Latent diffusion for generative visual attribution in medical image diagnostics

Siddiqui, A.

LATENT DIFFUSION FOR GENERATIVE VISUAL ATTRIBUTION IN MEDICAL IMAGE DIAGNOSTICS

AMMAR ADEEL SIDDIQUI

M00869323

A thesis submitted to Middlesex University in
partial fulfilment of the requirements for the degree of *MSc By Research*

Department of Computer Science

19th September 2023

Abstract

Visual attribution in medical imaging seeks to make evident the *diagnostically-relevant* components of a medical image, in contrast to the more common detection of diseased tissue deployed in conventional machine vision pipelines (due to the inherent learning nature of these latter models, they are typically not easily interpretable/explainable to clinicians). State-of-the-art techniques in visual attribution generally consist of different variants of deep neural networks, implemented as classifiers, or segmenters. However, they have not thus far included an explicit linguistic component.

We here present a novel generative visual attribution technique, one that leverages latent diffusion models in combination with domain-specific large language models, in order to generate *normal counterparts* of abnormal images. The discrepancy between the two hence gives rise to a mapping indicating the diagnostically-relevant image components. To achieve this, we deploy image priors in conjunction with appropriate conditioning mechanisms in order to control the image generative process, including natural language text prompts acquired from medical science and applied radiology. We perform experiments and quantitatively evaluate our results on the COVID-19 Radiography Database containing labelled chest X-rays with differing pathologies via the Frechet Inception Distance (FID), Structural Similarity (SSIM) and Multi Scale Structural Similarity Metric (MS-SSIM) metrics obtained between real and generated images.

The resulting system also exhibits a range of latent capabilities including *super-resolution* and *zero-shot localized disease induction*, which are evaluated with real examples from the cheXpert dataset.

Visual Attribution, Explainable AI, Diffusion models, Medical imaging

Acknowledgements

I would like to express my heartfelt gratitude to Prof. David Windridge and Prof. Raja Nagarajan for motivating me to start my journey as a researcher and kindly guiding and mentoring me throughout the course of the research.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Medical imaging has become increasingly important in modern medical settings for patient stratification, assessing disease progression, evaluating treatment response, and grading disease severity Holzinger et al. (2019). However, medical image diagnosis tends to involve far more than simple disease detection. Visual Attribution (VA) is the detection, identification and visualization of *evidence* of a particular class or category of images Baumgartner et al. (2018). It is a specific part of explainability of learned models i.e using visualization techniques to investigate the decisions made by a model, and attribute the decisions to distinct parts of an image. This opens the model to interpretation, a key aspect of XAI (Explainable AI) machine learning research, especially in relation to deep learning models Vellido et al. (2012).

As it manifests, in medical imaging, VA is hence the process of educing evidence for medical conditions in relation to different parts of an image, such as pathological, psychological or disease related effects Zhu et al. (2017) Ge et al. (2017) Feng et al. (2017) Zhang, Bhalerao & Hutchinson (2017). As such, VA differs from the straightforward detection or segmentation of pathological regions in standard medical machine vision. These detected or segmented parts of the image are thus crucial biomarkers, and may serve as additional diagnostic and prognostic evidence Meena & Hasija (2022). Such models base their decisions on locally or globally perceived evidence components, and it is thus in these terms that the VA aspects of the models must be visually and semantically interpretable Zhang, Xie, Xing, McGough & Yang (2017). In clinical practice, these findings may then be used to diagnose and select treatment options, which may be surgical intervention, prescription of drugs etc. Interpretability is also key for scientific understanding of the system as a whole, and VA knowledge may thus sit on top of the explicit output of the model (for example, VA-based delineation of those regions *affected* by a tumor, typically extending significantly beyond the segmented tumor region itself). VA knowledge factors may also relate to the safety of the application, or to the ethics and a priori biases of the data, highlighting incomplete or mismatched objectives being optimized by the model Doshi-Velez & Kim (2017).

A lack of interpretability of one or more of these examples may lead to complete or partial system failure, the model failing to achieve some aspect of the complex targets provided by the user/clinincian, or to optimization of an objective different to that intended. Model explainability is hence of critical interest in the medical imaging domain, having been identified as crucial to increasing the trust of medical professionals in the automated diagnostic domain Holzinger et al. (2019). Visual attribution consequently provides a way to increase the confidence between the system, patient and clinician, leading to fewer misinformed results Gulum et al. (2021*b*). It may also serve to decrease cognitive load on the clinicians and medical practitioners via automated localization and segmentation of areas of interest Lee et al. (2018) Gulum et al. (2021*a*). However, it is important to consider the specific requirements and safety-criticalities of the application when developing a VA model (methods that directly manipulate images in the pixel space typically have to gain the acceptance of diagnosticians as part of their work process Singla et al. (2023)), and use-case flexible human-in-the-loop models are therefore to be preferred in the general case.

1.1 Generative Visual Attribution

The most recent techniques in visual attribution involve variants of deep neural networks (DNNs), and which tackle the problem in different ways, though typically centred on classification or segmentation Liu et al. (2022) Tropea & Fedele (2019). The need for VA is especially acute for DNNs in a clinical setting due to their intrinsic high complexity and low interpretability (they are are often termed 'black boxes') Petch et al. (2022) Li et al. (2021). However, DNNs, uniquely amongst machine learning VA approaches have the capacity to act in a *generative* manner. They hence have the capacity to mimic the actual clinical practice of a radiologist or practitioner is typically trained via the *difference* between healthy and non-healthy disease manifestations. As a result, the diagnosis of a condition or disease may be implicitly explained in terms of abnormalities of non-healthy tissue in relation to a hypothetical healthy version of the same tissue Sun et al. (2020).

Generative DNN-based machine learning therefore leads to the state-of-the-art strategy of *generative visual attribution* (developed in part by the authors) that leverages generative methods for counterfactual normal generation, in which abnormal images are translated into their *normal counterparts* for observation by a clinician. These methods hence perform visual attribution map generation via heatmaps taking the difference between the observed image of a patient and its healthy counterfactual Zia et al. (2022), Sun et al. (2020), Sanchez et al. (2022).

Previously, such techniques have used a specific DNN generative mechanism, *Generative Adversarial Networks* or *GANs* to carry out this mapping (cf the techniques ANT-GAN Sun et al. (2020) and VANT-GAN Zia et al. (2022)). This attribution process exploits the underlying properties of GANs to directly model the differences present between the normal and abnormal clinical images, as well as capture the complete structure of the individual classes in a learned latent representation. GANs in general have the advantage of requiring relatively fewer abnormal examples Xia et al. (2022) than standard supervised learning while still capturing underlying features of the surrounding areas of the higher density information regions. (Examples of these overlooked regions might be e.g. micro tumors in other parts of an organ that may not, in themselves, have a highly significant effect on the supervised decision boundary Baumgartner et al. (2018); it has been shown, especially for medical imaging DNNs, that such models typically disregard a significant fraction of these regions, which are essentially background evidence in relation to the underlying pathological condition Nguyen-Duc et al. (2020)).

However, GANs, while powerful, have faults that have led to the very recent development of a new state-of-the-art generative mechanism: *visual diffusion*. Diffusion models are typically able to operate at higher resolutions and image qualities than GANs. They are also superior to GANs in not suffering from ‘mode collapse’ arising from the adversarial process of distinguishing real from generated images reaching a convergence (Nash equilibrium) in which critical image classes are omitted Dhariwal & Nichol (2021). Diffusion models have been used for counterfactual generation as Diff-SCM Sanchez et al. (2022), and similar methods Sun et al. (2020) Wolleb, Bieder, Sandkühler & Cattin (2022) Özbey et al. (2023).

In this work, we shall use visual diffusion for counterpart normal generation. Our approach hence uses counterfactual generation with diffusion models directed at visual attribution in the medical imaging domain in a manner that builds on the conceptual foundations of generative visual attribution laid out in VANT-GAN Zia et al. (2022). In doing so, we will aim at to increase the interpretability of the model by using multi modal (text and image) inputs. We hence leverage prior control and conditioning techniques to reliably steer the mapping process in an interpretable manner utilising text prompts and control images. We achieve this by training domain-specific language and vision models on relevant medical imaging data allowing the generation of visual attribution maps for specific medical conditions, which can be quantitatively measured using relevant metrics in the domain.

As well as improving reliability, trustworthiness and utility in respect to the previous techniques of generative visual attribution, the approach of utilizing diffusion models in combination with domain adapted large language models with enhanced controllability and conditioning potentially also opens horizons to applications such as *post-surgery simulation of ageing, disease* etc by leveraging natural language instructions, as well as a host of additional ‘zero-shot’ latent use-case capabilities.

1.2 Diffusion Generative Models

Diffusion models consist of an autoencoder, which encodes the image into a latent space, and a diffusion process in which stochastic perturbations are performed incrementally in the latent space, such that a DNN can learn the reverse denoising process capable of transforming random noise images into images from the trained domain (a process which may may be guided by a suitable language model to introduce linguistic priors in the image generation). Depending on the autoencoder, the images generated by diffusion models are typically of relatively high resolution (compared with GANs) and the textual conditioning may include a wide range of textual encoders trained on specific domains, e.g. in the medical domain BioBERT Lee et al. (2020), RadBERT Yan et al. (2022) and PubmedCLIP Eslami et al. (2021). Such langauge encoders can hence be used to condition the generation in a much more flexible way than other generative models, and in particular GANs.

Other approaches use the metadata in the datasets to help learn models that take into account age, gender, intracranial and ventricular volume etc in parallel with image conditioning such as RoentGen Chambon et al. (2022) and LDM+DDIM Pinaya et al. (2022) for synthetic image generation. This meta-information can then be used to measure correlation among real images.

This ability to guide diffusion models via external semantic model make them potentially very powerful and relevant to visual attribution, especially in the medical imaging domain.

Chapter 2

Proposed Methodological Approach

The current research builds upon a particular conception of generative visual attribution set out in Zia et al. (2022) in the context of GAN generative models. In particular, it seeks to build on the notion of *counterpart normal generation*, but enriched via the use of visual diffusion and large language models.

