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

Ekaterina Sedykh

Imagining Infinity: Endless CT Datasets through Conditional Diffusion Models

Master's Thesis (30 ECTS)

Supervisor(s): Dmytro Fishman, PhD

Tartu 2024

Imagining Infinity: Endless CT Datasets through Conditional Diffusion Models

Abstract:
Medical imaging, the technique of creating visual representations of the interior of a body for clinical analysis and medical intervention, critically depends on the availability of extensive and high-quality datasets. However, the acquisition of such datasets is often limited by logistical, ethical, and privacy concerns. Diffusion models, known for their effectiveness in generating high-quality synthetic data, can mitigate these challenges by producing realistic medical images for model training and research purposes. This thesis addresses the challenge of data scarcity in medical imaging by exploring the efficacy of diffusion models in generating synthetic CT images that can be used to enhance dataset diversity and volume. Here, we demonstrate that diffusion models can effectively generate synthetic CT images that closely mimic real diagnostic images, thereby potentially expanding the breadth of available training data for medical AI applications. The results reveal that the synthetic images produced are not only almost indistinguishable from real images but also retain the necessary clinical details, which is an advancement over previous generative models that often sacrificed clinically relevant details. This work further exemplifies the utility of synthetic data generated by diffusion models in improving the training and performance of AI systems in diagnosing and analyzing medical images. The integration of diffusion models into medical imaging practices promises to significantly strengthen the AI-driven diagnostic tools. By providing a novel method for generating synthetic medical images, this research highlights the potential of advanced generative models in overcoming practical and ethical barriers in medical research.

Keywords: Diffusion Models, Synthetic Data, Medical Imaging, CT

CERCS: P176 Artificial intelligence; B110 Bioinformatics, medical informatics, biomathematics, biometrics

Kujutades lõpmatust: lõputud CT andmekogumid tingimuslike difusioonimudelite kaudu

Lühikokkuvõte: Meditsiiniline pildistamine, mis on tehnika keha sisemuse visuaalsete kujutiste loomiseks kliinilise analüüsi ja meditsiinilise sekkumise eesmärgil, sõltub olulisel määral ulatuslike ja kvaliteetsete andmekogumite kättesaadavusest. Selliste andmekogumite hankimist piiravad aga sageli logistilised, eetilised ja eraelu puutumatusega seotud probleemid. Diffusioonimudelid, mis on tuntud oma tõhususe poolest kvaliteetsete sünteetiliste andmete genereerimisel, võivad neid probleeme leevendada, tootes realistlikke meditsiinilisi pilte mudelite treenimiseks ja teadustööks. Käesolevas väitekirjas käsitletakse andmete vähesuse probleemi meditsiinilise pildistamise valdkonnas, uurides difusioonimudelite tõhusust sünteetiliste KT-kujutiste loomisel, mida saab kasutada andmekogumi mitmekesisuse ja mahu suurendamiseks. Siinkohal näitame, et difusioonimudelid suudavad tõhusalt luua sünteetilisi KT-pilte, mis jäljendavad täpselt tegelikke diagnostilisi pilte, laiendades seeläbi potentsiaalselt meditsiinilise tehisintellekti rakenduste jaoks kättesaadavate treeningandmete ulatust. Tulemused näitavad, et loodud sünteetilised pildid ei ole mitte ainult peaaegu eristamatud reaalsetest piltidest, vaid säilitavad ka vajalikud kliinilised üksikasjad, mis on edasiminek võrreldes varasemate geneerivate mudelitega, mis sageli ohverdasid kliiniliselt olulised üksikasjad. See uurimus näitab veel kord, kui kasulikud on difusioonimudelite abil genereeritud sünteetilised andmed, et parandada tehisintellekti süsteemide koolitust ja tulemuslikkust meditsiiniliste piltide diagnoosimisel ja analüüsimisel. Diffusioonimudelite integreerimine meditsiinilise kujutamise praktikasse tõotab märkimisväärselt tugevdada tehisintellektipõhiseid diagnostikavahendeid. Pakkudes uudset meetodit sünteetiliste meditsiiniliste piltide genereerimiseks, toob see uuring esile täiustatud genereerivate mudelite potentsiaali praktiliste ja eetiliste takistuste ületamisel meditsiinilistes uuringutes.

Võtmesõnad:
Difusioonimudelid, Sünteetilised Andmed, Meditsiiniline Pildistamine, KT

CERCS: P176 Tehisintellekt; B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Contents

1 Introduction

1.1 Background and Motivation

The realm of medical imaging is experiencing a significant evolution, primarily driven by advancements in image processing and machine learning technologies. A key aspect of this research is the generation of synthetic CT scans with a focus on diffusion models. The generation of these synthetic images holds the potential to significantly enlarge publicly available medical datasets.

The availability of diverse and extensive datasets is critical for training robust and accurate AI models in medical diagnostics. However, one of the primary challenges in medical image analysis has been the limited availability of high-quality, annotated medical images, due to privacy concerns and the high cost of data acquisition. By generating synthetic yet realistic CT scans, we can address this data scarcity, enabling the development and refinement of AI models without the constraints of data accessibility and patient privacy.

1.2 Research Objectives

The primary objective of this thesis is to investigate the capabilities of diffusion models in the conditional generation of synthetic CT scans. The specific goals include:

- Developing and evaluating diffusion models for their ability to generate synthetic CT images that can supplement existing medical datasets.
- Assessing the impact of synthetic CT datasets on the training, performance, and reliability of AI models used in medical imaging.

1.3 Thesis Contribution

This thesis contributes to the field of medical image processing and artificial intelligence by:

1. Investigating the use of diffusion models for the generation of synthetic CT scans, an area that has not been extensively explored.
2. Introducing a novel method for the generation of synthetic CT images, which has the potential to considerably broaden current datasets and assist in developing more robust AI models for medical diagnosis.
3. Providing evidence of the value of synthetic data in improving the performance of AI models in medical imaging.

1.4 Writing Assistance

This thesis was enhanced with the assistance of Grammarly and ChatGPT. Grammarly, a cloud-based writing aid, leverages artificial intelligence to review and improve various aspects of writing, including grammar, spelling, punctuation, clarity, engagement, delivery, and the detection of plagiarism. It was employed to eliminate grammatical errors, enhance sentence structure for improved clarity, and ensure appropriate punctuation. ChatGPT, powered by the Generative Pre-trained Transformer (GPT-4) language model, is a conversational agent capable of executing a range of linguistic tasks, including answering questions, generating text and summarizing content. In this thesis, ChatGPT was utilized to refine sentences to a more scholarly tone and to rectify grammatical inaccuracies.

1.5 Thesis Structure

The thesis is structured as follows:

1. **Literature Review:** This section reviews current methods in medical image generation, emphasizing the role of synthetic data and the utilization of diffusion models in artificial intelligence.

2. **Methods:** The section details all methodologies utilized for generating synthetic CT images, both in 2D and 3D. It covers the datasets used and details of the diffusion models and their implementation to ensure reproducibility.

3. **Results and Discussion:** This section presents an analysis of the outcomes obtained from various models and configurations. It also discusses the influence of synthetic CT data produced by diffusion models on the training of AI models and their performance in diagnostics.

4. **Conclusion:** A summary of the findings and their implications for AI in medical diagnostics, as well as a discussion on future research.

This thesis specifically focuses on leveraging diffusion models for the creation of high-quality synthetic CT scans, aiming to significantly improve the dataset diversity and volume available for AI training in healthcare. It has a potential to impact the development of more accurate and reliable AI-driven diagnostic tools, thereby facilitating enhanced patient outcomes and more efficient medical imaging practices.

2 Literature Review

The integration of artificial intelligence into medical imaging marks a significant advance in how we approach healthcare diagnostics and research. One of the main parts here is the generation of synthetic medical images, a field that offers solutions to some of the most pressing issues faced by researchers and clinicians today. This literature review examines the role of synthetic data generation in medical imaging, focusing on the methods used to create these images, their applications, and their potential to overcome obstacles related to data availability. These models, particularly Generative Adversarial Networks (GANs), Conditional GANs (cGANs), Variational Autoencoders (VAEs), and more recently, Denoising Diffusion Probabilistic Models (DDPMs) and Stable Diffusion models, have been instrumental in addressing some of the most significant challenges, such as data scarcity and privacy concerns. The generation of synthetic images, driven by the latest advancements in deep learning, exemplifies the profound impact that AI can have on enhancing and broadening the scope of medical imaging practices.

2.1 Synthetic Data in Medical Imaging

Synthetic data in medical imaging refers to artificially created images that closely mimic real medical images in terms of structure and appearance. This concept has gained substantial attention as a solution to the challenges of data scarcity and privacy in medical imaging research and AI model development. Synthetic medical images are typically generated using computational models that are trained on real medical image datasets. These models can include various machine learning techniques, with recent trends leaning heavily towards advanced deep learning methods. The primary purpose of synthetic medical images is to augment existing datasets, particularly in scenarios where collecting large volumes of real medical images is impractical due to ethical, legal, or logistical constraints [10, 19, 11].

The impact of synthetic data on AI model training and validation in medical imaging is substantial. Models trained on larger and more diverse datasets, including synthetic images, tend to be more robust and generalize better to new, unseen data. An example would be a a study by Thambawita et al. [32], where a novel synthetic data generation pipeline SinGAN-Seg was created, to produce synthetic medical images with corresponding masks for medical image segmentation tasks. The synthetic data generated by their pipeline significantly improved the performance of segmentation models when tested on real data, showcasing the effectiveness of synthetic data in enhancing downstream tasks in medical imaging research. Another case of synthetic data in medical imaging is the creation of artificial X-ray images GANs. Khosravi & Li et al. [17], have successfully utilized GAN models trained on real X-ray datasets to generate synthetic X-ray images that closely resemble authentic medical images in structure and appearance. It shows that that incorporating these synthetic X-ray images can improve the performance and

generalization of deep learning models for medical imaging tasks, such as disease detection and classification . The synthetic X-ray images were able to capture the nuanced characteristics and pathological patterns present in real X-ray images. This approach addresses data scarcity and class imbalances, showcasing the potential of synthetic data to enhance medical imaging research and model outcomes. Additionally, synthetic data can be used to validate AI models by providing a controlled environment to test their performance across a wide range of scenarios, including rare conditions or pathologies not commonly available in public datasets.

2.1.1 Data Scarcity and Privacy Concerns

One of the most significant advantages of synthetic data is its ability to circumvent issues related to data scarcity. In medical imaging, the availability of large, diverse datasets is crucial for developing robust AI models. Synthetic data can substantially expand the size of training datasets. Furthermore, since synthetic images do not correspond to real patients, they inherently address privacy concerns, making them a valuable asset in research areas where patient confidentiality is paramount [9].

While synthetic data offers many advantages, it also comes with challenges. Ensuring that synthetic images are sufficiently realistic and accurately represent the diversity of real-world medical scenarios is critical. There is also the risk of introducing biases into the synthetic data, which can affect the performance of AI models trained on these datasets.

2.1.2 Methods for Generating Synthetic Data

The generation of synthetic medical images has evolved from basic algorithmic approaches to more sophisticated methods. Early techniques involved geometric shapes and complex mathematical models to simulate anatomical structures. However, the advent of deep learning, especially Generative Adversarial Networks (GANs), has improved this field. GANs, for instance, have been particularly effective in producing high-quality, realistic medical images for various modalities, including CT, MRI, and X-ray images.

Deep learning approaches have revolutionized the field of medical imaging by enabling the generation of synthetic data that closely mimics real medical scans. This section explores the principles, applications, and innovations brought about by deep learning models that are able to generate high-quality synthetic images, highlighting their significance in enhancing medical imaging practices and AI model development. Each of these models brings a unique approach to synthesizing data that bears close resemblance to authentic medical images. Figure 2 provides an overview of these generative models, highlighting their core methodologies and their role in synthetic data creation.

Generative Adversarial Networks (GANs): GANs, introduced by Goodfellow et al., have shown considerable promise in generating synthetic medical images that closely

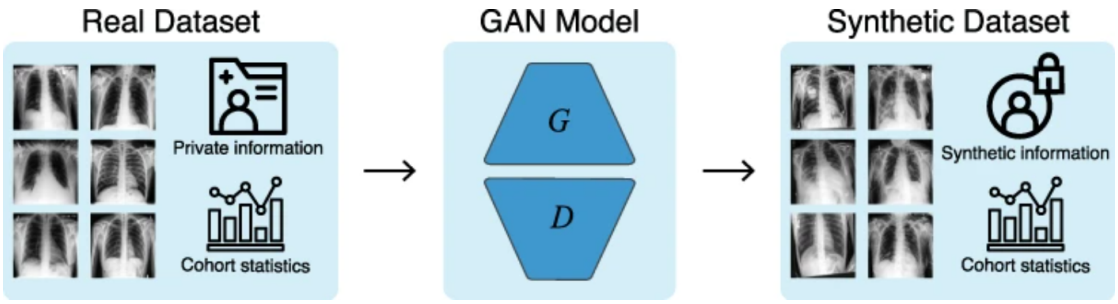

Figure 1. Synthetic Data Generation Using GANs. The process begins with a real dataset containing private information and cohort statistics. This dataset feeds into the GAN model, comprised of a generator (G) and a discriminator (D), which learns and simulates the distribution of the real data. The output is a synthetic dataset that retains the statistical properties of the original data while ensuring individual privacy. The synthetic images are designed to be indistinguishable from real images by the discriminator part of the GAN, thereby generating a new, artificial cohort for research and analysis purposes. Image taken from [9].

resemble authentic scans. A GAN is composed of two neural networks—the generator and the discriminator—that engage in a zero-sum game to produce data indistinguishable from real-life examples. In medical imaging, GANs have been instrumental in augmenting datasets, which is crucial for training robust AI models where data may be scarce or privacy concerns limit access to real patient scans [12].

Despite their capabilities, GANs face challenges like mode collapse and training instability, often attributed to the delicate balance required between the generator and discriminator. To combat these issues, innovative architectures and training methods have been proposed [30, 1, 2].

Conditional Generative Adversarial Networks (cGANs): The conditional variant of GANs introduces additional data into the training process, such as image labels or key features, to guide the generation of images. This approach enhances the specificity and relevance of the produced synthetic images, making cGANs exceptionally useful for generating medical imagery that meets particular clinical criteria [21, 16, 25].

Variational Autoencoders (VAEs): VAEs represent a different approach to synthetic image generation. By encoding input data into a latent space, VAEs maintain essential image features and utilize regularization to ensure consistency across generated images. These models are particularly adept at incorporating supplementary information, thereby tailoring the synthetic data more closely to actual patient data, which is vital for personalized medicine [18].

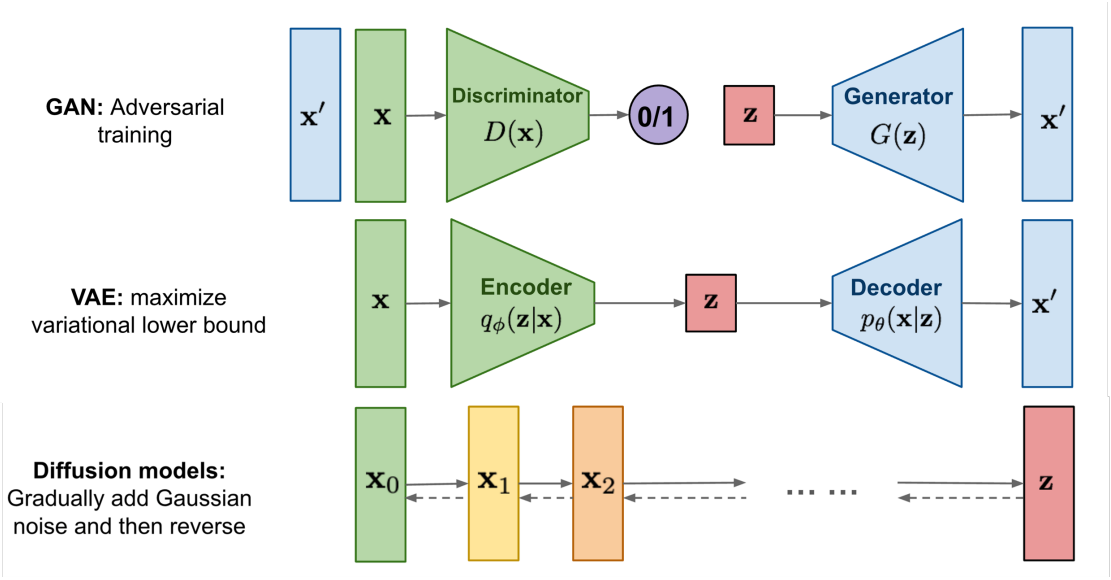

Figure 2. Comparative Overview of Generative Models. This figure showcases the fundamental operational strategies of different generative models. GANs utilize adversarial training, VAEs focus on maximizing the variational lower bound, and diffusion models employ gradual noise addition and reversal.

2.2 Diffusion Models

Diffusion is a natural phenomenon in which particles or substances travel from high concentration areas to low concentration areas due to their intrinsic random motion [5].

Diffusion is mathematically controlled by Fick's laws. These principles explain how the relationship between the concentration gradient and the rate of substance transfer (diffusion flux), which results in a change in concentration over time. To put it simply, particles gravitate toward locations where there are fewer of them in an effort to find equilibrium or balance. Specifically, Fick's first law states that the diffusion flux J is proportional to the concentration gradient ∇c, formalized as:

$$J = -D\nabla c$$

where D represents the diffusion coefficient. Fick's second law, or the diffusion equation, further describes the time-dependent change in concentration:

$$\frac{\partial c}{\partial t} = D\nabla^2 c$$

In these equations, c denotes the concentration of particles at a given location and t represents time. The negative sign in Fick's first law indicates that particles diffuse in the opposite direction of the concentration gradient, moving towards equilibrium [24].

