

A comparative study of Deep Learning architectures for Classification of Natural and Human-made Sea Events in SAR images

William Ramirez¹ · Pedro Achanccaray¹ · Marco Aurelio Pacheco¹

Received: 13 December 2021 / Accepted: 13 January 2022

Published online: 02 February 2022

© The Author(s) 2022 [OPEN](#)

Abstract

Sea monitoring is essential for a better understanding of its dynamics and to measure the impact of human activities. In this context, remote sensing plays an important role by providing satellite imagery every day, even in critical climate conditions, for the detection of sea events with a potential risk to the environment. The present work proposes a comparative study of Deep Learning architectures for classification of natural and man-made sea events using SAR imagery. The evaluated architectures comprises models based on convolutional networks, inception blocks, and attention modules. Two datasets are employed for this purpose: the first one encompasses a series of natural events (geophysical phenomena), while the second describes a real oil spill scenario in the Gulf of Mexico from 2018 to 2021. As a result, through experimental analysis, it is demonstrated how the Xception and Deep Attention sampling architectures obtained the highest performance metrics, presenting Recall values of 94.2% and 87.4% for the classification of natural and human-made events, respectively.

Keywords Deep learning · Remote sensing · Maritime events classification · Sentinel-1A · Convolutional neural networks

1 Introduction

At sea, natural and human-made events are constantly happening. The former is mainly related to oceanic and atmospheric phenomena, while the latter essentially refers to human activities. In this context, sea monitoring is crucial for a better understanding of its dynamics and to determine the impact of these activities, especially the human-made ones. A clear example is the Gulf of Mexico oil spill, the largest marine oil spill in history, caused by the explosion on the Deepwater Horizon oil rig in 2010, and whose consequences are still there [1–3].

Remote sensing plays an important role for sea monitoring by providing high-resolution and day-and-night imagery, even in critical weather conditions, using Synthetic Aperture Radar (SAR) sensors, by virtue of its susceptibility to the sea surface irregularities [4–6]. Sea monitoring using remote sensing can be a very demanding task in terms of processing time and human resources. For this reason, Deep Learning (DL) has been employed for the automatic detection of these events during the last years.

Remote sensing based on SAR, is currently commonly used in the study of marine areas due to its high performance in large-scale projects. Considering its principle of operation, SAR can return a large amount of information on the horizontal and vertical axis of a large area, having a correct operation in areas with extreme climates (presence of clouds, smog,

✉ William Ramirez, williamramirez@uan.edu.co; Pedro Achanccaray, pachanccarayd@uni.pe; Marco Aurelio Pacheco, marco@ele.puc-rio.br
| ¹Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil.



storms, among others). Among its main characteristics, SAR doesn't show dependence on sources of natural external lighting, unlike optical systems that limit their operation to hourly lighting cycles on earth.

Many different approaches have been proposed for sea events classification using ML and SAR images. Tong et al. [7] proposed a pixel-wise classification approach using Random Forest trained upon polarimetric attributes (multiple polarization and different incidence angles) for oil slick segmentation. Zhou and Peng [8] employed a feature selection algorithm (mRMR), for oil spills identification, to reduce the redundancy between SAR attributes and to improve Support Vector Machine (SVM) processing time and generalization. In spite of their good results, these approaches need a profound knowledge about the application to select the appropriate features.

On the other hand, Deep Learning architectures overcome this by an end-to-end learning from raw data to extract higher-level features. Wang et al. [9] employed an Inception v3 architecture for classification of geophysical phenomena from Sentinel-1 vignettes; concluding that adjusting all the layers is much better than just training the final classifier layer. Zeng and Wang [10] introduced the Oil Spill Convolutional Network (OSNet), a VGG16 based architecture adapted for dark patches (sea surface covered with oil films) classification. OSNet achieved an accuracy of 94%, surpassing traditional ML classifiers like Area weighted Adaboost Multi-Layer Perceptron (AAMLN). Song et al. [11] proposed a CNN-based framework with three main steps: data process—represented by the coherency matrix, CNN feature extraction, and SVM classification—trained upon the fusion of PCA of high-level features of the last two layers. The replacement of the softmax classifier by a SVM improved the classification accuracy and enhance the ability to identify oil spills.

The remaining parts of this work are organized as follows: Sect. 2 introduces the deep learning architectures evaluated, Sect. 3 presents the study area of interest, Sect. 4 methodology employed for our experiments, and Sect. 5 the results obtained. Finally, Sect. 6 summarizes the conclusions drawn from all the experiments performed.

2 CNNs for classification

2.1 VGG16

Presented by Oxford University's Visual Geometry Group, the VGG [12] networks achieved high performance in the ILSVRC-2014 challenge, achieving an error rate of 7.3%. Figure 1 illustrates VGG's architecture, where its main highlights are the employment of only small kernels (3×3), rather than large-sized ones (5×5 or 7×7 , like in AlexNet), the same number of kernels inside a block, and doubling them after each max-pooling layer (7×7), and the use of scale jittering as data augmentation during training. The VGG16, where 16 is the number of weight layers, was selected in this work due to its better results against the other VGG's variants.

2.2 MobileNet v2 (MNet)

MobileNet v2 [13] is a neural network designed to be used in mobile applications, as its architecture has a few parameters and low complexity. As shown in the Fig. 2 MobileNet v2 uses the concept of inverted residual bottleneck layers, where important information in depth is returned to the output. This architecture contains 32 filters represented in the initial fully convolution layers, followed by 19 residual bottlenecks, where ReLU6 is used for non-linearity, having a standard kernel's size of 3×3 .