We thus leverage domain-adapted language components combined with conditional generation to modify the latent diffusion in a manner suited to medical VA. The approach hence combines domain-adapted large language and vision models to enable broad medical understanding to be brought to bear on the problem of counterpart normal generation enabling generative visual attribution useful to understanding and pinpointing of visual evidence in the form of generated counterfactuals and visual maps. Additionally, the representative power of the domain adapted large language model alongside the image-domain representation of the vision model ensures that medical image concepts are grounded in medical language, such that counterfactual generation may be prompted via complex (natural language) text prompts including, potentially, location and intensity of disease or condition, or else constrained to the specific organs within a medical scan. Note that the vision model is not directly trained on such morphological concepts beforehand (e.g. the concept of an organ or the boundaries of an organ), yet is able to extrapolate from the combined multimodal knowledge using the data from the language and visual domain to discover these concepts latently.

Lastly, the model proposed shows zero-shot generation capabilities on disease concepts that are out of the training data distribution, but which also appear qualitatively valid in the generated counterfactuals. This is presumably the result of exploiting the different extrapolate capabilities of the respective vision and language models in a synergistic manner. The model thus latently encompasses the 'rules of biology' in generating counterfactuals, e.g not generating extra lung scar tissue where it could not exist, outside of the the chest cavity, irrespective of the language prompt.

This strengthens our argument for using latent diffusion models for visual attribution, since no direct perturbations are made in pixel space and neither is the model trained on synthetic data. We also need only use a dataset with a modest amount of images and basic one-word labels, relying on the text encoder (pretrained on domain-specific data, e.g. radiology reports) to supply additional linguistic concept relations.

The contributions of the study are as follows:

1. We illustrate the use of the visual diffusion pipeline for jointly fine-tuning the combination of a domain-adapted text encoder and a vision encoder with a modest amount of real medical scans and text prompts for conditional scan

generation (we thus eliminate the need for synthetic data).

2. We generate visually valid counterfactuals (non-healthy to healthy and vice versa) with minimal perturbations to the original real image guided by text prompts that employ complex natural language medical imaging concepts.

3. We explore the interpolation of knowledge in the text and vision domains using the composite text/vision models, evaluating the validity of the interpolations in the respective language and vision domains via their reflection into the other.

4. Using the generated counterfactuals, we generate visual maps by subtracting the generated counterfactual from the original image for visual attribution in the medical imaging domain, thereby enhancing diagnostic explainability in the manner of VANT-GAN (motivating the use of these models in safety-critical diagnostic applications in which visual explanations is critical for highlighting different areas of interest).

5. We show zero-shot generation capabilities in the visual domain for inducing diseases in healthy or non-healthy scans prompted by complex text prompts including medical imaging concepts using the text encoder.

6. Finally, we indicate the potential for future studies using such a combination of vision and language concepts for visual attribution using conditional generation.

Chapter 3

Related Work in Generative Visual Attribution

3.1 Generation of activation maps

Generative visual attribution includes a variety of classes of approach, each of which tackle the explainability problem in different ways. The particular class emphasised here, exemplified in a Sun et al. (2020) Baumgartner et al. (2018) and Zia et al. (2022), seek to generate complete or partial counterfactuals of the abnormal (i.e. diseased) image, and generate implicitly or explicitly a discrepancy map between the two. These maps are then visualize to highlight the attributing parts of the normal or abnormal image.

The ANT-GAN Sun et al. (2020) approach hence leverages GANs to generate normal or healthy-looking images from abnormal or unhealthy images and finds the difference between the two. These are then used to highlight local and global features from the image which otherwise might have been overlooked. The work in Baumgartner et al. (2018) learns a map generating function from the training data. This function then generates an instance specific visual attribution map highlighting the features unique for a class. The VANT-GAN Zia et al. (2022) approach generates VA maps directly from unhealthy images, which can then be used to generate healthy-looking images from unhealthy images. (This latter anticipates that the direct maps modelling learns *why* the image is unhealthy and captures the appropriate local and global visual attributes of the disease).

Charachon Charachon et al. (2021) generates a range of adversarial examples and tracks the gradient across the stable generation of the original image and the adversarial example. By mapping these gradients to image space, visual attribution maps are generated to find differences between the counterfactuals and the original image.

3.2 Generation of complete counterfactuals

The second (more common) class of generative visual attribution works generate complete subject/image counterfactuals, which are used for diagnostic findings and may or may not be used for explicit subtraction of images for highlighting the differences between the normal and generated counterfactual. STEEX Jacob et al. (2022) uses region-based selection of images and counterfactuals are generated only using semantic guidance. The regions are thus hoped to be meaningful (such as selecting a traffic signal with a green light and generating a counterfactual for a stop stop light within a complex image of a traffic junction). The counterfactuals are generated using semantic synthesis GAN, and the generation is constrained to keep the other regions unchanged. The Singla Singla et al. (2023) approach is a similar approach which uses perturbations in the original image controlled by a parameter. A counterfactual is generated for the perturbation such that the posterior probability of the image changes to the desired value of the parameter in the interval [0, 1].

Cutting edge methods of image generation, such as diffusion models, have significantly improved the resolution and quality of generated images. These models have been utilized in counterfactual generation techniques for the latter class of techniques such as Diff-SCM Sanchez & Tsaftaris (2022), “What is healthy” Sanchez et al. (2022) and other similar techniques Wolleb, Bieder, Sandkühler & Cattin (2022) Orgad et al. (2023). Diffusion models based generative VA techniques include Wolleb, Sandkühler, Bieder & Cattin (2022), which use noise encoding with reversed sampling and perform guidance using a class label and task-specific network. This combination is then denoised with a sampling scheme to generate a class conditional counterfactual. Unsupervised Medical Image Translation with Adversarial Diffusion Models Özbey et al. (2023) use a combination of diffusive and non diffusive models in an adversarial setup, to perform noising and transformation operations with the noised latents of the image to translate between two modalities of MRI scans, using class conditioning, such as transforming a T1 contrast image to T2. Diffusion Models for Medical Anomaly Detection Wolleb, Bieder, Sandkühler & Cattin (2022) use a weakly supervised setup for generating healthy counterfactuals of brain tumor images. The approach uses the noised latents from the diffusion model of the image and perform classifier guided denoising of the latent to produce a healthy image without a tumor. The What is Healthy? Sanchez et al. (2022) work similarly encodes the image into noised latents, using an unconditional model. The decoding of the latent can be done via class label or unconditionally, to generate a counterfactual of the starting input image. A heatmap of the region containing the lesion is then produced by taking the difference between the reconstructed healthy and starting image. The guidance is performed without a downstream classifier using conditional attention mechanism techniques.

In both of these broad classes of generative VA approach there is noticeable absence of a linguistic, natural language explanation or conditioning mechanism easily with which a domain expert could engage 'in the loop' (e.g. communicating with the system in domain specific terminologies via precise relational instructions for counterfactual generation). Such techniques require the use of classifier guidance for conditional descent of gradients mapping between the latent parameter space and the image space (for example, using weakly supervised decoding strategies or hyperparametric perturbation the image towards a healthy looking counterfactual). Furthermore, such techniques focus on regions of high information density, in most cases leaving the broad structure of the image remain changed. (An example would be a tumor causing exogenous pressure in the brain such that the surrounding tissue is displaced; this structural deformity would not be visually reversed by the above techniques, but rather just the tumor mass removed, and the unhealthy tissue converted into healthy tissue via transformations of pixel level features characteristic of the affected region).

Chapter 4

Diffusion Models

Diffusion models are probabilistic models which learn a data distribution by reversing a gradual noising process through sampling. Denoising thus proceeds from an assumed starting point of $x(t)$, where $x(t)$ is considered the final noisy version of the input x (which, being assumed to be equivalent to pure noise, can be treated as an easily sampled latent space). The model thus learns to denoise $x(t)$ into progressively less noisy versions $x(t-1), x(t-2)$.. until reaching a final version x(0) Dhariwal & Nichol (2021), representing a sample from the domain distribution. In transforming a (typically uniformly or Gaussian sampled) latent space into an observational domain, the process is thus one of generative machine learning, with the denoiser typical a deep neural network of learned parameter weights. The latest approaches, however, use the reweighted variant of the evidence lower bound, which estimates the gaussian noise added in the sample $x(t)$, using a parametrized function $\theta(x(t), t)$ rather than a denoised version of input x Rombach et al. (2022):

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(x_t, t\right)\|_2^2\right] \tag{4.1}$$

with $\epsilon_\theta\left(x_t, t\right)$ estimated via the diffusion model, such that the objective function is the difference between the predicted (latent paramter instantiation) noise and the actual noise instantiation (t is an arbitrary time step uniformly sampled from 1, . . . , T and E_x denotes the expected value over all examples x in the dataset).

4.1 Latent Diffusion models

To lower computational demands, latent diffusion models first seek to learn an appropriate latent space, one which, when decoded, is perceptually equivalent to the image space (a key assumption of latent diffusion is thus that noise perturbation of image and latent spaces are not intrinsically incompatible with regard to the generative process). Denoting the encoder by E, E hence learns to map images $x \in Dx$ into a spatial latent code $z = E(x)$. The essential mechanism of latent diffusion is then as indicated previously going forward - i.e. seeking to learn a model to correctly remove noise from an image, though this time in the latent space. The decoder D (which is usually a DNN) learns to map the latent codes back to images, such that $D(E(x))p \approx qx$. The objective function for the latent diffusion model now becomes

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1), t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t\right)\right\|_2^2\right] \tag{4.2}$$

where $z(t)$ is the latent noised to time step t Rombach et al. (2022) Gal et al. (2022).

4.2 Conditioning using a domain-specific encoder

In the following, the noise prediction function $\epsilon_\theta\left(x_t, t\right)$ is implemented using a time-conditioned Unet model Ronneberger et al. (2015), which can also be conditioned on class labels, segmentation masks, or outputs of a jointly trained domain specific encoder. Let y be the condition input and $T_{(\theta)}$ be a model which maps the condition y to an intermediate representation $T_{(\theta)}(y)$ which is then mapped to the intermediate layers of the UNet via a cross-attention layer Vaswani et al. (2017). The objective function for the class-conditional variant of latent diffusion Rombach et al. (2022) thus becomes

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\right\|_2^2\right] \tag{4.3}$$

4.2.1 Image Priors

In the above, any arbitrary image can be considered an instantiation of the generative latent parameters. Thus, instead of commencing from pure noise (i.e. purely stochastic latent parametric instantiantion), the latent diffusion process can instead be initiated from a given image, via application of the appropriate Stochastic Differential Equations (SDEs), as a form of prior conditioning in the image space. The given image (which may or may not be in the training data distribution, but which is presumed to lie within the manifold of natural images), is firstly perturbed with Gaussian noise ('lifting out the image manifold'. This noise is then removed progressively via the learned denoiser, which effectively acts to reproject the guide image back into the manifold of natural images; This may be thought of as a short random walk within *within the manifold* of a given metric distance.

