Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Satellite-based marine pollution monitoring for African coastal energy infrastructure: Transformer semantic segmentation of Sentinel-2 multispectral imagery on the MADOS benchmark

Temitope B. Waheeb

waheebtemitope@gmail.com

University of Ibadan

Williams O. Adaji-Agbane

University of Uyo

Adamson O. Oloyede

NASRDA

Research Article

Keywords: Marine pollution, Oil spill detection, Sentinel-2, Semantic segmentation, Vision transformers, MADOS

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11300519/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Chronic oil pollution and marine litter around coastal energy infrastructure remain poorly monitored in many African waters, where field access is limited and cloud cover restricts routine observation. Publicly available Sentinel-2 imagery offers a practical basis for operational monitoring, yet most published detectors rely on purpose-built architectures whose advantages over general-purpose models are unclear. In this study, two widely adopted transformer segmentation architectures, SegFormer-B2 and Mask2Former, were evaluated on the Marine Debris and Oil Spill (MADOS) dataset of 174 Sentinel-2 scenes with roughly 1.5 million annotated pixels across 15 classes. Both models were adapted for four-channel red, green, blue and near-infrared input by weight inflation of pretrained filters, trained on the official MADOS splits, and compared against the published MariNeXt baseline using mean intersection over union (mIoU) and per-class F1. SegFormer-B2 reached a mIoU of 0.726 and a macro-F1 of 0.807, and Mask2Former reached 0.623 and 0.714, against 0.643 and 0.760 for MariNeXt. Both transformers detected oil spills (F1 of 0.94 and 0.95 against 0.78) and marine debris (0.81 and 0.75 against 0.51) more accurately than the baseline, but lost ground on sparse floating algae, foam and jellyfish, a pattern consistent with the red-edge and shortwave infrared bands omitted from the four-channel input. A pretrained general-purpose transformer can therefore exceed a purpose-built multispectral model on aggregate accuracy and on the pollutant classes of greatest operational interest, offering a transferable reference point for satellite-based pollution monitoring around African coastal energy infrastructure.

1. Introduction

Oil production and transport along the African Atlantic seaboard concentrate environmental risk in a narrow coastal band where communities, fisheries and mangrove ecosystems sit beside wells, pipelines and terminals. The Niger Delta illustrates the scale of the problem. Satellite analysis of the two largest Ogoniland spills of 2008 and 2009 delineated 393 square kilometres of vegetation mortality and placed more than one million people inside the impact zone [1]. Recent Earth observation studies estimate that spills continue to remove around 5,644 hectares of mangrove forest each year, with damage clustered along pipeline corridors [2], and chronic exposure mapping with Sentinel-1 radar confirms that the risk is persistent rather than episodic [3].

Monitoring in the region, however, is still largely indirect. Most published work infers pollution from vegetation stress indices after the fact rather than segmenting the pollutant itself on the water surface [4, 5]. Direct detection studies rely heavily on radar, which is robust to cloud but struggles to separate slicks from wind-related look-alikes and says little about floating debris [6, 7]. The freely available Sentinel-2 constellation, with 13 spectral bands at 10 to 60 m resolution and a five-day revisit [8], can in principle distinguish oil, plastics, foam and biological material by their spectral behaviour, and the release of the Marine Debris and Oil Spill (MADOS) dataset has made rigorous benchmarking of such multi-class detectors possible for the first time [9].

A practical question remains open for anyone building an operational system: do purpose-built multispectral architectures such as MariNeXt hold a real advantage over general-purpose transformer segmentation models that benefit from years of engineering, pretraining and community support? General-purpose models are easier to maintain, reproduce and deploy, which matters for institutions with limited computing budgets. This study addresses that question directly.

The contributions of this paper are threefold. First, the first systematic evaluation of SegFormer-B2 [10] and Mask2Former [11] on MADOS is provided, with both models adapted for four-channel red, green and blue (RGB) and near-infrared (NIR) Sentinel-2 input and trained on the official dataset splits. Second, both models are benchmarked against the published MariNeXt baseline [9] using mean intersection over union (mIoU) and per-class F1; SegFormer-B2 exceeds the baseline on both aggregate metrics and on the oil spill and marine debris classes, while Mask2Former trails it on aggregate despite a substantially larger parameter count. Third, a per-class analysis links the detectability of individual pollutant classes to the spectral bands supplied to the models, distinguishing this satellite multispectral setting from the RGB aerial imagery literature [12], and the implications for monitoring around African coastal energy infrastructure are discussed.

2. Related work

2.1 Oil spill detection from radar imagery

Satellite oil spill detection matured first around synthetic aperture radar (SAR), where slicks damp capillary waves and appear as dark patches, and where the central difficulty is separating true spills from look-alikes such as low-wind areas [6, 13]. Krestenitis et al. [14] established the standard deep learning benchmark for this modality, comparing six segmentation networks on a five-class Sentinel-1 dataset, with later work addressing the severe class imbalance through two-stage designs and specialised losses [15] or boundary supervision [16]. Radar remains blind, though, to the spectral signatures that separate oil from debris and biological matter, which motivates the optical route taken here.

