A Guided Tour of Post-hoc XAI Techniques in Image Segmentation

Syed Nouman Hasany, Fabrice Mériaudeau, Caroline Petitjean

A Guided Tour of Post-hoc XAI Techniques in Image Segmentation

Syed Nouman Hasany[1][0000−0002−5915−4528], Fabrice Mériaudeau[2][0000−0002−8656−9913], and Caroline Petitjean[1][0000−0003−0013−5370]

[1] Université de Rouen Normandie
{syed-nouman,caroline.petitjean}@univ-rouen.fr
[2] Université de Bourgogne
fabrice.meriaudeau@u-bourgogne.fr

Abstract. Deep learning models have shown tremendous gains in computer vision in the last decade. However, given their highly non-linear nature, they are often seen as black-boxes. This has led to the development of eXplainable Artificial Intelligence (XAI) as a parallel field with the aim of investigating the behavior of deep learning models. Research in XAI, however, has almost exclusively been focused on image classification models. Dense prediction tasks such as image segmentation have received little attention. The last few years have seen a shift in this trend with works focusing on exploring XAI in the context of image segmentation. A fair number of these works have borrowed from XAI techniques proposed in the context of image classification. It is safe to assume that going forward the number of XAI techniques focused on image segmentation are bound to increase. Reviewing the journey of XAI in image segmentation thus far would therefore be ideal, and is the goal of the present work. This review aims at presenting an overview of the XAI techniques proposed in the context of image segmentation. Another goal is to highlight the lack of interest in this field and its potential causes as well as to comment on potentially underexplored avenues. Given the relative nascency of the field, no review papers currently exist, a gap this work aims to fill.

Keywords: image segmentation · explainability · interpretability · XAI · review.

1 Introduction

AlexNet's [47] breakthrough on ImageNet [13] in 2012 proved to be the pivotal moment in the evolution of computer vision. Going forward, the impact of classical computer vision declined, and that of deep learning increased. Since then, the field has been dominated by deep learning architectures, the most prominent of them being convolutional neural networks (CNNs). In recent years, owing to their competitive performance, transformer based architectures such as the vision transformer [15] have also started becoming popular. Whereas AlexNet was proposed to solve an image classification task, deep learning's influence has since

broadened to include various other computer vision tasks as well such as image segmentation [54], object detection [91], pose estimation [92], image generation [87], etc.

Given the highly non-linear nature of deep learning algorithms, a parallel field of research soon came into existence, that of XAI (eXplainable Artificial Intelligence). Research in XAI is broadly concerned with understanding a deep learning model's behavior. For example, one sub-field concerns itself with trying to understand why a model arrived at a particular decision. The earliest techniques in this regard were proposed as occlusion analysis [88] and image-specific class saliency [73] in 2013 and 2014 respectively. These techniques were focused on identifying regions in an input image which were important for the model to arrive at its decision. This sub-field has received significant traction as is evident from the popularity of algorithms such as LIME [63], SHAP [52], RISE [61], Grad-CAM [70], etc. Another sub-field concerns itself with investigating the kind of representations learned by a model. An example is [73] in which representative examples of a given class are generated for a CNN. This allows researchers to identify patterns which a CNN has associated with each class.

While research in XAI commenced within an year of the deep learning revolution, the initial focus of the field remained almost exclusively on image classification. In recent years, however, dense prediction tasks, such as image segmentation, have also started receiving attention in the context of XAI. Due to the attention XAI in image classification has received, multiple survey papers detail the field's progression in this context [67,2]. No review paper, however, exists for XAI in image segmentation. We aim to fill this gap by providing an overview of the post-hoc techniques which have been proposed in the context of XAI in image segmentation.

One of the main reasons as to why image segmentation has received relatively little attention lies with the utility of a saliency map in the context of a dense prediction task. Given that the prediction is already in the spatial domain, it is arguable as to whether the end-user will find the saliency map to be of any use [17]. Nevertheless, XAI research in image segmentation is still dominated by saliency generation methods. Other than that, research has also focused on detecting biases within a segmentation model [38], identifying useful concepts learned by a segmentation model [17], generating insights about the segmentation model [58], etc.

2 Categorization of XAI Techniques for Image Segmentation

In order to search for relevant papers, we used two approaches. In the first approach, we searched for abstracts which mentioned **XAI**, **explainability**, or **interpretability**, along with **segmentation**[3] on Google Scholar. For the second approach, we considered some of the most widely cited papers relevant

[3] as well as morphological variants of these words.

to XAI in image segmentation [38,78,84,65]. Following a recursive approach, we exhaustively searched for works which had cited these papers followed by those which had cited the ones found in the previous step, and so forth. Ending up with **81** papers, we classified these into two categories. The first category consists of works in which techniques are proposed (**39** papers) whereas the second category consists of works in which techniques are utilized as tools (**42** papers). Barring passing mentions of the second category, in this review paper we have primarily focused on the first category. A complete list of techniques belonging to the first category can be found in Appendix A.

Techniques in XAI can generally be classified as belonging to local XAI or global XAI. Local XAI concerns itself with techniques focused on understanding the model's decisions for individual input samples. In this context the terms interpretability and explainability are often interchangeably used. We have opted for 'explainability', and consider algorithms falling under this category as those aiming at explaining the model's decision to an end user. These explanations generally take the form of saliency maps (also referred to as attribution maps or heatmaps). On the other hand, global XAI concerns itself with techniques focused on investigating the model's overall characteristics; an example of this is the feature space carved out by the model's internal representations. Another category which overlaps with both local and global XAI is that concerned with the concepts a model has learned. Figure 1 shows a taxonomy of the XAI algorithm categories discussed in this review.

3 Review of XAI Techniques for Image Segmentation

3.1 Local XAI

For local XAI in image classification, we are interested in explaining the model's output score (y^c) for a target class given an input image. However, due to image segmentation being a dense prediction task, there are three possibilities when it comes to what we are interested in explaining: (i) the entire segmentation map for a target class, (ii) a region of a segmentation map for a target class, or (iii) an individual pixel from the segmentation map for a target class. Going forward, we shall refer to the region to be explained as $\mathcal{M}$. In the discussion to follow, we categorize local XAI in image segmentation into three categories: (i) methods relying on the intermediate feature space, (ii) methods relying on the input space, and (iii) perturbation based methods.

