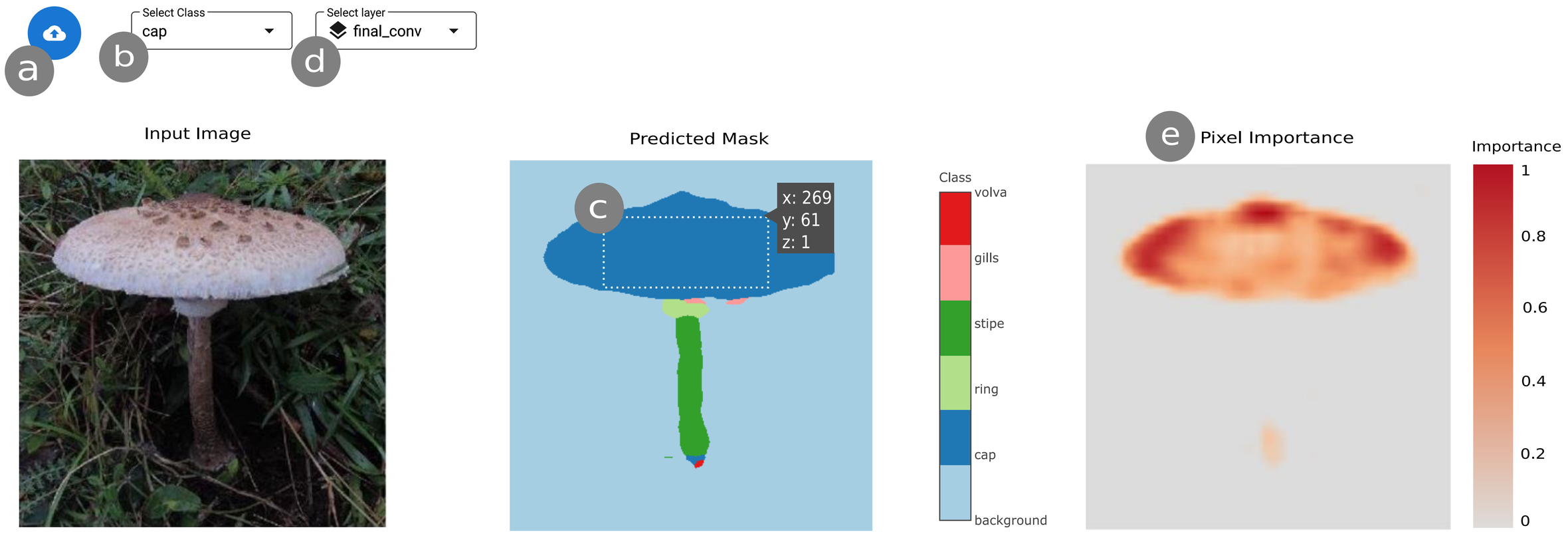
Interactive Attribution-based Explanations for Image Segmentation

Christina Humer, Mohamed Elharty, Andreas Hinterreiter, Marc Streit
Johannes Kepler University Linz

Part of the Institute of Computer Graphics at
Johannes Kepler University Linz, Austria



Explanations of deep neural networks (DNNs) give users a better understanding of the inner workings and generalizability of a network. While the majority of research focuses on explanations for classification networks, in this work we focus on explainability for image segmentation networks. As a first contribution, we introduce a lightweight framework that allows generalizing certain attribution-based explanations, originally developed for classification networks, to also work for segmentation networks. The second contribution is a web-based tool that utilizes this framework and allows users to interactively explore segmentation networks. We demonstrate the approach using a self-trained mushroom segmentation network.

Figure 1 *The web-based application allows users to interactively explore Grad-CAM explanations of segmentation networks. Users can* ***(a)*** *upload images,* ***(b)*** *choose a class for which the explanation is calculated,* ***(c)*** *choose single pixels or areas in the predicted segmentation mask to calculate the explanation, and* ***(d)*** *choose the convolutional layer for which the Grad-CAM is calculated. The* ***(e)*** *pixel importance shows the resulting Grad-CAM as a heatmap, which users can leverage for their analysis. The example of a mushroom segmentation shows that the network focuses mainly on the border pixels of the mushroom cap to segment the cap. Input image adapted from [2, 4].*

Segmentation network adaptation

Instead of adapting an *explanation method* to work with segmentation networks (like the approach proposed by Vinogradova et al. [3]), we propose to adapt the *segmentation networks* in such a way that they resemble classification networks. To this end the output shape of the network has to be reduced to give only one classification result instead of one result per pixel, which is achieved by using a global average pooling layer (GAP). Right before GAP, a custom binary mask is multiplied with the segmentation output of the network, which allows the specification of regions that are of interest to users (see **Fig 2**).

