https://doi.org/10.15388/vu.thesis.781
https://orcid.org/0000-0001-5166-0920

VILNIUS UNIVERSITY

Rokas Gipiškis

Post-Hoc Explainable Semantic Image Segmentation: Applications for Interpretability and Adversarial Attacks

DOCTORAL DISSERTATION

Natural Sciences,
Informatics (N 009)

VILNIUS 2025

The dissertation was prepared between 2020 and 2024 at Vilnius University.

Academic Supervisor – Prof. Dr. Olga Kurasova (Vilnius University, Natural Sciences, Informatics – N 009).

This doctoral dissertation will be defended at a public meeting of the Dissertation Defence Panel:

Chairman – Assoc. Prof. Dr. Algirdas Lančinskas (Vilnius University, Natural Sciences, Informatics – N 009).
Members:
Dr. Jolita Bernatavičienė (Vilnius University, Natural Sciences, Informatics – N 009),
Prof. Dr. Božena Kostek (Gdańsk University of Technology, Poland, Natural Sciences, Informatics – N 009),
Prof. Dr. Dalius Matuzevičius (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – T 007),
Assoc. Prof. Dr. Viktor Medvedev (Vilnius University, Natural Sciences, Informatics – N 009).

The dissertation will be defended at a public meeting of the Dissertation Defense Panel at 12:00 p.m. on June 26, 2025, in Room 203 of the Institute of Data Science and Digital Technologies of Vilnius University.
Address: Akademijos Street 4, LT-04812, Vilnius, Lithuania
Tel. +370 5 210 9300; e-mail: info@mii.vu.lt

The text of this dissertation can be accessed at the Library of Vilnius University, as well as on the website of Vilnius University: *https://www.vu.lt/lt/naujienos/ivykiu-kalendorius*.

https://doi.org/10.15388/vu.thesis.781
https://orcid.org/0000-0001-5166-0920

VILNIAUS UNIVERSITETAS

Rokas Gipiškis

Post-hoc paaiškinamasis vaizdų semantinis segmentavimas: taikymai interpretuojamumui ir priešiškoms atakoms

DAKTARO DISERTACIJA

Gamtos mokslai,
Informatika (N 009)

VILNIUS 2025

Disertacija rengta 2020-2024 metais Vilniaus universitete.

Mokslinė vadovė – prof. dr. Olga Kurasova (Vilniaus universitetas, gamtos mokslai, informatika – N 009).

Gynimo taryba:
Pirmininkas – doc. dr. Algirdas Lančinskas (Vilniaus universitetas, gamtos mokslai, informatika – N 009).
Nariai:
dr. Jolita Bernatavičienė (Vilniaus universitetas, gamtos mokslai, informatika – N 009),
prof. dr. Božena Kostek (Gdansko technologijos universitetas, Lenkija, gamtos mokslai, informatika – N 009),
prof. dr. Dalius Matuzevičius (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – T 007),
doc. dr. Viktor Medvedev (Vilniaus universitetas, gamtos mokslai, informatika – N 009).

Disertacija ginama viešame Gynimo tarybos posėdyje 2025 m. birželio 26 d. 12 val. Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų instituto 203 auditorijoje. Adresas: Akademijos g. 4, LT-04812, Vilnius, Lietuva, tel. +370 5 210 9300; el. paštas: info@mii.vu.lt.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir Vilniaus universiteto interneto svetainėje adresu: *https://www.vu.lt/lt/naujienos/ivykiu-kalendorius*.

ACKNOWLEDGEMENTS

I extend my gratitude to Prof. Olga Kurasova, my doctoral advisor, for her support and counsel. I am also thankful to my reviewers, Assoc. Prof. Viktor Medvedev and Assoc. Prof. Algirdas Lančinskas, for their valuable comments and constructive feedback.

To Adelė Varaneckienė, Rita Gipiškienė, Zenonas Gipiškis, Jonas Varaneckas, and Justina Petravičiūtė, thank you for believing in me, supporting me, and being there.

ABSTRACT

Explainable AI (XAI) has become increasingly important in computer vision applications. While substantial progress has been made in explainable image classification, XAI in semantic segmentation remains underexplored despite its critical role in healthcare, autonomous systems, and other high-stakes domains. Given the widespread use of image segmentation, a systematic investigation of its explainability is needed.

This dissertation bridges this gap by focusing on post-hoc interpretability in semantic segmentation and adversarial attack scenarios. It proposes and investigates three explainability method extensions: occlusion-based, activation perturbation-based, and gradient-based approaches, all specifically designed for segmentation tasks. These methods are assessed for their trade-offs between explanation noisiness and computational efficiency. The applications of post-hoc techniques are further evaluated in adversarial attack scenarios, demonstrating that semantic segmentation explainability techniques can be successfully attacked to generate arbitrary explanations. Key contributions also include a first survey of explainability techniques, not limiting itself to a particular type of explainability method or its application domain, a comprehensive taxonomy of XAI methods in segmentation, and insights into the broader implications of explainability in high-stakes applications.

AI	Artificial Intelligence
CAM	Class Activation Mapping
CL	Continual Learning
CNN	Convolutional Neural Network
CRP	Concept Relevance Propagation
DAG	Dense Adversary Generation
DL	Deep Learning
FCN	Fully Convolutional Network
Grad-CAM	Gradient-weighted Class Activation Mapping
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagation
ML	Machine Learning
MSE	Mean Squared Error
NAS	Neural Architecture Search
PCA	Principal Component Analysis
RISE	Randomized Input Sampling for Explanation
SHAP	SHapley Additive exPlanations
SSIM	Structural Similarity Index Measure
VAE	Variational Autoencoder
XAI	eXplainable AI

$g_{c,A}(x)$	A sum of logits for c in segmentation, $g_{c,A}(x) = \sum_{i,j \in A} g_c(x_{ij})$.
A	A set of pixel indices of interest.
$G(x, c)$	A saliency map of class c for classification, $G(x, c) = \frac{\partial g_c(x)}{\partial x}$.
$G_A(x, c)$	A saliency map of class c for segmentation, $G_A(x, c) = \frac{\partial g_{c,A}(x)}{\partial x}$.
c	A class of interest.
$g_c(x)$	A prediction score before the Softmax function for class c with respect to x, $g(x) = (g_1(x), ..., g_C(x)) \in \mathbb{R}^C$.
$l^c(x_{ij})$	The logit value for a single pixel x_{ij} for class c.
x_{adv}	A perturbed image.
x_{ij}	A single pixel of x, where i and j are the row and column indices, respectively.
x	An RGB image, where $x \in \mathbb{R}^{N \times M \times 3}$, with N and M as spatial dimensions, and 3 RGB channels.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

In the past decade, Artificial Intelligence (AI) systems have achieved impressive results, most notably in natural language processing and computer vision. The performance of such systems is typically measured by evaluation metrics that vary depending on the task but aim to assess the system's outputs. Today's leading AI systems largely rely on deep learning (DL) models, multi-layered neural networks that tend to exhibit increasingly complicated structures in terms of model parameters. The growing complexity of such systems resulted in them being labeled as "black boxes." This indicates that the evaluation metric does not show the full picture: even if its measurement is correct, it does not give insights into the inner workings of the model.

The field of explainable AI (XAI) encompasses different branches of methods that attempt to give insights into a model's inner workings, explain outputs, or make the entire system more interpretable to end users, such as human decision-makers. There is ongoing debate regarding XAI terminology. Concepts like interpretability, explainability, understanding, reasoning, and trustworthiness are challenging to formalize. While some authors use "interpretable" and "explainable" interchangeably [156], others distinguish between the two [175, 176]. When the distinction is made, it is usually to demarcate post-hoc explanations, a type of XAI techniques applied after the model has been trained, and inherently interpretable models [176]. Post-hoc can be understood as referring to the fact that XAI techniques are applied after training and interpret the results of the model, rather than its internal structure. In contrast, inherently interpretable models are designed in such a way that their inner workings can be understood directly without additional explanation techniques. This way interpretability becomes associated with the transparency of the model itself and depends on the ease with which one can interpret the model. For instance, a simple decision tree-based model might be considered more interpretable than a DL model composed of millions of parameters, provided that the former is not too deep. Explainability, in contrast, is often limited to understanding the model's results rather than the model as a whole. In this thesis, "interpretable" and "explainable" will be used interchangeably, while more specific "architecture-based" and "inherently interpretable" terms will be used when discussing model-specific XAI modifications. This is because not

many of the surveyed papers use the term interpretability in a second sense. Since most papers in explainable segmentation do not make this distinction, this might avoid unnecessary confusion when discussing their contents. It should also be noted that interpretability and ease of understanding vary according to the specific audience, whether it be the general public or a more specialized group with specific training, such as radiologists.

Explainable segmentation requires pixel-level explanations, making it a more complex and challenging task than explainable classification. Compared to explainable image classification, the number of explainability techniques available for image segmentation is limited. Many explainability methods developed for classification have yet to be extended to segmentation tasks, and even among the existing approaches, several lack comprehensive investigation. This gap is particularly noticeable given the widespread applications of image segmentation, ranging from medical to industrial domains. New explainability techniques for image segmentation are still being developed, and there is a lack of studies at the intersection of explainable segmentation and AI safety, particularly in evaluating their robustness against adversarial attacks. This dissertation seeks to expand the limited number of XAI techniques in image segmentation. It also investigates adversarial attacks targeting explainable segmentation. Furthermore, it provides a comprehensive survey of XAI in image segmentation and discusses how the proposed explainability techniques could contribute to XAI-driven model improvements.

Research Focus

The research focus is on post-hoc explainability methods for interpreting DL models, particularly convolutional neural networks (CNN), in semantic image segmentation.

Research Aim and Objectives

The research aim is to develop novel methods for explainable segmentation suitable for convolutional neural networks, and evaluate their susceptibility to adversarial attacks. To achieve this aim, the following objectives are set:

- Investigate existing interpretability methods for image classification and segmentation, identifying the most suitable solutions for convolutional neural networks. Based on this investigation, to prepare a comprehensive survey and taxonomy of XAI methods in image segmentation.
- Extend and implement new XAI techniques in segmentation based on XAI methods in classification, such as occlusion-based, activation perturbation-based, and gradient-based approaches, evaluating them both qualitatively and quantitatively.
- Investigate the potential deployment of interpretable semantic segmentation techniques in adversarial settings, evaluating both their defensive capabilities and susceptibility to adversarial attacks.

Scientific Novelty

1. This work proposes a comprehensive survey of explainability techniques in image segmentation, not limiting itself to a particular type of explainability method or its application domain, and includes a comprehensive taxonomy of explainable segmentation techniques.
2. This work presents an extension of Ablation-CAM [62], a widely used explainability technique, adapted for semantic image segmentation models.
3. This work provides a systematic investigation of input perturbation-based XAI techniques, evaluating the impact of varying input image occlusion sizes and colors on model outputs, with a focus on both qualitative and quantitative metrics.
4. This work demonstrates that it is possible to construct a successful adversarial attack against post-hoc explanation techniques in semantic segmentation, extending the original work [65] in image classification.

Practical Significance

XAI in image segmentation is a relatively new field, with the first articles on the subject appearing in the late 2010s [107, 212, 219]. Since then, the topic has gained more attention. Semantic image segmentation is an essential task in computer vision, with applications ranging from autonomous driving [76] to medical image analysis [17]. Its study is further motivated by the rapidly growing remote sensing and video data. Increasing deployments in medical AI are also contributing to the need for explainable segmentation. Both radiologists and surgeons need to know accurate boundaries for the anatomical structures of interest. Precise and reliable segmentation is required when working with most pathologies in different imaging modalities, ranging from magnetic resonance imaging (MRI) to computed tomography (CT).

Advances in XAI methods for image segmentation can address the growing demand for trustworthy AI in high-stakes domains. Improved post-hoc explainability can enhance user trust and facilitate regulatory compliance, particularly in medical and sensitive industrial applications. A thorough analysis of the use of XAI techniques in adversarial scenarios can aid in identifying and mitigating adversarial risks associated with explainable segmentation methods, contributing to the development of secure AI solutions. The taxonomy and methods proposed in this research can provide practitioners with tools to systematically evaluate and deploy explainability techniques, bridging the gap between theoretical advancements and real-world applications.

Statements to be Defended

- Perturbation-based post-hoc explainability methods, applied to both input and activation spaces, are suitable for interpreting the outputs of CNN-based semantic segmentation models.
- Post-hoc interpretable segmentation techniques are applicable in adversarial contexts, both in detecting and enabling adversarial attacks.

Peer-Reviewed Journal Publications and Conference Presentations

Articles in international research journals with a citation index in the Clarivate Analytics Web of Science (CA WoS) database.

1. Gipiškis, R., Chiaro, D., Preziosi, M., Prezioso, E. and Piccialli, F., 2023. The impact of adversarial attacks on interpretable semantic segmentation in cyber–physical systems. *IEEE Systems Journal*, 17(4), pp. 5327-5334. https://doi.org/10.1109/JSYST.2023.3281079

2. Gipiškis, R., Tsai, C.W. and Kurasova, O., 2024. Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey. *ICT Express.*, 10(6), pp. 1331-1354. https://doi.org/10.1016/j.icte.2024.09.008

International conference presentations:

1. Gipiškis, R. and Kurasova, O., Occlusion-based approach for interpretable semantic segmentation. *18th Iberian Conference on Information Systems and Technologies (CISTI)*, June 20–23, 2023, Aveiro, Portugal.

2. Gipiškis, R., XAI-driven Model Improvements in Interpretable Image Segmentation. *2nd World Conference on eXplainable Artificial Intelligence*, July 17–19, 2024, Valletta, Malta.

National conference presentations:

1. Gipiškis, R. and Kurasova, O., Application of CNNs for brain MRI image segmentation. *12th Conference on Data Analysis Methods for Software Systems*, December 2–4, 2021, Druskininkai, Lithuania.

2. Gipiškis, R. and Kurasova, O., Investigating post-hoc explainability techniques for image segmentation. *15th Conference on Data Analysis Methods for Software Systems*, November 28-30, 2024, Druskininkai, Lithuania.

Outline of the Thesis

This doctoral thesis consists of an introduction, four chapters, conclusions, and a summary in Lithuanian. The introduction provides an overview of XAI, emphasizing its importance in semantic image segmentation, and presents the research object, aim, objectives, scientific novelty, practical significance, and statements to be defended. Chapter 1 reviews the literature on XAI, discussing its development in computer vision, proposing a taxonomy for explainable semantic segmentation methods, and highlighting the limitations of existing techniques. Chapter 2 introduces the methods developed in this thesis, including perturbation-based, gradient-based, and adversarial approaches for explainable semantic segmentation, detailing their theoretical foundations and practical implementations. Chapter 3 presents the experimental evaluation of the proposed methods, focusing on their performance in interpretability, computational efficiency, and robustness under adversarial scenarios. Chapter 4 discusses the open challenges in XAI for semantic segmentation, evaluates trade-offs between interpretability and robustness, and suggests directions for future research, including hybrid and self-supervised methods. The dissertation concludes with a summary of key findings and contributions, implications for practical deployment, and recommendations for advancing the field. Bibliographic references are included at the end. The dissertation consists of 153 pages, 40 figures, and four tables.

1. SURVEY OF XAI IN IMAGE SEGMENTATION

This chapter provides an in-depth survey of XAI techniques in semantic image segmentation. It introduces the foundational concepts of XAI, examines the development of explainability methods in computer vision, and discusses their applications. A detailed taxonomy of XAI methods for segmentation is proposed. The chapter also highlights limitations in existing methods and underscores the need for advancements to address challenges in interpretability, particularly in high-stakes applications like healthcare and autonomous systems. The main results presented in this chapter have been published in [A.2].

XAI is not a new development, particularly in rule-based expert systems [190, 202] and machine learning (ML) [79], but it has experienced unprecedented growth since the revived interest in neural networks in 2012 [131]. This growth correlates with the increasing interest in DL and is further driven by: (1) the need for trustworthy models due to widely expanding industrial deployments [B.5]; (2) bureaucratic and top-down political emphasis on AI regulation [B.4]; and (3) concerns within the ML safety community [11] about the general trajectory of AI development in the short and long run. AI deployment is increasing across different sectors, and is significant both in terms of its size and impact. According to the AI Index Report 2023 [152], the proportion of companies adopting AI more than doubled from 2017 to 2022. In 2022, the medical and healthcare sectors attracted the most investment, with a total of 6.1 billion dollars [152]. The IBM Global AI Adoption Index 2023 [160], conducted by Morning Consult on behalf of IBM, indicates that about 42% of their surveyed ($> 1,000$ employees) enterprise-scale companies reported actively deploying AI, and an additional 40% exploring and experimenting with AI, out of which 59% reported an acceleration in their rollout or investments. Even with rapid deployment, critical high-impact sectors tend to move at a slower pace. One could expect even more healthcare-related applications and clinical deployments if AI methods were more interpretable. To a large extent, this applies to other industries as well. According to the same IBM report, most of the surveyed IT professionals (83% among companies already exploring or deploying AI) stated that it is important that their business explain how its AI reached a decision. Another accelerating trend is that of AI regulation (Fig. 1.1). The recent survey [56] indicates that 81% of re-

spondents ($N > 6{,}000$) expect some form of external AI regulation, with 57-66% of respondents reporting that they would be more willing to use AI systems if trustworthiness-assuring mechanisms were in place. AI trustworthiness and transparency are further emphasized in regulatory discussions, ranging from the EU's AI Act [52] to AI executive order [205] in the United States.

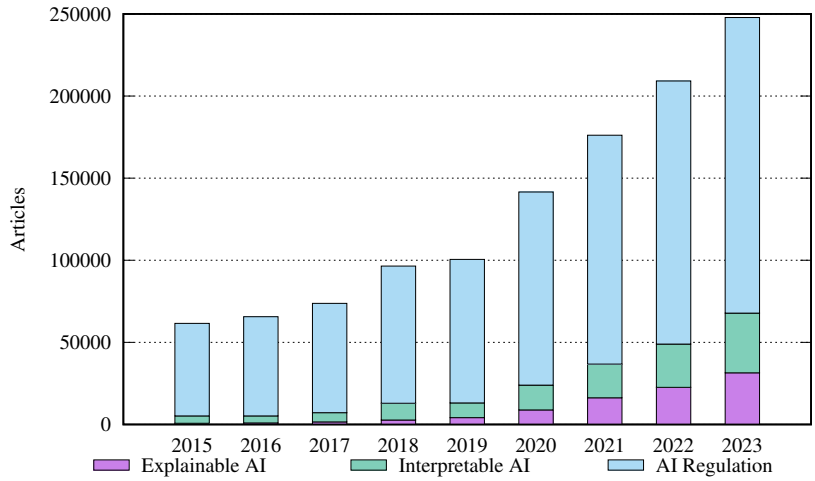


Figure 1.1: Publications with "explainable AI," "interpretable AI," and "AI regulation" as keywords. Publication data gathered from app.dimensions.ai, part of the Dimensions research analytics platform by Digital Science. The publication number is based on full-text data, not limited to titles and abstracts, and includes preprints from arXiv and SSRN.

Segmentation is commonly viewed as a dense prediction task where classification is performed on a pixel level. However, most XAI literature so far has focused on image classification tasks. Nonetheless, a growing number of works address the issue of interpreting semantic segmentation results by either extending classification-based methods or by proposing their own modifications. Two Ph.D. dissertations [162, 211] on XAI in image segmentation were written in 2023, but neither provided a comprehensive survey of the field, as they primarily focused on directly related works. Image segmentation methods have been reviewed in the medical domain [97], however, the focus has been just on the post-hoc techniques.

1.1. Background

1.1.1. Development of the Field of XAI in Computer Vision

There is a great variety of XAI methods in classification, with new techniques being proposed weekly. Typically, these methods employ some form of feature attribution, indicating the model's sensitivity or insensitivity to various features, such as certain pixel configurations in the input space. The most popular explainable classification methods, still influential in today's DL models, fall into gradient-based or perturbation-based categories. This subsection highlights key developments, with a particular focus on the methods that have influenced interpretable image segmentation. For an accessible introduction to and treatment of explainable classification, readers may refer to [158]. A more detailed survey on these topics can be found in [234].

The first gradient-based explainability techniques for classification in CNNs are proposed in [192]. The initial method generates artificial images that maximize the score for the selected class of interest. The second method, also referred to as vanilla gradient, produces a saliency map[1] that highlights important regions in the input space. This is based on the gradient for the class of interest with respect to the input. The authors also observe that this method can be used for weakly supervised segmentation. This marks the possibility of using XAI tools instrumentally, not just for the sake of explainability. In [188], the influential Grad-CAM technique is introduced. Its calculation is based on the gradient flow into the last convolutional layer. Since Grad-CAM is calculated for intermediate model activations, the resulting explanation needs to be upsampled. This upsampling process might negatively impact the quality of pixel-level explanations [129]. Similar to [192], the Grad-CAM technique also demonstrates the potential for instrumental use in weakly supervised localization.

Another area of explainable classification methods encompasses occlusion or perturbation-based techniques, which assess a model's decision-making by systematically occluding (or perturbing) the input and observing the impact on the output. This type of method was first

[1] Even though saliency maps are sometimes considered a specific type of heatmap, for the purposes of this dissertation, unless stated otherwise, the term "heatmap" will be used synonymously with "saliency map."

introduced in [227] under the name of occlusion sensitivity. It proposes systematically occluding the input image with a smaller grey filter and measuring the effect on the model's output. The likelihood of the model classifying the image as belonging to the actual class should decrease when the object of that class is occluded in the input space. Other perturbation-based techniques include LIME [173], SHAP [147], and RISE [169] which extend this idea in different ways.

Other noteworthy methods in explainable classification have focused on optimization. Activation maximization, previously proposed in [71], initially focused on restricted Boltzmann machines, a type of unsupervised models. In [192], it has been specifically implemented in supervised classification models. In [166], this technique was further popularized by demonstrating the results across different network layers. Unlike the previously discussed XAI techniques, this type of explanation method can be described as global because the generated image does not depend on a particular input image but rather on the model's internal weights.

XAI Research in Lithuania

At the national level, the XAI use has been primarily investigated in the financial sector [33, 34]. In [33], XAI is discussed in the context of multi-criteria decision-making methods, including their review and classification. In [34], another review covers the XAI applications in finance from 2005 to 2022, identifying LIME and SHAP as the most commonly used explainability methods. However, these studies do not directly engage with computer vision or image segmentation in particular. Medical applications of XAI have also been investigated for classification tasks [163, 204]. In [163], a soft attention mechanism is presented for CNNs, and its dermatological applications are investigated for the classification of skin cancer images. In [204], a spatial attention-based module is presented for the classification of brain MRI scans. In both cases, the proposed methods are not post hoc, and are based on architectural modifications. To the best of our knowledge, there has been no research at the intersection of XAI, image segmentation, and adversarial robustness.

1.1.2. Specifics of Semantic Segmentation

Most of the literature on interpretable computer vision focuses on classification. However, DL-based semantic segmentation techniques have achieved significant results. Classical encoder-decoder models such as U-Net [174] or SegNet [21] as well as their modifications, have been deployed in various fields. Vision transformer-based segmentation architectures have also been proposed [199]. There have even been attempts to combine these two approaches [37]. During semantic segmentation, class labels are assigned to each pixel, and the output is typically the same resolution as the input image. Modern segmentation models can be composed of millions of parameters, making their interpretation difficult and often resulting in their description as "black boxes."

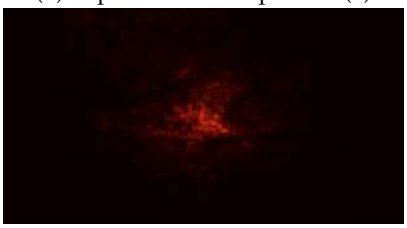
Interpretability in semantic image segmentation is a challenging area of study. On one hand, it can be viewed as an extension of a relatively intuitive interpretable classification. However, it requires combining the relative influence of each classified pixel of interest. On the other hand, interpreting its own explanations is not so straightforward or intuitive. One problem with interpretability methods, not limited to semantic segmentation, is the lack of ground truths for explanations. Furthermore, it is uncertain what the ideal explanation should look like or whether one interpretable instance can be limited to a single explanation. In classification, at least some candidates for good explanations exist, allowing for qualitative human-based studies [126]. Conducting a similar study for semantic segmentation is more complex, as it is less clear what constitutes good explanation candidates: should the interpretability saliency map focus on the entire area of the class of interest or just its boundaries? Can there be instances where the most salient features are outside the class area? What if the segmentation area is correct, but the attributed class is not? Moreover, semantic image segmentation is notorious for inter-observer variability, especially in manual delineations in medical images. One way to demonstrate the usefulness of explainable segmentation is to detect instances where the segmentation of one semantic class appears heavily dependent on the presence of different class pixels, whether nearby or otherwise. In [212], such a case is demonstrated when the U-Net detects the sky primarily due to the nearby trees, which belong to the *Nature* class. Interpretable semantic

segmentation techniques prove most useful when the segmentation is incorrect.

(a) Pixel of interest (top leftmost)

(b) Explanation for a pixel in (a)

(c) Pixel of interest (centermost)



(d) Explanation for a pixel in (c)

Figure 1.2: Explanation for single pixels: the selected pixels are shown on the left, with their corresponding gradient-based explanations on the right.

Since the segmentation task can be framed in terms of classification, applying explainable classification methods to it is relatively straightforward when focusing on a single pixel, as seen in Figure 1.2. For instance, a gradient for the selected output pixel of a chosen class can be calculated with respect to the entire input image. However, an explanation map for the classification of a single pixel is not particularly useful. It is less accessible to the human interpreter, as evaluating thousands of different explanations for just a single class in a single image would be required. Therefore, considering the effects of a larger number of pixels becomes necessary. Most popular explainable segmentation techniques operate under the underlying assumption of pixel importance. This assumption is particularly relevant to perturbation-based methods (Fig. 1.3)[2], where introducing noise to important pixels would degrade a model's performance more significantly than adding it to less critical pixels. To

[2]Here, and in Figures 1.5–1.10, only high-level frameworks are presented. Concrete implementation details might differ depending on the use case and the specific subtype of XAI method within each group in the provided taxonomy (Fig. 1.4).

explain the whole image (i.e., all pixels) instead of just a single pixel, most explainable segmentation techniques must visualize the relative contributions of all pixels simultaneously. Otherwise, the analysis of separate single-pixel-based explanation maps would be too tedious. The most popular way to do it involves using logit values, unnormalized probabilities before the Softmax layer, typically used in classification. This could be achieved, for instance, by summing up the logits of the class of interest for the pixels of interest. This new scalar value can then be used when generating a single explanation for the entire image, just like in the case of a single pixel.

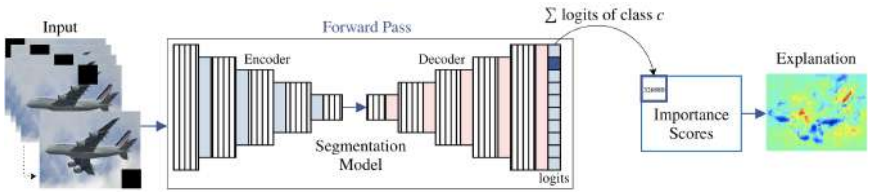


Figure 1.3: Explanation for multiple pixels: the application of a perturbation-based method to the COCO [141] dataset using summed up logit values.

1.1.3. Limitations

Feature attribution and saliency-based XAI methods in particular have faced criticism [5, 127, 182]. Although these criticisms have solely focused on explainable classification, they deserve a thorough examination as they could also extend to segmentation. Some of the XAI methods act as regular edge detectors, independently from the underlying model and training dataset. This independence is troubling because a local post-hoc XAI method should explain a specific model's prediction for a particular data point. In [29], limitations of feature attribution methods such as SHAP and integrated gradients are emphasized both theoretically and empirically, showing that they cannot reliably infer counterfactual model behavior. The authors observe that the analyzed attribution methods resemble random guessing in tasks like algorithmic recourse and spurious feature identification. Similar experimental results are observed with gradients, SmoothGrad [196], and LIME [173].

Attribution methods have also been criticized for confirmation bias [12]. An appealing but incorrect explanation might be judged more favorably than a more realistic one. A better understanding of the goals of an idealized attribution method is needed to develop improved quantitative tools for XAI evaluation [12]. In [6], the limitations of post-hoc explanations are investigated. The authors question their effectiveness in detecting unknown (to the user at test time) spurious correlations. These inefficiencies are detected in three types of post-hoc explanations: feature attribution, concept activation, and training point ranking. However, the authors acknowledge that these three classes do not fully cover all post-hoc explanation methods. Other methods have been criticized for their weak or untrustworthy causal relationships. In [19], saliency maps are criticized for their frequent unfalsifiability and high subjectivity. The study also highlights their causal unreliability in reflecting semantic concepts and agent behavior in reinforcement learning environments. In [164], it is argued that feature attribution techniques are not more effective than showing the nearest training-set data point when tested on humans. The limitations of attribution methods in cases of non-visible artifacts [239] have also been investigated.

Despite the critical studies on explainable classification and their potential extensions to segmentation, the widespread prevalence of image segmentation requires investigating different explainability tools and their working mechanisms. Although some studies point out the limitations of these techniques, better alternatives have yet to be developed. As observed in [85], the development of interpretability methods is dialectical: a new method is introduced, its failure modes are identified, and as a result, a new method is proposed, with the ongoing aim of making them more reliable. Current methods have much room for improvement, especially considering that the entire field is in the early stages of development. The above criticisms can serve as sanity checks for XAI methods. Despite the limitations, some techniques, such as gradients and Grad-CAM in the case of [5], do pass certain sanity checks. Even some critical literature [85] agrees that certain explainability techniques can be useful for exploratory use cases. No studies have yet explored the specifics of XAI limitations in image segmentation.

1.2. Taxonomy

Different XAI taxonomies have been introduced in classification, both with respect to specific subgroups of interpretability methods [16, 73] and more abstract conceptual terms [93]. Even meta-reviews of various existing taxonomies have been proposed [185, 197]. Since image segmentation can be seen as an extension of classification, many taxonomy-related aspects can be validly transferred from research in explainable classification. In most taxonomies, a particularly important role is played by three dichotomies: post-hoc vs ad-hoc (sometimes also referred to as inherent interpretability), model-specific vs model-agnostic, and local vs global explanations. This section provides a brief overview of these high-level dichotomies before introducing a low-level taxonomy based on the surveyed papers, categorizing various interpretability techniques into five subgroups based on how the explanation is generated.

Scope: Local vs. Global

The first prevalent dichotomy distinguishes between local and global explanations. Here, locality refers to the use of a single input image with respect to which the explanation is given. A global explanation, on the other hand, would aim to explain the model's behavior across a range of different images, not limiting itself to just one. According to meta-surveys [185, 197], the local-global dichotomy is prevalent in numerous XAI taxonomies. This distinction is essential in explainable segmentation as well, with most methods falling under local explanations.

Method and Its Timing: Post-Hoc vs. Ad-Hoc

The distinction between post-hoc and ad-hoc explanations highlights that one can either apply XAI techniques to an already-trained model without any interference or apply them during and as part of the training process. Sometimes, these explanations are also described as passive and active approaches [234]. Under this definition, active approaches require modifications to the network or the training process. Such changes influence both the model's performance in terms of evaluative metrics and its interpretability. Therefore, an accuracy-interpretability trade-off cannot be avoided in ad-hoc XAI methods, but it is avoided in the case of post-hoc applications.

This widely accepted dichotomy can nonetheless be slightly misleading, as both terms can be meant to emphasize different distinctive criteria. Post-hoc can be understood as referring to the fact that XAI techniques are applied after the training, hence "post". Naturally, it would seem that ad-hoc should be understood as referring to XAI techniques that are applied during training. However, sometimes, as a direct opposition to "post-hoc", terms like "inherent interpretability" [176] or "self-explainability" [178] are used, pointing to an entirely different aspect: the architecture or type of XAI method. In some cases, such interpretation could allow for XAI methods that are both inherently interpretable and post-hoc [158], which might cause confusion.

Range: Model-Specific vs. Model-Agnostic

The third distinction evaluates the flexibility of a given XAI technique in its application to different model architectures. Model-specific XAI methods heavily depend on the underlying model architecture, whereas model-agnostic methods are more universal in their compatibility with various models, and can be applied to different architectures without further modifications. The interpretation of inherently interpretable models is always model-specific [158].

XAI Taxonomy for Image Segmentation

Multiple compatible taxonomies are possible depending on the level of abstraction of interest. In [16], XAI methods in ML are divided into transparent models and post-hoc explainability methods, with post-hoc methods further categorized as model-specific or model-agnostic. In [73], interpretation methods are divided into post-hoc interpretability analysis and ad-hoc interpretable modeling. In [189], a higher-level taxonomy distinguishes between structural analysis, behavioral analysis, and explainability by design. In [45], a preliminary taxonomy of human subject evaluation in XAI is introduced, which might be particularly useful when using qualitative evaluations of XAI. Analysis of XAI taxonomies in classification suggests that they could also be applied to image segmentation. However, no specific framework has been introduced to address the ever-growing field of interpretable segmentation. A more detailed demarcation may be useful in navigating across different types of techniques.

