AI4EO: Task-Specific and Foundation Models

Project AI Report on Visiting Researcher period at ESA ϕ-lab (February - April)

Eva Claire Virginie Gmelich Meijling

University of Amsterdam
MSc Artificial Intelligence

Supervisors: Arnoud Visser, Roberto Del Prete, Tobia Armando La Marca

Abstract

This report presents three complementary research contributions to the field of AI for Earth Observation (AI4EO). First, a semantic segmentation framework was developed for mapping vegetation classes in dynamic wetland environments using optical satellite imagery. The methodology integrated self-supervised pretraining with supervised learning, achieving up to 88.23% accuracy on high-resolution Pleiades NEO imagery and reducing reliance on annotated data. Second, a benchmark evaluation of the TerraMind geospatial foundation model was conducted for flood and vessel detection from Sentinel-1 Synthetic Aperture Radar (SAR) data. Vessel detection was reframed from bounding boxes to segmentation maps, and the effect of a custom binary Seesaw Loss was assessed using scale-aware metrics. Although it showed improved Relaxed F1 scores, it did not consistently outperform Dice Loss across standard evaluation metrics. Third, a regression-based approach to small object detection in SAR imagery was proposed, showing promising results by incorporating object scale and aiming to condition the model to better account for small vessels during training. Together, these contributions address common EO challenges in label scarcity, domain adaptation and detection of small or sparse targets, while highlighting the performance of foundation models and alternative loss formulations in real-world EO applications.

1 Introduction

The work conducted during my three-month visiting researcher position at the European Space Agency ϕ-lab (ESA ESRIN) built upon the initially proposed plan submitted for this *Project AI*, before my arrival at ϕ-lab. The initial objective was the development of an advanced AI-driven framework for semantic segmentation of wetland ecosystems, specifically focusing on the Biesbosch floodplain in the Netherlands, leveraging both supervised and self-supervised learning paradigms.

While the original project centered on semantic segmentation using optical data, the dynamic research environment at ϕ-lab provided an opportunity to broaden the scope of my research. I expanded from developing a task-specific model toward exploring foundation models and incorporated not only optical but also SAR imagery. Specifically, I extended my research scope by benchmarking a geospatial foundation model for SAR based flood and vessel detection and explored regression based segmentation techniques to enhance small object detection in SAR imagery. Although the research topics evolved, the original goals were fully achieved and even expanded, leading to several deliverables, including datasets and valuable scientific insights for both ϕ-lab and myself, and perhaps most importantly, an amazing learning experience.

This report is structured around three main research directions conducted in collaboration with ESA ϕ-lab. Each is presented with its objective, methodology, results, and deliverables. The first covers semantic segmentation of wetland vegetation. The second, on benchmarking foundation models for SAR-based flood and vessel detection, includes more detailed data processing and methodological steps due to its broader scope. The third explores a regression-based approach to improve small vessel detection in SAR imagery.

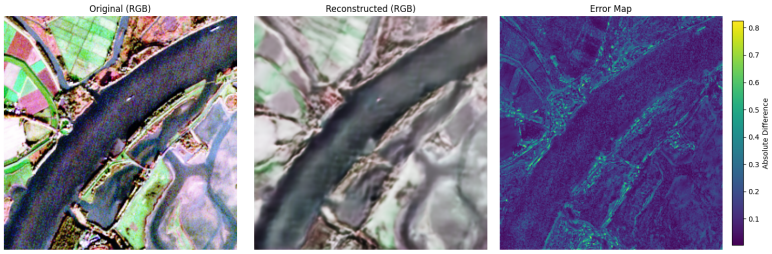
2 Supervised and self-supervised land-cover segmentation & classification of the Biesbosch wetlands

2.1 Objective

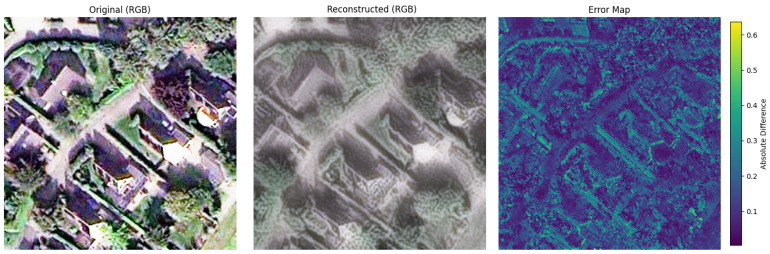
The objective of this research was to develop an AI-based semantic segmentation framework to classify wetland vegetation types in the Biesbosch floodplain using optical satellite imagery. These classifications are important for flood mitigation and ecological monitoring since vegetation roughness directly affects water flow and retention capacity. A key challenge that was tackled and common in the domain of remote sensing, is the scarcity of annotated data, especially for very-high-resolution imagery. To address this, the research aimed to combine supervised semantic segmentation with self-supervised learning to improve model performance under limited labeling conditions.

