Normal (86229 images)	FLOGA Anomalous (409 images)

Table 4.11: FLOGA experimental settings and data splits

The results are summarized in Table 4.12.

Backbone	Memory Bank	Normal Test	Anom. Test	Thr.	Acc.	Prec.	Rec.	F1	AUROC
HyperFree	Copernicus Pretrain	FLOGA Normal	FLOGA Anomalous	0.5434	0.1530	0.0025	0.8973	0.0050	0.3999
HyperFree	FLOGA Normal	FLOGA Normal	FLOGA Anomalous	0.5004	0.6503	0.0075	0.5575	0.0148	0.6126
Copernicus	Copernicus Pretrain	FLOGA Normal	FLOGA Anomalous	25.468	0.0062	0.0024	0.9951	0.0047	0.3769
Copernicus	FLOGA Normal	FLOGA Normal	FLOGA Anomalous	16.159	0.1286	0.0049	0.9095	0.0098	0.4588

Table 4.12: FLOGA analysis results.

Image-Level Comparative Discussion. At image level, the comparison between the Cross-Domain and In-Domain settings clearly highlights the impact of sensor consistency on the anomaly detection pipeline.

Starting from HyperFree, a substantial performance gap is observed when moving from the Cross-Domain configuration (AUROC = 0.3999) to the In-Domain one

(AUROC = 0.6126). When the memory bank is constructed using FLOGA normal samples, the embedding space becomes more coherent with the test distribution, resulting in a significantly better ranking of anomalous images.

A similar trend is observed for Copernicus-FM, although with overall lower performance. The Cross-Domain configuration reaches an AUROC of 0.3769, while the In-Domain setting improves to 0.4588. The gain, although smaller compared to HyperFree, still indicates that removing the sensor-shift effect leads to a more consistent anomaly score distribution.

Despite these improvements, both backbones exhibit extremely low precision values in all configurations. This behavior is explained by the strong class imbalance between normal and anomalous samples (409 anomalous images against 86229 normal ones), which makes the threshold selected through F1 optimization heavily biased toward high recall. Indeed, recall values remain consistently high (up to 0.9951), while precision remains close to zero, resulting in very low F1-scores.

Comparatively, HyperFree consistently outperforms Copernicus-FM in both Cross-Domain and In-Domain settings, suggesting a more stable embedding geometry for PatchCore-based anomaly ranking in the burned-area scenario.

Overall, the FLOGA analysis confirms that sensor alignment between memory bank and test data is a critical factor for image-level anomaly detection performance.

Image- and Patch-Level Analysis

To better understand the impact of sensor alignment in terms of anomaly score distributions, the image- and patch-level analysis further investigates the distributions through patch-level histograms, boxplots and ROC curves. The results are aggregated per setting, analyzing the cross-domain setup first, and then the in-domain one.

Cross-Domain Configuration

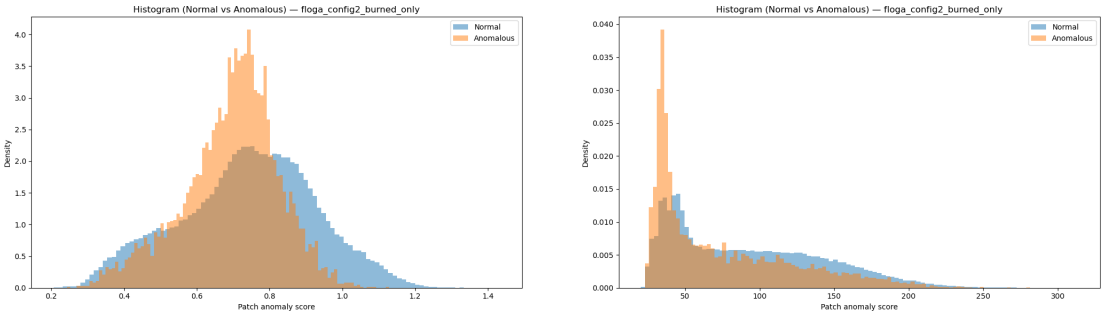
Table 4.13 summarizes the quantitative results.

Backbone	Memory Bank	Normal Test	Anom. Test	Thr.	Acc.	Prec.	Rec.	F1	AUROC
HyperFree	Copernicus Pretrain	FLOGA Normal	FLOGA Anomalous	0.5434	0.1530	0.0025	0.8973	0.0050	0.3999
Copernicus	Copernicus Pretrain	FLOGA Normal	FLOGA Anomalous	25.468	0.0062	0.0024	0.9951	0.0047	0.3769

Table 4.13: FLOGA analysis result - Cross-Domain Configuration

Patch-Level Histogram. Figure 4.18 highlights the patch-level anomaly score distributions for the normal and anomalous splits. The first observation that stands out, valid for both backbones, is the evident overlap between the

distributions, which matches the AUROC value (0.3999 and 0.3769) and confirms the poor separation capabilities of the embedding spaces in this configuration. Is also important to remember the high class imbalance that characterizes FLOGA, having only 409 anomalous samples.



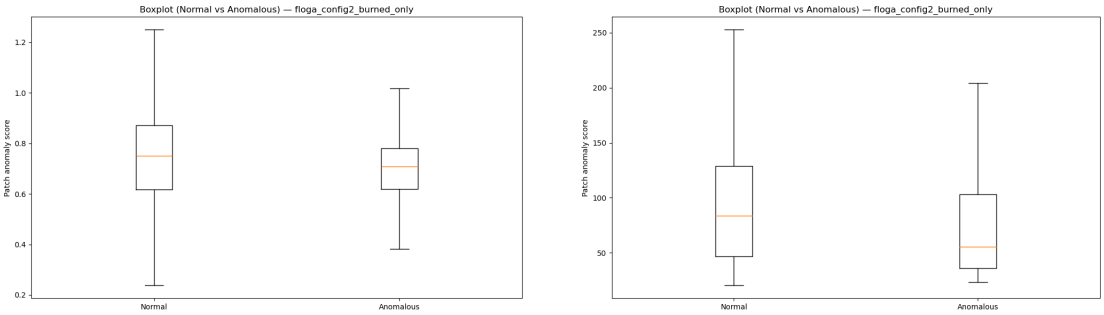
(a) HyperFree **(b)** Copernicus-FM

Figure 4.18: Patch-level anomaly score histogram (FLOGA - Cross-Domain Configuration)

Patch-Level Boxplot. Differently from histograms, Figure 4.19 allows a more precise analysis of the distributions. Even though it confirms the overlap, it allows an accurate comparison between normal and anomalous patch-level anomaly scores distributions.

HyperFree seems to have a much tighter representation for the anomalous samples, the distribution is less spread out than normal one, with a median value for the two splits that almost overlaps.

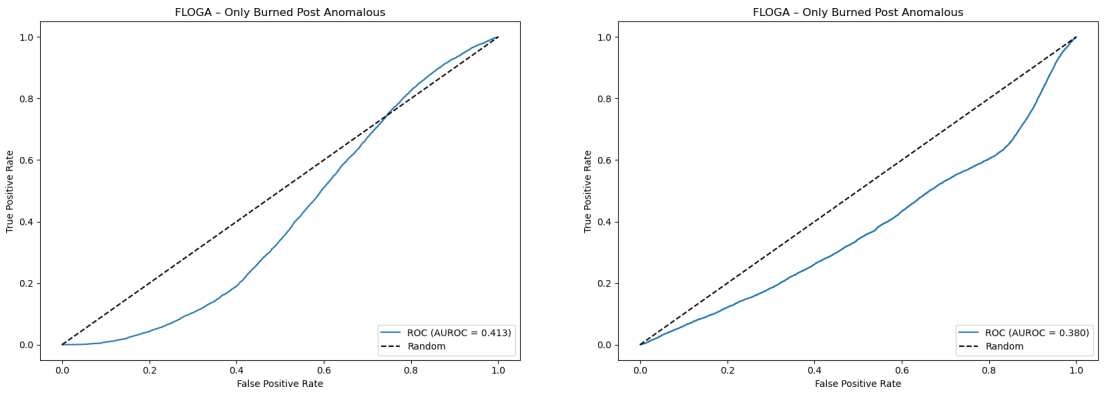
Copernicus-FM, on the other hand, provides wider distributions compared to the other backbone, although the difference between the two distributions is not particularly pronounced. The medians in this case are slightly different, with the anomalous cluster that seems to be represented closer to the memory bank in the embedding space.



(a) HyperFree **(b)** Copernicus-FM

Figure 4.19: Patch-level anomaly score boxplot (FLOGA - Cross-Domain Configuration)

Image-Level ROC Curve. ROC curves at Figure 4.20 reflects the expectations given by the AUROC values. Both curves tend to the diagonal, having HyperFree that slightly overtakes it towards the top-right part of the plot, while Copernicus-FM exhibits a concave trend.



(a) HyperFree **(b)** Copernicus-FM

Figure 4.20: Image-level ROC curves (FLOGA - Cross-Domain Configuration)

Overall, the cross-domain configuration confirms that the embedding learned on Copernicus Pretrain is not geometrically aligned with the burned-area distribution of FLOGA, leading to an almost inverted anomaly ranking.

In-Domain Configuration

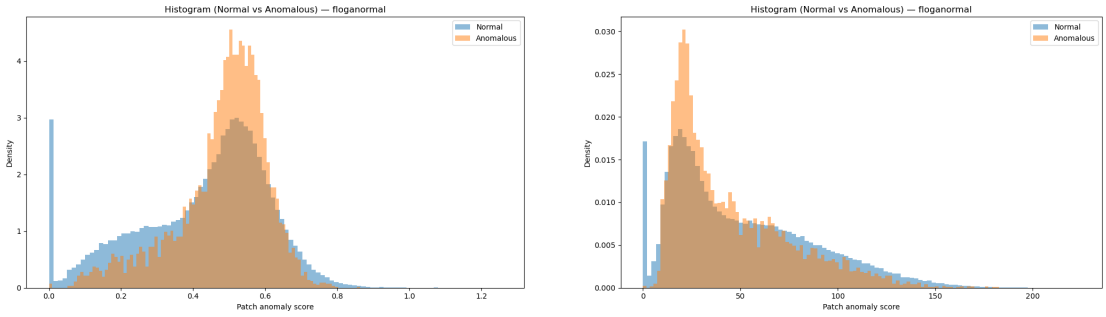
Table 4.14 summarizes the results.