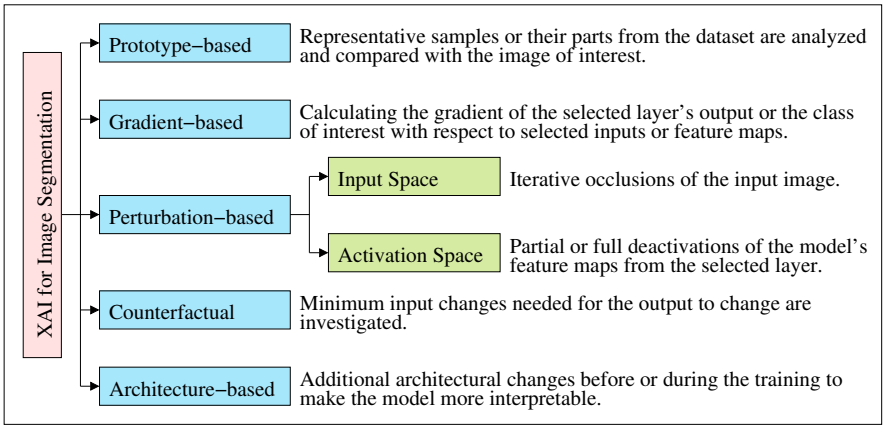


Figure 1.4: Method-centered taxonomy for explainable image segmentation.

This survey [A.2] proposes a taxonomy (Fig. 1.4) that is based on the reviewed literature in explainable image segmentation. The proposed method-centered taxonomy includes five method families: prototype-based, gradient-based, perturbation-based, counterfactual methods, and architecture-based techniques. Based on the previously discussed high-level taxonomies commonly presented in other surveys, most of the techniques fall under the local and post-hoc explainability categories. Exceptions are primarily found in the architecture-based category, which includes examples of ad-hoc, global, and model-agnostic explanations. Prototype-based methods employ representative samples or their parts from the dataset to analyze and compare with the input image. Gradient-based methods involve calculating the gradient of the output of a selected layer or the class of interest with respect to selected inputs or feature maps. Perturbation-based methods can be divided into two groups based on the perturbed space. Input space perturbations are iterative occlusions of the input image. Typically, they are based on a sliding filter, but different types of noise can also be introduced. Explanations are based on their effect on the model's outcome. Activation space perturbations involve partial or full deactivations of the model's feature maps from the selected layer. Once again, explanations are based on their effect on the model's outcome. Counterfactual methods employ the minimum input changes needed for the output to change. Finally, architecture-based techniques involve making additional architectural

changes either before or during training to enhance interpretability. Section 1.3 presents a more detailed analysis of each method group.

1.3. XAI for Image Segmentation

This section reviews the main methods representative of each subgroup in the taxonomy, as well as the metrics for explainable image segmentation.

1.3.1. Methods

Prototype-Based Methods

Prototype-based models [27] utilize typical representatives from the dataset, usually selected from the training set. These methods emphasize the intuitiveness of the provided explanations, presenting them in an easily understandable form of naturally occurring objects. Such features can be easily distinguished and discriminated by end users. Meanwhile, prototypical parts refer to specific regions within representative prototypes, also known as exemplars. In contrast to a prototype, a criticism is a data instance that is not well represented by the prototypes [158]. In terms of architecture, typical prototype-based methods require the insertion of a prototype layer into the segmentation model. Therefore, depending on the taxonomy, prototype-based methods could also be viewed as self-explainable and part of the architecture-based methods. However, due to their frequent mentions in the related classification literature under the same subgroup label, they are treated here as a separate group.

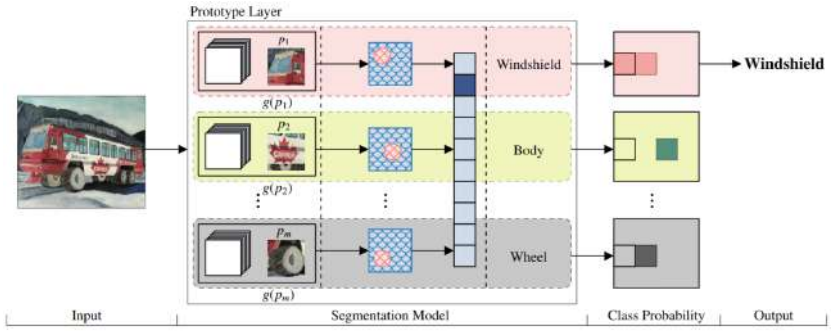


Figure 1.5: A framework for prototype-based methods.

Although prototypical methods are prevalent in classification [36, 66, 177], their extensions for segmentation are few. Typically, the prototype layer (Fig. 1.5) is a key component in prototype-based methods for both classification [36, 66, 177] and segmentation [178, 235]. Within a prototypical layer, different classes are represented by predefined or learned prototypes. In [178], a ProtoSeg model is proposed. The authors introduce a diversity loss function based on Jeffrey's divergence [112] to increase the prototype variability for each class. Better results are observed when the diversity loss component is introduced. The authors attribute this to the higher informativeness of a more diverse set of prototypes that leads to a better generalization. This could be related to the diversity hypothesis [104], first introduced in the context of reinforcement learning, and could be explored further. The experiments are performed using the Pascal VOC 2012 [72], Cityscapes [53], and EM Segmentation Challenge [1] datasets. The DeepLab [39] model is used as the backbone. In [235], a prototype-based method is used in combination with if-then rules for the interpretable segmentation of Earth observation data. The proposed approach is the extension of xDNN [13] and uses mini-batch K-mean [186] clustering. For the feature extraction part, the U-Net architecture is used. The experiments are performed using the Worldfloods [153] dataset.

Counterfactual Explanations

Counterfactual or contrastive explanations investigate the minimum input changes needed for the output to change. Unconditional counterfactual explanations were first introduced in [213]. This explainability subfield is related to adversarial attacks. Counterfactual images are similar to the original but can change the model's output. Counterfactual explanations can also be viewed as closely linked to perturbation-based explanations, which will be discussed in the next subsection. Counterfactual XAI techniques frequently fall into the local post-hoc category [95]. After the initial segmentation model, counterfactual-based interpretability methods typically employ additional networks for counterfactual generation. In the proposed pipeline (Fig. 1.6), this is depicted by additional encoder and decoder networks.

Generator-based counterfactual explanations are investigated in [229]. OCTET, a generative approach, produces object-aware counterfactual explanations for complex scenes. Counterfactual changes to the

image focus on road markings, such as changing the solid line into a dashed one, or the positions of cars by cropping and extending the relevant regions of the input image. The models are trained on the BDD100k [225] and BDD-OIA [223] datasets. Additional information can be found in the supplementary material [230]. In [194], segmentation results are qualitatively compared using counterfactual images. The experiments are performed on Kvasir-seg [113] and Kvasir-instrument [114] datasets. Counterfactual explanations are generated using the segmented area of interest, which is then replaced with the average pixel value of the rest of the image. In [109], counterfactual explanations are generated for complex scenes while preserving the semantic structure. The proposed method uses semantic-to-real image synthesis. Here, a noticeable contrast can be drawn between this approach and perturbation-based methods. In the latter, perturbations applied to the input space fail to produce semantically meaningful image regions. While perturbation-based techniques are related to counterfactual explanations, they form a distinct class, even though counterfactual explanations may also include perturbations.

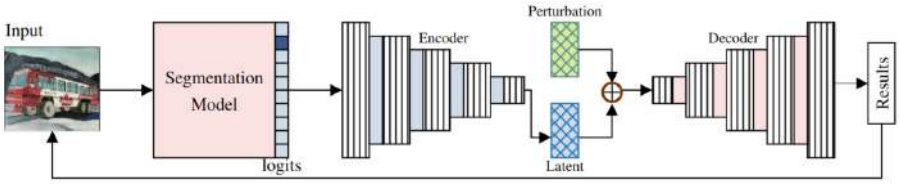


Figure 1.6: A framework for counterfactual methods.

Perturbation-Based Methods

Perturbation-based methods typically employ occlusions in the input (Fig. 1.7) or activation (Fig. 1.8) space, and then measure their influence on the model's output for a selected class. Here occlusions or perturbations could be understood as uninformative regions, transforming the input or its internal feature maps. Pixels in the occlusion filter can be set to 0 (representing black), as seen in Figure 1.7, or any other arbitrary value. Such sliding filter would occlude different regions of the input space. Its size and stride parameters are specified beforehand. Gaussian or any other random noise can also be used for these purposes. Multiple perturbative iterations are required for the generation

of an explanation map. During each inference, the score is calculated, measuring the perturbation's effect on the model's performance. This can be done by taking the difference between the score for the original image and that of its perturbed version. Such a score can be based on an evaluative metric or pre-Softmax prediction values. Since the same input image has to undergo multiple transformations, each requiring a separate forward pass through the model, perturbation-based XAI methods are considered computationally expensive. Another limitation is that perturbation-based input modifications can sometimes produce images that lie outside the original training distribution. These modified images may not resemble any data the model has previously encountered, which can lead to explanations that are not representative of the model's typical decision-making process. This, in turn, may generate misleading or irrelevant insights into how the model operates on actual, in-distribution data.

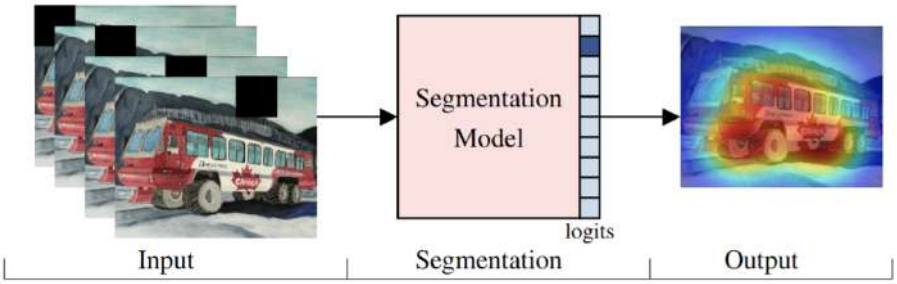


Figure 1.7: A framework for perturbation-based methods for the input space.

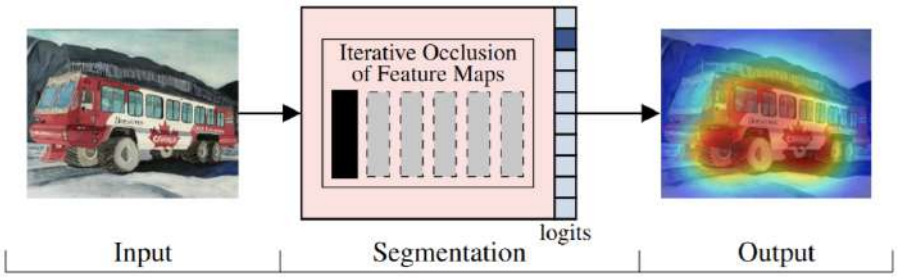


Figure 1.8: A framework for perturbation-based methods for the activation space.

The authors of [107] propose the first XAI solution for extending saliency techniques beyond classification. Their perturbation-based method is introduced for the detection of contextual biases. The experiments are performed using a synthetic toy dataset based on MNIST [137] as well as the Cityscapes [53] dataset. In [214], a hybrid SegNBDT approach is introduced, combining both decision trees and neural networks. This method falls under both the perturbation-based and self-explainable model categories. For the experimental part, the Pascal Context [161], Cityscapes [53], and Look Into Person [90] datasets are used. In [59], SHAP and RISE techniques are applied to image segmentation. SHAP is a popular post-hoc interpretability method, and the proposed approach is based on Kernel SHAP [147]. The experiments are performed on synthetic-aperture radar images from the unspecified dataset for oil slick detection at the sea surface and the Cityscapes [53] dataset. In [129], a perturbation-based occlusion sensitivity approach is used to measure the performance of the proposed interpretable semantic segmentation approach. Compared to occlusion sensitivity and Grad-CAM, their method achieves orders of magnitude lower inference time. However, it requires training an additional interpretability model. In [B.1], following [227], different types of input occlusions are investigated for applications in semantic segmentation. The paper discusses how occlusion filter sizes and colors can affect the generated explanations. It is observed that compared to image classification, input occlusions in segmentation models do not generate as much variance in evaluation metric scores. For the experimental investigation, the COCO [141] dataset is used. The proposed method is evaluated qualitatively, with select images also compared quantitatively using deletion curves.

Perturbations are not limited to the input space. For instance, Ablation-CAM [62] is a gradient-free method that systematically deactivates feature maps in a selected layer. In [B.2], Ablation-CAM is extended to semantic segmentation. It is a gradient-free interpretability technique based on ablating or perturbing activation maps. The experiments are performed on a private industrial dataset for fruit-cutting machines as well as on the COCO [141] dataset.

Gradient-Based Methods

Gradient-based methods (Fig. 1.9) typically use gradients of the outputs from later layers with respect to the input features. These techniques are

less computationally expensive compared to perturbation-based techniques because only a single backward pass is required. Perturbation techniques, on the other hand, require a separate forward pass for each perturbed image, increasing computational costs with each inference. However, gradient-based saliency maps can generate more noise.

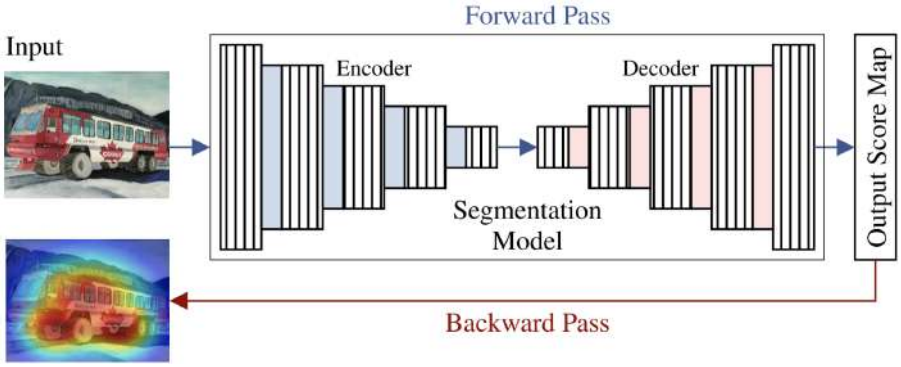


Figure 1.9: A framework for gradient-based methods.

In [212], Seg-Grad-CAM is proposed as the extension of Grad-CAM [188]. It is one of the best known explainability techniques in image segmentation. Just like in the case of regular Grad-CAM, the generated saliency is based on the weighted sum of the selected feature maps. Its application is demonstrated on a U-Net model, trained on the Cityscapes [53] dataset. In [42], the same method is applied for automatic rock joint trace mapping. The original Grad-CAM technique for classification, together with simple gradients, passes the previously discussed sanity checks, evaluating the reliability of XAI techniques. In [96], Seg-XRes-CAM is introduced. The authors criticize Seg-Grad-CAM [212] for not utilizing spatial information when generating saliency maps for a region of the segmentation map. The proposed approach draws inspiration from HiResCAM [69], a modification of the original Grad-CAM [188]. Subsequently, [88] adapts five CAM-based XAI methods from classification to the segmentation of high-resolution satellite images. Among the proposed extensions are Seg-Grad-CAM++, Seg-XGrad-CAM, Seg-Score-CAM, and Seg-Eigen-CAM. Just like in [B.2], Ablation-CAM, a gradient-free method, is also extended for segmentation. Besides using the drop in segmentation score to measure their methods' performance, the authors of [88] also propose an entropy-based XAI evaluation metric.

The implemented methods are tested on a WHU [115] building dataset. In [183], an interpretability and visualization toolbox is proposed for classification and segmentation networks. It includes several XAI extensions specifically for image segmentation. Among them are Guided Grad-CAM and segmented score mapping, extended from [118].

Architecture-Based Methods

This subgroup of methods introduces additional architectural changes (Fig. 1.10) that aim to make the models more interpretable. Instead of relying on post-hoc techniques that are added on top of the already trained models, these methods are typically employed as part of the training process. This class of XAI methods is sometimes described as interpretable by design, inherently interpretable, or interpretability as part of the architecture. In this case, interpretability is inherently linked to the specific model architecture or design, and is not easily transferable to other architectures.

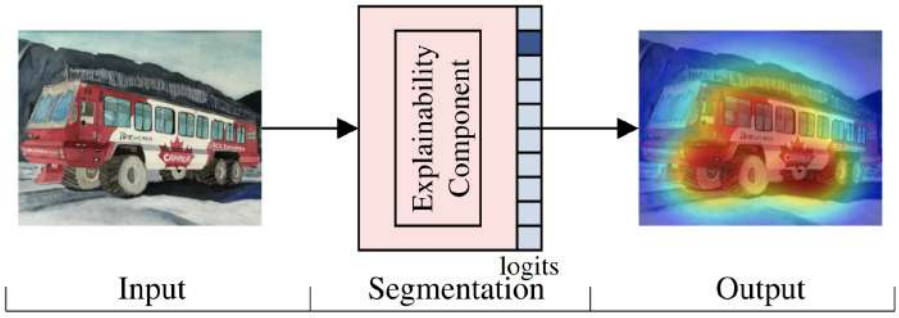


Figure 1.10: A framework for architecture-based methods.

One such example is the chimeric U-Net with an invertible decoder [184]. This approach introduces architectural constraints for the sake of explainability. The authors claim that it can achieve both local and global explainability. In [146], both supervised and unsupervised techniques of Semantic Bottlenecks (SB) are introduced for better inspectability of intermediate layers. This approach is proposed as an addition to the pre-trained networks. Unsupervised SBs are identified as offering greater inspectability compared to their supervised counterparts. The experiments are primarily performed on street scene segmentation images from the Cityscapes dataset. The results are also compared using two other datasets: Broden [24] and Cityscapes-Parts, a derivative of

Cityscapes. In [181], a framework for symbolic semantic segmentation is proposed. This work is at the intersection of image segmentation and emergent language models. The authors apply their research to medical images, specifically brain tumor scans. An Emergent Language model with a Sender and a Receiver is used for interpretable segmentation. The Sender is an agent responsible for generating a symbolic sentence based on information from the higher model layer, while the Receiver cogenerates the segmentation mask after receiving symbolic sentences. The Symbolic U-Net is trained on the Cancer Imaging Archive (TCGA) dataset[3] and used for providing inputs to the Sender network.

1.3.2. Metrics

XAI techniques are used in addition to standard evaluation metrics due to their limitations. However, to evaluate the performance of these techniques, they also need to be measured. Evaluations can be categorized into qualitative and quantitative assessments. Qualitative evaluation commonly refers to user-based evaluation and, based on the surveyed papers (Table 1.1 and Table 1.2), is the more prevalent of the two. To quantify subjective user results, various questionnaires have been proposed [105], such as the explanation goodness checklist, explanation satisfaction scale, trust scales, and ease of understanding when comparing different explainability techniques [92]. These methods still require polling multiple subjects, although, when surveying experts, in practice their number is limited to 2-5 [92]. This way, quantification still takes place, but it is based on subject-dependent evaluation. Since questionnaire studies require additional resources, most of the papers using qualitative evaluation only provide visual comparisons between different XAI techniques, leaving qualitative evaluation to the reader's eye.

Quantitative evaluation does not involve human subjects and can be more easily applied when comparing different interpretability methods. Infidelity and sensitivity [224] are the only two metrics that, as of 2024, are implemented in the Captum [130] interpretability library for PyTorch. Deletion and insertion metrics [169] are another type of quantitative evaluation, based on measuring the area under the curve

[3]https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=5309188

(AUC), generated after gradually deleting or inserting the most important pixels in the input space. A wider use of quantitative metrics could allow for a more standardized comparison across different models and explainability techniques, eliminating the element of human subjectivity. They could also be valuable when human studies are impractical due to resource constraints. This scalability is particularly relevant in industrial and research settings where numerous models need to be evaluated quickly and consistently. However, for some XAI methods, such as counterfactual explanations, it might be difficult to evaluate the usefulness of the explanation quantitatively. In the case of counterfactual explanations, it is possible to measure whether the generated images are realistic and how closely they resemble the query images, but for a more thorough evaluation of the explanation itself, user studies [229] might be required.

In [50], a psychophysics study (N = 1,150) is conducted to evaluate the performance of six explainable attribution methods on different neural network architectures. Shortcomings in the methods are detected when using them to explain failure cases. Comparative quantitative rankings of different saliency techniques can also be inaccurate. In [206], inspired by [5], sanity checks for saliency maps are investigated. The authors perform checks for inter-rater reliability, inter-method reliability, and internal consistency, and determine that the current saliency metrics are unreliable. It is observed that these metrics exhibit high variance and are sensitive to implementation details.

1.3.3. Applications

This subsection presents concrete XAI applications in medical and industrial domains, along with other use cases, primarily focusing on industry-related monitoring domains, such as remote sensing, environmental observation, and biometrics. Additionally, the potential uses of XAI for self-supervised image segmentation are reviewed.

Medical Applications

Most applications in explainable image segmentation have been investigated in the medical domain, using datasets from various medical fields (Table 1.1), ranging from cardiology to oncology. Proposed XAI solutions and applications are employed for diagnosing, monitoring,

and other clinical tasks. In some cases, there might be unavoidable overlaps between medical fields. For instance, overlaps occur at the intersection of oncology and histopathology when discussing microscopic tumor images, or between oncology and dermatology when considering melanoma [201]. Such overlaps can also arise from using multiple datasets, each associated with a different medical field. In these instances, the relevant details are specified in the method description.

Dermatology

Dermatology-centered XAI applications [201, 217] focus on skin lesions. Specifically, [201] discusses applications for interpreting melanoma diagnosis results. The proposed pipeline utilizes both classification and segmentation networks. Grad-CAM is employed to generate explainable heatmaps for the classifier, which are then used as inputs in the U-Net network. These heatmaps assist in generating indicator biomarker localization maps. The proposed approach can be used in self-supervised learning. Experiments are performed on the ISIC 2018 [48] and ISIC 2019 [49, 51, 208] datasets. In [217], a CAM-based explainability metric is proposed and incorporated into the loss function. This metric quantifies the difference between the CAM output and the segmentation ground truth for the targeted class. Both segmentation and explanation losses are considered during the model's training phase. The use of CAM with learnable weights enables a balance between segmentation performance and explainability. The proposed method belongs to the self-explainable XAI category. Similar to [201], the U-Net network is used. The experiments are conducted on the ISIC2018 [48] dataset. In [94], a comprehensive attention-based CNN is proposed for better interpretability in dermoscopic and fetal MRI images. This approach uses multiple attentions, combining the information about spatial regions, feature channels, and scales. The experiments are performed on ISIC 2018 [48] and a private fetal MRI dataset.

Forensic Medicine

The applications of explainable segmentation in forensic medicine are limited to iris segmentation. This can be more narrowly referred to as forensic ophthalmology. In [132], the investigation focuses on forensic postmortem iris segmentation. The authors apply a classical technique of Class Activation Mapping (CAM) [237]. The experiments are performed

on a private test dataset and publicly available post-mortem iris datasets collected by [207].

Table 1.1: Explainable image segmentation in medicine

Field	Imaging modality	Objects of interest	Datasets	Metric	Year	Ref.
G	IMG*	Colorectal polyps	EndoScene [210]	▷	2018	[219]
O	CT	Liver tumors	LiTS [28]	▶	2019	[55]
C	CMRI	Ventricular volumes	SUN09 [172], AC17 [26]	▷	2020	[200]
O	MRI	Brain tumors	TCGA	▷	2020	[181]
D	MRI/IMG	Skin lesions, multi-organ (incl. the fetal brain and the placenta)	ISIC2018 [48] and a private fetal MRI dataset	▷	2020	[94]
P	CT	Pancreatic region	Medical segmentation decathlon	▶	2021	[129]
C	CMRI	Ventricles, myocardium	Cardiac MRI dataset [26]	▷	2021	[111]
G	IMG	Polyps, med. instruments	Kvasir-SEG [113], Kvasir-Instrument [114]	▷	2021	[8]
O	MRI	Brain tumors	BraTS2018 [155]	▶	2021	[180]
FM	NIR	Iris	Private test dataset, post-mortem iris datasets, collected by [207]	▷	2022	[132]
V	CT/MRI/IMG	Skin lesions, abdomen multi-organ, brain tumors	HAM10000 [208], CHAOS 2019 [123], BraTS 2020 [155]	▷	2022	[121]
V	MRI	Brain tumors, human knees	BraTS 2017 [155], OAI ZIB [10]	▷	2022	[184]
O	MRI	Brain tumors	BraTS 2019 [155], BraTS 2021 [155]	▷	2022	[228]
N	MRA	Brain vessels	Private	▶	2022	[35]
H	IMG	Liver	Simulated dataset (Test Set 4) [22]	▷	2022	[23]
O	US/MG	Breast tumors	Private LE/DES datasets, and BUSI [9]	▷	2023	[216]
G	CT/IMG	Colorectal polyps, lung cancer	EndoScene [210], LIDC-IDRI [15]	▶	2023	[41]
O	CT/MRI	Prostate cancer	3D pelvis dataset [68]	▷	2023	[58]
G	CT	Abdominal organs	Synapse multi-organ CT dataset [136]	▷	2023	[96]
O	BUS	Breast tumors	BUSI [9], BUSIS [221], HMSS [84]	▷	2023	[120]
O	CT/PET	Non-small cell lung cancer, whole-body	NSCLC, AutoPET [83]	▶	2023	[117]
D	IMG	Skin lesions	ISIC2018 [48]	▷	2023	[217]
D	IMG	Melanoma	ISIC 2018 [48], ISIC 2019 [208]	▷	2023	[201]
O	MRI/IMG	Prostate tumors, optic disc and cup	Prostate** and fundus*** datasets	▷	2023	[231]
O	X-ray	Breast tumors	INbreast [159]	▶	2023	[74]
O	WSI	Head and neck tumors	Private	▷	2023	[67]
A	IR	Feet	ThermalFeet	▷	2023	[7]
V	CT/MRI/IMG	Brain tumors	BraTS 2018 [155], BraTS 2019 [155], BraTS 2020 [155], ISIC 2017	▷	2023	[99]
P	CT	Pancreas	Pancreas segmentation dataset [14]	▷	2023	[165]
Op	OCT	Retinal layers, glaucoma, diabetic macular edema	NR206, glaucoma dataset [140], DME dataset [43]	▷	2023	[100]
V	CT/MRI	Prostate, left ventricle, right ventricle, myocardium	NCI-ISBI 2013 [30], I2CVB [139], PROMISE12 [142]; MSCMR [240], EMIDEC [133], ACDC [26], MMWHS [241], CASDC 2013 [128]	▷	2023	[82]
Op	OCT	Retinal layers, glaucoma, diabetic macular edema	Vis-105H, glaucoma dataset [140], DME dataset [43]	▷	2024	[101]
V	CMRI/CT	Left atrium, thoracic organs	Atrium dataset [14], SegTHOR [135]	▷	2024	[134]

A: Anesthesiology, C: Cardiology, D: Dermatology, FM: Forensic Medicine, G: Gastroenterology, N: Neurology,
O: Oncology, Op: Ophthalmology, P: Pancreatology, and V: Various
* IMG: general-purpose digital image formats, such as JPEG
** Prostate datasets: RUNMC [30], BMC [30], HCRUDB [139], UCL [142], BIDMC [142], and HK [142]
*** Fundus datasets: DRISHTI-GS [195], RIM-ONE-r3 [81], and REFUGE [167]
▷: Qualitative XAI evaluation
▶: Quantitative XAI evaluation

Gastroenterology

XAI applications for endoscopic image segmentation primarily focus on polyps. In [219], the guided backpropagation [198] technique is extended to the semantic segmentation of colorectal polyps. Uncertainty in input feature importance is estimated, with higher uncertainty observed in inaccurate predictions. Uncertainty maps are generated using the Monte Carlo dropout method. The proposed solution is evaluated on the EndoScene [210] dataset. In [8], Layer-wise Relevance Propagation (LRP), a propagation-based explainability method, is applied to the endoscopic image segmentation of gastrointestinal polyps and medical instruments. LRP is specifically applied to the generator component within a generative adversarial network. The generated relevance maps are then qualitatively evaluated. The segmentation models are trained on the Kvasir-SEG [113] and Kvasir-Instrument [114] datasets.

Hepatology

In [23], two gradient-based post-hoc explanations, Grad-CAM and Grad-CAM++, are investigated for cross explanation of two DL models, U-Net and the Siamese/Stereo matching network, based on [22]. The experiments are performed on laparoscopic simulated stereo images [22], with a focus on liver segmentation.

Oncology

Most of the explainable medical AI applications in image segmentation are in oncology.

Liver:

A DeepDream-inspired method is proposed in [55] for the segmentation of liver tumors in CT scans, specifically focusing on binary segmentation. The study seeks to understand how human-understandable features influence the segmentation output and defines the network's sensitivity and robustness to these high-level features. High sensitivity indicates the importance of such features, while high network robustness shows its indifference to them. Radiomic features are also analyzed. The experiments are performed on the Liver Tumor Segmentation (LiTS) challenge[4] dataset [28]. Semantic segmentation in liver CT images is further investigated in [157], where the segmentation output is corrected

[4]https://competitions.codalab.org/competitions/17094

based on XAI. This approach is categorized as a global surrogate and is model-agnostic. However, its primary purpose is not interpretability but rather the improvement in the initial segmentation by using additional boundary validation and patch segmentation models. The authors of [64] investigate the segmentation of malignant melanoma lesions in 18-fluorodeoxyglucose (^{18}F-FDG) PET/CT modalities, focusing on metastasized tumors. The claim to interpretability is based on the visualization of the model's intermediate outcomes. The overall pipeline involves both segmentation and detection. Volumes of interest (VOI) are visualized for the liver as well as PET-positive regions classified as physiological uptake. This additional information is provided together with the final segmentation masks.

Brain:

An interpretable SUNet [181] architecture is proposed for the segmentation of brain tumors using the TCGA dataset. Experimental results and statistical analysis indicate that symbolic sentences can be associated with clinically relevant information, including tissue type, object localization, morphology, tumor histology and genomics data. In [180], 3D visual explanations are investigated for brain tumor segmentation models, using the quantitative deletion curve metric to compare the results with Grad-CAM and Guided Backpropagation [198] techniques. In [121], a region-guided attention mechanism is used for the explainability of dermoscopic, multi-organ abdomen CT, and brain tumor MRI images. The experiments are performed on HAM10000 [208], CHAOS 2019 [123], and BraTS 2020 [155] datasets. Another architecture-based solution is proposed in [184], where the U-Net architecture is modified and applied to two MRI datasets: BraTS 2017 [155] and OAI ZIB [10], respectively focusing on brain tumors and human knees. In [60], Grad-CAM results are compared to brain tumor segmentation results. The overall pipeline includes both classification and segmentation networks, where DenseNet is used for classification and Grad-CAM-based heatmaps are generated for different layers. However, Grad-CAM is not specifically tailored for segmentation but rather used as an explainable classification tool to evaluate segmentation results. In [228], a NeuroXAI framework is introduced, combining seven backpropagation-based explainability techniques, each suitable for both explainable classification and segmentation. Gliomas and their subregions are investigated using 2D and 3D

explainable sensitivity maps. A ProtoSeg method is proposed in [99] for interpreting the features of U-Net, presenting a segmentation ability score based on the Dice coefficient between the feature segmentation map and the ground truth. Experiments are performed on five medical datasets, including BraTS for brain tumors, each focusing on different medical fields or affected organs.

Pelvis:

In [58], a Generative Adversarial Segmentation Evolution (GASE) model is proposed for a multiclass 3D pelvis dataset [68]. The approach is based on adversarial training. Style-CAM is used to learn an explorable manifold. The interpretability part allows visualizing the manifold of learned features, which could be used to explain the training process (i.e. what features are seen by the discriminator during training).

Breast Cancer:

Oncological XAI applications for the segmentation of breast tumors are investigated in [74, 120, 216]. In [216], a multitask network is proposed for both breast cancer classification and segmentation. Its interpretations are based on contribution score maps, which are generated by the information bottleneck. Three datasets are used, each focusing on a different imaging modality. In [120], the SHAP explainability method is applied to the task of breast cancer detection and segmentation. The experiments are performed on the BUSI [9], BUSIS [221], and HMSS [84] datasets. In [74], explainability for mammogram tumor segmentation is investigated with the application of Grad-CAM and occlusion sensitivity, in both cases using Matlab implementations, and activation visualization. Their quantitative evaluation is based on image entropy, which gives additional information about the XAI method's complexity. Pixel-flipping techniques, which are directly related to deletion curves, are also employed. The experiments are performed on INbreast [159] dataset of X-ray images.

