



## OPEN ACCESS

EDITED BY  
Kutubuddin Ansari,  
Karadeniz Technical University, Türkiye

REVIEWED BY  
Afsheen Ameer,  
Warsaw University of Technology, Poland  
Maqusud Alam,  
Kyungpook National University, Republic  
of Korea

\*CORRESPONDENCE  
Soufyane Bouchelaghem  
✉ [sxb9983@miami.edu](mailto:sxb9983@miami.edu)

RECEIVED 10 December 2025  
REVISED 11 February 2026  
ACCEPTED 28 February 2026  
PUBLISHED 31 March 2026

CITATION  
Bouchelaghem S, Mamen L, Balsi M,  
Tibermacine A, Moroni M and  
Tibermacine IE (2026) Binary  
reformulation for marine debris  
detection in Sentinel-2 imagery: an  
empirical study on extreme class  
imbalance using the first benchmarks on  
combined MARIDA and MADOS datasets.  
*Front. Mar. Sci.* 13:1765021.  
doi: 10.3389/fmars.2026.1765021

COPYRIGHT  
© 2026 Bouchelaghem, Mamen, Balsi,  
Tibermacine, Moroni and Tibermacine.  
This is an open-access article distributed  
under the terms of the [Creative  
Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).  
The use, distribution or reproduction in  
other forums is permitted, provided the  
original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which does  
not comply with these terms.

# Binary reformulation for marine debris detection in Sentinel-2 imagery: an empirical study on extreme class imbalance using the first benchmarks on combined MARIDA and MADOS datasets

Soufyane Bouchelaghem<sup>1,2\*</sup>, Lahcene Mamen<sup>3</sup>, Marco Balsi<sup>2</sup>,  
Ahmed Tibermacine<sup>3</sup>, Monica Moroni<sup>4</sup>  
and Imad Eddine Tibermacine<sup>5</sup>

<sup>1</sup>Aircraft Center for Earth Studies (ACES), Rosenstiel School of Marine, Atmospheric, and Earth Sciences, University of Miami, Miami, FL, United States, <sup>2</sup>Department of Information Engineering, Electronics and Telecommunications (DIET), Sapienza University of Rome, Rome, Italy, <sup>3</sup>Department of Computer Science, University of Biskra, Biskra, Algeria, <sup>4</sup>Department of Civil, Building, and Environmental Engineering (DICEA), Sapienza University of Rome, Rome, Italy, <sup>5</sup>Department of Computer, Automation and Management Engineering (DIAG), Sapienza University of Rome, Rome, Italy

**Introduction:** Marine debris detection from satellite imagery is challenged by two major factors: extreme class imbalance, with debris pixels accounting for less than 0.01% of image content, and the need for robust generalization across diverse geographic and temporal domains for operational deployment. Although existing methods often report strong within-dataset performance, cross-dataset generalization, where models trained on one dataset are applied to entirely different geographic regions, remains insufficiently investigated.

**Methods:** To address this limitation, we conducted rigorous bidirectional cross-dataset validation experiments using the MARIDA and MADOS datasets. The problem was reformulated as a binary segmentation task and addressed using a standard U-Net architecture combined with a composite imbalance-aware loss and a rarity-aware sampling strategy. Two experimental settings were considered: training on MARIDA and testing on MADOS, and training on MADOS and testing on MARIDA.

**Results:** The experiments revealed asymmetric cross-dataset generalization. Models trained on the geographically diverse MADOS dataset achieved an F1-score of 0.890 when tested on MARIDA, corresponding to only a 1.25% decrease from the within-dataset baseline of 0.901. In contrast, models trained on MARIDA achieved an F1-score of 0.833 on MADOS, representing a 7.55% decrease. The average cross-dataset degradation was 4.38%, which is substantially lower than the typical 10–25% performance drops reported in remote sensing domain-shift scenarios. Despite comparable patch counts (2,529 for MADOS versus 2,173 for MARIDA), the superior transferability of MADOS-trained models indicates that geographic diversity across globally distributed tiles is more beneficial than exhaustive annotation within concentrated regions. Moreover, the MADOS-to-MARIDA cross-dataset F1-score of 0.890 exceeded MAP-Mapper's within-dataset

F1-score of 0.880 and closely approached MariNeXt's reported performance of 0.891.

**Discussion:** These findings show that careful data formulation and training design can enable standard architectures to achieve strong cross-domain performance under extreme class imbalance, approaching or even surpassing more specialized models in realistic deployment conditions. The results provide practical guidance for operational marine debris monitoring systems: spatially stratified sampling across diverse marine environments should be prioritized, F1-scores in the range of 0.86–0.89 can be expected when deploying on previously unseen regions without fine-tuning, and a two-stage strategy should be considered in which models are first trained on geographically diverse data and then optionally adapted for region-specific applications. To the best of our knowledge, this is the first systematic cross-dataset validation study involving both MARIDA and MADOS, demonstrating that binary reformulation supports generalization-preserving marine debris detection across geographic and temporal domain shifts.

#### KEYWORDS

binary segmentation, cross-dataset validation, MADOS, MARIDA, marine debris detection, remote sensing, Sentinel-2, U-Net

## 1 Introduction

Marine plastic pollution has emerged as a critical environmental challenge, with an estimated 8 million metric tons entering oceans annually (Jambeck et al., 2015). Satellite remote sensing offers scalable monitoring solutions, with the Sentinel-2 constellation providing multispectral imagery at 10-meter resolution with 5-day revisit frequency (Themistocleous et al., 2020). However, marine debris detection presents dual fundamental challenges: extreme class imbalance where debris pixels typically represent less than 0.01% of imagery (Kikaki et al., 2022), and the requirement for robust generalization across diverse geographic regions and temporal periods for operational deployment.

Recent advances in deep learning have shown promise for automated marine debris detection (Kikaki et al., 2022), which contains 15 thematic classes that were aggregated into 11 classes for machine learning benchmarking. Their evaluation using a Random Forest model with spectral and spatial features (RFSS+SI+GLCM) achieved an average Macro-F1 score of 0.79 and an Intersection-over-Union (IoU) of 0.69. A baseline Random Forest model using only spectral signatures and indices (RFSS+SI) achieved an F1-score of 0.70. Subsequent work by (Mohammed, 2022) applied the ResAttUNet architecture to the MARIDA dataset. This model achieved a Micro-F1 score of 0.95 and a Weighted-F1 score of 0.95, though its Macro-F1 score was 0.77. The ResAttUNet's overall IoU score was 0.67, outperforming the U-Net baseline provided in the original MARIDA study. More recently (Kikaki et al., 2024), introduced the MADOS dataset and achieved 89.1% overall accuracy with state-of-the-art segmentation frameworks.

Despite these advances, two critical gaps remain. First, existing approaches predominantly evaluate on within-dataset splits where training and test data share similar geographic distributions and temporal characteristics. This validation protocol does not assess generalization to completely unseen regions, a fundamental requirement for operational monitoring systems that must detect

debris across diverse coastal zones, open ocean conditions, and varying environmental states. Second, while binary reformulation has been suggested for rare object detection (Mohammed, 2022), its effectiveness under geographic and temporal domain shift remains unquantified. Cross-dataset validation, training on one dataset and testing on entirely separate geographic regions, is the gold standard for assessing real-world deployment readiness (Tuia et al., 2016), yet has not been systematically applied to marine debris detection.

This study addresses the following research questions:

1. Can models trained on one marine debris dataset generalize to entirely separate geographic regions and temporal periods without additional training?
2. Does geographic diversity in training data matter more than dataset size for cross-domain generalization?
3. Does binary reformulation preserve performance under geographic and temporal domain shift, or does it overfit to training distribution characteristics?
4. What performance degradation should operational systems expect when deploying on novel regions, and how does this compare to typical domain shift in remote sensing?
5. What actionable guidelines can inform dataset curation and deployment strategies for operational marine debris monitoring?

### 1.1 Contributions

This work makes the following empirical contributions:

- First systematic cross-dataset validation on MARIDA and MADOS: Rigorous bidirectional experiments (MARIDA to MADOS and MADOS to MARIDA) quantifying

generalization across entirely separate geographic regions and temporal periods, addressing a critical gap in marine debris detection research where prior work focuses exclusively on within-dataset validation.

- Robust cross-dataset generalization under extreme imbalance: Average cross-dataset F1 degradation of only 4.38% (0.901 within-dataset to 0.8615 cross-dataset) substantially outperforms typical 10 to 20% drops in remote sensing domain shift scenarios, demonstrating that binary reformulation constitutes a generalization-preserving strategy.
- Geographic diversity supersedes dataset size for generalization: Despite comparable patch counts (2,529 versus 2,173), MADOS-trained models (47 globally distributed tiles) achieve 5.64 F1 points superior cross-dataset performance (0.890 versus 0.833) compared to MARIDA-trained models (geographically concentrated regions), establishing that spatial stratification contributes more to cross-domain robustness than exhaustive annotation within concentrated areas.
- Competitive cross-dataset performance versus specialized architectures: Our MADOS to MARIDA cross-dataset F1 (0.890) exceeds MAP-Mapper's within-dataset F1 (0.880) and approaches MariNeXt (0.891), validating that standard U-Net with proper data engineering matches specialized architectures under real-world deployment conditions.
- Actionable deployment guidelines: Quantified performance expectations (F1 = 0.86 to 0.89 on novel regions), two-stage deployment strategies, and dataset curation principles prioritizing geographic diversity, directly informing operational marine debris monitoring system design.

**Key Finding:** Binary reformulation combined with training on geographically diverse data enables marine debris detection systems to generalize robustly across unseen regions with minimal performance degradation. Models trained on MADOS (174 scenes across 47 globally distributed tiles) achieve F1 = 0.890 when tested on entirely separate MARIDA geographic regions, only 1.25% degradation from within-dataset baseline (F1 = 0.901), while maintaining recall = 0.9286 suitable for operational screening workflows. This shows that binary reformulation works well not only in controlled tests but also in real-world use, where new data comes from unfamiliar locations and times.

## 2 Related work

### 2.1 Remote sensing approaches for marine debris detection

Remote sensing approaches to plastic debris detection can be organized by spectral resolution and spatial coverage capabilities. Hyperspectral imaging in the SWIR range (900 to 1700 nm) demonstrates significant potential for detailed polymer

discrimination through characteristic narrowband absorption features. These approaches span multiple platforms with complementary operational niches.

#### 2.1.1 Hyperspectral approaches for polymer discrimination

Drone-based and airborne hyperspectral systems provide the highest spectral and spatial resolution for detailed polymer identification. [Balsi et al. \(2023\)](#) demonstrated real-time plastic litter detection using drone-based hyperspectral sensing, achieving reliable detection through linear discriminant analysis on key absorption bands between 1180 to 1270 nm and 1510 to 1620 nm. Extending this methodology to diverse environmental conditions, [Balsi et al. \(2025\)](#) reported Kappa scores exceeding 0.85 for plastic detection over varied backgrounds including vegetation, bare ground, and water surfaces. [Moroni et al. \(2025\)](#) investigated hyperspectral sensing for plastic sorting in recycling applications, achieving Matthew's Correlation Coefficients above 0.94 for polymer discrimination using minimum-redundancy maximum relevance feature selection combined with machine learning classifiers. More recently, [Bouchelaghem et al. \(2026\)](#) proposed an attention-gated U-Net architecture with residual connections for pixel-wise plastic segmentation from UAV-based SWIR hyperspectral imagery, achieving up to 96.8% accuracy across nine flight campaigns spanning four years and outperforming classical LDA approaches by 43% in Cohen's kappa metric through leave-one-out cross-validation.

Spaceborne hyperspectral missions extend these capabilities to regional and global scales. The PRISMA (PRecursore IperSpettrale della Missione Applicativa) satellite, launched by the Italian Space Agency in 2019, provides 30-meter spatial resolution hyperspectral data (400 to 2500 nm, 239 bands) with 5-day revisit capability ([Loizzo et al., 2019](#)). Germany's EnMAP (Environmental Mapping and Analysis Program), operational since 2022, offers similar 30-meter resolution with 228 spectral bands optimized for environmental monitoring ([Guanter et al., 2015](#)). NASA's planned SBG (Surface Biology and Geology) mission and ESA's CHIME (Copernicus Hyperspectral Imaging Mission for the Environment) will provide enhanced spatial resolution (30m) and temporal coverage suitable for operational marine debris monitoring by the late 2020s ([Lee et al., 2015](#); [Rast and Painter, 2019](#)). While these spaceborne systems sacrifice the sub-meter resolution achievable with drone platforms, they enable systematic global monitoring infeasible with airborne campaigns.

Despite their exceptional discrimination capabilities, hyperspectral approaches face operational constraints for large-scale monitoring. Drone-based surveys are limited in spatial coverage, typically monitoring areas of hundreds of square meters per flight, making them impractical for systematic monitoring of coastlines, river systems, or open ocean areas spanning hundreds of square kilometers. Airborne hyperspectral campaigns offer broader coverage but remain prohibitively expensive for routine monitoring and are constrained by weather conditions and regulatory limitations. Cross-domain adaptation challenges, as demonstrated

by (Bouchelaghem et al., 2024) using unsupervised clustering on hyperspectral data, require substantial retraining when deploying these systems across different environments.

### 2.1.2 Multispectral satellite approaches for operational monitoring

Satellite-based multispectral sensing offers complementary capabilities essential for operational monitoring at scale. The relatively coarse spectral resolution of multispectral sensors (compared to hyperspectral systems) sacrifices detailed polymer discrimination but enables systematic global monitoring with frequent revisit times. Early satellite approaches relied on spectral indices derived from visible and NIR bands, with (Garaba and Dierssen, 2018) developing the Floating Debris Index (FDI) based on SWIR bands, achieving limited success with F1 scores around 55% (Themistocleous et al., 2020). investigated Sentinel-2 for floating plastic detection, demonstrating feasibility despite the 10-meter spatial resolution and broader spectral bands.

