UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Marko Lillemägi

Improving Counterfactual Image Generation for Weakly Supervised Tumour Segmentation through Theoretical and Architectural Alignment

Master's Thesis (30 ECTS)

Supervisors:
Joonas Ariva, MSc
Dmytro Fishman, PhD

Tartu 2025

Improving Counterfactual Image Generation for Weakly Supervised Tumour Segmentation through Theoretical and Architectural Alignment

Abstract:

As demand grows for scalable and interpretable AI tools in medical imaging, particularly for tumour detection, weakly supervised learning has emerged as a promising solution. It enables the use of existing datasets that include diagnostic labels without requiring labour-intensive, pixel-level annotations. This thesis explores how architectural alignment can better leverage image-level labels to enhance the performance and quality of existing segmentation frameworks. The work builds upon a classifier-guided conditional generative adversarial network (GAN) pipeline that segments tumours through counterfactual inpainting. Several architectural and theoretical modifications are proposed, including the removal of architectural bias, integration of classifier features into the generator, alignment of the discriminator with anatomically paired data, and the use of relativistic GAN loss with gradient normalisation. The effectiveness of these modifications is evaluated on the Tartu University Hospital (TUH) kidney tumour dataset, which contains annotated CT scans. Results show that the proposed modifications achieve segmentation performance comparable to the baseline while reducing model constraints and cutting training time by up to two-thirds. These findings suggest that aligning model components can enhance the performance of weakly supervised pipelines that rely on indirect learning signals.

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

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

Nõrgalt juhendatud kontrafaktuaalse pildi generaatori häälestamine kasvajate segmenteerimiseks läbi teoreetiliste ja arhitektuuriliste täiustuste

Lühikokkuvõte:

Meditsiinilises pildidiagnostikas, eriti kasvajate tuvastamisel, on kasvav nõudlus sklaeeritavate ja tõlgendatavate tehisintellektil põhinevate tööriistade järele. Nõrgalt juhendatud õpe on tõusnud üheks paljulubavaks lähenemiseks, mis võimaldab kasutada olemasolevaid diagnostilisi tõlgendusi sisaldavaid andmekogumeid, ilma et oleks vaja luua suurt töömahtu nõudvaid piksli tasandil märgendatud andmeid. Käesolevas töös uuritakse, kuidas eesmärgile paremini kohandatud ahitektuur võimaldab tõhusamalt kasutada pildi tasemel märgendeid, et tõsta olemasolevate segmenteerimisraamistike jõudlust ja kvaliteeti. Töös arendatakse edasi olemasolevat klassifitseerimismudelil põhinevat tingimuslikku generatiivset võistlevat võrgustikku (GAN), mis segmenteerib kasvajaid kontrafaktuaalse maalilmise kaudu. Esitatakse mitmeid täiustusi, sealhulgas arhitektuurilise kallutatuse eemaldamine, klassifitseerija tunnuste integreerimine generaatorisse, diskriminaatori kooskõlastamine anatoomiliselt oluliste andmetega ning relatiivse GAN-i kaofunktsiooni kasutamine koos gradiendi normaliseerimisega. Muudatuste tõhusust hinnatakse Tartu Ülikooli Kliinikumi neerukasvajate andmestikul, mis sisaldab märgendatud kompuutertomograafia pilte. Tulemused näitavad, et esitatud täiustused saavutavad baasmeetodiga võrreldava segmenteerimisjõudluse, vähendades mudeli piiranguid ja kasutades vaid üks kolmandik algsest treeningajast. Saadud tulemused viitavad sellele, et erinevate mudeli komponentide kooskõlastamine võib parandada nõrgalt juhendatud mudelite tõhusust, mis tuginevad kaudsetele õpisignaalidele.

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

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

Contents

1. Introduction

Medical imaging plays a crucial role in the early detection of cancer. Identifying cancer at an early stage, particularly malignant (cancerous) tumours, significantly improves the chances of successful treatment [1]. Various imaging modalities are used in clinical practice, such as computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound. However, these images alone do not provide diagnostic value without interpretation. The true value lies in the analysis and annotation of the images, a task performed by radiologists who translate visual data into actionable clinical insights.

As cancer incidence rates continue to rise [2], radiologists face increasing pressure to interpret a growing number of scans. In this context, artificial intelligence (AI) offers promising opportunities to automate diagnostic processes and streamline clinical workflows. Existing systems already work alongside radiologists to increase diagnostic accuracy and reduce the workload on clinicians [3].

To train AI systems for tumour detection, high-quality annotated datasets are essential. However, obtaining such datasets is difficult due to the confidential nature of medical data and the scarcity of publicly available datasets with detailed pixel-level annotations. Creating these annotations is especially challenging in medical imaging, as it requires radiologists to manually label and verify each scan, making the process both time-consuming and costly. To overcome these limitations, researchers are increasingly exploring approaches that reduce the reliance on pixel-level labels. Techniques such as weakly supervised and self-supervised learning aim to infer detailed annotations from less specific information, such as image-level labels.

In countries like Estonia, where healthcare systems are highly digitised, most medical scans are stored in a centralised database. In addition, a separate national database holds patients' complete medical histories, including diagnostic information. This allows researchers to link imaging data with clinical diagnoses without the need for radiologist involvement. Such datasets are particularly valuable for weakly supervised learning, where image-level labels, like diagnosis codes, can be used to train models that learn to detect and localise tumours without requiring precise, pixel-level annotations. This approach is not limited to only cancer since it can be applied to any disease or abnormality that presents visible patterns in medical imaging.

This thesis aims to improve weakly supervised tumour segmentation by exploring methods to better align components of an existing counterfactual image generation pipeline. The objective

is to maximise the information extracted from image-level labels, thereby enhancing both efficiency and performance. Section 2 provides an overview of key concepts and background knowledge, laying the foundation for this work. Section 3 explores the proposed modifications in detail, including their underlying motivation, as well as the dataset and evaluation metrics used. Section 4 presents the results of these modifications and concludes with an overall assessment of their effectiveness. Section 5 summarises the main contributions of the thesis, followed by reflections on the challenges encountered during the research and outlines potential directions for future work. Section 6 expresses gratitude to the individuals and organisations who supported and assisted this research.

2. Background

This section introduces the key concepts that form the foundation of the work presented in this thesis. It begins with describing the clinical workflow of tumour detection using computed tomography (CT), highlighting current challenges in manual interpretation. It then introduces weakly supervised semantic segmentation as a scalable alternative opposed to fully supervised and unsupervised methods, followed by counterfactual explanations as a tool for model interpretability and localisation. The section concludes with an overview of the architecture, which serves as the foundation for the modifications explored in this thesis.

2.1 Detecting Tumours with CT Imaging

Among the various imaging modalities used in clinical practice, CT plays a particularly important role in cancer detection. CT scans use a narrow beam of X-rays that rotate around the body, capturing multiple cross-sectional images (or ”slices”). These slices are then combined to form a detailed 3D image of the body. CT scans are relatively fast to acquire and provide high spatial resolution, making them effective for visualising anatomical structures. The high contrast in CT images allows for clear detection of tumours, which makes CT especially valuable for cancer detection. [4]

After a patient has been scanned, radiologist-driven interpretation involves manually examining the images for abnormalities, such as malignant lesions that could indicate the presence of cancer. This process is time-consuming and cognitively demanding, especially in high-volume clinical settings. As the number of scans continues to grow, there is increasing demand for automated tools that can support radiologists by identifying suspicious regions or even generating preliminary segmentations. The structured and high-resolution nature of CT data makes it a strong candidate for training AI models aimed at improving tumour detection and localisation.

2.2 Overview of Weakly Supervised Semantic Segmentation

In machine learning, the approach to model training is largely determined by the type and granularity of available labels. Supervised learning relies on detailed manual annotations like bounding boxes or segmentation masks. While this approach typically delivers high performance, obtaining pixel-level segmentation masks on a large scale may be impossible or impractical, as in the medical domain.

At the opposite end of the spectrum is unsupervised learning, where the model receives no labels at all and instead learns to uncover hidden structures and patterns within the data. While

promising in theory, such models are often large and require vast amounts of data to produce generalisable and meaningful results. This is particularly challenging in medical imaging, where scans are highly complex and vary significantly due to differences in human anatomy.

Weakly supervised semantic segmentation (WSSS) offers a promising middle ground between supervised and unsupervised learning by automatically generating pixel-level segmentation masks from image-level labels. Instead of requiring manual pixel-level annotations for each image, the model is trained using a limited number of image-level labels. For this thesis, a binary tumour label is used as an image-level label, dividing the dataset into two classes: images with tumours and healthy ones (Figure 1).

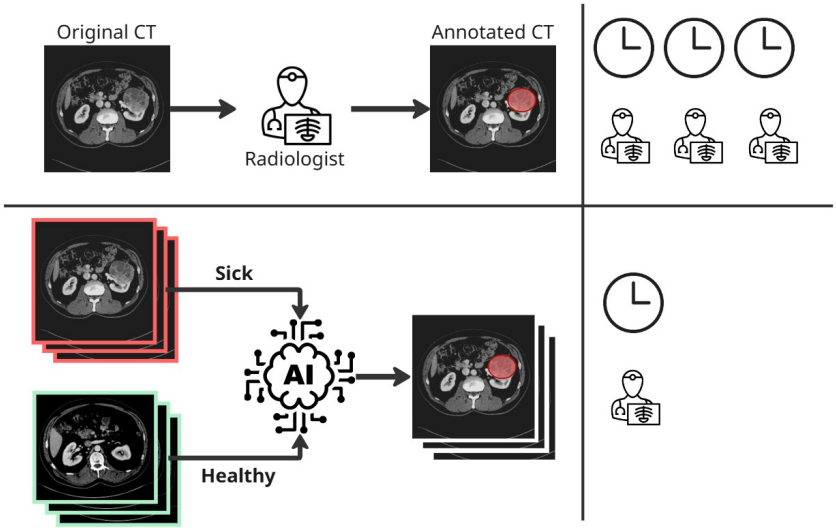

Figure 1. Visual example of WSSS benefits in the medical domain, highlighting the key differences compared to manual annotation. The top row indicates a process where a radiology specialist manually annotates pixel-level segmentation masks, while the bottom row shows a WSSS method that infers segmentation masks from image-level labels using artificial intelligence.

