Oil Spill Segmentation in SAR images: A Comparative Study of Loss Functions

Maximiliano Lexow[1,2], Claudio A. Delrieux[1], and Nicolás A. García[2]

[1] Laboratorio de las Imágenes, Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina
[2] Instituto de Física del Sur (IFISUR), Universidad Nacional del Sur – CONICET, Bahía Blanca, Argentina lexow.maximiliano@uns.edu.ar

Abstract. SAR images are a fundamental tool for monitoring the ocean because, unlike optical images, they are not affected by weather conditions or sunlight. However, oil spill segmentation remains a complex task due to confusion between semantically similar phenomena (look-alikes), as well as highly imbalanced data.
In this work, we analyze how loss functions affect segmentation using deep learning. We evaluated four losses configurations, including standard cross-entropy (CE), weighted versions and a combination of CE with Dice Loss. We assessed the model's performance using quantitative metrics—Dice, IoU, Precision, Recall, and F1-score— as well as confusion matrices. This allows us to determine whether the model performs better with certain classes or in a more general sense.
The results show that incorporating weights into the loss function significantly improves minority class detection when dealing with a severe class imbalance. However, using extremely high or low weights can be detrimental for the model. The best performing model was achieved by the combination of weighted CE with Dice Loss, which achieved a Dice Score of 71% and a F1-score for the oil spill class of 0.64.

Keywords: oil spill detection, loss functions, deep learning, SAR imagery

Segmentación de Derrames de Petróleo en Imágenes SAR: Un Estudio Comparativo de Funciones de Pérdida

Resumen. Las imágenes SAR constituyen una herramienta fundamental para monitorear el océano, ya que, al contrario que las imágenes ópticas, no son afectadas por las condiciones climáticas ni la luz solar. A pesar de esto, la segmentación de derrames de petróleo se mantiene como una tarea compleja debido a la confusión entre fenómenos semánticamente similares (look-alikes), así como también por datos muy desbalanceados. En este trabajo, analizamos como las funciones de pérdida afectan a la

segmentación mediante técnicas de aprendizaje profundo. Evaluamos cuatro configuraciones de pérdida, incluyendo cross-entropy estándar (CE), versiones pesadas y una combinación de CE con la pérdida de Dice. Evaluamos el rendimiento del modelo utilizando métricas cuantitativas— coeficiente de Dice, intersección sobre unión IoU, precisión, sensibilidad y medida F1— así como también matrices de confusión. Esto nos permite determinar si el modelo se desempeña mejor con ciertas clases, o en un sentido más general. Los resultados muestran que incorporar pesos a la función de pérdida mejora significativamente la detección de clases minoritarias cuando se tiene que lidiar con desbalance severo. Sin embargo, usar pesos extremadamente altos o bajos puede ser perjudicial para el modelo. El modelo con mejores métricas fue conseguido mediante la combinación de CE y la pérdida de Dice, el cual obtuvo un coeficiente de Dice de 71% y una medida F1 para la clase petróleo de 0.64.

Palabras clave: detección de derrames de petróleo, funciones de pérdida, aprendizaje profundo, imágenes SAR

1 Introduction

Oil spills, whether natural or man-made, have a major impact on the marine environment, as (Carpenter & Reichelt-Brushett, 2023) and (Chang et al., 2014) point out. Whether in terms of natural leaks from undersea oil reservoirs, or prosecuting illegal discharges from ships or oil rigs, it is vitally important to detect them in time. Currently, various satellites orbit the globe and gather images and other types of data all the time. However, manual inspection of this amount of data is impractical, so it is necessary to automate the analysis, making deep learning (DL) techniques, in particular Convolutional Neural Networks, an inviting alternative.

Among the various satellite sensing modalities, Synthetic Aperture Radar (SAR) is particularly suited for oil spill monitoring, as it operates independently of weather conditions and sunlight, enabling persistent surveillance of the same ocean areas over long periods of time. In this work we use the SAR dataset introduced in (Krestenitis et al., 2019b), further described in Section 2.