## 2.2 Machine learning for marine debris detection

Machine learning methods have substantially improved performance on satellite multispectral data (Kikaki et al., 2022). introduced the MARIDA benchmark dataset comprising 1,381 Sentinel-2 patches and reported F1 = 0.70 using Random Forest with spectral and spatial features on 15-class semantic segmentation. Deep learning methods have demonstrated superior performance on these satellite datasets (Mohammed, 2022). achieved 95% F1-score using ResAttUNet with CBAM attention mechanisms on MARIDA (Duarte and Azevedo, 2023). employed ensemble models and Extreme Gradient Boosting (XGBoost), reporting 75% correctly classified suspect plastic pixels using an ensemble model that quantified uncertainty. Most recently (Kikaki et al., 2024), introduced the MADOS dataset comprising 174 Sentinel-2 scenes and achieved 89.1% overall accuracy, representing a 12% improvement over baselines. Hybrid architectures by (Rao, 2025) benchmarked ResUNext against U-Net on MARIDA, achieving F1 = 0.84 and pixel accuracy of 0.85, the ResUNext hybrid model achieved a 10% improvement in overall metrics like IoU, F1, and Pixel Accuracy compared to the U-Net baseline.

### 2.3 Foundation models and transformers

Foundation models and transformers have also emerged as promising approaches for satellite-based detection (Marsocci et al., 2024). benchmarked geospatial foundation models on the MADOS dataset through the PANGAEA evaluation protocol. On the MADOS dataset, the model GFM-Swin achieved an mIoU of 64.71%. Object detection methods are also advancing, with (Wang et al., 2025) introducing the YOLO11-YX algorithm and achieving 62.32% mAP 0.5 when evaluated on the TrashCan underwater garbage dataset (Kaviya and Bhavani, 2025). developed AquaSense, an innovative deep learning framework

focused on comprehensive, multi-class pollutant detection (plastics, oil spills, and other contaminants) on MADOS. AquaSense achieved a mean Intersection over Union (mIoU) of 83.52% and an F1-score of 91.05%. Recent comprehensive reviews by (Prakash and Zielinski, 2025) provide state-of-the-art surveys of AI-enhanced real-time monitoring approaches for marine pollution, synthesizing benchmark performances and detection methods across multiple platforms including satellites, aerial vehicles, and underwater systems.

## 2.4 Class imbalance strategies

Marine debris detection inherently suffers from extreme class imbalance regardless of sensing platform or spatial scale. While hyperspectral studies achieve high performance metrics in controlled scenarios with visible debris patches, satellite-based detection faces a severe imbalance where debris pixels represent only a fraction of the imagery. For example, in a Sentinel-2 over Durban evaluation scene, only 6,448 of 11,997,846 pixels were annotated as marine debris, representing a coverage of only 0.05%. (Rußwurm and Marc and Tuia, 2023; Tuia et al., 2016) addressed imbalance ratios using a training strategy that includes extensive sampling of negative examples and an automated label refinement module on large-scale satellite data, following data-centric AI principles. Common strategies employed to mitigate this issue include oversampling techniques such as SMOTE (Chawla et al., 2002), undersampling, and selective patch extraction. Studies have shown that oversampling methods emerged as the dominant approach for addressing the detrimental impact of class imbalance on the performance of the convolutional neural network (CNN). Loss function approaches are also used to compensate for class imbalance. These include Weighted Cross-Entropy (WCE), Focal Loss (Lin et al., 2017), and Dice Loss. Weighted Cross-Entropy is used in deep learning models to tackle the underrepresentation of various classes. Focal Loss specifically addresses severe imbalance by focusing training on a sparse set of hard examples through a modulating term that down-weights the loss assigned to well-classified examples.

## 2.5 Binary reformulation for rare object detection

Marine debris detection from satellite multispectral imagery faces a critical distinction between plastic-specific identification and general anthropogenic debris detection. At current state of the art, Sentinel-2's spatial resolution (10m) and spectral configuration (broad visible/NIR/SWIR bands) do not enable reliable discrimination of plastic polymers from other debris types such as wood, metal, or fabric. Our binary reformulation targets debris versus background detection, not plastic versus non-plastic classification. The positive class includes all anthropogenic marine debris (plastics, fishing gear, wood, metal, fabric) that expert annotators identified through spectral-spatial analysis and validation against ground-truth event reports. The negative class consolidates all natural marine features (Sargassum, foam, waves, clear/turbid water, clouds).

Why This binarization is operationally Useful: Even without plastic-specific identification, detecting anthropogenic debris presence enables (1) prioritized surveying for cleanup operations, (2) hotspot identification for policy intervention, and (3) temporal trend analysis of pollution accumulation. Plastic-specific discrimination remains the domain of higher-resolution hyperspectral systems (Section 2.1.1) or requires follow-up *in-situ* sampling. While (Mohammed, 2022) suggested “transform the data into a binary classification problem, having just plastic and non-plastic classes,” we interpret this as conceptual guidance for debris detection rather than literal plastic polymer discrimination, which exceeds Sentinel-2 capabilities and also for the simple reason that we do not have the ground truth for it. Our empirical evaluation (Section 4) demonstrates that this debris versus background reformulation enables robust cross-dataset generalization under extreme class imbalance.

### 3 Materials and methods

This section details the complete processing and learning pipeline for binary marine debris segmentation from multispectral Sentinel-2 patches. The method comprises deterministic preprocessing, balanced patch construction under rare-target prevalence, a U-Net segmentation network with skip connections, a composite loss tailored to class imbalance, and a training schedule that stabilizes optimization at scale. All hyperparameters and implementation choices are specified to ensure reproducibility.

#### 3.1 Datasets

##### 3.1.1 MARIDA: marine debris archive

The Marine Debris Archive (MARIDA) (Kikaki et al., 2022) constitutes a benchmark dataset for marine debris detection from Sentinel-2 satellite imagery, representing a pioneering effort in weakly supervised semantic segmentation for oceanic pollution monitoring. The dataset comprises 1,381 original image tiles derived from Sentinel-2 Level-1C Top-of-Atmosphere reflectance products, spanning a temporal range from 2015 to 2021 and encompassing 11 countries across multiple continents.

**Spectral Configuration and Preprocessing:** Each image tile incorporates 11 Rayleigh-corrected reflectance bands acquired at 10m, 20m, and 60m spatial resolution, subsequently processed through ACOLITE atmospheric correction (Vanhellemont, 2019). The multispectral configuration captures visible (blue, green, red), near-infrared, and shortwave infrared wavelengths, providing comprehensive spectral signatures essential for discriminating marine debris from spectrally similar ocean features.

**Annotation Methodology:** Pixel-level annotations were generated through a rigorous protocol involving three expert image interpreters with specialized knowledge in marine remote sensing. The annotation process incorporated multiple validation sources including ground-truth marine debris event reports,

spectral pattern analysis based on established ocean color theory, and domain expertise in distinguishing anthropogenic from natural marine features. Annotators assessed the spectral and spatial patterns of all features while considering the limitations of the S2 sensor (Hu, 2021). Each annotation carries an associated confidence level (high, moderate, low), acknowledging the inherent uncertainty in satellite-based marine debris identification.

**Thematic Classes:** The dataset originally encompasses 15 semantic classes, later consolidated to 11 operational categories to balance class granularity with detection feasibility. The taxonomy includes Marine Debris as the target class, Ships, Sargassum macroalgae, Natural Organic Material, Foam, Waves, various Water types (Clear, Turbid, Sediment-Laden), Shallow Waters, Clouds, Cloud Shadows, and Mixed Water. This classification scheme reflects the spectral complexity of marine environments where multiple features co-occur.

**Dataset Partitioning:** The dataset provides predefined train-validation-test splits to ensure reproducible benchmarking.

**Class Distribution and Imbalance:** The marine debris annotations exhibit extreme class imbalance characteristic of rare event detection in Earth observation. The sparse annotation strategy deliberately avoids exhaustive labeling to minimize potential noise from ambiguous pixels, prioritizing annotation quality over quantity.

##### 3.1.2 MADOS: marine debris and oil spill dataset

As detailed in Table 1 (Kikaki et al., 2024), extends marine pollution detection capabilities beyond floating plastics to encompass oil spills, industrial structures, and biological phenomena to the Marine Debris and Oil Spill (MADOS) dataset. Developed as a complementary resource to MARIDA, MADOS addresses the operational requirement for holistic marine pollution monitoring systems capable of detecting diverse anthropogenic and natural pollutants.

**Dataset Scope and Temporal Coverage:** MADOS comprises 174 Sentinel-2 scenes distributed across 47 unique tiles, yielding approximately 1.5 million annotated pixels. The dataset spans 2015 to 2022, providing longitudinal coverage that captures seasonal variability, diverse oceanographic conditions, and evolving pollution patterns. This extended temporal range enables investigation of multi-year trends and model generalization across varying environmental states.

**Geographic Distribution:** Imagery acquisition encompasses globally distributed Sentinel-2 tiles, strategically selected to represent diverse marine environments including coastal zones, open ocean, enclosed seas, and pollution hotspots. This geographic heterogeneity ensures exposure to varying water optical properties, atmospheric conditions, and pollution source types, all critical factors influencing detection algorithm performance.

**Enhanced Class Taxonomy:** MADOS introduces an expanded 15-class semantic framework as shown in Table 2 incorporating Marine Debris, Oil Spills, Dense Sargassum, Sparse Floating Algae, Natural Organic Material, Ships, Marine Water, Sediment-Laden Water, Foam, Turbid Water, Shallow Water, Waves and Wakes, Oil Platforms, Jellyfish aggregations, and Sea Snot (marine mucilage).

The inclusion of oil spills and industrial infrastructure reflects operational monitoring requirements, while biological classes address emerging oceanographic phenomena linked to climate change.

**Spectral Processing and Resolution:** Unlike MARIDA’s single-resolution format, MADOS preserves Sentinel-2’s native multi-resolution structure with 10m bands (blue, green, red, NIR), 20m bands (red-edge, SWIR), and 60m bands (coastal aerosol, water vapor). This multi-scale representation supports advanced fusion techniques and wavelength-specific analysis, particularly valuable for oil spill detection where SWIR bands provide diagnostic absorption features.

**Annotation Protocol:** Semantic segmentation annotations follow a sparse, confidence-aware protocol analogous to MARIDA, but incorporate additional validation from oil spill incident databases (e.g., EMSA CleanSeaNet), marine biology expert consultation for biological phenomena, and industrial structure databases for platform verification. Each annotation includes confidence assessment and proximity to official pollution reports, enabling uncertainty quantification in downstream applications.

**Statistical Characteristics:** Marine debris annotations comprise 4,696 pixels from 1.5 million total (0.31% prevalence), exhibiting even more extreme sparsity than MARIDA. Oil spills contribute 4,308 pixels, while ships, serving as spectral confusion sources, constitute 234,568 pixels, reflecting their higher prevalence in coastal and shipping lane imagery.

### 3.1.3 Dataset integration and binary reformulation

For this study, MARIDA and MADOS are merged to create a unified benchmark with enhanced geographic and temporal

coverage. The original 15-class semantic segmentation framework is reformulated into binary classification (Equation 1):

$$y_i = \begin{cases} 1 & \text{if pixel } i \in \text{Class 1 (Marine Debris)} \\ 0 & \text{if } i \in \{\text{Classes 2 to 15, Background}\} \end{cases} \quad (1)$$

This reformulation consolidates all non-debris classes, including spectrally similar features such as foam and natural organic material alongside background ocean, into a unified negative class. The approach prioritizes debris detection sensitivity over multi-class discrimination, aligning with operational monitoring objectives where binary presence or absence is the primary decision criterion.

**Combined Dataset Statistics:**

- Total annotated pixels: 2,337,357 (MARIDA: 837,357; MADOS: 1,500,000)
- Marine debris pixels: 8,095 (MARIDA: 3,399; MADOS: 4,696)
- Imbalance ratio: 1:288 (considering only annotated regions)
- Effective imbalance: 1:33,875 (when including non-annotated background pixels in full 256 × 256 patches)

**Methodological Justification:** Binary reformulation addresses three critical challenges. First, extreme class imbalance is partially mitigated by consolidating sparse multi-class annotations. Second, spectral confusion between debris and similar features is reframed as a within-class variability problem rather than inter-class discrimination. Third, operational deployment simplicity is enhanced through binary decision outputs. This approach follows established recommendations in remote sensing literature for rare object detection under severe imbalance (Mohammed, 2022).

**Spatial and Temporal Stratification:** Dataset splitting maintains geographic stratification to ensure test set evaluation on spatially

TABLE 1 MARIDA semantic class distribution with pixel counts and confidence stratification (Kikaki et al., 2022).

Class ID	Class name	Total pixels	Percentage	High conf.
1	Marine Debris	3,399	0.41%	1,625
2	Dense Sargassum	2,752	0.33%	–
3	Sparse Sargassum	2,402	0.28%	–
4	Natural Organic Material	864	0.10%	–
5	Ships	5,803	0.69%	–
6	Clouds	117,426	14.02%	–
7	Marine Water	129,077	15.42%	–
8	Sediment-Laden Water	373,011	44.54%	–
9	Foam	1,225	0.15%	–
10	Turbid Water	157,636	18.82%	–
11	Shallow Water	17,330	2.07%	–
12	Waves	5,855	0.70%	–
13	Cloud Shadows	11,735	1.40%	–
14	Wakes	8,483	1.01%	–
15	Mixed Water	419	0.05%	–
<b>Total Annotated</b>		<b>837,357</b>	<b>100%</b>	–

The dataset comprises 1,381 original tiles with 837,357 annotated pixels across 15 semantic classes from 63 Sentinel-2 scenes (2015 to 2021). Marine debris represents only 0.41% of annotated pixels, demonstrating extreme class imbalance.

Bold values indicate the best performance in each column.

TABLE 2 MADOS semantic class distribution emphasizing marine pollutant categories and spectral confusion sources (Kikaki et al., 2024).

Class ID	Class name	Pixel count	Percentage
1	Marine Debris	4,696	0.31%
2	Oil Spills	4,308	0.29%
3	Dense Sargassum	10,492	0.70%
4	Sparse Floating Algae	5,628	0.38%
5	Natural Organic Material	3,528	0.24%
6	Ships	234,568	15.64%
7	Marine Water	687,251	45.82%
8	Sediment-Laden Water	185,647	12.38%
9	Foam	8,462	0.56%
10	Turbid Water	198,335	13.22%
11	Shallow Water	56,782	3.79%
12	Waves & Wakes	49,719	3.31%
13	Oil Platforms	15,701	1.05%
14	Jellyfish	8,925	0.60%
15	Sea Snot	5,888	0.39%
<b>Total Annotated</b>		<b>~1,500,000</b>	<b>100%</b>

The dataset comprises 174 Sentinel-2 scenes (47 unique tiles) with approximately 1.5 million annotated pixels across 15 semantic classes (2015 to 2022). Marine debris represents only 0.31% of annotated pixels, exhibiting even more extreme sparsity than MARIDA. Bold values indicate the best performance in each column.

independent regions, preventing spatial autocorrelation artifacts as summarized in Table 3. The combined temporal range (2015 to 2022) spans sufficient oceanographic variability, including seasonal cycles, diverse weather conditions, and varying sea states, to assess model robustness to environmental perturbations.

