Journal of
Marine Science and Engineering

MDPI

Article

Detection of Marine Oil Spill from PlanetScope Images Using CNN and Transformer Models

Jonggu Kang [1], Chansu Yang [2], Jonghyuk Yi [3] and Yangwon Lee [1,*]

[1] Major of Geomatics Engineering, Division of Earth Environmental System Sciences, Pukyong National University, Busan 48513, Republic of Korea; gu7529@pukyong.ac.kr
[2] Sea Power Reinforcement & Security Research Department, Korea Institute of Ocean Science and Technology, Busan 49111, Republic of Korea; yangcs@kiost.ac.kr
[3] SE Lab Incorporation, Seoul 06049, Republic of Korea; yi@selab.co.kr
* Correspondence: modconfi@pknu.ac.kr

Abstract: The contamination of marine ecosystems by oil spills poses a significant threat to the marine environment, necessitating the prompt and effective implementation of measures to mitigate the associated damage. Satellites offer a spatial and temporal advantage over aircraft and unmanned aerial vehicles (UAVs) in oil spill detection due to their wide-area monitoring capabilities. While oil spill detection has traditionally relied on synthetic aperture radar (SAR) images, the combined use of optical satellite sensors alongside SAR can significantly enhance monitoring capabilities, providing improved spatial and temporal coverage. The advent of deep learning methodologies, particularly convolutional neural networks (CNNs) and Transformer models, has generated considerable interest in their potential for oil spill detection. In this study, we conducted a comprehensive and objective comparison to evaluate the suitability of CNN and Transformer models for marine oil spill detection. High-resolution optical satellite images were used to optimize DeepLabV3+, a widely utilized CNN model; Swin-UPerNet, a representative Transformer model; and Mask2Former, which employs a Transformer-based architecture for both encoding and decoding. The results of cross-validation demonstrate a mean Intersection over Union (mIoU) of 0.740, 0.840 and 0.804 for all the models, respectively, indicating their potential for detecting oil spills in the ocean. Additionally, we performed a histogram analysis on the predicted oil spill pixels, which allowed us to classify the types of oil. These findings highlight the considerable promise of the Swin Transformer models for oil spill detection in the context of future marine disaster monitoring.

Keywords: oil spill detection; optical satellite image; deep learning; CNN; transformer

check for updates

Citation: Kang, J.; Yang, C.; Yi, J.; Lee, Y. Detection of Marine Oil Spill from PlanetScope Images Using CNN and Transformer Models. *J. Mar. Sci. Eng.* **2024**, *12*, 2095. https://doi.org/10.3390/jmse12112095

Academic Editors: Simona Verde, Virginia Zamparelli and Pietro Mastro

Received: 26 September 2024
Revised: 13 November 2024
Accepted: 16 November 2024
Published: 19 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Marine oil spills have been shown to cause significant environmental degradation, impacting ecosystems, fisheries, and the tourism industry. These effects, in turn, can lead to a wide range of economic, biological, and social consequences. Hence, it is crucial to detect marine oil spills rapidly and disseminate up-to-date information to minimize the damage. Given the challenges of accessing marine accidents and disasters from land, as well as the potential for these incidents to cover extensive areas, the use of remote sensing technologies—including satellites, aircraft, and unmanned aerial vehicles (UAVs)—offers distinct advantages. Satellite technology is particularly effective, on par with aircraft and UAVs, in detecting marine oil spills due to its capacity for rapid monitoring of vast maritime areas [1]. Several organizations, including the United States Coast Guard, the National Oceanic and Atmospheric Administration (NOAA), and the European Maritime Safety Agency (EMSA), utilize satellite-based oil spill detection. Synthetic aperture radar (SAR) images have traditionally been the primary tool for marine oil spill detection due to their ability to operate in all weather conditions. However, distinguishing oil spills in SAR images remains a significant challenge when ocean winds, internal waves, and biogenic

films are present [2,3]. Recently, deep learning technologies for image recognition from satellite images have been actively studied, offering high performance and accuracy in detecting oil spills.

Prior research on the satellite-based detection of marine oil spills has predominantly employed synthetic aperture radar (SAR) images. To date, studies on oil spill detection using optical satellite images are scarce. When using optical sensors, the satellite images are typically of medium and low resolution. Zhao et al. (2014) used Moderate Resolution Imaging Spectroradiometer (MODIS), Medium Resolution Imaging Spectrometer (MERIS), and Landsat 7/8 images to investigate the feasibility of detecting oil spills [4]. Mityagina et al. (2016) aimed to enhance the precision of oil spill detection by integrating SAR and optical images, including Envisat Advanced Synthetic Aperture Radar (ASAR), Sentinel-1 SAR, and Landsat 5/7/8 [5]. In a related study, Kolokoussis et al. (2018) applied the Object-Based Image Analysis (OBIA) method using Sentinel-2 optical images to examine an object-based oil spill analysis [6]. Arslan and Niyazi (2018) utilized Sentinel-1 SAR and Landsat 8 images [7], while Rajendran et al. (2021) employed a combination of Sentinel-1 and Sentinel-2 images, for oil spill detection [8]. Additionally, unmanned aerial vehicles (UAVs) have been increasingly used for oil spill detection in recent years due to their ability to capture high-resolution images. The studies by Mityagina et al. (2016) and Park et al. (2020) demonstrate how UAV imagery facilitates the differentiation between oil and similar materials [5,9]. Aznar et al. (2014) employed a UAV constellation for oil spill detection modeling [10], and Odonkor et al. (2019) conducted research into mapping coastal oil spills using a UAV team [11].

The advent of deep learning technology has been propelled by the accumulation of datasets and concurrent advancements in computing capabilities. Topouzelis et al. (2007) and Singha et al. (2013) applied deep learning techniques to differentiate between dark spots and oil spills [12,13]. Krestenitis et al. (2019) employed a variety of deep learning models, including U-Net, LinkNet, the Pyramid Scene Parsing Network (PSPNet), and three variants of the DeepLab model, to evaluate the efficacy and efficiency of oil spill detection [14]. Yekeen et al. (2020) developed a novel oil spill detection technique based on the Mask Regional Convolutional Neural Network (Mask R-CNN), an instance segmentation model [15]. Jiao et al. (2019) also utilized deep learning for the detection of oil spills from UAV images [16]. Additionally, Lalitha et al. (2024) proposed an AI-based system for precise oil spill detection, while Vekariya et al. (2024) conducted a survey on the You Only Look Once (YOLO) model for SAR image analysis [17,18]. Liao et al. (2023) examined polarimetric SAR satellite images using a DeepLabV3+-based model to monitor oil spill risks in coastal areas [19]. Ding et al. (2023) introduced Sw-YoloX, which enhances the detection performance for sea surface objects [20]. Kang et al. (2022) demonstrated the effectiveness of the DeepLabV3+ model for detecting oil spills using PlanetScope images [21]. These deep learning technologies are increasingly utilized in oil spill detection through remote sensing, supporting environmental monitoring and disaster response via complex data analysis.

Convolutional neural networks (CNNs) are a deep learning algorithm particularly adept at recognizing patterns in images. Recently, Transformer models have been adopted and have demonstrated superior performance compared to traditional CNNs. However, objective comparisons of CNN and Transformer-based models for oil spill detection using high-resolution optical images have been infrequent. With the increasing number of small cube satellites and the continuous refinement of optical image resolution, it is necessary to implement a collaborative monitoring system that combines satellite SAR and optical sensors to achieve enhanced spatial and temporal coverage for detecting oil spills in high-resolution images. In this context, the objective of this study is to evaluate the prediction performance of existing CNN models and novel Transformer-based models for oil spill detection using high-resolution optical satellite images. PlanetScope images were utilized for deep learning semantic segmentation with DeepLabV3+, a widely utilized CNN model; Swin-UPerNet, a representative Transformer model; and Mask2Former, which employs a

Transformer-based architecture for both encoding and decoding. To ensure an objective evaluation, the three models were subjected to k-fold cross-validation tests. Additionally, the pixels of the predicted oil spills were classified into oil types using histogram analysis.

2. Materials

2.1. PlanetScope Satellite Image

The PlanetScope satellite is a constellation of more than 430 small cube satellites, with a combined mass of 5.8 kg, operated by Planet Labs Inc. (San Francisco, CA, USA). These satellites maintain a sun-synchronous orbit, crossing the equator between 9:30 and 11:30 am. PlanetScope images cover a range of spectral bands, including blue (455–515 nm), green (500–590 nm), red (590–670 nm), and near-infrared (NIR, 780–860 nm) wavelengths. The spatial resolution is 3 m, with a revisit period of one day, and the positional error is less than 10 m (see Table 1). Due to its one-day revisit period, PlanetScope is less susceptible to the effects of inclement weather compared to other optical sensors. Moreover, the detection of minor oil spills is facilitated by the use of high-resolution images. The 3-m resolution of PlanetScope is a significant advantage in detecting marine oil spills. The enhanced spatial and temporal resolution of PlanetScope increases the probability of detecting marine oil spills and enables the acquisition of a greater quantity of data for deep learning modeling.

Table 1. Specification of the PlanetScope satellite [22].

Satellite	PlanetScope (DOVE)
Orbit Altitude	450–580 km
Field of View	3.0° (swath) 1.0° (scene length)
Sensor Type	Four-band frame imager with VIS + NIR filter (Planet Labs Inc., San Francisco, CA, USA)
Spectral bands	Blue: 455–515 nm Green: 500–590 nm Red: 590–670 nm NIR: 780–860 nm
Ground Sample Distance	3.0–4.1 m
Image Capture Capacity	200 million sq km/day
Revisit Time	Daily at nadir
Imagery Bit Depth	12-bit

2.2. Labeling for Semantic Segmentation

A total of 16 PlanetScope images from eight marine oil spill incidents were obtained through a literature review and by referencing relevant articles (Table 2). The Magic Wand Tool in Adobe Photoshop was used to create labels indicating the presence of oil spills. This tool is capable of grouping pixels that exhibit similar spectral characteristics. In Figure 1, row (a) shows the original true color satellite image, where oil spills are not clearly discernible. To improve visibility, brightness was adjusted using gamma correction, and contrast was enhanced through histogram adjustment, making it easier to distinguish oil spills from other pixels. After applying gamma correction and histogram adjustment, as shown in row (b), the visibility of oil spills improved significantly. These enhanced images were then used to create labeled images, with the resulting labels displayed in row (c). Pixels corresponding to oil spills were assigned a value of 1, while background pixels were assigned a value of 0.

Table 2. PlanetScope images used in this study.

