SYNTHETIC LABEL MASKS MAPPED ON OCEAN SATELLITE BACKGROUND FOR OIL SEEPAGE DETECTION

Lionel Boillot
TotalEnergies OneTech
CSTJF, Avenue Larribau
Pau 64000, FRANCE
lionel.boillot@totalenergies.com

Frédérik Pivot
TotalEnergies OneTech
CSTJF, Avenue Larribau
Pau 64000, FRANCE
frederik.pivot@totalenergies.com

Félix Klein
TotalEnergies OneTech
CSTJF, Avenue Larribau
Pau 64000, FRANCE
felix.klein@external.totalenergies.com

ABSTRACT

Using satellite radar images, signal deflection differences make oil slicks appearing darker on top of ocean surface. They form puddles that drift according to tides and winds before disintegrating. The rare occurrence of such stains and their similarity with some biochemical events, make them difficult to detect even using artificial intelligence. The main issue lies in the labelling part that is not able to represent the wide variety of shapes and lengths on the one hand, and the regional characteristics of ocean background on the other hand. We propose a method that circumvents these two limitations. First, by creating synthetic label masks based on mathematical simulation of the puddle drifting possibilities. Second, by mapping these masks on targeted ocean surface background images. Using such database, an unsupervised learning with a strong data augmentation allows for the detection of many stains, drastically reducing the geographic areas to investigate. In practice, we recommend an active learning strategy where the first model is fine-tuned with minimal geospatial expert labeling done on the delimited zones where evidence have been found. We illustrate our solution on a real dataset where only few hours of human labeling provided similar results to 8 days of manual interpretation.

Keywords Synthetic label · Active learning · satellite images

1 Introduction

Geospatial remote sensing interest recently intensifies for a lot of different purposes like forest growing, fire prevention, gas leakage, flooding rescue, etc. Thanks to the launch of many new satellites that reduces the delay between consecutive data acquisition and, in the same time, that brings new types of sensors at higher resolutions. This results in a data explosion, making mandatory the use of automatic processes like artificial intelligence. Indeed, the revolution of deep learning techniques for images opened new opportunities for remote sensing applications, especially open-source models based on technologies like YOLO [10] or more recently SAM [7].

Our context concerns the oil seepage detection that forms stains on top of the ocean surface, visible from radar satellites [3]. Unfortunately, usual deep learning approaches suffer from limitations [1, 9]. First, there is a very large diversity in the combination of the oil slicks' shapes and the ocean surface textures. This makes quite complex the building of generic models. Second, there are extremely rare occurrences of such stains, with also some similarities with biochemical events. This slows the labelling task down so that this task represents the vast majority of time in a study.

In this paper we propose to tackle this issue by generating a representative synthetic database that can be used for unsupervised training and then in an active learning strategy. We illustrate our solution performance on a real dataset.

2 Synthetic database

Oil seepage corresponds to one or several continuous oil slicks on ocean surface that may come from two origins:

- human activities like boats cleaning out their tank or offshore industrial platforms that accidentally leak;
- natural earth crust leakage through small cracks, due to subsurface pressure pushing hydrocarbon fluids up.

Radar satellites are able to capture high-frequency deflection from the waves inside these puddles that differs from surrounding waves, allowing to see stains on black and white images (the amount of brightness in that zones being significantly reduced). The human origin case generally leads to isolated and quite simple shaped stains whereas the natural origin case gives form to repetitive (through time) and complex stains. Indeed, surface accumulation point continuously moves due to the ocean ascending process and in the same time, waves and winds effects spread the puddle over changing directions.

For a deep learning perspective, typically the supervised training of a generic model, one idea could be to apply strong data augmentation on manual labelled collection as an attempt to represent the variety of stains' shapes and lengths. Unfortunately, the amount of labelled data is quite limited because of two reasons:

- the extreme rarity of such texture appearances, similar to some meteorological, coastal or seaweed footprints;
- the ocean background diversity that makes even more difficult the data augmentation process.

To circumvent these limitations, we propose to build a synthetic database so as to enable unsupervised training. For that, we separately tackle the diversity issues by forming label masks of slicks on the one hand, and by mapping them on background of ocean textures on the other hand.

In details, we designed a mathematical simulator of oil puddle drifting over an ocean surface, using physical parameters of tides and winds:

- Means $\{\mu_w,\mu_t\}$ and standard deviations $\{\sigma_w,\sigma_t\}$ of wind and tide velocities (Gaussian distribution);
- Angles $\{(\alpha_{w,min}, \alpha_{w,max}),\ (\alpha_{t,min}, \alpha_{t,max})\}$ of wind and tide orientations (uniform distribution).