More formally, if $x(0) \sim p_0$ is a sample from the data distribution, the forward SDE produces $x(t)$ for $t \in (0, 1]$ via Gaussian diffusion. Given $x(0)$, $x(t)$ is distributed as:

$$x(t) = \alpha(t)x(0) + \sigma(t)z, \quad z \sim N(0, I) \tag{4.4}$$

where the magnitude of the noise z is defined by the scalar function $\sigma(t) : [0, 1] \to [0, \infty)$. The magnitude of the data $x(0)$ is defined by the scalar function $\alpha(t) : [0, 1] \to [0, 1]$. The probability density function of $x(t)$ as a whole is denoted p_t.

The usually considered SDE are of two types. One is Variance Exploding SDE, where $\alpha(t) = 1$ for all t and $\sigma(1)$ is a large constant, which makes p_1 close to $N(0, \sigma^2(1)I)$. The second type is the Variance Preserving SDE, satisfying $\alpha^2(t) + \sigma^2(t) = 1$ for all t with $\alpha(t) \to 0$ as $t \to 1$, so that p_1 equals to $N(0, 1)$ Meng et al. (2021).

Image synthesis is then performed via a reverse SDE Anderson (1982) Song et al. (2020) from the noisy observation of $x(t)$ in order to recover $x(0)$, given knowledge of the noise-perturbed score function $\nabla x \log p_t(x)$. The learned score model as $s_\theta(x(t), t)$, the learning objective for time t is:

$$L_t = \mathbb{E}_{\mathbf{x}(0) \sim p_{\text{data}}, \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\| \sigma_t \boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{x}(t), t) - \mathbf{z} \|_2^2 \right] \tag{4.5}$$

with $s_\theta(x(t), t)$ a parametrized score model to approximate $\nabla x \log p_t(x)$; the SDE solution can be approximated with the Euler-Maruyama method Meng et al. (2021). The update rule from $(t + \Delta t)$ to t is:

$$\mathbf{x}(t) = \mathbf{x}(t+\Delta t) + \left(\sigma^2(t) - \sigma^2(t+\Delta t)\right) \boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{x}(t), t) + \sqrt{\sigma^2(t) - \sigma^2(t+\Delta t)}\mathbf{z} \tag{4.6}$$

A selection can be made on a discretization of the time interval from 1 to 0 and after the initialization $x(0) \sim \mathcal{N}(0, \sigma^2(1)I)$, Equation 4.4 can be iterated to produce an image $x(0)$ Meng et al. (2021).

4.2.2 Additional Control Priors

Additional conditioning mechanisms can be introduced to add further control to the generation e.g. ControlNet Zhang & Agrawala (2023) adds intermediate layers to the feature maps at each step of the downscaling operation while transitioning from image to latent space. Thus it becomes possible to add a task-specific image-conditioning mechanism to the model:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, t, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}, \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta\left(z_t, t, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}\right)) \|_2^2 \right] \tag{4.7}$$

Where given an image z_0, noised latents z_t are produced by progressively adding gaussian noise to the initial image after time steps t. Given the time step t, text prompts c_t, and task specific conditions c_f, the model learns a network to predict the added noise ϵ_θ. Some examples of the task specific conditions include Canny edge maps, Semantic Segmentaion, User sketching, and human pose Zhang & Agrawala (2023) etc.

The conditioning mechanisms of input text, image priors, depth and segmentation maps can thus be used in combination with each other, complementing or adding to the image generation for further generative control as required on a task-by-task basis.

Chapter 5

Methodology

In the following, we indicate normal medical images by I^n and abnormal images by I^a. We make the assumption that I^n and I^a are sampled from distributions $p_n(I)$ and $p_a(I)$ respectively. Additionally, we assume that the differences between an abnormal image and its corresponding normal image (from the same patient) are only the characteristic disease markers or indicators of diagnostically relevant abnormality, and no other structural differences are present. In this setup, given an input abnormal image I^a, we wish to produce a visual attribution map $M(I_i^a)$ that contains all the features that differentiate an abnormal image I_i^a from its normal counterfactual I_i^n.

Mathematically,

$$M(I_i^a) = I_i^a - I_i^n \tag{5.1}$$

To generate the normal counter part I_i^n we use a conditioned stable diffusion model which combines a text and an image condition or input of the forms set out in sections 4.2 and 4.2.1 . Using an image to image synthesis setting similar to SDEdit Meng et al. (2021), we start with the abnormal image as the guide $x^{(g)} = I_i^a$ and add Gaussian noise to form the noised latents $z_t = x^{(g)}(t_0) \sim \mathcal{N}(x^{(g)}; \sigma^2(t_0)I)$ which are then used to produce $x(0)$ using **Equation 4.6**, conditioned on $T_\theta(y)$, where T_θ is a domain adapted text encoder which maps the conditional prompt y to an intermediate representation $T_\theta(y)$. Hence the normal corresponding image $I_i^n = x(0)$ is synthesized as the denoised version of $\epsilon_\theta(z_t, t, T_\theta(y))$. The mask $M(I_i^a)$ is then explicitly produced by subtracting the generated normal counterpart from the abnormal image.

5.1 Implementation

5.1.1 Model Architecture

The conditioned latent diffusion model pipeline that we utilise consists in an encoder/decoder network of the form of a variational autoencoder (VAE), a time conditioned Unet model Ronneberger et al. (2015) conditioned on a domain-specific encoder in the textual domain (specifically a Bert based model trained on radiology reports called RadBERT Yan et al. (2022)) and, finally, the additional system fine tuning detailed below. Furthermore, we use an image-to-image conditioning mechanism similar to SDEdit Meng et al. (2021), such that the model takes two inputs, a text prompt and an image, and generates the counterfactual image from which a VA map is derived. The network architecture is shown in Figure 5.1.

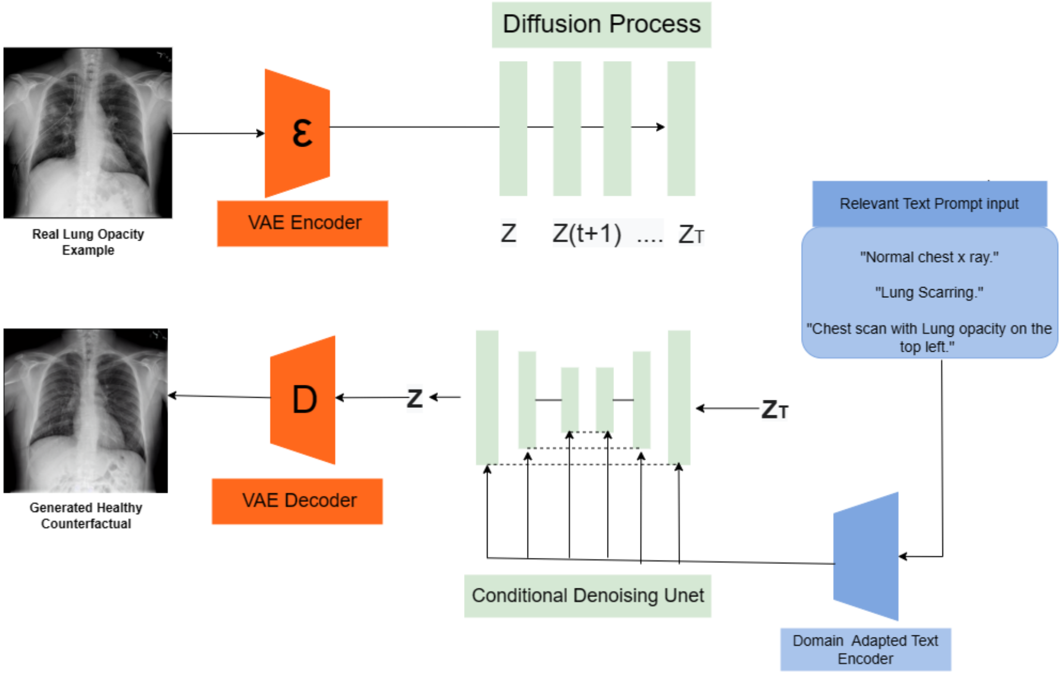

Diffusion Process
ε
VAE Encoder
Real Lung Opacity Example
Z
Z(t+1)
....
ZT
Relevant Text Prompt input
"Normal chest x ray."
"Lung Scarring."
"Chest scan with Lung opacity on the top left."
D
VAE Decoder
Z
ZT
Generated Healthy Counterfactual
Conditional Denoising Unet
Domain Adapted Text Encoder

Figure 5.1: The counterfactual generation pipeline takes as input the starting abnormal image x^a, which is encoded by the VAE encoder (ϵ) to form the encoded image latents Z and passed through the diffusion process to form noised latents of the image Z_T after incremental t steps. The fine-tuned conditional U-net denoises the latents into the conditioned latent Z, decoded by the VAE decoder D into the final generated counterfactual x^n. The loss function for the Unet conditioned on the domain specific encoder is used for joint fine tuning of the Unet and domain adapted text encoder i.e $L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t,t,\tau_\theta(y)\right)\|_2^2\right]$ (Equation 4.3)

5.1.2 Training Details

The pretrained latent diffusion model *CompVis/stable-diffusionv1-4* and the Bert based model *RadBERT* are obtained from Huggingface https://huggingface.co/StanfordAIMI/RadBERT. These were jointly fine-tuned using a single Quadro RTX 8000 at bf16 precision, with batch size = 2, at a resolution of 512x512px. The models were fine-tuned on the diffusers library using an approach for binding a unique identifier to a specific subject via a class-specific prior preservation loss, Dreambooth Ruiz et al. (2022), with 1200 training steps used for the Normal class, after which 500 training steps are applied for each of the non-healthy classes, namely Lung Opacity, COVID-19, and Viral Pneumonia, making a total number of training steps of 2700. The greater preponderance of the normal class ameliorates the intrinsic imbalance in dataset, with model convergence inherently slower for the X-ray image domain, being out of the initial distribution. The learning rate was 5e-05 and, for sampling, the PNDM scheduler strength is set at 0.55 with Guidance Scale=4 found to be most effective across all classes for counterfactual generation.