One popular way to convert class labels into segmentation masks is by using class activation maps (CAM), which highlight areas associated with the image-level label [5]. In this method, a classifier is trained on the image-level labels, and class activation maps are then extracted using a specific CAM technique.

The original CAM method requires the penultimate layer of the classifier to be a global average pooling layer. The activations from this layer are combined with the predicted class score to produce a saliency map that highlights class-specific discriminative regions [5]. To overcome

the architectural constraints of this approach, Grad-CAM was introduced. Grad-CAM uses the gradients of the class score with respect to the feature maps of a convolutional layer to generate the class activation map [6]. These gradients are globally average-pooled and used to weight the feature map activations, producing a saliency map. This method enables saliency maps to be extracted from any convolutional layer where gradients can be computed. However, deeper (higher-level) layers typically provide more semantically meaningful information for identifying class-specific regions.

Another method, called Score-CAM, eliminates the need for gradient computation by using input masking based on the feature maps from the last convolutional layer [7]. The feature maps are normalised and used as masks on the input image. For each masked image, the resulting class score is recorded and used as a weight to sum the feature maps, producing a saliency map for the target class. If the resulting activation value is close to zero, it indicates that the corresponding region does not contribute significantly to the classification decision.

By extracting these important regions, segmentation masks can be derived. CAM-based methods are also widely used as tools for model interpretability, helping to visualise what the model is ”looking at” when making its decisions.

In most cases, class activation maps highlight overly broad or general regions and must be further refined to achieve the level of detail required for the specific task. Post-processing is typically performed using classical computer vision techniques (such as thresholding and smoothing) or with more recent foundational models like SAM (Segment Anything Model) [8, 9] or nnInteractive [10].

SAM is a zero-shot segmentation model that also supports prompt-based guidance. NnInteractive is a newer state-of-the-art interactive segmentation model like SAM, but it operates on 3D volumes. By using class activation maps as prompts, SAM and nnInteractive can refine the segmentation mask by sharpening edges and segmenting the object at a more contextual level. Although these models were not originally designed as post-processing tools for refining class activation maps, this is one way they can be effectively integrated into a weakly supervised pipeline. Their primary purpose is to assist in generating preliminary segmentation masks in a flexible and interactive manner.

2.3 Counterfactual Explanation

Another way to gain insight into which parts of the input are important in the decision-making process of a classification model is through counterfactual explanations [11], which aim to change the model's output by modifying the input. This is done by generating alterations to the input that attempt to ”flip” the classifier's prediction, potentially revealing informative regions such as the location of a tumour. Furthermore, the goal is not only to change the outcome but to do so in a way that keeps the modified input logical and interpretable, reflecting a plausible ”what-if” scenario. This approach emphasises understanding the model's decision-making process and explaining its decision boundaries.

In the medical imaging domain, this can be represented by transforming an abnormality in an image back to a normal state. By generating counterfactual images, it is possible to detect and segment abnormalities by comparing the original and modified images [12]. This method of detection and segmentation has the added benefit of making a black-box model more transparent, contributing to a clearer decision-making pipeline. This is particularly important in the healthcare domain, where the stakes are high and all forms of supporting evidence are valuable for making the final decision.

Perturbation-based techniques [13] can be used to generate counterfactual explanations and to gain further insight into what the model considers most important. By modifying the model inputs and observing the resulting changes in outputs, it is possible to quantify the importance of specific regions of the input. However, this approach also presents challenges. In scenarios where the classifier is responsible for detecting a tumour, some automated or unsupervised modifications may act as adversarial attacks on the model, potentially generating false positives.

An adversarial attack is a method used to evaluate the robustness of models. This technique aims to confuse the model into misclassifying an image while keeping the image as visually similar to the original as possible. In medical imaging tasks, if the classifier is highly susceptible to adversarial attacks, it can lead to poor downstream results. For example, a generator may produce modifications that ”flip” the classifier's prediction but are not semantically meaningful or relevant to the true task of tumour localisation. One way to improve a classifier's robustness against adversarial attacks is to apply augmentations during training [14], enabling the model to learn to generalise. This also has the added benefit of effectively enlarging the dataset. However, this is not a silver bullet. Certain augmentations may overly distort the input, leading to unrealistic examples that can negatively affect training performance.

2.4 COIN

”COIN: Counterfactual Inpainting for Weakly Supervised Semantic Segmentation of Medical Images” [15] is a perturbation-based conditional counterfactual image generator (Figure 2) that serves as starting point architecture for this thesis. It builds upon the method proposed by Singla et al. for black-box classifier explanation [12]. Similar to its predecessor, COIN consists of a black-box classifier f and an explanation model $\mathscr{E}$. Given an input image X, the pre-trained classifier f produces a binarised output $f(X) \in 0, 1$, indicating whether the image is normal $(f(X) = 0)$ or abnormal $(f(X) = 1)$. The classification result is used as a condition c for the explanation model $\mathscr{E}$, which generates a counterfactual image $X_{cf} = \mathscr{E}(X, c)$ with the goal of flipping the classification outcome. Thus, when the original image X is classified as abnormal $(f(X) = 1)$, the counterfactual image should be classified as normal $(f(X_{cf}) = 0)$. The regions of abnormality are identified by subtracting the counterfactual image X_{cf} from the original image X, resulting in the segmentation mask $|X - X_{cf}|$.

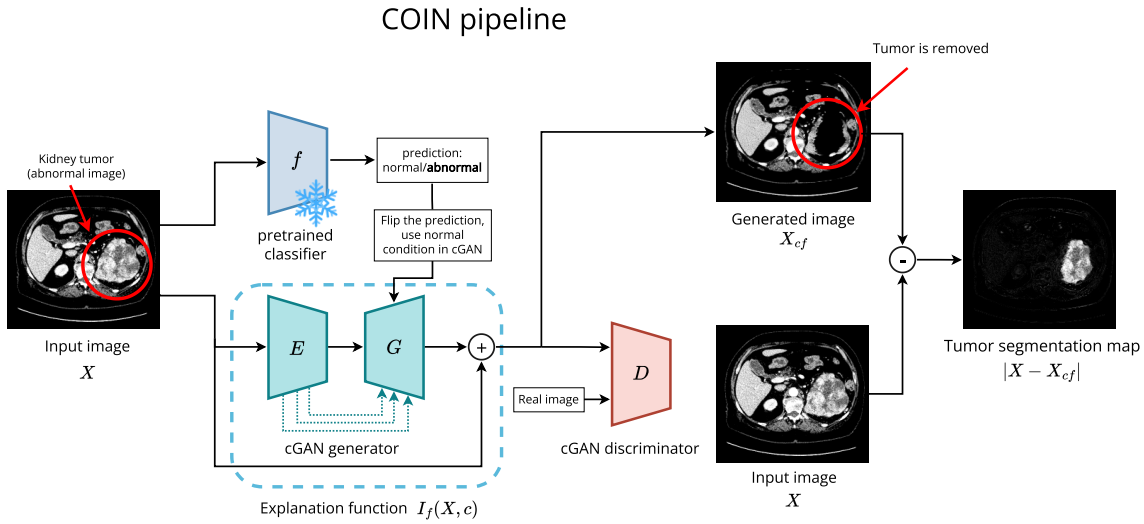

Figure 2. Overview of the COIN pipeline [15]

COIN extends this pipeline by employing perturbation-based techniques in the explanation model $\mathscr{E}$ to modify only specific parts of the input image that are likely to affect the classification output. This allows the generator model to focus less on reproducing the original image and more on generating impactful changes. Compared to the approach of Singla et al., COIN also integrates skip connections to produce more accurate perturbations and enhance reconstruction quality. Furthermore, the conditioning process is simplified by transforming images only from abnormal and normal to normal. Additionally a total variation loss term [16] is introduced to

improve the smoothness of the segmentation masks. Since Singla et al. did not publish the code used in their paper [12], the authors of COIN [15] had to reimplement everything from scratch. This may have led to some minor differences between the two models, which they acknowledge in their paper.

2.5 GAN and cGAN

The explanation model $\mathscr{E}$ used in the COIN architecture consists of three components: an encoder Z, a decoder G, and a discriminator D. This type of architecture is generally referred to as a Generative Adversarial Network (GAN) [17]. However, in this case, the decoder and discriminator have been modified by Singla et al. [12] to incorporate the condition label c, transforming it into a conditional GAN (cGAN) [18]. The image X is first encoded by the encoder Z into a fixed-size embedding vector $Z(X)$. This embedded representation is then passed to the decoder G, which generates an image conditioned on c. The generated image is evaluated by the discriminator D, which compares it to real, unaltered samples from the dataset. The objective of the discriminator is to distinguish between real and generated images, while the decoder aims to produce images realistic enough to fool the discriminator. This adversarial process between the decoder and the discriminator drives the learning in the GAN framework.

To accommodate the condition c, additional parameters are introduced into the decoder G and the discriminator D. By incorporating conditional vectors that are activated depending on the condition, the model gains the flexibility to learn condition-specific knowledge [18]. This can be viewed as having layers in the decoder G and discriminator D that are only utilised when their corresponding discrete condition is applied, while the remaining layers are shared. This approach is also preferable to training separate models for each condition, as it enables the sharing of information across conditions. However, a drawback of using conditional modelling is the requirement for balanced data across conditions to ensure optimal training. Otherwise, the shared (non-conditional) components may become biased toward the more prevalent condition.

3. Methods

This section provides a detailed explanation of the modified components within the COIN architecture, along with the proposed improvements. It begins with a description of the dataset used, followed by the architectural and training modifications introduced in this work. Additionally, the underlying motivation for each change is discussed to clarify its intended role. The section concludes with an overview of the evaluation metrics used to assess model performance.