Backbone	Memory Bank	Normal Test	Anom. Test	Thr.	Acc.	Prec.	Rec.	F1	AUROC
HyperFree	FLOGA Normal	FLOGA Normal	FLOGA Anomalous	0.5004	0.6503	0.0075	0.5575	0.0148	0.6126
Copernicus	FLOGA Normal	FLOGA Normal	FLOGA Anomalous	16.159	0.1286	0.0049	0.9095	0.0098	0.4588

Table 4.14: FLOGA analysis results - In-Domain Configuration

Patch-Level Histogram. Using FLOGA Normal for the memory bank construction, even if numerically increases the AUROC values (0.6126 and 0.4588), do not involve radical changes in the shape of patch-level anomaly score distributions.

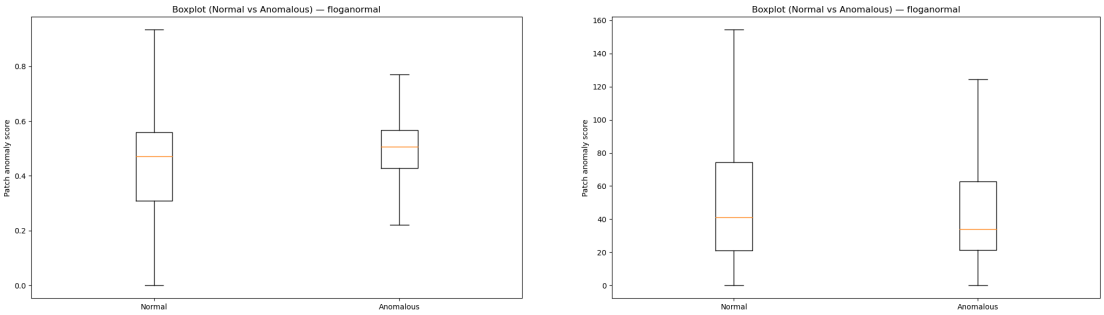
Both backbones still record a severe overlap of the distributions, but a common change is a singular peak close to anomaly score equal to zero, that depends on the different memory bank used, in this case by the fact that both memory bank and normal test set come from FLOGA Normal.



(a) HyperFree **(b)** Copernicus-FM

Figure 4.21: Patch-level anomaly score histogram (FLOGA - In-Domain Configuration)

Patch-Level Boxplot. While Copernicus-FM boxplot perseveres to show spread and overlap distributions, HyperFree keeps the anomalous split tighter, and this time, rather than the cross-domain configuration, its scores are slightly higher than the normal split. This explains the improvement in AUROC, from 0.3999 to 0.6126.

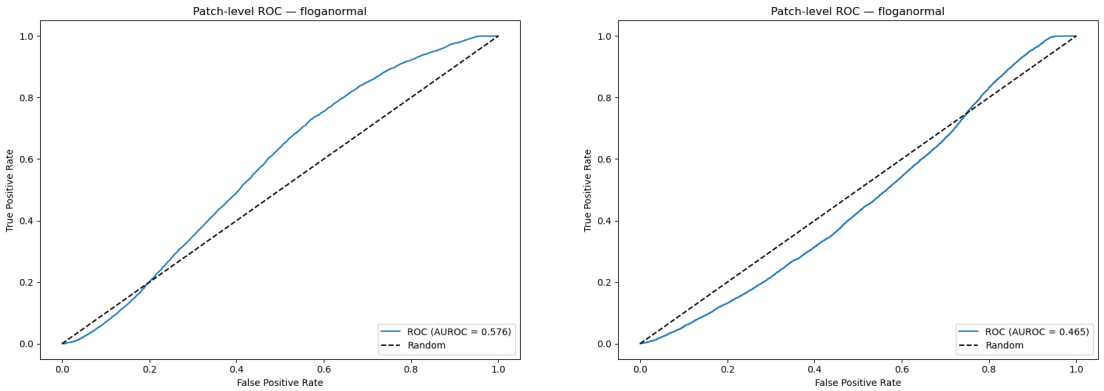


(a) HyperFree **(b)** Copernicus-FM

Figure 4.22: Patch-level anomaly score boxplot (FLOGA - In-Domain Configuration)

Image-Level ROC Curve. Figure 4.23 finally shows the HyperFree curve, although it is still close to the diagonal, slightly convex, suggesting the in-domain configuration improves the separation capability of the pipeline, deleting completely the sensor-shift problem.

Copernicus-FM registers on the other hand, a ROC curve quite aligned to the diagonal, converging to a random behavior.



(a) HyperFree **(b)** Copernicus-FM

Figure 4.23: Image-level ROC curves (FLOGA - In-Domain Configuration)

4.6 HyperFree Spectral Invariance Test: Hyperspectral vs Multispectral

A key question for the downstream anomaly detection pipeline is whether the HyperFree backbone produces consistent representations when the same physical

scene is observed under different spectral configurations. This question is also motivated by the claim of the HyperFree authors that their model can handle both modalities. In particular, the goal of this analysis is to understand if features extracted from a hyperspectral image remain comparable to features extracted from a multispectral version of the same image.

In order to answer this research question, two different analysis configurations were considered:

- **SpectralEarth vs SpectralEarth 7 Bands**, comparing the original hyperspectral version of the dataset with the synthesized, multispectral one.
- **PRISMA vs Sentinel-2**, comparing three groups of co-registered images acquired with sensor PRISMA (hyperspectral), Sentinel-2 (multispectral) and synthetized Sentinel-2 from PRISMA (multispectral), representing the same physical scene with short acquisition time difference.

The same analyses could not be performed on Copernicus-FM because the model cannot natively handle hyperspectral inputs.

4.6.1 SpectralEarth vs SpectralEarth 7 Bands

This configuration proposes the comparison between the hyperspectral SpectralEarth dataset, and its multispectral version, which was built to emulate the spectral response function (SRF) of Landsat-8. A detailed explanation of the synthetization process can be found in Section 4.4.2. The need to use an artificial version of the dataset comes from the lack of public datasets composed of paired and co-registered hyperspectral and multispectral images.

Experimental Setup. After the feature maps extraction with HyperFree, for each paired sample ($78k$ total pairs), the distances between modalities were computed (HS–MS_{SRF}). Specifically, the analysis relies on the cosine distance between the global pooled feature vectors:

$$d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|}$$

To interpret the magnitude of paired distances, a random baseline was computed by sampling mismatched pairs (different filenames) within the same modality (e.g., HS–HS) and computing the same distances. Intuitively, invariance would manifest as paired cross-modality distances that are much smaller than distances between unrelated scenes. This is measured with the ratio between pair mean cosine distance and random in-domain cosine distance.

Finally, linear CKA (Centered Kernel Alignment) is computed on a subset of 5000 samples as a similarity score between two representation spaces. CKA values close to 1 indicate very similar representations.

Quantitative Results. Table 4.15 summarizes the statistics.

Metric	**Mean**	**Median**	**p75**
$d_{\cos}(\mathrm{HS}, \mathrm{MS}_{\mathrm{SRF}})$	0.580454	0.578438	0.588825
$d_{\cos}(\mathrm{HS}, \mathrm{HS})$ random	0.077659	0.064645	0.103398
$d_{\cos}(\mathrm{MS}_{\mathrm{SRF}}, \mathrm{MS}_{\mathrm{SRF}})$ random	0.082358	0.068083	0.109576
Ratio $\frac{d_{\cos}(\mathrm{HS},\mathrm{MS}_{\mathrm{SRF}})}{\mathbb{E}[d_{\cos}(\mathrm{HS},\mathrm{HS})]}$	7.47439	7.44843	7.58218
Linear CKA (5,000 samples): HS–$\mathrm{MS}_{\mathrm{SRF}}$		0.583134	

Table 4.15: SpectralEarth vs SpectralEarth 7 Bands - HyperFree Invariance Test Results

Looking at the results, the paired cosine distance (HS–$\mathrm{MS}_{\mathrm{SRF}}$) is ≈ 0.58, while the random same-modality (HS–HS) baseline is much smaller (mean ≈ 0.078). The paired-to-random ratio is ≈ 7.47, indicating that the same scene seen in HS and MS appears more dissimilar than two random HS scenes in feature space. This is also confirmed by the CKA value, that shows a moderate similarity between the representations, far from a perfect match (0.58 vs 1).

These results are also supported by the histogram (Figure 4.24), which shows the cosine distance distributions of the cross-modality (HS–$\mathrm{MS}_{\mathrm{SRF}}$) and random same modality (HS–HS) pairs.

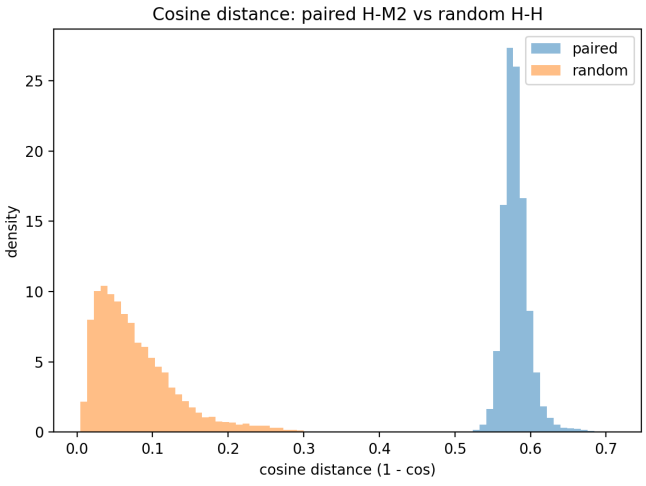


Figure 4.24: Cosine Distance Histogram - SpectralEarth vs SpectralEarth 7 Bands

Specifically, no overlap can be identified, with the random pairs distribution having significant lower values than the cross-modality pairs. Moreover, the random pairs appear to be more spread in the histogram, while cross-modality pairs have a higher peak over 0.6.

Conclusion. Overall, the experiment suggests that HyperFree features are not spectrally invariant between hyperspectral and multispectral inputs in this setting: the same scene mapped from HS to MS yields a much larger feature distance than typical distances between unrelated HS scenes.

This result motivates treating HS and MS as distinct domains at the feature level (e.g., via domain adaptation or modality-specific calibration) when mixing them in downstream tasks.