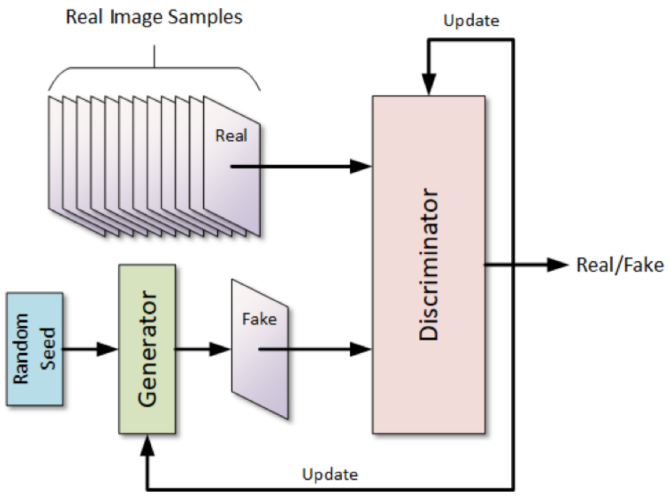

Figure 3. Generative Adversarial Network (GAN) structure, illustrating the interplay between the generator and discriminator networks.

In the context of image processing, diffusion models work by shifting the "concentration" or intensity of pixels within an image. The values of a pixel's intensity are handled similarly to concentrations [23, 6]. In image processing, diffusion models are typically applied iteratively. In each iteration, the model starts with the original image and uses a diffusion process to gradually improve it using the principles of diffusion. Depending on the intended result, the number of iterations and the particular diffusion parameters can be changed. The iterative diffusion process for an image can be expressed as:

$$I(x, y, t + \Delta t) = I(x, y, t) + \Delta t \cdot D\nabla^2 I(x, y, t)$$

where $I(x, y, t)$ is the image intensity at spatial location (x, y) and time t, and Δt is a discrete time step. This iterative process, applied over several timesteps, is capable of denoising or smoothing the image, akin to particle diffusion seeking equilibrium.

Diffusion models have the advantage of maintaining important image characteristics while carrying out operations like denoising or smoothing. Depending on the task at hand, they are flexible and can be adjusted to meet certain requirements. However, selecting the right settings for diffusion can sometimes be difficult. Under-diffusion might not produce the intended results while over-diffusion may cause the loss of image features.

The diffusion procedure can be integrated with neural networks, enabling the model to learn the ideal diffusion parameters directly from enormous volumes of data. Modern deep learning approaches combined with conventional diffusion theory have produced state-of-the-art results for tasks like picture denoising, restoration, and even generative tasks like image synthesis [23, 6].

2.3 Denoising Diffusion Probabilistic Models (DDPMs)

Denoising Diffusion Probabilistic Models (DDPMs) represent a class of generative models that have shown remarkable capabilities in generating high-quality images. They are part of the diffusion model family, which models the generation process as a Markov chain of gradual denoising steps. DDPMs are inspired by nonequilibrium thermodynamics, particularly the process of reversing diffusion, which is a physical process that degrades images. In the context of DDPMs, an image is gradually corrupted over a sequence of time steps by adding noise, and then a neural network is trained to reverse this process — to denoise the image. This is similar to learning the transition dynamics of particles from a higher entropy state back to a lower entropy state. During training, DDPMs learn to reverse a Markov chain that transforms data into noise. This is accomplished by starting with the data distribution and applying Gaussian noise in several steps to create a sample of pure noise. The neural network then learns to reverse these steps, effectively denoising the data to recreate the original image [15, 31].

The forward process of a DDPM corrupts the data gradually by adding noise over a series of discrete time steps t from 1 to T, where T is the total number of diffusion steps. This can be mathematically described as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Here, x_t represents the data at time t, β_t is a variance schedule defining how much noise to add at each step, and I is the identity matrix. The noise schedule β_t is carefully chosen to maintain a balance between the noise and the signal across each step.

DDPMs aim to learn the reverse of the forward diffusion process, which involves recovering the original data from the noise. The reverse process is modeled as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

In this equation, $\mu_\theta(x_t, t)$ is the mean function and $\Sigma_\theta(x_t, t)$ is the covariance function, both parameterized by the neural network with parameters θ. The network is trained to estimate these functions such that the noisy data x_t at each time step t can be reversed back to the data distribution x_0 [15].

Training a DDPM involves optimizing the neural network parameters θ using variational inference. The goal is to maximize the likelihood of the data by minimizing a variational lower bound on the negative log-likelihood. In practice, this often translates to minimizing a loss function that measures the difference between the original data and the data reconstructed by the reverse process at each time step.

DDPMs have considerable potential in medical imaging, particularly for tasks where high-quality image generation is crucial. They can be used for generating synthetic medical images, enhancing low-quality images, and potentially aiding in the diagnosis by sharpening subtle features that might not be easily discernible in noisy images [20].

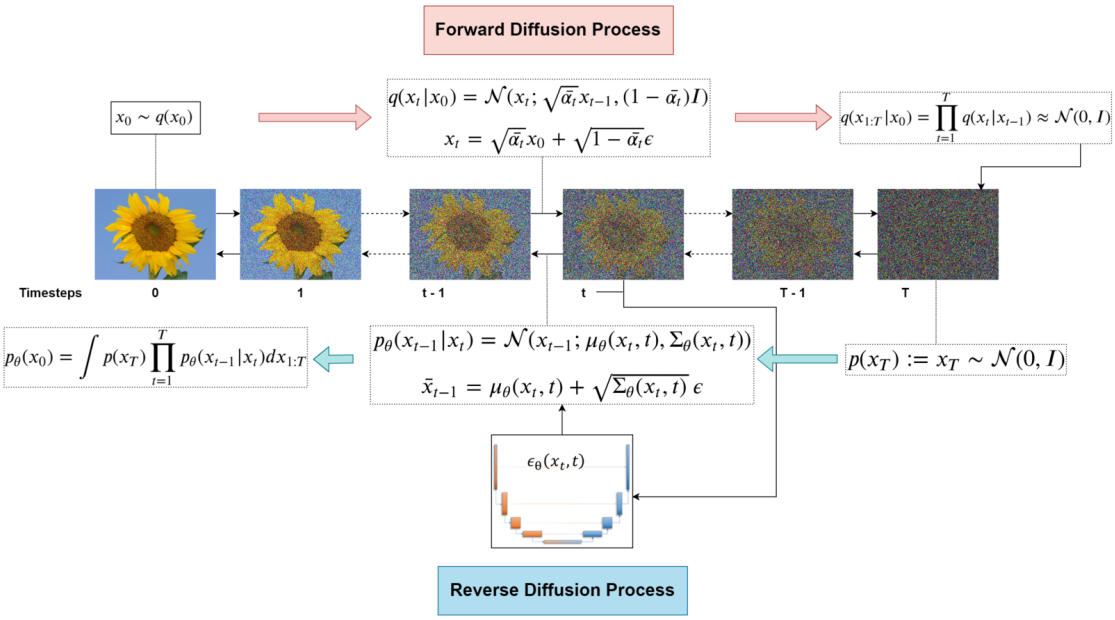

Figure 4. Denoising Diffusion Probabilistic Model (DDPM). The top sequence shows the forward diffusion process, where noise is incrementally added to an image across multiple timesteps, leading to a fully noised image. The bottom sequence shows the reverse diffusion process, where the model learns to iteratively denoise the image, recovering the data from noise. The equations describe the mathematical framework for both forward and reverse processes, highlighting the Markovian nature of diffusion and the iterative denoising strategy employed by DDPMs.

DDPMs offer several advantages over traditional generative models:

- They can generate images with higher quality and diversity.
- The generation process is stable and does not suffer from issues such as mode collapse, which is common in GANs.
- DDPMs have a well-defined likelihood, making them suitable for tasks that require precise probabilistic modeling.

While promising, DDPMs also come with their own set of challenges. The reverse diffusion process in DDPMs is computationally demanding due to the need for many iterative steps to transform the noise back into structured data. The computation required is directly related to the number of diffusion steps T, and the complexity of the neural network θ. Additionally, the quality of the generated images heavily relies on the number of diffusion steps and the capacity of the neural network.

2.4 Stable Diffusion

Stable diffusion is a recent advancement within diffusion models [26], focusing on stabilizing the training process and improving the quality of generated images. This approach typically involves modifying the diffusion process to ensure that the reverse diffusion path (from noise to data) remains stable and predictable across different runs and datasets. Stable diffusion models are designed to address the instability issues observed in some generative models, where small changes in input or parameters can lead to significant variations in output. By introducing stability in the diffusion process, these models can produce consistent and high-quality outputs, which is particularly crucial in medical imaging applications where precision is paramount [22].

Training stable diffusion models often requires careful calibration of the noise schedule—the rate at which noise is added during the forward diffusion process. This schedule is crucial to ensure that the reverse diffusion can effectively reconstruct the original data from the noise. To achieve stability, modifications to the noise schedule (see 2.3) are made. The noise schedule, which dictates the variance β_t at each timestep, is fine-tuned to maintain a balance that allows for reliable reconstruction of the original image during the reverse diffusion process. This ensures that the trajectory from the noised image back to the original data, parameterized by $p_\theta(x_{t-1}|x_t)$, is stable even with variations in input.

In medical imaging, stable diffusion can be leveraged for tasks such as creating detailed synthetic images for training diagnostic algorithms, reconstructing corrupted images, or enhancing features in images that assist in clinical diagnosis.

The main advantage of stable diffusion in medical imaging lies in its robustness and reliability. Unlike other generative models that may produce variable results, stable diffusion ensures that the generated images are consistent, which is vital for medical diagnosis and treatment planning. Moreover, the improved stability can potentially reduce the need for extensive datasets during model training, as the model can generalize better from limited data [22].

2.5 Integration of Neural Networks in Diffusion Processes

The integration of neural networks within diffusion models for synthetic medical image generation represents a confluence of two advanced technologies, aiming to leverage the precision of deep learning for enhancing the generative capabilities of diffusion-based techniques. This combination is particularly exemplified in the use of U-Net architectures, which have been pivotal in advancing the field of medical image analysis and synthesis [6].

2.5.1 U-Net Architecture for Diffusion Models

U-Net architecture, originally designed for biomedical image segmentation, has proven to be effective for tasks requiring the preservation of high-resolution details across images. Its architecture is characterized by a series of down-sampling layers that capture context and an equivalent series of up-sampling layers that enable precise localization, linked by skip connections that preserve spatial hierarchies across the network. This structure is inherently suited for the iterative refinement required in diffusion models, particularly in the context of medical imaging where the accurate representation of complex anatomical structures is paramount [27].

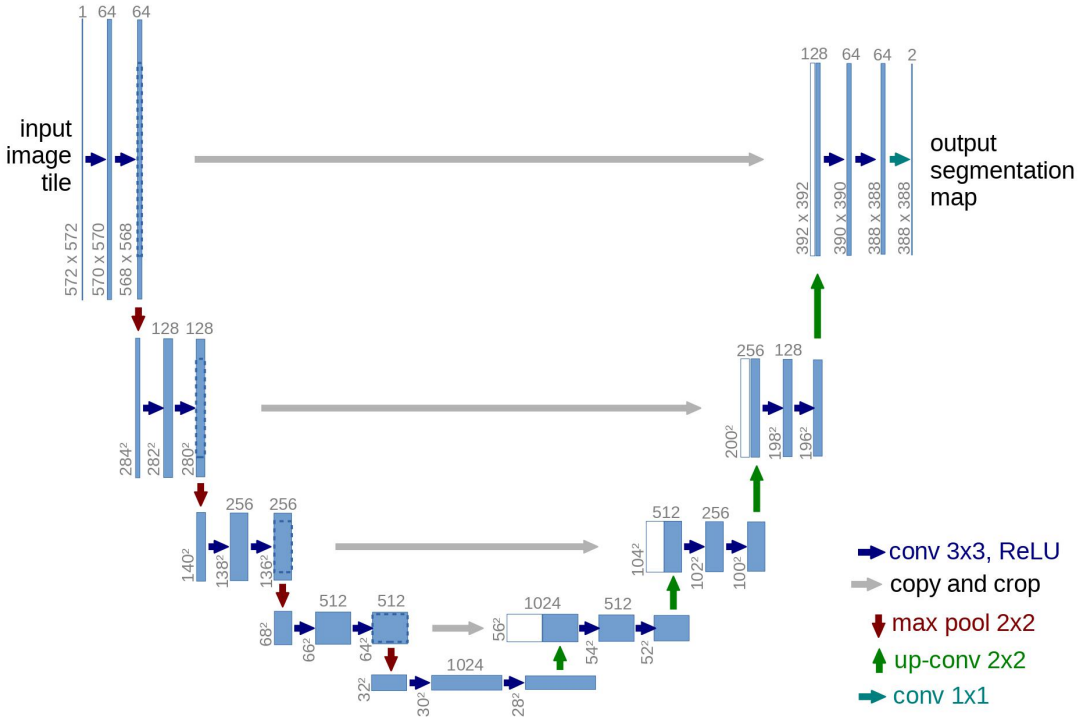

Figure 5. U-Net architecture [27]. The network's contracting path captures context through down-sampling, while the expansive path enhances localization via up-sampling, with skip connections bridging corresponding layers to preserve detailed spatial information.

In diffusion models, the U-Net architecture functions as the backbone for the denoising step, where the goal is to predict and reverse the added noise at each iteration, gradually moving towards the original, noise-free image. The model's ability to process and integrate information across multiple scales allows it to effectively reconstruct images with a high degree of fidelity, capturing both the macroscopic structures and the subtle textures that are characteristic of medical images.

The integration of neural networks with diffusion models necessitates specific architectural optimizations to ensure the effective synthesis of medical images. Key architectural optimizations include adjusting the depth and width to balance computational efficiency with the ability to capture complex image details. Enhancements to skip connections within the U-Net architecture improve information propagation, crucial for maintaining spatial details in image reconstruction. The integration of attention mechanisms [33], where attention scores focus processing on areas indicative of abnormalities, enhances diagnostic accuracy and the clinical utility of synthesized images. These modifications ensure that the networks not only generate high-quality images but also prioritize medically relevant features, increasing the reliability of the outputs.

2.5.2 Conditional Image Generation

A notable advantage of incorporating neural networks into diffusion models is the ability to perform conditional image generation. By conditioning the diffusion process on specific attributes, such as the presence of particular pathologies or patient demographics, neural networks can guide the synthesis of images that meet precise clinical criteria. This capability is crucial for creating diverse, realistic datasets that can support a wide range of diagnostic and therapeutic applications.

Overall, the integration of neural networks, especially the U-Net architecture, into diffusion processes for synthetic medical image generation, represents a significant advancement in the field of medical imaging. This approach not only enhances the quality and applicability of generated images but also opens new avenues for research and clinical application, promising to address some of the most pressing challenges in medical diagnostics and treatment planning. The combination of deep learning's nuanced understanding of image features with the iterative refinement offered by diffusion models creates a powerful tool for synthesizing medical images that are both realistic and anatomically precise. As research in this area progresses, further optimization of neural network architectures and training methodologies is expected to unlock even greater potential in synthetic data generation, contributing to advancements in personalized medicine, diagnostic accuracy, and the development of AI-driven healthcare solutions.

2.6 Metrics for Evaluating Synthetic Medical Images

Evaluating the quality of synthetic medical images is important for their application in clinical practice, research, and training. The assessment can include various metrics that collectively ensure the images' realism, accuracy, and utility.

2.6.1 Fréchet Inception Distance (FID)

FID evaluates the quality of synthetic images by comparing the distance between the feature vectors of real and generated images, computed using an Inception-v3 model pre-trained on ImageNet [14]. Lower FID scores indicate closer similarity between synthetic and real image distributions, suggesting higher quality of the generated dataset. FID is particularly useful in medical imaging to quantify the realism and variability of synthetic datasets, even when the generative model is not based on GANs. This metric is crucial for ensuring that synthetic medical images can effectively mimic the complexity and diversity of real anatomical features.

2.6.2 Inception Score (IS)

The Inception Score evaluates synthetic images based on two key aspects: the clarity of image classification and the diversity of generated images [30]. It uses a pre-trained Inception model to predict class labels for each image, calculating the score based on the distribution of these labels across the dataset. A higher IS indicates clearer, more diverse images. Despite its utility, the IS's ability to provide a comprehensive evaluation of image quality, especially in medical imaging, is limited by its focus on classification confidence and diversity, potentially overlooking the clinical relevance and anatomical accuracy of the generated images.

3 Methods

In this thesis, we employ a comprehensive approach to explore the capabilities of diffusion models in generating synthetic CT scans. Our methodology encompasses data acquisition, preprocessing, model selection, adjustment, and optimization, and the evaluation of synthetic image quality. This section outlines the systematic procedures and analytical frameworks adopted to ensure the reliability and validity of our research findings.

3.1 Data

The important part of any machine learning-based study is the quality and diversity of the data used. For this thesis, we focus on publicly available CT datasets, which provide rich information for training and testing of our models. The choice of CT data is motivated by its widespread use in medical diagnostics and the potential impact of synthetic data augmentation in this area.

3.1.1 TotalSegmentator

The TotalSegmentator [34] dataset comprises a diverse collection of CT scans from 1204 patients. This dataset includes annotations for 104 anatomical structures, including 27 organs, 59 bones, 10 muscles, and 8 vessels. Originating from varied scanners, institutions, and protocols, this dataset is based on 1204 CT examinations from 2012, 2016, and 2020. These images, randomly sampled from routine clinical studies, ensure representation across different age groups and abnormalities. The dataset's origins from varied scanners and institutions enhance its generalizability. By incorporating data from different sources, TotalSegmentator ensures that the developed segmentation models are not overly specialized to the particular characteristics of a single imaging protocol or scanner type.