3.1 Motivation

The primary motivation behind this thesis is to explore counterfactual image generation in depth, using the COIN pipeline as the starting point. COIN was recently developed and later published by members of our lab (Biomedical Computer Vision Lab at the University of Tartu), and the present work benefits from direct access to its implementation and dataset, as well as guidance from the original authors. The method demonstrated strong performance in weakly supervised semantic segmentation on CT images for cancer detection, showing that inpainting-based approaches can effectively recover spatial localisation from image-level labels. Its success has highlighted the potential of counterfactual reasoning in medical image analysis. The objective is to enhance the pipeline by aligning its various components toward a unified high-level goal of tumour segmentation and improving its overall efficiency. Central to this work is the belief that, although it might seem that a simple one-bit label indicating whether an image contains a tumour or not offers limited information, it actually holds much more value when leveraged effectively by the model architecture.

3.2 Dataset

The Tartu University Hospital (TUH) kidney tumour dataset consists of 291 cases of contrast-enhanced kidney tumours and 300 control cases. The dataset has been annotated by five radiologists into three pixel-level label classes: kidney, malignant lesion, and benign lesion. Each set of annotations was reviewed by at least two other radiologists from the same group. Any discrepancies were resolved through direct discussion, and the final labels were created by merging the two sets of annotations. Image-level labels used for classifier training are derived from the presence of malignant lesion annotations in the scans. Additionally, slices from the 3D scans are selected based on the presence of kidney tissue, determined using a thresholding-based

method. The scans are randomly divided into training and validation sets using an 80/20 split, stratified by tumour volume per scan.

3.3 Classifier

The ability to segment images using counterfactual inpainting is highly dependent on the classifier, as it provides the signal that guides counterfactual image generation. Training a reliable classifier is therefore a necessary prerequisite for using this pipeline. The COIN paper [15] achieved its results using EfficientNet V2-S [19]. Since this thesis focuses primarily on modifying the counterfactual image generation process, rather than maximising the Intersection over Union (IoU) score, the same classifier architecture will be retained to enable fair comparisons between different pipeline configurations. The classifier is initialised with pre-trained ImageNet weights and fine-tuned on CT scan slices from the TUH dataset. The optimal training epoch is selected based on the validation loss and F1 [20] score.

The pre-trained classifier is frozen during the training of the counterfactual image generation model and does not undergo further updates. The classifier takes a scan slice X as input and outputs a prediction, which is then discretised into either a normal or abnormal class label. Each scan is classified twice: once before counterfactual image generation to determine the original label, and once after, to assess whether the counterfactual image successfully flipped the classifier's prediction.

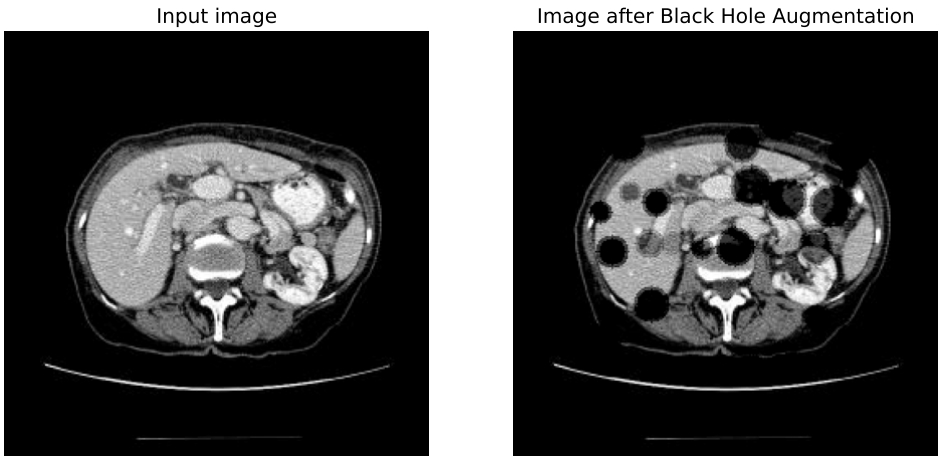

Figure 3. The effect of black hole augmentation. Black hole augmentation result (right) after being applied to the input image (left).

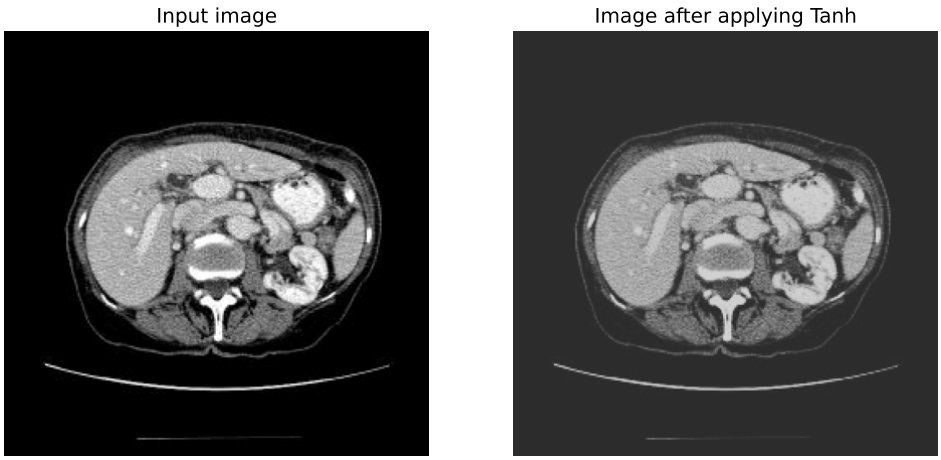

Figure 4. The effect of the hyperbolic tangent function. Hyperbolic tangent function result (right) after being applied to the input image (left).

As seen in Figure 2 and in the paper by Shvetsov et al. [15], COIN localises tumours by replacing them with black holes or by darkening the region. This behaviour may arise because it represents the simplest way to flip the classifier's prediction. Since the classifier has not encountered such artefacts in the training data, it may become confused and misclassify the image. To address this, this thesis proposes a data augmentation technique that introduces random circular regions with variable sizes and intensities, which are subtracted from the training images (Figure 3). This encourages the classifier to ignore such modifications and still correctly determine whether a tumour is present. The aim is to make the classifier more robust to these types of perturbations so that they do not trigger an adversarial effect, even when the modified region does not overlap with the tumour. This, in turn, would encourage the image generator to produce sharper and more confident inpaintings.

Another proposed augmentation involves applying the hyperbolic tangent function to all training images (Figure 4). This augmentation aims to better align the training and inference data distributions. During the normalisation step following perturbation in the counterfactual image generation pipeline, the underlying distribution of the image changes (as explained in Section 3.5). This change is architecturally necessary but might affect classifiers that, without the augmentation, are trained to operate on a different assumed data distribution, thereby impacting performance.

3.4 Turning cGAN Back to GAN

The counterfactual images are generated using a conditional GAN (cGAN) [12], which consists of an encoder, decoder, and discriminator. The discriminator is involved indirectly in the generation process, acting as an adversary to the encoder and decoder to encourage realistic image synthesis. The input image is first encoded into a vector representation, which is then used by the decoder to generate modifications based on the provided condition label c. This condition determines whether the generated image should resemble a normal or abnormal case. In this context, "abnormal" refers to an image with a tumour, while "normal" refers to an image without a tumour.

Since the ultimate goal of this pipeline is to segment tumours using a counterfactual approach, we focus solely on the condition that removes the tumour. The COIN paper already addresses this, but by only using one condition in the decoder. By removing the conditions altogether, effectively turning the cGAN back into a standard GAN, the decoder and discriminator can be further simplified. This modification also facilitates the implementation of other enhancements proposed in this thesis, such as the use of an alternative GAN loss and an integration between the classifier and GAN. Additionally, by eliminating the condition and leveraging discriminator alignment (as discussed in Section 3.6), the generator can be trained exclusively on abnormal scans, resulting in improved training efficiency. Since all of the pipelines described in this thesis require a classification model as a prerequisite, the segmentation pipeline can be used only on scans that have been classified as containing tumours.

3.5 Removing Architecture-Induced Bias

The COIN paper introduced perturbation-based techniques into the Singla et al. counterfactual image generation pipeline to modify only the semantically relevant regions of the image, rather than generating the entire image. This is achieved by adding the decoder's output to the original input image. In this way, the model only needs to generate the changes, not the whole image. Since the generated image is fed back into the classifier, with the expectation that it will be classified as normal, the model learns to effectively subtract the tumour from the image. In addition to the feedback from the classifier, the generator also receives guidance from the discriminator and the reconstruction loss. It is crucial that, during the addition of the input image and the generated modifications, the unchanged regions of the image remain intact.

$$X_{cf} = \tanh(X + G(Z(X))) \tag{1}$$

The addition method used in the COIN pipeline, which adds the decoder's output $G(Z(X))$ to the original input X, consists of two parts (Equation 1). The first part is the addition step, where the input image X and model output $G(Z(X))$ are summed elementwise. The second step involves normalising the sum to ensure that the result falls within the expected range. This is achieved by applying the hyperbolic tangent function $\tanh()$, which constrains the values to the range of -1 to 1. This is an important step, as without normalisation, the counterfactual image values would be unbounded, making it difficult to compute meaningful counterfactuals. Moreover, if the counterfactual image lies outside the expected input range, the classifier may be unable to process it reliably, as it assumes a fixed input distribution.

