# Explainability of Image Semantic Segmentation Through SHAP Values

Pierre Dardouillet, Alexandre Benoit, Emna Amri, Philippe Bolon, Dominique Dubucq, Anthony Crédoz

# Explainability of Image Semantic Segmentation Through SHAP Values.*

Pierre Dardouillet[1][0000−0002−3911−3446], Alexandre Benoit[1][0000−0002−0627−4948], Emna Amri[1,2][0000−0003−3173−1642], Philippe Bolon[1], Dominique Dubucq[2][0000−0002−1391−7606], and Anthony Credoz[2][0000−0002−2264−9025]

[1] LISTIC, Polytech Annecy-Chambery, University of Savoie Mont-Blanc, B.P. 80439, 74944 Annecy le Vieux Cedex France.
{firstname.lastname}@univ-smb.com
[2] TotalEnergies S.E, Avenue Larribau, F-64018, France.
{firstname.lastname}@totalenergies.com

**Abstract.** The introduction of Deep Neural Networks in high-level applications is significantly increasing. However, the understanding of such model decisions by humans is not straightforward and may limit their use for critical applications. In order to address this issue, recent research work has introduced explanation methods, typically for classification and captioning. Nevertheless, for some tasks, explainability methods need to be developed. This includes image segmentation that is an essential component for many high-level applications. In this paper, we propose a general workflow allowing for the adaptation of a state of the art explainability methods, especially SHAP, to image segmentation tasks. The approach allows for explanation of single pixels as well image areas. We show the relevance of the approach on a critical application such as oil slick pollution detection on the sea surface. We also show the applicability of the method on a more standard multimedia domain semantic segmentation task. The conducted experiments highlight the relevant features on which the models derive their local results and help identify general model behaviours.

**Keywords:** Model Explainability · Image Segmentation · Shapley Values · SAR Images

## 1 Introduction

Artificial intelligence (AI) models are increasingly used for many applications, as they have demonstrated their potential to solve complex tasks previously performed by humans. However, their high performance comes at a cost: AI models are often very complex and their decision processes cannot be clearly understood by humans, which impacts on their reliability and acceptability. To

* This work was supported by TotalEnergies company and also relied on HPC resources from GENCI–IDRIS (Grant 2021-AD011011418R1).

date, this major drawback is one of the obstacles in AI subfields such as Deep Learning. Thus, for tasks where confidence in the results obtained is as important as the results, the use of AI models is compromised. Then, the **eXplainable Artificial Intelligence** (XAI) field has been subject of growing interest and already gathers a multitude of methods designed to open those black boxes.

This paper focuses on image segmentation tasks for which explainability methods are for now limited. Image semantic segmentation is widely used as a preliminary process for various image types and applications, such as radar images for remote sensing, multimedia images for automatic driving, medical images for health diagnosis, and so on. This semantic segmentation task is complex and is nowadays addressed by deep neural networks. As a base application component, associated explanation methods become mandatory. However, few works have been dedicated to the understanding of such segmentation model decisions [7]. The main issue is related to the complexity of the explanation method since one expects any pixel or region-level decision to be explained with respect to the entire input image and maybe some metadata. In addition, it is thus required to provide relevant explanation in a timely manner.

To address this challenge, we propose an adaptation of an explainability method, called SHapley Additive exPlanations (SHAP). This method represents one of the most widely used post-hoc explainability methods [7] but its adaptation to semantic segmentation is not straightforward. Our approach can consider any type of image as input. It can identify features that inhibit or excite a model decision, i.e. negative or positive contributions to the decision. The resulting explanations are consistent with human intuition to the extent that they are built on Shapley values [13]. We base our approach on an agnostic implementation of SHAP, called Kernel SHAP [8], which we refer to hereafter as SHAP.

In this paper, two application domains are considered for experiments. We first focus on offshore oil slick detection illustrated in Fig. 1 for which we explain the predictions provided by a state-of-the-art semantic segmentation model proposed in [2]. This represents a typical critical application for environmental pollution monitoring for which detection results can induce strong and costly actions. In this context, oil slicks are generally detected from Synthetic Aperture Radar (SAR) images from which they appear as dark spots on the sea surface as shown in the left image of Fig. 1. Current detection methods rely on SAR analysis and is performed by photo-interpreters or automatically by deep neural network models [2]. In this context, automatic detection must be explained to decision makers. Our proposal is then to provide comprehensible explanations as coloured maps highlighting the input image areas that contributed to the model decision for a selected pixel or region. As illustrated in Fig. 1, the good detection related to the red region (no oil, left image) is explained on the right image. One observes that the local area close and within the region of interest contribute negatively to classification as oil slicks (red colours) while the dark neighbouring slick areas provide a positive influence (green colours). Then, the sum of these contributions yields the oil detection probability, here close to zero that explains classification as a sea area.