Other:

In [231], Importance Activation Mapping (IAM) is employed as an explainable visualization technique in continual learning. The generated heatmap shows which regions in the input space are activated by model parameters with high-importance weights, associated with the model's

memory. This approach is evaluated for the segmentation of prostate cancer. It also has applications in ophthalmology, specifically for segmenting the optic cup and disc. In [67], two CAM-based XAI techniques, Seg-Grad-CAM and High-Resolution CAM (HR-CAM), are applied to histopathological images of head and neck cancer. The explanations generated by both techniques appear to rely on the same features identified by professional pathologists. In [54], a solution based on Cartesian Genetic Programming is used to generate transparent and interpretable image processing pipelines. This method is applied to biomedical image processing, ranging from tissue histopathology to high-resolution microscopy images, and can be characterized as a few-shot learning approach. In [122], a classification-based version of Grad-CAM is used to enhance a U-Net-based segmentation network. The experiments are performed on the 3D-IRCADb-01 [44] dataset, comprised of 3D CT scans of venous phase CT patients. An Xception network generates 2D saliency maps for classification, which are then passed to the U-Net network together with the corresponding input images. This prior information enables more accurate segmentation. In [170], a framework for explainable classification and segmentation is presented. For segmentation, it relies on a feature hierarchy. The experiments are performed on the skin cancer dataset. The Factorizer architecture, introduced in [18], is based on nonnegative matrix factorization (NMF) components, which are argued to be more semantically meaningful compared to CNNs and Transformers. The proposed approach is categorized under architecture-based interpretability methods. The models are implemented for brain tumor and ischemic stroke lesion segmentation datasets. In [35], a framework for explainable semantic segmentation is presented, extending several classification techniques to segmentation. These methods are also applied to 3D models. Infidelity and sensitivity metrics are used, and the experiments are performed on vessel segmentation in human brain images using Time-of-Flight Magnetic Resonance Angiogram. The experimental data [154] is not publicly available. In [117], a new interpretation method is proposed for multi-modal segmentation of tumors in PET and CT scans. It introduces a novel loss function to facilitate the feature fusion process. The experiments are performed on two datasets: a private non-small cell lung cancer (NSCLC) dataset and AutoPET [83], a whole-body PET/CT dataset from the MICCAI 2022 challenge.

Ophthalmology

XAI is also employed in the segmentation of ophthalmological images. Optic disc and cup segmentation is explored in the setting of continual learning [231], where it is investigated in multi-site fundus datasets. Importance Activation Mapping is used to visualize the memorized content, facilitating an explanation of the model's memory. The focus is on reducing the model's forgetting. In [100], Seg-Grad-CAM is applied to ophthalmology for segmenting retinal layer boundaries. The study provides an entropy-based uncertainty visualization of segmentation probabilities. This offers more information about which retinal layers and regions exhibit higher uncertainty and allows for focusing on problematic areas. It is observed that higher uncertainty is associated with segmentation errors once it reaches a certain threshold. The experiments are performed on the NR206[5] dataset.

Pancreatology

In [129], an interpretable image segmentation approach is proposed for pancreas segmentation in CT scans. The method is also compared to Grad-CAM and occlusion sensitivity, demonstrating its superior inference time. This method identifies regions in the input images where noise can be applied without significantly affecting the model's performance. It relies on noisy image occlusion and can be classified as a perturbation-based technique. To directly parameterize the noise mask for each pixel without harming the model's performance, an additional small interpretability model is trained. Both interpretability and utility models are based on U-Net. Pixels that can be significantly perturbed without changing the model's performance are considered less important. Essentially, the proposed method involves training noise distributions. This approach allows training dynamic noise maps for individual images, differing from the typical static systematic occlusion. Experiments are performed on a pancreas dataset [193]. In [165], a smoothing loss is introduced to guide interpretability learning. The authors observe that the explanations produced by U-Noise are less continuous. Assuming that important pixels are likely to be spatially close, the proposed smoothing objective considers the correlation between pixels during optimization. The resulting explanations are compared to those generated

[5]https://github.com/Medical-Image-Analysis/Retinal-layer-segmentation

by Grad-CAM and U-Noise. Experiments are performed on a pancreas segmentation dataset [14] from the medical segmentation decathlon.

Urology

In [82], a Bayesian approach is proposed to address the problem of interpreting domain-invariant features. The experiments are performed for prostate and cardiac segmentation tasks. The experiments are performed on T2 prostate MRI images from NCI-ISBI 2013 [30], I2CVB [139], and PROMISE12 [142]. For cardiac segmentation, MSCMR [240], EMIDEC [133], ACDC [26], MMWHS [241], and CASDC 2013 [128] datasets are used.

Anesthesiology

In [7], an interpretable approach is investigated for regional neuraxial analgesia monitoring. The experiments focus on thermal foot images for patients who have received epidural anesthesia. The proposed method is based on Convolutional Random Fourier Features (CRFF) and layer-wise weighted CAM. CRFF and CAM-based techniques are investigated in three segmentation models: U-Net, FCN, and Res-U-Net. CRRF gradient layers are added at skip connections. The initial results indicate that the integration of CRRF gradient layers allows better differentiation between background and foreground classes. The experiments are performed on the ThermalFeet[6] dataset of infrared images.

Industry-Related Applications

Various industrial and industry-related activities require precise segmentation. These activities might range from precise manufacturing and processing [A.1] to structural health monitoring in infrastructure, particularly in evaluating damage [78]. Industrial applications of XAI are also related to risk management. For example, the AI TriSM (Trust, Risk and Security Management) framework [3, 89], has been adopted by several industrial organizations to enhance their risk management practices. This high-level framework includes both explainability and model monitoring as key components. Risk management is of particular importance in sensitive dynamic operational environments, where AI models need to be continuously assessed for performance and reliability.

[6] https://gcpds-image-segmentation.readthedocs.io/en/latest/notebooks/02-datasets.html

This subsection discusses both industrial processes and indirectly related tasks, such as environmental monitoring and remote sensing, which can have potential in industrial applications in a more narrow sense. Industry-related explainable segmentation solutions are divided into four categories: remote sensing, monitoring, scene understanding, and other more general applications.

Table 1.2: Explainable image segmentation in industry

Category	Domain	Datasets	Metric	Year	Ref.
Remote sensing	Building detection	IAIL [149]	▷	2019	[110]
Scene understanding	Autonomous driving	SYNTHIA [242], A2D2 [86]	▷	2021	[2]
Scene understanding	Pedestrian environments	PASCAL VOC 2012 [72], ADE20K [238], Cityscapes [53]	NA*	2021	[233]
Scene understanding	Autonomous driving	KITTI [80]	▷	2022	[150]
Environmental monitoring	Flood detection	Worldfloods [153]	▷	2022	[235]
Scene understanding/Biometrics	Driving scenes/Face recognition	BDD100k [225], CelebAMask-HQ [138], CelebA [144]	▷	2022	[109]
Monitoring/Scene understanding	Drones/Food processing	ICG drone dataset, private dataset	▷	2023	[A.1]
Monitoring/General applications	Food processing	COCO [141], private dataset	▷	2023	[B.2]
Biometrics	Facial emotions	Face recognition dataset [209]	▷	2023	[215]
Monitoring	Cracks in infrastructure	CrackInfra [143]	▷	2023	[143]
General applications	Common objects	COCO [141]	►	2023	[B.1]
Scene understanding/General applications	Street scenes/Common objects	Pascal VOC 2012 [72], Cityscapes [53]	▷	2023	[178]
Scene understanding	Driving scenes	BDD100k [225], BDD-OIA [223]	▷	2023	[229]
Scene understanding/General applications	Street scenes/Common objects	Cityscapes [53], Pascal VOC [72], COCO [141]	►	2023	[70]
General applications	Common objects	COCO [141]	▷	2023	[96]
General applications	Common objects	Pascal VOC [72]	►	2023	[41]
Scene understanding/Remote sensing	Street scenes/Building detection	Cityscapes [53], WHU [115]	►	2023	[191]

*The application focuses on introducing explainability to segmentation evaluation, rather than evaluating explainability techniques.
▷: Qualitative XAI evaluation
►: Quantitative XAI evaluation

Remote Sensing

One of the first applications of interpretable image segmentation is in remote sensing. In [110], the U-Net model is applied for building detection. The proposed method operates at the intersection of interpretability, representation learning, and interactive visualization, and is designed to explain U-Net's functionality. It employs Principal Component Analysis (PCA) on the activations in the bottleneck layer. PCA is the preferred method because it preserves the largest variance in the data. In the case of 3D visualizations, the first three components could be used. Fol-

lowing PCA, the new representations are clustered using k-means and the DBSCAN algorithm. This approach allows for the visualization of learned latent representations for all samples through an Intersection over Union (IoU)-based heatmap, enabling users to identify qualitatively different regions. The experiments are performed on the Inria Aerial Image Labeling (IAIL) [149] dataset. This technique can be applied to detect and evaluate damages in industrial disasters or humanitarian crises, extending beyond mere infrastructure and product monitoring in industry. Another remote sensing application [191], specifically focusing on high-resolution satellite images, employs a gradient-free Sobol method [75] and a U-Noise model [129]. The proposed method is also compared to the Seg-Grad-CAM++ classification extension.

Monitoring

This subsection reviews relevant papers that apply explainable segmentation-based monitoring in proximate environments. In [A.1], simple gradient [192] saliency maps and SmoothGrad-based [196] saliencies are implemented for semantic segmentation models to investigate the adversarial attack setting. The experiments are performed on two industry-related cyber-physical system datasets. A private dataset from CTI FoodTech, a manufacturer of fruit-pitting machines, is used. In [B.2], the same private dataset is used for experiments with a gradient-free XAI technique based on perturbations of intermediate activation maps.

In [143], the focus is on crack segmentation in critical infrastructures, such as tunnels and pavements. The U-Net model is used together with Grad-CAM, which is applied at the bottleneck, as in [212]. They investigate both simple and complex crack patterns as well as different backgrounds. Two other papers [78, 187] also investigate the segmentation of different crack types. However, the proposed XAI techniques are implemented in classification models, and used for weakly supervised segmentation. These techniques are discussed in the subsequent section. In [215], an interpretable Bayesian network is used for facial micro-expression recognition. The authors prefer these networks for segmentation over DL models, primarily because of their superior causal interpretability when dealing with uncertain information. This can make them better interpretability candidates when uncertain causal inference is involved. The experiments are performed on the database [209] of face images.

The trade-off between the noisiness of the explanation and its computation time is particularly important in monitoring. Gradient-based methods are suitable for monitoring environments that require low latency and rapid decision-making. However, if the monitoring task allows for more time, slower non-gradient methods, such as perturbation-based explanations, might be a better choice.

Scene Understanding

Scene understanding is an important area in applications for autonomous vehicles, monitoring of pedestrians and ambient objects, and surveillance. Precise real-time segmentation of road signs and obstructions is of particular importance. Explainable segmentation can be seen as part of explainable autonomous driving systems [2], which investigate events, environments, and engine operations. An explainable variational autoencoder (VAE) model is proposed in [2], focusing on neuron activations with the use of attention mapping. For the experiments, the SYNTHIA [242] and A2D2 [86] datasets are used. The results are analyzed both qualitatively and quantitatively, using the average area under the receiver operating characteristic curve (AUC-ROC) index. In [150], XAI techniques are employed to investigate pixel-wise road detection for autonomous vehicles. The experiments are performed on different segmentation models, using the KITTI [80] road dataset. The problem is formulated as a binary segmentation task, where the classes are limited to the road and its surroundings. Grad-CAM and saliency maps are used to generate explanations. Unmanned aerial vehicles can also fall under the category of autonomous driving systems. In [A.1], gradient-based XAI techniques are applied to semantic drone dataset[7] from Graz University of Technology.

Automated semantic understanding of pedestrian environments is investigated in [233]. Here the focus is not on a particular XAI technique, but on introducing some level of explainability to segmentation evaluation. The paper argues that popular pixel-wise segmentation metrics, such as IoU or Dice coefficient, do not sufficiently take into account region-based over- and under-segmentation. Here over-segmentation refers to those cases where the relevant ground-truth region is segmented into a lower number of regions than the predicted mask. For

[7] http://dronedataset.icg.tugraz.at/

instance, where there is only one bus in the segmented ground-truth, but the model segments it into three disjoint segments. In the case of under-segmentation, the opposite is true. Pixel-wise metrics do not accurately represent these differences in disjoint and joint regions as long as a large enough number of similar pixels is segmented in both the ground-truth image, and the corresponding prediction. The use of region-wise measures is proposed as a better way to explain the source of error in segmentation. The experiments are performed on PASCAL VOC 2012 [72], ADE20K [238], and Cityscapes [53]. In [63], the focus is on automatic semantic segmentation for sediment core analysis. To interpret the results, higher segmentation error regions and model prediction confidence are visualized. Here, the model confidence is defined as prediction probability, and the model error calculation is based on the normalized categorical cross-entropy.

The authors of [70] propose the Concept Relevance Propagation-based approach L-CRP as an extension of CRP [4]. By utilizing concept-based explanations, the study seeks to gain insights into both global and local aspects of explainability. The proposed approach seeks to understand the contribution of latent concepts to particular detections by identifying them, finding them in the input space, and evaluating their effect on relevance. Context scores are computed for different concepts. The experiments are performed on Cityscapes [53], Pascal VOC [72], and COCO [141] datasets.

General Applications

Some of the XAI-related experiments focus on more general datasets, which are typically used for evaluating the performance of segmentation models and do not fall into remote sensing, monitoring, or scene understanding categories. This includes [B.1] and [B.2], which will be briefly covered here, with concrete implementation details provided in Chapters 2 and 3. The COCO [141] dataset has been used as a benchmark in [B.1] and [B.2]. The dataset is composed of 21 classes of everyday objects, including several types of vehicles. Both [B.1] and [B.2] apply perturbation-based gradient-free methods. Input perturbations are used in [B.1], while feature map perturbations in pre-selected intermediate layers are used in [B.2].

The Tendency-and-Assignment Explainer (TAX) framework is introduced in [41]. It aims to explain what contributes to the segmenta-

tion output and the reasoning behind it (i.e. the why question). For this, a multi-annotator scenario is considered. The learned annotator-dependent prototype bank indicates the segmentation tendency, with a particular focus on uncertain regions. The experimental results on the Pascal VOC [72] dataset demonstrate that TAX predicts oversegmentation consistent with the annotator tendencies.

XAI Applications in Self-Supervised and Weakly Supervised Segmentation

Manual image labeling is an expensive operation, especially when pixel-wise labeling is involved. It requires significant time and financial resources, and depending on the dataset being annotated, may also require particularly narrow expertise. With this in mind, it has been suggested that XAI techniques could be employed for automated labeling, which could also help reduce some forms of annotation bias.

In [226], a new explainable transformer architecture is proposed for model-inherent interpretability and is investigated for weakly supervised segmentation. An explainable vision transformer is used as a Siamese network, where two branches process input images for the self-supervised learning of interpretable attention maps. For enhanced interpretability, model representations are regularized using an attribute-guided loss function. Higher-layer attention maps are fused and used alongside attribute features. Qualitative segmentation results are compared with Self-supervised Image-specific Prototype Exploration (SIPE) [40] for weakly supervised semantic segmentation. However, the model's limitation is its inability to incorporate attribute-level ground truth labels. Another application for weakly supervised segmentation employs LRP-based classification explanations [187]. These explanations are used to generate pixel-wise binary segmentations, which are then thresholded. The experiments are conducted on two datasets: one for cracks in sewer pipes and another for cracks in magnetic tiles [108].

In [78], surface crack detection and growth monitoring are investigated as part of structural health evaluation in infrastructure. Although no specific technique for explainable segmentation is proposed, explainable classification is used for weakly supervised segmentation, allowing for the quantification of crack severity. Six post-hoc techniques are implemented: InputXGradient, LRP, Integrated Gradients, DeepLift, DeepLiftShap, and GradientShap. Additionally, B-cos networks and the

Neural Network Explainer are employed. In [25], GradCAM is used for semantic segmentation. An additional classification model is used with masked inputs, based on the given class. The classifier is trained on all the masked images across all classes. Explainable classification can also be used to enhance the data efficiency of segmentation models. For instance, in [220], Grad-CAM is employed to extract data-efficient features from the classification model, which are then used for segmentation. The results indicate that this approach generalizes across different segmentation methods.

1.4. Adversarial Attacks

XAI techniques are used in addition to standard evaluation metrics to improve models' reliability and trustworthiness. However, with their increasing adoption, it is also important to evaluate the explainability techniques and their outputs themselves, in order to better understand their limitations and the extent to which they can be trusted. Especially in scenarios where potential XAI weaknesses could be exploited by adversaries through deliberate, targeted manipulation. This section discusses these concerns. First, it provides a brief overview of adversarial attacks, followed by a discussion of literature relevant to both interpretability and adversarial attacks in semantic segmentation. The main results presented in this section have been previously published in [A.1].

DL models are vulnerable to adversarial attacks where subtle modifications to input images that are imperceptible to the human eye can lead to confidently incorrect predictions. Since their introduction in 2013 [203], adversarial perturbations or adversarial noise have been extensively studied in both theoretical and practical contexts.

Models deployed in publicly accessible cyber environments are particularly vulnerable to adversarial techniques, including model exploratory and data poisoning attacks [171]. These attacks can pose significant safety hazards, compromising safety-critical cyber-physical systems (CPS) [168]. The most targeted industrial sectors currently include manufacturing and the supply of electricity, gas, steam, and air conditioning [124]. Mitigating these risks requires a deeper understanding of adversarial attacks, their scope, and their limitations.

For AI systems to be trustworthy and reliable in real-world applications, robustness is needed not only in the model's output but also in its interpretability. In recent years, it has been shown [87] that fragility to perturbations is not just limited to the model's predictions, but that it also extends to their explanations. Consequently, fragile interpretations could be measured in terms of their robustness to perturbations in the input image.

A prevalent branch of interpretability methods concerns itself with instance-based model-agnostic post-hoc explanations, where, within the context of computer vision, the explanation is limited to one image on the already trained model, irrespective of its underlying architecture, and without retraining it again. Some of the best-known examples of this type of approach come from gradient-based methods, where pixel importance maps, also known as saliency maps, are generated using the gradient of the prediction score with respect to the input image.

This dissertation investigates the effects of adversarial attacks on gradient-based saliency map techniques in semantic segmentation. The focus is on targeted attacks where the segmentation output of the original image is moved to a different place, without changing its area. Such attacks are particularly concerning because they are more difficult for human operators, even experts, to detect. Even though the adversarial attack literature tends to focus on vivid examples, such as turning the model's classification of a panda into that of a gibbon [91], similar scenarios would be more easily detectable in an industrial setting. However, the shift of the original segmentation output by a few pixels could frequently go unnoticed while, in practice, still causing substantial damage. The intersection of interpretability and adversarial robustness remains an underexplored area in semantic segmentation.

1.4.1. Interpretability and Adversarial Attacks in Semantic Segmentation

Both [107] and [212] introduced interpretable semantic segmentation methods and tested them on the Cityscapes [53] dataset. In [212], SegGradCam was proposed as an extension of Grad-CAM [188], an interpretability method for classification, to semantic segmentation. In [107], grid saliency was introduced to explain contextual information in semantic segmentation. Its main objective is to detect context biases.

Vanilla Gradient [192], as well as its SmoothGrad [196] extension, was also applied to a synthetic semantic segmentation dataset [107]. There is no research, however, on how these or similar interpretable semantic segmentation methods would work under adversarial attacks.

Similar to interpretability methods, most adversarial approaches focus on classification tasks. This can be explained by their higher prevalence compared to segmentation tasks as well as the lower computational resources needed to arrive at a satisfactory solution. In the field of CPS, [116] proposed an adversarial attack against anomaly detector, but the attack targeted a classification-based system. Nonetheless, several important methods [77, 222] have been proposed for implementing adversarial attacks in dense predictions. In [222], a Dense Adversary Generation (DAG) algorithm was proposed for generating adversarial examples in semantic segmentation and object detection tasks. However, the study does not explore interpretability-related areas. [65] presented a method for attacking different explanation techniques while minimizing changes to the model's output. However, its implications for semantic segmentation have not been investigated, and no attack on interpretability in this domain has been conducted.

When it comes to the intersection of adversarial attacks and interpretability, some of the recent works [65, 103, 232] have shifted the focus from adversarial attacks on a model's output to adversarial attacks on its saliency. In [65], a loss function optimization approach was proposed to manipulate the explanations generated by different XAI methods while keeping the output of the model relatively unchanged. In [103], passive and active attacks targeting an input image's saliency were introduced, demonstrating the transferability of such attacks across different interpretability methods, such as between LRP [20] and Grad-CAM or Vanilla Gradient. In contrast, [232] demonstrated a low transferability of such attacks against interpretability techniques. However, in all cases, the proposed approaches focused on classification tasks.

1.5. Chapter Conclusions

In this chapter, the first comprehensive survey of XAI methods for semantic image segmentation was presented, along with a taxonomy categorizing existing techniques into prototype-based, gradient-based, perturbation-based, counterfactual methods, and architecture-based

methods. The limitations of current methods were critically evaluated, demonstrating the need for more robust and effective solutions. This survey establishes the groundwork for developing new explainability approaches and integrating them with adversarial robustness, which will be explored in subsequent chapters.

Based on the literature review, it can be observed that most of the XAI methods in image segmentation fall into post-hoc and local explanation categories, with most of the conducted experiments only using qualitative evaluation. Furthermore, the review highlighted a clear gap at the intersection of explainable segmentation and adversarial attacks. Perturbation-based and gradient-based approaches appear particularly promising due to their transferability across different model architectures. These methods are also of interest because in certain contexts, such as real-time industrial applications, they might present a trade-off between computational cost and the robustness or noisiness of the generated explanations. For these reasons, they have been identified as suitable candidates for explaining the outputs of semantic segmentation models.

This chapter outlines the methods developed to advance post-hoc explainability in semantic segmentation. It introduces and provides theoretical foundations for three primary techniques: perturbation-based methods in input and activation spaces, and gradient-based approaches. The chapter also discusses XAI-driven model improvements in neural architecture search (NAS) and continual learning (CL). The main results presented in this chapter have been previously published in [A.1], [B.1], [B.2], and [B.3].

2.1. Perturbations in the Input Space

Input perturbation XAI techniques are based on the systematic occlusion of parts of an input image, generating explanations by measuring how these occlusions affect the model's output. For more details and a high-level framework for perturbation-based methods for the input space (Fig. 1.7), refer to Subsection 1.3.1. The images presented in this section are from the COCO dataset [141], which is further described in Section 3.1 when discussing experimental results.

Figure 2.1: Three 30×30 occlusions in the upper-left corner correspond to gray, black, and Gaussian filters.

Following [227], the proposed approach measures the effect of occlusions in terms of either the Dice coefficient or Intersection over Union (IoU) metric. The Dice coefficient, defined as

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|}$$

quantifies the overlap between the predicted segmentation A and the ground truth B. Similarly, the IoU, also known as the Jaccard index, is given by

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|},$$

which measures the ratio of intersection to the union of the predicted and actual segmentations. Just like in the case of classification, the input image is occluded by sliding the equilateral (although the implementation with different occlusion dimensions is possible) uninformative patch over it and measuring its effect on the Dice coefficient for the segmentation class of our choice. Suitable candidates for occlusion patches include but are not limited to gray, black, and Gaussian filters (Fig. 2.1). In the case of blurring the image with Gaussian filters, the standard deviation value was set to 7.

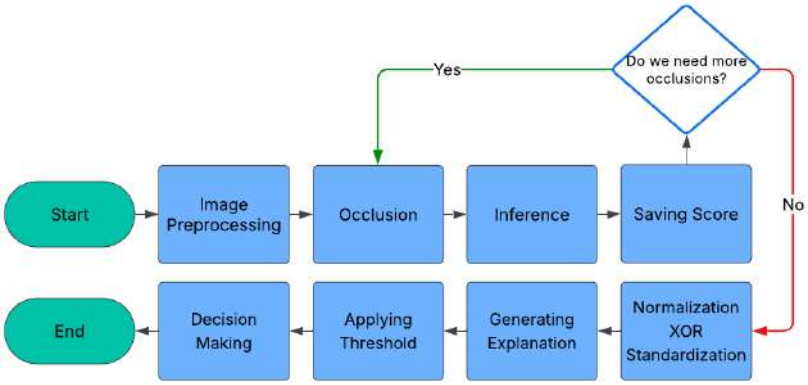


Figure 2.2: The workflow of an occlusion-based approach for interpretable semantic segmentation.

The occlusion process (Fig. 2.2) starts with a regular image preprocessing step that includes normalization and, in the case of limited computational resources, resizing. Then the selected type of occlusion filter is gradually slid along the whole image. A smaller slide size can be used as well but then filters would overlap. The inference is made for each newly occluded image by passing it through a trained neural network. Each time a new segmentation output (Fig. 2.3) is generated and is then used to calculate the evaluation metric's (typically the Dice coefficient or IoU, but a sum of logits for a class of interest can also

be used) score. The score is saved, and the process is repeated until the sliding occlusion filter has fully covered each corresponding image region.

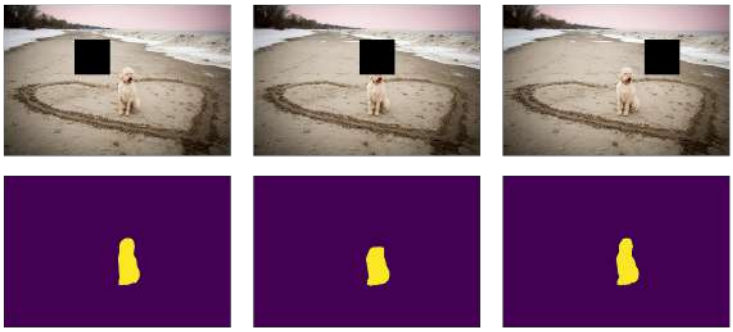


Figure 2.3: Occluded images and their corresponding segmentation output, generated by DeepLabV3.

The collected scores of all occlusions are used in creating a saliency map that shows the importance of each region for the successful segmentation of the class of interest. In the Dice score-based visualizations (Fig. 2.4), the image background was chosen as such class. To generate more color intensities, min-max normalization is employed.

Figure 2.4: DeepLabV3 (first row) and FCN (second row) results, using 294 30×30 occlusions. Min-max normalized maps are in the second column while z-score standardized maps are in the third. Either one of these techniques can be used to generate more color intensities in the final explanation.

A threshold can then be selected for the normalized scores to be visualized. Different thresholds will generate different maps, which

might be useful for the end-user in drawing attention to different regions in the image. Also, focusing on just the most salient regions will help to reduce the noisiness of the initial saliency map. Related experimental results are discussed in Section 3.2.

2.2. Perturbations in the Activation Space

Ablation-CAM was introduced in [62] for image classification tasks. It uses a full deactivation of feature maps within a particular layer to measure its impact on the model's performance with respect to a particular class. Based on that, the importance scores are calculated for each of the feature maps and they are then used as weights in a linear combination of activation maps. In this section, the application of Ablation-CAM is extended for dense predictions (Figs. 2.5 and 2.6). For more details and a high-level framework for perturbation-based methods for the activation space (Fig. 1.8), refer to Subsection 1.3.1. The images presented in this section are from a private industrial dataset from a manufacturer of fruit processing machinery and the COCO dataset [141], both of which are further described in Section 3.1 when discussing experimental results.

To extend ablation-based interpretability to segmentation, the ablation impact on a class of interest is computed in terms of its effect on cumulative logits, obtained by summing up the majority class (*argmax*) logits for each pixel. In the proposed extension, only the logits of c for pixels that were classified as c are accumulated.

Given the RGB image $x \in \mathbb{R}^{N \times M \times 3}$, the logit value for a single pixel x_{ij} for a class of interest c is defined as $l^c(x_{ij})$. Then the sum of logits for c, conditioned on x_{ij} being classified as c, is:

$$L^c(x) = \sum_{i,j} [\hat{c}_{ij} = c] l^c(x_{ij}), \tag{2.1}$$

where $\hat{c}_{ij}$ in the Iverson brackets is the predicted class for the x_{ij} pixel.

These accumulated class logit values are used to compute the importance score for each activation map, like in the original implementation. Following [62], the importance weight w_k^c for each activation unit k is defined as:

$$w_k^c = \frac{L^c(x) - L_k^c(x)}{L^c(x)}, \tag{2.2}$$

where $L_k^c(x)$ is the sum of logits for c after the ablation of activation map A_k.

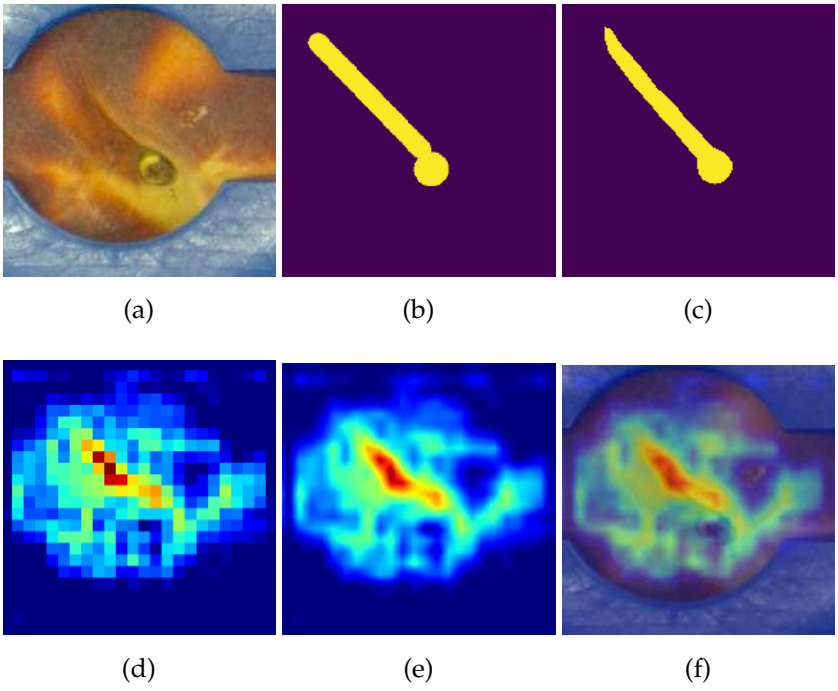
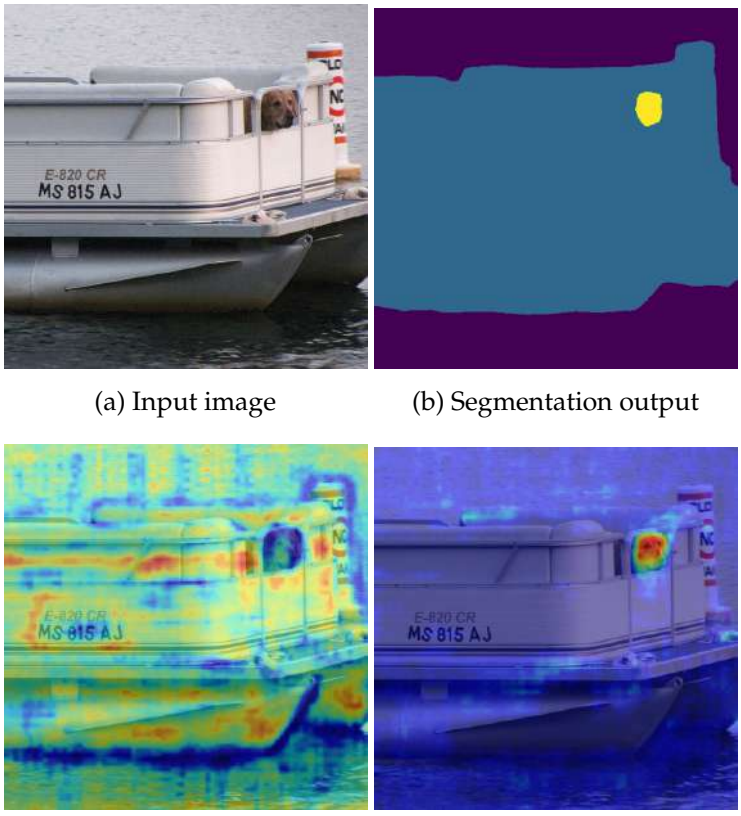


Figure 2.5: Ablation-CAM for semantic segmentation. (a) and (b) show the original input image and its corresponding ground truth (for a more detailed description of the dataset, see Section 3.1); (c) shows the U-Net's predicted segmentation output; (d) shows the output of Ablation-CAM for semantic segmentation, when applied on the last encoder layer; (e) shows resized and smoothed Ablation-CAM output; (f) is the Ablation-CAM output (e) overlayed on the original input image (a).

Then the calculated weights can be used in a linear combination of feature maps. Before the ablation, an encoder layer of interest has to be selected (Fig. 2.7). Within the U-Net architecture, the second convolutional layer of the last encoder block was selected[1]. Then each of its 256 feature maps was separately deactivated by setting its values to 0. The Ablation-CAM extension was also investigated in a more general multi-class setting (Fig. 2.6) with different-sized objects. In this case, the third encoder layer of FCN-ResNet-101 was ablated by separately

[1]Different layers can be selected but the preliminary experiments showed the most promising results with layers close to the U-Net bottleneck.

deactivating 1024 feature maps. Ablation-CAM was also implemented using the Dice score for a class of interest as an importance score instead of a sum of pre-normalized probabilities, but the visual results were better when using the latter.



