Universiteit Leiden
The Netherlands

Opleiding Informatica

Generating abdominal CT scan data for pancreatic cancer detection using latent diffusion

Ole Drenth s2863057

Supervisors: Prof. dr. ir Joost Batenburg, ir Șerban Vădineanu and Dr ir Meike Nauta

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

27/08/2024

Abstract

This study is driven by the critical need for high-quality medical data to improve machine learning algorithms for pancreatic cancer detection, given the severity of the disease and the challenges in early diagnosis. This thesis addresses the challenge of scarce high-quality medical data by using generative artificial intelligence to create synthetic abdominal CT scans with a latent diffusion model. We investigated which evaluation metrics could be used for assessing the quality of the synthetic samples including the Fréchet Inception Distance (FID) score, grey-scale histograms and a segmentation model named TotalSegmentator. The study concludes that while the FID score is unsuitable as a quantitative evaluation metric for 3D medical images, the grey-scale histograms provide a more reliable measure for assessing the quality of synthetic scans. Additionally, the TotalSegmentator model proves to be an effective tool for the evaluation of pancreatic representations in synthetic scans. The study evaluated the effectiveness of generating synthetic CT scans using a latent diffusion model originally designed for brain MRI scans. Results showed that the model successfully generated synthetic CT scans, with significant improvements in image quality during neural network training, though longer training yields minimal enhancements and may indicate overfitting. This research also has studied the impact of different cropping sizes on its ability to include accurate representations of the pancreas. What we found is that the role of surrounding context is crucial; while the synthetic cropped pancreas images were noisy and lacked structure, semi-cropped and full-body images provided sufficient context for accurate pancreas segmentation by the TotalSegmentator model. The grey-scale histograms revealed that more contextual information results in distributions closer to real CT scans. Thus, employing a larger cropping size provided the model with more contextual information, which improved the quality and realism of the generated pancreas in the synthetic scans. Conversely, a smaller cropping size offered less context, impacting the model's performance. By identifying that larger cropping sizes provide more contextual information and enhance the realism of synthetic scans, our findings offer valuable insights for improving the accuracy of medical image generation. This approach can be leveraged in future research to address the critical need for high-quality medical data in pancreatic cancer detection, potentially leading to more effective diagnostic tools and methods.

Contents

1 Introduction

The medical world has seen an increase in AI usage. Machine learning is revolutionizing many fields of medicine like accurate diagnosis and staging of diseases [LST+16]. This is achieved by analyzing vast amounts of data to identify patterns and anomalies that may be missed by human eyes, continuously learning and improving from new data and providing rapid, consistent and objective assessments. Especially in the field of diagnostic test in healthcare that involve pattern recognition such as pathology and radiology [Loh18]. In the case of pancreatic cancer, after detecting a tumor, surgery often removes part of the pancreas. Despite potentially curative surgery, most patients with pancreatic cancer experience a recurrence, and only up to 25% survive for five years after the procedure [SMZJ14]. Early detection of potential regrowth of tumorous tissue after a surgical operation can significantly enhance the prognosis and quality of life of the patient. But there is no standard program for screening patients at high risk of pancreatic cancer and they often fall short in providing necessary details for early detection because the tumors are either very small or are confused with fibrous tissue [KWIT16]. Since distinguishing between fibrosis and tumorous tissue is challenging, a second screening is typically conducted after a few months to check for tissue changes indicating possible tumor regrowth. This is not ideal and thus of utmost importance to find a reliable and fast method of classifying post-operative pancreatic cancer without the patient having to wait for follow up screenings. Therefore, it is important to improve these diagnoses when screening techniques like CT scans are used. Machine learning algorithms for segmentation and classification can aid doctors, by accurately depicting the boundaries of tumors and differentiating between various types of tissues, thereby enhancing the precision of diagnoses. This allows both the data driven decision making systems and the experience of professionals to reliably identify early signs of recurrence. Despite the promising applicability in the medical field, in order for these machine learning algorithms to correctly function it is crucial to have sufficient quality training data. Yet, the confidentiality of patient information significantly contributes to data scarcity in medical research and analysis. Plus, generating medical data solely for research purposes involves substantial costs. Consequently, obtaining adequate medical data remains challenging. This thesis aims to address the scarcity of high-quality data by providing synthetic CT scans using generative artificial intelligence. Specifically, it employs a generation technique called the latent diffusion model [RBL+22], which compresses data into a lower-dimensional latent space, making it computationally more feasible and capable of generating high-quality images. However, since the architecture of the latent diffusion model was originally designed for brain MRI scans, which differ from CT scans in both imaging modality and data characteristics, two key aspects will be evaluated: its effectiveness in generating synthetic CT scans and how different cropping sizes, including those focusing on specific regions like the pancreas, affect the model's performance in generating synthetic scans that include the pancreas.