Finally, in order to show the applicability of the method to other domains, a second experimental case study is proposed. We consider semantic segmentation in urban scene from RGB images relying on the CityScapes dataset [4] and a state-of-the-art model, HardNet-MSeg [5]. More specifically, we show the interest of the method to explain the competition between probable classes in different situations.


Fig. 1: An input Synthetic Aperture Radar image (left) is processed by a model for oil slick detection (centre). In order to explain prediction for a given region (red polygon on the left), the proposed SHAP based method provides a coloured map showing each image area's contribution (right).

The article is organized as follows: first, a state of the art in post-hoc methods for model applicability in machine learning is presented. Then, a general framework for the adaptation of occlusion based explanation methods to semantic segmentation is presented. We then integrate the kernel SHAP method as well as RISE [10] as a comparison baseline. Finally, results are presented and discussed, demonstrating the relevance of our approach and its sensitivity to hyper-parameter choices.

## 2 Related Works

### 2.1 Model Explainability Methods

As shown in recent surveys such as [3], the field of AI explainability includes methods having different approaches, such as post-hoc methods that aims to explain complex models, intrinsic methods that aims to create understandable models, methods used to enhance model fairness, or methods used to test model sensitivity. Also, most of these methods are applicable to tasks that provide a prediction that is global with respect to the input data i.e. image classification, captioning and so on. Explanation on local predictions as for pixel level or region classification is scarce. This work focuses exclusively on post-hoc interpretation methods applicable to images. Related methods do not modify or influence the model process nor apply some specific processing on the optimized model. In this context, three main categories of post-hoc explanation methods can be identified:

**Back propagation based methods** that are typically suitable for neural networks models. Several methods based on backpropagation are reported in the literature, such as Guided Backpropagation [15], LRP [9], or DeepLIFT [14]. These methods aim to produce explanations by back propagating a network output score (e.g. a class probability) through the network to the first layer. In this way, the input image pixels that contributed the most to the network decision are highlighted and thus produce a heat map also referred to as saliency maps in some papers. These methods compute a rather fast and precise explanation (at the pixel level). However, they have some limitations in terms of flexibility, as they are mainly used for classification neural networks. Moreover, the obtained saliency map often results from a tradeoff between human understandability and fidelity to the network decision process.

**Activation based methods** combine the feature maps of a considered neural network layer to produce explanations presented as a coarse heat map. One of the best-known methods in this category is the popular Grad-CAM [12]. The intuition behind activation-based explainability methods is to combine only feature maps that have patterns considered important by the network for an output. Selecting only these feature maps highlights the relevant areas of the input image with respect to the network. The obtained heat maps are relatively easy to interpret since they are coarse. However, they are also inaccurate and not suitable for fine-grained explanations.

**Occlusion based methods** are the only type of model-agnostic methods and rely on perturbation approaches. The intuition is that if a sample feature, for instance an image area, contains relevant information, then occluding such area will harm the model output. Thus, occlusion based methods, such as LIME [11], RISE [10] and SHAP [8] compute input feature importance estimates relying on model response when masking them. For this purpose, occluded versions of the input image are computed and passed through the model to compute an output value. This process is longer than any other explanation methods but has the advantage of creating more global, understandable and relevant results. Further, it is independent of model type and architecture and is easy to implement.

In this work, we focus on occlusion based methods, as they present more advantages than other methods. Most importantly, they allow for comparison between different models while not being dependent on their internal processes.

### 2.2 Comparison of Occlusion Based Explanation Methods

**LIME and SHAP** are based on the same algorithm described in [11], create a linear decision model $g_x$, that aims to approximate the black-box model $f$ for a given input $x$. Applied to image analysis problems, such approach relies on the input image division into super-pixels (image regions), which are further occluded to examine their impacts $\phi$ on the model output. The general formulation

has been introduced in [8] as:

$$g_x(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{1}$$

Where $x'$ corresponds to the mapping of the input $x$ through the function $h_x$, such that $x' = h_x(x)$. Formally, $x' \in \{0,1\}^M$ is a vector representing the presence or absence of input super-pixels and $M$ is the number of super-pixels.