Explainability Methods relying on the Intermediate Feature Space One of the most popular saliency generation methods in image classification has been Grad-CAM [71]. Grad-CAM generates saliency as the linear combination of the activation maps of a given model layer. This saliency for class c can be expressed as:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k \cdot A^k\right) \tag{1}$$

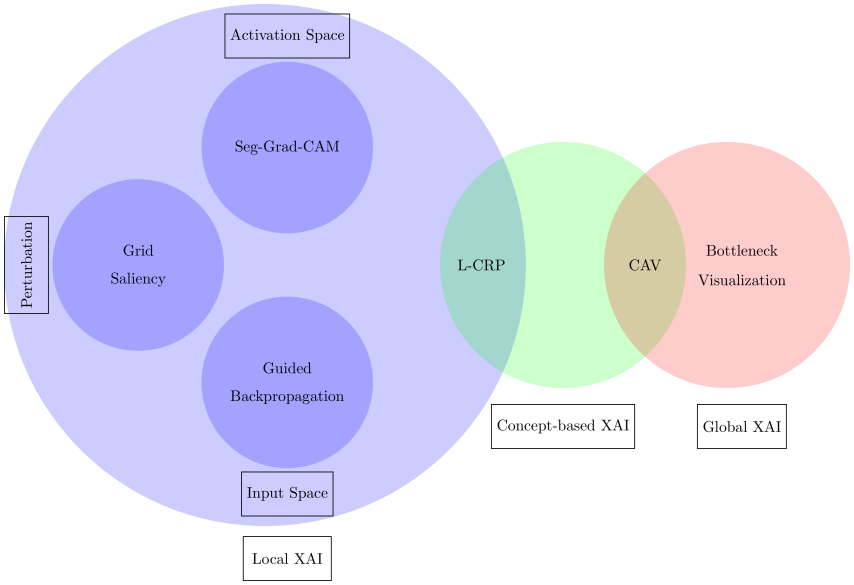

Fig. 1. Taxonomy of discussed XAI algorithms with a representative example for each category. Boxes indicates categories (and subcategories) whereas free text indicates instances of said categories.

A^k represents the kth activation map and α_k^c represents the kth coefficient. α_k^c is calculated in two steps. First, the gradient of the target class score with respect to the activation map (A^k) is computed, and this is followed by applying global average pooling (GAP) to this gradient matrix:

$$\alpha_k^c = GAP\left(\frac{\partial y^c}{\partial A^k}\right) \tag{2}$$

Since image segmentation returns a dense prediction instead of a single value, extending Grad-CAM to image segmentation required a necessary modification[4]. Primarily, two approaches can be tracked which aimed at extending Grad-CAM to image segmentation.

A pixel-wise application of Grad-CAM was proposed [80] leading to multiple saliency maps associated with individual pixels followed by an aggregation of these maps. Additionally, [80] also proposed the modification of Eq. (1) such that α_k^c is replaced by $\frac{\partial y_{i,j}^c}{\partial A^k}$ which is the gradient matrix, and the multiplication $(\cdot)$ is replaced by the element-wise multiplication $(\odot)$. This modification is proposed in order to avoid global average pooling (GAP) as GAP leads to a loss of spatial information owing to it collapsing the entire gradient matrix into a scalar. It is obvious that for a task such as image segmentation, this spatial information could have been crucial towards explaining the model's decision. This approach,

[4] Unless the goal is to explain the model's prediction for an individual output pixel.

named Grad-PAM, computes a map $L^c_{Grad-PAM}(i,j)$ for an individual pixel at location (i,j):

$$L^c_{Grad-PAM}(i,j) = ReLU\left(\sum_k \frac{\partial y^c_{i,j}}{\partial A^k} \odot A^k\right) \tag{3}$$

The most popular modification, however, is Seg-Grad-CAM [78] which works by replacing y^c in Eq. (2) with a sum of the target class' scores in the region of interest ($\mathcal{M}$). This leads us to α^c_k as being:

$$\alpha^c_k = GAP\left(\frac{\partial \sum_{(i,j)\in\mathcal{M}} y^c_{ij}}{\partial A^k}\right) \tag{4}$$

Grad-CAM has been considerably well-received in image classification and this has led to many derivatives such as Grad-CAM++ [1], Guided Grad-CAM [70], LayerCAM [42], etc. aimed at improving upon the original technique. Similarly, in image segmentation, Seg-Grad-CAM has led to techniques such as Seg-GradCAM++ [49,56,30], Seg-XGrad-CAM [30], Seg-Eigen-CAM [30], Seg-ScoreCAM [56,30], Seg-Ablation-CAM [30,27], and Seg-XRes-CAM [34]. As is obvious, these techniques are primarily extensions of their namesakes in image classification to image segmentation. It is worth noting that not all derivatives of Grad-CAM are gradient based; some such as Score-CAM, and Eigen-CAM bypass gradients in order to avoid the associated problems such as the noisy gradient problem [75].

Even though Seg-Grad-CAM (and its gradient-based derivatives) can be used to generate saliency maps when it comes to explaining an entire segmentation map for a target class, it is best avoided when it comes to the explanation of a region of interest within that segmentation map for a target class. This is attributable to the issue of spatial information collapsing due to the presence of global average pooling (GAP) [80]. If one wishes an explanation for the model's prediction of a region located near the top right of an image, it is unlikely that the bottom left region of the image would have made a significant contribution towards the model's decision. GAP, however, ensures that all spatial locations of an activation map (A^k) in Eq. (1) are multiplied by the same scalar α^c_k without any regard for their spatial relationship in the image.[5] Similar to the formulation of Eq. 3, Seg-XRes-CAM [34] and Seg-HiRes-Grad CAM [62] extend Seg-Grad-CAM such that the GAP is avoided when generating saliency maps for a region of interest given a target class.

Grad-CAM in image classification is generally applied to the final feature extraction convolutional layer. This layer is a good choice as its representations not only contain the effective summary of the original image, but these are also the sole convolutional representations responsible for the model's output decision. The situation in image segmentation, however, is not as straightforward.

[5] Non-gradient-based algorithms such as Seg-Ablation-CAM fail too since they are computing linear coefficients based on the complete spatial presence or absence of an activation map.

For a standard segmentation network such as a U-Net [64], the convolutional representations responsible for the final prediction (near the end of the decoder) do not contain a summary of the original image. A more likely candidate to contain the summary is the bottleneck layer (end of the encoder). The bottleneck's representations, however, are not directly responsible for the final segmentation output. These reservations notwithstanding, the bottleneck has been a common choice when it comes to applying Grad-CAM to image segmentation. Aiming to bypass this issue, [56] proposed an 'Adapted' alternative in which the technique is individually applied to all of the decoder's layers, and the obtained saliency maps are then aggregated by summing them up (appropriate transformations are applied in order to account for the differing spatial dimensions of the saliency maps). This adapted approach was utilized with Adapted Seg-Grad-CAM, Adapted Seg-GradCAM++, and Adapted Seg-ScoreCAM. Additionally, [56] also proposed a gradient-agnostic method of determining the linear coefficients such that the coefficient associated with each activation map is a product of the dependent and independent contribution of the activation map to the final segmentation. Dependence and independence, in this case, are defined such that they take into account the contribution of the activation map to the final prediction given its presence.

Explainability Methods relying on the Input Image Space Whereas Grad-CAM requires access to the intermediate activation space, many XAI techniques proposed in the context of image classification work directly with the input image space instead. Computing the saliency map as the derivative of the model's output score with respect to the input image was the earliest such method [73]. Owing to the noisy nature of the gradient signal in a deep network, techniques have been proposed which aim towards 'cleaning' this signal. An early technique proposed the inhibition of negative gradients in order to only highlight those gradient signals which contributed positively towards the model's decisions [76]. This method was one of the earliest XAI methods from image classification to be extended to image segmentation [84]. Other techniques proposed in the context of cleaning image gradient include SmoothGrad [75] which aims at generating saliency maps for multiple noisy versions of the input image followed by an averaging based aggregation. Similarly, Integrated Gradients [77] works by generating saliency maps for multiple versions of the input image each of which differs from the other in terms of the image brightness. This, too, is followed by an averaging based aggregation. Both of these method have been extended to image segmentation ([89], [33]). Similar to the problem of extending Grad-CAM to image segmentation, the necessary output modifications are applied to the final segmentation prediction before the application to these methods. Such modifications include applying global average pooling [39] to the final segmentation prediction or taking a summation over the target class' scores in the region of interest ($\mathcal{M}$).