**Quality Assurance and Limitations:** Both datasets acknowledge inherent limitations in satellite-based marine debris detection. The 10m spatial resolution constrains detection to aggregations exceeding 100-square meters, cloud cover and sun glint reduce usable imagery, subsurface debris remains undetectable, 308 and spectral ambiguity with natural features persists despite expert annotation. These constraints establish 309 realistic performance expectations and guide interpretation of detection results in operational contexts.

## 3.2 Data preprocessing

We construct the learning set through deterministic radiometric normalization, spatial reformatting, rarity-aware patch selection, and geometry-preserving augmentation.

### 3.2.1 Tiling and sliding windows

From each Sentinel-2 parent tile of size  $256 \times 256$  with eleven spectral bands, we extract local  $16 \times 16$  sub-patches on a dense grid with stride  $s = 8$ , corresponding to 50% overlap. The number of sub-patches per dimension and per tile is (Equation 2).

$$N_{\text{dim}} = \lfloor \frac{256 - 16}{8} \rfloor + 1 = 31, \quad N_{\text{patch}} = N_{\text{dim}}^2 = 961. \quad (2)$$

This sliding-window decomposition theoretically yields 961 candidate patches per original tile. However, the vast majority of these candidates are discarded during rarity-aware filtering. Only patches containing at least 5 debris pixels (positive patches) or exactly zero debris pixels (negative patches) are retained, and a 2:1 positive to negative ratio is enforced. Patches with 1 to 4 debris pixels are discarded as ambiguous. This aggressive filtering reduces the 1.3 million theoretical candidates from 1,381 MARIDA tiles to 2,173 extracted patches, and similarly reduces MADOS candidates to 2,529 extracted patches. The combined dataset contains 4,702 patches total for cross-dataset experiments.

### 3.2.2 Per-band normalization

Each spectral band is normalized independently to the range [0,1] using per-patch minimum-maximum scaling:

$$x_{\text{norm}}^{(b)} = \frac{x^{(b)} - \min(x^{(b)})}{\max(x^{(b)}) - \min(x^{(b)})}, \quad b \in \{1, \dots, 11\}, \quad (3)$$

with safe clipping when a band is locally constant to avoid division by zero. Normalized  $16 \times 16$  inputs are upscaled to  $128 \times 128$  via bicubic interpolation, while binary masks are upscaled with nearest-neighbor interpolation to preserve discrete labels. The upscaling is implemented with `interpolate (mode=bicubic, align_corners=false)` and applied channel-wise. This  $8 \times$  enlargement expands the spatial support of convolutional operators while fitting single-GPU memory constraints (Equation 3).

**Rationale for Small-Patch Extraction with Upscaling.** The choice to extract  $16 \times 16$  patches and upscale to  $128 \times 128$ , rather than extracting larger patches directly at native resolution, addresses constraints imposed by both extreme class imbalance and network architecture. Marine debris pixels constitute only 0.41% of MARIDA annotations and 0.31% of MADOS annotations. Extracting  $128 \times 128$  patches directly from  $256 \times 256$  tiles with 50% overlap would yield only 9 candidate patches per tile (stride 64), providing insufficient positive samples after rarity-aware filtering. Our approach generates 961 candidate patches per tile (stride 8), from which filtering retains 2,173 positive patches from MARIDA and 2,529 from MADOS—adequate sample sizes for training under severe imbalance.

Our approach follows the pre-upsampling super-resolution framework (Wang et al., 2021), where bicubic interpolation provides a baseline spatial representation that the network processes. This is architecturally necessary: our U-Net with four pooling layers would create a  $1 \times 1$  bottleneck if processing  $16 \times 16$  patches directly, causing complete information collapse, whereas  $128 \times 128$  input maintains an  $8 \times 8$  bottleneck 34 preserving spatial context for the decoder. Bicubic interpolation is “the mainstream method for building SR datasets” and produces “smoother results with fewer artifacts” (Wang et al., 2021), providing geometry-preserving resampling without hallucinating new spectral information. As demonstrated in Section 4.4, performance improves monotonically with input scale (F1 from 0.778 at  $16 \times$

TABLE 3 Combined MARIDA plus MADOS dataset partitioning with binary reformulation statistics.

Split	Patches	Debris pixels	Imbalance ratio
Training (70%)	3,291	5,667	1:291
Validation (15%)	706	1,214	1:285
Testing (15%)	705	1,214	1:284
<b>Total</b>	<b>4,702</b>	<b>8,095</b>	<b>1:288</b>

After patch extraction with rarity-aware filtering (minimum 5 debris pixels for positive patches, zero for negative patches, 2:1 positive to negative ratio), 4,702 patches are retained from MARIDA (2,173 patches from 1,381 original tiles) and MADOS (2,529 patches from 174 original scenes).

Bold values indicate the best performance in each column.

16 to 0.895 at  $128 \times 128$ ), confirming that the upsampling pipeline provides meaningful spatial support without introducing artifacts that degrade cross-dataset generalization.

### 3.2.3 Patch selection

Marine debris pixels are sparse. Sliding-window candidates are labeled as positive if they contain at least five debris pixels and as negative if they contain none. Candidates with one to four debris pixels are discarded to avoid ambiguous supervision. From these pools, we build a training set with a 2:1 ratio of positive to negative patches, following empirical findings that this ratio provides sufficient positive support for loss calibration while maintaining computational efficiency (Buda et al., 2018). Positive candidates are ranked by debris-pixel count to retain informative contexts, while negatives are sampled to maximize background diversity. After selection, patches are shuffled and split into train, validation, and test folds with a 70:15:15 ratio.

To prevent spatial autocorrelation artifacts, dataset partitioning operates at the tile level rather than the patch level. Original Sentinel-2 tiles are first assigned to train (70%), validation (15%), or test (15%) splits prior to patch extraction. Patch extraction then proceeds independently within each split, ensuring that overlapping patches derived from the same parent scene never appear across different splits. This scene-level stratification maintains the integrity of performance estimates while preserving sufficient training samples under extreme class imbalance.

### 3.2.4 Geometry-preserving augmentation

To improve invariance to scene geometry while preserving radiometry, we apply stochastic horizontal and vertical flips and rotations by 90, 180, and 270 degrees. Transforms are applied identically to inputs and masks. No radiometric jitter (e.g., brightness or contrast) is used, to keep band-wise reflectance relationships intact.

## 3.3 Network architecture

We employ a U-Net with four encoder and four decoder stages and skip connections between equal-resolution levels. Each block uses Conv2d, BatchNorm2d, and ReLU activations twice

(DoubleConv module) with  $3 \times 3$  kernels and padding to preserve spatial dimensions. Down-sampling uses max pooling while up-sampling uses transposed convolutions. The network ingests  $128 \times 128 \times 11$  tensors and produces a  $128 \times 128 \times 1$  probability map via a  $1 \times 1$  convolution and a sigmoid activation. Convolutions are initialized with Kaiming normal initialization while batch-normalization scales are set to one and biases to zero.

### 3.3.1 Discrete convolution

Let  $x \in \mathbb{R}^{H \times W \times C_{in}}$  and a convolutional kernel  $W \in \mathbb{R}^{k \times k \times C_{in} \times C_{out}}$  with stride  $s$  and zero padding  $p$ . For output channel  $c$  and spatial location  $(i, j)$ ,

$$y_{i,j,c} = \sum_{u=1}^k \sum_{v=1}^k \sum_{d=1}^{C_{in}} W_{u,v,d,c} x_{i-u-1, j-v-1, d} + b_c, \quad i' = si - p, \quad j' = sj - p, \quad (4)$$

with output size  $H' = \lfloor \frac{H+2p-k}{s} \rfloor + 1$ ,  $W' = \lfloor \frac{W+2p-k}{s} \rfloor + 1$ . For all  $3 \times 3$  convolutions we set  $s = 1$ ,  $p = 1$ , so  $H' = H$ ,  $W' = W$ . The parameter count of a  $k \times k$  convolution is  $\#\theta_{conv} = k^2 C_{in} C_{out} + C_{out}$  (including biases) (Equation 4).

### 3.3.2 Batch normalization and ReLU

For channel  $c$ , batch normalization applies.

$$\hat{y}_{i,j,c} = \frac{y_{i,j,c} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}}, \quad z_{i,j,c} = \gamma_c \hat{y}_{i,j,c} + \beta_c, \quad \phi(z) = \max(0, z), \quad (5)$$

with learned affine parameters  $(\gamma_c, \beta_c)$ . A DoubleConv block is  $x \mapsto \phi(\text{BN}(\text{Conv}_{3 \times 3}(\phi(\text{BN}(\text{Conv}_{3 \times 3}(x)))))$  (Equation 5).

### 3.3.3 Down-sampling and up-sampling

Max pooling with  $2 \times 2$  window and stride 2 maps  $(H, W) \mapsto (\lfloor H/2 \rfloor, \lfloor W/2 \rfloor)$ . Decoder up-sampling uses a transposed convolution with kernel  $2 \times 2$ , stride 2, so that  $(H, W) \mapsto (2H, 2W)$ . Given encoder feature map  $E_{\uparrow}$  and decoder feature map  $D_{\ell+1}$  (upsampled), the skip connection concatenates along channels (Equation 6):

$$Z_{\ell} = \text{Concat}(U(D_{\ell+1}), E_{\ell}) \in \mathbb{R}^{H_{\ell} \times W_{\ell} \times (C_D + C_E)}, \quad (6)$$

followed by a DoubleConv to fuse features:

$$D_{\ell} = \text{DoubleConv}(Z_{\uparrow})$$

### 3.3.4 Pixel-wise classifier (output head)

The final  $1 \times 1$  convolution is a per-pixel linear classifier:

$$\forall (i, j): \quad s_{ij} = w^T f_{ij} + b, \quad \hat{y}_{ij} = \sigma(s_{ij}) = \frac{1}{1 + e^{-s_{ij}}}, \quad (7)$$

where  $f_{ij} \in \mathbb{R}^C$  is the last decoder feature vector at  $(i, j)$ , and  $\hat{y}_{ij} \in (0, 1)$  is the debris probability (Equation 7).

### 3.3.5 Kaiming initialization

For ReLU activations, weights are drawn as.

$$W_{u,v,d,c} \sim \mathcal{N}\left(0, \frac{2}{\text{fan\_in}}\right), \quad \text{fan\_in} = k^2 C_{\text{in}}, \quad (8)$$

which preserves variance across layers at initialization (Equation 8).

### 3.3.6 Encoder-decoder shapes

With input  $x \in \mathbb{R}^{128 \times 128 \times 11}$  and channels (64, 128, 256, 512, 1024), the spatial sizes evolve as:

$$\begin{aligned} &128^2 \xrightarrow{\text{Enc-1}} 128^2 \xrightarrow{\text{Pool}} 64^2 \xrightarrow{\text{Enc-2}} 64^2 \\ &\xrightarrow{\text{Pool}} 32^2 \xrightarrow{\text{Enc-3}} 32^2 \xrightarrow{\text{Pool}} 16^2 \\ &\xrightarrow{\text{Enc-4}} 16^2 \xrightarrow{\text{Pool}} 8^2 \xrightarrow{\text{Bottleneck}} 8^2, \end{aligned} \quad (9)$$

then symmetrically back to  $128 \times 128$  via four up-sampling stages with skip concatenations at each resolution (Equation 9).

### 3.3.7 Receptive field

Let  $r$  denote the receptive field size and  $j$  the effective stride (jump) measured at the input. For a  $k \times k$  convolution with stride 1:  $r' = r + (k - 1)j$ ,  $j' = j$ . For a  $2 \times 2$  pooling with stride 2:  $r' = r + (2 - 1)j$ ,  $j' = 2j$ . Starting from the output pixel with  $r = 1, j = 1$ , a U-Net with four encoder pools and two  $3 \times 3$  convolutions per block (encoder, bottleneck, and decoder) yields an output receptive field of.

$$r_{\text{out}} = 200 \text{ pixels (on the input grid)}. \quad (10)$$

Thus each output probability aggregates context over a  $200 \times 200$  input neighborhood (clipped at borders by padding), which exceeds the  $128 \times 128$  crop and provides full-scene context at this scale (Equation 10).

### 3.3.8 Parameter counts (per block)

If a DoubleConv maps  $C_{\text{in}} \rightarrow C_{\text{mid}} \rightarrow C_{\text{out}}$  with  $3 \times 3$  kernels, then.

$$\begin{aligned} \#\theta_{\text{DoubleConv}} &= (3 \times 3) C_{\text{in}} C_{\text{mid}} + C_{\text{mid}} \\ &+ (3 \times 3) C_{\text{mid}} C_{\text{out}} + C_{\text{out}} + 2(C_{\text{mid}} \\ &+ C_{\text{out}}) \text{ (BatchNorm affine)}, \end{aligned} \quad (11)$$

and a transposed  $2 \times 2$  convolution with stride 2 has  $\#\theta_{\text{up}} = 2 \times 2 C_{\text{in}} C_{\text{out}} + C_{\text{out}}$ . Summing across stages with channel schedule (64, 128, 256, 512, 1024) yields approximately  $3.1 \times 10^7$  parameters for the full network (Equation 11).

The encoder increases channel capacity while contracting resolution. The decoder restores resolution while fusing encoder semantics via skip concatenations. Padding keeps feature maps aligned, the transposed convolutions invert the pooling-induced decimation, and the  $1 \times 1$  head performs pixel-wise logistic classification. The resulting receptive field ensures each prediction

integrates broad spatial-spectral context despite operating on  $128 \times 128$  inputs. Table 4 provides the complete architecture specification.