Differences between LIME and SHAP reside in the importance value attributed to each super-pixel: while LIME use heuristic coefficients to compute its contribution values, SHAP relies on Shapley values [13], a game-theory approach that leads contribution values to be better aligned with human intuition, and results in more relevant explanations. More into the details, Shapley values are defined to satisfy three properties, *Local Accuracy, Missingness and Consistency*. The derived equation detailed in [8] is:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z') - f_x(z' \setminus i)] \tag{2}$$

Where $|z'|$ is the number of non-zeros entries in $z'$, $z' \subseteq x'$ represents all $z'$ vectors where the non-zero entries are a subset of the non-zero entries in $x'$, and $f_x(z')$ the black-box model considering simplified inputs $z' = h_x(z)$.

**RISE** [10] is another occlusion based explanation method. It does not rely on a rigid super-pixel structure but applies occlusions relying on random mask generation. This has the advantage of reducing the potential bias caused by a rigid organization of image features. However, it also leads to coarser explanation maps comparable to those produced by Grad-CAM. Given a set of $s$ masks $M$ and the model output value *scalar* $o_i$, when the input image is masked with $M_i$, the final $Heatmap$ is computed as the normalized sum of the image masks weighted by the corresponding model output maps.

$$Heatmap = \frac{1}{\mathbb{E}[M].s} \sum_{i \in s} o_i \times M_i \tag{3}$$

A limitation of RISE is the fact that this method does not provide information on the type of feature contribution i.e. excitation or inhibition effect on the prediction.

## 3 Occlusion methods Adaptation to Semantic Segmentation

From state of the art, the SHAP method appears the most relevant. However, its adaptation to image segmentation is not straightforward. We first propose a general framework for the adaptation of any occlusion based method, from which we detail some steps, specific to our SHAP adaptation.

### 3.1 Proposed Approach

The proposed general workflow is illustrated in Fig. 2. It relies on four steps.

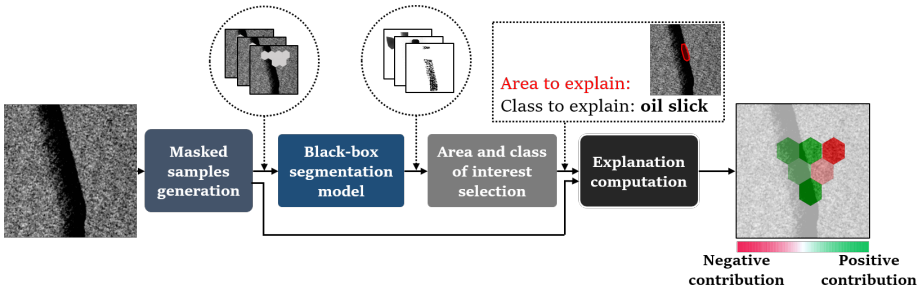


Fig. 2: Workflow of occlusion based explanation methods adapted to image segmentation.

The first step generates masked samples of the original image. Masks are occluding super-pixels whose shape and number are controlled by dedicated hyper-parameters. Resulting masked images are processed in the second step by a black-box image semantic segmentation model (e.g. a deep neural network) that provides one prediction for each masked image. For the applications presented in this paper, we always consider probability maps as the model output but model output logits or binary classification results could also be considered. The third step consists in the selection of the regions and classes of interest (RoIs) for which model output explanation must be computed. Lastly, occlusion based explanation methods are applied on the RoIs making use of the mask configurations and the selected model outputs. Finally, an explanation is generated, presented as a heat-map pointing areas of the input image that contributed the most to the model decision.

As illustrated in Fig. 2, the model decision for the red polygonal RoI, is explained. The red colour range is assigned to areas decreasing the target class probability value for the given RoI (negative contribution, or *inhibition*), while the green colour range is assigned to areas increasing this value (positive contribution, or *excitation*). Colour saturation is related to the amplitude of the contribution value.

Any occlusion method can be involved in this framework. RISE, as an example, only inputs the area and class of interest selection for its adaptation to segmentation: the masking step and explanation computation are already defined in the function to follow Eq. 3. As a comparison, SHAP method requires more information about the image sample mask configurations as described in the following.

### 3.2 SHAP Case Study

**Implementation**

The application of the kernel SHAP method to the framework is described in

Fig. 3. The adaptation consists in associating the explanation computation with the mask sample generation steps in order to comply with Eq. 2. SHAP indeed relies on a set of predefined and static features, here super-pixels. Also, for each masked image sample, feature state (i.e. masked or unmodified) must be known.