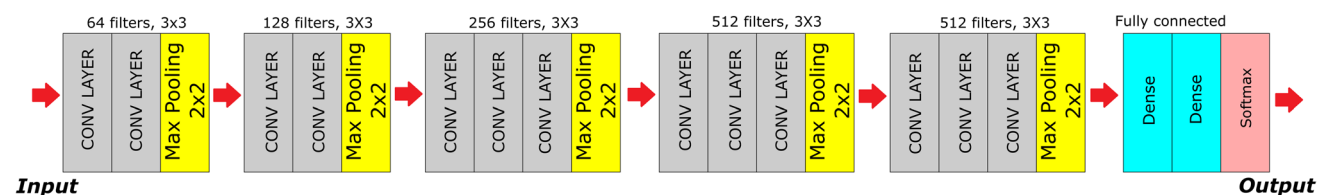
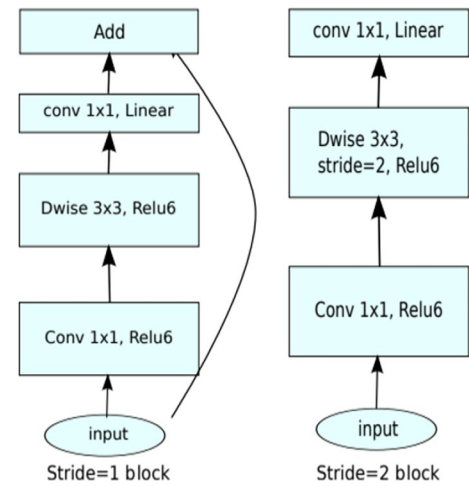


Fig. 1 VGG16 architecture

Fig. 2 MobileNet v2 architecture, obtained from [13]



2.3 Inception V3 (Incep)

Inception V3 [14] architecture, presented in 2016 by the University College London, expands the concept of Inception, improving the scores imposed by its previous versions in the ILSVRC 2012 challenge. It aims to combat the bottlenecks of computational cost, and achieve a deeper network with better computational efficiency by reducing the number of parameters. Inception V3 main highlights are the inclusion of smaller, asymmetric, and factorized convolutions, to reduce the number of parameters for faster training, auxiliary classifiers, which works as a regularizer, and more efficient grid size reductions, as shown in Fig. 3.

2.4 Xception (Xcep)

Based on Inception, Xception [15] proposes extreme blocks of Inception. Figure 4 shows how Xception takes advantage of the concept of separable convolutions (normal convolutions at the same level), where the convolution operation is divided into smaller convolutions located at the same level. This helps to have a computationally faster network. In Xception, 36 convolutional layers are structured into 14 modules, all with residual linear connections around them, except for the first and last modules, resulting in a total of 22.8 million parameters, approximately.

2.5 Inception ResNet v2 (Incep-R)

Inception ResNet v2 [16] is a convolutional neural network that belongs to the Inception family of architectures but incorporates residual connections, in order to have a deeper network. Figure 5 presents the proposed block, where the inclusion of the residual connection next to the scaled activation blocks is implemented. The architecture is focused on

Fig. 3 Original Inception module (A), and the Inception module proposed in Inception V3 (B), taken from [14]

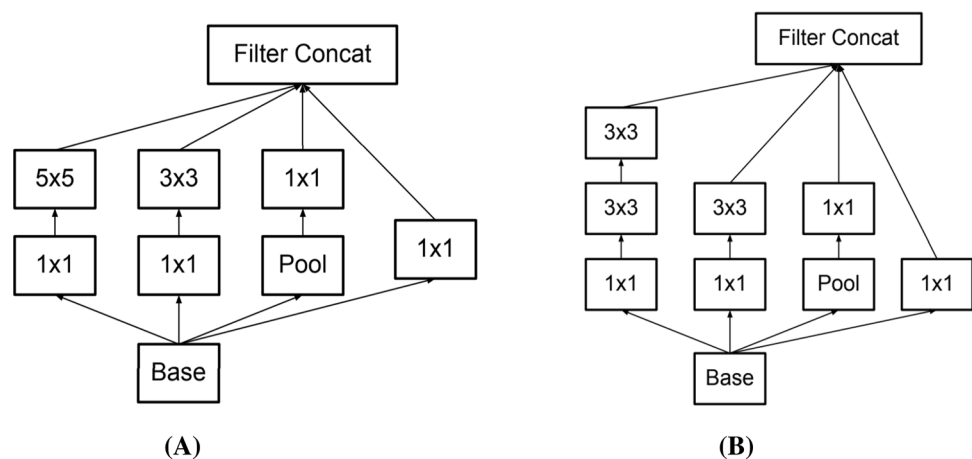


Fig. 4 Xception's separable convolutions, obtained from [15]

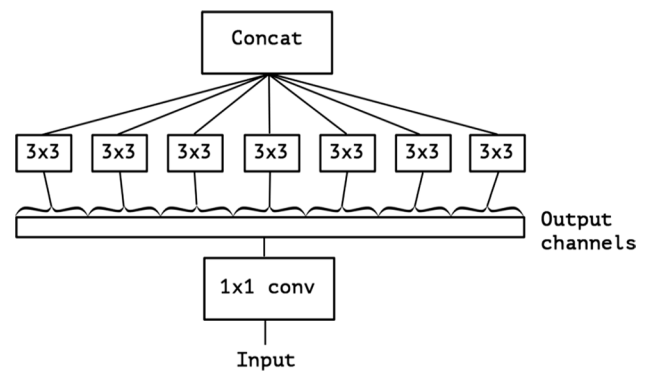
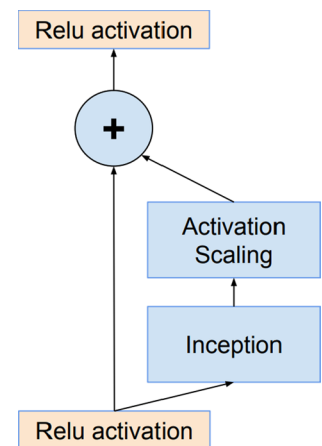


Fig. 5 Residual dimensioning using a scaling activation, obtained from [16]



have a deeper network while preventing the vanishing gradient (gradient death, where the global minimum cannot be reached).

2.6 Deep Attention Sampling (DAS)

The Deep Attention Sampling [17] architecture proposes a model able to process very large images without the need to re-sampling them. The model works with only a fraction of the whole image, where these locations are sampled from an attention distribution calculated from a lower resolution version of the original image (see Fig. 6). This sampling method aims to find the most informative regions of the input, and it is an optimal approximation, in terms of variance, of the model that processes the whole image.