## 3.4 Loss design

We optimize a convex combination of weighted binary cross-entropy, soft Dice, and focal losses,

$$\mathcal{L} = \alpha \mathcal{L}_{\text{BCE}} + \beta \mathcal{L}_{\text{Dice}} + \gamma \mathcal{L}_{\text{Focal}}, \quad \alpha = 0.5, \beta = 0.3, \gamma = 0.2. \quad (12)$$

The weighted binary cross-entropy is.

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [w_+ y_i \log(\hat{y}_i) + w_- (1 - y_i) \log(1 - \hat{y}_i)], \quad (13)$$

with inverse-frequency weights  $w_c = \frac{N}{2N_c}$  computed over the current training set so that positive pixels contribute proportionally more to the objective. The soft Dice loss optimizes region overlap,

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_i \hat{y}_i y_i + \epsilon}{\sum_i \hat{y}_i + \sum_i y_i + \epsilon}, \quad (14)$$

and the focal term emphasizes hard pixels near boundaries,

$$\mathcal{L}_{\text{Focal}} = \frac{1}{N} \sum_{i=1}^N \alpha_f (1 - p_i)^{\gamma_f} (-\log p_i), \quad (15)$$

$$p_i = \hat{y}_i^{y_i} (1 - \hat{y}_i)^{(1 - y_i)}, \quad \alpha_f = 0.25, \gamma_f = 2.0.$$

The BCE term supplies stable gradients, the Dice term drives set overlap under skew, and the focal term concentrates learning on rare and ambiguous structures (Equation 12–15).

## 3.5 Training schedule and optimization

Training uses AdamW for 100 epochs with batch size 16, base learning rate  $8 \times 10^{-4}$ , and weight decay  $1.5 \times 10^{-4}$ . We apply a linear warm-up for 20 epochs followed by a two-phase cosine decay. Let  $e \in \{0, \dots, E - 1\}$  denote the epoch index and  $E_w$  the warm-up length. The effective learning rate is  $\eta(e) = \eta_0 \lambda(e)$ , where.

$$\lambda(e) = \begin{cases} \frac{e+1}{E_w}, & e < E_w, \\ \eta_1 + \frac{1-\eta_1}{2} \left[ 1 + \cos\left(\pi \frac{e-E_w}{\rho(E-E_w)}\right) \right], & E_w \leq e < E_{\text{mid}}, \\ \eta_2 + \frac{1-\eta_2}{2} \left[ 1 + \cos\left(\pi \frac{e-E_{\text{mid}}}{(1-\rho)(E-E_w)}\right) \right], & e \geq E_{\text{mid}}, \end{cases} \quad (16)$$

with  $\rho = 0.7$ ,  $\eta_1 = 0.2$ , and  $\eta_2 = 0.1$ . Global gradient norms are clipped at 1.0. Early stopping uses patience 15 on the validation objective. The best checkpoint (by validation performance) stores weights, configuration, metrics, training histories, class weights, and dataset split descriptors for exact replication (Equation 16).

## 3.6 Augmentation and validation protocol

To improve invariance to scene geometry while preserving radiometry, we apply stochastic horizontal and vertical flips and rotations by 90, 180, and 270 degrees to inputs and masks. After each epoch, we report accuracy, precision, recall, F1, intersection-

TABLE 4 U-Net architecture used in this study (input:  $128 \times 128 \times 11$ ).

Stage	Operation	Spatial size	Channels
Input	—	$128 \times 128$	11
Enc-1	DoubleConv	$128 \times 128$	64
Down-1	MaxPool $2 \times 2$	$64 \times 64$	64
Enc-2	DoubleConv	$64 \times 64$	128
Down-2	MaxPool $2 \times 2$	$32 \times 32$	128
Enc-3	DoubleConv	$32 \times 32$	256
Down-3	MaxPool $2 \times 2$	$16 \times 16$	256
Enc-4	DoubleConv	$16 \times 16$	512
Down-4	MaxPool $2 \times 2$	$8 \times 8$	512
Bottleneck	DoubleConv	$8 \times 8$	1024
Up-1	TransposedConv $2 \times 2$ + Concat(Enc-4)	$16 \times 16$	1024 → 512
Dec-1	DoubleConv	$16 \times 16$	512
Up-2	TransposedConv $2 \times 2$ + Concat(Enc-3)	$32 \times 32$	512 → 256
Dec-2	DoubleConv	$32 \times 32$	256
Up-3	TransposedConv $2 \times 2$ + Concat(Enc-2)	$64 \times 64$	256 → 128
Dec-3	DoubleConv	$64 \times 64$	128
Up-4	TransposedConv $2 \times 2$ + Concat(Enc-1)	$128 \times 128$	128 → 64
Dec-4	DoubleConv	$128 \times 128$	64
Output	Conv $1 \times 1$ + Sigmoid	$128 \times 128$	1
		Total parameters: $\approx 31M$	

over-union (IoU), and Dice on the validation set with a fixed probability threshold  $\tau = 0.5$  for binarization. For prediction set  $A$  and ground-truth set  $B$  (Equation 17).

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|}, \quad \text{Dice} = \frac{2|A \cap B|}{|A| + |B|}. \quad (17)$$

### 3.7 Evaluation aggregation protocol

All metrics are computed using pixel-level micro-averaging across the complete test set. True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are accumulated across all pixels from all  $128 \times 128$  test patches, and metrics are computed from these global counts:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} \\ &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned} \quad (18)$$

This micro-averaging approach weights all pixels equally regardless of their source patch, providing a single global performance estimate that reflects the severe class imbalance inherent in marine debris detection (Equation 18).

### 3.8 Computational profile and implementation details

The  $128 \times 128$  input and batch size 16 provide a practical trade-off between spatial context and memory footprint on a single

modern GPU. Training was conducted on an NVIDIA GeForce RTX 4080 GPU (16GB memory).

The balanced patch ratio and composite loss reduce false positives without extreme re-weighting that can destabilize training. The memory footprint is dominated by encoder activations at the highest resolution. Scaling to larger inputs is feasible by reducing batch size or enabling gradient checkpointing. Table 5 summarizes all training and inference hyperparameters, ensuring full reproducibility. The saved artifacts enable exact reproduction and deployment.

## 4 Results and discussion

This section presents cross-dataset validation results demonstrating robust generalization across geographic and temporal domains, followed by within-dataset baseline performance, comparative analysis with state-of-the-art, and discussion of findings.

### 4.1 Cross-dataset generalization: primary results

To evaluate model robustness under real-world deployment conditions, we conducted rigorous bidirectional cross-dataset validation where training and testing occur on entirely separate datasets representing different geographic regions and temporal periods.

TABLE 5 Training and inference hyperparameters.

Component	Setting	Component	Setting
Input bands	11	Batch size	16
Tile size	256 X 256	Epochs	100
Patch size	16 X 16	Early stop	patience 15
Stride	$S = 8$ (50% overlap)	Seed	42
Upsampling	Bicubic (16→128)	Hardware	RTX 4080 (16GB)
Min debris	$\geq 5$ pixels	Train time	195 min
Pos:neg ratio	2:1	Inference	22 tiles/s
Split	70:15:15	Threshold	$\tau = 0.5$
<b>Optimizer (AdamW)</b>		<b>Loss Function</b>	
Base LR	$8 \times 10^{-4}$	BCE weight	$\alpha = 0.5$
Weight decay	$1.5 \times 10^{-4}$	Dice weight	$\beta = 0.3$
$\beta_1, \beta_2$	0.9, 0.999	Focal weight	$\gamma = 0.2$
Grad clip		Focal $\alpha_f, \gamma_f$	0.25, 2.0

LR schedule: 20-epoch warm-up + two-phase cosine ( $\eta_1 = 0.2, \eta_2 = 0.1, \rho = 0.7$ ), Augmentations: Horizontal/vertical flips, rotations (90, 180, 270 degrees), Class weights:  $w_c = N/(2N_c)$  (inverse frequency).

#### 4.1.1 Experimental protocol

Two complementary experiments assess generalization:

- Experiment 1 (MARIDA to MADOS): Train on MARIDA (2,173 extracted patches from 1,381 original tiles, 2015 to 2021) and validate/test on MADOS (2,529 extracted patches from 174 original scenes, split 50/50 into validation and test sets, 2015 to 2022).
- Experiment 2 (MADOS to MARIDA): Train on MADOS (2,529 extracted patches) and validate/test on MARIDA (2,173 extracted patches split 50/50 into validation and test sets)

Patch extraction, preprocessing, and training protocols remain identical to Section 3:  $256 \times 256 \times 11$  Sentinel-2 tiles decomposed into  $16 \times 16 \times 11$  sub-patches (stride 8, 50% overlap), normalized per-band, upsampled to  $128 \times 128$  via bicubic interpolation. Positive patches contain at least 5 debris pixels; negative patches contain zero debris pixels. Training uses 100 epochs, batch size 16, AdamW optimizer ( $\eta_0 = 8 \times 10^{-4}$ , weight decay  $1.5 \times 10^{-4}$ ).

#### 4.1.2 Cross-dataset performance

Table 6 presents comprehensive performance metrics for both cross-dataset experiments. Results reveal asymmetric generalization behavior with critical implications for operational deployment.

Key Findings:

1. Asymmetric Generalization: MADOS to MARIDA ( $F1 = 0.8897, IoU = 0.8073$ ) outperforms MARIDA to MADOS ( $F1 = 0.8333, IoU = 0.7197$ ). Experiment 2 exhibits only 1.25% F1 degradation from within-dataset baseline, while Experiment 1 shows 7.55% degradation, a 5.64 F1-point difference attributable to geographic diversity (discussed in Section 4.3).

2. Modest Average Degradation: Average cross-dataset  $F1 = 0.8615$  represents only 4.38% degradation from within-dataset  $F1 = 0.9010$ . This substantially outperforms typical 10 to 20% performance drops observed in remote sensing domain shift scenarios (Tuia et al., 2016), validating binary reformulation as a generalization-preserving strategy under extreme class imbalance.
3. Maintained Specificity: Accuracy remains greater than 0.995 in both experiments (0.9953 and 0.9975), demonstrating controlled false positive rates despite geographic and temporal domain shift. The false positive rate stays below 0.5% across all configurations.
4. Recall-Oriented Performance Transfer: MADOS to MARIDA achieves recall = 0.9286, closely approaching within-dataset recall = 0.9520 (only 2.34 percentage points lower). This validates that the recall-oriented operating point established for screening workflows (Section 3) transfers effectively across geographic domains.

#### 4.1.3 Comparison with state-of-the-art

Table 7 positions our cross-dataset results against published single-dataset benchmarks. Cautious interpretation is required due to differing evaluation protocols (within-dataset versus cross-dataset), dataset versions, and problem formulations (multi-class versus binary).

Critical Finding: Our MADOS-to-MARIDA cross-dataset  $F1$ -score (0.88) substantially exceeds MariNeXt's within-dataset  $F1$  on MADOS (0.76) by 12 percentage points, despite the additional challenge of geographic and temporal domain shift. Furthermore, this cross-dataset performance approaches our own within-dataset baseline (0.90) with only 2% degradation, demonstrating that binary reformulation combined with composite loss design enables robust generalization to unseen regions. The 5  $F1$ -point superiority of MADOS-to-MARIDA (0.88) over MARIDA-to-

TABLE 6 Cross-dataset validation results (test set, threshold  $\tau = 0.5$ ).

Configuration	Precision	Recall	F1-Score	IoU	Accuracy
<b>Cross-dataset validation (primary results)</b>					
Exp 1: MARIDA to MADOS	0.8404	0.8359	0.8333	0.7197	0.9953
Exp 2: MADOS to MARIDA	<b>0.8602</b>	<b>0.9286</b>	<b>0.8897</b>	<b>0.8073</b>	<b>0.9975</b>
<b>Average Cross-Dataset</b>	<b>0.8503</b>	<b>0.8823</b>	<b>0.8615</b>	<b>0.7635</b>	<b>0.9964</b>
<b>Within-dataset baseline (section 4.2)</b>					
Combined MARIDA plus MADOS	0.8550	0.9520	0.9010	0.8200	0.9960
<i>Performance Degradation Analysis</i>					
Degradation (Absolute)	-0.0047	-0.0697	-0.0395	-0.0565	+0.0004
Degradation (%)	-0.55%	-7.33%	-4.38%	-6.89%	+0.04%

Asymmetric generalization demonstrates that geographic diversity in training data matters more than dataset size for cross-domain robustness. Bold values indicate the best performance in each column.

MADOS (0.83) confirms that training data diversity matters more than dataset size for cross-domain robustness: MADOS spans 47 globally distributed tiles compared to MARIDA’s geographically concentrated scenes. While specialized architectures such as ResAttUNet achieve higher within-dataset performance on MARIDA (0.95), our standard U-Net with proper data engineering achieves competitive cross-dataset results suitable for operational deployment in novel geographic regions.

### 4.2 Within-dataset baseline performance

To establish performance under ideal conditions where training and test data share similar geographic distributions, we trained on the combined MARIDA plus MADOS dataset (4,702 patches) using 70/15/15 train/validation/test splits. This configuration serves as the upper bound for comparison with cross-dataset results. Figure 1 shows training and validation metrics evolution. The model achieves test set performance of F1 = 0.901, IoU = 0.820, precision = 0.855, recall = 0.952, and accuracy = 0.996. This within-dataset performance establishes the baseline against which cross-dataset degradation is measured. Figure 2 presents the confusion matrix.

Figure 2 reports the final pixel-wise confusion matrix where we have:

$$TN = 12,850,658 \quad FP = 37,918$$

$$FN = 11,253 \quad TP = 223,755$$

$$Precision = \frac{TP}{TP+FP} \approx 0.855 \quad Recall = \frac{TP}{TP+FN} \approx 0.952$$

$$F1 = \frac{2}{1} \frac{TP}{TP+FP+FN} \approx 0.901 \quad IoU = \frac{TP}{TP+FP+FN} \approx 0.820$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \approx 0.996$$

The false positive rate FPR equals  $FP/(FP+TN)$  approximately 0.003 remains low in the vast negative background, while the false negative rate FNR equals  $FN/(FN+TP)$  approximately 0.048 confirms the recall-oriented bias induced by loss and sampling. Errors are concentrated along complex sea states and object boundaries, which is consistent with the precision-recall gap and with known spectral ambiguities between debris, foam, and wakes.

