# $H^2O$: Heatmap by Hierarchical Occlusion

Luc-Etienne Pommé, Romain Bourqui, Romain Giot

# $H^2O$: Heatmap by Hierarchical Occlusion

Luc-Etienne Pommé
LaBRI UMR CNRS 5800, University of Bordeaux
Talence, France
luc.pomme-cassierou@u-bordeaux.fr

Romain Bourqui
LaBRI UMR CNRS 5800, University of Bordeaux
Talence, France
romain.bourqui@u-bordeaux.fr

Romain Giot
LaBRI UMR CNRS 5800, University of Bordeaux
Talence, France
romain.giot@u-bordeaux.fr

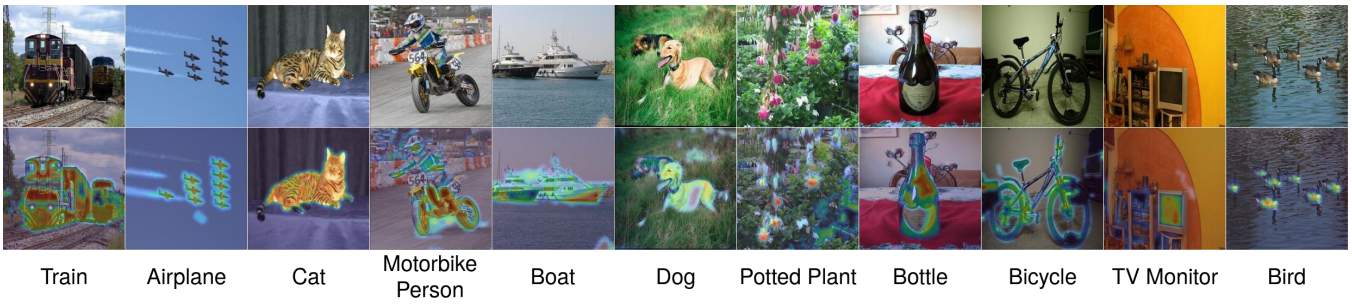


**Figure 1: Saliency maps generated with $H^2O$ for images sampled from the PASCAL VOC2007 [7] dataset. The first row shows input images while the second row displays the superimposition with the corresponding saliency map.**

## ABSTRACT

The rise of Deep Learning (DL) has led to a breakthrough in the research field of content-based multimedia indexing. Newly developed systems based on complex models outperform classic machine learning algorithms in object detection, image segmentation or classification tasks. However, despite their high performance, these systems still make mistakes. To be used in industrial conditions, these systems must be able to provide trustworthy decisions with guarantees or justifications. Therefore, it is crucial to provide means to analyze and comprehend the decision process that leads a model to its decision. Image classification implies tracking and understanding which input features the model relies on to make its prediction. This paper focuses on features attribution techniques and proposes *Heatmaps by Hierarchical Occlusion* ($H^2O$), a novel method for detecting pattern-relevant features in an image. We also propose two new pairs of metrics that overcome some evaluation issues: (a) Insertion and Deletion Spearman correlation coefficients which both estimate a correlation between the computed scores in a saliency map and the importance for the model of the associated pixels in the image. (b) Insertion Positive and Deletion Negative Gradient Sums both estimate the coherence of the scores in the saliency maps. Both visual inspection and evaluation on 7 metrics show that $H^2O$ is competitive against state-of-the-art methods.


*CBMI2023, September 20–22, 2023, Orléans, France*

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

## CCS CONCEPTS

• **Computing methodologies** → **Interest point and salient region detections**.

## KEYWORDS

XAI, Hierarchical Features Attribution, Image classification

**ACM Reference Format:**
Luc-Etienne Pommé, Romain Bourqui, and Romain Giot. 2023. **H²O**: Heatmap by Hierarchical Occlusion. In *Proceedings of Content-based Multimedia Indexing (CBMI2023).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Content-based multimedia indexing is used in our daily life activities [6]. Such systems strongly rely on various machine learning tasks, especially classification techniques [13]. As these systems are not infallible, designers need to evaluate them [8] to assess their actual performance. However, the commonly used metrics only enable to compare models or to have a general idea of their efficiency. Some metrics and approaches are able to measure the failures and successes [18] but do not enable understanding *why* they fail or succeed. This is especially problematic when using deep learning models [15] that are commonly considered to be black boxes [12, 22] due to their internal complexity and number of parameters. This is where eXplainable Artificial Intelligence (XAI) comes into play. This field aims at providing explanations of models with still images or interactive applications [14]. Among the possible categories of explanation methods, we are interested in *features attribution* methods that aim at explaining which input features influenced the decision process of an image classifier.