3 Study area

3.1 TenGeoP-SARwv dataset

Through Sentinel-1A Data, this Dataset [5] presents ten events present around the Antarctic sea, Indian and Arabic oceans, providing a total of 37 K SAR vignettes of 547×490 pixels. As shown in Table 1, this dataset provides information related to 10 classes of great relevance during Radar mapping, which can be of great risk to humans.

3.2 Oil Spill dataset

The incident at Deepwater Horizon in 2010 established a radical change in the world and the industry related to this sector. Almost 3.19 million barrels were dumped into the sea, causing one of the largest catastrophes in the world. This incident continued to affect the environment over time, due to the magnitude of the incident [1, 3].

Fig. 6 Deep Attention Sampling methodology, obtained from [17]

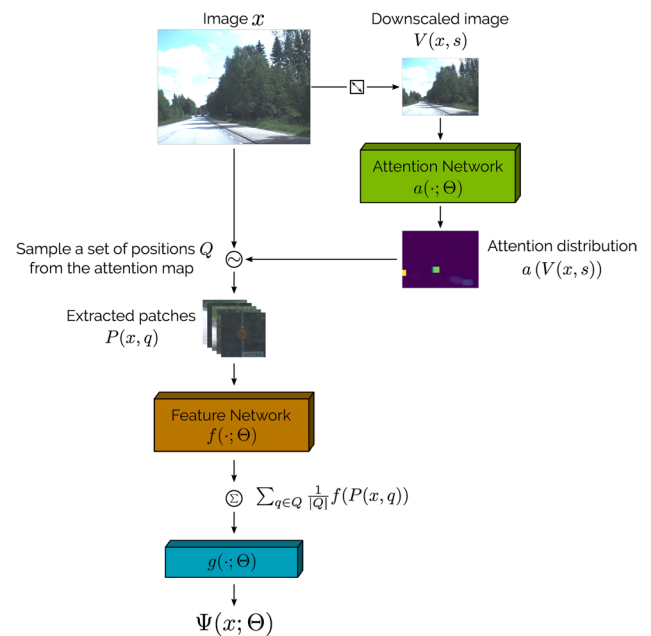


Table 1 Number of images per class in the TenGeoP-SARwv dataset

ID	Class	# Patches
0	Pure Ocean Waves	4900
1	Wind streaks	4797
2	Micro convective cells	4598
3	Rain cells	4738
4	Biological slicks	4709
5	Sea Ice	4369
6	Icebergs	1980
7	Low wind area	2159
8	Atmospheric front	4097
9	Oceanic front	1199

The Oil Spill dataset consists of Sentinel-1A scenes in VV polarization acquired, in the Copernicus Open Access Hub, from 2018 to 2020 in the Gulf of Mexico, as shown in Fig. 7 Each scene was radiometrically corrected, applied orbit and thermal noise filtering, and finally converted to dB [18]. The annotations were obtained through the reports of the National Oceanic and Atmospheric Administration—NOAA. Table 2 presents the acquisition dates of each scene, and the distribution of images for training (bold) and testing (normal).

4 Experimental protocol

The two aforementioned datasets were employed for image classification of sea events with the following architectures: VGG16, Inception V3 (Incep), Xception (Xcep), Inception ResNet v2 (Incep-R), MobileNet (MNet), and Deep Attention Sampling (DAS). For the selection of the architectures, recent concepts around the classification of SAR images were considered [9, 10]. Concepts focused on attention for sampling, Inception and MobileNets were studied seeking to direct the approach to the optimization of computational resources, deep and multiscale extraction of attributes.

For the TenGeoP-SARwv dataset, three sets were generated randomly for training, validation, and testing, with the following distribution: 70%, 20%, and 10%, respectively. Each vignette was resized to a size of 512×512 for training the models. Furthermore, in the Oil Spill dataset, as the annotations were for semantic segmentation, it was necessary to convert them for image classification. Then, a patch was assigned to the class Oil Spill, if at least 8% of the pixels belonged to this class. This

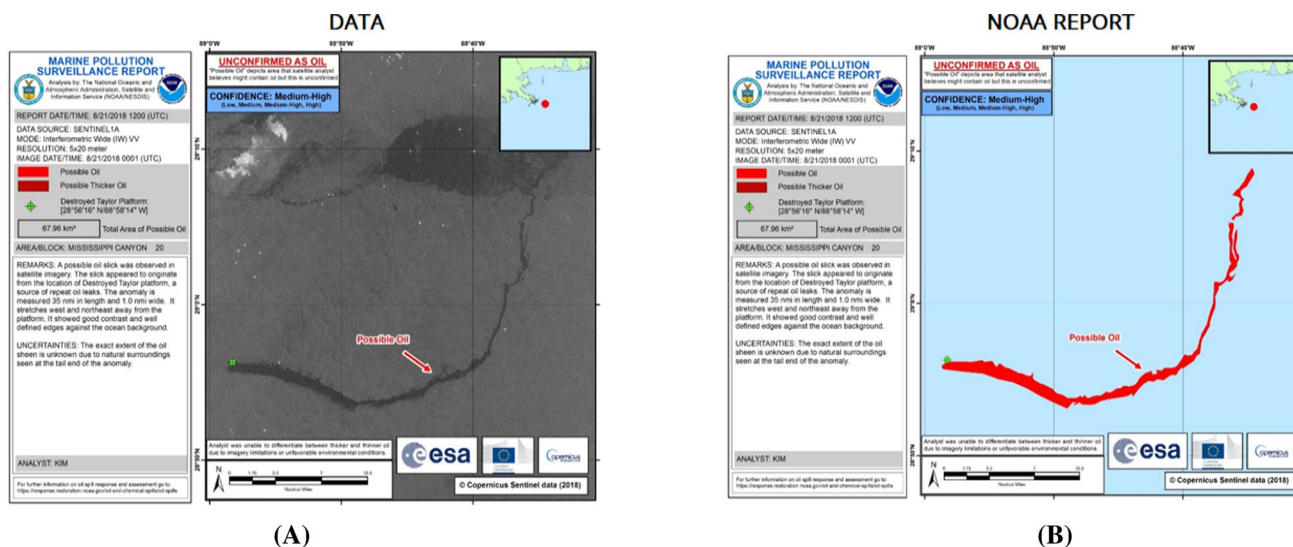


Fig. 7 NOAA's report, where the oil slicks are presented in the original SAR image (A), and its corresponding polygon (B).

