After the Storm: Foundation Models and Classical Baselines for SAR-based Windthrow Detection

Marcus Melander & Alex Thelander

June 10, 2026

Supervisor: Professor Anders Heyden
Centre for Mathematical Sciences, Lund University

Industrial supervisor: Dr. Pol del Aguila Pla
Qamcom Research and Technology, Stockholm

Centre for Mathematical Sciences
Lund University

Qamcom Research and Technology

Abstract

Wind damage is one of the most economically and ecologically significant natural disturbances in European forests, yet the cloud cover that follows major storms routinely blocks optical satellite acquisitions for weeks, precisely when rapid damage mapping is most valuable. Synthetic aperture radar (SAR) circumvents this limitation through all-weather imaging, but pixel-level windthrow detection from Sentinel-1 remains challenging due to speckle noise, heterogeneous forest structure, and inconsistent labelling regimes across regions.

This thesis investigates whether remote sensing foundation models (RSFMs) offer a meaningful advantage over established methods for SAR-based windthrow detection, and whether they constitute a viable approach for operational Earth observation. Two multi-modal RSFMs, Galileo and SkySense++, are evaluated against three classical baselines: a backscatter threshold detector, a support vector machine, and a CNN-UPerNet, on four geographically distinct datasets covering Sweden, Russia, Ireland, and Germany. All models share an identical preprocessing pipeline based on five pre-event and five post-event Sentinel-1 acquisitions, and are compared under bi-temporal composite and multi-temporal time-series input configurations, with complementary studies on label efficiency, temporal depth, cross-dataset generalisation, augmentation strategy, and stand-level aggregation.

The results show that performance is strongly governed by annotation regime and event geometry rather than by model capacity alone. On the cleanest dataset, SkySense++ reaches 90.1% pixel-level damaged-class F_1, substantially exceeding both Galileo and the CNN. On datasets with annotation finer than the sensor can resolve, or with less consistent labels, foundation model gains are modest or negated, and pixel-level performance is bounded primarily by label fidelity. Temporal stacking improves performance on the noisier datasets and plateaus around $T = 8$ acquisitions, while object-level stand aggregation lifts every learned model above its pixel-level score and substantially above the threshold baseline, reaching 93.1% object-level F_1 on the cleanest dataset. These results suggest that SAR-based windthrow mapping is most credibly deployed as a stand-level prioritisation tool, with foundation models delivering decisive gains where annotation quality permits.

Acknowledgements

We would like to extend our deepest appreciation to our industrial supervisor, Dr. Pol del Aguila Pla, whose expertise, encouragement, and thoughtful feedback have been invaluable throughout this thesis. His research contributions and technical guidance played an important role in the development of this work. We further wish to thank Qamcom Research and Technology AB in collaboration with whom this thesis was carried out, for their support and for providing a stimulating environment in which to conduct our research.

We are likewise grateful to our supervisor, Professor Anders Heyden, for his support, insightful discussions, and valuable perspectives throughout the project.

We also wish to thank Anders Persson of Skogsstyrelsen, and Marius Rüetschi of the Swiss Federal Institute for Forest, Snow and Landscape Research for generously providing additional data and resources that made this research possible.

We further acknowledge the use of the MareNostrum 5 supercomputer at the Barcelona Supercomputing Center (BSC-CNS), whose high-performance computing resources supported the computational aspects of this research.

During the preparation of this thesis, AI-assisted tools such as ChatGPT and Claude Code were used in a limited supporting capacity, primarily for language editing, proofreading, figure generation, and programming assistance. All written material, analyses, implementations, and conclusions were critically reviewed and finalized by the authors.

Contents

List of Abbreviations **6**

1 Introduction **7**
1.1 Scope and limitations 9

2 Theory **10**
2.1 Synthetic aperture radar 10
2.1.1 Technology 10
2.1.2 Geometry and spatial resolution 10
2.1.3 Scatter 11
2.1.4 Speckle noise 12
2.1.5 Advanced SAR techniques 13
2.1.6 SAR applications 14
2.2 Windthrow 14
2.2.1 Windthrow in European forests 14
2.2.2 Ecological and economic consequences 15
2.2.3 SAR signatures of windthrown forest 15
2.2.4 Remote sensing approaches to windthrow detection 16
2.2.5 Object-based stand-level evaluation 17
2.3 Machine learning 17
2.3.1 Classical classifiers 18
2.3.2 Neural networks 19
2.3.3 Convolutional neural networks 21
2.3.4 Transformers 22
2.3.5 Evaluation metrics 23
2.4 Self-supervised learning 25
2.4.1 Types of self-supervised learning 25
2.5 Remote sensing foundation models 28
2.5.1 SUMMIT 28
2.5.2 Galileo 28
2.5.3 SkySense++ 29

3 Data **32**
3.1 Storm Johannes, Sweden . . . 33
3.2 The European Russia Windthrow Database, Russia . . . 37
3.3 Private Forest Wind Damage Assessment, Ireland . . . 37
3.4 Storm Xavier, Germany . . . 37

4 Methods **39**
4.1 Baseline classifiers . . . 39
4.1.1 Threshold . . . 39
4.1.2 Support vector machine classifier . . . 40
4.1.3 CNN-UPerNet segmentation . . . 40
4.2 Foundation models . . . 42
4.2.1 Shared downstream pipeline . . . 42
4.2.2 Bi-temporal fusion pipeline . . . 43
4.2.3 Primary bi-temporal experiment . . . 44
4.2.4 Temporal extension . . . 46
4.2.5 Controlled sensitivity studies . . . 48
4.2.6 Stand-level evaluation . . . 49
4.3 Computational environment . . . 49

5 Results **51**
5.1 Baseline performance . . . 51
5.1.1 Threshold . . . 51
5.1.2 Support vector machine . . . 51
5.1.3 CNN-UPerNet classifier . . . 52
5.2 Foundation model results . . . 53
5.2.1 Primary bi-temporal experiment . . . 53
5.2.2 Ten-acquisition temporal extension . . . 53
5.2.3 Stand-level evaluation . . . 55
5.3 Additional experiment results . . . 57
5.3.1 Experiment A: label efficiency . . . 57
5.3.2 Experiment B: temporal depth . . . 59
5.3.3 Experiment C: cross-dataset generalisation . . . 59
5.3.4 Experiment D: augmentation strategy . . . 61

6 Discussion **63**
6.1 Baseline performance across datasets . . . 63
6.2 Foundation models versus the CNN baseline . . . 66
6.3 Effect of temporal information . . . 68
6.4 Adaptation regime and label efficiency . . . 69
6.5 Cross-dataset generalisation . . . 70
6.6 Augmentation strategy . . . 71
6.7 Operational implications . . . 72
6.8 Limitations . . . 72
6.9 Future work . . . 73

7 Conclusion **76**

List of abbreviations

Abbreviation	Definition
SAR	Synthetic Aperture Radar
InSAR	Interferometric Synthetic Aperture Radar
PolSAR	Polarimetric Synthetic Aperture Radar
Pol-InSAR	Polarimetric Interferometric Synthetic Aperture Radar
S1	Sentinel-1
S2	Sentinel-2
ESA	European Space Agency
VV	Vertical-Vertical (polarization)
VH	Vertical-Horizontal (polarization)
HH	Horizontal-Horizontal (polarization)
HV	Horizontal-Vertical (polarization)
ML	Machine Learning
CNN	Convolutional Neural Network
ViT	Vision Transformer
SVM	Support Vector Machine
SSL	Self-Supervised Learning
MIM	Masked Image Modeling
FM	Foundation Model
RSFM	Remote Sensing Foundation Model
IoU	Intersection over Union (also known as Jaccard Score)
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
P/R	Precision / Recall

1 Introduction

Due to its all-weather active microwave imaging, synthetic aperture radar (SAR) has become an essential tool in remote sensing applications. This makes SAR uniquely suited for time-critical applications such as disaster response, maritime surveillance, windthrow, and change detection in regions with persistent cloud cover. However, interpreting SAR imagery remains challenging: unlike optical sensors, SAR captures coherent backscatter intensity, which is influenced by surface roughness, dielectric properties, and acquisition geometry rather than visual appearance.

Recent years have seen rapid progress in large-scale self-supervised learning (SSL), where models are pretrained on massive collections of unlabeled data and subsequently adapted to many downstream tasks with limited supervision [52]. In remote sensing, the appeal of this is especially strong. Satellite archives grow continuously, whereas high-quality manual annotations remain scarce. Manual annotation of SAR imagery is considered costly and requires expert knowledge [38, 56]. This data imbalance has made remote sensing a natural application area for self-supervised representation learning, often referred to as remote sensing foundation models (RSFMs), which aim to generalize across sensors, geographies, and tasks [52].

However, many popular vision backbones are pretrained on natural-image datasets, so their learned features and training assumptions may not transfer cleanly to remote sensing data [52]. This is particularly relevant for SAR, which exhibits modality-specific phenomena such as speckle and a more indirect relationship between pixel intensities and semantic scene content, increasing both labelling difficulty and the risk of negative transfer from optical pretraining [32]. These considerations motivate evaluating foundation models that are trained on SAR, or explicitly include SAR during pretraining, as they are more likely to learn representations aligned with SAR scattering characteristics [52].

An important enabler for developing and evaluating such models is the availability of large, open Earth observation archives. In particular, the Copernicus Sentinel missions provide globally-consistent, freely-accessible data products at scale. Sentinel-1 (S1) delivers C-band SAR imagery, while Sentinel-2 (S2) provides multi-spectral optical observations with complementary information content [15]. This open-access setting supports reproducible benchmarking and enables large-scale self-supervised pretraining on realistic, multi-modal remote sensing data [52].

In parallel, advances in deep learning architectures have substantially improved the quality of learned visual representations. Early progress in remote sensing and computer vision was driven by convolutional neural networks (CNNs) [57], whose locality and translation-equivariance made them effective feature extractors for imagery [53]. More recently, transformer-based architectures, including vision transformers (ViTs) and hierarchical variants, have emerged as increasingly popular alternatives due to their ability to model long-range dependencies and capture global image context through self-attention [57]. This growing adoption of transformer models has been observed across remote sensing applications, where they are receiving widespread attention as complementary or alternative feature extraction backbones [57]. These architectural developments have, in turn, helped accelerate self-supervised learning, as transformer backbones paired with masking or distillation-style objectives often produce transferable representations that improve label efficiency and downstream robustness [35, 51].

Despite the growing interest in RSFMs, the field currently lacks standardised benchmarks, making it difficult to assess how well these models perform relative to one another and to established machine learning baselines. This uncertainty is compounded by the practical demands of foundation models, which often have large memory and compute requirements, limiting accessibility and reproducibility. For SAR, the problem is more specific. Most foundation models are pretrained on optical imagery, and it remains unclear whether models that explicitly incorporate SAR during pretraining offer a meaningful advantage for SAR-specific downstream tasks.

One such SAR-specific task is wind damage to forests (windthrow) which is one of the most economically and ecologically significant natural disturbances in Europe, with major storm events capable of felling timber volumes equivalent to several years of normal harvest and triggering secondary cascades of fire risk and bark beetle outbreaks [19, 42]. Detecting and mapping windthrow rapidly after a storm is operationally critical, yet optical satellite imagery is unreliable for this purpose when it is most needed. The frontal weather systems that produce damaging winds leave persistent cloud cover that can block acquisition for weeks, as illustrated by the Irish post-storm mapping campaign following Storm Éowyn in early 2025, where cloud-free imagery was unavailable until several weeks after the event [10]. SAR is unaffected by cloud cover and responds to the characteristic collapse of canopy structure caused by windthrow through a measurable change in backscatter intensity [39, 47], making it a natural modality for operational post-storm monitoring.

Therefore this thesis investigates whether RSFMs constitute a viable approach for operational Earth observation using SAR data by evaluating them on the downstream application of windthrow detection. Two foundation models, Galileo [49] and SkySense++ [54], are evaluated against three classical machine learning baselines: a backscatter threshold detector, a support vector machine (SVM), and a UPerNet [55] convolutional neural network. A third foundation model, SUMMIT [13], was considered but excluded after preliminary experiments; see Section 1.1. Rather than comparing the foundation models against one another, the question is whether their learned representations offer a meaningful

advantage over already established methods on a task of industrial relevance. The evaluation spans four geographically distinct datasets covering parts of Sweden, Russia, Ireland, and Germany.

This leads to the following research question:

> *Can remote sensing foundation models provide a meaningful advantage over established baselines for SAR-based windthrow detection, and do they constitute a viable approach for operational Earth observation monitoring of windthrow?*

1.1 Scope and limitations

This thesis is scoped to binary semantic segmentation of windthrow using Sentinel-1 SAR imagery. The evaluation covers four geographically distinct datasets from European forest types. Conclusions about generalisability to tropical, subtropical, or non-European boreal forests cannot be drawn from the results reported here. Only Sentinel-1 C-band backscatter intensity in the VV and VH polarisations is used.

Three foundation models were considered: Galileo [49], SkySense++ [54], and SUMMIT [13]. SUMMIT was excluded from the main evaluation after preliminary experiments yielded uniformly poor results using our method across all dataset variants. Both Galileo and SkySense++ was pretrained on multimodal imagery including optical and multispectral sensors, but is evaluated here in a SAR-only configuration; results therefore do not reflect the full multimodal capability described by their authors [49, 54].

2 Theory

2.1 Synthetic aperture radar

2.1.1 Technology

Unlike optical sensors that rely on reflected sunlight, SAR is an active microwave imaging system. The sensor carries its own illumination source, transmitting electromagnetic pulses toward the Earth's surface and recording the backscattered echoes. This independence from external light sources, combined with the ability of microwaves to penetrate clouds, makes SAR an essential tool for consistent, all-weather Earth observation.

The raw data recorded by a SAR system is captured as a complex-valued signal, which characterizes the surface through three fundamental components: intensity, phase, and polarization. The intensity, or amplitude, represents the magnitude of the backscattered energy and is primarily dictated by the physical roughness and dielectric properties of the observed target. The phase component records the specific state of the electromagnetic wave upon its return, effectively encoding the precise distance between the sensor and the target to within a fraction of the wavelength. The polarization describes the orientation of the electromagnetic field's oscillation as it interacts with the terrain. By controlling the plane in which the wave is transmitted and received, typically in horizontal or vertical orientations, the SAR system can get an understanding of the structural geometry of the target [32].

2.1.2 Geometry and spatial resolution

A SAR sensor views the world through a specific geometric lens that differs significantly from the nadir view (the perspective looking directly down at the Earth's surface, perpendicular to the ground) of most optical satellites. The platform moves along a flight path, known as the azimuth direction, while looking off to the side toward the cross-track or range direction, as shown in Figure 2.1. To achieve high-resolution images, the system must resolve targets in these two dimensions. The range resolution is determined by the pulse bandwidth, allowing the radar to distinguish between targets based on the time delay of their echoes.

The azimuth resolution presents a greater challenge. In a traditional real aperture radar, this resolution depends on the physical length of the antenna.

To achieve a 10-meter resolution from space, an antenna several kilometers long would be required. SAR overcomes this physical limit through synthetic aperture processing. As the platform moves with velocity v, it transmits thousands of pulses that illuminate the same target from different positions along the orbit. By coherently combining these echoes and exploiting the phase history, the system synthesizes a massive virtual antenna. This enables high azimuth resolution that is independent of the distance to the target [32].

Beyond geometry, the choice of operating frequency fundamentally shapes what a SAR sensor can observe. SAR systems transmit at microwave frequencies that are conventionally grouped into bands, each associated with a characteristic wavelength range. Because penetration depth scales with wavelength, the band determines how the signal interacts with vegetation and the ground. Longer wavelengths penetrate deeper into a volume and carry a more pronounced volume contribution in the backscattered signal, whereas shorter wavelengths interact predominantly with the upper surface of the canopy [32]. The most commonly used bands in Earth observation are L, C, and X, summarised in Table 2.1.

Table 2.1: Commonly used SAR frequency bands and their characteristics. Wavelength and frequency ranges follow [32]

Band	**Frequency (GHz)**	**Wavelength (cm)**
X	12–7.5	2.5–4
C	7.5–3.75	4–8
L	2–1	15–30
P	0.5–0.25	60–120

2.1.3 Scatter

In SAR, the backscattered signal from each resolution cell is measured as a complex quantity, containing both amplitude and phase, and this measurement is made separately for different polarization combinations. While phase encodes the propagation path and coherent interaction of the radar wave with the target, the polarization describes how the target transforms the incident electromagnetic field.

Modern systems achieve this by transmitting and receiving waves in specific orthogonal planes, horizontal (H) and vertical (V). A fully polarimetric (quad-pol) system captures four combinations: HH and VV (co-polarized), and HV and VH (cross-polarized). These channels are useful for identifying the geometric properties of a target, as different surfaces "twist" or reflect the signal in predictable ways.

Together, phase and polarization determine the observed scattering behavior. There are three main types of scatter; surface scattering, dihedral scattering and volume scattering, summarised in Table 2.2.

Surface scattering is characterised by a single dominant reflection from a rough interface. It largely preserves the incident polarization, resulting in strong

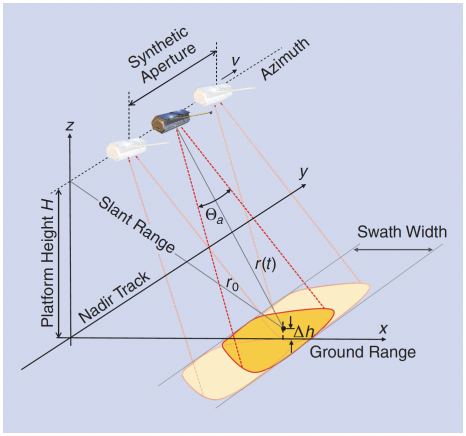


Figure 2.1: Illustration of the SAR imaging geometry. r_0 stands for the shortest approach distance, Θ_a for the azimuth beam width, and v the sensor velocity. Figure reprinted from A. Moreira et al., 'A Tutorial on Synthetic Aperture Radar,' IEEE Geoscience and Remote Sensing Magazine, March 2013. © 2013 IEEE [32].

co-polarized returns (HH and VV) with a stable phase relationship and weak cross-polarized components. The relatively low randomness of the phase and the polarization preservation indicate a deterministic scattering mechanism.

Dihedral scattering arises from double-bounce interactions, such as ground to wall or trunk to ground geometries. This mechanism also produces strong co-polarized returns, but with a distinctive phase difference between HH and VV caused by the two reflections. Polarimetric phase relationships therefore play a key role in identifying dihedral scatterers, which are common in urban and forested environments.

Volume scattering occurs when the radar wave penetrates a medium containing many randomly oriented scatterers. The coherent summation of many scattering contributions leads to significant depolarization, strong cross-polarized returns, and random phase behavior within the resolution cell. This combination of high cross-polarized power and phase randomness is common in volumetric media such as vegetation, snow, or dry soil [32].

2.1.4 Speckle noise

In SAR imagery, the dominant form of noise is speckle. Speckle arises because SAR is a coherent imaging system, where within each resolution cell, many small elemental scatterers contribute echoes with different amplitudes and phases. When these echoes are coherently summed, they interfere constructively and destructively, producing strong pixel-to-pixel fluctuations in both intensity and phase, even over physically uniform surfaces. As a result, SAR image intensity follows a random statistical distribution rather than a deterministic one. Speckle

Table 2.2: Interpretation of radar polarizations and dominant scattering mechanisms.

Polarization	**Scattering Mechanism and Typical Targets**
HH	**Surface & Dihedral Scattering** Calm water, road surfaces, and tree trunks (double-bounce).
VV	**Surface Scattering** Rough sea surfaces (waves/wind) and vertical structures (no double bounce).
HV	**Volume Scattering** Forest canopies, complex ice ridges, and vegetation (the signal is "twisted" by the target).
VH	**Volume Scattering** Physically identical to HV. Usually averaged with HV to improve the detection of forests/biomass.

is not caused by instrument errors and cannot be reduced by increasing transmit power. Instead, it is multiplicative in nature, its variance increases with signal strength. Although commonly referred to as "noise," speckle contains physical information about the scatterer distribution within each resolution cell. In practice, speckle is mitigated using techniques such as multi-looking, adaptive filtering or mean composites, which improve radiometric interpretability, but at the cost of some spatial resolution [32].

In the era of deep learning, more advanced methods such as convolutional neural networks and self-supervised learning have been deployed to denoise SAR images while better preserving fine structural details and edges than classical filters [59].

2.1.5 Advanced SAR techniques

By using the multi-dimensional nature of the radar signal, specifically through the combination of polarizations and multiple viewing geometries, techniques such as Polarimetric SAR (PolSAR), Interferometric SAR (InSAR), and Polarimetric Interferometric SAR (Pol-InSAR) are used.

PolSAR focuses on the analysis of the full scattering matrix $\mathbf{S}$, which relates the incident and scattered electric fields through the horizontal (H) and vertical (V) polarization states. In a fully polarimetric system, the complex scattering behavior of a target is captured as:

$$\mathbf{S} = \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix}$$

PolSAR techniques utilize mathematical decomposition theorems, such as the Pauli or Freeman-Durden decompositions [7, 16], to separate the total

backscattered energy into the physical mechanisms discussed in the previous sections: surface, double-bounce, and volume scattering.

InSAR uses the phase component of the complex signal by comparing two or more images of the same scene acquired from slightly different orbital positions or at different times [37]. The primary output of this technique is an interferogram, which represents the phase difference between the signals. This phase difference is for example very sensitive to the topography of the terrain and any subtle displacements of the surface that occurred between acquisitions. InSAR is the standard for monitoring millimeter-scale ground deformations, such as those caused by seismic activity, volcanic inflation, or urban subsidence [29].

Pol-InSAR is the fusion of polarimetry and interferometry. It represents one of the most sophisticated levels of radar remote sensing. While InSAR provides information regarding the effective height of a scatterer, PolSAR provides sensitivity to the scatterer's shape and orientation. By performing interferometric measurements across different polarization channels, Pol-InSAR can distinguish between scattering contributions originating from different vertical layers within a single resolution cell [33]. This is particularly useful for forest monitoring, where the technique can separate the volume scattering signal returning from the canopy, the dihedral scattering of the trunk, and the surface scattering from the ground beneath it [32].

2.1.6 SAR applications

SAR is used across a wide range of Earth observation tasks owing to its independence from solar illumination and cloud cover [32]. Single-scene applications include ship detection, sea ice monitoring, and ocean wind-field retrieval, where VV backscatter responds to surface roughness in characteristic ways. In forestry and agriculture, PolSAR decompositions provide estimates of biomass and canopy structure that optical sensors cannot directly observe [33]. Multi-temporal analysis further expands these capabilities: by comparing acquisitions over time, InSAR enables millimetre-scale surface deformation mapping for seismic, volcanic, and subsidence monitoring [29], while backscatter change detection supports flood mapping, deforestation monitoring, and windthrow detection in forests [39, 47].