The resulting piecewise segments are then deformed to get slicks with morphological parameters:

- erosion and dilatation kernel sizes $\{N_e,N_d\}$.

We can then map the label masks into ocean surface tiles, chosen where there is statistically no stains. The blending is corrected with diffusion filter parameters:

- white noise deviation σ_n;
- Gaussian blur deviation σ_b;
- contrast coefficient γ.

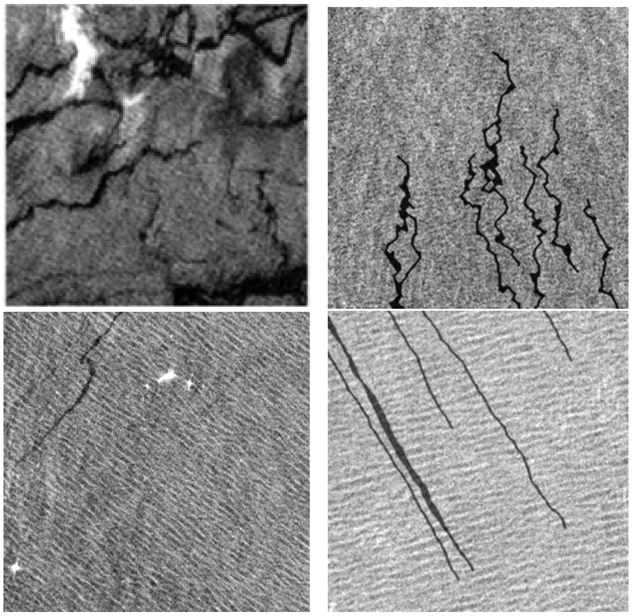

Figure 1: Two examples of real data (left) and two examples of synthetic label masks mixed with real ocean background (right), they correspond to parameters detailed in Table 1 (A on top, B on bottom).

In practice, all these parameters are randomly generated, according to their respective distributions and bounded by their realistic ranges. Figure 1 illustrate two different choices of parameters (listed Table 1), where low wind and high tide draw winding and squashed stains, on the contrary to high wind and low tide that draw long and streamlined stains. Thus, it is possible to create unlimited synthetic database, repeating this procedure, resumed on Algorithm 1.

Table 1: Complete list of parameters used to generate the two examples illustrated in Figure 1.

	μ_w	σ_w	μ_t	σ_t	$\sigma_{w,min}$	$\sigma_{w,max}$	$\sigma_{t,min}$	$\sigma_{t,max}$	σ_n	σ_b	γ	N_d	N_e
Example A	1	5	17	16	-3°	0°	100°	230°	2	1	14	13	9
Example B	7	5	16	1	3°	4°	306°	310°	12	2	5	12	8

Algorithm 1 Synthetic database generation

```
Require: $\mathcal{N}_s, \mathcal{N}_t, \delta_s$    ▷ number of slicks per tile, number of segment points per stick, spatial step for segments
/* Step 1 - create a label-only image (simulating puddle drifting) */
for $s = 1 : N_s$ do
    stickPoints = ones($N_t$)*randomUniform([0,0],imageDimensions)    ▷ Segment points' initialization
    {WindLengthes,TileLengthes} = randomNormal({$\mu_w$,$\mu_t$},{$\sigma_w$,$\sigma_t$},$N_t$)    ▷ Physical velocities
    {WindAngles,TileAngles} = randomUniform({$\sigma_{w,min}$,$\sigma_{t,min}$},{$\sigma_{w,max}$,$\sigma_{t,max}$},$N_t$)    ▷ Orientation angles
    {WindShifts,TileShifts} = {WindLengthes,TileLengthes}.*cos({WindAngles,TileAngles})
    stickPoints .+= cumulativeSum({WindShifts,TileShifts})    ▷ Segment points' displacements
    syntheticLabel = drawSegmentsInEmptyImage(stickPoints,$\delta_s$)    ▷ Segments drawing into piecewise lines
    syntheticLabel = erode(dilate(syntheticLabelImage,$N_d$),$N_e$)    ▷ Morphological segments' deformations
end for
/* Step 2 - combine label image with real background */
background = pickSatelitteOceanImage()    ▷ Choose an ocean tile (use statistics to be sure there in no stains)
syntheticLabel = syntheticLabel.*background/(randomUniform(0.5,1)*$\gamma$)    ▷ Randomize label contrast
synthetic = background .+(randomUniform(0,100)>$\sigma_n$)* syntheticLabel    ▷ Blend label with noise (i.e. misses)
synthetic = filterBlur(synthetic,$\sigma_b$)    ▷ Apply blur deviation
```

3 Active learning strategy

Using the synthetic database, we can then consider to train a deep learning model, without any need of human labelling. In practice it is sufficient to use a small U-net architecture [8], based on Adam optimizer [5], optionally extended with a Feature Pyramid Network (FPN [6]) that may improve low-frequency feature detection. One key to fully exploit the database is to add non-linear data augmentation, like grid distortion or elastic transformation, for instance using '*albumentation*' Python library [2].

