Benchmarking Machine-Learning Methods for Marine Debris Detection in Satellite Imagery

Brina Erznožnik, Filip Wolf, Luka Čehovin Zajc

University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana
E-pošta: luka.cehovin@fri.uni-lj.si

Abstract

Marine debris monitoring is an important yet underrepresented application of Earth observation, with existing benchmarks and foundation models primarily focused on terrestrial environments. This study presents a comprehensive evaluation of machine learning approaches for the semantic segmentation of floating marine debris using the MARIDA dataset. We compare classical machine learning, convolutional neural networks, and Earth Observation Foundation Models under a unified evaluation framework. Our results indicate that a task-specific, optimised convolutional model outperforms pretrained foundation models on this dataset, particularly in detecting small and spatially complex debris regions. We hypothesise that this performance gap on the chosen dataset stems from terrestrial pre-training biases and aggressive spatial downsampling in transformer backbones, which may hinder their direct transferability to aquatic environments. We highlight the continued importance of domain-specific model optimisation and emphasise the need for foundation models and benchmarks that better represent marine applications.

1 Introduction

Although water covers over 71% of the Earth's surface and is immensely important for various processes on the planet, marine environments, even surface ones, remain significantly under-monitored. Despite advances in modern Earth observation (EO) techniques using satellites, aquatic scenes are often not observed at the same spatial and temporal resolutions, leaving the sparse data available rarely utilised to its full potential. This is also apparent in the development of automatic monitoring techniques using machine learning, where large-scale water scenes are rarely used for evaluation. One such task is marine debris monitoring [1], where the challenge is detecting small regions of different types of debris in scenes predominantly composed of water with varying turbidity and surface properties, which is unlike typical tasks such as land-use classification or agricultural monitoring [2].

This study addresses this gap by evaluating modern machine learning methods for marine debris segmentation using the publicly available MARIDA dataset [3], thereby extending the assessment of contemporary EO models to an aquatic application that is poorly represented in existing benchmarks. Specifically, we compare classical machine learning methods, an optimised convolutional neural network (CNN), and several advanced Earth Observation Foundation Models (EOFMs). We further adapt multiple EOFMs that have not previously been evaluated for this task and assess their performance against established baselines to examine their suitability for marine debris monitoring.

Detecting and segmenting floating marine debris presents a severe spatial challenge. The combination of extreme class imbalance and the small spatial extent of debris motivates the inclusion and evaluation of approaches ranging from classical machine learning methods to modern deep learning architectures and EOFMs. Our exploration reveals a notable performance gap between a localized, domain-optimized CNN architecture and current EOFMs. This leads us to hypothesize that general-purpose foundation models are constrained on aquatic tasks by a terrestrial pre-training bias and aggressive spatial downsampling.

2 Related Work

The rapid development of machine learning (ML) methods for EO has created an increasing need for standardised benchmarks that enable fair and reproducible comparisons of models across diverse tasks and datasets. Recent benchmark suites such as GeoBench [4] and PANGAEA [5] address this need by providing unified evaluation protocols spanning land-cover classification, crop mapping, change detection, and other terrestrial applications, while also facilitating the assessment of modern EOFMs. However, these benchmarks provide only limited coverage of marine debris segmentation and other applications in open-water environments.

Existing marine debris datasets, therefore, remain the primary means of evaluating methods for this application. Early datasets, such as FloatingObjects [6], established initial ML baselines using classical classifiers, including Support Vector Machines (SVM), Random Forests (RF), and Naive Bayes. With the introduction of the MARIDA dataset, Kikaki et al. [3] provided a more comprehensive benchmark by evaluating RF classifiers alongside a baseline UNet [7] architecture.

Subsequent research has primarily focused on improving segmentation performance through architectural innovations. Priyadarshini et al. [8] proposed the Residual Block Attention-Activated UNet to improve boundary

preservation for small debris regions, while Rußwurm et al. [9] refined MARIDA annotations and re-evaluated established methods, including RF [10], UNet [7], and UNet++ [11]. Despite the growing interest in EOFMs, their performance on marine debris segmentation remains largely unexplored, as this application is still underrepresented in existing EO benchmark suites.