The literature proposes several methods (see next section) to produce features attribution of image classifiers. The most popular

ones are either gradient-based methods [21, 23] or occlusion methods [17, 19]. In this paper, we focus on some of the problems of existing occlusion methods. Occlusion methods usually produce an explanation of a single class per input sample, regardless the predicted class(es) for that instance. Such an explanation is not entirely representative of a model's prediction. This can be problematic for single-label classification tasks, where models are built with a softmax function. Because of this function, a model can predict a specific class because the other classes are less likely, and not because the sample belongs to a class. This can also be problematic for multi-label tasks where each sample may belong to multiple classes. Currently, there is no consensus to determine whether the ground truth class(es) or the predicted class(es) should be explained. In both cases, it is not reasonable to display a heatmap for every single class, due to their number, but existing methods do not provide any way to aggregate the heatmaps of the target classes.

Moreover, existing occlusion methods strongly depend on the strategy used to replace the masked zones (usually a single color). Depending on the chosen color(s) the image features can be either masked or enhanced, which may alter the quality of the explanation.

In this paper, we focus on these two problems. We propose a novel class-agnostic method to generate an explanation for the whole prediction vector of an image classifier. It relies on a hierarchical superpixel segmentation of the image and a new masking color strategy to increase the quality and stability of the explanation. We suggest alternative metrics to evaluate their quality alongside a strategy to compare class-agnostic and class-specific explanations.

The paper is organized as follows. We first present a brief state-of-the-art of existing methods that generate an explanation in Section 2. In Section 3, we then describe our method. Then, we establish the evaluation protocol of our study (Sec. 4) and proceed with quantitative and qualitative results (Sec. 5). We also discuss the strengths, the weaknesses and the future improvements on this work (Sec. 6), and draw conclusions in Sec. 7.

## 2 PREVIOUS WORK

Among the possible families of explanations, we are interested in *feature explanations*. Feature explanations aim at highlighting input features of interest: those strongly correlated with the final decision of the classifier. For the classification of RGB images represented by tensors of shape $(w, h, 3)$, the explanation corresponds to a *saliency map* (also named *explanability map* or *feature attribution map*) of shape $(w, h)$ containing high values for pixels of interest and low values for the others. This saliency map is usually presented to the user with a heatmap overlaid on the input image. This section overviews the main concepts behind existing methods as well as various evaluation methods.

### 2.1 Features Attribution for Image Classifiers

Several methods [10, 17, 19, 21] have been proposed in the literature to generate features explanations that can be used in the context of image classification. As shown in the summary in Table 2 in the Appendix, key contributions fall into four main categories. *Gradient-based* [21] and *Activation-based* [3, 10] methods are both model-specific methods. They either use the weights gradients of a differentiable model [21] or aggregate the outputs [3, 10] produced when feeding a model with samples, to explain individual classes. In the former category, Deconvnet [24] or back-propagation-based methods [24, 25] try to reverse the operations processed in a neural network to highlight the input features detected by the model. *Alteration-based* methods [20] modify an input image to determine which input features impact the most a model's prediction. *Occlusion-based* methods [17, 19] can be seen as a special case of *alteration* methods when the input images are masked to deteriorate the prediction. While the first two categories are model-specific, the last two are model-agnostic as they work with any model. Another distinction can be made between class-specific methods that fix a class to explain (*e.g.* GradCAM [21], LIME [19], RISE [17]), and class-agnostic methods (*e.g.* FEM [10], MLFEM [3]) that have the advantage of explaining all the prediction components at once.

For occlusion-based methods, which are the focus of this paper, evaluating precisely the impact of each feature of an image in the model's prediction would require to consider every possible subset of pixels, which is too costly to consider. To tackle this problem, LIME [19] groups similar pixels into clusters (superpixels), and only masks a subset of the possible combinations of superpixels. A model then estimates the importance of each superpixel based on the deterioration induced in the prediction vectors. RISE [17] randomly mask an input image with $N$ masks. The saliency scores are computed from the deterioration scores of each masked image.

Inspired by these works, we propose a *hierarchical*, *model-agnostic*, *class-agnostic*, and *occlusion*-based method.

### 2.2 Evaluation of Features Explanations

Evaluating explanations is one of the key problems the XAI community has to face. To the best of our knowledge, there is not yet a consensus on the evaluation procedure: most methods are evaluated in different ways and new protocols are not broadly adopted (see Table 2). Still, two categories of metrics can be established from the literature: *human-centered* and *automatic* metrics.

*Human-centered* metrics measure how close an explanation provided by an algorithm is to a ground truth explanation provided by humans. In this category, the Pointing Game (PG) [25] measures whether or not the maximum value of class-specific explanations falls into one of the human-labeled bounding boxes of that class, in the whole dataset. Though, this metric only measures the accuracy of the highest score in an explanation, and does not take into account the entire distribution of scores. In MLFEM [3], the explanations are compared to Gaze Fixation Density Maps (GFDM) using a Similarity (SIM) metric and a Pearson Correlation Coefficient (PCC). These maps represent the zones of interest that humans used when asked to look at an image and classify it. Both of these approaches evaluate an interpretability criterion [21], but Adebayo et al. [2] identified that features attribution methods cannot be evaluated only on visual assessments. These metrics do not measure the faithfulness to the model, which is the ability to accurately represent the features selected by the model to make its decision [21].