This unsupervised training should be able to detect a wide variety of oil slicks but not all of them. Despite missing some stains, it highlights the areas of potential other evidence of natural seepage, significantly accelerating this research step over thousands of images. Carefully looking at these areas, the geospatial expert may find other sticks. From that it is possible to improve the model by transfer learning, meaning re-using the first model as starting point, and doing a fine-tuning based on the few new manual labels. This strategy, called active learning, is iterative and only requires minimal user work while checking the results.

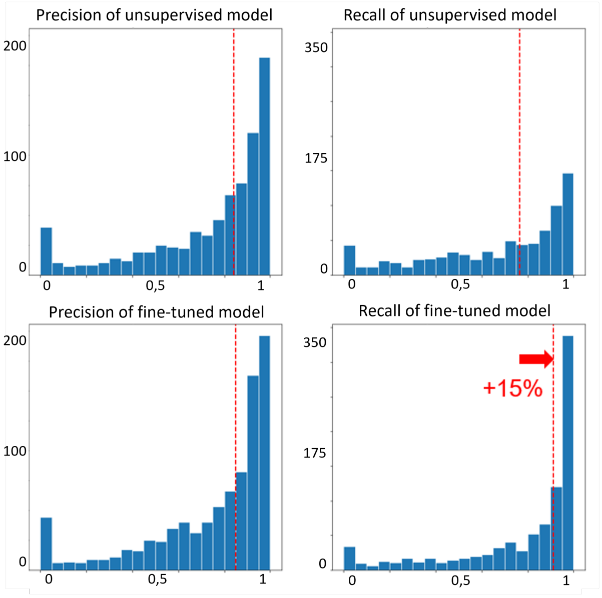

Figure 2: Comparison of metrics: precision (on left) and recall (on right), obtained using the unsupervised learning (on top) and the active learning (on bottom). Axes correspond to score (horizontal) and number of images (vertical). The red line correpond to the median score.

4 Real dataset example

We applied our solution on a real regional study where sticks have been manually labelled. The dataset is composed of 21 Sentinel-1 radar images, of size 25,500 × 16,800 pixels each. We split these images into 12,000 tiles of size 1024 × 1024 pixels each.

For the unsupervised dataset, we created a synthetic database, using ocean background tiles coming from this targeted area. We created 1200 tiles with synthetic labels and 300 without anyone. Indeed, we decided to also learn the local geochemical textures as not being oil slicks.

The neural network is a U-net of 5 layers depth with 48 filters. The unsupervised training was done over 800 epochs with a 10^{-4} learning rate and a batch size equals to 16. Using a Nvidia V100 graphic processor with 32GB of memory, this first training over 1500 tiles took almost a day and the inference on 12'000 tiles took an hour. We then fine-tuned the model through active learning with only 50 tiles coming from the labelled solution. This second training was done over 350 epochs with a 10^{-5} learning rate and a batch size equals to 2. On the same machine it took an hour and half.

When stacking the images over time (satellite goes back over every couple of days), the label masks make flower-shaped structure apparent. This is the case of natural leaks that come from the same cracks. The final goal is to find these crack positions. For that reason, the most important metric we will consider is the recall, that allows over-detection whereas precision metric penalizes false positives. Figure 2 displays these metrics for the unsupervised model and the fine-tuned model. In both cases the precision is quite good, meaning there is only few wrong prediction among the detected sticks. Regarding the recall, the unsupervised model score is average, so it was not able to find many sticks, but the fine-tuned model score is significantly high to provide usable results, as depicted Figure 3. Indeed, even if all the sticks are not detected, there are enough to catch the flower-shape structures.

Manual interpretation on this dataset took 8 days. Even with relatively limited computational power, our user workflow is more than five times faster.