The COVID-19 Radiography Database Chowdhury et al. (2020) contains 10192 normal, 3616 COVID-19, 4945 Lung Opacity and 1345 Viral pneumonia chest x-ray images. The dataset is obtained from:
https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database.
The model is fine-tuned on the images using their respective labels as text prompts i.e *Normal chest scan*, *Lung Opacity*, *Viral Pneumonia*, and *COVID 19*.

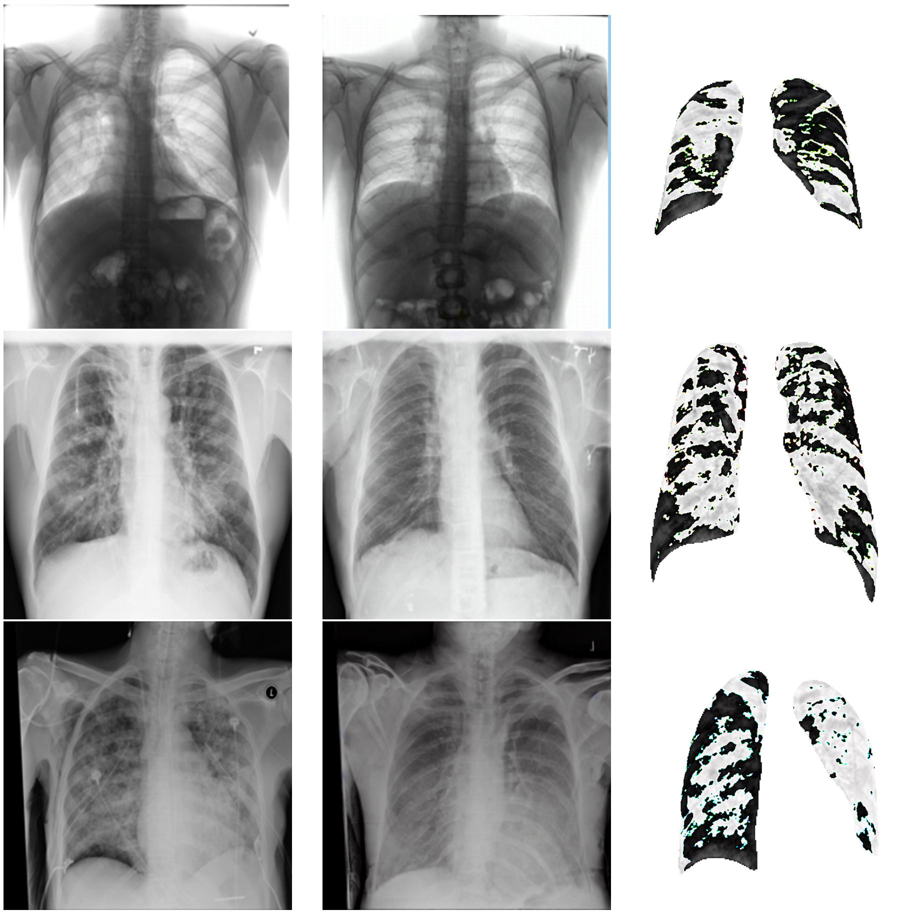
Chapter 6

Experiments

We firstly evaluate counterfactual generation –the generation of healthy counterparts to unhealthy scans– via an investigation of its qualitative impact i.e. the overall *visual plausibility* of the generated counterpart. Following this, we seek to quantitatively analyze the generative perturbation of the tested unhealthy scans in order to determine the utility of the method in its primary mode of VA application. Finally, we explore the latent capacity of the trained system to carry out a series of zero-shot counterfactual generation exercises, in particular: *localized disease induction* and the *induction of diseases from outside the training data* in relation to input healthy scans.

6.1 Qualitative evaluation

Example images from the disease COVID-19 Radiography Database and their generative healthy counterparts are given in figure 6.1. The images on the far left are instances of the lung opacity class from the real images in the dataset. The images on the right are examples of the generated healthy counterfactuals obtained via latent



((a)) Lung Opacity ((b)) Generated Normal ((c)) Generated Healthy Tissue via difference

Figure 6.1: Healthy Counterfactual Generation for a case of lung opacity (White parts in the difference indicate generated tissue by the model)

space diffusion, with RadBERT-guided textual-conditioning via a “normal chest x ray” conditional prompt. A total of 75 diffusion inference steps are used with image conditioning strength=0.85 and guidance scale=7.5. (The former indicates the level of constraint on changes to the original input image and the latter is the weight given to the textual encoder conditioning in the generation of the image, ranging over [0,1] and [0,9], respectively).

Side-by-side inspection suggests that, as required, only minimal perturbations are made to the original image with respect to healthy pixels, and to localised image sites without structural medical defects. In the top row, the medical structural defect in the original image is due to a lung opacity, and characterized via a relatively complex interaction between the scanner & subject manifesting as ‘gaps’ in the corresponding portions of the lung scan.

The healthy/non-healthy discrepancy maps in the above cases are obtained via masked subtraction of the original image from the generated image. (Ground truth segmentation masks are present for each image in the dataset corresponding to the broad area of interest –i.e. the complete lung). The generated healthy tissue is thus a subset of the mask and is shown in fig. 6.1 for the respective cases.

In the context of a VANTGAN Zia et al. (2022)-based approach, this highlighted material constitutes the diagnostic counterfactual visual attribution, i.e. the selection of material relevant to the diagnosis of the unhealthy condition. Corresponding healthy counterfactual generation was performed for the complete datasets in the three unhealthy classes, i.e *Lung opacity*, *Viral Pneumonia* and *COVID*, examples of which are given in fig. 6.2 for the three classes (all of the generated healthy counterfactuals from this experiment can be found on `https://huggingface.co/ammaradeel/diffusionVA`). Visual inspection indicates that the generated counterfactuals are, in general, visually plausible with minimal perturbations made to the unhealthy image overall. At the detail-level, two further key aspects of visual plausibility are apparent: firstly, the model does not change those aspects of the images unrelated to the medical condition and, secondly, the model makes changes to the unhealthy regions selectively, and in a structurally sound manner, e.g. generating missing portions of the lung while refraining from generating extraneous lung material where it would not normally exist (e.g. in the abdominal cavity).

6.2 Quantitative evaluation

6.2.1 Fréchet Inception Distance (FID)

For quantitative evaluation on the COVID19 dataset, the Fréchet Inception Distance (FID) Heusel et al. (2017) was calculated for the generated healthy counterfactuals for each class in order to measure the level of realism of the images, and also how distant the generated counterpart normal distribution is from that of the healthy and diseased image sets. The results are as shown in Table 6.1.

The Fréchet Inception Distance (FID) proposed by Heusel et al. Heusel et al. (2017) is a quantitative measure to measure the quality of generated samples by generative models. The measure requires embedding the samples to a feature space by a specific layer of the Inception Net. This layer is then viewed as a continuous multivariate Gaussian, and the mean and covariance are estimated for the complete generated and real samples of the data, resulting in two Gaussians. The FID between these two Gaussians is then defined as

$$FID(x, g) = \|\mu_x - \mu_g\|_2^2 + \mathrm{Tr}(\Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{1/2}) \tag{6.1}$$

Where (μ_x, Σ_x) is the mean and covariance of the sample embeddings from the data distribution (real data), and (μ_g, Σ_g) are the mean and covariance of the sample embeddings from the model (generated samples) Lucic et al. (2018). The study Heusel et al. (2017) shows that the score is consistent with human judgement, and is relatively more robust to noise than other similar measures.

The FID scores are calculated with default characterisations i.e activations of the pool3 layer of the InceptionV3 model with 2048 dimensions. The particular implementation deployed is sourced from the Pytorch FID package Seitzer (2020). A lower FID would indicate that distribution of the two image sets are similar. Obtained results indicate that the real healthy and the generated healthy counterfactuals have relatively similar distributions, with the exception of the Viral Pneumonia class, which has a significantly larger absolute relative difference in FID scores.

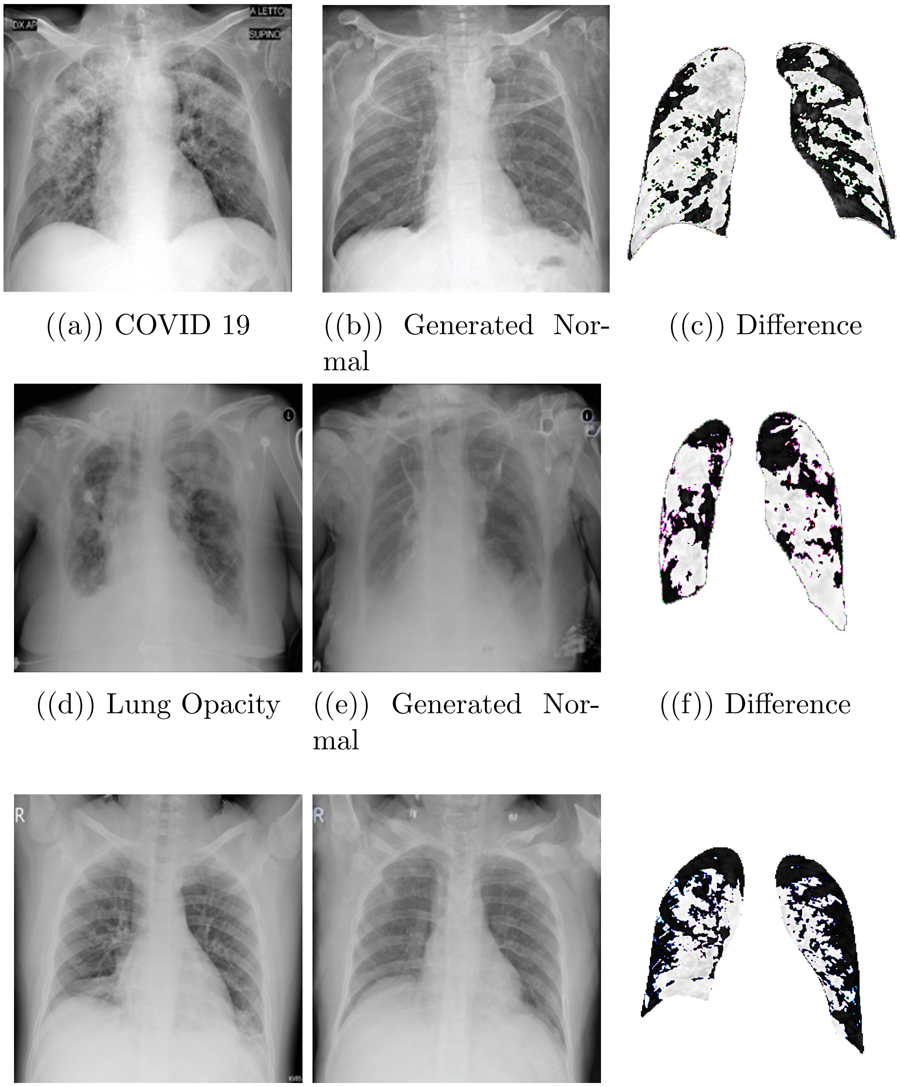
An “ImageSet“ indicates all images of a real class or a generated class. E.g. In the first row of Table 1, ImageSet 1 is Lung Opacity, referring to all images of the Lung Opacity class from the original dataset, while ImageSet 2 contains all ***generated*** healthy images corresponding to ImageSet1. ImageSet 1 and ImageSet2 in the second row correspond to the images of the Lung Opacity and Healthy classes of the ***original*** dataset respectively. In the FID paradigm, ImageSet1 would be referring to (μ_x, Σ_x), and ImageSet2 to (μ_g, Σ_g).