Overall, evaluation on marine debris segmentation remains limited compared to other EO tasks. Existing studies have primarily introduced datasets, established baseline methods, or proposed individual architectural improvements, while comprehensive comparisons and evaluations across diverse ML approaches remain scarce.

3 Methodology

To provide a comprehensive evaluation of marine debris segmentation methods, we consider three categories of remote sensing image analysis approaches: classical ML, CNNs, and EOFMs.

3.1 Classical Machine Learning

As a representative classical ML approach, we include the best-performing Random Forest (RF) variant from the original MARIDA benchmark [3], denoted as $RF_{ss+si+glcm}$ in their paper. In addition to the Sentinel-2 spectral bands, the model incorporates spectral indices and Grey-Level Co-occurrence Matrix (GLCM) texture features. The spectral indices enhance discrimination of floating materials by emphasising characteristic spectral responses, while the GLCM features capture local spatial texture not represented by spectral information alone.

3.2 Segmentation using Deep Learning

To establish a deep learning baseline, we reproduce the UNet model proposed in the original MARIDA benchmark [3], which serves as the primary CNN reference for semantic segmentation. We additionally evaluate UNet++, an extension of the original UNet architecture that introduces nested skip connections to improve multi-scale feature fusion and preserve fine-grained spatial information.

Beyond reproducing the published implementation, we train an optimised UNet (UNet-O) configuration through extensive hyperparameter tuning and modifications to the loss function. This optimised model provides a substantially stronger convolutional baseline, enabling a more rigorous comparison with modern foundation models.

3.3 Earth Observation Foundation Models

To evaluate recent advances in remote sensing representation learning, we include four state-of-the-art EOFMs. These build upon the foundation model paradigm established in computer vision, adapting it to EO data. Like Vision Foundation Models (VFMs), they are pretrained on large-scale datasets to learn transferable feature representations across a wide range of downstream tasks. However, unlike general-purpose VFMs, EOFMs are pre-trained on large collections of multispectral satellite imagery and are designed to capture the spectral and spatial characteristics of EO data. Since these models provide only a feature-extraction backbone, a task-specific segmentation head must be trained for downstream semantic segmentation. In our experiments, we employ a UPerNet [12] decoder while keeping the pre-trained backbone frozen.

The selected models represent diverse approaches to EOFM pretraining, differing in their learning objectives, architectural design, and supported sensor modalities. **Copernicus-FM** is pretrained on Sentinel-2 imagery using masked image modeling combined with DINOv2-based knowledge distillation, producing robust feature representations for optical EO tasks [13]. **DEO** employs a DINO-based self-supervised learning framework combined with DINOv3 [14] distillation to learn transferable multispectral representations from Sentinel-2 imagery, with a particular emphasis on dense downstream performance [15]. **TerraFM** is a Vision Transformer designed for multimodal Earth observation, jointly processing Sentinel-1 SAR and Sentinel-2 optical imagery through modality-specific encoders and cross-attention fusion [16]. **CROMA** combines masked image modelling with cross-modal contrastive learning to jointly learn feature representations from aligned optical and radar satellite imagery [17].

4 Experiments

This study addresses floating marine debris detection as a semantic segmentation task using the publicly available MARIDA dataset [3]. MARIDA consists of multispectral Sentinel-2 image patches with sparse pixel-level annotations for marine debris and other marine surface materials, providing a standardised benchmark for evaluating segmentation algorithms.

Marine debris segmentation is particularly challenging because debris occupies only a small fraction of an image, resulting in severe class imbalance. In addition, several marine classes exhibit similar spectral characteristics, making it difficult to distinguish debris from naturally occurring materials such as algae, foam, and sediments. These challenges require segmentation models to effectively exploit both spatial and spectral information.

4.1 Dataset and Evaluation Protocol

The dataset contains Sentinel-2 imagery with 11 spectral bands across 256×256 pixel tiles. Marine water and sediment account for approximately 98% of all pixels, while marine debris represents only 0.41%. We merged four highly correlated or ambiguous classes (mixed water, waves, cloud shadows, and ship wakes) into the marine water class, resulting in 11 final classes. Data was split into the train/val/test set using the official MARIDA split (50/25/25%). Training data was augmented using random horizontal flips and orthogonal rotations.