Perturbation based Methods Another class of explainability methods are based on tracking the model's behavior as systematic modifications (perturbations) are applied to the input image. Perturbations which significantly impacted the model's output serve as indicators towards identifying salient regions in the original image. These techniques are categorized as perturbation based techniques. Given that they treat a model as a black box, they possess the distinct advantage of being model agnostic. This is advantageous because techniques belonging to the CAM family [93] are not guaranteed to work for non-convolutional architectures, a pressing example of which are architectures utilizing transformer blocks [8,35]. Grid saliency [38] was the earliest attempt of utilizing a perturbation based technique to image segmentation. This was an extension of the Meaningful Perturbation technique [21] originally proposed in the context of image classification. Grid saliency iteratively modifies an input image by removing irrelevant regions in order to finally arrive at the necessary context required by the segmentation model for its correct segmentation. This modification is guided by the gradient of a loss function, the aim of which is to simultaneously preserve the model's prediction as well as to get rid of unneccesary regions. Examining necessary contexts for various images allowed [38] to identify potential biases which the segmentation model might have learnt during its training.

Another popular perturbation based approach in image classification is SHAP [52] which works by dividing an image into superpixels, and then aims at distributing the overall contribution of each superpixel to the model's final decision. This is achieved by feeding the input image to the model multiple times such that on each occasion certain superpixels are masked and the remaining unmasked. Upon multiple iterations, this allows us to quantify the contribution of each superpixel towards the model's final decision. This approach was extended to image segmentation as well [12] with a slight modification. Instead of using algorithms such as SLIC [4] to identify superpixels, the authors reported better results if the input image is divided into a hexagonal grid structure, followed by treating each hexagon as a superpixel. The same work also extended RISE [61], another perturbation based approach, to image segmentation. For image classification, RISE starts by generating multiple random masks which are multiplied with the original image leading to multiple masked versions of the original image. The saliency map is then defined as the linear combination of these masks with the linear coefficients obtained using the model's classification scores associated with each masked image. Extending both RISE and SHAP to image segmentation is straightforward with the minor modification of averaging over the prediction scores of our target class in a region of interest ($\mathcal{M}$) in order to convert it to a scalar.

U-Noise [45] proposed a relatively novel approach in which an independent model is learned on top of the existing segmentation model (frozen) with the aim of directly predicting a saliency map given an input image. The goal of the independent model is to generate a noise mask, the addition of which to the original input image would not lead to a deterioration of the segmentation model's prediction. The rationale is that the predicted noise mask will contain more noise

for regions in the input image that are less important to the segmentation model and less noise for more important regions. An identical approach does not exist in image classification. Some similarities, however, can be seen in [11] in which a model is learned to predict a mask which is then element-wise multiplied with the input image before feeding it to the classification model. The goal in both cases [11,45] is to learn an additional model that can predict a mask (noise mask in the case of U-Noise) in real-time such that the application of this mask to the input image does not harm the original prediction. A modification to U-Noise's optimization procedure was recently proposed arguing for the inclusion of bilateral filtering to generate smooth noise masks [59].

Concept based Methods Instead of generating a single saliency map, concept based methods aim at identifying the useful concepts present in the image which might have contributed towards the final segmentation decision. [80] extended [79] to image segmentation in which the segmentation model's decision for a single pixel is decomposed into a decision tree. A saliency map can be generated for each decision node allowing us to track the conceptual journey undertaken by the model in order to arrive at its final decision. Another approach was proposed by [17] in which the explanation of a segmentation model's prediction is decomposed in terms of its concepts. Filters responsible for the detection of specific concepts are first identified in the segmentation model followed by a conditional application of Layer-wise Relevance Propagation (LRP) [6]. This allows us to highlight the contribution of individual filters leading to a disentangled concept-specific explanation.

3.2 Evaluation of Local XAI methods

The results of local XAI methods are visual, and therefore relatively qualitative in nature, a consequence of which has been a lack of an agreed upon evaluation methodology. Occasionally, proposed techniques have found it sufficient to simply display the saliency maps without providing a quantitative qualification of the proposed technique. Quantitative evaluation, however, is not completely absent, and some evaluation metrics have been proposed. One popular evaluation technique is faithfulness [66,67] which tracks the model's performance as the most significant pixels from the input image - as identified from the local XAI technique - are removed.

Specifically in the context of image segmentation, a pair of metrics were proposed [56] in order to evaluate the generated saliency maps. The saliency map is first used to mask the input image such that the unimportant pixels - as per the saliency map - are removed. This masked image is then fed to the segmentation model. The first of our metrics, 'Prediction Preserved Score' (PPS), records the percentage of the prediction which is preserved given this masked image as compared to the model's prediction on the original image. The second metric, 'Image Preserved Score' (IPS), acting as a complement to the first, records the percentage of the original image which is retained in the

masked image. The idea behind this pair is that a good saliency map would score highly on the PPS and low on the IPS.

An evaluation strategy utilizing the U-Noise model was recently proposed [72]. The saliency map is first utilized to mask the input image. This masked image is then fed to the U-Noise model. The statistics of the generated noise mask serve as our evaluation metric with the rationale being that less noise would be added if the saliency map was correct in its identification of the important regions.

3.3 A Comparative Analysis of Local XAI Methods

Figure 2 shows the saliency maps generated using Seg-Grad-CAM, Seg-XRes-CAM, Seg-Eigen-CAM, Seg-AblationCAM, and RISE on a few samples from the COCO-2017 dataset [50]. The first four of these algorithms belong to the intermediate feature space category whereas RISE is a perturbation based method. Between the first four methods, Seg-Grad-CAM and Seg-XRes-CAM are gradient dependent whereas Seg-Eigen-CAM[6] and Seg-AblationCAM are gradient independent. The segmentation model was a pre-trained DeepLabv3[7], and the intermediate feature space methods were applied to the bottleneck layer.

From a qualitative point of view, a few things are observable. First, some parts of the object appear more salient as compared to other parts. For example, in the case of 'Cat', the ears and the nose appear more prominent compared to other regions of the cat. Similarly, for the 'Dog' case, the nose is comparatively brighter than the rest of the body in all saliency maps barring that of Seg-Eigen-CAM. Secondly, except for Seg-Eigen-CAM, the three intermediate feature space algorithms broadly agree with each other. For Seg-Eigen-CAM, in the cases of 'Dog', 'Bike', and 'Train', the saliency maps, interestingly, highlight the background instead of the object itself - almost a complement of the other saliency maps in the same row. Lastly, even though saliency maps for RISE broadly agree with the rest as far as general localization is concerned, the maps are much coarser as compared to the others. Where the intermediate feature space methods might highlight specific regions of interest in some detail, the same cannot be expected from RISE.

For the same dataset and segmentation model, Figure 3 shows the saliency maps generated using RISE with varying number of masks. Unsurprisingly, the number of masks play an important role in the saliency map generation. It appears that between 100 masks and 2000 masks, the saliency map becomes more refined in the sense of being more concentrated around the object of interest. An interesting case presents itself in the third row where the segmentation model has predicted two cats whereas it is obvious that the left one is a dog. Saliency maps generated from RISE prominently highlight the correct cat whereas the incorrect dog is almost completely ignored.

