UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Dmytro Shvetsov

Weakly Supervised Segmentation in Medical Imaging: A Counterfactual Approach

Master's Thesis (30 ECTS)

Supervisor(s): Joonas Ariva, MSc
Marharyta Domnich, MSc
Dmytro Fishman, PhD

Tartu 2024

Weakly Supervised Segmentation in Medical Imaging: A Counterfactual Approach

Abstract:
Deep learning is transforming medical imaging and radiology, notably improving pathology detection in CT and X-ray images. However, the effectiveness of deep learning models in segmentation tasks is often complicated by the requirement for large, annotated datasets. Addressing this challenge, this thesis explores Weakly Supervised Semantic Segmentation enabled by Explainable AI techniques, particularly through generating counterfactual explanations. Our innovative approach, termed Counterfactual Inpainting (COIN), flips classification predictions from abnormal to normal given the input image by erasing identified pathology regions in medical images. For example, when a classifier labels an image as abnormal, COIN produces minimally modified counterfactual example, effectively inverting the original prediction. Our method allows for accurate pathology segmentation without relying on detailed pre-existing masks, using instead much simpler to obtain image-level labels. We validate the effectiveness of COIN by segmenting both synthetic targets and real kidney tumors in TotalSegmentator and Tartu University Hospital CT scan datasets. Our results show that COIN significantly outperforms traditional methods like ScoreCAM, LayerCAM, and RISE, as well as the baseline counterfactual approach by Singla et al. These findings confirm that COIN is a promising method for kidney tumor segmentation in CT images, making a significant advancement in applying deep learning in healthcare domain that is limited by the scarcity of annotated data. The developed approach not only strives for high accuracy required in the medical domain but also reduces reliance on extensive annotated datasets by leveraging semi-automated labeling techniques.

Keywords:
Explainable AI, Counterfactual explanations, GANs, Semantic Segmentation, Medical Imaging, CT scans, Kidney Tumour.

CERCS:
T111 - Imaging, image processing; P176 - Artificial intelligence; B110 - Bioinformatics, medical informatics, biomathematics biometrics

Nõrga Juhendatud Segmenteerimine Meditsiinilises Kujutamises: Kontrafaktuaalne Meetod

Lühikokkuvõte:
Süvaõpe on muutmas meditsiinilise piltdiagnostika valdkonda. Näiteks täiustatakse süvaõppe mudelitega patoloogiate tuvastamist kompuutertomograafia- ja röntgenpiltidelt. Kahjuks on selliste mudelite treenimine keerukas kuna need eeldavad suurte märgendatud andmestike olemasolu, mida meditsiinilise piltdiagnostika valdkonnas on raske koguda. Käesolev lõputöö proovib seda probleemi lahendada läbi nõrgalt juhendatud semantilise segmenteerimise ja seletatava tehisintellekti tehnikate. Magistritöös kirjeldatakse innovaatilist meetodit nimega kontrafaktuaalne maalimine (Counterfactual Inpainting - COIN), mille käigus pööratakse klassifitseerimismudeli prognoos meditsiinilisele pildile anomaalsest normaalseks läbi patoloogilise piirkonna "üle maalimise" (kustutamisele). Näiteks kui klassifitseerimismudel hindab pildi anomaalseks (ehk patoloogiliseks), siis COIN loob pildist võimalikult minimaalsete muudatustega kontrafaktuaalse pildi, mis pööraks ümber klassifitseerija esialgse ennustuse. Sellise meetodiga saab segmenteerida patoloogiaid vaid pildi tasemel märgenditega, hüljates nii täpsete segmenteerimismaskide kasutamise, mida on keeruline hankida. COIN mudeli tõhusust hinnati Totalsegmentori ja Tartu Ülikooli Kliinikumi kompuutertomograafiapiltide andmestikel, kus segmenteeriti sellega nii tehislike kasvajaid kui ka päris kasvajaid neerudes. Tulemused näitavad et COIN ületab oluliselt traditsioonilisi seletatava tehisintellekti meetodeid nagu ScoreCAM, LayerCAM ja RISE ja ka võrdluseks võetud tavalist kontrafaktuaalmeetodit. Need leiud kinnitavad, et COIN on paljulubav meetod neerukasvajate segmenteerimiseks kompuutertomograafia piltidel, tehes olulise edusammu sügavõppe rakendamisel tervishoiuvaldkonnas, mida piirab märgendatud andmete nappus. Väljatöötatud metoodika mitte ainult ei püüdle meditsiinivaldkonnas nõutava kõrge täpsuse poole, vaid vähendab ka sõltuvust ulatuslikest annoteeritud andmestikest, kasutades ära poolautomaatseid märgistamistehnikaid.

Võtmesõnad:
Selgitav tehisintellekt, kontrafaktuaalsed selgitused, GAN-id, semantiline segmenteerimine, meditsiiniline pildistamine, CT-skaneeringud, neerukasvaja.

CERCS: T111 – Pilditehnika; P176 – Tehisintellekt; B110 - Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Contents

1 Introduction

The realm of medical imaging is witnessing a transformative era with the integration of artificial intelligence (AI) [1, 2, 3]. One of the pivotal challenges in this domain is the development of efficient and reliable methods for effective localization and classification of various diseases in human organs. Cui et al. highlights that training of modern deep learning models, particularly the highly accurate ones, requires a substantial amount of labeled data to function effectively [4]. Labeling medical images accurately requires highly specialized knowledge, often the domain of trained medical professionals such as radiologists. Their expertise is essential to ensure that the annotations are correct, as any error in labeling can lead to inaccurate training of the model and, consequently, unreliable outcomes. Tajbakhsh et al. discusses that the process of manually reviewing and labeling medical images is labor-intensive [5], since given the complex nature of these images, a single image can take a considerable amount of time to annotate correctly. The survey by Chaddad et al. [6] demonstrates that these problems are further amplified by the limited access to already labeled medical image datasets due to data protection laws. Another significant challenge in the application of deep learning models to medical imaging arises from the diversity of imaging tools and techniques used in collecting these images. Medical images, such as Computer Tomography (CT) scans and X-rays, are often captured using various types of equipment and software, each with its own characteristics and specifications. Guan et al. demonstrates well how this variability can lead to inconsistencies in the images [7], which is known as data shift phenomenon that poses a problem for deep learning models. Naturally, the need for methods arises to automate or simplify the manual data labelling efficiently.

One of the ways to simplify labelling for medical images is to try to extract meaningful and detailed pixel-level annotations from simpler and more general image labels. This task is categorized as Weakly Supervised Semantic Segmentation (WSSS) and it is often tackled with methods known from Explainable AI (XAI) [8, 9, 10, 11]. In computer vision, XAI methods are typically used to audit the pre-trained models and identify the image regions most critical for the classification decisions. Saliency maps output using these methods, can serve as preliminary segmentation masks. However, Ghassemi et al. [12] highlights that while saliency maps are visually straightforward, they often merge relevant and irrelevant information. This fusion complicates the task of pinpointing specific image features crucial for the model's decisions, thus challenging the creation of accurate segmentations from the saliency maps.

On the contrary to classical XAI methods, counterfactual explanation approaches [13, 14, 15] have emerged, aiming for minimal changes in the input to alter the decision output of a classifier. Originating from the work of Wachter et al. [13], these methods explore a series of 'what if' scenarios that help to visualize how slight changes in the presented data can lead to different predictive outcomes. Although counterfactual explanations have gained popularity across various tasks beyond computer vision, their

application from the perspective of WSSS for medical imaging remains unexplored.

Building on these findings, this thesis introduces an innovative approach to tackle the challenges of medical imaging data labelling, focusing on the use of WSSS and counterfactual explanations. The goal is to build a comprehensive computer vision pipeline that can effectively generate weak segmentation labels for existing unlabeled datasets from as little supervision annotations as possible. The contribution to this field is anchored in the principles established by the paper "Explaining the black-box smoothly - Counterfactual Approach" by Singla et al. [15]. While this paper laid the foundation for using a conditional Generative Adversarial Network model for generating counterfactual explanations, this thesis takes a step further. In addition to the replicated results from Singla et al., the methodology is advanced, focusing on enhancing the quality of the generated counterfactual images. The re-designed counterfactual pipeline builds a technology to generate accurate segmentation labels from having only a basic classification model. These advancements are not just theoretical but are backed by detailed experimentation and validation on the datasets with synthetic and real diseases. Finally, the developed counterfactual pipeline is tested against classical attribution methods [16, 17, 18], as well as baseline approach [15], to verify its superiority and practical purpose.

The thesis consist of the following parts: Section 2 outlines concepts involved in XAI, WSSS and main approaches used nowadays. Section 3 gives an overview of the main increments made to the model of Singla et al. to obtain the proposed Counterfactual Inpainting (COIN) pipeline. Section 4 presents the main experiments and results of the COIN to prove its effectiveness. Section 5 discusses achieved results and key learnings. Section 6 highlights conclusions, limitations and potential future works. Section 7 acknowledges the contributions of individuals and organizations that have supported this research.

2 Background

In this chapter, the overview of the XAI field is provided, emphasizing its importance in making AI systems more understandable and trustworthy. Various XAI methods, such as traditional attribution techniques and advanced trainable models, are explored and discussed in the context of various domains, including healthcare. Additionally, counterfactual explanation concepts are introduced. This discussion highlights how XAI can enhance WSSS and improve effectiveness and practical usage of AI applications.

2.1 Explainable AI

Explainable AI involves methods and techniques that help make the outcomes of AI systems clear to humans. This area has grown because there's a need to understand how complex AI models, especially those using deep learning, make decisions in many domains like finance, automated transportation, legal systems, healthcare and others. In finance, XAI is used to enhance transparency services related to credit scoring and algorithmic trading to facilitate regulatory and ethical considerations [19]. In the transportation domain, explainable methods help in understanding decisions made by autonomous systems, which is crucial for safety and regulatory compliance [20]. In the legal domain, XAI is applied to ensure fairness and transparency in automated judicial decision-making systems [21]. Lastly, XAI assists in diagnostic processes and treatment decisions, where understanding AI decisions can impact patient outcomes significantly [7].

2.2 Techniques used in XAI

In this chapter, the overview of the most popular methods are given for explaining the black-box classification models. The key topologies of algorithms in XAI are highlighted, including class activation mapping, perturbation-based, and counterfactual explanation models.