Table 2 Acquisition dates for each scene in the Oil Spill dataset. Training images are in bold

Data	Size	Data	Size
21/08/2018	5800 × 4600	19/12/2018	5300 × 2500
14/09/2018	3800 × 3200		4700 × 2200
26/09/2018	5200 × 2900		2500 × 2000
13/11/2018	4600 × 2600		3900 × 2400
07/12/2018	4900 × 2600		3600 × 2600
	4900 × 2300		2900 × 3600
31/12/2018	5700 × 3000	7/03/2020	3900 × 2400
16/08/2019	3900 × 2000	19/03/2020	3800 × 2100
08/09/2019	4100 × 1900		4100 × 2000
15/10/2019	5100 × 2100	31/03/2020	3700 × 2100
24/02/2020	4400 × 2200	22/08/2020	3600 × 2300
	4000 × 2100		

threshold was selected empirically after different experiments, seeking to have an appropriate context. The distribution of images for training, validation, and testing was the same as in the previous dataset, where the proportions were computed at patch level extracted with sizes of 256 × 256 and 512 × 512.

All models were implemented using Keras with TensorFlow as backend. A standard normalization was employed to pre-process the input images. Then, each model was trained for a maximum of 100 epochs, with Adam optimizer (momentum = 0.9), a learning rate of 0.001, and an early stop to break after 6 epochs without improvement in the F1-score of the validation set.

Additionally, data augmentation was performed to increase the models robustness with vertical and horizontal flips, and zoom in of 2–3%. A fine-tuning of each model was performed with each dataset, where pre-trained weights with ImageNet were applied. For the DAS network, a VGG16 was employed as feature extractor, the scale for the initial downsampling was 20%, and the size of the patches for the attention sampling is 15% of the input size.

The assessment of each architecture is performed in two ways: quantitatively and qualitatively. For the quantitative analysis, each experiment was carried out 50 times, and the mean and standard deviation of the Precision, Recall, and F1-score were computed.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

where TP is the true positives samples of the class of interest correctly classified, and FP is the false positives samples that do not belong to the class but are classified as part of that class. Finally, FN represents the false negatives samples that belong to the class but were miss-classified.

We focused on the Recall values to compare the models as it is more important to reduce the number of false negatives, especially in the Oil Spill dataset, where it is critical to miss Oil samples. On the other hand, the qualitative evaluation is performed based on the generated Class Activation Maps (CAM), which indicates the regions that were more relevant in the classification of the evaluated image. In this work, the Gradient-weighted Class Activation Mapping (Grad-CAM) [19] are generated for each model.

5 Results

5.1 TenGeoP-SARwv dataset

Figure 8 shows the evolution of the F1-score during the training of each model, where we can see that after 40 epochs all models have converged. Each curve represents the mean of all F1-score along with the 50 experiments for each model.

Inception ResNet, Xception and MobileNet architectures achieved the highest F1-score values: 0.91, 0.90 and 0.89, respectively. Then, DAS and Inception v3 obtained values of 0.85 and 0.84, while VGG16 reached the lowest F1-score value: 0.73. Table 3 summarizes the performance of each model in terms of mean and standard deviation of Recall in the test set after running each experiment 50 times.

Xception and MobileNet architectures achieved the highest Recall values for different classes. On the one hand, Xception got the best results in classes that are defined by almost the whole image such as 0 (Pure Ocean Waves), 1 (Wind Streaks), 2 (Micro Convective Cells), 7 (Low Wind Area), and 8(Oceanic Front). On the other hand, MobileNet obtained better results in classes that are defined by a fraction of the image like 3 (Rain Cells), 4 (Biological Slicks), 5 (Sea Ice), 6 (Icebergs), and 9 (Oceanic Front).

Moreover, we can see a significant difference between VGG16 and DAS, where a VGG16 is employed as a feature extractor trained upon samples provided by an attention distribution. The addition of the attention sampling brought Recall gains of up to 8%, demonstrating the benefits of using this kind of sampling. Similarly, the usage of many Inception blocks by Xception produced Recall improvements of up to 5% compared to Inception.

In order to verify that the models have learnt the regions of the image that best characterize each class, the Grad-CAMs were generated. Due to the nature of each of the geophysical events, a variation can be found with respect to attributes related to periodicity, impact area, and density in the water. In this way, our inference will be able to indicate to us,

Fig. 8 Evolution of the F1-score during training of each model in the TenGeoP-SARwv dataset. Each curve represents the average of the F1-score along with the 50 experiments

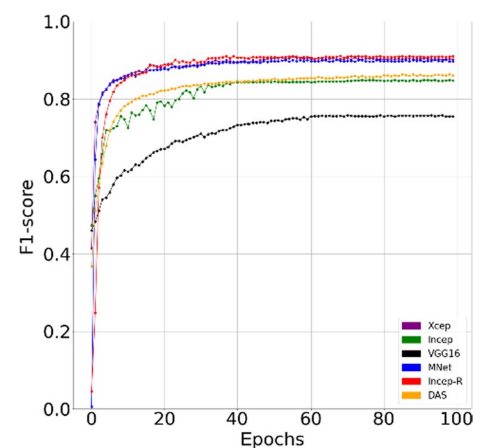


Table 3 Mean and Standard Deviation of Recall as a percentage for each model obtained in the test set after running each experiment 50 times with the TenGeoP-SARwv dataset

