Seg-XRes-CAM: Explaining Spatially Local Regions in Image Segmentation

Syed Nouman Hasany[1] Caroline Petitjean[1] Fabrice Mériaudeau[2]
[1]LITIS UR 4108, Normandie Université, INSA de Rouen,
UNIROUEN, UNIHAVRE, Rouen, France
[2]Institut de Chimie Moléculaire de l'Université de Bourgogne, ICMUB UMR CNRS 6302,
Université Bourgogne, Dijon 21000, France
{syed-nouman.hasany, caroline.petitjean}@univ-rouen.fr
fabrice.meriaudeau@u-bourgogne.fr

Abstract

While many post-hoc model interpretability techniques exist for image classification, image segmentation has not received the same attention. An extension of Grad-CAM, Seg-Grad-CAM was proposed as a local interpretability technique for image segmentation. In this paper, we highlight that by virtue of its design, Seg-Grad-CAM does not utilize spatial information when it comes to generating explanations for regions within a segmentation map. Taking inspiration from HiResCAM, we propose Seg-XRes-CAM in order to solve this problem. We verify the utility of our proposed method by visually comparing explanations generated from Seg-Grad-CAM and Seg-XRes-CAM against a model-agnostic, perturbation-based method, RISE. The code is available at https://github.com/Nouman97/Seg_XRes_CAM.

1. Introduction

Stemming from their groundbreaking success in image classification in 2012 [11], deep learning algorithms have quickly become the standard when it comes to approaching computer vision problems. Their unparalleled predictive performance is often a product of them being highly non-linear, and hence not inherently interpretable. However, it is often desirable to be able to understand why an algorithm arrived at its decision for a particular example. This can be useful in multiple contexts such as debugging the model when it comes to the developer or improving our confidence in the model's prediction when it comes to the end user.

A separate field has since developed with the aim of developing tools that can be utilized in order to explain individual predictions, *i.e.*, local interpretability. Primarily, two streams can be identified in this regard: (a) gradient-based approaches, and (b) perturbation-based approaches. Gradient-based approaches utilize the differentiable nature of neural networks in order to obtain a gradient-based saliency map. One of the earliest works in this regard was of Simonyan [20] which extracted saliency for a target class as the gradient of its score with respect to the input image. SmoothGrad [22] and IntegratedGradients [23] followed a similar methodology but aimed at refining the otherwise noisy gradient. An alternate approach was pursued in Grad-CAM [19] and its derivatives which backpropagated gradients only up to an intermediate layer. Perturbation-based approaches, on the other hand, are generally model-agnostic and treat models as black boxes. They work on the premise of observing changes to a model's prediction as the input is methodically modified. Early work in this regard by Zeiler and Fergus [25] was based on utilizing occlusion. Further approaches have since been proposed such as LIME [17], SHAP [14], and RISE [16] based on generating multiple modified instances of the original image which are then modeled using an inherently interpretable model. Most of these techniques, however, are usually developed in the context of image classification, and other problem domains have received much less attention.

The domain of image segmentation, for example, is considerably sparse when it comes to interpretability and relatively few techniques have been developed in its context [2,3,9,15,24]. One such example is that of Seg-Grad-CAM introduced by Vinogradova et al. [24] which is an extension of Grad-CAM [19] towards explaining segmentation results. Unlike image classification where, for any particular image, we often only need to explain a single score from a target class, in image segmentation we might want to explain either (a) the entire segmentation map for a target class or (b) a region of that segmentation map for a target class. In the present work, we demonstrate that, by virtue of its utilization of Grad-CAM, Seg-Grad-CAM cannot solve the latter problem i.e. of explaining a region of

the segmentation map. Inspired by HiResCAM [5], we propose Seg-XRes-CAM as a modification to Seg-Grad-CAM which solves this problem and can be utilized in both cases. HiResCAM has been utilized in image classification as an alternate to Grad-CAM in order to provide location aware explanations [6].

