OSDMAMBA: ENHANCING OIL SPILL DETECTION FROM REMOTE SENSING IMAGES USING SELECTIVE STATE SPACE MODEL

Shuaiyu Chen[1]*, **Fu Wang**[1], **Peng Ren**[2], **Chunbo Luo**[1] **& Zeyu Fu**[1]
[1] Department of Computer Science, University of Exeter
[2] College of Oceanography and Space Informatics, China, University of Petroleum (East China)

ABSTRACT

Oil Spill Detection (OSD) in remote sensing images is commonly viewed as a semantic segmentation task. However, due to the small proportion of oil spill area within images, existing OSD datasets tend to be class-unbalanced, posing significant challenges for convolutional neural network (CNN)-based segmentation models with limited receptive fields. In this study, we introduce OSDMamba, which is the first Mamba-based architecture specifically designed for oil spill detection. Compared to CNN, OSDMamba leverages Mamba's selective scanning mechanism to effectively expand the model's receptive field while preserving critical details. To further enhance performance, we propose an asymmetric decoder that integrates state-space modelling and deep supervision, improving multi-scale feature fusion and increasing sensitivity to minority-class samples. In our experiments, the proposed OSDMamba achieves state-of-the-art performance, outperforming CNN-based models with improvements of 8.9% and 11.8% in terms of mIoU across two OSD datasets, demonstrating its effectiveness. Our code will be available to the public at `https://github.com/Chenshuaiyu1120/Oil-Spill-detection`.

1 INTRODUCTION

Marine oil spills, caused by drilling blowouts Bentz (1976), pipeline leaks Doerffer (2013), and tanker spills Al-Ruzouq et al. (2020), threaten marine ecosystems. Effective detection is crucial for mitigation Solberg et al. (2007). Advances in synthetic aperture radar (SAR) and sensors provide vast earth observation data, enabling deep learning-based solutions Satyanarayana & Dhali (2023). Convolutional Neural Networks (CNN) based approaches Ronneberger et al. (2015b) have shown

*This work was supported by the China Scholarship Council and University of Exeter PhD Scholarships.

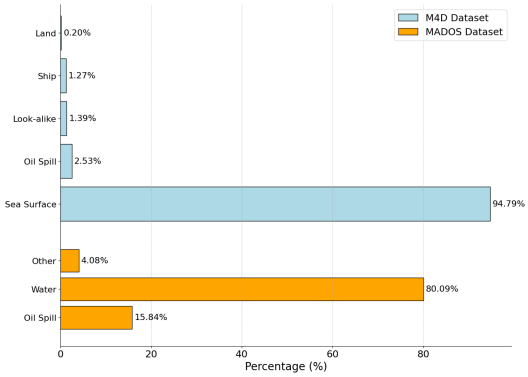

Figure 1: The distribution of semantic classes in the M4D Dataset and MADOS Dataset.

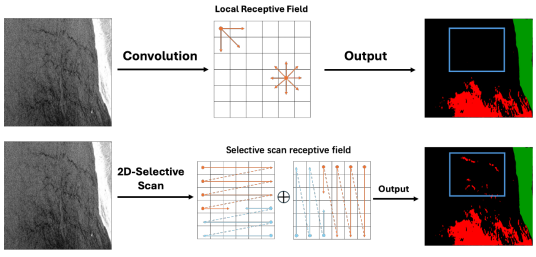

Figure 2: Comparison of receptive fields between convolution and 2D selective scan used in OSDMamba.

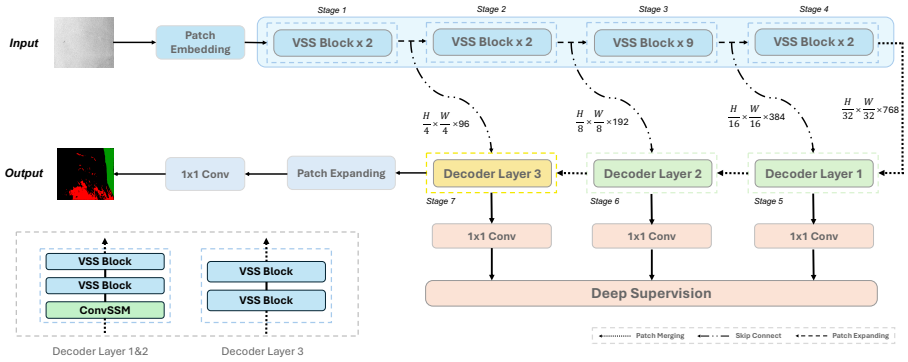

Figure 3: Architectural overview of OSDMamba.

promising performance in oil spill detection, outperforming traditional threshold and machine learning methods Fiscella et al. (2000).