Each CT scan comprises individual segmentation files corresponding to specific anatomical structures, such as separate files for the right and left kidneys. In our training, the 'kidney' class was established by merging the annotations for the right and left kidneys. Similarly, the 'lung' class was synthesized by combining annotations from six distinct lung regions: lung lower lobe left, lung lower lobe right, lung middle lobe right, lung middle lobe left, lung upper lobe left, and lung upper lobe right. For liver-related studies, we utilized the dataset's 'liver' class. For each case, we focused on slices that exclusively contained the organ of interest, thereby ensuring specificity in our analysis and skipping the unrelated data.

3.1.2 KiTS

The KiTS dataset, also known as the Kidney Tumor Segmentation Challenge dataset, focuses only on kidneys [13]. The KiTS Challenge is a bilennial competition that invites researchers to develop their segmentation algorithms, which are then evaluated against a hidden test set for accuracy, precision, and other metrics. This initiative has significantly contributed to advancing the field of medical image analysis, particularly in automating and improving the accuracy of kidney and tumor segmentation tasks. The dataset typically includes 599 cases, each provided with corresponding segmentation masks of three distinct classe: kidneys, tumors and cysts present in the kidneys. While the primary aim of the KiTS Challenge is to propel advancements in medical image analysis, particularly by enhancing the automation and precision of segmenting kidneys, tumors, and cysts, our research adopts the KiTS 2023 dataset for conditinal scans generation.

In the KiTS dataset processing stage, contrary to combining tumor, cyst, and kidney annotations into a single class, we opted to focus exclusively on the tumor class since we already had 'kidney' class in TotalSegmentator. This decision was driven by our specific research aim to enhance the generation of synthetic tumor data, which can lead to significant increase of data amount for tumor detection and segmentation.

Artificial KiTS dataset To check the generalization capabilities of our tumor inpainting model, we developed an artificial KiTS dataset. This process involved insertion of tumor masks from one patient's CT scan onto the kidney region of another, thereby creating synthetic data. Each CT scan in the KiTS dataset was paired with another randomly chosen scan. For each pair, we aligned the kidney-containing slices, ensuring that the introduced tumors would reside within the kidney tissue. We identified slices containing the kidneys by utilizing the segmentation masks provided within the KiTS dataset. If the source scan had a broader range of kidney slices with tumors than the target, we mapped the source's tumor masks to a corresponding subset of the target's kidney slices, ensuring the most possible overlap. Post-generation, each synthetic image underwent a manual verification process to ensure the tumors were appropriately placed within the kidneys, and if not, they were removed from the artificial dataset. Similar to the original KiTS data, the artificially augmented images were normalized and resized to maintain consistency across the dataset. This introduced a new level of complexity and variability in tumor generation, thereby providing a possibility of creating new synthetic data and expanding the dataset size.

Creation of 2D dataset We started experiments from 2D dataset, concentrating on specific cross-sectional views. Although it is possible to slice 3D data across axial, sagittal, and coronal planes (Figure 6), we opted for simplicity in our dataset by exclusively utilizing axial plane slicing.

In medical imaging, axial slices are particularly valued for their detailed portrayal of cross-sectional anatomy, offering a distinct and comprehensive view that is paramount

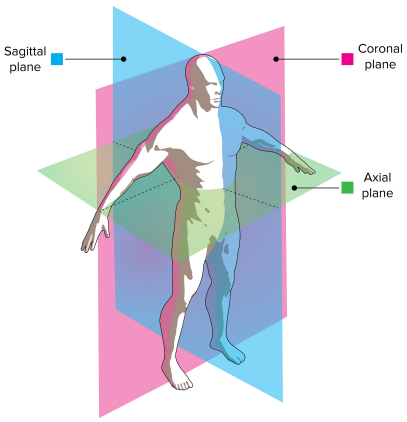

Figure 6. Three primary anatomical planes used in CT scans: the sagittal plane (in blue), coronal plane (in pink), and axial plane (in green). The sagittal plane bisects the body into left and right halves, the coronal plane divides it into anterior (front) and posterior (back) parts, and the axial plane separates the body into upper (superior) and lower (inferior) sections.

in clinical scenarios. The axial plane is especially beneficial in clinical settings for the detailed examination and identification of various pathologies or anatomical structures, making it a pivotal component in the diagnostic process.

Although our work is not focused on segmentation, the choice of axial slices significantly impacts the data analysis process. Axial plane images are crucial in providing clear and detailed views of anatomical structures, which aids in enhancing the accuracy and efficacy of data interpretation. This focus on the axial plane enables a more targeted and nuanced analysis, which is essential in applications where precise understanding and visualization of anatomy are necessary, such as in the generation of synthetic data or other forms of medical image analysis.

Creation of 3D dataset The 3D dataset is highly beneficial because it maintains the volumetric spatial relationships that are essential for understanding and analyzing complex anatomical structures. In medical imaging, the three-dimensional context provides crucial information that can be lost when considering only 2D slices. This perspective is particularly important for tasks such as segmentation, where accurately delineating the boundaries of structures requires knowledge of their shape and position within the entire volume.

Processing 3D scans and creating masks in a 3D context allows for a more accurate representation of anatomical features, which is critical for applications such as surgical planning, disease diagnosis, and treatment monitoring. It supports the identification of irregularities and pathologies that may not be as apparent when viewed on a single 2D plane. Moreover, it aids in the development of more sophisticated models that can make

better use of the available spatial data, leading to improved performance and reliability.

In our research, while we are not engaged in segmentation, the creation of a 3D dataset is instrumental for other analytical purposes. Our focus is on generating and utilizing 3D data to support a detailed understanding of anatomical structures, which is essential for tasks like synthetic data generation and other forms of medical image analysis. The capacity to handle 3D data directly makes our methods particularly suited for producing realistic and anatomically coherent synthetic images that can be used for a multitude of medical applications, including but not limited to enhanced data augmentation, preoperative planning, and medical education.

3.2 Diffusion Training

Diffusion models have shown great promise in various image generation tasks. In this thesis, we adapted two existing diffusion-based frameworks: the Palette framework for 2D inpainting [28], and the Med-DDPM model for 3D data generation and inpainting [8].

3.2.1 Palette Framework

The Palette framework, introduced by Saharia et al. [28], serves as the foundational model for our 2D image processing tasks, specifically in the domains of image uncropping, colorization, and inpainting. Palette is a versatile framework that leverages the capabilities of diffusion models to handle diverse and complex tasks like inpainting with remarkable fidelity and coherence.

Palette utilizes a U-Net architecture, which has been adapted from the class-conditional U-Net model proposed by Dhariwal and Nichol [7] with several critical modifications. Unlike the original 256×256 class-conditional model, Palette's architecture omits class-conditioning and incorporates additional source image conditioning through concatenation, as suggested by Saharia et al. [29], who recommend concatenating the source image with intermediate feature maps at various levels of the U-Net. This adjustment allows for a more focused and efficient training process, tailored to the specific requirements of image inpainting and colorization without the need for class labels.

In this modified U-Net architecture, the model consists of residual blocks, attention mechanisms, and specific channel multipliers at different levels of the network. Each residual block features a dropout rate of 0.2, enabling the model to prevent overfitting during training. Attention blocks are placed in the network to allow for long-range dependencies within the image, specifically at a downsampling rate of 16. This is facilitated by setting the number of channels per attention head to 32, allowing the model to adjust the granularity of the attention mechanism, meaning that the attention mechanism divides the feature map into groups of 32 channels, with each group processed by a separate attention head. The U-Net employs channel multipliers at different levels,

scaling the channels by factors of 1, 2, 4, and 8 as the network processes the input deeper, enhancing the model's capacity to capture and synthesize complex image features. As the network depth increases, the increased channel capacity allows for richer feature representations, which is crucial for the network's ability to perform complex image transformations.

The foundation of the Palette framework is a conditional diffusion model, a concept introduced by Chen et al. [3] and further developed for image-to-image applications by Saharia et al. [29]. These models are pivotal in making the denoising process contingent upon an input signal, thereby equipping the framework with the capability to undertake a wide array of complex image transformations. Here, both the input (x) and output (y) are images, facilitating the model's ability to perform transformations that are crucial for tasks like restoring damaged or incomplete medical images and generating high-fidelity visual content.

The models consist of a forward diffusion process and a reverse denoising process. The forward diffusion process, a Markovian process, iteratively adds Gaussian noise to a data point over T iterations, following the equation:

$$q(x_{t+1}|x_t) = \mathcal{N}(x_{t+1}; \sqrt{1-\beta_t}x_t, \beta_t I) \tag{1}$$

where β_t are the hyperparameters defining the noise schedule. The process is designed so that at the final step T, the data distribution is indistinguishable from Gaussian noise.

The reverse process is learned and aims to invert this forward process. Given a noisy image x_t, the goal is to recover the original image x_0. This is achieved by parameterizing a neural network to predict the noise vector added at each step, optimizing the objective:

$$\mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right] \tag{2}$$

where ϵ is the noise vector and ϵ_θ is the neural network's estimate of this noise.

Inpainting with Palette The Palette framework, adapted for our purposes, uses conditional diffusion models not for reconstructing or repairing medical images but for generating new data within medical imaging. Specifically, the framework is employed to synthesize realistic anatomical structures, like organs, within predefined boundaries in medical images. This approach allows us to create detailed and accurate medical images from scratch or enhance existing ones by adding new, synthetic elements like tumors that are indistinguishable from real anatomical features (example of training and inference pipelines are depicted in 7).

During the training phase, the Palette model was exposed to a variety of medical images, learning to generate specific anatomical classes such as the liver for example, within their accurate ground truth boundaries. The model's ability to understand and interpret the spatial and contextual relationships in medical images enables it to predict

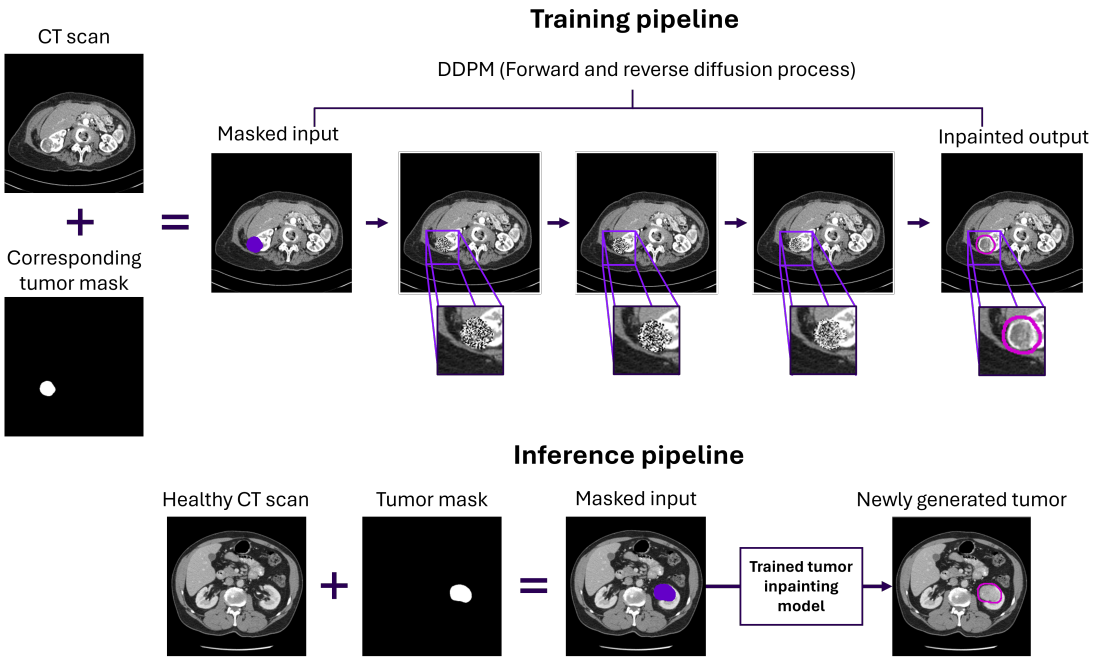

Figure 7. Workflow of the tumor inpainting model training and inference process. The upper section illustrates the training phase, where an initial CT image is combined with a mask to create the input image. The model then iteratively refines this region to generate a synthetic tumor. The lower section depicts the inference phase, where a trained model uses a given CT image and mask from another CT to create a realistic tumor in a new place.

and generate pixel values for new, unseen areas within these boundaries, effectively creating new data where none existed before.

This generation process is particularly valuable in medical imaging, where having a diverse set of accurate and detailed images is crucial for various applications, including training machine learning models, enhancing diagnostic processes, and supporting medical education.

3.2.2 Diffusion Model Algorithms

The Palette framework employs two main algorithms for training and inference:

Algorithm 1: Training the Denoising Model f_θ

1: **repeat**
2: Sample $(x_0, y_0) \sim p(x, y)$
3: Sample $y \sim p(y)$
4: Sample $\epsilon \sim \mathcal{N}(0, I)$
5: Take a gradient descent step on $\nabla_\theta \log f_\theta(x, \sqrt{y_t}y_0 + \sqrt{1 - y_t}\epsilon)$
6: **until** converged

The training algorithm iteratively updates the model parameters by sampling data points and corresponding noise. The objective is to minimize the difference between the sampled noise and the noise estimated by the model.

- (x_0, y_0): A pair of data samples drawn from the joint distribution $p(x, y)$, where x represents the input data and y represents the corresponding noisy data.
- y: A noisy version of the data sampled from the noise distribution $p(y)$.
- ϵ: A noise sample drawn from a standard normal distribution $\mathcal{N}(0, I)$.
- y_t: The noise level at timestep t in the diffusion process.
- $\nabla_\theta \log f_\theta$: The gradient of the log-probability of the model's prediction with respect to the model parameters θ.
- $\sqrt{y_t}y_0 + \sqrt{1 - y_t}\epsilon$: The noisy version of the original data y_0 at timestep t.

Algorithm 2: Inference in T Iterative Refinement Steps

1: Initialize $y_T \sim \mathcal{N}(0, I)$
2: **for** $t = T, \ldots, 1$ **do**
3: Sample $z \sim \mathcal{N}(0, I)$ if $t > 1$, else set $z = 0$
4: Update y_{t-1} using the reverse process formula
5: **end for**

6: **return** y_0

In the inference algorithm, we start with a sample from the noise distribution and iteratively refine it through the reverse process to approach the original data distribution.

- y_T: The initial noisy sample drawn from a standard normal distribution $\mathcal{N}(0, I)$.
- T: The total number of iterative refinement steps.
- t: The current step in the iterative refinement process, starting from T and decreasing to 1.
- z: A noise sample drawn from a standard normal distribution $\mathcal{N}(0, I)$. If $t = 1$, z is set to 0 to prevent adding additional noise in the final step.
- y_{t-1}: The refined sample at step $t - 1$, updated using the reverse process formula which aims to reduce the noise progressively.
- y_0: The final denoised output obtained after T iterative refinement steps.

Optimization Objective The optimization objective is formalized as:

$$\mathbb{E}_{(x,y)}\mathbb{E}_{\epsilon}\left[\|\epsilon - f_\theta(x, \sqrt{y_t}y_0 + \sqrt{1-y_t}\epsilon)\|_p\right] \tag{3}$$

This objective, also known as L_{simple} in [15], is equivalent to maximizing a weighted variational lower-bound on the likelihood. Here, x represents the input, y_t the target at time t, ϵ the noise, and f_θ the model's noise prediction function. This equation represents the expected value on the data and noise distributions of the p-norm of the difference between the actual noise and the noise predicted by the model.

Inference Process The inference process is defined by the following equation:

$$\hat{y_0} = \frac{1}{\sqrt{y_t}}(y_t - \sqrt{1-y_t}f_\theta(x, y_t)) \tag{4}$$

In this equation, $\hat{y_0}$ represents the estimated original image, and the model corrects the noisy image y_t by subtracting the scaled noise predicted by f_θ. The inference equation calculates the estimated original image by correcting the noisy image with the prediction from the model, effectively reversing the noise added during the forward process.

3.2.3 Med-DDPM for 3D Data

In addition to the Palette framework, we employed the Med-DDPM model for processing 3D medical data. The Med-DDPM, a conditional diffusion model detailed by Dorjsembe et al. [8], is specially designed for medical image processing tasks in three dimensions. Our implementation focuses on generating and inpainting 3D CT scans, leveraging diffusion models' ability to model the distribution of medical image data effectively.

A key feature of Med-DDPM is its conditional generation capability, which allows the model to generate images based on certain specified conditions. In our case, this feature was leveraged to create CT scans conditioned on masks of anatomical structures. During inpainting, Med-DDPM focuses on the specified regions, filling them with plausible structures that align with the surrounding anatomy, thus generating synthetic yet anatomically coherent data segments within the scans.

Med-DDPM's architecture is based on a 3D U-Net structure, capable of processing the volumetric data inherent in 3D medical imaging. This architecture captures spatial relationships and anatomical coherence, crucial for maintaining the integrity of medical images. The model includes 3D residual blocks that facilitate the flow of information and gradients through the network, enhancing the learning process and enabling the model to capture complex patterns in the data. Med-DDPM integrates 3D attention mechanisms within the U-Net architecture similarly to our 2D model. The architecture employs channel multipliers at different levels, scaling the number of channels as the network processes the input deeper. This scaling is crucial for the model to build a rich representation of the data, aiding in the accurate generation of 3D structures. The training of Med-DDPM incorporated a forward diffusion process that progressively introduces noise into the data, followed by a reverse process where the model learns to denoise, aiming to recreate the original image from the noisy input. This process is mathematically governed by a set of equations that describe the noise addition and the prediction of the denoising steps:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{5}$$

$$\hat{x}_0 = \frac{x_t - \sqrt{1 - \alpha_t}\hat{\epsilon}_\theta(x_t, t)}{\sqrt{\alpha_t}} \tag{6}$$

where x_0 is the original image, x_t is the noised image at timestep t, α_t is a variance schedule, and $\hat{\epsilon}_\theta$ is the predicted noise by the model.