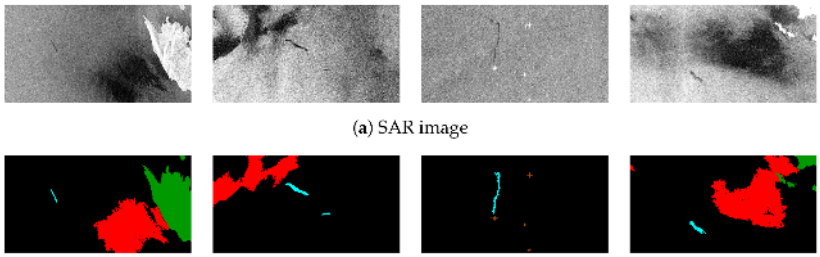
Datasets often suffer from severe class imbalance, as oil spills constitute small portions of the images. In the dataset used for this work, 1% of total pixels are oil slicks. Another problem is distinguishing oil slicks from other dark spots (look-alikes), such as low-wind areas, wave shadows, internal waves, or biogenic slicks. DL models implicitly learn the physical characteristics (shape, size, contrast) of verified oil spills to discriminate them from look-alikes, making them appropriate for this task.

This work seeks to evaluate the effect of different loss functions on multi-class SAR oil spill semantic segmentation, with emphasis on improving the class imbalance problem and oil/look-alike discrimination.

2 Materials and methods

2.1 Dataset

The used dataset contains JPG images derived from the Sentinel-1 European Satellite missions, acquired from the Copernicus Open Access Hub, of the European Space Agency (ESA) database, as described in (Krestenitis et al., 2019b), (Krestenitis et al., 2019a). The images contain 5 semantic classes: oil spill, look-alike, land, ship and ocean (background), as well as their corresponding ground-truth segmentation masks. The dataset contains 1002 images for training and 110 images for testing. We further split the training set into 801 for training and 201 for validation following an 80/20 split. The original resolution is 1250×650 pixels; we resized them to 512×256 to preserve the original aspect ratio as much as possible while satisfying the input constraints of the segmentation architecture (size must be a multiple of 32). Aspect ratio is important to maintain because oil spills have a characteristic elongated geometry, which otherwise will be lost.



(a) SAR image

(b) Ground truth

Fig. 1: Samples of dataset: (a) grayscale SAR images and (b) are ground truth segmentation masks. Cyan color corresponds to oil spills, red to look-alikes, brown to ships, green to land and black is for sea surface.

Table 1 shows the pixel distribution across classes in the training set: The dataset

Table 1: Pixel distribution per class in the training set.

Class	Pixels (Millions)	Percentage
Ocean	573.3 M	88.09%
Oil Spill	6.6 M	1.02%
Look-alike	36.6 M	5.63%
Ship	0.3 M	0.04%
Land	34.0 M	5.22%

exhibits severe class imbalance, a common problem in semantic segmentation. Oil spill and ship classes are heavily underrepresented, 1.02% and 0.04% of the

training set respectively. This motivates the use of weighted loss functions, as described in Section 2.3.

To further improve the variability of the dataset, data augmentation (Buslaev et al., 2020) was used, introducing random flips and brightness and contrast jitter.

2.2 Framework

All experiments were implemented in Python, utilizing PyTorch as the Deep Learning framework. The hardware used to train the models is composed of a single NVIDIA GeForce RTX 3060 with 12 GB of VRAM, a 12th Gen Intel(R) Core(TM) i9-12900K and 32 GB of RAM. To ensure reproducibility, a fixed random seed of 42 was set prior to each experiment, controlling randomness in data augmentation and weight initialization.

The segmentation architecture was implemented using the publicly available python library; Segmentation Models Pytorch (Iakubovskii, 2019). This library provides modular implementations of segmentation architectures, encoders, and loss functions, making experimentation much easier.

2.3 Experimental design

Loss functions are one of the most important aspects of neural networks, as they are directly responsible for fitting the model to the given training data. As noted in the introduction, in semantic segmentation, where class imbalance is a common problem, the loss needs to take into account small details in the image (pixel level), but also have regional coherence.