On the contrary, *automatic* metrics evaluate specific properties that measure how faithful the explanation is to the model decision. Deletion Area Under Curve (DAUC) and Insertion Area Under Curve (IAUC) are two complementary metrics that estimate the quality of the distribution of the scores in an explanation. The

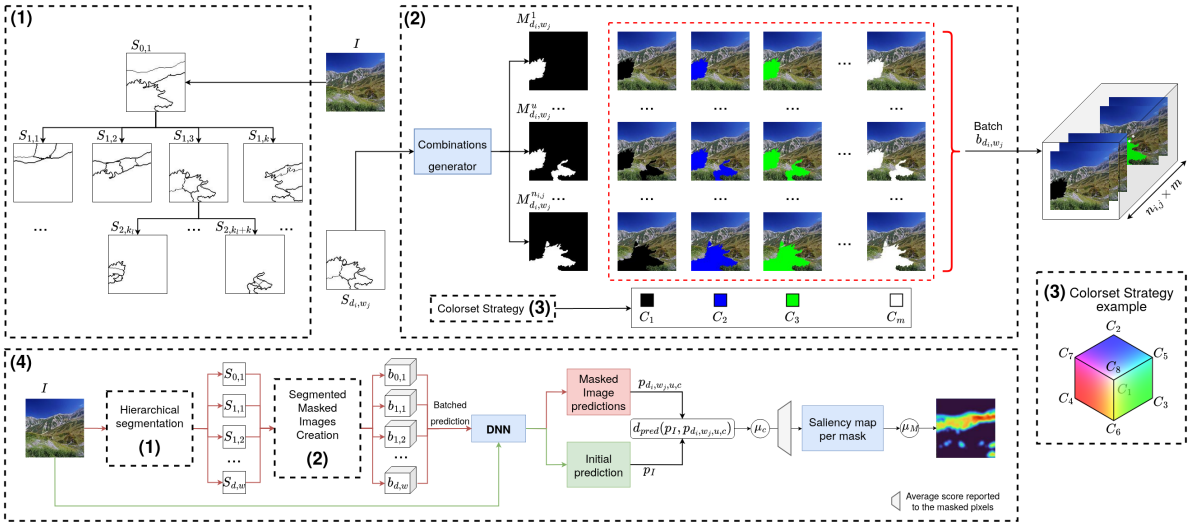


**Figure 2: Overview of $H^2O$, a model and class-agnostic occlusion-based explanation method. (1) An input image is hierarchically segmented into superpixels. (2) In each level of the hierarchy, all superpixels combinations form a set of masks, applied on the input image. (2), (3) The masked pixels are replaced with one color at a time. (4) All altered images are predicted, leading to a potential deterioration that is used to compute the importance scores of each pixel in the input image.**

DAUC metric iteratively removes pixels from an image in decreasing order of importance (w.r.t to the saliency map), while computing a class-wise prediction for target classes. DAUC is the area under the curve of the deterioration according to the number of pixels. Its IAUC counterpart starts from a blurred image and iteratively adds the pixels of the original image in decreasing order of importance (w.r.t to the saliency map), before computing the area under the curve.

However, these two metrics only consider the ranking of the scores and not their values. To tackle this problem, Gomez et al. [11] propose the Deletion and Insertion Correlation metrics to measure a correlation between the explanation scores and the class-wise deterioration or improvement.

## 3 H²O: HEATMAP BY HIERARCHICAL OCCLUSION

This section details how $H^2O$ computes the saliency map[1].

### 3.1 Process Overview

$H^2O$ (see Fig. 2) is a method that explains the output prediction vector of an image classifier (*i.e.* what the network has globally detected in the image). The method is (a) model agnostic as it considers the model as a black box, and (b) class agnostic as it aims at explaining all values of a prediction vector at once and not how the model may have seen a specific class in an image, which can be beneficial in both single-label and multi-label task scenarios. $H^2O$ belongs to the occlusion family as the importance of the pixels in the image to explain is computed by altering the image in different ways and measuring the variation (usually a deterioration) induced in the prediction vector. The method leverages a hierarchical segmentation algorithm to measure different levels of pixels relevance to the model. $H^2O$ uses a colorset strategy to alleviate the problems of masked features enhancement while masking pixels in the image with an arbitrary color (see Sec. 1). The deterioration scores obtained in all levels of the hierarchical segmentation are then aggregated to compute the final saliency map.