The bachelor thesis at LIACS was conduced under the guidance of Professor Joost Batenburg at LIACS, PhD student Șerban Vădineanu at LIACS and PhD supervisor Dr. Meike Nauta at Datacation. It is structured as follows: after the introduction, we discuss background knowledge in Section 2 to explain key concepts of this research together with an overview of related approaches. We then describe our methodology in Section 3, followed by the experiments and their evaluations in Section 4. After that, we have a discussion together with future work in Section 5 and conclude with a conclusion of our findings in Section 6.

2 Background knowledge

This section includes background information for this research. Firstly, we discuss pancreatic cancer, covering its prevalence, risk factors, current diagnostic and treatment approaches. After that it explains what images are and it provides an overview of the different medical imaging modalities there are. Thirdly, we will explain the basics of machine learning and deep learning, and hereafter we discuss cancer detection using machine learning. We then discuss briefly what synthetic data is, what its applications can be and how it can be evaluated. Lastly, we delve into generative AI techniques, how they generate synthetic data, explaining their underlying principles and looking at related work regarding using synthetic data in medical imaging.

2.1 Pancreatic cancer

Medical imaging modalities, such as MRI, ultrasound, PET, and CT scans, play a critical role in diagnosing and monitoring various diseases, including pancreatic cancer. The pancreas is a crucial organ of the body as it functions both as an exocrine gland, producing digestive enzymes and an endocrine gland, regulating blood sugar levels by releasing hormones such as insulin and glucagon [LL10]. However, this crucial organ, is susceptible to a highly lethal form of cancer. Pancreatic cancer remains an incurable condition and is the leading cause of deaths of cancer [PON+20]. This is partly because most patients remain asymptomatic until the disease reaches an advanced stage [KWIT16]. On top of that, even after potentially curative surgery, most patients experience a recurrence of pancreatic cancer and only up to 25% survive for five years post-surgery [SMZJ14].

The imaging technique used for screening the pancreas is the CT scan because it provides detailed cross-sectional images of soft tissue that help detect abnormalities. What often happens after the detection of a tumor in the pancreas is a surgical operation that removes part of the pancreas. Like every form of cancer, it is crucial to find out whether the tumor has recurred after the surgery. In the case of pancreatic cancer it is often hard for doctors to spot the difference between fibrosis and tumorous tissue. Usually, an additional screening is done after a period of a few months to see if the tissue has changed, indicating tumor regrowth or not. This is challenge that advanced machine learning techniques and generative AI can address by creating synthetic data for more accurate and efficient post-operative screening. The need for multiple screenings makes MRI less ideal due to its high cost, while PET scans are also costly and involve repeated exposure to radioactivity from the necessary injections.

This challenge highlights the importance of choosing the appropriate imaging modality for effective post-operative monitoring, further emphasizing the critical role that various medical imaging techniques play in providing accurate diagnoses and guiding treatment decisions.

2.2 Imaging modalities

Medical imaging plays a crucial role in modern healthcare, enabling clinicians to visualize internal anatomical structures and physiological processes non-invasively. A variety of imaging modalities are used, each offering unique advantages in terms of resolution, contrast, and the specific clinical

questions they address.

Before discussing these imaging modalities, let us first discuss what an image is. An image in mathematical terms is essentially a function that maps coordinates to color values. In a more formal description, an image $\mathcal{I}$ can be viewed as a two-dimensional matrix where each element in the matrix represents a pixel. For greyscale images, each pixel's value is a single scalar representing the intensity of light, while for color images, each pixel is typically represented by a vector of values corresponding to different color channels, often known as RGB channels. The dimensions of an image are defined by its width and height. If an image $\mathcal{I}$ has dimensions $\mathcal{W} \times \mathcal{H}$, this means the image has $\mathcal{W}$ pixels in the horizontal direction and $\mathcal{H}$ pixels in the vertical direction. The dimension $\mathcal{W}$ corresponds to the number of columns, and $\mathcal{H}$ corresponds to the number of rows in the pixel matrix. For 3D images, this means the dimensions are $\mathcal{W} \times \mathcal{H} \times \mathcal{D}$, where $\mathcal{D}$ corresponds to the number of slices or layers. For instance, a 3D medical image like a CT or MRI scan can be thought of as a stack of 2D slices. Each slice has a width $\mathcal{W}$ and height $\mathcal{H}$, and there are $\mathcal{D}$ such slices in the stack.