To make things reproducible, some parameters need to be set constant, in this case we fixed:

- Architecture: UNet with ImageNet pretrained weights (Ronneberger et al., 2015) and ResNet-34 encoder (He et al., 2015)
- Optimizer: AdamW, with weight decay 10^{-2}
- Learning rate: 10^{-3}
- Batch size: 16
- Epochs: 100, with early stopping if validation loss does not improve for 10 consecutive epochs

Table 2: Notation used in segmentation loss (from Azad et al., 2023).

Symbol	Description
N	Number of pixels.
C	Number of target classes.
t_{nc}	Binary indicator: 1 if the nth pixel belongs to class c, otherwise 0.
$\hat{p}_{nc}$	Predicted class probabilities for the nth pixel.
w_c	Weight assigned to class c.

We evaluate four loss function configurations:

1. **CE**: Standard cross-entropy loss, no weighting, used as baseline

$$L_{CE}(w_c) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c \times t_{nc} \times log\left(\hat{p}_{nc}\right), \qquad w_c = 1$$

2. **CE-IFW**: Cross-entropy loss with inverse frequency weights:

$$L_{CE-IFW} = L_{CE}, \qquad w_c^{IFW} = \frac{N}{C \times n_c},$$

where n_c is is the number of pixels belonging to class c.

3. **CE-Norm**: Cross-entropy loss with normalized weights:

$$L_{CE-Norm} = L_{CE}, \qquad w_c^{Norm} = \frac{\sqrt{w_c^{IFW}}}{\frac{1}{C}\sum_{c=1}^{C}\sqrt{w_c^{IFW}}}$$

4. **Combo-Loss**: Combination of CE-Norm and Dice loss (1 - Dice coefficient):

$$L_{Combo-Loss} = L_{CE-Norm} + L_{dice},$$

$$L_{dice} = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{n=1}^{N} t_{nc} \times \hat{p}_{nc}}{\sum_{n=1}^{N} t_{nc} + \sum_{n=1}^{N} \hat{p}_{nc}}$$

Table 3 shows the differences in both weighting strategies:

Table 3: Class weight comparison for the two selected weighting strategies.

Class	w_c^{IFW}	w_c^{Norm}
Ocean	0.23	0.08
Oil Spill	19.66	0.74
Look-alike	3.55	0.31
Ship	456.38	3.55
Land	3.83	0.32

We can see that CE-IFW produces very large weights for the ship and oil spill class ($w_{\text{ship}} = 456.38$, $w_{\text{oil spill}} = 19.66$) which can destabilize training. CE-Norm has reduced dynamic range while preserving the relative ordering of class priorities.

3 Results and discussion

Regarding metrics, we utilized:

– Dice Score

– Intersection over Union (IoU) Score

– Precision

– Recall

– F1-score

Dice and IoU scores are essential metrics for evaluating segmentation performance, each with its strengths and limitations. IoU is stricter, weighing false positives, false negatives, and true positives equally, while Dice emphasizes overlap, making it generally less harsh, especially for small or imperfect segmentations. Reporting both metrics together provides a more comprehensive view of model performance. Precision is the proportion of all positive classificated pixels that are truly positive. Recall is the proportion of actual positive pixels, that were correctly classified as positives. F1-score is the harmonic mean of these two metrics and is high only when both precision and recall are high.

In the oil spill detection problem, a false positive (predicting oil spill when its not) is more acceptable than a false negative (missing an oil spill), therefore, high recall is more important than high precision. Nevertheless, a good model needs to be balanced, that's why we consider F1-score as the go-to metric.

We are going to discuss results on the most relevant classes: Oil Spill, Look-alike and Ship. Oil Spills are the main focus of this work. Look-alikes are important because the model should be able to differentiate between textures and shapes similar to that of oil spills. And ships are important to detect because they are sometimes related to the spills.

Confusion matrices (CM) were used to visualize class-wise model performance, and identify the predominant confusion patterns. Each CM is row-normalized, meaning the sum of each row is 100%, which aids interpretation under heavy class imbalance. Rows correspond to ground truth labels, while columns represent model predictions. The main diagonal accounts for the recall of each class.

In Figure 2 we can see 4 SAR images, its respective segmentation masks, and the predicted masks for each of the 4 experiments.