### 3.2 Hierarchical Segmentation

To compute any deterioration, some portions of an image must be occluded first. To determine the portions to mask, the initial image is segmented into $k$ superpixels (see Sec. 2.1). When $k$ is large, the number of combinations of superpixels grows exponentially to $n_{i,j} = \sum_{i=1}^{k-1} \binom{k}{i}$. This leads state-of-the-art methods to only generate a random subset of the possible altered images (*e.g* LIME [19]). In comparison, we aim to generate the $n_{i,j}$ combinations of superpixels, which is achievable with a small $k$. However, a small $k$ produces large superpixels. To enable a fine-grained segmentation, $H^2O$ hierarchically segments an input image with an algorithm that guarantees that any segmentation produces at most $k$ superpixels at each level. The segmentation process can be represented as a $k-$way tree (see Fig. 2(1)) where the root node $S_{0,1}$ is the segmentation applied on the whole image. Each superpixel from a parent node $S_{d_i,w_j}$ is re-segmented into $k$ superpixels at most, if its size exceeds $T$ pixels, $T$ modeling an ideal and fixed minimum number of pixels an important object could occupy in an image.

### 3.3 Segmented Masked Images creation

Each tree node $S_{d_i,w_j}$ represents the $k$ or fewer superpixels resulting from a segmentation step on a portion of the initial image. As our method aims at generating all possible combinations of input features, for each node, a generator exhaustively enumerates the

[1] $H^2O$'s implementation is available at https://github.com/labribkb/h2o.

$n_{i,j}$ combinations of the $k$ or fewer superpixels obtained for that node (see Fig. 2(2)). For every combination of superpixels of a node $S_{d_i,w_j}$, a boolean mask, noted $M^u_{d_i,w_j}$ is built, where $u$ is the $u^{th}$ combination of that node. This mask contains a 1 for each pixel that belongs to one of the superpixels selected in the combination, and 0 elsewhere. Each mask is individually applied to the input to only alter the pixels corresponding to a 1. Ideally, when predicting an altered image, its masked pixels should not be involved in the decision-making process at all. To simulate this alteration without changing the model, existing methods replace the masked pixels with an alternative color (either black [17] or the superpixels color mean [19]). However, input images may already contain such colors for important objects. It is then unclear whether the model would be interpreting the masked pixels as important parts of the objects or not. In other words, it is probable that some colors enhance the presence of an object while some others reduce it. To try to reduce the impact of the replacement color, we suggest increasing their number from 1 to $m$. Thus, for each mask $M$, the input image is altered $m$ times with all replacement colors $C_c, \forall c \in [1, m]$.

### 3.4 Colorset Strategy

We describe here one strategy to choose this set of replacement colors $[C_1, C_m]$. This strategy consists in linearly subsampling the RGB cube (see Fig. 2 (3)) restricted to the colors of the considered image. To achieve that, the minimum and maximum values of all pixels along each axis (red, green, blue) are individually computed. This minimum and maximum values define a potentially smaller color cube. This restricted color cube is split into a three-dimensional grid of $z-1$ regular intervals. Each intersection point in the grid defines one of the $m$ colors that replace the masked pixels.

### 3.5 Importance score computation

All previously masked images are fed to the model to determine how much the masked pixels impact the prediction vector of the input image. Commonly in the literature, a set of pixels is considered important if *removing* them induce a high deterioration in the prediction vector. For each altered image's prediction vector $p_{d_i,w_j,u,c}$, a distance to the prediction vector $p_I$ of the input image $I$ is computed. As our approach is class-agnostic, this distance applies to the whole prediction vectors of $N$ classes, instead of on a unique class. For simplification purpose, in the following equation of the distance, the prediction vector $p_{d_i,w_j,u,c}$ is noted $p_{I'}$, where $I'$ denotes the image $I$ altered with the mask $M'$:

$$d_{pred}\left(p_I, p_{I'}\right) = \sum_{r=0}^{N} \max\left(0, \left(p_I[r] - p_{I'}[r]\right)\right) \cdot p_I[r] \qquad (1)$$

This distance can be seen as a Manhattan distance where only deteriorations are measured (max function). Weighting this distance with the scores of the initial prediction vector gives more importance to classes with high probability scores without completely ignoring the classes with a low probability score.

For a pixel $\rho$, we define $\widetilde{M}$ as the set of all masks of all nodes $S_{d_i,w_j}$. Then, we define $M^1(\rho) = \{M \in \widetilde{M}, M[\rho] = 1\}$, the set of masks $\widetilde{M}$ that contain 1 on position $\rho$. Given $\widetilde{I}$, the set of all altered images for all nodes, we compute the final saliency map $H$ of image $I$ as follows:

$$H[\rho] = \frac{1}{\left|M^1(\rho)\right|} \sum_{M' \in M^1(\rho)} \frac{d_{pred}\left(p_I, p_{I'}\right)}{\left|\{\rho' \in M', M'\left[\rho'\right] = 1\}\right|} \qquad (2)$$

To highlight the most important features and improve the locality of the explanations, the saliency maps are finally normalized with a min-max normalization and thresholded up to $\mu + s \cdot \sigma$.