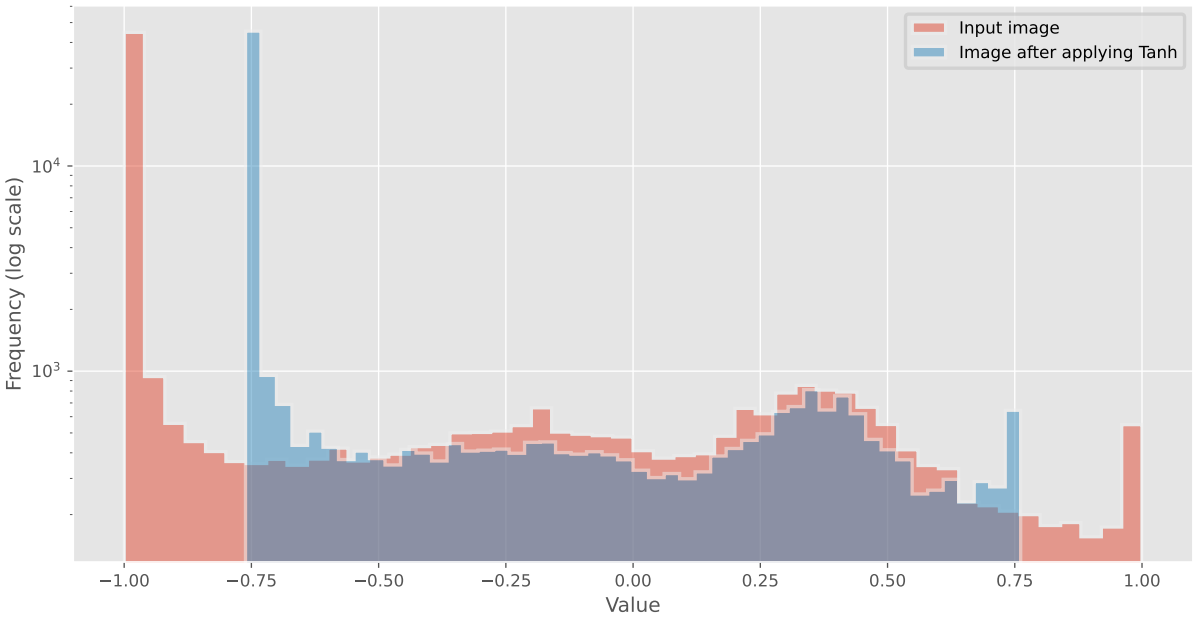

Figure 5. Distribution of image values. Histogram showing image distribution before (red) and after (blue) applying the hyperbolic tangent function.

However, applying the hyperbolic tangent function also modifies the distribution of the underlying original image by compressing it into a narrower range (Figure 5). This introduces bias that can be easily detected by both the discriminator and the reconstruction loss. As a result, the generator is placed at an inherent disadvantage, as it must learn how to generate outputs that,

when added to the input image and subsequently passed through the hyperbolic tangent function, still resemble the distribution of real images.

The solution proposed and implemented in this thesis is to apply the hyperbolic tangent function also to the real images, which are used for comparison by the discriminator and for difference map calculation. In this way, all images involved in any comparison are processed uniformly through the hyperbolic tangent function, thereby eliminating architecture-induced bias.

3.6 Discriminator Alignment

Since the discriminator evaluates the realism of the images, this thesis proposes aligning its objectives with those of the entire pipeline. In the COIN setup, the discriminator compares the input image used for counterfactual generation with the generated result during each training iteration. This gives the discriminator an inherent advantage over the generator, as there will always be visible changes when inpainting is expected.

While it could be argued that the COIN pipeline incorporates the condition c, and thus the images being compared are assessed differently, this argument no longer holds after the removal of conditions, as described in Section 3.4. This becomes even more evident in one of the modifications proposed in this thesis (see Section 3.7.4), where the loss function is changed to a more relativistic form.

This thesis proposes that the discriminator should compare the generated counterfactual image to a healthy scan, thereby aligning its objective with the overall pipeline. By having the discriminator evaluate generator outputs against healthy (tumour-free) scan slices, the generator is encouraged to produce realistic healthy kidneys rather than inpainting tumours with black holes. Additionally, the randomly selected healthy scan slice is chosen from a similar relative height, as slices taken from different heights can exhibit significant anatomical and pathological variation. This height-based matching is intended to help the discriminator focus on meaningful differences between healthy and abnormal tissue. Without this constraint, anatomical differences across varying heights could distract the discriminator and force it to interpret complex 3D structures from a limited 2D viewpoint.

3.7 Communicating Goals to the Model via Loss

The loss function is the foundation of every machine learning model, as it defines what the model is optimising for. In the case of GANs, it's a minimax game between the generator and the discriminator, where each tries to outperform the other, leading to meaningful learning as a

by-product. However, this adversarial push and pull can make GAN training unstable. If one component becomes too dominant, it can cause the loss function to diverge or lead to mode collapse.

The COIN architecture introduces even more complexity, as the model also aims to flip the classifier's prediction to generate counterfactual images without tumours. This additional objective is incorporated through an extra loss term, and the model optimises a linear combination of several losses, as shown in Equation (2). The resulting total loss $\mathcal{L}_{total}$ includes the adversarial GAN loss $\mathcal{L}_{GAN}$, the classifier loss $\mathcal{L}_f$, which is computed as the Kullback–Leibler (KL) divergence between the classifier's prediction and the desired outcome, and two regularisation terms: the reconstruction loss $\mathcal{L}_{rec}$ and the total variation loss $\mathcal{L}_{tv}$. Each loss term is scaled by a corresponding weight λ in the linear combination.

$$\mathcal{L}_{total} = \lambda_{GAN}\mathcal{L}_{GAN} + \lambda_f\mathcal{L}_f + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{tv}\mathcal{L}_{tv} \tag{2}$$

3.7.1 Aiming Without a Clear Target

If one loss term is weighted more heavily, the model tends to prioritise that objective and learn in that direction more rapidly than others. Additionally, the COIN pipeline is not directly optimising for the objective on which it is ultimately evaluated. The core GAN loss encourages the model to generate images that closely resemble those presented to the discriminator. At the same time, the generated images are fed to the classifier, which is expected to classify them as normal (i.e., without tumours). However, this does not necessarily ensure that the image is truly tumour-free, as classifiers can be deceived by adversarial-like images.

The ideal outcome is for the generated image to appear realistic while also being classified as normal. In traditional segmentation tasks, the loss function compares the model's output against discrete segmentation masks, explicitly indicating whether the entire tumour was detected or partially missed. In the current case, there is no clear discrete target metric, so the training relies entirely on balancing the different loss components to find the optimal trade-off.

If the balance favours image realism too strongly, the classifier's requirements may be ignored. On the other hand, if satisfying the classifier becomes the main goal, the generator might produce images that no longer resemble realistic human anatomy. Furthermore, the COIN pipeline includes reconstruction loss and total variation loss, which aim to help improve image

realism and maintain smoothness in the segmentation maps, but can also suppress meaningful counterfactual changes if not properly balanced.

3.7.2 Reconstruction Loss

In the COIN architecture, reconstruction loss is used to encourage the GAN model to generate images that closely resemble the input. This is essential for enabling the use of the counterfactual image, denoted as $\mathcal{E}(X)$, to produce a pixel-wise difference map that highlights regions of abnormality, such as tumours. The mean absolute error (MAE) loss, denoted by $\mathcal{L}_{mae}$, is used to calculate the difference between the unchanged input image and the generated counterfactual. It is computed as:

$$\mathcal{L}_{mae}(X, \hat{X}) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} |x_{i,j} - \hat{x}_{i,j}|$$

where H and W denote the height and width of the image, respectively.

Furthermore, the reconstruction loss is applied in a cyclical manner. This means that the generator performs two forward passes: one to produce the counterfactual image $\mathcal{E}(X)$, and another resulting in $\mathcal{E}(\mathcal{E}(X))$. The total reconstruction loss is therefore defined as:

$$\mathcal{L}_{rec} = \mathcal{L}_{mae}(X, \mathcal{E}(X)) + \mathcal{L}_{mae}(X, \mathcal{E}(\mathcal{E}(X)))$$

The author of this thesis argues that, after modifying the perturbation mechanism (see Section 3.5) in the generator, the cyclical reconstruction loss becomes redundant and adds unnecessary complexity to the overall architecture. Using MAE as the reconstruction loss function penalises any changes made to the counterfactual image, which can suppress the signal coming from the classifier. Additionally, introducing another loss term increases the complexity of balancing all the losses and makes it more difficult to achieve stable convergence during training. However, an alternative reconstruction loss was explored, focusing on penalising overly sparse predictions.

The proposed reconstruction loss calculates a difference map between the input image and the generated counterfactual, which is then dilated using a square kernel. The resulting array is summed and divided by the total number of pixels, normalising the loss between 0 and 1. Including dilation in the loss term ensures that sparsely spaced predictions are penalised more heavily than clustered ones. Furthermore, the loss is not used cyclically, so there is no need to perform inference twice with the image generation model. Nevertheless, this loss still penalises any differences between the input and the counterfactual. These modifications to

the reconstruction loss were explored prior to removing the architecture-induced bias by the perturbation mechanism, so it was still required to generate realistic counterfactuals.

3.7.3 Total Variation Loss

In addition to the reconstruction loss, the COIN architecture also uses total variation (TV) loss to improve the smoothness of the segmentation masks. The authors of the COIN paper state that TV loss is used as a regularisation term to enhance consistency in the difference maps and, in doing so, encourages the model to perturb only densely located regions. It can still be argued that TV loss also pushes the generated counterfactual images to remain similar to the input, in order to minimise differences in the difference map. However, since the way the loss $\mathcal{L}_{tv}$ is calculated (see Equation 3), it focuses more on the differences between neighbouring pixels, making it particularly effective at suppressing random noise. As a result, the model is encouraged to make changes only where necessary, rather than introducing scattered predictions that might confuse the classifier. For this reason, the TV loss was not modified during the experiments in this thesis and is included in all evaluation pipelines.

$$\mathcal{L}_{tv} = \frac{1}{HW}\left(\sum_{i=1}^{H-1}\sum_{j=1}^{W}\left(x_{i+1,j} - x_{i,j}\right)^2 + \sum_{i=1}^{H}\sum_{j=1}^{W-1}\left(x_{i,j+1} - x_{i,j}\right)^2\right) \tag{3}$$

3.7.4 Relativistic GAN Loss with Gradient Normalisation

The paper by Huang et al. [21] proposes a modern loss function designed to circumvent non-convergence and mode collapse, as an alternative to relying on ad-hoc tricks for stabilising GAN training. The COIN pipeline uses binary cross-entropy (BCE) as its discriminator loss function. These two losses compare generated and real images in fundamentally different ways. BCE-based loss assesses images independently, comparing the discriminator's output to an expected discrete outcome, similar to how a classifier operates. In contrast, the loss proposed by Huang et al. builds on the idea of RpGAN by Jolicoeur-Martineau et al. [22], which calculates loss based on the difference between the discriminator outputs for real and generated samples. Huang et al. propose the addition of gradient normalisation to prevent gradient explosion and ensure convergence. This approach is more relativistic, resulting in a smoother and more continuous decision boundary. As part of the experiments, this new GAN loss was implemented and evaluated.