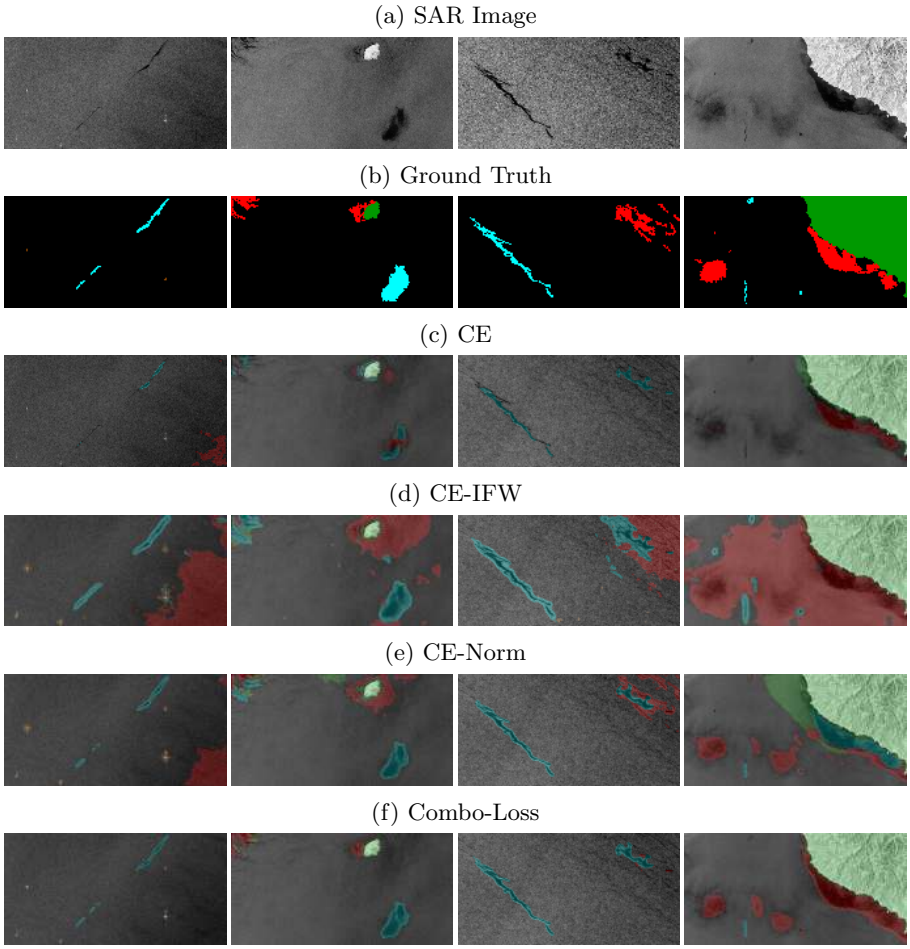


Fig. 2: Segmentation results on 4 test samples. (a) SAR image, (b) ground truth mask, (c)–(f) predicted segmentation overlaid on SAR image for each loss function.

Before looking at more exact results with metrics and confusion matrices, we can have a general sense of how each loss work. CE is overlooking minority classes, missing ships entirely, and poorly detecting oil spills. CE-IFW is placing too much emphasis on rare classes, predicting ships on every bright spot. CE-Norm considerably improves, and Combo-Loss seems to be the best overall.

Tables 4 to 7 shows the obtained metrics of all experiments for all classes. Note that the best model for each class is bold.

Table 4: Dice and IoU metrics for each of the tested losses.

Experiment	Dice	Iou
CE	0.52	0.39
CE-IFW	0.47	0.28
CE-Norm	0.65	0.46
Combo-Loss	**0.71**	**0.52**

Table 5: F1 metrics for each of the tested losses.

Experiment	Background	Oil Spill	Look Alike	Ship	Land
CE	0.97	0.45	0.33	0.00	0.84
CE-IFW	0.89	0.38	0.42	0.03	0.64
CE-Norm	0.97	0.63	**0.67**	0.16	**0.84**
Combo-Loss	**0.97**	**0.64**	**0.67**	**0.49**	0.80

Table 6: Precision for each of the tested losses.