(a) Input image (b) Segmentation output

(c) The boat class (d) The dog class

Figure 2.6: Ablation-CAM for the multi-class dataset.

Warmer colors indicate greater importance of those regions in segmenting a class of interest. The generated heatmaps (Fig. 2.5 and Fig. 2.8) show that areas with warmer colors tend to correspond to the fruit's cutting line. The 24×24 image was resized to the original pre-processed input size of 192×192 for overlaying onto the input image, as seen in Fig. 2.5 (f). Related experimental results are discussed in Section 3.3.

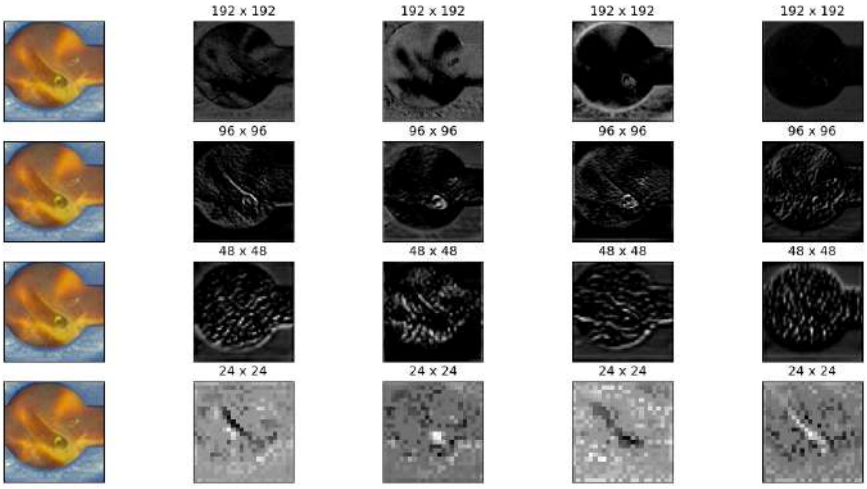


Figure 2.7: Randomly selected feature maps with their corresponding spatial dimensions from the four encoder blocks (one per each row). Feature map occlusions for the background class with a t value of 0.5 can be seen in the last row, corresponding to the last convolutional layer before the U-Net bottleneck.

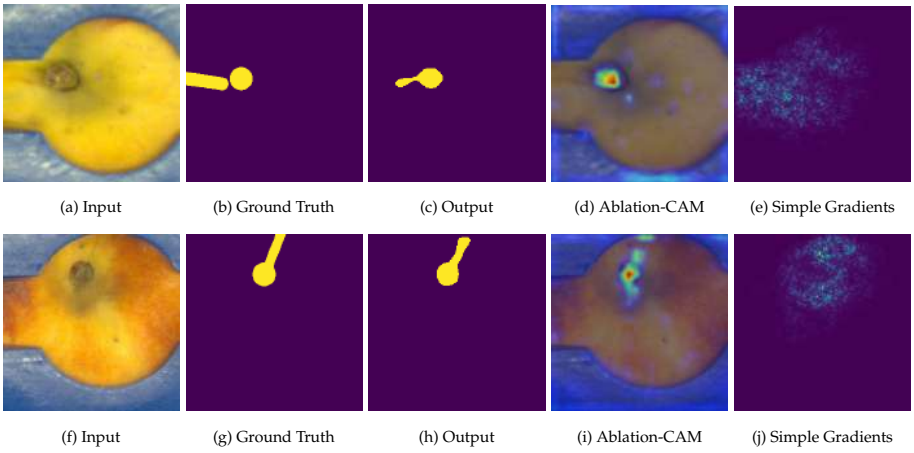


Figure 2.8: Comparison of Ablation-CAM and gradient-based saliency maps for two input images. The first row presents an instance with a worse segmentation performance (c). This can be explained by the corresponding Ablation-CAM map (d), where one can see the network's failure to detect the fruit's suture line.

2.3. Gradient-Based Explanations in Adversarial Contexts

To investigate model outputs, two gradient-based interpretability techniques are extended to semantic segmentation: a simple gradient saliency map, also referred to as Vanilla Gradient, and the same saliency map enhanced with SmoothGrad, which can be applied to any gradient-based interpretability method to reduce noise. For more details and a high-level framework for gradient-based methods (Fig. 1.9), refer to Subsection 1.3.1. The images presented in this section are from a private industrial dataset from a manufacturer of fruit processing machinery and a public semantic drone dataset[2] from Graz University of Technology, both of which are further described in Section 3.1 when discussing experimental results.

2.3.1. Gradient-Based Saliency Maps

Saliency maps in classification

In classification, a simple gradient-based saliency map is calculated by taking the gradient of the predicted score with respect to the input image. Formally, given a set of classes $\{1, 2, ..., C\}$, where C is the number of classes, and given the RGB image $x \in \mathbb{R}^{N \times M \times 3}$, $g(x) = (g_1(x), ..., g_C(x)) \in \mathbb{R}^C$ is defined, where $g_c(x)$ is the prediction score before the Softmax function for class $c = 1, 2, ..., C$ with respect to x. Then the saliency map of class c is $G(x, c) = \frac{\partial g_c(x)}{\partial x}$.

Saliency maps in semantic segmentation

When calculating saliency map values with Vanilla Gradient for semantic segmentation, the logits (not-normalized probability scores before the Softmax function) of the class of interest are summed up to obtain a scalar value. The gradient of that value is then calculated with respect to all input pixels. This way, saliency maps can be generated for each segmentation class of interest. It is possible to calculate the absolute value of each pixel's score if we are indifferent between its positive or negative contribution to the segmentation score of the class of interest.

Formally, given a set of classes $\{1, 2, ..., C\}$, where C is the number of classes, and given the RGB image $x \in \mathbb{R}^{N \times M \times 3}$, the function

[2] http://dronedataset.icg.tugraz.at/

$g(x) = (g_1(x_{ij}), ..., g_C(x_{ij})) \in \mathbb{R}^{N \times M \times C}$ is defined, where $g_c(x_{ij})$ is the prediction score before the Softmax function for class $c = 1, 2, ..., C$ with respect to the pixel x_{ij} of x. The notation $\overline{g}(x)$ is used for the prediction score after applying the Softmax function. The sum of logits for c is then defined as: $g_{c,A}(x) = \sum_{i,j \in A} g_c(x_{ij})$, where A is a set of pixel indices of interest. Then the saliency map of class c is $G_A(x, c) = \frac{\partial g_{c,A}(x)}{\partial x}$.

The sensitivity of the saliency map can be controlled by setting a threshold t to the gradient values of each pixel.

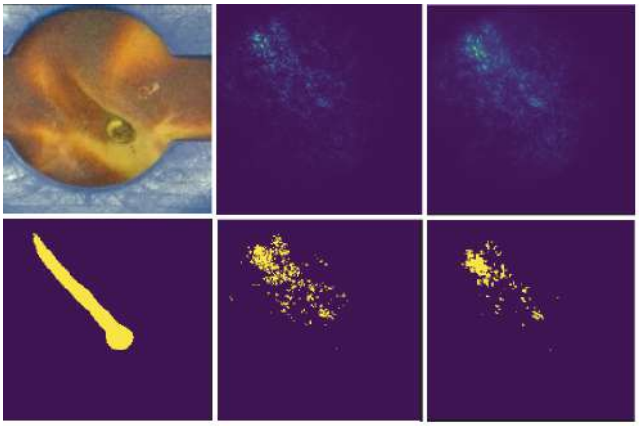
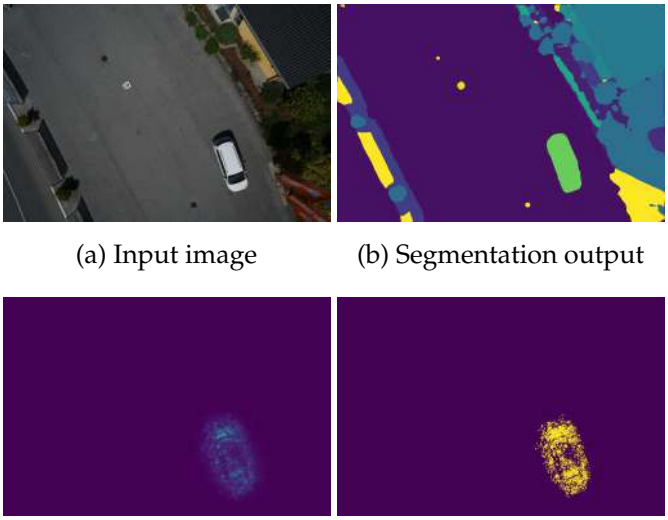


Figure 2.9: The input image and U-Net's segmentation output in the first column. Vanilla Gradient saliencies without any threshold (top) and with t set to 100 (bottom) in the second column. SmoothGrad saliencies without any threshold (top) and with t set to 100 (bottom) in the third column.

To illustrate this, saliency maps for the cutting line in Figure 2.9 were calculated using the threshold of 100. In this way, the corresponding map values are simplified into two values based on the thresholding condition. SmoothGrad saliency was calculated using 50 noise sampling[3] iterations. This value worked well with respect to the computational resources, as larger value would usually not provide clearer visualizations, but would take longer to compute. In all cases, the absolute values were used and saliency map values were calculated with respect to the suture line class.

[3]SmoothGrad reduces the noise in gradient-based saliency maps by adding random noise to the input image during each iteration, calculating the gradient-based explanation for each noisy image, and then averaging these explanations.



(a) Input image (b) Segmentation output

(c) Saliency (d) Thresholded saliency

Figure 2.10: The input image and U-Net's segmentation output in the first row. Vanilla Gradient saliencies for the car class without any threshold (left) and with t set to 100 (right) in the second row.

As can be seen in Figure 2.9 and Figure 2.10, salient values tend to correlate with the direction of the classes of interest which, in this case, are the cutting line and the car. Due to the higher computational cost of SmoothGrad, its application can be omitted when the generated saliency maps are sufficiently clear and not overly noisy (Fig. 2.10).

2.3.2. Adversarial Attacks in Semantic Segmentation

Gradient-based XAI techniques were systematically evaluated in adversarial scenarios to determine their susceptibility to attacks. To better control the perturbation process and prevent excessive distortion in the adversarial examples, an additional loss term was introduced. This term ensures that the applied adversarial noise remains constrained, preserving both the segmentation output and the perceptual similarity of the perturbed image to the original. The formal definition of the loss function is presented below.

Given two images $x, y \in \mathbb{R}^{N \times M \times 3}$ and two classes of interest $c_1, c_2 \in \{1, 2, ..., C\}$, the goal is to attack the saliency map of x, $G_{A_1}(x, c_1)$, to

get the target saliency map of y, $G_{A_2}(y, c_2)$, where A_i is the area of the image classified as c_i, $\forall i = 1, 2$. At the same time, the segmentation output $\overline{g}(x_{adv})$ of the perturbed image x_{adv} should remain similar to the segmentation output $\overline{g}(x)$, and the perturbed image should not change too much from x. The following loss function is proposed [A.1] and defined as:

$$L = \gamma_1 L_{exp} + \gamma_2 L_{out} + \gamma_3 L_{im}, \tag{2.3}$$

where

$L_{exp} = ||G_{A_2}(y, c_2) - G_{A_1}(x_{adv}, c_1)||^2$,
$L_{out} = ||\overline{g}(x) - \overline{g}(x_{adv})||^2$,
$L_{im} = ||x - x_{adv}||^2$,

and where γ_1, γ_2, γ_3 are fixed parameters that control the relative importance of each loss term during optimization. L_{exp} measures the distance between the generated explanation and its adversarial target (i.e., the explanation that we want the attacked model to generate), L_{out} measures the distance between the model's original output and the attacked model's output, and L_{im} measures the distance between the original input image and the attacked input image.

Section 3.4 presents related experimental results, focusing on two types of adversarial attacks on semantic segmentation networks using industry-related datasets. First, the Dense Adversary Generation (DAG) [222] attack is applied to input images and its impact is evaluated on the corresponding saliency maps. Second, adversarial attacks are extended to saliency maps in semantic segmentation. Gradient-based interpretability extensions are also explored. To date, there is no research on interpretable semantic segmentation under adversarial attacks.

2.4. XAI-Driven Model Improvements

Little attention has been paid to the use of XAI in non-explainability-related scenarios, where XAI methods are applied not for interpretability *per se*, but rather for other instrumental reasons, such as improving a model's performance. Such use cases can potentially extend to AI safety, specifically in the case of adversarial attacks, self-supervised learning, NAS, and CL. Most of these areas have never been investigated in the context of interpretable segmentation. This section considers concrete

frameworks for potential uses of interpretable image segmentation for model compression in the case of NAS, and instance-based memory compression in the case of CL.

Based on the literature review, there are no XAI-driven model improvements specifically for NAS or CL in segmentation. In classification, [236] proposes a NAS model based on class activation mapping (CAM). The teacher and student models are incorporated into the evolutionary search. The less complex student model has to generate an explanation map that closely approximates the one generated by the teacher model, as measured by the inverse of the Euclidean distance. In [106], an input saliency-based NAS is introduced as a way of reweighing different data points. However, the proposed solution only focuses on the features in the input space, leaving investigation of the activation space features for further research. This approach is suitable for differentiable NAS methods, but further investigation is needed for non-differentiable methods, such as evolutionary-algorithm-based NAS. Additional modifications or selecting a non-gradient based optimization algorithm would be required.

The underlying assumption is that efficient explainable segmentation techniques can identify those regions in the input space or those feature maps in the activation space that are most important for the decision-making of a model and, by extension, its accuracy. Since XAI techniques primarily focus on these areas, their results could be used for model compression in NAS, or memory compression in CL. To better investigate XAI-driven segmentation model enhancements, the following objectives can be defined:

- Identify the most suitable XAI techniques in segmentation based on computational requirements and quantitative XAI metrics.
- Investigate whether the CAM-NAS application in classification can be successfully extended to segmentation.
- Evaluate the performance of various explainable segmentation techniques, focusing on their potential uses in NAS.
- Explore the use of explainable segmentation techniques for memory compression in experience replay by storing only the image areas centered around the most important input features, as identified by selected XAI techniques.

2.4.1. Neural Architecture Search

Firstly, suitable XAI techniques have to be selected for the experiments. Based on previous research [A.1], [B.1], [B.2], gradient-based methods are preferable for the proposed use cases due to their lower computational costs. Processing time is an important factor when extracting saliency maps, especially when multiple iterations are required. This is further supported by the CAM-NAS [236] experiments in classification, where gradient-based methods achieve the best results. A simple gradient-based saliency map technique can be used as a baseline. Different implementations of Seg-Grad-CAM [212] can also be investigated. Since gradient-based techniques can generate a lot of noise, noise-reduction techniques, such as thresholding a certain percentage of pixels based on their importance, might also be considered. Especially when manual human-in-the-loop supervision is involved.

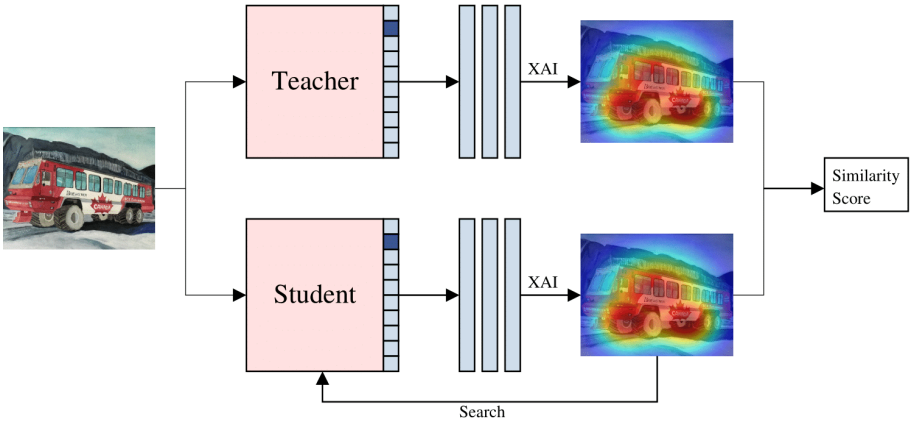


Figure 2.11: The pipeline for CAM-NAS in segmentation, based on the original implementation [236] for classification tasks. This is an idealized scenario where the saliency maps generated by both models are identical.

NAS focuses on automating the design of neural network architectures. Following [236], the initial teacher-student model will be extended to semantic segmentation models. The explanations will be generated based on the summed-up pre-Softmax prediction scores for the selected class of interest (Fig. 2.11). A well-trained segmentation model (the teacher) will be paired up with a less complex model (the student).

Then, explainable segmentation maps will be generated for the same input images and compared in terms of a similarity score. If the teacher model has truly learned the most important representations in an unbiased way, and if the selected XAI technique can capture the most important features for the model's decision-making, then the student model should ideally become sensitive to the same features. This could be viewed as a knowledge transfer from the teacher model to the student. The original CAM-NAS implementation [236] uses evolutionary algorithms for the generation of search submodels, and it could serve as an initial starting point for the experiments.

2.4.2. Continual Learning

It is less clear whether XAI-driven model enhancements can be implemented in the case of CL, specifically for memory compression in experience replay. CL focuses on how an already trained model can learn new tasks without forgetting the previous ones. Experience replay is an efficient CL strategy that allows storing the most important examples from old tasks inside the memory so that the model can still be exposed to them in the future. In classification, it is possible to reduce memory utilization by storing just the most salient regions of the data samples [179]. By cropping the image so that it is centered around the most important regions, memory can be utilized more efficiently. However, it is unclear if the cropping strategy could work in the case of segmentation, as it is a dense prediction task that, unlike image classification, could not be completed if part of the image was missing. In this particular context, compared to compression in classification, segmentation appears to be more sensitive to partial data. Enough critical contextual information would have to be stored for the segmentation to be successful. Perhaps less salient regions could be downsampled, as described in [148] in the case of classification. Then, enough contextual information could still be preserved to complete segmentation, especially if the right contextual information was identified by the explainable segmentation technique. Following [148], once the most important image regions are identified, they can be occluded by a bounding box. The resulting image with unoccluded non-discriminative pixels is then downsampled. Then, the previously occluded salient region is summed up to the downsampled image. The final image occupies significantly less space in memory. To

date, similar experiments have not yet been conducted for CL in image segmentation.

2.5. Chapter Conclusions

This chapter introduced three novel extensions to XAI techniques tailored for semantic segmentation: occlusion-based, activation perturbation-based, and gradient-based methods. An adaptation of occlusion-based methods, traditionally used in classification, to the pixel-dense domain of semantic segmentation was proposed. This included designing customized occlusion strategies and assessing their computational and interpretive trade-offs. Additionally, activation perturbation techniques were introduced, enabling the selective deactivation of feature maps to evaluate their contributions to segmentation outputs. Gradient-based methods were also extended for segmentation tasks, addressing common issues such as noise and instability in saliency maps to produce more reliable explanations. A three-term loss function was introduced to further investigate adversarial attacks against segmentation models in an experimental setting. Furthermore, this chapter provided theoretical frameworks for extending explainable segmentation techniques to XAI-driven model improvements, particularly in the case of neural architecture search and continual learning.

The following conclusions can be drawn:

1. Perturbations in the input space can be used for explainable segmentation if either min-max normalization or z-score standardization is applied first to disperse the Dice or IoU scores and generate more color intensities in the explanation. Alternatively, logit values can be used to generate explanations.

2. Ablation-CAM can be extended for segmentation tasks, and it is possible to generate explanations using either full or partial occlusions of their activation maps.

3. It is possible to attack explanations of segmentation models by optimizing a three-part loss function that minimizes the similarity between the generated explanation and its adversarial target (i.e., the explanation that we want the attacked model to generate), the similarity between the model's original output and the attacked

model's output, and the similarity between the original input image and the attacked input image.

This chapter presents a systematic evaluation of the proposed XAI methods, focusing on their interpretability and computational efficiency. Experiments are conducted using state-of-the-art segmentation models and datasets to assess the performance of perturbation-based and gradient-based methods. The chapter also explores the impact of adversarial attacks on explainability and proposes mitigation strategies. The main results presented in this chapter have been published in [A.1], [B.1], [B.2].

3.1. Datasets

Three datasets were used for the experiments: a private dataset obtained from CTI FoodTech[1], a leading manufacturer in industrial food processing machinery, a public semantic drone dataset[2] from Graz University of Technology, which focuses on increasing the safety of autonomous drones, and the COCO dataset [141] of common objects.

The private dataset is comprised of 752 peach images with the corresponding masks, showing the pit and the suture line, as can be seen in Figure 3.1. The ground-truths for the first dataset were manually labeled by CTI FoodTech operators.

Figure 3.1: Representative input image and its corresponding mask.

During preprocessing, a mean image was generated, and the images were cropped along its farthest contours while visually double-checking that the fruit remained within the frame. Input images and their corresponding masks were resized to 192×192. Extensive data augmentation was used, applying a sequence of transformations: horizontal and vertical flips, 90-degree rotations, the application of Gaussian noise, and

[1]https://ctifoodtech.com/en/
[2]http://dronedataset.icg.tugraz.at/

random changes in image brightness and contrast. Also, since the primary objective was to identify the cutting line, the pit and suture classes were merged into a single class, i.e., the masks were binarized, as seen in Figure 3.2.

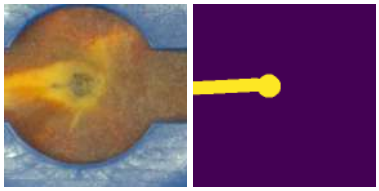


Figure 3.2: The pre-processed image with the corresponding mask.

The first dataset is limited by the size of its data points. Nonetheless, the simplicity of its geometric structures in its primary label allows for a methodic investigation of its response to adversarial attacks. When working with the first dataset, the classical U-Net [174] architecture with four encoder layers and 32 initial filters was trained, reaching a 0.778 mean Dice coefficient score on the test dataset.

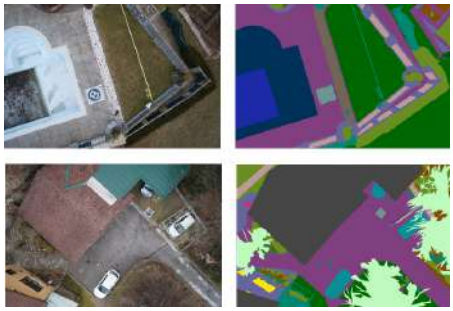


Figure 3.3: Representative input images and their RGB masks.

The second dataset is comprised of 400 6000×4000 high-resolution aerial segmentation images. It includes 20 semantic categories for the segmentation task, encompassing, among others, such classes as people, cars, and obstacles. The representative images can be seen in Figure 3.3. Given computational limitations, the images were resized to 1008×672 during training, offering a reasonable trade-off between image quality and memory efficiency. The augmentation techniques used on the peach dataset were similarly implemented on the drone dataset. The same four-encoder layer U-Net architecture was used, just like when working

with the peach dataset. However, for better performance, the model's backbone was pre-trained on the Imagenet [61] dataset. This way the model was able to achieve a 0.693 mean Dice coefficient score.

For the perturbation experiments, the COCO dataset [141] was also used. For the segmentation task, it consists of 21 classes of everyday objects, including different types of vehicles. Perturbation experiments in the input space, discussed in Section 3.2, were performed using the COCO dataset. Perturbation experiments in the activation space, discussed in Section 3.3, were conducted using both a private CTI FoodTech dataset and the COCO dataset. The experiments involving gradient-based saliency maps and their susceptibility to adversarial attacks, discussed in Section 3.4, were conducted using the CTI FoodTech dataset and the semantic drone dataset.

3.2. Experiments with Perturbations in the Input Space

3.2.1. Occlusion Approach for Semantic Segmentation

For further investigation, two pre-trained segmentation networks were used: FCN [145] with a ResNet-101 [98] backbone and DeepLabV3 [38]. COCO [141] dataset with 21 segmentation classes was chosen due to its focus on everyday objects, making the visual qualitative evaluation part easier by not requiring domain-specific knowledge. In most cases, the segmentation results of both models do not differ significantly.

Usually, in the case of non-background classes, occlusions generate the same segmentation output, just without the occluded region. Some larger occlusions, however, cause segmentation output to encompass areas that previously were not a part of it (Fig. 3.4).

Figure 3.4: The occluded airplane fuselage and the corresponding output, generated by FCN.

Different types of occlusion filters can generate different segmentation outputs, even when applied to the same image region (Fig. 3.5). Segmentations based on Gaussian filter occlusions seem to be less prone to confuse the occlusion filter with the foreground object. They also seem to be more resilient in terms of recovering the occluded part of the image. Black filter occlusions can generate segmentation outputs where the filter is treated as an extension of the object, especially when the background is light. In such cases, those occlusion areas would also appear as darker regions within the saliency map. The visualizations (Fig. 3.5) have been generated using 100×100 occlusions in order to see the pronounced effect. In most cases, such an occlusion size will be too large for a useful saliency map.

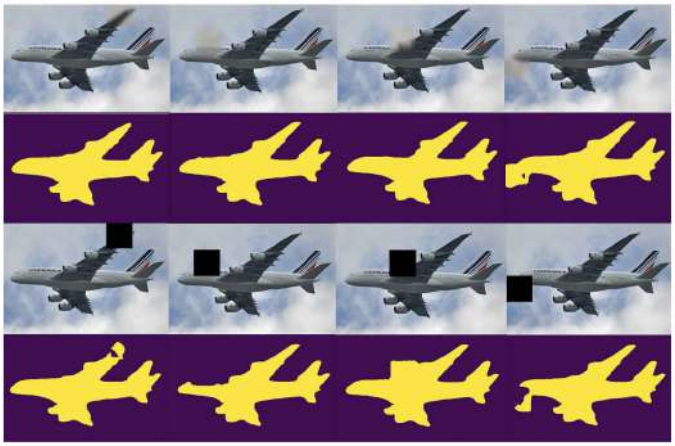


Figure 3.5: The effect of Gaussian (top two rows) and black (bottom two rows) occlusion filters on the segmentation output (FCN).

Just like in the classification task, it is noticeable that a small occlusion size usually does not have a significant effect on the model's output. Probably because it is easy for the network to recreate the image from the contextual information that is not occluded. However, a more detailed saliency map can be obtained by using a smaller occlusion size. Unlike in the classification task, most images show only a minimal difference between various Dice scores after occlusions, often just in the third decimal place. If these scores are used to generate saliency maps directly, different regions within such maps will be almost indistinguishable to the human eye (Fig. 3.6).

Figure 3.6: The original image, its segmentation output, and the saliency map based on non-normalized scores.

In such cases, min-max normalization was used to ensure that values were more dispersed within the interval [0, 1] rather than centered around a particular Dice coefficient value. In the experiments, the normalization step helped to increase the standard deviation of collected scores by 2-3 orders of magnitude. As a result, a greater range of color intensities was achieved when visualizing saliency maps (Fig. 3.7).

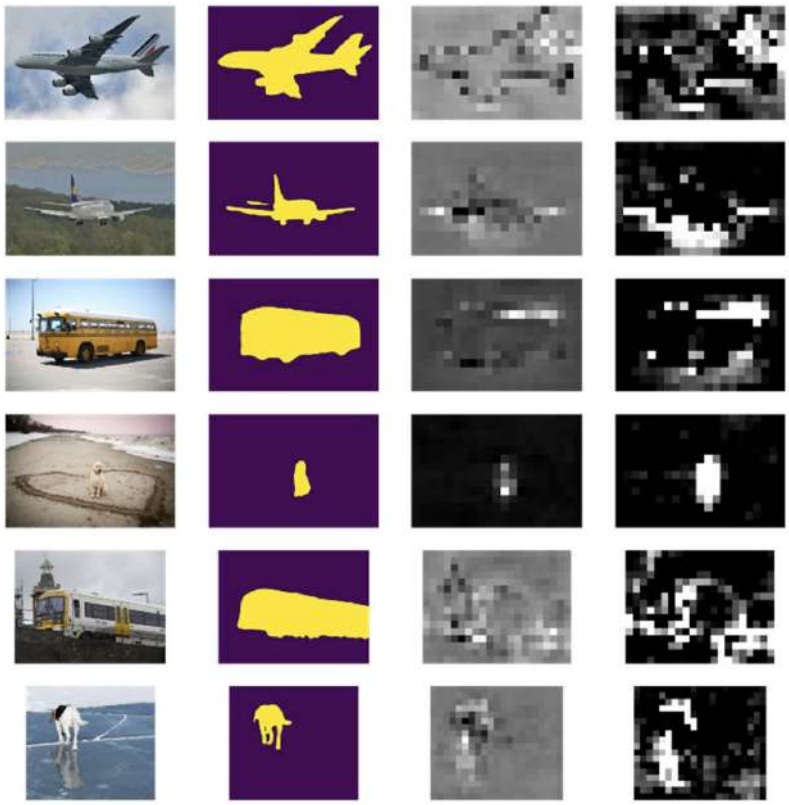


Figure 3.7: Each row shows an input image, its predicted mask, normalized saliency map, and standardized saliency map (DeepLabV3).

Z-score standardization with clipped negative values was also tested before visualizing the image. Based on the results, z-score-based saliency maps have starker contrasts between light and dark regions due to the less gradual color transitions. However, in most cases the results are clearer and less noisy when using normalization. Normalization was not required when visualizing occlusion scores for the classification task.

Visualizing the saliency map on top of the original image (Fig. 3.8) can be useful in detecting important features. However, some of the generated maps can be noisy. Color intensity thresholds can be selected to make the generated saliency maps clearer. For example, in the case of an airplane image (Fig. 3.8), the threshold for normalized scores was set to 0.3, so that only the most important features would be represented.

Figure 3.8: Overlayed saliency maps with (bottom image) and without (top image) thresholding.

The importance scores used for generating heatmaps were also calculated using pre-normalized probability scores before the Softmax layer, computed with respect to a class of interest. All logit values for a class of interest were summed up for each pixel classified as belonging to that class, and the resulting scalar value was used to measure the impact of occlusion. Their corresponding visualizations (Fig. 3.9) appeared more sensitive compared to those generated using the Dice score (Fig. 3.7).

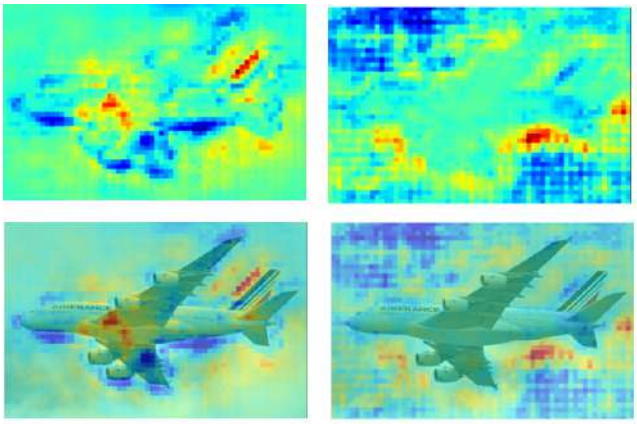


Figure 3.9: Heatmaps, generated using 2752 10×10 filters. Explanations correspond to the airplane (left) and the background (right) classes.

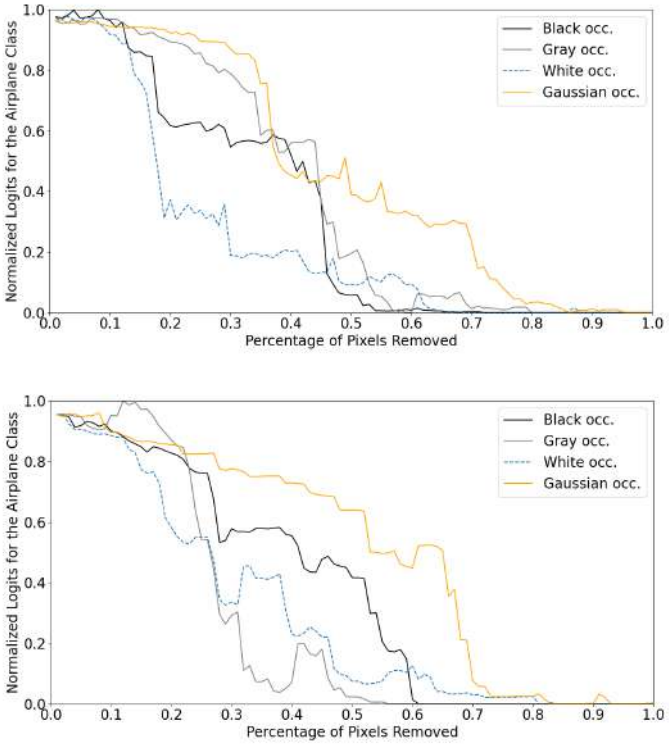


Figure 3.10: Deletion curves for the airplane image. 30×30 (top image) and 50×50 (bottom image) occlusion filters were used.