ID	Incep	Xcep	Incep-R	MNet	VGG16	DAS
0	98.6±0.3	99.2±0.2	98.4±0.9	95.0±0.6	98.5±0.5	98.6±0.4
1	96.2±0.5	98.3±0.3	96.8±0.6	90.0±1.1	95.3±0.9	97.3±0.5
2	98.7±0.6	99.1±0.5	98.9±0.5	98.5±0.8	96.0±0.8	98.8±0.4
3	97.1±0.8	89.9±0.6	94.2±0.9	98.6±0.9	95.4±1.2	97.2±0.7
4	92.9±0.7	98.3±0.3	98.5±0.8	98.5±0.7	93.5±0.8	98.7±0.5
5	95.0±0.7	96.2±0.5	89.8±0.6	98.2±0.9	94.2±0.7	97.4±0.4
6	96.6±0.9	91.0±0.7	94.5±1.0	97.9±1.2	95.2±1.0	96.7±0.8
7	94.7±0.8	97.0±0.5	93.2±0.9	96.4±1.3	92.3±0.8	94.9±0.6
8	93.8±0.5	94.6±0.4	92.5±0.7	97.5±0.4	96.6±0.9	92.7±0.7
9	88.3±1.0	92.1±0.8	84.7±1.0	86.8±1.1	75.9±1.1	83.6±0.9
Total	84.5±0.7	94.2±0.6	92.5±0.8	93.9±1.0	87.9±0.8	93.4±0.7

through the GradCAM [19], which aspects the model considers important during the classification. Figure 9 illustrates the Grad-CAM obtained by all architectures for each class, where the most intense tonality close to the red color will imply a higher value in the activation, and the lowest tonality close to the blue color, will imply activation values close to 0.

VGG16 and DAS produced activation maps closer to the definition of classes 0 (Pure Ocean Waves), 1 (Wind Streaks), 2 (Micro Convective cells), and 7 (Low Wind Areas). Notice that these classes are characterized by having a presence of the class spread over the whole scene, with small patterns with a high periodicity, completely covering each vignette.

Regarding class 6 (Icerbergs), where this class is represented by small regions with ice in the scene, DAS achieved the most accurate classification maps, focusing only on those regions. This demonstrates the benefits of learning by attention sampling. Classes 8 (Atmospheric Front) and 9 (Oceanic Front) were more challenging due to the absence of a clear definition of the presence of these classes in the scene. They are mainly defined by the transition between two masses of air or water at different conditions.

For Class 5 (Sea Ice), Xception showed a high performance, having activation maps in brighter areas with the presence of cracks on the surface. Meanwhile, in classes 3 (Rain Cells) and 4 (Biological Slicks), DAS produced activation maps closer to each class presence in the scene. This effect can be observed in the activation map for class 4 (Biological Slicks), where small lines are correctly represented and detected.

As shown in Fig. 9, only Xception managed to obtain a more detailed activation map describing this transition, observing a clear dependency with the contextual information to the definition of this class. In summary, Xception achieved the best performance due to the deep extraction of attributes raised in this architecture, demonstrating high performance in scenes with a high presence of the class and robust feature extraction, being able to distinguish classes with high variability in terms of texture and intensity. Moreover, Xception obtained the highest accuracy in terms of Recall, 94%.

5.2 Oil Spill dataset

Similar to the previous section, Fig. 10 shows the evolution of the F1-score during the training of each model, where each curve represents the mean of along the 50 experiments. Xception and DAS architectures achieved the highest F1-score values: 0.95 and 0.96, respectively.

Table 4 outlines the results obtained by each model in terms of mean and standard deviation of Recall in the test set after running each experiment 50 times. Notice the 0 represents the class Non-Oil, while 1 the class Oil. DAS reached the highest Recall for the class 1 (Oil), with a value of 87.4%, followed by Xception and Inception, with 80.1% and 79.9%, respectively. Analogous to the previous section, the inclusion of attention sampling to train a VGG16 increases the Recall values by 22.5%, considerably reducing the number of false negatives.

A similar behaviour can be seen in Fig. 11 where the inference results of Xception, VGG16 and DAS are presented with the original image, and its corresponding references at pixel- and patch-wise levels. While VGG16 got many false positives, confusing Oil with different look-alikes (dark spots), Xception and DAS managed to get just a few of them. It is noteworthy to mention that Oil Spill classification is a challenging task due to the similarities, in terms of back-scattering response, geometry and texture, between oil slicks and other kind of natural sea events. In addition, DAS obtained less false negatives than Xception.