Model performance was evaluated using mean Intersection over Union (mIoU), mean pixel accuracy (mPA), and macro F1-score.

4.2 Implementation Details and Training Pipeline

The training and evaluation pipeline was implemented in PyTorch, building upon the open-source code released with MARIDA. The baseline UNet [7], as evaluated in the original paper, and UNet++ [11] serve as standard baselines. UNet++ [11] was trained for 134 epochs using the original MARIDA optimization settings (batch size of 5 and initial learning rate of 2×10^{-4}) and feature widths of $[32, 64, 128, 256, 512]$. Our optimised version of the UNet model, denoted as UNet-O, was designed to better address class imbalance and prevent overfitting. The UNet-O architecture uses 64 base feature channels, compared to 16 in the original MARIDA baseline. In addition, the loss function combined Cross-Entropy ($\mathcal{L}_{\text{CE}}$) and multi-class Dice loss ($\mathcal{L}_{\text{MultiDice}}$):

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{MultiDice}},$$

with $\lambda = 0.6$. UNet-O was trained for 134 epochs with a batch size of 8 and an initial learning rate of 1×10^{-4}.

All foundation models used the same training pipeline, with frozen backbones and a UperNet [12] segmentation head. These models were trained for 50 epochs with a batch size of 5, using the Adam optimiser with an initial learning rate of 2×10^{-4} and no weight decay.

To accommodate differences in architecture and pretrained inputs, we applied model-specific adaptations to the input data. DEO excludes Sentinel-2 Band 1 (443 nm). Copernicus-FM requires auxiliary metadata, which was appended to the image inputs without modifying the spectral bands. CROMA and TerraFM require 12 spectral bands; therefore, the missing Band 9 was padded with zeros.

For all models, the learning rate was reduced by a factor of 0.1 after 10 epochs without validation loss improvement using `reduce_lr_on_plateau`. Pixels marked with the ignore index (-1) were excluded from both Cross-Entropy and Dice loss calculations.

5 Results

The performance of eight evaluated methods is compared across three categories: classical machine learning (RF), convolutional neural networks (UNet, UNet++, UNet-O), and Earth Observation Foundation Models (TerraFM, CROMA, Copernicus-FM, DEO). The quantitative results are presented in Table 1.

The optimised convolutional baseline (UNet-O) outperforms all tested pretrained foundation models across all evaluation metrics, achieving a peak mIoU of 0.76 and a macro F1-score of 0.85. It also significantly improves over the original UNet baseline (0.57 mIoU), demonstrating the benefit of targeted architectural modifications, hyperparameter optimisation, and loss function design for this domain-specific task. These results suggest that specialised convolutional architectures remain highly competitive against generalised pretrained representations in challenging aquatic environments. The UNet++ architecture also improves upon the UNet by enhancing feature fusion through architectural modifications, although it does not match the performance of the UNet-O configuration. The RF baseline achieves competitive performance, outperforming the original UNet and several EOFMs, while remaining slightly below the optimised UNet-O and DEO.

Model	mIoU	mPA	F1-macro
Classical machine learning			
RF [10]	0.69	0.79	0.79
Convolutional neural networks			
UNet [7]	0.57	0.69	0.69
UNet++ [11]	0.68	0.79	0.77
UNet-O	**0.76**	**0.86**	**0.85**
Earth Observation Foundation Models			
Terra [16]	0.54	0.67	0.64
CROMA [17]	0.66	0.77	0.75
Cop.-FM [13]	0.66	<u>0.83</u>	0.76
DEO [15]	<u>0.70</u>	0.80	<u>0.79</u>

Table 1: Semantic segmentation performance on the MARIDA dataset. Models are grouped by methodology and sorted from worst to best by mIoU. Best results are marked in **bold**, and second best are <u>underlined</u>.

Figure 2 provides a more detailed view of the per-class performance on six challenging minority classes that are particularly relevant for marine debris monitoring. While all evaluated methods struggle with the rarest categories, especially Marine Debris and Natural Organic Material, the optimised UNet-O generally achieves the strongest and most consistent performance across the selected classes. Among the foundation models, DEO remains the strongest overall performer, whereas TerraFM consistently yields the lowest scores. CROMA and CopernicusFM perform competitively on certain classes, such as Ship and Foam, but do not match the overall performance of the optimised UNet-O across the selected minority classes.