[6] A gradient driven version of Seg-Eigen-CAM is also possible

[7] `https://pytorch.org/hub/pytorch_vision_deeplabv3_resnet101/`

For a quantitative comparison we utilize a slighlty modified version of the pair of metrics from [56]. The saliency map is first binarized, and this binarized version is then used to mask out the original image. This masked image is then passed on to the segmentation model. We define our first metric, Dice Explained, as the dice score between the segmentation model's prediction on this masked image and the segmentation model's prediction on the original image. Our second metric is defined as the ratio of the number of pixels in the binarized saliency map to the number of pixels in the prediction. Figure 4 summarizes our metric calculation process. For binarization purposes, thresholds of 0.05 and 0.1 are experimented with for intermediate feature space methods, and thresholds of 0.2 and 0.4 are used with RISE.

Table 1 summarizes the results of applying our methods on a subset of the COCO-2017 dataset. RISE with 500 masks reports the best dice explained of **0.948**, whereas for intermediate feature space methods, Seg-AblationCAM leads at **0.843**. The saliency ratios for the two are **10.59** and **6.68** respectively. The worst dice explained with RISE is **0.747** (2000 masks) with a saliency ratio of **2.196**, and for intermediate feature space methods, it is that of Seg-Eigen-CAM at **0.354** with a saliency ratio of **4.83**. While it might be tempting to decide in favor of RISE or Seg-AblationCAM when it comes to generating saliency maps for image segmentation, the computational time paints a disheartening picture with average times around **2.5** minutes for Seg-AblationCAM and **1.5** to **5.5** minutes for RISE. Both Seg-Grad-CAM and Seg-XRes-CAM, operate much more swiftly taking less than a second on average. This is easy to explain as both RISE (a perturbation based method) and Seg-AblationCAM (an intermediate feature space method) require multiple iterations before they can compute the necessary coefficients in order to generate a saliency map. Seg-Grad-CAM and Seg-XRes-CAM, on the other hand, only require a single forward and backward pass.

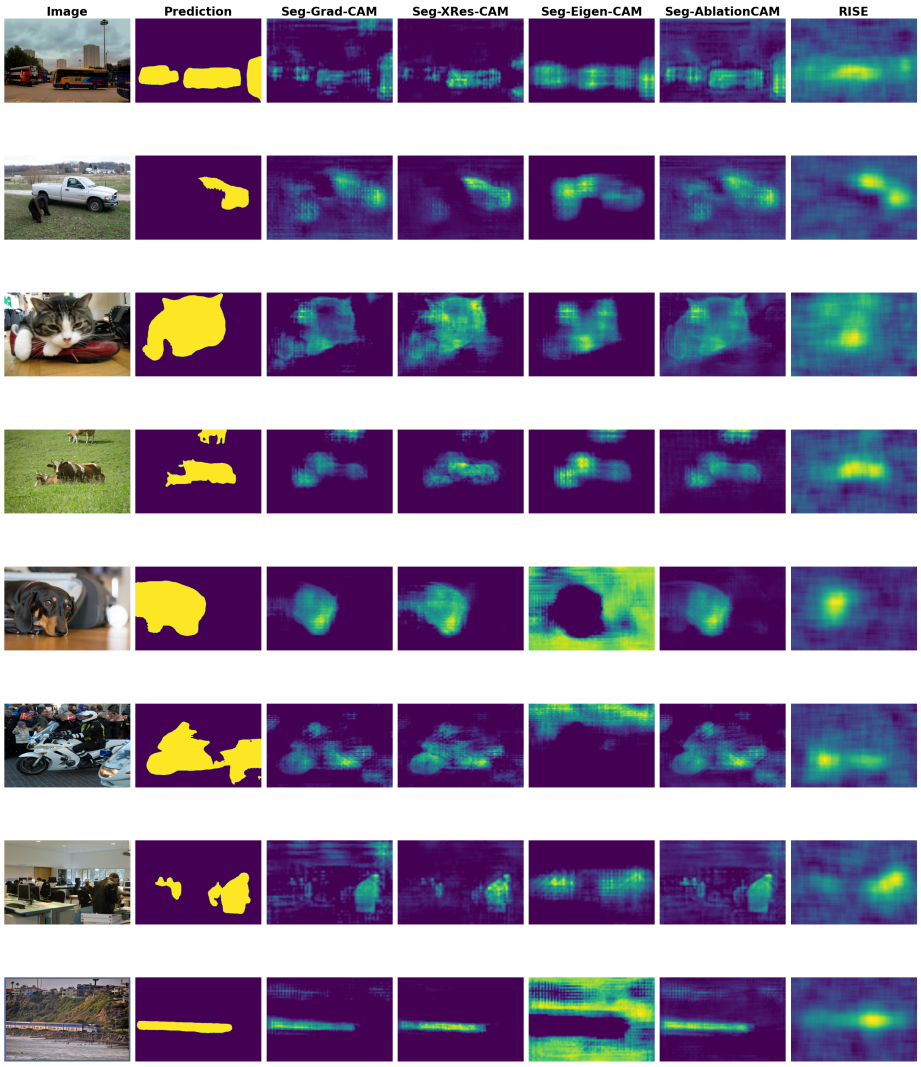

Fig. 2. Sample saliency maps generated from the application of Seg-Grad-CAM, Seg-XRes-CAM, Seg-Eigen-CAM, Seg-AblationCAM, and RISE on a few samples from the COCO-2017 dataset. For RISE, 2000 masks were utilized.

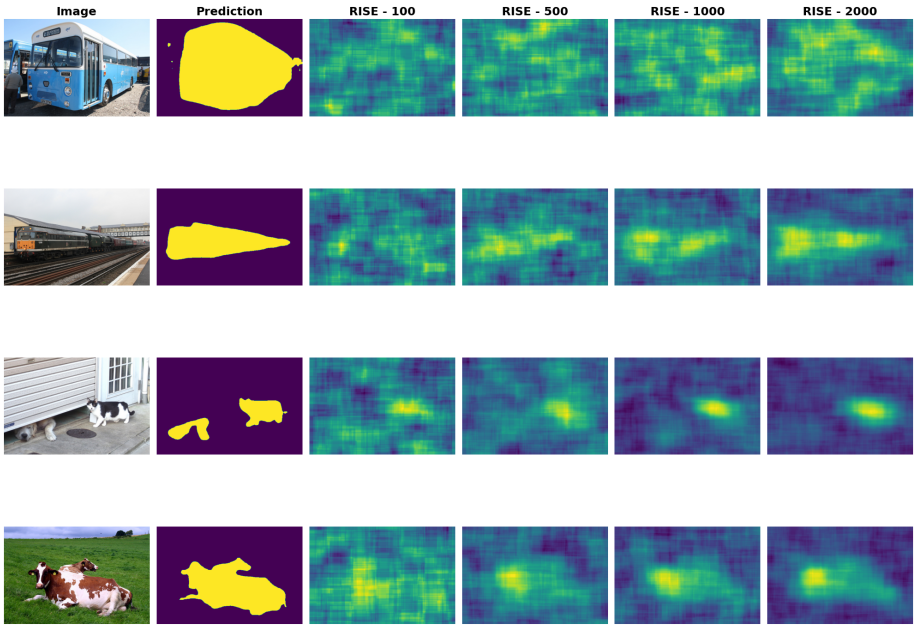

Fig. 3. Sample saliency maps generated from the application of RISE with various number of masks (100, 500, 1000, 2000) on a few samples from the COCO-2017 dataset.