2.2 Marine pollutant detection from Sentinel-2

Optical detection of floating pollutants was demonstrated by Biermann et al. [17], who showed that macroplastic patches are separable from seaweed in Sentinel-2 data through spectral shape and a floating debris index. Controlled target experiments from the Plastic Litter Project quantified how biofouling and submersion modify those signatures [18], and case studies traced debris transport in coastal waters [19]. The MARIDA benchmark formalised marine debris detection as a machine learning task on Sentinel-2 [20], and detectors trained on it have been scaled to coastal areas worldwide, including an evaluation site at Accra, Ghana [21]. MADOS then broadened the task to 15 classes spanning oil spills, debris, sargassum, ships and water types across 174 scenes, and introduced MariNeXt, a SegNeXt-based architecture [22] that reached a macro-F1 of 0.76 and a mIoU of 0.64 [9]. Follow-up studies confirm these baseline figures and treat MADOS as the reference benchmark for multi-class marine pollution segmentation [23]. A recent framework, AquaSense, reports higher MADOS

scores [24], but its evaluation protocol is not described in enough detail to confirm comparability with the official splits, so MariNeXt is treated here as the authoritative baseline.

2.3 Aerial imagery for oil spill segmentation

A parallel line of work performs oil spill segmentation from low-altitude RGB imagery. The LADOS dataset provides 3,388 unmanned aerial vehicle (UAV) images across six classes, including emulsions and sheens that matter for early leak response [12], building on earlier segmentation work by the same group [25]. This aerial literature is related to the present work but distinct in sensor modality and scope: UAV platforms offer centimetre detail over small areas on demand, whereas Sentinel-2 trades resolution for free, repeated, wide-area coverage and, critically, for spectral bands beyond the visible range. The present study should therefore be read as complementary to, not competing with, the aerial line.

2.4 Transformer architectures for semantic segmentation

Transformers [26] entered vision through the Vision Transformer (ViT) [27] and now anchor two widely used segmentation families. SegFormer pairs a hierarchical mix transformer (MiT) encoder with a lightweight multilayer perceptron (MLP) decoder, giving strong accuracy at modest cost [10]. Mask2Former reframes segmentation as mask classification, constraining cross-attention to predicted mask regions [11], extending the earlier MaskFormer design [28]. Both have been applied to Earth observation tasks [29], but neither has been evaluated on MADOS, and it is notable that MariNeXt is built on SegNeXt, an architecture that argues convolutional attention beats self-attention for segmentation [22]. The comparison reported here therefore also tests that claim in the multispectral marine setting.

3. Materials and methods

3.1 System overview and pipeline

Figure 1 summarises the methodological workflow of the study, from data acquisition and preprocessing through four-channel model adaptation and training to test-set evaluation against the published baseline. A four-channel Sentinel-2 input patch (RGB and NIR) is passed through two independently adapted transformer-based segmentation architectures, SegFormer-B2 and Mask2Former, each with its patch-embedding layer modified to accept four input channels rather than the standard three. Each architecture retains its native encoder-decoder structure (the hierarchical MiT-B2 encoder with an all-MLP decoder for SegFormer-B2, and the Swin-Base encoder with a query-based mask decoder for Mask2Former) before producing a 16-class segmentation map (15 MADOS thematic classes plus the non-annotated background class) at the native 240 × 240 patch resolution.

The data preparation stage of this workflow (Fig. 1) proceeds as follows. Raw MADOS scenes were obtained from Zenodo as per-band GeoTIFF files at three native Sentinel-2 resolutions (10 m, 20 m, 60 m); four bands at 10 m resolution (red, green, blue and NIR) were selected and stacked to form the model input, following the band-selection rationale detailed in Section 3.3. Patches containing corrupted

pixel data (not-a-number, NaN, values) were identified and dropped, the remaining data were normalised per channel, and the dataset was partitioned using the official MADOS train/validation/test split.

3.2 The MADOS dataset

MADOS comprises 174 Sentinel-2 scenes acquired between 2015 and 2022, with about 1.5 million sparsely annotated pixels across 15 classes covering marine debris, dense and sparse sargassum, natural organic material, ships, oil spills, several water types, foam, waves and wakes, oil platforms, jellyfish, and sea snot [9]. The imagery carries 11 multispectral bands at 10 to 60 m resolution [8]. The official location- and date-based train, validation and test splits, in an approximate 50/25/25 ratio, were used exactly as released [9], which keeps the reported figures directly comparable with the published baseline. The dataset is openly available from Zenodo (https://doi.org/10.5281/zenodo.10664073).

3.3 Preprocessing and band handling

Four of the eleven available Sentinel-2 bands were used at native 10 m resolution: Red (B04, approximately 665 nm), Green (B03, approximately 559 nm), Blue (B02, approximately 492 nm), and Near-Infrared (B08, approximately 833 nm), rather than resampling and stacking the full 11-band set. NIR was included alongside the visible bands for its established discriminative value in oil–water spectral separation. Pixel values (Rayleigh-corrected reflectance) were normalised per-channel using dataset-specific mean and standard deviation statistics computed from a 200-patch sample of the training set (μ = [0.0520, 0.0439, 0.0351, 0.0363], σ = [0.0268, 0.0285, 0.0286, 0.0537] for R, G, B, NIR respectively), rather than reusing ImageNet statistics, since Sentinel-2 reflectance values occupy a substantially narrower numerical range than standard 8-bit RGB imagery. Training used the native 240×240 MADOS patch size without cropping or resizing, with horizontal flip ($p = 0.5$), vertical flip ($p = 0.3$) and random 90° rotation ($p = 0.3$). Patches containing no-data pixels were excluded, removing 12 of 1,433 training patches.