In Sec. 2.1, we elaborate on Seg-Grad-CAM and demonstrate why it cannot provide us with explanations for a region within the segmentation map. In Sec. 2.2, we introduce Seg-XRes-CAM and show why it is more suited to the task. In Sec. 3.1 we mention our experimental configuration. In Sec. 3.2, we give a brief explanation of a model-agnostic, perturbation-based interpretability method, RISE [16] in the context of image segmentation. RISE is chosen for comparison due to its model-agnostic and gradient independent nature. In Sec. 3.3 we apply Seg-Grad-CAM, Seg-XRes-CAM, and RISE to sample images and compare their visualizations. In particular, we focus on the visual agreeability of RISE with Seg-XRes-CAM. In Sec. 3.4 we compare the impact of mean and max pooling on the explanations generated by Seg-XRes-CAM on sample images.

2. Methods

2.1. Revisiting Seg-Grad-CAM

Grad-CAM [19] (an extension and generalization of CAM [26]) computes a linear combination of activation maps from a chosen block of the model (Eq. 1), producing a class-discriminative localization map $L^c_{Grad-CAM}$ defined as:

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

where A^k represents the k'th activation map and α^c_k represents the k'th coefficient. c is reflective of the class we are interested in. $ReLU$ is applied in order to only retain the positive contributions.

For classification, α^c_k is calculated by backpropagating the gradient from the target class score back to the activation maps, and for each activation map, the coefficient is a normalized sum (the global average pooled value) of its corresponding gradient matrix, as follows:

$$\alpha^c_k = GAP\left(\frac{\partial Y^c}{\partial A^k}\right) \tag{2}$$

where GAP refers to global average pooling.

In order to extend this approach to segmentation, Seg-Grad-CAM proposes utilizing a mask in order to highlight the desired region of the target class followed by summing up the scores contained within the masked region. This sum is then backpropagated in order to calculate the Grad-CAM linear coefficients, as follows:

$$\alpha^c_k = GAP\left(\frac{\partial \sum_{(i,j)\in\mathcal{M}} Y^c_{ij}}{\partial A^k}\right) \tag{3}$$

where $\mathcal{M}$ indicates the region of interest. Depending on what one wants to interpret, this region can be the entire segmentation map of the target class or a certain portion of that map or, in fact, just a single pixel from the target class map.

A consequence of each activation map having been assigned a single coefficient (obtained after the global average pooling of the gradient matrix) is that the spatial distribution of the activation map plays no part in the eventual linear combination. This can prove to be undesirable when it comes to explaining segmentation results for a region of interest belonging to a target class. If a region of interest lies in the bottom right corner of the image, it is unlikely that the top left corner of the image would have had any impact on its segmentation result. Yet, by virtue of its design, Seg-Grad-CAM would not take this spatial consideration into account and give the same weight to each spatial location of an individual activation map.

2.2. Seg-XRes-CAM

HiResCAM [5] proposed a modification to Grad-CAM in order to incorporate spatial information into the linear combination of activation maps. It modifies the original Grad-CAM equation such that each activation map is element-wise multiplied (Hadamard product) by its corresponding gradient matrix, yielding the following localization map $L^c_{HiResCAM}$:

$$L^c_{HiResCAM} = ReLU\left(\sum_k \frac{\partial Y^c}{\partial A^k} \odot A^k\right) \tag{4}$$

This allows each spatial location within a feature map to be weighted differently. While it was originally proposed in the context of image classification, its design hints towards its utility in image segmentation.

We propose a generalization in order to have control over how fine or coarse we want our gradient matrix to be. This can simply be achieved by applying a pooling operation (max or mean) with a window size of $h \times w$ over the gradient matrix. A window size of 1×1 corresponds to the HiResCAM formulation whereas a window size of $H \times W$ where H is the height and W is the width of the activation map corresponds to a Grad-CAM formulation. The pooled gradient matrix is then upsampled in order to perform the Hadamard product. The resulting localization map $L^c_{XRes-CAM}$ for image classification can then be written as:

$$L^c_{XRes-CAM} = ReLU\left(\sum_k Up[Pool\left[\frac{\partial Y^c}{\partial A^k}\right]] \odot A^k\right) \tag{5}$$

where $Pool$ indicates the pooling operation and Up indicates upsampling (with the resulting dimension being the same as A^k).