3.4 Global XAI

Similar to image classification, global XAI approaches in image segmentation are considerably sparse compared to local XAI approaches. [10] investigated the sensitivity of a segmentation model to image features using an approach borrowed from activation maximization. An image for which the segmentation model predicts the background class for a region of interest is used as the initial image. The goal is for this image to be modified such that the model predicts the desired foreground class for the region of interest. The derivative of the foreground class' activation with respect to the input image serves as the guiding signal which gradient ascent uses in order to modify the initial image. This can be formulated as follows:

$$X_{j+1} = X_j + \alpha \cdot \frac{\partial i}{\partial X_j} \tag{5}$$

X_j represents the image at iteration j, whereas i is the output activation one wishes to maximize. α is the learning rate.

Table 1. Comparison between Seg-Grad-CAM, Seg-XRes-CAM, Seg-Eigen-CAM, Seg-Ablation-CAM, and RISE on a subset of the COCO-2017 Dataset. The first four algorithms are representatives of the intermediate feature space methods whereas RISE represents the perturbation based methods. For dice explained, saliency ratio, and time, the mean value has been reported.

XAI Method	**Binary Threshold**	**Gradient-based**	**No. of Masks**	**Dice Explained**	**Saliency Ratio**	**Time (s)**
Seg-Grad-CAM	0.05	✓	-	0.606	3.47	0.25
Seg-Grad-CAM	0.1	✓	-	0.515	2.403	0.25
Seg-XRes-CAM	0.05	✓	-	0.671	1.788	0.26
Seg-XRes-CAM	0.1	✓	-	0.636	**1.25**	0.26
Seg-Eigen-CAM	0.05	✗	-	0.421	5.166	7.67
Seg-Eigen-CAM	0.1	✗	-	0.354	4.831	7.67
Seg-AblationCAM	0.05	✗	-	**0.843**	6.675	163.4
Seg-AblationCAM	0.1	✗	-	0.793	3.859	163.4
RISE	0.2	✗	100	0.927	11.155	16.1
RISE	0.4	✗	100	0.853	7.773	16.1
RISE	0.2	✗	500	**0.948**	10.59	80.4
RISE	0.4	✗	500	0.766	3.884	80.4
RISE	0.2	✗	1000	0.942	9.282	160.8
RISE	0.4	✗	1000	0.762	3.154	160.8
RISE	0.2	✗	2000	0.913	7.053	321.2
RISE	0.4	✗	2000	0.747	**2.196**	321.2

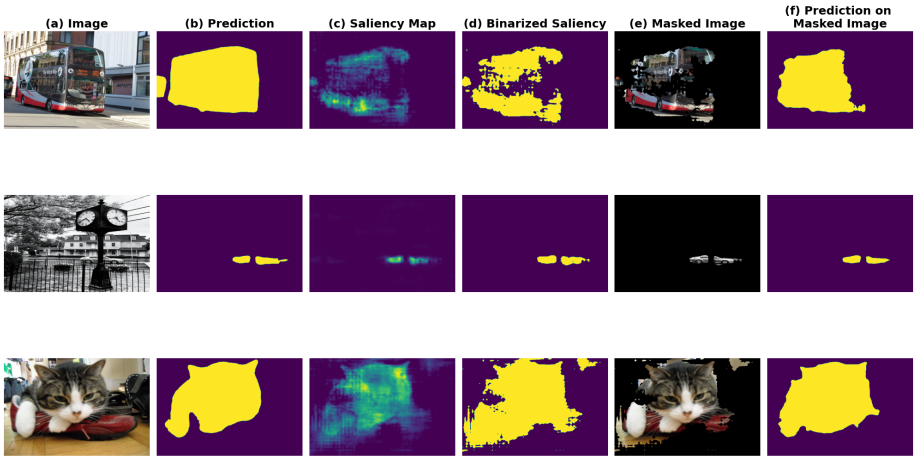

Fig. 4. Examples (from the COCO-2017 dataset) in order to explain the process of computing metrics for evaluating Saliency Maps for Image Segmentation. (a) Original Image, (b) Prediction obtained after passing (a) through the segmentation model, (c) Generated Saliency Map, (d) Binarizing the Saliency Map with a threshold (e) Masking the original image using the binarized Saliency Map (f) Obtaining the segmentation model's prediction on this Masked Image. The Dice Explained is then defined as the dice between (b) and (f). The Saliency Ratio is defined as the ratio of the number of non-zero pixels in (d) to the number of non-zero pixels in (b). For the current examples, the Dice Explained are: **0.81**, **0.86**, and **0.94** respectively. The Saliency Ratios are: **0.8**, **1.42**, and **1.45** respectively. For the current examples, the saliency maps were generated using Seg-XRes-CAM, and a threshold of **0.1** was utilized for binarization.

The path taken by the image is referred to as the DeepDream path with the idea being that by taking the steepest route, the image would have followed a path whereby only the most important features would have been modified. Accross multiple images, this analysis allows one to identify features to which a model is most or least sensitive to.

[58] too utilized activation maximization in order to investigate the kind of concepts being learned in various filters of an image segmentation model. An image is randomly initialized, and in order to determine the kind of concepts a filter might be looking for in an input image, gradient ascent is again utilized such that it is now guided by the derivative of an intermediate activation (that of the filter of interest) instead of the prediction space. Gradually, the random image transforms such that it is now dominated by the concept which the filter of interest is most activated by.

[41] explored the relationship of the segmentation model's bottleneck representations with the segmentation model's Intersection over Union (IoU) scores. Bottleneck representations of sample images followed by an application of dimensionality reduction serve as the input whereas the corresponding IoU scores for each of those images serves as the output. A simple regression model is then trained utilizing these input-output pairs allowing us to assess the segmentation model's predictions for unseen images whose ground truths are not available.

Another work [40] focused on the identification of useful concepts learned by an image segmentation model on a cardiac dataset. This was achieved with the help of Automated Concept-based Explanation (ACE) [25] which, in turn, is an extension of Concept Activation Vectors (CAV) [44]. Superpixels are first generated from individual images using SLIC [4]. These are then fed to the segmentation model in order to obtain their bottleneck representations. Clustering is performed on these representations, and each cluster center is identified as a concept. A Concept Activation Vector (CAV) is then learned for each concept in order to determine its global importance for each segmentation category.