2.2 Windthrow

2.2.1 Windthrow in European forests

Windthrow refers to the uprooting or stem breakage of trees caused by wind loading that exceeds the mechanical resistance of the root system or the stem itself [22]. The primary meteorological drivers are extratropical cyclones tracking across the Atlantic into the European continent, capable of producing sustained winds and gusts over large areas of forest in a matter of hours. Forest susceptibility is governed by stand height and structure, species composition, edge exposure,

and soil and drainage conditions, with conifer stands on exposed and wet sites among the most vulnerable [18].

The historical record documents a pattern of recurring, large-scale damage across Europe, of which the following events represent only a selection of the most severe. Storm Lothar in December 1999, together with its sister storm Martin, felled an estimated $228 \times 10^6\,\mathrm{m}^3$ of timber across France, Switzerland, and Germany [22]. Storm Gudrun in January 2005 destroyed approximately $77 \times 10^6\,\mathrm{m}^3$ in Sweden alone. Storm Kyrill in January 2007 caused around $49 \times 10^6\,\mathrm{m}^3$ of loss across Central Europe [22]. More recently, the storms Darragh and Éowyn during the winter of 2024/2025 caused widespread forest damage in Ireland and the United Kingdom [11, 30]. Gregow et al. [22] show that reported windstorm damage to European forests increased significantly over the period 1951 to 2010. Seidl et al. [42] project further increases under future climate scenarios.

2.2.2 Ecological and economic consequences

The operational urgency of rapid windthrow assessment arises from two compounding pressures. The first is economic: the volume of timber damaged in a single large storm has in some cases exceeded the entire annual harvest of the affected country, and fallen material depreciates rapidly through fungal infection in the weeks following the event [39]. The scale of losses is substantial. Windstorm Ciarán, which struck France, the United Kingdom, Belgium, and the Netherlands in November 2023, resulted in a loss of over € 2 billion [34]. The second pressure is biological: fallen and damaged timber provides ideal breeding habitat for bark beetles, whose populations can increase rapidly in the years following a storm and attack adjacent standing trees if windthrown material is not removed promptly [39]. Forest managers therefore require spatially explicit damage maps as quickly as possible after a storm to prioritise salvage harvesting, plan road access, and schedule phytosanitary treatments. Seidl et al. [42] further document that intensifying disturbance regimes reduce the capacity of European forests to act as carbon sinks, with projections indicating that damage-related losses in carbon storage will continue to grow under future climate scenarios.

2.2.3 SAR signatures of windthrown forest

The physical basis for detecting windthrow with SAR backscatter lies in the structural reorganisation of the forest canopy following a storm. An intact forest produces a characteristic backscatter regime in which the canopy contributes strongly to volume scattering in cross-polarised returns, while co-polarised returns receive contributions from both the canopy volume and double-bounce interactions between vertically oriented trunks and the ground surface, as described in Section 2.1.3. The spatial regularity and upright orientation of living trees collectively produce a stable multi-temporal signature that changes slowly over the phenological cycle [39].

Windthrow disrupts this regime through two concurrent mechanisms. First, the collapse of the canopy removes the volumetric scattering layer and exposes the underlying ground surface. Second, the fallen trunks and debris create a chaotically arranged assembly of scattering from fallen trees, which increases local surface roughness and introduces diffuse returns across both polarisation channels. Rüetschi et al. [39] report statistically significant post-event backscatter increases in both VV and VH polarisations for windthrown mixed temperate forest in Switzerland and Germany, with the effect more pronounced in VH due to the sensitivity of cross-polarised returns to the randomised orientation of fallen material. An important complication is that the sign and magnitude of the response depend on damage severity, with Tomppo et al. [47] noting that large stand-replacing events can produce a response resembling clear-cutting, with an overall decrease in backscatter driven by canopy loss, whereas smaller damage patches may instead increase local returns through edge diffraction and debris roughness. Tanase et al. [46] found that L-band backscatter decreased after windthrow, in contrast to the positive responses reported in Sentinel-1 studies at shorter wavelengths (C-band). Environmental conditions such as soil moisture and freeze-thaw state further modulate the signal, as frozen ground suppresses the ground-scattering contribution and can partially mask the windthrow signature in winter events [47].

2.2.4 Remote sensing approaches to windthrow detection

Windthrow mapping from space has been approached through both optical and SAR imagery, with each modality presenting distinct trade-offs between spatial detail, spectral information content, and availability. Optical sensors have been applied across European storm events with strong results under cloud-free conditions. Einzmann et al. [14] applied a hybrid object- and pixel-based change detection workflow using 5 m spatial resolution RapidEye imagery to sites damaged by Storm Niklas in 2015, identifying over 90% of windthrow areas larger than 0.5 ha. The critical operational limitation of optical sensors is their dependence on cloud-free daylight conditions, as cloud cover following a storm can substantially delay the first usable acquisition, removing much of the value of early damage mapping [39]. Vaglio Laurin et al. [50] found that optical data outperformed SAR for mapping the 2018 Vaia storm in northern Italy, while SAR remained valuable for rapid first assessment because of its weather independence.

These limitations motivate the use of SAR as the primary tool for rapid operational windthrow assessment. Rüetschi et al. [39] established the methodological baseline for S1 windthrow detection using a bi-temporal change-detection approach applied to composite backscatter images built from five pre-event and up to ten post-event acquisitions. A heuristic windthrow index, defined as the sum of the post-minus-pre differences in both VV and VH polarisations, is thresholded within a forest mask, after which a connected-component size filter removes noise-driven isolated detections. The method was validated across a Swiss development area and a German study area affected by Storm Xavier in October 2017, achieving strong detection of areal windthrow while struggling

with scattered individual fallen trees. The study further demonstrated that five to six post-event acquisitions were sufficient to deliver windthrow maps within approximately two weeks of a storm event [39]. The object-based evaluation protocol used by Rüetschi et al. is described in detail in Section 2.2.5, as the same protocol is adopted in this thesis to score the learned models at the stand level.

Tomppo et al. [47] extended the S1 detection framework to the boreal context by evaluating three supervised classifiers on forest stands in southern Finland affected by a summer 2017 storm. Their support vector machine achieved an overall accuracy of 0.79, and already 0.75 accuracy was attainable from a single post-event acquisition acquired two days after the storm, implying that near-real-time localisation is feasible within the first S1 repeat cycle [47].

A common observation across these studies is that accuracy metrics based on object-level overlap, such as producer's and user's accuracy, tend to present a more optimistic picture of detection capability than pixel-level evaluation. Pixel-level F_1 provides a stricter measure of how well predicted damage boundaries agree with reference annotations and is therefore used as the primary metric throughout this thesis.

2.2.5 Object-based stand-level evaluation

The evaluation protocol of Rüetschi et al. [39] operates at the level of connected damage objects rather than individual pixels. Reference windthrow stands are defined as 8-connected components of the binary damage mask, and predicted windthrow stands are defined as 8-connected components of the thresholded prediction raster. A reference stand is considered detected if it shares at least one pixel with any predicted stand, and a predicted stand is considered correct if it shares at least one pixel with any reference stand. These overlap matches feed two complementary object-level accuracy measures. Producer's accuracy (PA) measures the fraction of reference stands that were detected and user's accuracy (UA) measures the fraction of predicted stands that are confirmed by a reference stand. In the remote sensing literature these correspond to object-level recall and object-level precision, respectively. The equations for precision and recall can be found at Equation 2.10. Rüetschi et al. optimise parameter choices by the arithmetic mean of PA and UA.

2.3 Machine learning

Machine learning (ML) is a subfield of artificial intelligence in which a model learns a mapping from input data to outputs by optimising a parametric function with respect to a loss, rather than through explicit hand-crafted rules. For image-based tasks such as SAR analysis, the inputs are high-dimensional arrays of pixel values and the outputs may be class labels (classification) or per-pixel predictions (segmentation). Methods differ in how much structure they impose on this mapping and, correspondingly, in how many parameters they fit from

data. This work spans that range, progressing from a hand-crafted detection rule, through classical classifiers that fit a decision boundary over fixed input features, to neural networks that learn the relevant features directly from the imagery. These methods form the baselines against which the foundation models are later assessed.

2.3.1 Classical classifiers

The first two baselines, the threshold-based detection and the support vector machine, sit at the lower end of this range. Together they establish how far the damage signal can be recovered before any features are learned from the data itself.

Threshold-based detection

The simplest possible detector assigns a binary label based on whether a single scalar feature exceeds a fixed threshold t. The threshold is selected by sweeping candidate values on a training split and choosing the one that maximises a validation metric. This is a grid search over a hand-crafted rule rather than the optimisation of a parametric model against a loss, so the detector falls outside the ML definition above. It is retained because its simplicity and interpretability make it a natural lower bound against which the learning-based methods are compared.

Support vector machine

A support vector machine (SVM) is a kernel-based classifier that constructs a decision boundary by finding the hyperplane that maximises the margin between two classes in a, possibly high-dimensional, feature space. Rather than fitting a flexible curve directly through the training data, the SVM anchors its boundary to the small subset of training points that lie closest to the decision boundary, known as support vectors, and pushes the boundary as far as possible from both classes simultaneously. This maximum-margin principle gives the SVM strong generalisation properties even when training data are scarce.

A key strength of the SVM is the kernel trick: instead of explicitly mapping inputs into a high-dimensional feature space, the classifier computes only inner products between pairs of points through a kernel function $\kappa(\mathbf{x}_i, \mathbf{x}_j)$, making non-linear decision boundaries tractable without increasing the cost of training. A regularisation parameter C controls the trade-off between maximising the margin and tolerating misclassified training points, with larger values of C producing a tighter fit to the training data at the risk of overfitting. This work uses both a linear kernel, which is computationally efficient and well suited to high-dimensional feature spaces, and a radial basis function (RBF) kernel

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right), \tag{2.1}$$

which measures similarity by Euclidean distance in the original input space and can model curved decision boundaries [21].

2.3.2 Neural networks

A neural network is a parametric function composed of layers of simple computational units called neurons. Each neuron computes a weighted sum of its inputs and passes the result through a non-linear activation function σ:

$$z = \sigma\big(\mathbf{w}^\top \mathbf{x} + b\big)\,, \tag{2.2}$$

where $\mathbf{w}$ are learnable weights, b is a bias term, and $\mathbf{x}$ is the input vector. Common activation functions include the Rectified Linear Unit (ReLU), $\sigma(z) = \max(0, z)$, and the sigmoid, $\sigma(z) = 1/(1 + e^{-z})$. Without non-linearity, stacking multiple layers would collapse to a single affine transformation, and the network could not represent complex functions. Layers are stacked so that each layer receives the activations of the previous one as input, building progressively more abstract representations of the data. The final layer produces the network's output, for example, a probability over classes or a per-pixel damage score [21].

Loss functions

To train a network, a scalar loss $\mathcal{L}$ measures the discrepancy between predictions and ground truth labels. The following subsections introduce the loss functions used to train the segmentation models in this work, covering standard formulations and the class-imbalance-aware variants that are necessary when the target class occupies only a small fraction of each tile.

Cross-Entropy Loss. Standard cross-entropy (CE) treats all pixels equally, which is poorly suited to imbalanced datasets [36]. To compensate, class weights were introduced such that the damaged class receives a substantially higher penalty:

$$\mathcal{L}_{\text{CE}} = -\sum_i w_{y_i} \log \hat{p}_{i,y_i}, \quad w = [1.0,\ w_{\text{event}}], \tag{2.3}$$

where the event weight w_{event} is derived from the pixel-level class frequency in the training set:

$$w_{\text{event}} = \sqrt{N_{\text{healthy}}/N_{\text{windthrow}}}, \tag{2.4}$$

with N_{healthy} the number of healthy forest pixels and $N_{\text{windthrow}}$ the number of windthrow pixels. This ratio scales the minority-class penalty inversely with its prevalence, producing a large weight when windthrow is rare.

Focal Loss. Focal loss modulates the per-pixel CE contribution by a factor that down-weights easy, well-classified examples and focuses training on hard or misclassified pixels [28]:

$$\mathcal{L}_{\text{Focal}} = -\alpha_t (1 - \hat{p}_t)^\gamma \log \hat{p}_t, \tag{2.5}$$

where $\hat{p}_t$ is the predicted probability for the ground-truth class, α_t is a class-balancing factor, and $\gamma \geq 0$ is the focusing parameter. Higher values of γ suppress easy negatives more aggressively.

Dice loss. Dice loss is an overlap-based loss that encourages the predicted mask to match the ground-truth mask. It is well suited to imbalanced segmentation problems because the loss is normalised by the sizes of the predicted and reference regions, rather than being dominated by the number of background pixels [31]:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon}, \tag{2.6}$$

where $p_i \in [0, 1]$ denotes the predicted probability for pixel i, $g_i \in \{0, 1\}$ denotes the corresponding ground-truth label, and ϵ is a small smoothing constant used to avoid division by zero.

Focal Dice Loss. The compound Focal Dice objective combines both losses with equal weighting:

$$\mathcal{L}_{\text{FocalDice}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{Focal}}. \tag{2.7}$$

This formulation was the primary configuration explored in this work, as it benefits from the class-agnostic overlap sensitivity of Dice loss together with the hard-example emphasis of focal loss.

Focal Tversky Loss. The Focal Tversky loss generalises the Dice loss by introducing asymmetric penalisation of false positives and false negatives through parameters α and β [40]:

$$T(\alpha, \beta) = \frac{\text{TP} + \epsilon}{\text{TP} + \alpha\,\text{FP} + \beta\,\text{FN} + \epsilon}, \qquad \mathcal{L}_{\text{FT}} = (1 - T)^{\gamma}. \tag{2.8}$$

Setting $\beta > \alpha$ imposes a higher cost on false negatives, which is desirable when missing a windthrow event is more costly than a spurious detection. The default configuration used $\alpha = 0.3$, $\beta = 0.7$, and $\gamma = 0.75$.

Backpropagation and optimisation

Network parameters θ are updated by gradient descent. Computing the gradient $\nabla_\theta \mathcal{L}$ efficiently across all layers is achieved by backpropagation, which applies the chain rule recursively from the output layer back toward the inputs. For a composed function $\mathcal{L} = f_n \circ \cdots \circ f_1(\mathbf{x})$, the gradient with respect to parameters in layer k is:

$$\frac{\partial \mathcal{L}}{\partial \theta_k} = \frac{\partial \mathcal{L}}{\partial f_n} \cdot \frac{\partial f_n}{\partial f_{n-1}} \cdots \frac{\partial f_{k+1}}{\partial f_k} \cdot \frac{\partial f_k}{\partial \theta_k} \tag{2.9}$$

In practice, the Adam optimiser is commonly used. It maintains per-parameter adaptive learning rates based on running estimates of the first and second moments of the gradient, making training less sensitive to the choice of global learning rate than vanilla SGD [26].

2.3.3 Convolutional neural networks

Rather than connecting every input pixel to every neuron, which would be computationally intractable for high-resolution images, a CNN applies a set of learned filters $\mathbf{W} \in \mathbb{R}^{k \times k \times C_{\text{in}} \times C_{\text{out}}}$ that slide over the spatial dimensions of the input. Each filter computes a local dot product, producing a feature map that highlights where a specific spatial pattern appears in the image. Several design choices make CNNs well-suited to imagery: weight sharing (the same filter is applied at all spatial positions, greatly reducing the parameter count), local receptive fields (each neuron responds to only a $k \times k$ neighbourhood), and spatial pooling (progressive downsampling that builds hierarchically abstract representations).

A standard CNN encoder applies repeated blocks of convolution, non-linear activation, and pooling. The successive downsampling increases the effective receptive field, later layers respond to increasingly large image regions. Allowing the network to capture both fine-grained texture and coarse global structure. This is particularly relevant for SAR-based wind damage detection, where damage manifests as backscatter change at multiple spatial scales, with individual fallen trees at fine scale and large clearings at coarse scale [21].

UPerNet

UPerNet [55] is a general-purpose dense prediction decoder designed to aggregate multi-scale feature representations produced by a hierarchical backbone. At the deepest feature level, a pyramid pooling module (PPM) as seen in Figure 2.2 applies adaptive average pooling at several fixed scales and concatenates the resulting context descriptors with the original feature map, capturing global semantic context that local convolutional filters cannot reach. The resulting representation is fused with shallower backbone features through a feature pyramid network (FPN) pathway [55], where lateral 1×1 projections align channel dimensions at each scale, and a top-down pathway progressively upsamples and adds deeper context to finer spatial levels, preserving high-frequency spatial detail while incorporating large-scale semantic information. The multi-scale outputs are concatenated and passed through a bottleneck convolution before a final 1×1 classifier produces per-pixel logits.

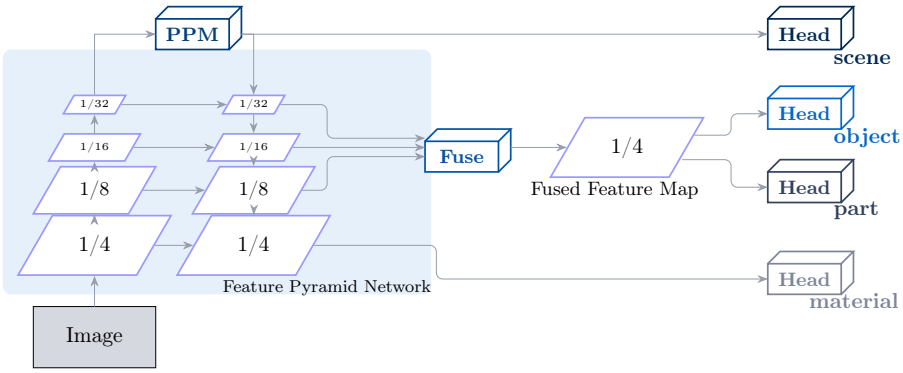


Figure 2.2: UPerNet architecture as proposed by Xiao et al. [55]. The pyramid pooling module (PPM) at the deepest feature level captures global context; a feature pyramid network (FPN) fuses this with shallower backbone features at four spatial scales. In this work the four task-specific heads are replaced by a single two-channel classifier for binary windthrow segmentation.

2.3.4 Transformers

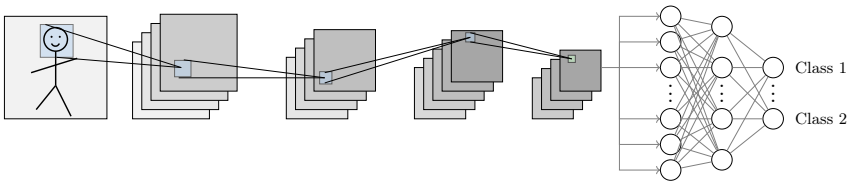
Transformers are a type of neural network architecture originally designed for processing sequences of data, such as sentences in natural language processing. Their core innovation is the self-attention mechanism, which allows the model to weigh the importance of different parts of the input relative to each other rather than processing information strictly in order or through local filters. This enables transformers to capture long-range relationships efficiently and in parallel, making them highly effective for tasks involving complex dependencies [51].

Vision transformers

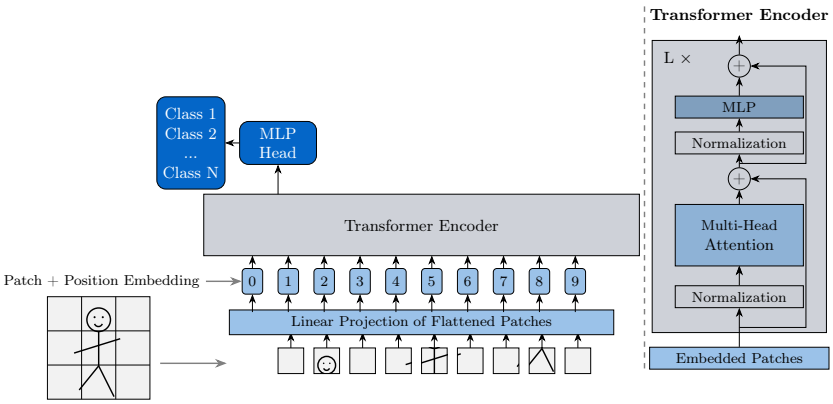
Vision transformers (ViT) adapt the transformer architecture from natural language processing to computer vision by rethinking how visual data is represented. Instead of processing words or subwords, an image is split into fixed-size patches [12]. Each patch is linearly embedded and treated as a sequence token, making the input compatible with standard self-attention mechanisms. Because transformers do not inherently account for the spatial arrangement of these tokens, positional embeddings are added to preserve the structural context of the image [12].

Self-attention allows every patch to directly interact with every other patch, enabling the model to capture long-range dependencies across the entire image in a single layer, a capability that CNNs only achieve indirectly through successive layers and pooling. Unlike CNNs, ViTs lack strong inductive biases such as locality and translation invariance. Consequently, they typically require large-scale pre-training on extensive datasets to learn spatial hierarchies effectively. However, this trade-off results in a highly flexible, globally aware representation

that transfers exceptionally well across various downstream tasks [12]. To further optimize this approach, extensions such as hierarchical vision transformers introduce multi-scale representations, merging the architectural strengths of CNNs with the expressive power of attention-based models.



(a) Common CNN architecture



(b) Vision Transformer architecture

Figure 2.3: Comparison of a standard CNN architecture (a) and a Vision Transformer (b). The CNN applies learned filters locally at every spatial position; the ViT splits the image into fixed-size patches, embeds each as a token, and uses self-attention to model relationships across the full image in a single layer.

2.3.5 Evaluation metrics

Windthrow detection is a binary segmentation problem with severe class imbalance, damaged pixels are a small fraction of the total forested area. Standard accuracy (fraction of correctly classified pixels) is therefore misleading. A classifier that predicts no damage everywhere can achieve high accuracy simply because most pixels are undamaged. The metrics used in this work are instead derived from the confusion matrix, which counts four outcomes for each pixel: true positives (TP, correctly predicted damage), true negatives (TN, correctly predicted no damage), false positives (FP, undamaged pixels incorrectly flagged),

and false negatives (FN, damaged pixels missed) [21].

Precision and recall

Precision measures what fraction of pixels flagged as damaged are truly damaged, while recall measures what fraction of truly damaged pixels are detected:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \ , \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.10}$$

The two metrics capture complementary failure modes. A model that flags everything as damaged achieves perfect recall but near-zero precision. A model that only flags the most obvious damage achieves high precision but low recall. Both failure modes are costly in an operational context, where false positives waste field inspection resources, while false negatives leave undetected damage unaddressed.

F1 score

The F1 score is the harmonic mean of precision and recall, providing a single scalar that balances both:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}} \tag{2.11}$$

The harmonic mean is used rather than the arithmetic mean because it penalises large imbalances between precision and recall, a model with precision 1.0 and recall 0.01 scores F1 = 0.02, not 0.5.