3.8 Classifier-Driven Conditioning via Latent Integration

In the COIN architecture, the classifier and the counterfactual image generator operate independently and do not share information during inference. During training, the GAN model receives an input image and generates a counterfactual. This counterfactual is then evaluated by the classifier, which provides feedback to the GAN through back-propagation. Based on the classifier's output, a KL divergence loss term $\mathcal{L}_f$ is calculated and incorporated into the GAN's overall loss function, as shown in formula 2.

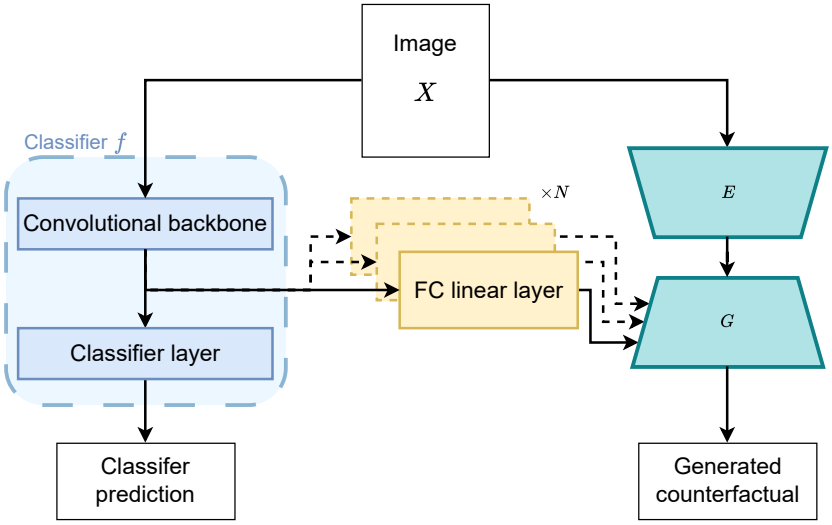

Figure 6. High-level overview of the proposed connection between the classifier and the counterfactual generator.

In this thesis, a connection between the classifier and the GAN is proposed by injecting the classifier's penultimate layer activations into the conditional batch normalisation (CBN) layers of the GAN model. The CBN layers were originally used in the cGAN framework to incorporate conditional embeddings for accommodating a given condition c. Depending on the condition, different embeddings were used as weights for the CBN layers. In this work, those conditional embeddings are replaced with the classifier's penultimate layer activations. These activations are not injected directly; instead, they are passed through a fully connected linear layer to account for variations in embedding size, as the dimensions of the conditional embeddings differ depending on the convolutional layers used (Figure 6). Additionally, using a fully connected linear layer as a bridge gives the model greater flexibility in determining when and where to use the classifier signal.

3.9 Evaluation Metrics

To evaluate the counterfactual image generator, this thesis uses both quantitative and qualitative methods. Intersection over Union (IoU) and Pixel-wise Error Rate (PER) measure localisation accuracy based on ground truth tumour masks. Since these metrics may not fully capture visual quality or stylistic differences in inpainting, visual inspection is also included to provide complementary insight.

3.9.1 Intersection over Union

Intersection over Union is used throughout this thesis as a quantitative metric to evaluate how effectively the counterfactual image generator localises abnormal regions. To enable this, both the original and counterfactual images are first shifted from the $[-1, 1]$ range to $[0, 1]$. A difference map is then computed by taking the pixel-wise absolute difference between the two images. This map is binarised using a fixed threshold of 0.25, under the assumption that pixels with greater changes are more likely to indicate altered regions related to the classifier's decision, and to filter out minimal changes that might be artefacts. The IoU is then calculated between the resulting binary masks using the standard formulation shown in Equation 4.

$$\text{IoU}(S, S_c) = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{4}$$

In this context, S denotes the ground truth mask and S_c represents the predicted mask, where the terms TP, FP, and FN correspond to true positive, false positive, and false negative predictions, respectively.

3.9.2 Pixel-Wise Error Rate

Pixel-wise Error Rate is employed in this thesis as a supplementary quantitative metric alongside IoU to assess the behaviour of the counterfactual image generator. Specifically, with low IoU scores, it helps evaluate whether the generator is making small, targeted modifications or simply inpainting large regions.

$$\text{Pixel-wise Error Rate} = \frac{\text{FP} + \text{FN}}{\text{Total Pixels}} \tag{5}$$

To compute this, both the original and counterfactual images are first normalised from the $[-1, 1]$ range to $[0, 1]$. A binary difference mask is then created by thresholding the pixel-wise absolute difference at 0.25, consistent with the IoU preprocessing step. This binary mask is compared

against the corresponding ground truth mask, and the total number of misclassified pixels is summed. The pixel-wise error rate is then defined as the ratio of these misclassified pixels to the total number of pixels in the image, as shown in Equation 5. A lower error rate with a high IoU score indicates more precise localisation, whereas a high value may suggest that the model is altering irrelevant regions.

3.9.3 Visual Evaluation

In addition to quantitative evaluation using IoU and PER, qualitative visual inspection was performed on a subset of the test set to better understand the inpainting strategies used by different models. One generated counterfactual image per batch was selected for manual review totalling in 136 images per model. Since this work does not involve expert medical interpretation, the goal of visual inspection was not to evaluate anatomical accuracy, but rather to assess how well the generated images mimicked the original structure and texture of the input, and how the model chose to indicate abnormal regions. Particular attention was paid to whether the model inpainted tumours by removing them entirely (e.g., with black holes) or subtly highlighting them. This process also helped identify issues such as artefacts, blurring, or unrealistic shading patterns that would not be captured by numerical metrics alone. Moreover, because both IoU and PER rely on thresholding difference maps, which assume fixed intensity changes, these metrics may not fairly compare models that use different inpainting styles or operate over different value ranges. Visual inspection thus provided crucial complementary insight, especially in cases where high or low metric scores might not align with the apparent quality or interpretability of the generated outputs.

4. Experiments and Results

This section describes the experimental setup, followed by an overview of the conducted experiments and their results. An overall analysis is provided at the end, summarising the key findings from all experiments carried out in this thesis.

4.1 Experimental Setup

All models were developed using PyTorch [23] and trained on the Large Unified Modern Infrastructure (LUMI) High Performance Computing (HPC) cluster, which is located in Kajaani, Finland. Training was done on a single AMD MI250X GPU with 64 GB of memory. The input slices are selected with at least 100 pixels in the kidney mask area and are resized to 256x256 format using bilinear interpolation. The images are then grouped into batches of 256 images for classifier training and batches of 16 for the counterfactual image generator training. The code used in this thesis is based on a fork of the original COIN implementation [24] from the paper by Shvetsov et al. [15]. The modified version, incorporating all changes made during this research, is available at:

https://github.com/Markoxyz/COIN_Aligned

4.2 Experimental Results

The experiments conducted in this thesis were carried out in an iterative manner, as exhaustively evaluating all possible combinations of model modifications would have been computationally prohibitive. Additionally, some experiments depended on prior modifications, making a staged approach both practical and necessary. The final experimental setup is divided into three main subgroups. The first subgroup explores the impact of a different reconstruction loss function, as the reconstruction loss is not included in any of the other experiments. The second focuses on modifications to the image generation pipeline, and the third on changes to the classifier fine-tuning. Modifications to the generation pipeline and classifiers are cumulative. Meaning each successive change builds upon the previous ones, ultimately resulting in the proposed aligned pipeline for counterfactual image generation in tumour detection. For comparison, the COIN model was also trained as a baseline.

All models were trained for at least 100 epochs, with the best-performing epoch selected for evaluation based on the highest average IoU score on the test set. The test set consists of 2,191 tumour-containing images, and evaluation was performed exclusively on this subset. Quantitative evaluation was conducted using IoU and PER as primary metrics. In addition, qualitative analysis

was conducted by visually inspecting a subset of the generated counterfactuals. Visual inspection was performed by selecting one image per batch, totalling 136 counterfactuals per pipeline, which is about 1/16th of the test set.

4.2.1 COIN Baseline

The baseline model was implemented using the original configuration provided in the COIN paper's GitHub repository [24]. It achieved an average IoU of 0.343 at epoch 39, with a standard deviation of 0.205. The mean PER was 0.019, with a standard deviation of 0.011. Visual inspection showed that the model consistently inpainted tumours by generating black holes. The counterfactuals retained textures with slight blurring (see Figure 7) but still closely resembled the original images. Some anatomical shapes appeared distorted, and occasional artefacts were present in the generated outputs.