4.6.2 PRISMA vs Sentinel-2

In this configuration the analysis is performed over three groups of images, which are co-registered and representing the same physical scene with a close-to-identical acquisition time. In particular:

- **PRISMA**, hyperspectral images having 230 bands.
- **Sentinel-2**, multispectral images having 10 bands.
- **Sentinel-2 Synthetized**, also multispectral and with 10 bands, were obtained starting from the PRISMA images simulating the spectral response function (SRF) of Sentinel-2, with the process explained in Section 4.4.2.

The goal of this specific analysis, even if performed only on three groups of images, was to first remove the eventual synthetization bias introduced with the artificial generation of data, and then measure it, having a comparison between the real Sentinel-2 images and the artificial ones.

Experimental Setup. After projecting each of the three images for each group with HyperFree backbone, the analysis relies on the cosine distance between each pair, defining it as:

$$d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|}$$

Having single images instead of a bigger distribution, as it was for the previous analysis, makes impossible to use the same ratio used before, in order to perform a proper comparison between the analysis.

Quantitative Results. Table 4.16 reports the cosine distance values for the three acquisitions, considering the following pairs:

- PRISMA vs Sentinel-2 (real)
- PRISMA vs Sentinel-2 (synthetized)
- Sentinel-2 (real) vs Sentinel-2 (synthetized)

Acquisition	PRISMA–S2 (real)	PRISMA–S2 (synth)	S2 real–S2 synth
02-06-2025	0.1014	0.1398	0.0257
16-06-2025	0.1486	0.1455	0.0274
29-08-2025	0.1547	0.1650	0.0176

Table 4.16: Cosine distance between embedding vectors produced by HyperFree for co-registered PRISMA and Sentinel-2 images.

First, it can be observed that the cosine distance between the real Sentinel-2 images and their synthetized counterpart is consistently very low, ranging between 0.0176 and 0.0274. This indicates that, in the embedding space induced by HyperFree, the artificial SRF-based conversion from PRISMA to Sentinel-2 produces representations that are extremely close to those obtained from the real Sentinel-2 images.

Second, when comparing PRISMA directly with real Sentinel-2 images, the cosine distance ranges between 0.1014 and 0.1547. When comparing PRISMA with synthetized Sentinel-2, the values range between 0.1398 and 0.1650. Across the three acquisitions, the difference between these two quantities is relatively small and does not follow a consistent trend: in one case the synthetized image is slightly further from PRISMA (02-06-2025 and 29-08-2025), while in another case it is marginally closer (16-06-2025).

The variation between PRISMA–S2 (real) and PRISMA–S2 (synth) remains within approximately 0.01–0.04 in cosine distance, which is modest compared to the overall PRISMA–Sentinel gap (around 0.10–0.16). This suggests that the dominant source of embedding discrepancy is the intrinsic difference between hyperspectral (230 bands) and multispectral (10 bands) representations, rather than the SRF-based synthetization procedure itself.

A closer look at the differences between the pairs can be taken looking at Figure 4.25, which shows the difference in cosine similarity ($1 - d_{cos}$).

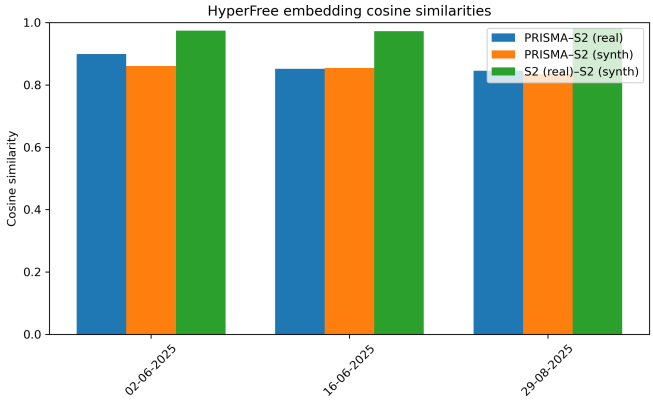


Figure 4.25: Cosine Similarities - PRISMA vs Sentinel-2

Conclusion. The extremely low cosine distance between Sentinel-2 (real) and Sentinel-2 (synthetized) embeddings confirms that the SRF simulation process preserves the spectral characteristics that are most relevant for the HyperFree backbone. In other words, from the perspective of the learned representation, the synthetized multispectral image behaves almost identically to a real Sentinel-2 observation.

Moreover, the similarity between PRISMA–S2 (real) and PRISMA–S2 (synth) distances indicates that the artificial conversion does not introduce a significant systematic bias in the embedding space. If a strong synthetization bias were present, one would expect a consistently larger distance between PRISMA and S2 (synth) compared to PRISMA and S2 (real), which is not observed.

4.6.3 Discussion

The two analyses provide complementary insights into the spectral invariance properties of HyperFree.

On the large-scale SpectralEarth experiment (78k paired samples), the cosine distance between hyperspectral and multispectral versions of the same scene (≈ 0.58) is substantially larger than the random same-modality baseline (≈ 0.08), with a paired-to-random ratio of ≈ 7.47 and a moderate CKA similarity (0.58). This clearly indicates that HyperFree does not produce spectrally invariant representations when switching from hyperspectral to multispectral inputs. In feature space, HS and MS behave as distinct domains.

The PRISMA vs Sentinel-2 analysis, although limited to three co-registered acquisitions, further clarifies the origin of this discrepancy. The cosine distance between real Sentinel-2 and SRF-synthesized Sentinel-2 images remains extremely low (0.0176–0.0274), confirming that the synthetization process itself does not introduce

significant distortion in the embedding space. In contrast, the PRISMA–Sentinel distance remains consistently higher (0.10–0.16), regardless of whether the multispectral image is real or synthetized.

Taken together, these results suggest that the lack of invariance is primarily driven by the intrinsic spectral reduction from hyperspectral (230 bands) to multispectral (10 bands), rather than by artifacts introduced by SRF simulation. Consequently, when mixing hyperspectral and multispectral data within the same downstream pipeline, HyperFree representations should be treated as modality-dependent, and additional alignment strategies may be required to ensure cross-sensor consistency.

4.7 Fine-Tuning Motivation and Experiments

4.7.1 Fine-Tuning Motivation

Starting from the takeaways of the baseline experiments (Section 4.5) and the spectral invariance test (Section 4.6), three main limitations emerged:

- **Sensor-Shift**, introduced when jointly analyzing datasets acquired by different satellite sensors, leading to different spectral signatures.
- **Modality-Shift**, arising from the combined processing of multispectral and hyperspectral data.
- **Embedding Space Representation for Anomaly Detection**, referring to how samples are positioned in the embedding space, having overlapping normal and anomalous clusters.

These results indicate that, although pretrained foundation models provide strong general-purpose representations, they are not fully aligned with the specific requirements of the anomaly detection task.

To address these issues, the idea is to perform a semi-supervised fine-tuning on both backbones, with the objective of reshaping the embedding space to better fit the anomaly detection task. In order to delete sensor and modality shift, the fine-tuning is performed exclusively on FLOGA, which is based on Sentinel-2 imagery. Specifically, reshaping the embedding space means acting differently on normal and anomalous samples, after computing the centroid of normal training set:

- **Normal Samples**, are pulled towards the centroid, to obtain a more compact normal cluster.
- **Anomalous Samples**, are pushed away from the centroid, enforcing a minimum separation margin.

This is done introducing also a distillation term to keep the spatial distribution semantically similar to the baseline models, preventing excessive drift during adaptation. The complete loss formulation is reported in Section 3.4.

4.7.2 Fine-Tuning Setup

Fine-tuning is performed using the FLOGA splits defined in Table 4.1, and in particular:

- **Training Set**, used only to update the model weights.
- **Validation Set**, used to select the best checkpoint, based on PatchCore AUROC.
- **Test Set**, used to compute the final metrics.

Both HyperFree and Copernicus-FM are initialized with their pretrained weights. During fine-tuning, all backbone parameters are kept frozen, except for the last four ViT blocks out of twelve.

On top of the backbones, a PatchCore classification head is added. Its role is, starting from the validation set, to build the memory bank with half of the normal split, and then compute anomaly scores for the other half of the normal set and the anomalous set. After anomaly score computation, AUROC is measured and used to choose the best checkpoint. Optimization details are listed in Table 4.17. Note the difference in epochs between grid search and final tuning: to reduce computational cost, the grid search is performed with half the number of epochs used in the final runs.

Optimizer	AdamW
Learning rate	$1 \cdot 10^{-4}$
Weight decay	$1 \cdot 10^{-4}$
Batch size	4 (with automatic halving on OOM)
Epochs (sweep / final)	5 / 10
Precision	Mixed FP16/FP32 (AMP)

Table 4.17: Fine-Tuning Optimization Hyperparameters

Moreover, to find the optimal loss-weight hyperparameters (reported in Section 3.4.4), two separate grid searches, one for each backbone, are conducted. The search space explored is defined in Table 4.18.

Hyperparameter	Symbol	Values
Center loss weight	λ_{center}	$\{0.1, 0.3, 0.5\}$
Distillation weight	λ_{distill}	$\{0.3, 0.5\}$
Anomaly margin weight	λ_{anom}	$\{1.0, 3.0\}$
Anomaly margin	m_{anom}	1.3 (HyperFree), 100.0 (Copernicus-FM)

Table 4.18: Loss-weight hyperparameter grid

The search space is built in order to explore a range of configurations that varies from a more conservative setup, in terms of center-loss and distillation, to a more aggressive one. The anomaly margin is fixed per backbone, in particular it is chosen by computing the mean distance of the training normal samples and training anomalous samples to the centroid, and selecting a value that provides a smooth compromise between the two. The resulting margins differ significantly (1.3 vs. 100.0) because the distance magnitudes in the HyperFree and Copernicus-FM embedding spaces are on very different scales. In all cases, PatchCore AUROC on the validation set is used as the model-selection criterion across the grid.

After the grid search, the best setup (shared across both backbones) is

$$\lambda_{\text{center}} = 0.5, \quad \lambda_{\text{distill}} = 0.3, \quad \lambda_{\text{anom}} = 3.0.$$

4.7.3 Fine-Tuning Results

The proposed analysis relies on four different configurations per backbone, starting from the baseline results on the FLOGA test set, there are two intermediate fine-tuning configurations and the final best configuration using the grid search results. Table 4.19 summarizes the explored configurations.