Date	Region	File Name	Width	Height
11 August 2017	Persian Gulf	20170811_064051_1001_3B_AnalyticMS	1784	3305
		20170811_064052_1001_3B_AnalyticMS	3139	1986
		20170811_064053_1001_3B_AnalyticMS	1790	2621
22 October 2017	Texas	20171022_162437_102c_3B_AnalyticMS	3270	2164
8 June 2019	California	20190608_172426_1050_3B_AnalyticMS	408	1362
		20190608_172427_1050_3B_AnalyticMS	998	1602
		20190608_182133_101f_3B_AnalyticMS	677	3852
		20190609_172600_0f36_3B_AnalyticMS	2062	4191
		20190622_181926_1040_3B_AnalyticMS	2016	6759
1 August 2020	Venezuela	20200801_151700_73_106a_3B_AnalyticMS	1681	1068
30 November 2020	Venezuela	20201130_144119_1002_3B_AnalyticMS	2345	953
		20201130_151041_42_1069_3B_AnalyticMS	1674	877
6 June 2021	Syria	20210606_074514_75_222f_3B_AnalyticMS	1452	1234
24 August 2021	Syria	20210824_082044_59_2424_3B_AnalyticMS	3613	2279
		20210824_082046_90_2424_3B_AnalyticMS	5042	2611
2 September 2021	Louisiana	20210902_164144_98_2274_3B_AnalyticMS	5246	9632

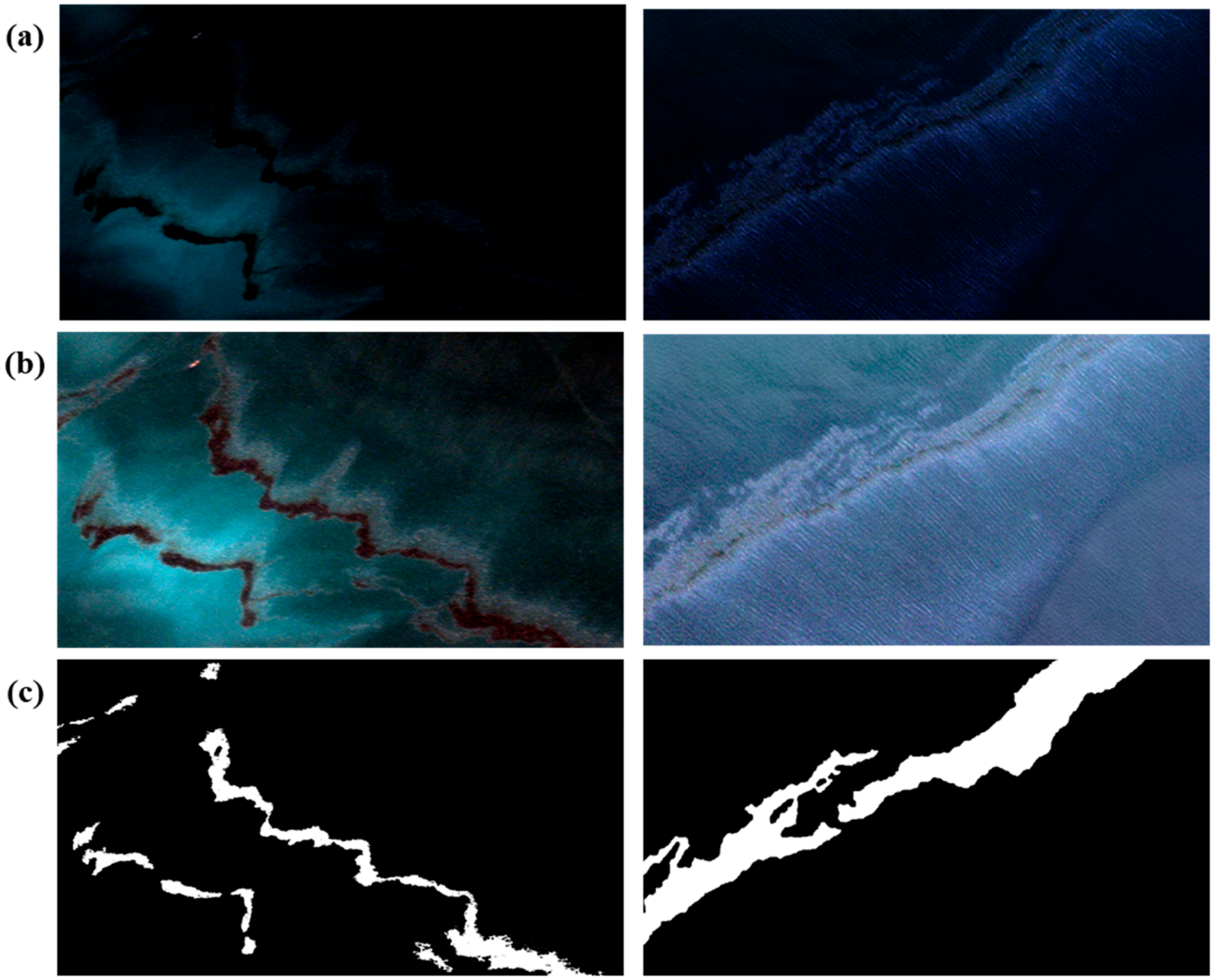

Figure 1. Examples of image processing steps: (**a**) original satellite images, (**b**) images after gamma correction and histogram adjustment, and (**c**) labeled images.

3. Method

In this section, we present the methodology used for detecting oil spills with CNN and Transformer models (Figure 2). The flowchart below provides an overview of the steps involved, such as labeling, modeling, optimization, and evaluation.

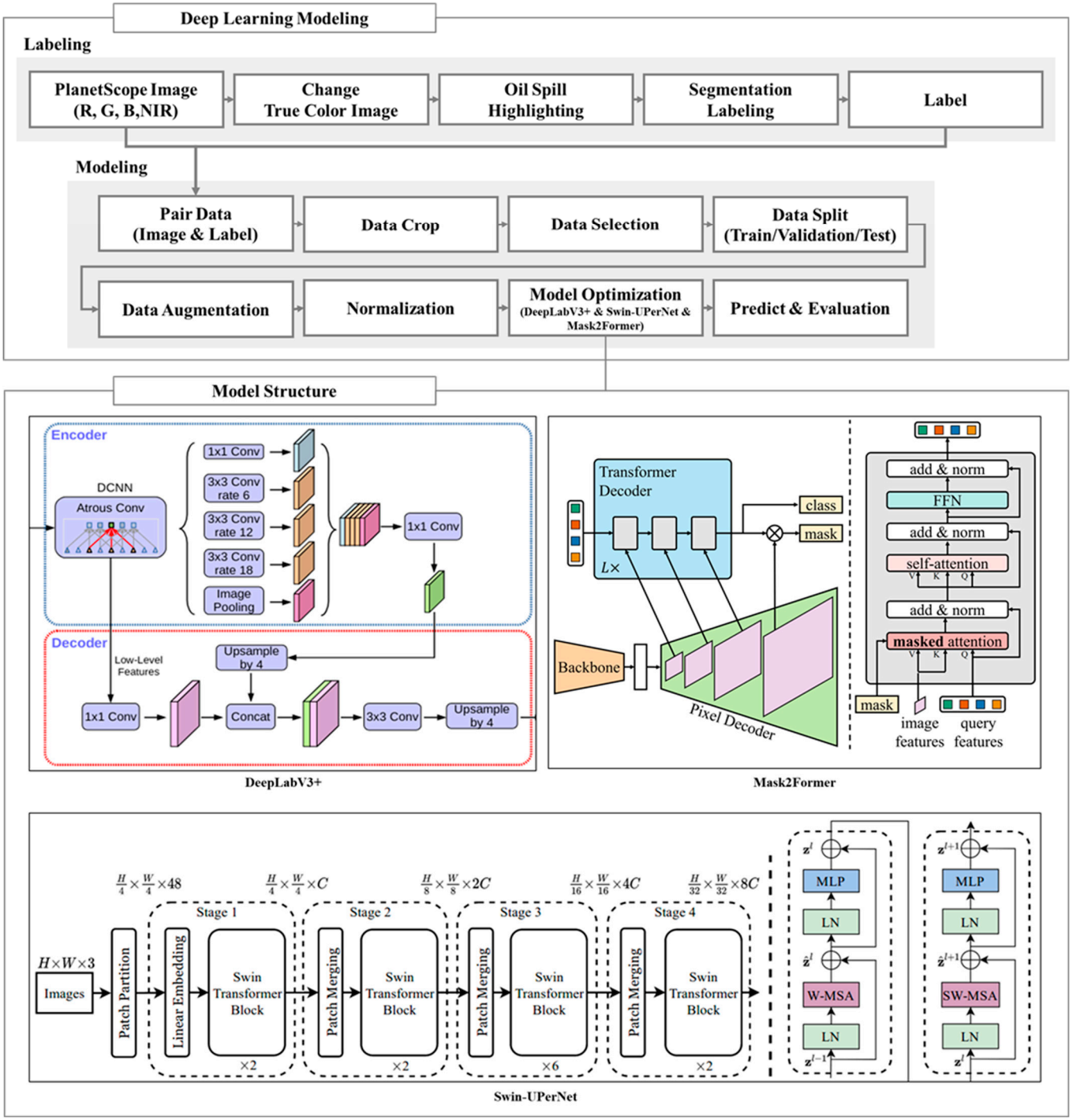

Figure 2. Flowchart of this study, illustrating the processes of labeling, modeling, optimization, and evaluation using the DeepLabV3+, Swin-UPerNet, and Mask2Former models [23–25].

3.1. Deep Learning Semantic Segmentation

The growth of datasets and the evolution of graphics processing units (GPUs) have significantly advanced the field of neural networks. These advancements have enabled neural networks to process large amounts of data, learn complex relationships, and perform

excellent generalizations, thereby making substantial contributions to various computer vision tasks. As Lateef and Ruichek (2019) mentioned, deep learning-based semantic segmentation is a technique that classifies pixels into specific object classes, making it a crucial step in image interpretation and analysis tasks [26]. In this study, we employed DeepLabV3+, Swin-UPerNet, and Mask2Former as representative CNN and Transformer-based models, chosen for their strong performance in satellite remote sensing and their ability to handle complex image features for environmental monitoring.

The DeepLab model utilizes atrous convolution to facilitate more efficient semantic image segmentation. The DeepLabV1 model introduced atrous convolution, while DeepLabV2 implemented the Atrous Spatial Pyramid Pooling (ASPP) technique for multi-scale context, allowing for the parallel processing of multiple atrous convolutions from the feature map and the subsequent merging of the results. In DeepLabV3, atrous convolution was integrated into the existing ResNet architecture to produce a denser feature map. DeepLabV3+ further advanced this approach by incorporating atrous separable convolution, a combination of separable convolution and atrous convolution. Atrous convolution increases the field of view (FOV) by creating additional space within the filter while maintaining a computational cost comparable to that of the original convolution. The receptive field, defined as the area of the input image covered by a single pixel in the feature map, is a crucial factor influencing the performance of semantic image segmentation [23].

A novel approach, known as the Vision Transformer technique, has recently been developed. This technique applies the Transformer model, originally designed for sequence processing, to image recognition, thereby enhancing accuracy. It achieves this by efficiently selecting and focusing on input information through the self-attention mechanism [27]. The self-attention mechanism, which selectively references the most relevant regions and channels from images, was further refined in the Swin Transformer, proposed by Microsoft (Redmond, WA, USA) in 2021. This model demonstrates superior performance compared to conventional CNN and Vision Transformer models. Its effectiveness lies in its ability to locally apply self-attention within a hierarchical shifted windows structure, where input values are processed sequentially and alternately shifted by the window size set in overlapping blocks [24].