Class imbalance, with Marine Water and Sediment-Laden Water dominating the pixel-level label distribution and rare classes such as Jellyfish and Foam comprising a small fraction of annotated pixels, was not addressed through loss weighting in the final reported runs. Class-weighted cross-entropy was tested during development in both a raw inverse-frequency form and a capped, log-dampened variant, but neither improved early-epoch validation mIoU relative to unweighted cross-entropy. This decision also reflects the primary objective of the study, which centres on Oil Spill detection rather than uniform performance across all fifteen MADOS classes. Oil Spill is itself moderately well represented, with 39,128 pixels in the training sample used for weight estimation, so inverse-frequency schemes designed to lift severely underrepresented classes such as Jellyfish and Foam would have suppressed its loss contribution, working against the central detection objective. Both models were therefore trained with standard unweighted losses: cross-entropy for SegFormer-B2 and the native Mask2Former classification and mask losses computed after Hungarian matching.

3.4 Model architectures and adaptation

SegFormer-B2 uses the MiT-B2 hierarchical encoder with an all-MLP decoder [10]. Mask2Former was configured with a Swin-Base backbone [11]. For both models the input stem was adapted from 3 to 4 channels, corresponding to the RGB and NIR bands selected in Section 3.3, by weight inflation of pretrained RGB filters: the pretrained convolutional weights for the red, green, and blue channels were copied directly into the corresponding channels of the new stem, and the additional NIR channel was initialised as the mean of the three pretrained RGB channel weights. This preserved the pretrained representational quality of the visible-spectrum filters while providing a principled, non-random initialisation for the new channel, rather than a full re-initialisation of the input layer. The classification head was set to 16 output classes, corresponding to the 15 MADOS thematic classes plus the non-annotated background class.

Encoders were initialised from ImageNet-1k weights for SegFormer-B2 (nvidia/mit-b2) and from ImageNet-21k weights fine-tuned on ADE20K semantic segmentation for Mask2Former (facebook/mask2former-swin-base-IN21k-ade-semantic). The same channel-adaptation strategy was applied identically to both models so that the comparison isolates encoder and decoder design rather than the treatment of the input channels.

3.5 Training configuration

Both models were trained for 40 epochs using the AdamW optimiser, with a linear warmup over the first 10% of training steps followed by linear decay, batch size 8, on an NVIDIA Tesla T4 GPU. SegFormer-B2 used an initial learning rate of 6×10^{-5} with a weight decay of 0.01, and was trained with unweighted cross-entropy loss. Mask2Former used an initial learning rate of 5×10^{-5} with the same weight decay, and was trained with its native combined loss, comprising classification and mask losses computed after Hungarian bipartite matching, following the standard Mask2Former training formulation. Gradient norms were clipped at 1.0 for both models. Random seeds were fixed at 42 across NumPy, PyTorch, and the Python random module, and the checkpoint achieving the highest validation mIoU was retained for final test-set evaluation in each case.

The MariNeXt baseline figures are taken directly from the original publication, which trained from scratch for 80 epochs with Adam at an initial learning rate of 2×10^{-4}, applying learning rate decay at epochs 45 and 65, and using a batch size of 5 [9]. Unlike the present study, the models in that work were initialised from scratch rather than from pretrained checkpoints, and used the full 11-band Sentinel-2 input rather than the 4-band RGB and NIR configuration used here; these differences in initialisation and input dimensionality should be considered when interpreting the comparison against this baseline.

3.6 Evaluation metrics

Following the MADOS protocol, per-class F1 and macro-averaged F1 are reported together with mIoU on the held-out test split. Per-class F1 exposes performance on rare but operationally important classes such as marine debris, which aggregate metrics conceal.

3.7 Reproducibility

Table 1 summarises the implementation details necessary to reproduce the reported results. All experiments were conducted using PyTorch 2.10 (CUDA 12.8) and the Hugging Face Transformers library, on a single NVIDIA Tesla T4 GPU via the Kaggle computing platform. A fixed random seed (42) was applied across NumPy, PyTorch, and the Python random module for all experiments. For each model, the checkpoint corresponding to the highest validation mIoU across the 40 training epochs was selected for final test-set evaluation, rather than the final-epoch checkpoint, in order to mitigate the effects of overfitting observed after peak validation performance. Code, configuration files and preprocessing scripts are openly available (see Code availability).