Similar to Seg-Grad-CAM, in the context of segmentation Y^c is replaced with $\sum_{(i,j)\in\mathcal{M}} Y^c_{ij}$ in order to represent the target region we are interested in obtaining the explanation for.

3. Experiments

3.1. Configuration

We experiment with three different model architectures, two of which are standard convolutional networks (DeepLabV3 [1] and UNet [18]) whereas the third one (UNETR [8]) utilizes a pre-trained version of the recently proposed vision transformer model [4] as its encoder backbone. We experiment on both natural and medical images, natural images coming from the COCO 2017 dataset [13] and medical images coming from the Synapse multi-organ CT dataset [12].

In the case of natural images, the utilized model was a DeepLabV3 [1] model already trained on a subset of the COCO 2017 data[1] [13]. Samples for visualization were also taken from the same dataset. For the medical dataset, two custom models were utilized: (a) UNet [18] with a pre-trained VGG-16 backbone [21] and (b) 2D UNETR [8] with a pre-trained ViT backbone [4]. These models were trained on the Synapse multi-organ CT dataset [12] from which the visualization samples were also taken from. In all experiments, we extracted saliency maps from the models' bottleneck.

3.2. RISE

RISE [16] is a perturbation-based model agnostic interpretability method. It works by generating multiple random masks, and a linear combination of those masks serves as our explanation. In order to explain a classification decision, the masks are pointwise multiplied by the input image and fed into the model. The target class score corresponding to each modified input serves as the linear combination coefficient for that particular mask. RISE can be extended from classification to segmentation [3]. In this case, for a region of interest of the target class, the Dice score (for that region) between a masked input and the original (unmodified) input image can serve as the linear combination coefficient for that particular mask. In all experiments, we utilized RISE [10] with 2000 random masks in order for it to generate each explanation.

3.3. Visual comparison of saliency maps between Seg-Grad-CAM, Seg-XRes-CAM, and RISE

Visualizations obtained for sample images can be seen in Fig. 1. In general, saliency maps obtained from Seg-Grad-CAM [7] display little awareness of the particular regions they were supposed to be explanations for. Seg-XRes-CAM, on the other hand, appears to be taking that into account and accordingly makes use of spatial information making its saliency maps more localized. Additionally, maps generated by RISE seem to have a higher degree of agreement with Seg-XRes-CAM as compared to Seg-Grad-CAM.

3.4. Impact of pooling mechanism on Seg-XRes-CAM

In contrast to Seg-Grad-CAM, window size and pooling mechanism are additional hyperparameters for Seg-XRes-CAM. We experiment with mean and max pooling using windows of size 2 and 4. Visualizations for sample images can be seen in Fig. 2 and Fig. 3. In order to determine whether the region highlighted by our explainability method is sufficient for the model to predict the target class in our selected region, we utilize a binarized version of the generated saliency map to mask out the remaining image. In all experiments, we used a threshold of 0.2 to binarize the saliency map. This masked image is then fed to the model, and a Dice score between the prediction for the selected region of the modified image versus the prediction from the original image is calculated. Following Mullan and Sonka [15] we record the dice score as well as the percentage of the image which was retained following masking. The retained portion in the image should be as small as possible but at the same time it should have enough context to allow the model to make the correct prediction for the selected region.

From Fig. 2 and Fig. 3 it appears that saliency maps obtained from max pooling preserve a comparatively higher percentage of the image as compared to mean pooling. However, this extra portion leads to a higher dice score and can potentially avoid cases where mean pooling completely fails (Fig. 3).