Mask2Former, a recent advancement in instance segmentation, builds upon the Transformer model to adapt it effectively for object segmentation in images. This model leverages a masked attention mechanism to focus on specific regions with fine granularity, improving segmentation accuracy. Unlike traditional instance segmentation models, Mask2Former treats segmentation as a mask prediction task, using a unified approach across tasks such as instance, semantic, and panoptic segmentation. By dynamically refining masks through each Transformer layer, Mask2Former achieves higher accuracy and efficiency, surpassing conventional models in handling complex scenes [25].

3.2. Dataset Preparation

The satellite and labeled images were divided into 256 × 256 pixel segments, considering the dimensions of the oil spills and the efficacy of the model learning process. To avoid biased learning toward a specific class, it is important to ensure that the ratio of pixels between classes does not differ significantly. Images containing a significant proportion of land, clouds, or null values were excluded to maintain a relatively consistent ratio between marine oil spill and background classes. As a result, 260 sets of satellite images with blue, green, red, and NIR channels, along with their corresponding labeled images, were prepared.

The test results of a deep learning model trained with a limited dataset can vary depending on how the dataset is partitioned, potentially leading to an inaccurate assessment of the model's performance. To evaluate the model's generalization performance, we conducted five rounds of training and assessment using k-fold cross-validation. The 260 datasets were initially divided into five random samples, which were then used for training, validation, and testing. Each round was configured with three folds for training,

one for validation, and one for testing, resulting in a ratio of 6:2:2 for the training, validation, and test sets, respectively (Figure 3).

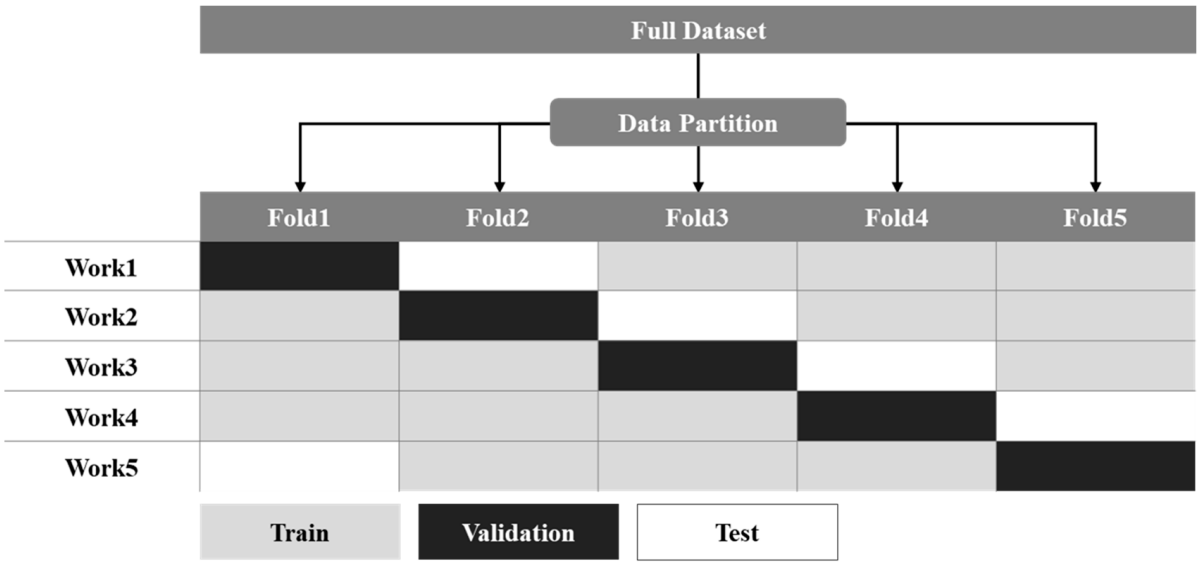

Figure 3. Concept of the 5-fold cross-validation in this study.

Overfitting can occur when the training dataset is insufficient, leading to predictions biased toward specific data subsets. To mitigate this issue, we employed data augmentation techniques to expand the datasets and help the model learn diverse patterns. Data augmentation can prevent overfitting, enhance the generalization performance, and contribute to the development of more robust models. It is crucial to ensure that the original dataset and its augmented versions are not included in the training and test groups simultaneously, as this could lead to data leakage, resulting in an overestimation of the model's performance and distortion of the actual results. Therefore, data augmentation was implemented within each separated fold during the cross-validation process. The open-source data augmentation library Albumentations was used to apply geometrical and spectral transformations to the images, including rotation, flipping, optical distortion, grid distortion, RGB shift, and brightness/contrast adjustments. The original 260 images were augmented tenfold, resulting in 2600 datasets. Of these, 1560 were used for training, 520 for validation, and 520 for testing (Figure 4).

To enhance the stability of the model, the pixel values of the input data were normalized. Instead of using min–max normalization, which can be susceptible to issues arising from extreme values, we utilized z-standardization. This method is more robust in handling outliers by using the mean (μ) and standard deviation (σ) of the data. The mean and standard deviation for each channel were calculated, and all pixel values were normalized using the z-score formula. This normalization process ensures that the input data has a mean of zero and a standard deviation of one, enhancing the model's convergence during training. Each pixel value z is calculated using Formula (1):

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where z represents the z-score, x is the value being standardized, μ is the mean of the dataset, and σ is the standard deviation of the dataset.

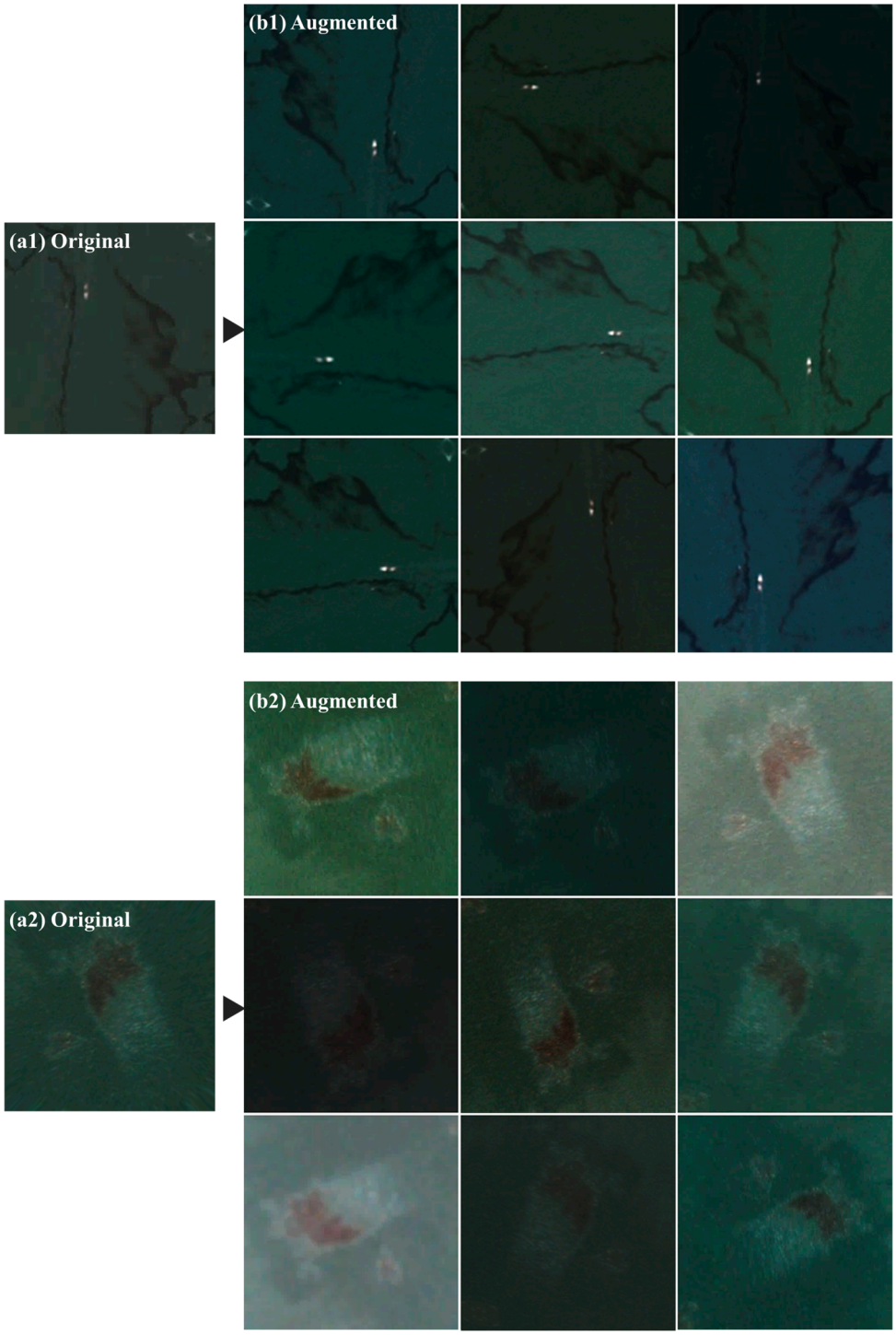

Figure 4. Examples of image data augmentation using the Albumentations library. The example images include random 90-degree rotation, horizontal flip, vertical flip, optical distortion, grid distortion, RGB shift, and random brightness/contrast adjustment.

3.3. Model Optimization

The DeepLabV3+, Swin-UPerNet, and Mask2Former models were trained on the cross-validation datasets, resulting in the creation of five models for each approach. All models were configured with commonly used backbone networks, loss functions, and optimizers. The DeepLabV3+ model employed ResNet-50 and ResNet-101 as backbones, used the Binary Cross Entropy (BCE) loss function, and was optimized with the Adaptive Moment Estimation (Adam) optimizer. In the Swin-UPerNet architecture, the Swin Trans-

former serves as the backbone, available in small (S), base (B), and large (L) versions, and is combined with the UPerNet model, using the BCE loss function and the Adam optimizer with weight decay (AdamW). Similarly, Mask2Former uses the Swin Transformer as its backbone, along with the BCE loss function and the AdamW optimizer. The remaining hyperparameters were initially configured using a grid search technique and then consistently applied across the remaining four folds. The batch size was set to 8 for DeepLabV3+ and Swin-UPerNet and 4 for Mask2Former. The training ran for 150 epochs for DeepLabV3+ and 100 epochs each for Swin-UPerNet and Mask2Former. The learning rates were set at 6×10^{-7} for DeepLabV3+, 6×10^{-5} for Swin-UPerNet, and 1×10^{-5} for Mask2Former. To effectively handle unusual cases, the dropout technique was employed in both models, with a dropout ratio of 0.3. Table 3 illustrates the hyperparameter settings.