A relative comparison of the scores is needed for a counterfactual generation with semantic meanings attached such as “COVID” as a direct comparison between the **generated healthy set** and the **real healthy set** would assume the anatomical-structural differences between the images is due to the effects of a medical condition **all else being equal**, which is not the case. The underlying biases of the data collection of the image sets may have a relatively large unwanted effect on the measures, such as age (Structural changes due to size of chest frame, bone density etc.), sex (Breast tissue), ethnicity etc. An instance drawback of such a comparison is pointed out in a found bias in the Viral Pneumonia set discussed in this section.

Given the insight on the comparison, the direct differences between generated healthy and real healthy images are presented in Table 6.2 for respective classes. This difference may be used as a measure of fidelity or quality to chest X-rays in general. These differences are relatively good in terms of fidelity as the differences using the original stable diffusion without any training or fine-tuning may go up to 275.0 as pointed out in the Roentgen Chambon et al. (2022) study.

The overall visual soundness of the generated images, as validated via the absolute and relative FID scores obtained for each of the classes, is broadly consistent with the previous qualitative interpretation that the tested image distributions are minimally perturbed in order to transform them into healthy counterfactuals, while refraining from making changes to the healthy local regions of the image (the scores of the COVID19 class are the closest in this respect among the tested disease conditions, with a relative absolute difference of **6.0** in FID scores between real and generated images.

The scores for the viral pneumonia class appear to be in a large part attributable



((a)) COVID 19 ((b)) Generated Normal ((c)) Difference

((d)) Lung Opacity ((e)) Generated Normal ((f)) Difference

((g)) Viral Pneumonia ((h)) Generated Normal ((i)) Difference

Figure 6.2: Healthy Counterfactual Generation (White parts in the difference indicate generated tissue by the model)

Table 6.1: FID as a measure of minimum valid perturbations across classes to generate healthy counterfactuals

Image Set 1	ImageSet 2	FID
Lung Opacity (4945)	Generated Healthy (4945)	27.8
Lung Opacity (4945)	Real Healthy (10192)	46.9
	Relative Absolute Difference	**19.1**
Viral Pneumonia (1345)	Generated Healthy (1345)	37.63
Viral Pneumonia (1345)	Real Healthy (10192)	97.6
	Relative Absolute Difference	**59.97**
COVID 19 (3616)	Generated Healthy (3616)	32.2
COVID 19 (3616)	Real Healthy (10192)	38.2
	Relative Absolute Difference	**6.0**

Table 6.2: FID as a measure of image quality

Image Set 1	ImageSet 2	FID
Real Healthy (10192)	Generated Healthy from Lung Opacity (4945)	60.60
Real Healthy (10192)	Generated Healthy from Viral Pneumonia (1345)	110.72
Real Healthy (10192)	Generated Healthy from COVID19 (3616)	45.11

to the relatively larger magnitude of fundamental structural differences between healthy and viral pneumonia images in the training set: in particular, the viral pneumonia image set mostly had scans from children and infants, while the healthy class was of adult majority. (This data bias would break the basic assumption that differences between class image sets is due only to structural defects of disease).

SSIM and MS-SSIM

For further quantitative evaluation of the generated counterfactuals, the Structural Similarity (SSIM) and the Multi Scale Structural Similarity Metric (MS-SSIM) Rouse & Hemami (2008) were calculated between the unhealthy images and their respective generated counterparts, and averaged across the classes.

SSIM or the Structural Similarity index Wang et al. (2004) is a measure to quantify the differences between a processed/distorted image and a reference image, using studied properties of the human visual system. It is based on the assumption that human visual perception is highly adapted for extracting structural information from a vision and hence is designed to capture changes in structural information, luminance and contrast.

The SSIM is a combination of three key comparisons between x and y, namely **luminance comparison** $l(x, y)$, **contrast comparison** $c(x, y)$ and **structure comparison** $s(x, y)$.

For computing these functions, the mean intensity and the standard deviation of the images or signals x and y are required.

The luminance of each signal, estimated as the mean intensity is given as μ_x and the standard deviation as an estimate of the signal contrast as σ_x.

$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2\right)^{\frac{1}{2}}$

The luminance $l(x, y)$, contrast $c(x, y)$ and structure comparisons $s(x, y)$ are then defined as

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{6.2}$$

where the constant is included to avoid instability when $\mu_x^2 + \mu_y^2$ is very close to zero. C_1 is defined as

$$C_1 = (K_1 L)^2 \tag{6.3}$$

where L is the dynamic range of the pixel values, and $K_1 << 1$ is a small constant.

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{6.4}$$

where $C_2 = (K_2 L)^2$, and $K_2 << 1$.

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \tag{6.5}$$

where C_3 is a small constant. In discrete form, σ_{xy} is defined as

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \,. \tag{6.6}$$

The SSIM(x,y) between two signals or images x and y is then defined as

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^\alpha \cdot [c(\mathbf{x}, \mathbf{y})]^\beta \cdot [s(\mathbf{x}, \mathbf{y})]^\gamma \tag{6.7}$$

where α, β and γ are weighting variables, used to control the relative importance of luminance, contrast and structure in the measure. We use the general form of the measure where $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$, resulting in

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{6.8}$$

The Multi-Scale Structural Similarity Wang et al. (2003) (MS-SSIM) is a method developed to incorporate image details at different resolutions, as it has been studied that the distance to the image plane from the observer, and the perceptual capabilities of the observer have an effect on the percievability of the image details. The original scale is denoted as Scale 1, and the highest scale is Scale M, obtained after $M - 1$ iterations. The MS-SSIM method downsamples the x and y signals using a low-pass filter for every iteration by a factor of 2. The j-th contrast comparison Eq. 14 and structure comparison Eq. 15 are denoted as $c_j(x, y)$ and $s_j(x, y)$ respectively. The luminance comparison Eq.12 is made at only the largest scale (original size) at scale M. The Multiscale SSIM is then defined as

$$\mathrm{MultiscaleSSIM}(\mathbf{x}, \mathbf{y}) = [l_M(\mathbf{x}, \mathbf{y})]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(\mathbf{x}, \mathbf{y})]^{\beta_j} [s_j(\mathbf{x}, \mathbf{y})]^{\gamma_j} \tag{6.9}$$

We interpret the metrics as displaying the appropriate amount of structural similarity between the generated counterfactuals and the unhealthy real images, as the overall structure of the unhealthy images should not completely change, but should also not remain unchanged; and hence only the required perturbations should be made. A low structural similarity indicates larger perturbations to the unhealthy image, and a higher structural similarity indicates smaller perturbations. In the extreme cases, 0 would indicate no structural similarity at all, and 1 would indicate identical images.

The SSIM and the MS-SSIM measures for the classes are displayed in Table 6.3.

Table 6.3: MS-SSIM and SSIM as a measure of minimum valid perturbations across classes to generate healthy counterfactuals

Image Set 1	Image Set 2	MS-SSIM	SSIM
COVID (3616)	Generated Healthy (3616)	0.830	0.798
Lung Opacity (4945)	Generated Healthy (4945)	0.813	0.780
Viral Pneumonia (1345)	Generated Healthy (1345)	0.802	0.768

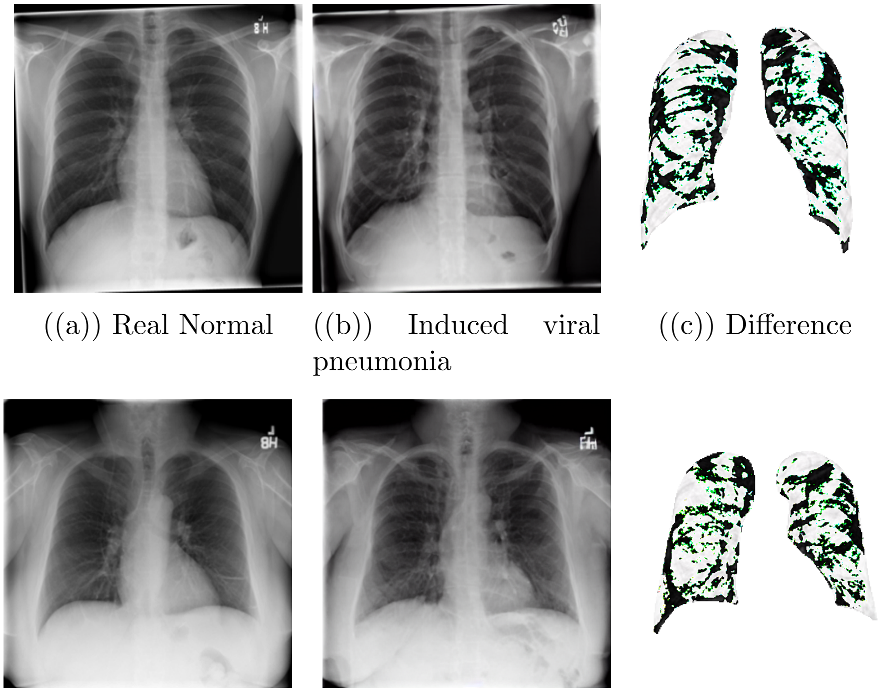
Chapter 7

Localized disease induction

The model was also tested in regard to its latent capability to induce disease in specific locations via LLM-guided text conditioning, for example, conditioning on “lung opacity on the top left of the chest” and general conditions such as “COVID 19“ and “Viral Pneumonia” on which the model was trained on. The model performs well visually and is sensitive to the strength and guidance scale parameters. The induction of the condition and its severity which may manifest itself in a specific amount of structural damage in the generation, and the fidelity to the original input image are factors which can be controlled by the domain expert. The effect of hyperparameters is illustrated and explained in future sections.