[1] https://pytorch.org/hub/pytorch_vision_deeplabv3_resnet101/

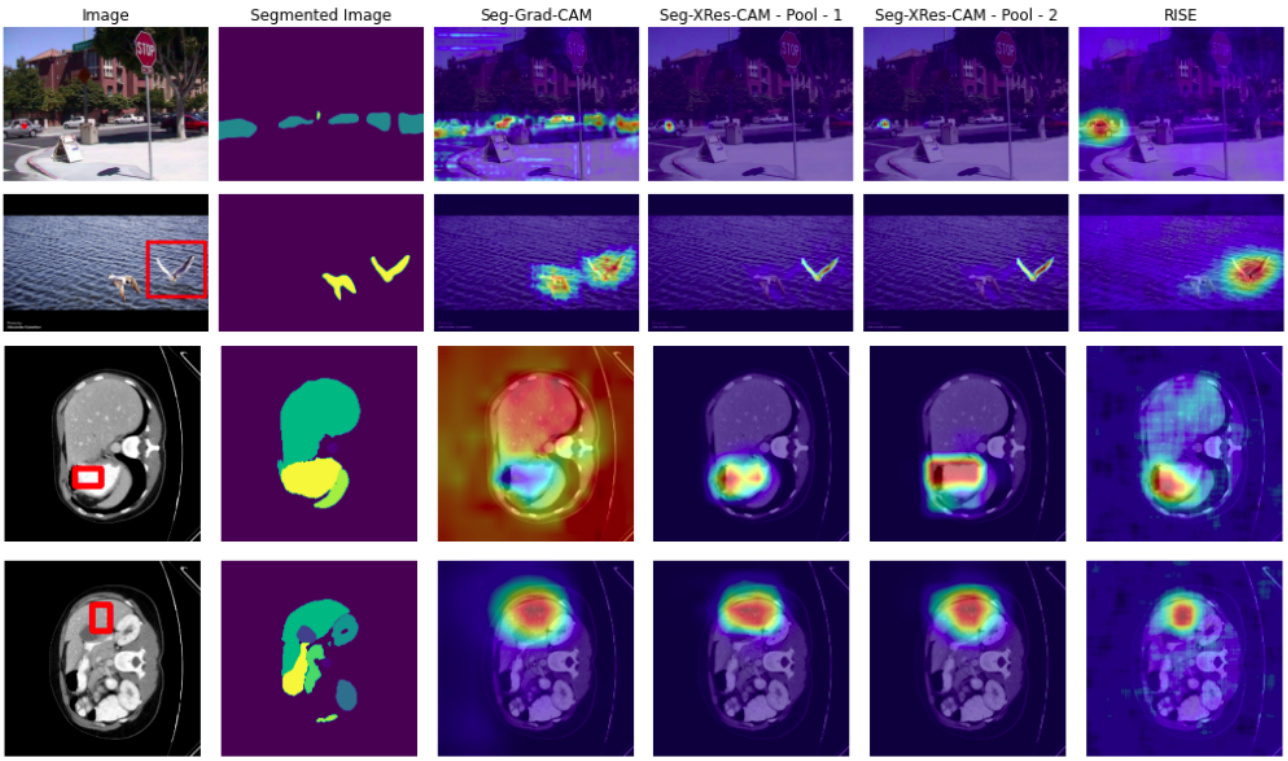

Figure 1. Saliency maps generated from interpretability methods in order to explain a region (in red) within the segmentation map. (First Row) the explanation for a point (in red) on the leftmost car - the model being DeepLabV3 (Second Row) explanation for the right bird - the model being DeepLabV3 (Third Row) explanation for a region in the stomach - the model being 2D UNETR (Fourth Row) explanation for a region in the liver - the model being UNet with pre-trained VGG-16 backbone. For Seg-XRes-CAM, a pooling window of size 1 and 2 were utilized with mean average pooling