4 Tools for Practitioners

Existing tools for XAI in image segmentation are mostly focused on local methods. One of the earliest toolbox was developed by [69] which allows for the application of various local methods belonging to the intermediate feature space as well as the input image space. Another similar, but more recent, toolbox is [89] which provides seven local XAI algorithms. [26] provides a library which contains most of the CAM-based methods (intermediate feature space). The library is designed to work for both image classification as well as image segmentation[8]

5 Discussion

In terms of their application, the task for which XAI algorithms in image segmentation have been most utilized for is the generation of saliency maps, with Seg-Grad-CAM [78] being the favored choice of practitioners - see [32,85] for medical examples, and [43,9] for examples from natural images. However, another application of these saliency generation methods has been found in investigating the intermediate representations of segmentation models. The application of a saliency generation method to various layers of a segmentation model allows one to inspect the flow of information [48,53]. Additionally, if a custom layer is added to a segmentation model, these XAI methods can allow one to monitor whether the layer's representations follow the outcome one expected from the layer's design [82,51]. Another interesting application has been the identification of how much context is required by a segmentation model in order to successfully segment an object. An analysis of context information can allow one to

[8] See also: `https://jacobgil.github.io/pytorch-gradcam-book/Class%20Activation%20Maps%20for%20Semantic%20Segmentation.html`

identify whether the segmentation model has associated an object with spurious context [38]. Association with spurious context can be harmful as the segmentation model would fail to successfully segment the object in the absence of that context. Identifying such associations can help developers refine the training pipeline by introducing enough variation into the training data in order to minimize the possibility of a segmentation model learning irrelevant contexts.

While intermediate feature space dependent saliency generation methods have enjoyed considerable popularity for convolutional architectures, it is clear that with the advent of alternate architecture designs such as those incorporating transformer blocks [8,35], these methods would have to be accordingly adjusted or new methods be proposed. Given that the receptive field of transformer models is fundamentally different from convolutional models, it remains an open question as to which layers need to be utilized in order to generate a saliency map.

With concept-based XAI, researchers have mostly been focused on identifying the various concepts which have been learned by the intermediate activations of the segmentation model [58,40]. However, a major issue in this regard, particularly for medical datasets, is concept identification. Concepts identified by automatic concept generation algorithms are hard to associate with tangible medical concepts, whereas labelled medical concept datasets using which a segmentation model can be probed are almost non-existent. For natural images, the problem is considerably alleviated due to the availability of a generous amount of online data regarding virtually any concept as well as due to the fact that one often doesn't require special training in order to interpret concepts associated with natural images.

6 Conclusion

The methods discussed above consistently reveal a common theme: XAI techniques used in image segmentation are derived from those employed in image classification. [78]'s Seg-Grad-CAM is an extension of Grad-CAM [71], [12] being an extension of RISE [61] and SHAP [52], and [17] being an extension of [5], for example. Some exceptions to this rule include [41] and [45]. Despite its widespread use, the heavy dependence on XAI in classification is not straightforward to justify. Classification and segmentation are inherently distinct tasks: the former seeks to assign a single label to the entire image, while the latter involves a dense prediction task, aiming to assign a label to each pixel. Since the output of image segmentation is a spatial map, the effectiveness of generating saliency maps is diminished, as they often merely emphasize the object they were meant to clarify [17].

Instead of merely extending XAI approaches from image classification to image segmentation, it would be better if XAI in image segmentation was treated as an independent field by researchers. The initial focus should then be towards identifying potential applications of XAI in image segmentation as that would organically lead to the proposal of XAI methods fine-tuned to image segmen-

tation. In the medical field, for example, saliency maps in image classification can aid a medical practitioner identify important regions in an image, whether the same role can be played by saliency maps of image segmentation is hardly obvious. Given the ubiquitous nature of image segmentation for medical images, the involvement of medical practitioners can help identify useful end goals for the field of XAI in medical image segmentation. Once potential applications have been identified, the design of XAI algorithms would then naturally be tailored towards fulfilling those roles allowing for informed borrowing from XAI algorithms in image classification. Taking inspiration from XAI in image classification is only natural, however, simple imitation should be discouraged owing to the fundamental difference between the two sets of computer vision problems.

Acknowledgments. The authors would like to acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant Project-ANR-21-CE23-0013 (project MediSEG).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. A. Chattopadhay, A. Sarkar, P.H., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (2018)
2. Abhishek, K., Kamath, D.: Attribution-based xai methods in computer vision: A review. arXiv preprint arXiv:2211.14736 (2022)
3. Abtahi, M., Le, D., Lim, J.I., Yao, X.: Mf-av-net: an open-source deep learning network with multimodal fusion options for artery-vein segmentation in oct angiography. Biomed. Opt. Express (2022)
4. 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 (2012)
5. Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From" where" to" what": Towards human-understandable explanations through concept relevance propagation. arXiv preprint arXiv:2206.03208 (2022)
6. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE (2015)
7. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. Proceedings of the IEEE conference on computer vision and pattern recognition (2017)
8. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arxiv preprint, arXiv:2102.04306 (2021)
9. Chen, T., Jiang, D., Li, R.: Swin transformers make strong contextual encoders for vhr image road extraction. IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium (2022)

10. Couteaux, V., Nempont, O., Pizaine, G., Bloch, I.: Towards interpretability of segmentation networks by analyzing deepdreams. MICCAI Workshop on Interpretability of Machine Intelligence in Medical Image Computing (2019)
11. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. Advances in neural information processing systems (2017)
12. Dardouillet, P., Benoit, A., Amri, E., Bolon, P., Dubucq, D., Crédoz, A.: Explainability of image semantic segmentation through shap values. ICPR Workshop on Explainable and Ethical AI (2022)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database (2009)
14. Desai, S., Ramaswamy, G.H.: Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) (2020)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
16. Draelos, R.L., Carin, L.: Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. arxiv preprint, arXiv:2011.08891 (2020)
17. Dreyer, M., Achtibat, R., Wiegand, T., Samek, W., Lapuschkin, S.: Revealing hidden context bias in segmentation and object detection through concept-specific explanations. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023)
18. Dreyer, M., Achtibat, R., Wiegand, T., Samek, W., Lapuschkin, S.: Revealing hidden context bias in segmentation and object detection through concept-specific explanations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 3828–3838 (2023)
19. Fel, T., Cadène, R., Chalvidal, M., Cord, M., Vigouroux, D., Serre, T.: Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. Advances in neural information processing systems (2021)
20. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research (2019)
21. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
22. Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., Li, B.: Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. arXiv preprint arXiv:2008.02312 (2020)
23. Gan, Y., Mao, Y., Zhang, X., Ji, S., Pu, Y., Han, M., Yin, J., Wang, T.: " is your explanation stable?" a robustness evaluation framework for feature attribution. Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (2022)
24. Garret, G., Vacavant, A., Frindel, C.: Xai-vesselnet: explain liver vessel segmentation by a graph-based approach (2023)
25. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. Advances in neural information processing systems (2019)
26. Gildenblat, J., contributors: Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam` (2021)
27. Gipiškis, R., Chiaro, D., Annunziata, D., Piccialli, F.: Ablation studies in activation maps for explainable semantic segmentation in industry 4.0. IEEE EUROCON 2023 - 20th International Conference on Smart Technologies (2023)