Intersection over union

Intersection over Union (IoU), also known as the Jaccard index, measures the overlap between the predicted damage mask $\hat{M}$ and the reference mask M:

$$\text{IoU} = \frac{|M \cap \hat{M}|}{|M \cup \hat{M}|} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{2.12}$$

IoU is the standard metric for segmentation benchmarks and is more stringent than F1, since it penalises both false positives and false negatives in the denominator without the factor of two that softens the F1 denominator. A model must achieve precise spatial alignment with the reference mask to score well on IoU, making it a particularly informative metric for the spatially structured task of forest damage mapping. mIoU is calculated as the arithmetic mean of the IoU scores across both classes.

2.4 Self-supervised learning

Self-supervised learning (SSL) represents a fundamental shift in machine learning, enabling models to extract rich feature representations from unlabeled data without human supervision [35]. By constructing learning tasks directly from the input itself, such as reconstructing masked components or contrasting augmented views, SSL overcomes the scalability bottlenecks present in supervised learning. This approach enables the creation of robust FMs that often surpass supervised counterparts [52], particularly in their ability to generalize across diverse tasks. SSL has consequently become the standard for handling large-scale datasets where annotation is impractical or prohibitively expensive. This challenge is especially acute in remote sensing, where satellites generate petabytes of imagery daily, creating a setting where raw data is effectively unlimited but labeled ground truth remains scarce. For SAR in particular, the problem is compounded by the non-literal nature of the imagery. Rather than visual appearance, SAR captures complex electromagnetic scattering properties, making human interpretation difficult and requiring expert-level domain knowledge for reliable labelling [38, 56].

2.4.1 Types of self-supervised learning

The roots of SSL extend to the early deep learning era, evolving from pretext tasks such as image colorization, jigsaw puzzle solving, and layer-wise pretraining. Despite this history, the field historically lacked a unified taxonomy, with definitions frequently overlapping as methods share mechanisms across different objectives. To provide a structured foundation for the analysis that follows, this work adopts the classification framework proposed by Balestriero et al. [1], which unifies the field into four distinct families based on their training objectives: contrastive learning, self-distillation, redundancy reduction, and masked image modeling.

Contrastive

Contrastive learning, rooted in deep metric learning, operates on the principle of instance discrimination. The training objective encourages the encoder to map semantically equivalent augmented views of the same image, termed positive pairs, to nearby points in the embedding space, while simultaneously repelling representations of different images, referred to as negative pairs. This dual pressure prevents the model from collapsing to a trivial constant solution [1].

Modern contrastive methods commonly implement this principle using the InfoNCE loss. Given a batch of N images, two augmented views are generated for each image, resulting in $2N$ samples. For a positive pair $(\mathbf{z}_i, \mathbf{z}_j)$, the loss encourages their similarity to be high relative to the similarity between $\mathbf{z}_i$ and all other views in the batch:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)}, \tag{2.13}$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ denote latent representations of two augmented views of the same image, sim denotes cosine similarity, τ is a temperature parameter controlling the sharpness of the similarity distribution, and $\mathbf{1}_{[k \neq i]}$ excludes the anchor sample from the denominator. The loss can be interpreted as a classification objective in which the model must identify the correct positive view among all other candidate views in the batch.

Prominent contrastive frameworks such as SimCLR [5] therefore rely on large batch sizes, or alternatively memory banks, to provide a sufficient number of negative examples. They also depend strongly on stochastic data augmentation, since the choice of augmentations defines which image properties should be treated as invariant. While effective, this dependence on negative samples increases memory requirements and can make the learned representation sensitive to augmentation design. These limitations motivated the development of SSL methods that remove the need for explicit negative pairs altogether [6, 23].

Self-distillation

The self-distillation family, also referred to as non-contrastive or bootstrap learning, eliminates the requirement for explicit negative pairs entirely. These methods adopt a teacher-student architecture in which two networks, an online student encoder and a momentum teacher encoder, each process a different augmented view of the same image, and the training signal is defined as the distance between the student's projection and the teacher's projection. Because both networks could trivially minimize this objective by outputting a constant vector, structural asymmetries are introduced to prevent collapse. In BYOL [23], shown in Figure 2.4, the teacher's weights are not updated by gradient descent but instead by an exponential moving average (EMA) of the student's weights, producing a slowly evolving target that forces the student to continuously refine its representations to match a moving reference. SimSiam [6] achieves a related effect by applying a stop-gradient operation to the target branch, which prevents gradient flow through the teacher and breaks the symmetry that would otherwise lead to collapse. Self-distillation methods are computationally attractive because they remove the large-batch requirement of contrastive approaches, making them more practical in memory-constrained settings.

Redundancy reduction

The redundancy reduction family approaches the collapse problem from a statistical perspective, focusing on the cross-covariance structure of the learned representations rather than on direct sample-pair comparisons. Methods such as Barlow Twins [58] and VICReg [3] operate by driving the cross-correlation matrix of embeddings produced from two augmented views toward the identity. The Barlow Twins objective achieves this through two complementary terms, where the diagonal invariance term aligns the same feature dimension across views, enforcing view-invariant representations, while the off-diagonal redundancy term penalizes correlations between distinct feature dimensions, decorrelating the

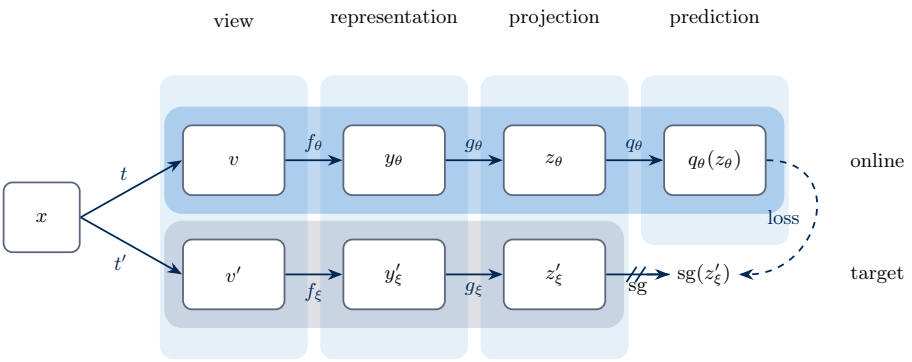


Figure 2.4: BYOL architecture

embedding space and preventing dimensional collapse. Dimensional collapse is a failure mode in which the model learns to encode information in only a low-dimensional subspace of the available embedding, discarding representational capacity. By explicitly penalizing such inter-neuron correlations, redundancy reduction methods force information to be distributed uniformly across all embedding dimensions. Unlike contrastive methods, this family requires no large batches or memory banks, as the objective is defined over feature statistics computed within the standard training batch.

Masked image modeling

Masked image modeling (MIM) departs from the invariance-based framework shared by the three preceding families and instead adopts a generative reconstruction objective. A proportion of the input tokens is randomly withheld from the encoder, and the model is trained to reconstruct the content of those masked tokens from the visible context. By learning to predict missing image regions, the encoder implicitly acquires an understanding of local textures, global scene structure, and the statistical regularities of the domain, without augmentation pipelines or pairwise objectives [25]. Because the model must predict fine-grained token statistics, MIM objectives naturally preserve high-frequency spatial detail that tends to be discarded by the global pooling operations used in contrastive frameworks, which has made MIM-pretrained models particularly strong on dense prediction tasks such as segmentation. Remote sensing adaptations of MIM, such as SatMAE [8], extend this masking strategy to the spectral dimension by treating inputs as three-dimensional volumes defined by height, width, and spectral band, forcing the model to reconstruct missing channels from the remaining ones and thus encouraging cross-spectral reasoning directly relevant to multi-sensor Earth observation data.

2.5 Remote sensing foundation models

The SSL methods described above form the backbone of a growing class of RSFMs, which apply these techniques at scale to satellite imagery. While early RSFMs focused primarily on optical data, more recent models explicitly incorporate SAR, either as the sole modality or as part of a multimodal pretraining scheme, making them more applicable to SAR-specific downstream tasks. Despite this progress, the field currently lacks standardized benchmarks, making it difficult to assess how well these models perform relative to one another and to established machine learning baselines [52]. For SAR in particular, most foundation models are pretrained predominantly on optical imagery, and it remains unclear whether models that explicitly incorporate SAR during pretraining offer a meaningful advantage for SAR-specific tasks, such as windthrow detection. The three models evaluated in this thesis are: SUMMIT, which applies a domain-specialized MIM objective exclusively to SAR data, Galileo, which combines two complementary MIM strategies across nine Earth observation modalities, and SkySense++, which sequences a contrastive pretraining stage with a semantic MIM stage to build progressively richer representations.

2.5.1 SUMMIT

SUMMIT is a foundation model designed exclusively for SAR imagery, targeting intrinsic modality properties that general-purpose vision models tend to overlook, including coherent speckle, electromagnetic scattering mechanisms, and the indirect relationship between backscatter intensity and scene semantics. The model adopts the MIM framework as its primary pretraining objective, training the encoder to reconstruct randomly masked patches from the remaining visible context. Critically, this reconstruction loss is augmented with three domain-specific auxiliary tasks, each formulated as an additional self-supervised signal. These are speckle denoising, edge feature extraction, and scattering point detection. Each auxiliary task requires the model to reason about a physically grounded property of SAR image, steering the representations toward features aligned with electromagnetic scattering theory rather than generic image statistics. The model is pretrained on the custom MuSID dataset containing over 560,000 multi-source SAR images. In downstream evaluations, SUMMIT achieves state-of-the-art performance across image classification, object detection, and instance segmentation, with improvements attributed to the domain-specific auxiliary objectives rather than scale alone. Pretrained weights and code are publicly available [13].

2.5.2 Galileo

Galileo is a multimodal Earth observation foundation model that processes a broad set of remote sensing modalities jointly across space and time, including multispectral optical imagery, SAR, elevation models, weather reanalysis, and land cover maps, through a shared ViT backbone receiving modality-specific patch embeddings. The pretraining objective combines two complementary MIM

strategies that address a fundamental tension in masked reconstruction, in which structured space-time masking, where entire spatial blocks or temporal frames are withheld, forces the model to acquire long-range semantic context while unstructured random token masking, applied independently per token, preserves fine-grained spatial detail by requiring reconstruction of locally missing patches. By combining both regimes within a single dual-objective framework, Galileo aims to produce representations that are simultaneously semantically rich and spatially faithful, a balance particularly important for dense prediction tasks such as pixel-level windthrow segmentation. Pretrained on a globally sampled dataset of over 127,000 multi-modal instances spanning nine modalities at 10 m resolution, Galileo achieves state-of-the-art or near-state-of-the-art performance across the GeoBench benchmark suite, covering classification, segmentation, SAR flood mapping, and multimodal crop classification. Code, pretrained weights, and pretraining data are openly released [49].

Additionally, Galileo can use time-series data. For a multi-temporal input, each spatial patch at each timestep is embedded independently into a token of dimension d. Temporal position is encoded by a learned month embedding that maps calendar month indices to dense vectors added to the patch tokens prior to self-attention, allowing the model to explicitly represent the acquisition schedule rather than relying on only positional order. The full set of spatio-temporal tokens is then processed by the shared transformer in a single forward pass, so that self-attention operates jointly across space and time at every layer. The resulting representations therefore reflect the full temporal context of the acquisition sequence rather than the spatial content of individual timesteps in isolation [49]. For dense prediction over multi-temporal inputs, the Galileo authors aggregate the per-timestep features produced by the encoder via simple temporal pooling rather than through a learned post-encoder fusion module, as demonstrated on the PASTIS image-time-series segmentation benchmark [49].

2.5.3 SkySense++

SkySense++ is a multimodal RSFM built around a factorised architecture that disentangles sensor-specific spatial encoding from cross-modal temporal fusion [54]. Modality-specific encoders process the high-resolution optical, multispectral, and SAR inputs independently for each timestep, and the resulting per-modality feature maps are concatenated into a joint multi-modal temporal representation. A date-specific temporal positional encoding indexed by actual acquisition dates is then added before a 24-layer multi-modal temporal fusion transformer attends jointly across modalities and time, producing a single spatio-temporal feature per spatial location. Because the spatial encoders operate independently per timestep and the fusion transformer receives the full set of dated feature tokens, the model natively supports inputs of varying temporal length and irregular acquisition schedules, making it sensitive to the actual ordering and timing of observations rather than their positional index alone. This built-in multi-temporal capacity is what the temporal extension in this work exploits, as described in Section 4.2.4, and is described in detail by Wu et al. [54]. In the implementation used here,

the date-specific positional encoding is not active, and temporal information is conveyed to the fusion transformer only by the fixed ordering of the input timesteps.

Wu et al. [54] pretrain SkySense++ in two sequential stages on a curated dataset of 27 million multimodal remote sensing images, as seen in Figure 2.5. The first stage applies multi-granularity contrastive learning, aligning representations of spatially co-located but sensor-diverse image patches at token, object, and image level, instilling broad cross-modal correspondence and geographic invariance. The second stage introduces masked semantic learning, in which the model is trained to predict discrete semantic tokens generated by a pretrained segmentation network rather than reconstructing raw pixel values, shifting the reconstruction target from low-level texture to semantic category labels. This combination produces representations rich enough to support few-shot generalisation to unseen task categories with minimal labelled examples, and SkySense++ demonstrates consistent improvements over prior state-of-the-art models across 12 Earth observation benchmarks spanning seven application domains, including forestry and disaster management [54].

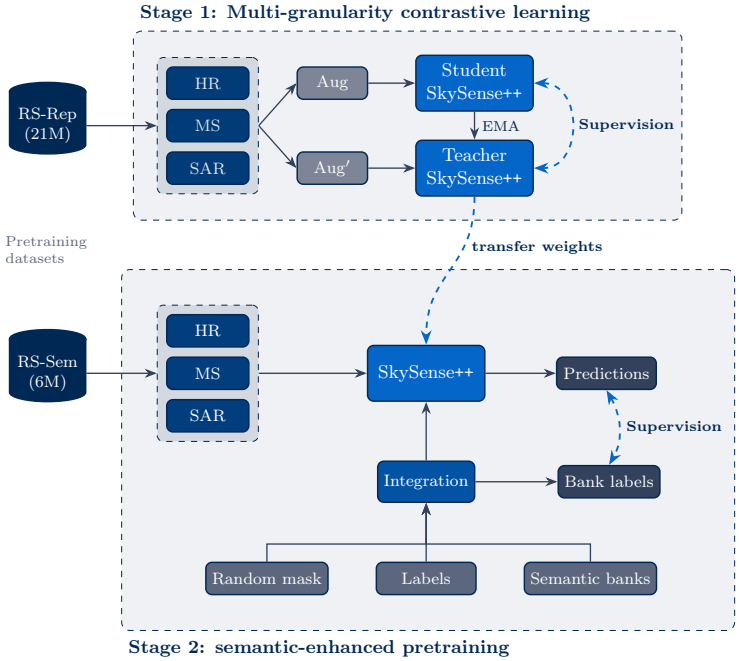


Figure 2.5: Overview of the pretraining of SkySense++

In this work the Sentinel-1 SAR encoder is used in isolation, exploiting the factorised design to operate on the two SAR polarisations (VV, VH) without requiring the optical modalities present during pretraining. Each acquisition is forwarded independently through the shared ViT-Large backbone, producing

per-timestep feature maps at layers 5, 11, 17, and 23. A dedicated fusion transformer then aggregates these per-timestep features across the temporal axis: for each spatial location, the features from all acquisitions are treated as a short sequence of tokens, a learned class token is added at the beginning of the sequence, and self-attention is applied to let every timestep attend to every other. The resulting class token is read out as a single fused feature vector that summarises the full temporal context at that location. This is specifically used in the temporal extension experiment, described in Section 4.2.4.

3 Data

The preprocessing pipeline transforms raw vector annotations, acquired from the different datasets, of storm-damaged forest into training-ready numerical arrays. The process begins with damage polygons sourced from each country dataset. These polygons are rasterized onto a regular spatial grid and used to generate spatially distinct tiles of 448×448 pixels. This tile size was chosen to match the input requirements of the foundation models used in this study and to allow the spatial reconstruction of forest patches of varying extent.

For each tile, five S1 SAR acquisitions acquired before the storm event and five acquired after are stored as individual GeoTIFF files in the two-channel format (VV, VH). All S1 data were retrieved from Google Earth Engine using the Copernicus S1 collection. Since VV and VH were the only available polarization channels for the chosen satellite S1, these two channels were used throughout the study. Alongside the SAR imagery, two binary mask files are generated per tile. The file `mask.tif` contains the raw windthrow reference map, where a pixel value of 1 indicates a confirmed windthrow event and 0 indicates no detected damage. The file `forest_mask.tif` defines the target segmentation mask $\mathbf{Y} \in \{0, 1\}^{448 \times 448}$, where 1 denotes damaged forest, 0 denotes undamaged forest and non-forest pixels are represented as `NaN`, so that only valid forest pixels are considered, in line with [39, 47]. For most datasets the forest mask was derived from the ESA WorldCover 10 m 2021 product using class 10 to identify forest. For the Swedish dataset, non-forest areas had already been marked as `NaN` by Skogsstyrelsen prior to publication and no additional forest masking was therefore required. The resulting per-country directory structure is as follows:

```
<country>/
└── tiles/
    └── <tile_id>/
        ├── pre_1.tif
        ├── pre_2.tif
        ├── pre_3.tif
        ├── pre_4.tif
        ├── pre_5.tif
        ├── post_1.tif
        ├── post_2.tif
        ├── post_3.tif
        ├── post_4.tif
        ├── post_5.tif
        ├── mask.tif
        └── forest_mask.tif
```

From this tile structure, two distinct datasets are produced, each pairing the same segmentation target $\mathbf{Y}$ with a different SAR input representation.

The first dataset uses a mean composite input. To reduce the effect of speckle noise inherent in SAR imagery, a pixel-wise mean is computed separately across the five pre-event acquisitions and the five post-event acquisitions. The four resulting channels are concatenated to form the composite representation

$$\mathbf{X}_{\text{comp}} \in \mathbb{R}^{448\times448\times4} = \left[\overline{VV}_{\text{pre}},\ \overline{VH}_{\text{pre}},\ \overline{VV}_{\text{post}},\ \overline{VH}_{\text{post}}\right].$$

Each sample in this dataset therefore consists of the pair $(\mathbf{X}_{\text{comp}},\ \mathbf{Y})$. Examples of how the data looks can be seen in Figure 3.2, with additional Sentinel-2 (RGB) and a post − pre SAR delta for clarity.

The second dataset retains the full temporal stack without temporal averaging, preserving all ten acquisitions in chronological order. With two polarization channels per acquisition, each tile is represented as

$$\mathbf{X}_{\text{ts}} \in \mathbb{R}^{448\times448\times10\times2},$$

where the temporal dimension contains the five pre-event acquisitions followed by the five post-event acquisitions, and the last dimension holds the (VV, VH) polarization pair. Each sample in this dataset consists of the pair $(\mathbf{X}_{\text{ts}},\ \mathbf{Y})$.

Due to the large class imbalance, seen in Table 3.1, filtered versions of the datasets were created in which only tiles containing more than a certain percentage of the event class were retained for training.

3.1 Storm Johannes, Sweden

Storm Johannes struck Sweden in late December 2025 and caused extensive forest damage [45]. The following description is how Skogsstyrelsen created the dataset recived from direct contact with representatives from Skogsstyrelsen.

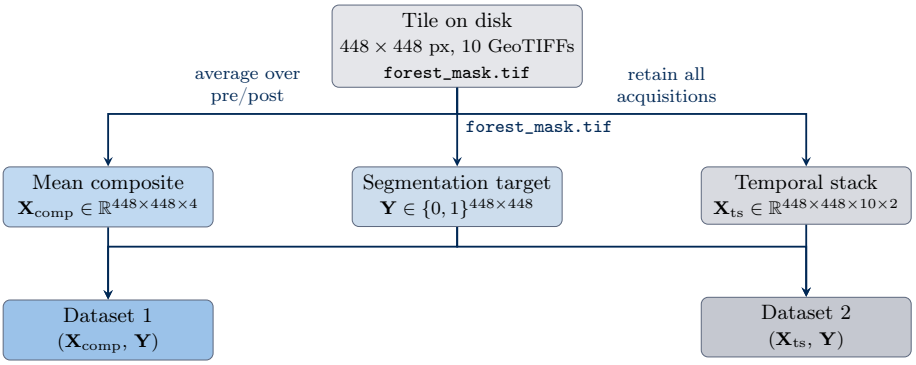


Figure 3.1: Overview of the preprocessing pipeline. Each tile on disk is used to produce two training datasets sharing the same segmentation target $\mathbf{Y}$ derived from the forest mask.

Table 3.1: Tile counts and class balance per dataset for the unfiltered base variant and the 5% event-fraction filtered variant. The base variant contains all forest tiles, while the 5% variant retains only tiles in which at least 5% of forest pixels are labelled as damaged. The damage fraction reports the mean percentage of damaged pixels among all valid forest pixels in each variant.

	Base		5% filtered	
Dataset	Tiles	Damage frac.	Tiles	Damage frac.
Germany	54	2.75%	9	6.98%
Ireland	1627	2.19%	215	9.85%
Russia	205	3.66%	41	9.69%
Sweden	2941	2.37%	99	16.07%

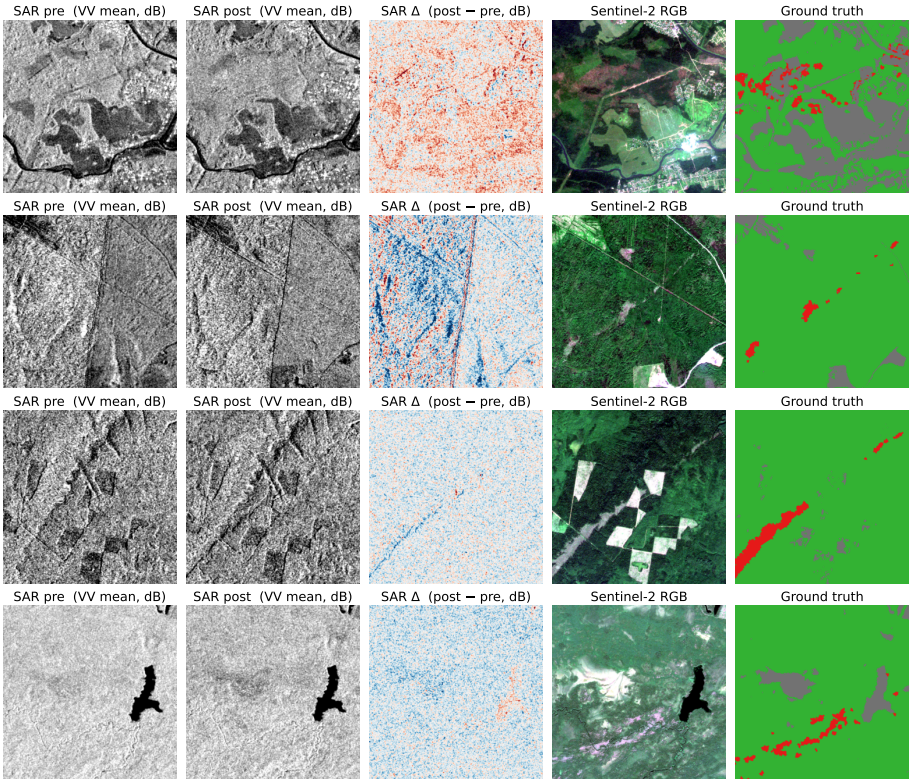


Figure 3.2: Pre- and post-storm SAR composites, their dB difference (post − pre), a near-storm Sentinel-2 RGB image with their respective forest and damage mask (red marks damaged forest, green undamaged forest and grey non-forest pixels) for four 448x448 tile examples from the Russian dataset.