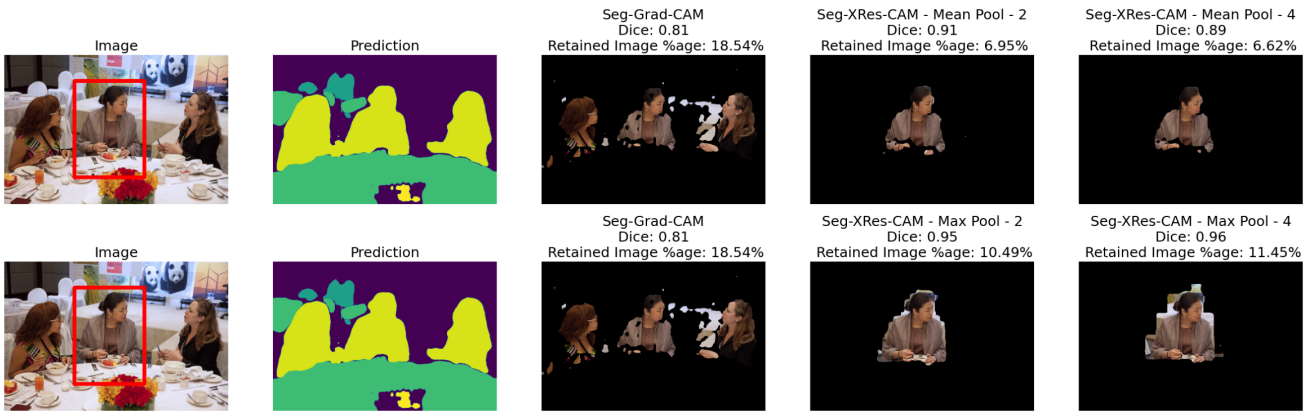

Figure 2. Impact of pooling mechanism - For a sample image, its prediction, and the region to be explained, saliency maps are generated from Seg-Grad-CAM and Seg-XRes-CAM. Masked versions of the original image are obtained utilizing the saliency maps. Here, the image is masked based on the explanation generated for the middle woman. The percentage of the retained image as well as the dice score for the selected region obtained from the masked image are provided (First Row) masking based on the explanations generated by Seg-Grad-CAM and Seg-XRes-CAM (mean pooling , window sizes 2 and 4) (Second Row) masking based on the explanations generated by Seg-Grad-CAM and Seg-XRes-CAM (max pooling , window sizes 2 and 4).

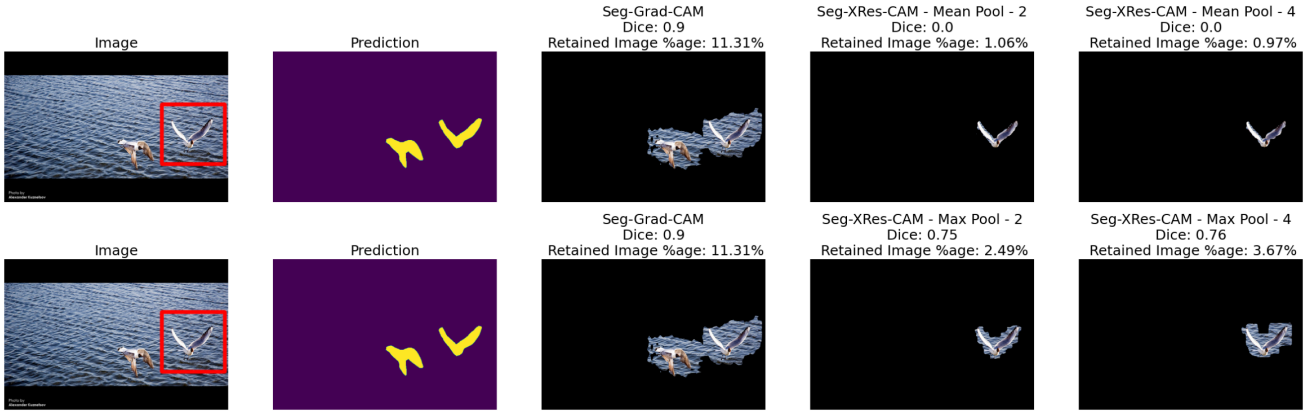

Figure 3. Impact of pooling mechanism - For a sample image, its prediction, and the region to be explained, saliency maps are generated from Seg-Grad-CAM and Seg-XRes-CAM. Masked versions of the original image are obtained utilizing the saliency maps. Here, the image is masked based on the explanation generated for the right bird. The percentage of the retained image as well as the dice score for the selected region obtained from the masked image are provided (First Row) masking based on the explanations generated by Seg-Grad-CAM and Seg-XRes-CAM (mean pooling , window sizes 2 and 4) (Second Row) masking based on the explanations generated by Seg-Grad-CAM and Seg-XRes-CAM (max pooling , window sizes 2 and 4).