Furthermore, the results were evaluated quantitatively using deletion curves [169]. The impact of different occlusions was investigated on two input images where the foreground class of interest varied significantly in terms of occupied area (Fig. 3.8). The most important input features were gradually occluded, starting from the 99th percentile based on the previously calculated importance scores, and the impact on the model's prediction was measured. A lower AUC value, corresponding to a sharper decrease in the deletion curve, indicates a more discriminative interpretability technique capable of distinguishing more important input features.

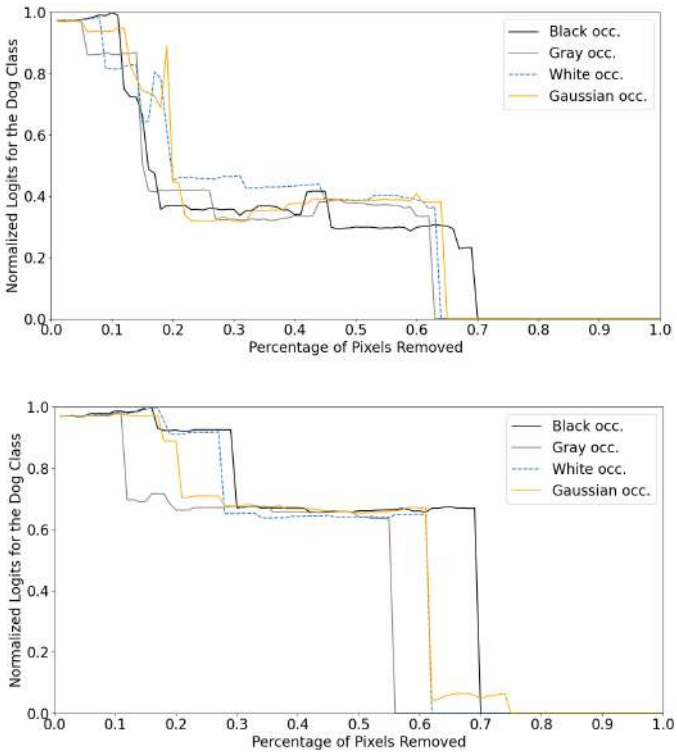


Figure 3.11: Deletion curves for the dog image. 30×30 (top image) and 50×50 (bottom image) occlusion filters were used.

The influence of occlusion color on overall segmentation was also investigated. As seen in Figure 3.10 and Figure 3.11, gray occlusion filters produced better results compared to white or black occlusions when larger 50×50 occlusions were used. Here, better results refer to a

sharper decline in the deletion curves, indicating that the explanation method effectively identifies the most important regions of the input and that the model's predictions are highly sensitive to occlusions in these areas. Specifically, a lower area under the deletion curve reflects greater sensitivity to important features, and provides evidence that the model relies appropriately on meaningful input regions. In most cases, black occlusions resulted in the worst results, consistent with previous qualitative observations (Fig. 3.5). For different input images, the corresponding AUC tends to decrease as the size of the occlusion filter decreases. The segmentation logits for the airplane class drop to zero after occluding 20-30% of input pixels when using 10×10 black occlusions (Fig. 3.12). A sharper decline in the deletion curves can be observed for the dog image (Fig. 3.11), which could be explained by its smaller relative area compared to the background.

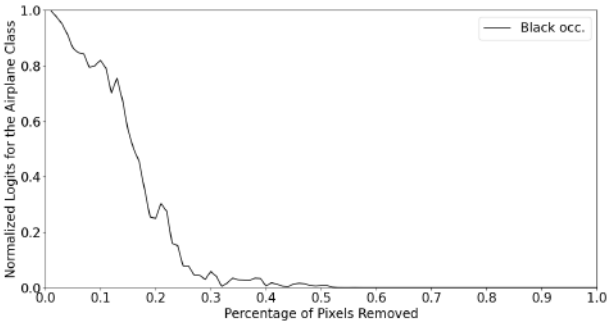


Figure 3.12: Deletion curve for the airplane class. 10×10 black filters were used.

However, the quantitative evaluation of interpretability methods is still an ongoing research area because of the previously discussed difficulty of interpretability-related concept formalization. The reliability of such evaluation might be questioned due to the generation of out-of-distribution samples when masking the input image [21].

3.3. Experiments with Perturbations in the Activation Space

In the perturbation-based experiments in the activation space, a U-Net [174] architecture with four encoder layers was employed, along with an FCN [145] model with a ResNet-101 [98] backbone for additional

experiments on the COCO [141] dataset. This section describes the implementation of the Ablation-CAM extension for semantic segmentation and provides a qualitative evaluation of its results.

The results of Ablation-CAM segmentation interpretability were compared with those of gradient-based saliency maps [192] extended for segmentation. Gradient-based saliency maps generated significantly more noise. Even though the main areas of interest were highlighted correctly, a wide additional region was salient as well. Gradient-free method, on the other hand, provided a clearer visualization of the important regions. The most influential areas were centered around fruit pits. For an image with lower segmentation accuracy (Fig. 2.8 (c)), the heatmap primarily captured the oval pit shape. The fruit's suture line, which posed challenges in segmentation, lacked visible activations. The gradient-based saliencies for this image were more scattered compared to those of the more accurately segmented images.

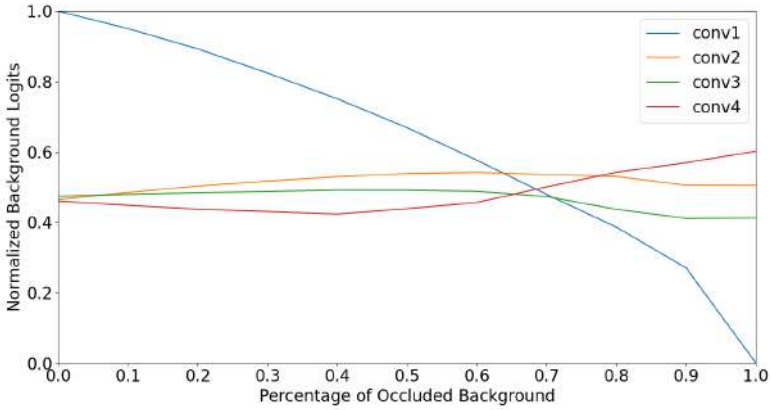
Partial Deactivation of Feature Maps

Partial deactivations of feature map regions were also performed based on their belonging either to the background or the foreground class.

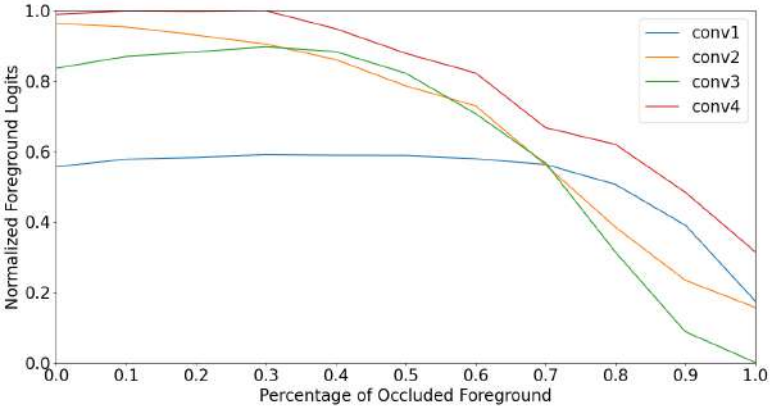
Prior research [125] suggests that the feature map occlusion inside the background area has a lesser influence on the network's classification when compared to the occlusion inside the foreground area (i.e. the class of interest). To investigate these results in the context of semantic segmentation, both the background and the cutting line were separately occluded for each of the four hidden layers (Fig. 2.7) inside the encoder block and the effect of such occlusions on the network's segmentation output was measured in terms of Dice score and min-max normalized logits for a class of interest. The results were calculated using different occlusion threshold values t, ranging from 0.0 to 1.0 in increments of 0.1. Results also indicate a more drastic change in the network's performance in the case of the cutting line occlusion (Fig. 3.13) for different encoder layers.

Experiments were also conducted with the partial deactivation of feature maps in order to explore other possible visualization strategies. The activation foreground was completely occluded while keeping the background intact, and vice versa. First, the importance score of each feature map was evaluated with its foreground region deactivated. The same evaluation was then performed for its background region. How-

ever, the resulting interpretability visualization did not have significant qualitative differences in either case. It is important to note that the results of interpretability experiments can vary depending on the specific task, the model architecture, and the data used. Therefore, it is always a good idea to experiment with different methods and compare the results to determine the best approach for a given task. In any case, partial deactivation of feature maps can still provide valuable information about the model's behavior and can help improve the transparency and accountability of AI models in various industrial applications.



(a) Background occlusions.



(b) Foreground occlusions.

Figure 3.13: The effect of feature map occlusions on the background and the foreground class. 50 random occlusion iterations were used on the image from Fig. 3.2.

To further analyze the impact of partial occlusion, the resulting importance scores were subtracted from the initial scores, which were generated as described in Section 2.2. The motivation behind this was to visualize which regions were affected the most by partial deactivations. If sensitivity to occlusions alone is of interest, rather than their direction, the absolute value of the initial difference can be taken. At the same time, this also allows visualizing which regions are the most resilient to partial ablations. In Figure 3.14, such regions correlate with colder colors. The visualization also shows that a larger region is affected in the case of background occlusion. In Figure 3.14 (d)-(f) activation maps indicate the highest sensitivity to background occlusion around the outer regions. While the fruit is also considered part of the background in areas where there are no cutting lines, the peach body does not seem to be affected as much. By visualizing the differences in the importance scores, a deeper understanding of how the model is using different input regions for prediction can be gained. This information can be useful for fine-tuning the model, improving its robustness, and enhancing its interpretability.

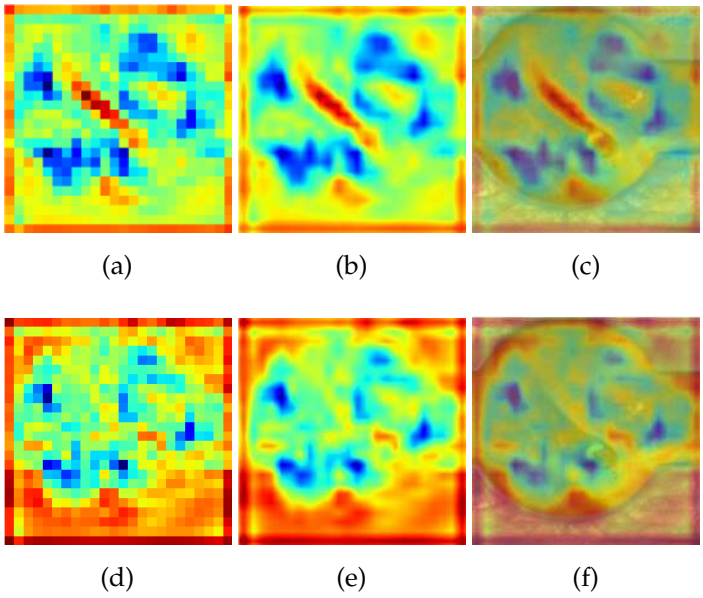


Figure 3.14: Occlusion difference maps. (a)-(c) refer to the foreground occlusion; (d)-(f) refer to the background occlusion. In each row, the occlusion difference map is shown together with its resized version and the input image overlay.

3.4. Experiments with Gradient-Based Explanations in Adversarial Contexts

3.4.1. Attack Against Segmentation Outputs

In this experiment, different transformations of the segmentation output are created by shifting it in four directions (vertically up, vertically down, horizontally left, and horizontally right) one pixel at a time, until the first pixel belonging to the mask reaches the edge of the image. If the segmentation output is already adjacent to the edge in a given direction, no new images are generated. The area of the segmented output remains invariant throughout.

Then, all generated images serve as adversarial dense prediction targets in the DAG algorithm [222] to produce the corresponding adversarial inputs. DAG allows finding the noise that, when applied to the input image, results in the model's output being close to the adversarial target. For the experiments, the number of iterations was set to 20, 30, or 50, while γ, which controls the change in the input image, was set to 0.1.

For the next step, saliency maps – both with and without a threshold – were calculated for each DAG-perturbed image. They were then compared to the saliency maps generated for the original unperturbed image using the structural similarity index measure (SSIM) [218] and mean squared error (MSE) metrics. Following [218], the SSIM is defined as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{3.1}$$

where

- μ_x and μ_y are the mean intensities of images x and y.
- σ_x^2 and σ_y^2 are the variances of x and y.
- σ_{xy} is the covariance between x and y.
- C_1 and C_2 are small constants to avoid instability, defined as:

  $$C_1 = (K_1 L)^2, \quad C_2 = (K_2 L)^2 \tag{3.2}$$

  where L is the dynamic range of pixel values, and $K_1 = 0.01$, $K_2 = 0.03$.

This experiment was conducted on 25 images from the peaches test set. The simple geometric shapes of the cutting lines allow for a systematic study of the impact of adversarial attacks on saliency maps in each of the selected spatial directions. The investigation focused on shifts along one axis at a time, although targeted attacks based on other transformations, such as rotations, would also work using this approach. When working with limited computational resources or larger images, transformations for adversarial targets can be created by moving the segmentation output by more than one pixel at a time.

In Figure 3.15, the SSIM dependency on shift magnitude in pixels in four directions is represented. The number of data points in each direction depends on the previous one pixel transformations up to the edge. Each subsequent transformation moves spatially further away from the original input image and that affects the corresponding saliencies. Therefore, as could be expected, the SSIM score for saliencies tends to decline with each shift. However, the decline of the curve is not as steep in each direction, and the AUC up to a selected value on a shift magnitude axis differs as well.

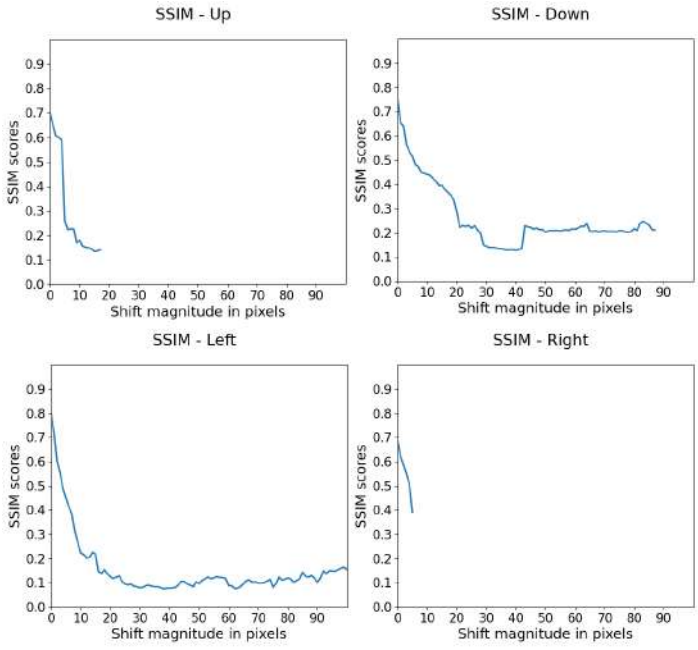


Figure 3.15: SSIM dependency on shift magnitude in four directions.

Similarity score distributions are visualized using box plots (Fig. 3.16). For the test images, most not-thresholded SSIM values fall within the 0.1-0.3 range. SSIM values for thresholded ($t = 100$) images are significantly higher, which can be explained by binarized saliencies. Based on scores from all four directions, a corresponding heatmap (Fig. 3.17) illustrates changes in similarity between attacked saliency maps and the original with respect to the targeted attack's spatial direction. Lighter colors correspond to higher SSIM values. The opposite is true for the normalized MSE values. In areas where the vertical and horizontal transformations intersect, pixel values are calculated as the average of intersecting values for that particular pixel. Most inspected cases exhibit a thin light buffer area around the segmentation output, where the similarity score does not change too drastically.

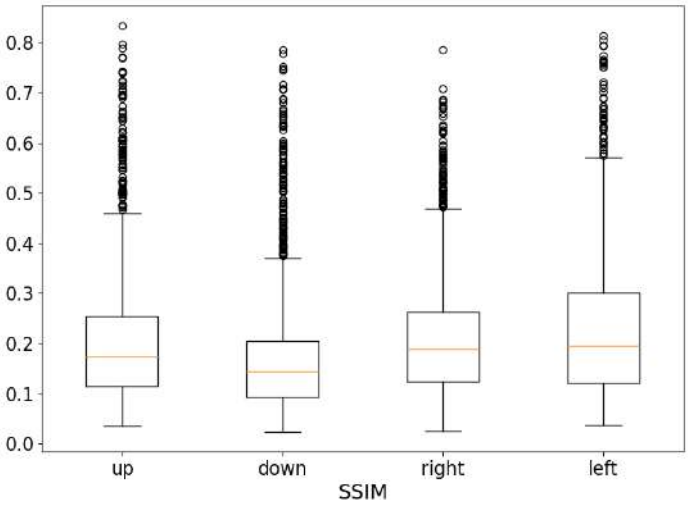


Figure 3.16: SSIM score distributions for all transformation directions for 25 test images from the peaches dataset.

The saliency similarity visualization approach was also implemented for the aerial segmentation dataset (Fig. 3.18). In this case, segmentation outputs are more complex in terms of their geometry and overlapping sections. When shifting the mask of a selected class, its previous location, which is not currently covered by the newly shifted mask, is filled with the dominant class of that image. In Figure 3.18 such class corresponds

to the paved area. However, after 50 DAG iterations, the selected image did not change too much even under the significant transformation, when it was moved right to the edge of the image. This indicates that, despite providing the DAG with the correct targets, the attack might fail due to an insufficient number of iterations or image-related conditions. In such cases, the target will not be reached, and saliency will be calculated for the output that was changed too little. The relative homogeneity of the generated heatmap (Fig. 3.18) would indicate an attack failure under the current settings. Another observation is that throughout the DAG attack, thresholded ($t = 100$) saliencies tend to become more visible compared to the saliency of the original unperturbed image.

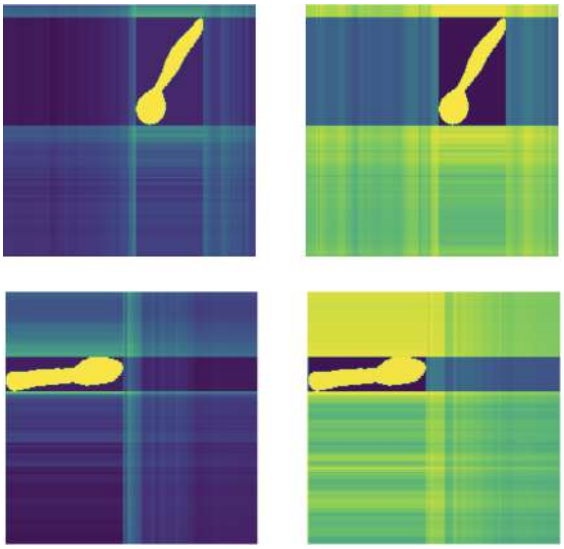
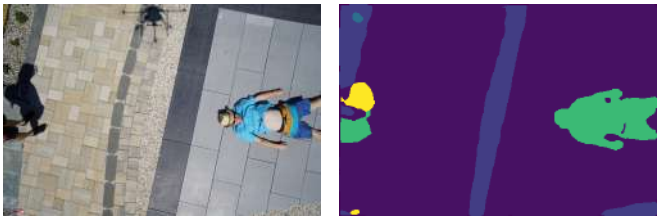


Figure 3.17: Saliency sensitivity heatmaps, generated using SSIM scores calculated for saliencies with (right) and without (left) a threshold.

These visualizations can be used to investigate how similar the saliency map of an attacked image will be to the original saliency map with respect to the targeted attack's direction. This approach is not limited to the DAG and the same visualization technique can be applied to other types of adversarial attacks. Further studies are needed to determine the extent to which the similarity scores depend on the underlying data and model as opposed to the failures of the attack itself. However, the visualizations can be useful in either case.



(a) Input image (b) Segmentation

(c) Shifting down (d) Shifting up

(e) Sensitivity heatmap (f) Thresholded heatmap

Figure 3.18: Saliency similarities after 50 DAG iterations.

3.4.2. Attack Against Segmentation Saliencies

In most cases, adversarial attacks seek to change the model's output. However, it is possible to further extend the reach of such attacks by focusing on manipulations in saliency while keeping the model's output close to the original. So far, the research on such attacks has been limited to classification. This subsection investigates its extension to semantic segmentation.

Compared to attacks on classification saliencies, it takes more time for segmentation saliency attacks to achieve the desired result. And even when this part of the attack manipulates the model's saliency successfully without changing its output drastically, the added noise

might render the input distortion perceptible to the human eye, thus invalidating the whole attack, as seen in Figure 3.19.

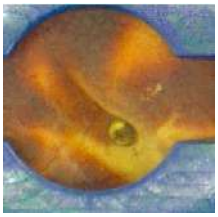


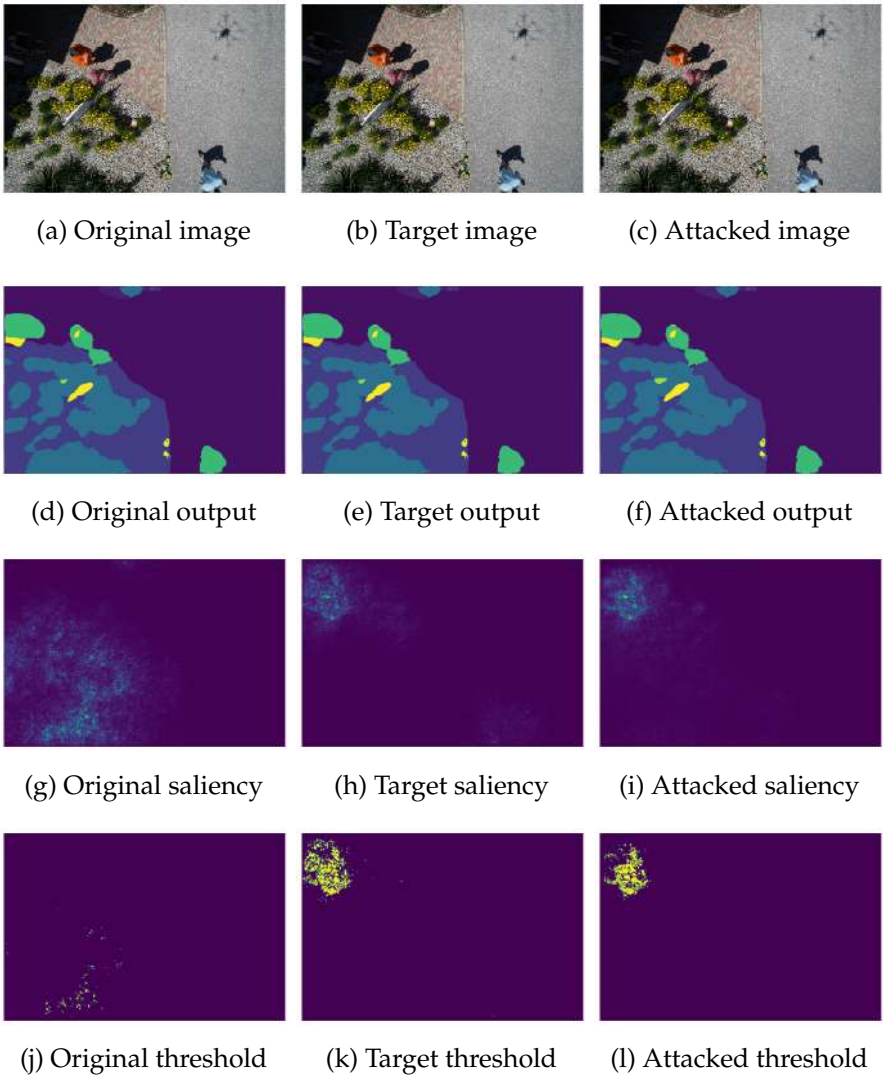
Figure 3.19: Perceptible distortions near the corners.

Based on [65], the saliency attack was implemented with the goal of changing the original image's explanation map while also trying to keep any changes in the model's output to a minimum. This can be achieved by constructing a loss function that controls changes in segmentation outputs and corresponding saliencies. This type of attack becomes easier if significant changes in the input image are allowed.

Unlike [65], where specific control of noise was not necessary to achieve a successful attack (possibly due to the relative ease of such attacks in classification tasks), experiments showed that the lack of control on perturbation often led to distorted images. To address this issue, a third loss term was introduced (see Subsection 2.3.2) to control the application of adversarial noise to the input image. This term constrains the adversarial noise, preserving the similarity between the original input image and the perturbed image.

Similar to [65], this study encountered the vanishing second derivative problem for ReLU non-linearities. To address this, the network's activation functions were changed from ReLU to Softplus during optimization. High beta values allow Softplus to approximate ReLU. The Adam optimizer was used to minimize the loss function. In practice, this type of attack could be more easily implemented on smaller images with a lower number of classes. For the aerial segmentation dataset, input images were resized to 432×288 to accommodate computational constraints. This resolution was chosen as it provided a reasonable trade-off between image quality and memory efficiency. As a result, the quality of the segmentation output decreased in terms of the mean Dice score, but the optimizing step for finding the adversarial noise became less computationally expensive. Based on experimental results, this image

size provided a good trade-off between the segmentation output and the time it takes to perform the attack.



(a) Original image (b) Target image (c) Attacked image

(d) Original output (e) Target output (f) Attacked output

(g) Original saliency (h) Target saliency (i) Attacked saliency

(j) Original threshold (k) Target threshold (l) Attacked threshold

Figure 3.20: A successful saliency attack on the drone dataset. The input image was perturbed to generate the saliency of the class *person* from the saliency of the class *vegetation*.

For the aerial segmentation dataset, the saliency of a vegetation class was successfully changed into the saliency of a person (Fig. 3.20). Both original and target saliencies were selected from the same input image.

The proposed attack allowed to change the saliency maps significantly while limiting perturbations of the input image. In the process, the segmentation output changed by a small amount. However, as can be seen in Figure 3.20, the segmentation output for both the person and vegetation classes does not look too different.

For the peaches dataset, two different images were selected and the attack changed the saliency of a cutting line of one peach into that of another (Fig. 3.21). The selected adversarial saliency targets were not similar to the original image saliencies. This made the attack more difficult. A larger number of iterations typically leads to a better MSE value between the target and attacked saliency maps, resulting in a higher degree of similarity. However, the experiments revealed that in certain cases, a relatively effective attack could be achieved with just 100 iterations. A quantitative analysis of 30 randomly selected images was performed to investigate the mutual influence of each term in the loss function on MSE results (Table 3.1).

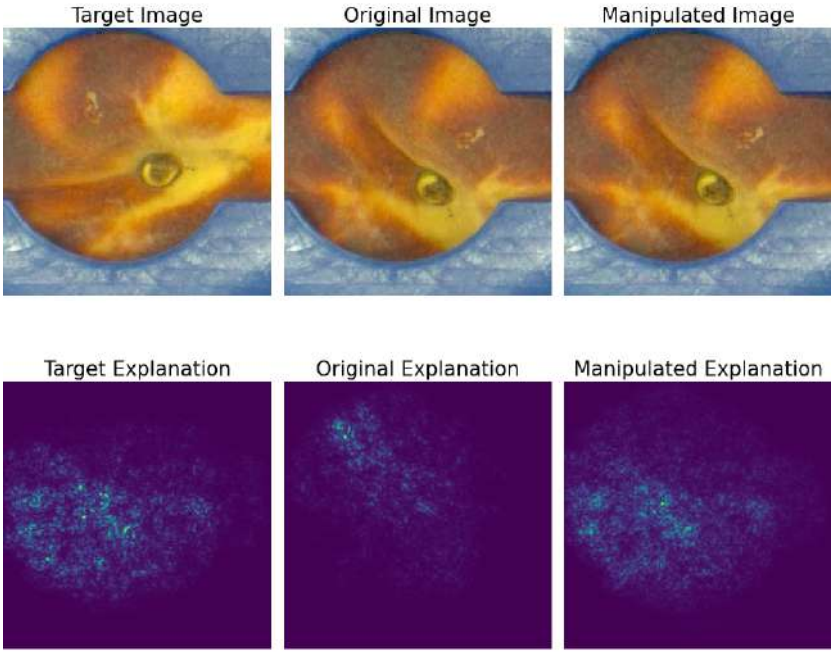


Figure 3.21: A successful saliency attack on the peaches dataset.

Table 3.1: Ablation study for the peach and drone datasets.

This table presents the mean MSE values along with their standard deviations.
For the peach dataset:
$iterations = 500, learning\,rate = 10^{-5}, \gamma_1 = 10^{11}, \gamma_2 = 10^5, \gamma_3 = 5 \cdot 10^6$;
For the drone dataset:
$iterations = 500, learning\,rate = 10^{-4}, \gamma_1 = 10^{11}, \gamma_2 = 5 \cdot 10^4, \gamma_3 = 10^6$.

	Peach Dataset	**Drone Dataset**
I $L = \gamma_1 L_{exp}$	$L_{exp} = (3.73 \pm 1.27) \cdot 10^{-9}$ $L_{out} = (6.91 \pm 8.25) \cdot 10^{-3}$ $L_{im} = (6.28 \pm 1.24) \cdot 10^{-6}$	$L_{exp} = (3.15 \pm 1.10) \cdot 10^{-10}$ $L_{out} = (3.58 \pm 1.33) \cdot 10^{-3}$ $L_{im} = (5.70 \pm 2.81) \cdot 10^{-4}$
I+II $L = \gamma_1 L_{exp} + \gamma_2 L_{out}$	$L_{exp} = (3.74 \pm 1.16) \cdot 10^{-9}$ $L_{out} = (1.03 \pm 1.17) \cdot 10^{-4}$ $L_{im} = (6.79 \pm 1.33) \cdot 10^{-6}$	$L_{exp} = (4.82 \pm 1.12) \cdot 10^{-10}$ $L_{out} = (2.51 \pm 1.20) \cdot 10^{-5}$ $L_{im} = (7.01 \pm 1.57) \cdot 10^{-5}$
I+III $L = \gamma_1 L_{exp} + \gamma_3 L_{im}$	$L_{exp} = (3.69 \pm 1.11) \cdot 10^{-9}$ $L_{out} = (6.29 \pm 7.45) \cdot 10^{-3}$ $L_{im} = (2.03 \pm 0.51) \cdot 10^{-6}$	$L_{exp} = (4.05 \pm 1.58) \cdot 10^{-10}$ $L_{out} = (5.05 \pm 1.54) \cdot 10^{-5}$ $L_{im} = (1.90 \pm 1.88) \cdot 10^{-5}$
I+II+III $L = \gamma_1 L_{exp} + \gamma_2 L_{out} + \gamma_3 L_{im}$	$L_{exp} = (3.79 \pm 1.16) \cdot 10^{-9}$ $L_{out} = (1.07 \pm 1.22) \cdot 10^{-4}$ $L_{im} = (1.98 \pm 0.45) \cdot 10^{-6}$	$L_{exp} = (5.59 \pm 1.38) \cdot 10^{-10}$ $L_{out} = (1.15 \pm 1.20) \cdot 10^{-5}$ $L_{im} = (6.22 \pm 2.35) \cdot 10^{-6}$

The MSE values for the peaches dataset are less dispersed than the MSE values for the drone dataset, and, based on the experimental results, the good adversarial attack loss parameters for one image in the peaches dataset are not that different from the other one. This could be explained by the fact that all the inputs, the outputs, and the saliency maps are relatively similar to each other compared to their corresponding counterparts in the drone dataset. Another contributing factor could be the significantly lower number of segmentation classes in the case of the peaches dataset. Ablation results showed that for the peaches dataset, the addition of each new term to optimize the loss function drastically lowered the value of the corresponding term without affecting the others significantly. For example, in the second row, L_{out} decreased by more than 50 times, in the third row, L_{im} decreased by more than 3

times, and in the fourth row, the benefits of the previous two methods were obtained without significant increases in L_{exp}. This indicates that the proposed three-term loss function performs better than the two-term loss function in [65]. In the drone dataset, the loss terms were more interconnected, and modifying each loss term had a substantial impact on the others. Therefore, a three-term loss function could be advantageous because of a greater number of possible combinations with more parameters.

3.5. Chapter Conclusions

The experimental results demonstrate the effectiveness of the proposed XAI methods but also highlight the trade-offs inherent in various explainability methods for semantic segmentation. While gradient-based techniques proved computationally efficient and well-suited for real-time or resource-constrained tasks, they exhibited limitations in robustness and stability. On the other hand, gradient-free perturbation-based methods offered more detailed and interpretable explanations but at the cost of significantly higher computational overhead, making them less practical for time-sensitive applications. The findings also revealed vulnerabilities to adversarial manipulation, underscoring the need for robust evaluation frameworks. These observations validate the potential of the methods while highlighting areas for further improvement, particularly in adversarial contexts.

4. DISCUSSION

This chapter discusses the broader implications of the findings from the experimental evaluation, focusing on the trade-offs between interpretability, robustness, and computational efficiency. It identifies open challenges in XAI for semantic segmentation and outlines future research directions. The main results presented in this chapter have been published in [A.2].