2.2.1 Class Activation Mapping

Class Activation Mapping (CAM) techniques [22, 16, 17] generate visual explanations by mapping the activations within the network to specific regions of the input image. These methods are founded on the premise that if a small change in a particular feature significantly affects the output, then this feature holds considerable influence over the decision. For instance, GradCAM [22] uses the gradient information from the last convolutional layer to create a heatmap highlighting the most important regions in the image for predicting a target class. In contrast, ScoreCAM [16] uses class activation maps derived from the forward activation outputs rather than relying on gradients. This approach avoids potential issues related to gradient saturation and noisy gradients, leading

to more stable and reliable visual explanations. LayerCAM [17] extends these capabilities by integrating both gradient and activation information across multiple layers, allowing for a more comprehensive exploration of how both deep and shallow features in a network influence its output.

2.2.2 Perturbation-based techniques

Classical perturbation-based techniques [23, 18, 24] systematically modify input images to analyze how these changes affect model outputs. The underlying principle is that if a modification in a specific part of input image changes the model's output, that region is crucial to the model's decision-making process. A notable implementation of this technique is the Randomized Input Sampling for Explanation (RISE) method. RISE involves generating thousands of random masks to obscure different parts of an input image and then observing how each masked version impacts the model's prediction. This approach allows RISE to compute importance scores for various image regions, effectively illustrating the model's focal points through visual saliency maps. In contrast, Extremal Perturbation [24], optimizes simple perturbations that are most impactful on the model's output. This method focuses on identifying the smallest or most extreme parts of the image that, when altered, significantly affect the output, offering a more targeted approach to understanding feature importance. Another aproach in this area is the Meaningful Perturbation technique [23], which computes the minimal perturbation needed to change a model's decision. Unlike RISE, which uses random masks, Meaningful Perturbation iteratively learns to mask out the least amount of image area while maximally confusing the model.

2.2.3 Counterfactual Explanations

Counterfactual explanations have their roots in cognitive science [25], psychology [26], and causality research [27], offering a profound way to dissect and understand AI decision-making processes. The formal counterfactual optimization function introduced in by Wachter et al. [13] marks a pivotal development in XAI, establishing a structured approach to generate these explanations in continuous data settings. Tim Miller's research [28] underscores the importance of contrastive explanations in human reasoning, suggesting that counterfactual reasoning is essential for making AI systems comprehensible and decisions justifiable.

In the context of tabular data, a counterfactual method can describe what attribute changes need to happen to influence the decision-making process. For instance, in a home loan application, a counterfactual explanation might highlight what adjustments, such as increasing annual income or improving credit score, would be necessary for an application to be approved. The Figure 1 illustrates this process. In the context of computer vision, this paradigm of models does more than simply flip the decision

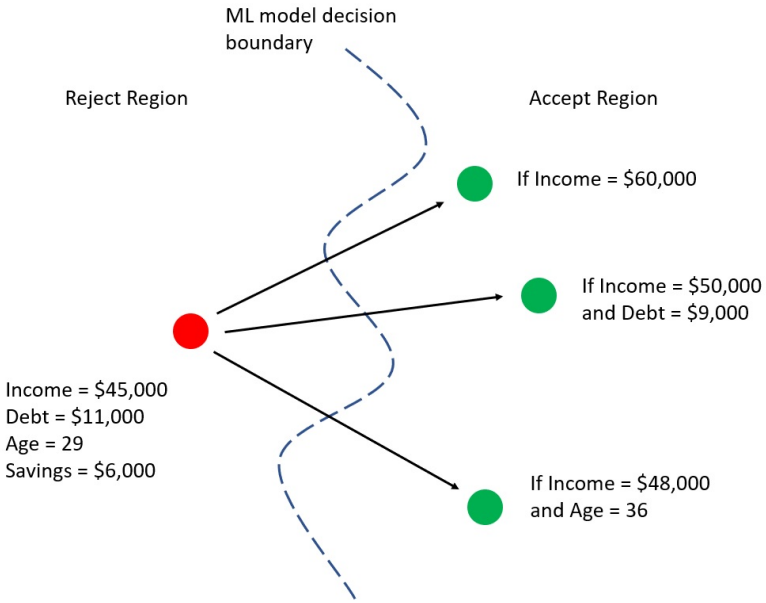

Figure 1. Visualization of an example how a counterfactual method can highlight how a rejected loan can become accepted given a pre-trained classifier [29].

of a classifier in an adversarial attack style but also ensures that the modifications to the input images are meaningful and interpretable within a real-world context, as discussed in studies by Zemni et al. [30] and Jeanneret et al. [31]. Initial models trained with the use of a conditional Generative Adversarial Network (cGAN) [32] have shown significant promise in clarifying the decision-making processes of classifiers and identifying potential biases or failure points.

While classical perturbation-based techniques are effective, they often lack the flexibility to be fine-tuned for specific datasets or to incorporate training dynamics that might yield higher quality saliency maps. In contrast, counterfactual explanations can be seen as a special case of perturbation-based methods that are more data-driven and adaptable. These techniques are model-agnostic and instance-based, meaning they can be applied across different models and work with individual data points (e.g a single image). They are designed to identify the minimal changes needed to an input to alter the model's prediction outcome.

Extensive research by Guidotti et al. [33] and Karimi et al. [34] surveys a variety of counterfactual frameworks, which generally tackle optimization problems or apply heuristic search strategies to derive explanations. These methods have been particularly enhanced in the domain of image processing. For example, Akula et al. [35] emphasized the role of Theory of Mind in explainability, developing techniques to generate counterfactuals by modifying minimal semantic-level features in images — such as changing

the stripes on a zebra — to investigate how such alterations influence model predictions.

In medical imaging, counterfactual explanations have been instrumental for diagnostic applications. For instance, Atad et al. [36] utilized a StyleGAN-based technique to adjust specific latent directions in chest X-ray images, facilitating an analysis of the underlying patterns that diagnostic models depend on. This approach was clinically evaluated with radiologists, demonstrating its practical benefits in medical diagnostics. Similarly, Singla et al. explored counterfactual explanations for chest X-ray imaging by training a generative model that can produce images with slight modifications that change classification outcomes. This model maintains the context of the original images, ensuring the relevance of the generated counterfactuals. Their work, validated in clinical settings, confirms that these generated explanations are not only academically interesting but also practically useful, offering clinicians new insights into the interpretative processes of AI in healthcare.

2.3 Weakly Supervised Semantic Segmentation

Semantic segmentation is a one of the fundamental task in computer vision, especially in medical applications. However, obtaining detailed segmentation masks can be prohibitively expensive. As a result, Weakly Supervised Semantic Segmentation (WSSS) has become popular for its ability to derive these masks from simpler image-level labels. Recently, many WSSS approaches [37, 38, 39] have adopted CAM methods to generate saliency maps, which are later used to indicate the most important image regions for the classifiers decision for each class. Frequently used WSSS workflow consist of the following steps:

1. Training image classifier for one or multiple classes.
2. Extracting saliency maps with a CAM methods.
3. Refining the saliency maps by postprocessing techniques (e.g morphology operations).
4. Train a segmentation model to produce the final segmentation labels (optional).

However, CAM methods have several drawbacks for WSSS. A significant issue is that they tend to highlight only the most distinctive regions of an image rather than the entire object relevant to the class, which can result in inaccurate segmentations. Another problem is that the resolution of saliency maps produced by CAMs depends on the resolution of the activation maps from the classifier model. The deeper layers of the model, while providing more meaningful semantic information, typically generate lower-resolution activation maps, leading to poorer resolution in the resulting saliency maps. Additionally, it has been observed that saliency maps might not change even

when the model's predictions are substantially altered by adversarial attacks, as noted by Ghassemi et al., which is questioning their reliability as explanations.

Given these challenges, a more promising approach to enhance WSSS could be the adoption of counterfactual explanations instead of relying on CAMs.

2.4 Generative Adversarial Neural Networks

One of the crucial steps to implementing a data-driven WSSS in computer vision is a choice of an image generation approach. Popular image generation frameworks include GAN-based and diffusion-based methods.

Generative Adversarial Networks (GANs) are one of the most established image generation methods originally introduced by Goodfellow et al. [40]. GANs operate by competing two neural networks against each other: a generator that creates images and a discriminator that evaluates them. While the discriminator is optimized to distinguish between real and fake images, the generator learns to create images that resemble real ones to deceive the discriminator. This adversarial process leads to high-quality image synthesis.

Compared to diffusion-based image generation [41], which gradually constructs images by adding noise and refining through reverse processes, GANs typically are more computationally efficient, since they can generate images in single feedforward pass [42]. While diffusion models are praised for their robustness and diversity in generated images, GANs often excel in scenarios where fine control over the generated output is crucial [43, 42, 44], such as in the creation of counterfactual visual explanations where precise changes are needed. This makes GANs particularly suitable for tasks requiring detailed manipulation of image attributes, enhancing their applicability in fields like medical imaging and visual data analysis.

A fundamental aspect of this thesis is the counterfactual explanation pipeline developed by Singla et al. The authors employ a conditional GAN framework as the image generation model to audit a pre-trained classifier. The cGAN is a specialized form of the foundational generative adversarial network, which incorporates classification labels as additional input. This inclusion facilitates more targeted and controlled image generation. Within counterfactual explanations, the cGAN is tasked with creating images that are slightly different from a given input yet lead to a different outcome from the pre-trained classifier. These modified outputs represent counterfactuals – the closest variations that could shift the classifier's decision. By analyzing how minor changes in these generated images influence the classifier's responses, insights can be drawn in a manner analogous to using saliency maps in attribution methods. Therefore, one can extract weak segmentation labels from the explanation maps.

2.5 Advancements in GAN-Driven Counterfactual Explanations

Generative Adversarial Networks have become increasingly popular for creating counterfactual explanations, showing their versatility in different areas. For example, Kenny et al. [45] used the PIECE method on MNIST data to change specific image features to produce realistic counterfactuals for CNN classifiers. In the field of autonomous driving, where clear explanations are essential due to safety concerns, Zemni et al. introduced a framework tailored for images containing multiple objects, like those typically found in urban driving scenes. Their approach involves encoding images into a structured latent space to allow detailed adjustments at the object level, making it ideal for complex settings. This method was even tested in driving scenarios to show how it can go beyond simple classification tasks to explain models handling semantic segmentation.

Jeanneret et al. took a different approach by turning adversarial attacks into meaningful changes for images showing facial expressions. They suggested that Denoising Diffusion Probabilistic Models [41] could be used to adjust these attacks to create useful and clear changes in images, such as altering a sad expression to a happy one. Bischof et al. [46] developed a comprehensive framework using image-to-image translation GANs aimed at enhancing both the interpretability and robustness of image classifiers. They tested this framework on tasks like detecting cracks in concrete and spotting defects in fruit, showing that it not only helps in understanding the images better but also makes models more resistant to adversarial attacks.