Table 1
Reproducibility details

Parameter	Details
Framework	PyTorch 2.10 (CUDA 12.8); HuggingFace Transformers
Hardware	NVIDIA Tesla T4 GPU, Kaggle platform
Random seed	42 (NumPy, PyTorch, Python random)
Input resolution	240x240 px, 4 channels (RGB + NIR)
SegFormer-B2 training	40 epochs, batch size 8, AdamW, lr 6e-5
Mask2Former training	40 epochs, batch size 8, AdamW, lr 5e-5
Checkpoint selection	Best validation mIoU across 40 epochs

4. Results

4.1 Comparison with the MariNeXt baseline

Table 2 summarises test-set performance. SegFormer-B2 reached a mIoU of 0.726 and a macro-F1 of 0.807; Mask2Former reached 0.623 and 0.714. Relative to MariNeXt (mIoU 0.643, macro-F1 0.760), SegFormer-B2 represents a gain of 8.3 points on mIoU and 4.7 points on macro-F1, while Mask2Former represents a deficit of 2.0 points on mIoU and 4.6 points on macro-F1. SegFormer-B2 therefore outperforms both the MariNeXt baseline and Mask2Former on both metrics, with the largest architectural gap emerging on classes with sparse or texturally ambiguous annotations, discussed further in Section 4.2.

Table 2
Test-set performance on MADOS compared with the published MariNeXt baseline [9]

Model	mIoU	Macro-F1	Params (M)
MariNeXt (baseline)	0.643	0.760	Not reported
SegFormer-B2 (ours)	0.726	0.807	24.7
Mask2Former (ours)	0.623	0.714	106.9

4.2 Per-class analysis

Table 3 reports per-class F1 for all 15 MADOS classes, Fig. 2 presents the same F1 values graphically, and Fig. 3 gives the corresponding per-class IoU. SegFormer-B2 achieved the highest F1 on nine of the fifteen classes, two of them shared (Marine Water with Mask2Former, and Sediment-Laden Water with both other models), including near-perfect performance on Sediment-Laden Water (F1 = 1.00) and Oil Platform (F1 = 0.99) and strong results on Turbid Water (F1 = 0.97) and Shallow Water (F1 = 0.95). Mask2Former performed best on Oil Spill (F1 = 0.95) and Waves & Wakes (F1 = 0.78). The MariNeXt baseline remained the strongest on four classes: Dense Sargassum (F1 = 0.81, marginally above 0.80 for Mask2Former), Sparse Floating Algae (F1 = 0.84), Foam (F1 = 0.87) and, most notably, Jellyfish (F1 = 0.90), where both architectures in this study failed entirely (F1 = 0.00). The Jellyfish result is likely to reflect the copy-paste augmentation MariNeXt applies to severely underrepresented classes rather than a limitation specific to either transformer. Apart from Jellyfish, the weakest SegFormer-B2 class was Sparse Floating Algae (F1 = 0.61), whereas Mask2Former collapsed on Natural Organic Material (F1 = 0.03), a class on which SegFormer-B2 reached 0.78, indicating that this failure is tied to the query-based decoder of Mask2Former rather than to the class being unlearnable. On the two classes central to the objective of this study, both architectures performed strongly and exceeded the baseline: Oil Spill F1 reached 0.94 (SegFormer-B2) and 0.95 (Mask2Former), both well above the 0.78 of MariNeXt, and Marine Debris F1 reached 0.81 (SegFormer-B2) and 0.75 (Mask2Former), both above the 0.51 of MariNeXt, confirming that the unweighted training strategy adopted here did not compromise performance on the primary target class.

Table 3
Per-class F1 on the MADOS test split. MariNeXt values are from Kikaki et al. [9]

Class	MariNeXt F1	SegFormer-B2 F1	Mask2Former F1
Marine Debris	0.51	0.81	0.75
Dense Sargassum	0.81	0.78	0.80
Sparse Floating Algae	0.84	0.61	0.52
Natural Organic Material	0.57	0.78	0.03
Ship	0.63	0.90	0.84
Oil Spill	0.78	0.94	0.95
Marine Water	0.91	0.95	0.95
Sediment-Laden Water	1.00	1.00	1.00
Foam	0.87	0.84	0.63
Turbid Water	0.90	0.97	0.93
Shallow Water	0.64	0.95	0.89
Waves & Wakes	0.38	0.72	0.78
Oil Platform	0.84	0.99	0.89
Jellyfish	0.90	0.00	0.00
Sea Snot	0.81	0.88	0.78

4.3 Qualitative results

4.3.1 SegFormer-B2

Figure 4 shows predictions for representative test scenes covering Natural Organic Material and Oil Spill, the class central to the objective of this study. For Oil Spill, SegFormer-B2 agrees closely with the ground truth in the simpler single-slick scene, correctly localising the annotated region against open water, although it also assigns surrounding water pixels to a broad, spatially uniform class across most of the frame. In the more complex coastal scene the model separates several distinct surface classes, correctly segmenting the coastline and open water even where the sparse annotation protocol left most of the patch unlabelled. For Natural Organic Material the ground truth consists of a handful of scattered annotated pixels per patch, and predictions do not closely track those points, reflecting the difficulty of localising a class of very limited spatial extent. Throughout, the model produces spatially coherent predictions across the full patch while the ground truth annotates only a small fraction of pixels, a property of the sparse annotation protocol rather than a modelling artefact, so disagreement in unannotated regions does not necessarily indicate error.

4.3.2 Mask2Former

Figure 5 shows the corresponding Mask2Former predictions for the same two classes. Oil spill localization is comparable to SegFormer-B2, consistent with the near-identical F1 values in Table 3, whereas the collapse on Natural Organic Material recorded quantitatively (F1 = 0.03) leaves that class effectively unrepresented in the prediction maps.