28. Gipiškis, R., Chiaro, D., Preziosi, M., Prezioso, E., Piccialli, F.: The impact of adversarial attacks on interpretable semantic segmentation in cyber–physical systems. IEEE Systems Journal (2023)
29. Gipiškis, R., Kurasova, O.: Occlusion-based approach for interpretable semantic segmentation. 2023 18th Iberian Conference on Information Systems and Technologies (CISTI) (2023)
30. Gizzini, A.K., Shukor, M., Ghandour, A.J.: Extending cam-based xai methods for remote sensing imagery segmentation. arXiv preprint arXiv:2310.01837 (2023)
31. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
32. Gunashekar, D.D., Bielak, L., Hägele, L., Berlin, A., Oerther, B., Benndorf, M., Grosu, A., Zamboglou, C., Bock, M.: Explainable ai for cnn-based prostate tumor segmentation in multi-parametric mri correlated to whole mount histopathology (2022)
33. Habib, N.: Cascaded u-net++ for segmentation of lung (2021)
34. Hasany, S.N., Petitjean, C., Mériaudeau, F.: Seg-xres-cam: Explaining spatially local regions in image segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2023)
35. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: Unetr: Transformers for 3d medical image segmentation. WACV (2022)
36. He, S., Feng, Y., Grant, P.E., Ou, Y.: Segmentation ability map: Interpret deep features for medical image segmentation. Medical image analysis (2023)
37. Heide, N.F., Müller, E., Petereit, J., Heizmann, M.: X3seg: Model-agnostic explanations for the semantic segmentation of 3d point clouds with prototypes and criticism. 2021 IEEE International Conference on Image Processing (ICIP) (2021)
38. Hoyer, L., Munoz, M., Katiyar, P., Khoreva, A., Fischer, V.: Grid saliency for context explanations of semantic segmentation. NeurIPS (2019)
39. Humer, C., Elharty, M., Hinterreiter, A., Streit, M.: Interactive Attribution-based Explanations for Image Segmentation. EuroVis 2022 - Posters (2022)
40. Janik, A., Dodd, J., Ifrim, G., Sankaran, K., Curran, K.: Interpretability of a deep learning model in the application of cardiac mri segmentation with an acdc challenge dataset. Medical Imaging 2021: Image Processing (2021)
41. Janik, A., Sankaran, K., Ortiz, A.: Interpreting Black-Box Semantic Segmentation Models in Remote Sensing Applications. Machine Learning Methods in Visualisation for Big Data (2019)
42. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: Exploring hierarchical class activation maps for localization. IEEE Transactions on Image Processing (2021)
43. Joshi, I., Utkarsh, A., Kothari, R., Kurmi, V.K., Dantcheva, A., Roy, S.D., Kalra, P.K.: Sensor-invariant fingerprint roi segmentation using recurrent adversarial learning. 2021 International Joint Conference on Neural Networks (IJCNN) (2021)
44. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). International conference on machine learning (2018)
45. Koker, T., Mireshghallah, F., Titcombe, T., Kaissis, G.: U-noise: Learnable noise masks for interpretable image segmentation. 2021 IEEE International Conference on Image Processing (ICIP) (2021)
46. Kori, A., Natekar, P., Srinivasan, B., Krishnamurthi, G.: Interpreting deep neural networks for medical imaging using concept graphs. AI for Disease Surveillance and Pandemic Intelligence: Intelligent Disease Detection in Action (2022)

47. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NIPS (2012)
48. Lai, Z., Guo, R., Xu, W., Hu, Z., Mifflin, K., DeCarli, C., Dugger, B.N., ching Cheung, S., Chuah, C.N.: Automated segmentation of amyloid-βstained whole slide images of brain tissue. bioRxiv (2021)
49. Lei, J.: Interpretation of semantic urban scene segmentation for autonomous vehicles (2022)
50. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context. ECCV (2014)
51. Liu, Z., Guo, F., Liu, H., Xiao, X., Tang, J.: Cmlocate: A cross-modal automatic visual geo-localization framework for a natural environment without gnss information. IET Image Processing (2023)
52. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions. NIPS (2017)
53. Melching, D., Strohmann, T., Requena, G., Breitbarth, E.: Explainable machine learning for precise fatigue crack tip detection. Scientific Reports (2022)
54. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence (2021)
55. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. 2020 international joint conference on neural networks (IJCNN) (2020)
56. Mullan, S., Sonka, M.: Visual attribution for deep learning segmentation in medical imaging. Medical Imaging 2022: Image Processing (2022)
57. Mullan, S., Sonka, M.: Kernel-weighted contribution: a method of visual attribution for 3d deep learning segmentation in medical imaging. Journal of Medical Imaging (2023)
58. Natekar, P., Kori, A., Krishnamurthi, G.: Demystifying brain tumour segmentation networks: Interpretability and uncertainty analysis. Frontiers in Computational Neuroscience (2020)
59. Okamoto, T., Gu, C., Yu, J., Zhang, C.: Generating smooth interpretability map for explainable image segmentation. 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE) (2023)
60. O'Sullivan, C., Coveney, S., Monteys, X., Dev, S.: Interpreting a semantic segmentation model for coastline detection. 2023 Photonics & Electromagnetics Research Symposium (PIERS) (2023)
61. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. BMVC (2018)
62. Rheude1, T., Wirtz, A., Wesarg, S., Kuijper, A.: Leveraging cam algorithms for explaining medical semantic segmentation. iMIMIC at MICCAI 2023 (2023)
63. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. ACM SIGKDD (2016)
64. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. MICCAI (2015)
65. Saleem, H., Shahid, A.R., Raza, B.: Visual interpretability in 3d brain tumor segmentation network. Computers in Biology and Medicine (2021)
66. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems (2016)

67. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE (2021)
68. Santamaria-Pang, A., Kubricht, J., Chowdhury, A., Bhushan, C., Tu, P.: Towards emergent language symbolic semantic segmentation and model interpretability. Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23 (2020)
69. Schorr, C., Goodarzi, P., Chen, F., Dahmen, T.: Neuroscope: An explainable ai toolbox for semantic segmentation and image classification of convolutional neural nets. Applied Sciences (2021)
70. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision (2020)
71. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. IJCV (2019)
72. Shreim, H., Gizzini, A.K., Ghandour, A.J.: Trainable Noise Model as an XAI evaluation method: application on Sobol for remote sensing image segmentation. arXiv e-prints (2023)
73. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. Workshop at International Conference on Learning Representations (2014)
74. Singh, D., Somani, A., Horsch, A., Prasad, D.K.: Counterfactual explainable gastrointestinal and colonoscopy image segmentation. 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) (2022)
75. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. ICML Workshop on Visualization for Deep Learning (2017)
76. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for simplicity: The all convolutional net. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings (2015)
77. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. ICML (2017)
78. Vinogradova, K., Dibrov, A., Myers, G.: Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). AAAI (2020)
79. Wan, A., Dunlap, L., Ho, D., Yin, J., Lee, S., Jin, H., Petryk, S., Bargal, S.A., Gonzalez, J.E.: Nbdt: neural-backed decision trees. arXiv preprint arXiv:2004.00221 (2020)
80. Wan, A., Ho, D., Song, Y., Tillman, H., Bargal, S.A., Gonzalez, J.E.: Segnbdt: Visual decision rules for segmentation. arXiv preprint arXiv:2006.06868 (2020)
81. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020)
82. Wang, L., Huang, J., Xing, X., Yang, G.: Swin deformable attention hybrid u-net for medical image segmentation. arXiv preprint arXiv:2302.14450 (2023)

83. Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. 2018 ieee 28th international workshop on machine learning for signal processing (mlsp) (2018)
84. Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Medical image analysis (2020)
85. Wu, J., Liu, Z., Gou, F., Zhu, J., Tang, H., Zhou, X., Xiong, W.: Ba-gca net: Boundary-aware grid contextual attention net in osteosarcoma mri image segmentation. Computational Intelligence and Neuroscience (2022)
86. Xiao, M., Zhang, L., Shi, W., Liu, J., He, W., Jiang, Z.: A visualization method based on the grad-cam for medical image segmentation model. 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS) (2021)
87. Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: A review. Medical image analysis (2019)
88. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. ArXiv (2013)
89. Zeineldin, R.A., Karar, M.E., Elshaer, Z., Coburger, J., Wirtz, C.R., Burgert, O., Mathis-Ullrich, F.: Explainability of deep neural networks for mri analysis of brain tumors. International Journal of Computer Assisted Radiology and Surgery (2022)
90. Zemni, M., Chen, M., Zablocki, É., Ben-Younes, H., Pérez, P., Cord, M.: Octet: Object-aware counterfactual explanations. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2023)
91. Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X.: Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems (2019)
92. Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M.: Deep learning-based human pose estimation: A survey. ACM Computing Surveys (2023)
93. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. CVPR (2016)

A Reviewed XAI Algorithms

Our search yielded **81** papers out of which **39** were works responsible for initially proposing a technique, and **42** were works which utilized these techniques as tools. Tables 2,3,4 list works belonging to the former category. This ends up being **44** proposed techniques (as some papers proposed more than one technique). Additionally, the datasets on which these techniques were applied in their parent papers are also mentioned with a symmetrical distribution of **31** datasets each for both medical as well as natural images.

Table 2. Local XAI algorithms considered for this paper. Bold is used to indicate medical datasets. Additionally, if the algorithm is based on an existing algorithm from image classification, the predecessor is also identified. Key: **IFS** = Intermediate Feature Space.

Algorithm	Family	Gradient-based	Dataset	Precursor in Image Classification
Grad-PAM [80]	Local (IFS)	✓	Pascal-Context Cityscapes Look Into Person	Grad-CAM [71]
Seg-Grad-CAM [78,58]	Local (IFS)	✓	Cityscapes	Grad-CAM [71]
improved Grad-CAM [86]	Local (IFS)	✓	**Polyp (Colonoscopy)** **Liver Tumor (CT)** **Skin Lesion**	Grad-CAM [71]
Seg-GradCAM++ [49,56]	Local (IFS)	✓	Cityscapes **Brain Tumor (MRI)**	Grad-CAM++ [1]
Seg-ScoreCAM [56]	Local (IFS)	✗	**Brain Tumor (MRI)** Satellite Imaging	Score-CAM [81]
Seg-XGrad-CAM [30]	Local (IFS)	✓	Satellite Imaging	XGrad-CAM [22]
Seg-Eigen-CAM [30]	Local (IFS)	✗	Satellite Imaging	Eigen-CAM [55]
Seg-AblationCAM [30,27]	Local (IFS)	✗	Satellite Imaging Industrial Dataset	Ablation-CAM [14]
Seg-XRes-CAM [34]	Local (IFS)	✓	**Multi-organ (CT)** COCO-2017	HiResCAM [16]
Seg-Hires-Grad CAM [62]	Local (IFS)	✓	**Teeth (X-Ray)** **Kidney (CT)** Cityscapes	HiResCAM [16]
Seg-Sobol [72]	Local (IFS)	✗	Cityscapes Satellite Imaging	Sobol XAI [19]
Adapted Seg-Grad-CAM [56]	Local (IFS)	✓	**Brain Tumor (MRI)**	-
Adapted Seg-GradCAM++ [56]	Local (IFS)	✓	**Brain Tumor (MRI)**	-
Adapted Seg-ScoreCAM [56]	Local (IFS)	✗	**Brain Tumor (MRI)**	-
Kernel-Weighted [56,57]	Local (IFS)	✗	**Brain Tumor (MRI)**	-
Activations × Predictions [65]	Local (IFS)	✗	**Brain Tumor (MRI)**	-
Attribution-based Explanation [39]	Local (IFS)	-	Fungi	-
Score Maps [69]	Local (IFS)	✓	nuScenes	-

Table 3. Local XAI algorithms considered for this paper (cont.). Bold is used to indicate medical datasets. Additionally, if the algorithm is based on an existing algorithm from image classification, the predecessor is also identified. Key: **IS** = Input Space, **P** = Perturbation, **C** = Concept.

Algorithm	Family	Gradient-based	Dataset	Precursor in Image Classification
Guided backpropagation[83,84]	Local (IS)	✓	**Polyp (colonoscopy)**	Guided backpropagation [76]
Integrated Gradient [33]	Local (IS)	✓	**Lungs (CT)**	Integrated Gradient [77]
SmoothGrad [89]	Local (IS)	✓	**Brain Tumor (MRI)**	SmoothGrad [75]
Vanilla Saliency [3]	Local (IS)	✓	**Artery-Vein (OCT Angiography)**	Vanilla Saliency [73]
Grid Saliency [38]	Local (P)	✓	Cityscapes	Meaningful Perturbation [21]
SHAP [12]	Local (P)	✗	SAR Cityscapes	SHAP [52]
RISE [12]	Local (P)	✗	SAR	RISE [61]
Stable Explanation [23]	Local (P)	-	Cityscapes	-
U-Noise [45,59]	Local (P)	✓	**Pancreas (CT)**	-
Occlusion [29]	Local (P)	✗	COCO-2017	Occlusion [88]
SegNBDT [80]	Local (C)	✓	Pascal-Context Ctiyscapes Look Into Person	NBDT [79]
Concept Graphs [46]	Local (C)	✓	**Brain Tumor (MRI)**	-
L-CRP [18]	Local (C)	✓	Cityscapes	CRP
Example-based Explanation[37]	Local	✗	SemanticKITTI	-
Graph-based [24]	Local	-	**Liver Vessel (CT)**	-
Adversarial attacks on Saliency Maps[28]	Local	✓	Industrial	Adversarial Perturbation [31]
Counterfactual [74]	Local	✗	**Instrument (Endoscopy) Polyp (colonoscopy)**	-
Counterfactual - Generative [90]	Local	✓	BDD100k	-

Table 4. Global XAI algorithms considered for this paper. Bold is used to indicate medical datasets. Additionally, if the algorithm is based on an existing algorithm from image classification, the predecessor is also identified. Key: **C** = Concept.

Algorithm	Family	Gradient-based	Dataset	Precursor in Image Classification
Network Dissection [58]	Global (C)	✗	**Brain Tumor (MRI)**	Network Dissection [7]
Activation Maximization [58]	Global (C)	✓	**Brain Tumor (MRI)**	Activation Maximization [73]
CAV [40]	Global (C)	✓	**Heart (MRI)**	CAV [44]
DeepDream [12]	Global	✓	**Liver Tumor (CT)**	Activation Maximization [73]
Visualizing Bottleneck [41] Representations	Global	✗	Satellite Imaging	-
Emergent Language [68]	Global	✓	**Brain Tumor (MRI)**	-
Segmentation Ability [36]	Global	-	**Brain Tumor (MRI) COVID-19 (CT) Prostate (MRI) Pancreatic Mass (CT)**	-
Permutation Feature[60] Importance	Global	✗	Satellite Imaging	Permutation Feature[20] Importance