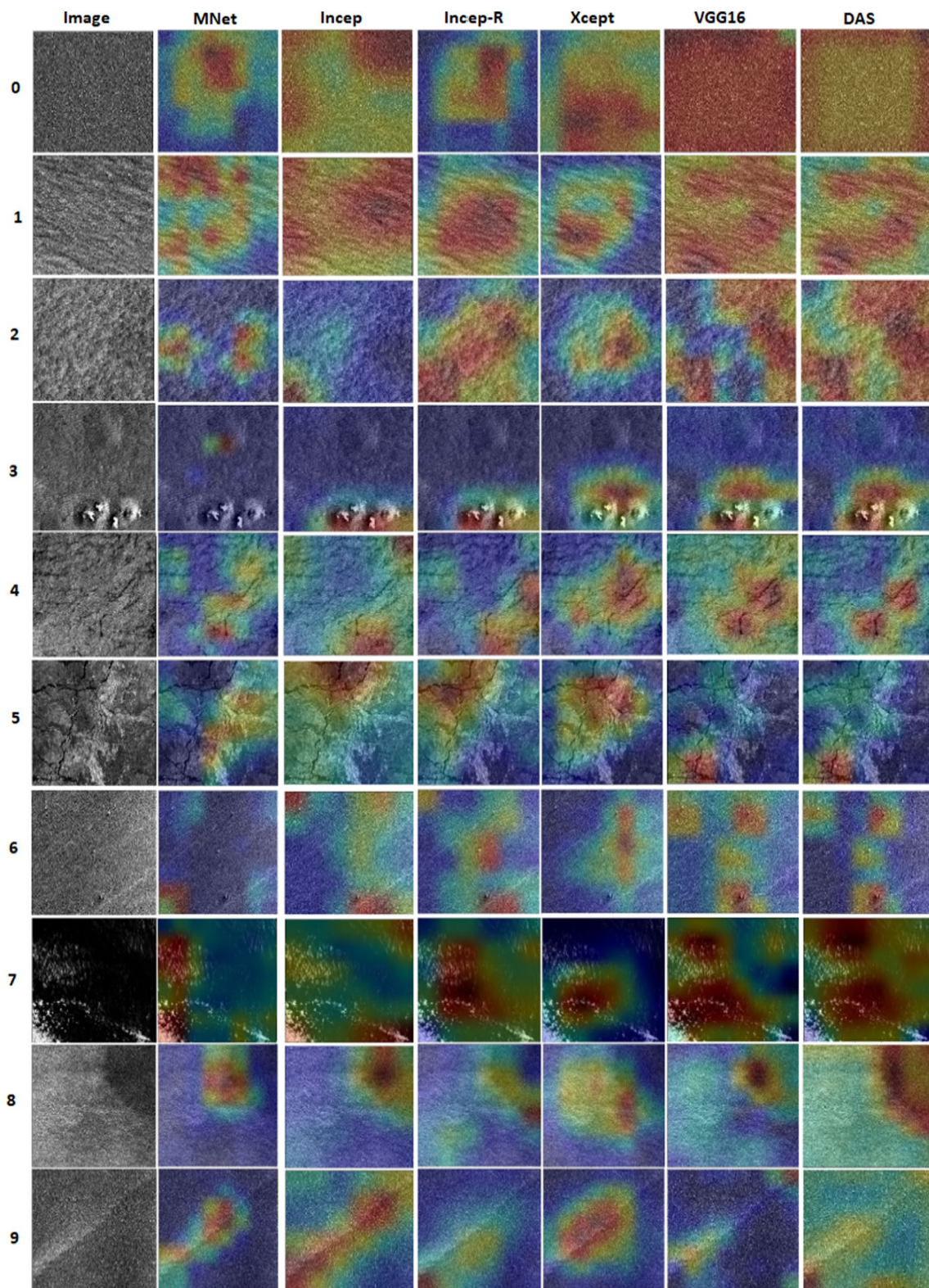


Fig. 9 Grad-CAMs produced by MobileNet V2, Inception V3, Inception-ResNet, Xception, VGG16 and Deep Attention Sampling in the TenGeoP-SARwv dataset for classes 0 (Pure Ocean Waves), 1 (Wind Streaks), 2 (Micro Convective Cells), 3 (Rain Cells), 4 (Biological Slicks), 5 (Sea Ice), 6 (Icebergs), 7 (Low Wind Area), 8 (Atmospheric Front) and 9 (Oceanic Front)

Fig. 10 Evolution of the F1-score during training of each model in the Oil Spill dataset. Each curve represents the average of the F1-score along the 50 experiments

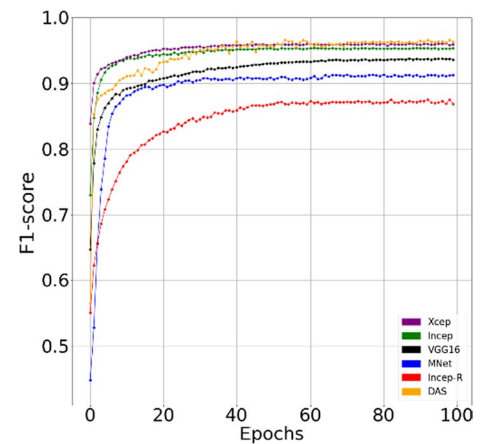
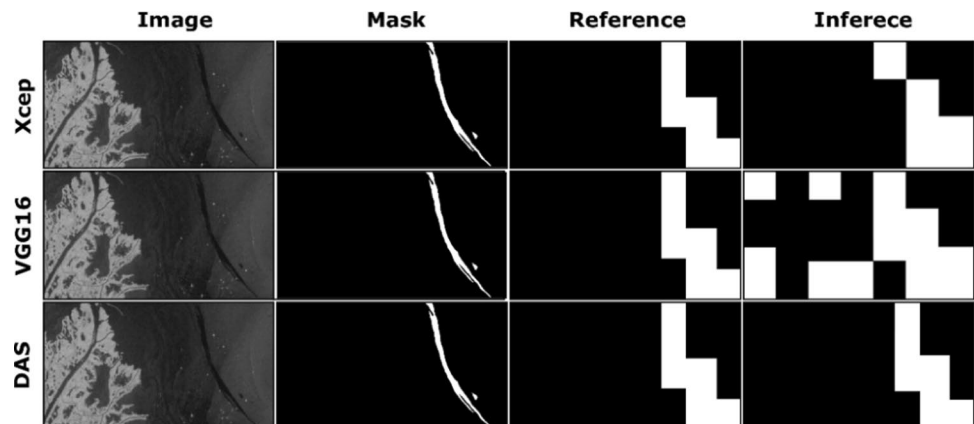


Table 4 Mean and Standard Deviation of Recall as percentage for each model obtained in the test set after running each experiment 50 times with the Oil Spill dataset

ID	Incep	Xcep	Incep-R	MNet	VGG16	DAS
0	92.3 ± 0.8	94.5 ± 0.5	95.6 ± 0.9	95.2 ± 1.1	96.4 ± 0.8	93.1 ± 0.6
1	79.9 ± 1.0	80.1 ± 0.8	68.2 ± 1.6	74.5 ± 1.3	64.9 ± 1.1	87.4 ± 0.7

Fig. 11 Inference results using the Xception, VGG16 and DAS models with an image from the Oil Spill dataset. From left to right, the original SAR image (VV polarization, 19/12/2018)



In Fig. 12, it can be observed how DAS achieved a better performance than other architectures. In this scenario, it's observed how smaller oil slicks manage to be correctly classified even in the presence of pixels with a back-scattering value close to the Oil pixels. In this way, DAS demonstrated high performance in scenarios with high variability in back-scattering values, mainly in the presence of low wind areas.