Table 3. Hyperparameter settings for oil spill detection using the DeepLabV3+, Swin-UPerNet, and Mask2Former models.

Model	DeepLabV3+	Swin-UPerNet	Mask2Former
Input image size	256×256		
Input channels	4 (Blue, Green, Red, and NIR)		
Backbone	ResNet101	Swin Transformer (Swin-B)	
Model	DeepLabV3+	UPerNet	Mask2Former
Loss function	Binary Cross Entropy		
Optimizer	Adam	AdamW	
Batch size	8	8	4
Epoch	150	100	100
Learning rate	6×10^{-7}	6×10^{-5}	1×10^{-5}
Dropout ratio	0.3		
Output	Probability map		

4. Results

Table 4 presents the evaluation metrics calculated by comparing the predictions and segmentation labels of the test datasets from the five cross-validation folds on the 2600 images. Intersection over Union (IoU) is defined as the ratio of the intersection between the predictions and the labels to their union. The mean Intersection over Union (mIoU) across multiple classes is considered a crucial evaluation metric for image segmentation. The mIoU of the DeepLabV3+ model was 0.740, while the Swin-UPerNet model achieved a mIoU of 0.840, and Mask2Former achieved 0.804.

Table 4. Performance comparison of 5-fold cross-validation results for DeepLabV3+ (DL), Swin-UPerNet (Swin), and Mask2Former (M2F).

		mIoU	IoU (Oil)	IoU (Not Oil)	Accuracy	Precision	Recall	F1_Score	Kappa
Fold 1	DL	0.733	0.602	0.864	0.887	0.857	0.825	0.839	0.679
	Swin	0.836	0.756	0.916	0.933	0.912	0.906	0.909	0.817
	M2F	0.775	0.663	0.887	0.908	0.886	0.855	0.869	0.738
Fold 2	DL	0.723	0.576	0.870	0.889	0.837	0.825	0.830	0.661
	Swin	0.832	0.744	0.919	0.935	0.898	0.914	0.906	0.811
	M2F	0.798	0.693	0.904	0.921	0.876	0.893	0.884	0.768

Table 4. *Cont.*

		mIoU	IoU (Oil)	IoU (Not Oil)	Accuracy	Precision	Recall	F1_Score	Kappa
Fold 3	DL	0.794	0.713	0.876	0.905	0.899	0.871	0.883	0.766
	Swin	0.874	0.827	0.920	0.942	0.934	0.930	0.932	0.864
	M2F	0.811	0.743	0.879	0.910	0.891	0.897	0.894	0.788
Fold 4	DL	0.720	0.568	0.872	0.890	0.814	0.846	0.828	0.657
	Swin	0.811	0.699	0.922	0.934	0.893	0.889	0.891	0.782
	M2F	0.789	0.666	0.912	0.925	0.875	0.878	0.876	0.753
Fold 5	DL	0.732	0.582	0.882	0.899	0.829	0.845	0.837	0.673
	Swin	0.850	0.763	0.937	0.948	0.912	0.921	0.916	0.833
	M2F	0.845	0.754	0.937	0.947	0.910	0.918	0.913	0.827
Avg.	DL	0.740	0.608	0.873	0.894	0.847	0.842	0.843	0.687
	Swin	0.840	0.758	0.923	0.938	0.910	0.912	0.911	0.821
	M2F	0.804	0.704	0.904	0.922	0.888	0.888	0.887	0.775

Precision is defined as the ratio of correctly classified pixels within the predicted image to the total number of pixels in that image, whereas recall is defined as the ratio of correctly classified pixels within the label image to the total number of pixels in that image. The precision was 0.847 for DeepLabV3+, 0.910 for Swin-UPerNet, and 0.888 for Mask2Former, while the recall was 0.842, 0.912, and 0.888, respectively. Generally, low precision indicates overestimation, whereas low recall suggests underestimation. The F1 score, which is the harmonic mean of precision and recall, was 0.843 for DeepLabV3+, 0.911 for Swin-UPerNet, and 0.887 for Mask2Former. The Kappa coefficient, which quantifies the degree of agreement between the predicted and labeled images, was 0.687 for DeepLabV3+, 0.821 for Swin-UPerNet, and 0.775 for Mask2Former.

Figures 5–9 show the input datasets and predicted images for the three models. The first column displays the input images, with pixel values adjusted to a range of 0 to 255 across the red, green, and blue channels. The remaining columns present the segmentation labels alongside predictions from DeepLabV3+, Swin-UPerNet, and Mask2Former. While all models demonstrated proficiency in identifying the location and shape of the oil spill class, Swin-UPerNet and Mask2Former exhibited better accuracy and precision, with fewer instances of missing data compared to DeepLabV3+. Moreover, the Transformer models offered superior edge detail representation, with a more precise delineation of oil spill boundaries compared to DeepLabV3+, which produced coarser edges and missed finer details. This improvement is due to the shifted window-based attention mechanism, which captures both local and global features and adapts to various sizes of irregular objects.

Table 5 presents the results of a comparative analysis of the performance of different model sizes. DeepLabV3+ was trained with ResNet50 as the backbone network, while Swin-UPerNet and Mask2Former were trained in both small and large versions. The difference in performance between the various model sizes was negligible.

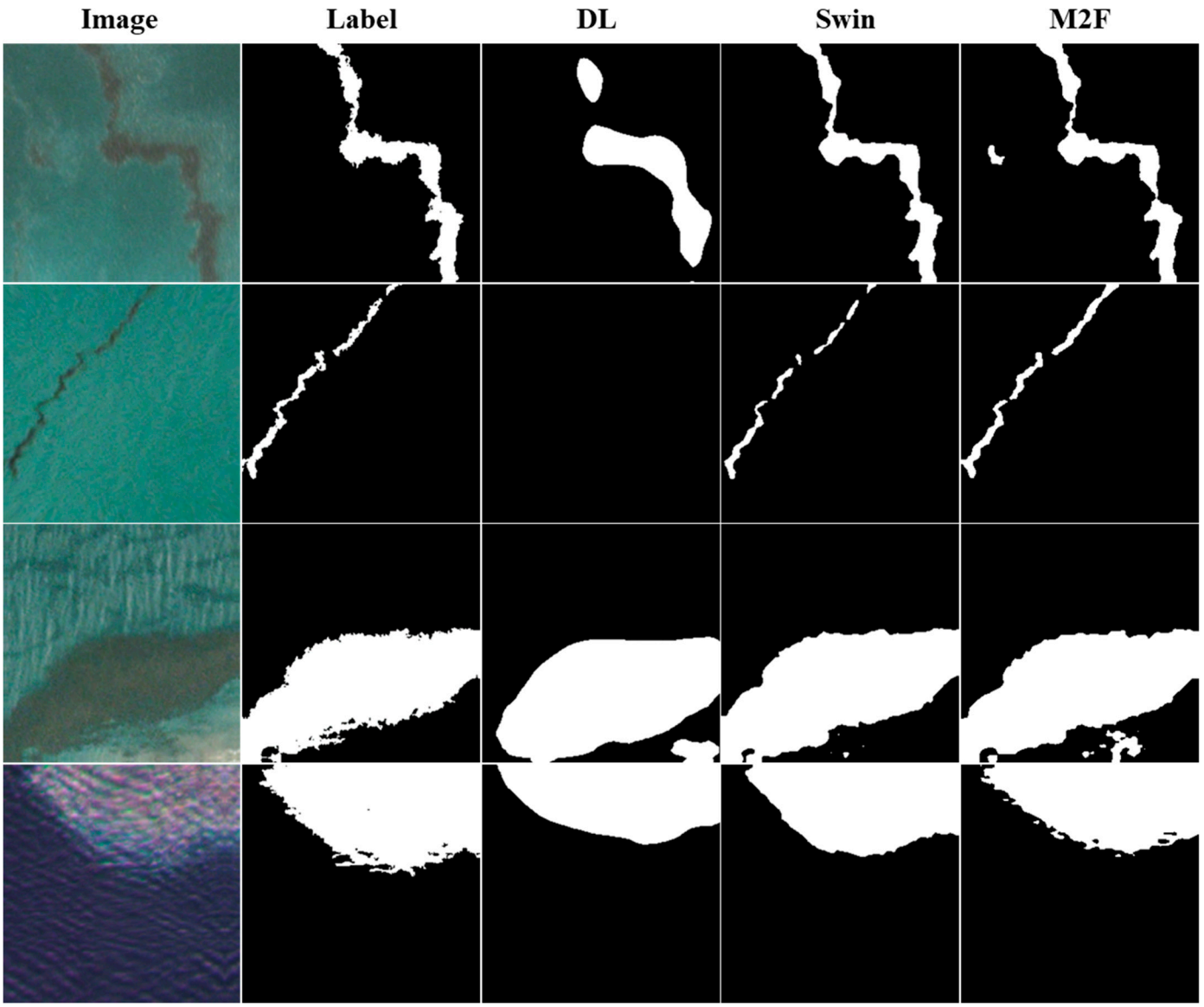

Figure 5. Randomly selected examples from fold 1, including PlanetScope RGB images, segmentation labels, and predictions from DeepLabV3+ (DL), Swin-UPerNet (Swin), and Mask2Former (M2F).

Table 5. Performance comparison for model versions between DeepLabV3+ (DL), Swin-UPerNet (Swin), and Mask2Former (M2F).

		mIoU	IoU (Oil)	IoU (Not Oil)	Accuracy	Precision	Recall	F1_Score	Kappa
DL	r50	0.726	0.587	0.864	0.886	0.863	0.813	0.833	0.668
	r101	0.733	0.602	0.864	0.887	0.857	0.825	0.839	0.679
Swin	SwinS	0.840	0.763	0.918	0.935	0.913	0.910	0.911	0.823
	SwinB	0.836	0.756	0.916	0.933	0.912	0.906	0.909	0.817
	SwinL	0.841	0.763	0.918	0.935	0.914	0.909	0.912	0.823
M2F	SwinS	0.786	0.680	0.893	0.913	0.891	0.864	0.876	0.753
	SwinB	0.775	0.663	0.887	0.908	0.886	0.855	0.869	0.738
	SwinL	0.774	0.661	0.887	0.907	0.886	0.853	0.868	0.736