3.3 GAN models

Generative Adversarial Networks (GANs) have been a breakthrough in generative models, renowned for their ability to create high-fidelity synthetic images. In our work, we utilize

GANs as a comparative measure to the diffusion model's performance in synthetic medical image generation. For 2D image synthesis, we employ the Pix2Pix framework [16], and for 3D data generation, we utilize the Vox2Vox model [4].

The Pix2Pix model by Isola et al. [16] provides a framework for image-to-image translation in a supervised context, adept at translating input images into output images within a paired dataset framework. This capacity makes it especially suited for a range of tasks, including medical image synthesis. Pix2pix uses a conditional GAN setup, comprising a generator and a discriminator. The generator is based on a U-Net architecture, which excels at preserving context, crucial for generating detailed images. The discriminator is a convolutional PatchGAN, which assesses the authenticity of image patches rather than the whole image, thereby concentrating on fine details. The objective function is a blend of adversarial and L1 losses, ensuring that the images are not only convincing but also closely resemble the ground truth on a pixel level.

Extending the Pix2Pix approach into three dimensions, the Vox2Vox model by Cirillo et al. [4] is tailored for processing volumetric data, making it well-suited for generating 3D images such as organ synthesis and brain tumor segmentation within CT and MRI scans. Mirroring its 2D counterpart, Vox2Vox features a 3D U-Net-like generator and a 3D PatchGAN discriminator. The generator captures spatial dependencies critical for realistic 3D image generation, while the discriminator operates on 3D patches, discerning the local textures and structures within the medical images.

3.4 CT Window Setting

The CT window, or windowing, is a technique used in medical imaging to enhance the visibility of specific tissues or structures within CT scans. By adjusting the window level (WL) and window width (WW), radiologists can highlight particular ranges of Hounsfield units (HU), which represent different densities within the body. In the context of image processing and inpainting, it's essential to apply the appropriate CT window settings when providing 2D visualizations. The selected window directly influences the range of pixel values in the image and can impact the effectiveness of algorithms in detecting features, segmenting tissues, or inpainting missing segments.

Choosing the correct CT window is crucial when working with medical images, especially when different tissues or organs are of interest. Different tissues require different window settings to optimize their visibility. For example, lung tissues and soft tissues like the liver or kidneys have significantly different densities, and thus, distinct window settings are necessary to visualize them effectively (Figure 8).

Lung Window: Typically set with a WL of -600 to -700 HU and a WW of 1500 to 2000 HU, the lung window allows for optimal visualization of the pulmonary structures, emphasizing air-filled areas and lung parenchyma while deemphasizing denser structures like bones and soft tissues.

Soft Tissue Window: Commonly used for visualizing organs like the liver and kidneys, the soft tissue window is usually set with a WL of around 40 to 60 HU and a WW of 350 to 400 HU. This setting enhances the contrast between different soft tissues, allowing for better differentiation of organs, muscles, and blood vessels.

We have experimented with changing the CT window as both a preprocessing and postprocessing step. Adjusting the CT window in preprocessing modifies the data to the needed window, allowing the model to better understand and emphasize the structure of interest. In contrast, adjusting the CT window in postprocessing aids only in visualization, without affecting the model's understanding or processing of the data.

In preprocessing variant, before feeding the images into the inpainting model, we adjusted the CT window to focus on the relevant anatomical structures. For soft tissues, we used a window width (WW) of 400 Hounsfield Units (HU) and a window level (WL) of 50 HU, which is typical for soft tissue visualization. For lung structures, the window settings were adjusted to a WW of 1500 HU and a WL of -600 HU, optimizing the visibility of lung parenchyma (functional tissue of the lungs that is involved in gas exchange) and associated pathologies.

In postprocessing version, after the inpainting process, we applied CT window settings again to enhance the visualization of the specific structures in the output images. The same window settings used in preprocessing were applied in postprocessing to ensure consistency in visualization and to allow for an accurate assessment of the inpainting results.

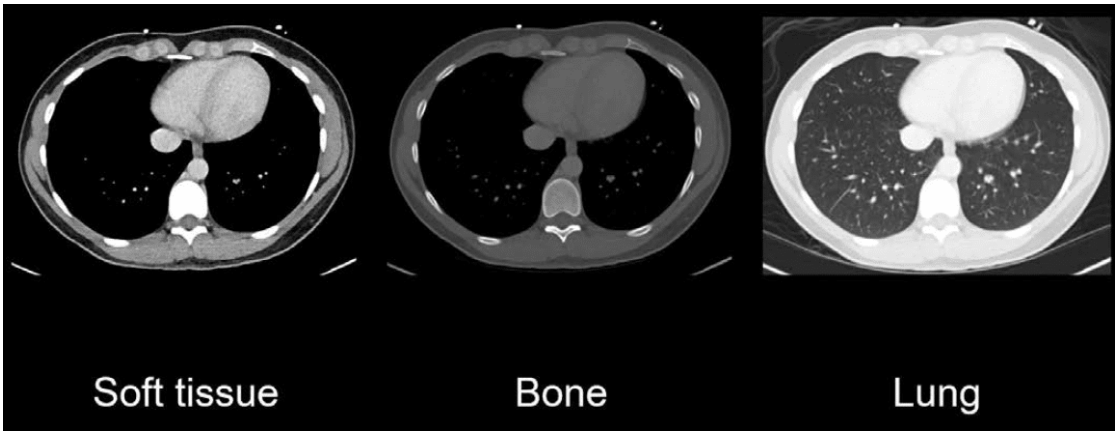

Figure 8. Example of different CT window settings.The image displays three axial CT scans of the chest, each optimized for different tissue contrasts by adjusting the CT window settings. The first scan is set to the soft tissue window, highlighting the heart, aorta, and other soft tissue structures with clear delineation. The second scan applies the bone window, enhancing the visibility of the spine and other bones. The last scan uses the lung window, which is ideal for assessing the lungs and airways, revealing the pulmonary textures, vascular markings, and air-filled spaces distinctly.

3.5 Segmentation model

To assess the impact of synthetic data on segmentation performance, we employed nnUNet, a highly adaptive framework designed for medical image segmentation. nnUNet, short for "no new U-Net," automatically adjusts its configuration based on the data it is given, optimizing network architecture, preprocessing, and training strategies to maximize segmentation accuracy.

The nnUNet operates with a dynamic architecture that adapts depending on the specifics of the dataset, including image dimensions, spacing, and the number of segmentation classes. The model determines the most effective U-Net configuration by analyzing these dataset attributes, then sets parameters such as the number of layers, features, and training strategies without manual intervention. This approach allows nnUNet to achieve state-of-the-art performance by tailoring its network to best fit the segmentation task at hand.

For our work, nnUNet was applied in three distinct experimental setups to evaluate segmentation performance: training on pure real data, pure synthetic data, mix of real and synthetic data. Each configuration was trained under the same conditions to maintain consistency, using a fixed set of hyperparameters prescribed by the nnUNet framework based on the input data characteristics. The effectiveness of nnUNet in these different training environments was evaluated against a consistent test set composed exclusively of real images, ensuring that the evaluation reflects real-world applicability. Performance metrics such as the Dice coefficient and Jaccard index were used to quantify the segmentation accuracy, providing insights into the utility of synthetic data in training segmentation models.

4 Results and Discussion

The section will examine the performance of diffusion models we have adapted and how they adapt to the complexities of medical data. We will detail the process of our modifications and adaptations, rather than focusing solely on the final model, including the steps of preprocessing, model training, and post-processing. The quality of the generated images will be assessed, with particular emphasis on their anatomical accuracy. The augmentation of datasets with synthetic images generated by diffusion models presents a promising solution to the pressing issues of data scarcity and privacy in medical research.

Moreover, the discussion will address the practicality of diffusion models application in a clinical setting. We will also consider the limitations and potential pitfalls of relying on synthetic data, from the accuracy of the models to the ethical considerations of their use.

Figure 9 illustrates the primary useful scenarios we tested in our model's capability to generate 2D medical images. Each scenario shows the generated output alongside its ground truth and mask, demonstrating the model's ability to accurately fill in masked regions with plausible anatomical details.

4.1 2D diffusion generation

Initially, the model was designed for general RGB images, which provided a foundation for training but was not optimized for the unique requirements of medical imaging. Training the model in this manner yielded relatively good results for some structures but failed for others, such as the lungs. To address this, we modified the model to predict a single channel instead of three. This approach reduced the complexity of the task, allowing the model to focus on producing accurate results for the one channel we needed. Consequently, we began to see improved results for structures like the problematic lungs.

Kidneys From the initial epochs, synthesized kidneys had a satisfactory level of detail in terms on having texture and not being homogeneous, but they were a bit too bright. This was particularly evident when using the three channel approach. However, switching to grayscale processing, which aligns more closely with the nature of CT scans, resulted in a significant improvement. As the model underwent further training, the synthetic representations gradually approached the fidelity of actual CT scans so it became harder to differentiate between real and generated. This advancement in the model's capabilities is captured in Figure 10, which depicts the final output of the model alongside the initial masked input and the ground truth. The development across epochs is further illustrated in Figure 11. The progression of epochs showcased a notable evolution in the synthetic images, with the later epochs producing outputs that closely mirrored the authentic scans.

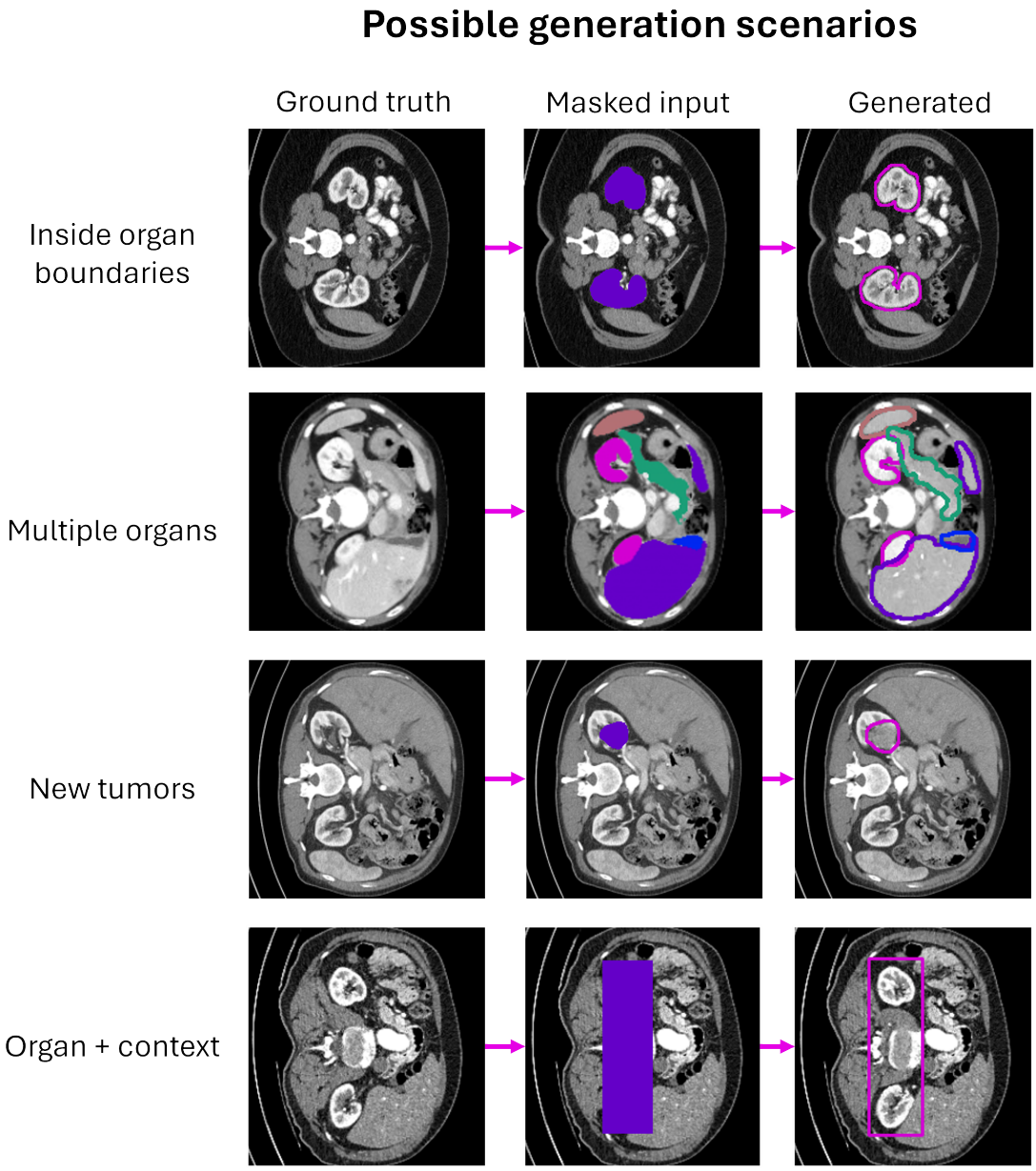

Figure 9. Demonstration of the model's generation capabilities across various scenarios: generating structures inside organ boundaries, generating multiple organs from one binary mask, generating new tumors, and generating organs with context around it. Each case includes the ground truth, masked input, and the generated output.

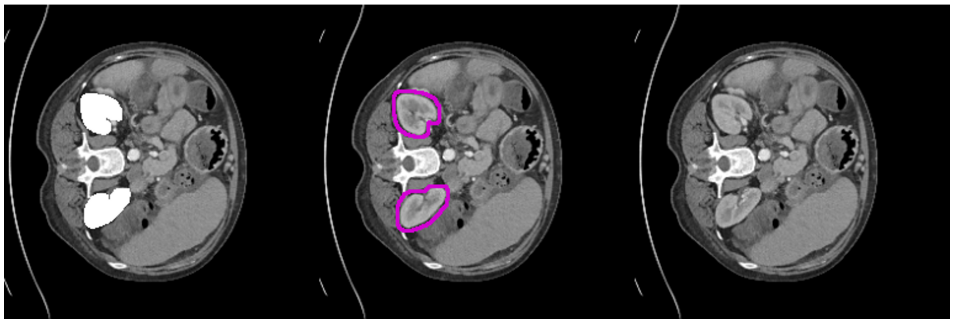

Figure 10. Synthetic kidney inpainting in a CT image in 2D slice. From left to right: input for inpainting with the kidney region masked out; synthetic kidney generated by the model, with the area of interest outlined in pink; the ground truth (GT). The model showed capability to generate anatomically coherent structures that closely align with the real organ appearance.

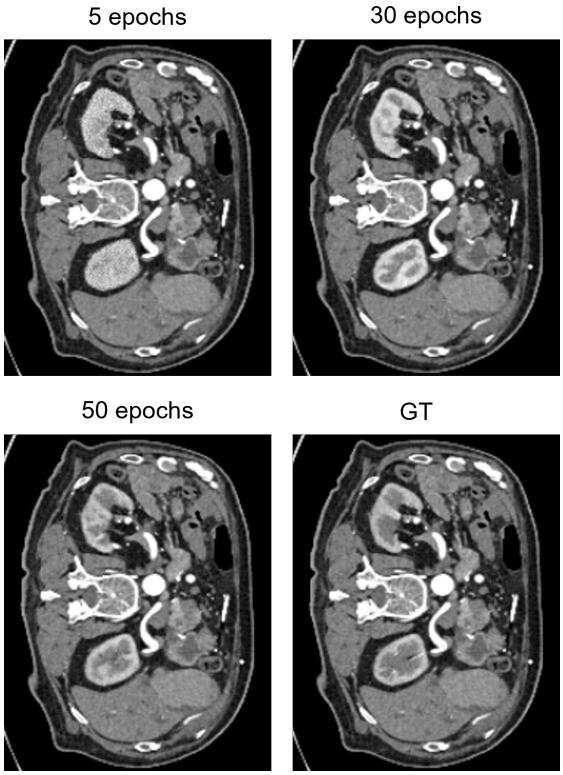

Figure 11. Progression of kidney inpainting over training epochs. The top left image represents the outcome after 5 epochs, indicating the early stage of learning with less detail. Progression to 30 epochs, shown top right, yields more refined details. The bottom left image displays the results after 50 epochs, where the inpainting quality has markedly improved. The bottom right image is the ground truth (GT) against which the inpainted images are compared.