4.1. Open Issues

Plenty of unresolved challenges remain in explainable semantic segmentation, most of which are also applicable to image classification tasks. Below is a non-exhaustive list of these challenges:

- **Evaluation metrics for XAI**
  Most of the literature on XAI in image classification focuses on introducing new explainability techniques and their modifications, rather than proposing new evaluative frameworks or benchmark datasets. This tendency is even more visible in explainable semantic segmentation. Currently, there are no papers dedicated solely to evaluating XAI results in image segmentation. The investigation of XAI metrics remains limited to the experimental results sections, and only in those few cases where quantitative evaluation is used. There is no consensus on which evaluation metrics are most crucial for capturing the key aspects of explainability, largely due to the difficulty in formalizing explainability-related concepts. A better theoretical understanding of the problem should inform the creation of evaluative XAI metrics and benchmarks. Such foundations would likely result in more efficient explainable segmentation methods that are better adapted to the problem at hand.

- **Safety and robustness of XAI methods**
  With the rapid deployment of DL models in medical, military, and industrial settings, XAI techniques are set to play an even more important role. Their primary use is driven by the need to determine if the model is reliable and trustworthy. However, similar questions can also be raised about the XAI techniques themselves. It is important to investigate their vulnerabilities and loopholes.

Both deployers and end-users need to know whether they are secure against intentional attacks directed at XAI techniques or the model. Even if there is no direct threat, the robustness of each specific XAI method needs to be investigated on a case-by-case basis.

Just like classification models, semantic segmentation models are susceptible to adversarial attacks. Different attack methods have been proposed [47, 77, 222]. When discussing adversarial attacks, it is common to focus on the model's output as the primary target. However, it is also possible to attack the output's explanation saliency while leaving both the input and the output perceptibly unchanged. Such attacks have been introduced and investigated in the context of image classification [65]. It has also been demonstrated that these second-level attacks can be extended to image segmentation [A.1]. More research is needed to find the best ways to combat them, especially since new adversarial attacks are constantly being developed, and comprehensive safety guarantees are challenging to ensure. Systematic investigations need to be undertaken for both white-box attacks, where the attacked model is known to the attacker, and black-box attacks, where it is unknown. Similar investigations into the robustness of interpretable segmentation could contribute to the overall security of AI systems.

Adversarial examples are typically not part of the training and testing datasets. This omission can lead to vulnerabilities in deployed models. Another critical issue is the presence of biases. When the most salient regions of the explanation map fall outside the boundaries of the object of interest, this might signal not just a misguided prediction but also the potential presence of adversarial influences [107]. Natural adversarial examples [102] and their influence on XAI in segmentation could be investigated as well.

- **XAI for video segmentation**
  As semantic scene segmentation is not limited to 2D images, new interpretability techniques could be investigated for video data, where temporal semantic segmentation is carried out. Video object segmentation requires significantly more computational resources. To date, no studies have specifically investigated explainable im-

age segmentation in a dynamic setting. The nature of dynamic scenes could introduce novel challenges not previously encountered in 2D segmentation contexts. For instance, one would need to add a temporal explanation axis to account for differences in interpretability maps across video frames. This task could be further extended to real-time semantic segmentation by focusing on how to reduce the latency of the generated explanations.

- **Computational complexity**
  Different XAI techniques require different computational resources, which may not be readily available in certain environments. Deployment constraints may include issues related to both software and hardware, where real-time services have to be ensured for edge devices and online service platforms [46]. Further experimental studies are needed to investigate methods for reducing the computational complexity involved in generating explanations. This includes evaluating and optimizing the trade-offs between explanation quality and generation latency, particularly within diverse industrial contexts. Post-hoc methods, with perturbation-based techniques in particular, are rather inefficient in terms of explanation generation time. The total cost of generating local explanations increases with each new input image that requires interpretation, making it crucial to understand your use case and carefully assess available resources when selecting an XAI technique. Further research can help develop more efficient and scalable XAI methods, ensuring that high-quality explanations are provided within the resource constraints of specific applications.

4.2. Future Directions

- **XAI benchmarking and evaluations**
  Given that most literature primarily focuses on qualitative metrics, the need for a well-defined benchmark and evaluation strategy for XAI methods in image segmentation should be emphasized. To date, no studies have specifically addressed evaluations or benchmarks for XAI methods in this area. Moreover, research focusing on the formal aspects of quantitative metrics in XAI is limited.

- **Mechanistic interpretability and other XAI approaches**
  Mechanistic interpretability [31] is a promising research area. This approach seeks to reverse engineer how models function. Furthermore, there have been no significant contributions in formal explainability [151] or argumentative XAI [57] within the context of image segmentation.

- **XAI for transformers**
  Additionally, there has not been much research [119] into the interpretability of transformers for segmentation, especially compared to convolutional networks. This is of particular interest given the growing popularity of transformer architectures in various applications. Conducting more comparative studies between XAI techniques for different architectures or different convolutional operations, such as atrous convolutions, could also be explored.

- **Failure Modes**
  This area is related to evaluation metrics. However, it covers problematic areas that could not be identified by the commonly used metrics. Specifically, XAI could be used to identify and mitigate bias in segmentation models. A systematic analysis of failure cases and potential failure modes could better determine the scope of applicability for XAI methods. Several studies [5] have critically evaluated different groups of explainability techniques in classification. However, a similar investigation has not yet been conducted in image segmentation.

- **Neural architecture search**
  NAS explores automating neural architecture designs. XAI techniques can be applied in NAS in at least two distinct ways. First, existing XAI methods can be incorporated into NAS algorithms to improve their performance. For example, in [236], an explainable CAM technique is integrated with the NAS algorithm to avoid fully training submodels. Second, NAS algorithms can include interpretability aspects as one of the metrics to be optimized in multi-objective optimization. In [32], a surrogate interpretability metric has been used for multi-objective optimization in image classification. However, currently, no similar approaches exist for semantic segmentation tasks.

- **Continual Learning**
  CL refers to the research area that investigates techniques allowing models to learn new tasks without forgetting the previously learned ones. This strong tendency for DL models to forget previously learned information upon acquiring new knowledge is commonly described as catastrophic forgetting. More efficient solutions to CL problems would allow the models to be used more resourcefully, without retraining them from scratch when new data arrives. The intersection of XAI and CL presents an interesting area for investigation. XAI methods can be employed in CL to: (1) improve the model's performance; (2) better understand and explain the model's predictions; and (3) investigate the phenomenon of catastrophic forgetting. The exploration of XAI and CL could also lead to improved model understanding when either a shift in data distribution or concept drift occurs.

This dissertation presents a comprehensive view of XAI in image segmentation. It provides an up-to-date literature survey of various types of interpretability methods applied in semantic segmentation, and clarifies conceptual misunderstandings by proposing a taxonomy for explainable segmentation and general frameworks for different types of interpretability techniques. XAI methods in segmentation were categorized into five major subgroups: prototype-based, gradient-based, perturbation-based, counterfactual methods, and architecture-based techniques. Based on the surveyed literature on explainable image segmentation, it is evident that most of the methods focus on local explanations and rely on qualitative evaluation.

Occlusion sensitivity techniques were investigated for interpretable semantic segmentation. Contrary to their application for image classification, the occlusion-based methods in semantic segmentation do not seem to generate that much variance in the evaluation metric scores. Therefore, min-max normalization can be employed to generate cleaner saliency maps with more color intensities. Based on the qualitative results, the logits-based approach appears to be more sensitive compared to the Dice score-based approach and might be a better choice for generating saliency maps. The quantitative evaluation demonstrates that occlusions with colors that are more similar to the ones found in the image of interest are more suitable for generating interpretable heatmaps. Further research could systematically investigate input occlusions in multi-class segmentation scenarios as well as experiment with different occlusion slide sizes.

Qualitative results show a successful extension of Ablation-CAM to dense prediction tasks. A recreation of Foreground vs Background occlusion for different encoder layer activation maps supports the observations of [125], showing that the foreground occlusions have a greater impact on the network's output compared to the background occlusions. Partial occlusion sensitivities can be useful in showing regions that are the most and the least resilient to foreground or background occlusions.

Compared to simple gradients, the Ablation-CAM is less noisy, but more computationally demanding. Ablation-CAM-based approaches might not be as suitable for real-time or time-sensitive large-scale applications as those methods that only require a single inference and

backpropagation. The most suitable XAI method will depend on the specific requirements of the application and the trade-off between computational efficiency and the explanation noisiness. In some instances, it might be beneficial to use a combination of methods. Future work could investigate multi-class segmentation scenarios in the industrial setting.

This study represents the first investigation into the impact of adversarial attacks on the interpretability of semantic segmentation models. The proposed approach allows for visual analysis of adversarial attack effects on model explanations, particularly in scenarios with simpler segmentation shapes and fewer target classes. This dissertation also draws attention to the possibility of adversarial attacks on interpretable semantic segmentation when saliencies are targeted directly. Further research could explore physical adversarial attacks that take place under real-life conditions, as well as their transferability to black-box models.

The key conclusions from this dissertation are as follows:

1. Five distinct categories of XAI methods in segmentation have been identified (prototype- based, gradient-based, perturbation-based, counterfactual methods, and architecture-based techniques), with most of the field relying on qualitative XAI evaluations and local, and post-hoc, explanations.

2. The extension of Ablation-CAM to dense prediction tasks demonstrates that foreground occlusions have a greater impact than background occlusions.

3. Occlusion sensitivity techniques are applicable to semantic segmentation, with min-max normalization improving saliency maps and logits-based approaches proving more sensitive than Dice score-based methods.

4. Semantic segmentation models are susceptible to adversarial attacks that manipulate explanations. This highlights the need for research into attack transferability and defenses in black-box settings.

[1] Segmentation of neuronal structures in EM stacks challenge-ISBI2012-imagej.net. https://imagej.net/events/isbi-2012-segmentation-challenge. Last Accessed on July 4, 2024.

[2] M. Abukmeil, A. Genovese, V. Piuri, F. Rundo, and F. Scotti. Towards explainable semantic segmentation for autonomous driving systems by multi-scale variational attention. In *Proceedings of the IEEE International Conference on Autonomous Systems*, pages 1–5, 2021.

[3] G. Aceto, D. Ciuonzo, A. Montieri, V. Persico, and A. Pescapé. AI-powered internet traffic classification: Past, present, and future. *IEEE Communications Magazine*, 62(9):168–175, 2023.

[4] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin. From" where" to" what": Towards human-understandable explanations through concept relevance propagation. *arXiv preprint arXiv:2206.03208*, pages 1–87, 2022.

[5] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *Proceedings of the Advances in Advances in Neural Information Processing Systems*, volume 31, pages 1–11, 2018.

[6] J. Adebayo, M. Muelly, H. Abelson, and B. Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *Proceedings of the International Conference on Learning Representations*, pages 1–13, 2021.

[7] J. C. Aguirre-Arango, A. M. Álvarez-Meza, and G. Castellanos-Dominguez. Feet segmentation for regional analgesia monitoring using convolutional RFF and layer-wise weighted CAM interpretability. *Computation*, 11(6):113, 2023.

[8] A. M. Ahmed and L. A. Ali. Explainable medical image segmentation via generative adversarial networks and layer-wise relevance propagation. *Nordic Machine Intelligence*, 1(1):20–22, 2021.

[9] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020.

[10] F. Ambellan, A. Tack, M. Ehlke, and S. Zachow. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the

Osteoarthritis Initiative. *Medical Image Analysis*, 52:109–118, 2019.

[11] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, pages 1–29, 2016.

[12] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, pages 1–16, 2017.

[13] P. Angelov and E. Soares. Towards explainable deep neural networks (xDNN). *Neural Networks*, 130:185–194, 2020.

[14] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.

[15] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2): 915–931, 2011.

[16] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[17] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh. Deep semantic segmentation of natural and medical images: A review. *Artificial Intelligence Review*, 54:137–178, 2021.

[18] P. Ashtari, D. M. Sima, L. De Lathauwer, D. Sappey-Marinier, F. Maes, and S. Van Huffel. Factorizer: A scalable interpretable approach to context modeling for medical image segmentation. *Medical Image Analysis*, 84:102706, 2023.

[19] A. Atrey, K. Clary, and D. Jensen. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. *arXiv preprint arXiv:1912.05743*, pages 1–23, 2019.

[20] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier

decisions by layer-wise relevance propagation. *PloS one*, 10(7): e0130140, 2015.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[22] F. Bardozzo, T. Collins, A. Forgione, A. Hostettler, and R. Tagliaferri. StaSiS-Net: A stacked and siamese disparity estimation network for depth reconstruction in modern 3d laparoscopy. *Medical Image Analysis*, 77:102380, 2022.

[23] F. Bardozzo, M. D. Priscoli, T. Collins, A. Forgione, A. Hostettler, and R. Tagliaferri. Cross X-AI: Explainable semantic segmentation of laparoscopic images in relation to depth estimation. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, pages 1–8, 2022.

[24] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017.

[25] M. S. Bedmutha and S. Raman. Using class activations to investigate semantic segmentation. In *Proceedings of the Computer Vision and Image Processing Conference, Prayagraj, India, Revised Selected Papers, Part III 5*, pages 151–161, 2021.

[26] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[27] M. Biehl, B. Hammer, and T. Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2):92–111, 2016.

[28] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand, et al. The liver tumor segmentation benchmark (LiTS). *Medical Image Analysis*, 84:102680, 2023.

[29] B. Bilodeau, N. Jaques, P. W. Koh, and B. Kim. Impossibility theorems for feature attribution. *arXiv preprint arXiv:2212.11870*,

pages 1–38, 2022.

[30] N. Bloch, A. Madabhushi, H. Huisman, J. Freymann, J. Kirby, M. Grauer, A. Enquobahrie, C. Jaffe, L. Clarke, and K. Farahani. NCI-ISBI 2013 challenge: Automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370(6):5, 2015.

[31] N. Cammarata, G. Goh, S. Carter, C. Voss, L. Schubert, and C. Olah. Curve circuits. *Distill*, 2021. doi: 10.23915/distill.00024.006. https://distill.pub/2020/circuits/curve-circuits.

[32] Z. Carmichael, T. Moon, and S. A. Jacobs. Learning interpretable models through multi-objective neural architecture search. *arXiv preprint arXiv:2112.08645*, pages 1–25, 2021.

[33] J. Černevičienė and A. Kabašinskas. Review of multi-criteria decision-making methods in finance using explainable artificial intelligence. *Frontiers in artificial intelligence*, 5:827584, 2022.

[34] J. Černevičienė and A. Kabašinskas. Explainable artificial intelligence (xai) in finance: a systematic literature review. *Artificial Intelligence Review*, 57(8):216, 2024.

[35] S. Chatterjee, A. Das, C. Mandal, B. Mukhopadhyay, M. Vipinraj, A. Shukla, R. Nagaraja Rao, C. Sarasaen, O. Speck, and A. Nürnberger. TorchEsegeta: Framework for interpretability and explainability of image-based deep learning models. *Applied Sciences*, 12 (4):1834, 2022.

[36] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: Deep learning for interpretable image recognition. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 32, pages 8930–8941, 2019.

[37] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, pages 1–13, 2021.

[38] L.-C. Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40 (4):834–848, 2017.

[40] Q. Chen, L. Yang, J.-H. Lai, and X. Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4288–4298, 2022.

[41] Y.-C. Cheng, Z.-Y. Shiau, F.-E. Yang, and Y.-C. F. Wang. TAX: Tendency-and-assignment explainer for semantic segmentation with multi-annotators. *arXiv preprint arXiv:2302.09561*, pages 1–10, 2023.

[42] J. Chiu, C. C. Li, and O. J. Mengshoel. Potential applications of deep learning in automatic rock joint trace mapping in a rock mass. In *Proceedings of the IOP Conference*, volume 1124, pages 1–8, 2023.

[43] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu. Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. *Biomedical Optics Express*, 6(4):1172–1194, 2015.

[44] P. F. Christ, F. Ettlinger, F. Grün, M. E. A. Elshaera, J. Lipkova, S. Schlecht, F. Ahmaddy, S. Tatavarty, M. Bickel, P. Bilic, et al. Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. *arXiv preprint arXiv:1702.05970*, pages 1–20, 2017.

[45] M. Chromik and M. Schuessler. A taxonomy for human subject evaluation of black-box explanations in XAI. *ExSS-ATEC@IUI*, 1: 1–7, 2020.

[46] Y.-N. Chuang, G. Wang, F. Yang, Z. Liu, X. Cai, M. Du, and X. Hu. Efficient XAI techniques: A taxonomic survey. *arXiv preprint arXiv:2302.03225*, 2023.

[47] M. Cisse, Y. Adi, N. Neverova, and J. Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, pages 1–12, 2017.

[48] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, pages 1–12, 2019.

[49] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al.

Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pages 168–172, 2018.

[50] J. Colin, T. Fel, R. Cadène, and T. Serre. What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pages 2832–2845, 2022.

[51] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, et al. BCN20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, pages 1–3, 2019.

[52] E. Commission. Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021. Available Online at: `https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A52021PC0206`, Last Accessed on April 22, 2024.

[53] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[54] K. Cortacero, B. McKenzie, S. Müller, R. Khazen, F. Lafouresse, G. Corsaut, N. Van Acker, F.-X. Frenois, L. Lamant, N. Meyer, et al. Evolutionary design of explainable algorithms for biomedical image segmentation. *Nature Communications*, 14(1):7112, 2023.

[55] V. Couteaux, O. Nempont, G. Pizaine, and I. Bloch. Towards interpretability of segmentation networks by analyzing DeepDreams. In *Proceedings of the Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI Conference, Shenzhen, China*, pages 56–63, 2019.

[56] C. Curtis, N. Gillespie, and S. Lockey. AI-deploying organizations are key to addressing 'perfect storm' of AI risks. *AI and Ethics*, 3 (1):145–153, 2023.

[57] K. Čyras, A. Rago, E. Albini, P. Baroni, and F. Toni. Argumentative XAI: A survey. *arXiv preprint arXiv:2105.11266*, pages 1–8, 2021.

[58] W. Dai, S. Liu, C. B. Engstrom, and S. S. Chandra. Explainable semantic medical image segmentation with style. *arXiv preprint arXiv:2303.05696*, pages 1–12, 2023.

[59] P. Dardouillet, A. Benoit, E. Amri, P. Bolon, D. Dubucq, and A. Crédoz. Explainability of image semantic segmentation through SHAP values. In *Proceedings of the ICPR Workshops of the International Conference on Pattern Recognition Workshops*, pages 188–202, 2022.

[60] S. Dasanayaka, S. Silva, V. Shantha, D. Meedeniya, and T. Ambegoda. Interpretable machine learning for brain tumor analysis using MRI. In *Proceedings of the IEEE International Conference on Advanced Research in Computing*, pages 212–217, 2022.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[62] S. Desai and H. G. Ramaswamy. Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020.

[63] A. Di Martino, G. Carlini, G. Castellani, D. Remondini, and A. Amorosi. Sediment core analysis using artificial intelligence. *Scientific Reports*, 13(1):20409, 2023.

[64] I. Dirks, M. Keyaerts, B. Neyns, and J. Vandemeulebroucke. Computer-aided detection and segmentation of malignant melanoma lesions on whole-body 18F-FDG PET/CT using an interpretable deep learning approach. *Computer Methods and Programs in Biomedicine*, 221:106902, 2022.

[65] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel. Explanations can be manipulated and geometry is to blame. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 32, pages 1–12, 2019.

[66] J. Donnelly, A. J. Barnett, and C. Chen. Deformable ProtoPNet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10265–10275, 2022.

[67] M. Dörrich, M. Hecht, R. Fietkau, A. Hartmann, H. Iro, A.-O. Gostian, M. Eckstein, and A. M. Kist. Explainable convolutional neural networks for assessing head and neck cancer histopathology. *Diagnostic Pathology*, 18(1):121, 2023.

[68] J. A. Dowling, J. Sun, P. Pichler, D. Rivest-Hénault, S. Ghose, H. Richardson, C. Wratten, J. Martin, J. Arm, L. Best, et al. Automatic substitute computed tomography generation and contouring for magnetic resonance imaging (MRI)-alone external beam radiation therapy from standard MRI sequences. *International Journal of Radiation Oncology* Biology* Physics*, 93(5):1144–1153, 2015.

[69] R. L. Draelos and L. Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, pages 1–20, 2020.

[70] M. Dreyer, R. Achtibat, T. Wiegand, W. Samek, and S. Lapuschkin. Revealing hidden context bias in segmentation and object detection through concept-specific explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3838, 2023.

[71] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. Technical report, University of Montreal, 2009. 1341.

[72] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88:303–338, 2010.

[73] F.-L. Fan, J. Xiong, M. Li, and G. Wang. On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(6):741–760, 2021.

[74] A. Farrag, G. Gad, Z. M. Fadlullah, M. M. Fouda, and M. Alsabaan. An explainable AI system for medical image segmentation with preserved local resolution: Mammogram tumor segmentation. *IEEE Access*, pages 125543–125561, 2023.

[75] T. Fel, R. Cadène, M. Chalvidal, M. Cord, D. Vigouroux, and T. Serre. Look at the variance! Efficient black-box explanations with Sobol-based sensitivity analysis. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pages 26005–26014, 2021.

[76] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

[77] V. Fischer, M. C. Kumar, J. H. Metzen, and T. Brox. Adversarial examples for semantic image segmentation. *arXiv preprint arXiv:1703.01101*, pages 1–4, 2017.

[78] F. Forest, H. Porta, D. Tuia, and O. Fink. From classification to segmentation with explainable AI: A study on crack detection and growth monitoring. *arXiv preprint arXiv:2309.11267*, pages 1–43, 2023.

[79] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[80] J. Fritsch, T. Kuehnl, and A. Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *Proceedings of the IEEE Conference on Intelligent Transportation Systems*, pages 1693–1700, 2013.

[81] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez. RIM-ONE: An open retinal image database for optic nerve evaluation. In *Proceedings of the International Symposium on Computer-based Medical Systems*, pages 1–6, 2011.

[82] S. Gao, H. Zhou, Y. Gao, and X. Zhuang. BayeSeg: Bayesian modeling for medical image segmentation with interpretable generalizability. *arXiv preprint arXiv:2303.01710*, pages 1–15, 2023.

[83] S. Gatidis, T. Hepp, M. Früh, C. La Fougère, K. Nikolaou, C. Pfannenberg, B. Schölkopf, T. Küstner, C. Cyran, and D. Rubin. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022.

[84] T. Geertsma. Ultrasoundcases. info, 2014.

[85] R. Geirhos, R. S. Zimmermann, B. Bilodeau, W. Brendel, and B. Kim. Don't trust your eyes: on the (un) reliability of feature visualizations. *arXiv preprint arXiv:2306.04719*, pages 1–32, 2023.

[86] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. H. Pham, M. Mühlegg, S. Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, pages 1–10, 2020.

[87] A. Ghorbani, A. Abid, and J. Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019.

[88] A. K. Gizzini, M. Shukor, and A. J. Ghandour. Extending CAM-based XAI methods for remote sensing imagery segmentation. *arXiv preprint arXiv:2310.01837*, pages 1–7, 2023.

[89] L. Goasduff. Gartner says cisos need to champion ai trism to improve ai results, 2023. Available online at: https://www.gartner.com/en/newsroom/press-releases/2023-09-27-gartner-says-cisos-need-to-champion-ai-trism-to-improve-ai-results. Last accessed: August 28, 2024.

[90] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017.

[91] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[92] M. Graziani, I. Palatnik de Sousa, M. M. Vellasco, E. Costa da Silva, H. Müller, and V. Andrearczyk. Sharpening local interpretable model-agnostic explanations for histopathology: Improved understandability and reliability. In *Proceedings of the MICCAI Conference, Strasbourg, France, Part III 24*, pages 540–549, 2021.

[93] M. Graziani, L. Dutkiewicz, D. Calvaresi, J. P. Amorim, K. Yordanova, M. Vered, R. Nair, P. H. Abreu, T. Blanke, V. Pulignano, J. O. Prior, L. Lauwaert, W. Reijers, A. Depeursinge, V. Andrearczyk, and H. M*ü*ller. A global taxonomy of interpretable AI: Unifying the terminology for the technical and social sciences. *Artificial Intelligence Review*, 56(4):3473–3504, 2023.

[94] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(2):699–711, 2020.

[95] R. Guidotti. Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

[96] S. N. Hasany, C. Petitjean, and F. Mériaudeau. Seg-XRes-CAM:

Explaining spatially local regions in image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3732–3737, 2023.

[97] S. N. Hasany, F. Mériaudeau, and C. Petitjean. Post-hoc XAI in medical image segmentation: The journey thus far. In *Proceedings of the Medical Imaging with Deep Learning*, pages 1–17, 2024.

[98] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[99] S. He, Y. Feng, P. E. Grant, and Y. Ou. Segmentation ability map: Interpret deep features for medical image segmentation. *Medical Image Analysis*, 84:102726, 2023.

[100] X. He, Y. Wang, F. Poiesi, W. Song, Q. Xu, Z. Feng, and Y. Wan. Exploiting multi-granularity visual features for retinal layer segmentation in human eyes. *Frontiers in Bioengineering and Biotechnology*, 11:1–14, 2023.

[101] X. He, W. Song, Y. Wang, F. Poiesi, J. Yi, M. Desai, Q. Xu, K. Yang, and Y. Wan. Light-weight retinal layer segmentation with global reasoning. *arXiv preprint arXiv:2404.16346*, 2024.

[102] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[103] J. Heo, S. Joo, and T. Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in Neural Information Processing Systems*, 32, 2019.

[104] J. Hilton, N. Cammarata, S. Carter, G. Goh, and C. Olah. Understanding RL vision. *Distill*, 5(11):e29, 2020.

[105] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, pages 1–50, 2018.

[106] R. Hosseini and P. Xie. Saliency-aware neural architecture search. *Proceedings of the Advances in Neural Information Processing Systems*, 35:14743–14757, 2022.

[107] L. Hoyer, M. Munoz, P. Katiyar, A. Khoreva, and V. Fischer. Grid saliency for context explanations of semantic segmentation. In *Proceedings of the Advances in Neural Information Processing Systems*,

volume 32, pages 1–12, 2019.

[108] Y. Huang, C. Qiu, and K. Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 36(1):85–96, 2020.

[109] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord. STEEX: Steering counterfactual explanations with semantics. In *Proceedings of the European Conference on Computer Vision*, pages 387–403, 2022.

[110] A. Janik, K. Sankaran, and A. Ortiz. Interpreting black-box semantic segmentation models in remote sensing applications. *Machine Learning Methods in Visualisation for Big Data*, pages 7–11, 2019.

[111] A. Janik, J. Dodd, G. Ifrim, K. Sankaran, and K. Curran. Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset. In *Proceedings of SPIE Medical Imaging Conference*, volume 11596, pages 861–872, 2021.

[112] H. Jeffreys. *The Theory of Probability*. OUP Oxford, 1998.

[113] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen. Kvasir-seg: A segmented polyp dataset. In *Proceedings of the MultiMedia Modeling Conference, Daejeon, South Korea, January 5–8, Part II 26*, pages 451–462, 2020.

[114] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. de Lange, P. T. Schmidt, H. D. Johansen, et al. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *Proceedings of the MultiMedia Modeling Conference, Prague, Czech Republic, June 22–24, Part II 27*, pages 218–229, 2021.

[115] S. Ji, S. Wei, and M. Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2018.

[116] Y. Jia, J. Wang, C. M. Poskitt, S. Chattopadhyay, J. Sun, and Y. Chen. Adversarial attacks and mitigation for anomaly detectors of cyber-physical systems. *International Journal of Critical Infrastructure Protection*, 34:100452, 2021.

[117] S. Kang, Z. Chen, L. Li, W. Lu, X. S. Qi, and S. Tan. Learning feature fusion via an interpretation method for tumor segmentation on PET/CT. *Applied Soft Computing*, 148:110825, 2023.

[118] A. Kapishnikov, T. Bolukbasi, F. Viégas, and M. Terry. XRAI: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4948–4957, 2019.

[119] R. Karim and R. P. Wildes. Understanding video transformers for segmentation: A survey of application and interpretability. *arXiv preprint arXiv:2310.12296*, pages 1–113, 2023.

[120] M. Karimzadeh, A. Vakanski, M. Xian, and B. Zhang. Post-hoc explainability of BI-RADS descriptors in a multi-task framework for breast cancer detection and segmentation. *arXiv preprint arXiv:2308.14213*, pages 1–11, 2023.

[121] M. Karri, C. S. R. Annavarapu, and U. R. Acharya. Explainable multi-module semantic guided attention based network for medical image segmentation. *Computers in Biology and Medicine*, 151: 106231, 2022.

[122] A. Kaur, G. Dong, and A. Basu. GradXcepUNet: Explainable AI based medical image segmentation. In *Proceedings of the International Conference on Smart Multimedia*, pages 174–188, 2022.

[123] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, et al. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[124] H. Kayan, M. Nunes, O. Rana, P. Burnap, and C. Perera. Cybersecurity of industrial cyber-physical systems: a review. *ACM Computing Surveys (CSUR)*, 54(11s):1–35, 2022.

[125] B. Kim, J. Seo, S. Jeon, J. Koo, J. Choe, and T. Jeon. Why are saliency maps noisy? cause of and solution to noisy saliency maps. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4149–4157. IEEE, 2019.

[126] S. S. Kim, N. Meister, V. V. Ramaswamy, R. Fong, and O. Russakovsky. Hive: evaluating the human interpretability of visual explanations. In *Proceedings of the European Conference on Computer Vision*, pages 280–298, 2022.

[127] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un) reliability of saliency methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280, 2019.

[128] H. Kirişli, M. Schaap, C. Metz, A. Dharampal, W. B. Meijboom, S.-

L. Papadopoulou, A. Dedic, K. Nieman, M. A. de Graaf, M. Meijs, et al. Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography. *Medical Image Analysis*, 17(8):859–876, 2013.

[129] T. Koker, F. Mireshghallah, T. Titcombe, and G. Kaissis. U-noise: Learnable noise masks for interpretable image segmentation. In *2021 IEEE International Conference on Image Processing*, pages 394–398, 2021.

[130] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al. Captum: A unified and generic model interpretability library for PyTorch. pages 1–11, 2020.

[131] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 25, pages 1–9, 2012.

[132] A. Kuehlkamp, A. Boyd, A. Czajka, K. Bowyer, P. Flynn, D. Chute, and E. Benjamin. Interpretable deep learning-based forensic iris segmentation and recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 359–368, 2022.

[133] A. Lalande, Z. Chen, T. Decourselle, A. Qayyum, T. Pommier, L. Lorgis, E. de La Rosa, A. Cochet, Y. Cottin, D. Ginhac, et al. Emidec: A database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac MRI. *Data*, 5(4):89, 2020.

[134] Z. Lambert and C. Le Guyader. About the incorporation of topological prescriptions in CNNs for medical image semantic segmentation. *Journal of Mathematical Imaging and Vision*, pages 1–28, 2024.

[135] Z. Lambert, C. Petitjean, B. Dubray, and S. Kuan. SegTHOR: Segmentation of thoracic organs at risk in CT images. In *Proceedings of the International Conference on Image Processing Theory, Tools and Applications*, pages 1–6, 2020.

[136] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein. MICCAI multi-atlas labeling beyond the cranial vault–workshop

and challenge. In *Proceedings of the MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015.

[137] Y. LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[138] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.

[139] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review. *Computers in Biology and Medicine*, 60:8–31, 2015.

[140] J. Li, P. Jin, J. Zhu, H. Zou, X. Xu, M. Tang, M. Zhou, Y. Gan, J. He, Y. Ling, et al. Multi-scale GCN-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary OCT images. *Biomedical Optics Express*, 12(4):2204–2220, 2021.

[141] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the ECCV 2014, Zurich, Switzerland, September 6-12, Part V 13*, pages 740–755, 2014.

[142] G. Litjens, R. Toth, W. Van De Ven, C. Hoeks, S. Kerkstra, B. Van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, et al. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Medical Image Analysis*, 18(2):359–373, 2014.

[143] F. Liu, W. Ding, Y. Qiao, and L. Wang. Transfer learning-based encoder-decoder model with visual explanations for infrastructure crack segmentation: New open database and comprehensive evaluation. *Underground Space*, pages 60–81, 2023.

[144] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[145] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[146] M. Losch, M. Fritz, and B. Schiele. Semantic bottlenecks: Quantifying and improving inspectability of deep representations. *Inter-*

national Journal of Computer Vision, 129:3136–3153, 2021.

[147] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 30, pages 4768—-4777, 2017.

[148] Z. Luo, Y. Liu, B. Schiele, and Q. Sun. Class-incremental exemplar compression for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11371–11380, 2023.

[149] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 3226–3229, 2017.

[150] H. Mankodiya, D. Jadav, R. Gupta, S. Tanwar, W.-C. Hong, and R. Sharma. OD-XAI: Explainable AI-based semantic object detection for autonomous vehicles. *Applied Sciences*, 12(11):5310, 2022.

[151] J. Marques-Silva. Logic-based explainability in machine learning. In *Reasoning Web. Causality, Explanations and Declarative Knowledge: 18th International Summer School 2022, Berlin, Germany, September 27–30, 2022, Tutorial Lectures*, pages 24–104. 2023.