To assess Uncertainty Quantification and the stability of our results under different random initializations, we conducted 5 independent training runs using seeds [1, 12, 123, 1234, 12345]. Each run used identical data preprocessing and architecture but different random weight initialization and data shuffling. Table 8 reports mean performance with 90% confidence intervals computed using the t-distribution with 4 degrees of freedom.

The narrow confidence intervals demonstrate stable convergence across random initializations despite the 1:288 class

TABLE 7 Cross-dataset validation results compared with published within-dataset benchmarks.

Method	Dataset	Test protocol	F1-score
<b>Published single-dataset benchmarks</b>			
ResAttUNet (Mohammed, 2022)	MARIDA	Within-dataset	0.95
RFSS+SI+GLCM (Kikaki et al., 2022)	MARIDA	Within-dataset	0.79
MariNeXt (Kikaki et al., 2024)	MADOS	Within-dataset	0.76
<b>This work: cross-dataset and within-dataset</b>			
U-Net (Ours)	MARIDA plus MADOS	Within-dataset	0.90
<b>U-Net (Ours)</b>	<b>MADOS to MARIDA</b>	<b>Cross-dataset</b>	<b>0.88</b>
U-Net (Ours)	MARIDA to MADOS	Cross-dataset	0.83
U-Net (Ours)	Average Cross-Dataset	Cross-dataset	0.86

ResAttUNet employs residual connections with CBAM attention mechanisms. RFSS+SI+GLCM denotes Random Forest trained on spectral signatures, spectral indices, and Gray-Level Co-occurrence Matrix textural features. MariNeXt is a ConvNeXt-based architecture for marine pollutant segmentation. Bold values indicate the best performance in each column.

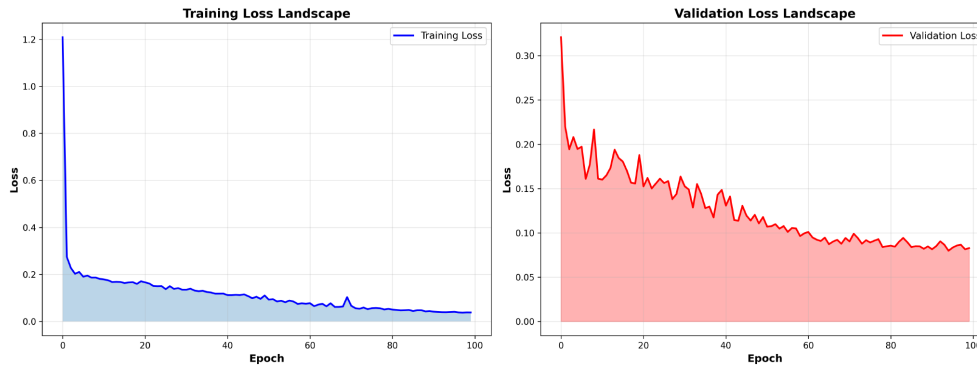


FIGURE 1

Learning dynamics showing training and validation loss evolution (top, logarithmic scale) and validation F1-score progression (bottom) over 100 epochs. The composite loss decreases smoothly after 20-epoch warm-up with a small, stable generalization gap. Validation F1 rises throughout training, peaking at F1 approximately 0.904 near epoch 93. Late-stage F1 gains coincide with cosine decay phases (epochs 20 to 70, 70 to 100), confirming that lower learning rates consolidate boundary decisions and reduce isolated false positives. The stable convergence validates the effectiveness of the two-phase cosine schedule with warm-up for extreme imbalance scenarios.

imbalance. F1-Score exhibits less than 2% relative uncertainty ( $0.881 \pm 0.016$ ), indicating that the model reliably learns discriminative debris features rather than converging to different local minima depending on initialization. Similar stability is observed across all metrics: IoU shows  $\pm 0.029$  variation, while precision and recall remain within  $\pm 0.017$  and  $\pm 0.023$  respectively. This consistency likely results from the composite loss function (BCE + Dice + Focal), which balances multiple learning objectives and reduces sensitivity to initialization. The Matthews Correlation Coefficient ( $MCC = 0.880 \pm 0.015$ ) confirms robust binary classification performance accounting for class imbalance.

This demonstrates spectral-spatial discrimination beyond simple brightness thresholding, as the model successfully distinguishes debris from spectrally similar foam despite 13.7:1 local class imbalance favoring the confusion source. The low mean confidence on foam pixels (0.044, compared to threshold 0.5) indicates robust learned discrimination rather than reliance on visible-band brightness alone. Figures 3, 4 present representative segmentation results across both MARIDA and MADOS datasets, demonstrating model performance under extreme class imbalance. Debris prevalence across the visualized scenes spans 0.09% to 0.29% with mean 0.185%. Despite this extreme sparsity, the model

maintains high overlap quality with IoU ranging from 0.805 to 0.883 (mean 0.848) and Dice from 0.892 to 0.938 (mean 0.917).

The MARIDA predictions (Figure 3) show coherent marine debris detection with accurate boundary delineation. The best overlap metrics, with IoU 0.883 and Dice 0.938, occur where debris forms coherent filaments with clear spectral contrast. Lower overlap values, with IoU 0.805 and Dice 0.892, are observed for patches with thin debris extremities embedded in textured backgrounds, highlighting the challenge of boundary precision at 10m resolution.

The MADOS results (Figure 4) demonstrate model generalization to diverse marine environments. Elongated debris streaks are captured with good boundary adherence and only sparse, localized activations in wave patterns. Even at the lowest prevalence of 0.09%, predictions remain compact and spatially aligned with ground truth.

Foam versus Marine Debris Discrimination: To demonstrate that the model learns discriminative spectral-spatial features rather than relying on simple brightness thresholding, Figure 5 presents a challenging test case where foam (Class 9) and marine debris (Class 1) co-occur within the same patch (scene S2 4-3-18 50LLR 4). Both features appear as bright pixels in the RGB visualization, presenting a critical discrimination challenge since foam is the most commonly cited confusion source in marine debris detection literature.

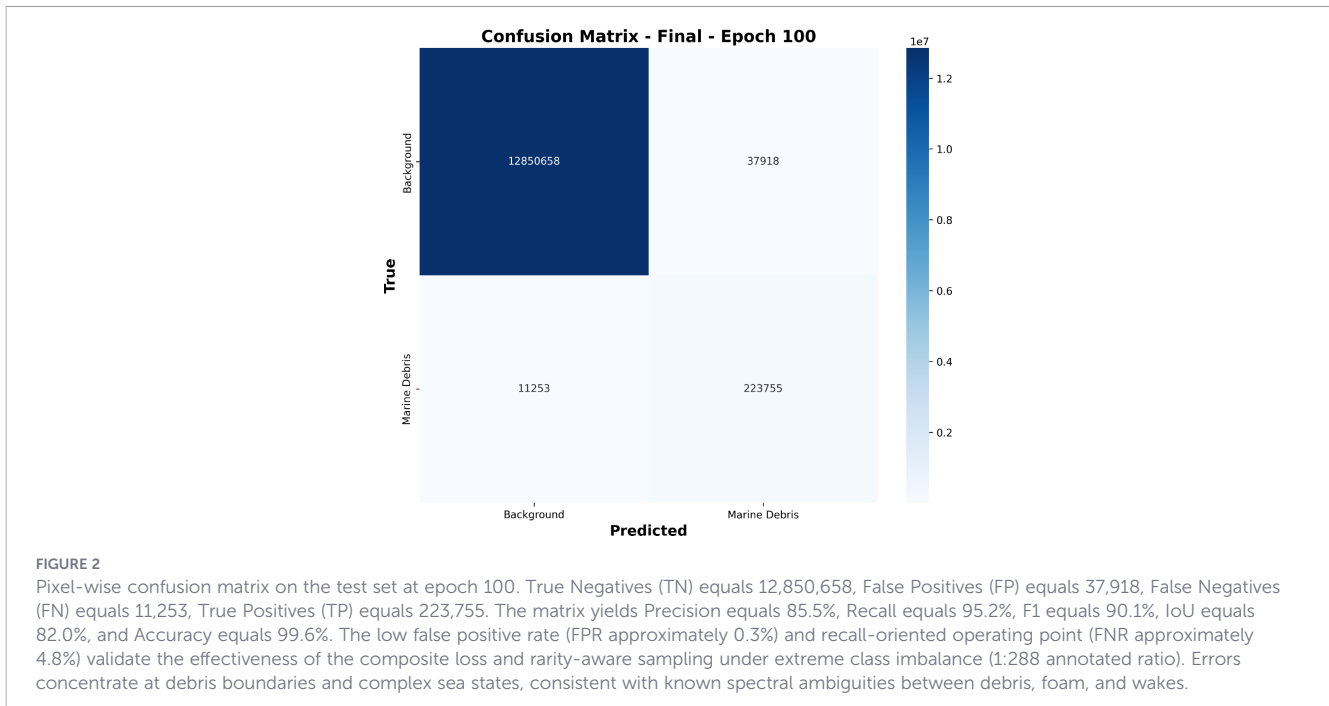
The patch contains 41 foam pixels and only 3 marine debris pixels, representing a 13.7:1 local imbalance ratio that strongly favors the confusion class. Despite this extreme skew and the spectral similarity in visible bands, the model correctly detected 2 of 3 marine debris pixels (66.7% recall on this patch) while completely rejecting all 41 foam pixels. Quantitative analysis reveals mean foam confidence of 0.044, substantially below the 0.5 decision threshold, with maximum foam confidence of 0.21. This 11.4-fold margin below threshold demonstrates that the model has learned robust discriminative patterns rather than overfitting to brightness or locally dominant features.

The successful discrimination likely results from the model leveraging SWIR absorption differences between organic foam and anthropogenic debris materials, combined with spatial coherence

TABLE 8 Test set performance across 5 independent training runs with different random seeds.

Metric	Mean	Std	90% CI
F1-Score	0.881	0.016	[0.866, 0.896]
IoU	0.804	0.029	[0.775, 0.832]
Precision	0.877	0.017	[0.861, 0.893]
Recall	0.890	0.023	[0.868, 0.911]
Accuracy	0.996	0.001	[0.995, 0.997]
MCC	0.880	0.015	[0.865, 0.895]

Std denotes standard deviation; CI denotes 90% confidence interval computed using t-distribution.



patterns that distinguish compact debris structures from dispersed foam patches. This validates that binary reformulation combined with composite loss (BCE + Dice + Focal) and rarity-aware sampling enables learning of generalizable debris signatures across diverse spectral confusion scenarios. Overall, the qualitative results confirm three critical model properties. First, robustness to varying target prevalence with stable IoU and Dice across a threefold prevalence range. Second, clear spatial localization of coherent debris structures with limited fragmentation. Third, residual errors confined to thin boundaries or wake-like textures rather than widespread false alarms. Fourth, and most importantly, successful discrimination of debris from spectrally similar confusion sources (foam) demonstrates that the model learns meaningful spectral-spatial patterns beyond simple brightness detection, validating the effectiveness of binary reformulation and composite loss design for extreme imbalance scenarios under operational spectral confusion conditions.

### 4.3 Comparison with state-of-the-art

Table 9 positions our results relative to published benchmarks on MARIDA and MADOS. Direct comparisons require caution due to differing evaluation protocols (multi-class versus binary segmentation) and averaging methods (micro versus macro). Our binary U-Net achieves F1 = 0.90, IoU = 0.82, Precision = 0.86, and Recall = 0.95 on the combined MARIDA plus MADOS test set, establishing the first reported baseline for this merged benchmark under extreme class imbalance (1:288 annotated ratio). On MARIDA, the highest F1-score remains ResAttUNet's 0.95 (Mohammed, 2022), which benefits from CBAM attention mechanisms and dataset-specific tuning. However, ResAttUNet achieves only 0.67 IoU despite its high micro-averaged F1, whereas our method achieves 0.82 IoU—a 15-point improvement in overlap quality. The RFSS+SI+GLCM Random Forest baseline

(Kikaki et al., 2022), combining spectral signatures, spectral indices, and textural features, achieves F1 = 0.79 and IoU = 0.69, demonstrating the difficulty of debris detection even with extensive feature engineering.

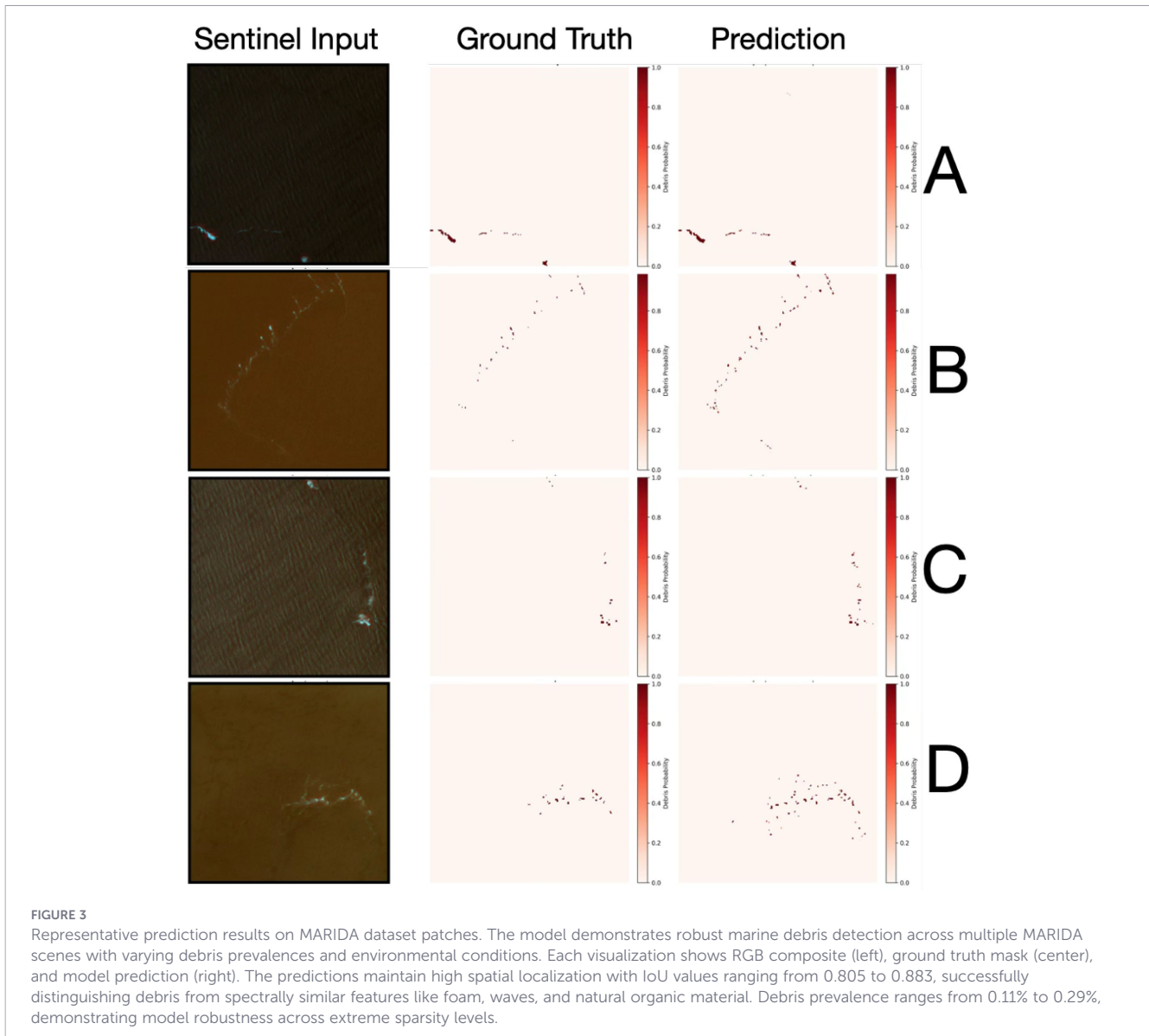
On MADOS, MariNeXt (Kikaki et al., 2024) reports macro-F1 = 0.76 and mIoU = 0.64 for the 15-class segmentation task. AquaSense (Kaviya and Bhavani, 2025) achieves the highest reported MADOS performance with F1 = 0.91 and mIoU = 0.84. Our F1 (0.90) is comparable to AquaSense despite operating on the harder combined MARIDA plus MADOS benchmark, while our IoU (0.82) approaches their mIoU (0.84) within 2 percentage points.