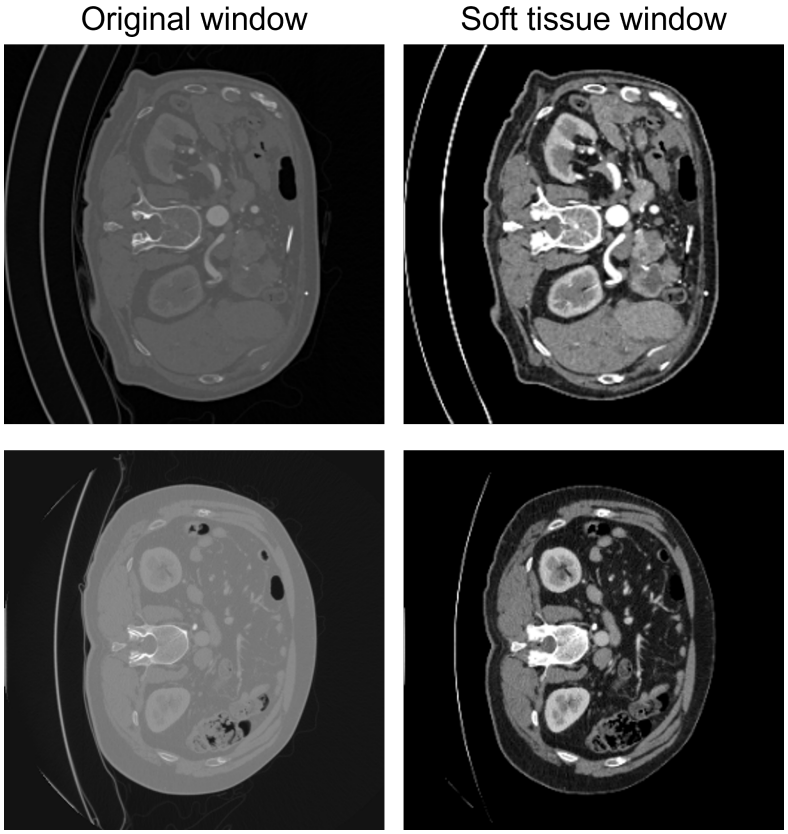

Figure 12. Comparative visualization of CT scans, showing kidneys with and without the application of soft tissue window settings. Left column images display non-windowed slices, where the kidneys appear as uniform structures with limited visible detail. Right column images are corresponding windowed, highlighting the enhanced visibility of kidney texture and anatomical structures.

The subsequent experiments involved utilizing windowed slices specifically tuned for soft tissues, which show the proper density range where kidneys are best visualized which can be seen in Figure 12. This approach further refined the model's output, enhancing the visibility of subtle textural details and anatomical structures within the kidney.

Lungs In the early attempts at lung synthesis, the model produced incorrect contrast, where the generated regions appeared overly bright, lacking the intricate network of blood vessels characteristic of lung tissue. The model made attempts to introduce texture that was distinct from other organs, such as the liver or kidneys, aiming for a more heterogeneous representation which can be seen in Figure 13 A and B. However, the complexity of the lung's structure was not adequately captured, leading to synthetic outputs that were not in line with the real lungs in CT scans. Potentially, the model was trying to get more context from the rest of the CT scan than needed, leading to overly

bright regions.

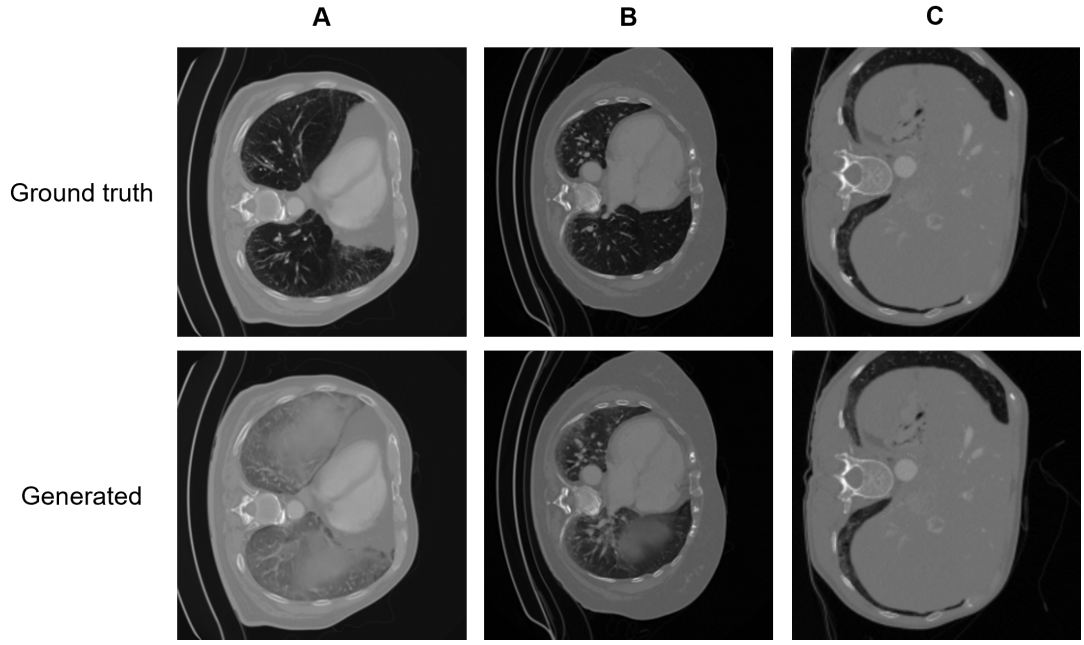

Figure 13. **A** depicts the sample with incorrect contrast in the generated region. **B** shows the sample where the model introduced a texture to the lungs that is more similar to soft tissue. **C** presents the example of smaller region of lungs where the contrast was correct from the beginning epochs.

Interestingly, the model performed visually better when dealing with smaller lung regions. In scenarios where the lung tissue occupied a lesser portion of the axial slice, the model managed to achieve a more accurate contrast level, suggesting its potential to handle localized lung features more effectively than larger expanses of lung tissue (example can be seen in Figure 13 column C).

Later, during training, the results started to get much better, revealing the structures similar to vessels, as well as correct contrast (more examples can be seen in Figure 23). It is worth noting that to get to the desired state in lung generation, we had to train the model for more epochs than other organs. While there were some problematic images, there were a lot of samples with satisfying synthetic regions, still showing a potential to use diffusion model to generate lungs.

Liver During the initial phase of liver image generation, the model produced images with improper contrast, resulting in a depiction of the liver that stood out unnaturally from the surrounding tissues. The liver appeared overly luminous and the texture was insufficiently developed, making it relatively simple to identify the images as synthetic. As the model continued to train, there was a noticeable improvement in the contrast and

texture of the generated liver images. Over time, the synthetic livers began to integrate more seamlessly with the surrounding tissues, enhancing the overall realism of the generated CT scans.

A significant improvement was the application of CT window settings for liver. When processing non-windowed liver samples, the output resembled a uniform gray mass without discernible internal structures, which was especially evident when these 2D slices were compiled into a 3D volume and observed in software like Slicer with different windowing, indicating that the essential textural information was not captured during the generation process. However, when the model was provided with liver samples preprocessed with appropriate window settings for soft tissues, there was a notable improvement in the visualization of internal structures. The windowed input allowed the model to capture and replicate the liver's intricate vascular architecture, resulting in synthetic images where the liver tissue exhibited realistic variations in density and texture (Figure 24).

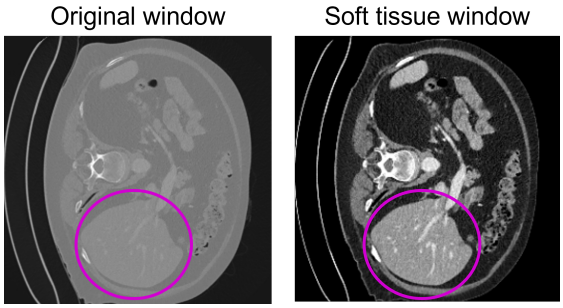

Figure 14. Contrast in liver visualization on CT images with and without soft tissue window settings. The left image presents a non-windowed liver slice, displaying a more homogenous appearance. The right image shows the same slice with soft tissue window settings applied, enhancing the visibility of the liver's vascular structures. The outlined region in pink highlights the liver area.

Multi-organ: kidney, liver, gallbladder, spleen, pancreas The multi-organ generation task posed a more substantial challenge, leading to a slower training progression. Initial results were not good, with organ brightness not adhering to the natural variance expected within a CT scan. Nevertheless, as the epochs advanced, improvements were observed. Certain organs began to exhibit a more realistic appearance, incrementally aligning closer to their genuine counterparts in terms of brightness, texture, and overall anatomical accuracy (Figure 15). Important to notice is the fact that we gave just one binary mask without any organ separation and the model was able to differentiate them properly.

Tumor In this thesis, we employed the KiTS dataset to focus on the task of tumor inpainting within kidney CT scans. Our approach had two parts: first, we inpainted

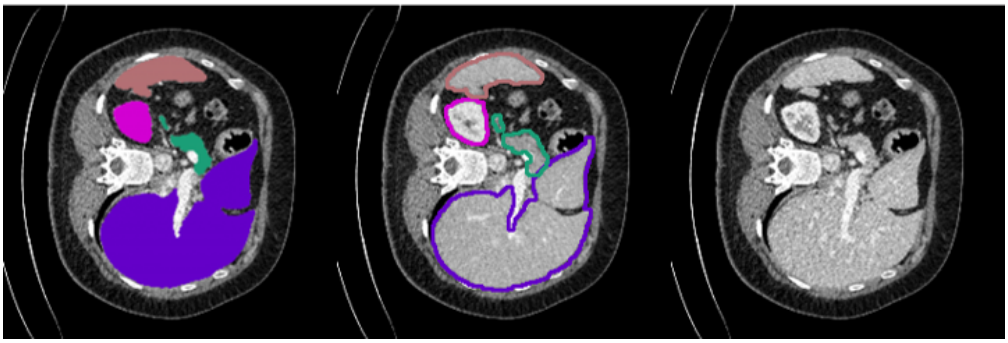

Figure 15. Synthetic multi-organ inpainting in a CT image in 2D slice. From left to right: input for inpainting with multiple organ regions masked out; synthetic organs generated by the model, with the areas of interest outlined in distinct colors; the ground truth (GT). Each organ is highlighted using a specific color: kidneys in magenta, liver in purple, spleen in blue, gallbladder in a orange, and pancreas in green. The model demonstrated its capability to generate anatomically coherent structures that closely align with the real organ appearances, effectively handling the complexity of multiple organ inpainting simultaneously.

original tumors from the dataset; second, we utilized our artificially augmented dataset, where tumor masks were placed within the kidney regions of different scans. Our model demonstrated proficiency in inpainting original tumors from the KiTS dataset, producing results that were anatomically plausible and closely aligned with the ground truth, while still providing some variability (see example in Figure 16). The model's success in this context suggests it has effectively learned the textural and structural nuances of kidney tumors, which are critical for generating realistic synthetic images. We believe that given the variability in tumor appearance and structure — far more than in organs like the liver or kidneys — the model's ability to generate plausible tumor images was better. The synthetic tumors varied in shape, size, and texture, reflecting the diversity observed in actual clinical scenarios.

The results indicate significant potential for using our approach to create synthetic tumor datasets. Unlike other organs, tumors do not have a set structure, such as specific vessels or ducts, allowing greater variability in their synthetic representation. This flexibility suggests that our method could be a valuable tool for generating diverse, realistic datasets for tumor analysis, potentially aiding in the development of more robust models.

In further tests of tumor generation, we explored the model's ability to synthesize entirely new tumor instances by integrating tumor masks with CT scans from different subjects. This approach not only demonstrated the model's capability to replicate the intricate details of tumor structures as observed in prior tests but also its potential to create authentic-looking, novel tumor images (example in Figure 17). The generated tumors

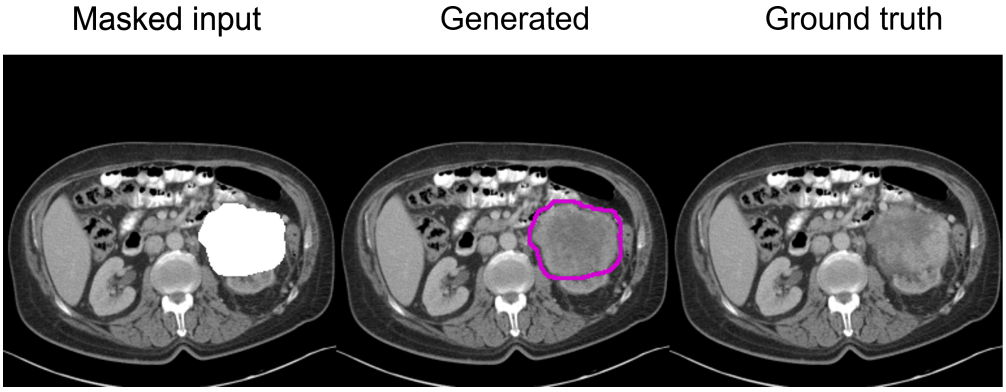

Figure 16. Synthetic tumor inpainting in a 2D CT slice. From left to right: the original scan with the tumor region masked for inpainting; the synthetic tumor generated by the model, delineated by a pink outline; and the ground truth (GT) for reference. The model successfully produced plausible tumor structures that exhibit an appropriate degree of variability, reflecting the diverse and irregular forms found in actual tumor presentations, which is beneficial given that tumors do not possess a uniform structure like other organs such as the kidneys.

exhibited realistic textures and variations, suggesting that the model can effectively use cross-subject data to enhance the diversity and realism of synthetic tumor samples.

Boxed kidneys All previous experiments were conducted using predefined boundaries of an organ. While this approach is useful, to create a more diverse dataset, we need to consider inpainting over a larger region. For this purpose, we introduced an expanded bounding box around the kidney regions in the CT scans. Specifically, this approach can help us to generate negative samples for control purpose for dataset, which have only positive cases (like KiTS dataset which has scans only with tumors and cysts). By covering larger areas, we can occlude the tumors that can be inside the region and generate healthy kidneys or other organs. The larger inpainting regions required the model to interpolate over more substantial gaps in the image data, meaning not only create one organ texture, but also any other part that lied in the mask not directly bordered by the target organ tissues. This model to not only fill in the missing kidney areas but also accurately extend the soft tissue environments around them.

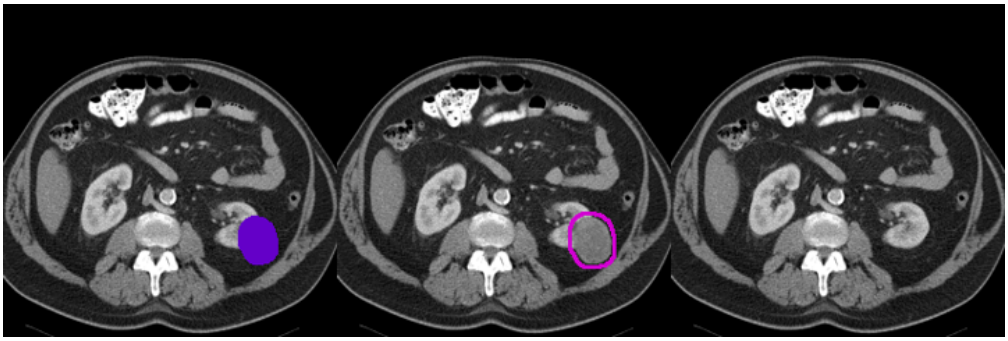

Figure 17. Newly generated tumor. **A** shows the area subjected to inpaint with original scan overlayed with tumor mask taken from another scan, **B** illustrates the inpainted tumor, demonstrating the model's ability to synthesize realistic tumor textures, and **C** represents the original CT scan without the tumor

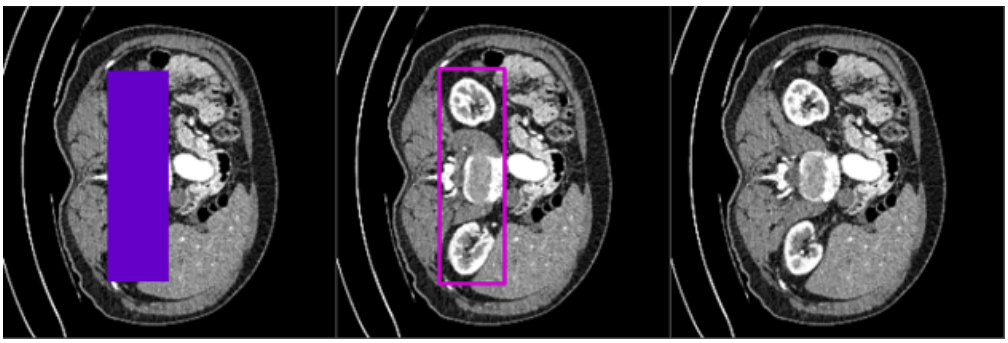

Figure 18. Visualization of the inpainting process for boxed kidneys using expanded bounding boxes. The left image shows the masked input where the kidney and surrounding tissues are hidden; the middle displays the inpainted output where the model has generated the missing areas; the right panel provides the ground truth.

Exploration of Original Hounsfield Unit (HU) Scale To enhance the realism of synthetic CT scans, we went beyond conventional image-based inputs to experiment with original HU scale arrays. This approach was aimed to work with the full spectrum of CT scan data, which inherently operates on the HU scale, diverging significantly from the standard 0-255 intensity range of typical image formats (Figure 19). This transition was motivated by the desire to maintain the intrinsic radiodensity information present in CT data, which is can be compressed or lost when converting to standard image formats. Despite the theoretical advantages of working directly with HU values, the practical outcomes were not as anticipated. The model struggled to effectively process the broad range of HU values, which led to the generation of regions of interest that appeared as gray masses. These outputs lacked the necessary textural and contrast variation. This challenge may be attributed to the model's difficulty in adapting to the wide and

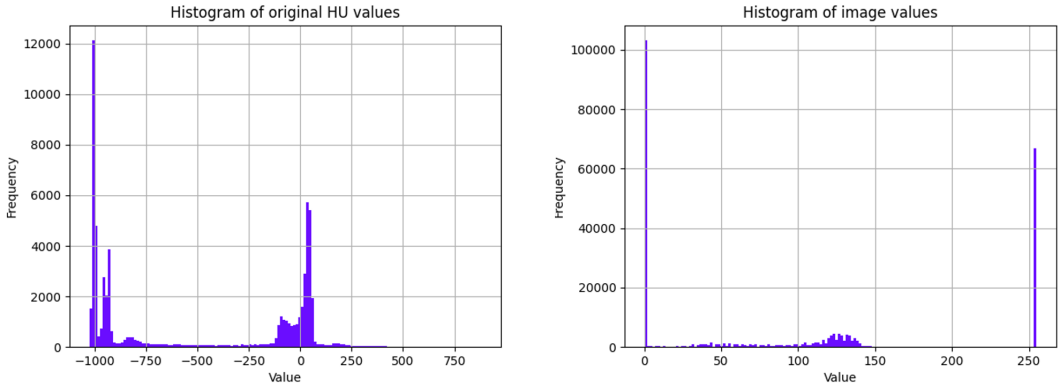

Figure 19. Distribution of values in the original CT scan slice with HU scale (left); distribution of values in image with set CT window for soft tissues and scaled to 8-bit (right).

unevenly distributed HU value range. Unlike the more uniform distribution of values in standard image formats, HU values in CT scans can vary significantly, with distinct peaks corresponding to different tissues or substances. The model's existing architecture and training methodologies might not have been optimized to handle this complexity, leading to an inability to discern meaningful patterns within the data.

