Leveraging CAM Algorithms for Explaining Medical Semantic Segmentation

Tillmann RHEUDE
TU Darmstadt, Darmstadt, Germany
Fraunhofer Institute for Computer Graphics Research (IGD), Darmstadt, Germany

Andreas WIRTZ
Fraunhofer Institute for Computer Graphics Research (IGD), Darmstadt, Germany

Arjan KUIJPER
TU Darmstadt, Darmstadt, Germany
Fraunhofer Institute for Computer Graphics Research (IGD), Darmstadt, Germany

Stefan WESARG
Fraunhofer Institute for Computer Graphics Research (IGD), Darmstadt, Germany

Abstract

Convolutional neural networks (CNNs) achieve prevailing results in segmentation tasks nowadays and represent the state-of-the-art for image-based analysis. However, the understanding of the accurate decision-making process of a CNN is rather unknown. The research area of explainable artificial intelligence (xAI) primarily revolves around understanding and interpreting this black-box behavior. One way of interpreting a CNN is the use of class activation maps (CAMs) that represent heatmaps to indicate the importance of image areas for the prediction of the CNN. For classification tasks, a variety of CAM algorithms exist. But for segmentation tasks, only two CAM algorithms for the interpretation of the output of a CNN exist. We propose a transfer between existing classification- and segmentation-based methods for more detailed, explainable, and consistent results which show salient pixels in semantic segmentation tasks. The resulting *Seg-HiRes-Grad CAM* is an extension of the segmentation-based *Seg-Grad CAM* with the transfer to the classification-based *HiRes CAM*. Our method improves the previously-mentioned existing segmentation-based method by adjusting it to recently published classification-based methods. Especially for medical image segmentation, this transfer solves existing explainability disadvantages. The code is available at `https://github.com/TillmannRheude/SegHiResGrad_CAM`.

Keywords: Deep Learning, Explainable Artificial Intelligence, Gradient-Based Methods, Medical Image Segmentation

1. Introduction

Even if the (mathematical) theory of neural networks' training process is well known, the exact reasoning behind why neural networks derive at a particular prediction is rather hard to interpret (Selvaraju et al., 2017). The black box behavior of neural networks complicates the understanding of their decision-making process. Especially for medical tasks, understanding this process is essential (Chen et al., 2022). Here, the question would be, *e.g.*, where does the prediction that there is a tumor in the image come from? Ideally, the prediction comes from image regions containing the tumor. However, it would also

be possible that the prediction is triggered by other factors that do not contribute to the presence or absence of the tumor.

Algorithms for explainable artificial intelligence (xAI) can be classified in different ways. Local explanation algorithms derive a reasoning of individual predictions $f(x)$ of a model f in contrast to global explanation algorithms which derive a reasoning with only the model f and without the need of any predictions $f(x)$ of the model (Agarwal et al., 2021). Model-specific algorithms are only applicable to certain models (Agarwal et al., 2021). On the other hand, model-agnostic algorithms like SHAP (Lundberg and Lee, 2017) are applicable regardless of the model choice (Agarwal et al., 2021). Gradient-based methods use the gradients of the neural network as a proxy in comparison to pertubation-based methods which derive the explanation by altering the input data (Ivanovs et al., 2021). In summary, algorithms which use class activation maps (CAMs)[1] are therefore classified as local because of the use of individual input datapoints on which a heatmap is placed and they are model-specific since they cannot be applied to any model like a simple regression model. It has to be noted that CAM-based algorithms do not have to be gradient-based but they are most of the times nowadays as explained in Section 2. With our work, we want to improve the state-of-the-art (SOTA) for explainable, local, model-specific and gradient-based algorithms for image segmentation via CAMs.

To generate a heatmap of an input image that interprets the model output, two well-known methods were proposed in the past: saliency maps (Zeiler and Fergus, 2014) and CAMs (Zhou et al., 2016). A saliency map, *e.g.*, created by Saliency Mapping (Simonyan et al., 2014), DeconvNets (Zeiler and Fergus, 2014) or Guided Backpropagation (Springenberg et al., 2015), is a heatmap for showing the importance of individual image pixels. These maps are based on different calculcations of the gradients regarding the input features. CAMs are also heatmaps but they are based on the activations of the last or one specific convolutional layer of a convolutional neural network (CNN). They are able to visualize the spatial positioning and expansion and are used as a weighting factor for the activations of the feature maps. In summary, both methods produce similar heatmaps for interpretation, but they differ in how they are calculated. While saliency maps are classified as input-level approaches, CAM algorithms are classified as output-level approaches (Draelos and Carin, 2021). The advantage of CAM algorithms is that they do not need the propagation through all layers and that they are more robust in contrast to saliency maps (Draelos and Carin, 2021). Robustness is especially important in domains where the precise localization and visualization of influential image regions play a pivotal role, such as medical imaging applications.