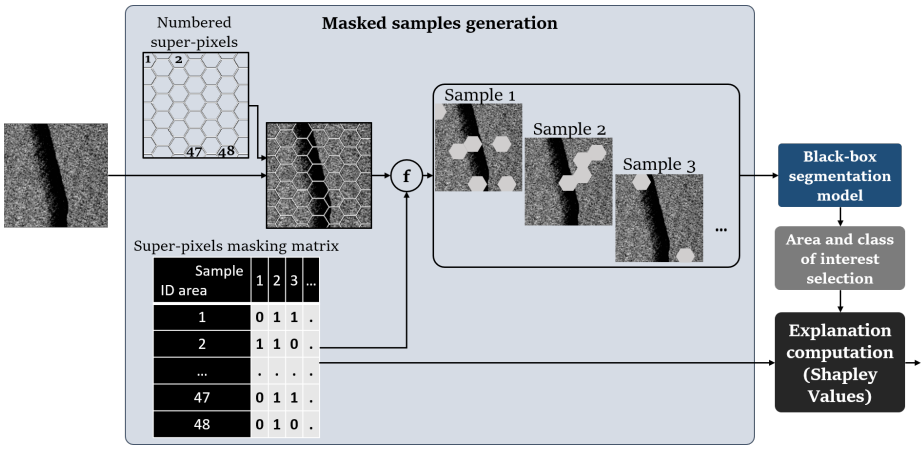


Fig. 3: The *Masked samples generation* step from Fig. 2 detailed for our SHAP adaptation. Each super-pixel is delimited following a hexagonal grid. Then, masked samples are generated via function $f$, using a masking matrix to occlude a given super-pixel, on a given sample.

The input image is then clustered into a set of uniformly organized and non-overlapping hexagons, having an identical area. This choice is more detailed in the next section and facilitates the readability of the result relying on super-pixels of equal importance and more homogeneous neighbourhood relations. In our experiments, considering images of size $512 \times 512$, the input image is typically clustered in $M = 224$ super-pixels of about 1170 pixels.

**On the Relevance of Super-Pixel Shapes**

The most critical parameter for the SHAP method applied to image analysis models is the delimitation of the input super-pixels. Several experiments were conducted as presented in Fig. 4. We first considered a method based on a configurable k-means-based clustering algorithm, SLIC [1]. Tests were performed applying SLIC on the input data with different parameters as illustrated in Fig. 4.A and 4.B. Clustering has also been applied from ground truth images for more homogeneous clustering while making use of class boundaries as shown in Fig. 4.C. From preliminary results and visual analysis, we conclude that all those automatic clustering based methods cannot provide consistent and stable clustering and could not provide homogeneous super-pixel delimitation of dark patches and sea areas. In addition, SHAP values also depend on super-pixels surface, meaning that more homogeneity in super-pixel shapes would facilitate

human interpretation. We thus suggest clustering pixels regions, not relying on the image content but rather making use of regular grids that yield super-pixels homogeneous in shape and size. More specifically, we propose to rely on a hexagonal grid (Fig. 4.D), which present more regular connectivity patterns that could help create more natural explanations.

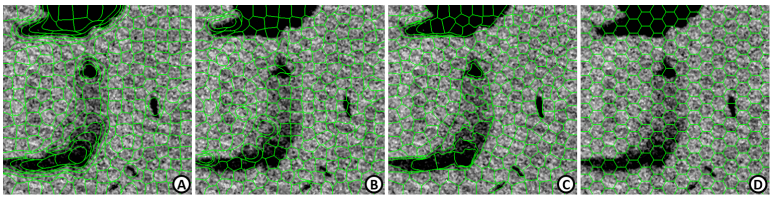


Fig. 4: Tested super-pixels for SHAP method, delimited in green. One compares automatic clustering approaches (A-C) and a predefined hexagonal grid (D).

### 3.3 Explaining Predictions on Images Regions

From a practical point of view, user consider RoIs as image regions while limiting to a single pixel is a scarce case study. As an example, on the oil slick detection problem, photo-interpreters for now manually delimit large regions surrounding the slicks they detect. Conversely, when it comes to assessing automatic detection, such experts, expect models to follow similar behaviours. Finally, explanations that would assist experts in their assessment should also be compliant with such behaviours and thus provide a regular approach for all case studies.