Configuration	λ_{center}	λ_{distill}	λ_{anom}	m_{anom}
Baseline	None	None	None	None
Intermediate-1	0.3	1.0	1.0	1.3 (HF), 100.0 (C-FM)
Intermediate-2	0.3	0.5	3.0	1.3 (HF), 100.0 (C-FM)
Best	0.5	0.3	3.0	1.3 (HF), 100.0 (C-FM)

Table 4.19: Fine-Tuning configurations

The results are summarized in Table 4.20.

Backbone	Configuration	Thr.	Acc.	Prec.	Rec.	F1	AUROC
HyperFree	Baseline	0.5263	0.6025	0.0063	0.5385	0.0125	0.5592
HyperFree	Intermediate-1	0.4057	0.9910	0.0526	0.0538	0.0532	0.6215
HyperFree	Intermediate-2	0.4170	0.9954	0.5319	0.1923	0.2825	0.7068
HyperFree	Best	0.2064	0.9953	0.5333	0.0615	0.1103	**0.7514**
Copernicus	Baseline	12.475	0.0380	0.0047	0.9769	0.0094	0.4423
Copernicus	Intermediate-1	24.760	0.3911	0.0057	0.7385	0.0112	0.5235
Copernicus	Intermediate-2	61.588	0.8957	0.0089	0.1923	0.0170	0.6188
Copernicus	Best	56.358	0.9938	0.1587	0.0769	0.1036	**0.7007**

Table 4.20: Fine-Tuning results

Image-Level Comparative Discussion. At image level, fine-tuning consistently improves anomaly detection performance for both backbones. For HyperFree, the AUROC increases from 0.5592 in the frozen baseline to 0.7514 in the grid-selected configuration, corresponding to an absolute improvement of approximately +0.19. A similar trend is observed for Copernicus-FM, where the AUROC rises from 0.4423 to 0.7007, yielding an even larger absolute gain of approximately +0.26.

The intermediate configurations highlight a progressive improvement pattern. For both backbones, moving from the baseline to Intermediate-1 results in a moderate AUROC increase, while Intermediate-2 produces a more substantial gain. The final grid-selected configuration achieves the highest AUROC in both cases. This behavior supports the hypothesis that reshaping the embedding space through the center, distillation, and margin components gradually enhances class separability.

From a precision–recall perspective, fine-tuning significantly alters the decision dynamics. In the frozen setting, both backbones exhibit extremely low precision with relatively high recall, indicating a strong tendency to over-detect anomalies. After fine-tuning, precision increases substantially, while recall decreases. This shift reflects a more selective anomaly detection behavior, reducing false positives and improving overall ranking quality, as confirmed by the AUROC improvements.

Importantly, the objective of this analysis is not to identify a universally optimal configuration, but to demonstrate that backbone fine-tuning on a specific sensor and data modality can effectively adapt the embedding space to the target domain.

Image- and Patch-Level Analysis

This section deeply analyzes the patch and image anomaly scores distributions, with the goal of understand if and how the embedding space is reshaped with the fine-tuning. This is done analyzing the patch-level histograms and boxplots, and

the image-level ROC curve. Initially there will be a specific analysis on HyperFree and Copernicus-FM, followed by a comparative analysis.

HyperFree Table 4.21 summarizes the quantitative results.

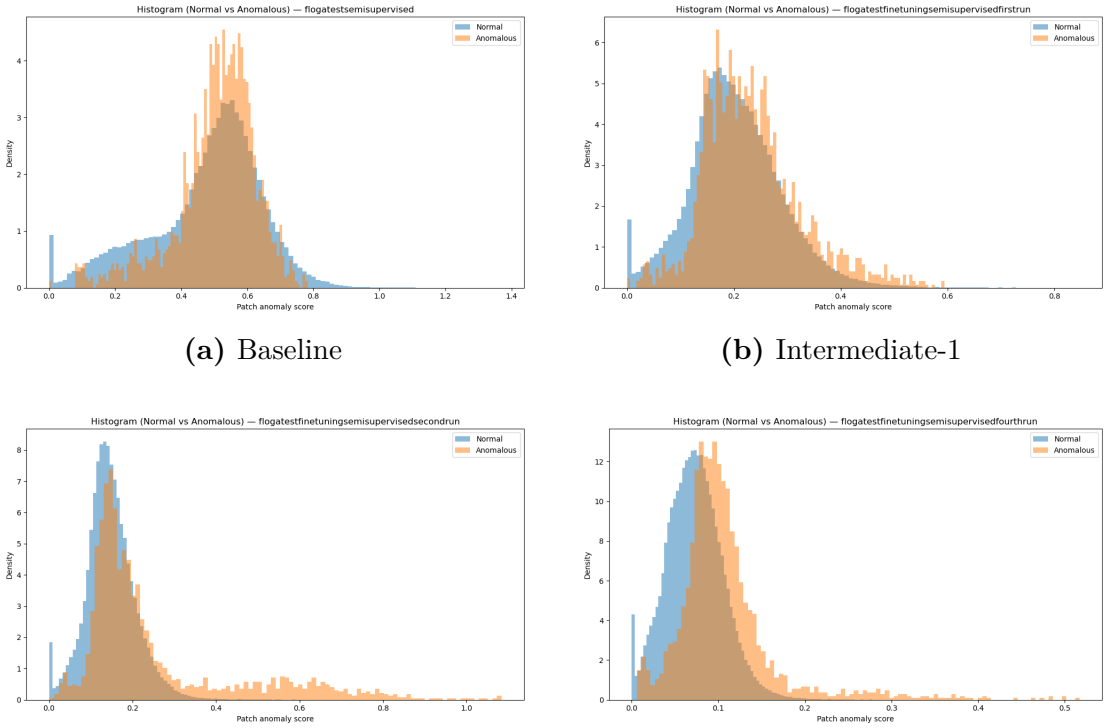
Backbone	Configuration	Thr.	Acc.	Prec.	Rec.	F1	AUROC
HyperFree	Baseline	0.5263	0.6025	0.0063	0.5385	0.0125	0.5592
HyperFree	Intermediate-1	0.4057	0.9910	0.0526	0.0538	0.0532	0.6215
HyperFree	Intermediate-2	0.4170	0.9954	0.5319	0.1923	0.2825	0.7068
HyperFree	Best	0.2064	0.9953	0.5333	0.0615	0.1103	**0.7514**

Table 4.21: HyperFree Fine-Tuning results

Patch-Level Histogram. Figure 4.26 illustrates the evolution of anomaly score distributions for normal and anomalous sets across fine-tuning configurations. Starting from the baseline configuration, the clusters are highly overlapped, with a common peak at about 0.6, indicating limited separability and explaining the relatively low AUROC of 0.5592.

The first intermediate configuration already shows an improvement compared to the baseline. The normal cluster shifts towards lower scores, and, even if there is still overlap with the anomalous cluster, the peaks of the distributions are not coincident anymore. Intermediate-2 presents a tighter normal cluster, while the anomalous one is spread to the right. This improved separation is reflected by the AUROC rising to 0.7068.

The grid-selected best configuration is the one that vividly changes compared to the baseline. While perfect separation is not achieved, the anomalous cluster is clearly shifted to higher scores compared to the normal one, leading to an AUROC value of 0.7514, an increase of +0.1922 compared to the baseline.



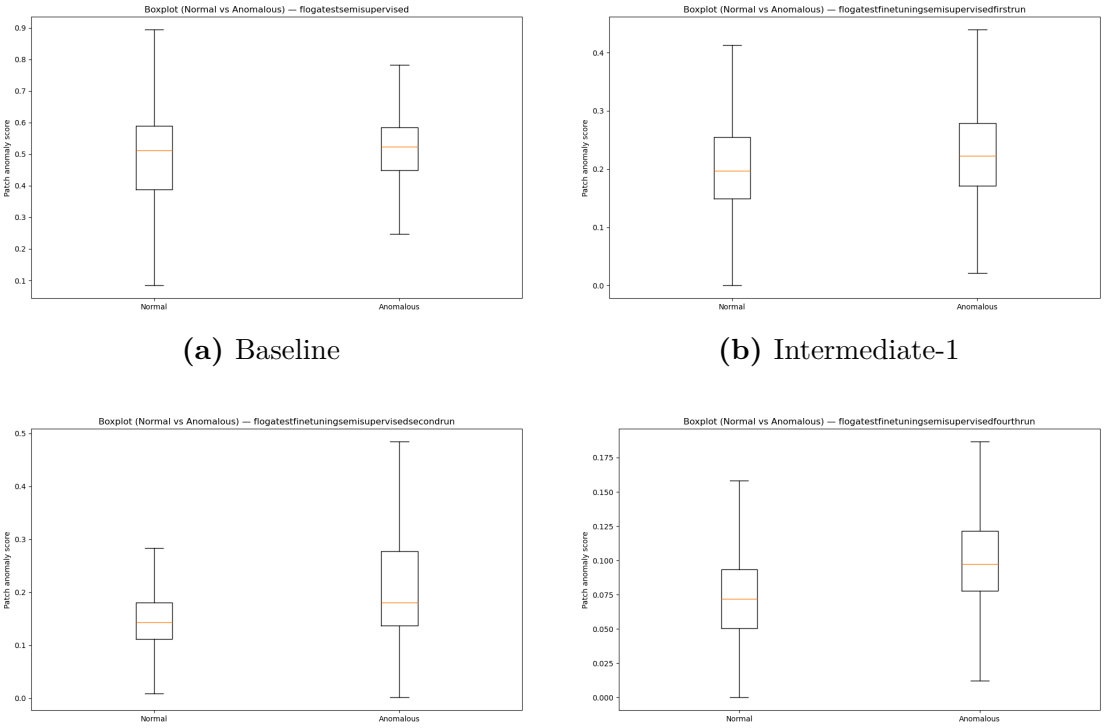
(a) Baseline **(b)** Intermediate-1

(c) Intermediate-2 **(d)** Best (Grid-selected)

Figure 4.26: Patch-level anomaly score histograms for HyperFree across fine-tuning configurations.

Patch-Level Boxplot. The boxplots in Figure 4.27 provide a complementary view of the distributional shift. Looking at the four plots together, it is clear that the anomalous distribution moves towards higher values while the normal distribution shifts close to zero.