The qualitative evaluation in Figure 1 further illustrates the differences between the evaluated approaches. All foundation models tend to produce overly smooth, blob-like segmentation masks and struggle to recover sharp object boundaries. In contrast, UNet-O generates significantly more detailed and spatially precise predictions, better aligning with the ground truth structure. This difference becomes particularly evident in scenarios where small objects are densely packed or spatially overlapping, such as regions of mixed algae and debris. In such cases, the reduced spatial resolution of foundation model embeddings limits their ability to separate fine-grained classes, whereas convolutional architectures such as UNet retain better localisation capabilities, albeit facing inherent challenges under these conditions. The RF baseline also behaves differently, owing to its per-pixel classification using handcrafted spectral and texture features, yielding comparatively sharper predictions.

5.1 Computational Efficiency and Complexity

In addition to predictive accuracy, operational feasibility depends heavily on computational overhead during training and inference. Table 2 summarizes the parameter

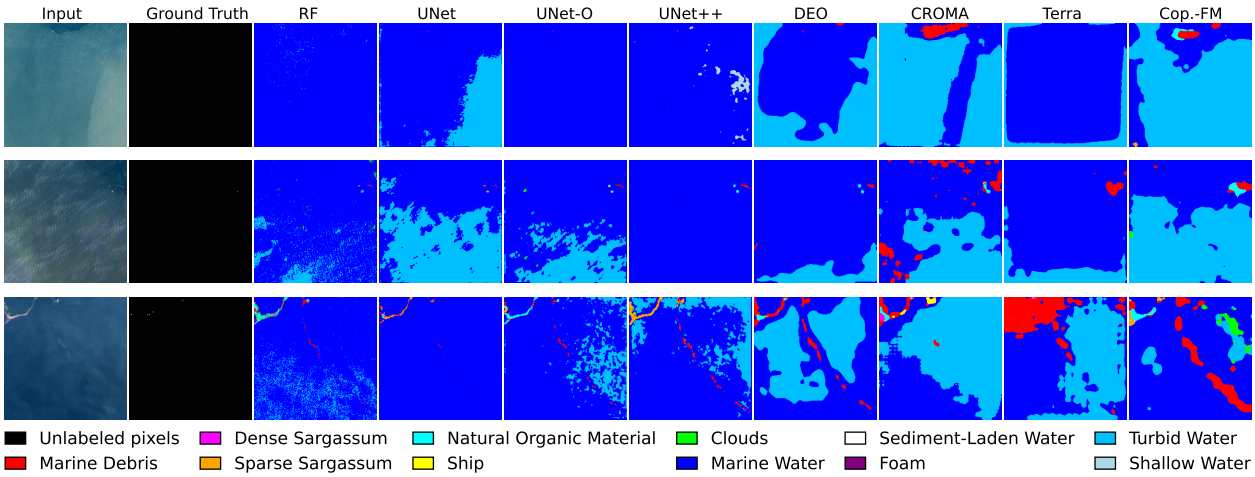


Figure 1: Qualitative results of semantic segmentation. The first row shows a patch where only the foundation models (DEO, CROMA, Terra, Cop.-FM) perform poorly, whereas the non-FM models produce more accurate predictions. The second row shows a challenging patch where all evaluated models fail. The third row shows a patch where the enhanced UNet-O outperforms all other models. Note that the labels are sparse and only a small percentage of pixels are actually labelled.

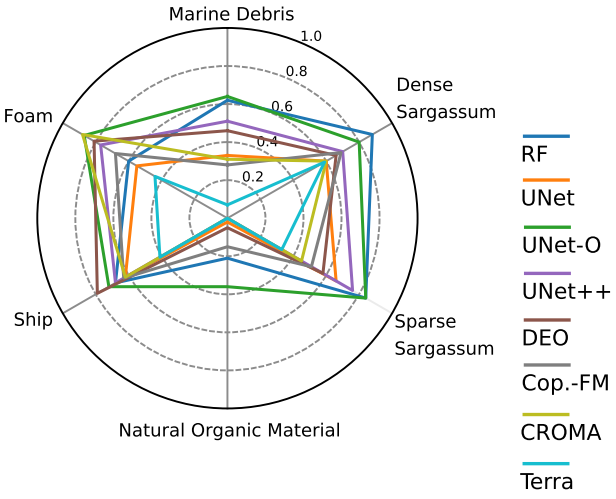


Figure 2: Per-class IoU achieved by the evaluated methods on six challenging minority classes particularly relevant to marine debris monitoring.

footprint, training/fine-tuning times, and GPU inference latencies for all evaluated deep learning models.