Analyzing the class activation maps using the Grad-CAMs produced by DAS architecture (see Fig. 13) we can see how the model focused on the dark pixels of each scene. The intensity expresses the activation values in each pixel, higher intensity represents a higher value. Moreover, it is worth noting how DAS was able to detect Oil Spill patches in the presence of other sea events like low wind areas or atmospheric fronts (see bottom image in Fig. 13) demonstrating its robustness against look-alikes.

In the oil spill classification, DAS showed a high performance in dark patches, clearly differentiating geophysical events, as well as, the presence of oil spill with just a few pixels in the scene. Figure 14 presents a scenario where there is little presence of oil spill, having a larger scene with the presence of other areas with a back-scattering value close to Oil.

Fig. 12 Inference results using the Xception, VGG16 and DAS models with an image from the Oil Spill dataset. From left to right, the original SAR image (VV polarization, 19/03/2018)

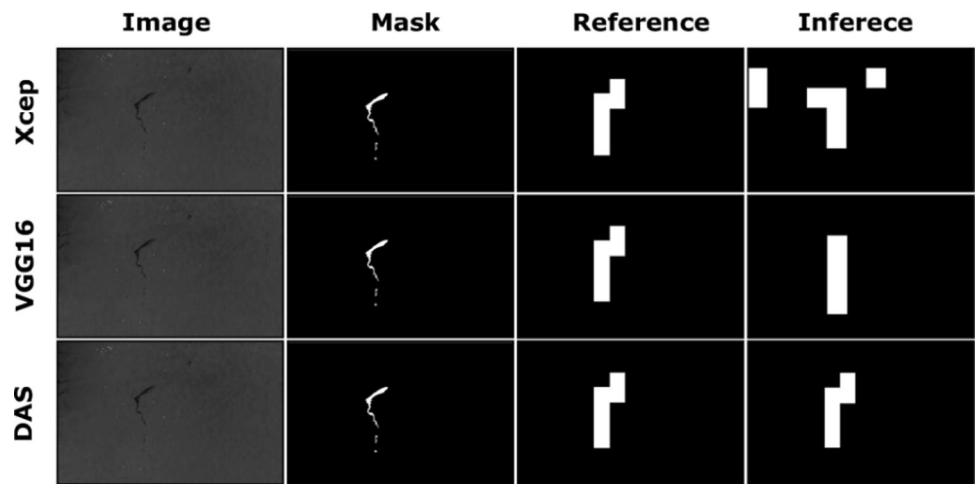


Fig. 13 Grad-CAMs produced by DAS with images from the Oil Spill dataset (19/12/2018) for class Oil. Only patches predicted by the model as belonging to this class are presented

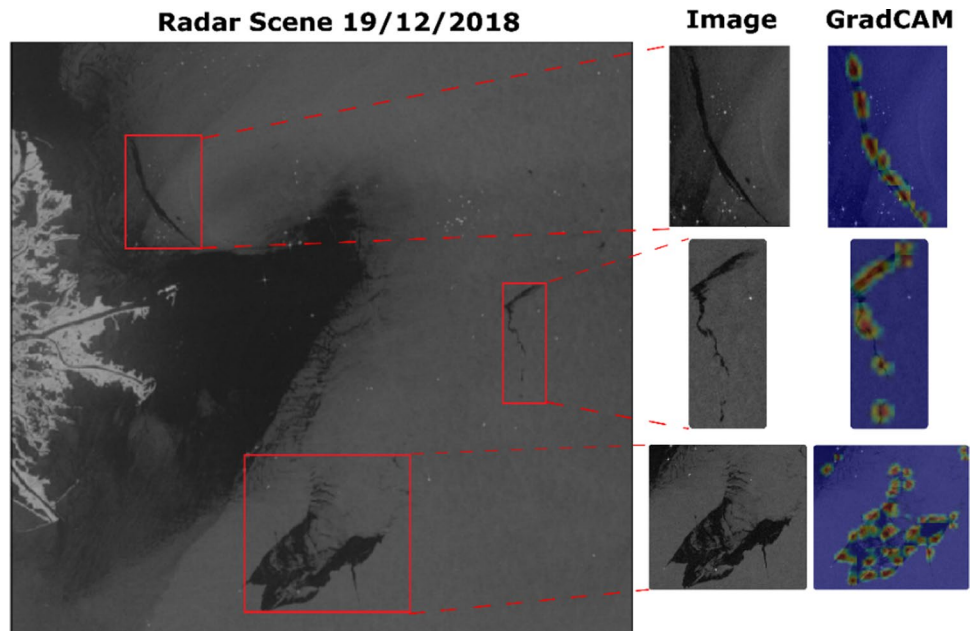
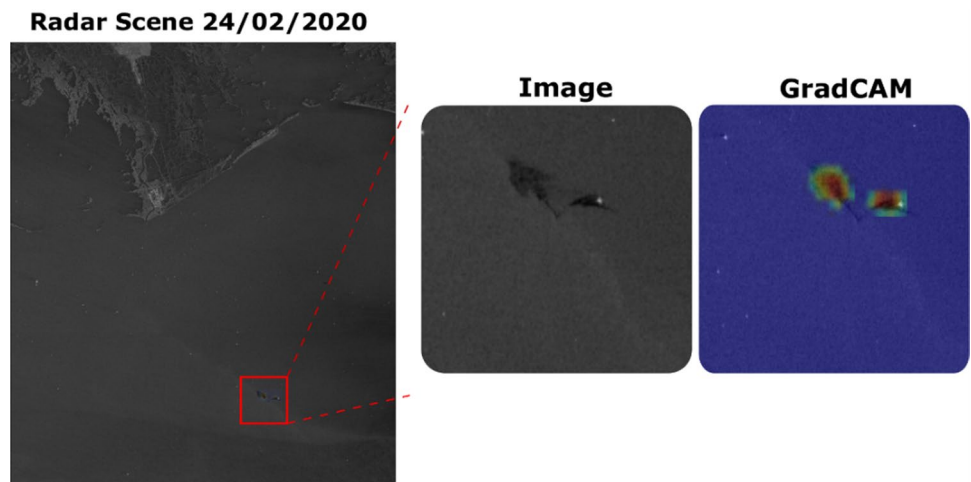