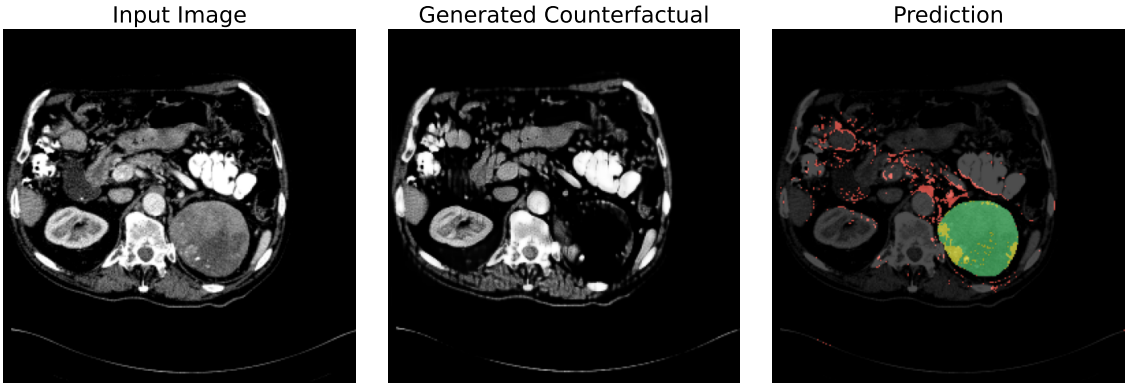

Figure 7. Model Results: Input image (left), generated counterfactual (middle), and thresholded predictions overlaying the input image (right). Green denotes correct predictions, yellow indicates missed ground truth, and red represents misclassified pixels.

4.2.2 Modified Reconstruction Loss with Dilation

To improve the generation of realistic counterfactual images in the COIN architecture, an alternative reconstruction loss was proposed. Unlike the original cyclical MAE-based loss, this new formulation penalises overly sparse difference maps by applying dilation with a square kernel, aiming to enhance the spatial coherence of predictions without requiring a second inference pass.

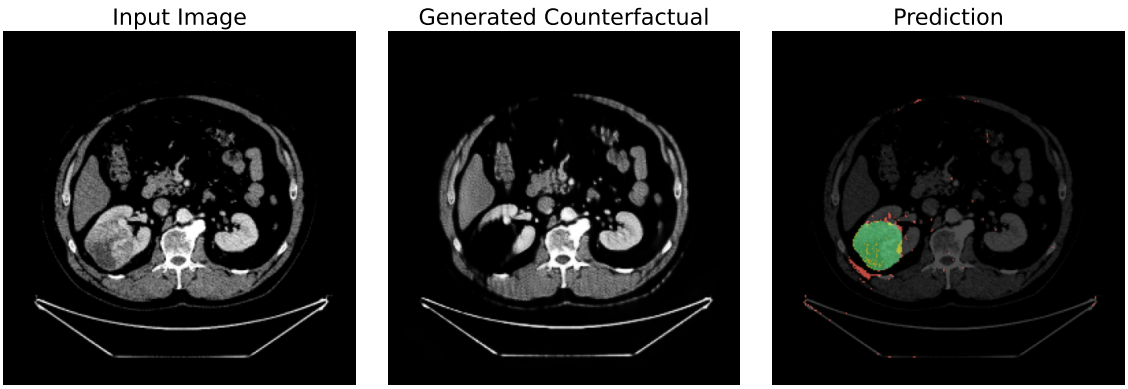

Figure 8. Model Results: Input image (left), generated counterfactual (middle), and thresholded predictions overlaying the input image (right). Green denotes correct predictions, yellow indicates missed ground truth, and red represents misclassified pixels.

Using a dilation kernel of 3 pixels, the model achieved an average IoU of 0.368 with a standard deviation of 0.222 at epoch 18. The average PER was 0.016 with a standard deviation of 0.010. Visual inspection indicated that while the predictions were denser, the generated counterfactuals exhibited vertical line artefacts (see Figure 8), potentially introduced by the square dilation kernel. The model consistently inpainted tumours by generating black holes.

4.2.3 Uniform Transformation for Distribution Alignment

To address distributional bias introduced by the hyperbolic tangent $tanh()$ function in the COIN pipeline, this experiment applied the $tanh()$ transformation uniformly to both real and generated images. This ensures that all images involved in comparisons, such as for the discriminator or total variation loss, share the same value range, thereby reducing architecture-induced bias and allowing the generator to focus on semantically meaningful modifications. It is important to note that this experiment, along with all subsequent experiments, is conducted without reconstruction loss.

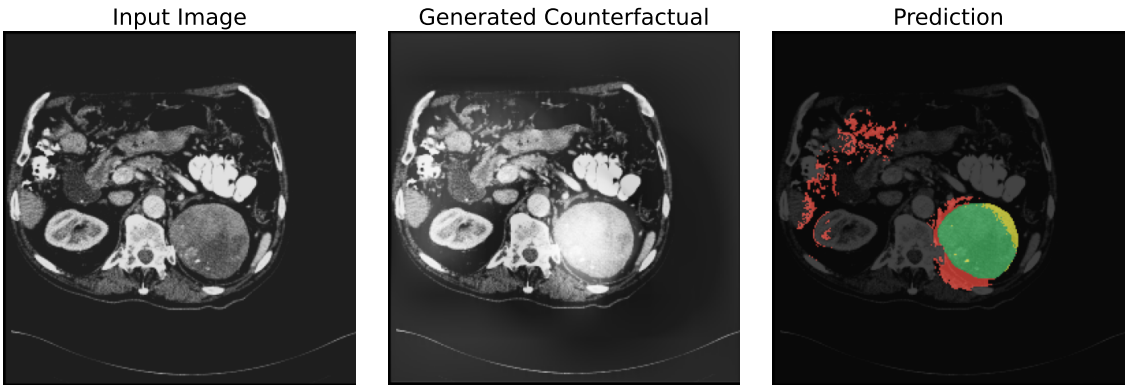

Figure 9. Model Results: Input image (left), generated counterfactual (middle), and thresholded predictions overlaying the input image (right). Green denotes correct predictions, yellow indicates missed ground truth, and red represents misclassified pixels.

With this adjustment, the model achieved a highest average IoU of 0.220 at epoch 37, with a standard deviation of 0.216. The mean PER was 0.040, with a standard deviation of 0.040. Visual inspection showed that the model began localising tumours by highlighting regions (see Figure 9) instead of generating blacked-out areas. The generator's outputs accurately preserved organ shapes and textures, modifying only intensity, consistent with the overlay-based perturbation mechanism. However, the inpaintings themselves appeared broad and lacked the sharp borders that would be ideally expected for high-quality predictions.

4.2.4 Discriminator Alignment with Anatomically Paired Data

This experiment combines two modifications to the COIN pipeline. Firstly, the removal of the condition label from the cGAN architecture to simplify the decoder and discriminator. Secondly, alignment of the discriminator's objective by comparing generated counterfactuals to real healthy kidney scans taken from similar anatomical heights. These changes aim to simplify training and encourage the generator to produce realistic, tumour-free counterfactuals by focusing discriminator feedback on meaningful visual differences.

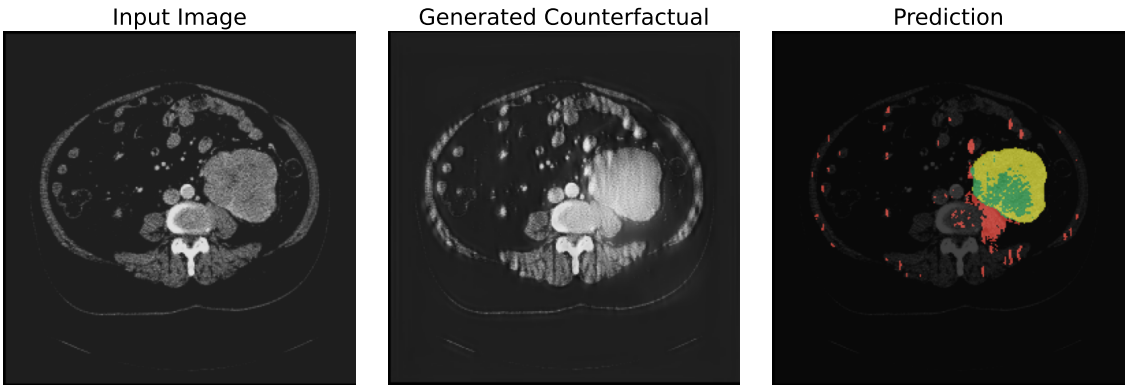

Figure 10. Model Results: Input image (left), generated counterfactual (middle), and thresholded predictions overlaying the input image (right). Green denotes correct predictions, yellow indicates missed ground truth, and red represents misclassified pixels.

The model achieved its highest average IoU of 0.157 at epoch 75, with a standard deviation of 0.148. The mean PER was 0.022, with a standard deviation of 0.014. Visual inspection showed that the generator preserved the shapes and borders of organs well, but the textures appeared blurred and contained unnatural zebra-like patterns (see Figure 10). The model inpainted tumour regions by highlighting, which remained broad and lacked clearly defined borders, similar to the previous experiment. Furthermore, there were no clear attempts by the model to generate healthy kidneys instead of tumours.

4.2.5 Relativistic GAN Loss with Gradient Normalisation

This experiment evaluates a modern GAN loss function proposed by Huang et al., which replaces the BCE loss used in the COIN pipeline. Unlike BCE, which treats real and generated images independently, the new loss introduces a relativistic comparison between the two, with gradient normalisation to ensure stability and convergence. The aim is to achieve smoother decision boundaries and more stable training.

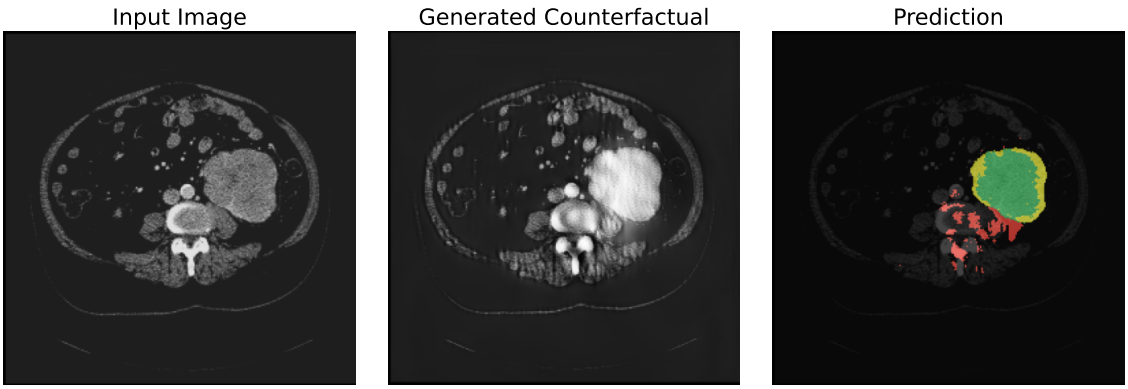

Figure 11. Model Results: Input image (left), generated counterfactual (middle), and thresholded predictions overlaying the input image (right). Green denotes correct predictions, yellow indicates missed ground truth, and red represents misclassified pixels.