5. Discussion

5.1 Spectral bands and class detectability

The per-class pattern in Table 3 divides along an informative line. Both transformers exceed the baseline on classes carrying strong spatial or brightness signatures, including Oil Spill, Marine Debris, Ship, Oil Platform, Shallow Water, Turbid Water and Waves and Wakes. Oil films alter both surface roughness and near-infrared reflectance relative to clean water, so the four-channel configuration retains most of the signal that class depends on, which is consistent with the F1 values of 0.94 and 0.95 obtained here.

The baseline holds its advantage on precisely those classes whose separability rests on the bands that were omitted. Sparse Floating Algae (0.84 against 0.61 and 0.52) and Foam (0.87 against 0.84 and 0.63) involve floating matter that often occupies only part of a pixel and must be distinguished from water by spectral shape rather than brightness or texture. The indices established for that purpose depend on red-edge and shortwave infrared reflectance: the floating debris index of Biermann et al. [17] combines red-edge, near-infrared and shortwave infrared bands, and the unmixing experiments of Papageorgiou et al. [18] show that biofouling and partial submersion shift exactly those regions of the spectrum. Excluding B05 to B07 and B11 to B12 therefore removes the cues on which photosynthetic and biogenic surfaces rely, while leaving hydrocarbon and structural classes largely intact. This reading follows from the per-class pattern rather than a controlled ablation, and a systematic band ablation is the natural next experiment.

5.2 General-purpose transformers against a purpose-built model

SegFormer-B2 exceeds MariNeXt by 8.3 points of mIoU and 4.7 of macro-F1 while using fewer input bands, half the training epochs and no task-specific architectural design. The most plausible explanation is transfer: large-scale pretraining supplies spatial and textural priors that a model trained from scratch must learn from 1,421 sparsely annotated patches. Where the specialised design of the baseline should matter most, in sub-pixel spectral discrimination, it does retain an advantage; where general visual priors dominate, pretraining wins. Input dimensionality, initialisation, epoch budget and batch size all differ between the two studies, so this is evidence that a pretrained general-purpose model is a strong and inexpensive default rather than proof of architectural superiority.

The comparison between the two transformers is better controlled, since they share bands, splits, schedule and channel adaptation. Mask2Former trails SegFormer-B2 by 10.3 points of mIoU despite roughly four times the parameters, converges more slowly, and collapses on Natural Organic Material. Its mask-classification formulation must first propose coherent object-like regions before labelling them [11], which suits densely annotated instances but is poorly matched to supervision that is sparse, scattered and rarely object-shaped. The simpler hierarchical encoder with an all-MLP decoder is the better default in this regime, and within this setting no support is found for the claim that convolutional attention outperforms self-attention for segmentation [22], although the confounds noted above prevent a strong architectural conclusion.

5.3 Implications for African coastal monitoring

For agencies monitoring the Gulf of Guinea, the practical reading is that freely available imagery and openly maintained architectures can support multi-class pollution screening without bespoke model development. Persistent cloud cover in the region will still interrupt optical coverage, so an operational system would pair this detector with radar-based screening [3, 7], using the multispectral pass to classify pollutant type once a candidate area is flagged. Because Russwurm et al. [21] already demonstrate Sentinel-2 debris detection at Accra, extending validated multi-class segmentation to West African waters is a realistic near-term step.

5.4 Limitations

MADOS scenes are globally distributed, but none capture Niger Delta events, so regional transfer remains untested and is the immediate next step of this work. Annotations are sparse rather than dense, which the official metrics account for but which limits boundary-quality assessment. Each configuration was trained once with a fixed seed on a single T4 GPU, so no confidence intervals accompany the reported figures, and validation mIoU oscillating between 0.73 and 0.79 during the SegFormer-B2 run suggests that run-to-run variation of a few points is plausible. The 40-epoch budget and a limited hyperparameter search mean neither model is necessarily at its ceiling, and the best SegFormer-B2 checkpoint at epoch 12 indicates that overfitting sets in early at this dataset size. Finally, the four-band configuration was selected for the oil spill objective and pretraining compatibility rather than optimality, so the spectral interpretation in Section 5.1 remains inferential; per-class values should also be read alongside annotated test-pixel counts, since classes with very few test pixels produce extreme values in both directions.

6. Conclusions

This study benchmarked SegFormer-B2 and Mask2Former on the MADOS Sentinel-2 marine pollution dataset against the purpose-built MariNeXt baseline. SegFormer-B2 exceeded the baseline on both aggregate metrics, reaching a mIoU of 0.726 and a macro-F1 of 0.807 against 0.643 and 0.760, while Mask2Former trailed it on aggregate despite roughly four times the parameters. On the classes central to monitoring around energy infrastructure both transformers were the stronger detectors, with oil spill

F1 of 0.94 and 0.95 and marine debris F1 of 0.81 and 0.75 against 0.78 and 0.51 for the baseline. The per-class analysis indicates that the omitted red-edge and shortwave infrared bands, rather than architecture alone, set the ceiling on biogenic and sub-pixel classes, and the work provides a documented, reproducible reference for satellite-based pollution monitoring in African coastal waters. Future work will assemble a Niger Delta validation set, run a systematic band ablation, and test radar-optical fusion for cloud-robust operational screening.