The storm damage dataset consisted of a Sentinel-2-based (optical satelite) change image provided as a georeferenced raster product. The product was divided into index tiles covering approximately 50×50 km and was generated from satellite data acquired on 2025-02-12 and 2026-02-14. The raster had a spatial resolution of 10 m. It consisted of three bands: `NIR-first`, containing near-infrared reflectance from the first date; `Red-last`, containing red-band reflectance from the last date; and `Green-first`, containing green-band reflectance from the first date. When displayed in GIS software, these three bands were assigned to the red, green, and blue display channels, respectively.

To enhance visual interpretation, the original satellite values were contrast-stretched and rescaled independently in each band to a pixel value range of 1–255. The interpretation of the composite image was based on differences in snow reflectance between the two acquisition dates. Since snow has high reflectance in all included spectral bands, areas with snow present on both dates appear bright or nearly white in the resulting image. In contrast, areas where snow is visible only on the later date appear distinctly green. According to Skogsstyrelsen, such light green areas indicate locations where forest was likely blown down during the storm, thereby exposing snow-covered ground that had previously been obscured by standing trees.

A potential limitation of this approach is that areas harvested between the two acquisition dates may produce a similar visual response. However, Skogsstyrelsen stated that such areas were masked out as far as possible during preprocessing.

To derive training labels for the Sentinel-1-based model, the Sentinel-2 change image described above was processed using a custom Python script. The three-band composite encodes storm damage as elevated values in Band 2 (the green display channel) relative to Bands 1 and 3, which are assigned to the red and blue display channels respectively. To exploit this spectral signal and have a index, a pixel-wise damage index was computed as the difference between Band 2 and the maximum of the other two bands:

$$\text{index} = B_2 - \max(B_1,\, B_3) \tag{3.1}$$

Positive values of this index indicate pixels where the green channel dominates, which, following the interpretation provided by Skogsstyrelsen, corresponds to locations where snow-covered ground has become visible in the later acquisition that was previously obscured by standing forest canopy, suggesting that the canopy has been removed by the storm. A threshold of 50 was applied to binarise the index into a damage mask this will serve as our ground truth. This value was selected through visual inspection of the resulting mask against the original composite, where lower thresholds produced excessive false detections in non-damaged areas, while higher thresholds caused clear damage patches to be missed. The value of 50 represented the best balance between sensitivity and specificity under visual assessment. The resulting binary mask was subsequently filtered using 8-connected component labelling, and connected components smaller than 10 pixels were removed to suppress noise and eliminate small isolated detections. The final mask was used as the reference label for all training and evaluation of the Sweden dataset.

3.2 The European Russia Windthrow Database, Russia

The European Russia Windthrow Database [44] is a GIS dataset of stand-replacing windthrow events in the forest zone of European Russia. The original dataset was produced from the Landsat archive combined with the Global Forest Change and Eastern Europe's Forest Cover Change products. The annotation methodology was based on manual delineation of windthrow areas from satellite imagery, using Landsat time series in combination with the Global Forest Change and Eastern Europe's Forest Cover Change products to identify and map stand-replacing windthrow patches. It contains 102,747 elementary damaged areas, grouped into 700 windthrow events associated with 486 storm events [43]. For this study, the dataset was filtered to events dated 2016 or later. Events before 2016 were excluded to ensure availability of S1 SAR observations for all samples.

3.3 Private Forest Wind Damage Assessment, Ireland

The Ireland dataset was developed in response to widespread forest windthrow following Storm Darragh and Storm Éowyn during the winter of 2024/2025 [11, 30], and published by the Irish Department of Agriculture, Food and the Marine (DAFM). The dataset provides a spatial inventory of wind-damaged privately owned forest areas across Ireland at stand level, and is distributed primarily as an ESRI Shapefile polygon layer.

The mapping was derived from an Earth Observation workflow based primarily on high-resolution SkySat satellite imagery, supplemented by pre- and post-storm S2 and PlanetScope imagery. For a smaller number of sites, drone imagery was also acquired to improve interpretation in specific locations. DAFM reports that approximately 75% of the mapping relied on SkySat imagery (around 0.5 m spatial resolution), 20% on Sentinel-2 and PlanetScope imagery (around 10 m and 3 m, respectively), and 5% on drone imagery (around 0.2 m resolution). The imagery acquisition depended on cloud-free conditions, and the mapping exercise was conducted largely between early February and the beginning of April 2025. The final output is a polygon database of wind-damaged private forest areas greater than or equal to 0.1 ha [10].

3.4 Storm Xavier, Germany

For this dataset, windthrow reference polygons obtained by direct contact with the authors of Rüetschi et al. [39] was used. The polygons document damage caused by Storm Xavier, which struck northern Germany on 5 October 2017 with wind speeds of up to 117 km/h [39]. The reference data were produced by the department of forestry of Mecklenburg-Vorpommern through interpretation

of aerial imagery acquired between 16 and 29 October 2017 using a Nikon D300 camera at a flight altitude of approximately 800 m, yielding a spatial resolution of 12 cm. The survey covered approximately 125 km^2 of forested area east of Lüneburg. Each delineated polygon belonged to one of three damage severity classes based on the estimated proportion of windthrown stems: single windthrown trees (1–50%), single standing trees (51–90%), and areal windthrow (more than 90%). Only areas exceeding 0.5 ha were retained, resulting in 250 reference polygons [39]. In our experiments, these three classes were merged into a single binary class (windthrow vs. no windthrow) to match the label format used across datasets.

4 Methods

4.1 Baseline classifiers

Three baseline classifiers were selected to span the methodological range from rule-based detection to supervised deep learning, providing a structured reference against which the foundation models are evaluated. Intensity thresholding has previously been applied to storm damage detection in SAR imagery [39] and is retained here as a simple lower bound. Among the supervised classical methods compared by Tomppo et al. [47] on the same task, SVM achieved the highest accuracy and was therefore selected as a second baseline. A UPerNet segmentation network was additionally included as a supervised deep learning reference, representing the class of task-specific convolutional models that foundation models are intended to match or surpass without task-specific pretraining.

4.1.1 Threshold

As a classical baseline, a threshold-based wind-damage detector was implemented inspired by Rüetschi et al. [39] and applied across all datasets. Training and testing were performed independently for each dataset.

The calculations were done on the mean images $\mathbf{X}_{\text{comp}}$. The reference masks were preprocessed by removing all connected damage components smaller than 10 pixels (8-neighbour connectivity), consistent with the minimum detectable group size swept during threshold optimisation. Additionally, any tile containing no remaining event pixels after this filtering was excluded from both training and evaluation, ensuring that only tiles with confirmed windthrow events contributed to the sweep and metrics. The same component-size filtering was applied consistently in both training and testing. The windthrow index was then computed as:

$$\text{WI} = \left(\overline{VV}_{\text{post}} - \overline{VV}_{\text{pre}}\right) + \left(\overline{VH}_{\text{post}} - \overline{VH}_{\text{pre}}\right).$$

A pixel was classified as potential wind damage when $\text{WI} \geq t$. On the training split, a range of thresholds t and an additional 8-neighbour connected-component filter (minimum group size) were swept to convert potential detections into final damage predictions.

For the thresholds, the bound spanned from the 2nd to the 100th percentile in 50 evenly spaced steps. Group sizes were tested across the set $\{10, 15, 20, 30, 50, 75, 100, 150, 200, 300\}$ to cover a wide and realistic range. For each dataset, the best training configuration was selected via a joint threshold and group-size sweep.

4.1.2 Support vector machine classifier

A support vector machine (SVM) classifier was developed to detect windthrow events from S1 SAR imagery, replicating Tomppo et al. [47] but adapted to match the available datasets. Tomppo et al. [47] use stands defined by the Finnish forest department, as this data was unavailable, patches of 16×16 pixels were chosen instead, which closely matched the mean size of damaged stands reported by Tomppo et al. [47] while remaining compatible with the available data. A patch was used if a majority of it was labelled as forest, and it was labelled as windthrow if more than 50% of its reference pixels were positive. A cap of 200 000 stands was set because of compute constraints, which only affected the Swedish dataset since the other ones were smaller than that limit. Following Tomppo et al. [47], a radial basis function (RBF) kernel was used, allowing non-linear separations as described in Section 2.3.1. Prior to training, all features were standardised to zero mean and unit variance [41]. For the kernel, the bandwidth parameter γ was swept over $\{0.001, 0.01, 0.1, 1.0, 10.0\}$ following Tomppo et al. [47], with model selection performed on an internal 80/20 train–validation split retaining the value that maximised validation F_1.

4.1.3 CNN-UPerNet segmentation

As an intermediate step between the SVM baseline and the foundation models, a CNN-UPerNet segmentation model was trained from scratch on the same standardised SAR composites $\mathbf{X}_{\text{comp}}$. This model serves as the main deep-learning baseline prior to the foundation model experiments.

CNN backbone and UPerNet decoder

The model consists of a lightweight hierarchical convolutional backbone followed by a UPerNet decoder [55]. The backbone comprises a sequence of D convolutional stages, each containing two convolution–batch-normalisation–ReLU blocks; all but the final stage are followed by 2×2 max-pooling, yielding a four-level feature hierarchy $\{f_s\}_{s=0}^{3}$ with base channel width C_0 doubling at each stage up to a maximum of 512. The feature maps are passed directly to the UPerNet decode head described in Section 2.3.3, which fuses multi-scale context through its pyramid pooling module and feature pyramid pathway before projecting to two output channels. The logits are upsampled bilinearly to the original 448×448 tile resolution, yielding dense pixel-wise class scores. After softmax normalisation, the damaged-class probability map p_{ij}^{damaged} is thresholded at a

decision threshold t to obtain the final binary prediction mask:

$$\hat{Y}_{ij} = \begin{cases} 1, & p_{ij}^{\text{damaged}} \geq t, \\ 0, & \text{otherwise.} \end{cases}$$

The threshold t is selected on the validation set as described in Section 4.1.3.

Training setup

For each dataset variant, tiles are partitioned into training, validation, and test subsets using stratified tile-level splits with proportions 0.8/0.1/0.1. Stratification is based on tile labels so that the relative frequency of damaged and undamaged samples is preserved across splits. The optimization objective is a compound segmentation loss combining Dice loss and focal loss over valid pixels, as seen in Equation 2.7. This loss was chosen because windthrow damage is spatially sparse and strongly imbalanced relative to healthy forest, making plain cross-entropy less suitable. Weighted cross-entropy loss (Equation 2.3) and Focal Tversky loss (Equation 2.8) were also tested but did not yield results as strong as the compound focal-Dice loss.

Hyperparameter search strategy

The CNN-UPerNet baseline was evaluated on the same eight dataset variants used throughout the standardised pipeline: Ireland, Sweden, Russia, and Germany, together with their corresponding 5% filtered versions.

The search space included both architectural and optimisation parameters. Candidate learning rates η were drawn from $\{10^{-5}, 3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$; the base channel width C_0 from $\{24, 32, 48\}$; the backbone depth D from $\{4, 5\}$; the decoder channel width C_{dec} from $\{128, 256, 384\}$; the weight decay λ_{wd} from $\{10^{-5}, 10^{-4}, 5 \times 10^{-4}\}$.

Because the full Cartesian product of this space is large, a fixed sample of 24 hyperparameter configurations was drawn from the candidate grid using a fixed random seed. The same 24 configurations were then evaluated on each of the eight dataset variants, yielding a total of 192 training runs. This design keeps the comparison across study areas controlled, since differences in performance reflect dataset effects rather than differences in the sampled search spaces.

Post-processing and model selection

Model selection followed a validation-only protocol. After training, each candidate run was evaluated over a threshold sweep from $t = 0.00$ to $t = 1.00$ in steps of 0.01, and the threshold $t^\star$ that maximised the validation damaged-class F_1 score was retained. The final model for each dataset variant was then chosen as the run with the highest validation F_1 after threshold calibration. Test-set metrics were reported once for that selected run using the threshold $t^\star$ determined on the validation split, ensuring that the test set was used neither for checkpoint selection nor for threshold tuning. The primary ranking metric throughout is

the damaged-class F_1 score, while damaged-class IoU and mean IoU are reported as complementary measures of spatial segmentation quality.

4.2 Foundation models

4.2.1 Shared downstream pipeline

Both foundation models evaluated in this work, SkySense++ and Galileo, are adapted to the windthrow segmentation task through a downstream pipeline that fixes the encoder pretraining, the segmentation decoder, the input data, and the training protocol. SUMMIT [13], a third foundation model considered for this evaluation, was excluded from the main comparison following a preliminary assessment in which its representations failed to transfer meaningfully to the target domain, yielding segmentation outputs that did not improve upon the rule-based threshold baseline. Given the associated computational cost of full fine-tuning runs, exhaustive hyperparameter exploration for all three foundation models was not feasible, and resources were therefore concentrated on the two candidates whose preliminary results indicated a genuine capacity for learning discriminative features on this task. Each model serves as a backbone whose representations are decoded by a UPerNet head identical in architecture to the one described in Section 2.3.3. This design isolates the contribution of pretraining, where differences in segmentation performance between the foundation models and the CNN baseline can be attributed to the quality of the pretrained representations rather than to differing decoder capacities or task formulations.

The two models differ in one aspect that the rest of this chapter assumes: the post-encoder stage that combines pre-event and post-event feature representations into a single tensor before decoding. SkySense++ ships as an integrated factorised pipeline in which the temporal-fusion module, a 24-layer `FusionTransformer` sitting inside the `FusionMultiLevelNeck`, is an integral part of the model [54]. Galileo ships as a pretrained backbone only, with its authors evaluating downstream performance via linear probing, kNN probing, or plain fine-tuning, and only aggregating per-timestep features by simple temporal pooling [49]. For Galileo, the post-encoder fusion mechanism is therefore a design choice made in this work, where the standard concat-diff fusion from the SAR change-detection literature [4, 9] is adopted in the bi-temporal experiment (Section 4.2.3), and PASTIS-style temporal mean-pooling is adopted in the temporal experiment (Section 4.2.4). Each model is therefore evaluated using the configuration that most closely matches the setup made available by its original authors.

Backbone-specific feature extraction

The two backbones differ in how they expose multi-scale spatial feature maps to the decoder. SkySense++ was designed for dense prediction tasks with its ViT-Large backbone, with 1024-dimensional embeddings and 24 transformer layers, natively exposes intermediate feature maps at layer indices 5, 11, 17, and 23,

which form the four-level pyramid consumed by the `FusionMultiLevelNeck`. Galileo was designed as a general-purpose multimodal encoder and does not natively expose spatial feature maps; its transformer returns a flat sequence of patch tokens. To adapt it for dense prediction, forward hooks are registered on four selected transformer blocks at indices 2, 5, 8, and 11, and the captured token sequences are reshaped from $(N_{\text{sub}},\ n_p \cdot n_p,\ d)$ into 2-D spatial grids of shape $(N_{\text{sub}},\ n_p,\ n_p,\ d)$, projected to a common channel width by 1×1 convolutions, and passed to the shared UPerNet decoder. In both cases the decoder follows the architecture described in Section 2.3.3, skip connections forming a feature pyramid network, and a final bilinear upsampling to the original tile resolution.

4.2.2 Bi-temporal fusion pipeline

Windthrow detection is inherently a change-detection problem, where the discriminative signal lies in the difference between pre-event and post-event backscatter rather than in either acquisition in isolation. The bi-temporal experiment provides each model with the four-channel mean composite $\mathbf{X}_{\text{comp}}$ and processes the pre-event and post-event tiles through the shared backbone, with the per-acquisition feature representations combined by each model's post-encoder fusion mechanism.

For Galileo, the four-channel composite is split into a pre-event tensor $\mathbf{X}_{\text{pre}} \in \mathbb{R}^{H\times W\times 2}$ and a post-event tensor $\mathbf{X}_{\text{post}} \in \mathbb{R}^{H\times W\times 2}$, where the two channels are (VV, VH). Each acquisition is encoded independently with the temporal dimension set to one. Following the hook-based multi-scale extraction described in Section 4.2.1, intermediate representations are captured at transformer blocks 2, 5, 8, and 11, producing a four-level feature pyramid per acquisition:

$$\mathbf{f}^{(l)}_{\text{pre}} = E^{(l)}_{\theta}(\mathbf{X}_{\text{pre}}), \qquad \mathbf{f}^{(l)}_{\text{post}} = E^{(l)}_{\theta}(\mathbf{X}_{\text{post}}), \quad l \in \{2, 5, 8, 11\}, \tag{4.1}$$

where E_θ denotes the frozen encoder. At each pyramid level, the two feature maps are fused by the concat-diff operation of the SAR change-detection literature [4, 9]:

$$\mathbf{f}^{(l)}_{\text{fused}} = \left[\, |\mathbf{f}^{(l)}_{\text{post}} - \mathbf{f}^{(l)}_{\text{pre}}|,\ \mathbf{f}^{(l)}_{\text{post}},\ \mathbf{f}^{(l)}_{\text{pre}} \,\right]. \tag{4.2}$$

If each feature map has C channels, the fused representation has $3C$ channels. The absolute difference term makes the temporal change signal explicit to the decoder, while the retained pre- and post-event maps preserve contextual information about each acquisition. A dedicated 1×1 convolution per pyramid level then projects each $3C$-channel fused map back to the common channel width, after which the four projected maps are passed as the feature pyramid to the UPerNet decode head. This is visualised in Figure 4.1.

For SkySense++, each acquisition is processed independently by the shared S1 ViT-Large backbone as seen in Figure 4.2, and the resulting two-step feature stack is passed to the `FusionMultiLevelNeck`. For every spatial location, the per-acquisition features are reshaped into a length-two token sequence and fed to the 24-layer `FusionTransformer`, which prepends a learned class token, attends across the temporal axis, and outputs a single fused feature vector per spatial

location from the final class-token embedding. No explicit concat-diff projection is applied since the change-aware combination is learned implicitly by the fusion transformer. The same `FusionTransformer` is used in the temporal extension of Section 4.2.4, with only the input sequence length changing between the two experiments.

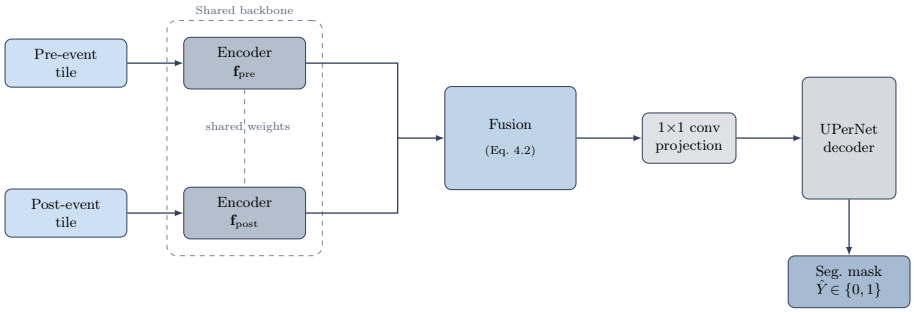


Figure 4.1: Visualization of the Siamese bi-temporal pipeline for Galileo.

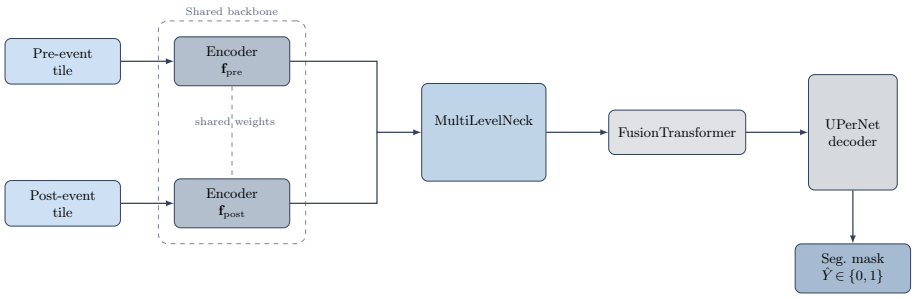


Figure 4.2: Visualization of the bi-temporal pipeline for SkySense++.

4.2.3 Primary bi-temporal experiment

The primary experiment establishes the baseline performance reference for both SkySense++ and Galileo under the bi-temporal pipeline, using the mean images $\mathbf{X}_{\text{comp}}$ described in Section 4.2.1. The central question it addresses is how much of the pretrained backbone needs to be adapted before the downstream windthrow task can be solved, and the experiment therefore varies the fine-tuning regime across two configurations, `lora` and `full`, alongside the loss function and learning rate.

In the `lora` configuration, the backbone remains frozen but Low-Rank Adaptation (LoRA) adapters are inserted into the transformer layers, so that only the LoRA parameters, the post-encoder module, and the decode head are updated. This enables lightweight task-specific adaptation while leaving the pretrained

backbone weights unchanged. In the `full` configuration, all backbone parameters are unfrozen and trained jointly with the post-encoder module and the decode head. Layer-wise learning rate decay is applied so that shallower layers, which tend to encode more general features, are updated more conservatively than deeper layers.

Loss functions

Five loss functions were evaluated for their ability to handle the class imbalance inherent to binary windthrow segmentation: weighted cross-entropy $\mathcal{L}_{\mathrm{CE}}$ (Equation 2.3), focal loss $\mathcal{L}_{\mathrm{Focal}}$ (Equation 2.5), Dice loss $\mathcal{L}_{\mathrm{Dice}}$ (Equation 2.6), the compound focal Dice loss $\mathcal{L}_{\mathrm{FocalDice}}$ (Equation 2.7), and the focal Tversky loss $\mathcal{L}_{\mathrm{FT}}$ (Equation 2.8). Initial experiments across both models and dataset variants indicated that standalone focal loss, Dice loss and focal Tversky loss did not produce competitive validation F_1-scores for the damaged class, and were therefore excluded from the full sweep. Three loss functions were carried forward: weighted cross-entropy $\mathcal{L}_{\mathrm{CE}}$ with per-dataset class weights computed using Equation 2.4; a compound focal-Dice loss with $\alpha = 0.25$ and $\gamma = 1.0$ (FD-γ1); and the same compound form with $\gamma = 2.0$ (FD-γ2).