Despite the recent success, there remain two key challenges that affect detection performance: class imbalance and limited receptive fields of convolution. Oil spill incidents are rare, resulting in a limited number of available image samples Brekke & Solberg (2005). In these images, oil spills occupy only a small portion of the scene, leading to a significant class imbalance between the sea surface and oil spill regions, as shown in Figure 1. Additionally, convolution-based models have limited receptive fields. When the receptive field is small, the model may fail to capture global information and the contextual relationships of the target, leading to misclassification or complete omission of small objects, as shown in Figure 2.

Recently, state-space models (SSMs), such as the Mamba models Gu & Dao (2023), have shown promise in addressing various imaging tasks Ruan & Xiang (2024). Mamba excels at capturing long-range dependencies and extracting meaningful features at both local and global scales, leveraging bidirectional scanning and large convolutional kernels for enhanced performance. In this paper, we study the effectiveness of Mamba models to the above challenges of oil spill detection. To achieve this, we propose the first Mamba-based architecture tailored for detecting oil spills, called **OSDMamba**, as shown in Figure 3. By leveraging Mamba's sliding window mechanism for local self-attention, it ensures the dependence of information in each local region, helping the model more accurately learn minority class features in imbalanced datasets Zhu et al. (2024). Additionally, we develop an asymmetric decoder with ConvSSM Smith et al. (2024) and deep supervision Liu et al. (2024) for more effective feature fusion at multiple scales. By combining convolution operations with SSM, the model preserves global state information while extracting local features, enhancing feature representations in complex scenarios. Experimental results show that the proposed OSDMamba achieves state-of-the-art performance, yielding mIoU improvements of 8.9% and 11.8% in oil spill detection across two publicly available datasets.

2 METHODOLOGY

Figure 3 illustrates our proposed OSDMamba. A SAR image is first divided into small patches using a Patch Embedding technique. These patches are transformed into a linear embedding sequence, which is further processed through Visual State Space (VSS) Blocks with a configuration of $\{2, 2, 9, 2\}$ Zhu et al. (2024). Features with downsampling rates of $\{4, 8, 16\}$, generated at the corresponding three stages, are transferred to the decoder for upsampling. We employ an asymmetric decoder structure, where ConvSSM modules are applied in the first two stages to achieve spatial feature fusion.

Encoder We based the OSDMamba encoder on design principles from VMamba-Tiny Zhu et al. (2024) and Swin-UMamba Liu et al. (2024). Similar to Swin-UMamba, OSDMamba performs 2× downsampling at each stage to preserve low-level details. The later stages follow a similar design of VMamba-Tiny, with 2, 2, 9, and 2 VSS blocks in stages 1 to 4, respectively. This configuration balances model depth and efficiency, allowing for robust feature extraction across multiple scales.

Table 1: The quantitative performance comparison and ablation experiments of the proposed method (on M4D Dataset).

Model	SeaSurface(%, ↑)	Oil Spill(%, ↑)	Look-alike(%, ↑)	Ship(%, ↑)	Land(%, ↑)	mIoU(%, ↑)
Unet Ronneberger et al. (2015a)	93.90	53.79	39.55	44.93	92.68	64.97
LinkNet Chaurasia & Culurciello (2017)	94.99	51.53	43.24	40.23	93.97	64.79
PSPNet Zhao et al. (2017)	92.78	40.10	33.79	24.42	86.90	55.60
Deeplabv2 Chen et al. (2017)	94.09	25.57	40.30	11.41	74.99	49.27
Deeplabv2(msc) Chen et al. (2017)	95.39	49.28	31.26	88.65	93.97	62.83
Deeplabv3+ Chen et al. (2018)	96.43	53.38	55.40	27.63	92.44	65.06
YOLOv8-SAM Wu et al. (2024)	94.34	41.84	48.15	52.48	87.65	64.89
SAM-OIL Wu et al. (2024)	96.05	51.60	**55.60**	**52.55**	91.81	69.52
OSDMamba without Decoder	95.60	64.76	46.30	50.32	92.07	67.60
OSDMamba without Deep Supervision	95.50	65.97	53.61	43.20	91.93	69.01
OSDMamba (ours)	**96.47**	**65.59**	47.57	46.85	**94.76**	**70.25**

Additionally, both the VSS blocks and patch merging layers can be initialized with ImageNet Deng et al. (2009) pre-trained weights to enhance performance and accelerate convergence.