Declarations

Ethics approval and consent to participate

Not applicable. This study used only publicly available satellite imagery and annotations and involved no human participants or animals.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution

T. B.W.: Conceptualization, Investigation, Writing - original draft, Writing - review and editing. W.O. A: Methodology, Software, Formal analysis, Validation, Data curation, Visualization, Writing - review and editing. A.O.O.: Writing - review and editing. All authors read and approved the final manuscript.

Data Availability

The MADOS dataset analysed in this study is openly available in the Zenodo repository at https://doi.org/10.5281/zenodo.10664073 [9]. All results reported in this study were derived from that dataset using its official splits.

Code availability

The training code, configuration files and preprocessing scripts used in this study are openly available on GitHub at https://github.com/Bherney/Cross-Sensor-Oil-Spill-Detection.

References

1. Obida CB, Blackburn GA, Whyatt JD, Semple KT. Counting the cost of the Niger Delta's largest oil spills: satellite remote sensing reveals extensive environmental damage with more than 1 million people in the impact zone. Sci Total Environ. 2021;775:145854. https://doi.org/10.1016/j.scitotenv.2021.145854.
2. O'Farrell J, O'Fionnagain D, Babatunde AO, Geever M, Codyre P, Murphy PC, Spillane C, Golden A. Quantifying the impact of crude oil spills on the mangrove ecosystem in the Niger Delta using AI and Earth observation. Remote Sens. 2025;17(3):358. https://doi.org/10.3390/rs17030358.
3. Oseni RD, Omosanya KOL, Mosuro GO, Saka HO, Oludipe VD, Oshomoji A. Mapping persistent oil spill risk in the Niger Delta using Sentinel-1 SAR and environmental exposure indicators. Mar Pollut Bull. 2026;229:119674. https://doi.org/10.1016/j.marpolbul.2026.119674.
4. Kuta AA, Grebby S, Boyd DS. Remote monitoring of the impact of oil spills on vegetation in the Niger Delta, Nigeria. Appl Sci. 2025;15(1):338. https://doi.org/10.3390/app15010338.
5. Adebangbe SA, Dixon DP, Barrett B. Evaluating contaminated land and the environmental impact of oil spills in the Niger Delta region: a remote sensing-based approach. Environ Monit Assess. 2025. https://doi.org/10.1007/s10661-025-14584-4.
6. Brekke C, Solberg AHS. Oil spill detection by satellite remote sensing. Remote Sens Environ. 2005;95(1):1–13. https://doi.org/10.1016/j.rse.2004.11.015.
7. Kim M, Park J, Hyun C-U. A review of oil spill detection and monitoring techniques using satellite remote sensing data and the Google Earth Engine platform. J Mar Sci Eng. 2026;14:565. https://doi.org/10.3390/jmse14060565.
8. Drusch M, Del Bello U, Carlier S, Colin O, Fernandez V, Gascon F, et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens Environ. 2012;120:25–36. https://doi.org/10.1016/j.rse.2011.11.026.
9. Kikaki K, Kakogeorgiou I, Hoteit I, Karantzalos K. Detecting marine pollutants and sea surface features with deep learning in Sentinel-2 imagery. ISPRS J Photogramm Remote Sens. 2024;210:39–54. https://doi.org/10.1016/j.isprsjprs.2024.02.017.
10. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems 34. 2021. p. 12077–12090.
11. Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. p. 1290–1299. https://doi.org/10.1109/CVPR52688.2022.00135.

12. Gkountakos K, Melitou M, Ioannidis K, Demestichas K, Vrochidis S, Kompatsiaris I. LADOS: aerial imagery dataset for oil spill detection, classification, and localization using semantic segmentation. Data. 2025;10(7):117. https://doi.org/10.3390/data10070117.
13. Fingas M, Brown CE. A review of oil spill remote sensing. Sensors. 2018;18(1):91. https://doi.org/10.3390/s18010091.
14. Krestenitis M, Orfanidis G, Ioannidis K, Avgerinakis K, Vrochidis S, Kompatsiaris I. Oil spill identification from satellite images using deep neural networks. Remote Sens. 2019;11(15):1762. https://doi.org/10.3390/rs11151762.
15. Shaban M, Salim R, Abu Khalifeh H, Khelifi A, Shalaby A, El-Mashad S, et al. A deep-learning framework for the detection of oil spills from SAR data. Sensors. 2021;21(7):2351. https://doi.org/10.3390/s21072351.
16. Zhu Q, Zhang Y, Li Z, Yan X, Guan Q, Zhong Y, et al. Oil spill contextual and boundary-supervised detection network based on marine SAR images. IEEE Trans Geosci Remote Sens. 2022;60:1–10. https://doi.org/10.1109/TGRS.2021.3115492.
17. Biermann L, Clewley D, Martinez-Vicente V, Topouzelis K. Finding plastic patches in coastal waters using optical satellite data. Sci Rep. 2020;10:5364. https://doi.org/10.1038/s41598-020-62298-z.
18. Papageorgiou D, Topouzelis K, Suaria G, Aliani S, Corradi P. Sentinel-2 detection of floating marine litter targets with partial spectral unmixing and spectral comparison with other floating materials (Plastic Litter Project 2021). Remote Sens. 2022;14(23):5997. https://doi.org/10.3390/rs14235997.
19. Kikaki A, Karantzalos K, Power CA, Raitsos DE. Remotely sensing the source and transport of marine plastic debris in Bay Islands of Honduras (Caribbean Sea). Remote Sens. 2020;12(11):1727. https://doi.org/10.3390/rs12111727.
20. Kikaki K, Kakogeorgiou I, Mikeli P, Raitsos DE, Karantzalos K. MARIDA: a benchmark for marine debris detection from Sentinel-2 remote sensing data. PLoS ONE. 2022;17(1):e0262247. https://doi.org/10.1371/journal.pone.0262247.
21. Russwurm M, Venkatesa SJ, Tuia D. Large-scale detection of marine debris in coastal areas with Sentinel-2. iScience. 2023;26(12):108402. https://doi.org/10.1016/j.isci.2023.108402.
22. Guo M-H, Lu C-Z, Hou Q, Liu Z, Cheng M-M, Hu S-M. SegNeXt: rethinking convolutional attention design for semantic segmentation. In: Advances in Neural Information Processing Systems 35. 2022. p. 1140–1156.
23. Bouchelaghem S, Tibermacine IE, Moroni D, Tibermacine A. 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. 2026;13:1765021. https://doi.org/10.3389/fmars.2026.1765021.
24. Kaviya K, Bhavani R. Advanced deep learning framework AquaSense for comprehensive marine pollutant detection using Sentinel-2 multispectral data. J Inf Syst Eng Manag. 2025;10(51s):37–50. https://doi.org/10.52783/jisem.v10i51s.10363.