Learning rate

The learning rate η was varied across a logarithmically spaced grid to determine a suitable range for fine-tuning the foundation model backbone. Tested values spanned from 10^{-5} to 10^{-3}, covering both conservative settings appropriate for a pretrained backbone and more aggressive values intended to adapt the decoder head more rapidly. All experiments used the AdamW optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a head learning-rate multiplier of $\times 10$ over the backbone rate, and layer-wise learning-rate decay $\gamma = 0.9$. Based on these experiments, backbone learning rates were treated as dataset-dependent in the final sweep, where Ireland and Sweden were assigned a conservative pair $\eta \in \{10^{-5}, 3{\times}10^{-5}\}$, while Russia and Germany were assigned $\eta \in \{3{\times}10^{-5}, 10^{-4}\}$.

Experimental grid and model selection

The findings from the loss and learning-rate investigations are combined into a single factorial sweep. The varied factors are: fine-tuning regime {`lora`, `full`}; loss function {CE, FD-γ1 (α=0.25, γ=1.0), FD-γ2 (α=0.25, γ=2.0)}; learning rate $\mathrm{LR}_{\mathrm{IE/SE}}$ $\{10^{-5}, 3{\times}10^{-5}\}$; and learning rate $\mathrm{LR}_{\mathrm{RU/DE}}$ $\{3{\times}10^{-5}, 10^{-4}\}$.

Training proceeds for a maximum of 50 epochs with a uniform early-stopping protocol where the training is stopped if the validation damaged-class F_1 has not improved by at least 0.5 absolute points over any 10 consecutive epochs, subject to a minimum of 15 epochs before stopping can trigger. This minimum prevents premature termination on smaller datasets such as Russia and Germany, where the model may require some epochs before producing any windthrow predictions at all. After training, a threshold sweep over $t \in [0.00, 1.00]$ in steps of 0.01

is performed on the validation split, and the value $t^\star$ maximising validation F_1^{damaged} is retained.

4.2.4 Temporal extension

The temporal experiments investigate whether the downstream segmentation model benefits from operating on the full ordered SAR acquisition sequence rather than on the bi-temporal mean composite $\mathbf{X}_{\text{comp}}$. Both Galileo and SkySense++ natively support multi-temporal inputs through their internal temporal attention mechanisms, as described in Section 2.5.2 and Section 2.5.3 respectively, and the temporal extension exploits this capacity by providing the full ten-acquisition time series $\mathbf{X}_{\text{ts}}$ at temporal depth $T = 10$ directly to the backbone rather than a pre-averaged composite.

Galileo temporal

The bi-temporal Galileo pipeline encodes the pre-event and post-event composites as two independent forward passes through the shared backbone, and only afterwards merges the resulting features via the change-aware fusion of Equation 4.2. Temporal context within each composite is therefore invisible to the encoder, which means the backbone has no means of attending across acquisitions. The temporal variant removes this limitation by packing all $T = 10$ acquisitions into a single forward pass, allowing Galileo's joint spatio-temporal attention to model the full pre-to-post transition directly inside the encoder.

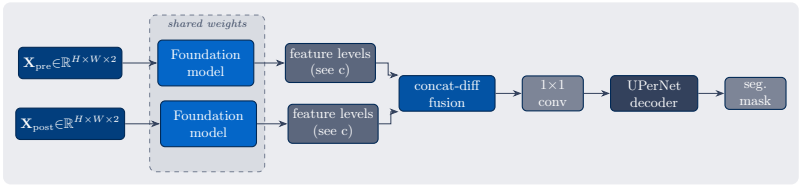
Each tile is assembled into Galileo's native multimodal input format: the (VV, VH) values for every timestep t are written into Galileo's spatio-temporal input tensor $\mathbf{s}_t$ at the Sentinel-1 band positions, and a corresponding month-index tensor of shape (N_{sub}, T) provides the calendar month for each acquisition. These month indices are consumed by Galileo's learned temporal position embedding, enabling the backbone to differentiate pre-event from post-event timesteps when attending across the sequence. The tile batch is then passed to Galileo in a single forward pass, with joint self-attention operating over the full set of spatio-temporal tokens at every transformer layer.

The post-encoder pipeline follows the temporal-aggregation strategy used by the Galileo authors for image-time-series segmentation on PASTIS [49]. The per-timestep feature maps captured at hook indices 2, 5, 8, and 11, each of shape $(N_{\text{sub}}, n_p, n_p, T, d)$, are reduced along the temporal axis by an arithmetic mean:

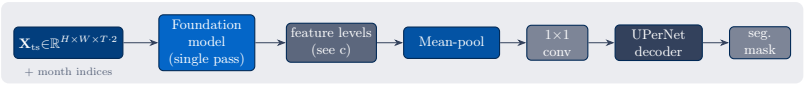
$$\mathbf{h}_{\text{pooled}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_t \ \in \ \mathbb{R}^{N_{\text{sub}} \times n_p \times n_p \times d}, \tag{4.3}$$

producing a single per-spatial-location feature vector at each pyramid level. A 1×1 convolution then projects each pooled level to the common channel width consumed by the UPerNet decoder of Section 4.2.1. Pre/post structure is therefore not imposed as an explicit post-encoder operation, it is instead left to the joint-attention encoder, which has access to the calendar month of

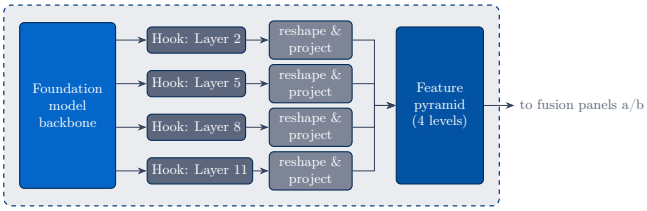
every acquisition through Galileo's temporal position embedding. The contrast against the bi-temporal Galileo pipeline of Section 4.2.2 is clear. In the bi-temporal pipeline the change signal is exposed explicitly through concat-diff fusion applied after two independent encoder passes, whereas in the temporal pipeline the change signal is exposed implicitly through joint attention across the full acquisition sequence, with no post-encoder fusion module beyond temporal mean-pooling. The full pipeline is visualised in Figure 4.3.



(a) Bi-temporal pipeline



(b) Temporal pipeline



(c) Feature extraction detail shared by both pipelines

Figure 4.3: Overview of the bi-temporal and temporal Galileo model pipelines and the shared feature extraction mechanism.

SkySense++ temporal

The temporal SkySense++ pipeline follows the same overall structure as the Galileo temporal pipeline described in Section 4.2.4, with all $T = 10$ acquisitions are packed into a single forward pass, the first five forming the pre-event group and the last five the post-event group, and the resulting feature maps are passed to the shared UPerNet decoder. The implementation differs from Galileo in two aspects.

First, SkySense++ keeps spatial encoding and temporal fusion architecturally separate, as described in Section 2.5.3. Second, no acquisition-date or month-index embedding is provided. Where Galileo receives an explicit month index tensor that its temporal position embedding uses to exploit the known acquisition schedule, SkySense++ conveys temporal information solely through the fixed

chronological ordering of the timesteps. As a consequence, the fused feature maps are passed directly to the UPerNet decoder without a subsequent change-aware concatenation step, where the temporal change information is encoded implicitly by the fusion transformer rather than made explicit by the differencing operation of Equation 4.2.

4.2.5 Controlled sensitivity studies

Beyond the primary experiment and temporal encoding sweeps, four controlled sensitivity studies are conducted to characterise specific aspects of the foundation model pipelines. Each study holds all hyperparameters fixed to the best identified in the primary experiment and varies a single factor of interest, allowing individual design choices to be evaluated in isolation.

Label efficiency

The label-efficiency study examines how performance scales with the fraction of available training data, generating learning curves that reveal whether pretrained representations confer a practical advantage in low-data regimes. Four training fractions are evaluated: 1%, 10%, 50%, and 100% of the training split, with sampling stratified by tile label to preserve the proportion of damaged tiles at each fraction. SkySense++, Galileo and the CNN baseline are compared, providing a direct measure of relative label efficiency across model families. Ireland was chosen as a large operational dataset, while Russia was chosen as a smaller and cleaner stand-replacing windthrow dataset.

Temporal depth

The temporal depth study measures how segmentation performance varies as the number of SAR acquisitions provided to the backbone increases from 2 to 10. Temporal depths $T \in \{2, 4, 6, 8, 10\}$ are evaluated, with each stack composed of $T/2$ pre-event acquisitions followed by $T/2$ post-event acquisitions in chronological order. Both SkySense++ and Galileo are evaluated under full fine-tuning across all four datasets.

Cross-dataset generalisation

The cross-dataset generalisation study assesses whether the models can transfer to unseen geographic domains under balanced multi-country training. Two related protocols are evaluated. In the full hold-out setting, the target dataset is excluded entirely from training, and the model is trained only on a balanced pool drawn from the remaining datasets. In the partial hold-out setting, the target dataset contributes training tiles to the balanced pool, but validation and testing remain target-dataset only. In both cases, balancing is enforced so that no single dataset dominates the training distribution. Germany is excluded from this study because its training set is much smaller than the others, and including it would force the balanced pool to be limited by a very low tile count.

Augmentation strategy

The augmentation strategy study evaluates whether SAR-compatible augmentations provide a consistent benefit across model families and geographic domains. Three conditions are compared: no augmentation, geometric augmentation only (horizontal flip with $p = 0.5$ and 180° rotation with $p = 0.5$), and geometric augmentation combined with intensity jitter (additive Gaussian noise applied independently to VV and VH channels with $\sigma \approx 0.1$ dB, simulating inter-acquisition calibration differences and speckle variation). Vertical flip and 90°/270° rotation are excluded throughout, as these transformations are physically invalid for Sentinel-1 orbit imagery.

4.2.6 Stand-level evaluation

All primary results in this thesis are reported at the pixel level. To connect those results to the operational unit that forestry managers act on, a complementary stand-level evaluation is conducted following the object-based protocol of Rüetschi et al. [39], whose formal definitions are given in Section 2.2.5. The evaluation reuses the trained checkpoints from the primary bi-temporal experiments without any additional training.

For each model and dataset combination, the best checkpoint selected by validation F_1 in the primary experiment is reloaded together with its locked probability threshold. A single forward pass is performed on the test set, and per-tile binary prediction masks are obtained by applying that threshold to the class-one softmax probability. Reference stands are defined as 8-connected components of the binary damage reference mask, retaining only components of at least 50 pixels (approximately 0.5 ha at 10 m resolution), matching the minimum stand area in Rüetschi et al. Predicted stands are defined as 8-connected components of the binary prediction mask, intersected with the valid forest mask, and any component with pixel count $\leq n$ is discarded. A reference stand is counted as detected (true positive toward PA) if it overlaps any surviving predicted stand, and a predicted stand is counted as correct (true positive toward UA) if it overlaps any reference stand.

The minimum predicted-object size n is calibrated per model and dataset pair on the validation split by sweeping $n \in \{5, 10, 15, \ldots, 200\}$ pixels and selecting the value that maximises the object-level F_1 score. The locked n^* is then applied once to the test set, and the resulting PA, UA, and F_1^{obj} are recorded alongside the pixel-level F_1 from the same checkpoint. For a point of comparison, a reference result using WI, as described in Rüetschi et al. [39], is also made, where a is swept over $\{2.00, 2.05, \ldots, 4.00\}$ and the most optimal validation a^* is selected.

4.3 Computational environment

All experiments in this thesis were executed on H100 GPUs on the MareNostrum 5 supercomputer at the Barcelona Supercomputing Center (BSC-CNS).

MareNostrum 5 is a EuroHPC supercomputer hosted by BSC-CNS [2]. The use of a high-performance computing environment was necessary due to the volume of S1 data, the number of dataset variants, size of the models and the large number of model configurations evaluated throughout the study.

Jobs were submitted as separate batch runs on the cluster. For the baseline experiments, these jobs included threshold and connected-component sweeps as well as SVM training and evaluation across the different datasets and filtered variants. For the deep learning experiments, the cluster was used to train and evaluate the CNN-UPerNet baseline and the foundation model pipelines based on Galileo and SkySense++. The computational resources were particularly important for the factorial sweeps over adaptation strategy, loss function, learning rate, dataset variant, and temporal input configuration described in the previous sections.

The implementations were based on a shared Python workflow. The classical baselines used the same preprocessed arrays, train/validation/test splits, and evaluation scripts as the neural network models. The CNN and foundation model experiments were implemented using Python based deep learning frameworks, including the framework code required by the respective foundation models. In particular, the SkySense++ experiments relied on its MMSegmentation-based implementation. Across all experiments, model outputs, checkpoints, validation metrics, and prediction masks were stored and evaluated using the same metric pipeline to ensure comparability between methods.

MareNostrum 5 enabled the scale of the experiments, but did not define the methodological differences between the evaluated approaches.

5 Results

5.1 Baseline performance

This chapter reports test-set performance for all baseline classifiers, both foundation models, and the four controlled sensitivity studies introduced in Section 4.2.5. The primary ranking metric throughout is the damaged-class F_1 score; damaged-class IoU and mIoU are reported as complementary measures of spatial segmentation quality, and damaged-class precision and recall are reported as P/R to expose the operating point selected by the validation damaged-class F_1 scpre. All numbers, if not stated otherwise, are pixel-level metrics computed on the held-out test split, evaluated only over valid forest pixels as defined in Section 3.

5.1.1 Threshold

Table 5.1 reports the windthrow-index threshold detector with the threshold and minimum-group-size jointly tuned on the training split as described in Section 4.1.1. Damaged-class F_1 stays below 7% on every unfiltered dataset and rises to between 11% and 22% on the 5% filtered variants, where the elevated event prior makes near-saturated recall easier to achieve at the cost of very low precision. The 5% variants for Russia, Ireland, and Germany reach a recall of 100.0% paired with single-digit precision, consistent with the detector firing on essentially every forest pixel that exceeds the swept intensity threshold once the prevalence of windthrow has been raised by tile filtering. Sweden is the one dataset where the unfiltered detector exhibits a balanced but very weak operating point (5.3% precision, 10.0% recall), with the threshold sweep unable to find a setting that recovers more than a small fraction of the labelled positives.

5.1.2 Support vector machine

Table 5.2 reports the patch-level RBF-SVM classifier of Section 4.1.2. Compared with the threshold baseline, the SVM generally improves damaged-class F_1, although the magnitude varies substantially across dataset variants, where the gain is negligible on unfiltered Sweden, increasing only from 6.9% to 7.4%, while Russia and Russia 5% reach 30.4% and 39.7%, respectively. The main exception is Germany 5%, where performance collapses to 0.8% F_1, consistent with the

Table 5.1: Threshold results on the held-out test split for all eight dataset variants. F_1^{dmg}, IoU_{dmg}, and mIoU are pixel-level metrics evaluated over valid forest pixels only; P/R denotes damaged-class precision and recall at the threshold selected by validation F_1.

Dataset variant	F_1^{dmg}	$\mathbf{IoU_{dmg}}$	**mIoU**	**P/R**
Sweden	6.9%	3.6%	**49.7%**	5.3% / 10.0%
Sweden 5%	**22.1%**	**12.4%**	6.6%	**12.4%** / 99.2%
Russia	5.5%	2.8%	45.7%	3.6% / 12.0%
Russia 5%	11.1%	5.9%	3.0%	5.9% / **100.0%**
Ireland	5.3%	2.7%	1.5%	2.7% / **100.0%**
Ireland 5%	16.4%	8.9%	4.5%	8.9% / **100.0%**
Germany	5.0%	2.6%	45.0%	3.3% / 10.6%
Germany 5%	11.4%	6.0%	3.0%	6.0% / **100.0%**

dataset retaining only nine tiles after filtering. The precision/recall balance shifts towards recall on nearly all datasets, indicating that the kernel-bandwidth sweep selected operating points that prefer to over-flag damage rather than miss it.

Table 5.2: SVM patch-level classifier results on the held-out test split.

Dataset variant	F_1^{dmg}	$\mathbf{IoU_{dmg}}$	**mIoU**	**P/R**
Sweden	7.4%	3.8%	26.1%	4.1% / 39.9%
Sweden 5%	22.3%	12.6%	26.9%	13.5% / 65.1%
Russia	30.4%	17.9%	54.7%	19.6% / 67.1%
Russia 5%	**39.7%**	**24.8%**	**55.2%**	26.7% / **77.1%**
Ireland	16.3%	8.9%	45.9%	9.4% / 59.9%
Ireland 5%	37.7%	23.2%	53.6%	**29.6%** / 51.9%
Germany	15.0%	8.1%	43.5%	8.6% / 59.5%
Germany 5%	0.8%	0.4%	43.3%	0.7% / 0.9%

5.1.3 CNN-UPerNet classifier

Table 5.3 reports the CNN-UPerNet baseline trained from scratch under the hyperparameter sweep of Section 4.1.3. The supervised CNN clearly separates from both classical baselines on every dataset where windthrow is geometrically well-defined, reaching 62.0% damaged-class F_1 on the unfiltered Russia variant and 41.3% on Germany. Performance on Sweden remains modest at 21.2% F_1 while the Germany 5% row collapses to 9.3% F_1 score.

Table 5.3: CNN-UPerNet baseline test-set performance for the selected configuration on each dataset variant. P/R denotes damaged-class precision / recall.

Dataset variant	F_1^{dmg}	$\mathbf{IoU_{dmg}}$	**mIoU**	**P/R**
Sweden	21.2%	11.8%	54.9%	25.8% / 17.9%
Sweden 5%	24.7%	14.1%	35.8%	16.2% / 52.4%
Russia	**62.0%**	**44.9%**	**71.3%**	**59.0%** / 65.3%
Russia 5%	44.9%	29.0%	58.2%	31.7% / **77.3%**
Ireland	42.0%	26.6%	61.6%	39.8% / 44.4%
Ireland 5%	41.6%	26.3%	56.0%	39.7% / 43.7%
Germany	41.3%	26.0%	61.2%	43.2% / 39.5%
Germany 5%	9.3%	4.9%	48.6%	15.7% / 6.6%

5.2 Foundation model results

Tables 5.4 and 5.5 report test-set performance for the selected primary configuration for each foundation model and dataset, as defined in Section 4.2.3.

5.2.1 Primary bi-temporal experiment

Table 5.4 reports Galileo under the Siamese bi-temporal pipeline. The qualitative ordering across datasets matches the CNN baseline, with Russia best (59.0% damaged-class F_1), Ireland and Germany in the middle, and Sweden weakest. On the unfiltered datasets, Galileo lands within a couple of percentage points of the CNN on Sweden, Russia, and Ireland, and roughly 10 points below the CNN on Germany.

Table 5.5 reports SkySense++ under the bi-temporal pipeline. The qualitative ordering across datasets diverges from both Galileo and the CNN. Russia is again strongest, reaching 90.1% damaged-class F_1 with a near-symmetric 89.8%/90.4% precision/recall pair, approximately 28 percentage points above the CNN and 31 percentage points above Galileo. Ireland follows at 51.6%, above both the CNN and Galileo. Sweden lands at 15.1% and Germany collapses to 6.9%, in both cases below the CNN and Galileo. Among the 5% filtered variants, Sweden 5% and Germany 5% rise above their unfiltered counterparts, while Russia 5% drops to 48.1% and Ireland 5% remains close to its unfiltered value. Figure 5.1 compares the five models directly.

Two examples of a side-by-side comparison of Galileo and SkySense++ predictions against the ground-truth mask is shown in Figure 5.2.

5.2.2 Ten-acquisition temporal extension

Table 5.6 reports Galileo under the temporal pipeline of Section 4.2.4, with all ten acquisitions packed into a single forward pass. Compared to its bi-temporal counterpart in Table 5.4, damaged-class F_1 rises by approximately 9

Table 5.4: Galileo primary bi-temporal segmentation results on the held-out test set.

Dataset variant	F_1^{dmg}	IoU_{dmg}	mIoU	P/R
Sweden	21.2%	11.9%	54.8%	22.0% / 20.5%
Sweden 5%	25.5%	14.6%	38.7%	17.3% / 48.1%
Russia	**59.0%**	**41.8%**	**69.6%**	51.4% / 69.2%
Russia 5%	56.0%	38.9%	65.7%	45.5% / **73.0%**
Ireland	37.6%	23.1%	60.0%	41.6% / 34.3%
Ireland 5%	37.1%	22.8%	55.3%	35.2% / 39.2%
Germany	31.1%	18.4%	56.8%	28.2% / 34.7%
Germany 5%	6.8%	3.5%	47.0%	8.5% / 5.6%

Table 5.5: SkySense++ primary bi-temporal segmentation results on the held-out test set.

Dataset variant	F_1^{dmg}	IoU_{dmg}	mIoU	P/R
Sweden	15.1%	8.2%	50.7%	10.4% / 27.6%
Sweden 5%	36.1%	22.0%	39.7%	29.9% / 45.4%
Russia	**90.1%**	**81.9%**	**90.5%**	**89.8%** / **90.4%**
Russia 5%	48.1%	31.7%	60.4%	34.1% / 81.9%
Ireland	51.6%	34.7%	66.5%	63.1% / 43.6%
Ireland 5%	47.9%	31.5%	59.0%	42.7% / 54.6%
Germany	6.9%	3.6%	50.0%	10.4% / 5.2%
Germany 5%	18.3%	10.1%	49.1%	21.5% / 16.0%

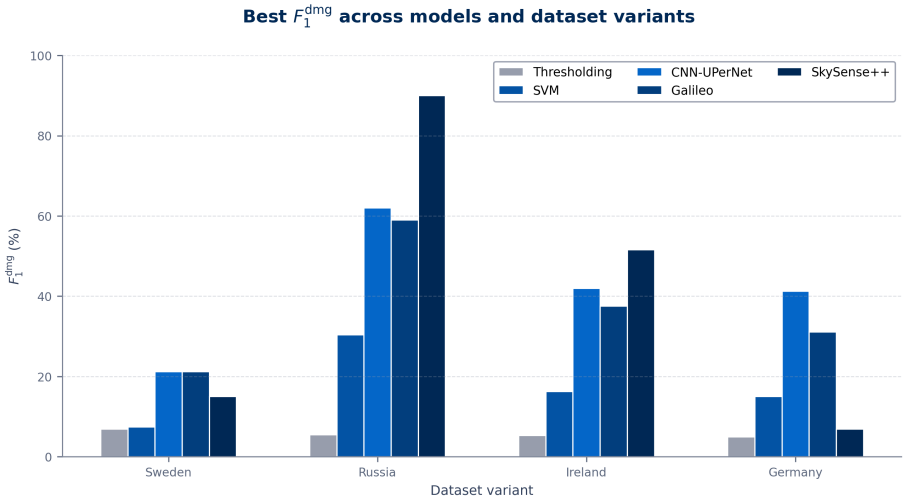


Figure 5.1: Damaged-class F_1 on the test set for each baseline, Galileo and SkySense++ under the primary bi-temporal pipeline.