2.2 Methodology

The model architecture used for this research was a U-Net, a convolutional neural network designed for pixel-wise classification. For the baseline model, the network was trained from scratch on medium-resolution Sentinel-2 imagery (10m per pixel) using Dynamic World land-cover labels [3]. To address the issue of label scarcity, a self-supervised learning (SSL) pipeline was introduced. An autoencoder was trained to reconstruct the original satellite images, such that its trained encoder weights could then be transferred to initialize the U-Net's encoder. The same method was applied for very-high-resolution (VHR) Pleiades NEO imagery (0.3m per pixel) with manually annotated labels. The output of this SSL-based pretraining approach is visually demonstrated in Figure 1. Furthermore, a controlled experiment was set up to assess how both resolution and pretraining impacted performance.



(a) Reconstruction result for medium-resolution Sentinel-2 data.



(b) Reconstruction result for high-resolution Pléiades Neo.

Figure 1: *Reconstruction results from the autoencoder for medium-resolution Sentinel-2 data (a) and high-resolution data (b). The first column shows the histogram-equalized original RGB image. The second column presents the reconstructed image from the autoencoder, while the last column displays the error map, where blue indicates minimal pixel differences and yellow highlights larger discrepancies.*

2.3 Results

The U-Net trained from scratch on Sentinel-2 data, without pretraining, achieved an overall accuracy of 85.26% and a Dice score of 0.648%. When applied to VHR imagery, performance without pretraining yielded only 60.35% accuracy. However, after SSL-based pretraining, accuracy increased significantly to 88.23%, confirming the usefulness of representation learning in low-label, high-resolution scenarios. By comparing the results in Table 1 for the medium-resolution data and Table 2 for the high-resolution

data, it can be seen that the pretraining process was especially helpful for fine-grained high-resolution data

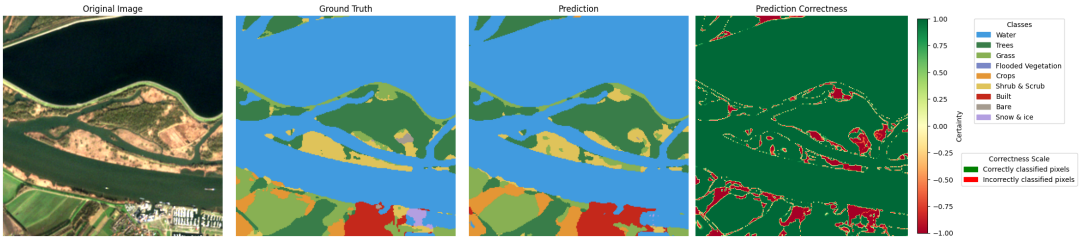
Training method	Accuracy ↑	Dice ↑	IoU ↑	Precision ↑	Recall ↑	Dice Loss ↓
Scratch	0.8526	0.6480	0.5346	0.6616	**0.6694**	**0.4865**
Pretrained	**0.8542**	**0.6518**	**0.5378**	**0.6923**	0.6483	0.4905

Table 1: *Land-use classification performance for the U-Net model on the Sentinel-2 dataset with and without pretraining. The U-Net was trained for 300 epochs on the dataset. The pretrained model started with weights extracted from the autoencoder.*

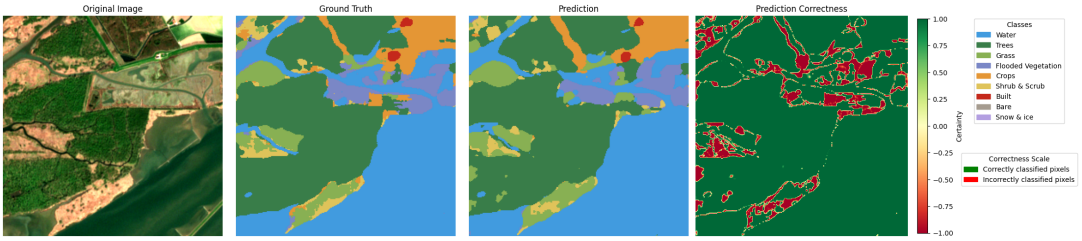
Training method	Accuracy ↑	Dice ↑	IoU ↑	Precision ↑	Recall ↑	Dice Loss ↓
Scratch	0.6035	0.2827	0.2243	0.3889	0.3158	**0.5114**
Pretrained	**0.8823**	**0.4457**	**0.3919**	**0.5079**	**0.4551**	0.5457

Table 2: *Performance comparison of non-pretrained and pretrained U-Net on high-resolution imagery and labels.*

Figure 2 illustrates the qualitative improvements in segmentation accuracy when using a pretrained model compared to a model trained from scratch on medium-resolution data and labels. More clearly, Figure 3 shows the enhanced precision and detail provided by high-resolution imagery compared to medium-resolution imagery, highlighting its effectiveness for detailed ecological applications such as detecting small objects like the ship, and transitions between mixed classes.



(a) Land-cover classification for Sentinel-2 data trained from scratch.



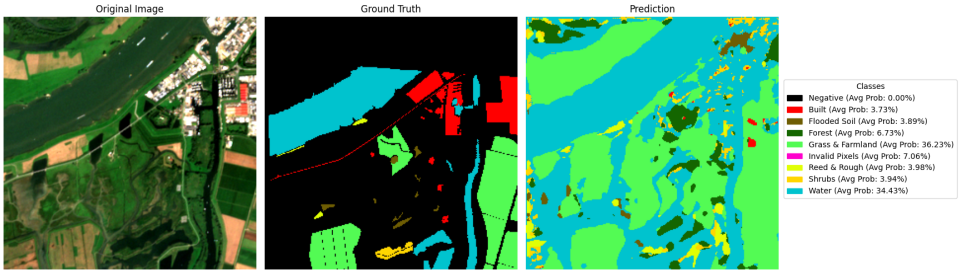
(b) Land-cover classification for Sentinel-2 data with a pretrained model.