4. Discussion

Visualization of saliency maps generated by Seg-Grad-CAM confirms our initial hypothesis that by virtue of its formulation, it cannot take spatial information into account when explaining regions within a segmentation map. On the other hand, visualizations obtained from the proposed Seg-XRes-CAM seem to be utilizing spatial information in order to generate its explanations. Additionally, explanations generated by Seg-XRes-CAM tend to agree with those generated by RISE as far as the localization of the explanation is concerned.

However, Seg-XRes-CAM, too, seems to be suffering from a few flaws which require further investigation. Firstly, explanations generated from Seg-XRes-CAM are quite fine and tend not to produce sufficient explanations (dilation can be used as a potential post-processing step). Secondly, while we investigated the bottleneck layer, the choice of the layer is for the user to decide, and is therefore a hyperparameter.

The other two hyperparameters for Seg-XRes-CAM are its pooling mechanism and the window size. It appears that compared to mean pooling, max pooling tends to produce comparatively coarser and more reliable explanations. For the window size, the bottleneck dimensions need to be taken into account because while a larger window size can produce a coarser explanation, it might do so at the expense of localization.

As far as a comparison of RISE and Seg-XRes-CAM is concerned, RISE is model agnostic and produces much coarser explanations compared to Seg-XRes-CAM. However, RISE is computationally more expensive as Seg-XRes-CAM (and Seg-Grad-CAM) requires only a single forward and backward pass through the network whereas RISE requires as many forward passes as the number of masks it uses (this can be sped up by batching). Additionally, the number of masks is a relatively important hyperparameter of RISE as fewer masks generally tend to generate spurious explanations which are not very informative.

5. Conclusion

Consequently, in this work we identified a potential flaw in Seg-Grad-CAM when it comes to explaining regions within a segmentation map, and proposed Seg-XRes-CAM as a potential improvement. We also highlighted some of the shortcomings of our proposed method which we plan on investigating further in the future.

6. Acknowledgments

The authors would like to thank the reviewers for their deeply insightful comments and suggestions. Additionally, the authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant Project-ANR-21-CE23-0013 (project MediSEG).

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *arxiv preprint, arXiv:1706.05587*, 2017. 3

[2] Vincent Couteaux, O. Nempont, Guillaume Pizaine, and Isabelle Bloch. Towards interpretability of segmentation networks by analyzing deepdreams. In *MICCAI Workshop on Interpretability of Machine Intelligence in Medical Image Computing*, 2019. 1

[3] Pierre Dardouillet, Alexandre Benoit, Emna Amri, Philippe Bolon, Dominique Dubucq, and Anthony Crédoz. Explainability of image semantic segmentation through shap values. In *ICPR Workshop on Explainable and Ethical AI*, 2022. 1, 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[5] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. In *arxiv preprint, arXiv:2011.08891*, 2020. 2

[6] Rachel Lea Draelos and Lawrence Carin. Explainable multiple abnormality classification of chest ct volumes. *Artificial Intelligence in Medicine*, 2022. 2

[7] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 3

[8] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *WACV*, 2022. 3

[9] Lukas Hoyer, Mauricio Munoz, Prateek Katiyar, Anna Khoreva, and Volker Fischer. Grid saliency for context explanations of semantic segmentation. In *NeurIPS*, 2019. 1

[10] Yuchi Ishikawa. Randomized input sampling for explanation of black-box models (rise). https://github.com/yiskw713/RISE, 2020. 3

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[12] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *MICCAI Workshop Challenge on Multi-Atlas Labeling Beyond Cranial Vault*, 2015. 3

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[14] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, 2017. 1

[15] Sean Mullan and Milan Sonka. Visual attribution for deep learning segmentation in medical imaging. In *Medical Imaging 2022: Image Processing*, 2022. 1, 3

[16] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018. 1, 2, 3

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD*, 2016. 1

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[19] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 2019. 1, 2

[20] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2014. 1

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[22] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *ICML Workshop on Visualization for Deep Learning*, 2017. 1

[23] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 1

[24] Kira Vinogradova, Alexandr Dibrov, and Gene Myers. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In *AAAI*, 2020. 1

[25] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1

[26] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2