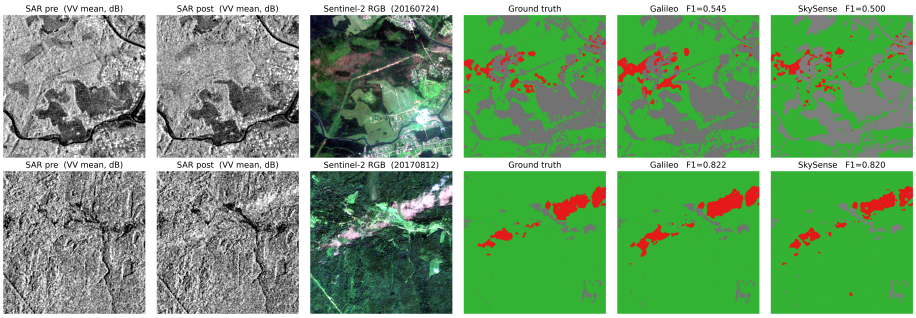


Figure 5.2: Galileo and SkySense++ damage predictions compared against the ground-truth mask, alongside the pre- and post-storm SAR composites and a S2 RGB image; red marks damaged forest, green undamaged forest, and grey non-forest pixels excluded from evaluation.

percentage points on Sweden (21.2% → 30.4%), 4 percentage points on Ireland (37.6% → 41.1%), drops by 3 percentage points on Russia (59.0% → 56.4%), and falls by 11 percentage points on Germany (31.1% → 20.5%).

Table 5.7 reports SkySense++ under the same temporal pipeline. Compared to its bi-temporal counterpart in Table 5.5, damaged-class F_1 rises on Sweden (15.07% → 25.38%) and Germany (6.90% → 18.18%), gaining roughly 10 and 11 percentage points respectively. On the two datasets where its bi-temporal performance is strongest, damaged-class F_1 falls instead, dropping by approximately 6 percentage points on Ireland (51.57% → 45.14%) and by approximately 40 percentage points on Russia (90.05% → 50.19%).

Table 5.6: Galileo temporal-extension results at $T = 10$ acquisitions on the held-out test set (unfiltered datasets only).

Dataset variant	F_1^{dmg}	$\mathbf{IoU_{dmg}}$	**mIoU**	**P/R**
Sweden	30.4%	17.9%	55.6%	28.9% / 32.1%
Russia	**56.4%**	**39.2%**	**63.4%**	**57.1%** / **55.6%**
Ireland	41.1%	25.9%	54.7%	33.9% / 52.2%
Germany	20.5%	11.4%	46.0%	13.0% / 48.3%

5.2.3 Stand-level evaluation

Table 5.8 reports the stand-level evaluation, with WI being the baseline method of Rüetschi et al. [39]. The strongest result is obtained on Russia, where SkySense++ reaches 93.1% object-level F_1, compared with 43.0% for WI. On Ireland, SkySense++ gives the best result at 57.7%, while on Sweden it improves over WI from 11.5% to 27.7%. On Germany, CNN-UPerNet performs best with 43.6% object-level F_1, compared with 31.1% for WI.

Table 5.7: SkySense++ temporal-extension results at $T = 10$ acquisitions on the held-out test set (unfiltered datasets only).

Dataset variant	F_1^{dmg}	$\mathbf{IoU_{dmg}}$	**mIoU**	**P/R**
Sweden	25.38%	14.53%	53.03%	20.49% / 33.34%
Russia	**50.19%**	**33.50%**	**59.25%**	**47.82%** / 52.79%
Ireland	45.14%	29.15%	57.23%	38.45% / **54.64%**
Germany	18.18%	10.00%	44.90%	11.48% / 43.65%

Table 5.8: Stand-level evaluation following the object-based protocol of Rüetschi et al. [39]. PA is producer's accuracy (object-level recall), UA is user's accuracy (object-level precision), and F_1^{obj} is their harmonic mean.

Dataset	**Model**	n^*	a^*	**PA**	**UA**	F_1^{obj}	$F_1^{\text{dmg pixel}}$
Sweden	CNN-UPerNet	15	–	11.6%	30.8%	16.8%	14.8%
	Galileo bi-temporal	55	–	15.9%	26.0%	19.7%	14.7%
	SkySense++ bi-temporal	100	–	29.9%	25.7%	27.7%	15.1%
	WI (Rüetschi et al. [39])	200	2.05	9.7%	14.0%	11.5%	5.4%
Russia	CNN-UPerNet	45	–	74.8%	67.0%	70.7%	62.0%
	Galileo bi-temporal	115	–	68.9%	65.6%	67.2%	59.0%
	SkySense++ bi-temporal	50	–	94.5%	91.8%	93.1%	90.1%
	WI (Rüetschi et al. [39])	50	3.25	53.1%	36.1%	43.0%	12.9%
Ireland	CNN-UPerNet	65	–	38.7%	46.0%	42.1%	42.0%
	Galileo bi-temporal	100	–	30.9%	50.6%	38.4%	37.6%
	SkySense++ bi-temporal	50	–	47.1%	74.4%	57.7%	51.6%
	WI (Rüetschi et al. [39])	90	2.05	25.9%	19.2%	22.1%	8.2%
Germany	CNN-UPerNet	75	–	44.9%	42.3%	43.6%	41.3%
	Galileo bi-temporal	65	–	42.9%	32.6%	37.0%	31.1%
	SkySense++ bi-temporal	40	–	12.9%	20.4%	15.8%	6.9%
	WI (Rüetschi et al. [39])	50	2.60	44.9%	23.8%	31.1%	9.9%

5.3 Additional experiment results

5.3.1 Experiment A: label efficiency

Tables 5.9, 5.10 and 5.11 report damaged-class F_1, IoU_{dmg}, mIoU and P/R as a function of the training-data fraction for CNN-UPerNet, Galileo and SkySense++ on the unfiltered Ireland and Russia datasets. Galileo and SkySense++ are each reported under both the `lora` and `full` regimes, while CNN-UPerNet is trained conventionally. Figure 5.3 visualises the curves on a logarithmic horizontal axis.

At the 1% fraction every model on both datasets is in a non-working regime, with low damaged-class F_1 and precision in the single digits paired with high recall, indicating near-indiscriminate positive predictions on a training split that contains only a handful of damaged tiles. All models recover by the 10% fraction. On Russia at 10%, CNN-UPerNet reaches 38.9% F_1, Galileo `full` reaches 48.0%, and SkySense++ `full` trails both at 21.0%; on Ireland at 10% the ordering differs, with CNN-UPerNet at 29.3%, Galileo `full` at 25.6%, and SkySense++ `lora` at 30.8%.

On Russia the three models diverge at full data. CNN-UPerNet rises monotonically to 62.0% F_1 at 100%, Galileo `full` plateaus at 57.6%, and SkySense++ `full` climbs steeply from 21.0% at 10% to 72.4% at 50% and 90.1% at 100%, the highest damaged-class F_1 recorded in this experiment. On Ireland the spread is much narrower, with CNN-UPerNet at 42.0%, Galileo `full` at 38.3%, and SkySense++ `full` at 51.6% at full data, and the gap between the 50% and 100% fractions are marginal for all three models.

The `lora` regime for SkySense++ does not scale with the training fraction in the same way as `full`. On both datasets SkySense++ `lora` peaks at the 50% fraction and then declines at 100%. On Russia at full data the SkySense++ `lora` score trails the corresponding `full` score by more than 50 percentage points. Galileo shows no comparable reversal, where its `lora` and `full` scores remain within a few percentage points of one another at the 50% and 100% fractions on both datasets.

Table 5.9: CNN-UPerNet label-efficiency results on the Ireland and Russia datasets.

Dataset	**Training fraction**	F_1^{dmg}	$\mathbf{IoU_{dmg}}$	**mIoU**	**P/R**
Ireland	1%	11.2%	5.9%	44.5%	6.5% / 39.3%
	10%	29.3%	17.2%	56.6%	28.9% / 29.7%
	50%	38.4%	23.8%	60.3%	40.3% / 36.7%
	100%	42.0%	26.6%	61.6%	39.8% / 44.4%
Russia	1%	5.7%	3.0%	9.1%	3.0% / 91.4%
	10%	38.9%	24.1%	58.9%	26.6% / 72.4%
	50%	59.2%	42.0%	69.9%	58.2% / 60.2%
	100%	62.0%	44.9%	71.3%	59.0% / 65.3%

Table 5.10: Galileo label-efficiency results on the Ireland and Russia datasets.

		lora				full			
Dataset	**Training fraction**	F_1	**IoU**	**mIoU**	**P/R**	F_1	**IoU**	**mIoU**	**P/R**
Ireland	1%	5.4%	2.8%	37.8%	3.0% / 29.0%	5.9%	3.1%	14.6%	3.1% / 85.0%
	10%	6.2%	3.2%	32.2%	3.3% / 46.7%	25.6%	14.7%	55.3%	24.6% / 26.8%
	50%	33.4%	20.1%	58.4%	36.5% / 30.8%	30.6%	18.1%	57.4%	34.4% / 27.6%
	100%	37.6%	23.1%	60.0%	41.6% / 34.3%	38.3%	23.6%	60.2%	40.3% / 36.4%
Russia	1%	5.4%	2.8%	2.6%	2.8% / 98.3%	5.4%	2.8%	3.6%	2.8% / 96.4%
	10%	43.4%	27.7%	61.4%	31.8% / 68.2%	48.0%	31.5%	63.9%	38.7% / 63.0%
	50%	56.2%	39.1%	68.2%	50.5% / 63.3%	56.9%	39.8%	68.5%	51.6% / 63.4%
	100%	59.0%	41.8%	69.6%	51.4% / 69.2%	57.6%	40.5%	68.9%	52.0% / 64.6%

Table 5.11: SkySense++ label-efficiency results on the Ireland and Russia datasets.

		lora				full			
Dataset	**Training fraction**	F_1	**IoU**	**mIoU**	**P/R**	F_1	**IoU**	**mIoU**	**P/R**
Ireland	1%	7.3%	3.8%	43.2%	4.2% / 30.5%	5.9%	3.0%	43.3%	3.3% / 23.6%
	10%	30.8%	18.2%	57.7%	34.5% / 27.9%	13.1%	7.0%	52.0%	16.5% / 10.8%
	50%	54.8%	37.7%	68.0%	69.5% / 45.2%	51.5%	34.6%	66.4%	62.9% / 43.6%
	100%	34.7%	21.0%	59.0%	35.0% / 34.4%	51.6%	34.7%	66.5%	63.1% / 43.6%
Russia	1%	4.6%	2.3%	14.9%	2.4% / 83.2%	9.1%	4.8%	10.6%	4.8% / 86.4%
	10%	33.3%	19.9%	58.5%	30.5% / 36.5%	21.0%	11.7%	53.3%	37.5% / 14.5%
	50%	60.4%	43.3%	70.9%	66.9% / 55.1%	72.4%	56.7%	77.1%	78.3% / 67.3%
	100%	39.9%	24.9%	61.1%	37.5% / 42.7%	90.1%	81.9%	90.5%	89.8% / 90.4%

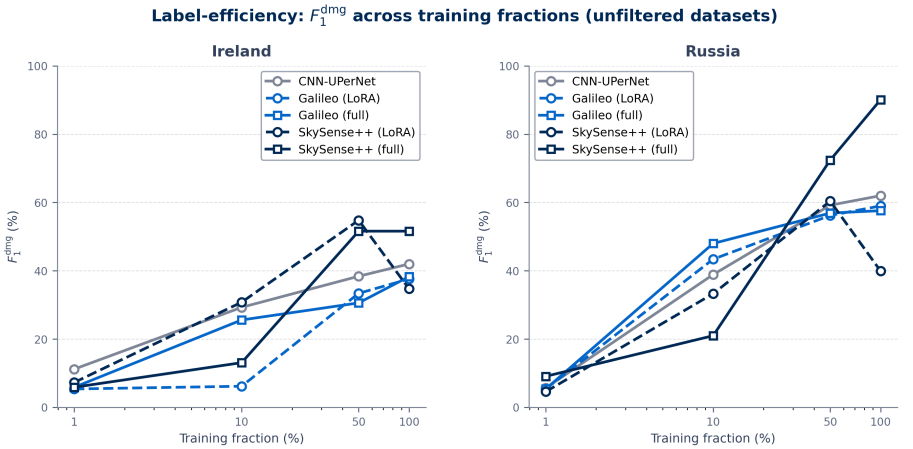


Figure 5.3: Label-efficiency learning curves on the unfiltered Ireland and Russia datasets. Damaged-class F_1 as a function of the fraction of available training data, shown on a logarithmic horizontal axis. Numbers are taken from Tables 5.9, 5.10 and 5.11.

5.3.2 Experiment B: temporal depth

Table 5.12 reports damaged-class F_1 as a function of the input temporal depth T for Galileo and SkySense++, with all other hyperparameters fixed to the configuration selected in the temporal sweep of Section 4.2.5. Both models improve monotonically with T on Sweden, Russia, and Ireland. On Sweden, Galileo rises from 17.3% to 30.4% and SkySense++ from 14.0% to 25.4%, with SkySense++ trailing at every depth. On Russia, Galileo rises from 43.4% to a peak of 57.7% at $T = 8$ before easing to 56.4% at $T = 10$, while SkySense++ rises more gradually from 41.9% to its peak of 50.2% at $T = 10$. On Ireland the ordering reverses, with SkySense++ ahead of Galileo from $T = 2$ onward and reaching 45.1% at $T = 10$ against Galileo's 43.3% peak at $T = 8$. The two models therefore differ in where they peak, as Galileo plateaus or turns over at $T = 8$ on Russia and Ireland, whereas SkySense++ continues to its maximum at $T = 10$ on all three of these datasets. Germany is irregular for both, with Galileo peaking at $T = 4$ (21.7%) and SkySense++ at $T = 8$ (21.6%) before dropping to 18.2% at $T = 10$. Figure 5.4 visualises the data.

Table 5.12: Temporal-depth F_1^{dmg} results on the unfiltered datasets for Galileo and SkySense++. T denotes the number of acquisitions.

		F_1^{dmg}				
Dataset	**Model**	**T=2**	**T=4**	**T=6**	**T=8**	**T=10**
Sweden	Galileo	17.3%	23.2%	26.7%	30.1%	30.4%
	SkySense++	14.0%	19.0%	19.8%	23.8%	25.4%
Russia	Galileo	43.4%	48.7%	53.4%	57.7%	56.4%
	SkySense++	41.9%	45.7%	47.8%	48.9%	50.2%
Ireland	Galileo	35.9%	40.2%	42.9%	43.3%	41.1%
	SkySense++	38.4%	40.2%	42.1%	42.3%	45.1%
Germany	Galileo	16.5%	21.7%	18.3%	19.1%	20.5%
	SkySense++	16.0%	19.4%	18.6%	21.6%	18.2%

5.3.3 Experiment C: cross-dataset generalisation

Tables 5.13 and 5.14 report the cross-dataset generalisation results described in Section 4.2.5. In the full hold-out setting, where the target country is excluded entirely from training, all three models perform weakly on Sweden: the CNN-UPerNet reaches 0.3%, Galileo 0.4%, and SkySense++ 6.3%, where the latter is recall-driven (17.2%) at very low precision (3.9%), indicating broad low-confidence activations rather than reliable detection. Russia gives the strongest full hold-out results across the board, with 4.6% for the CNN, 30.4% for Galileo, and 11.8% for SkySense++. Ireland falls between these extremes at 13.9%, 23.0%, and 9.5% respectively.

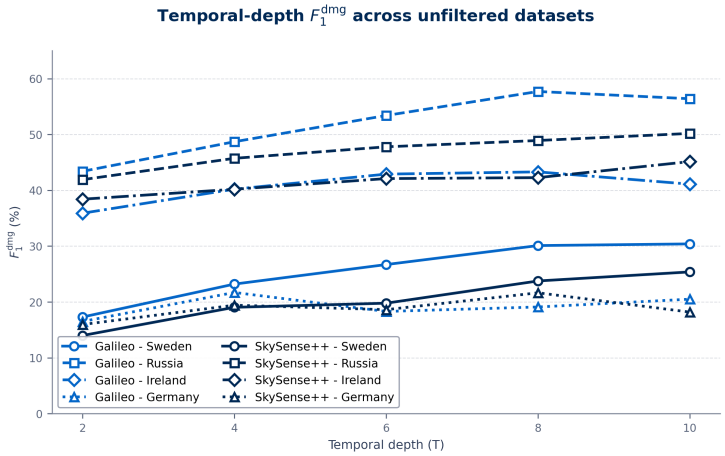


Figure 5.4: Damaged-class F_1 as a function of input temporal depth T for Galileo and Skysense++, evaluated on the unfiltered variant of each country dataset.

In the partial hold-out setting, where the target country contributes training tiles to the balanced pool, performance improves substantially for all three models on all three targets. Russia reaches 61.8% for the CNN, 59.1% for Galileo, and 88.7% for SkySense++. Ireland reaches 22.1%, 30.1%, and 32.3%, and Sweden reaches 8.9%, 15.5%, and 14.5% respectively. The relative ordering across models shifts between protocols, where SkySense++ leads by a wide margin in partial hold-out on Russia, but falls below Galileo in full hold-out on the same target, a gap of 77 percentage points between its own two conditions compared to 29 for Galileo and 57 for the CNN.

Table 5.13: CNN-UPerNet cross-dataset generalisation on the balanced three-country study. Partial hold-out includes target-country training tiles in the balanced pool, while full hold-out excludes the target country entirely from training.

Target country	**Training setting**	F_1^{dmg}	$\mathbf{IoU_{dmg}}$	**mIoU**	**P/R**
Sweden	Partial hold-out	8.9%	4.7%	50.2%	8.2% / 9.7%
	Full hold-out	0.3%	0.1%	49.0%	4.9% / 0.1%
Russia	Partial hold-out	61.8%	44.7%	71.2%	59.5% / 64.2%
	Full hold-out	4.6%	2.4%	49.4%	9.3% / 3.1%
Ireland	Partial hold-out	22.1%	12.4%	53.7%	19.3% / 25.8%
	Full hold-out	13.9%	7.5%	51.7%	17.3% / 11.6%

Table 5.14: Foundation-model cross-dataset generalisation on the balanced three-country study with full fine-tuning. Partial hold-out includes target-country training tiles in the balanced pool, while full hold-out excludes the target country entirely from training.

		Galileo				SkySense++			
Target country	**Training setting**	F_1	**IoU**	**mIoU**	**P/R**	F_1	**IoU**	**mIoU**	**P/R**
Sweden	Partial hold-out	15.5%	8.4%	52.2%	13.9% / 17.4%	14.5%	7.8%	50.9%	11.5% / 19.5%
	Full hold-out	0.4%	0.2%	49.0%	3.4% / 0.2%	6.3%	3.3%	45.0%	3.9% / 17.2%
Russia	Partial hold-out	59.1%	42.0%	69.7%	53.8% / 65.7%	88.7%	79.7%	89.6%	89.7% / 87.7%
	Full hold-out	30.4%	17.9%	57.0%	29.5% / 31.3%	11.8%	6.3%	49.6%	8.0% / 22.8%
Ireland	Partial hold-out	30.1%	17.7%	57.0%	29.8% / 30.4%	32.3%	19.3%	58.3%	37.4% / 28.4%
	Full hold-out	23.0%	13.0%	54.3%	21.9% / 24.3%	9.5%	5.0%	49.0%	6.7% / 16.5%

5.3.4 Experiment D: augmentation strategy

Table 5.15 summarises the augmentation sweep from Section 4.2.5. CNN-UPerNet is more sensitive to augmentation than Galileo on most datasets. Geometric augmentation reduces CNN performance on Russia and Germany, while geometric-plus-jitter partly recovers the loss. Ireland is mixed, with a small gain under geometric augmentation followed by a large drop when jitter is added. Galileo is comparatively stable on Sweden, Russia, and Ireland, varying by less than 4 percentage points across augmentation settings, but drops sharply on Germany under both augmentation conditions. Sweden is the most stable dataset overall, with all models remaining near 20% damaged-class F_1.

SkySense++ is the most augmentation-sensitive model overall. On Russia it achieves the highest score of the experiment without augmentation (90.1%) but falls to around 57% once augmentation is applied. Ireland follows the same pattern, decreasing from 51.6% without augmentation to around 31% with augmentation. Sweden remains relatively unchanged across conditions, while Germany shows the opposite trend to the other models, improving from 6.9% without augmentation to 11.9% under geometric-plus-jitter. Overall, no single augmentation strategy consistently benefits all models and datasets.

Table 5.15: Augmentation-strategy results on the unfiltered datasets.

		CNN-UPerNet		**Galileo**		**SkySense++**	
Dataset	**Augmentation**	F_1^{dmg}	**IoU**	F_1^{dmg}	**IoU**	F_1^{dmg}	**IoU**
Sweden	None	21.2%	11.8%	21.2%	11.9%	15.1%	8.2%
	Geometric	20.4%	11.4%	17.2%	9.4%	14.8%	8.0%
	Geometric + jitter	20.6%	11.5%	19.6%	10.9%	16.3%	8.9%
Russia	None	62.0%	44.9%	59.0%	41.8%	90.1%	81.9%
	Geometric	47.8%	31.4%	55.0%	37.9%	57.6%	40.4%
	Geometric + jitter	54.4%	37.4%	54.7%	37.7%	56.9%	39.7%
Ireland	None	42.0%	26.6%	37.6%	23.1%	51.6%	34.7%
	Geometric	43.5%	27.8%	36.6%	22.4%	31.0%	18.3%
	Geometric + jitter	31.8%	18.9%	35.6%	21.7%	30.9%	18.3%
Germany	None	41.3%	26.0%	31.1%	18.4%	6.9%	3.6%
	Geometric	33.8%	20.3%	6.0%	3.1%	8.9%	4.7%
	Geometric + jitter	38.8%	24.1%	6.5%	3.4%	11.9%	6.3%

6 Discussion

6.1 Baseline performance across datasets

The baseline results reveal an important asymmetry across the four datasets that cannot be explained by modeling choices alone. The Russian dataset consistently has the highest damaged-class F_1 across all three baselines (Tables 5.1, 5.2, 5.3), and the pattern holds after filtering. The Irish and German dataset has mediocre results for the CNN but substantially lower ones for the other baseline methods. The Swedish dataset is worse on every model on the unfiltered variant, where the CNN achieves only 21.2% F_1, and the threshold detector achieves 6.9% (Table 5.1).

The most natural explanation for Russia's advantage is annotation quality and event geometry. The European Russia Windthrow Database labels were produced by manual delineation of stand-replacing windthrow patches from Landsat time series combined with the Global Forest Change product [44], a method that prefers to capture large, high-severity clearings. Stand-replacing events produce the largest and most spatially coherent SAR backscatter change signal and sometimes complete canopy removal, which leads to a near total loss of volume scattering and a shift toward surface scattering from the ground. A model trained on such labels therefore learns a comparatively clean mapping from the SAR signal to a damage indicator, and the same model evaluated on a test set drawn from the same distribution encounters the same type of clear, geometrically compact targets. This is also visible in the precision and recall balance. At 59.0% precision and 65.3% recall for the CNN, Russia is the only dataset where the model does not exhibit the severe precision-recall imbalance that characterizes the harder datasets.