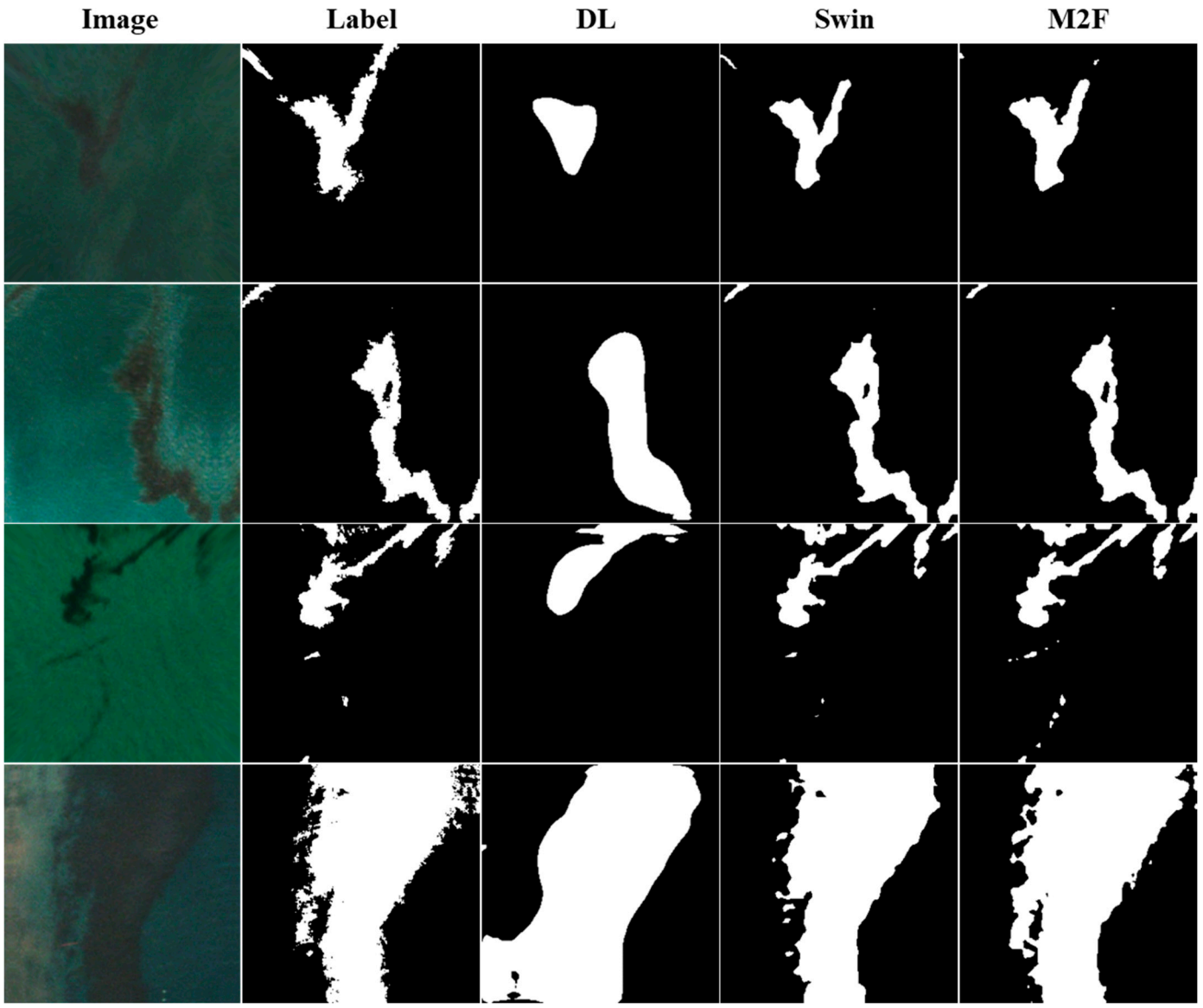

Figure 6. Randomly selected examples from fold 2, including PlanetScope RGB images, segmentation labels, and predictions from DeepLabV3+ (DL), Swin-UPerNet (Swin), and Mask2Former (M2F).

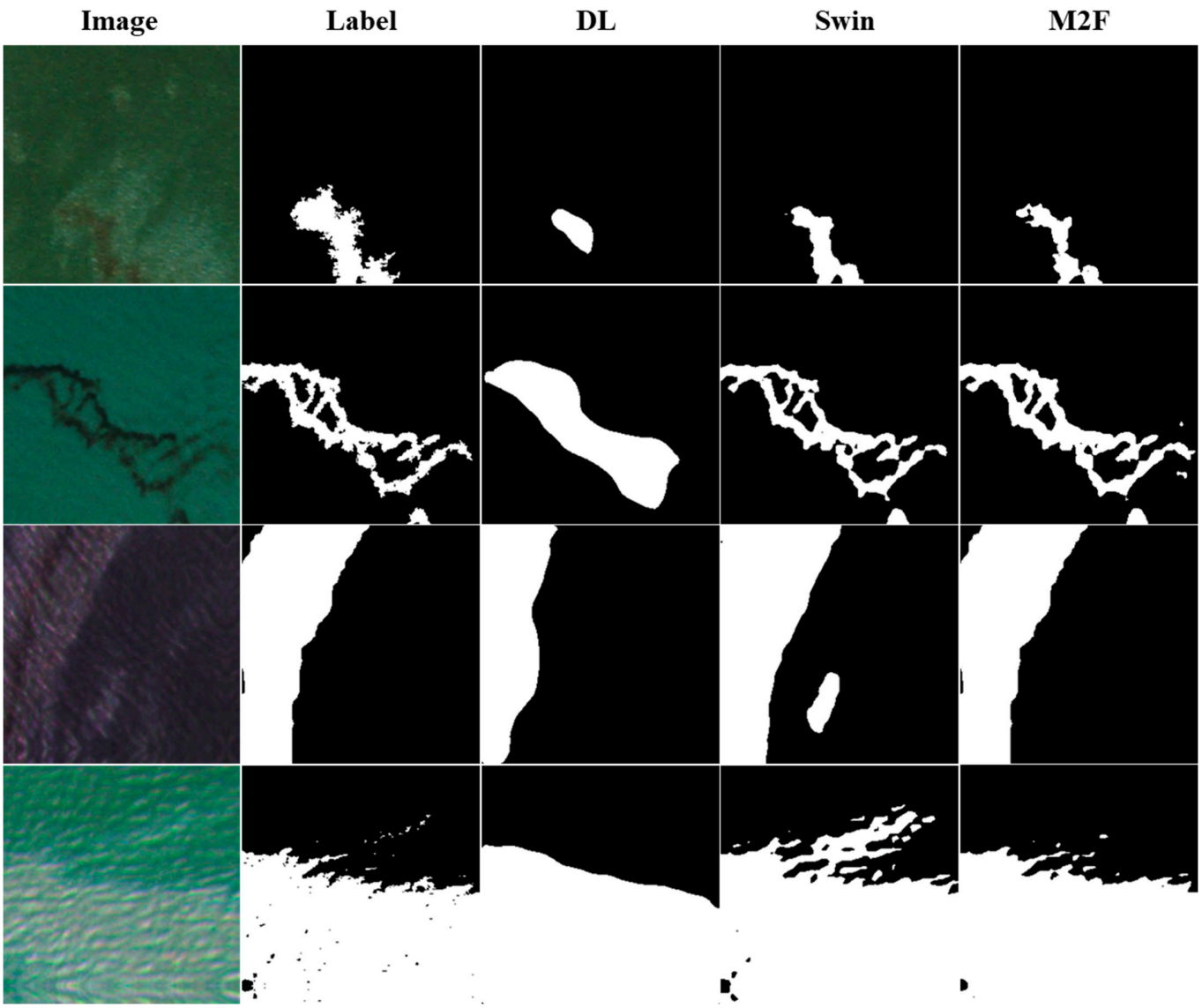

Figure 7. Randomly selected examples from fold 3, including PlanetScope RGB images, segmentation labels, and predictions from DeepLabV3+ (DL), Swin-UPerNet (Swin), and Mask2Former (M2F).

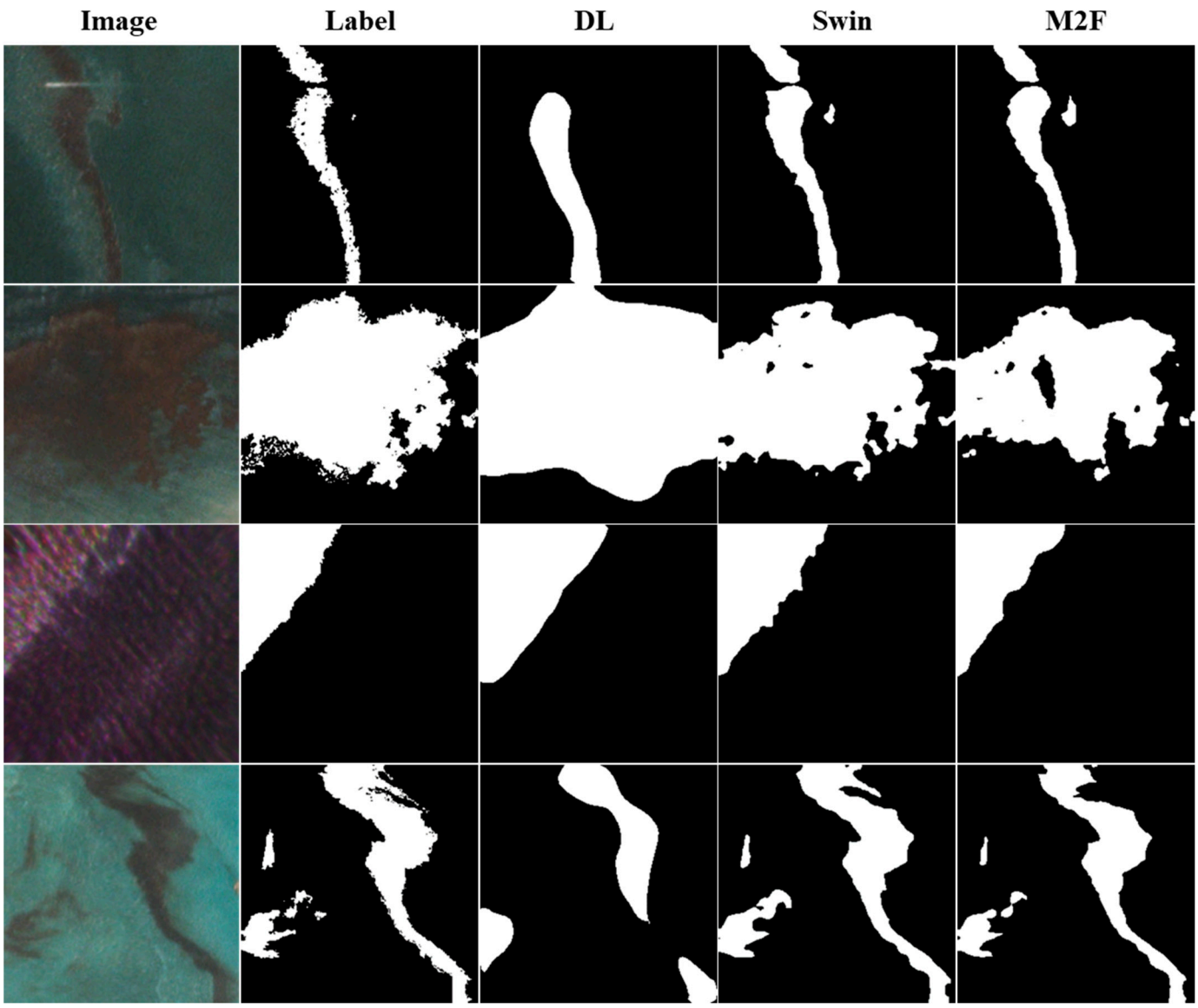

Figure 8. Randomly selected examples from fold 4, including PlanetScope RGB images, segmentation labels, and predictions from DeepLabV3+ (DL), Swin-UPerNet (Swin), and Mask2Former (M2F).