Although these examples highlight how effective counterfactual explanations can be for various applications, their use in WSSS for the medical domain has not been fully explored. Many current methods depend on having segmentation masks during GAN training, which might not always be possible. This thesis seeks to fill this gap by adapting the counterfactual approach for segmentation. It particularly builds on the techniques developed by Singla et al., modifying them to generate segmentation labels without relying on existing masks. This adaptation aims to leverage the strengths of counterfactual explanations to enhance the segmentation process, potentially leading to better accuracy and understanding in image analysis in healthcare.

3 Methods

In this section, we review the explanation pipeline from the original study by Singla et al., discuss its limitations, and suggest improvements to boost its performance. These enhancements include a better architecture, a new perturbation-based generator, added skip connections, a different loss measure, and removing the need for segmentation masks to better suit WSSS purposes. Finally, we introduce a new Counterfactual Inpainting method designed to summarize all the improvements for weak semantic segmentation. Furthermore, the main metrics, datasets and alternative methods for experimentation are

discussed.

3.1 Base Counterfactual Explanation Pipeline

In the original study by Singla et al., a counterfactual method for explaining the decisions of a medical image classifier is proposed. The method consists of two parts, namely a black-box classifier f and an explanation model $\mathscr{E}$. The pre-trained classifier accepts an input image X and outputs whether an image X is **normal** ($y = 0$) if $f(X) < t$ or **abnormal** ($y = 1$) if $f(X) \geq t$, where t is a threshold used to binarize the prediction output of f. In contrast, the generative model (cGAN) accepts an input image X and a tweakable parameter δ, such that the counterfactual image $X_{cf} = \mathscr{E}(X, \delta)$ flips the prediction class y. The auxiliary input δ is derived as the inverse of the classifier prediction. Specifically, if the classifier predicts **abnormal** for an image X, the generator aims to inpaint or remove the abnormal region, effectively **flipping** the classifier's prediction. Similarly, if the classifier predicts **normal**, the generated counterfactual image should add the **abnormal** region, **flipping** the classifier's prediction.

3.1.1 Image generation architecture

Singla et al. adapts the SNGAN [47] architecture for the generator and discriminator networks. The original generative model (cGAN) consists of an encoder $E(X) = z$ and a decoder $G(z, \delta) = X_{cf}$. The encoder transforms an input image X into a latent representation z, which is then passed into the decoder along with a condition label δ to produce a counterfactual image X_{cf}. Since the parameter $\delta \in [0; 1]$ is a continuous variable, the $[0; 1]$ range is discretized into $N = 10$ equal bins:

$$\{[0; 0.1)_{c\,=\,0},\ [0.1; 0.2)_{c\,=\,1},\ \ldots,\ [0.9, 1]_{c\,=\,10}\},$$

each corresponding to a unique condition vector c. The higher the number of conditions N, the more granular interface is exposed to audit the classifier's decisions. However, this comes with a serious requirement for the amount of data needed to train the explanation model, since the dataset for training cGAN is expected to be condition-balanced to achieve stable training. In addition, we outline that such break-down of the conditions has no practical purpose for the WSSS in the later sections. Therefore, only dual-conditioning with $N = 2$ is studied throughout this research.

3.1.2 Loss functions overview

In order to achieve realism among the generated images that influence the classifier decision, the explanation model is optimized according to the following objective function:

$$\min_{E,G} \max_{D} \left(\lambda_{GAN} L_{\text{GAN}} + \lambda_f L_f + \lambda_{idt} \mathcal{L}_{\text{idt}}\right),$$

where E, G and D are the Encoder, Decoder and Discriminator of GAN; f is a pre-trained classifier; L_{GAN} is a data consistency loss term; $\mathcal{L}_f$ is a classification model consistency loss term; $\mathcal{L}_{\text{idt}}$ is a domain-aware self-consistency loss term; λ_{GAN}, λ_f, λ_{idt}, λ_{tv} are respective hyper-parameters to configure contribution of each term. The overview of each loss term is provided in the following sections.

3.1.3 Data consistency loss term

The intuition behind the data consistency loss is that generated images should look similar to the images in the training dataset. For cGAN, the two networks are trained: a discriminator $D(\cdot)$ and a generator, which consists of the encoder $E(\cdot)$ and decoder $G(\cdot)$. The function itself computes a standard adversarial loss on the real/fake labels as follows:

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{X \sim P(X)} \left[\log \left(D(X)\right)\right] \; + \; \mathbb{E}_{X \sim P(X_{cf})} \big[\log(1 - D(E(G(X))\big],$$

where $P(X)$ and $P(X_{cf})$ denote distributions of real and generated images respectively.

3.1.4 Classification model consistency loss term

Classification loss ensures that the explanation model produces counterfactual images that effectively alter the classifier's predictions as intended. The general formulation of the objective function computes the distance measure between the predicted and expected distributions of probabilities.

$$\mathcal{L}_f = D_{\text{KL}}(f(X_{cf}) \;||\; 1 - f(X)),$$

where D_{KL} is the Kullback-Leibler divergence between observed (predicted) and reference (ground truth) probability distributions.

3.1.5 Domain-aware self-consistency loss term

The objective function is designed to guide the model in learning to produce counterfactual images that are consistent even when subjected to multiple counterfactual generations. The function includes two main components that encourage the model to return an identity (unchanged) image depending on the condition used. The complete function is formulated as follows:

$$\mathcal{L}_{\text{idt}} = \mathcal{L}_{\text{rec}}(X, \mathscr{E}(X, f(X))) \; + \; \mathcal{L}_{\text{rec}}(X, \mathscr{E}(\mathscr{E}(X, 1 - f(X)), f(X)),$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction loss between input and generated images. The first component of the function penalizes the model to generate identity image if conditioned

on the $f(X)$, whereas the second one should generate the identity when doing counterfactual generation two times in a row. For the experiments where we refer that segmentation masks are used, the $\mathcal{L}_{\text{rec}}$ function is defined as the average L_1 distance computed between foreground pixels belonging to each class class:

$$\mathcal{L}_{\text{rec}}(X, X') = \sum_j \frac{\sum S_j(X) \cdot |X - X'|}{\sum S_j(X)},$$

where X and X' represent the grayscale input and generated images, and S_j is the segmentation mask for the j-th label. For the binary segmentation scenario, the function will have two iterations of computation for the background and the main classes respectively.

In the case of the experiment with no masks used, $\mathcal{L}_{\text{rec}}$ is defined as L_1 distance over all the pixels of the images:

$$\mathcal{L}_1(X, X') = \frac{1}{HW} \sum |X - X'|,$$

where H and W are the height and width of the images X and X'.

3.2 Model Architecture Designs

In this chapter, the methodology is provided describing the threefold improvements made to the original counterfactual pipeline [15]. Firstly, perturbation-based image generation has been introduced to minimize the occurrence of artifacts typically associated with traditional encoder-decoder image generation methods. Secondly, the incorporation of skip connections, similar to those used in U-Net [48], enhances the integrity and detail of the generated images. Lastly, the conditioning process has been simplified, focusing the model specifically for WSSS. These enhancements collectively refine the pipeline's effectiveness and reliability resulting in the unique modeling framework denoted as Counterfactual Inpainting.

3.2.1 Perturbation-based Image Generation

A core advancement in the methodology over the original counterfactual explanation pipeline is the implementation of perturbation-based image generation. In conventional counterfactual generation models, the entire image is typically reconstructed. While effective in some scenarios, this method often leads to the introduction of artifacts – unintended alterations in the image that can skew the classification model's interpretation and analysis. The concept of perturbation-based image generation is closely related to the principles underlying adversarial attacks in AI [49] [50]. Both counterfactual explanation and adversarial attacks explore how minor alterations to an input can lead to significantly different outputs from a model. In adversarial attacks, these alterations are designed to

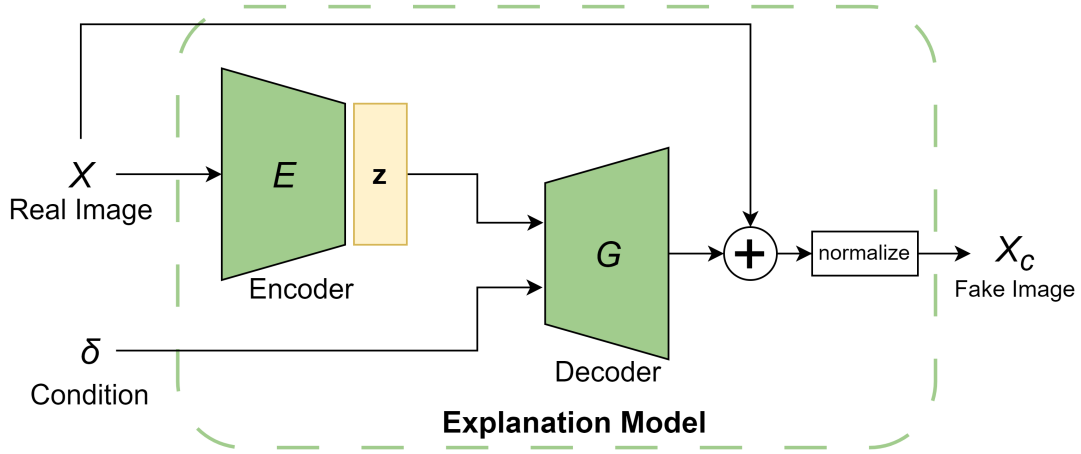

Figure 2. Overview of the perturbation-based counterfactual explanation model. The decoder generates only a perturbation map fused into the input image to produce a counterfactual example.

deliberately "fool" the model, whereas in counterfactual explanations, they are used to understand and explain the model's decision-making process. The Figure 2 describes a modified counterfactual generation pipeline.

3.2.2 Skip-connections in the generator

The U-Net architecture [48], originally developed for biomedical image segmentation, introduced the concept of skip-connections as a way to improve the flow of information across the network. The Figure 3 describes the idea of fusing intermediate features between different encoder-decoder layers.

In the context of generating counterfactual explanations, maintaining the integrity and identity of the original input image is crucial. The encoder-decoder architecture of the cGAN can sometimes lead to a loss of fine-grained details during the encoding process because of pooling or other downsampling layers that reduce the spacial resolution of activation maps. By integrating skip-connections, we facilitate more accurate perturbations, enhance reconstruction quality and preserve the input image details as much as possible.