25. Krestenitis M, Orfanidis G, Ioannidis K, Avgerinakis K, Vrochidis S, Kompatsiaris I. A deep neural network for oil spill semantic segmentation in SAR images. In: Proceedings of the 25th IEEE International Conference on Image Processing (ICIP). 2018. p. 3773–3777. https://doi.org/10.1109/ICIP.2018.8451113.
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30. 2017. p. 5998–6008.
27. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations. 2021.
28. Cheng B, Schwing AG, Kirillov A. Per-pixel classification is not all you need for semantic segmentation. In: Advances in Neural Information Processing Systems 34. 2021. p. 17864–17875.
29. He X, Zhou Y, Zhao J, Zhang D, Yao R, Xue Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens. 2022;60:1–15. https://doi.org/10.1109/TGRS.2022.3144165.

Figures

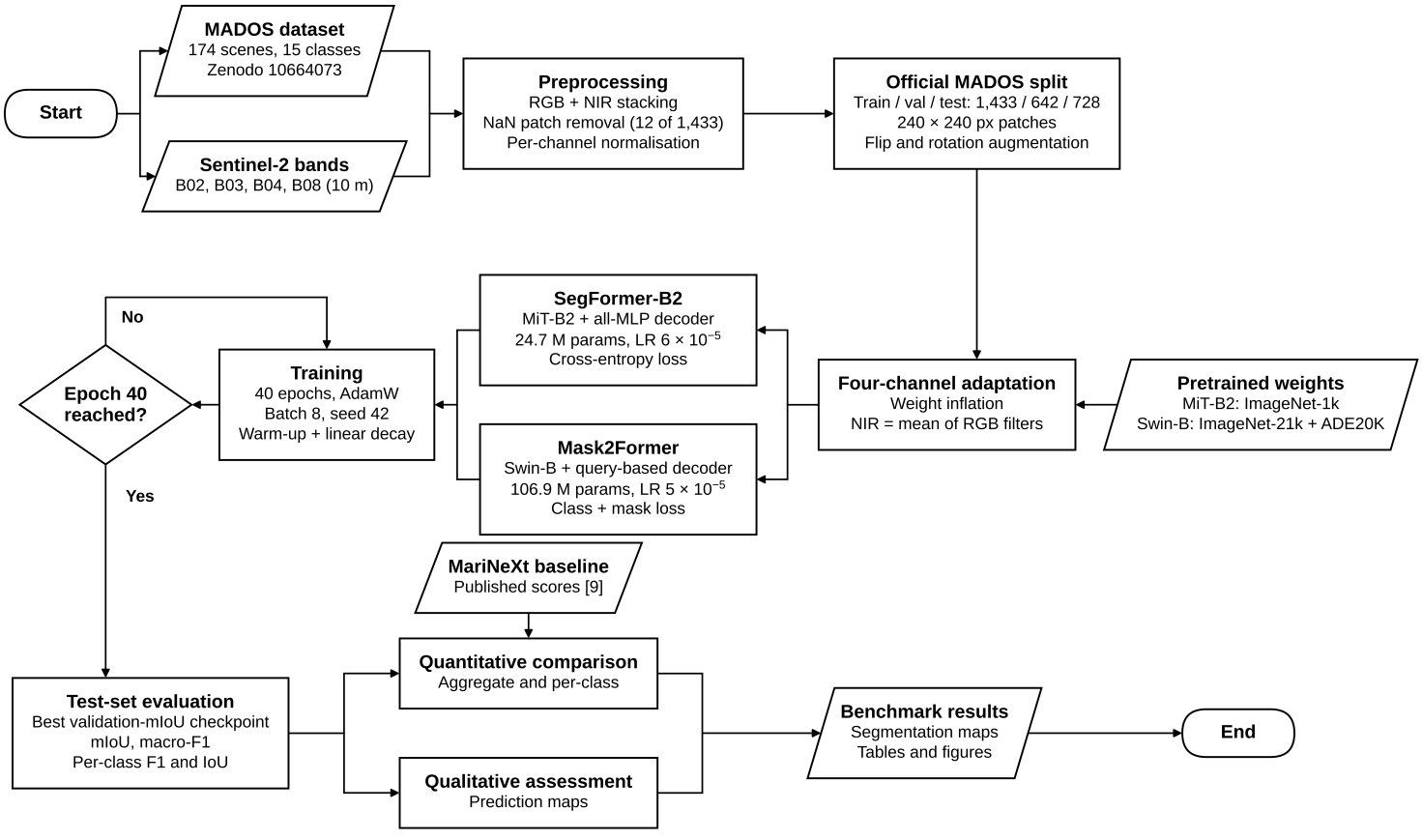


Figure 1

Methodological workflow of the study: data inputs, preprocessing, four-channel adaptation of SegFormer-B2 and Mask2Former, training with best-checkpoint selection, and test-set evaluation against the published MariNeXt baseline

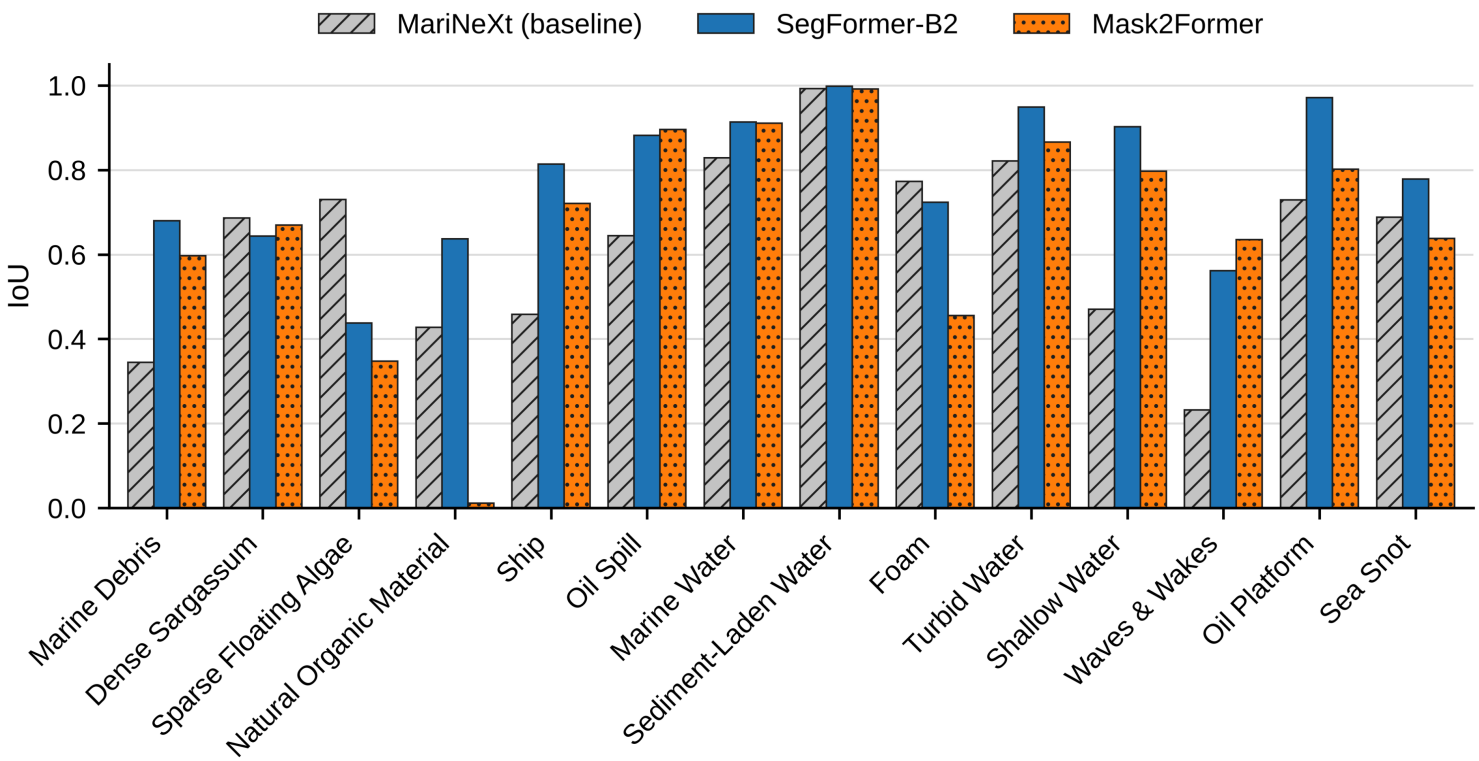


Figure 3

Per-class intersection over union (IoU) on the MADOS test split for MariNeXt (hatched), SegFormer-B2 (solid) and Mask2Former (dotted) across 14 classes, with Jellyfish omitted as in Fig. 2 (MariNeXt values from Kikaki et al. [9])