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision (ICCV'17)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74

[2] Statens Naturhistoriske Museum et al. Danmarks svampeatlas, 2022. Accessed March 29, 2022.

[3] K. Vinogradova, A. Dibrov, and G. Myers. Towards Interpretable Semantic Segmentation via Gradient-Weighted Class Activation Mapping (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13943–13944, 2020. doi: 10.1609/aaai.v34i10.7244

[4] Visipedia. 2018 FGCVx fungi classification challenge, 2018. Accessed March 29, 2022.

Interactive tool

Having attribution-based explanations is important for better understanding deep neural networks (DNNs). However, static explanations alone are insufficient for gaining a deeper understanding of the inner workings of such networks. We developed an interactive tool that facilitates the exploration of attribution-based explanations for image segmentation networks as shown in **Fig 1**. The application builds on top of the previously described network adaptation framework to calculate Grad-CAM [1] explanations for segmentation networks.

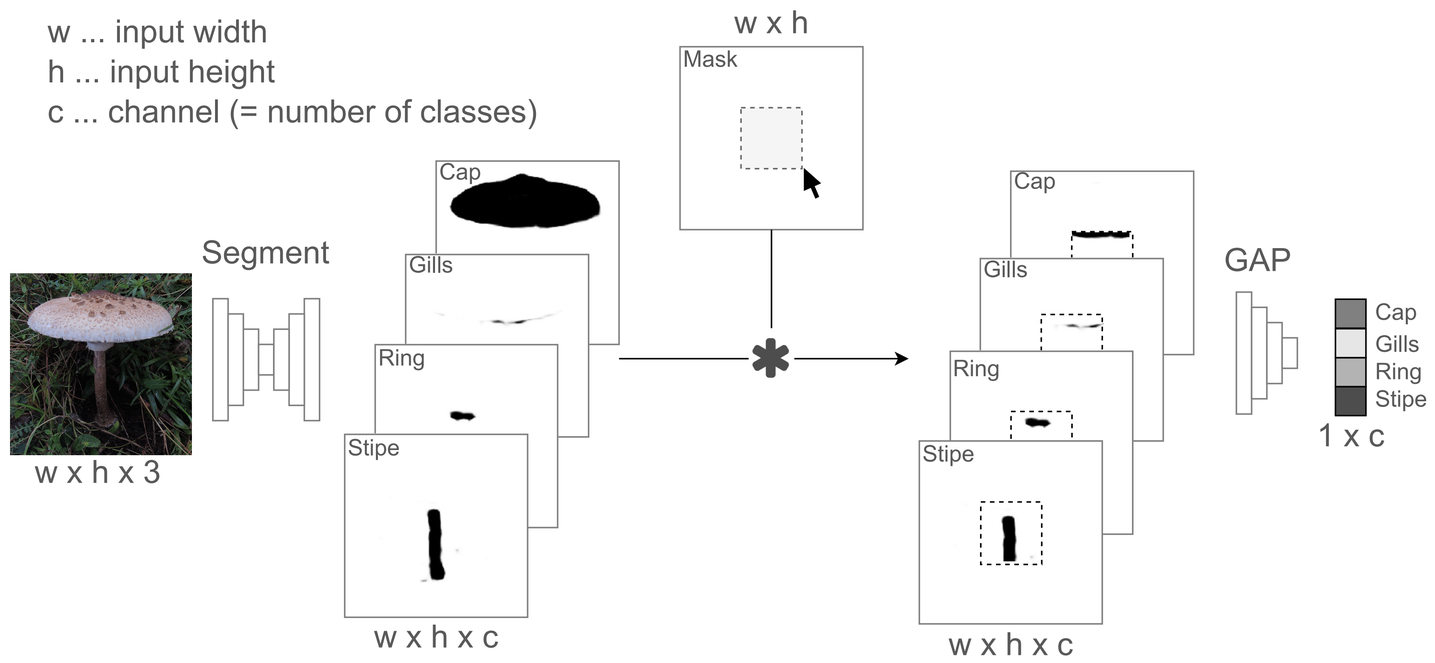


Figure 2 *Schematic overview of a framework to adapt segmentation networks in a way that makes them resemble classification networks. As usual the input image is processed by the segmentation network (****Segment****), which results in one classification per pixel ($w \times h \times c$). A custom binary-mask is then multiplied with the segmentation output to select a region of interest. Using the global average pooling layer (****GAP****) the dimension of the result is then reduced to $1 \times c$ (i.e., one output for each class), which resembles the output of a classifier. Input image adapted from [2, 4].*

Contact
christina.humer@jku.at
https://jku-vds-lab.at/publications/#posters

Acknowledgements
State of Upper Austria and the Austrian Federal Ministry of Education, Science and Research via the LIT – Linz Institute of Technology (LIT-20-7-SEE-117)
Austrian Science Fund (FWF DFH 23–N)