Figure 2: *Land-use classification with U-Net for medium-resolution Sentinel-2 data with a model trained from scratch (a) and pretrained (b). The first column shows the histogram-equalized original RGB image. The second column presents the ground truth classification by Dynamic World. The third column shows the prediction of the U-Net. The last column displays the error map, where red represents misclassified pixels, and the color intensity reflects the model's certainty.*

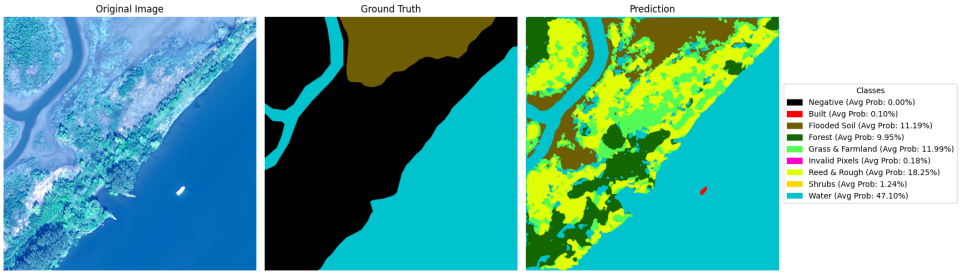
2.4 Deliverables

This research produced several outputs. First of all, a publicly available dataset of Sentinel-2 imagery with Dynamic World labels was released on Zenodo[1]. The annotated Pleiades NEO dataset is available

[1] `https://doi.org/10.5281/zenodo.15125549`



(a) Medium-resolution segmentation results



(b) High-resolution segmentation results

Figure 3: *Comparison of medium-resolution and high-resolution segmentation results. While performance metrics appear similar, high-resolution imagery (b) provides finer details and more precise segmentation. The legend shows the average predicted probability for each class across the image, highlighting model confidence in class assignments.*

upon request. All source code for the segmentation pipeline, including training and evaluation scripts, is hosted at `https://github.com/Evameijling/WetlandSemanticSegmentation` and includes full documentation and usage instructions. A shortened version[2] of the thesis summarizing the methodology and findings was submitted for the NCCV 2025 conference in Utrecht, The Netherlands.

3 Benchmarking Foundation Models for SAR-based Flood and Vessel Detection

3.1 Objective

This research aimed to evaluate the TerraMind geospatial foundation model [2], developed collaboratively by IBM and ESA, for its suitability in detecting floods and small vessels in synthetic aperture radar (SAR) data. Foundation models are pretrained on large multimodal datasets and are expected to generalize well across downstream tasks. In this context, the goal was to assess how well TerraMind could adapt to SAR imagery and whether it offered improvements over conventional task-specific models when applied to Earth Observation tasks. Figure 4 illustrates the architecture of the TerraMind foundation model, highlighting its multimodal design and training approach.

This benchmarking effort was part of a broader research direction focused on evaluating foundation model performance across different processing levels of Sentinel-1 SAR data. Namely, RAW (Level-0), SLC, and GRD products, with the long-term goal to achieve robust detection performance directly on lower-level SAR inputs. This could potentially reduce reliance on computationally expensive preprocessing and would allow faster, lower-latency applications (e.g., onboard AI systems). Figure 5 illustrates this multi-level evaluation strategy, showing how the TerraMind model is fine-tuned separately on RAW, SLC, and GRD SAR inputs for downstream tasks such as flood and vessel detection.

[2] `https://staff.fnwi.uva.nl/a.visser/publications/SegmentationClassificationBiesboschWetlands.pdf`