4.2 2D GAN Generation with Pix2Pix

The exploration with the Pix2Pix model, initially in grayscale format, presented a different set of challenges and insights. While the model was adept at capturing broader structural details, it struggled with finer textures and details, particularly in the context of kidney and tumor synthesis.

Kidneys: The Pix2Pix model, despite its initial promise, often produced kidney images that were somewhat blurry and lacked the detailed fidelity required for clinical utility. This blurriness was consistent across different training epochs, suggesting a fundamental limitation in the model's capacity to replicate the intricate details of kidney anatomy in inpainting task.

Lungs: Lung synthesis with Pix2Pix showed a slightly better contrast adaptation compared to the kidneys. However, the model often interpreted the vascular structures of the lungs as sporadic bright spots, failing to capture the interconnected network of vessels and airways.

Tumor: The challenges in tumor synthesis with Pix2Pix were similar to those observed with kidney synthesis. The generated tumor images often lacked the necessary detail and variability, appearing blurry and less defined compared to the diffusion model outputs.

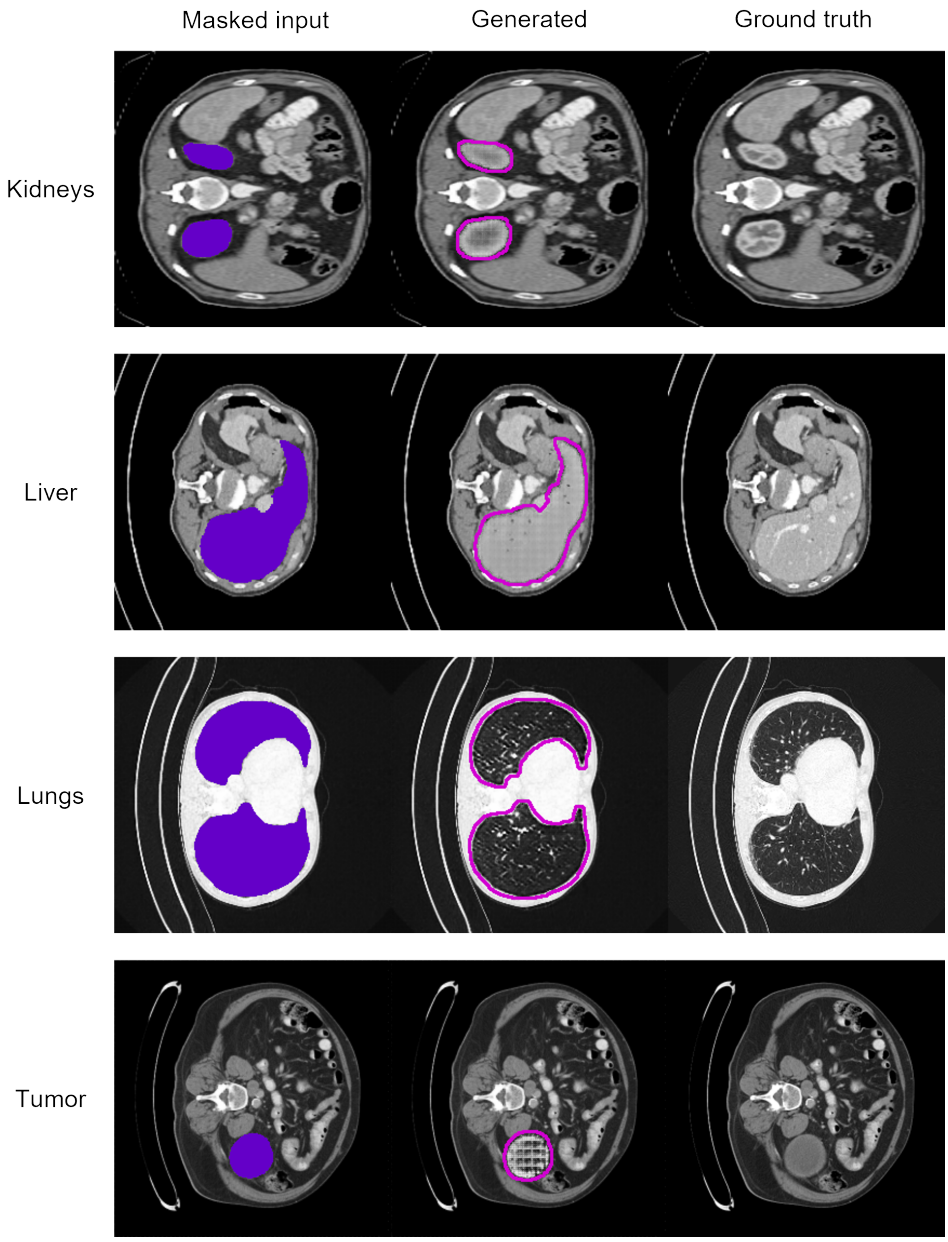

Figure 20. Example outputs from the Pix2Pix model for multiple tissue types in CT scans. The first column displays the masked CT slices with the tissues of interest in purple. The second column showcases the tissues generated by the Pix2Pix model, outlined in pink, which illustrate common issues encountered during this task with GAN model, such as blurriness and patchiness.

4.3 3D diffusion generation

Moving from 2D to 3D generation in medical imaging greatly improves our ability to understand complex anatomical structures. While 2D images can be helpful, they often miss the depth and continuity needed for a complete view. 3D generation solves this by providing consistent, detailed representations across all slices.

Initially, we aimed to generate full 3D CT scans from just binary masks representing all organs presence against a background. This ambitious task involved providing the model with only the masks of organs, without distinguishing between different organ types. However, given the model's initial underperformance in generating realistic full CT images, we shifted our focus to a more constrained inpainting task. Instead of generating entire CT scans, we concentrated on generating specific organs from their masks and integrating these into the original scans.

A significant enhancement in our Med-DDPM implementation is the introduction of conditional generation for inpainting tasks. Specifically, we condition the generation process on the available portions of the CT scan, allowing the model to inpaint masked organs. This is achieved by modifying the model's input to include a binary mask indicating the regions to be inpainted, alongside the CT scan. The model then focuses on these masked areas during the reverse diffusion process, effectively filling in the gaps with plausible data that aligns with the surrounding anatomical structure. When we decided to move to the inpainting task rather than trying to generate the full CT scan as we saw that part was not fully successful and was not producing medically meaningful data.

We modified the loss function to compute the loss only in the masked regions. Applying the input mask to the output of the denoising function and comparing it only to the corresponding masked region in the ground truth. However, this does not restrict the model's output to only the masked regions, but it optimizes the model's parameters to accurately generate the content within those regions. The following logical step would be to combine initial CT and generated organ. To produce outputs that fill only the masked regions and rest of the CT intact, additional steps in the sampling process are needed. Specifically, using the mask to blend the generated content with the original unmasked parts of the image.

First trial of changing our current model to inpainting included:

1. Generating noise for the current sampling step
2. Predicting the denoised image for the timestep t which will be the inpainting result for the masked region
3. Applying a binary mask to blend the predicted inpainted region with the original unmasked parts of the image
4. Adds noise to the sample depending on the timestep

These steps did not yield the desired results. The model's inability to learn the identity function led to poor integration of the generated content. The inpainted regions often did not align well with the surrounding anatomical structures, resulting in inconsistencies.

Despite the overall poor quality of the inpainted regions, we noted some consistency throughout the slices, which was an improvement compared to our 2D experiments. This consistency indicated that the model was starting to learn some aspects of the 3D structure. However, the generated content within the masked regions remained anatomically inaccurate, and the integration with the existing CT scan was not smooth.

Our experiments with generating full 3D CT scans from binary masks and inpainting specific organs revealed significant challenges. The generated images from masks alone were not realistic, and the inpainting results, while consistent in slices, were anatomically inaccurate.

4.4 3D GAN Generation with Vox2Vox

For comparative analysis, we also evaluated the performance of a GAN-based model, Vox2Vox, for the task of 3D CT scan inpainting. The Vox2Vox model also did not show any promising results, struggling significantly to generate 3D scans. The main issues were the higher level of noise, blurriness, patchiness and poor contrast in the images, making it difficult to distinguish if some structure details were present, which was similar to Pix2Pix results on 2D slices. Additionally, the model often failed to accurately replicate the identity function, resulting in noisier images. These issues also highlighted that to work with 3D for this task we need to make changes in model architecture and increase dataset size. Figure 21 shows an example of listed issues.

4.5 Comparison between diffusion and GANs results

The diffusion-based models demonstrated a robust ability to capture the nuanced textures and contrasts of various organs, showing significant improvements in realism over training epochs. This was particularly evident in the synthesis of kidney and liver images, where the diffusion models replicated the intricate details that are crucial for clinical relevance. The models' performance benefitted from the incorporation of CT window settings, enhancing the visual fidelity of the generated images.

On the other hand, GANs models, while capable of generating structurally coherent images and learning identity function for the rest of CT, often struggled to match the level of detail and texture fidelity achieved by the diffusion models. The GAN-generated images, specifically for the kidneys and tumors, sometimes appeared blurrier and lacked the precise textural nuances present in real CT scans. Although Pix2Pix showed some advantage in providing lung images with proper contrast, it generally fell short of the diffusion model's performance.

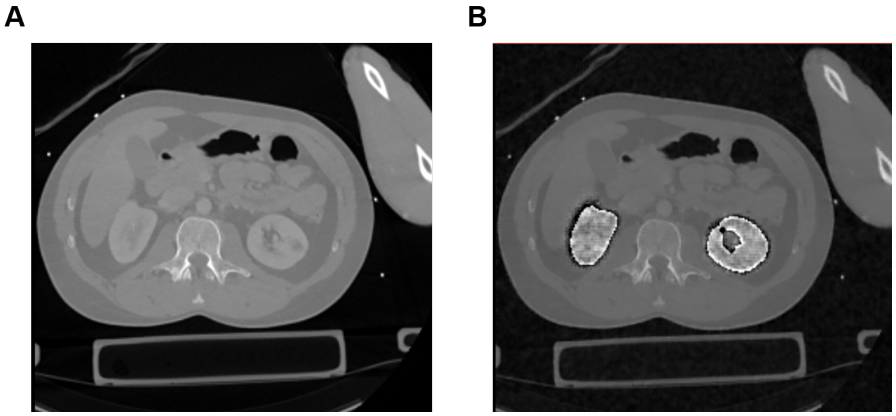

Figure 21. Slice from 3D CT scan, showing the Vox2Vox results. **A** Slice taken from original CT scan. **B** Slice taken from CT generated by Vox2Vox for kidney inpainting task. Kidneys are too bright, showing that model was not able to generate correct contrast, as well as kidney structure is not present.

4.6 FID and IS metrics

The Fréchet Inception Distance (FID) and the Inception Score (IS) are important metrics for evaluating the quality of images generated by computational models in medical imaging. FID measures the distance between the feature vectors extracted from real and generated images using an InceptionV3 model, indicating how closely the generated images resemble real ones. Lower FID scores suggest higher quality synthetic images. The Inception Score evaluates the clarity and diversity of generated images based on the predicted probability distributions of image classes by the same Inception model. Higher IS values indicate better clarity and diversity. To ensure accuracy and relevance, the FID and IS scores were calculated by matching each generated image with its corresponding real image based on specific slices from which they were derived. This matching process ensures that the evaluations are directly relevant and meaningful, providing a true measure of the model's performance in recreating or synthesizing realistic images.

While we can measure the metrics on our data, we still need to have some baseline to understand the metrics better. In our case, benchmarking against the "next slice" involves using adjacent or sequential slices from the same CT scan as a form of ground truth. This approach assumes that adjacent slices are similar enough to provide a realistic standard for comparison but distinct enough to maintain the challenge of accurate prediction or generation.

It's important to note that while FID and IS provide useful insights, they may not fully capture the nuances of medical image quality. In medical imaging, slight variations can be clinically significant, and these metrics might not fully account for such subtleties. Also, their calculation uses the InceptionV3 model, which was originally trained on the

ImageNet dataset — a large-scale dataset primarily composed of everyday objects and scenes, rather than medical images. This discrepancy can affect the applicability of these metrics for medical imaging.

Organ	**↓FID (Pix2Pix)**	**↓FID (Diffusion)**	**↓FID (Neighbouring slice)**
Kidneys	11.8897	0.4678	0.0689
Lungs	28.1489	2.4213	0.0465
Liver	19.7494	0.8220	0.0653
Tumor	15.5298	1.0933	1.2491

Table 1. FID scores for different organs generated by the diffusion model, Pix2Pix model, and benchmarking against the next slice.

Organ	**↑IS (Diffusion)**	**↑IS (Pix2Pix)**	**↑IS (Neighbouring slice)**
Kidneys	1.243 ± 0.0100	1.219 ± 0.0041	1.219 ± 0.0039
Lungs	1.265 ± 0.0025	1.257 ± 0.0016	1.258 ± 0.0014
Liver	1.249 ± 0.0032	1.245 ± 0.0024	1.246 ± 0.0022
Tumor	1.182 ± 0.0076	1.185 ± 0.0041	1.171 ± 0.0048

Table 2. Inception Scores (IS) for different organs generated by the diffusion model, Pix2Pix model, and benchmarking against the next slice.

The results from FID and IS metrics demonstrate varying degrees of success in synthetic image generation:

- **Kidneys:** The diffusion model achieved the lowest FID score (0.4678) compared to other tissues and a relatively high IS, indicating high-quality synthetic image generation with good clarity and diversity, effectively replicating kidney anatomy. The Pix2Pix model, achieved much higher FID (11.8897), not producing reasonably good images as it was less effective at capturing the fine details of kidney anatomy.

- **Lungs:** The diffusion model recorded a high FID score (2.4213) but also a high IS (1.265 ± 0.0025), reflecting challenges in capturing the intricate vascular structures and heterogeneous texture of lung tissue while still generating diverse images. The Pix2Pix model's higher FID (28.1489) indicates more significant challenge for the GAN model, though the IS remains around the same value. Benchmarking against the neighboring slice provides a very low FID (0.0465) and a comparable IS (1.258 ± 0.0014), emphasizing the difficulty of lung image generation when we compare this benchmarking value to the generated images.

- **Liver:** The diffusion model demonstrated moderate success with both FID (0.8220) and IS (1.249 ± 0.0032), suggesting good quality in synthetic images with room for improvement in both realism and textural diversity. The Pix2Pix model, with a higher FID (19.7494), shows that model struggled with finer textures and details.

- **Tumor:** The diffusion model exhibited reasonable FID and IS scores, suggesting a moderate replication of tumor textures with reasonable diversity, despite the complexity and variability in tumor appearance. The Pix2Pix model shows a high FID, indicating challenges in realism but reasonable diversity. Benchmarking against the neighboring slice results in a higher FID (1.2491), which is different from other tissues, potentially because tumors are less consistent in structure between each other and have higher variability in general.

4.7 Downstream task with synthetic data

The evaluation of nnUNet's segmentation performance on different training configurations - pure real data, and a mix of real and synthetic data - provided some insights into the efficacy of synthetic data in medical image segmentation. The evaluation was performed on the same real data for both training configurations. The Dice coefficient for the model trained on real data was 0.9608, while the model trained on a mix of real and synthetic data achieved a Dice score of 0.9683. These results were expected because the synthetic data used in this test was generated by inpainting within the organ boundary, producing data that is not significantly different from the real data. This outcome demonstrates that our approach can work, even though it does not yet leverage the full potential of synthetic data.

Ideally, downstream tasks would be tested on data produced by masking larger regions and not inside organ boundaries, allowing the model to generate more context around the organ, like we did with box masking experiment. Although we achieved great results in generating such data, we lack the segmentations necessary for proper evaluation. However, this technique allows us to create new tumors by placing a mask in a specific region — using a tumor mask from another scan, resizing, and deforming it. We can then inpaint this region to generate new synthetic tumors.

We could also perform classification tasks with this synthetic data, though this would still require some manual annotation verification. The TotalSegmentator dataset includes tumors and cysts, but the default dataset annotations do not provide detailed information on these structures. Consequently, the model trained on this data sometimes produces random tumors and cysts, making it challenging to evaluate classification tasks without thorough annotations.

4.8 Ethical and Regulatory Considerations

The integration of AI in medical imaging, especially through the generation of synthetic images, raises significant ethical and regulatory considerations that must be addressed. As we advance our capabilities in creating highly realistic and clinically applicable synthetic medical images, it becomes imperative to navigate the ethical part with a commitment to patient welfare, data integrity, and transparency.

Our work benefited significantly from accessing the TotalSegmentator dataset, which provided a comprehensive and balanced collection of medical images. This diversity in data is crucial to avoid biases in the synthetic images generated, ensuring they are representative and applicable across varied medical scenarios.

5 Conclusion

In this thesis, we have explored the use of diffusion models as a means for synthetic data generation in the field of medical imaging. The primary focus has been on creating synthetic CT scans using diffusion models, which have the potential to address data limitations in training medical AI systems by producing diverse and high-quality datasets, thereby enhancing the development and implementation of AI tools in medical diagnostics.