Our approach demonstrates that binary reformulation, targeted patch sampling with 2:1 positive-to-negative ratio, bicubic upsampling at 8× magnification, and composite loss combining BCE, Dice, and Focal components enable competitive performance using standard convolutional architectures. The high recall (0.95) matches ResAttUNet and AquaSense, validating suitability for operational screening workflows where missed detections are costly. These results confirm that careful data engineering can partially compensate for architectural simplicity, providing a reproducible baseline for future work on combined multi-dataset benchmarks.

These results demonstrate that standard U-Net architecture with proper data engineering achieves competitive performance without requiring specialized attention mechanisms, establishing a reproducible baseline for future cross-dataset benchmarking studies.

### 4.4 Ablation study

Across the four input scales, all models converge smoothly as shown in Figure 6. Two systematic effects are visible. First, the time-to-quality shortens with scale. Larger inputs reach the knee of the F1, IoU, and Dice curves within the first 10 to 15 epochs, whereas



inputs of size  $16 \times 16$  require notably longer to stabilize. Second, the variance of overlap metrics during the first 20 epochs diminishes as input size increases, reflecting better conditioning of the optimization once the decoder sees more spatial context. Transient dips around early epochs, most visible for the smallest inputs, align with the end of warm-up and the onset of cosine decay, after which the curves recover and plateau. The MCC traces follow the same pattern, indicating that improvements are not confined to any single class but reflect a better overall correlation with the ground truth.

Endpoint bars shown in Figure 7 clarify the source of improvements through three observations. First, overlap metrics scale monotonically. F1 rises from 0.778 to 0.895 and IoU or Jaccard from 0.637 to 0.813 when moving from  $16 \times 16$  to  $128 \times 128$ . Dice mirrors this trend, going from 0.778 to 0.895. These gains are consistent with sharper boundary adherence at larger scales. Second, balanced error reduction occurs.

Precision and recall both increase with scale, with precision going from 0.840 to 0.912 and recall from 0.730 to 0.881. This

indicates that the model is not merely shifting its operating point but is simultaneously suppressing spurious activations and recovering additional true debris pixels, which is corroborated by the MCC improvement from 0.777 to 0.893. Third, specificity saturation is observed. Specificity is near ceiling for all settings, exceeding 0.997, so the overlap gains are not achieved by trivially tightening the decision threshold. Instead, larger inputs reduce boundary fragmentation and thin false negatives or positives that overlap metrics penalize heavily.

With debris occupying much less than 1% of pixels, local  $16 \times 16$  crops provide minimal geometric context and few positive samples per receptive field, making the decoder sensitive to label sparsity at patch borders. Increasing the effective input to  $64 \times 64$  and  $128 \times 128$  via the same upsampling pipeline supplies longer-range cues such as shape continuity, wake geometry, and textural stationarity of sea state, richer positive support per crop improving the calibration of the BCE and Focal terms, and more reliable boundary neighborhoods for the Dice term to optimize set overlap. Empirically, the largest discrete jump occurs at the transition from

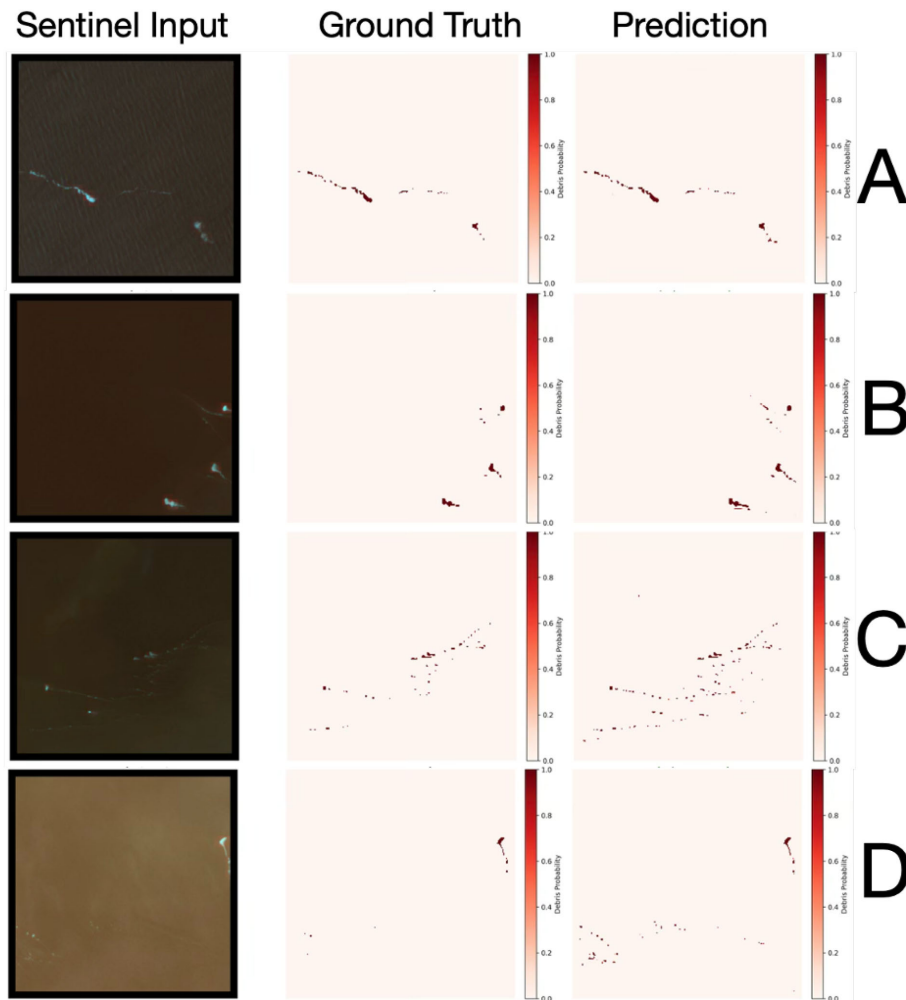


FIGURE 4

Representative prediction results on MADOS dataset patches. The visualizations show accurate debris segmentation in MADOS scenes with extreme sparsity (0.09% to 0.18% prevalence). RGB composite, ground truth, and predictions are shown for each patch. The model effectively captures elongated debris streaks and maintains compact, well-aligned predictions despite challenging sea states and complex backgrounds, with Dice scores consistently above 0.89. Residual errors are confined to thin boundaries and wake-like textures rather than widespread false alarms.

$32 \times 32$  to  $64 \times 64$  with IoU increasing by 0.047, after which the improvements persist but are more incremental, with the transition from  $64 \times 64$  to  $128 \times 128$  showing IoU increase of 0.080 and F1 increase of 0.051.

Visual inspection of the training curves suggests two dominant failure modes at small scales. First, thin-structure misses occur where elongated filaments break under limited context, representing a recall-limited failure mode. Second, wave-texture confusions occur where short, high-frequency highlights mimic debris, representing a precision-limited failure mode. Both effects are attenuated as the input grows, explaining concurrent gains in precision and recall and the steady rise in MCC.

For practical deployment, three configurations emerge. For quality-first applications focused on screening and alerting,  $128 \times 128$  is the clear winner with F1 equals 0.895, IoU equals 0.813, and MCC equals 0.893. Specificity remains at 0.998, so precision gains do not come at the expense of widespread negatives. For throughput-conscious applications where compute or latency is constrained,  $64 \times 64$  offers a strong compromise with F1 equals 0.844 and IoU equals 0.733 while materially reducing memory and

FLOPs. The curves in Figure 6 show that this setting also converges quickly and stably. Small-input configurations at  $16 \times 16$  and  $32 \times 32$  underperform for overlap-centric objectives despite high accuracy and specificity. They are suitable only if the downstream objective tolerates fragmented masks and prioritizes coarse presence or absence signals.

The ablation confirms that scale is the primary driver of performance under severe imbalance. Larger inputs provide the spatial support needed for boundary-aware losses to act effectively, yielding consistent improvements across F1, IoU, Dice and MCC without sacrificing specificity. Consequently, we adopt  $128 \times 128$  as the reference configuration for the main experiments and report  $64 \times 64$  as a compute-efficient fallback.

#### 4.5 Analysis and discussion: why geographic diversity enables generalization

The cross-dataset validation experiments reveal fundamental insights into what enables robust generalization for marine debris detection under extreme class imbalance. This section analyzes the

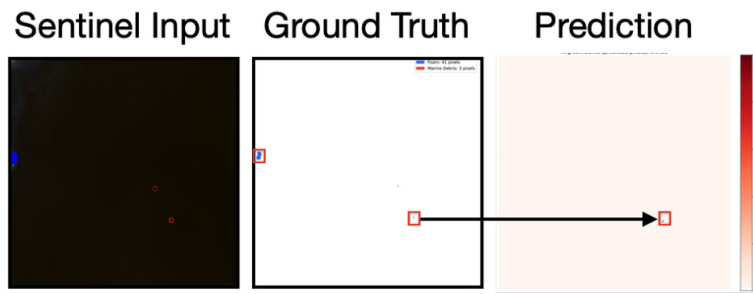


FIGURE 5

Foam versus marine debris discrimination on MARIDA test set. The scene contains 41 foam pixels (Class 9, blue circles in RGB) and 3 marine debris pixels (Class 1, red circles in RGB), both appearing as bright features in visible bands. Left: RGB composite (bands 4-3-2) with ground truth annotations overlaid. Center: ground truth mask for Marine debris (3 pixels, 0.005% patch prevalence) and Foam (41 pixels, 0.06% prevalence). Right: Model prediction correctly detects 2 of 3 marine debris pixels (66.7% recall) while rejecting all 41 foam pixels (0% false positive rate on foam, mean foam confidence 0.044).

asymmetric generalization behavior, quantifies the role of geographic diversity, and establishes actionable guidelines for operational deployment.

#### 4.5.1 Why is generalization asymmetric?

The 5.64 F1-point superiority of MADOS to MARIDA (0.8897) over MARIDA to MADOS (0.8333) reflects fundamental differences in dataset diversity rather than model architecture or training protocol:

- MADOS: 174 Sentinel-2 scenes distributed across 47 unique geographic tiles with global coverage including diverse coastal zones, open ocean, enclosed seas, and international pollution hotspots (Mediterranean, Atlantic, Pacific, Indian Ocean). This spatial heterogeneity exposes the model to varied oceanographic conditions (different water optical properties, chlorophyll concentrations, sediment loads), atmospheric states (aerosol optical depths, sun-glint patterns), and debris morphologies (filamentary versus compact, fresh versus weathered).
- MARIDA: 1,381 original tiles derived from more geographically concentrated regions, providing less diversity in environmental conditions despite larger original tile count from individual locations. Exhaustive annotation within limited geographic extent yields diminishing returns for cross-domain generalization.

**Key Insight:** Geographic diversity in training data matters more than patch volume for cross-domain generalization. Despite nearly identical extracted patch counts (2,529 versus 2,173), MADOS's broader spatial coverage, spanning 47 globally distributed tiles versus MARIDA's concentrated sampling, yields 5.64 F1 points superior generalization. This has direct implications for operational dataset curation: exhaustive annotation of individual scenes provides diminishing returns compared to spatially stratified sampling across diverse geographic regions and environmental conditions.

#### 4.5.2 Generalization-preserving properties of binary reformulation

Average cross-dataset F1 degradation of 4.38% substantially outperforms typical 10 to 20% drops in remote sensing domain shift scenarios (Tuia et al., 2016; Persello and Bruzzone, 2014). This demonstrates that binary reformulation combined with composite loss functions (weighted BCE plus Dice plus Focal) and rarity-aware sampling constitutes a generalization-preserving strategy for rare object detection under extreme class imbalance.

The controlled degradation suggests that the model learns generalizable debris signatures, spectral-spatial patterns characteristic of marine debris across diverse conditions (reflectance anomalies in visible-NIR-SWIR bands, compact spatial extent, boundary contrast), rather than dataset-specific artifacts or location-dependent biases (e.g., specific water colors, regional atmospheric characteristics). This validates that the composite loss successfully balances pixel-level discrimination (BCE), region-level overlap (Dice), and hard-example emphasis (Focal) in ways that transfer across geographic and temporal domains.