While the EOFMs benefit from frozen feature extractors by requiring optimization over only 34.43M to 41.12M trainable parameters during downstream fine-tuning, their total model capacity (122.35M to 180.30M parameters) introduces substantial computational latency during forward passes. Specifically, the inference latencies for the evaluated foundation models range from 23.03 ms (TerraFM) up to 78.61 ms (CROMA). In contrast, UNet-O achieves peak predictive accuracy with a total parameter footprint of 13.40M while maintaining an inference latency of 5.80 ms. Overall, these results demonstrate that domain-optimized convolutional networks offer a more operationally viable balance of predictive precision and computational efficiency for large-scale, real-time marine monitoring.

Table 2: Computational complexity and runtime analysis on the MARIDA dataset. All inference latencies are measured per 256×256 tile on a unified GPU environment. Total training times reflect training up to the best validation checkpoint (Epoch 134 for UNet-O/UNet++, Epoch 50 for EOFMs). RF is omitted as parameter counts and GPU latencies are not directly comparable for non deep-learning baselines.

Model	Total Params (M)	Trainable Params (M)	Training Time (min)	Inference Latency (ms)
Convolutional neural networks				
UNet [7]	0.84	0.84	–	4.77
UNet++ [11]	9.17	9.17	66	8.44
UNet-O	13.40	13.40	44.64	5.80
Earth Observation Foundation Models				
Terra [16]	154.11	41.12	26.51	23.03
CROMA [17]	131.46	41.12	94.8	78.61
Cop.-FM [13]	180.30	41.12	29.49	29.25
DEO [15]	122.35	34.43	26.74	55.04

6 Discussion

The comparison between classical ML and deep learning methods shows that increased model complexity alone does not guarantee improved performance. The Random Forest baseline achieves competitive results, highlighting the effectiveness of handcrafted spectral and texture features for distinguishing floating materials. In contrast, the standard UNet provides only modest improvements, suggesting that generic convolutional architectures are not well suited to small, highly imbalanced remote sensing datasets. By adapting the model capacity, training strategy, and optimisation objective, UNet-O improves mIoU from 0.57 to 0.76, demonstrating that task-specific modifications can be more beneficial than adopting more complex architectures. In particular, combining Cross-Entropy and Dice loss helps mitigate severe class imbalance.

The results also highlight the challenges of evaluating marine debris segmentation. The scarcity of annotated debris pixels and the dominance of water-related classes bias quantitative metrics towards majority classes, allowing

models to achieve competitive scores despite inaccurate boundaries or missed small debris regions. This limitation is particularly evident for EOFMs, whose pretrained representations yield competitive overall performance but often produce smooth, spatially coarse predictions and fail to outperform the optimised CNN baseline.

A possible explanation for this performance gap on the MARIDA dataset is the nature of their pretraining data. Large-scale EO foundation models are predominantly trained on terrestrial imagery, with limited coverage of coastal and open-water environments. Consequently, their learned representations may not fully capture the spectral and spatial characteristics of marine debris. In addition, the spatial downsampling used in many foundation models reduces sensitivity to small objects, a limitation that is only partially reflected in standard evaluation metrics due to the dataset's severe class imbalance.

7 Conclusion

This study presented a comprehensive evaluation of machine learning approaches for marine debris segmentation using the MARIDA dataset [3]. Although Earth Observation Foundation Models have demonstrated strong performance across many remote sensing tasks, our results show that a carefully optimised convolutional network can remain more effective for this challenging aquatic application, particularly for detecting small and spatially complex debris regions. These findings lead us to hypothesise that current foundation models are constrained by terrestrial pretraining biases and limited spatial resolution, highlighting the continued importance of domain-specific optimisation. Future work should focus on developing foundation models for aquatic applications and more representative marine benchmarks to improve automated marine monitoring.