Particularly, looking at the baseline and the best configuration, the normal cluster reduces its standard deviation, archiving a tighter cluster. This structural reshaping of the score distributions aligns with the observed AUROC improvements and supports the hypothesis that fine-tuning enhances embedding separability.



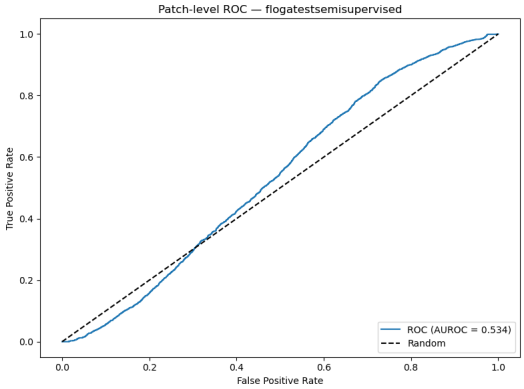
(a) Baseline **(b)** Intermediate-1

(c) Intermediate-2 **(d)** Best (Grid-selected)

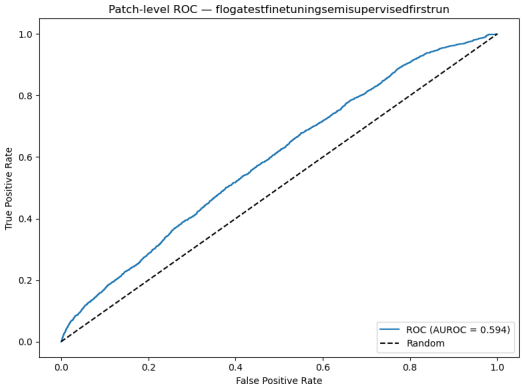
Figure 4.27: Patch-level anomaly score boxplots for HyperFree across fine-tuning configurations.

Image-Level ROC Curve. Figure 4.28 best represents the separability increase reached with fine-tuning. Baseline configuration curve lie close to diagonal, reflecting almost a random behavior. As fine-tuning progresses, the ROC curves bend toward the top-left corner, reflecting an improved separability archived. The grid-selected configuration achieves the most convex curve, consistent with the highest AUROC value of 0.7514.

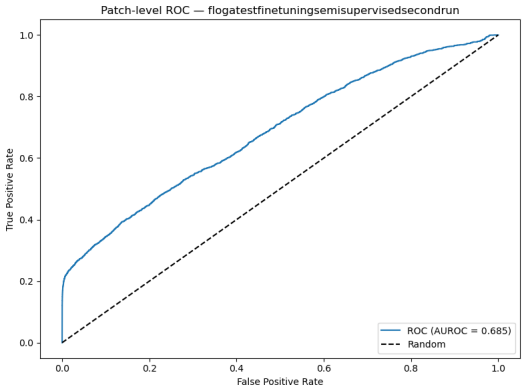
This evolution confirms that fine-tuning does not merely shift score distributions, but effectively enhances the ranking capability of the anomaly detection pipeline.



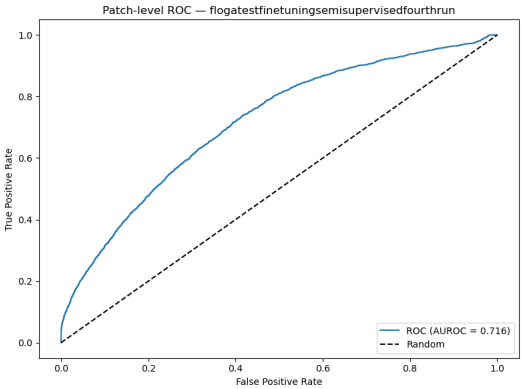
(a) Baseline



(b) Intermediate-1



(c) Intermediate-2



(d) Best (Grid-selected)

Figure 4.28: Image-level ROC curves for HyperFree across fine-tuning configurations.

Copernicus-FM

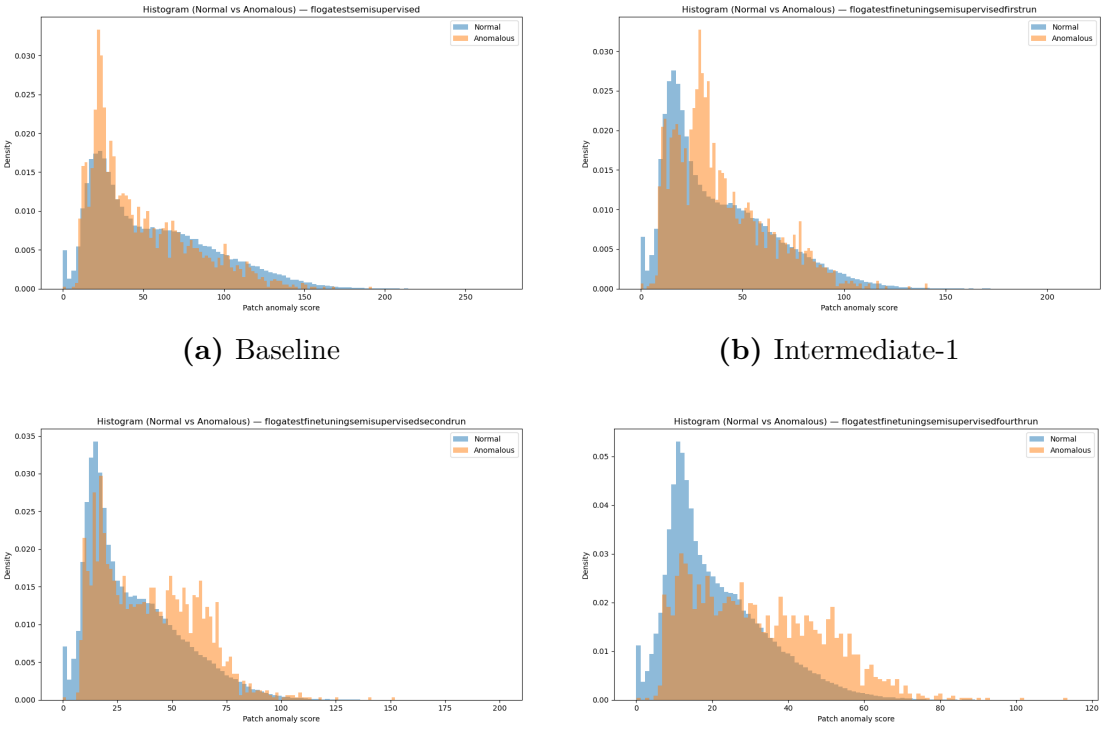
Table 4.22 summarizes the quantitative results.

Backbone	Configuration	Thr.	Acc.	Prec.	Rec.	F1	AUROC
Copernicus	Baseline	12.475	0.0380	0.0047	0.9769	0.0094	0.4423
Copernicus	Intermediate-1	24.760	0.3911	0.0057	0.7385	0.0112	0.5235
Copernicus	Intermediate-2	61.588	0.8957	0.0089	0.1923	0.0170	0.6188
Copernicus	Best	56.358	0.9938	0.1587	0.0769	0.1036	**0.7007**

Table 4.22: Copernicus-FM Fine-Tuning results

Patch-Level Histogram. Figure 4.29 presents the distribution of normal

and anomalous samples in terms of patch-level anomaly scores. Looking toward fine-tuning histograms, comparing them with baseline distributions, it can be noticed how the anomalous cluster slowly loses a defined peak. If the baseline presents a strong anomalous peak in correspondence to the normal one, the best configuration shows how, even if the normal cluster keeps the peak close to lower anomaly score values, the anomalous distribution is spread more homogeneously, losing the sharp peak.



(a) Baseline

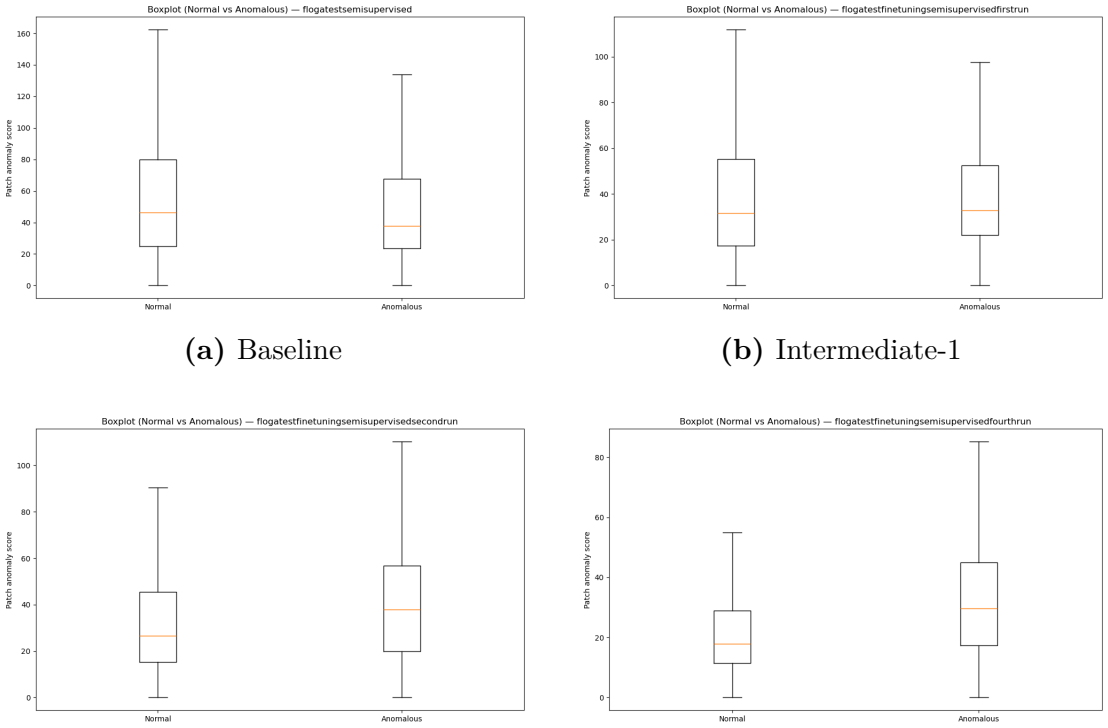
(b) Intermediate-1

(c) Intermediate-2

(d) Best (Grid-selected)

Figure 4.29: Patch-level anomaly score histograms for Copernicus-FM across fine-tuning configurations.