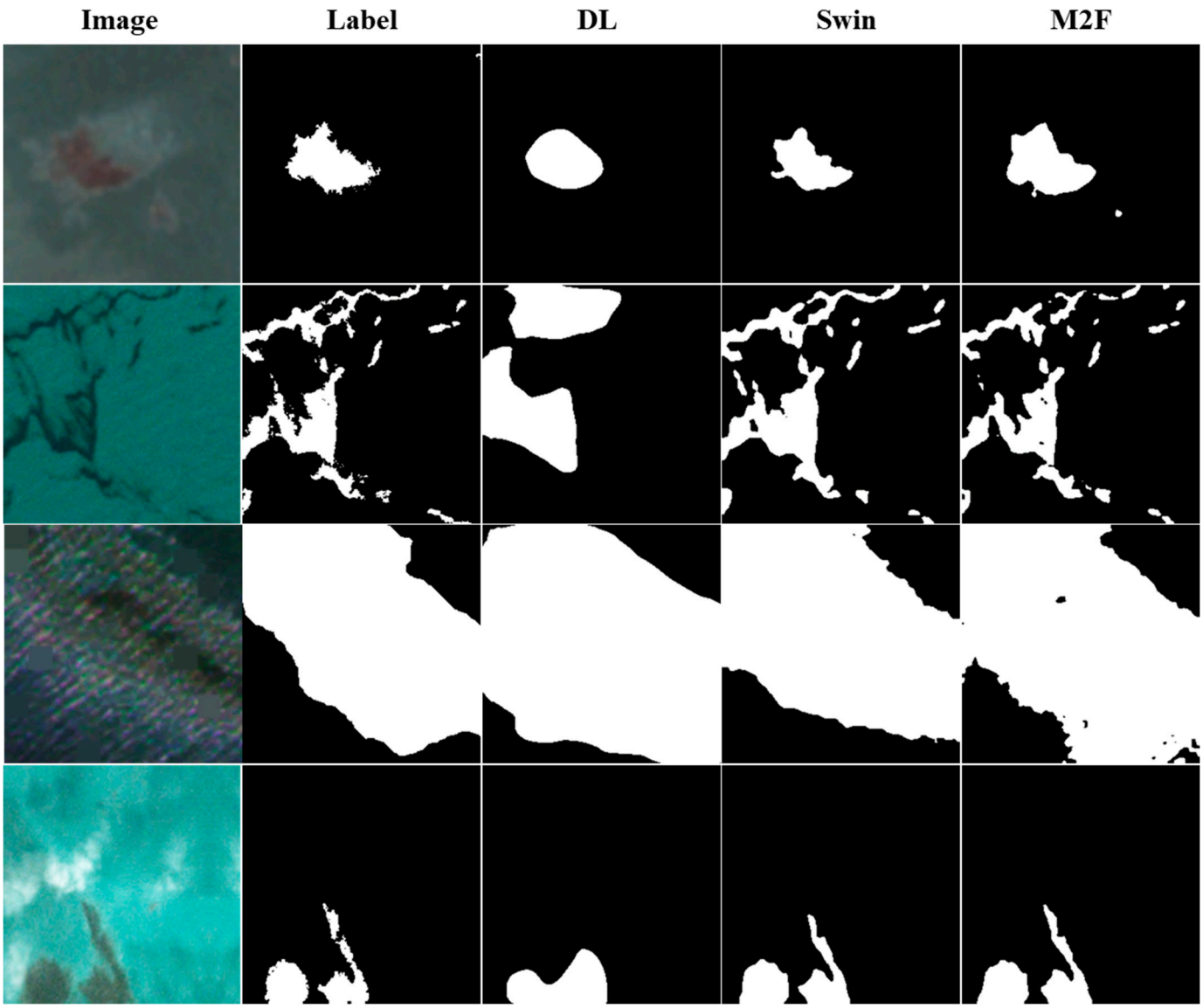

Figure 9. Randomly selected examples from fold 5, including PlanetScope RGB images, segmentation labels, and predictions from DeepLabV3+ (DL), Swin-UPerNet (Swin), and Mask2Former (M2F).

5. Discussion

To date, optical satellite images have rarely been used for deep learning-based detection of marine oil spills. Furthermore, objective comparisons between CNN and Transformer models for oil spill detection are lacking. In this study, we used high-resolution optical satellite images to compare the performance of DeepLabV3+, Swin-UPerNet, and Mask2Former, which are representative models for CNN and Transformer architectures, respectively. Figures 5–9 demonstrate that all the models are capable of detecting the location and shape of oil spills. Table 4 shows that DeepLabV3+, Swin-UPerNet, and Mask2Former achieved high scores across evaluation metrics such as mIoU, accuracy, precision, recall, F1 score, and Kappa coefficient. Notably, the mIoU performance indicates that all models can effectively detect marine oil spills, highlighting its importance as a critical evaluation metric in semantic segmentation. Additionally, the similar precision and recall values suggest that there was no significant tendency for overestimation or underestimation. The five-fold cross-validation yielded consistent results across each fold, indicating that the models are relatively stable, with minimal risk of overfitting or bias.

However, the quantitative and qualitative results indicate that the Swin-UPerNet and Mask2Former models, both utilizing the Transformer backbone, outperform DeepLabV3+,

although detecting oil spills with non-uniform shapes and spectral patterns remains challenging. When the proportion of the oil spill class in the image was larger, all the models were able to detect oil spills stably. However, when the proportion of the oil spill was small, DeepLabV3+ occasionally failed to detect it, whereas both Swin-UPerNet and Mask2Former consistently and reliably identified these spills. This suggests that models utilizing the Transformer backbone are more suitable for oil spill detection due to their attention mechanism, which enables more effective feature mapping, despite a slightly higher computational resource requirement. Additionally, the architecture of Transformer models enables them to retain fine details in edge areas. In CNN-based models, edges often become smoothed during the convolution and pooling processes used for downsampling, as these operations tend to blur spatial boundaries when the image resolution decreases. However, the attention mechanisms in Swin-UPerNet and Mask2Former allow these models to capture and preserve local details, making them more effective at delineating the intricate borders of oil spills. Their scale adaptability ensures both global context awareness and fine-grained spatial attention. This ability to represent detailed edges in Transformer models is especially beneficial in cases like oil spills, where complex, irregular boundaries are common.

Both Transformer models, Swin-UPerNet and Mask2Former, outperformed the CNN-based DeepLabV3+ model. In particular, Swin-UPerNet consistently showed better performance than Mask2Former in our experiment. This may be because (1) the shifted-window (Swin) self-attention mechanism effectively captures both global and local features; (2) the UPerNet model employs pyramid pooling to combine multi-scale features effectively and incorporates a head structure for detailed binary delineation, whereas Mask2Former is optimized for instance segmentation; and (3) Swin-UPerNet's simpler design, compared to Mask2Former, enables better performance under constrained computing resources.

Despite the advantages of optical satellite images, there are still some limitations. For instance, optical satellite images can be obscured by clouds, making it challenging to collect a diverse range of oil spill cases that account for variations in color and thickness based on the characteristics of the oil spills. The IoU for the oil spill class was 0.608 for DeepLabV3+, 0.758 for Swin-UPerNet, and 0.704 for Mask2Former, while the IoU for the non-oil class was 0.873 for DeepLabV3+, 0.923 for Swin-UPerNet, and 0.904 for Mask2Former. This difference is likely due to the challenge of obtaining sufficient training data for marine oil spills. Although DeepLabV3+ does not perform as well as Swin-UPerNet and Mask2Former across multiple performance metrics, all the models could benefit from training on image datasets with a broader spatiotemporal range, enabling them to learn a wider variety of oil spill patterns. The superior performance metrics of Swin-UPerNet and Mask2Former, both utilizing the Swin Transformer backbone, highlight their advanced adaptability to diverse oil spill scenarios. This suggests that optimizing model architectures for specific environmental conditions could further enhance their effectiveness. Additionally, providing more detailed information on oil types could further improve model performance. Future works should also explore more state-of-the-art (SOTA) models beyond Swin-UPerNet and Mask2Former and consider using an ensemble of multiple models.

Moreover, experiments were conducted on color characteristics and oil-type classification through a histogram analysis of spilled oil pixels. Figures 10–12 show the histogram distribution graphs and box plots of the oil spill pixels extracted from the labels, DeepLabV3+, Swin-UPerNet, and Mask2Former, respectively. In the histogram graphs, the x-axis represents the pixel value, and the y-axis represents the number of pixels. In the box plots, the x-axis indicates the pixel value, and the y-axis indicates the RGB band. The histogram and box plot for Figure 10 show that the red, green, and blue band values are all low and clustered together, with little variation. In contrast, the histograms and box plots for Figures 11 and 12 show a higher distribution of pixel values in the red, green, and blue bands compared to Figure 10. The histograms for the red, green, and blue bands in Figure 11 almost overlap, while those in Figure 12 are somewhat misaligned. This suggests that the oil in Figure 10, with relatively low RGB pixel values, is close to dark black, while

the oil in Figures 11 and 12, with relatively high RGB pixel values, has a bright silver or rainbow color. The histograms and box plots for the same cases exhibit similar patterns.