Algorithms for the visualization of CAMs are dominated by classification-based work (Table 1). For segmentation purposes, a transfer is needed and already proposed with *Seg-Grad CAM* (Vinogradova et al., 2020). Nevertheless, this transfer is based on *Grad CAM* (Selvaraju et al., 2017) and for classification tasks *Grad CAM* has the disadvantage of certain inaccuracies (Draelos and Carin, 2021). These inaccuracies include, *e.g.*, the visualization of regions that are not used by the CNN for the prediction (Draelos and Carin, 2021).

We propose *Seg-HiRes-Grad CAM* to solve the before-mentioned inaccuracies in the seg-

---

1. For a particular algorithm, we use the italic font while it is not italicized when we refer to the general idea of class activation maps (CAMs).

mentation case, too. *Seg-HiRes-Grad CAM* is based on *Seg-Grad CAM* and transfers the classification-based *HiRes CAM* (Draelos and Carin, 2021) to segmentation tasks (Fig. 1).
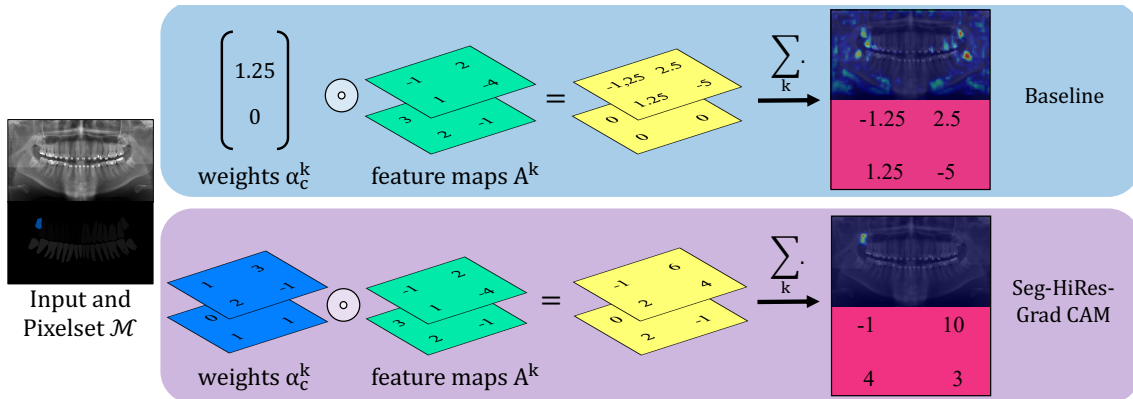

Figure 1: Calculation flow of *Seg-Grad CAM* (Vinogradova et al., 2020) (**top row**) and our proposed *Seg-HiRes-Grad CAM* (**bottom row**) based on the computations and flowchart of *HiRes CAM* (Draelos and Carin, 2021) with K feature maps, dimensions D_1 and D_2 of the feature maps and the average weight values α_c^k. Gradients (blue) are multiplied with activation maps (green). The sum (red) of the product (yellow) is upscaled for the final heatmap before ReLU and after ReLU (above red square). On the left, the respective input image and semantic segmentation of the U-Net (Ronneberger et al., 2015) is shown. It is striking that the output CAMs are very different. ReLU is not visualized to ensure a better comparability.

The paper is structured as follows: We start by stating related work in Section 2 for being able to understand the method derivation in Section 3. Afterwards, we show qualitative results in Section 4 and discuss possible caveats in Section 5. At the end, we draw a conclusion and point out possible future work in Section 6.

2. Related Work

The first algorithm of visualizing CAMs after the before-mentioned methods for saliency maps is called *CAM* (Zhou et al., 2016). The idea of *CAM* is the use of the weights of the last feed-forward network (FFN) in a CNN as weightings for the feature maps of the last convolutional layer. For a more complex calculation and more interpretable visual results, *Grad CAM* (Selvaraju et al., 2017) is proposed which complements the weight calculation with the average of the respective gradients of a neural network. But, *Grad CAM* still visualizes certain inaccuracies in some cases (Draelos and Carin, 2021). One of the most present inaccuracies is that there are cases in which *Grad CAM* visualizes regions the CNN did not actually use (Draelos and Carin, 2021). *HiRes CAM* (Draelos and Carin, 2021) solves these inaccuracies. In *HiRes CAM*, the gradients are not averaged anymore as in *Grad CAM*. *HiRes CAM* uses the raw gradients for calculating the weights, which are multiplied with the activations of the CNN. The before-mentioned algorithms represent the most important publications for visualizing CAMs. Further publications which are related

Table 1: Excerpt of popular algorithms for CAMs in classification tasks and the respective publications for segmentation tasks. Our proposed algorithm (bold) represents the transfer of *HiRes CAM* (Draelos and Carin, 2021) to segmentation tasks.