[152] N. Maslej, L. Fattorini, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, H. Ngo, J. C. Niebles, V. Parli, et al. Artificial Intelligence Index Report 2023. *arXiv preprint arXiv:2310.03715*, pages 1–386, 2023.

[153] G. Mateo-Garcia, J. Veitch-Michaelis, L. Smith, S. V. Oprea, G. Schumann, Y. Gal, A. G. Baydin, and D. Backes. Towards global flood mapping onboard low cost satellites with machine learning. *Scientific Reports*, 11(1):7249, 2021.

[154] H. Mattern, A. Sciarra, F. Godenschweger, D. Stucht, F. Lüsebrink, G. Rose, and O. Speck. Prospective motion correction enables highest resolution time-of-flight angiography at 7t. *Magnetic Resonance in Medicine*, 80(1):248–258, 2018.

[155] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

[156] T. Miller. Explanation in artificial intelligence: Insights from the

social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[157] S. Mohagheghi and A. H. Foruzan. Developing an explainable deep learning boundary correction method by incorporating cascaded x-Dim models to improve segmentation defects in liver CT images. *Computers in Biology and Medicine*, 140:105106, 2022.

[158] C. Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Independently published, 2022.

[159] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso. INbreast: Toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248, 2012.

[160] Morning Consult. IBM global AI adoption index - enterprise report, 2023. Available Online at: `https://filecache.mediaroom.com/mr5mr_ibmspgi/179414/download/IBM%20Global%20AI%20Adoption%20Index%20Report%20Dec.%202023.pdf/`, Last Accessed on April 24, 2024.

[161] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[162] S. Mullan. *Deep Learning and Explainable AI in Medical Image Segmentation*. PhD thesis, The University of Iowa, 2023.

[163] I. M. Nasir, S. Tehsin, R. Damaševičius, and R. Maskeliūnas. Integrating explanations into cnns by adopting spiking attention block for skin cancer detection. *Algorithms*, 17(12):557, 2024.

[164] G. Nguyen, D. Kim, and A. Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pages 26422–26436, 2021.

[165] T. Okamoto, C. Gu, J. Yu, and C. Zhang. Generating smooth interpretability map for explainable image segmentation. In *Proceedings of the IEEE Global Conference on Consumer Electronics*, pages 1023–1025, 2023.

[166] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

[167] J. I. Orlando, H. Fu, J. B. Breda, K. Van Keer, D. R. Bathula, A. Diaz-

Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59:101570, 2020.

[168] A. Pereira and C. Thomas. Challenges of machine learning applied to safety-critical cyber-physical systems. *Machine Learning and Knowledge Extraction*, 2(4):579–602, 2020.

[169] V. Petsiuk, A. Das, and K. Saenko. RISE: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, pages 1–17, 2018.

[170] E. Pintelas and I. E. Livieris. XSC—An explainable image segmentation and classification framework: A case study on skin cancer. *Electronics*, 12(17):3551, 2023.

[171] B. Qian, J. Su, Z. Wen, D. N. Jha, Y. Li, Y. Guan, D. Puthal, P. James, R. Yang, A. Y. Zomaya, et al. Orchestrating the development lifecycle of machine learning-based iot applications: A taxonomy and survey. *ACM Computing Surveys (CSUR)*, 53(4):1–47, 2020.

[172] P. Radau, Y. Lu, K. Connelly, G. Paul, A. J. Dick, and G. A. Wright. Evaluation framework for algorithms segmenting short axis cardiac MRI. *The MIDAS Journal*, 2009.

[173] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[174] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[175] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.

[176] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[177] D. Rymarczyk, Ł. Struski, M. Górszczak, K. Lewandowska, J. Tabor, and B. Zieliński. Interpretable image classification with differentiable prototypes assignment. In *Proceedings of the European*

Conference on Computer Vision, pages 351–368, 2022.

[178] M. Sacha, D. Rymarczyk, Ł. Struski, J. Tabor, and B. Zieliński. ProtoSeg: Interpretable semantic segmentation with prototypical parts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1481–1492, 2023.

[179] G. Saha and K. Roy. Online continual learning with saliency-guided experience replay using tiny episodic memory. *Machine Vision and Applications*, 34(4):65, 2023.

[180] H. Saleem, A. R. Shahid, and B. Raza. Visual interpretability in 3d brain tumor segmentation network. *Computers in Biology and Medicine*, 133:104410, 2021.

[181] A. Santamaria-Pang, J. Kubricht, A. Chowdhury, C. Bhushan, and P. Tu. Towards emergent language symbolic semantic segmentation and model interpretability. In *Proceedings of the MICCAI Conference, Lima, Peru, October 4–8, Part I 23*, pages 326–334, 2020.

[182] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S. Q. Truong, C. D. Nguyen, V.-D. Ngo, J. Seekins, F. G. Blankenberg, A. Y. Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022.

[183] C. Schorr, P. Goodarzi, F. Chen, and T. Dahmen. Neuroscope: An explainable ai toolbox for semantic segmentation and image classification of convolutional neural nets. *Applied Sciences*, 11(5): 2199, 2021.

[184] K. Schulze, F. Peppert, C. Schütte, and V. Sunkara. Chimeric U-net–modifying the standard U-net towards explainability. *bioRxiv*, pages 1–12, 2022.

[185] G. Schwalbe and B. Finzel. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59, 2023.

[186] D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1177–1178, 2010.

[187] C. Seibold, J. Künzel, A. Hilsmann, and P. Eisert. From explanations to segmentation: Using explainable AI for image segmentation. *arXiv preprint arXiv:2202.00315*, pages 1–10, 2022.

[188] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[189] A. Shahroudnejad. A survey on understanding, visualizations, and explanation of deep neural networks. *arXiv preprint arXiv:2102.01792*, 2021.

[190] E. H. Shortliffe and B. G. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3-4):351–379, 1975.

[191] H. Shreim, A. K. Gizzini, and A. J. Ghandour. Trainable noise model as an XAI evaluation method: application on sobol for remote sensing image segmentation. *arXiv preprint arXiv:2310.01828*, pages 1–7, 2023.

[192] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, pages 1–8, 2013.

[193] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, pages 1–15, 2019.

[194] D. Singh, A. Somani, A. Horsch, and D. K. Prasad. Counterfactual explainable gastrointestinal and colonoscopy image segmentation. In *Proceedings of the IEEE 19th International Symposium on Biomedical Imaging*, pages 1–5, 2022.

[195] J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, A. S. Tabish, et al. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomedical Imaging Data Papers*, 2(1):1004, 2015.

[196] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, pages 1–10, 2017.

[197] T. Speith. A review of taxonomies of explainable artificial intelligence (XAI) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, 2022.

[198] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, pages 1–14, 2014.

[199] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Trans-

former for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[200] J. Sun, F. Darbehani, M. Zaidi, and B. Wang. SAUNet: Shape attentive U-net for interpretable medical image segmentation. In *Proceedings of the MICCAI Conference, Lima, Peru, Part IV 23*, pages 797–806, 2020.

[201] R. Sun and M. Rostami. Explainable artificial intelligence architecture for melanoma diagnosis using indicator localization and self-supervised learning. *arXiv preprint arXiv:2303.14615*, pages 1–16, 2023.

[202] W. Swartout, C. Paris, and J. Moore. Explanations in knowledge systems: Design for explainable expert systems. *IEEE Expert*, 6(3): 58–64, 1991.

[203] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[204] S. Tehsin, I. M. Nasir, R. Damaševičius, and R. Maskeliūnas. Dasam: Disease and spatial attention module-based explainable model for brain tumor detection. *Big Data and Cognitive Computing*, 8(9):97, 2024.

[205] The White House. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, 2023. Available online at: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. Last accessed: April 22, 2024.

[206] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6021–6029, 2020.

[207] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Post-mortem iris recognition resistant to biological eye decay processes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2307–2315, 2020.

[208] P. Tschandl, C. Rosendahl, and H. Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018.

[209] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[210] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, A. Courville, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, pages 1–9, 2017.

[211] K. Vinogradova. *Explainable Artificial Intelligence for Image Segmentation and for Estimation of Optical Aberrations*. PhD thesis, Dresden University of Technology, Germany, 2023.

[212] K. Vinogradova, A. Dibrov, and G. Myers. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13943–13944, 2020.

[213] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:841, 2017.

[214] A. Wan, D. Ho, Y. Song, H. Tillman, S. A. Bargal, and J. E. Gonzalez. SegNBDT: Visual decision rules for segmentation. *arXiv preprint arXiv:2006.06868*, pages 1–15, 2020.

[215] C. Wang, X. Gao, and X. Li. An interpretable deep Bayesian model for facial micro-expression recognition. In *Proceedings of the IEEE International Conference on Control and Robotics Engineering*, pages 91–94, 2023.

[216] J. Wang, Y. Zheng, J. Ma, X. Li, C. Wang, J. Gee, H. Wang, and W. Huang. Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation. *Medical Image Analysis*, 83:102687, 2023.

[217] K. Wang, S. Yin, Y. Wang, and S. Li. Explainable deep learning for medical image segmentation with learnable class activation mapping. In *Proceedings of the 2023 2nd Asia Conference on Algorithms, Computing and Machine Learning*, pages 210–215, 2023.

[218] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[219] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical Image Analysis*, 60:

101619, 2020.

[220] X. Wu, Z. Li, C. Tao, X. Han, Y.-W. Chen, J. Yao, J. Zhang, Q. Sun, W. Li, Y. Liu, et al. DEA: Data-efficient augmentation for interpretable medical image segmentation. *Biomedical Signal Processing and Control*, 89:105748, 2024.

[221] M. Xian, Y. Zhang, H.-D. Cheng, F. Xu, K. Huang, B. Zhang, J. Ding, C. Ning, and Y. Wang. *A benchmark for breast ultrasound image segmentation (BUSIS)*. Infinite Study, 2018.

[222] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1369–1378, 2017.

[223] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9523–9532, 2020.

[224] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 32, pages 1–12, 2019.

[225] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2020.

[226] L. Yu, W. Xiang, J. Fang, Y.-P. P. Chen, and L. Chi. eX-ViT: A novel explainable vision transformer for weakly supervised semantic segmentation. *Pattern Recognition*, 142:109666, 2023.

[227] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, pages 818–833, 2014.

[228] R. A. Zeineldin, M. E. Karar, Z. Elshaer, ·. J. Coburger, C. R. Wirtz, O. Burgert, and F. Mathis-Ullrich. Explainability of deep neural networks for MRI analysis of brain tumors. *International Journal of Computer Assisted Radiology and Surgery*, 17(9):1673–1683, 2022.

[229] M. Zemni, M. Chen, É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord. OCTET: Object-aware counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-*

tern Recognition, pages 15062–15071, 2023.

[230] M. Zemni, M. Chen, E. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord. OCTET: Object-aware counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15062–15071, 2023.

[231] J. Zhang, R. Gu, P. Xue, M. Liu, H. Zheng, Y. Zheng, L. Ma, G. Wang, and L. Gu. S3R: Shape and semantics-based selective regularization for explainable continual segmentation across multiple sites. *IEEE Transactions on Medical Imaging*, pages 1–13, 2023.

[232] X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020.

[233] Y. Zhang, S. Mehta, and A. Caspi. Rethinking semantic segmentation evaluation for explainability and model selection. *arXiv preprint arXiv:2101.08418*, pages 1–14, 2021.

[234] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.

[235] Z. Zhang, P. Angelov, E. Soares, N. Longepe, and P. P. Mathieu. An interpretable deep semantic segmentation method for Earth observation. In *Proceedings of the IEEE International Conference on Intelligent Systems*, pages 1–8, 2022.

[236] Z. Zhang, Z. Wang, and I. Joe. CAM-NAS: An efficient and interpretable neural architecture search model based on class activation mapping. *Applied Sciences*, 13(17):9686, 2023.

[237] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[238] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127: 302–321, 2019.

[239] Y. Zhou, S. Booth, M. T. Ribeiro, and J. Shah. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9623–9633, 2022.

[240] X. Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2933–2946, 2018.

[241] X. Zhuang and J. Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Medical Image Analysis*, 31:77–87, 2016.

[242] J. Zolfaghari Bengar, A. Gonzalez-Garcia, G. Villalonga, B. Raducanu, H. Habibi Aghdam, M. Mozerov, A. M. Lopez, and J. Van de Weijer. Temporal coherence for active learning in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1–10, 2019.

LIST OF AUTHOR PUBLICATIONS

Below is a list of articles in international research journals with a citation index in the Clarivate Analytics Web of Science (CA WoS) database, as well as other publications.

Articles in CA WoS journals:

[A.1] Gipiškis, R., Chiaro, D., Preziosi, M., Prezioso, E. and Piccialli, F., 2023. The impact of adversarial attacks on interpretable semantic segmentation in cyber–physical systems. *IEEE Systems Journal*, 17(4), pp. 5327-5334. https://doi.org/10.1109/JSYST.2023.3281079

[A.2] Gipiškis, R., Tsai, C.W. and Kurasova, O., 2024. Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey. *ICT Express.*, 10(6), pp. 1331-1354. https://doi.org/10.1016/j.icte.2024.09.008

Other publications:

[B.1] Gipiškis, R. and Kurasova, O., 2023, June. Occlusion-based approach for interpretable semantic segmentation. In *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE. https://doi.org/10.23919/CISTI58278.2023.10212017

[B.2] Gipiškis, R., Chiaro, D., Annunziata, D. and Piccialli, F., 2023, July. Ablation Studies in Activation Maps for Explainable Semantic Segmentation in *Industry 4.0. In IEEE EUROCON 2023-20th International Conference on Smart Technologies* (pp. 36-41). IEEE. https://doi.org/10.1109/EUROCON56442.2023.10199094

[B.3] Gipiškis, R., 2024. XAI-driven Model Improvements in Interpretable Image Segmentation. xAI-2024 Late-breaking work, demos and doctoral consortium joint proceedings, Valletta, Malta, July 17-19, 2024., pp. 369-376.

[B.4] Gipiškis, R., Joaquin, A.S., Chin, Z.S., Regenfuß, A., Gil, A. and Holtman, K., 2024. Risk Sources and Risk Management Measures in Support of Standards for General-Purpose AI Systems. arXiv preprint arXiv:2410.23472. https://doi.org/10.48550/arXiv.2410.23472

[B.5] Stelling, L., Yang, M., Gipiškis, R., Staufer, L., Chin, Z.S., Campos, S. and Chen, M., 2025. Existing Industry Practice for the EU AI Act's General-Purpose AI Code of Practice Safety and Security Measures. arXiv preprint arXiv:2504.15181. https://doi.org/10.48550/arXiv.2504.15181

CURRICULUM VITAE

Rokas Gipiškis received a Bachelor's degree in Philosophy with a minor in Computer Science from Vytautas Magnus University in 2018 and a Master's degree in Computer Science from Vilnius University in 2020. From 2020 to 2024, he was a Ph.D. student at Vilnius University.

SUMMARY IN LITHUANIAN

Per pastarąjį dešimtmetį dirbtinio intelekto (DI) sistemos pasiekė reikšmingų rezultatų, ypač natūralios kalbos apdorojimo ir kompiuterinės regos srityse. Šių sistemų veikimas įprastai matuojamas įverčio metrikomis, kurios skiriasi priklausomai nuo sprendžiamo uždavinio. Šiuo metu pažangiausios DI sistemos daugiausia remiasi giliojo mokymosi modeliais – daugiasluoksniais neuroniniais tinklais, sudarytais iš vis didesnio modelio parametrų skaičiaus. Dėl didėjančio savo sudėtingumo neretai šios sistemos įvardijamos kaip „juodosios dėžės“. Tuo pabrėžiama, kad įverčio metrikos neatskleidžia visumos: net jei išvesties duomenys yra teisingi, jie nesuteikia informacijos apie vidinį modelio veikimą.

Paaiškinamojo DI sritis apima įvairius metodus, kuriais siekiama paaiškinti modelio vidinį veikimą, jo rezultatus arba padaryti visą sistemą suprantamesnę galutiniams vartotojams ir sprendimų priėmėjams. Šiuo metu tebevyksta diskusijos dėl paaiškinamojo DI srities terminologijos. Tokios sąvokos kaip „interpretuojamumas“, „paaiškinamumas“, „supratimas“ ir „patikimumas“ yra sunkiai apibrėžiamos. Kai kurie autoriai vartoja „interpretuojamas“ ir „paaiškinamas“ sinonimiškai [156], o kiti – skirtingai [175, 176]. Kai šie terminai nėra vartojami sinonimiškai, įprastai daroma skirtis tarp post-hoc paaiškinimų, taikomų jau apmokytam modeliui, ir architektūriškai interpretuojamų modelių [176]. Tokiu būdu interpretuojamumas siejamas su paties modelio skaidrumu (angl. *transparency*) ir priklauso nuo to, kaip lengvai galima suprasti jo veikimą. Pavyzdžiui, nesudėtingas sprendimų medžiu paremtas modelis gali būti lengviau interpretuojamas nei giliojo mokymosi modelis, sudarytas iš milijonų parametrų. O štai paaiškinamumas dažnai apsiriboja modelio rezultatų, o ne viso modelio paaiškinimu.

Šioje disertacijoje terminai „interpretuojamas“ ir „paaiškinamas“ vartojami kaip sinonimai, o konkretesni terminai „architektūra pagrįsti“ ir „architektūriškai interpretuojami“ vartojami aptariant konkretaus modelio paaiškinamumo modifikacijas. Taip yra todėl, kad tik nedidelėje apžvelgtos literatūros dalyje paaiškinamumo terminas vartojamas antrąja prasme. Kadangi dauguma paaiškinamo segmentavimo darbų neakcentuoja šio skirtumo, tai gali padėti išvengti nereikalingos painiavos apžvelgiant jų turinį.

Vaizdų klasifikavimo srityje jau yra pasiūlyta įvairių paaiškinamumo metodų, o semantinio segmentavimo srityje pastebimas jų trūkumas.

Nauji vaizdų segmentavimo paaiškinamumo metodai vis dar kuriami, ir trūksta tyrimų, kuriuose būtų nagrinėjama paaiškinamojo segmentavimo ir DI saugumo (angl. *AI safety*) sankirta, ypač įvertinant tokių metodų atsparumą priešiškoms atakoms. Šioje disertacijoje siekiama išplėsti šiuo metu ribotą paaiškinamojo DI metodų skaičių vaizdų segmentavimo srityje. Disertacijoje taip pat tiriamos priešiškos atakos, nukreiptos prieš paaiškinamąjį segmentavimą. Taip pat pateikiama išsami paaiškinamojo DI literatūros apžvalga vaizdų segmentavimo srityje, siūloma metodų taksonomija ir aptariama, kaip siūlomi paaiškinamumo metodai galėtų prisidėti prie paaiškinamuoju DI paremtų modelių tobulinimo.

Tyrimo objektas

Tyrimo objektas – giliojo mokymosi modelių *post hoc* paaiškinamumo metodai semantinio vaizdų segmentavimo uždaviniui.

Tyrimo tikslas ir uždaviniai

Tyrimo tikslas – sukurti naujus paaiškinamojo segmentavimo metodus, tinkamus konvoliuciniams neuroniniams tinklams, ir įvertinti jų pažeidžiamumą priešiškoms atakoms. Šiam tikslui pasiekti keliami šie uždaviniai:

- Ištirti esamus vaizdų interpretuojamumo metodus klasifikavimo ir segmentavimo uždaviniuose, nustatant tinkamiausius sprendimus konvoliuciniams neuroniniams tinklams. Remiantis šiuo tyrimu parengti išsamią paaiškinamojo DI metodų apžvalgą ir taksonomiją vaizdų segmentavimo srityje.
- Pasiūlyti ir pritaikyti naujus paaiškinamojo DI metodus (pavyzdžiui, įvesties vaizdo uždengimais, aktyvacijų perturbacijomis ir gradientais paremtus metodus), tinkamus segmentavimo uždaviniui, juos įvertinant kokybiškai ir kiekybiškai.
- Ištirti interpretuojamuosius semantinio segmentavimo metodus priešiškų atakų kontekste, įvertinant jų gynybines galimybes ir atsparumą priešiškoms atakoms.

Mokslinis darbo naujumas

1. Šioje disertacijoje pateikiama išsami vaizdų segmentavimo paaiškinamumo metodų apžvalga, neapsiribojanti viena paaiškinamumo metodų rūšimi ar taikymo sritimi, ir pristatoma išsami paaiškinamojo segmentavimo metodų taksonomija.

2. Pasiūlomas Ablation-CAM [62], plačiai klasifikavime naudojamo paaiškinamumo metodo, pritaikymas semantinio vaizdų segmentavimo uždaviniui.

3. Pateikiamas sisteminis perturbacijų įvesties erdvėje tyrimas, įvertinantis skirtingų įvesties vaizdo uždengimų filtrų dydžių ir spalvų poveikį modelių išvestims, atsižvelgiant į kokybines ir kiekybines metrikas.

4. Parodoma, kad galima sėkmingai sukonstruoti priešišką ataką, nukreiptą prieš post-hoc paaiškinamumo metodus semantiniame segmentavime, išplečiant pirmąjį šios srities tyrimą [65] vaizdų klasifikavime.

Praktinė darbo vertė

Paaiškinamasis vaizdų segmentavimas yra palyginti nauja sritis – pirmieji straipsniai šia tema pasirodė 2019–2020 metais [107, 212, 219]. Ši tyrimo sritis sulaukia vis daugiau dėmesio, o pats semantinis segmentavimas yra kertinis kompiuterinės regos uždavinys, kurio taikymas apima įvairias sritis: nuo autonominių automobilių [76] iki medicininių vaizdų analizės [17]. Dėmesys vaizdų segmentavimo paaiškinamumo metodams yra susijęs ir su augančiu patikimo DI (angl. *trustworthy AI*) poreikiu ypatingos svarbos taikymo srityse. Patobulintas post-hoc paaiškinamumas gali pagerinti naudotojų pasitikėjimą ir palengvinti teisės aktų laikymąsi, ypač medicinoje ar svarbiose pramonės šakose. Išsami paaiškinamumo metodų naudojimo priešiškų atakų scenarijuose analizė gali padėti geriau įvertinti ir suvaldyti su priešiškomis atakomis susijusią riziką, taip prisidedant prie saugių DI sprendimų kūrimo. Šioje disertacijoje pasiūlyta taksonomija ir metodai gali suteikti tyrėjams įrankių sistemingai įvertinti ir diegti paaiškinamumo metodus, mažinant atotrūkį tarp teorinės pažangos ir realaus taikymo.

Ginamieji teiginiai

- Perturbacijomis paremti post-hoc paaiškinamumo metodai, taikomi tiek įvesties, tiek aktyvacijos erdvėse, yra tinkami semantinio segmentavimo modelių rezultatams aiškinti.
- Post-hoc paaiškinamieji segmentavimo metodai yra pritaikomi priešiškų atakų kontekste tiek identifikuojant, tiek įgalinant priešiškas atakas.

Tyrimo rezultatų aprobavimas

Straipsniai tarptautiniuose moksliniuose žurnaluose su cituojamumo indeksu *Clarivate Analytics Web of Science* (*CA WoS*) duomenų bazėje:

1. Gipiškis, R., Chiaro, D., Preziosi, M., Prezioso, E. and Piccialli, F., 2023. The impact of adversarial attacks on interpretable semantic segmentation in cyber–physical systems. *IEEE Systems Journal*, 17(4), pp. 5327-5334. `https://doi.org/10.1109/JSYST.2023.3281079`
2. Gipiškis, R., Tsai, C.W. and Kurasova, O., 2024. Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey. *ICT Express.*, 10(6), pp. 1331-1354. `https://doi.org/10.1016/j.icte.2024.09.008`

Pranešimai tarptautinėse konferencijose:

1. Gipiškis, R. and Kurasova, O., Occlusion-based approach for interpretable semantic segmentation. *18th Iberian Conference on Information Systems and Technologies (CISTI)*, June 20–23, 2023, Aveiro, Portugal.
2. Gipiškis, R., XAI-driven Model Improvements in Interpretable Image Segmentation. *2nd World Conference on eXplainable Artificial Intelligence*, July 17–19, 2024, Valletta, Malta.

Pranešimai nacionalinėse konferencijose:

1. Gipiškis, R. and Kurasova, O., Application of CNNs for brain MRI image segmentation. *12th Conference on Data Analysis Methods for Software Systems*, December 2–4, 2021, Druskininkai, Lithuania.

2. Gipiškis, R. and Kurasova, O., Investigating post-hoc explainability techniques for image segmentation. *15th Conference on Data Analysis Methods for Software Systems*, November 28-30, 2024, Druskininkai, Lithuania.

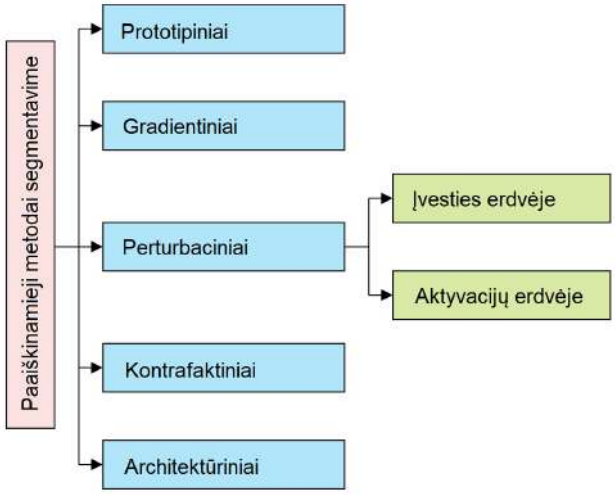
Disertacijos struktūra

Disertaciją sudaro įvadas, keturi skyriai, išvados ir santrauka lietuvių kalba. Įvade apžvelgiamas paaiškinamasis DI, pabrėžiant jo svarbą semantiniam vaizdų segmentavimui, pristatomas tyrimo objektas, tikslas, uždaviniai, mokslinis naujumas, praktinė reikšmė ir ginamieji teiginiai. Pirmajame skyriuje apžvelgiama paaiškinamojo DI srities mokslinė literatūra, raida kompiuterinėje regoje, pristatoma paaiškinamojo semantinio segmentavimo metodų taksonomija ir pabrėžiami esamų metodų trūkumai. Antrajame skyriuje pristatomi rengiant šią disertaciją sukurti metodai, įskaitant perturbacijomis ir gradientais paremtus bei priešiškų atakų generavimo semantinio segmentavimo kontekste metodus, išsamiai aprašomi jų teoriniai pagrindai ir praktinis įgyvendinimas. Trečiajame skyriuje pateikiamas eksperimentinis pasiūlytų metodų vertinimas, daugiausia dėmesio skiriant paaiškinamumo, skaičiavimo efektyvumo ir atsparumo priešiškiems scenarijams rezultatams. Ketvirtajame skyriuje aptariami paaiškinamojo DI semantinio segmentavimo srityje uždaviniai, vertinami paaiškinamumo ir patikimumo kompromisai ir siūlomos būsimų mokslinių tyrimų kryptys, įskaitant hibridinius metodus. Disertacijos pabaigoje pateikiamos pagrindinės išvados, praktinio taikymo prielaidos ir rekomendacijos tolesniems tyrimams šioje srityje. Disertaciją sudaro 153 puslapiai, 40 paveikslėlių ir keturios lentelės.

S.1. Paaiškinamojo DI vaizdų segmentavimo srityje literatūros apžvalga

S.1.1. Taksonomija

Paaiškinamumas mašininio mokymosi kontekste įprastai yra skirstomas į skaidrius modelius ir post-hoc paaiškinamumo metodus, kurie savo ruožtu skirstomi į konkrečiam modeliui būdingas ir nuo modelio nepriklausomas kategorijas [16]. Priklausomai nuo norimo abstrakcijos lygio, galima naudoti kelias viena su kita suderinamas taksonomijas. Paaiškinamieji metodai gali būti skirstomi į post-hoc ir ad-hoc paaiškinamumo metodus [73]. Bendresnėje taksonomijoje [189] išskiriama struktūrinė analizė, elgsenos analizė ir nuo dizaino priklausomas paaiškinamumas. Paaiškinamojo DI taksonomijų analizė klasifikavimo srityje rodo, kad jos taip pat galėtų būti pritaikytos vaizdų segmentavimo srityje. Visgi iki šiol nebuvo pristatyta jokia taksonomija, skirta vis augančiai interpretuojamojo segmentavimo sričiai.



S.1 pav.: Paaiškinamojo DI metodai semantiniame segmentavime.

Šioje disertacijoje siūloma taksonomija (S.1 pav.) remiasi apžvelgta paaiškinamojo vaizdo segmentavimo literatūra. Siūloma taksonomija apima penkias metodų grupes: prototipais, gradientais ir perturbacijomis paremtus metodus, kontrafaktinius metodus ir nuo architektūros priklausančius metodus. Remiantis anksčiau aptartomis

bendresnėmis taksonomijomis, įprastai siūlomose kitose apžvalgose, dauguma metodų patenka į lokalaus ir post-hoc paaiškinamumo kategorijas. Prototipais besiremiantys metodai naudoja reprezentatyvius pavyzdžius ar jų dalis iš duomenų rinkinio, kuriuos analizuoja ir lygina su įvesties vaizdu. Gradientais paremti metodai apima pasirinkto tinklo sluoksnio arba dominančios klasės išvesties gradiento skaičiavimą pasirinktų įvesties duomenų arba požymių žemėlapių atžvilgiu. Perturbacijomis paremtus metodus galima suskirstyti į dvi grupes, priklausomai nuo perturbacijų erdvės. Perturbacijos įvesties erdvėje – tai iteraciniai įvesties vaizdo uždengimai. Dažniausiai šioms vaizdo transformacijoms naudojamas slankusis filtras, tačiau taip pat gali būti naudojami ir įvairių tipų triukšmai. Perturbacijos metodų generuojami paaiškinimai remiasi perturbacijų poveikiu modelio rezultatams. Perturbacijoms aktyvacijos erdvėje naudojamas dalinis arba visiškas pasirinkto modelio sluoksnio aktyvacijos žemėlapių (angl. *activation maps*) deaktyvavimas. Kaip ir perturbacijų įvesties erdvėje atveju, paaiškinimai remiasi perturbacijų poveikiu modelio rezultatams. Kontrafaktiniai metodai siekia identifikuoti mažiausius įvesties pokyčius, kurių reikia, kad pasikeistų modelio išvestis. Nuo architektūros priklausantys metodai apima papildomas architektūrines modifikacijas, atliekamas dar prieš modelio apmokymą arba jo metu, siekiant pagerinti paaiškinamumą.

S.2. Paaiškinamojo DI metodai vaizdų segmentavime

S.2.1. Perturbacijos įvesties erdvėje

Perturbaciniai paaiškinamojo DI metodai įvesties erdvėje remiasi sistemingu įvesties vaizdo dalių uždengimu, o paaiškinimai generuojami išmatavus, kaip šie uždengimai veikia modelio išvestį. Remiantis [227], pasiūlytas metodas matuoja uždengimų poveikį Dice koeficientui arba *Intersection over Union* (*IoU*) metrikai. Kaip ir klasifikavimo atveju, įvesties vaizdas dalinai uždengiamas juo sistematiškai praslenkant filtrą ir matuojant šio uždengimo poveikį Dice koeficientui iš anksto pasirinktai segmentavimo klasei. Atliekant eksperimentus tiriami filtrų dydžiai ir spalvos gali skirtis. Pavyzdžiui, gali būti naudojami pilki, juodi arba Gauso uždengimo filtrai.

Uždengimo procesas prasideda nuo įprasto vaizdo apdorojimo etapo, apimančio normalizavimą ir, esant ribotiems skaičiavimo ištekliams,

vaizdo dydžio keitimą. Tada pasirinktas filtro tipas palaipsniui slenkamas išilgai viso vaizdo. Galima naudoti ir mažesnio dydžio poslinkio reikšmę, tačiau tada filtrai persidengtų. Kiekvienas naujai uždengtas įvesties vaizdas yra leidžiamas pro pasirinktą konvoliucinį neuroninį tinklą. Gauta segmentavimo išvestis naudojama modelio įverčio reikšmei apskaičiuoti. Rezultatas išsaugomas ir procesas kartojamas tol, kol uždengimo filtras yra visiškai praslenkamas pro visą įvesties vaizdo sritį.

S.2.2. Perturbacijos aktyvacijos erdvėje

Siekiant pritaikyti perturbacijomis aktyvacijos erdvėje (dar žinomomis kaip abliacijomis) paremtus metodus paaiškinamajam segmentavimui, perturbacijos poveikis dominančiai klasei apskaičiuojamas pagal jos poveikį *logits* reikšmėms, gautoms sudėjus kiekvieno pikselio daugumos klasės (*argmax*) *logits* reikšmes. Siūlomoje modifikacijoje kaupiami tik tų pikselių, kurie buvo klasifikuoti kaip priklausantys c klasei, c klasės *logits* reikšmės.