The model achieved an average IoU of 0.274 at epoch 79, with a standard deviation of 0.173. The mean PER was 0.026, with a standard deviation of 0.010. Visual inspection showed that the model still continues to highlight tumour regions instead of blacking them out. The resulting counterfactual images had preserved organ shapes and sharp boundaries, but the textures appeared cartoonish, smudged, and blurred (see Figure 11). Zebra-like artefacts similar to those in previous experiments were present, and the highlighted tumour regions remained broad and lacked sharp, localised borders.

4.2.6 Classifier-Driven Conditioning via Latent Integration

This experiment investigates a tighter integration between the classifier and the GAN by replacing the conditional embeddings in the generator's conditional batch normalisation (CBN) layers with activations from the classifier's penultimate layer. These activations are passed through a fully connected linear layer before injection, allowing the generator to leverage classifier-informed context during image synthesis.

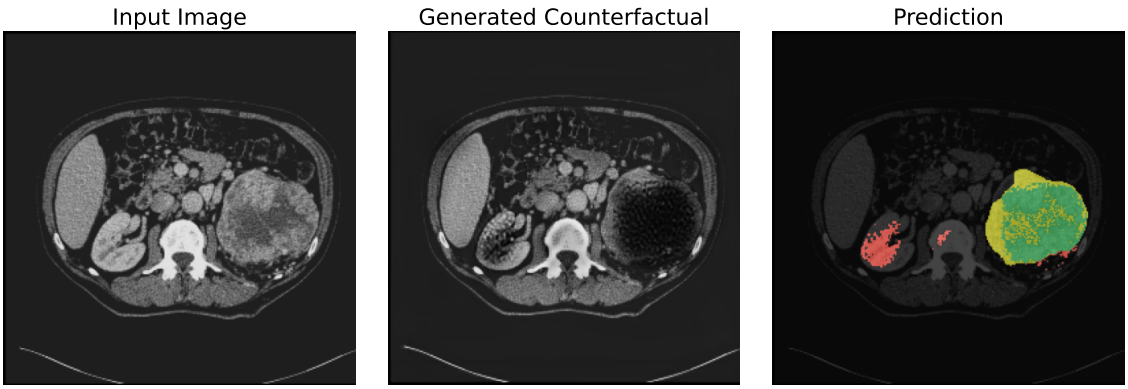

Figure 12. Model Results: Input image (left), generated counterfactual (middle), and thresholded predictions overlaying the input image (right). Green denotes correct predictions, yellow indicates missed ground truth, and red represents misclassified pixels.

The model achieved an average IoU of 0.238 at epoch 88, with a standard deviation of 0.199. The mean PER was 0.020, with a standard deviation of 0.012. Visual inspection showed that the model often indicated tumour presence by generating sparse black holes, and occasionally by highlighting affected regions. The generated images retained accurate organ contours and shapes. Compared to the previous experiment without classifier-GAN integration, textures appeared more realistic and less cartoonish or smudged. However, intricate textures were still not fully recovered. The generated holes had sharp edges but were not unified, appearing instead as scattered small voids (see Figure 12). In several instances, both kidneys were perturbed, suggesting possible mode collapse.

4.2.7 Classifier with Distribution Normalisation

This experiment evaluates the counterfactual inpainting pipeline using a pre-trained EfficientNet V2-S classifier fine-tuned on CT scan slices from Tartu University Hospital (TUH). The classifier, frozen during GAN training, guides counterfactual generation by providing prediction feedback before and after image perturbation. Additionally, a hyperbolic tangent transformation is applied uniformly to all training images to align data distributions between training and inference to minimise architectural bias.

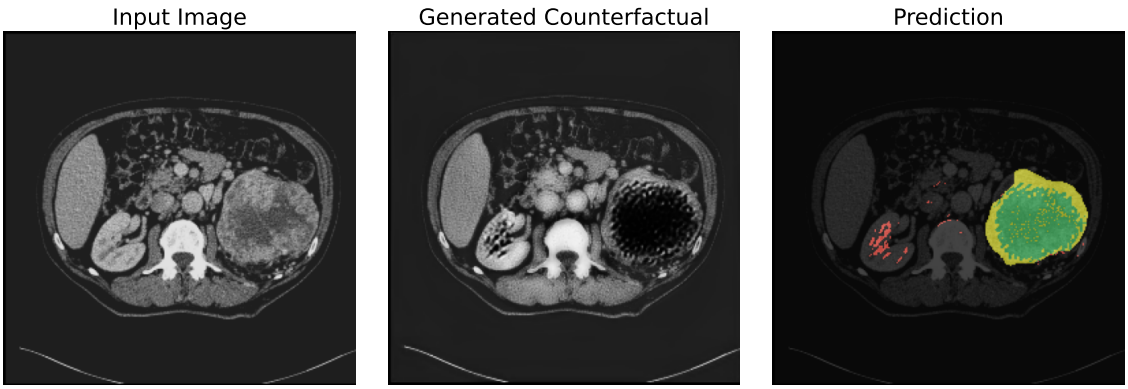

Figure 13. Model Results: Input image (left), generated counterfactual (middle), and thresholded predictions overlaying the input image (right). Green denotes correct predictions, yellow indicates missed ground truth, and red represents misclassified pixels.

The model achieved an average IoU of 0.351 at epoch 74, with a standard deviation of 0.178. The mean PER was 0.016, with a standard deviation of 0.009. Visual inspection showed that the model consistently localised tumours by generating black holes, while unrelated highlighting appeared in areas near the spine (see Figure 13). The generated images retained accurate shapes and borders of organs, with textures that matched the original images where no inpaintings were applied. The highlighted regions, however, exhibited blurred textures with uniform intensities.

4.2.8 Robust Classifier via Augmentation and Distribution Alignment

In this experiment, a data augmentation technique was applied to improve the robustness of the classifier, which guides counterfactual image generation in the COIN pipeline. Specifically, random circular regions with variable sizes and intensities were subtracted from training images to prevent the classifier from being misled by black holes introduced during counterfactual generation. Additionally, a hyperbolic tangent function was applied to align the training and inference data distributions. These modifications aimed to encourage the generator to produce sharper and more realistic inpaintings.

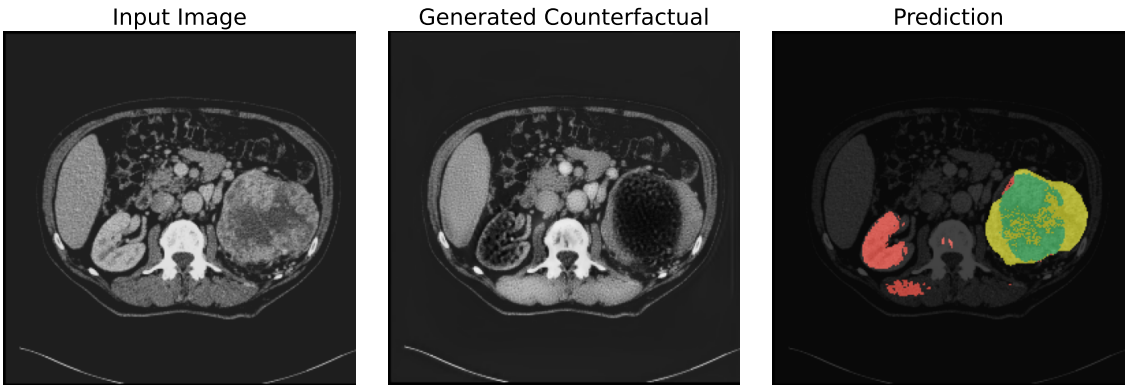

Figure 14. Model results: Input image (left), generated counterfactual (middle), and thresholded predictions overlaying the input image (right). Green denotes correct predictions, yellow indicates missed ground truth, and red represents misclassified pixels.

The model achieved an average IoU of 0.260 at epoch 87, with a standard deviation of 0.151. The mean PER was 0.028, with a standard deviation of 0.012. Visual inspection showed that the model consistently indicated tumours by generating black holes. However, it also highlighted regions near the spine, which did not match the textures of the original image. The shapes and borders of organs were closely matched, but the generated regions appeared darker and larger than before (see Figure 14), suggesting that the classifier's increased resilience to black holes may have led the model to generate even bigger, darker regions in an attempt to fool the classifier.

4.3 Overall Analysis

The experiments in this thesis assess modifications to the COIN architecture aimed at improving the quality of counterfactual image generation. IoU and PER scores varied across configurations, while experiments 1 and 6 in Table 1 achieved better average IoU and PER scores than the reproduced baseline. However, as the original baseline performance reported in the COIN paper could not be replicated, it is not possible to definitively conclude that the proposed approach outperforms the original. Nonetheless, experiment 6 achieved comparable results without using reconstruction loss, while also significantly reducing training time. The COIN baseline reached 100 epochs in approximately 6 days, whereas the "Classifier Guidance with Distribution Normalisation" configuration completed training in approximately 2.2 days.

Visual inspection of 136 counterfactuals per experiment showed that different configurations produced varying inpainting styles. Some models generated sharper segmentations but with

visual artefacts, while others produced more realistic textures with lower metric scores. To illustrate this point, a side experiment was conducted using the ”Classifier-Driven Conditioning via Latent Integration” model. This model was re-evaluated at epoch 90, with the IoU and PER binarisation threshold lowered from 0.25 to 0.15. As a result, the IoU score increased to 0.304, with a PER of 0.021, which is better than the reported score for this experiment (see Table 1 experiment 5). This demonstrates that performance assessment of these models cannot rely solely on strict quantitative metrics. Additionally, the figures for each experiment were selected to highlight key visual findings. All pipelines (including the COIN baseline) share a failure case where nothing is inpainted or highlighted, suggesting that the model sometimes fails to recognise the tumour. This likely results from classifier errors, which propagate to the counterfactual generator that depends on its performance.