3.2.3 Simplified conditioning

Another pivotal advancement of this thesis is the re-designing the counterfactual pipeline to solve the problems essential for weak segmentation labelling. At the inference stage of WSSS, our goals are to use the trained generator G to "inpaint" or remove the abnormality only when the classifier predicts that the image is **abnormal** to be able to annotate

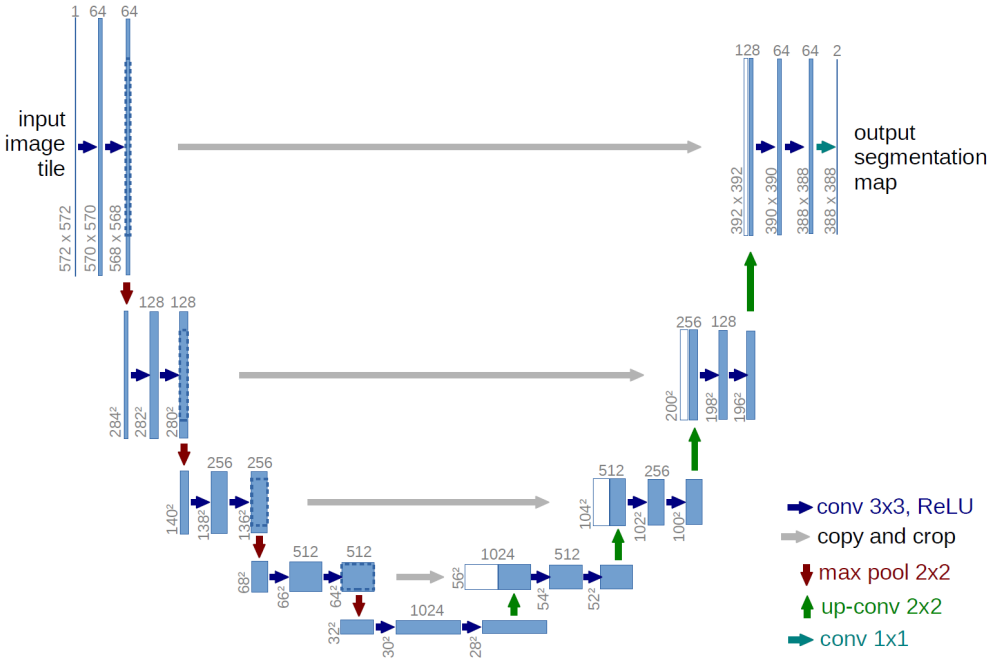

Figure 3. High-level overview of the original U-Net architecture [48].

existing anomalies. Therefore, once counterfactual image X_{cf} is generated, its absolute difference with original image X provides weak segmentation label. By applying a dual-conditioning strategy with $N = 2$ from Singla et al.'s framework, four distinct combinations (y, c) of the classifier predicted label and the condition used are present during inference, such as $(0; 0), (0; 1), (1; 0)$ and $(1; 1)$. However, only the combination $(1; 0)$ correlates with our goals at inference time. The combinations $(0; 0)$ and $(1; 1)$ refer to generating images without any perturbations (identity), which serves a purpose to only enforce the model with self consistency during training. The combination $(0, 1)$ refers to introducing synthetic anomalies to **normal** images, which also does not meet objectives of WSSS. With this said, we re-design the strategy to incorporate only single condition in the generator model. In the next section, we give a complete overview of our final WSSS pipeline.

3.3 Counterfactual Inpainting

In this chapter, we describe the main difference between the original Singla et al. and our COIN method in terms of loss functions for training the image generation model. COIN simplifies the architecture to handle only one condition to inpaint or remove the abnormal region. We train the model as a simple GAN that produces counterfactual image X_{cf} where $f(X_{cf}) < t$. In this case, the flipping happens only in one direction

when the X is **abnormal** and X_{cf} becomes **normal**, removing the area that affects the classifier's prediction. This approach addresses the challenge of needing a condition-balanced training set for the cGAN by reducing the complexity and data requirements. The overview of our Counterfactual Inpainting method is depicted in Figure 4.

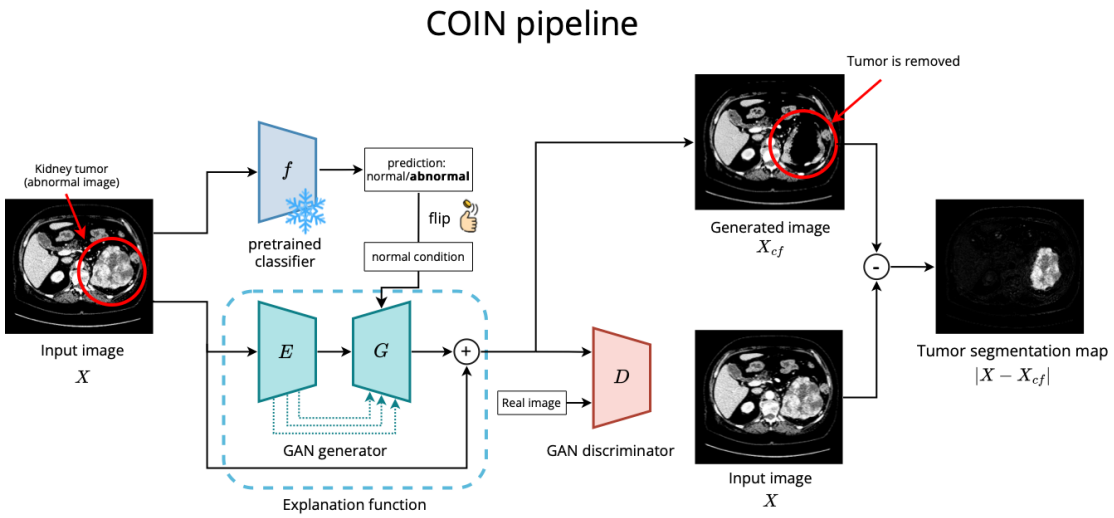

Figure 4. Overview of the proposed Counterfactual Inpainting pipeline [51]. Given the input image X and black-box classifier f that produces a classification label, the image-to-image model (GAN) generates a counterfactual image X_{cf} with $y = 0$. If X is **abnormal**, it is expected that X_{cf} no longer contains the abnormal part of the input image. Computing the absolute difference of the original image X and counterfactual image X_{cf} results in a weak tumor segmentation map. While training the pipeline, only GAN weights are updated. Classifier predictions are used for classifier consistency loss calculation.

3.3.1 Loss functions overview

We inherit the same loss functions described by Singla et al., and introduce an additional Total-Variation loss [52] to enforce smoothness in the generated images. The complete objective function for the Counterfactual Inpainting pipeline is given as follows:

$$\min_{E,G} \max_{D} \left(\lambda_{GAN} L_{\text{GAN}} + \lambda_f L_f + \lambda_{idt} \mathcal{L}_{\text{idt}} + \lambda_{tv} \mathcal{L}_{\text{tv}}\right),$$

where E, G and D is the Encoder, Decoder and Discriminator of GAN; f is a pre-trained classifier; L_{GAN} is a data consistency loss term; $\mathcal{L}_f$ is a classification model consistency loss term; $\mathcal{L}_{\text{idt}}$ is a domain-aware self-consistency loss term, and $\mathcal{L}_{\text{tv}}$ is a Total-variation loss term; λ_{GAN}, λ_f, λ_{idt}, λ_{tv} are respective hyper-parameters to configure contribution of each term.

3.3.2 Classification model consistency loss term.

Having introduced a new conditioning strategy, our intended classifier probabilities on the generated images are should be as close to 0 as possible. To optimize this between predicted and ground truth distributions, we still need a KL divergence, however in our reformulation for the inpainting pipeline, the condition-aware loss simplifies to:

$$\mathcal{L}_f = D_{\mathrm{KL}}(f(\mathscr{E}(X)) \,||\, 0),$$

where $\mathscr{E}$ is the image generation model.

3.3.3 Domain-aware self-consistency loss term.

Similarly to Singla et al., the main idea behind the objective function is to let the model learn cyclically consistent counterfactual images when applying a series of counterfactual generations. We compute one cycle of generations to produce $\mathscr{E}(X)$ and $\mathscr{E}(\mathscr{E}(X))$. Both counterfactual images should retain as much details as possible from the input image perturbing it only if it is **abnormal**. Additionally, Singla et al. uses segmentation masks to enforce local consistency over the foreground pixels of different segmentation labels present in the image. However, we employ simpler supervision, not requiring segmentation masks, to compute the reconstruction loss over the whole image instead of local regions. The domain-aware self-consistency loss for our Counterfactual Inpainting pipeline is given by:

$$\mathcal{L}_{\mathrm{idt}} = \mathcal{L}_1(X, \mathscr{E}(X)) \; + \; \mathcal{L}_1(X, \mathscr{E}(\mathscr{E}(X)))$$

3.3.4 Total-Variation loss term.

To further improve the smoothness of the segmentation masks, we adopt a Total-Variation (TV) loss [52] computed directly from the difference maps. TV loss enforces smoothness for the generated counterfactuals suppressing the noise and preserving the edges at the same time. The formula for the objective function is as follows:

$$L_{tv} = \frac{1}{HW}\left(\sum_{i=1}^{H-1}\sum_{j=1}^{W}(x_{i+1,j} - x_{i,j})^2 + \sum_{i=1}^{H}\sum_{j=1}^{W-1}(x_{i,j+1} - x_{i,j})^2\right),$$

where $x_{i,j}$ is the intensity of a pixel at position (i, j) in the input image X.

TV loss serves as a regularization term and improves consistency in the raw difference maps enforcing the model to perturb only densely located regions.

3.4 Singla et al.* method

To set up a baseline for comparison, we wanted to compare our method with that of Singla et al. However, a direct comparison was difficult because the model's code was not available requiring our own implementation. For a fair comparison, we also needed to modify the loss function to not rely on segmentation masks, since our goal is to detect these masks for WSSS without directly using them. We adjusted the loss terms and used a simple L_1 loss, calculated by averaging over all pixels in the generated and input images. This change initially led to poor segmentation performance. Suspecting that the original cGAN model might have unreported skip connections, we added these connections to our baseline. In our results, we refer to this enhanced version as the modified Singla et al*.