Publications for classification tasks	**Respective publications for segmentation tasks**
CAM (Zhou et al., 2016)	*n/a*
Grad CAM (Selvaraju et al., 2017)	Seg-Grad CAM (Vinogradova et al., 2020)
Grad CAM++ (Chattopadhyay et al., 2018)	*n/a*
XGrad CAM (Fu et al., 2020)	*n/a*
Ablation CAM (Desai and Ramaswamy, 2020)	*n/a*
Score CAM (Wang et al., 2020)	*n/a*
Eigen CAM (Muhammad and Yeasin, 2020)	*n/a*
Layer CAM (Jiang et al., 2021)	*n/a*
FullGrad CAM (Srinivas and Fleuret, 2019)	*n/a*
LIFT CAM (Jung and Oh, 2021)	*n/a*
HiRes CAM (Draelos and Carin, 2021)	Seg-XRes-CAM (Hasany et al., 2023) **Seg-HiRes-Grad CAM** (ours)

to the topic of visualizing CAMs but which are not necessary to understand our proposed approach are listed in Table 1 (left column).

The before-mentioned methods refer to the visualization of CAMs in classification tasks but not to the visualization in segmentation tasks. For the latter, *Seg-Grad CAM* is the first method for visualizing CAMs. However, the inaccuracies that occur in *Grad CAM*, which were tackled by *HiRes CAM*, are also present in *Seg-Grad CAM*, since *Seg-Grad CAM* is a modification to *Grad CAM*. Consequently, we propose the transfer of the classification-based *HiRes CAM* and the segmentation-based *Seg-Grad CAM* resulting in *Seg-HiRes-Grad CAM* (Fig. 1). The simultaneously published method *Seg-XRes-CAM* (Hasany et al., 2023) is also based on the combination of *Seg-Grad CAM* and *HiRes CAM*. However, *Seg-XRes-CAM* differs from *Seg-HiRes-Grad CAM* as it includes a pooling layer in the calculation. This results in additional hyperparameters such as a window size. An evaluation for medical images and an ablation study regarding the pooling layer are not provided for *Seg-XRes-CAM*.

3. Method

With our work, we want to improve the SOTA of CAM algorithms for (medical) image segmentation. As elaborated in the previous section, CAMs can be interpreted as heatmaps which can be placed upon the input image illustrating the areas which are more critical for the decision-making process. Since most of the work is done for classification tasks (Table 1), we start empirically by explaining the mathematical theory behind these methods first and draw the transfer to segmentation-based algorithms afterwards. CAM is described mathematically as follows with the heatmap L_{CAM}^c, weights α of the FFN, the respective class c, feature maps A and the respective number of the channel k in the feature map (Simonyan et al., 2014):

$$L_{CAM}^c = \sum_k \alpha_c^k A^k. \tag{1}$$

Improving the visual explainability, *Grad CAM* (Selvaraju et al., 2017) is proposed based on *CAM*. The weighting (α in Eq. (1)) is calculated with the respective gradients instead of only the weights of the FFN. Consequently, the mathematical description changes as follows with the heatmap $L_{GradCAM}^c$ and the ReLU function:

$$L_{GradCAM}^c = \text{ReLU}\left(\sum_k \alpha_c^k A^k\right), \tag{2}$$

including the weights α with the number of pixels N, individual pixels u, v, and outputs y^c:

$$\alpha_c^k = \frac{1}{N} \sum_{u,v} \frac{\partial y^c}{\partial A_{uv}^k}. \tag{3}$$

But recently, it was shown that *Grad CAM* visualizes regions in the image which do not contribute to the outcoming prediction (Draelos and Carin, 2021). Thus, *e.g.*, in the classification task of the atelectasis (collapsed lung), *Grad CAM* suggests that the neural network uses pixels which are located at the heart (Draelos and Carin, 2021). This is not explainable since the CNN should look at the lung instead. To solve this inaccuracy, *HiRes CAM* (Draelos and Carin, 2021) is proposed which correctly visualizes the lung region in this example. The difference between these two CAM visualization methods is the mean calculation for the weights α (Eq. (3)) (Draelos and Carin, 2021). It is proposed to calculate the weights with the same notation (Eq. (3)) as follows instead (Draelos and Carin, 2021):

$$\alpha_c^k = \frac{\partial y^c}{\partial A^k}. \tag{4}$$

So far, all equations refer to classification-based CNNs. For segmentation-based CNNs, a transfer is needed: In segmentation tasks, there is a label for every pixel and not only for every picture, as in classification tasks. As a consequence, Vinogradova et al. (Vinogradova et al., 2020) propose to modify y^c in Eq. (2) and Eq. (3) as follows with $\mathcal{M}$, the "set of pixel indices of interest in the output mask" (Vinogradova et al., 2020) and the respective, individual pixels i, j:

$$y^{c,new} = \sum_{i,j \in \mathcal{M}} y_{i,j}^c. \tag{5}$$

With this modification, the pixel set $\mathcal{M}$ can be chosen in a flexible way and the classification-based *Grad CAM* is transferred to the segmentation-based *Seg-Grad CAM* (Vinogradova et al., 2020). For instance, $\mathcal{M}$ can define all pixels of the image or only a certain amount of pixels up to only one certain pixel (Vinogradova et al., 2020). On the other hand, this method still suffers from the inaccuracies of the original proposed *Grad CAM* revealed by *HiRes CAM* for classification purposes. To address this limitation of *Seg-Grad CAM*, we propose *Seg-HiRes-Grad CAM* by combining the segmentation-based *Seg-Grad CAM* and the classification-based *HiRes CAM* (Fig. 1). The weights α_c^k are not calculated by using the mean anymore, but the formula for the weights still involves the modification with the pixel set $\mathcal{M}$. This results in the following weight calculation formula:

$$\alpha_c^k = \frac{y^{c,new}}{\partial A^k} = \frac{\sum_{i,j \in \mathcal{M}} y_{i,j}^c}{\partial A^k}. \tag{6}$$

For the segmentation task, every layer with feature maps (*i.e.*, A^k) of a CNN can be used, but it is common to use the deepest layer which contains the highest number of feature maps (Vinogradova et al., 2020). The deepest layer represents most of the feature information, while higher layers represent more edge-like structures (Vinogradova et al., 2020). Since our method is affected by the model performance, we have attached corresponding details and training results (even for further datasets) in Table 2 for improved transparency. We use commonly known, multi-class and semantic segmentation datasets such as Cityscapes (Cordts et al., 2016), Kits23 (Heller et al., 2020) and OPG (Jader et al., 2018) to prove the method in medical and non-medical settings (Table 2). These heterogeneous datasets contain varying amounts of data, classes, and also challenges such as homogeneous gray levels (OPG) or fine-grained details (Cityscapes, Kits23).

4. Results

Comparing our proposed *Seg-HiRes-Grad CAM* with the *Seg-Grad CAM* baseline, the resulting heatmaps differ regarding their accuracy and explainability. In Fig. 2, the qualitative comparison with the Cityscapes dataset (Cordts et al., 2016) for the car-class is illustrated.

Seg-HiRes-Grad CAM can highlight pixels that are car-related only. In contrast, *Seg-Grad CAM* highlights coarser regions, including the sky, trees, and street parts. Two medical examples demonstrate the difference between both algorithms more distinctly (Fig. 3, Fig. 4). In these cases, the baseline method does not produce explainable results, while the result of *Seg-HiRes-Grad CAM* indicates the region around the segmented tooth / tumor accurately. Summarizing these examples, especially segmentations that are close together (such as the teeth) are highlighted more accurately with *Seg-HiRes-Grad CAM* in comparison to *Seg-Grad CAM*. Especially for medical images, the differences of the CAMs are striking and not as subtle as for natural images. Nevertheless, not all segmentations can be explained - even with our proposed *Seg-HiRes-Grad CAM*.
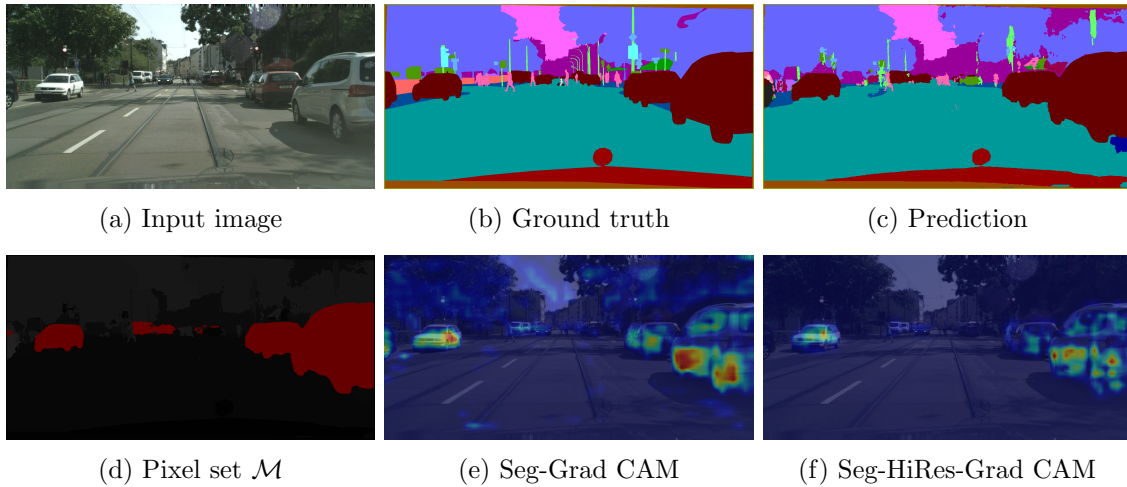
(a) Input image   (b) Ground truth   (c) Prediction