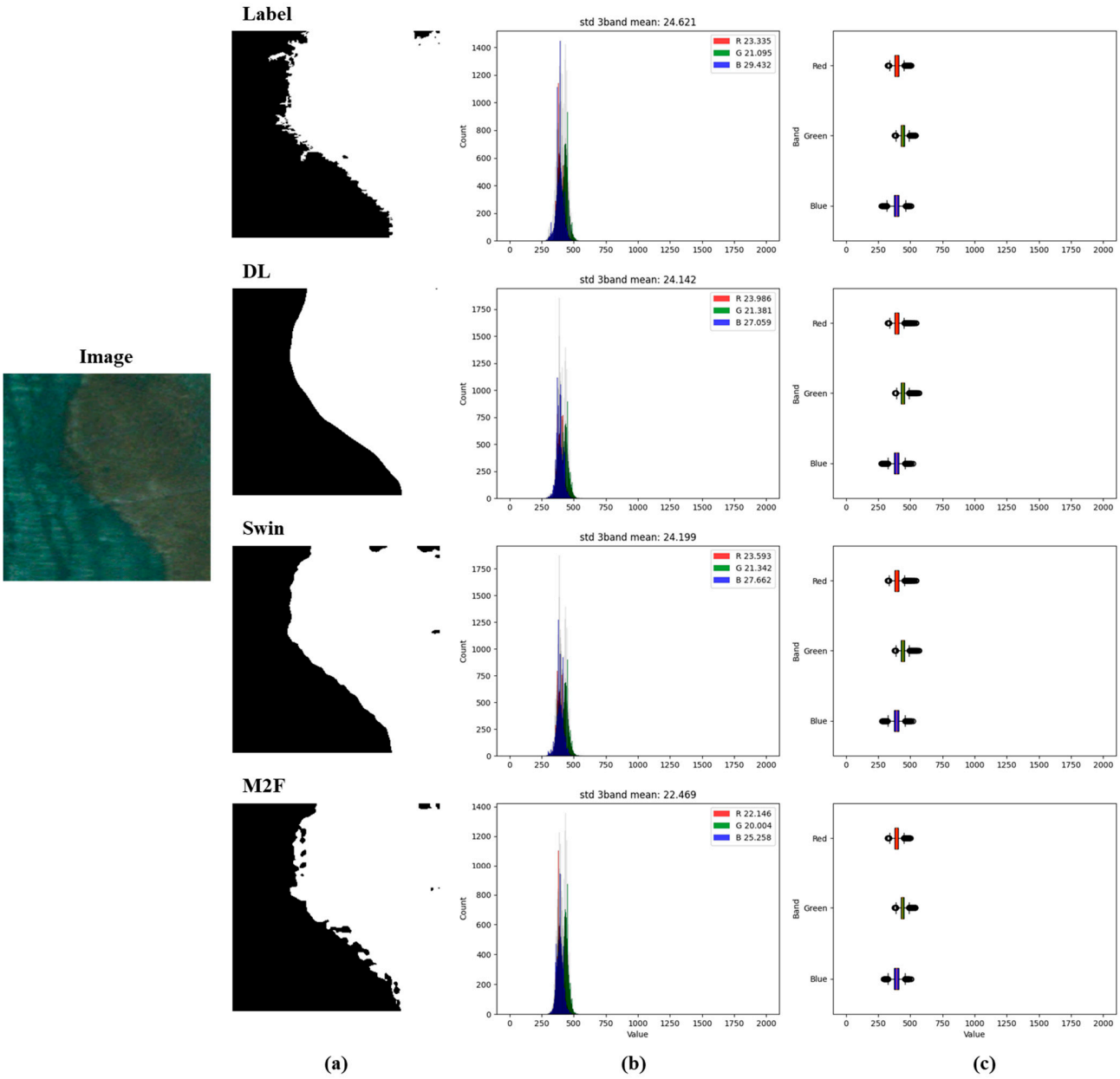

Figure 10. Thick oil layers with a dark black tone: histogram distribution graph and box plot of oil spill pixels extracted from the labels, DeepLabV3+, Swin-UPerNet, and Mask2Former. The *x*-axis values represent the digital numbers (DNs) from PlanetScope images. (**a**) Oil mask, (**b**) histogram, and (**c**) box plot.

Thick oil layers are often brown or black, while thin oil layers appear silver [21]. Based on this, it is inferred that the oil in Figure 10 represents a thick oil layer, while the oils in Figures 11 and 12 can be thin oil layers. The rainbow colors in the thin oil layers are caused by light interference, where light reflected from the top and bottom surfaces of the oil film interferes with each other. Additionally, oil refracts light differently depending on the wavelength, resulting in various colors depending on the viewing angle. The thin oil layers in Figures 11 and 12 exhibit a silver or rainbow color, with Figure 11 being closer to a silver tone and Figure 12 showing a more remarkable rainbow hue. This difference is reflected in the histograms. In Figure 11 of the silver tone, the histograms of the red, green,

and blue bands almost overlap, while, in Figure 12 of the rainbow tone, the histograms are somewhat misaligned. These oil spill color characteristics can be classified into (1) thick oil layers with a dark black tone, (2) thin oil layers with a bright silver tone, and (3) thin oil layers with a bright rainbow tone, based on the histogram analysis. The thick oil layer with a dark black tone is likely heavy oil, such as bunker C oil, while the thin oil layers with bright silver and rainbow tones are likely diesel oil.

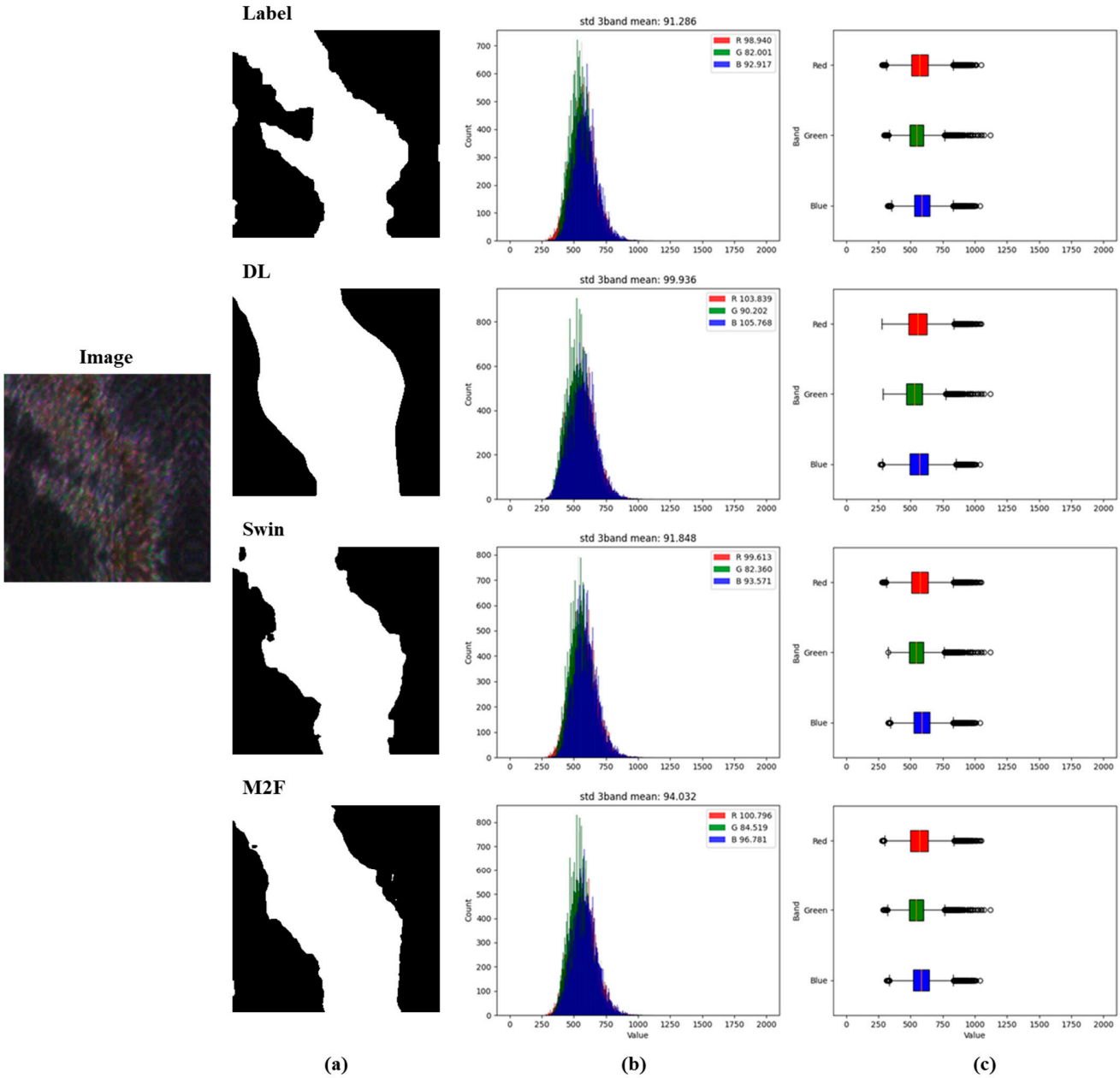

Figure 11. Thin oil layers with a bright silver tone: histogram distribution graph and box plot of oil spill pixels extracted from the labels, DeepLabV3+, Swin-UPerNet, and Mask2Former. The *x*-axis values represent the digital numbers (DNs) from PlanetScope images. (**a**) Oil mask, (**b**) histogram, and (**c**) box plot.

Distinguishing oil types through color characteristics and histogram analysis offers substantial practical value for environmental monitoring and response. By categorizing oil spills based on color characteristics, such as dark black tones for heavy oils and silver or rainbow tones for lighter oils, responders can quickly identify the type of oil involved. This distinction is crucial, as different oil types require tailored cleanup approaches. Heavy

oils, often persistent in marine environments, require robust removal strategies [28,29], while lighter, volatile oils may disperse more readily but pose acute toxicity risks [28,30]. Additionally, color analysis enables the detection of oil spill thickness and condition over time, aiding in the monitoring of oil dispersion or degradation. Utilizing color analysis enhances the ability to rapidly assess and respond to oil spill incidents, supporting more effective and responsive environmental management. Accurate extraction of oil spill pixels is also essential for oil spill color classification, as the inclusion of non-oil pixels can distort the results. Transformer-based models, such as Swin-UPerNet and Mask2Former, excel at precisely delineating oil spill edges, making them more effective at isolating only the oil spill pixels, leading to reliable histogram analyses of oil characteristics and clearer classification of oil types based on color properties.