3.5 Alternative XAI approaches

In addition to the baseline Singla et al.* approach, we select three classical XAI methods for comparison with our Counterfactual Inpainting pipeline including ScoreCAM, LayerCAM and RISE. All the three methods do not require any model training or thorough hyperparameters tuning to be applied to an arbitrary image classification model. ScoreCAM and LayerCAM were chosen for their established roles within the class activation mapping category of XAI methods relying mostly on the intermediate features extracted from different layers of the model to be explained. As for RISE, the method is chosen for its close similarity with the counterfactual explanation approaches since it directly perturbs the input image to generate the final explanation map of which image regions are the most important.

3.6 Datasets and data preparation

In this chapter, we outline the main datasets data preparation techniques used for experimentation. In particular, we explore both synthetic anomalies and real tumors to test the developed counterfactual pipeline.

3.6.1 TotalSegmentator

The TotalSegmentator [53] dataset is a large collection of CT images specifically created for training and evaluating image segmentation algorithms. It includes a variety of medical scans from different body areas such as the chest, abdomen, pelvis, etc. For our experimentation, we used the metadata file present in the dataset to persist only the scans analyzing abdomen area. While the dataset has manually labeled ground truth masks, many of its annotations are made using pre-trained segmentation models. In our inspection of the quality of the kidney masks, we found many labelling inconsistencies,

so we used the latest nnU-Net [54] model to generate more accurate ones. We split the dataset's scans randomly divided into training and validation sets with an 80/20% split.

3.6.2 TUH

The Tartu University Hospital (TUH) kidney tumor dataset includes contrast-enhanced CT scans of 291 kidney tumor cases and 300 control cases. It features pixel-level annotations for three classes: kidney, malignant lesion, and benign lesion. Five radiologists annotated the dataset, with each scan reviewed by at least two of them. Discrepancies were resolved through direct discussion among the radiologists. The final labels were created by merging two sets of annotations. We used image-level labels, which indicate the presence or absence of a malignant lesion, for classifier training. Pixel-level labels were reserved for evaluating segmentation quality. The dataset's scans are randomly split into training and validation sets at an 80/20% ratio, stratified by the total tumor area in voxels per scan.

3.6.3 Synthetic anomaly augmentation

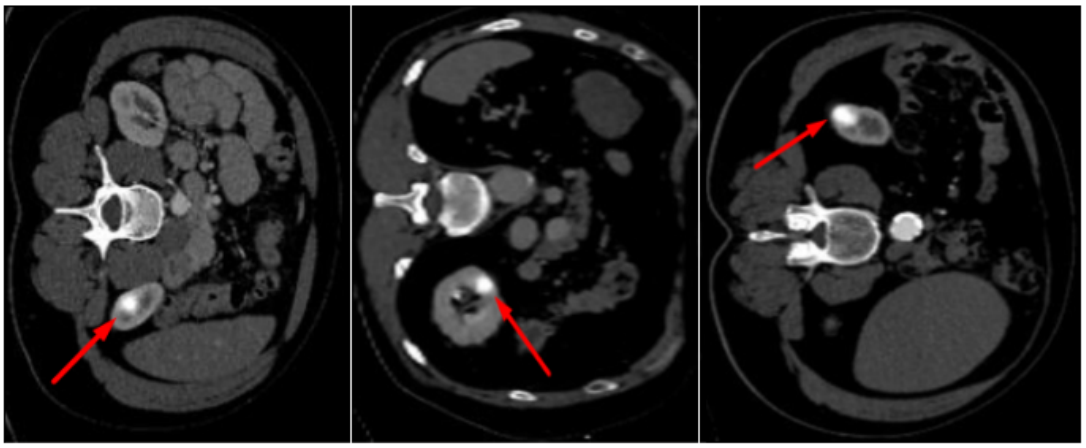

Figure 5. Examples of the synthetic anomalies injected randomly inside kidneys.

Given the limitations of real-world medical datasets, which often have few CT scans for each class, we begin our development with synthetic data to effectively assess modeling decisions, and later apply our findings to real-world kidney tumors. This thesis introduces a synthetic anomaly designed specifically for integration into CT scan datasets. With this approach, we set a solid benchmark dataset that allows us to control the quality, quantity and diversity of the instances to be segmented. This approach helps simulate a controlled environment to specifically test and improve the models' capability.

The synthetic anomaly is designed as a Gaussian blob with specific characteristics — fixed sigma and radius — chosen to resemble typical radiological findings like circular or ellipsoid structures seen in medical images. This makes our test conditions relevant to real-world scenarios.

We randomly sample a synthetic anomaly inside of one of the kidney kidney slices within the TotalSegmentator dataset to mimic the unpredictability and variety of real medical conditions. We also increase the challenge for our models by applying various transformations to the anomaly, such as random grid distortions, scaling, and rotations, to see how well the models can adapt to and process these changes. This method prepares us to then move to analyzing real-world kidney tumors. The examples of anomalous slices are presented in the Figure 5.

3.7 Evaluation

3.7.1 Realism of generated images

The Fréchet Inception Distance (FID) score [55] is a metric to evaluate how similar generated images are to real images. It uses a pre-trained neural network to get feature vectors from both image sets, capturing details like textures and shapes. The similarity is calculated using a formula that measures the difference between the distribution of features from real images x and generated images (counterfactual explanations) x_c. The formula is:

$$FID(x, x_\delta) = ||\mu_{\mathbf{x}} - \mu_{\mathbf{x}_\delta}||_2^2 + \mathrm{Tr}(\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{x}_\delta} - 2\big(\Sigma_{\mathbf{x}}\Sigma_{\mathbf{x}_\delta}\big)^{\frac{1}{2}}),$$

where the means and covariance of the activation vectors are denoted as μ's and Σ's respectively. The Tr notation is a function that computes a trace of a matrix, which is a sum of its diagonal elements. In this work, the features for computing the FID score are taken from the penultimate layer of an Inception v3 model [56].

3.7.2 Classifier consistency

The **Counterfactual Validity (CV)** score measures how often counterfactual explanations change a model's decision. For instance, if the original image is classified as **positive**, the counterfactual explanation should make the model classify it as **negative**. This change is counted as successful if the difference in predictions $|f(X) - f(X_{cf})|$ exceeds a specific threshold τ. The formula for the metric is given by:

$$\mathrm{CV}(X, X_{cf}) = \frac{\mathbf{1}(||f(X) - f(X_{cf})|| > \tau)}{N},$$

where input and counterfactual images are denoted as X and X_{cf}, τ is is the minimum required probability difference, N is the number of input images, and $\mathbf{1}$ that sums all

instances where the difference exceeds τ. Following Singla et al., we use $\tau = 0.8$ is used in the evaluations.

3.7.3 Segmentation metric

The Intersection Over Union (IoU) score is a common metric used to measure the accuracy of segmentation models. It effectively quantifies how much the predicted segmentation overlaps with the actual segmentation, making it an objective and straightforward evaluation metric. The formula is given by:

$$IoU(S, S_c) = \frac{|S \cap S_c|}{|S \cup S_c|} = \frac{TP}{TP + FP + FN},$$

where S is a ground truth mask, S_c is a predicted mask and TP, FP, FN represent predictions including true positives, false positives and false negatives respectively.

4 Experiments & Results

We developed all neural networks using PyTorch [57] and trained them at the University of Tartu's High Performance Computing (HPC) cluster on a single Nvidia A100-SXM-40GB GPU. We selected kidney image slices with at least 32 pixels in the kidney mask area, resized them to 256x256 using bilinear interpolation, and grouped them into batches of 16 for training. Our training process involved two stages: training a classifier and an explanation model. During the explanation model's training, the classifier's parameters were kept unchanged.

For the classifier, we used ResNet18 [58] for synthetic anomalies and EfficientNet-V2 [59] for real tumors. We chose ResNet18 for synthetic data because these anomalies are less complex, usually clear-cut, and strictly localized to the kidneys. On the other hand, real tumors vary widely in size, shape, and texture, requiring the more advanced EfficientNet-V2 architecture. We used the Adam optimizer with parameters $\alpha = 0.0002$, $\beta_1 = 0$ and $\beta_2 = 0.9$. to optimize our models.

The segmentation masks are generated by taking the absolute difference between the counterfactual and input images and applying a threshold. The code for our project is available at:

`https://github.com/Dmytro-Shvetsov/counterfactual-search`

4.1 Iterative COIN development

In this chapter, we describe our process of iteratively improving the Singla et al. method to develop our COIN pipeline. The Table 1 summarizes all the milestones (experiments

Table 1. Metric results for the iterative improvements of the original Singla et al. and our COIN methods [51]. Scenarios A and H refer to the original Singla et al. and modified Singla et al.* experiments respectively. The remaining cases B-G demonstrate positive impact of the proposed architectural changes.

ID & iterative changes	uses masks	perturbations	skip connections	conditions	**FID** ↓	**CV** ↑	**IoU** ↑
A	✓		0	2	1.6190	0.992	0.020
B (+ perturbations)	✓	✓	0	2	**0.5500**	**0.934**	**0.086**
C (+ skip connection)	✓	✓	1	2	0.3254	0.968	0.284
D (+ skip connection)	✓	✓	2	2	0.1751	0.933	0.480
E (+ skip connection)	✓	✓	3	2	0.0849	0.961	0.463
F (+ skip connection)	✓	✓	4	2	**0.0253**	0.9542	**0.509**
G (- masks)		✓	4	2	0.0493	0.996	0.528
H (- perturbations)			4	2	0.0466	0.998	0.445
COIN		✓	4	1	**0.0029**	**0.997**	**0.646**

A-H) and the results achieved on TotalSegmentator dataset and synthetic anomalies. First, we equipped the Singla et al.'s method with perturbation-based generator (experiment B). Next, we test different numbers of skip connections (experiments C-F). Finally, we removed the dependency on the segmentation masks in model training to make WSSS applicable (experiment G). The experiments A and H correspond to the original Singla et al. and modified Singla et al.* methods respectively.

4.1.1 Influence of perturbation-based generator

In this experiment, we used the counterfactual explainer from Singla et al. and tested its improvement with perturbation-based image generation, focusing on FID and IoU scores. Perturbation-based approach produces higher quality images by not rebuilding the entire input image from scratch. Instead, the decoder focuses on making just the necessary changes to alter the classifier's decision. We provide a qualitative comparison of images generated by both methods in the Figure 6.