Germany presents the most differing pattern in the baselines. The CNN reaches 41.3% F_1 on the unfiltered variant, comparable to Ireland, yet the SVM collapses to 15.0% on the same split and then to 0.8% on the 5% filtered variant, the latter can be compared in effectiveness as being random classified due to its low result. The Germany dataset derives from aerial photography at 12 cm ground resolution, manually interpreted by the state forestry authority of Mecklenburg-Vorpommern [39]. This very high annotation resolution means that small and isolated damage patches are included in the reference, including single windthrown trees, which constitute the majority of the damage polygons in the

dataset, even when they cover as little as 0.5 ha. At the 10 m resolution of S1 and within a 448×448 pixel tile, such sub-hectare patches occupy only a few pixels, their backscatter signature is diluted by surrounding intact canopy and barely distinguishable from speckle. The SVM, which operates on patch-level texture statistics without any spatial context, therefore cannot reliably discriminate damaged from undamaged forest in the Germany feature space, especially after the 5% filtering discards all but nine tiles. The CNN, by contrast, exploits spatial context through convolution and can detect the characteristic spatial configuration of a damage patch even when individual pixels are ambiguous. Germany therefore illustrates how a resolution mismatch between the annotation source and the SAR input degrades detection, where features that are apparent to the annotation hardware can be impossible to detect in the coarser SAR imagery.

Sweden occupies the opposite end of the label-quality spectrum. The reference mask for Storm Johannes was derived algorithmically from a S2 change composite. While this procedure captures damage at 10 m resolution and avoids the sub-pixel dilution problem that affects Germany, it introduces a different source of noise. The index (Equation 3.1) responds to any change in snow reflectance between the two acquisition dates, including harvesting activity, phenological differences, and variation in snow conditions, none of which constitute windthrow. The labelling procedure relies on snow being present on the ground on both acquisition dates, since the entire Skogsstyrelsen method depends on newly exposed snow appearing in openings created by removed canopy, and this assumption holds in our case. The complication is that even when snow is present on both dates, the SAR signal does not respond to canopy removal in the same simple way that the optical labelling procedure assumes. Snow depth, sub-canopy distribution, freeze-thaw state, and interception geometry each modulate S1 backscatter independently of canopy structure. Pixels that the optical procedure correctly flags as damaged because newly exposed snow becomes visible can therefore lack a corresponding SAR signature, and conversely, snow-state variation over intact forest can mimic the backscatter change associated with damage. Skogsstyrelsen noted that harvest areas were masked where possible, but errors are unavoidable in an automated workflow. In the context of supervised learning, these mislabelled pixels act as noisy supervision, penalising the model for correctly detecting intact forest and rewarding it for matching spurious positives. This is consistent with the very low precision of the threshold detector on Sweden (5.3%) and the relatively modest CNN performance (21.2% F_1 versus 62.0% for Russia), which both suggest that a substantial fraction of the positive labels do not correspond to SAR-detectable windthrow. Sweden has the largest tile count of all four countries (2941 unfiltered, Table 3.1), so the problem is not insufficient training data but insufficient label accuracy.

Ireland lies between the two. Its labels were derived from a multi-source Earth Observation campaign at resolutions ranging from 0.5 m (SkySat) to 10 m (S2), conducted by DAFM in the weeks following Storms Darragh and Éowyn [10]. The dominant annotation source being sub-metre imagery gives Ireland a geometric precision closer to Germany than to Sweden, but the stand-

level minimum mapping unit of 0.1 ha is smaller than Germany's 0.5 ha threshold and therefore admits more marginal damage patches. The CNN F_1 of 42.0% and the relatively balanced precision and recall (39.8% / 44.4%) suggest a moderately well-defined learning target, with performance limited primarily by the inherent difficulty of detecting sub-hectare damage at S1 resolution rather than by label noise.

A further structural factor that complicates the comparison is dataset size. Germany has only 54 base tiles, of which 9 satisfy the 5% damage fraction criterion. This is insufficient to train a stable model on the filtered variant, and the collapse of SVM performance from 15.0% to 0.8% after filtering suggests that the sweep was fitting to noise rather than to a stable signal. The results for the Germany 5% variant should therefore be interpreted with great caution throughout all experiments.

The 5% filtered variants also deserve a clarification. Their primary motivation was not to enable cross-dataset comparison but to mitigate the very strong class imbalance present in all four base datasets, where damaged pixels make up a small fraction of the total forest area (Table 3.1). By restricting training to tiles in which at least 5% of pixels are labelled damaged, the filtered variants raise the prior probability of the positive class and provide a denser learning signal per gradient update, with the intent of giving the lower-capacity baselines an easier representation of the task. The results show that this expectation is not met. The filtered variants do not produce significantly better damaged-class F_1 than their unfiltered counterparts on the strong learners, and in several cases (Germany 5%, Russia 5% for the CNN) the filtered variant performs worse. Class imbalance is therefore not the dominant bottleneck for this task.

Taken together, these observations argue for treating the four datasets not as four equivalent replications of the same task but as differing on both annotation quality and the amount of data. Russia represents a high-quality dataset, where SAR-based detection is possible even when using classical methods. Sweden represents a low-quality dataset where label noise fundamentally limits achievable performance. Germany and Ireland occupy intermediate positions, separated primarily by annotation resolution and minimum patch size. Any claim about relative model performance should therefore be conditioned on where a dataset sits along this spectrum. This label-quality-over-quantity pattern recurs throughout the experiments below, where Russia, with only 205 tiles, reaches roughly three times the damaged-class F_1 of Sweden at 2941 tiles across every model.

Comparison with prior published work

A direct numerical comparison with the two studies that most closely match our setting, Rüetschi et al. [39] and Tomppo et al. [47], requires several caveats but is still informative.

Rüetschi et al. work on the same Mecklenburg-Vorpommern reference data we use for Germany, but their evaluation is object-based rather than pixel-based, enumerates no negative class (so $\mathrm{IoU}_{\mathrm{dmg}}$ and mIoU are undefined), and distinguishes three windthrow severity classes. They report producer's accuracy

(PA, object recall) and user's accuracy (UA, object precision), as described in Section 2.2.5, per class on the German validation area: areal windthrow PA 0.88, UA 0.21 ($F_1 = 0.34$, $N_{\text{ref}} = 8$); single standing trees PA 0.29, UA 0.15 ($F_1 = 0.20$, $N_{\text{ref}} = 17$); single windthrown trees PA 0.07, UA 0.48 ($F_1 = 0.12$, $N_{\text{ref}} = 225$).

The CNN reaches 41.3% pixel-level and 43.6% object-level damaged-class F_1 on the unfiltered Germany variant when all three Rüetschi classes are merged, which exceeds each of their per-class object-based F_1 values on the same reference and approaches the $F_1 = 0.74$ they achieve on their Swiss development area under the easier two-class setup. Pixel-level and object-based F_1 are not directly comparable, since object-based evaluation counts any pixel overlap as a correct detection and is therefore more lenient, so this comparison is indicative rather than definitive. Both their evaluation and ours agree that areal windthrow is detectable and scattered single-tree damage essentially is not.

Tomppo et al. [47] target the 2010 Finnish storm Asta with varying stand-level classifiers on a reference defined by ground-survey volume losses above 20% and 50% thresholds, reporting stand-level overall accuracies in the 70–80% range. The setting, boreal stand-replacing damage, is closest to our Russia dataset, where the CNN achieves 62.0% pixel-level F_1 and 70.7% object-level F_1. These are not directly comparable, since both the dataset and methodology differ, but both setups agree that boreal stand-replacing windthrow is the regime where S1 detection works best.

The published S1 windthrow literature has overwhelmingly evaluated its methods at the object or stand level rather than at the pixel level, and the apparent gap between their headline accuracies and our pixel level F_1 values is largely an artifact of the choice of evaluation unit. The systematic difference between pixel-level and object-level F_1 in our own results mirrors this pattern and is examined quantitatively in Section 6.7.

6.2 Foundation models versus the CNN baseline

The progression from the threshold detector to the SVM to the CNN-UPerNet baseline gives a consistent improvement on every dataset. The further step from the CNN to a foundation model under the bi-temporal pipeline depends strongly on which foundation model is used. Galileo lands within a few absolute percentage points of the CNN on Russia, Ireland, and Sweden, and roughly ten points below it on Germany (Tables 5.4, 5.3). SkySense++ diverges from this pattern on the two cleanest datasets. On Russia it reaches 90.1% damaged-class F_1 against 62.0% for the CNN and 59.0% for Galileo, and on Ireland it reaches 51.6% against 42.0% and 37.6% respectively (Table 5.5). On Sweden and Germany the picture inverts and SkySense++ sits below both the CNN and Galileo. The ordering of datasets (Russia best, Sweden and Germany weakest) is preserved across all three model families, but the spread between models on Russia and Ireland is far larger for SkySense++ than for Galileo.

Two mechanisms can explain why the foundation model advantage is muted

for Galileo on all datasets and asymmetric for SkySense++ across them. The first is the structure of the discriminative signal. The dominant component of bi-temporal windthrow detection in S1 is a low-level intensity and texture difference between pre-event and post-event composites, which a CNN with relatively few learnable filters can already extract, consistent with evidence that pretrained representations confer their largest gains on tasks with substantial semantic structure beyond low-level statistics [24, 27]. This explains the CNN matching or exceeding Galileo on three of four datasets and the saturation observed when Galileo's larger pretrained backbone is applied to the same input. The Russia result for SkySense++ shows that this is not the whole story. Beyond the dominant change component, the bi-temporal composite still contains residual structure that separates damaged from undamaged forest, such as spatially extended texture patterns and speckle realisations that a small convolutional receptive field averages away. Extracting this residual signal requires both substantial capacity and pretraining that yields SAR-relevant features, and only the largest model with SAR-aware pretraining is positioned to do so. That the improvement materialises specifically on Russia and Ireland, the two datasets with the cleanest stand-replacing labels, supports this reading, where on Sweden and Germany the label noise and event geometry cap the achievable F_1 at a level any sufficiently expressive model can already reach (Section 6.1).

The second mechanism is pretraining-distribution mismatch. Galileo's pretraining mix is dominated by S2 and other optical sources, with S1 forming one of nine modalities and contributing a relatively small share of the global and local masked-modelling objectives [49]. The pretrained features are therefore optimised primarily for optical reflectance statistics rather than for SAR backscatter, and the gap has to be closed during fine-tuning. SkySense++ adopts a different recipe in which the SAR encoder is one of three modality-specific spatial encoders, kept architecturally separate from the optical and multispectral inputs, with a dedicated cross-modal fusion transformer producing the spatio-temporal representation [54]. In the S1-only configuration used here the optical encoders are inactive, but the SAR encoder retains the representations learned during multi-granularity contrastive and masked semantic learning over 27 million remote sensing images. This yields a stronger SAR feature extractor than Galileo's on the two datasets with the cleanest labels, where the residual signal is recoverable, and a weaker or comparable one where the supervision signal itself is noisier and the residual signal is masked by label noise.

A consequence is that parameter count alone does not predict performance on the bi-temporal task, while scale combined with the right pretraining and fusion architecture does. The CNN-UPerNet has approximately 16.3 M parameters trained from scratch, Galileo-Base roughly 124.3 M, and SkySense++ around 2.03 B . The gap in damaged-class F_1 between the smallest and largest model is small and inconsistent in sign for the Galileo-CNN comparison, but reaches roughly 28 percentage points on Russia and 14 on Ireland in favour of SkySense++. The CNN trained from scratch is enough to extract the discriminative bi-temporal signal on the smaller and noisier datasets, where neither additional pretraining nor a larger backbone helps, but on the largest dataset with the cleanest labels

the bi-temporal task is evidently not yet saturated and the larger pretrained model continues to extract additional information.

6.3 Effect of temporal information

Moving from the bi-temporal composite to the full ten-acquisition time series helps Galileo on the two noisier datasets and leaves the cleaner ones roughly unchanged (Table 5.6). Sweden gains approximately nine percentage points of damaged-class F_1 (21.2% $\rightarrow$ 30.4%) and Ireland roughly four (37.6% $\rightarrow$ 41.1%), while Russia is stable to within single-seed variance (59.0% $\rightarrow$ 56.4%) and Germany declines by about eleven points (31.1% $\rightarrow$ 20.5%), consistent with the high-variance behaviour of its 54-tile set discussed in Section 6.1.

SkySense++ shows the opposite dataset dependence under the same extension (Table 5.7). It gains on Sweden and Germany alongside Galileo, by roughly 10 and 11 percentage points respectively, though the Germany numbers carry the same small-dataset caution. It loses heavily on the two datasets where its bi-temporal performance is strongest, dropping by about six percentage points on Ireland and close to forty on Russia, collapsing toward the same 45–50% band that Galileo and the CNN occupy. The most likely explanation is that the bi-temporal pipeline hands SkySense++ a pre-event and post-event composite that already isolates the change signal cleanly on the clean-label datasets, whereas the temporal pipeline forces the fusion transformer to recover that signal from ten individually noisy acquisitions, discarding the very structure that drove its strong bi-temporal Russia result (Section 6.2). The data dependence noted in Section 6.4 compounds this, since spreading supervision across ten timesteps may further weaken an already data-needy model.

Two complementary mechanisms explain why temporal information helps where it does. The first is statistical denoising. Each S1 acquisition is a noisy single-look estimate of the underlying backscatter, and stacking ten acquisitions reduces speckle variance under the standard multi-look assumption [32], exposing the change signal more cleanly. This argument also applies in part to the bi-temporal pipeline, where the four-channel composite $\mathbf{X}_{\text{comp}}$ is itself the average of five pre-event and five post-event acquisitions. What the bi-temporal composite cannot recover is the second mechanism, where rather than receiving a single pre-collapsed composite, the encoder under the temporal pipeline receives individual acquisitions and can learn to weight them according to their discriminative content. Under PASTIS-style [20] mean-pooling the temporal order is not preserved, but pooling the learned representations rather than the raw backscatter values allows the model to suppress uninformative acquisitions in representation space in a way that pre-event intensity averaging cannot.

The temporal results do not show the stronger event-localisation effect that the SkySense++ architecture might be expected to deliver. Because the Fusion-Transformer attends explicitly across the temporal axis it preserves ordering information, unlike the mean-pooling used by Galileo, and in principle it could separate slow seasonal change (canopy moisture, snow accumulation, freeze-thaw

cycles) from the abrupt step at the storm date more sharply than a pooled representation can. In practice no such advantage is visible: SkySense++ does not exceed Galileo at $T = 10$ on any of the four datasets and is substantially worse on Russia and Ireland. One contributing factor is that the S1 implementation does not supply acquisition-date or month-index embeddings (Section 4.2.4), so the only temporal signal available to the fusion transformer is the fixed chronological ordering of the timesteps.

The depth-sensitivity curve flattens around $T \in [8, 10]$. Galileo on Russia rises steadily through $T = 8$ (Table 5.6) and then drops slightly at $T = 10$, Ireland follows the same shape, and Sweden plateaus around $T = 8$. The plateau is consistent with diminishing returns once enough acquisitions are stacked to denoise speckle and bracket the event in time. Rüetschi et al. [39] reach a similar conclusion in their multi-temporal windthrow study, reporting that approximately five post-event acquisitions are sufficient to stabilise the detection. Our T-depth curves are aligned with this finding, with the plateau emerging at five post-event acquisitions in both setups. The Germany curve is non-monotone, which most likely reflects the small 54-tile dataset producing high-variance test estimates rather than a real failure of the temporal mechanism.

The plateau also has an operational dimension. With a two-satellite S1 constellation, the exact-repeat cycle is six days and the average effective revisit time using both ascending and descending tracks at northern-European latitudes is roughly three days [48]. With $T = 10$ acquisitions arranged as five pre-event and five post-event, the post-event window spans 24 days using same-geometry passes, or roughly 12 days using mixed-geometry passes. After this window, optical sensors such as S2 typically begin to provide cloud-free observations even in northern Europe, and optical change-detection becomes a stronger detector than SAR. The flat point of the T-depth curve is therefore close to the natural operational regime for SAR-only windthrow detection, the period when optical confirmation is not yet available.

6.4 Adaptation regime and label efficiency

On Russia under full fine-tuning, all three models collapse to roughly 5–9% damaged-class F_1 at the 1% training fraction, recover to the 21–48% range at 10%, and by 50% the CNN and Galileo converge near 57–59% while SkySense++ accelerates sharply to 72.4% and then to 90.1% at full data, a substantial margin above both (Section 6.2). On Ireland the picture is more uniform, where under full fine-tuning all three recover from near-degenerate performance at 1% to the 13–29% range at 10%, with SkySense++ recovering most slowly despite eventually surpassing Galileo and matching the CNN by 50–100% training data, converging near 51%.

The LoRA experiments reveal that parameter-efficient adaptation affects the two foundation models differently. For Galileo, LoRA largely preserves the same label-efficiency pattern as full fine-tuning, suggesting that Galileo's ceiling is set by the suitability of its pretrained representation for the SAR

damage-segmentation task rather than by the number of trainable parameters used during adaptation. For SkySense++, LoRA changes the learning curve more. A plausible reading is that LoRA acts as a useful regulariser when labels are scarce, limiting over-adaptation of the larger SkySense++ model, but becomes a bottleneck once enough data are available to support deeper task-specific realignment of the pretrained representation.

The behaviour of SkySense++ on Russia under full fine-tuning is qualitatively different from the other models, where rather than saturating early, its label-efficiency curve steepens between 10% and 100%, suggesting that its pretrained representations require a broader coverage of storm-patch diversity before they can be reliably fine-tuned to the target domain. This is consistent with the relative size of the three model, with the CNN being the smallest and SkySense++ the largest. Although foundation models in remote sensing are often expected to perform better when only limited labelled data is available [8, 17], our results suggest that the models need some data to yield good results, rather than uniformly improving on baselines in low-data settings. A plausible reading of the very low end of the curve is that at 1% the Russia split contains only a handful of damaged tiles, below the absolute count of positives needed for the model to see a representative range of storm patches, and pretrained representations cannot manufacture positives where none exist. The 1% fraction is included as a stress test rather than a realistic operating point.

The Irish dataset plateau across all models warrants a separate observation. Despite Ireland being the larger dataset by tile count, no model exceeds 52% damaged-class F_1 at full data under full fine-tuning, and the gap between the 50% and 100% fractions is marginal for all three. This indicates that additional labelled data within the same domain offers diminishing returns, and that annotation quality, label consistency, or the inherent difficulty of the Irish storm signature is a more binding constraint than data volume. This is the same label-quality-over-quantity pattern established in Section 6.1, where SkySense++ on Russia at full data exceeds even the best Ireland result by a wide margin despite a much smaller data set (Table 3.1).

6.5 Cross-dataset generalisation

The cross-dataset generalisation results show that transfer across countries is limited even when the training pool is balanced. The foundation models use the bi-temporal composite here, so the partial hold-out values are directly comparable to the bi-temporal ceilings reported in Section 6.2. In the full hold-out setting, where the target country is excluded entirely from training, performance decreases sharply across all three model families: Sweden becomes effectively non-detectable (damaged-class F_1 below 7% for every model), and Russia, although the least poor transfer direction, still drops to 4.6% for the CNN, 30.4% for Galileo, and 11.8% for SkySense++. This reinforces the discussion in Section 6.1, where although the datasets share the same sensor and nominal task, they differ too strongly in annotation regime, event geometry, and class

definition for a model trained on one mixture of countries to generalise cleanly to another.

The partial hold-out setting is more informative. Once the target country contributes training tiles, performance recovers substantially, and performance on Russia nearly matches the per-country baseline across all three models. The models can therefore make effective use of multi-country training, but only when the target-domain label distribution is already represented in the pool. Ireland also benefits from the target-country contribution though it remains below its single-country upper bound, while Sweden remains weak even in partial hold-out, consistent with the label-quality pattern established in Section 6.1.

The two foundation models differ in the full hold-out setting. Galileo generalises considerably better than SkySense++ on Russia (30.4% versus 11.8%) and Ireland (23.0% versus 9.5%), despite SkySense++ leading in the single-country and partial hold-out settings. This inversion suggests that the larger capacity of SkySense++ comes at the cost of stronger overfitting to target-domain statistics when any target-country signal is present during training, making it more sensitive to their removal, whereas Galileo's more modest capacity produces representations that transfer better across domain boundaries.

The practical implication is that a model trained on data from one storm event and one forest type should not be expected to generalise to structurally different events without local fine-tuning. The minimum viable deployment scenario is therefore in-country fine-tuning on a small set of annotated tiles from a recent local storm, rather than zero-shot transfer from a geographically distant training pool. As the label-efficiency results in Section 6.4 show, all three models recover much of their in-distribution performance with as little as 50% of the full training set on Russia and Ireland, a modest manual annotation effort following a new storm event.

6.6 Augmentation strategy

The augmentation results in Table 5.15 do not support a single recommendation that holds across the models and datasets. SkySense++ is the most augmentation-sensitive of the three, and the sensitivity is concentrated on exactly the clean-label datasets where it is strongest, where adding augmentation costs it the most. On the noisier datasets the effect reverses or vanishes. This mirrors the temporal-extension finding in Section 6.3. Augmentation disturbs the clean bi-temporal change signal that is the source of SkySense++'s advantage on Russia and Ireland, whereas on Sweden and Germany there is little clean signal left for it to corrupt. The CNN shows a milder version of the same dataset dependence, benefiting from intensity jitter on some datasets and losing on others, while Galileo is largely insensitive to jitter, consistent with its higher-capacity backbone being robust to a change that is small relative to the variation in its pretraining distribution.

Geometric augmentation alone is therefore the safer default across models and datasets, as it never harms performance substantially in our results, while intensity jitter should be treated as a dataset-specific hyperparameter rather than

a universal improvement. The case against vertical flip and $90°/270°$ rotation is independent and physical. These transformations are inconsistent with the fixed right-looking orbital geometry of S1 acquisitions [32] and would corrupt the geometric relationship between the VV and VH polarisation channels and the range look direction.

6.7 Operational implications

The pixel-level segmentation framing used throughout this thesis is a demanding formulation of the windthrow detection problem. Per-pixel binary classification at 10 m resolution requires the model to commit to a damaged-or-intact decision at every cell, and our results suggest that for most datasets and models this commitment is too aggressive for the signal-to-noise ratio of bi-temporal S1 imagery, particularly outside the cleanest stand-replacing regime represented by Russia. Damaged-class F_1 values in the 30–40% range are not a strong basis for operational deployment, even though the same models often localise the broad spatial extent of damage correctly when inspected visually.