Table 1. Experimental results based on IoU and PER metrics. The star symbol ($\star$) indicates that this modification was not included in subsequent pipelines, while others build upon each other.

Index	**Experiment name**	**IOU ↑**	**PER ↓**
0	COIN (baseline)	0.343	0.019
1	Modified Reconstruction Loss with Dilation $\star$	0.368	0.016
2	Distribution Alignment through Normalisation	0.220	0.040
3	Discriminator Alignment with Anatomically Paired Data	0.157	0.022
4	Relativistic GAN Loss with Gradient Normalisation	0.274	0.026
5	Classifier-Driven Conditioning via Latent Integration	0.238	0.020
6	Classifier Guidance with Distribution Normalisation	0.351	0.016
7	Robust Classifier via Augmentation and Distribution Alignment	0.260	0.028

5. Conclusion

The motivation for this thesis stems from the growing need to enhance cancer detection in medical imaging through more efficient and scalable means. As cancer incidence continues to rise, so too does the demand on radiologists to interpret an increasing volume of complex imaging data. Traditional approaches to training artificial intelligence models for tumour segmentation rely heavily on pixel-level annotations, which are both time-consuming and resource-intensive to produce. To address this, the thesis investigates weakly supervised learning as a feasible alternative, utilising image-level labels to train segmentation models with minimal reliance on manual annotation. Specifically, the research focuses on improving alignment within an existing counterfactual image generation pipeline, with the aim of making better use of image-level labels to support weakly supervised learning in a manner that is both practical and scalable.

We developed a more aligned tumour segmentation pipeline that achieved comparable results to COIN while using only a third of the training time. The inability to reproduce the results reported in the COIN paper is likely due to differences in classifier selection, as the model appears to be highly sensitive to this component. But with the modification of the perturbation normalisation, the model no longer has to re-create the entire image and can just offer its inputs as an overlay on top of the input images. This finding underscores the importance of careful selection and application of normalisation functions in perturbation-based image generation to minimise the risk of reduced learning efficiency.

5.1 Limitations and Future works

After modifying nearly every component of the pipeline except the encoder, it appears that the encoder may be its weakest link. This limitation is closely tied to the pipeline's reliance on the classifier, as both depend on semantically meaningful image embeddings to perform effectively. If the encoder fails to produce a meaningful vector representation of the input, the performance of all downstream tasks is negatively affected. Furthermore, having an encoder capable of generating semantically rich embeddings would open the door to developing improved evaluation metrics for medical counterfactual images. As it turned out during the evaluation, strict quantitative metrics may fall short when assessing texture fidelity and fine-grained details in generated counterfactuals. This is particularly important to perform hyperparameter tuning in a reliable manner. A stronger encoder could therefore enable not only better generation and classification but also support the development of more nuanced, embedding-based evaluation criteria that would align with clinical or visual expectations.

6. Acknowledgements

I would like to express my gratitude to my supervisors, Joonas Ariva and Dmytro Fishman, for their guidance, support and constructive feedback throughout the course of this research. I am also deeply thankful to everyone at the Biomedical Computer Vision Lab, with whom I had the pleasure of engaging in insightful discussions that greatly enriched my academic experience. I would also like to acknowledge the team at the High Performance Computing Center of the University of Tartu for their support in accessing and managing hardware resources and for generously sharing their expert knowledge.

References

[1] Crosby D., Bhatia S., Brindle K. M., Coussens L. M., Dive C., Emberton M., Esener S., Fitzgerald R. C., Gambhir S. S., Kuhn P., Rebbeck T. R., and Balasubramanian S. Early detection of cancer. *Science* 375.6586 (2022), eaay9040. DOI: 10.1126/science.aay9040. https://www.science.org/doi/10.1126/science.aay9040.

[2] World Health Organization. Global Cancer Burden Growing Amidst Mounting Need for Services. Accessed: 2025-05-10. World Health Organization. Feb. 2024. https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services.

[3] Better Medicine. BMVision Kidney: AI Kidney Cancer Detection Software. Accessed: 2025-05-10. 2025. https://bettermedicine.ai/bmvision/kidney/.

[4] Morshid A., Duran E. S., Choi W. J., and Duran C. A Concise Review of the Multimodality Imaging Features of Renal Cell Carcinoma. *Cureus* 13.2 (Feb. 2021). Published 2021 Feb 8, e13231. DOI: 10.7759/cureus.13231. https://doi.org/10.7759/cureus.13231.

[5] Zhou B., Khosla A., Lapedriza A., Oliva A., and Torralba A. Learning Deep Features for Discriminative Localization. 2015. arXiv: 1512.04150 [cs.CV]. https://arxiv.org/abs/1512.04150.

[6] Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., and Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. DOI: 10.1007/s11263-019-01228-7. http://dx.doi.org/10.1007/s11263-019-01228-7.

[7] Wang H., Wang Z., Du M., Yang F., Zhang Z., Ding S., Mardziel P., and Hu X. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. 2020. arXiv: 1910.01279 [cs.CV]. https://arxiv.org/abs/1910.01279.

[8] Kirillov A., Mintun E., Ravi N., Mao H., Rolland C., Gustafson L., Xiao T., Whitehead S., Berg A. C., Lo W.-Y., Dollár P., and Girshick R. Segment Anything. 2023. arXiv: 2304.02643 [cs.CV]. https://arxiv.org/abs/2304.02643.

[9] Kweon H. and Yoon K.-J. From SAM to CAMs: Exploring Segment Anything Model for Weakly Supervised Semantic Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 19499–19509.

[10] Isensee F., Rokuss M., Krämer L., Dinkelacker S., Ravindran A., Stritzke F., Hamm B., Wald T., Langenberg M., Ulrich C., Deissler J., Floca R., and Maier-Hein K. nnInteractive:

Redefining 3D Promptable Segmentation. 2025. arXiv: 2503.08373 [cs.CV]. https://arxiv.org/abs/2503.08373.

[11] Dhurandhar A., Chen P.-Y., Luss R., Tu C.-C., Ting P., Shanmugam K., and Das P. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. 2018. arXiv: 1802.07623 [cs.AI]. https://arxiv.org/abs/1802.07623.

[12] Singla S., Eslami M., Pollack B., Wallace S., and Batmanghelich K. Explaining the Black-box Smoothly- A Counterfactual Approach. 2022. arXiv: 2101.04230 [cs.CV]. https://arxiv.org/abs/2101.04230.

[13] Ribeiro M. T., Singh S., and Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 2016. arXiv: 1602.04938 [cs.LG]. https://arxiv.org/abs/1602.04938.

[14] Goodfellow I. J., Shlens J., and Szegedy C. Explaining and Harnessing Adversarial Examples. 2015. arXiv: 1412.6572 [stat.ML]. https://arxiv.org/abs/1412.6572.

[15] Shvetsov D., Ariva J., Domnich M., Vicente R., and Fishman D. COIN: Counterfactual Inpainting for Weakly Supervised Semantic Segmentation for Medical Images. *Explainable Artificial Intelligence*. Springer Nature Switzerland, 2024, pp. 39–59. DOI: 10.1007/978-3-031-63800-8_3. http://dx.doi.org/10.1007/978-3-031-63800-8_3.

[16] Javanmardi M., Sajjadi M., Liu T., and Tasdizen T. Unsupervised Total Variation Loss for Semi-supervised Deep Learning of Semantic Segmentation. 2018. arXiv: 1605.01368 [cs.CV]. https://arxiv.org/abs/1605.01368.

[17] Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., and Bengio Y. Generative Adversarial Networks. 2014. arXiv: 1406.2661 [stat.ML]. https://arxiv.org/abs/1406.2661.

[18] Mirza M. and Osindero S. Conditional Generative Adversarial Nets. 2014. arXiv: 1411.1784 [cs.LG]. https://arxiv.org/abs/1411.1784.

[19] Tan M. and Le Q. V. EfficientNetV2: Smaller Models and Faster Training. 2021. arXiv: 2104.00298 [cs.CV]. https://arxiv.org/abs/2104.00298.

[20] PyTorch Lightning Team. F1 Score — TorchMetrics Documentation. Accessed: 2025-05-13. 2024. https://lightning.ai/docs/torchmetrics/stable/classification/f1_score.html.

[21] Huang Y., Gokaslan A., Kuleshov V., and Tompkin J. The GAN is dead; long live the GAN! A Modern GAN Baseline. 2025. arXiv: 2501.05441 [cs.LG]. https://arxiv.org/abs/2501.05441.

[22] Jolicoeur-Martineau A. The relativistic discriminator: a key element missing from standard GAN. 2018. arXiv: 1807.00734 [cs.LG]. https://arxiv.org/abs/1807.00734.

[23] Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N., Antiga L., Desmaison A., Kopf A., Yang E., DeVito Z., Raison M., Tejani A., Chilamkurthy S., Steiner B., Fang L., Bai J., and Chintala S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. NeurIPS 2019. 2019. arXiv: 1912.01703 [cs.LG]. https://arxiv.org/abs/1912.01703.

[24] Shvetsov D. Counterfactual Search. GitHub repository. 2024. https://github.com/Dmytro-Shvetsov/counterfactual-search (05/11/2025).

Appendices

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.

I. Licence

I, Marko Lillemägi,

1. grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis

   Improving Counterfactual Image Generation for Weakly Supervised Tumour Segmentation through Theoretical and Architectural Alignment,

   supervised by Joonas Ariva and Dmytro Fishman

2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, 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. am aware of the fact that the author retains the rights specified in points 1 and 2;

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

Marko Lillemägi
15/05/2025

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