4.1.2 Influence of skip connections on the generated image quality

In this experiment, we build on Singla et al.'s perturbation-based counterfactual generation to demonstrate how skip connections in the network help prevent severe distortions

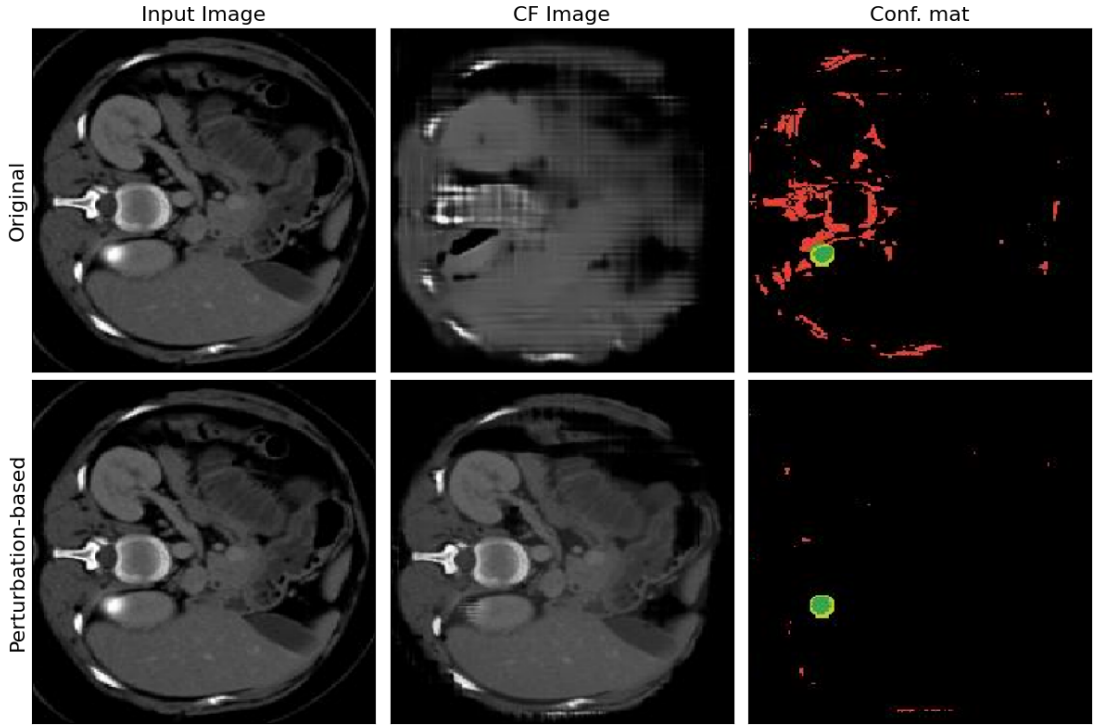

Figure 6. Examples of images generated with original and perturbation-based Singla et al.* pipelines [51]. The first and second columns denote input and counterfactual images respectively. The third column visualizes confusion matrix with colors: green for **true positives**, red for **false positives**, and yellow for **false negatives**.

in the input images. The encoding process naturally leads to information loss, making it difficult to recreate counterfactual images with minimal changes. By progressively adding skip connections between the encoding and decoding layers, we show improvements in FID and IoU scores. Using skip connections (sc), the images are less distorted, which also lowers the FID score. The Figure 7 visualizes the qualitative comparison of images generated with and without skip connections.

4.1.3 Perturbation-based Singla et al.* vs COIN

This experiment shows that our proposed Counterfactual Inpainting pipeline performs better than the base Singla et al.* method. We trained both approaches and assessed their ability to accurately segment using weak labels derived from the counterfactual images. The Counterfactual Inpainting offers two main advantages: it doesn't need segmentation masks to maintain local consistency, and it achieves higher IoU scores because the model simply decides whether to inpaint the anomaly to create the segmentation mask. A

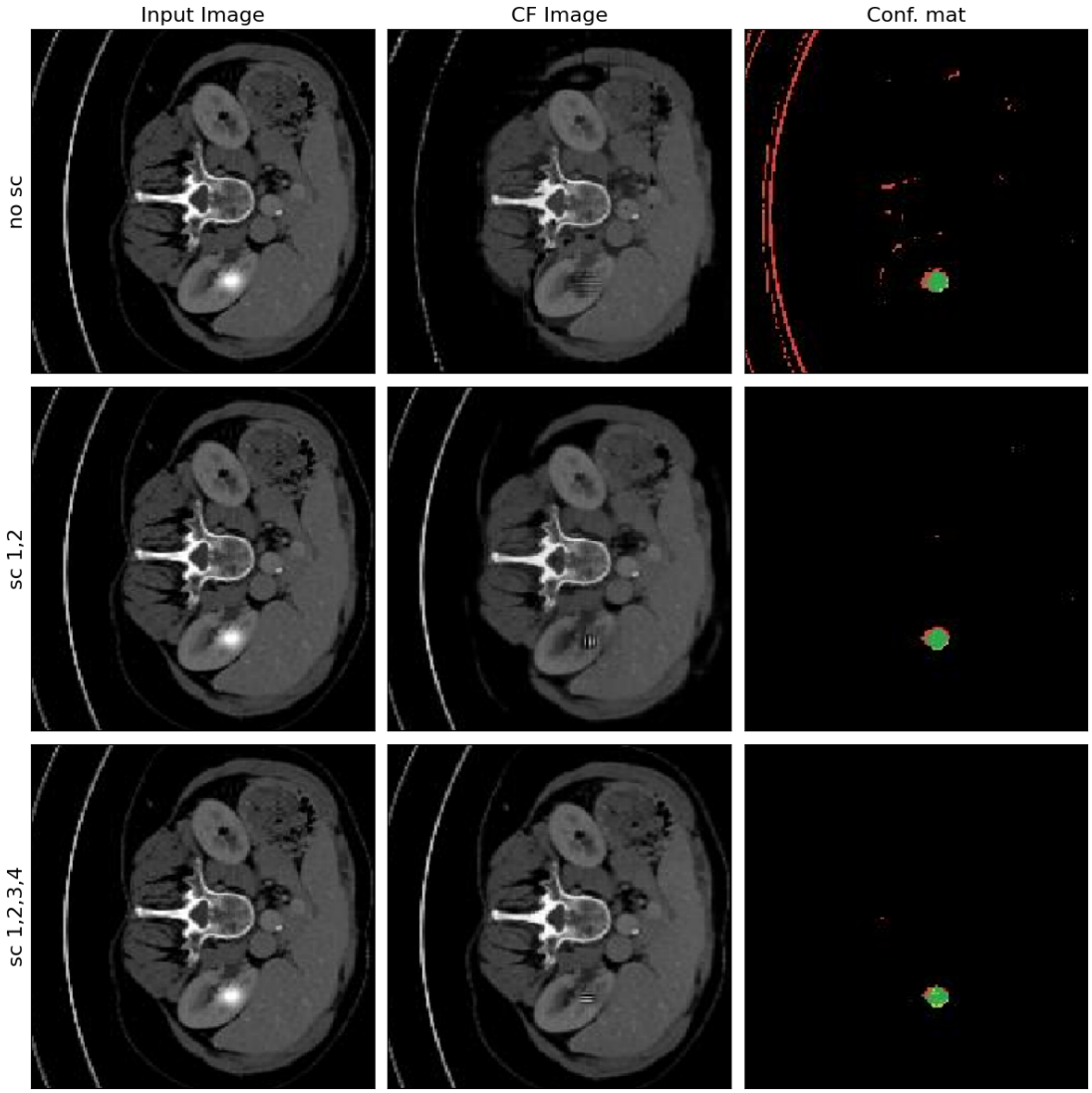

Figure 7. Examples of images generated with and without skip-connections (sc) between encoder-decoder layers of the perturbation-based Singla et al.* pipeline [51]. The first and second columns denote input and counterfactual images respectively. The third column visualizes confusion matrix with colors: green for **true positives**, red for **false positives**, and yellow for **false negatives**.

qualitative comparison of the counterfactuals produced by both methods is shown in the Figure 8.

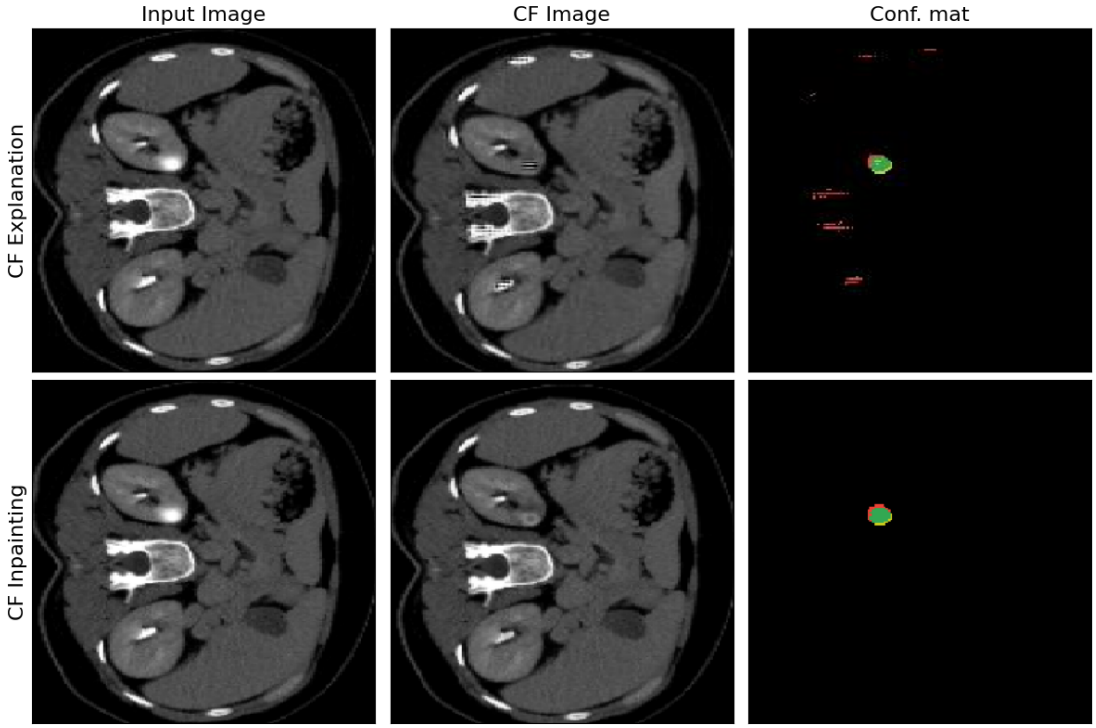

Figure 8. Examples of images generated with perturbation based Singla et al.* method equipped with skip connections and with the proposed Counterfactual Inpainting approach [51]. The first and second columns denote input and counterfactual images respectively. The third column visualizes confusion matrix with colors: green for **true positives**, red for **false positives**, and yellow for **false negatives**.