(d) Pixel set $\mathcal{M}$ (e) Seg-Grad CAM   (f) Seg-HiRes-Grad CAM

Figure 2: Comparison between *Seg-Grad CAM* (Vinogradova et al., 2020) **(e)** and *Seg-HiRes-Grad CAM* **(f)**. In this case, $\mathcal{M}$ equals the respective pixels of the prediction **(c)** for the car class **(d)**, which is similar to the ground truth **(b)**. The input image **(a)** from the Cityscapes dataset (Cordts et al., 2016) is used since Vinogradova et al. (Vinogradova et al., 2020) use it.
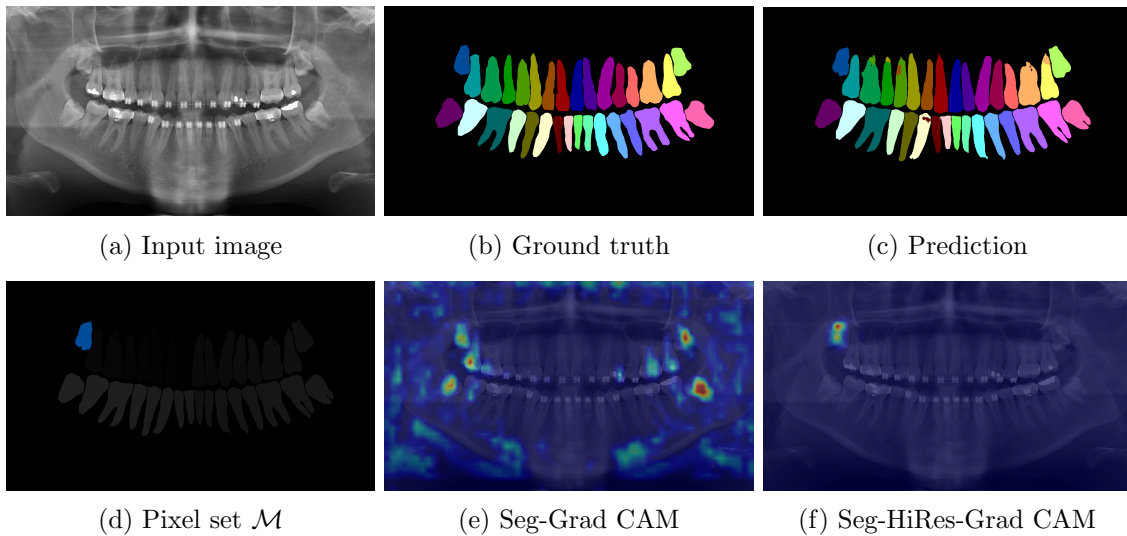

(a) Input image   (b) Ground truth   (c) Prediction

(d) Pixel set $\mathcal{M}$ (e) Seg-Grad CAM   (f) Seg-HiRes-Grad CAM

Figure 3: Comparison between *Seg-Grad CAM* (Vinogradova et al., 2020) **(e)** and *Seg-HiRes-Grad CAM* **(f)** for the upper right wisdom tooth (blue segmentation) **(d)**. In this case, $\mathcal{M}$ equals the respective pixels of the prediction **(c, d)**, which is similar to the ground truth **(b)**. The input image **(a)** comes from the orthopantogram (OPG) dataset (Jader et al., 2018).