The explanation of regions can be performed in a variety of ways and we focus on an approach that makes sense with respect to the application context while not increasing the computational cost compared to a single pixel explanation. When explaining a pixel classification, SHAP estimates the sensitivity of its target class prediction probability with respect to the super-pixel coalition changes. Similarly, when willing to explain the prediction on a group of connected pixels, i.e. a region, we propose to estimate the sensitivity of the average target class predicted probability over that region. The semantics remain the same but for a wider region of interest. From an implementation point of view, it consists in a limited change at the third step of the workflow depicted in Fig. 2, 'Area and class of interest selection': the SHAP implementation remains the same but receives either a single pixel of interest probability or the average probability of the pixels within the region of interest.

## 4 Experiments

The proposed method has been evaluated on real application case studies. The first one relates to oil slick detection at the sea surface. It involves the application of a semantic segmentation model applied to SAR images used on operations.

This is a critical environmental and safety case study where oil detection can generate very strong and costly responses. Thus model predictions explanations make real sense. The second case study relates to semantic segmentation of multiple object categories in urban scenes on the standard Cityscapes dataset [4], relying on RGB images. In this section, we push emphasis on the first case study and show the applicability of the same method on the more classical second case study. We then first detail the experimental setup for oil slick detection with the data and the models considered. Second, we present and discuss the results obtained from different perspectives.

### 4.1 Oil Slicks Detection Experimental Setup

We build on the model and data collections presented in [2] that are dedicated to offshore oil slick semantic segmentation at the sea surface from SAR data. The model involved is based on the FC-Densenet architecture [6] and is trained in a supervised manner on a large collection of images extracted from real monitoring scenarios and annotated by photo-interpreters. SAR Imagery allows for day-and-night detection of oil slicks that appear as patches darker than their neighbourhood thanks to radar response on their surface.

In this paper, no more details on the model and related optimization are provided in order to keep the focus on the explanation methods. Paper [2] provides more details on the model that we consider here as black box. The aim is indeed to bring transparency to such models from an application point of view. Then, one considers the predictions of this preliminary trained model on new images not involved in the training process as for real monitoring scenarios. In this specific case study, the background colour applied to masked pixels cannot be black: It would induce a bias in explanations, as masked super-pixels would be detected as an oil slick. Then, background value is set to the input image average grey level.

### 4.2 Oil Slicks Detection Explanation Results and Discussions

First, we compare our adapted SHAP and RISE approaches for the explanation of semantic segmentation results of the same model on the same samples. Next, focusing on the SHAP approach, we study the impact of the super-pixel size on the explanation relevance. We finally show the consistency of the pixel and region-level explanations.

**SHAP vs. RISE**

RISE adaptation to semantic segmentation is made following the workflow presented in Fig. 2. Default RISE hyper-parameters are kept, such that $\frac{1}{\mathbb{E}[M]} = 2$, typical masks (occlusion cells) are of size $64 \times 64$, i.e. 4096 pixels at most, and the number of samples remains $S = 2000$.

Considering the same model and the same output selection, a comparison of the explanations provided by the SHAP and RISE based methods is presented in Fig. 5.