4.2 WSSS comparison of COIN and other methods

In this chapter, we present the evaluation results of our COIN method against attribution methods [17, 18, 16] and the modified Singla et al.* counterfactual pipeline. We normalized the output attention maps from all methods to a range of $[0; 1]$ and applied a fixed threshold. The parameter sweep is performed over the 0-1 range to select the best binarization threshold that maximizes the IoU score for each method. In addition, for the counterfactual methods, we used morphological operations like closing and opening to reduce noise and keep only the largest connected component in the masks. The results displayed in the Table 2 show that COIN significantly outperforms the other methods in IoU scores, achieving up to **60%** on synthetic anomalies and **14%** on real tumors.

Additionally, the Figure 10 shows visual examples of counterfactuals generated using our COIN method from the TUH dataset's validation set. We demonstrate that the COIN pipeline is effective for medical imaging WSSS applications, as it can learn useful weak segmentation labels with only classifier supervision.

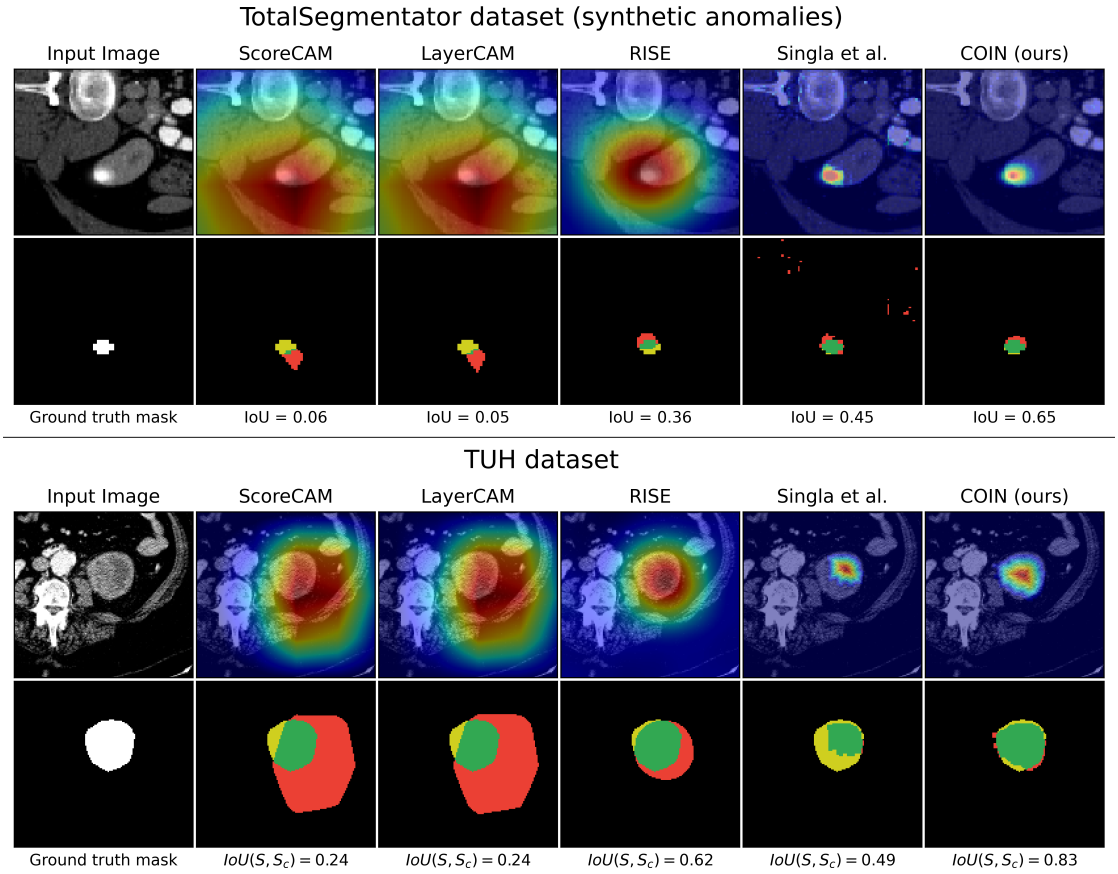

Figure 9. Visualization of the attribution and our Counterfactual Inpainting pipeline methods' predictions on TotalSegmentator and TUH datasets [51]. For each dataset and each method, the top row shows overlaid heatmap of the predicted saliency image, while the bottom row depicts thresholded masks obtained from saliency maps from each method. The colors in segmentation masks represent **ground truth** (white), **true positives** (green), **false positives** (red), and **false negatives** (yellow). Images are zoomed in for better clarity.

4.3 Ablation experiments for the final architecture

In the ablation study of this research, we examine how each loss term affects image realism (FID), the rate at which classifier predictions flip (CV), and segmentation accuracy (IoU) in our Counterfactual Inpainting pipeline. We conducted experiments where we independently removed the effect of each loss term by setting its weight to zero and then compared the outcomes. The results are shown in the Table 3.

Classifier consistency loss is key for producing effective counterfactual images. Removing this loss reduces the classifier's impact on the generator's outputs, which

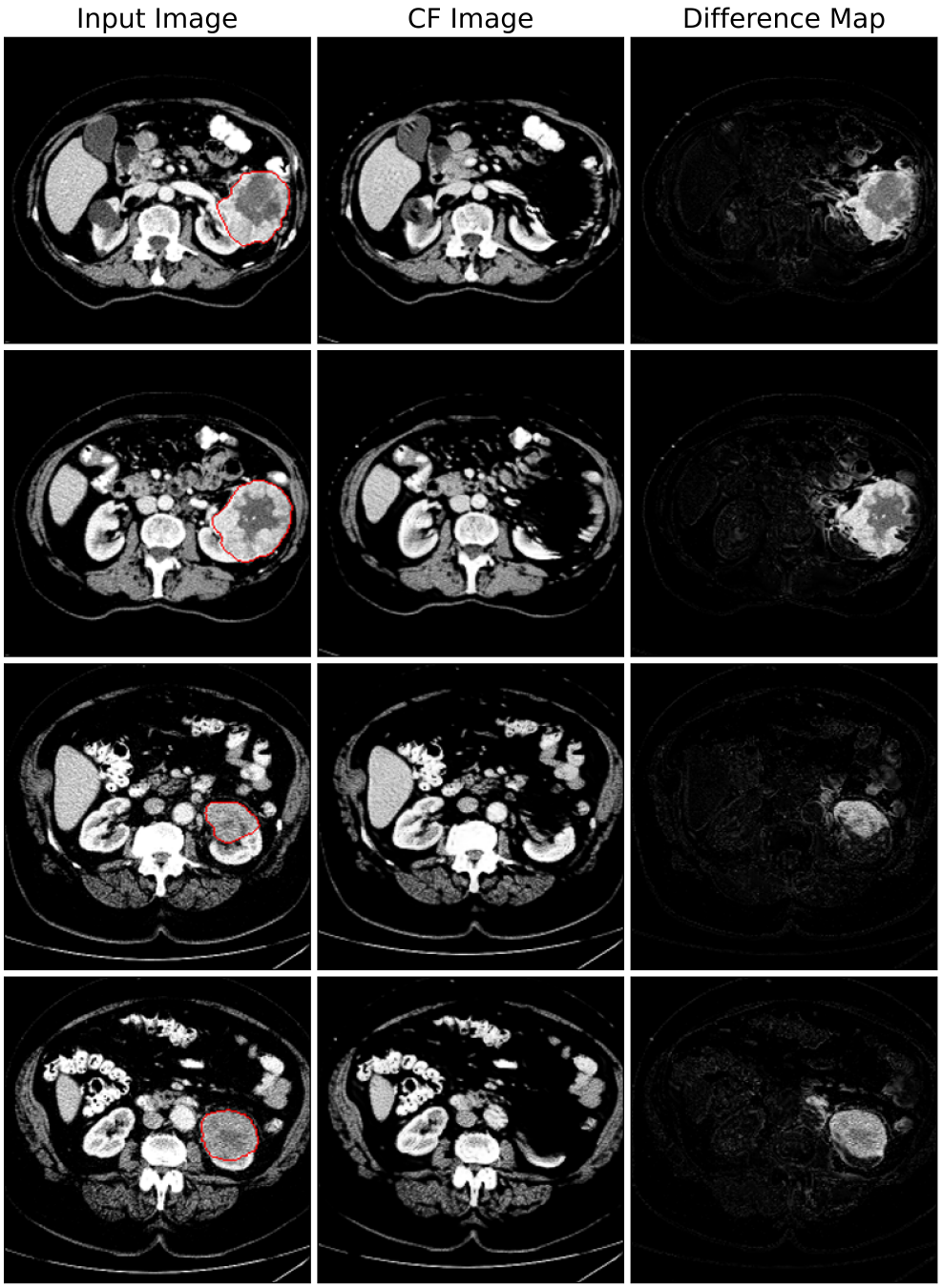

Figure 10. Visualization of counterfactual examples generated with COIN for TUH validation set [51]. The first and second columns refer to input and counterfactual images respectively. The third column depicts the absolute difference image providing the explanation map. Red contour corresponds to ground truth annotation of kidney real tumor. The model correctly inpaints majority of the tumors resulting in rich difference map making it convertible into a weak segmentation label.

Table 2. Metric results for the attribution methods and our Counterfactual Inpainting pipeline on TUH dataset [51]. Since CAMs and RISE do not create counterfactual images, FID and CV metrics cannot be computed for these methods.

Datasets	Methods	FID ↓	CV ↑	IoU ↑
TotalSegmentator	ScoreCAM	-	-	0.030
	LayerCAM	-	-	0.026
	RISE	-	-	0.397
	Singla et al.*	0.047	0.998	0.445
	COIN	**0.003**	**0.997**	**0.646**
Tartu University Hospital	ScoreCAM	-	-	0.293
	LayerCAM	-	-	0.296
	RISE	-	-	0.294
	Singla et al.*	0.203	0.992	0.352
	COIN	**0.036**	**0.980**	**0.432**

Table 3. Ablation study results on each loss term contribution based on TotalSegmentator dataset with synthetic anomalies [51].

Experiment	FID ↓	CV ↑	IoU ↑
$\lambda_{idt} = 0$	0.0024	0.997	0.500
$\lambda_{f} = 0$	0.0032	0.642	0.424
$\lambda_{tv} = 0$	0.0178	0.997	0.427
COIN baseline	**0.0029**	**0.997**	**0.646**

lowers the CV score by 35%. The Self-Consistency and TV losses help the model avoid random changes and concentrate on densely located regions, which forces the model to learn minimal perturbations and increases the IoU by up to **22%**.