The model shows the effects of sample bias in lung opacity examples with respect to the left and right side of the lungs, and also due to there being more examples of Viral pneumonia scans for children than adults.

These capabilities also highlight the model’s understanding of the structural attribution of the disease, e.g it generated lung scarring, lung opacity and structural defects for the respective disease in the accurate regions, i.e not outside the lung or



((a)) Real Normal ((b)) Induced viral pneumonia ((c)) Difference

((d)) Real Normal ((e)) Induced COVID19 ((f)) Difference

Figure 7.1: Induction of diseases in real healthy scans (Black in the difference indicates induced scarring or damage)

in an orientation in which they could not anatomically exist. The disease induction mechanism is illustrated in figure 7.1.

Chapter 8

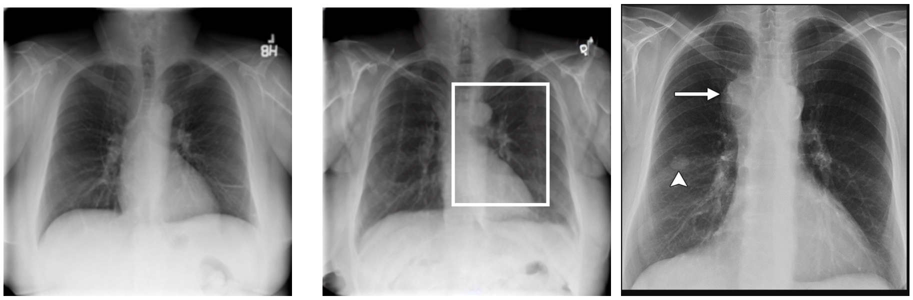
Zero Shot generation - unseen disease induction

The trained model was prompted for localized generative counterparts, particularly lung scarring, small cell carcinoma and cardiomegaly. The results show distinct and overlapping differences in conditions, i.e. only lung scarring, conditions which include lung scarring and other structural defects in combination, such as lung cancer.

An example of carcinoma is shown in Figure 8.1 for reference with the real healthy and generated carcinoma. We suspect this capability arises as a result of knowledge adaptation from the domain-adapted text encoder to the visual domain via the visual model, given that the domain-adapted text encoder is trained on the full panoply of Radiology reports.

8.1 Zero shot evaluation and the use of hyperparameters

The disease cardiomegaly was not present in the training data, and to evaluate zero shot induction, the real images from the small version of the Chexpert Irvin et al. (2019) dataset were used from https://www.kaggle.com/datasets/ashery/



((a)) Real Normal

((b)) Model Induced Carcinoma

((c)) Real example with expert markings Carter et al. (2014)

Figure 8.1: Zero shot carcinoma induction with a real example marked by experts

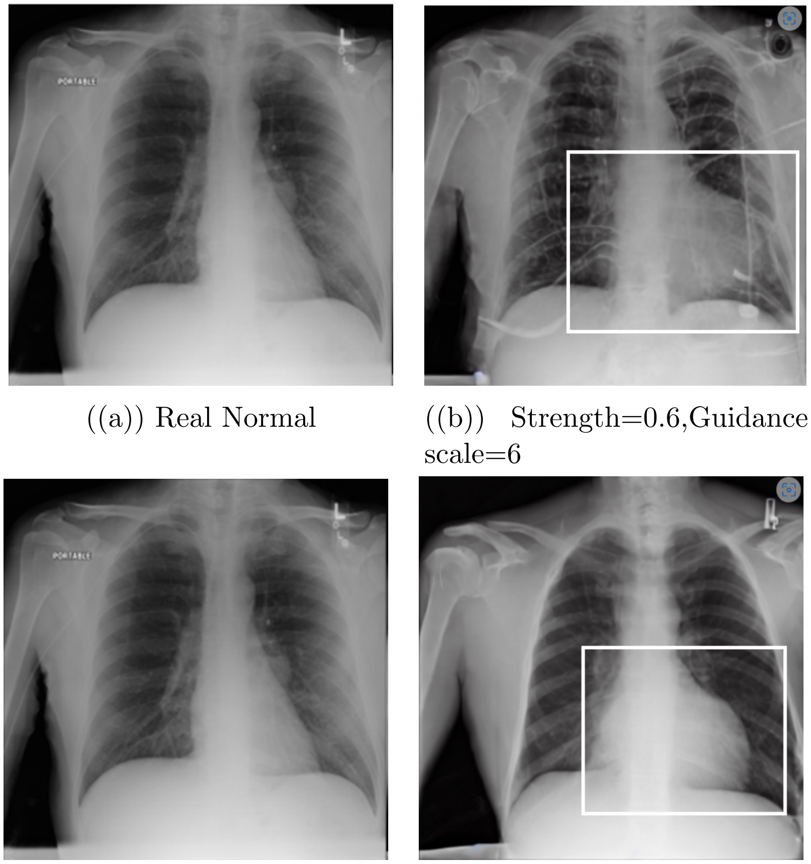
chexpert. 8060 images of positively identified cases of cardiomegaly were used as the image set of real cardiomegaly, and for each of the healthy images from the COVID 19 database, an induced version was generated by the model with the prompt "Cardiomegaly". Some of the generated images are displayed in figure 8.2.

The visual plausibility is shown in the images for the disease, and to evaluate the zero-shot generations quantitively, the FID score was used between the real cases of cardiomegaly from the Chexpert dataset and the generated images.

Table 8.1: FID as a measure of minimum valid perturbations for zero-shot cardiomegaly induction

Image Set 1	Image Set 2	FID
Real Cardiomegaly (8060)	Generated Cardiomegaly (10192)	52.08
Real Healthy (10192)	Generated Cardiomegaly (10192)	17.71

The scores in Table 8.1 show that the generated cardiomegaly images do not have a large distance from the real images using which they were generated, suggesting appropriate perturbations were made and the generations were reasonably close to the



((a)) Real Normal

((b)) Strength=0.6,Guidance scale=6

((c)) Real Normal

((d)) Strength=0.9, Guidance scale=7

Figure 8.2: Induction of Cardiomegaly in real healthy scans

real cardiomegaly set from the Chexpert dataset. The generations were performed for different settings of hyperparameters, which did not have a large difference in FID scores when evaluated on complete image sets, yet the visual differences for the same image are significant as highlighted in Figure 8.2 for different sets of hyperparameters. This is due to the different aspects specific to the patient image, such as prior health of the patient, structural variances due to age, recording equipment, size etc. which which may have varying differences in generations for different patient scans, but not for the complete image set. The depiction conveys the use of the Strength and Guidance-scale hyperparameters and the trade-off between them.

A high value of Strength provides flexibility for larger perturbations to the original image, but the generated image is as a result far from the original image. Using this flexibility paired with the text prompt conditioning (guidance scale) conditions which require different intensity of structural changes can be generated. An example of healthy to non-healthy may be extreme scarring and lesions throughout the chest generated from a scan of a very healthy lung. Similarly, larger flexibility would be required to restore tissue from an extremely damaged lung to generate a normal image. For smaller values of strength, the generated counterpart would be much closer to the real image, and subtle prompts may require this setting such as inducing a lesion or a micro-tumour for the unhealthy generation, and restoring a small amount of damage to lung tissue for a healthy counterfactual generation.

Chapter 9

Conclusion

In this work, we present a novel generative visual attribution technique for improving explainability in the medical imaging domain, leveraging a fusion of vision and large language models via the stable diffusion pipeline, built on foundational generative VA concepts from the VANT-GAN approach. The model developed on the technique can be used to generate normal counterparts of scans affected with different medical conditions to provide contrast between the real affected and generated normal scan, providing insight into the inference made by the model in a style synonymous with human radiologists.

The pre-trained domain-adapted text and vision encoder are jointly fine-tuned using a modest number of image and one-word text training examples from the medical imaging domain for image-to-image generations. The generation capabilities include the induction of different medical conditions in healthy examples induced with varying severity, controlled by hyperparameters.

The inputs to the text encoder support advanced medical domain language and terminology, with specific geographical locations in organs. By harnessing the model's learned multimodal knowledge from the domain-adapted text encoder and the vision model, out-of-training data distribution or zero-shot generations can be made for unseen medical conditions.

In the medical diagnostics domain, future work in the study opens horizons to complex disease-interaction induction, providing simulations on the combined effects of age, lifestyle choices and medical tests from different areas of the body, spread of disease and metastasis which may prove revolutionary in differential diagnosis. Coupled with the discussed advanced control methods, it can also be used for real-time surgery simulations, such as incisions, haemorrhages etc. The modest need for data may also prove helpful for few-shot learning of applied concepts, such as rare diseases with limited examples and infant scans etc.

The study applies the joint potential of image and natural language to medical knowledge, but the fundamental multimodal knowledge learning concepts can be used in any domain, such as criminology (e.g. facial composite), psychology (e.g. behaviour analysis), satellite imagery (e.g climate and disaster simulation) the most powerful of which may prove to be using the zero-shot inference capabilities of the model.

Bibliography

Anderson, B. D. (1982), 'Reverse-time diffusion equation models', *Stochastic Processes and their Applications* **12**(3), 313–326.

Baumgartner, C. F., Koch, L. M., Tezcan, K. C., Ang, J. X. & Konukoglu, E. (2018), Visual feature attribution using wasserstein gans, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 8309–8319.

Carter, B. W., Glisson, B. S., Truong, M. T. & Erasmus, J. J. (2014), 'Small cell lung carcinoma: staging, imaging, and treatment considerations', *Radiographics* **34**(6), 1707–1721.

Chambon, P., Bluethgen, C., Delbrouck, J.-B., Van der Sluijs, R., Połacin, M., Chaves, J. M. Z., Abraham, T. M., Purohit, S., Langlotz, C. P. & Chaudhari, A. (2022), 'Roentgen: Vision-language foundation model for chest x-ray generation', *arXiv preprint arXiv:2211.12737* .

Charachon, M., Cournède, P.-H., Hudelot, C. & Ardon, R. (2021), Visual explanation by unifying adversarial generation and feature importance attributions, *in* 'Interpretability of Machine Intelligence in Medical Image Computing, and Topological

Data Analysis and Its Applications for Medical Data: 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 4', Springer, pp. 44–55.