Aggregating the same pixel-level outputs into coarser decision units offers a more viable framing. Forest departments typically act on stands, compartments, or coherent damage patches rather than individual 10 m pixels, so false positives on isolated pixels matter less than whether affected areas are flagged reliably enough for prioritisation, inspection, or follow-up mapping. Aggregation also acts as a low-pass filter, suppressing some of the residual speckle and label-noise errors that dominate the per-pixel errors discussed in Section 6.1. The stand-level results in Table 5.8 support this. WI serves as the threshold baseline, and every learned model improves over it on every dataset. The strongest case remains Russia, where SkySense++ bi-temporal reaches 93.1% object-level F_1 against 43.0% for WI, with CNN-UPerNet at 70.7%. Ireland also improves substantially, SkySense++'s results on Sweden still almost triples the WI result, and on Germany CNN-UPerNet is strongest.

SAR-based windthrow mapping is therefore more viable as a stand-level prioritisation tool than as a strict pixel-level segmentation system. The situation where this matters most operationally is also where performance is strongest, in northern boreal storms, where optical sensors are often blocked by cloud cover and limited daylight and rapid ground access is constrained. These are precisely the conditions where SAR's all-weather imaging is very useful and where the learned models deliver stand-level F_1 values high enough to support prioritised inspection and rapid damage assessment.

6.8 Limitations

A general limitation of the experimental design is that each country dataset corresponds to a single biogeographic context, so storm intensity, forest ecosystem, and annotation regime are different within each country and cannot be separated

from the available data. The four datasets should therefore be treated as four separate problems rather than as four replications of a single windthrow-detection task.

The Irish dataset carries additional caveats. DAFM describes the data as provisional and indicative rather than legally definitive. The mapping focused on conifer stands aged 15 years or older, so damage in younger and broadleaf stands is likely under-represented, and the exercise included some older windblown areas estimated at 750–1 000 ha of the national total. Despite these limitations, the dataset is sufficiently reliable and spatially detailed for the purposes of this study.

All metrics depend on a per-tile decision threshold tuned to maximise validation damaged-class F_1. We chose F_1 because it balances precision and recall and matches the imbalanced nature of the task, but a different operating point, for example maximising recall at fixed precision, would shift the precision-recall trade-offs in our tables and could change qualitative comparisons between models. The reported numbers should be read as F_1-balanced operating points, not as the only reasonable choice.

The forest mask used to restrict evaluation for Russia, Ireland, and Germany is itself approximate. Damaged-class F_1 on these three datasets is computed only over forest pixels identified by the ESA WorldCover product, which is roughly 90% accurate at the pixel level, so some background pixels are actually forest and vice versa, adding noise to the evaluation. The effect is more pronounced for mIoU, where the background class dominates, than for damaged-class F_1. Sweden is exempt from this limitation since its forest mask was provided directly by Skogsstyrelsen.

Finally, the experimental scope was limited by available compute. The full set of foundation model sweeps spanned two models and four datasets, each with their respective hyperparameter variations, and we did not have the budget to repeat runs with multiple seeds. Single-seed point estimates are reported throughout, and small differences between configurations on the order of one to two F_1 score points are likely within seed variance. We have drawn qualitative conclusions only when the gaps are substantially larger than this scale.

6.9 Future work

The most direct extension of this work is a standardised, multi-location dataset comparable in label quality to Russia but with greater geographic and ecological diversity. The Russian reference is what makes Russia the strongest dataset in our comparison, not the storm itself, and replicating its annotation protocol , with manual delineation of stand-replacing windthrow patches at a consistent minimum mapping unit, applied to multiple events across the boreal and temperate zones, would produce a benchmark on which both classical and pretrained models could be compared without the problem of differing labelling regimes that limits the cross-dataset analysis in this thesis.

The most important methodological direction we identify is a SAR-only

foundation model trained on a large amount of diverse, high-quality SAR pretraining data. Both Galileo [49] and SkySense++ [54] are multi-modal models in which SAR is one of several input modalities and the pretraining objective is dominated by optical reflectance statistics rather than SAR backscatter, so the pretrained features carry only a fraction of the SAR-specific structure a dedicated single-modality model would learn. A SAR-only foundation model trained at a comparable scale, with a pretraining objective tuned to SAR properties (speckle statistics, polarimetric channels, temporal coherence), would allow a direct test of whether the varying bi-temporal advantage we observe is an intrinsic property of the task or an artefact of the pretraining-distribution mismatch. The architectural separation between modality-specific encoders in SkySense++ already points this way, where its dedicated SAR encoder retains representations learned across 27 million images even when optical modalities are absent suggests dedicated SAR pretraining capacity is a probable binding constraint. We regard this as one of the most consequential follow-ups to this work.

A more extensive temporal study of SkySense++ is a natural and self-contained follow-up. The temporal results reported here rest on a limited configuration search, constrained by the model's large size and the cost of each temporal run, and the SkySense++ temporal pipeline was additionally run without the acquisition-date embeddings that its architecture is designed to exploit (Section 4.2.4). A dedicated study with a broader temporal-depth and learning-rate sweep, multiple seeds, and the date-specific positional encoding activated would establish whether the absence of a temporal advantage observed here is intrinsic to the model on this task or an artefact of the restricted search. Given that the FusionTransformer preserves temporal ordering, this is the single experiment most likely to reveal a genuine event-localisation benefit if one exists.

Polarimetric and interferometric extensions are also relevant. S1 acquires Single Look Complex products that retain dual-polarisation amplitude and phase, while the Ground Range Detected (GRD) products used here discard the phase. We therefore already use the full polarimetric content of GRD in the form of VV and VH channels, and the natural extension is to recover the discarded phase. Interferometric coherence loss across a windthrow event is a textbook indicator of canopy disruption that the GRD pipeline cannot access. Neither Galileo nor SkySense++ ingests phase or interferometric coherence, so adding a coherence channel would not be a drop-in extension but would require either a SAR-specific foundation model pretrained on coherence data or a from-scratch architecture that accepts coherence alongside backscatter. Even as an additional coherence channel concatenated with VV and VH in a CNN baseline.

A final direction is fusion with optical data. Once cloud cover lifts, typically three to four weeks post-storm in northern Europe, S2 NDVI delta provides a strong confirmation signal for windthrow. A natural pipeline is a SAR detector for the cloud-blocked early window (the regime targeted in this thesis) followed by an optical confirmation step once S2 becomes usable. The same logic extends to the multi-modal foundation models studied here, where both can in principle ingest optical and SAR streams in their native configuration, and a late-fusion variant of the temporal pipeline that adds S2 acquisitions where available is a

natural extension that would exploit the full input of these models.

7 Conclusion

This thesis evaluated whether remote sensing foundation models offer a meaningful operational advantage over established baselines for SAR-based windthrow detection across four European country datasets. The results give a nuanced but consistent answer, that foundation models can deliver decisive gains, but only where the underlying label and event regime is clean enough for those gains to surface, and the operationally useful conclusions are tied more to evaluation unit and temporal configuration than to the model alone.

The clearest positive finding is SkySense++ on Russia. With 90.1% pixel-level damaged-class F_1 and 93.1% object-level F_1, it sits roughly 28 percentage points above the CNN-UPerNet and 31 points above Galileo on the same dataset, and reaches an accuracy band that is operationally viable for essentially any windthrow task. The Russia regime, stand-replacing windthrow in boreal forest annotated through manual delineation, also represents precisely the condition where SAR's all-weather imaging is most indispensable, since northern boreal storms occur under persistent cloud cover that block optical acquisition for weeks. At this combination of model, dataset, and operational setting, remote sensing foundation models clearly justify their cost.

Outside that regime, pixel-level segmentation at 10 m resolution from Sentinel-1 is difficult. Damaged-class F_1 values in the 15–45% range, which characterise most country-model combinations on Sweden, Ireland, and Germany, are too low to support a strict per-pixel operational deployment. The dominant bottleneck is not model capacity but label fidelity, event geometry, and the signal-to-noise ratio of bi-temporal C-band backscatter on smaller, more diffuse damage patches, with the consequence that more training data within the same domain offers diminishing returns once label quality is binding.

Aggregating pixel-level outputs to the stand level reframes the problem and improves operational viability. Every learned model improves over the threshold baseline on every dataset under object-level evaluation, and the improvements are often large. SkySense++ rises substantially on Ireland, reaching 57.7% object-level F_1, while the Galileo bi-temporal model reaches 67.2% on Russia. Forestry managers act on stands rather than individual 10 m pixels, and stand-level aggregation acts as a natural low-pass filter that suppresses residual speckle and label-noise errors. SAR-based windthrow mapping is therefore more credibly positioned as a stand-level prioritisation tool than as a strict pixel-level segmentation system, with the exception of the Russia-SkySense++ combination, where

pixel-level deployment is also defensible.

The temporal experiments add a complementary operational result. Galileo's temporal pipeline lifts damaged-class F_1 on Sweden and Ireland, with the depth-sensitivity curve flattening around $T = 8$ acquisitions. With Sentinel-1's effective revisit time of roughly three days, $T = 8$ acquisitions arranged as four pre-event and four post-event corresponds to a post-event window of approximately twelve days under mixed-geometry passes. The plateau at $T = 8$ therefore aligns closely with the natural SAR-only operational regime, where the cloud-blocked window before Sentinel-2 becomes reliably usable, during which the temporal Galileo pipeline coupled with stand-level aggregation provides a strong early-warning detector, most clearly demonstrated on Russia and Ireland.

Cross-dataset generalisation remains the main constraint. Full hold-out transfer collapses across all models, while partial hold-out recovers most of the in-distribution performance once the target country contributes training tiles. The minimum viable deployment scenario is therefore in-country fine-tuning on a modest number of annotated tiles from a recent local storm, rather than zero-shot transfer from a geographically distant training pool. The label-efficiency results indicate that this is a tractable amount of annotation work in practice.

Taken together, the answer to the research question is that foundation models do provide a meaningful advantage for SAR-based windthrow detection, but the magnitude of that advantage is dataset-dependent, the operationally relevant evaluation unit is the stand rather than the pixel outside the cleanest stand-replacing regime, and the most viable deployment is a temporal SAR pipeline at $T \approx 8$ feeding into stand-level aggregation, with light in-country fine-tuning.

Bibliography

[1] Randall Balestriero et al. *A Cookbook of Self-Supervised Learning*. 2023. DOI: 10.48550/ARXIV.2304.12210.

[2] Barcelona Supercomputing Center. *MareNostrum 5*. https://www.bsc.es/marenostrum/marenostrum-5. Accessed: 2026-05-19.

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning". In: (May 2021). DOI: 10.48550/ARXIV.2105.04906. arXiv: 2105.04906 [cs.CV].

[4] Hao Chen, Zipeng Qi, and Zhenwei Shi. "Remote Sensing Image Change Detection with Transformers". In: *IEEE Transactions on Geoscience and Remote Sensing* (2021). DOI: 10.48550/arXiv.2103.00208. arXiv: 2103.00208.

[5] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: (Feb. 2020). DOI: 10.48550/ARXIV.2002.05709. arXiv: 2002.05709 [cs.LG].

[6] Xinlei Chen and Kaiming He. "Exploring Simple Siamese Representation Learning". In: (Nov. 2020). DOI: 10.48550/ARXIV.2011.10566. arXiv: 2011.10566 [cs.CV].

[7] S.R. Cloude and E. Pottier. "A review of target decomposition theorems in radar polarimetry". In: *IEEE Transactions on Geoscience and Remote Sensing* 34.2 (Mar. 1996), pp. 498–518. ISSN: 0196-2892. DOI: 10.1109/36.485127.

[8] Yezhen Cong et al. "SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery". In: (July 2022). DOI: 10.48550/ARXIV.2207.08051. arXiv: 2207.08051 [cs.CV].

[9] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. "Fully Convolutional Siamese Networks for Change Detection". In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. 2018. DOI: 10.48550/arXiv.1810.08462. arXiv: 1810.08462.

[10] Department of Agriculture, Food and the Marine. *Private Forest Wind Damage Assessment Spatial Database – May 2025*. DAFM Open Data Portal. Version 01, accessed 2026-02-20. 2025. URL: https://opendata.agriculture.gov.ie/en_GB/dataset/private-forest-wind-damage-assessment-spatial-database-may-2025.

[11] Department of Housing, Local Government and Heritage. *Review of Storm Eowyn*. Tech. rep. Accessed 2026-04-20. Department of Housing, Local Government and Heritage, 2025. URL: https://assets.gov.ie/static/documents/a4d6b12a/Review_of_Storm_%C3%89owyn_161025_2.pdf.

[12] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: (Oct. 2020). DOI: 10.48550/ARXIV.2010.11929. arXiv: 2010.11929 [cs.CV].

[13] Yuntao Du et al. "SUMMIT: A SAR foundation model with multiple auxiliary tasks enhanced intrinsic characteristics". In: *International Journal of Applied Earth Observation and Geoinformation* 141 (July 2025), p. 104624. ISSN: 1569-8432. DOI: 10.1016/j.jag.2025.104624.

[14] Kathrin Einzmann et al. "Windthrow Detection in European Forests with Very High-Resolution Optical Data". In: *Forests* 8.1 (2017). ISSN: 1999-4907. DOI: 10.3390/f8010021. URL: https://www.mdpi.com/1999-4907/8/1/21.

[15] European Space Agency. *Free access to Copernicus Sentinel satellite data*. Accessed: 2026-02-18. 2013. URL: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Free_access_to_Copernicus_Sentinel_satellite_data.

[16] A. Freeman and S.L. Durden. "A three-component scattering model for polarimetric SAR data". In: *IEEE Transactions on Geoscience and Remote Sensing* 36.3 (May 1998), pp. 963–973. ISSN: 0196-2892. DOI: 10.1109/36.673687.

[17] Anthony Fuller, Koreen Millard, and James Green. "CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders". In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 5506–5538. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/11822e84689e631615199db3b75cd0e4-Paper-Conference.pdf.

[18] Barry Gardiner et al., eds. *Living with Storm Damage to Forests*. What Science Can Tell Us. Editors listed on title page; print ISBN used. European Forest Institute, 2013. ISBN: 978-952-5980-08-0.

[19] GARDINER et al. *Destructive storms in European forests: past and forthcoming impacts*. Jan. 2010. DOI: 10.13140/RG.2.1.1420.4006.

[20] Vivien Sainte Fare Garnot and Loic Landrieu. "Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 4872–4881. DOI: 10.48550/arXiv.2107.07933.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[22] H. Gregow, A. Laaksonen, and M. E. Alper. "Increasing large scale windstorm damage in Western, Central and Northern European forests, 1951–2010". In: *Scientific Reports* 7 (2017), p. 46397. DOI: 10.1038/srep46397.

[23] Jean-Bastien Grill et al. *Bootstrap your own latent: A new approach to self-supervised Learning*. 2020. DOI: 10.48550/ARXIV.2006.07733.

[24] Kaiming He, Ross Girshick, and Piotr Dollár. "Rethinking ImageNet Pre-Training". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 4918–4927.

[25] Kaiming He et al. *Masked Autoencoders Are Scalable Vision Learners*. 2021. DOI: 10.48550/ARXIV.2111.06377.

[26] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: (Dec. 2014). DOI: 10.48550/ARXIV.1412.6980. arXiv: 1412.6980 [cs.LG].

[27] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. "Do Better ImageNet Models Transfer Better?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2661–2671.

[28] Tsung-Yi Lin et al. "Focal Loss for Dense Object Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2020), pp. 318–327. DOI: 10.1109/TPAMI.2018.2858826.

[29] Didier Massonnet et al. "The displacement field of the Landers earthquake mapped by radar interferometry". In: *Nature* 364.6433 (July 1993), pp. 138–142. ISSN: 1476-4687. DOI: 10.1038/364138a0.

[30] Met Office. *Storm Darragh, 6 to 7 December 2024*. Tech. rep. Accessed 2026-04-20. Met Office, 2024. URL: https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/weather/learn-about/uk-past-events/interesting/2024/2024_10_storm_darragh_v1.pdf.

[31] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, Oct. 2016, pp. 565–571. DOI: 10.1109/3dv.2016.79.

[32] Alberto Moreira et al. "A tutorial on synthetic aperture radar". In: *IEEE Geoscience and Remote Sensing Magazine* 1.1 (Mar. 2013), pp. 6–43. ISSN: 2168-6831. DOI: 10.1109/mgrs.2013.2248301.

[33] K.P. Papathanassiou and S.R. Cloude. "Single-baseline polarimetric SAR interferometry". In: *IEEE Transactions on Geoscience and Remote Sensing* 39.11 (2001), pp. 2352–2363. ISSN: 0196-2892. DOI: 10.1109/36.964971.

[34] PERILS AG. *EUR 2,067m - PERILS releases final industry loss estimate for Windstorm Ciarán of November 2023*. PERILS. Nov. 2024. URL: https://www.perils.org/news/eur-2-067m-perils-releases-final-industry-loss-estimate-for-windstorm-ciaran-of-november-2023 (visited on 04/23/2026).

[35] Veenu Rani et al. "Self-supervised Learning: A Succinct Review". In: *Archives of Computational Methods in Engineering* 30.4 (Jan. 2023), pp. 2761–2775. ISSN: 1886-1784. DOI: 10.1007/s11831-023-09884-2.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241. ISBN: 9783319245744. DOI: 10.1007/978-3-319-24574-4_28.

[37] P.A. Rosen et al. "Synthetic aperture radar interferometry". In: *Proceedings of the IEEE* 88.3 (Mar. 2000), pp. 333–382. ISSN: 1558-2256. DOI: 10.1109/5.838084.

[38] Mohammad Rostami et al. "SAR Image Classification Using Few-Shot Cross-Domain Transfer Learning". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2019, pp. 907–915. DOI: 10.1109/cvprw.2019.00120.

[39] Marius Rüetschi, David Small, and Lars T. Waser. "Rapid Detection of Windthrows Using Sentinel-1 C-Band SAR Data". In: *Remote Sensing* 11.2 (Jan. 2019), p. 115. ISSN: 2072-4292. DOI: 10.3390/rs11020115.

[40] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. "Tversky loss function for image segmentation using 3D fully convolutional deep networks". In: *CoRR* abs/1706.05721 (2017). arXiv: 1706.05721. URL: http://arxiv.org/abs/1706.05721.

[41] scikit-learn developers. *7.3. Preprocessing data — scikit-learn 1.8.0 documentation*. scikit-learn. 2026. URL: https://scikit-learn.org/stable/modules/preprocessing.html (visited on 04/14/2026).

[42] Rupert Seidl et al. "Increasing forest disturbances in Europe and their impact on carbon storage". In: *Nature Climate Change* 4.9 (2014), pp. 806–810. ISSN: 1758-6798. DOI: 10.1038/nclimate2318. URL: https://doi.org/10.1038/nclimate2318.

[43] A. Shikhov et al. *A satellite-derived database for stand-replacing windthrow events in boreal forests of European Russia in 1986–2017*. Version 6. figshare, 2020. DOI: 10.6084/m9.figshare.12073278.v6. URL: https://doi.org/10.6084/m9.figshare.12073278.v6.

[44] Andrey N. Shikhov et al. "A satellite-derived database for stand-replacing windthrow events in boreal forests of European Russia in 1986–2017". In: *Earth System Science Data* 12.4 (Dec. 2020), pp. 3489–3513. ISSN: 1866-3516. DOI: 10.5194/essd-12-3489-2020.

[45] Skogsstyrelsen. *Stormen Johannes kan ha orsakat största stormskadorna på tio år*. Pressmeddelande, accessed 2026-04-20. Jan. 2026. URL: https://www.skogsstyrelsen.se/nyhetslista/stormen-johannes-kan-ha-medfort-storsta-stormskadorna-pa-tio-ar/.

[46] Mihai A. Tanase et al. "Detection of windthrows and insect outbreaks by L-band SAR: A case study in the Bavarian Forest National Park". In: *Remote Sensing of Environment* 209 (2018), pp. 700–711. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2018.03.009. URL: https://www.sciencedirect.com/science/article/pii/S0034425718301020.

[47] Erkki Tomppo et al. "Detection of Forest Windstorm Damages with Multitemporal SAR Data—A Case Study: Finland". In: *Remote Sensing* 13.3 (Jan. 2021), p. 383. ISSN: 2072-4292. DOI: 10.3390/rs13030383.

[48] Ramon Torres et al. "GMES Sentinel-1 mission". In: *Remote Sensing of Environment* 120 (2012), pp. 9–24. DOI: 10.1016/j.rse.2011.05.028. URL: https://doi.org/10.1016/j.rse.2011.05.028.

[49] Gabriel Tseng et al. "Galileo: Learning Global and Local Features of Many Remote Sensing Modalities". In: (Feb. 2025). DOI: 10.48550/ARXIV.2502.09356. arXiv: 2502.09356 [cs.CV].

[50] Gaia Vaglio Laurin et al. "Satellite open data to monitor forest damage caused by extreme climate-induced events: a case study of the Vaia storm in Northern Italy". In: *Forestry: An International Journal of Forest Research* 94.3 (July 2021), pp. 407–416. ISSN: 0015-752X. DOI: 10.1093/forestry/cpaa043. eprint: https://academic.oup.com/forestry/article-pdf/94/3/407/37775265/cpaa043.pdf. URL: https://doi.org/10.1093/forestry/cpaa043.

[51] Ashish Vaswani et al. "Attention Is All You Need". In: (June 2017). DOI: 10.48550/ARXIV.1706.03762. arXiv: 1706.03762 [cs.CL].

[52] Yi Wang et al. "Self-supervised Learning in Remote Sensing: A Review". In: (June 2022). DOI: 10.48550/ARXIV.2206.13188. arXiv: 2206.13188 [cs.CV].

[53] Zihao Wang and Lei Wu. *Theoretical Analysis of Inductive Biases in Deep Convolutional Networks*. 2023. DOI: 10.48550/ARXIV.2305.08404.

[54] Kang Wu et al. "A semantic-enhanced multi-modal remote sensing foundation model for Earth observation". In: *Nature Machine Intelligence* 7.8 (Aug. 2025), pp. 1235–1249. ISSN: 2522-5839. DOI: 10.1038/s42256-025-01078-8.

[55] Tete Xiao et al. "Unified Perceptual Parsing for Scene Understanding". In: *arXiv preprint arXiv:1807.10221* (2018). URL: https://arxiv.org/pdf/1807.10221.

[56] Yi Yang et al. “SAR-KnowLIP: Towards Multimodal Foundation Models for Remote Sensing”. In: (Sept. 2025). DOI: 10.48550/ARXIV.2509.23927. arXiv: 2509.23927 [cs.CV].

[57] Chaojie Yu et al. “Deep Learning-Based Change Detection in Remote Sensing: A Comprehensive Review”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18 (2025), pp. 24415–24437. ISSN: 2151-1535. DOI: 10.1109/jstars.2025.3605170.

[58] Jure Zbontar et al. “Barlow Twins: Self-Supervised Learning via Redundancy Reduction”. In: (Mar. 2021). DOI: 10.48550/ARXIV.2103.03230. arXiv: 2103.03230 [cs.CV].

[59] Xiao Xiang Zhu et al. *Deep Learning Meets SAR*. 2020. DOI: 10.48550/ARXIV.2006.10027.