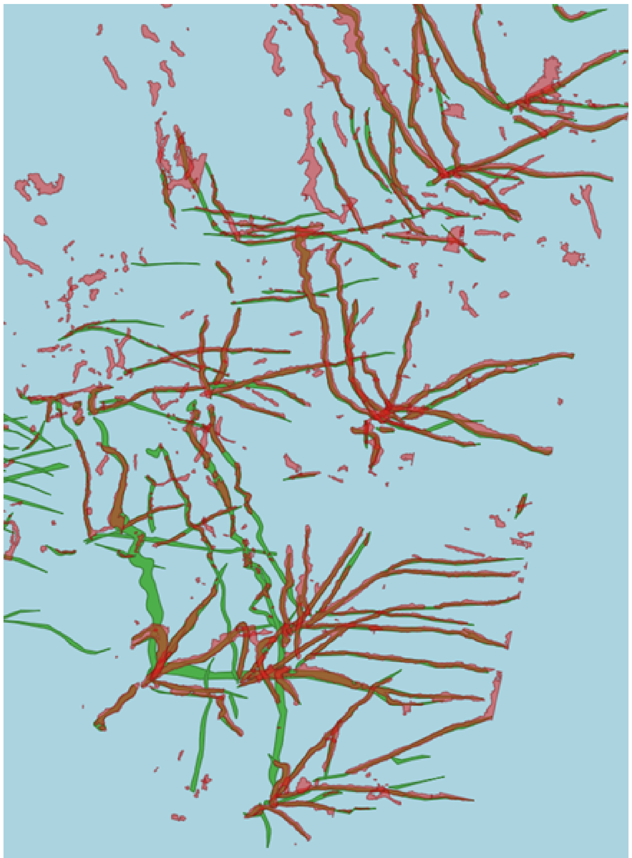

Figure 3: Comparison of oil slicks detection over a time period, from only manual interpretation (green) and active learning process (red).

5 Conclusion

In the deep learning era, neural networks have solved many vision problems when given enough data to encode a particular texture. In particular, satellite remote sensing is a field where artificial intelligence have achieved outstanding improvements.

This paper proposed a new approach for the delicate application of oil seepage detection that suffers from specific constraints. Indeed, the wide diversity of shapes and lengths, and the rarity of occurrences on huge images, prevent the building of supervised generic models. To tackle this issue, we developed an algorithm in two steps. First, a mathematical simulator of puddle drifting based on tide and wind effects that creates unlimited realistic label masks. Second, the blending of these masks on real ocean surface images, including the targeted regional area, that represents the diversity of ocean textures.

This synthetic database generator does not require human intervention and can quickly be used for an unsupervised learning in order to detect areas with evidence of oil seepage. Then, a transfer learning strategy using very few images from a fast and localized labelling done by geospatial experts is able to reproduce the full human interpretation results. Considering industrial computing resources, our method typically divides the total time of work by one order of magnitude.

The next step is to link this solution directly in professional geospatial softwares like QGIS, using Python script [4]. This will allow to continuously improve the synthetic database with new sticks shapes and ocean textures, and to collect small but highly-valuable expert labelling over regional studies. At some point, we will finally obtain a robust enough generic model for automatic oil slicks deep learning detection.

References

[1] Rami Al-Ruzouq, Mohamed Barakat A Gibril, Abdallah Shanableh, Abubakir Kais, Osman Hamed, Saeed Al-Mansoori, and Mohamad Ali Khalil. Sensors, features, and machine learning for oil spill detection and monitoring: A review. *Remote Sensing*, 12(20):3338, 2020.

[2] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[3] Fanny Girard-Ardhuin, Grégoire Mercier, and René Garello. Oil slick detection by sar imagery: potential and limitation. In *Oceans 2003. Celebrating the Past... Teaming Toward the Future (IEEE Cat. No. 03CH37492)*, volume 1, pages 164–169. IEEE, 2003.

[4] Anita Graser and Victor Olaya. Processing: A python framework for the seamless integration of geoprocessing tools in qgis. *ISPRS International Journal of Geo-Information*, 4(4):2219–2245, 2015.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[7] Lucas Prado Osco, Qiusheng Wu, Eduardo Lopes de Lemos, Wesley Nunes Gonçalves, Ana Paula Marques Ramos, Jonathan Li, and José Marcato Junior. The segment anything model (sam) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124:103540, 2023.

[8] 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*, pages 234–241. Springer, 2015.

[9] Mohamed Shaban, Reem Salim, Hadil Abu Khalifeh, Adel Khelifi, Ahmed Shalaby, Shady El-Mashad, Ali Mahmoud, Mohammed Ghazal, and Ayman El-Baz. A deep-learning framework for the detection of oil spills from sar data. *Sensors*, 21(7):2351, 2021.

[10] Zhihuan Wu, Xiangning Chen, Y Gao, and Yuntao Li. Rapid target detection in high resolution remote sensing images using yolo model. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:1915–1920, 2018.