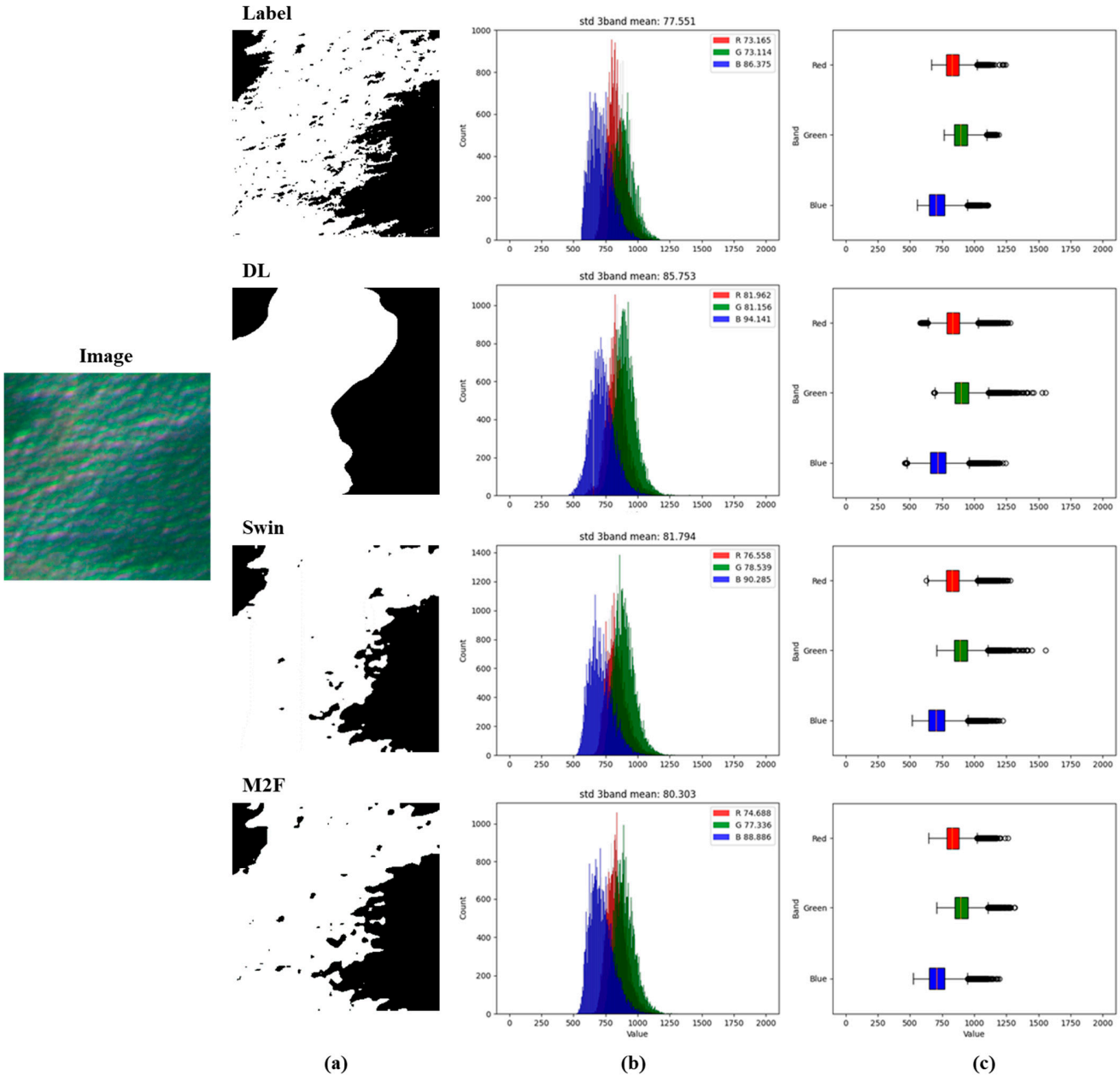

Figure 12. Thin oil layers with a bright rainbow tone: histogram distribution graph and box plot of oil spill pixels extracted from the labels, DeepLabV3+, Swin-UPerNet, and Mask2Former. The *x*-axis values represent the digital numbers (DNs) from PlanetScope images. (**a**) Oil mask, (**b**) histogram, and (**c**) box plot.

6. Conclusions

While SAR images are primarily used for oil spill detection, combining SAR and optical images could improve the spatial and temporal coverage of oil spill detection. In this context, this paper introduces a new approach to assess the feasibility of optical satellite images for marine oil spill detection. This study utilized high-resolution optical satellite imagery from PlanetScope to perform marine oil spill detection using DeepLabV3+, Swin-UPerNet, and Mask2Former, representing CNN and Transformer models, respectively. We conducted five blind tests to objectively compare and evaluate the applicability of these models. Additionally, we were able to classify the type of oil by analyzing the histogram of the oil spill pixels predicted by the deep learning models. A total of 260 images with four channels and 256 × 256 pixel dimensions were augmented to 2600 images for training and evaluation. Overall, the results suggest that both the CNN and Transformer models can detect oil spills with high accuracy, but the Transformer models (Swin-UPerNet and Mask2Former) outperformed the CNN model (DeepLabV3+). Particularly, Swin-UPerNet consistently demonstrated the highest performance, attributed to its shifted-window self-attention mechanism, pyramid pooling for multi-scale feature integration, and head structure for detailed binary delineation. Furthermore, Transformer models are particularly effective in detecting oil spills with complex, irregular shapes, as their ability to preserve fine edge details makes them more adept at delineating the intricate forms of oil spills. To address the limitations of dataset volume, future research should incorporate additional datasets and explore novel image augmentation techniques using generative adversarial networks (GANs). These efforts could further enhance the performance of deep learning models for marine oil spill detection. Future work will be needed to develop a robust segmentation model using a multi-model ensemble with automated hyperparameter optimization. Additionally, the combination of SAR and optical images should be explored to ensure broader spatial and temporal coverage for oil spill detection.

Author Contributions: Conceptualization, J.K. and Y.L.; methodology, J.K. and Y.L.; formal analysis, J.K.; data curation, J.K.; writing—original draft preparation, J.K.; writing—review and editing, J.K., C.Y., J.Y. and Y.L.; project administration, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the "Monitoring System of Spilled Oils Using Multiple Remote Sensing Techniques" funded by the Korea Coast Guard, Korea. This research was supported by a grant (2021-MOIS37-002) of the Intelligent Technology Development Program on Disaster Response and Emergency Management funded by the Ministry of Interior and Safety (MOIS, the Republic of Korea).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Author Jonghyuk Yi was employed by the company SE Lab Incorporation. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Brekke, C.; Solberg, A.H.S. Oil spill detection by satellite remote sensing. *Remote Sens. Environ.* **2005**, *95*, 1–13. [CrossRef]
2. Solberg, A.H.S. Remote sensing of ocean oil-spill pollution. *Proc. IEEE* **2012**, *100*, 2931–2945. [CrossRef]
3. Alpers, W.; Benjamin, H.; Kan, Z. Oil spill detection by imaging radars: Challenges and pitfalls. *Remote Sens. Environ.* **2017**, *201*, 133–147. [CrossRef]
4. Zhao, J.; Temimi, M.; Ghedira, H.; Hu, C. Exploring the potential of optical remote sensing for oil spill detection in shallow coastal waters-a case study in the Arabian Gulf. *Opt. Express* **2014**, *22*, 13755–13772. [CrossRef] [PubMed]
5. Mityagina, M.; Lavrova, O. Satellite survey of inner seas: Oil pollution in the Black and Caspian seas. *Remote Sens.* **2016**, *8*, 875. [CrossRef]
6. Kolokoussis, P.; Karathanassi, V. Oil spill detection and mapping using sentinel two imagery. *J. Mar. Sci. Eng.* **2018**, *6*, 4. [CrossRef]

7. Arslan, N. Assessment of oil spills using Sentinel 1 C-band SAR and Landsat 8 multispectral sensors. *Environ. Monit. Assess.* **2018**, *190*, 637. [CrossRef]
8. Rajendran, S.; Vethamony, P.; Sadooni, F.N.; Al-Kuwari, H.A.S.; Al-Khayat, J.A.; Seegobin, V.O.; Govil, H.; Nasir, S. Detection of Wakashio oil spill off Mauritius using Sentinel-1 and 2 data: Capability of sensors, image transformation methods, and mapping. *Environ. Pollut.* **2021**, *274*, 116618. [CrossRef]
9. Park, S.H.; Jung, H.S.; Lee, M.J. Oil spill mapping from Kompsat-2 high-resolution image using directional median filtering and artificial neural network. *Remote Sens.* **2020**, *12*, 253. [CrossRef]
10. Aznar, F.; Sempere, M.; Pujol, M.; Rizo, R.; Pujol, M.J. Modelling oil-spill detection with swarm drones. *Abstr. Appl. Anal.* **2014**, *2014*, 949407. [CrossRef]
11. Odonkor, P.; Ball, Z.; Chowdhury, S. Distributed operation of collaborating unmanned aerial vehicles for time-sensitive oil spill mapping. *Swarm Evol. Comput.* **2019**, *46*, 52–68. [CrossRef]
12. Topouzelis, K.; Karathanassi, V.; Pavlakis, P.; Rokos, D. Detection and discrimination between oil spills and look-alike phenomena through neural networks. *ISPRS J. Photogramm. Remote Sens.* **2007**, *62*, 264–270. [CrossRef]
13. Singha, S.; Bellerby, T.J.; Trieschmann, O. Satellite oil spill detection using artificial neural networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 2355–2363. [CrossRef]
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*, 1762. [CrossRef]
15. Yekeen, S.T.; Balogun, A.L.; Yusof, K.B.W. A novel deep learning instance segmentation model for automated marine oil spill detection. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 190–200. [CrossRef]
16. Jiao, Z.; Jia, G.; Cai, Y. A new approach to oil spill detection that combines deep learning with unmanned aerial vehicles. *Comput. Ind. Eng.* **2019**, *135*, 1300–1311. [CrossRef]
17. Lalitha, S.D.; Senthilkumar, S.; Kumar, B.P.; Jambulingam, U.; Anish, T.P.; Kalpana, A.V. Advanced AI-based System for Precision Oil Spill Detection in Marine Environments. In *Maintaining a Sustainable World in the Nexus of Environmental Science and AI.*; IGI Global: Hershey, PA, USA, 2024; pp. 131–160.
18. Vekariya, D.; Vaghasiya, M.; Tomar, Y.; Laad, P. A Survey on Oil Spill Detection using SAR images in Machine Learning. In Proceedings of the 2024 Parul International Conference on Engineering and Technology (PICET), Vadodara, India, 3–4 May 2024.
19. Liao, L.; Zhao, Q.; Song, W. Monitoring of oil spill risk in coastal areas based on polarimetric SAR satellite images and deep learning theory. *Sustainability* **2023**, *15*, 14504. [CrossRef]
20. Ding, J.; Li, W.; Pei, L.; Yang, M.; Ye, C.; Yuan, B. Sw-YoloX: An anchor-free detector based transformer for sea surface object detection. *Expert Syst. Appl.* **2023**, *217*, 119560. [CrossRef]
21. Kang, J.; Youn, Y.; Kim, G.; Park, G.; Choi, S.; Yang, C.; Yi, J.; Lee, Y. Detection of marine oil spills from PlanetScope images using DeepLabV3+ model. *Korean J. Remote Sens.* **2022**, *38*, 1623–1631.
22. Planet Imagery and Archive. Available online: https://www.planet.com/products/planet-imagery (accessed on 6 November 2022).
23. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
24. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
25. Misra, I.; Schwing, A.G.; 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 (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
26. Lateef, F.; Ruichek, Y. Survey on semantic segmentation using deep learning techniques. *Neurocomputing* **2019**, *338*, 321–348. [CrossRef]
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
28. Johann, S.; Goßen, M.; Mueller, L.; Selja, V.; Gustavson, K.; Fritt-Rasmussen, J.; Wegeberg, S.; Ciesielski, T.M.; Jenssen, B.M.; Hollert, H.; et al. Comparative toxicity assessment of in situ burn residues to initial and dispersed heavy fuel oil using zebrafish embryos as test organisms. *Environ. Sci. Pollut. Res.* **2021**, *28*, 16198–16213. [CrossRef] [PubMed]
29. Zhang, B.; Matchinski, E.J.; Chen, B.; Ye, X.; Jing, L.; Lee, K. Marine oil spills—Oil pollution, sources and effects. In *World Seas: An Environmental Evaluation*; Academic Press: Cambridge, MA, USA, 2019; pp. 391–406.
30. Redman, A.D.; Parkerton, T.F. Guidance for improving comparability and relevance of oil toxicity tests. *Mar. Pollut. Bull.* **2015**, *98*, 156–170. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.