In our comparative analysis, diffusion models consistently outperformed GANs, the closest competitors of diffusion models, especially in the domain of 2D image generation, producing more realistic and detailed images of various organs. For kidneys, liver, and other organs, as well as tumors, the diffusion-based approach yielded synthetic images that closely resembled actual CT scans, highlighting the potential of these models in enhancing medical imaging analysis.

However, the transition from 2D to 3D image generation introduced complexities that were not fully overcome, as evidenced by the less successful attempts at creating 3D synthetic data. This outcome, while partially anticipated due to the limited availability of 3D data and the inherent challenges of 3D image synthesis, motivates the need for further research and optimization in this area.

5.1 Future Work

As we look to the future, several avenues appear particularly promising. Future studies should aim to refine the computational efficiency of diffusion models, making them more accessible and faster for broader use in the medical field. There is also an opportunity to extend the application of diffusion models to a wider range of medical imaging modalities such as MRI, PET, and others. Further investigation is needed to determine if and how synthetic data generated through the approach presented in this thesis can improve the performance of downstream tasks, such as classification of CT scans or segmentation of various structures within CT scans. In future plans, we aim to test these downstream tasks using data produced by masking larger regions, allowing the model to generate more context around the organ. Although we achieved great results in generating such data, we currently lack the segmentations necessary for proper evaluation. Additionally, we plan to explore classification tasks with synthetic data, by producing negative or positive cases with models that, for example, creates tumors from a mask or occludes them by generating larger healthy regions. We have extensive plans for testing various downstream tasks, highlighting the potential and versatility of our synthetic data generation approach.

A key area for future work is the development of 3D image generation or 3D approximations, such as generating multiple 2D slices conditionally based on previous slices. This approach could help bridge the gap between 2D and 3D data synthesis, providing

more contextually accurate synthetic datasets.

References

[1] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[2] Agisilaos Chartsias, Thomas Joyce, Rohan Dharmakumar, and Sotirios A Tsaftaris. Adversarial image synthesis for unpaired multi-modal cardiac data. In *Simulation and Synthesis in Medical Imaging: Second International Workshop, SASHIMI 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 10, 2017, Proceedings 2*, pages 3–13. Springer, 2017.

[3] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.

[4] Marco Domenico Cirillo, David Abramian, and Anders Eklund. Vox2vox: 3d-gan for brain tumour segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6*, pages 274–284. Springer, 2021.

[5] John Crank. *The mathematics of diffusion*. Oxford university press, 1979.

[6] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[8] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d medical image synthesis. *arXiv preprint arXiv:2305.18453*, 2023.

[9] August DuMont Schütte, Jürgen Hetzel, Sergios Gatidis, Tobias Hepp, Benedikt Dietz, Stefan Bauer, and Patrick Schwab. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ digital medicine*, 4(1):141, 2021.

[10] Mauro Giuffrè and Dennis L Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digital Medicine*, 6(1):186, 2023.

[11] Aldren Gonzales, Guruprabha Guruswamy, and Scott R Smith. Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1):e0000082, 2023.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[13] Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, Yoel Shoshan, Flora Gilboa-Solomon, Yasmeen George, Xi Yang, Jianpeng Zhang, Jing Zhang, Yong Xia, Mengran Wu, Zhiyang Liu, Ed Walczak, Sean McSweeney, Ranveer Vasdev, Chris Hornung, Rafat Solaiman, Jamee Schoephoerster, Bailey Abernathy, David Wu, Safa Abdulkadir, Ben Byun, Justice Spriggs, Griffin Struyk, Alexandra Austin, Ben Simpson, Michael Hagstrom, Sierra Virnig, John French, Nitin Venkatesh, Sarah Chan, Keenan Moore, Anna Jacobsen, Susan Austin, Mark Austin, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct, 2023.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[17] Bardia Khosravi, Frank Li, Theo Dapamede, Pouria Rouzrokh, Cooper U Gamble, Hari M Trivedi, Cody C Wyles, Andrew B Sellergren, Saptarshi Purkayastha, Bradley J Erickson, et al. Synthetically enhanced: Unveiling synthetic data's potential in medical imaging research. *arXiv preprint arXiv:2311.09402*, 2023.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] Theodora Kokosi and Katie Harron. Synthetic data in medical research. *BMJ medicine*, 1(1), 2022.

[20] Yunchuan Li. Theory introduction and application analysis of ddpm. *Highlights in Science, Engineering and Technology*, 57:27–31, 2023.

[21] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[22] Loc X Nguyen, Pyae Sone Aung, Huy Q Le, Seong-Bae Park, and Choong Seon Hong. A new chapter for medical image generation: The stable diffusion method. In *2023 International Conference on Information Networking (ICOIN)*, pages 483–486. IEEE, 2023.

[23] Sibam Parida, Vignesh Srinivas, Bhavishya Jain, Rajesh Naik, and Neeraj Rao. Survey on diverse image inpainting using diffusion models. In *2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, pages 1–5. IEEE, 2023.

[24] Aloke Paul, Tomi Laurila, Vesa Vuorinen, Sergiy V Divinski, Aloke Paul, Tomi Laurila, Vesa Vuorinen, and Sergiy V Divinski. Fick's laws of diffusion. *Thermodynamics, diffusion and the kirkendall effect in solids*, pages 115–139, 2014.

[25] Amit Ranjan, Debanshu Lalwani, and Rajiv Misra. Gan for synthesizing ct from t2-weighted mri data towards mr-guided radiation treatment. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 35(3):449–457, 2022.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer International Publishing, 2015.

[28] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[29] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[32] Vajira Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven A Hicks, Hugo L Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and Michael A Riegler. Singan-seg: Synthetic training data generation for medical image segmentation. *PloS one*, 17(5):e0267976, 2022.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[34] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. *arXiv preprint arXiv:2208.05868*, 2022.

Appendix

I. Glossary

Conditional Generative Adversarial Network (cGAN) An extension of the Generative Adversarial Network that uses labeled data to generate images that meet certain conditions. The conditioning can be on any type of auxiliary information, such as class labels or data from other modalities, to guide the generation process.

CT Scan Computed Tomography Scan, a medical imaging technique used to visualize internal structures of the body in a non-invasive manner.

Diffusion Models In the context of this thesis, diffusion models refer to a class of generative models used for image processing tasks, such as image translation, denoising, and super-resolution. These models simulate the process of diffusing an image to a state of pure noise and then learning to reverse this process to generate or modify images.

Denoising Diffusion Probabilistic Model (DDPM) A type of generative model that gradually adds noise to an image and then learns to reverse this process. The model is trained to convert a noisy image back into a clean one, simulating the reverse of the diffusion process. DDPMs have shown success in generating high-fidelity images and are notable for their stability during training.

Generative Adversarial Network (GAN) A class of machine learning frameworks designed to generate new data samples that resemble a given dataset. GANs consist of two parts: the generator, which creates samples, and the discriminator, which evaluates them.

Image Segmentation The process of partitioning a digital image into multiple segments (sets of pixels) to simplify the representation of an image into something more meaningful and easier to analyze.

Variational Autoencoder (VAE) A type of autoencoder used for generative tasks in machine learning. VAEs are used to compress data into a latent space and then reconstruct it back into the original space.

II. Generated kidneys

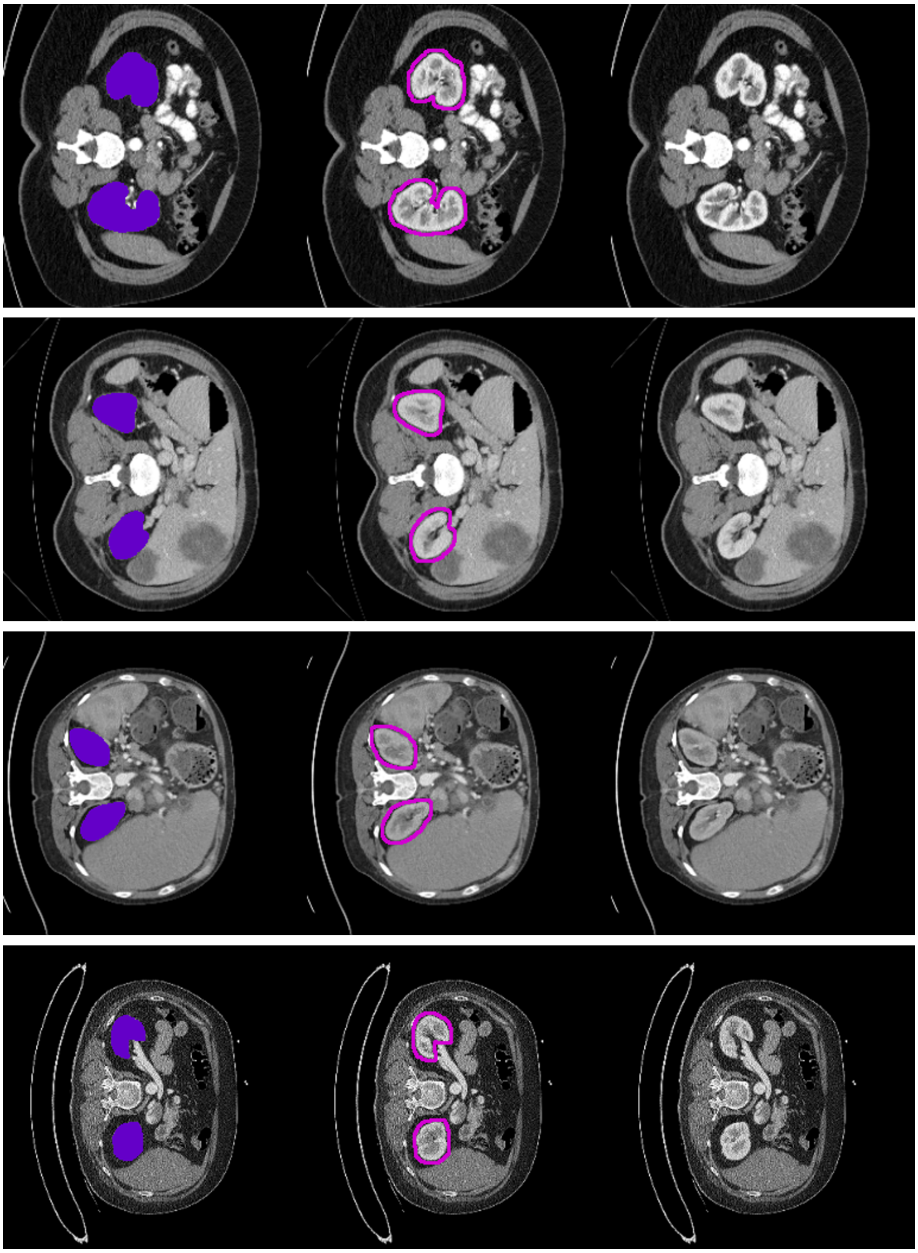

Figure 22. Representative samples of synthetic kidney generation. Column 1 depicts the masked input images with the kidney region masked out, column 2 shows the corresponding synthetic kidney generated by the diffusion model with the boundary of the inpainted region highlighted in pink, and column 3 presents the ground truth (GT) images for reference.

III. Generated lungs

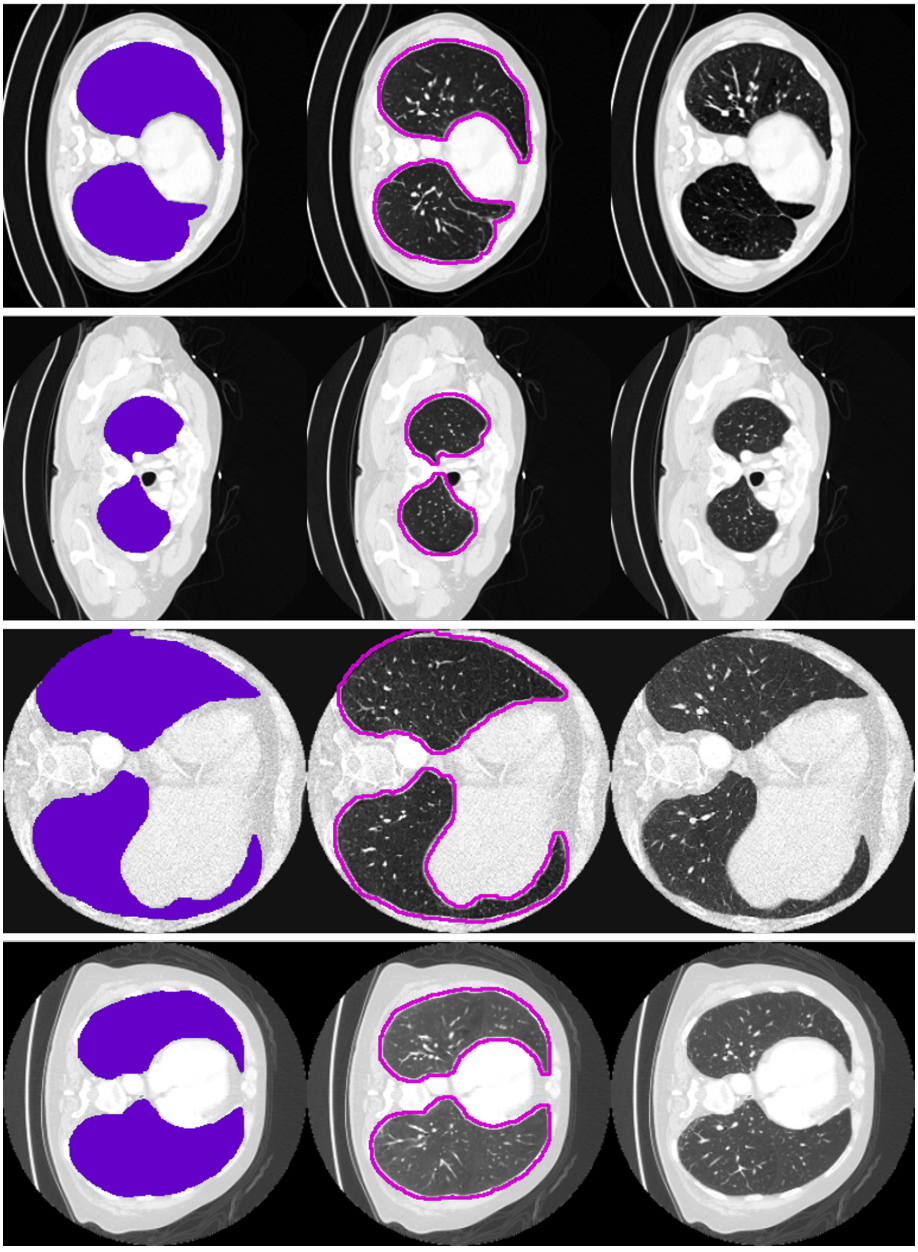

Figure 23. Illustrative examples of synthetic lung tissue generation. Column 1 illustrates the masked input images with the lung area masked out, column 2 displays the diffusion model's synthetic lung tissue output with the boundary of the inpainted area highlighted in pink, and column 3 shows the ground truth (GT) images.

IV. Generated liver

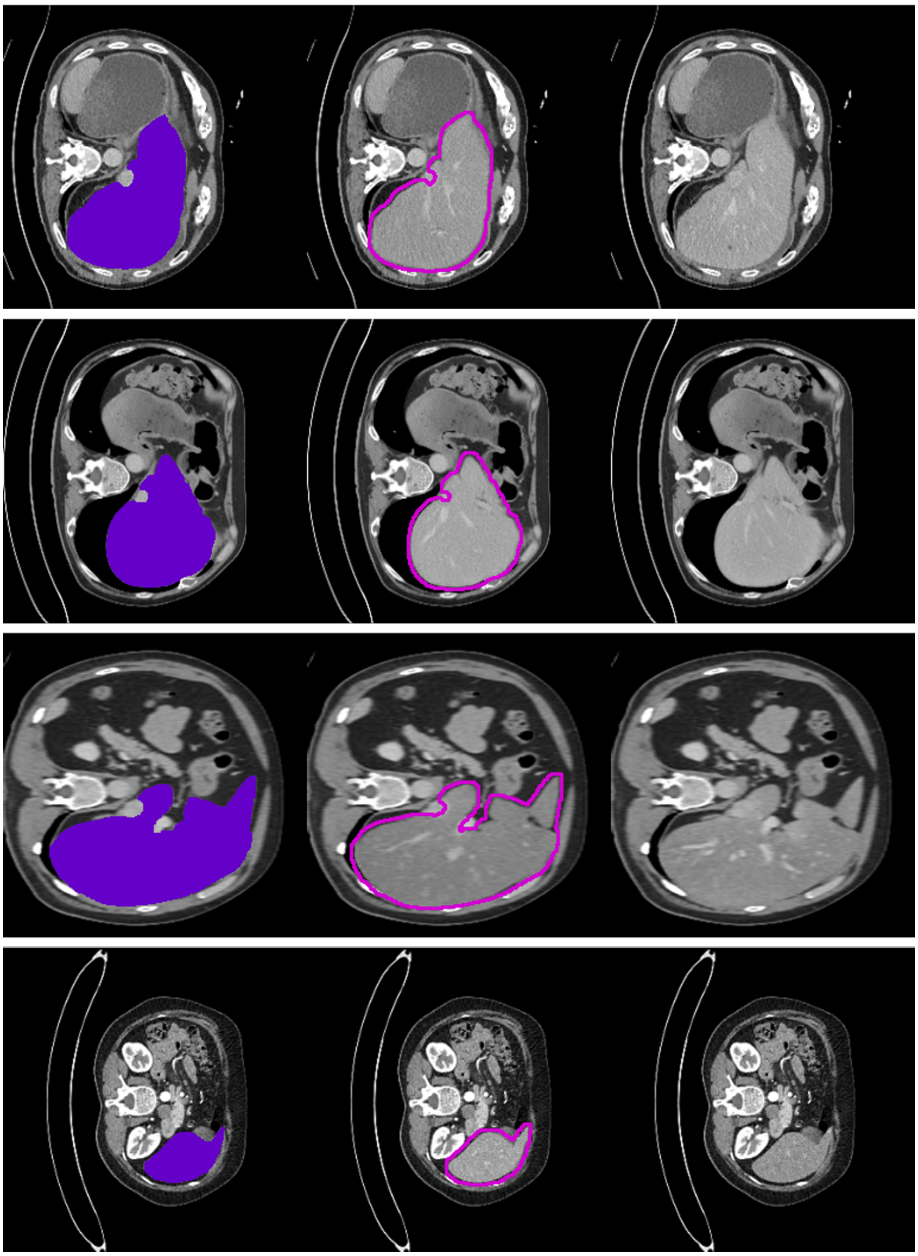

Figure 24. Selected examples of synthetic liver tissue synthesis. Column 1 presents the masked input images with the liver region masked out, column 2 reveals the generated liver tissue by the diffusion model with the contour of the synthetic area lined in pink, and column 3 offers the ground truth (GT) images.