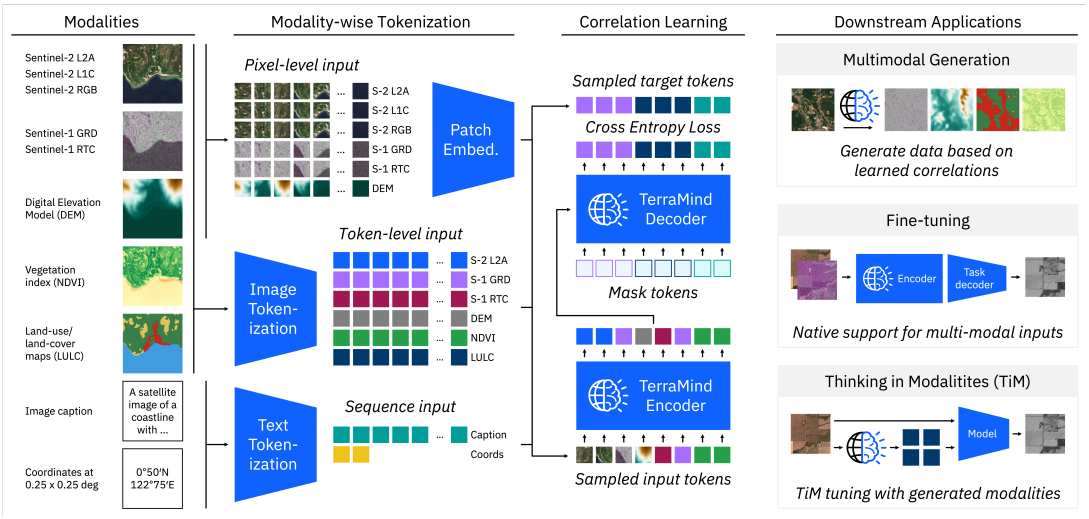


Figure 4: *Architecture of the TerraMind foundation model, reproduced from Jakubik et al. [2]. The model integrates nine EO modalities, including Sentinel-1 SAR data (GRD, RTC), Sentinel-2 optical data (L1C, L2A, RGB), DEM, NDVI, LULC, image captions, and coordinates. Inputs are tokenized at both pixel-level and token-level before being processed through a masked correlation learning objective. Associated geolocation coordinates are tokenized and included as sequence input. TerraMind supports a range of downstream EO tasks, including flood and vessel detection, through fine-tuning, multimodal generation, and inference using Thinking-in-Modalities (TiM).*

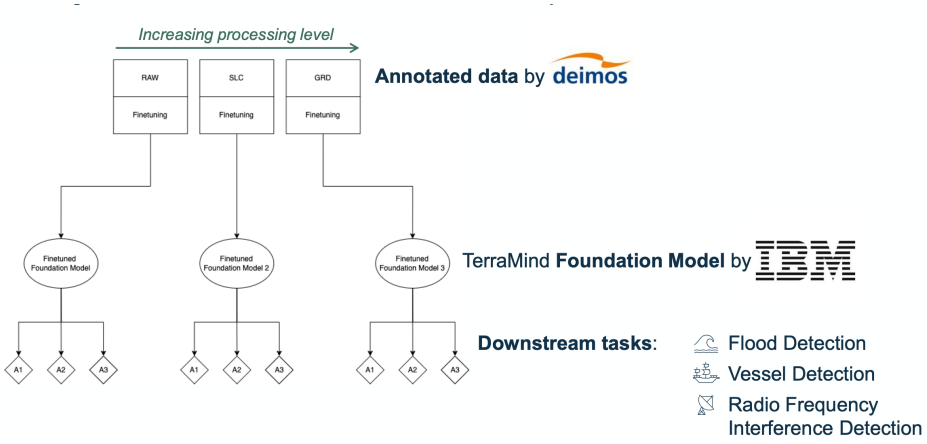


Figure 5: *Schematic overview of the broader research direction underlying this work. The TerraMind foundation model is fine-tuned separately on different SAR processing levels (RAW, SLC, GRD), with each version adapted for downstream tasks such as flood detection, vessel detection, and radio frequency interference (RFI) detection.*

3.2 Data Quality Analysis of the Deimos GRD Dataset

One of the preparatory steps in this research was the inspection of annotation quality in the GRD SAR dataset provided by Deimos. To assist this process, I visualized randomly sampled SAR images side-by-side with their corresponding segmentation masks (as will be discussed in the methodology section; it was an iterative process), derived from XML annotation files. This visual approach helped to identify inconsistencies between the imagery and the provided labels.

Through this visual analysis and manual review of the files and dataset structure, several issues were discovered and compiled into a dedicated internal report, shared with ESA ϕ-lab. Notable findings included:

- **Duplicate annotation files:** One of the GRD scenes had no unique annotation because its file was an exact duplicate of another, resulting in a missing label entry.

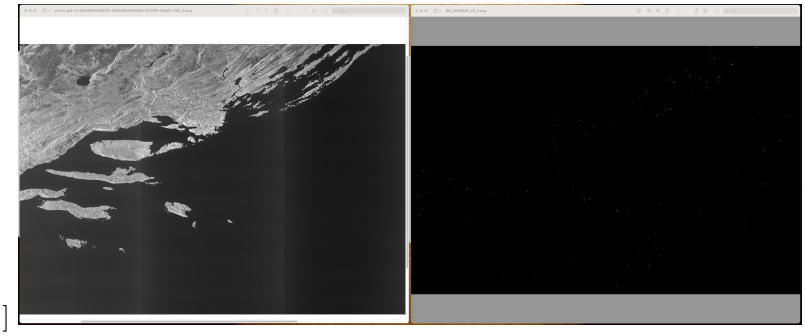


Figure 6: *Example of visual inspection: SAR image (left) and segmentation mask derived from XML annotations (right). Discrepancies between white radar signatures and labeled vessel locations were used to assess label quality.*

- **Invalid bounding boxes:** Several ships had undefined or non-numeric bounding box coordinates, resulting in unusable labels.
- **Mislabeled features:** Some annotated regions aligned poorly with the SAR image content, either omitting prominent bright features or labeling land structures (e.g., coastal islands) as vessels.
- **Coordinate logic errors:** In some SLC annotations, the bounding box `Top` coordinate was greater than the `Bottom` coordinate, contradicting expected image coordinate conventions.

3.3 Methodology

Due to the scope and complexity of this study, the methodology is subdivided into four parts: task reformulation, training setup, evaluation metrics, and benchmarking procedures.