However, the latter is still way more consistent, so it is considered the better visualization method. Exemplary results for, *e.g.*, different pixel sets $\mathcal{M}$, datasets and activation map levels are presented in Appendix A.
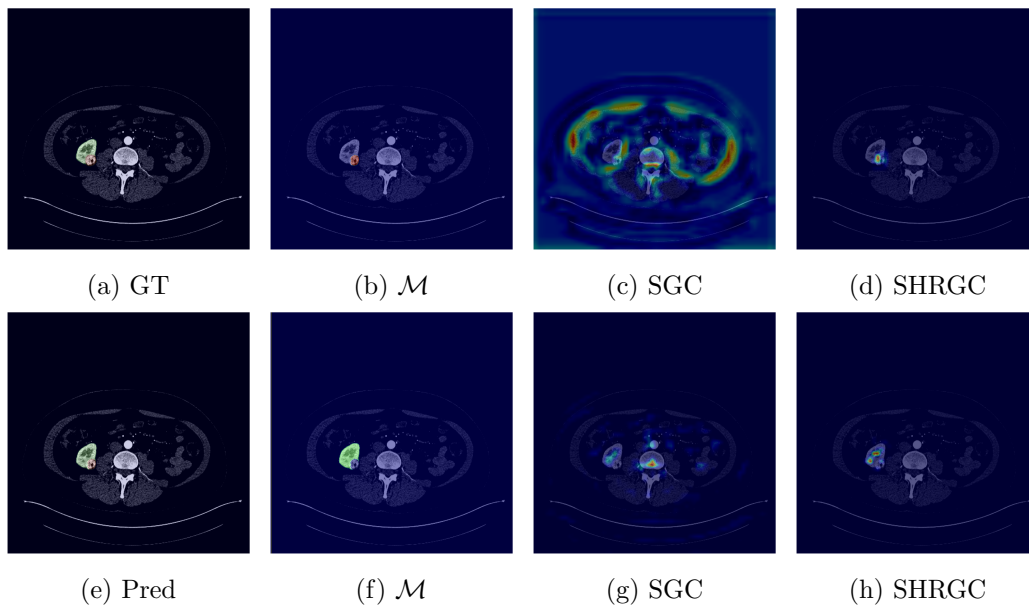
| (a) GT | (b) $\mathcal{M}$ | (c) SGC | (d) SHRGC |

| (e) Pred | (f) $\mathcal{M}$ | (g) SGC | (h) SHRGC |

Figure 4: Comparison between *Seg-Grad CAM* (Vinogradova et al., 2020) **(c, g)** and *Seg-HiRes-Grad CAM* **(d, h)** for a tumor **(b)** and kidney **(f)**. The input image comes from the Kits23 dataset (Heller et al., 2020).

Table 2: Validation-/Test-split results and details of the U-Net for the different datasets (2D slices in case of 3D dataset) we used for our experiments. Hyperparameters are selected empirically, splits are pre-defined or chosen randomly (80, 10, 10), learning rate is set to $3e - 3$ and the U-Net depth is four (512, 256, 128, 64). The U-Nets are trained for 300 epochs.

Dataset	F1-Score	IoU	Resolution	Augmentation
Cityscapes (Cordts et al., 2016)	0.865	0.774	512×1024	None
OPG (Jader et al., 2018)	0.959	0.921	560×992	Vertical Mirroring
Kits23 (Heller et al., 2020)	0.996	0.993	512×512	None

5. Discussion

Our proposed *Seg-HiRes-Grad CAM* explains salient regions more precisely and accurately in comparison to *Seg-Grad CAM*. Misinterpretations due to the visualization method can thus be better excluded as existing work describes for classification tasks (Draelos and Carin, 2021). Accordingly, *HiRes CAM* can also be applied to segmentation tasks by implementing *Seg-HiRes-Grad CAM*, which provides more transparent results than *Seg-Grad CAM*. Particularly in a medical setting, this difference of explanation can be of strong importance. On the other hand, certain limitations have to be addressed. First, the runtime is a general lim-

itation of CAM algorithms for segmentation tasks. Compared to classification-based CAM visualization algorithms, the segmentation-based algorithms need more time to produce the heatmap(s) due to the pixel-level probabilities in semantic segmentation tasks instead of a single class distribution for an entire image in the case of classification tasks. Second, it should be noted that the success of CAM algorithms depends on the input image resolution: The method fails for minimal image resolutions in combination with CNNs. Small image resolutions are common practice due to graphics processing unit (GPU) limitations, especially for 3D data. These cause a minimal resolution of the feature map of the deepest layer of a U-Net. If this feature map is spatially too small, too much detail is lost for an accurate representation of the gradients to be possible. Last, the CAM visualization depends on the segmentation result. Consequently, false negative or false positive segmented pixels will result in less explainable results when *Seg-HiRes-Grad CAM* or similar methods are used.

6. Conclusion and Future Work

In this work, we propose a semantic segmentation CAM visualization method (*Seg-HiRes-Grad CAM*), an extension of *Seg-Grad CAM* by combining it with the classification-based *HiRes CAM*. The proposed method accurately highlights salient regions and delivers more explainable results especially when medical datasets are used. We demonstrate that the on *Grad CAM* based method *Seg-Grad CAM* has the same disadvantage of highlighting misleading regions as Draelos and Carin (Draelos and Carin, 2021) elaborated for classification tasks. In contrast, our transfer of *HiRes CAM* (Draelos and Carin, 2021) to segmentation tasks generates more consistent results. Also, further transfers to the variety of classification-based methods could be drawn and quantitative methods such as Remove and Retrain (ROAR) (Hooker et al., 2019) could be used for enhanced comparisons.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

All datasets utilized in this study are publicly available and can be accessed freely. However, access to some of these datasets requires prior registration or application for access due to the sensitive nature of the data or to comply with data protection regulations. Despite the public availability of the datasets, direct sharing of the data by the authors is not permissible due to copyright and licensing restrictions imposed by the data providers. We encourage

interested researchers to obtain the data directly from the respective repositories, adhering to the specified access procedures and usage policies.