Patch-Level Boxplot. Figure 4.30 further highlights the distributions shift. In the baseline configuration, distributions are highly overlapped, the normal cluster is more spread than the anomalous one. As fine-tuning progresses, the anomalous distribution is vividly shifted towards higher anomaly score values than the normal distribution. Specifically, it is noticeable that the normal configuration shrinks as we move towards the best fine-tuning configuration, reflecting the objective of the fine-tuning, which goal for the normal cluster is to compact the distribution around the centroid.

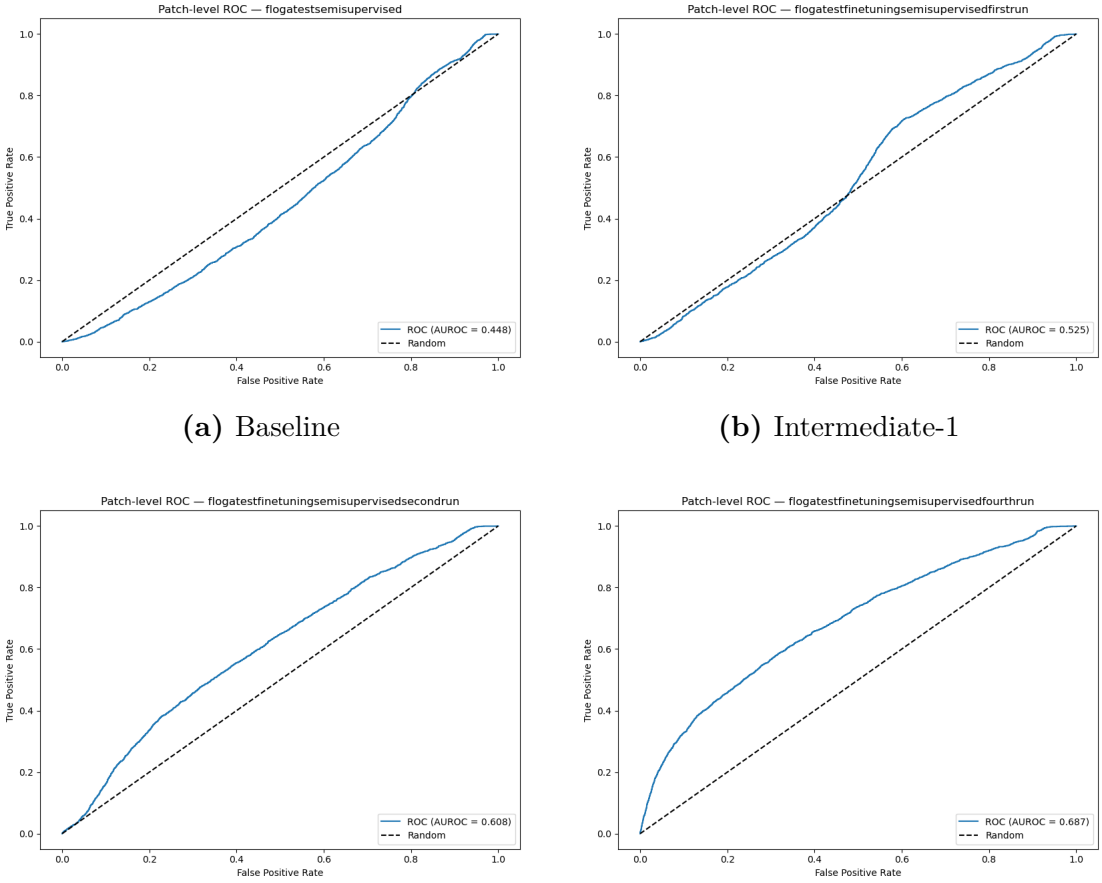


(a) Baseline **(b)** Intermediate-1

(c) Intermediate-2 **(d)** Best (Grid-selected)

Figure 4.30: Patch-level anomaly score boxplots for Copernicus-FM across fine-tuning configurations.

Image-Level ROC Curve. Lastly, Figure 4.31 reflects the separability increase registered with the fine-tuning. The baseline configuration presents a slightly concave curve, which lies close to the diagonal, reflecting the AUROC value (0.4423). If the first improvements in term of curve bands are noticeable in the intermediate configurations, the best configuration shows a clear improvement than the baseline. The ROC curve appears convex and banded toward the top-left corner. This means that there is a substantial separability improvement, measured with the AUROC value, which increases to 0.7007, leading to a final enhancement compared to the baseline of +0.2584.



(a) Baseline **(b)** Intermediate-1

(c) Intermediate-2 **(d)** Best (Grid-selected)

Figure 4.31: Image-level ROC curves for Copernicus-FM across fine-tuning configurations.

Comparative Analysis

Baseline behavior. At the frozen-backbone level, HyperFree and Copernicus-FM start from markedly different regimes. HyperFree achieves an AUROC of 0.5592 on FLOGA (Table 4.21), slightly above random and consistent with the partially overlapped but already asymmetric score distributions observed in Figures 4.26 and 4.27. In contrast, Copernicus-FM exhibits a baseline AUROC of 0.4423 (Table 4.22), with a ROC curve close to the diagonal (Figure 4.31) and very low precision despite almost perfect recall. This indicates a severe imbalance in the ranking of normal and anomalous samples, and suggests a stronger domain- and sensor-shift for Copernicus-FM in the frozen setting.

Fine-tuning gains. Under the proposed semi-supervised objective, both backbones benefit from fine-tuning, but with different magnitudes. HyperFree improves

from an AUROC of 0.5592 to 0.7514 in the best configuration, corresponding to an absolute gain of +0.1922. Copernicus-FM, starting from a weaker baseline, increases from 0.4423 to 0.7007, with a larger gain of +0.2584. Patch-level histograms and boxplots (Figures 4.26, 4.27, 4.29, 4.30) consistently show normal scores being compressed toward lower values and anomalous scores being pushed toward higher values for both models, while image-level ROC curves become progressively more convex and closer to the top-left corner. Overall, Copernicus-FM is more sensitive to fine-tuning (larger relative improvement), whereas HyperFree retains the best absolute performance after adaptation.

Embedding reshaping across architectures. Qualitatively, HyperFree transitions from partially overlapping clusters to a configuration where normal samples form a compact low-score mode and anomalies occupy a broader, higher-score region. Copernicus-FM undergoes a more radical transformation: score distributions that are almost indistinguishable at baseline become clearly separated after fine-tuning, with normal samples sharply concentrated near zero and anomalous samples spread over higher anomaly scores. These trends are coherent with the design of the loss in Section 3.4, which explicitly enforces compactness of the normal cluster and a margin-based repulsion of anomalies from the normal centroid. The fact that these effects are observed for two structurally different backbones supports the generality of the proposed fine-tuning strategy for adapting geospatial foundation models to anomaly detection on satellite data.

4.8 Computational and Structural Analysis

Beyond predictive performance, a practical requirement for the proposed anomaly detection pipeline is computational efficiency during both finetuning and inference. In preliminary experiments, Copernicus-FM was observed to be significantly faster and less memory-intensive than HyperFree. In order to understand the origin of this difference, a detailed structural and runtime analysis of both backbones was conducted on the FLOGA test split.

Architectural Overview. Both HyperFree and Copernicus-FM are based on a 12-layer Vision Transformer (ViT) backbone. Each transformer block contains approximately 7.1M parameters, resulting in roughly 85M parameters devoted to the transformer stack in both models.

However, the two models differ substantially in their input resolution and auxiliary modules.

- **HyperFree** operates at an image resolution of 1024×1024 with patch size 16, resulting in $64 \times 64 = 4096$ tokens per image.

- **Copernicus-FM** operates at 224×224 with patch size 16, resulting in $14 \times 14 = 196$ tokens per image.

Since self-attention scales quadratically with the number of tokens, the attention mechanism in HyperFree processes:

$$4096^2 \approx 1.68 \times 10^7$$

token interactions per layer, whereas Copernicus-FM processes:

$$196^2 \approx 3.84 \times 10^4$$

interactions per layer. This corresponds to more than two orders of magnitude difference in attention complexity per transformer block.

Parameter Distribution. Although the total parameter counts are comparable:

- HyperFree: 155.9M parameters
- Copernicus-FM: 139.5M parameters

their internal distribution differs significantly.
HyperFree includes several additional high-capacity components:

- Block spectral weight banks: 41.5M parameters
- Point spectral weight banks: 16.7M parameters
- Multi-scale convolutional modules: 9.5M parameters
- Additional contrastive modules and neck layers

In contrast, Copernicus-FM relies primarily on:

- A ViT backbone (12 blocks, $\sim$ 85M parameters)
- Lightweight dynamic MLP-based spectral patch embedding
- Fourier-based coordinate, scale, and time expansions

Importantly, Copernicus-FM's additional modules operate at the patch level rather than at full spatial resolution, which substantially reduces computational overhead.

Runtime and Memory Analysis. Both models were evaluated on the same FLOGA test configuration using identical hardware and dataloaders. Measurements were taken over 30 batches after a warm-up phase.

Metric	HyperFree	Copernicus-FM
Image resolution	1024×1024	224×224
Tokens per image	4096	196
Total parameters	155.9M	139.5M
Avg time per image	141.1 ms	8.3 ms
Peak GPU memory	5647 MB	568 MB

Table 4.23: Structural and runtime comparison between HyperFree and Copernicus-FM on FLOGA.

Despite having a similar number of transformer blocks and comparable total parameter counts, HyperFree is approximately:

- 17× slower per image during inference,
- 10× more memory-demanding in terms of peak allocated GPU memory.

Discussion. The observed computational gap is primarily explained by two factors:

1. **Token count explosion:** HyperFree processes $\sim 21\times$ more tokens per image. Given the quadratic complexity of self-attention, this leads to dramatically higher computational cost per layer.
2. **Additional spectral and multi-scale modules:** HyperFree includes large spectral weight banks and multi-scale convolutional structures that operate at high spatial resolution, further increasing both memory usage and runtime.

Copernicus-FM, while architecturally comparable at the transformer level, operates at lower spatial resolution and uses lighter patch-level conditioning mechanisms, resulting in significantly better computational efficiency.