3.3.1 Reframing Detection as Segmentation

Since TerraMind's architecture is optimized for dense token-level predictions over spatial patches, the vessel detection problem was reformulated from a traditional object detection setup into a semantic segmentation task. Georeferenced binary segmentation masks were generated by applying multiple thresholding methods to SAR image patches centered on annotated vessels, followed by morphological operations such as dilation and erosion to refine the shape. Overlapping results were combined to produce the final binary segmentation maps. This allowed the model to predict dense vessel presence maps, aligning more naturally with its pretraining objective. Figure 7 visualizes this pipeline of transforming bounding box annotations into georeferenced binary segmentation maps.

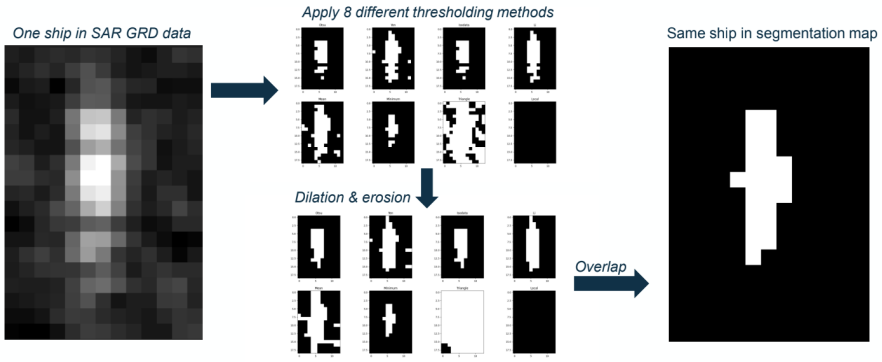


Figure 7: *Transformation from bounding box annotations (XML) to binary segmentation maps for SAR data, enabling semantic segmentation evaluation. The example is based on the VH polarization band, which was chosen due to its improved contrast between vessel signatures and background clutter in the Deimos dataset.*

While Figure 7 shows a representative example of the transformation from bounding boxes to segmentation masks, it should be noted that no quantitative metric could be computed to assess the accuracy of this transformation, as no ground truth segmentation masks were available. The original bounding boxes were the only provided labels, so validation was instead carried out through visual inspection by comparing the resulting binary masks with the underlying SAR images.

These inspections indicated that, in most cases, the generated masks correctly overlapped with vessel signatures in the SAR VH-band imagery. However, challenges did arise in scenes with strong wave interference or surface clutter, where radar backscatter around the vessel could lead to over-segmentation. This effect explains why some masks deviate from the expected rectangular boat shapes, and instead contain irregular or elongated regions. Such ambiguities, although inherent to radar imagery, were reduced by focusing on the VH polarization band, which empirically provided better vessel-background contrast than the VV band in the dataset used. The VH channel was therefore selected as the input for segmentation mask generation throughout this study.

3.3.2 Training Setup and Loss Function Adaptation

TerraMind was fine-tuned on the generated segmentation maps using a dedicated, non-public SAR (GRD) dataset provided by Deimos[3]. For comparison, a baseline U-Net model was trained from scratch under similar experimental conditions.

A challenge identified was the class imbalance and small size of the objects (i.e., vessels) relative to the background. To address this, a custom binary version of the Seesaw Loss [4] was implemented, where a mitigation and compensation factor is used to adjust the contribution of easy and hard examples dynamically. Originally designed for multi-class instance segmentation, Seesaw Loss was adapted to a binary semantic segmentation context, incorporating class imbalance-awareness on a pixel level instead of an instance leve, by penalizing dominant background predictions and emphasizing minority class recall.

In this binary version, logits for vessel and background classes were normalized, and the loss emphasized penalizing dominant background predictions while promoting vessel detections.

The detailed mathematical reformulation of the Binary Seesaw Loss is as follows:

$$\textbf{Logits: } z_{v,i},\, z_{b,i} \quad \text{(for vessel and background at pixel } i\text{)}.$$

$$\hat{p}_{v,i} = \frac{\exp(z_{v,i})}{\exp(z_{v,i}) + \exp(z_{b,i})}, \quad \hat{p}_{b,i} = \frac{\exp(z_{b,i})}{\exp(z_{v,i}) + \exp(z_{b,i})}.$$

$$\textbf{Ground truth labels: } y_i \in \{0, 1\}, \quad \text{where } y_i = 1 \text{ for vessel and } y_i = 0 \text{ for background.}$$

$$\textbf{Binary Seesaw Loss: } \quad L_{\text{binary\_seesaw}} = -\frac{1}{|D|} \sum_{i \in D} [y_i \log(\hat{p}_{v,i}) S_{v,i} + (1 - y_i) \log(\hat{p}_{b,i}) S_{b,i}]$$

where $S_{v,i}$ and $S_{b,i}$ are seesaw multipliers based on class frequency and prediction ratios; they are the product of mitigation and compensation factors weighting the pixel's contribution to the loss. Further details can be found in the original Seesaw Loss paper [4].

3.3.3 Evaluation Metrics

To better evaluate segmentation performance, particularly for sparse and small targets, two task-specific evaluation metrics were introduced: the relaxed F1 score (allowing small spatial tolerance) and the Scale-adaptive Intersection over Union (SIoU), which adjusts overlap sensitivity based on object size and is better suited for evaluating thin or sparse structures [4, 1].