V. Generated tumors

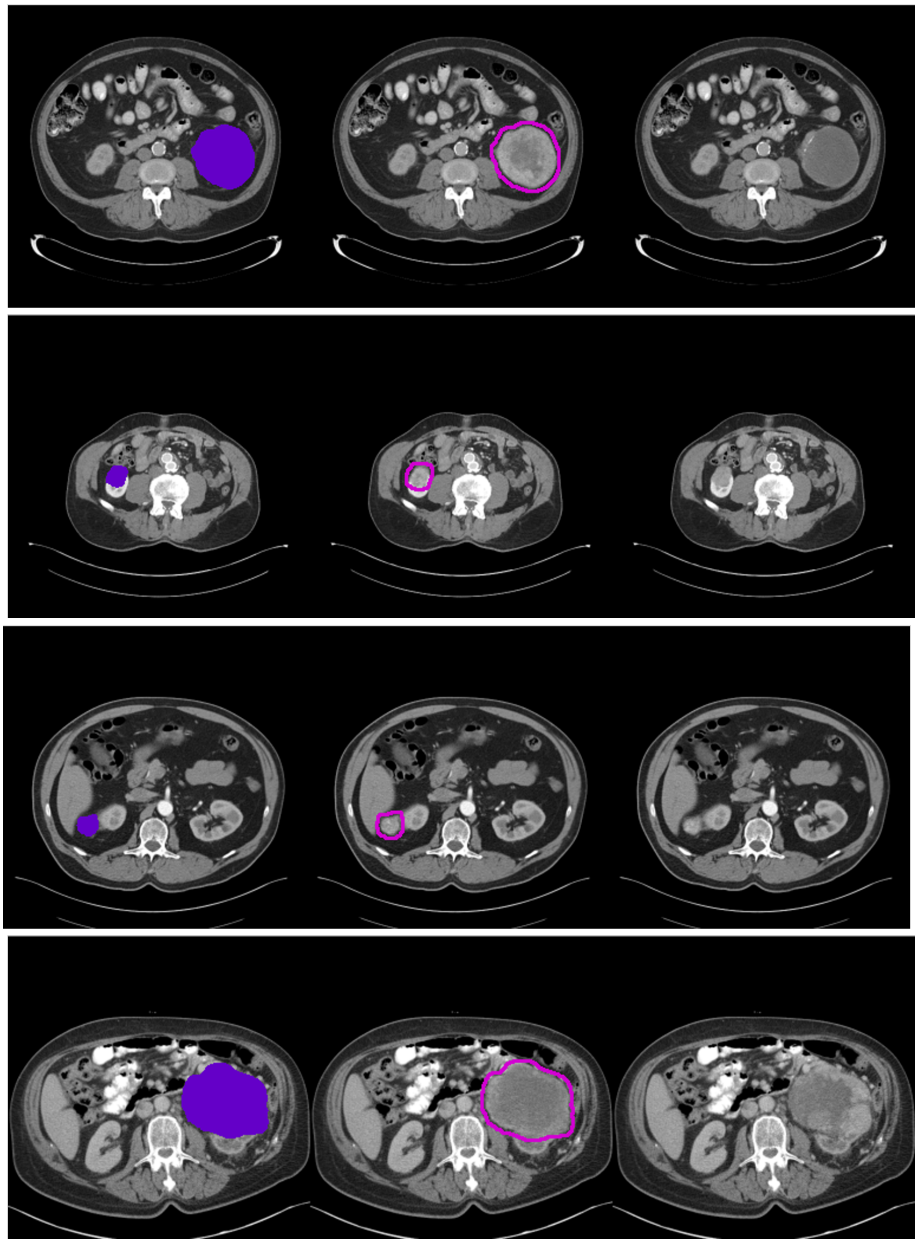

Figure 25. Selected examples of synthetic liver tissue synthesis. Column 1 presents the masked input images with the liver region masked out, column 2 shows the generated liver tissue by the diffusion model with the contour of the synthetic area lined in pink, and in the column 3 are the ground truth (GT) images.

VI. Generated tumors artificial

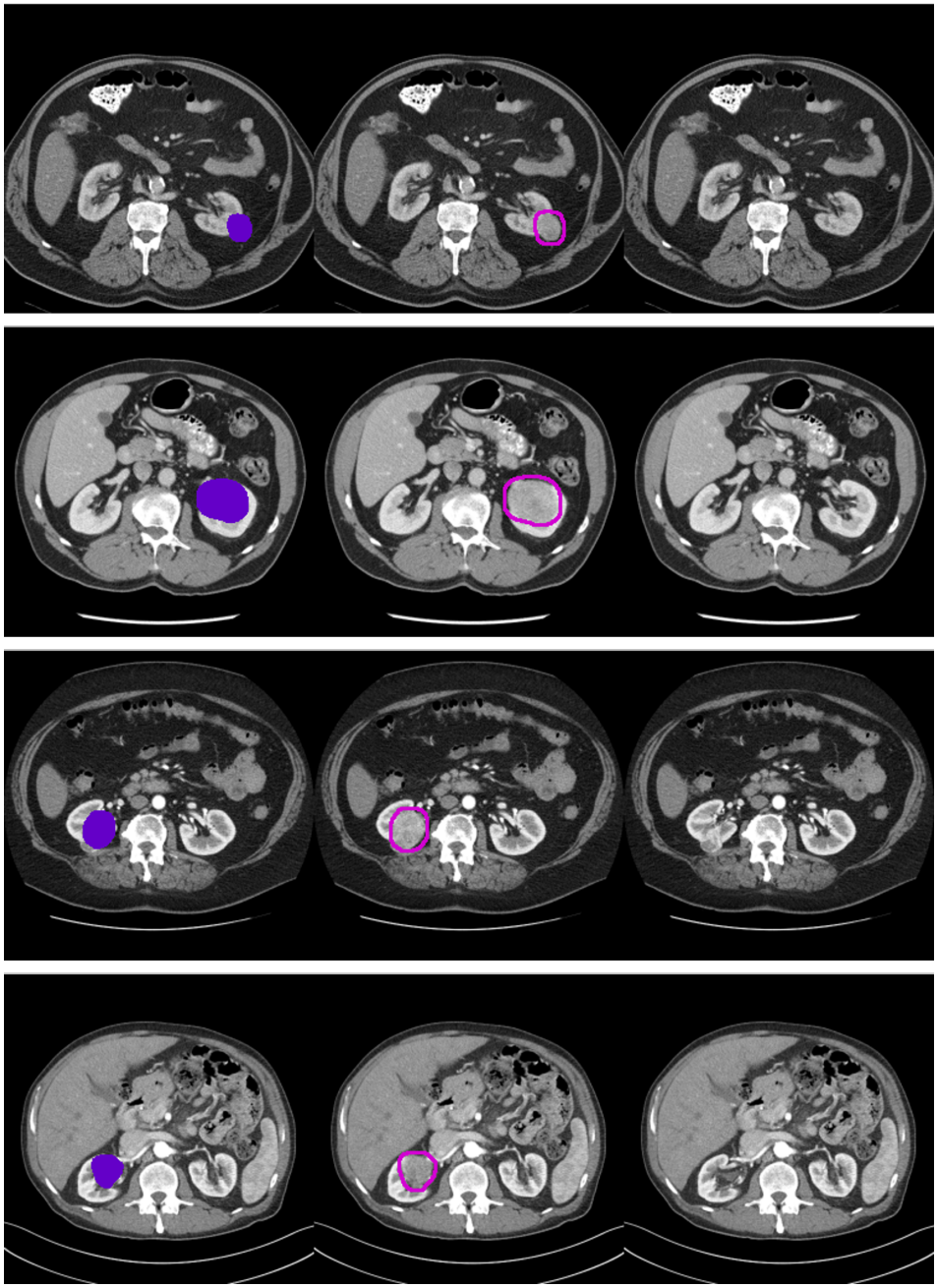

Figure 26. Selected examples of synthetic tumor generation in kidney regions. Column 1 presents the masked input images with tumor masks inserted from a different patient's CT scan, column 2 shows the generated tumors by the diffusion model, and column 3 are the ground truth (GT) images from the original scans without tumors.

VII. Generated multiorgans

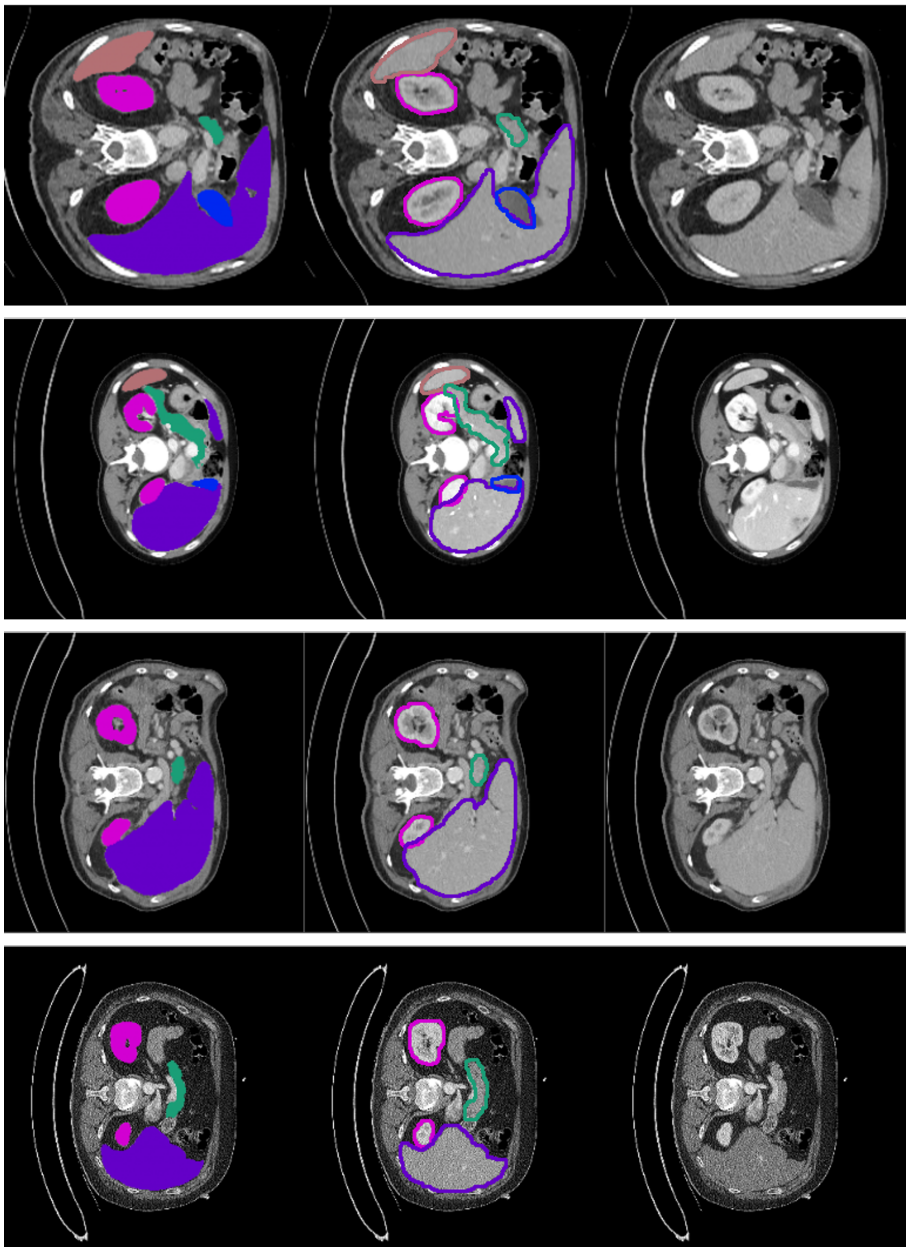

Figure 27. Selected examples of scans with generated kidneys, liver, gallbladder, spleen and pancreas. Column 1 shows the masked input images with each organ in a different color: kidneys in magenta, liver in purple, spleen in blue, gallbladder in a orange, and pancreas in green, column 2 shows the generated tissue by the diffusion model with the contour of the synthetic area lined in same colors as masks, and in the column 3 are the ground truth (GT) images.

VIII. Generated regions around kidneys

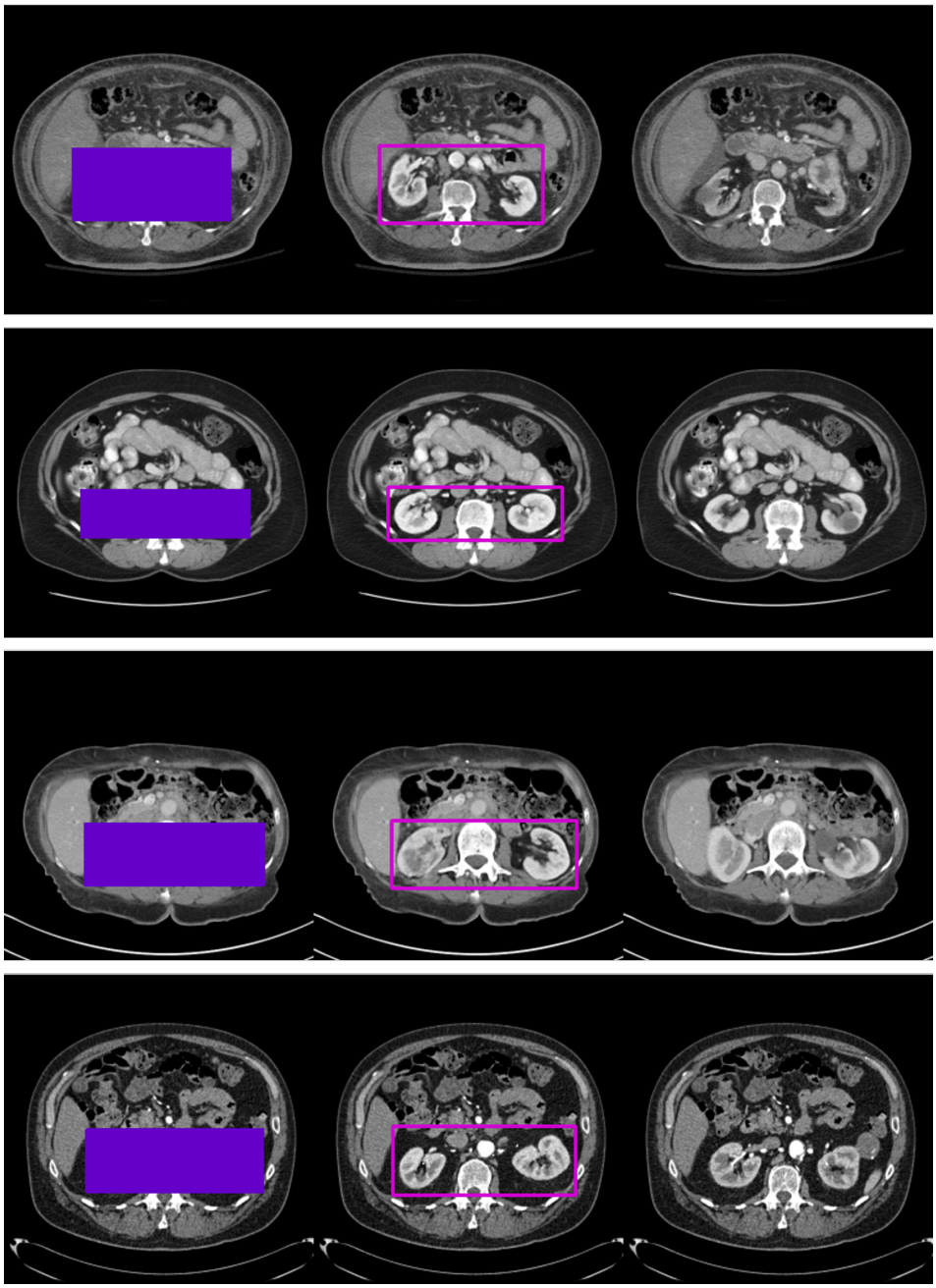

Figure 28. Selected examples of generating regions around kidneys using expanded bounding boxes. Column 1 shows the masked input images with an expanded bounding box around the kidney regions. Column 2 presents the generated images, filling in the kidney areas and surrounding tissues. Column 3 displays the ground truth images.

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Ekaterina Sedykh**,
*(*author's name)

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

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

   Imagining Infinity: Endless CT Datasets through Conditional Diffusion Models,
   *(*title of thesis)

   supervised by Dmytro Fishman.
   *(*supervisor's name)

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

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

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

Ekaterina Sedykh
15/05/2024