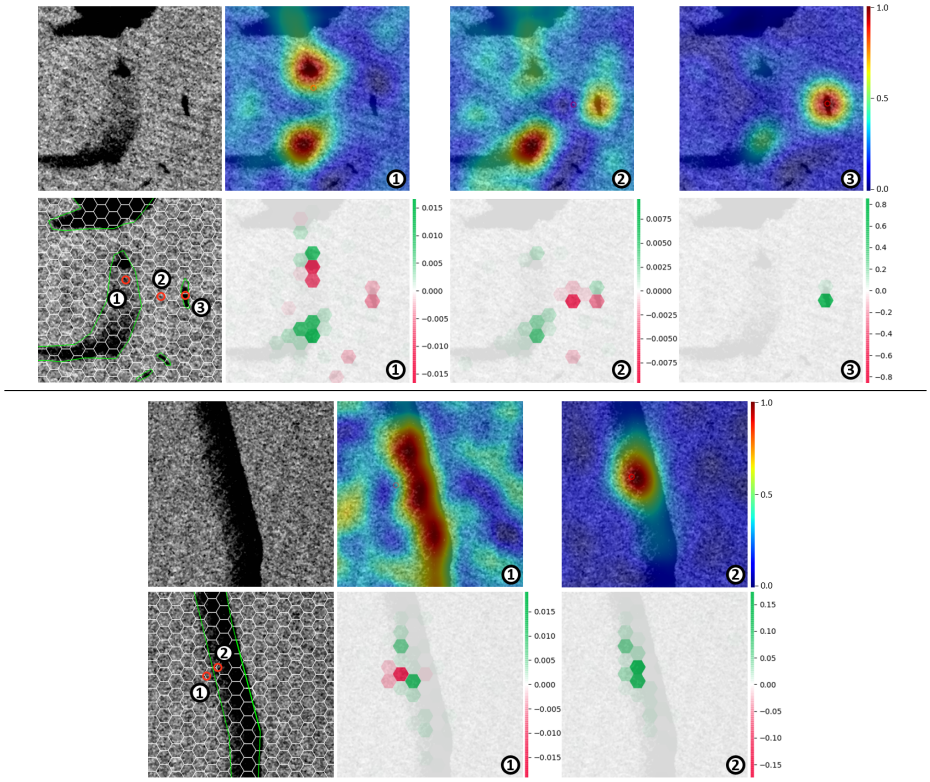

Fig. 5: For 2 image examples (top, bottom), explanation maps on the model decision for some pixels of interest (red enumerated circles) with either RISE (top-right images) or SHAP (bottom-right) adapted methods.

First, one can observe that high RISE explanation values tend to correspond to the super-pixels with a positive contribution obtained with SHAP. Surprisingly, RISE additionally reports diffuse areas with low contribution values that can be very distant from the RoIs. However, negatively contributing features reported by SHAP are not highlighted by RISE. Also, relevant regions reported by RISE have more spatial extent and are poorly contrasted such that this reduces the explanation precision. This can be explained by internal mask subsampling process, and by RISE occlusion masks, four times bigger than the ones used with SHAP.

As a first conclusion, while RISE and SHAP report highly positive contributions to the decision in a consistent way, SHAP provides more detailed and more relevant explanations both in terms of resolution and contribution type.

**Super-Pixel Size Impact on Explanation**

Focusing on the SHAP based method, we examine the impact of the super-pixel size on the explanation relevance. Fig. 6 shows explanation maps on the same model prediction but with different super-pixel sizes.

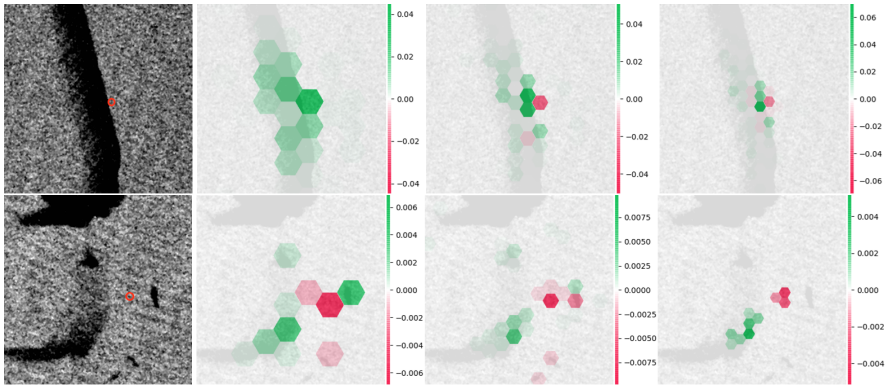

Fig. 6: Two examples (top, bottom) of adapted SHAP explanation of model predictions on the red circled pixel considering large, medium and small super-pixels.

One observes that large super-pixels may cover patterns of different target categories and thus yields a loss of information regarding the contribution type (excitation or inhibition). On the other hand, too small super-pixels may also cause a loss of information, when they provide a partial view of large objects in the visual scene. The relevance of the explanation then actually depends on the image content and the super-pixel spatial distribution but rigid super-pixel grids already provide a good compromise.

**From Pixel Level to Region Explanation**

As described in section 3.1, the proposed approach allows for the explanation of single pixels and regions in a unified way. Considering the same model and the same test image shown in Fig.5, explanations of predicted regions are shown in Fig. 7 and Fig. 1. The considered regions actually surround a single pixel explained previously and one can observe the consistency of the explanations: in the case of the sea area, the region affecting the model decision is large, and considers slick regions as either inhibition or excitation; in the case of an oil slick area, the contributing regions are mostly restricted to the neighbouring oil slick super-pixels that positively contribute to oil classification; in the case of an in between area, the model combines both behaviours.

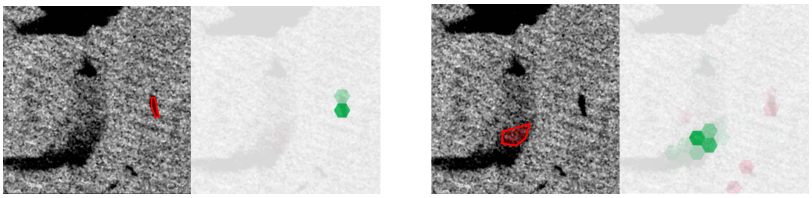

Fig. 7: Explanation provided for two regions: oil (left) and oil and sea mixture (right). Fig. 1 shows a clean sea area in the same experimental conditions.