SIoU was originally proposed for small object detection in bounding box regression tasks [4]. It introduces a dynamic scaling factor that reduces the sensitivity of the IoU metric to small localization errors for small-sized objects, thereby mitigating high variance in evaluation.

The original SIoU for two bounding boxes b_1 and b_2 is defined as:

$$\text{SIoU}(b_1, b_2) = (\text{IoU}(b_1, b_2))^p, \quad p = 1 - \gamma \exp\left(-\frac{\sqrt{w_1 h_1 + w_2 h_2}}{\sqrt{2\kappa}}\right) \tag{1}$$

[3] `https://deimos-space.com`

where w_1, h_1 and w_2, h_2 are the width and height of the two bounding boxes, γ is a scaling parameter (typically 0.5), and κ controls the rate of scaling with respect to object size.

Since in semantic segmentation no bounding boxes are used, I reformulated SIoU for binary segmentation tasks. In this adaptation, the object size is approximated based on confusion matrix components:

- $A = \text{TP} + \text{FN}$: number of positive pixels in the ground truth,
- $B = \text{TP} + \text{FP}$: number of positive pixels in the prediction,
- $s = A + B$: total mass of positives,
- $s_{\text{norm}} = \frac{s}{\text{Total Pixels}}$: normalized size measure.

The scaling factor p is then computed as:

$$p = 1 - \gamma \exp\left(-\sqrt{\frac{s_{\text{norm}}}{2\kappa}}\right) \tag{2}$$

and the final Scale-adaptive IoU is calculated as:

$$\text{SIoU}(A, B) = \left(\frac{|A \cap B|}{|A \cup B|}\right)^p \tag{3}$$

Furthermore, a relaxed F1 score was applied to account for minor localization inaccuracies in the segmentation results. This metric was implemented by applying a morphological dilation to the ground truth mask prior to evaluation, allowing predictions within a small spatial tolerance to be considered correct. Specifically, the ground truth mask was dilated using a max-pooling operation with a kernel size of $2 \times \text{tolerance} + 1$ pixels. In these experiments, a tolerance of 2 pixels was used.

3.3.4 Benchmarking Setup

In addition to the Deimos SAR dataset, the performance of the TerraMind foundation model was also evaluated on the High-Resolution SAR Imagery Dataset (HRSID) [5], which provides high-quality SAR images with vessel annotations. This was motivated by the deficiencies identified in the Deimos data, as discussed in Section 3.2.

For both datasets, the foundation model's segmentation performance was benchmarked against a task-specific baseline architecture: a ResNeXt50_32x4d-based U-Net pretrained on ImageNet. Multiple loss functions were evaluated, including the adapted Seesaw Loss. This experimental design allowed for a direct comparison between a general-purpose foundation model and a specialized supervised model under equivalent training conditions.

3.4 Results

Fine-tuning TerraMind on SAR datasets resulted in reasonable segmentation performance, particularly when evaluated with relaxed metrics suitable for small target detection. The foundation model was able to converge in relatively few training epochs and demonstrated satisfactory recall for vessels, particularly under relaxed evaluation settings. While the model trained with Seesaw Loss showed competitive performance in some cases, as can be seen in Table 3, it did not consistently outperform models trained with Dice Loss across standard evaluation metrics.

Loss Function	**F1** ↑	**Relaxed F1** ↑	**IoU** ↑	**SIoU** ↑	**mAcc** ↑	**mF1** ↑	**mIoU** ↑
Weighted Cross-Entropy	12.16	34.61	6.47	22.70	**97.65**	55.11	51.34
Dice Loss	**47.44**	45.67	**31.10**	**54.75**	75.75	**73.65**	**65.40**
Focal Loss	41.76	25.26	26.39	50.48	65.79	70.82	63.07
Dice + Focal Loss	35.93	24.22	21.90	45.87	63.67	67.90	60.82
Seesaw Loss	40.45	**58.14**	25.36	49.00	85.64	70.09	62.40

Table 3: *Comparison of model performance metrics (%) for vessel segmentation using different loss functions.*

Nevertheless, when compared against a task-specific ResNeXt50_32x4d-based U-Net architecture pretrained on ImageNet, TerraMind did not consistently achieve superior segmentation results. Table 4 provides a detailed overview of the comparative performance between TerraMind and the U-Net baselines

across the two SAR datasets. On both the Deimos and HRSID datasets, the U-Net baseline exhibited higher F1 scores and intersection-over-union (IoU) values under standard evaluation metrics. This indicates that, despite TerraMind's multimodal pretraining, domain adaptation to SAR-specific dense segmentation tasks remains challenging.