5 Discussion

We developed a new Counterfactual Inpainting approach for weakly supervised semantic segmentation that outperforms existing attribution methods and the baseline Singla et al.* counterfactual method in segmenting synthetic anomalies on TotalSegmentator dataset and real tumors on the Tartu University Hospital dataset. For a fair comparison, we carefully chose the best threshold for all attribution methods. However, CAM methods did not perform as well because they focus on high-level classification-relevant areas of the image and often miss the central parts of anomalies. RISE was more effective

than ScoreCAM and LayerCAM because it creates saliency maps by directly perturbing the input image with pre-generated masks. We intended to compare our method with the original approach by Singla et al., but their code was unavailable. We implemented their method ourselves, adding skip connections to improve it. We suspect the original method may have omitted some architectural details because it initially performed poorly. Our modified version of Singla et al.'s approach showed improved segmentation results after postprocessing but still resulted in images with low fidelity, leading to a high FID score. Overall, our COIN method produces accurate difference maps, confirming its suitability for the WSSS task. Moreover, the figure 9 provides a qualitative evaluation of the counterfactuals generated by our method and others discussed.

5.1 Limitations and Future works.

A major limitation of our Counterfactual Inpainting pipeline is its reliance on the underlying black-box classifier. Training a classifier, though simpler than training a segmentation model, can become a complex task since it may require a lot of labeled data, significant computational resources, and meticulous parameter tuning. Any issues with the classifier, such as training data biases or problems with overfitting or underfitting, can negatively impact the quality of the counterfactuals that are generated. This may result in poor segmentation labels that do not accurately represent the data or highlight the relevant features for analysis.

Another drawback is that our current method only works with 2D data, even though CT scans are inherently 3D. This limitation can result in a loss of important spatial information, which is crucial for accurately segmenting and analyzing medical images. To address this, we plan to develop our pipeline to handle 3D data, which would improve the accuracy and relevance of the segmentation masks by providing a more detailed view of the structures in the scans.

Furthermore, while our pipeline currently focuses on tumor data within the medical field, we aim to broaden its application. Future research will test its effectiveness on different kinds of data and in various settings, especially where acquiring detailed segmentation masks is harder than obtaining classification labels. This could be particularly useful in fields where such data is scarce or has intricate structures of the objects to be segmented, which results in difficulties with data labelling.

6 Conclusion

In this study, we introduce a new strategy that uses XAI techniques for the weakly supervised semantic segmentation task in the medical field. We built on a counterfactual method from Singla et al., enhancing it with a perturbation-based generator, simplifying the process for either inpainting or removing abnormalities, and removing the need for segmentation masks. Additionally, we incorporated a new loss term to increase the quality of the counterfactual images. These improvements led to more accurate segmentation masks and outperformed both attribution methods and the original counterfactual approach. Our innovative method produces more detailed and nuanced counterfactual examples, making a significant contribution to the field of weakly supervised learning. By overcoming the challenges of sparse annotations and utilizing counterfactual reasoning, our inpainting pipeline provides a strong solution for enhancing semantic segmentation models without relying on extensive manually labeled datasets.

7 Acknowledgements

Special thanks are due to all the supervisors of this research, notably Joonas Ariva, Marharyta Domnich, and Dmytro Fishman, for their invaluable guidance and support. The data has been collected as part of the Clinical Investigation that has been approved by The University of Tartu Research Ethics Committee (no 332/T-9, dated 21.12.2020), we thank BetterMedicine for providing the dataset. Additionally, we acknowledge the HPC Center of the University of Tartu for supplying the necessary hardware resources for our research. The gratitude is extended to the Natural & Artificial Intelligence Lab (NAIL) led by Raul Vincente and Jaan Aru for the opportunity to present the research findings and receive insightful feedback, contributing significantly to the preparation of our paper for the XAI conference.

References

[1] Ross J Burton, Mahableshwar Albur, Matthias Eberl, and Simone M Cuff. Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. 19:1–11. Publisher: Springer.

[2] Mark A Musen, Blackford Middleton, and Robert A Greenes. Clinical decision-support systems. In *Biomedical informatics: computer applications in health care and biomedicine*, pages 795–840. Springer.

[3] Prashant Sadashiv Gidde, Shyam Sunder Prasad, Ajay Pratap Singh, Nitin Bhatheja, Satyartha Prakash, Prateek Singh, Aakash Saboo, Rohit Takhar, Salil Gupta, Sumeet Saurav, and others. Validation of expert system enhanced deep learning algorithm for automated screening for COVID-pneumonia on chest x-rays. 11(1):23210. Publisher: Nature Publishing Group UK London.

[4] Hengji Cui, Dong Wei, Kai Ma, Shi Gu, and Yefeng Zheng. A unified framework for generalized low-shot medical image segmentation with scarce data. 40(10):2656–2671. Publisher: IEEE.

[5] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. 63:101693. Publisher: Elsevier.

[6] Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable AI techniques in healthcare. 23(2):634. Publisher: MDPI.

[7] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. 69(3):1173–1185.

[8] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990.

[9] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 347–362. Springer.

[10] Zhang Chen, Zhiqiang Tian, Jihua Zhu, Ce Li, and Shaoyi Du. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11676–11685.

[11] Wei Shen, Zelin Peng, Xuehui Wang, Huayu Wang, Jiazhong Cen, Dongsheng Jiang, Lingxi Xie, Xiaokang Yang, and Q Tian. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. Publisher: IEEE.

[12] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. 3(11):e745–e750. Publisher: Elsevier.

[13] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. 31:841. Publisher: HeinOnline.

[14] Tim Miller. Contrastive explanation: A structural-model approach. 36:e14. Publisher: Cambridge University Press.

[15] Sumedha Singla, Motahhare Eslami, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly- a counterfactual approach.

[16] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks.

[17] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. LayerCAM: Exploring hierarchical class activation maps for localization. 30:5875–5888.

[18] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models.

[19] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.

[20] Towards explainable artificial intelligence.

[21] Nadia Burkart and Marco F. Huber. A survey on the explainability of supervised machine learning. 70:245–317.

[22] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

[23] Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457.

[24] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks.

[25] Ruth MJ Byrne. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *IJCAI*, pages 6276–6282.

[26] Frank C Keil. Explanation and understanding. 57:227–254. Publisher: Annual Reviews.

[27] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. 62(3):54–60. Publisher: ACM New York, NY, USA.

[28] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. 267:1–38. Publisher: Elsevier.

[29] jamesdmccaffrey. Researchers release open source counterfactual machine learning library.

[30] Mehdi Zemni, Mickaël Chen, Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. OCTET: Object-aware counterfactual explanations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15062–15071. IEEE.

[31] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16425–16435. IEEE.

[32] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets.

[33] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. pages 1–55. Publisher: Springer.

[34] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. 55(5):1–29. Publisher: ACM New York, NY.

[35] Arjun R Akula, Keze Wang, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Chai, and Song-Chun Zhu. CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. 25(1). Publisher: Elsevier.

[36] Matan Atad, Vitalii Dmytrenko, Yitong Li, Xinyue Zhang, Matthias Keicher, Jan Kirschke, Bene Wiestler, Ashkan Khakzar, and Nassir Navab. Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan.

[37] Lian Xu, Mohammed Bennamoun, Farid Boussaid, Wanli Ouyang, Ferdous Sohel, and Dan Xu. Auxiliary tasks enhanced dual-affinity learning for weakly supervised semantic segmentation.

[38] Jingxuan He, Lechao Cheng, Chaowei Fang, Zunlei Feng, Tingting Mu, and Mingli Song. Progressive feature self-reinforcement for weakly supervised semantic segmentation.

[39] Arvi Jonnarth, Yushan Zhang, and Michael Felsberg. High-fidelity pseudo-labels for boosting weakly-supervised segmentation.

[40] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks.

[41] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models.

[42] Yiran Xu, Taesung Park, Richard Zhang, Yang Zhou, Eli Shechtman, Feng Liu, Jia-Bin Huang, and Difan Liu. VideoGigaGAN: Towards detail-rich video super-resolution.

[43] Mengfei Xia, Yujun Shen, Ceyuan Yang, Ran Yi, Wenping Wang, and Yong-jin Liu. SMaRt: Improving GANs with score matching regularity.

[44] Taesun Yeom and Minhyeok Lee. DuDGAN: Improving class-conditional GANs via dual-diffusion.

[45] Eoin M Kenny and Mark T Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11575–11585. Issue: 13.

[46] Rafael Bischof, Florian Scheidegger, Michael A. Kraus, and A. Cristiano I. Malossi. Counterfactual image generation for adversarially robust and interpretable classifiers.

[47] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks.

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation.

[49] Joana C. Costa, Tiago Roxo, Hugo Proença, and Pedro R. M. Inácio. How deep learning sees the world: A survey on adversarial attacks & defenses.

[50] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4422–4431. IEEE.

[51] Dmytro Shvetsov, Joonas Ariva, Marharyta Domnich, Raul Vicente, and Dmytro Fishman. COIN: Counterfactual inpainting for weakly supervised semantic segmentation for medical images.

[52] Mehran Javanmardi, Mehdi Sajjadi, Ting Liu, and Tolga Tasdizen. Unsupervised total variation loss for semi-supervised deep learning of semantic segmentation.

[53] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: robust segmentation of 104 anatomical structures in CT images. 5(5):e230024.

[54] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnU-net: Self-adapting framework for u-net-based medical image segmentation.

[55] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium.

[56] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. version: 3.

[57] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition.

[59] Mingxing Tan and Quoc V. Le. EfficientNetV2: Smaller models and faster training.

Appendix

I. Usage of ChatGPT language model in academic writing

In the course of our research, we employed a natural language processing model, namely ChatGPT[1], to facilitate the academic writing process for this thesis. Specifically, ChatGPT was used to validate novel ideas and suggest alternative phrasings for sentences and paragraphs. Additionally, it was employed to answer our inquiries related to the thesis, however the responses provided by the model were verified against academic literature.

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Dmytro Shvetsov**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Weakly Supervised Segmentation in Medical Imaging: A Counterfactual Approach,

   supervised by Joonas Ariva, Marharyta Domnich, Dmytro Fishman.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Dmytro Shvetsov
14/05/2024

[1] https://openai.com/blog/chatgpt