#### 4.5.3 Recall-oriented operating point transfers across domains

MADOS to MARIDA achieves recall = 0.9286, closely approaching within-dataset recall = 0.9520 (only 2.34 percentage points lower). This validates that the recall-oriented operating point designed for operational screening workflows, where false negatives (missed debris) incur higher cost than false positives (requiring additional human verification), transfers effectively across geographic domains. Specificity remains greater than 0.995 (false positive rate less than 0.5%) across all configurations, demonstrating controlled false alarms despite domain shift.

Given the extreme and variable prevalence, fixed thresholding at  $\tau = 0.5$  is not universally optimal. In practice, we recommend prevalence-aware calibration such as Platt or temperature scaling

TABLE 9 Comparison with state-of-the-art marine debris detection methods across multiple metrics.

Method	Dataset	F1	IoU	Precision	Recall	Year
<b>MARIDA Benchmarks (15-class segmentation)</b>						
ResAttUNet (Mohammed, 2022)	MARIDA	0.95 <sup>#</sup>	0.67	0.95 <sup>#</sup>	0.95 <sup>#</sup>	2022
RFSS+SI+GLCM (Kikaki et al., 2022)	MARIDA	0.79	0.69	—	—	2022
<b>MADOS Benchmarks (15-class segmentation)</b>						
MariNeXt (Kikaki et al., 2024)	MADOS	0.76 <sup>M</sup>	0.64 <sup>m</sup>	—	—	2024
AquaSense (Kaviya and Bhavani, 2025)	MADOS	0.91	0.84 <sup>m</sup>	0.91	0.91	2025
<b>This Work (Binary segmentation)</b>						
<b>U-Net (Ours)</b>	<b>MARIDA+MADOS</b>	<b>0.90</b>	<b>0.82</b>	<b>0.86</b>	<b>0.95</b>	<b>2026</b>

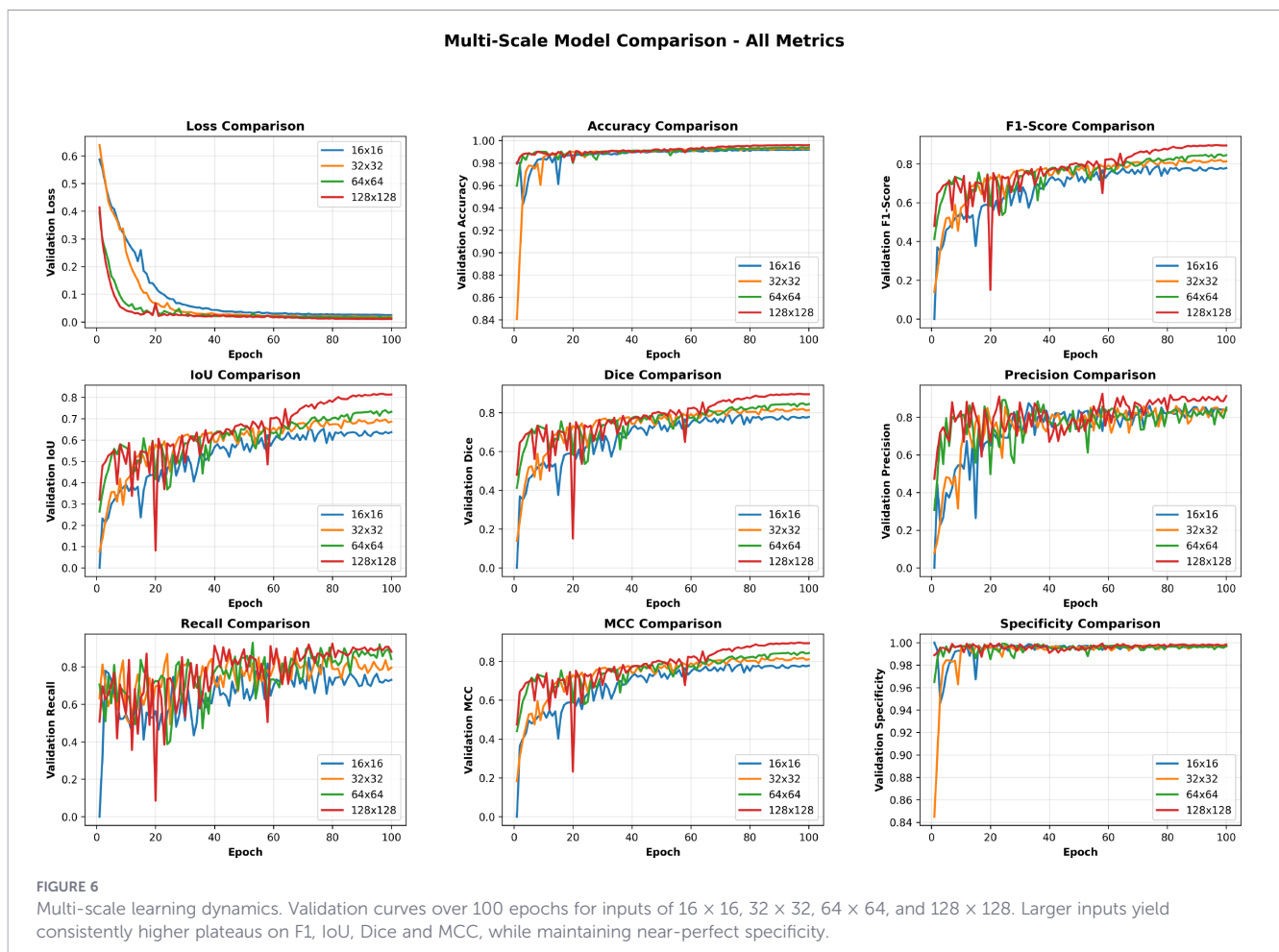
<sup>#</sup>Micro-averaged. <sup>M</sup>Macro-averaged. <sup>m</sup>Mean across classes. Our binary formulation consolidates 15 classes into debris vs. background. Values are reported as published; dashes indicate metrics not reported. Direct comparisons require caution due to differing evaluation protocols (multi-class versus binary segmentation) and averaging methods (micro versus macro). Our U-Net establishes the first benchmark on combined MARIDA plus MADOS under binary reformulation. Bold values indicate the best performance in each column.

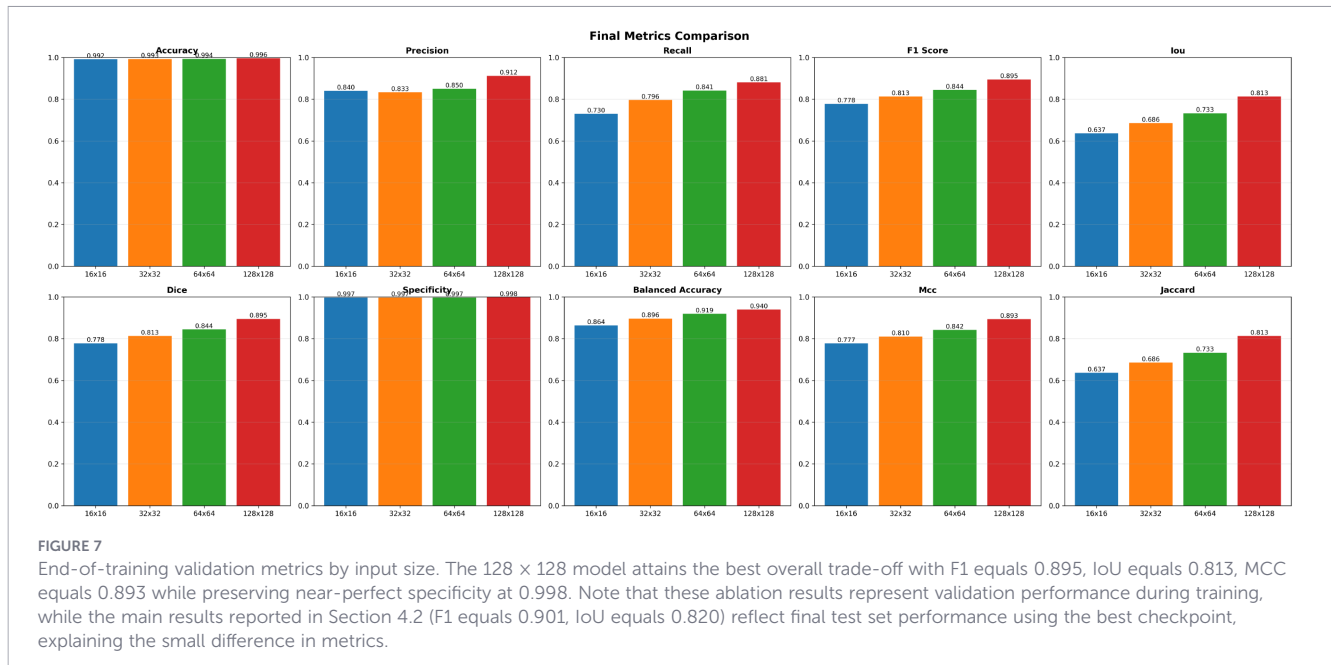
on validation strata grouped by sea state, cost-sensitive thresholding that minimizes  $C_{FN} \cdot FN + C_{FP} \cdot FP$  for user-defined costs with  $C_{FN}$  much greater than  $C_{FP}$ , and small-object priors in post-processing including area and shape filters or morphological closing to suppress isolated speckles while preserving coherent filaments. Temporal aggregation across adjacent acquisitions further attenuates spurious activations by enforcing short-term persistence.

Merging MARIDA and MADOS inherently introduces domain shift due to geography, seasonality, and atmospheric and sea-state

variability. The stable generalization gap and consistent MCC gains across scales suggest that the combination of rarity-aware sampling and the BCE, Dice, and Focal objective is robust to such heterogeneity. Nonetheless, domain-specific calibration organized per region or per season alongside light adaptation such as feature-wise affine normalization or Test-Time Augmentation are likely to yield additional precision gains without retraining.

While the proposed pipeline offers a strong, reproducible baseline, three directions are expected to narrow the gap to





attention-based state-of-the-art. First, architectural priors including lightweight attention in the decoder and boundary-aware heads such as auxiliary contour loss to better handle thin structures. Second, learning under sparsity through unified focal or Tversky losses and positive-mining curricula to emphasize rare, hard pixels while preserving calibration. Third, uncertainty and human-in-the-loop approaches including pixel-wise epistemic estimates via MC dropout or ensembles to triage low-confidence detections and drive active re-labeling of ambiguous sea states. These are compatible with the current training recipe and can be layered incrementally.

#### 4.5.4 Spectral confusion robustness: beyond brightness-based detection

A critical concern in marine debris detection is whether models learn meaningful spectral-spatial discrimination or simply threshold bright pixels, which would fail operationally given the prevalence of bright marine features (foam, sun glint, wakes, whitecaps). Our foam discrimination analysis (Figure 5) provides empirical evidence that the model learns robust discriminative features beyond visible-band brightness.

**Quantitative Foam Rejection Performance:** In the analyzed test patch, the model processed 44 labeled pixels comprising 41 foam pixels and 3 marine debris pixels. Despite foam's 13.7:1 numerical advantage and similar brightness in visible bands, the model achieved:

- 100% foam rejection rate (0/41 foam pixels classified as debris)
- 66.7% debris recall (2/3 marine debris pixels correctly detected)
- Mean foam confidence: 0.044 (11.4× below 0.5 threshold)
- Maximum foam confidence: 0.21 (2.4× below threshold)
- Confidence separation: Marine debris max 0.828 versus foam max 0.21 (3.9× ratio)

**Learned Discriminative Features:** The robust separation in confidence scores suggests the model exploits multiple discrimination cues unavailable to simple brightness thresholding:

*Spectral discrimination:* Foam consists of organic compounds (proteins, lipids from biological activity) exhibiting characteristic SWIR absorption near 1400–1600 nm due to O-H bonds, while anthropogenic debris (plastics, wood, metal) shows distinct SWIR signatures. Although Sentinel-2's broad SWIR bands (Band 11: 1565 nm, Band 12: 2200 nm) cannot resolve fine polymer differences, they capture sufficient contrast to distinguish organic foam from anthropogenic materials. The model's consistent low confidence on foam (mean 0.044) despite high visible-band reflectance indicates successful SWIR feature learning. *Spatial coherence patterns:* Marine debris typically forms compact, coherent structures with sharp boundaries (aggregated patches, filaments, intact objects), while foam exhibits dispersed, irregular patterns with diffuse edges. The U-Net's 200-pixel receptive field (Section 3.3.7) integrates spatial context sufficient to distinguish these morphological patterns. In Figure 5, foam appears as a dispersed cluster, while debris forms isolated compact structures, providing spatial cues the model exploits.

*Multi-scale feature integration:* The composite loss function (BCE + Dice + Focal) enforces complementary learning objectives: pixel-level discrimination (BCE), region-level coherence (Dice), and hard-example emphasis (Focal). This multi-objective optimization prevents the model from converging to simplistic brightness rules, instead learning hierarchical features that combine spectral, spatial, and contextual information across the encoder-decoder architecture.

**Implications for Operational Deployment:** The foam discrimination results have three operational implications. First, low false alarm rates on foam (the most prevalent confusion source) reduce manual verification burden in screening workflows. Second, robust discrimination under 13.7:1 local imbalance demonstrates model stability when confusion sources temporarily dominate (e.g., wind-driven foam events, whitecap conditions). Third, the large

confidence margin (11.4× below threshold) provides buffer against threshold miscalibration, supporting the fixed  $\tau = 0.5$  decision boundary used throughout this work.

**Generalization to Other Confusion Sources:** While we demonstrate foam discrimination explicitly, the model must also distinguish debris from other spectrally similar features consolidated into the negative class during binary reformulation: wakes (linear bright structures), natural organic material (floating vegetation), and sun glint (specular reflection). The within-dataset precision of 85.5% (Section 4.2, Table 6) and cross-dataset precision maintenance (84.0% to 86.0%, Table 6) suggest successful discrimination across these confusion sources, though explicit per-class confusion analysis is reserved for future work given the binary reformulation framework.