## 4 EXPERIMENTS

This section presents the protocol to evaluate the quality of the generated saliency maps by $H^2O$, with the following parameters. We used Slic [1] to hierarchically segment the initial images into $k = 4$ superpixels. For an image $I$ of size $W = H = 224$, we fixed $T = \sqrt{W \times H} = 224$, the number of pixels below which to stop segmenting a superpixel in the hierarchy. Overall, the number of combinations of superpixels per image is between 2 400 and 2 600. For the chosen colorset strategy (see Sec. 3.4), we fixed the number of linear splits in the rgb cube to $z = 3$. The final saliency maps $H$ are thresholded with $s = 1$.

### 4.1 Dataset, network, baseline methods

We used the PASCAL VOC 2007 [7] test subset to evaluate our method, as it is widely used in the literature. All images are resized to $224 \times 224$ pixels. An instance of Resnet50 was fine-tuned on the train subset, from the pretrained Imagenet weights available with the Keras framework [4].

We compare $H^2O$ to four methods from the literature. We selected the commonly used GradCAM [21] method to represent the *Gradient* family and FEM [10] to represent the *Activation* family. The *Occlusion* family is represented by both LIME [19] and RISE [17], the former leveraging a superpixel approach as $H^2O$ and the latter using a random masking technique. Thus, we compare $H^2O$ to model-specific, model-agnostic, class-specific and class-agnostic approaches. The implementation of each method follows the default authors' recommendations. GradCAM and FEM are computed on the Activation of the last Convolution layer. We generated 1 000 altered samples for LIME masking superpixels with their color average. We generated 8 000 masks of size $7 \times 7$ for RISE, with a probability of 0.5 that each pixel in the mask is deleted.

### 4.2 Class-agnostic Saliency Maps Computation

For the class-specific methods, a saliency map can be computed either for the predicted class to reveal the features that have led the model to predict a class or the ground truth class to show the features that resemble the most the ground truth class, regardless the prediction. However, the multi-label property of common datasets makes it difficult to choose which class to explain. To tackle this problem, we choose to compute a linear combination of all class-specific saliency maps to create a class-agnostic one for a given method. The coefficients of such combination are the prediction scores of each class, so the saliency maps of the classes that have a high (resp. low) probability score have high (resp. low) impact on the final map. Such a saliency map is then used to compare class-specific methods to $H^2O$.

**Table 1: Comparative evaluation relatively to seven metrics averaged over the whole dataset. Bold values (resp. underlined values) denote the best (resp. the second-best) method.**

| | H²O | GradCAM | FEM | LIME | RISE |
|---|---|---|---|---|---|
| **IAUC [↓]** | **0.418** | 0.556 | 0.569 | 0.649 | <u>0.527</u> |
| **DAUC [↑]** | 0.801 | **0.917** | <u>0.909</u> | 0.817 | 0.905 |
| **IS [↓]** | **-0.760** | -0.620 | -0.648 | <u>-0.675</u> | -0.592 |
| **DS [↑]** | <u>0.638</u> | 0.549 | 0.545 | **0.652** | 0.515 |
| **IPGS [↓]** | **0.852** | 1.057 | <u>1.047</u> | 1.111 | 1.123 |
| **DNGS [↓]** | 1.476 | **0.990** | 1.051 | <u>1.036</u> | 1.095 |
| **Sparsity [↑]** | **12.793** | 6.049 | <u>6.736</u> | 6.474 | 3.227 |

### 4.3 Evaluation metrics

The following metrics are computed on the class-agnostic saliency maps of all methods. Even though the PG cannot be adapted to class-agnostic methods without any knowledge on the explained classes, the Insertion and Deletion metrics can, since they measure a variation in the prediction scores. Class-specific approaches measure a score difference on the prediction of a class while we suggest measuring a distance on the entire prediction vector. We argue that altering classes with high prediction scores is more critical than altering other classes, so we compute both IAUC and DAUC with Eq. 1. As the deterioration score is a distance, the lower IAUC and the higher DAUC, the better.

Intuitively, the saliency map of a perfect explanation should produce monotonic Insertion and Deletion curves. To measure this property, we consider two complementary pairs of metrics: Insertion/Deletion Spearman correlation (IS, DS), Insertion Positive Gradients Sum (IPGS) and Deletion Negative Gradients Sum (DNGS). The first two metrics calculate a Spearman rank-order correlation coefficient of the monotonic relationship between the improvement (resp. deterioration) in the prediction scores and the cumulative scores of the saliency map (ordered by decreasing importance). IPGS (resp. DNGS) computes the absolute sum of the strictly positive (resp. negative) gradients in the Insertion (resp. Deletion) curve. As the Insertion (resp. Deletion) curve should be monotonic and decreasing (resp. increasing), this pair of metrics penalize an explanation when adding (resp. removing) pixels increases (resp. reduces) the distance $d_{pred}(p_I, p_{I'})$.

We also evaluate the interpretability of all explanations with the Sparsity metric [11] that measures how local an explanation is with a ratio between its maximum and its mean values.

## 5 RESULTS

The quantitative results on the chosen metrics are summarized in Table. 1. In this table, bold and underlined numbers respectively emphasize the best and the second-best method according to each metric. A down (resp. up) arrow next to a metric name means that the lower (resp. higher), the better. IS and DS are correlation metrics so their scores vary between −1 and 1, 0 meaning no correlation. A positive (resp. negative) score indicates that as one variable increases the other increases (resp. decreases).