Chowdhury, M. E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., Islam, K. R., Khan, M. S., Iqbal, A., Al Emadi, N. et al. (2020), 'Can ai help in screening viral and covid-19 pneumonia?', *Ieee Access* **8**, 132665–132676.

Dhariwal, P. & Nichol, A. (2021), 'Diffusion models beat gans on image synthesis', *Advances in Neural Information Processing Systems* **34**, 8780–8794.

Doshi-Velez, F. & Kim, B. (2017), 'Towards a rigorous science of interpretable machine learning', *arXiv preprint arXiv:1702.08608* .

Eslami, S., de Melo, G. & Meinel, C. (2021), 'Does clip benefit visual question answering in the medical domain as much as it does in the general domain?', *arXiv preprint arXiv:2112.13906* .

Feng, X., Yang, J., Laine, A. F. & Angelini, E. D. (2017), Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules, *in* 'Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20', Springer, pp. 568–576.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G. &

Cohen-Or, D. (2022), ‘An image is worth one word: Personalizing text-to-image generation using textual inversion’, *arXiv preprint arXiv:2208.01618* .

Ge, Z., Demyanov, S., Chakravorty, R., Bowling, A. & Garnavi, R. (2017), Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images, *in* ‘Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20’, Springer, pp. 250–258.

Gulum, M. A., Trombley, C. M. & Kantardzic, M. (2021*a*), Multiple interpretations improve deep learning transparency for prostate lesion detection, *in* ‘Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB Workshops, Poly 2020 and DMAH 2020, Virtual Event, August 31 and September 4, 2020, Revised Selected Papers 6’, Springer, pp. 120–137.

Gulum, M. A., Trombley, C. M. & Kantardzic, M. (2021*b*), ‘A review of explainable deep learning cancer detection models in medical imaging’, *Applied Sciences* **11**(10), 4573.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. (2017), ‘Gans trained by a two time-scale update rule converge to a local nash equilibrium’, *Advances in neural information processing systems* **30**.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. (2019), ‘Causability and explainability of artificial intelligence in medicine’, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **9**(4), e1312.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K. et al. (2019), Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, *in* 'Proceedings of the AAAI conference on artificial intelligence', Vol. 33, pp. 590–597.

Jacob, P., Zablocki, É., Ben-Younes, H., Chen, M., Pérez, P. & Cord, M. (2022), Steex: steering counterfactual explanations with semantics, *in* 'Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII', Springer, pp. 387–403.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. & Kang, J. (2020), 'Biobert: a pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics* **36**(4), 1234–1240.

Lee, S., Lee, J., Lee, J., Park, C.-K. & Yoon, S. (2018), 'Robust tumor localization with pyramid grad-cam', *arXiv preprint arXiv:1805.11393* .

Li, R., Wang, Z. & Zhang, L. (2021), Image caption and medical report generation based on deep learning: a review and algorithm analysis, *in* '2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI)', IEEE, pp. 373–379.

Liu, X., Yang, L., Chen, J., Yu, S. & Li, K. (2022), 'Region-to-boundary deep learning model with multi-scale feature fusion for medical image segmentation', *Biomedical Signal Processing and Control* **71**, 103165.

Lucic, M., Kurach, K., Michalski, M., Gelly, S. & Bousquet, O. (2018), 'Are gans created equal? a large-scale study', *Advances in neural information processing systems* **31**.

Meena, J. & Hasija, Y. (2022), 'Application of explainable artificial intelligence in the identification of squamous cell carcinoma biomarkers', *Computers in Biology and Medicine* **146**, 105505.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y. & Ermon, S. (2021), Sdedit: Guided image synthesis and editing with stochastic differential equations, *in* 'International Conference on Learning Representations'.

Nguyen-Duc, T., Zhao, H., Cai, J. & Phung, D. (2020), 'Med-tex: Transferring and explaining knowledge with less data from pretrained medical imaging models', *arXiv preprint arXiv:2008.02593* .

Orgad, H., Kawar, B. & Belinkov, Y. (2023), 'Editing implicit assumptions in text-to-image diffusion models', *arXiv preprint arXiv:2303.08084* .

Özbey, M., Dalmaz, O., Dar, S. U., Bedel, H. A., Özturk, Ş., Güngör, A. & Çukur, T. (2023), 'Unsupervised medical image translation with adversarial diffusion models', *IEEE Transactions on Medical Imaging* .

Petch, J., Di, S. & Nelson, W. (2022), 'Opening the black box: the promise and limitations of explainable machine learning in cardiology', *Canadian Journal of Cardiology* **38**(2), 204–213.

Pinaya, W. H., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S. & Cardoso, M. J. (2022), Brain imaging generation with latent diffusion models, *in* ‘Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings’, Springer, pp. 117–126.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021), Learning transferable visual models from natural language supervision, *in* ‘International conference on machine learning’, PMLR, pp. 8748–8763.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. (2022), High-resolution image synthesis with latent diffusion models, *in* ‘Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition’, pp. 10684–10695.

Ronneberger, O., Fischer, P. & Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* ‘Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18’, Springer, pp. 234–241.

Rouse, D. M. & Hemami, S. S. (2008), Analyzing the role of visual structure in the recognition of natural image content with multi-scale ssim, *in* ‘Human Vision and Electronic Imaging XIII’, Vol. 6806, SPIE, pp. 410–423.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M. & Aberman, K. (2022), ‘Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation’.

Sanchez, P., Kascenas, A., Liu, X., O'Neil, A. Q. & Tsaftaris, S. A. (2022), What is healthy? generative counterfactual diffusion for lesion localization, *in* 'Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings', Springer, pp. 34–44.

Sanchez, P. & Tsaftaris, S. A. (2022), 'Diffusion causal models for counterfactual estimation', *arXiv preprint arXiv:2202.10166* .

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. et al. (2022), 'Laion-5b: An open large-scale dataset for training next generation image-text models', *Advances in Neural Information Processing Systems* **35**, 25278–25294.

Seitzer, M. (2020), 'pytorch-fid: FID Score for PyTorch', https://github.com/mseitzer/pytorch-fid. Version 0.3.0.

Singla, S., Eslami, M., Pollack, B., Wallace, S. & Batmanghelich, K. (2023), 'Explaining the black-box smoothly—a counterfactual approach', *Medical Image Analysis* **84**, 102721.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S. & Poole, B. (2020), 'Score-based generative modeling through stochastic differential equations', *arXiv preprint arXiv:2011.13456* .

Sun, L., Wang, J., Huang, Y., Ding, X., Greenspan, H. & Paisley, J. (2020), 'An adversarial learning approach to medical image synthesis for lesion detection', *IEEE journal of biomedical and health informatics* **24**(8), 2303–2314.

Tropea, M. & Fedele, G. (2019), Classifiers comparison for convolutional neural networks (cnns) in image classification, *in* '2019 IEEE/ACM 23rd International Symposium on Distributed Simulation and Real Time Applications (DS-RT)', IEEE, pp. 1–4.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017), 'Attention is all you need', *Advances in neural information processing systems* **30**.

Vellido, A., Martín-Guerrero, J. D. & Lisboa, P. J. (2012), Making machine learning models interpretable., *in* 'ESANN', Vol. 12, Citeseer, pp. 163–172.

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

Wang, Z., Simoncelli, E. P. & Bovik, A. C. (2003), Multiscale structural similarity for image quality assessment, *in* 'The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003', Vol. 2, Ieee, pp. 1398–1402.

Wolleb, J., Bieder, F., Sandkühler, R. & Cattin, P. C. (2022), Diffusion models for medical anomaly detection, *in* 'International Conference on Medical image computing and computer-assisted intervention', Springer, pp. 35–45.

Wolleb, J., Sandkühler, R., Bieder, F. & Cattin, P. C. (2022), 'The swiss army knife for image-to-image translation: Multi-task diffusion models', *arXiv preprint arXiv:2204.02641* .

Xia, X., Pan, X., Li, N., He, X., Ma, L., Zhang, X. & Ding, N. (2022), 'Gan-based anomaly detection: a review', *Neurocomputing* .

Yan, A., McAuley, J., Lu, X., Du, J., Chang, E. Y., Gentili, A. & Hsu, C.-N. (2022), 'Radbert: Adapting transformer-based language models to radiology', *Radiology: Artificial Intelligence* **4**(4), e210258.

Zhang, L. & Agrawala, M. (2023), 'Adding conditional control to text-to-image diffusion models', *arXiv preprint arXiv:2302.05543* .

Zhang, Q., Bhalerao, A. & Hutchinson, C. (2017), Weakly-supervised evidence pinpointing and description, *in* 'Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25', Springer, pp. 210–222.

Zhang, Z., Xie, Y., Xing, F., McGough, M. & Yang, L. (2017), Mdnet: A semantically and visually interpretable medical image diagnosis network, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 6428–6436.

Zhu, W., Lou, Q., Vang, Y. S. & Xie, X. (2017), Deep multi-instance networks with sparse label assignment for whole mammogram classification, *in* 'International conference on medical image computing and computer-assisted intervention', Springer, pp. 603–611.

Zia, T., Murtaza, S., Bashir, N., Windridge, D. & Nisar, Z. (2022), 'Vant-gan: adversarial learning for discrepancy-based visual attribution in medical imaging', *Pattern Recognition Letters* **156**, 112–118.

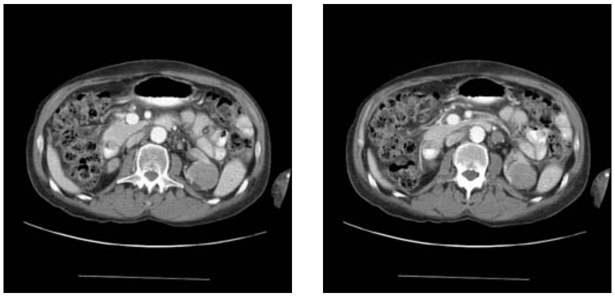
Appendix A

Experimental Design and Results

A.1 Experiments with the PEIR Dataset, fine-tuning the textual embedding

Experiments were performed on the PEIR dataset https://peir.path.uab.edu/library/ using the Textual inversionGal et al. (2022) strategy for fine-tuning the textual embeddings of the model. The Textual inversion strategy involves freezing the weights of the visual and textual encoder, and hence no gradient updates are performed on either of the models, rather only the word embeddings for the concept fed to the model are changed. This results in a search in the already learnt embedding space for the new concept, minimizing error between the generated images and the input images. The optimization function is the same as mentioned in Equation 3, keeping both τ_θ and ϵ_θ fixedGal et al. (2022).