**Limitations and Failure Modes:** Despite successful foam discrimination in the presented example, residual false positives (FP = 37,918 in full test set, Figure 2) indicate remaining confusion challenges. Visual inspection of false positives (not shown) reveals concentration along ship wakes and wave crests where geometric similarity to debris filaments causes misclassification. Future work incorporating wake-suppression priors (e.g., penalizing linear high-aspect-ratio detections near ship tracks) or temporal persistence filtering (debris persists across acquisitions, wakes dissipate) could reduce these failure modes while preserving foam discrimination capability.

**False-Positive Characterization and Mitigation.** Of the 37,918 false positive pixels in the test set (FPR  $\approx 0.3\%$ ), visual inspection reveals two dominant error sources: (1) ship wakes, whose linear high-reflectance patterns exhibit geometric similarity to elongated debris filaments, and (2) wave crests and whitecaps, where transient surface brightness mimics debris spectral signatures. Notably, foam, often cited as the primary confusion source is successfully rejected (mean confidence 0.044, Section 4.2), indicating that learned SWIR-based discrimination is effective for organic bright features but less reliable for geometric mimics. Regarding thresholding, the current decision boundary ( $\tau = 0.5$ ) reflects a recall-oriented operating point appropriate for screening workflows where missed debris carries higher cost than false alarms. Adjusting  $\tau$  involves a direct precision-recall trade-off: higher thresholds would reduce wake-related false positives but at the cost of missing thin debris structures at patch boundaries.

Post-processing offers more targeted mitigation without sacrificing recall: morphological opening can remove isolated false-positive pixels, shape-based filtering can exploit the high aspect ratio of linear wakes versus amorphous debris, and temporal persistence filtering across consecutive acquisitions can suppress transient wake signatures while preserving persistent debris detections. These strategies are identified as promising directions for operational deployment but are considered beyond the scope of the current binary reformulation study.

#### 4.5.5 Actionable guidelines for operational deployment

Cross-dataset validation results establish concrete guidelines for operational marine debris monitoring systems:

1. **Prioritize Geographic Diversity Over Exhaustive Annotation:** Training on spatially diverse datasets (e.g., MADOS: 47 globally distributed tiles) yields superior generalization compared to exhaustive annotation of geographically concentrated regions. Operational dataset curation should emphasize spatial stratification, ensuring representation of diverse coastal zones, open ocean conditions, pollution contexts, and oceanographic regimes, over maximizing patch counts from individual locations. Our results demonstrate that 2,529 globally diverse patches outperform 2,173 geographically concentrated patches by 5.64 F1 points for cross-domain generalization.
2. **Expected Performance on Novel Regions:** Operational systems can expect F1-scores of 0.86 to 0.89 when deployed on geographically novel regions without additional training or fine-tuning. This performance level is sufficient for operational screening and alerting workflows where false negatives are costly (recall = 0.8359 to 0.9286 captures 83.6 to 92.9% of debris) while maintaining high specificity (greater than 0.995, false positive rate less than 0.5%). The 4.38% average degradation provides realistic expectations for deployment planning.
3. **Two-Stage Deployment Strategy:** For global operational monitoring: (a) Train initial models on geographically heterogeneous datasets analogous to MADOS (diverse tiles, varied environmental conditions, global spatial coverage); (b) Optionally fine-tune on region-specific data analogous to MARIDA for applications requiring F1 greater than 0.90. The 5.64 F1-point improvement from MARIDA to MADOS (0.8333) to MADOS to MARIDA (0.8897) demonstrates the value of diverse initial training before specialization.
4. **Architecture Matters Less Than Data Diversity:** Standard U-Net with proper data engineering (geographic diversity, binary reformulation, composite loss) achieves cross-dataset F1 = 0.8897, exceeding specialized architectures' within-dataset performance (MAP-Mapper: 0.8800). Investment in diverse training data yields better returns than architectural complexity for cross-domain robustness.

For large-area monitoring, we recommend  $128 \times 128$  inputs when resources permit, or  $64 \times 64$  for real-time or embedded settings, paired with prevalence-aware thresholding and light post-processing. This configuration preserves the recall advantages evidenced by the confusion-matrix analysis while keeping the rate of false alarms manageable, thereby aligning the model's behavior with operational risk tolerance in marine-debris surveillance.

## 5 Conclusion

This study presents the first systematic evaluation of a binary reformulation for marine debris detection on the combined MARIDA plus MADOS benchmark under extreme class imbalance, approximately 1:33,875. By consolidating non-debris

categories into a single background class, extracting multi-scale sub-patches with 50% overlap, and training a standard U-Net with weighted loss, we obtain a competitive baseline despite the rarity of positive pixels. On the held-out test set, the model reaches F1 equals 90.1%, IoU equals 82.0%, with precision equals 85.5% and recall equals 95.2%, demonstrating that careful data engineering including binary reformulation, sliding-window extraction, and class-weighted optimization can yield strong performance even with a baseline architecture.

At the same time, the gap to recent attention-based models on MARIDA, such as ResAttUNet with 95% F1, underscores the value of architectural advances for this task. Our ablations indicate that an 8 times upsampling stage and 50% overlap provide the best accuracy-efficiency trade-off, while denser overlaps deliver diminishing returns at substantially higher cost. The confusion-matrix analysis shows a recall-oriented operating point with controlled false positives, suggesting that lightweight post-processing including area filters, morphology, and temporal aggregation can further raise precision without retraining.

Cross-dataset validation experiments provide critical evidence of model robustness for operational deployment. Bidirectional experiments training on one dataset and testing on the other reveal asymmetric generalization: MADOS-trained models achieve F1 = 0.890 on MARIDA (only 1.25% degradation), while MARIDA-trained models achieve F1 = 0.833 on MADOS (7.55% degradation). The average 4.38% cross-dataset performance drop is substantially lower than typical 10 to 20% degradation in remote sensing domain shift scenarios, validating that binary reformulation with composite loss functions constitutes a generalization-preserving strategy. A key finding is that geographic diversity in training data matters more than patch volume: MADOS's 47 globally distributed tiles yield 5.64 F1 points superior generalization compared to MARIDA's more concentrated coverage, despite nearly identical patch counts. This establishes actionable guidelines for operational systems, prioritize spatially stratified sampling across diverse marine environments over exhaustive annotation of individual locations, and expect F1-scores of 0.86 to 0.89 when deploying on novel geographic regions without fine-tuning.

The results establish a reproducible baseline for binary debris detection across MARIDA plus MADOS and validate the effectiveness of problem reformulation under severe imbalance. Future work should integrate attention mechanisms and boundary-aware objectives, explore focal or unified focal variants for rare structures, adapt thresholds or calibrate decisions by sea state, and leverage temporal ensembling to suppress isolated false alarms. Extensions to multi-class segmentation with class-specific imbalance handling and uncertainty estimation are also promising directions toward operational deployment

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

## Author contributions

SB: Writing – review & editing, Formal analysis, Writing – original draft, Visualization, Data curation, Conceptualization, Methodology, Validation, Investigation. LM: Software, Writing – review & editing, Data curation, Methodology, Formal analysis, Writing – original draft. MB: Validation, Writing – review & editing, Supervision, Writing – original draft. AT: Validation, Supervision, Writing – review & editing, Writing – original draft. MM: Writing – original draft, Validation, Writing – review & editing, Supervision, Software. IT: Validation, Writing – review & editing, Writing – original draft.

## Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was partially funded by Sapienza Università di Roma “Progetti Piccoli e Medi 2024”.

## Acknowledgments

The authors thank the developers of the MARIDA and MADOS datasets for making these valuable resources publicly available.

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Balsi, M., Bouchelaghem, S., Conti, L., Moroni, M., and Scalia, R. (2023). "Real-time plastic litter detection using hyperspectral sensing on drone," in *2023 13th workshop on hyperspectral imaging and signal processing: evolution in remote sensing (WHISPERS)* (Athens, Greece: IEEE), 1–4.
- Balsi, M., Moroni, M., and Bouchelaghem, S. (2025). Plastic litter detection in the environment using hyperspectral aerial remote sensing and machine learning (Athens, Greece). *Remote Sens.* 17, 938. doi: 10.3390/rs17050938
- Bouchelaghem, S., Balsi, M., and Moroni, M. (2026). Attention-gated u-net for robust cross-domain plastic waste segmentation using a uav-based hyperspectral swir sensor. *Remote Sens.* 18, 182. doi: 10.3390/rs18010182
- Bouchelaghem, S., Tibermacine, I. E., Balsi, M., Moroni, M., and Napoli, C. (2024). "Cross-domain machine learning approaches using hyperspectral imaging for plastics litter detection," in *2024 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, (IEEE Proceedings) Vol. 36–40. doi: 10.1109/M2GARSS57310.2024.10537535
- Buda, M., Maki, A., and Mazurkowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106, 249–259. doi: 10.1016/j.neunet.2018.07.011
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953
- Duarte, M. M., and Azevedo, L. (2023). Automatic detection and identification of floating marine debris using multispectral satellite imagery. *IEEE Trans. Geosci. Remote Sens.* 61, 1–15. doi: 10.1109/TGRS.2023.3283607
- Garaba, S. P., and Dierssen, H. M. (2018). An airborne remote sensing case study of synthetic hydrocarbon detection using short wave infrared absorption features identified from marine-harvested macro- and microplastics. *Remote Sens. Environ.* 205, 224–235. doi: 10.1016/j.rse.2017.11.023
- Guanter, L., Kaufmann, H., Segl, K., Foerster, S., Rogass, C., Chabrillat, S., et al. (2015). The enMAP spaceborne imaging spectroscopy mission for earth observation. *Remote Sens.* 7, 8830–8857. doi: 10.3390/rs70708830
- Hu, C. (2021). Remote detection of marine debris using satellite observations in the visible and near infrared spectral range: Challenges and potentials. *Remote Sens. Environ.* 259, 112414. doi: 10.1016/j.rse.2021.112414
- Jambeck, J. R., Geyer, R., Wilcox, C., Siegler, T. R., Perryman, M., Andrady, A., et al. (2015). Plastic waste inputs from land into the ocean. *Science* 347, 768–771. doi: 10.1126/science.1260352
- Kaviya, K., and Bhavani, R. (2025). Advanced deep learning framework aquasense for comprehensive marine pollutant detection using sentinel-2 multispectral data. *J. Inf. Syst. Eng. Manage.* 10, 37–50. doi: 10.52783/jisem.v10i51s.10363
- Kikaki, K., Kakogeorgiou, I., Hoteit, I., and Karantzas, K. (2024). Detecting marine pollutants and sea surface features with deep learning in sentinel-2 imagery. *ISPRS J. Photogrammetry Remote Sens.* 212, 338–354. doi: 10.1016/j.isprs.2024.02.017
- Kikaki, K., Kakogeorgiou, I., Mikeli, P., Raitos, D. E., and Karantzas, K. (2022). MARIDA: A benchmark for marine debris detection from sentinel-2 remote sensing data. *PLoS One* 17, e0262247. doi: 10.1371/journal.pone.0262247
- Lee, C. M., Cable, M. L., Hook, S. J., Green, R. O., Ustin, S. L., Mandl, D. J., et al. (2015). An introduction to the NASA hyperspectral infrared imager (HypIRI) mission and preparatory activities. *Remote Sens. Environ.* 167, 6–19. doi: 10.1016/j.rse.2015.06.012
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision (ICCV)*. (ICCV Conference Proceedings), 2980–2988. doi: 10.1109/ICCV.2017.324
- Loizzo, R., Guarini, R., Longo, F., Scopa, T., Formaro, R., Facchinetti, C., et al. (2019). "PRISMA: the italian hyperspectral mission," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. (IEEE IGARSS Conference), 175–178. doi: 10.1109/IGARSS.2019.8899272
- Marsocci, V., Jia, X., Caro-Cuenca, M., Owers, C. J., Roscher, R., Schmitt, M., et al. (2024). PANGAEA: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint*. arXiv:2412.04204. doi: 10.48550/arXiv.2412.04204
- Mohammed, A. (2022). Resattunet: Detecting marine debris using an attention activated residual unet. *ARXIV PREPRINTS*.
- Moroni, M., Balsi, M., and Bouchelaghem, S. (2025). Plastics detection and sorting using hyperspectral sensing and machine learning algorithms. *Waste Manage.* 203, 114854. doi: 10.1016/j.wasman.2025.114854
- Persello, C., and Bruzzone, L. (2014). Active and semisupervised learning for the classification of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 52, 6937–6956. doi: 10.1109/TGRS.2014.2305805
- Prakash, A., and Zielinski, O. (2025). Ai-enhanced real-time monitoring of marine pollution: part 1 - a state-of-the-art and scoping review. *Front. Mar. Sci.* 12. doi: 10.3389/fmars.2025.1486615
- Rao, M. S. (2025). Hybrid deep learning approach for marine debris detection in satellite imagery using unet with resnext50 backbone. *J. Appl. Sci. Technol. Trends* 6, 50–60. doi: 10.38094/jastt61243
- Rast, M., and Painter, T. H. (2019). Earth observation imaging spectroscopy for terrestrial systems: an overview of its history, techniques, and applications of its missions. *Surveys Geophysics* 40, 303–331. doi: 10.1007/s10712-019-09517-z
- Rußwurm, S. J. V., and Marc and Tuia, D. (2023). Large-scale detection of marine debris in coastal areas with sentinel-2. 26, 108402. doi: 10.1016/j.isci.2023.108402
- Themistocleous, K., Papoutsas, C., Michaelides, S., and Hadjimitsis, D. (2020). Investigating detection of floating plastic litter from space using sentinel-2 imagery. *Remote Sens.* 12, 2648. doi: 10.3390/rs12162648
- Tuia, D., Persello, C., and Bruzzone, L. (2016). Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geosci. Remote Sens. Magazine* 4, 41–57. doi: 10.1109/MGRS.2016.2548504
- Vanhellemont, Q. (2019). Adaptation of the dark spectrum fitting atmospheric correction for aquatic applications of the landsat and sentinel-2 archives. *Remote Sens. Environ.* 225, 175–192. doi: 10.1016/j.rse.2019.03.010
- Wang, Y., Liu, S., He, Y., and Zhang, Y. (2025). Yolo11-yx: An efficient algorithm for marine debris target detection. *Mar. Pollut. Bull.* 221, 118511. doi: 10.1016/j.marpolbul.2025.118511
- Wang, Z., Chen, J., and Hoi, S. C. H. (2021). Deep learning for image super-resolution: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 3365–3387. doi: 10.1109/TPAMI.2020.2982166