Model	F1 ↑	Relaxed F1 ↑	IoU ↑	SIoU ↑	mAcc ↑	mF1 ↑	mIoU ↑
GRD FM	40.45	**58.14**	25.36	49.00	85.64	70.09	62.40
GRD U-Net	77.64	–	63.45	–	99.92	83.38	71.49
GRD U-Net Pretrained	**83.33**	–	**71.43**	–	**99.96**	**90.59**	**82.79**
HRSID FM	80.55	81.66	67.43	81.32	97.33	90.20	83.57
HRSID U-Net	82.20	–	69.77	–	99.94	**95.34**	**91.10**
HRSID U-Net Pretrained	**96.53**	–	**93.28**	–	**99.94**	94.68	89.89

Table 4: *Performance comparison of TerraMind foundation model (FM) and U-Net baselines on GRD and HRSID datasets using Seesaw Loss. Missing values (–) indicate metrics not computed for that configuration.*

These results show that, while foundation models offer flexibility and transferability, specialized architectures trained directly for the target domain can still outperform foundation models.

It is important to note that the experiments presented here were conducted using an early, non-final version of the TerraMind model and its accompanying codebase. Since that time, further improvements and refinements have been made to the model architecture, training procedures, and release versions. Therefore, the reported results may not fully represent the current capabilities of the finalized TerraMind foundation model [2]. Future evaluations using the latest version are expected to yield improved performance.

3.5 Deliverables

The adapted binary Seesaw Loss implementation was made publicly available on GitHub at `https://github.com/ESA-PhiLab/SegSeeSawLoss`, including documentation and a training example. In addition, a custom Python script was developed to convert XML annotations from both GRD and SLC SAR datasets into COCO format. Although not used in the segmentation pipeline, this conversion enables compatibility with object detection frameworks and supports future experiments on detection-based models.

A dedicated internal report analyzing the quality of the GRD annotation data was also produced. This document highlighted annotation inconsistencies (e.g., duplicate files, NaN bounding boxes, and mislabeled objects) and helped guide the filtering and preparation of the dataset before model training.

Internal benchmarking reports, data transformation pipelines, and configuration files for fine-tuning TerraMind are open to ESA ϕ-lab. A more detailed comparison of model outputs using segmentation visualizations and metric outputs might be included in possible future research.

4 Enhancing Small Vessel Detection via Regression-based Semantic Segmentation

4.1 Objective

The objective of this research stream was to explore an alternative modeling strategy for small object detection in synthetic aperture radar (SAR) imagery. While the previous research approach framed vessel detection as a classification problem using binary segmentation maps, this investigation sought to capture scale and object presence through continuous regression targets. The goal was to improve sensitivity to smaller vessels by using scale-aware spatial outputs rather than discrete class boundaries.

4.2 Methodology

A modified U-Net architecture was employed, in which the output layer was adapted to predict continuous regression values instead of binary class probabilities. The regression targets were derived from binary segmentation masks and encoded both object scale and spatial decay using a Gaussian-like distribution centered on each vessel, as illustrated in Figure 8, on the HRSID dataset. The objective was to produce

soft activation maps where high-confidence regions correspond to probable vessel centers, with decreasing confidence toward the periphery.

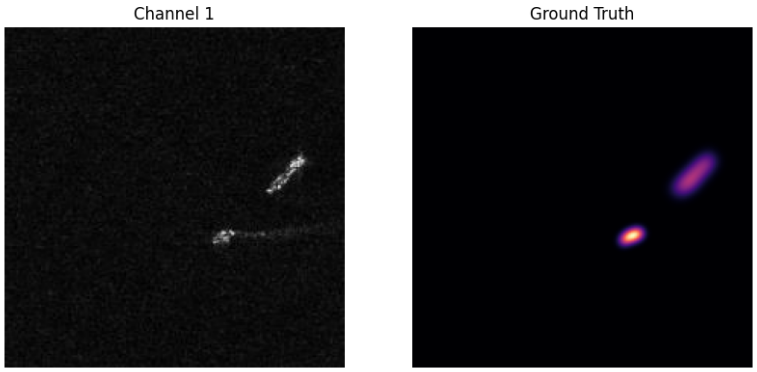


Figure 8: Sample SAR image used for evaluating the regression-based vessel detection approach.

This heatmap representation was controlled by two hyperparameters: the *sigma factor* (σ_{factor}), which determines the spread of the Gaussian by scaling with the square root of the vessel's area, and the *weight exponent* (w_{exp}), which adjusts the central amplitude inversely with vessel size. As a result, small vessels are represented with sharper peaks and narrower spatial spread, while larger vessels are assigned broader, flatter activation profiles. This continuous representation was intended to improve the network's sensitivity to small vessels, which may otherwise be lost in traditional thresholded binary segmentation.

The following pseudocode outlines the process of generating these scale-aware heatmaps from binary vessel masks:

Algorithm 1 Conversion of Binary Vessel Mask to Scale-Aware Heatmap

procedure SegToHeatmap(bin_mask, σ_{factor}, w_{exp}, normalize)
  Label connected components in bin_mask as *labels*, count n
  Initialize *heatmap* $\leftarrow$ zero matrix of same shape as bin_mask
  for *comp_id* $\leftarrow 1$ to n **do**
    comp $\leftarrow$ pixels belonging to component *comp_id*
    $A \leftarrow$ area of *comp*
    if $A = 0$ **then continue**
    end if
    weight $\leftarrow A^{-w_{\text{exp}}}$
    $\sigma \leftarrow \sigma_{\text{factor}} \cdot \sqrt{A}$
    $d \leftarrow$ normalized distance transform of *comp*
    blurred $\leftarrow$ Gaussian blur of $(comp \cdot d)$ with σ
    heatmap $\leftarrow$ *heatmap* + *weight* $\cdot$ *blurred*
  end for
  if normalize **then**
    Rescale *heatmap* to range $[0, 1]$
  end if
  return *heatmap*
end procedure