References

Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Upadhyay, Steven Wu, and Himabindu Lakkaraju. Towards the unification and robustness of perturbation and gradient based explanations. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 110–119. PMLR, 2021.

Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 839–847. IEEE Computer Society, 2018. .

Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Medicine*, 5(1):156, Oct 2022. ISSN 2398-6352. .

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223. IEEE Computer Society, 2016. .

Saurabh Desai and Harish G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 972–980. IEEE, 2020. .

Rachel Lea Draelos and Lawrence Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *ArXiv*, November 2021.

Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

Syed Nouman Hasany, Caroline Petitjean, and Fabrice Mériaudeau. Seg-xres-cam: Explaining spatially local regions in image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, page 3733–3738, June 2023.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in

kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, page 101821, 2020.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9734–9745, 2019.

Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognit. Lett.*, 150:228–234, 2021. .

Gil Jader, Jefferson Fontineli, Marco Ruiz, Kalyf Abdalla, Matheus Pithon, and Luciano Oliveira. Deep instance segmentation of teeth in panoramic x-ray images. In *31st SIB-GRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2018, Paraná, Brazil, October 29 - Nov. 1, 2018*, pages 400–407. IEEE Computer Society, 2018. .

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Trans. Image Process.*, 30:5875–5888, 2021. .

Hyungsik Jung and Youngrock Oh. Towards better explanations of class activation mapping. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1316–1324. IEEE, 2021. .

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–7. IEEE, 2020. .

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. .

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. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017. .

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4126–4135, 2019.

Kira Vinogradova, Alexandr Dibrov, and Gene Myers. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13943–13944. AAAI Press, 2020.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 111–119. Computer Vision Foundation / IEEE, 2020. .

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer, 2014. .

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2921–2929. IEEE Computer Society, 2016. .
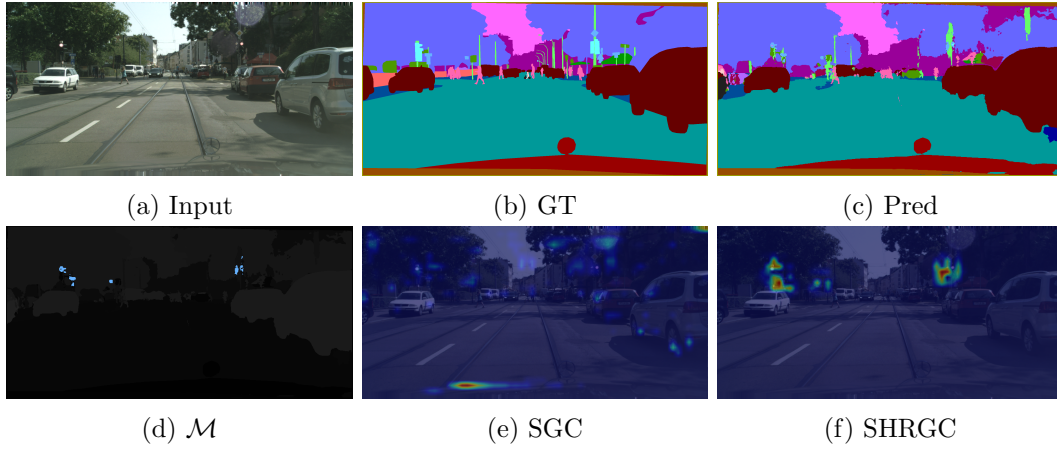
Appendix A. Additional Experiments

(a) Input   (b) GT   (c) Pred

(d) $\mathcal{M}$ (e) SGC   (f) SHRGC

Figure 5: *SGC* (Vinogradova et al., 2020) **(e)** and *SHRGC* **(f)** for the traffic sign-class (Cordts et al., 2016).

Seg-Grad CAM (Vinogradova et al., 2020) ($\mathcal{M}$ = image)   Seg-Grad CAM (Vinogradova et al., 2020) ($\mathcal{M}$ = class)   Seg-HiRes-Grad CAM ($\mathcal{M}$ = image)   Seg-HiRes-Grad CAM ($\mathcal{M}$ = class)
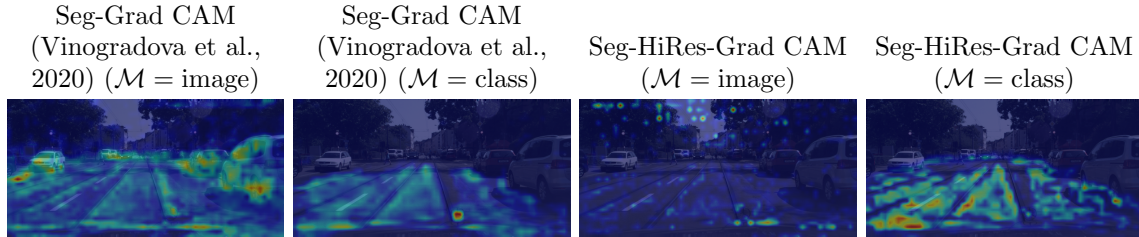
Figure 6: CAMs for the road-class with different pixel sets and the same ground truth as in Fig. 2.