**Conclusions on Model Behaviours**

Using SHAP explanations presented in this section, one can understand the model behaviour when confronted to different cases :

For RoIs classified as sea, close sea super-pixels have inhibitory effects to classify as oil while neighbouring super-pixels containing oil have an excitatory effect. Overall, these effects are extremely low and balance each other, making the final classification as sea.

For RoIs classified as slick, both model decisions are based on a limited number of super-pixels containing oil, close to the explained RoIs. They almost always have an excitatory effect, with high intensity.

For ROIs in between slick and sea, models tend to consider contextual information over the whole image, and specifically contrasted patterns. However, if no salient features are present in the image, the model relies on a very local area.

### 4.3 Urban Scene Segmentation Results and Discussions

One considers an implementation of the state-of-the-art HardNet-Mseg model [5] trained to perform semantic segmentation of the 34 visual concepts of the Cityscapes dataset [4] (people, cars, road signs and so on) from RGB images. Fig. 8 shows a typical visual scene from the validation set with associated ground truth and a model prediction. Compared to the previous case study, this multi-class segmentation problem must consider more numerous and diverse overlapping class instances. Then, relying on the same configuration (image size of 512x512 pixels, same hexagonal grid), we show the applicability of the method for a very different context.

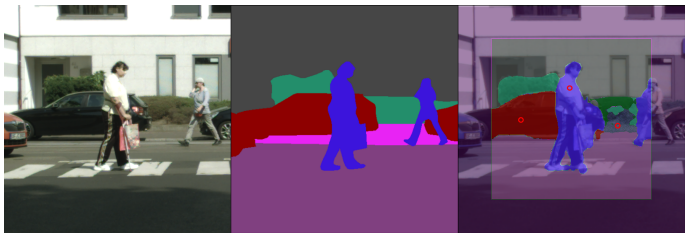

Fig. 8: Urban scene semantic segmentation example: input (left), coloured ground truth (centre), predicted segmentation (right), model does not predict on crop boundaries, 3 circled pixels in the predicted area are subject to explanation.

SHAP super-pixel size is kept medium for these experiments, and the background masking value is set to zero (black) as usually done for such multimedia data. Three pixels of interest are considered for explanation, one on a car that is well predicted, one on a person also predicted well. The last one is more ambiguous. It is annotated as sidewalks but lies at the frontier between sidewalk, concrete, ground and vegetation and is predicted as *static* (a class regrouping

indistinguishable objects that correspond to none of the other classes). For each pixel, two explanation maps are provided, for two different classes of interest.

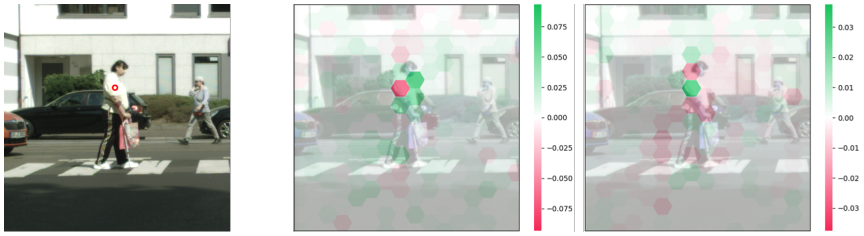


Fig. 9: SHAP explanation for the pixel circled in red in the input image. Middle image corresponds to the class 'person', and right image to the class 'dynamic'.

Fig. 9 first presents explanation maps for the selected pixel classified as 'person' (probability=94%). It shows a significant positive contribution focused on the front and upper part of the body, which is an expected phenomenon. However, the hexagon containing the explained pixel has a negative contribution. It covers a homogeneous white area around the person shoulder with some contours on its left boundaries. This is explained by the second explanation map, related to the 'dynamic' class that gathers movable objects. This class has a similar spatial distribution than the 'person' class, but with opposed signs. From these two explanations, we can conclude that the shoulder super-pixel alone help the model detect the pixel as a dynamic object, but neighbouring contextual information (head, arm and low chest) contribute more significantly to classification as a person.