$H^2O$ gives the best results according to all insertion-based metrics (IAUC, IS and IPGS). Though, $H^2O$ does not give the best results according to both DAUC and DNGS. According to DS, $H^2O$ gives the second-best ranking-order correlation between the distances in the prediction vectors and the cumulative scores of the saliency maps. According to DAUC, GradCAM outperforms the other methods, closely followed by FEM and RISE, while according to DNGS, LIME, FEM and RISE follow GradCAM with a relatively small gap. The Sparsity metric shows that $H^2O$ produces an explanation almost twice as localized as the second-best method.

From a quantitative point of view, $H^2O$ performs best on four out of seven metrics, and second best on one additional metric.

On a qualitative point of view, $H^2O$ explanations are displayed alongside GradCAM, FEM, LIME and RISE class-agnostic explanations on four images of the PASCAL VOC dataset [7], in Fig. 3. This figure displays correctly classified samples with high probability. Globally, there is no strict consensus on the explanations, despite the fact that GradCAM and FEM (and RISE on the last row) tend to produce similar heatmaps. In the first row, all methods find value in some of the pixels of the horse. However, while $H^2O$ precisely highlight the horse's neck and tail, the horseman's trousers and the right background pillar, other methods imprecisely highlight a large portion of the horse and the background grass. In the second row, almost the whole body of the bird is crucial to the model according to $H^2O$. Comparatively, LIME and RISE mostly highlight the water, while GradCAM and FEM produce a circular shape covering parts of the bird and the water. For the airplane and the cows (rows 3-4), all methods but $H^2O$ highlight a relatively large part of the important objects with many background pixels. Only $H^2O$ precisely highlights the objects that correspond to existing classes in the dataset. In the third row, few pixels are highlighted outside of the plane for $H^2O$, while in the fourth row, $H^2O$ finds some value in the background tree to predict cows (and not a potted plant).

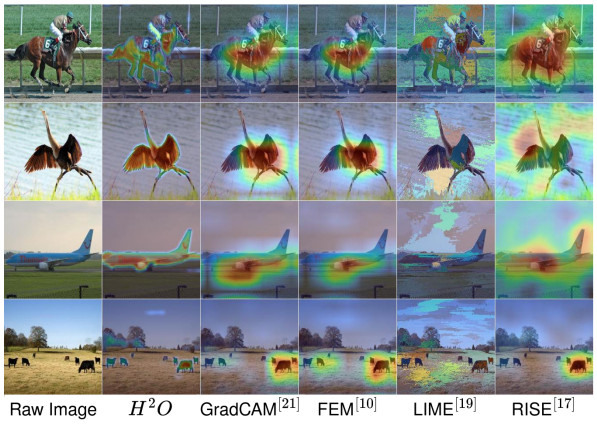


**Figure 3: Examples of class-agnostic saliency maps generated through five methods. H²O produces more localized explanations than any other method, regardless of the size and number of objects or classes in an image.**

## 6 DISCUSSION AND LIMITATIONS

As shown with the sparsity metric and the image examples, $H^2O$ produces more localized explanations than other methods considered in this experiment. Globally, $H^2O$ produces better results for

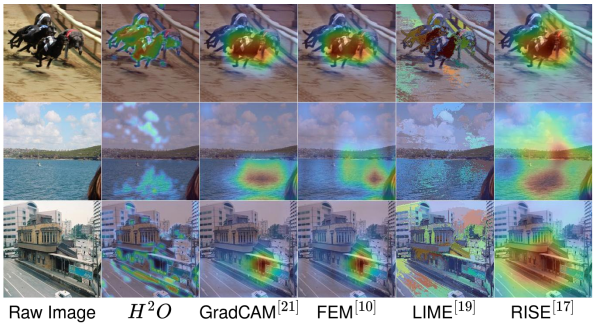


**Figure 4: Examples of saliency maps generated on samples with incorrect or uncertain prediction.**

Insertion-based metrics and worse results for Deletion-based metrics, except DS. Intuitively, insertion-based metrics rely on the most important pixels as these are iteratively inserted in the blurred input image while deletion-based metrics rely on the least important pixels as they are iteratively replaced with black. The greater results of $H^2O$ tend to indicate that $H^2O$ designated important pixels are sufficient to make a consistent prediction that is relatively unaffected by the less important pixels. However, once the most important pixels are masked out for deletion-based metrics, the prediction relies either on the remaining pixels or on the strong gradients created by black-masked pixels. For that reason, since a threshold is applied on the least important pixels of $H^2O$ saliency maps, we argue that different pixel orders may produce different results. As other methods do not apply any threshold to the heatmap, their better results according to both DAUC and DNGS are not surprising. On the opposite, IAUC and IPGS may not be as much affected by the thresholds if the important pixels are sufficient for the model to make the same prediction as on the input image. Moreover, the removed (resp. restored) pixels in DAUC (resp. IAUC) tend to better follow the objects' shape in $H^2O$ than in GradCAM or FEM, where objects borders are reached faster, on average (*e.g.* the bird in Fig. 3). Then, for deletion-based metrics with $H^2O$, it is possible that the strong gradients created by black masks preserve the shape of key objects, detected by the model. This could also explain the worse results for DAUC and DNGS compared to other methods. This phenomenon would not occur for IAUC and IPGS because the masked pixels are the blurred ones from the input image, and the gradients between masked pixels and input pixels are smoother than with black.