Fig. 14 Grad-CAMs produced by DAS with images from the Oil Spill dataset (24/02/2020) for class Oil. Only the patches predicted by the model as belonging to this class are presented



6 Conclusion

A comparative study of Deep Learning architectures for image classification in SAR images has been performed. Two datasets were employed, with natural geophysical phenomena (TenGeoP-SARwv) and human-made (Oil Spill) sea events. The comparison has been carried out considering the performance of each model during training, as well as its Recall values in the test set. Additionally, the class activation maps, using Grad-CAM, were analyzed to confirm that each model has learnt the regions of the image that best characterize each class.

Xception achieved the highest Recall for the natural events, with a value of 94.2% in the test set. Meanwhile, Deep Attention Sampling (DAS) obtained the best Recall for the oil class, with a value of 87.4%. The inclusion of attention sampling by DAS, to train a VGG16 upon the most representative regions of the image, was reflected in Recall gains of up to 8% and 22.5%, for the TenGeoP-SARwv and Oil Spill datasets, respectively. Furthermore, based on the Grad-CAMs of DAS and VGG16, the former generated better activation maps in terms of precision to delineate the most representative regions for each class.

As an extension of this work, it is planned to consider different architectures such as those based on dense blocks, recurrent networks or transformers, with/without attention sampling, for classification and semantic segmentation. Moreover, the analysis of architectures for semantic segmentation will be considered for a more precise detection of these events. Architectures based on Fully Convolutional Neural (FCNs) networks will be assessed, such as U-Net or DeepLab. This work is fully reproducible as the source code is publicly available in our repository (https://github.com/williamalbert94/Oil_spill_classification).

Acknowledgements The authors would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) and Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) for their financial support.

Authors' contributions Conceptualization, Formal analysis, Methodology and Software WR and P.D; Supervision P.D. and M.A.P.; Writing WR and P.D.; All authors have read and agreed to the published version of the manuscript. All authors read and approved the final manuscript.

Data availability Repository: This dataset contains sentinel-1A GRD pre-processed images together with their corresponding masks for use in applications related to machine learning and deep learning. <https://doi.org/10.5281/zenodo.4672426>. This project contains the following underlying data: Images. (Sentinel 1-A pre-processed image); Mask. (Oil spill annotations for each image); CSV distribution. (Distribution paths for training stage). Data are available under the terms of the Creative Commons Attribution 4.0 International.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Klapp J, editor. Experimental and theoretical advances in fluid dynamics. Berlin Heidelberg: Springer; 2012.
2. Rabalais NN, Smith LM, Eugene-Turner R. The Deepwater Horizon oil spill and Gulf of Mexico shelf hypoxia. *Continental Shelf Res.* 2018;152:98–107.
3. Beyer J. Environmental effects of the Deepwater Horizon oil spill: a review. *Mar Pollut Bull.* 2016;110(1):28–51.
4. Wang X. Comparison of Arctic sea ice thickness from satellites, aircraft, and PIOMAS data. *Remote Sens.* 2016;8(9):713.
5. Wang, C. Labeled SAR imagery dataset of ten geophysical phenomena from Sentinel-1 wave mode (TenGeoP-SARwv). 2018.
6. Mahmud MS. Seasonal evolution of L-band SAR backscatter over land fast Arctic sea ice. *Remote Sens Environ.* 2020;251:112049.
7. Tong S. Multi-feature based ocean oil spill detection for polarimetric SAR data using random forest and the self-similarity parameter. *Remote Sens.* 2019;11(4):451.
8. Zhou H, Peng C. Oil spills identification in SAR image using mRMR and SVM model. In: 2018 5th International Conference on Information Science and Control Engineering (ICISCE). IEEE; 2018.
9. Wang C. Classification of the global Sentinel-1 SAR vignettes for ocean surface process studies. *Remote Sens Environ.* 2019;234:111457.

10. Zeng K, Wang Y. A deep convolutional neural network for oil spill detection from spaceborne SAR images. *Remote Sens.* 2020;12(6):1015.
11. Song D, et al. A novel marine oil spillage identification scheme based on convolution neural network feature extraction from fully polarimetric SAR imagery. *IEEE Access.* 2020;8:59801–20.
12. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y, editors. 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings; 2015.
13. Sandler M, Howard A, Zhu M. Mobilenetv2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF conference on computer vision and pattern recognition; 2018, p. 4510–20. <https://doi.org/10.1109/CVPR.2018.00474>.
14. Szegedy C, Vanhoucke V, Ioffe S. Rethinking the inception architecture for computer vision. In 2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016, p. 2818–26; <https://doi.org/10.1109/CVPR.2016.308>.
15. Chollet F. Xception: deep learning with depthwise separable convolutions. *IEEE Conf Comput Vis Pattern Recognit.* 2017;2017:1251–8.
16. Szegedy C, Ioffe S, Vanhoucke V. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2017, vol. 31, no. 1.
17. Katharopoulos A, Fleuret F. Processing megapixel images with deep attention-sampling models. In: Proceedings of the International Conference on Machine Learning (ICML); 2019.
18. Filipponi F. Sentinel-1 GRD preprocessing workflow, 3rd international electronic conference on remote sensing. Proceedings. 2019. <https://doi.org/10.3390/ECRS-3-06201>.
19. RR Selvaraju, M Cogswell, A Das Grad-CAM. Visual explanations from deepnetworks via gradient-based localization. In: 2017 IEEE international conference on computer vision (ICCV). 2017, p. 618–626. <https://doi.org/10.1109/ICCV.2017.74>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.