Experiment	Background	Oil Spill	Look Alike	Ship	Land
CE	0.95	**0.66**	**0.68**	0.00	**0.73**
CE-IFW	**1.00**	0.25	0.28	0.02	0.47
CE-Norm	0.99	0.53	0.58	0.09	0.73
Combo-Loss	0.98	0.57	0.65	**0.40**	0.67

Table 7: Recall for each of the tested losses.

Experiment	Background	Oil Spill	Look Alike	Ship	Land
CE	**0.98**	0.35	0.22	0.00	0.97
CE-IFW	0.80	**0.86**	**0.84**	**0.98**	**1.00**
CE-Norm	0.95	0.78	0.78	0.75	0.99
Combo-Loss	0.96	0.72	0.68	0.64	0.99

Table 4 shows that Combo-Loss has the best Dice and IoU score, meaning that the sensation we got from Figure 2 was correct. CE-IFW is the worst performing in this regional metrics, which also follows what we saw visually in Figure 2.

Combo-Loss has the highest F1-score for the three relevant classes, that is the most important metric, as we discussed. It means it has balance between being confident in a detection, while not missing too many instances.

CE has the best precision for Oil Spill and look-alike, which is important, but is has very low recall. That is because its very certain that a pixel is one of those two classes, but it misses a lot of them. It fails completely to recognize ships, as a consequence of the extremely low count of ship pixels in data.

CE-IFW has extremely low precision and the highest recall for the three relevant classes. That means it seldom fails to recognize one of those instances, but it tends to over-segment, labeling many background pixels as one of the relevant classes.

Weight normalization in CE-Norm partially helps, it doesn't get any top metric, but it gets the second highest recall for oil spill, maintaining a high enough value of F1-score. It shouldn't be discarded, although probably performs better when combined with other losses, as in Combo-Loss.

Figures 3 to 6 show row-normalized confusion matrices for the four loss configurations. As we discussed, the main diagonal shows the recall of each class, it can be seen that it shares the same values as the recall Table 7.

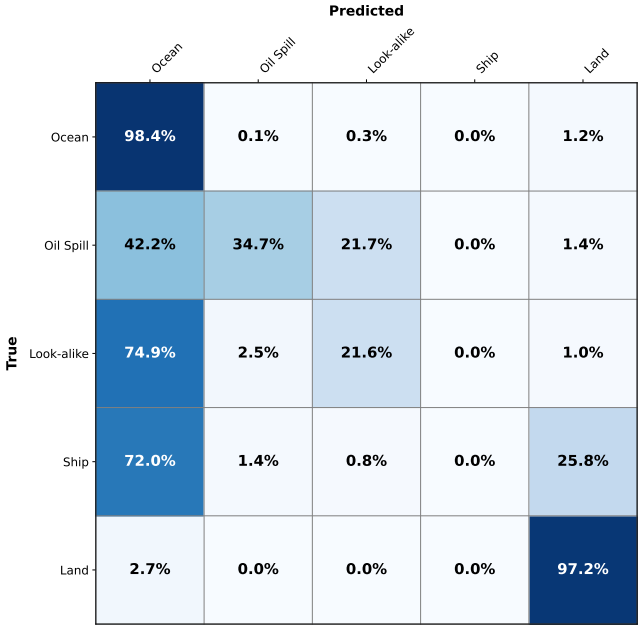


Fig. 3: Confusion matrix for Cross-Entropy (CE). Rows represent ground truth and columns model predictions.

Figure 3 (CE) confuses most of relevant classes pixels with background, which is a clear indicator that weights are necessary in imbalanced datasets.

Figure 4 (CE-IFW) shows a very clean CM, with a strong main diagonal. This corresponds with Table 7, which shows it has the best recall of all experiments, across all classes except background. Moreover, CE-IFW presents the worse background recall, which is a clear indicator of overrating rare classes, a consequence of its high aggressive weighting strategy.