(a) Input image   (b) Prediction   (c) Pixel set $\mathcal{M}$
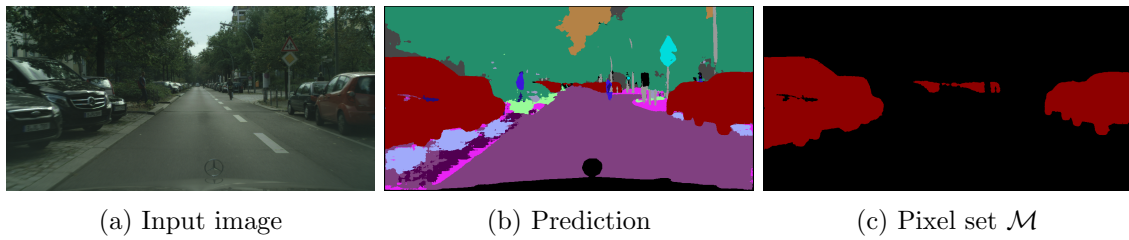
Figure 7: Input image (**a**), the predicted segmentation (**b**) and the pixel set $\mathcal{M}$ for the car-class (**c**). The ground truth is not available since the Cityscapes (Cordts et al., 2016) dataset does not provide the ground truth for the test set.
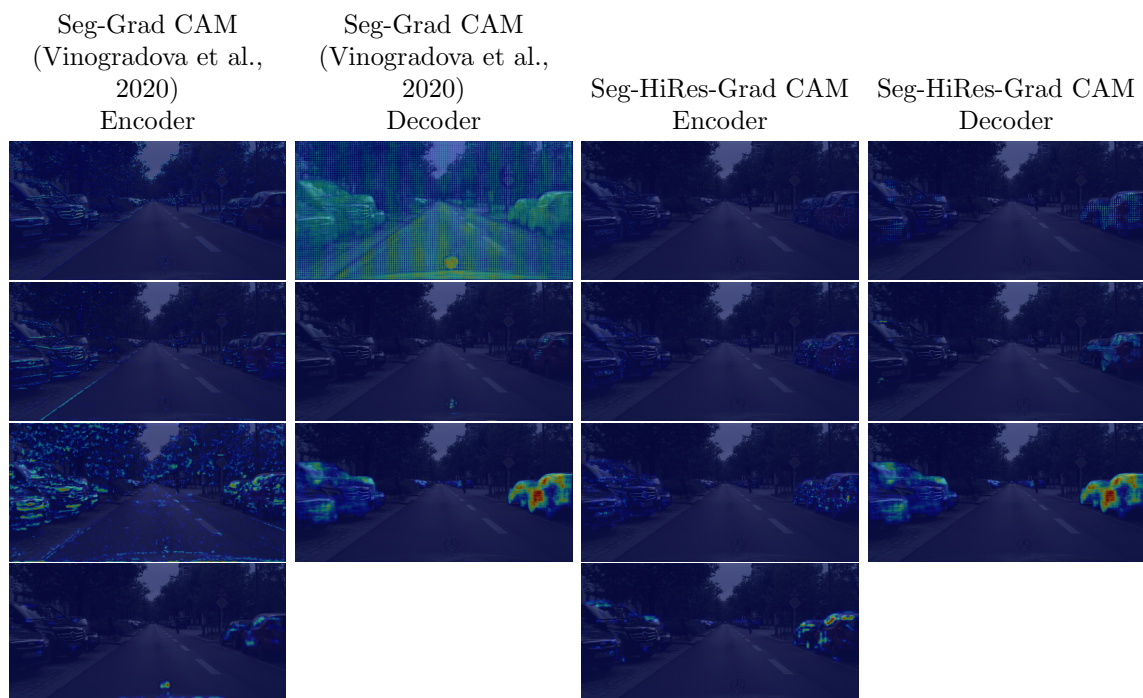
Figure 8: *Seg-Grad CAM* (Vinogradova et al., 2020) and *Seg-HiRes-Grad CAM* for the image and the pixel set shown in Fig. 7. The different rows represent the levels of the U-Net (Ronneberger et al., 2015) split into encoder and the decoder. We used the activation maps before the pooling operation but after the convolutional operations of a layer. The last row represents the lowest layer of the U-Net. In this case, the U-Net has a depth of four.