A modified U-Net architecture was implemented to regress continuous heatmaps from SAR image input, using the ground truth heatmaps as described earlier. The predicted output of the U-Net consists of smooth activation maps that similarly highlight vessel center as the ground truth, as can be seen in Figure 9. To convert these predictions back into binary segmentation masks, a thresholding-and-watershed-based approach was applied. First, low-confidence areas were filtered out using a fixed

probability threshold. Then, local maxima were identified to initialize foreground seeds, and the watershed algorithm was used to delineate object boundaries. This postprocessing strategy allowed for more spatially consistent segmentations and avoided over-fragmentation of small vessel detections.

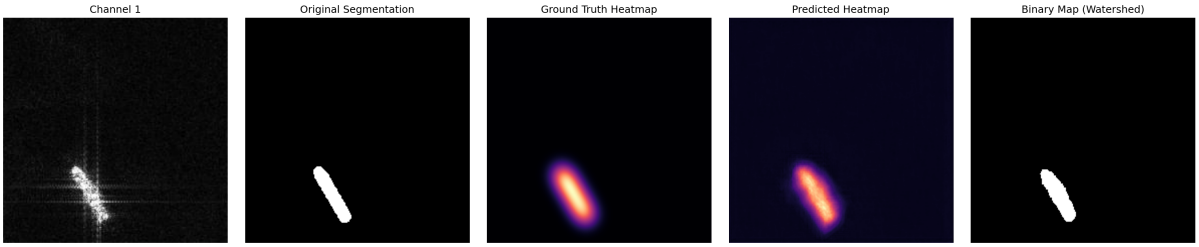


Figure 9: End-to-end output of the regression-based segmentation pipeline. From left to right: input SAR image (from the HRSID dataset), original binary segmentation mask, ground truth regression heatmap, predicted regression heatmap, and final binary mask obtained via thresholding and watershed postprocessing. The predicted heatmaps produce smooth and scale-aware activations that enhance small vessel detection.

4.3 Preliminary Results

Initial experiments demonstrated promising results. The regression-based U-Net exhibited notable sensitivity to very small vessels, with the predicted heatmaps capturing smooth, localized activations that aligned well with vessel centers. As illustrated in Figures 8 and 9, the continuous output produced by the model effectively highlighted targets that might be missed by conventional binary segmentation.

However, due to the characteristics of the HRSID dataset (specifically, the lack of scenes containing both small and large vessels within the same tile) it was not possible to rigorously benchmark the regression approach against traditional binary segmentation methods in terms of scale-aware performance. This limited the ability to quantify improvements in detecting smaller or underrepresented vessel types. Future work will address this limitation by evaluating the regression-based approach on more suitable datasets, such as xView3-SAR and SSDD, which offer greater diversity in ship size and context.

4.4 Deliverables

The codebase for the regression-based segmentation pipeline is currently under development and will be published in a dedicated GitHub repository: `https://github.com/Evameijling/`. The repository will include code for generating regression targets, training scripts, and evaluation metrics customized for small object detection in SAR.

5 Conclusion and Reflection

During the three-month research period, I achieved all of the originally planned deliverables and was able to extend the work in two new directions. The main objective, which was to develop a semantic segmentation model for wetland vegetation, was completed successfully and documented, and the results were submitted to the NCCV 2025 conference. In addition, the dynamic and collaborative environment at ESA ϕ-lab created the opportunity to explore two additional research topics: benchmarking a geospatial foundation model for SAR-based detection and developing a regression-based approach for identifying small vessels in SAR imagery.

While the operational deployment of the wetland segmentation model will be carried out by FREE Nature and Accenture, the full AI pipeline is publicly available via GitHub at `https://github.com/Evameijling/WetlandSemanticSegmentation`, supporting open research and reproducibility.

The three-month period at ESA ϕ-lab in ESRIN (Frascati) was an incredibly valuable learning experience. It gave me the chance to strengthen my thesis work with input from experts and also to dive into new areas I hadn't worked with before, especially foundation models and SAR data. Weekly research meetings in the Explore Office, combined with informal discussions with other researchers, helped me get feedback, share ideas, and learn from ongoing work around me. This experience really helped me grow my knowledge in AI for Earth Observation and gave me a solid foundation to build on in the future.

References

[1] Roberto Del Prete, Manuel Salvoldi, Domenico Barretta, Nicolas Longépé, Gabriele Meoni, Arnon Karnieli, Maria Daniela Graziano, and Alfredo Renga. Enhancing maritime situational awareness through end-to-end onboard raw data analysis. *arXiv preprint arXiv:2411.03403*, 2024.

[2] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. *arXiv preprint arXiv:2504.11171*, 2025.

[3] Zander S Venter, David N Barton, Tirthankar Chakraborty, Trond Simensen, and Geethen Singh. Global 10 m land use land cover datasets: A comparison of dynamic world, world cover and esri land cover. *Remote Sensing*, 14(16):4101, 2022.

[4] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9695–9704, 2021.

[5] Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation. *Ieee Access*, 8:120234–120254, 2020.