Convolutional State Space Model ConvSSM combines convolution and state-space modelling to extract spatial features and capture spatiotemporal dependencies, enabling efficient and accurate feature extraction and fusion in complex scenarios. Below is a detailed derivation of the method: Consider a continuous tensor-valued input $\mathbf{U}(t) \in \mathbb{R}^{H' \times W' \times U}$ with height H', width W', and the number of input features U. We define a continuous-time, linear convolutional state space model (ConvSSM) with state $\mathbf{X}(t) \in \mathbb{R}^{H \times W \times P}$, derivative $\mathbf{X}'(t) \in \mathbb{R}^{H \times W \times P}$, and output $\mathbf{Y}(t) \in \mathbb{R}^{H \times W \times U}$, using the differential equations:

$$\mathbf{X}'(t) = \mathbf{A} * \mathbf{X}(t) + \mathbf{B} * \mathbf{U}(t) \tag{1}$$

$$\mathbf{Y}(t) = \mathbf{C} * \mathbf{X}(t) + \mathbf{D} * \mathbf{U}(t) \tag{2}$$

where $*$ denotes the convolution operator, $A \in R^{P \times P \times k_A \times k_A}$ is the state kernel, $B \in R^{P \times U \times k_B \times k_B}$ is the input kernel, $C \in R^{U \times P \times k_C \times k_C}$ is the output kernel, and $D \in R^{U \times U \times k_D \times k_D}$ is the feedthrough kernel. For simplicity, we pad the convolution to ensure the same spatial resolution, $H \times W$, is maintained in the states and outputs. The definition of the discrete-time convolutional state space model is similar to the above steps.

Decoder Although existing decoders Liu et al. (2021) can effectively retain both fine details and global contextual information, additional strategies are necessary to further enhance model performance, particularly when addressing highly imbalanced class distributions. To tackle this challenge, we implemented a locally asymmetric decoder structure and designed two distinct types of upsampling blocks tailored for different stages of the decoding process. In the context of marine oil spill detection tasks, early-stage upsampling blocks play a critical role in preserving substantial local information Fiscella et al. (2000). Accurately capturing and analysing this local information is essential for detecting small targets and subtle features within SAR images.

To optimize this, we integrated the VSS Block and ConvSSM in stages 1 and 2 of the decoder, as shown in Fig. 3. The VSS Block ensures the preservation of spatial boundaries and minority class features, while ConvSSM's self-attention mechanism enhances global context awareness. Simultaneously, its convolutional layers focus on local feature extraction, improving the fusion of both global and local features for enhanced detection outcomes. For the third and fourth decoder layers, we adopted a more streamlined approach by employing patch expansion and incorporating two VSS blocks as the upsampling components. Such a combination of locally asymmetric design and tailored upsampling strategies can enhance the model's ability to detect oil spills, particularly in scenarios with imbalanced datasets and small target regions. To enhance the model's multi-level feature representation capability, we employed a deep supervision module Liu et al. (2024) by applying 1×1 convolution mapping to the decoder layers with sizes of 1/4, 1/8, and 1/16.

3 EXPERIMENTS

We use the M4D dataset Krestenitis et al. (2019) and the MADOS dataset Kikaki et al. (2024) for evaluation. The proposed OSDMamba model was implemented using PyTorch on an NVIDIA

Table 2: Quantitative comparison of OSDMamba on the MADOS dataset.

Model	F1(%, ↑)	mIoU(%, ↑)	OA(%, ↑)
RF Breiman (2001)	56.6	43.9	67.1
RF++ Karpievitch et al.	64.4	52.4	83.8
U-Net Ronneberger et al. (2015a)	63.8	51.0	82.9
SegNext Guo et al. (2022)	60.6	49.2	**86.6**
MariNeXt (reproduced)Kikaki et al. (2024)	70.6	59.2	81.6
OSDMamba (ours)	**71.2**	**68.1**	82.3

Table 3: Comparison of false positives between the baseline and OSDMamba on the MADOS dataset.

Model	Oil Spill FP(%, ↓)	Overall FP(%, ↓)
Baseline (U-Net)	37.2	43.5
OSDMamba	**32.4**	**35.8**

L40 GPU, following the training strategy outlined in Satyanarayana & Dhali (2023). We trained the OSDMamba model using a hybrid loss function, as given by $L = -\alpha_t(1-p_t)^\gamma \log(p_t) + (1 - \frac{|A \cap B|}{|A \cup B|})$ where α_t is the class weight, used to balance class imbalance; P_t is the predicted probability of the true class by the model; γ is a modulation factor used to adjust the weight of hard and easy samples; A is the predicted positive region; B is the ground truth positive region. We utilized the AdamW optimizer Loshchilov (2017) with an initial learning rate of 0.01 and a weight decay of 0.0001. The model was initialized with pre-trained weights from ImageNet Deng et al. (2009). A batch size of 4 was used, and the model was trained for 100 epochs across all stages.