Implications for the Anomaly Detection Pipeline. From a deployment and experimentation standpoint, Copernicus-FM offers substantially faster finetuning and inference, enabling larger-scale experimentation under constrained hardware resources. HyperFree, although more computationally demanding, provides stronger hyperspectral modeling capabilities and higher spatial fidelity, which may justify its cost in specific remote sensing scenarios.

These results highlight a fundamental trade-off between spectral flexibility, spatial resolution, and computational efficiency in geospatial foundation models.

4.9 Experimental Conclusions

This chapter investigated the effectiveness and adaptability of geospatial foundation models for anomaly detection under domain and modality shifts.

Frozen-backbone analysis. In homogeneous settings, both HyperFree and Copernicus-FM demonstrate strong anomaly detection capabilities, even though these results can be addressed to sensor- and domain-shift. This is confirmed under heterogeneous conditions, where performance degrades significantly when comparing datasets acquired with different satellite sensors.

Spectral invariance. The spectral invariance study reveals that the lack of invariance is primarily driven by the intrinsic spectral reduction from hyperspectral to multispectral, rather than by artifacts introduced by SRF simulation. This is confirmed by experiments involving both synthetically generated data and native satellite data.

Fine-tuning effects. The proposed semi-supervised fine-tuning strategy consistently improves AUROC for both backbones. By enforcing centroid compactness for normal samples and margin-based separation for anomalies, the embedding space is reshaped toward improved ranking behavior. Notably, Copernicus-FM exhibits a larger relative improvement, while HyperFree achieves the highest absolute performance after adaptation.

Computational considerations. Copernicus-FM demonstrates significantly lower computational cost compared to HyperFree due to reduced token interactions. This highlights an important trade-off between representational capacity and scalability in practical deployments.

Overall, the experiments confirm that while frozen foundation models provide strong general-purpose features, their robustness under domain and sensor shifts is limited. Domain-specific fine-tuning therefore emerges as a crucial step for achieving reliable anomaly detection performance in satellite imagery.

Chapter 5

Conclusions

The goal of this thesis is to adapt an industrial anomaly detection technique, namely PatchCore, to the remote sensing domain using multispectral and hyperspectral imagery. The approach leverages remote sensing foundation model backbones to handle satellite imagery and identify anomalies corresponding to active wildfires and burned areas.

This study is motivated by the limited evidence regarding the application of embedding-based anomaly detection techniques in remote sensing. In this field, anomaly detection has traditionally been treated as an unsupervised task, treating as anomalous any spectrum that deviates significantly from the image background, leaving out any semantic information.

The proposed framework applies PatchCore on embeddings extracted from two foundation model backbones, HyperFree and Copernicus-FM, to detect burned areas and active wildfires in hyperspectral and multispectral satellite imagery, leveraging the complex and informative spectral domain of this type of imagery. In addition, the analysis comprises a semi-supervised fine-tuning on the backbones to reshape the generated embedding spaces to improve anomaly ranking, together with spectral invariance studies and computational and structural comparisons of the considered backbones.

Overall, the thesis investigates whether foundation model representations can provide semantically meaningful embedding spaces for anomaly detection in complex spectral domains, and how these representations behave under domain and sensor shifts.

5.1 Main Contributions

This thesis makes the following contributions to the study of anomaly detection in remote sensing:

- A systematic adaptation of the industrial anomaly detection method PatchCore to multispectral and hyperspectral satellite imagery using foundation model embeddings.
- A structured experimental analysis of domain and sensor shift effects under homogeneous and heterogeneous conditions, highlighting their impact on embedding separability and anomaly ranking.
- A spectral invariance study combining cosine distance and CKA analysis to quantify embedding drift under spectral response function transformations and hyperspectral-to-multispectral reduction.
- The introduction of a semi-supervised fine-tuning strategy designed to reshape embedding geometry by enforcing centroid compactness and margin-based anomaly separation.
- A computational and structural comparison of foundation model backbones, linking token complexity to runtime scalability in large-scale remote sensing applications.

5.2 Key Findings

The experimental analysis leads to several key findings regarding the behavior of foundation model embeddings for anomaly detection in remote sensing.

Domain robustness of frozen foundation models is limited. While frozen backbones provide strong semantic representations in homogeneous settings, their performance degrades significantly under cross-sensor and cross-domain conditions. This indicates that foundation model embeddings, although expressive, are not inherently robust to domain shift in spectral data and they are not able to distinguish fire-related anomalies without any fine-tuning.

Spectral compression is a dominant factor in representation drift. The spectral invariance study shows that embedding differences are primarily driven by the intrinsic reduction from hyperspectral to multispectral representations, rather than by artifacts introduced by spectral response function simulation. This suggests that spectral dimensionality plays a fundamental role in shaping embedding geometry.

Embedding geometry directly governs anomaly ranking performance. The experiments reveal a strong relationship between cluster structure in the embedding space and PatchCore effectiveness. Overlapping normal and anomalous distributions result in poor ranking behavior, whereas compact normal clusters combined with margin-separated anomalies significantly improve AUROC. This

confirms that embedding reshaping through fine-tuning has a predictable and controllable effect on anomaly detection.

5.3 Research Questions

This section revisits the research questions posed in the introduction (Section 1) and summarizes the conclusions drawn from the experimental analysis.

RQ1: Can pre-trained foundation models detect fire-related events (active fires and burned areas) as anomalies using a standard unsupervised anomaly detection pipeline?

Not reliably. Although homogeneous experiments initially suggest strong performance, these results are influenced by the use of the same normal dataset for both memory bank construction and normal testing, which introduces a bias in anomaly score distribution. When evaluated under heterogeneous conditions, where normal datasets differ between memory bank and test phases, a significant performance degradation is observed. The FLOGA analysis, which removes this bias, reveals the intrinsic behavior of frozen foundation model embeddings for fire anomaly detection. In this setting, performance approaches chance level, indicating that standard unsupervised pipelines are insufficient for robust fire-related anomaly detection across domains.

RQ2: Are foundation model embeddings consistent and comparable across different satellite sensors?

No, not intrinsically. The spectral invariance analysis shows that embeddings are sensitive to spectral dimensionality and sensor characteristics. While spectral response function simulation introduces limited distortion, the intrinsic reduction from hyperspectral to multispectral representations produces measurable embedding drift. This limits cross-sensor comparability and makes unified, sensor-agnostic memory banks impractical without adaptation.

RQ3: Can semi-supervised fine-tuning with limited labeled fire examples substantially improve foundation model embeddings for fire anomaly detection?

Yes. The proposed semi-supervised fine-tuning strategy consistently improves anomaly ranking performance for both evaluated backbones. By reshaping embedding geometry through centroid compactness and margin-based separation, fine-tuning mitigates domain shift effects and enhances separability between normal and anomalous samples, even when only limited labeled fire data are available.

5.4 Limitations and Future Work

Limitations. While the proposed framework provides valuable insights into embedding-based anomaly detection for wildfire monitoring, several limitations should be acknowledged.

First, the experimental analysis, especially the fine-tuning, is primarily conducted on the FLOGA dataset, which is based on Sentinel-2 imagery acquired in Greece. Although this allows for controlled evaluation of sensor and modality effects, thereby reducing the sensor- and modality-shift, the results may not directly generalize to other satellite platforms or geographic contexts.

Second, the fine-tuning strategy focuses on partial backbone adaptation (last four ViT blocks) combined with a centroid-based objective. Alternative approaches, such as contrastive learning and pseudo-labeling, are not explored, as well as different anomaly detection techniques.

Third, due to dataset availability, the anomaly definition is restricted to wildfire-related events, namely active fires and burned areas. The applicability of the proposed framework to other types of anomalies in remote sensing imagery remains to be investigated.

Finally, computational constraints limited the extent of hyperparameter exploration and multi-seed statistical validation. A more exhaustive study could provide deeper insights into training stability and robustness.

Future Work. Several directions could extend the findings of this thesis and further improve embedding-based anomaly detection in remote sensing.

First, future work should investigate cross-sensor adaptation strategies to improve robustness across different satellite platforms. Creating an extended dataset of hyperspectral-multispectral pairs would allow a more systematic investigation of fine-tuning strategies to reduce sensor-shift, potentially enabling the same anomaly detection pipeline to operate across heterogeneous satellite inputs.

Alternative fine-tuning strategies could be explored to further improve embedding separability. While this work demonstrates how fine-tuning is an effective solution to reshape the embedding space, improving anomaly detection performance, different techniques, such as contrastive learning, LoRA or pseudo-labeling, may improve embedding space separability and training efficiency.

The anomaly definition could be expanded beyond wildfire-related events. Evaluating the proposed framework on additional anomaly categories, such as floods, volcanic activity, or other rare environmental events, would help assess the general applicability of embedding-based anomaly detection techniques in remote sensing. Similarly, the availability of additional burned-area datasets could enable a better evaluation of how the proposed fine-tuning strategy generalizes beyond the FLOGA dataset.

Finally, a larger computational budget would enable deeper hyperparameter exploration and longer fine-tuning schedules. More extensive experiments could provide a clearer understanding of the performance limits and training stability of the proposed approach.

Bibliography

[1] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. *AnySat: One Earth Observation Model for Many Resolutions, Scales, and Modalities*. 2025. arXiv: 2412.14123 [cs.CV]. URL: https://arxiv.org/abs/2412.14123 (cit. on p. 10).

[2] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. «Sen1-Floods11: A Georeferenced Dataset to Train and Test Deep Learning Flood Algorithms for Sentinel-1». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2020 (cit. on pp. 9, 12).

[3] Nikolaos Ioannis Bountos, Arthur Ouaknine, Ioannis Papoutsis, and David Rolnick. *FoMo: Multi-Modal, Multi-Scale and Multi-Task Remote Sensing Foundation Models for Forest Monitoring*. 2025. arXiv: 2312.10114 [cs.CV]. URL: https://arxiv.org/abs/2312.10114 (cit. on p. 11).

[4] Nassim Ait Ali Braham, Conrad M. Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. «SpectralEarth: Training Hyperspectral Foundation Models at Scale». In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18 (2025), pp. 16780–16797. ISSN: 2151-1535. DOI: 10.1109/jstars.2025.3581451. URL: http://dx.doi.org/10.1109/JSTARS.2025.3581451 (cit. on pp. 12, 31).

[5] Christopher F. Brown et al. *AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data*. 2025. arXiv: 2507.22291 [cs.CV]. URL: https://arxiv.org/abs/2507.22291 (cit. on p. 11).