The images from the PEIR dataset A.1 were chosen as they were labelled by experts with detailed diagnosis reports, containing multiple instances of the same disease effects. An experiment was performed using “LYMPHOMA/PTLD POST



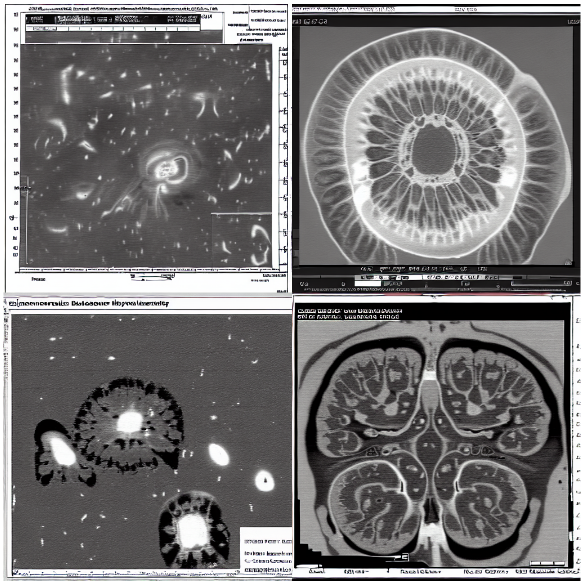
((a)) Example 1 ((b)) Example 2

Figure A.1: LYMPHOMA/PTLD POST RENAL TX ADRENAL AND RENAL MASSES

RENAL TX ADRENAL AND RENAL MASSES” instance images using the textual inversion strategy. The strategy was chosen as it needed a smaller amount of data, was swift to train (∼20 minutes), and did not require large computational resources. The experiments were performed on the NVIDIA Tesla K80 GPU using the free version of Google Colab. After results from different experiments A.2, it was concluded that fine-tuning other components of the model may be required, as the default model was trained on the Laion-5bSchuhmann et al. (2022) dataset, which was significantly different than scans used in medical imaging, and only finding embedding may not be ample to obtain good results.

A.2 Testing the VAE for domain adapted capabilities

Fine-tuning or training the model from scratch was a computationally expensive task, and hence the components were tested for domain-adapted capabilities to reduce the time and computational footprint, to be chosen for fine-tuning. The Variational Autoencoder was tested for its capabilities and the latent image encodings A.4, fed

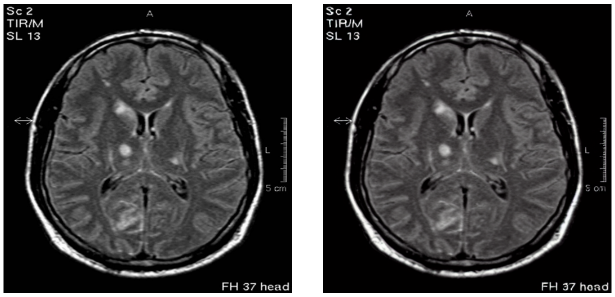


((a)) Multiple generations using the same prompt with textual inversion

Figure A.2: Renal masses in an abdomen ultrasound

to the model in the pipeline were visualized and the image was reconstructed using the decoder part of the VAE to analyze differences. Brain scans from the ADNI dataset were set as the working dataset at this point due to the limited amount of data present in the PEIR dataset for similar disease effects, and the brain scan being simpler to learn as compared to abdominal scans with a variety of different disease affects present across multiple organ systems.

The findings from the experiment showed that the VAE was not in need of fine-tuning for domain adaptation as the reconstruction was almost identical.



((a)) Original Brain Scan from the ADNI dataset

((b)) Reconstruction

Figure A.3: Reconstruction testing of the VAE for medical imaging domain

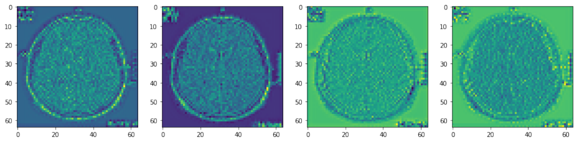


Figure A.4: Latent encodings from the VAE

A.3 Fine tuning the Unet with the ADNI Dataset

Fine-tuning experiments were performed using domain-adapted textual encoders such as Biobert Lee et al. (2020), and the results were not optimal, most of them being plain black or brown images. It was concluded that the Unet of the model needed to be experimented with fine-tuning, as replacing the CLIP Radford et al. (2021) text encoder with a domain-adapted text encoder did not yield results.

Experiments were performed by fine-tuning the Unet and Radbert, a domain-adapted textual encoder, with a smaller subset of the ADNI dataset. Radbert was preferred over Biobert as it was trained on a larger dataset. The fine-tuning was performed jointly on the textual encoder and Unet. The ADNI dataset required a significant amount of preprocessing, and hence experiments were performed on a

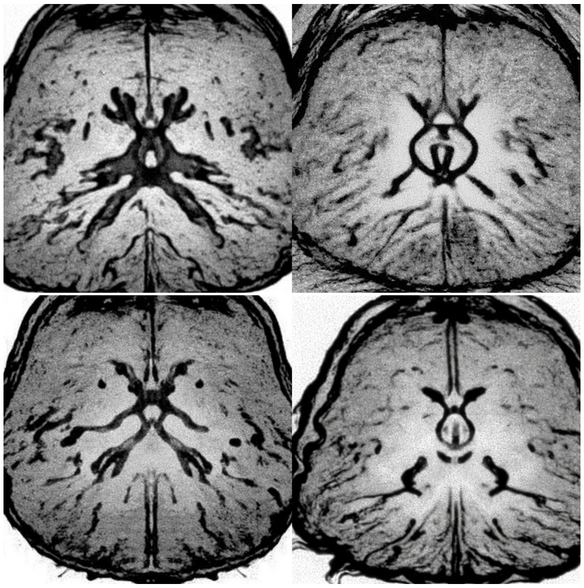


Figure A.5: Experimental results with the ADNI dataset

manually processed subset of the data, developing pipelines for automatic preprocessing in parallel. Some results are presented in Figure A.5.

A.4 Experiments with the COVID-19 Radiography Database

Experiments were performed with only the Normal class images of the COVID-19 Radiography database, containing 10192 images of chest scans, as a target to build a system capable of generating healthy counterfactuals given a non-healthy image from a different class. The dataset was chosen due to no requirement of extensive preprocessing and potentially a large enough size. Some of the results are presented in Figure A.6. The conclusion made from the results suggested some control over the generations may be required other than text encoder.

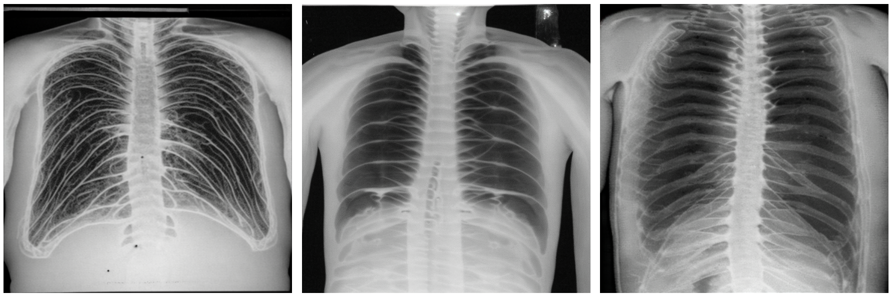


Figure A.6: Experimental results with the COVID-19 Radiography database

A.5 Experiments combining Stable diffusion and image priors using Control Net

To add control to the generation process, the Control Net Zhang & Agrawala (2023) mechanism was explored, with the fine-tuned model, combined with the domain-adapted textual encoder for generations. The control net mechanism required an input image, which could be used as an added condition to the textual encoder for the generation, as mathematically mentioned in Equation 7. The image can be added as a condition in different forms to achieve a variety of objectives. These forms may include edge maps via different image processing techniques or much more sophisticated forms such as pose, segmentation maps, and depth from other complex models, as per the application.

We explored the canny edge detector A.7 with different threshold settings to control the generations. The hope was to preserve the relevant structure of the image, such as the number of ribs, frame size etc. making changes in lung tissue - ideally, filling the gaps or "Lung Opacity" in unhealthy scans. The results were not optimal after experiments with different combinations of the hyperparameters.

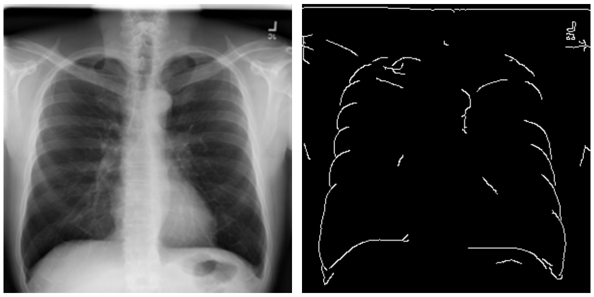


Figure A.7: Canny edge detection for control net conditional generations

Fine-tuning of the control net mechanism was considered to enhance the type of control needed for chest X-rays. Some generations are displayed in figure A.8.

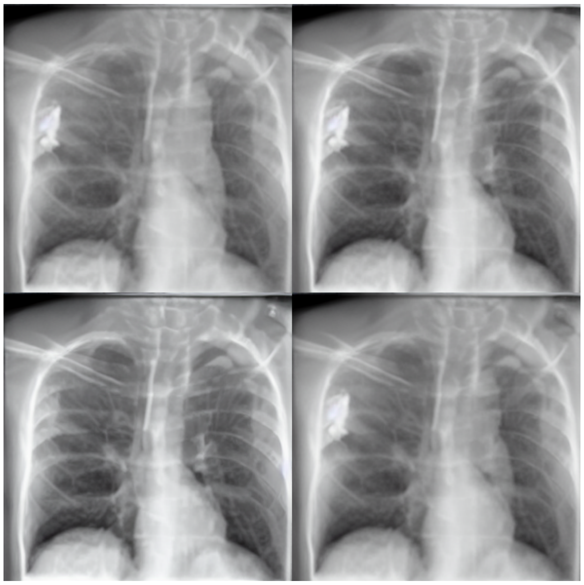


Figure A.8: Generations using the Canny Edge detections as a condition using Control Net