Turint RGB vaizdą $x \in \mathbb{R}^{N \times M \times 3}$, vieno pikselio x_{ij} *logits* reikšmė dominančiai klasei c apibrėžiama kaip $l^c(x_{ij})$. Tada *logits* suma c, priklausanti nuo to, ar x_{ij} klasifikuojamas kaip c, yra:

$$L^c(x) = \sum_{i,j} [\hat{c}_{ij} = c] l^c(x_{ij}), \tag{S.1}$$

kur $\hat{c}_{ij}$ yra modelio prognozuojama x_{ij} pikselio klasė.

Šios sukauptos klasių *logits* reikšmės naudojamos kiekvieno aktyvacijos žemėlapio svarbos koeficientui arba svoriui apskaičiuoti. Remiantis [62], kiekvieno aktyvacijos vieneto w_k^c svarbos koeficientas w_k^c apibrėžiamas kaip:

$$w_k^c = \frac{L^c(x) - L_k^c(x)}{L^c(x)}, \tag{S.2}$$

kur $L_k^c(x)$ yra klasės c *logits* suma po aktyvacijos žemėlapio A_k abliacijos.

Apskaičiuotus aktyvacijos žemėlapių svarbos svorius galima naudoti tiesiškoje požymių žemėlapių kombinacijoje. Prieš pradedant aktyvacijos žemėlapių perturbaciją, reikia pasirinkti dominantį modelio sluoksnį.

S.2.3. Gradientiniai metodai priešiškose atakose

Gradientais paremti paaiškinamojo DI metodai buvo tiriami priešiškų atakų sąlygomis, siekiant nustatyti jų atsparumą. Norint geriau kontroliuoti vaizdo perturbacijos procesą ir išvengti pernelyg didelių iškraipymų priešiškuose įvesties vaizdo pavyzdžiuose, pasiūlyta nuostolio funkcija su papildoma komponente. Ši komponentė užtikrina, kad įvesties vaizdui taikomas priešiškas triukšmas būtų apribotas, išsaugant modelio segmentavimo rezultatus ir užpulto vaizdo panašumą į originalą. Toliau pateikiamas formalus pasiūlytos nuostolių funkcijos apibrėžimas.

Turint du vaizdus $x, y \in \mathbb{R}^{N \times M \times 3}$ ir dvi dominančias klases $c_1, c_2 \in \{1, 2, ..., C\}$, siekiama užpulti vaizdo x paaiškinamąjį žemėlapį $G_{A_1}(x, c_1)$, kad vietoje jo gautume vaizdo y paaiškinamąjį žemėlapį $G_{A_2}(y, c_2)$, kur A_i yra vaizdo sritis klasifikuojama kaip c_i, $\forall i = 1, 2$. Sykiu užpulto vaizdo x_{adv} segmentavimo išvestis $\overline{g}(x_{adv})$ turėtų išlikti panaši į segmentavimo išvestį $\overline{g}(x)$, o užpultas vaizdas neturėtų ženkliai skirtis nuo x. Pasiūlyta [A.1] nuostolio funkcija apibrėžiama kaip:

$$L = \gamma_1 L_{exp} + \gamma_2 L_{out} + \gamma_3 L_{im}, \tag{S.3}$$

kur:

$L_{exp} = ||G_{A_2}(y, c_2) - G_{A_1}(x_{adv}, c_1)||^2$,
$L_{out} = ||\overline{g}(x) - \overline{g}(x_{adv})||^2$,
$L_{im} = ||x - x_{adv}||^2$,

ir kur γ_1, γ_2, γ_3 yra parametrai, kontroliuojantys santykinę kiekvienos nuostolio funkcijos komponentės svarbą optimizavimo metu. L_{exp} matuoja atstumą tarp gauto užpulto vaizdo paaiškinimo ir siektino paaiškinimo (t. y. paaiškinimo, kurį norėtume gauti iš užpulto modelio), L_{out} matuoja atstumą tarp neužpulto ir užpulto modelio išvesties tam pačiam vaizdui, o L_{im} matuoja atstumą tarp neužpulto ir užpulto įvesties vaizdo.

S.3. Eksperimentiniai rezultatai

S.3.1. Eksperimentai įvesties erdvėje

Šiame poskyryje aprašomi eksperimentiniai tyrimai atlikti naudojant visiškai konvoliucinį neuroninį tinklą [145] su ResNet-101 [98] pagrin-

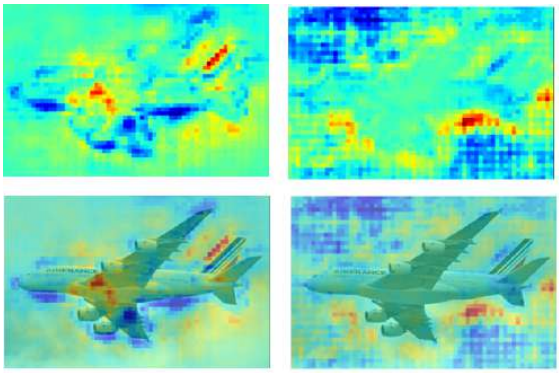
dine dalimi (angl. *backbone*) ir DeepLabV3 [38] modelį bei COCO [141] duomenų rinkinį.

Kaip ir klasifikavimo atveju, pastebėta, kad mažas uždengimo filtro dydis įprastai neturi didesnio poveikio modelio išvesčiai. Taip tikriausiai yra todėl, kad modelis geba atkurti vaizdą remdamasis kontekstine informacija iš likusios neuždengtos vaizdo dalies. Tačiau naudojant mažesnį uždengimo filtro dydį galima sugeneruoti detalesnį paaiškinimą. Kitaip nei klasifikavimo užduotyje, segmentavimas po uždengimų pasižymi nežymiu Dice koeficiento skirtumu, dažniausiai apsiribojančiu tik trečia vieta po kablelio. Jei koeficientai būtų naudojami paaiškinimams generuoti, skirtingos jų vaizdo sritys būtų beveik neatskiriamos viena nuo kitos.

Siekiant, kad reikšmės būtų labiau išsiskirsčiusios intervalo $[0, 1]$ ribose, o ne susitelkusios aplink konkrečią Dice koeficiento reikšmę, pritaikytas min-max normalizavimas. Normalizavimas leido padidinti surinktų koeficientų standartinį nuokrypį 2–3 dydžio eilėmis (angl. *orders of magnitude*). Todėl vizualizuojant paaiškinimus, pasiektas didesnis spalvų intensyvumo diapazonas. Taip pat išbandytas ir Z reikšmių standartizavimas atmetant neigiamas reikšmes prieš vaizdo vizualizavimą, tačiau daugeliu atveju aiškesni ir mažiau triukšmingi rezultatai gaunami pritaikius normalizavimą.

Paaiškinimams generuoti naudoti koeficientai taip pat apskaičiuoti naudojant *logits* reikšmes. Visos pasirinktai klasei priklausančios *logits* reikšmės susumuotos kiekvienam pikseliui, kuris klasifikuotas kaip priklausantis tai klasei. Gauta skaliarinė reikšmė buvo naudojama uždengimo poveikiui įvertinti. Gautos paaiškinimų vizualizacijos (S.2 pav.) buvo jautresnės lyginant su Dice koeficientu paremtomis vizualizacijomis.

Rezultatai kiekybiškai įvertinti naudojant ištrynimo kreives. Buvo tiriamas skirtingų uždengimo filtrų poveikis dviem įvesties vaizdams, kuriuose pagrindinio plano klasė ženkliai skyrėsi savo užimamu plotu. Svarbiausi įvesties vaizdo pikseliai buvo palaipsniui uždengiami, pradedant nuo 99-ojo procentilio pagal pirmiau apskaičiuotus svarbos koeficientus, o tada buvo matuojamas šių uždengimų poveikis išvesčiai. Mažesnė ploto po kreive (angl. *area under the curve, AUC*) reikšmė, atitinkanti staigesnį ištrynimo kreivės kritimą, yra siejama su paaiškinamumo metodo gebėjimu atskirti svarbiausius įvesties vaizdo požymius.



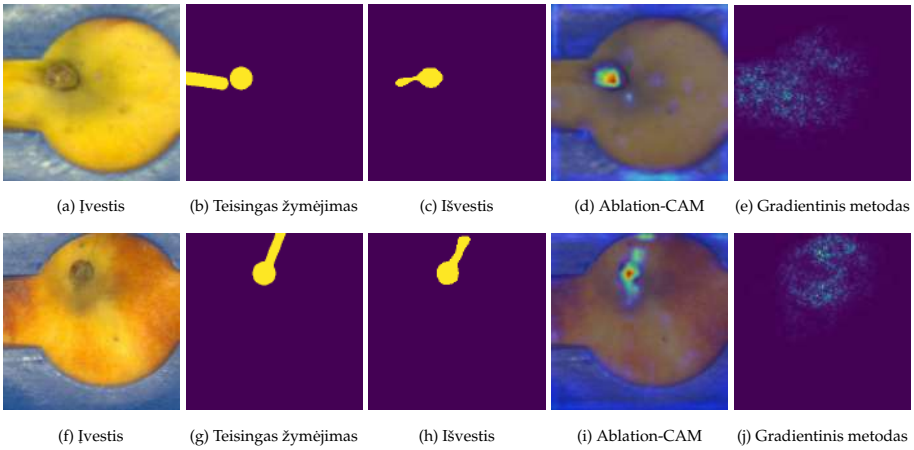
S.2 pav.: Paaiškinimai sugeneruoti naudojant 2752 10×10 dydžio uždengimo filtrus. Paaiškinimai lėktuvo klasei yra kairėje, o fono klasei – dešinėje.

Taip pat buvo tiriamas uždengimo filtro spalvos poveikis segmentavimo rezultatui. Pilki uždengimo filtrai davė geresnių rezultatų nei balti ar juodi, kai buvo naudojami didesni (50×50 dydžio) uždengimai. Daugeliu atvejų juodi uždengimo filtrai lemdavo prasčiausius rezultatus.

S.3.2. Eksperimentai aktyvacijų erdvėje

Šiame poskyryje aprašomi eksperimentiniai tyrimai atlikti naudojant U-Net [174] modelį su keturiais enkoderio sluoksniais ir visiškai konvoliucinį neuroninį tinklą [145] su ResNet-101 [98] pagrindine dalimi. Naudotas COCO [141] duomenų rinkinys ir privatus duomenų rinkinys iš pramoninius maisto apdorojimo aparatus gaminančios įmonės.

Ablation-CAM paaiškinamumo metodo rezultatai palyginti su segmentavimui pritaikytu gradientais paremtu metodu [192]. Gradientais paremtuose paaiškinimuose buvo pastebimai daugiau triukšmo. Nors svarbiausios įvesties vaizdo sritys buvo teisingai paryškintos, papildomai išryškėjo ir platesnė vaizdo sritis. O štai segmentavimui pritaikyto Ablation-CAM metodo sugeneruoti paaiškinimai buvo aiškesni ir mažiau išsisklaidę. Didžiausią įtaką modelio sprendimui turinčios sritys susitelkė ties vaisiaus kauliuku. Pavyzdžiui, išvesties su mažesniu segmentavimo tikslumu (S.3 (c) pav.) paaiškinimas daugiausia buvo sutelktas į ovalo formos kauliuką. O štai segmentavimo modelio prasčiau aptinkama vaisiaus pjovimo linija neturėjo matomų aktyvacijų.



(a) Įvestis (b) Teisingas žymėjimas (c) Išvestis (d) Ablation-CAM (e) Gradientinis metodas

(f) Įvestis (g) Teisingas žymėjimas (h) Išvestis (i) Ablation-CAM (j) Gradientinis metodas

S.3 pav.: Ablation-CAM ir gradientais paremto metodo palyginimas dviems įvesties vaizdams. Viršutinėje eilutėje pasirinktas vaizdas su prastesniu segmentavimo rezultatu (c). Sugeneravus paaiškinimą (d) su Ablation-CAM pastebima, kad segmentuojant šį vaizdą modeliui sunkiau aptikti vaisiaus pjovimo liniją.

Eksperimentiniai tyrimai atlikti ir su daliniais aktyvacijos žemėlapių sričių uždengimais, priklausomai nuo to, ar jos priklauso fono, ar pagrindinio plano klasei. Ankstesni tyrimai [125] klasifikavime parodė, kad fono klasės aktyvacijos žemėlapių sričių uždengimas turi mažesnį poveikį tinklo klasifikavimui lyginant su pagrindinio plano srities (t. y. mus dominančios klasės) uždengimais. Siekiant ištirti šiuos rezultatus semantinio segmentavimo kontekste, kiekviename iš keturių enkoderio bloko sluoksnių buvo atskirai uždengiami tiek fonas, tiek pjūvio linija, ir šių uždengimų poveikis modelio išvesčiai buvo vertinamas naudojant Dice koeficientą ir min-max normalizuotas *logits* reikšmes pasirinktai dominančiai klasei. Rezultatai apskaičiuoti naudojant skirtingas uždengimo slenksčio reikšmes t, kurios svyravo nuo 0,0 iki 1,0 pritaikant 0,1 žingsnį. Gauti rezultatai taip pat parodė ryškesnių pokyčių pagrindinio plano (t. y. pjūvio linijos) uždengimo atveju.

S.3.3. Eksperimentai su priešiškomis atakomis

Įprastai priešiškomis atakomis siekiama pakeisti modelio išvestį. Tačiau jų taikymą galima išplėsti paaiškinamajam DI siekiant užpulti paaiški-

namumo metodo generuojamą atsakymą konkrečiam modeliui taip, kad jo įvestis ir išvestis išliktų nepastebimai panašios į neužpulto modelio įvestį ir išvestį. Tokio tipo atakų tyrimas iki šiol apsiribojo vien klasifikavimo užduotimis. Šiame poskyryje nagrinėjamas jų išplėtimas semantiniam segmentavimui.

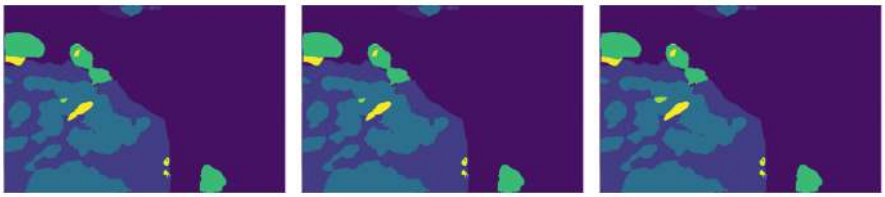
Lyginant su priešiškomis atakomis klasifikavime, norimam segmentavimo atakos rezultatui gauti reikia daugiau laiko. Pastebėta, kad net jei priešiška semantinio segmentavimo ataka sėkmingai užpuola modelio paaiškinimą ir išlaiko panašią modelio išvestį, dažnai yra pastebimai iškraipomos įvesties vaizdo sritys. Tokiu atveju priešiškos atakos nebūtų galima laikyti sėkminga.

Remiantis [65], įgyvendinta į paaiškinimus nukreipta priešiška ataka, stengiantis užtikrinti, kad modelio išvesties pokyčiai būtų kuo mažesni. Šiuo tikslu pasitelkta nuostolių funkcija, kontroliuojanti segmentavimo išvesties ir jos paaiškinimų pokyčius. Tokio tipo ataka tampa lengvesnė, jei leidžiami dideli įvesties vaizdo pokyčiai.

Kitaip nei [65], kur sėkmingai atakai pasiekti nebuvo būtina įvesties vaizdo triukšmo kontrolė (galbūt dėl santykinai lengvesnio tokių atakų taikymo klasifikavimo užduotyse), eksperimentai parodė, kad nekontroliuojamos įvesties vaizdo perturbacijos dažnai sugeneruoja pastebimai iškraipytus vaizdus. Šiai problemai spręsti įvesta trečioji nuostolių funkcijos komponentė (žr. poskyrį S.2.3), kuria kontroliuojamas priešiško triukšmo taikymas įvesties vaizdui.

Panašiai kaip ir [65], šiame tyrime susidurta su nykstančios antrosios išvestinės problema ReLU netiesiškumams. Siekiant išspręsti šią problemą, optimizavimo metu tinklo aktyvavimo funkcijos pakeistos iš ReLU į Softplus. Nuostolių funkcijai minimizuoti naudotas Adam optimizatorius. Praktiškai tokio tipo ataką būtų lengviau įgyvendinti mažesniuose vaizduose su mažesniu klasių skaičiumi. Dėl skaičiavimo apribojimų dronų segmentavimo duomenų rinkinio įvesties vaizdų dydis sumažintas iki 432×288. Remiantis eksperimentų rezultatais, toks vaizdo dydis leido pasiekti gerą kompromisą tarp segmentavimo išvesties ir atakos atlikimo laiko. Segmentuojant dronų duomenų rinkinį, augmenijos klasės paaiškinimas sėkmingai pakeistas į žmogaus klasės paaiškinimą (S.4 pav.).

(a) Pradinė įvestis (b) Siekiama įvestis (c) Užpulta įvestis



(d) Pradinė išvestis (e) Siekiama išvestis (f) Užpulta išvestis

(g) Pradinis paaiškinimas (h) Siekiamas paaiškinimas (i) Užpultas paaiškinimas

S.4 pav.: Sėkmingos priešiškos atakos nukreiptos prieš dronų rinkinio vaizdus pavyzdys.

Eksperimentams su persikų duomenų rinkiniu pasirinkti du skirtingi vaizdai, o ataka pakeitė vieno persiko pjovimo linijos paaiškinimą į kito persiko pjovimo linijos paaiškinimą. Pasirinkti priešiški paaiškinimo taikiniai nebuvo panašūs į originalių vaizdų paaiškinimus, o tai apsunkino ataką. Didesnis iteracijų skaičius įprastai lemia mažesnę MSE vertę tarp siekiamo užpuolimo taikinio ir užpulto paaiškinimo, rodančią didesnį vaizdų panašumą. Eksperimentai parodė, kad tam tikrais atvejais gana veiksmingą ataką galima įvykdyti atlikus vos 100 iteracijų. Siekiant ištirti kiekvienos nuostolių funkcijos komponentės tarpusavio įtaką MSE rezultatams, atlikta kiekybinė 30 atsitiktinai parinktų vaizdų analizė (žr. S.1 lentelę).

S.1 lentelė: Abliacijos tyrimas persikų ir dronų duomenų rinkiniams.

Lentelėje pateikiamos vidutinės MSE reikšmės ir jų standartiniai nuokrypiai.
Persikų rinkiniui:
$iteracijos = 500$, $mokymo\,greitis = 10^{-5}$, $\gamma_1 = 10^{11}$, $\gamma_2 = 10^5$, $\gamma_3 = 5 \cdot 10^6$;
Dronų rinkiniui:
$iteracijos = 500$, $mokymo\,greitis = 10^{-4}$, $\gamma_1 = 10^{11}$, $\gamma_2 = 5 \cdot 10^4$, $\gamma_3 = 10^6$.

	Persikų rinkinys	**Dronų rinkinys**
I $L = \gamma_1 L_{exp}$	$L_{exp} = (3{,}73 \pm 1{,}27) \cdot 10^{-9}$ $L_{out} = (6{,}91 \pm 8{,}25) \cdot 10^{-3}$ $L_{im} = (6{,}28 \pm 1{,}24) \cdot 10^{-6}$	$L_{exp} = (3{,}15 \pm 1{,}10) \cdot 10^{-10}$ $L_{out} = (3{,}58 \pm 1{,}33) \cdot 10^{-3}$ $L_{im} = (5{,}70 \pm 2{,}81) \cdot 10^{-4}$
I+II $L = \gamma_1 L_{exp} + \gamma_2 L_{out}$	$L_{exp} = (3{,}74 \pm 1{,}16) \cdot 10^{-9}$ $L_{out} = (1{,}03 \pm 1{,}17) \cdot 10^{-4}$ $L_{im} = (6{,}79 \pm 1{,}33) \cdot 10^{-6}$	$L_{exp} = (4{,}82 \pm 1{,}12) \cdot 10^{-10}$ $L_{out} = (2{,}51 \pm 1{,}20) \cdot 10^{-5}$ $L_{im} = (7{,}01 \pm 1{,}57) \cdot 10^{-5}$
I+III $L = \gamma_1 L_{exp} + \gamma_3 L_{im}$	$L_{exp} = (3{,}69 \pm 1{,}11) \cdot 10^{-9}$ $L_{out} = (6{,}29 \pm 7{,}45) \cdot 10^{-3}$ $L_{im} = (2{,}03 \pm 0{,}51) \cdot 10^{-6}$	$L_{exp} = (4{,}05 \pm 1{,}58) \cdot 10^{-10}$ $L_{out} = (5{,}05 \pm 1{,}54) \cdot 10^{-5}$ $L_{im} = (1{,}90 \pm 1{,}88) \cdot 10^{-5}$
I+II+III $L = \gamma_1 L_{exp} + \gamma_2 L_{out} + \gamma_3 L_{im}$	$L_{exp} = (3{,}79 \pm 1{,}16) \cdot 10^{-9}$ $L_{out} = (1{,}07 \pm 1{,}22) \cdot 10^{-4}$ $L_{im} = (1{,}98 \pm 0{,}45) \cdot 10^{-6}$	$L_{exp} = (5{,}59 \pm 1{,}38) \cdot 10^{-10}$ $L_{out} = (1{,}15 \pm 1{,}20) \cdot 10^{-5}$ $L_{im} = (6{,}22 \pm 2{,}35) \cdot 10^{-6}$

Persikų duomenų rinkinio MSE reikšmės yra mažiau išsklaidytos nei dronų duomenų rinkinio MSE reikšmės, o remiantis eksperimentų rezultatais, geri priešiškos atakos nuostolių parametrai vienam persikų duomenų rinkinio vaizdui nelabai skiriasi nuo kito. Tai galima paaiškinti tuo, kad visos modelio įvestys, išvestys ir paaiškinimai yra gana panašūs vieni į kitus, ypač lyginant su dronų duomenų rinkinio atitikmenimis. Kitas įtaką darantis veiksnys galėtų būti gerokai mažesnis segmentavimo klasių skaičius persikų duomenų rinkinyje.

S.4. Tolesni tyrimai

Paaiškinamojo semantinio segmentavimo srityje lieka daug neišspręstų uždavinių, kurių dauguma taip pat taikytini ir vaizdų klasifikavimo užduotims. Toliau pateikiamas nebaigtinis šių uždavinių sąrašas:

- **Paaiškinamojo DI įverčio metrikos**
  Paaiškinamojo DI vaizdų klasifikavimo srityje literatūroje daugiausia dėmesio skiriama naujiems paaiškinamumo metodams ir jų modifikacijoms pristatyti, o ne naujų įverčio metrikų ar paaiškinimų palyginimams siūlyti. Ši tendencija dar labiau išryškėja paaiškinamojo semantinio segmentavimo srityje. Šiuo metu nėra straipsnių, skirtų vien tik paaiškinamojo DI rezultatams vertinti segmentavimo kontekste. Nėra vieningos nuomonės dėl to, kurios įverčio metrikos yra svarbiausios pagrindiniams paaiškinamumo aspektams tirti. Iš dalies tai galima paaiškinti sunkumu apibrėžiant su paaiškinamumu susijusias sąvokas. Geresnis teorinis problemos supratimas galėtų padėti kurti paaiškinamojo DI įverčio metrikas.

- **Paaiškinamojo DI metodų saugumas ir atsparumas atakoms**
  Sparčiai diegiant giliojo mokymosi modelius medicinos, karinėje ir pramonės srityse, paaiškinamojo DI metodams tenka vis svarbesnis vaidmuo. Siekiama išsiaiškinti, ar naudojamas modelis yra patikimas. Tačiau panašų klausimą galima kelti ir dėl pačių paaiškinamojo DI metodų. Svarbu ištirti jų pažeidžiamumą ir esamas spragas. Tiek modelių tiekėjai, tiek galutiniai naudotojai turi žinoti, ar jie yra apsaugoti nuo tyčinių atakų, nukreiptų prieš paaiškinamojo DI metodus.

  Kaip ir klasifikavimo modeliai, semantinio segmentavimo modeliai gali būti pažeidžiami priešiškomis atakomis. Literatūroje jau yra pasiūlyta įvairių atakų metodų [47, 77, 222]. Aptariant priešiškas atakas, įprasta sutelkti dėmesį į modelio išvestį kaip pagrindinį taikinį. Tačiau taip pat galima užpulti išvesties paaiškinimą, įvestį ir išvestį paliekant nepakitusias. Tokios atakos jau ištirtos vaizdų klasifikavimo kontekste [65]. Taip pat parodyta, kad šias antrojo lygio atakas galima taikyti ir vaizdų segmentavimui [A.1]. Norint rasti geriausius kovos su jomis būdus, reikia atlikti daugiau tyrimų, ypač dėl to, kad nuolat kuriamos naujos priešiškos atakos,

o užtikrinti išsamias saugumo garantijas yra sudėtinga. Reikia sistemingai tirti tiek „baltosios dėžės“ atakas, kai užpuolikas žino atakuojamą modelį, tiek „juodosios dėžės“ atakas, kai modelis nežinomas. Panašūs interpretuojamo segmentavimo patikimumo tyrimai galėtų prisidėti prie bendro DI sistemų saugumo.

Priešiški pavyzdžiai paprastai nėra mokymo ir testavimo duomenų rinkinių dalis. Taigi į rinką paleistuose modeliuose gali atsirasti pažeidžiamumų. Kita svarbi problema – šališkumas. Kai svarbiausios paaiškinimų sritys patenka už dominančio objekto ribų, tai gali parodyti ne tik klaidingą prognozę, bet ir galimą priešiškos atakos atvejį [107]. Taip pat būtų galima ištirti natūralius priešiškus pavyzdžius [102] ir jų įtaką paaiškinamojo DI segmentavimui.

- **Paaiškinamasis DI vaizdo įrašams segmentuoti**
  Kadangi semantinis scenos segmentavimas neapsiriboja vien 2D vaizdais, būtų galima ištirti naujus interpretavimo metodus vaizdo duomenims, kai atliekamas laikinis semantinis segmentavimas. Vaizdo objektams segmentuoti reikia gerokai daugiau skaičiavimo išteklių. Iki šiol nė viename tyrime nebuvo nagrinėtas paaiškinamasis vaizdo segmentavimas dinaminėje aplinkoje. Dinaminių scenų pobūdis gali kelti naujų iššūkių, su kuriais anksčiau nebuvo susidurta 2D segmentavimo kontekste. Pavyzdžiui, norint atsižvelgti į paaiškinamumo žemėlapių skirtumus skirtinguose vaizdo kadruose, reikėtų pridėti papildomą laiko paaiškinimo ašį. Šią užduotį būtų galima dar labiau išplėsti iki realiuoju laiku atliekamo semantinio segmentavimo, sutelkiant dėmesį į būdus, kaip sumažinti generuojamų paaiškinimų vėlavimą.

- **Skaičiavimo kompleksiškumas**
  Skirtingiems paaiškinamojo DI metodams reikia skirtingų skaičiavimo išteklių, kurie tam tikrose aplinkose gali būti sunkiai prieinami. Diegimo apribojimai gali apimti klausimus, susijusius tiek su programine, tiek su techninė įranga, kai reikia užtikrinti realiojo laiko paslaugas mobiliesiems įrenginiams ir internetinių paslaugų platformoms [46]. Tolesniuose eksperimentiniuose tyrimuose turėtų būti nagrinėjami metodai, kaip sumažinti skaičiavimo kompleksiškumą, susijusį su paaiškinimų generavimu. Tai apima kompromisų tarp paaiškinimų kokybės ir generavimo vėla-

vimo vertinimą ir optimizavimą, ypač įvairiose pramonės srityse. Post-hoc metodai, ypač perturbacijomis grindžiami metodai, yra gana neefektyvūs paaiškinimų generavimo laiko požiūriu. Lokalių paaiškinimų generavimo sąnaudos didėja su kiekvienu nauju įvesties vaizdu, kurį reikia aiškinti, todėl renkantis paaiškinamojo DI metodą labai svarbu suprasti jo naudojimo sritis ir atidžiai įvertinti turimus išteklius.

Disertacijoje pristatomas išsamus paaiškinamojo DI vaizdų segmentavimo srityje tyrimas. Taip pat pristatoma išsami literatūros apžvalga, išskirianti skirtingų tipų paaiškinamumo metodus, taikomus semantiniame segmentavime, ir siūloma paaiškinamojo segmentavimo taksonomija. Paaiškinamojo segmentavimo metodai suskirstyti į penkis pagrindinius pogrupius: prototipinius, gradientinius, perturbacinius, kontrafaktinius ir architektūrinius metodus. Dauguma paaiškinamojo segmentavimo metodų remiasi lokaliais paaiškinimais ir kokybiniu rezultatų palyginimu.

Paaiškinamajam semantiniam segmentavimui ištirti perturbaciniai paaiškinamieji metodai. Priešingai nei vaizdų klasifikavimo atveju, vaizdo uždengimu paremtiems metodams semantiniame segmentavime būdinga didelė įverčio metrikų rezultatų variacija. Todėl norint sugeneruoti mažiau triukšmingus paaiškinimo žemėlapius su didesniu spalvų intensyvumu, siūloma pritaikyti min-max normalizavimą. Kokybiniai rezultatai rodo, kad *logits* reikšmėmis paremtas metodas yra jautresnis lyginant su Dice koeficientu paremtu metodu ir gali būti geresnis pasirinkimas generuojant paaiškinimus. Kiekybinis įvertinimas rodo, kad vaizdo uždengimai, kurių filtrų spalvos yra panašesnės į įvesties vaizdo, yra tinkamesnės paaiškinimams generuoti. Tolesniuose tyrimuose būtų galima sistemingai ištirti įvesties uždengimus keleto klasių segmentavimo uždaviniuose, taip pat eksperimentuojant su skirtingų dydžių uždengimo filtrais.

Kokybiniai rezultatai rodo sėkmingą Ablation-CAM pritaikymą segmentavimo užduotyse. Pirmojo plano ir fono uždengimų atkūrimas skirtingų sluoksnių aktyvacijų žemėlapiuose patvirtina ankstesnių tyrimų [125] klasifikavime rezultatus, rodančius, kad pirmojo plano uždengimai turi didesnį poveikį modelio išvesties rezultatams nei fono uždengimai. Dalinių uždengimų jautrumas gali būti naudingas parodant vaizdo sritis, kurios yra labiausiai arba mažiausiai atsparios pirmojo plano ar fono uždengimams.

Taip pat disertacijoje pristatytas pirmasis tyrimas priešiškų atakų poveikiui segmentavimo modelių paaiškinimams. Pasiūlytas metodas leidžia vizualiai analizuoti priešiškų atakų poveikį modelio rezultatų paaiškinimams, ypač scenarijuose su paprastesnėmis segmentavimo formomis ir mažesniu klasių skaičiumi. Disertacijos tyrimai taip pat parodo

prieš paaiškinamąjį semantinį segmentavimą nukreiptų priešiškų atakų galimybę. Tolesniuose tyrimuose būtų galima ištirti fizines priešiškas atakas realiomis sąlygomis, taip pat jų pritaikomumą juodosios dėžės modeliams.

Pagrindinės disertacijos išvados:

1. Identifikuotos penkios pagrindinės paaiškinamojo DI metodų semantinio segmentavimo kontekste kategorijos ir nustatyta, kad daugiausia šiuo metu literatūroje siūlomų metodų remiasi kokybiniu paaiškinamumo įverčiu ir generuoja lokalius post-hoc paaiškinimus.

2. Taikant darbe pasiūlytą Ablation-CAM išplėtimą segmentavimo uždaviniui, nustatyta, kad pirmojo plano uždengimai turi didesnę įtaką nei fono uždengimai.

3. Perturbaciniai vaizdo įvesties metodai pritaikyti semantiniam segmentavimui, min-max normalizavimas pagerina ryškumo žemėlapius, o *logits* reikšmėmis paremti metodai yra jautresni lyginant su Dice įverčiu paremtais metodais.

4. Prieš paaiškinimus nukreiptos ir jais manipuliuojančios priešiškos atakos gali paveikti semantinio segmentavimo modelius. Šis pažeidžiamumas pabrėžia tolesnių tyrimų svarbą eksperimentuojant su tokio tipų atakų perkėlimu kitiems modeliams ir jų tyrimu juodosios dėžės aplinkoje.

Rokas Gipiškis
Post-Hoc Explainable Semantic Image Segmentation: Applications for Interpretability and Adversarial Attacks
Doctoral Dissertation
Natural Sciences
Informatics (N 009)
Thesis Editor: Zuzana Šiušaitė

Rokas Gipiškis
Post-hoc paaiškinamasis vaizdų semantinis segmentavimas: taikymai interpretuojamumui ir priešiškoms atakoms
Daktaro disertacija
Gamtos mokslai
Informatika (N 009)
Santraukos redaktorė: Jorūnė Rimeisytė-Nekrašienė

Vilnius University Press
9 Saulėtekio Ave., Building III, LT-10222 Vilnius
Email: info@leidykla.vu.lt, www.leidykla.vu.lt
bookshop.vu.lt, journals.vu.lt
Print run of 20 copies