[6] L. Bruzzone and D.F. Prieto. «Automatic analysis of the difference image for unsupervised change detection». In: *IEEE Transactions on Geoscience and Remote Sensing* 38.3 (2000), pp. 1171–1182. DOI: 10.1109/36.843009 (cit. on p. 6).

[7] Gabriel Henrique de Almeida Pereira, Andre Minoro Fusioka, Bogdan Tomoyuki Nassu, and Rodrigo Minetto. «Active fire detection in Landsat-8 imagery: A large-scale dataset and a deep-learning study». In: *ISPRS Journal of Photogrammetry and Remote Sensing* 178 (2021), pp. 171–186. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2021.06.002. URL: https://www.sciencedirect.com/science/article/pii/S092427162100160X (cit. on pp. 12, 33).

[8] Minh Kha Do, Kang Han, Phu Lai, Khoa T. Phan, and Wei Xiang. «RobSense: A Robust Multi-modal Foundation Model for Remote Sensing with Static, Temporal, and Incomplete Data Adaptability». In: *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025, pp. 7427–7436. DOI: 10.1109/CVPR52734.2025.00696 (cit. on pp. 6, 10).

[9] Martin Hermann Paul Fuchs and Begüm Demir. «HySpecNet-11k: a Large-Scale Hyperspectral Dataset for Benchmarking Learning-Based Hyperspectral Image Compression Methods». In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, July 2023, pp. 1779–1782. DOI: 10.1109/igarss52108.2023.10283385. URL: http://dx.doi.org/10.1109/IGARSS52108.2023.10283385 (cit. on p. 12).

[10] Anthony Fuller, Koreen Millard, and James R. Green. *CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders*. 2023. arXiv: 2311.00566 [cs.CV]. URL: https://arxiv.org/abs/2311.00566 (cit. on p. 10).

[11] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. *Bridging Remote Sensors with Multisensor Geospatial Foundation Models*. 2024. arXiv: 2404.01260 [cs.CV]. URL: https://arxiv.org/abs/2404.01260 (cit. on p. 10).

[12] Masroor Hussain, Dongmei Chen, Angela Cheng, Hui Wei, and David Stanley. «Change detection from remotely sensed images: From pixel-based to object-based approaches». In: *International Journal of Photogrammetry and Remote Sensing* 80 (June 2013), pp. 91–106. DOI: 10.1016/j.isprsjprs.2013.03.006 (cit. on pp. 6, 7).

[13] Gabriele Inzerillo, Diego Valsesia, and Enrico Magli. «Efficient Onboard Multitask AI Architecture Based on Self-Supervised Learning». In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18 (2025), pp. 828–838. ISSN: 2151-1535. DOI: 10.1109/jstars.2024.3502776. URL: http://dx.doi.org/10.1109/JSTARS.2024.3502776 (cit. on p. 10).

[14] Johannes Jakubik et al. *Foundation Models for Generalist Geospatial Artificial Intelligence*. 2023. arXiv: 2310.18660 [cs.CV]. URL: https://arxiv.org/abs/2310.18660 (cit. on p. 8).

[15] Johannes Jakubik et al. *TerraMind: Large-Scale Generative Multimodality for Earth Observation*. 2025. arXiv: 2504.11171 [cs.CV]. URL: https://arxiv.org/abs/2504.11171 (cit. on p. 11).

[16] Alexander Kirillov et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV]. URL: https://arxiv.org/abs/2304.02643 (cit. on p. 14).

[17] Jingtao Li et al. *HyperFree: A Channel-adaptive and Tuning-free Foundation Model for Hyperspectral Remote Sensing Imagery*. 2025. arXiv: 2503.21841 [cs.CV]. URL: https://arxiv.org/abs/2503.21841 (cit. on pp. 6, 9, 11, 12, 14, 17).

[18] Sicong Liu, Daniele Marinelli, Lorenzo Bruzzone, and Francesca Bovolo. «A Review of Change Detection in Multitemporal Hyperspectral Images: Current Techniques, Applications, and Challenges». In: *IEEE Geoscience and Remote Sensing Magazine* 7.2 (2019), pp. 140–158. DOI: 10.1109/MGRS.2019.2898520 (cit. on p. 6).

[19] Siqi Lu, Junlin Guo, James R Zimmer-Dauphinee, Jordan M Nieusma, Xiao Wang, Parker VanValkenburgh, Steven A Wernke, and Yuankai Huo. *Vision Foundation Models in Remote Sensing: A Survey*. 2025. arXiv: 2408.03464 [cs.CV]. URL: https://arxiv.org/abs/2408.03464 (cit. on p. 7).

[20] Valerio Marsocci et al. *PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models*. 2025. arXiv: 2412.04204 [cs.CV]. URL: https://arxiv.org/abs/2412.04204 (cit. on pp. 6, 8, 9).

[21] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwar, Salman Khan, and Fahad Shahbaz Khan. *Rethinking Transformers Pre-training for Multi-Spectral Satellite Imagery*. 2024. arXiv: 2403.05419 [cs.CV]. URL: https://arxiv.org/abs/2403.05419 (cit. on p. 10).

[22] Christopher Phillips, Sujit Roy, Kumar Ankur, and Rahul Ramachandran. *HLS Foundation Burnscars Dataset*. Aug. 2023. DOI: 10.57967/hf/0956. URL: https://huggingface.co/ibm-nasa-geospatial/hls_burn_scars (cit. on p. 9).

[23] Eduard Portales-Julia, Gonzalo Mateo-Garcia, Christopher Purcell, Yuxuan Li, Yifang Ban, and Gustau Camps-Valls. «Global Flood Extent Segmentation in Optical Satellite Images». In: *Scientific Reports* 13 (2023), p. 20316. DOI: 10.1038/s41598-023-47595-7. URL: https://doi.org/10.1038/s41598-023-47595-7 (cit. on p. 12).

[24] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. *Towards Total Recall in Industrial Anomaly Detection*. 2022. arXiv: 2106.08265 [cs.CV]. URL: https://arxiv.org/abs/2106.08265 (cit. on pp. 13, 20).

[25] Maria Sdraka, Alkinoos Dimakos, Alexandros Malounis, Zisoula Ntasiou, Konstantinos Karantzalos, Dimitrios Michail, and Ioannis Papoutsis. «FLOGA: A Machine-Learning-Ready Dataset, a Benchmark, and a Novel Deep Learning Model for Burnt Area Mapping With Sentinel-2». In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17 (2024), pp. 7801–7824. ISSN: 2151-1535. DOI: 10.1109/jstars.2024.3381737. URL: http://dx.doi.org/10.1109/JSTARS.2024.3381737 (cit. on pp. 12, 34).

[26] Noman Raza Shah, Abdur Rahman M. Maud, Farrukh Aziz Bhatti, Muhammad Khizer Ali, Khurram Khurshid, Moazam Maqsood, and Muhammad Amin. «Hyperspectral anomaly detection: a performance comparison of existing techniques». In: *International Journal of Digital Earth* 15.1 (2022), pp. 2078–2125. DOI: 10.1080/17538947.2022.2146770. eprint: https://doi.org/10.1080/17538947.2022.2146770. URL: https://doi.org/10.1080/17538947.2022.2146770 (cit. on pp. 5, 6).

[27] Oriane Siméoni et al. *DINOv3*. 2025. arXiv: 2508.10104 [cs.CV]. URL: https://arxiv.org/abs/2508.10104 (cit. on p. 11).

[28] Anh Tran, Minh Tran, Esteban Marti, Jackson Cothren, Chase Rainwater, Sandra Eksioglu, and Ngan Le. «Land8Fire: A Complete Study on Wildfire Segmentation Through Comprehensive Review, Human-Annotated Multispectral Dataset, and Extensive Benchmarking». In: *Remote Sensing* 17.16 (2025). ISSN: 2072-4292. DOI: 10.3390/rs17162776. URL: https://www.mdpi.com/2072-4292/17/16/2776 (cit. on p. 12).

[29] Bing Tu, Xianchang Yang, Baoliang He, Yunyun Chen, Jun Li, and Antonio Plaza. «Anomaly Detection in Hyperspectral Images Using Adaptive Graph Frequency Location». In: *IEEE Transactions on Neural Networks and Learning Systems* 36.7 (2025), pp. 12565–12579. DOI: 10.1109/TNNLS.2024.3449573 (cit. on p. 6).

[30] Di Wang et al. *HyperSIGMA: Hyperspectral Intelligence Comprehension Foundation Model*. 2025. arXiv: 2406.11519 [cs.CV]. URL: https://arxiv.org/abs/2406.11519 (cit. on p. 12).

[31] Yi Wang et al. *Towards a Unified Copernicus Foundation Model for Earth Vision*. 2025. arXiv: 2503.11849 [cs.CV]. URL: https://arxiv.org/abs/2503.11849 (cit. on pp. 11, 12, 14, 18, 31).

[32] Yichu Xu, Lefei Zhang, Bo Du, and Liangpei Zhang. «Hyperspectral Anomaly Detection Based on Machine Learning: An Overview». In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), pp. 3351–3364. DOI: 10.1109/JSTARS.2022.3167830 (cit. on pp. 5, 6).

[33] Yonghao Xu, Amanda Berg, and Leif Haglund. «SEN2FIRE: A Challenging Benchmark Dataset for Wildfire Detection Using Sentinel Data». In: *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, July 2024, pp. 239–243. DOI: 10.1109/igarss53475.2024.10641441. URL: http://dx.doi.org/10.1109/IGARSS53475.2024.10641441 (cit. on p. 12).

[34] Jie Zhao, Zhitong Xiong, and Xiao Xiang Zhu. *UrbanSARFloods: Sentinel-1 SLC-Based Benchmark Dataset for Urban and Open-Area Flood Mapping*. 2024. arXiv: 2406.04111 [cs.CV]. URL: https://arxiv.org/abs/2406.04111 (cit. on p. 12).

[35] Yan Zhu, Jingyang Zhu, Ting Wang, Yuanming Shi, Chunxiao Jiang, and Khaled Ben Letaief. *Satellite Federated Fine-Tuning for Foundation Models in Space Computing Power Networks*. 2025. arXiv: 2504.10403 [cs.LG]. URL: https://arxiv.org/abs/2504.10403 (cit. on p. 9).