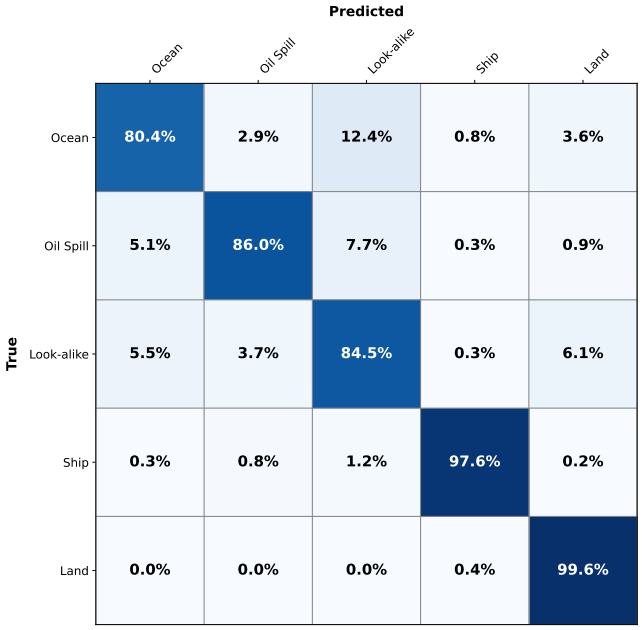


Fig. 4: Confusion matrix for Cross-Entropy with inverse frequency weighting (CE-IFW).

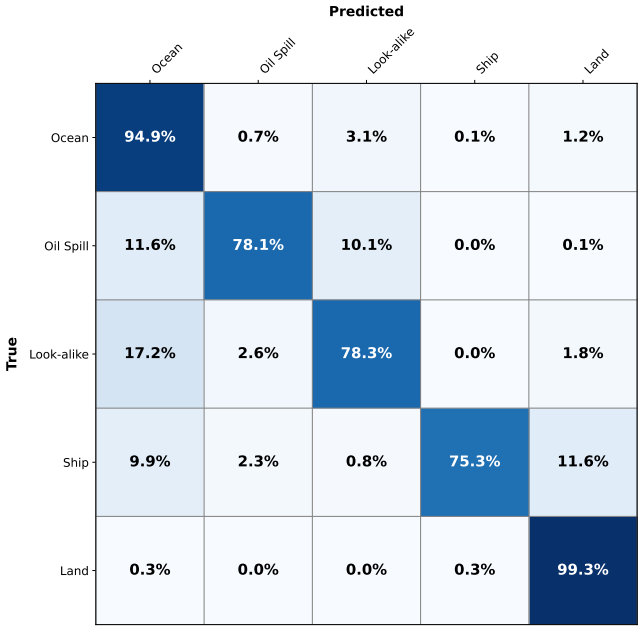


Fig. 5: Confusion matrix for normalized Cross-Entropy (CE-Norm).

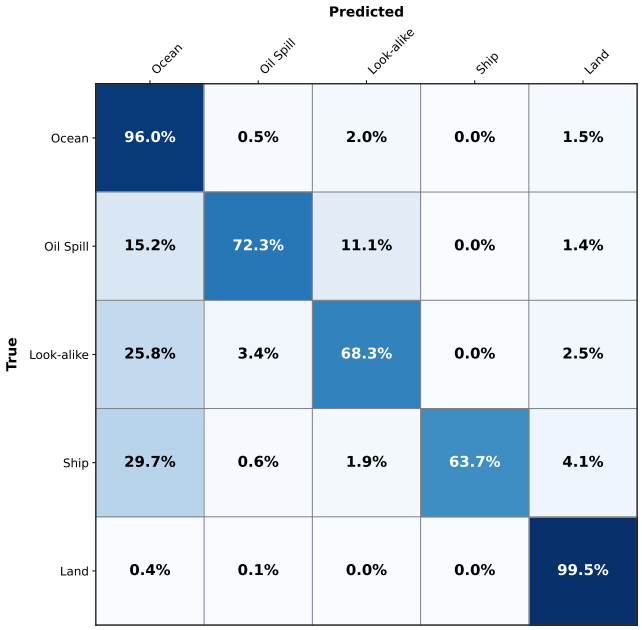


Fig. 6: Confusion matrix for the combined loss function (Combo-Loss).

Figures 5 (CE-Norm) and 6 (Combo-Loss) exhibit similar patterns, both predict a fraction of oil spills as background or look-alike; and a fraction of look-alike as background. Although CE-Norm has a more visually appealing CM, in terms of diagonal dominance, this doesn't imply better overall performance. As shown in 6, it achieves lower precision in every relevant class.