Acknowledgement

Research was in part funded by the ARIS research projects GC-0006 (GeoAI) and J2-60045 (RoDEO), as well as the ARIS research programme P2-0214.

References

[1] K. Topouzelis, D. Papageorgiou, G. Suaria, and S. Aliani, "Floating marine litter detection algorithms and techniques using optical remote sensing data: A review", *Marine Pollution Bulletin*, vol. 170, p. 112 675, 2021. DOI: 10.1016/j.marpolbul.2021.112675.

[2] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources", *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017. DOI: 10.1109/MGRS.2017.2762307.

[3] K. Kikaki, I. Kakogeorgiou, P. Mikeli, D. E. Raitsos, and K. Karantzalos, "Marida: A benchmark for marine debris detection from sentinel-2 remote sensing data", *PLOS ONE*, vol. 17, no. 1, pp. 1–20, 2022. DOI: 10.1371/journal.pone.0262247.

[4] A. Lacoste, N. Lehmann, P. Rodriguez, E. Sherwin, H. Kerner, B. Lütjens, J. Irvin, D. Dao, H. Alemohammad, A. Drouin, M. Gunturkun, G. Huang, D. Vazquez, D. Newman, Y. Bengio, S. Ermon, and X. X. Zhu, "Geo-bench: Toward foundation models for earth monitoring", in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 51 080–51 093.

[5] V. Marsocci, Y. Jia, G. Le Bellier, D. Kerekes, L. Zeng, S. Hafner, S. Gerard, E. Brune, R. Yadav, A. Shibli, H. Fang, Y. Ban, M. Vergauwen, N. Audebert, and A. Nascetti, "Pangaea: A global and inclusive benchmark for geospatial foundation models", *arXiv preprint arXiv:2412.04204*, 2024. arXiv: 2412.04204.

[6] J. Mifdal, N. Longépé, and M. Rußwurm, "Towards detecting floating objects on a global scale with learned spatial features using sentinel-2", *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-3-2021, pp. 285–293, 2021. DOI: 10.5194/isprs-annals-V-3-2021-285-2021.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation", in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2015, pp. 234–241.

[8] P. R. Priyadarshini, V. Arya, and S. S. Kamath, "Automated marine debris detection from sentinel-2 satellite imagery", in *Proceedings of the 2024 IEEE Space, Aerospace and Defence Conference (SPACE)*, 2024, pp. 454–458. DOI: 10.1109/SPACE63117.2024.10668301.

[9] M. Rußwurm, S. J. Venkatesa, and D. Tuia, "Large-scale detection of marine debris in coastal areas with sentinel-2", *iScience*, vol. 26, no. 12, p. 108 402, 2023. DOI: 10.1016/j.isci.2023.108402.

[10] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[11] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation", in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madab-

hushi, Eds., Cham: Springer International Publishing, 2018, pp. 3–11, ISBN: 978-3-030-00889-5.

[12] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding", in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018.

[13] Y. Wang, Z. Xiong, C. Liu, A. J. Stewart, T. Dujardin, N. I. Bountos, A. Zavras, F. Gerken, I. Papoutsis, L. Leal-Taixé, and X. X. Zhu, "Towards a unified copernicus foundation model for earth vision", *arXiv preprint arXiv:2503.11849*, 2025. arXiv: 2503.11849.

[14] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, *Dinov3*, 2025. arXiv: 2508.10104 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2508.10104.

[15] F. Wolf, B. Rolih, and L. Čehovin Zajc, "Brewing stronger features: Dual-teacher distillation for multispectral earth observation", *arXiv preprint arXiv:2602.19863*, 2026. arXiv: 2602.19863.

[16] M. S. Danish, M. A. Munir, S. R. A. Shah, M. H. Khan, R. M. Anwer, J. Laaksonen, F. S. Khan, and S. Khan, "Terrafm: A scalable foundation model for unified multisensor earth observation", *arXiv preprint arXiv:2506.06281*, 2025. arXiv: 2506.06281.

[17] A. Fuller, K. Millard, and J. R. Green, "Croma: Remote sensing representations with contrastive radar-optical masked autoencoders", in *Advances in Neural Information Processing Systems*, vol. 36, 2023.