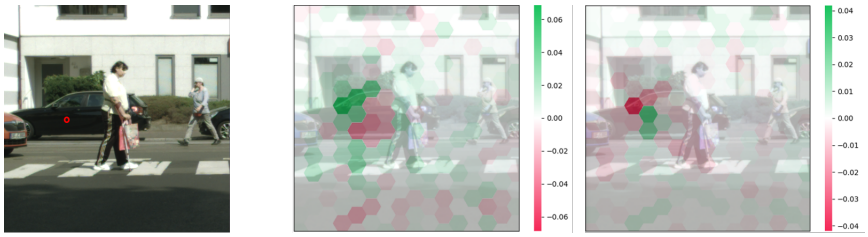


Fig. 10: SHAP explanation for the pixel circled in red in the input image. Middle image corresponds to the class 'car', and right image to the class 'sidewalk'.

Fig. 10 shows explanation on the pixel detected as a car (probability=69%). The first explanation map, representing the car class, shows that the vehicle windowed part increases the car class probability, while the car bodywork decreases it. However, in this region, 'sidewalk's explanation map class report opposite values. This shows that the car bodywork, without a larger view of the vehicle, can be interpreted as a sidewalk by the model, which may appear natural in light of its homogeneous dark colour.

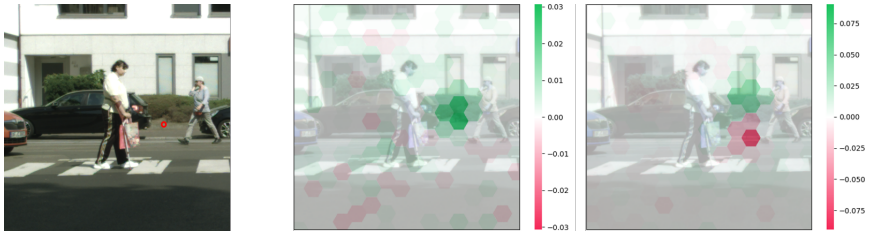


Fig. 11: SHAP explanation for the pixel circled in red in the input image. Middle image corresponds to the class 'static', and right image to the class 'dynamic'.

Fig. 11 focuses on a misclassified pixel as 'static' (probability=34%), at the frontier on the sidewalk and vegetation with visible concrete and ground. This is a typical area subject to difficult annotation. One can observe the explanation of the two dominant classes for this pixel, 'static' and 'dynamic'. Static class is excited by the pixel surroundings, while the dynamic class seems to be inhibited by the road below the pixel. On the other side, the explanation maps for sidewalks and vegetation do not report significant values. Then, the two most probable classes are finally relevant for such complex region and highlight the difficulty of the annotation.
These results show the relevance of the approach on a second very different case study but relying on the same explanation method hyper-parameters. Further refined analysis could be proposed to provide more details on the local patterns impact on the decision by adjusting super-pixel size and the number of samples on explanation maps. However, this depends on the expected explanation level. Typically, smaller super-pixels may lead to more intuitive explanations, as it would better fit to various small objects and features of input images. Moreover, increasing the number of masked samples along the explanation process has proven to reduce noise in the explanation maps particularly present in Fig. 10.

## 5 Conclusion

This work presents a general workflow that allows for the adaptation of SHAP and RISE explainability methods to the semantic segmentation task. SHAP based method provides more relevant explanations and allows for refined understanding of model behaviours. Experiments were conducted to assess the parameters choice of the presented method and detail its advantages and pitfalls. The developed method was tested on deep neural networks trained for remote sensing oil slick segmentation, as well as urban scene segmentation on the public Cityscapes dataset. Explanations permitted to identify general model rules for specific input data configurations. Future works will focus on the extraction of relevance metrics by involving human domain experts, as well as model behaviour comparison using SHAP explanation. Finally, SHAP super-pixel delimitation strategies need to be studied more deeply for different applications, as it may lead to better explanations.

## References


1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (2012). https://doi.org/10.1109/TPAMI.2012.120
2. Amri, E., Courteille, H., Benoit, A., Bolon, P., Dubucq, D., Poulain, G., Credoz, A.: Automatic offshore oil slick detection based on deep learning using sar data and contextual information. In: Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions 2021. vol. 11857, pp. 35–42. SPIE (2021)
3. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)
4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
5. Huang, C., Wu, H., Lin, Y.: Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 FPS. CoRR **abs/2101.07172** (2021), https://arxiv.org/abs/2101.07172
6. Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 11–19 (2017)
7. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. Entropy **23**(1), 18 (2021)
8. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. pp. 4768–4777 (2017)
9. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018)
10. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421 (2018)
11. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
12. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
13. Shapley, L.S.: A value for n-person games. Princeton University Press (2016)
14. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning. pp. 3145–3153. PMLR (2017)
15. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)