The explanation on why Combo-Loss has better F1-score than CE-Norm while having a weaker CM diagonal is that it handles background pixels more effectively. In particular, it reduces miss-classification of background pixels as oil spill, look-alikes or ships. This effect is critical due to the strong class imbalance. Background constitutes the majority of pixels, and even small relative errors can lead to a large number of false positives in rare classes. While CE-Norm achieves higher recall in most classes, it does so at the cost of incorrectly labeling background pixels.

4 Conclusions and future work

In this work, we evaluated four loss functions for semantic segmentation of oil spills in a SAR imagery dataset from Krestenitis et al., 2019b, using a UNet with Imagenet pretrained weights and a ResNet-34 encoder. The different losses were: Cross Entropy (CE), CE with inverse frequency weights (CE-IFW), CE with normalized weights (CE-Norm) and Combo-Loss, a combination of CE-Norm and Dice Loss.

Results show that Combo-Loss produces the best F1-score for the three relevant classes (oil spill, look-alike and ship), suggesting that combining a pixel focused loss and a regional loss is convenient for this task. It can handle class imbalance without sacrificing too much precision. CE without weights has the best precision, but lowest recall, and CE-IFW shows the opposite behaviour. This is because with the inverse frequency weights, too much focus is put on the rare classes; and if no weights are used, the rare classes are ignored. CE-Norm does not stand out in any metric, as it seems to work best when coupled with another loss like Dice.

For future work, an interesting addition is to use Tversky Loss (Salehi et al., 2017), which is designed to control the trade-off between false positives and false negatives. This allows finer tuning of the model, focusing more on precision or recall as needed.

Another idea that might be particularly useful in this work is introducing a loss function that weights specific cross-class misclassifications differently. For instance, it is more important to distinguish oil spill from look-alike than to correctly classify land.

References

Azad, R., Heidary, M., Yilmaz, K., Hüttemann, M., Karimijafarbigloo, S., Wu, Y., Schmeink, A., & Merhof, D. (2023, December 8). Loss functions in the era of semantic segmentation: A survey and outlook. https://doi.org/10.48550/arXiv.2312.05391

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, *11*(2), 125. https://doi.org/10.3390/info11020125

Carpenter, A., & Reichelt-Brushett, A. (2023). Oil and gas. In A. Reichelt-Brushett (Ed.), *Marine pollution – monitoring, management and mitigation* (pp. 129–153). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-10127-4_6

Chang, S. E., Stone, J., Demes, K., & Piscitelli, M. (2014). Consequences of oil spills: A review and framework for informing planning. *Ecology and Society*, *19*(2). https://doi.org/10.5751/ES-06406-190226

He, K., Zhang, X., Ren, S., & Sun, J. (2015, December 10). Deep residual learning for image recognition. https://doi.org/10.48550/arXiv.1512.03385

Iakubovskii, P. (2019). *Segmentation models pytorch*. https://github.com/qubvel-org/segmentation_models.pytorch

Krestenitis, M., Orfanidis, G., Ioannidis, K., Avgerinakis, K., Vrochidis, S., & Kompatsiaris, I. (2019a). Early identification of oil spills in satellite images using deep CNNs [Series Title: Lecture Notes in Computer Science]. In I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W.-H. Cheng, & S. Vrochidis (Eds.), *MultiMedia modeling* (pp. 424–435, Vol. 11295). Springer International Publishing. https://doi.org/10.1007/978-3-030-05710-7_35

Krestenitis, M., Orfanidis, G., Ioannidis, K., Avgerinakis, K., Vrochidis, S., & Kompatsiaris, I. (2019b). Oil spill identification from satellite images using deep neural networks. *Remote Sensing*, *11*(15), 1762. https://doi.org/10.3390/rs11151762

Ronneberger, O., Fischer, P., & Brox, T. (2015, May 18). U-net: Convolutional networks for biomedical image segmentation. https://doi.org/10.48550/arXiv.1505.04597

Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017, June 18). Tversky loss function for image segmentation using 3d fully convolutional deep networks. https://doi.org/10.48550/arXiv.1706.05721