The large gap between $H^2O$ and other methods on the sparsity metric may be explained by the threshold we apply on the heatmaps. We argue that no threshold should obtain lower sparsity scores, but better scores for the other metrics. We leave as future work the comparison of different threshold values. An in-depth study of the impact of the segmentation algorithm, the input image sizes, the objects' structures and color variability of the images, the correlation between the quality of the explanation and the confidence level of the model is also left as future work.

Lastly, key images are shown in Fig. 4 to illustrate $H^2O$ heatmaps when the model is uncertain of its predictions or makes mistakes. For example, the first row shows a *dogs* race, where only the dogs are highlighted ($H^2O$). These dogs are quite similar to most of the racehorses in the dataset, which could explain the horse prediction. We also noticed that 79% of the images containing a horse in the dataset also contain a person. This statistical fact may be enough to explain the prediction on the class *person*, along with the dog in the background that could be seen as a horseman. Though, it seems harder to interpret the prediction on Person with methods such as GradCAM or FEM since only the two foreground dogs are roughly highlighted. The model predicts a *boat* in the second image, with a probability of 0.61. All methods either highlight pixels of the sea or of the sky and clouds in their explanations. However, looking at GradCAM, FEM and RISE heatmaps, the large red zones seem to indicate the model is certain of its prediction, which is not the case. In comparison, both $H^2O$ and LIME produce heatmaps with almost no red shades. This may indicate a correlation between the heatmaps' scores and the predicted vector (IS, DS) that would need further statistical investigation to determine whether or not sea pixels only exist in the boat class. The last image contains small *cars* and *buses*. However, the model predicts a *train* with a probability of 0.99. We believe GradCAM and FEM explanations are less convincing than $H^2O$, RISE or even LIME. Indeed, RISE seems to indicate that the model confuses the central building with a station platform, a concept related to *trains*. Except the strong pixels of LIME in the sky that are hard to interpret, the road pixels may have been confused with a railroad, and the central building as a locomotive. Meanwhile, the strongest pixels of the building in $H^2O$ may have been confused with the locomotive wheels and side.

Despite its convincing visual results, $H^2O$ is computationally expensive because of the number of colors to replace the masked pixels which augment the number of images to predict. Our color strategy fixes to $z = 3$ the number of linear splits in each axis of the rgb cube, determining $3^3 = 27$ colors. Even if the predictions are batched, around $27 * 2500 = 67500$ altered images per image $I$ are processed through the model. On average, an explanation on a $224 \times 224$ image is computed in 55s of wall clock time, on a machine equipped with a 3.2 GHz processor, an NVIDIA GeForce RTX 3090 with a capacity of 24 GB, and 64 GB of RAM. We experimentally observed that choosing only a few colors randomly, drastically alters the quality of the output saliency maps. We leave as future work the search of a better color strategy to improve the explanations quality while reducing the number of colors.

## 7 CONCLUSION

This paper considers the problem of explaining which input features influence the decision process in an image classifier. To address this problem, we propose $H^2O$, an occlusion method to explain the prediction vector of an image classifier through a model and class-agnostic saliency map. Quantitatively, $H^2O$ outperforms existing techniques on most of the existing metrics. Qualitatively, $H^2O$ heatmaps are more localized than heatmaps of other methods. As future work, we could compare different colorset strategies to improve even more the quality of the saliency maps, or reduce their computation time. It may also be interesting to investigate the impact of the image segmentation algorithm on the resulting saliency maps. We finally plan to investigate the evaluation metrics to reduce the potential bias the masking color may have on them.

## ACKNOWLEDGMENTS

We acknowledge the Nouvelle-Aquitaine Region, Bordeaux Métropole and SUEZ, le LyRE for mainly funding and supporting this work through the Convention N°AAPR2020-2019-8171810.

## REFERENCES

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence* 34, 11 (2012), 2274–2282.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).

[3] Luca Bourroux, Jenny Benois-Pineau, Romain Bourqui, and Romain Giot. 2022. Multi Layered Feature Explanation Method for Convolutional Neural Networks. In *Pattern Recognition and Artificial Intelligence: Third International Conference, ICPRAI 2022, Paris, France, June 1–3, 2022, Proceedings, Part I*. Springer, 603–614.

[4] François Chollet et al. 2015. Keras. https://keras.io.