Table 1 presents a quantitative comparison of OSDMamba against other methods on the M4D dataset. Overall, OSDMamba achieves a mIoU of 70.25%, outperforming all other models in the comparison. This result highlights OSDMamba's superior ability to detect marine oil spills with enhanced generalization performance. Notably, OSDMamba delivers the best performance in the oil spill category, achieving a 12.18% improvement over the second-best model, U-Net. This significant gain indicates OSDMamba's capability to effectively learn from underrepresented classes in imbalanced datasets. Furthermore, OSDMamba also achieves top performance in detecting Sea Surface and Land, demonstrating its robustness across multiple categories. Table 2 provides detailed results on the MADOS dataset. Here, OSDMamba again surpasses all competing methods, with an improvement of 0.6% in F1-score and 8.9% in mIoU. These consistent improvements across different datasets confirm OSDMamba's ability to deliver reliable and accurate detection performance in various scenarios.

We also conducted an ablation study to evaluate the impact of each component of OSDMamba on overall performance, as shown in the last three rows of Table 1. OSDMamba without our decoder achieves a mIoU of 67.62%, demonstrating that the lack of multi-scale feature fusion negatively impacts performance, particularly in key categories like Oil Spill and Ship. OSDMamba without deep supervision reaches a mIoU of 65.85%, further highlighting the role of deep supervision in guiding feature learning and improving accuracy, as shown by the reduced performance in categories such as Oil Spill (64.79%) and Ship (43.00%). In contrast, OSDMamba (ours) achieves the best performance, with a mIoU of 70.25%, demonstrating superior segmentation across key categories like Oil Spill (65.59%) and Ship (46.85%). The corresponding qualitative analysis is provided in the appendix.

To highlight how our method addresses the two previously mentioned challenges, we present the false positive (FP) statistics both within the oil spill areas and overall, in Table 3. Compared to the baseline method, our approach significantly reduces the number of false positives (FP) within oil spill areas, i.e., the minority class. This reduction indicates that our method effectively mitigates the bias associated with class imbalance. Similarly, the overall FPs are significantly reduced by our approach, demonstrating its strength in minimizing misclassifications at both local and global scales and thereby ensuring more accurate and reliable detection outcomes.

4 Conclusion

In this paper, we presented a Mamba-based image segmentation framework for marine oil spill detection. By stacking VSS Blocks, the model is better able to distinguish limited samples. Additionally, we designed an asymmetric decoder and integrated ConvSSM into specific decoder layers to enhance feature fusion. Extensive experiments demonstrate that our model achieves state-of-the-art segmentation performance on two oil spill detection datasets. We believe that our design using the Mamba architecture offers a new perspective for marine oil spill detection tasks.

REFERENCES

R. Al-Ruzouq, M. B. A. Gibril, A. Shanableh, and et al. Sensors, features, and machine learning for oil spill detection and monitoring: A review. *Remote Sensing*, 12(20):3338, 2020.

A. P. Bentz. Oil spill identification. *Analytical Chemistry*, 48(6):454A–472A, 1976.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Camilla Brekke and Anne HS Solberg. Oil spill detection by satellite remote sensing. *Remote sensing of environment*, 95(1):1–13, 2005.

Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, pp. 1–4. IEEE, 2017.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, June 2009.

J. W. Doerffer. *Oil spill response in the marine environment*. Elsevier, 2013.

B Fiscella, A Giancaspro, F Nirchio, P Pavese, and Paolo Trivero. Oil spill detection using marine sar images. *International Journal of Remote Sensing*, 21(18):3561–3566, 2000.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022.

Yuliya V Karpievitch, Elizabeth G Hill, Anthony P Leclerc, Alan R Dabney, and Jonas S Almeida. Rf++: Generalized random forest-based classifier for cluster-correlated data.

K. Kikaki, I. Kakogeorgiou, I. Hoteit, and et al. Detecting marine pollutants and sea surface features with deep learning in sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 210:39–54, 2024.

Marios Krestenitis, Georgios Orfanidis, Konstantinos Ioannidis, Konstantinos Avgerinakis, Stefanos Vrochidis, and Ioannis Kompatsiaris. Oil spill identification from satellite images using deep neural networks. *Remote Sensing*, 11(15):1762, 2019.

Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. *arXiv preprint arXiv:2402.03302*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ilya Loshchilov. Decoupled weight decay regularization. *arXiv preprint*, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015a.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015b.

Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

Abhishek Ramanathapura Satyanarayana and Maruf A Dhali. Oil spill segmentation using deep encoder-decoder models. *arXiv preprint arXiv:2305.01386*, 2023.

Jimmy Smith, Shalini De Mello, Jan Kautz, Scott Linderman, and Wonmin Byeon. Convolutional state space models for long-range spatiotemporal modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

Anne HS Solberg, Camilla Brekke, and Per Ove Husoy. Oil spill detection in radarsat and envisat sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):746–755, 2007.

W. Wu, M. Sing Wong, X. Yu, G. Shi, C. Y. T. Kwok, and K. Zou. Compositional oil spill detection based on object detector and adapted segment anything model from sar images. *IEEE Geoscience and Remote Sensing Letters*, 2024.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

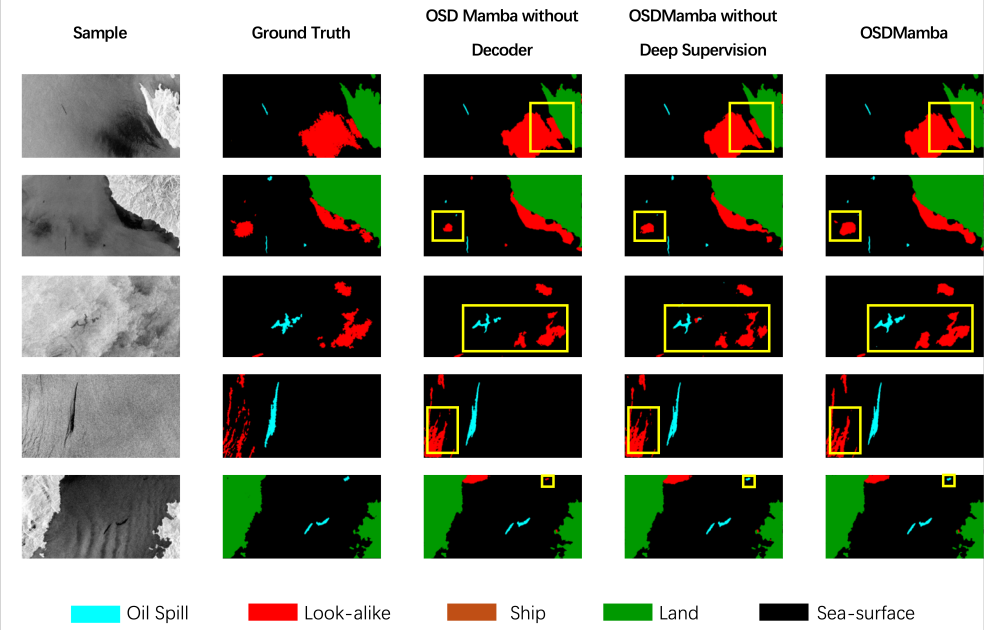

Figure 4: Qualitative analysis of ablation experiments

A APPENDIX

A.1 DATASETS

The Oil Spill Detection Dataset Krestenitis et al. (2019) was used in this study, including 1002 training SAR images and 110 testing images. The dataset consists of five semantic categories: sea surface, oil spill, oil spill look-alike, ship, and land. The images in the dataset have dimensions of 1250×650 pixels.

The MADOS (Marine Debris and Oil Spill) dataset Kikaki et al. (2024) is a globally distributed benchmark designed for detecting marine pollution, including oil spills and debris. It consists of 174 high-resolution multispectral Sentinel-2 satellite images collected between 2015 and 2022, with approximately 1.5 million labelled pixels spanning 15 thematic categories.

A.2 QUALITATIVE ANALYSIS

As shown in Figure 4, the qualitative analysis visualises the contributions of each module through a column-wise comparison. Using the CNN decoder, the model struggles with segmenting small or ambiguous regions, such as isolated oil spills and look-alike areas (e.g., third columns). Our decoder improves the segmentation of minority classes, particularly for oil spills and ships (e.g., 5th column), but some regions remain under-segmented or misclassified, leading to a moderate mIoU of 69.01%. The complete OSDMamba achieves the best segmentation, accurately capturing fine details and distinguishing small-scale targets, as depicted in the 5th column of Fig. 4. These results demonstrate the decoder's effectiveness in multi-scale feature fusion and the critical role in guiding the model to better handle class imbalance and small-scale features by introducing multi-level supervision signals in intermediate layers, enhancing the learning of small-scale features.