[5] Jessica Cooper, Ognjen Arandjelović, and David J Harrison. 2022. Believe the HiPe: Hierarchical perturbation for fast, robust, and model-agnostic saliency mapping. *Pattern Recognition* 129 (2022), 108743.

[6] Chabane Djeraba. 2002. Content-based multimedia indexing and retrieval. *IEEE multimedia* 9, 2 (2002), 18–22.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007 Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[8] Peter Flach. 2019. Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9808–9814.

[9] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2950–2958.

[10] Kazi Ahmed Asif Fuad, Pierre-Etienne Martin, Romain Giot, Romain Bourqui, Jenny Benois-Pineau, and Akka Zemmari. 2020. Features understanding in 3d cnns for actions recognition in video. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 1–6.

[11] Tristan Gomez, Thomas Fréour, and Harold Mouchère. 2022. Metrics for saliency map evaluation of deep learning explanation methods. In *Pattern Recognition and Artificial Intelligence: Third International Conference, ICPRAI 2022, Paris, France, June 1–3, 2022, Proceedings, Part I*. Springer, 84–95.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. 2011. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 6 (2011), 797–819.

[14] Biagio La Rosa, Graziano Blasilli, R Bourqui, D Auber, Giuseppe Santucci, Roberto Capobianco, Enrico Bertini, Romain Giot, and Marco Angelini. 2023. State of the Art of Visual Analytics for eXplainable Deep Learning. In *Computer Graphics Forum*. Wiley Online Library.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[16] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[17] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421* (2018).

[18] Luc-Etienne Pommé, Romain Bourqui, Romain Giot, and David Auber. 2022. Relative Confusion Matrix: Efficient Comparison of Decision Models. In *IV2022-IVE-Information Visualization Evaluation & User Studies*.

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.

[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[21] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[22] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[23] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

[24] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 818–833.

[25] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2018. Top-down neural attention by excitation backprop. *International Journal of Computer Vision* 126, 10 (2018), 1084–1102.

[26] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2018. Top-down neural attention by excitation backprop. *International Journal of Computer Vision* 126, 10 (2018), 1084–1102.

## A RELATED WORK SUMMARY

Table 2 summarizes the methods of the state-of-the-art.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

**Table 2: Summary of the key methods listed in Section 2.1, sorted by the family. Only MLFEM requires a training phase with an explanation ground truth. Only FEM and MLFEM generate explanations that do not depend on a class of interest.**

| | Explainer | | | | | Evaluation | | |
|---|---|---|---|---|---|---|---|---|
| **Paper** | **Fam.** | **MA** | **CA** | **GT** | **Hie.** | **Classifier** | **Dataset** | **Metric** |
| GradCAM [21] | Grad. | | | | | VGG16 | Imagenet | |
| Guided back-propagation. [23] | Grad. | | | | | | | |
| Deconvnet [24] | Grad. | | | | | Adhoc. network | Imagenet, VOC 2012, Caltech | Manual Occlusion |
| MWP [26] | Grad. | | | | | GoogLeNet, VGG16 | MSCOCO, VOC 2007, Imagenet | PG |
| FEM [10] | Act. | | ✓ | | | Adhoc. network | TTStroke-21 | SIM, PCC (against Grad-CAM) |
| MLFEM [3] | Act. | | ✓ | ✓ | | ResNet50 | MexCulture, Salicon, Cat2000 | SIM, PCC (against GFDM) |
| ANCHORS [20] | Alt. | ✓ | | | | MLP, Logistic Regression, Gradient boosted trees, Inception V3, Visual7W open-ended VQA system | Adult, Rcdv, lending, Imagenet, Visual7W Dataset | Precision, Coverage |
| LIME [19] | Occ. | ✓ | | | | Decision Tree, Logistic Regression, MLP, SVM, Random Forest | Books, DVDs, 20 newsgroups, Adhoc image dataset | Trustworthiness |
| RISE [17] | Occ. | ✓ | | | | ResNet50, VGG16 | MSCOCO, VOC 2007, Imagenet | PG, IAUC, DAUC |
| SHAP [16] | Occ. | ✓ | | | | Adhoc. network | MNIST | |
| HiPe [5] | Occ. | ✓ | | | ✓ | ResNet50 | MSCOCO 2014, VOC 2007 | PG, IAUC, DAUC |
| ExtP [9] | Occ. | ✓ | | | | GoogLeNet | MSCOCO, Imagenet | PG, Advers. SM detection |

*Fam.*: Method family (*Occ.*: Occlusion; *Grad.*: Gradient; *Act.*: Activation; *Alt.*: Alteration), *MA*: Model-Agnostic, *CA*: Class-Agnostic, *GT*: Ground Truth required to train the explainer, *Hie.*: Hierarchical method, *SIM*: Similarity, *PCC*: Pearson Correlation Coefficient, *PG*: Pointing Game, *GFDM*: Gaze Fixation Density Maps, *IAUC*: Insertion Area Under Curve, *DAUC*: Deletion Area Under Curve,