Resolution in the context of 3D images refers to the amount of detail present across all three dimensions. Higher resolution means more detail in the width, height, and depth, which corresponds to having more pixels in each of these dimensions. Resizing a 3D image involves changing its dimensions $\mathcal{W}$, $\mathcal{H}$ and $\mathcal{D}$. This process adjusts the number of pixels in the horizontal, vertical and depth directions. For example, resizing a 3D image from $240 \times 240 \times 155$ to $128 \times 128 \times 100$ reduces the number of pixels in all three dimensions, thereby decreasing the resolution. Conversely, resizing it to $512 \times 512 \times 200$ increases the number of pixels, but this increase may not necessarily enhance the true detail or clarity. It uses a technique called bilinear interpolation, this is a method used to resize images by calculating the pixel values at new positions based on the values of nearby pixels in the original image. Understanding these fundamental concepts provides insight into the complexity and capabilities of different imaging modalities.

In medical imaging, file formats like DICOM and NIfTI play crucial roles in how images are stored, shared, and analyzed. DICOM is the standard format for storing and sharing medical images, used widely in clinical settings due to its comprehensive metadata. However, its complexity makes it less suitable for research environments. NIfTI is a simpler, more specialized format favored in neuroimaging research. It retains only essential image data and minimal metadata, allowing for easier and faster processing. NIfTI's simplicity and compatibility with various research tools make it ideal for scientific use, especially in fields like neuroimaging [LMA+16].

Magnetic Resonance Imaging (MRI), utilizes the relaxation properties of magnetically-excited hydrogen nuclei in water molecules for imaging. Powerful magnets polarize these nuclei, producing a detectable signal that is spatially encoded. MRI uses non-ionizing radiation, making it safer for patients. Radio frequency energy elevates hydrogen nuclei to higher energy states within a magnetic field, and as molecules return to their normal state, they emit energy, a process known as relaxation. Variations in relaxation rates between tissues are used to generate images. MRI requires strong, uniform magnetic fields and produces 2D slices of the body and can create 3D images with modern instruments, making it a powerful medical imaging technique[EJ16]. MRI is often used for detailed imaging of soft tissues in the brain, spinal cord, joints, and internal organs, making it essential for diagnosing and monitoring conditions such as tumors, injuries, and neurological disorders. It is also

preferred for cardiac and vascular imaging, as well as in pediatric patients, due to its non-ionizing radiation.

Medical ultrasound, also known as diagnostic sonography or ultrasonography, is a technique used to visualize internal body structures through the application of ultrasound. Ultrasound refers to sound waves with frequencies higher than the audible range. Sonograms are created by sending ultrasound pulses into tissues using probes. Different sonographic instruments can produce various types of images, with the most common being the B-mode image, which displays a 2-D cross-section based on the tissue's acoustic impedance. Ultrasound offers advantages such as real-time imaging, portability, lower cost and the absence of harmful ionizing radiation, making it widely used for both diagnostic and therapeutic procedures[EJ16].

A positron emission tomography (PET) scan is an imaging test that produces images of organs and tissues in action. It involves injecting a safe radioactive chemical called a radio tracer and using a PET scanner to detect it. The scanner identifies diseased cells that absorb high amounts of the radiotracer, indicating potential health issues [Mos01].

Computed Tomography (CT) scan is a radiological imaging technique, it is the primary focus of this thesis due to their critical role in providing detailed cross-sectional images that are essential for accurate diagnosis and treatment planning, particularly in detecting and monitoring pancreatic abnormalities. A CT scan is essentially an X-ray study where a series of rays rotate around a specific body part to produce computer-generated cross-sectional images. These tomographic images offer a significant advantage over conventional X-rays by providing detailed cross-sectional information of a specific area, eliminating image superimposition. This results in superior diagnostic accuracy, making CT scans highly effective for correlating clinical and pathological findings in suspected illnesses [PDJ21]. One disadvantage of CT scans is that they expose patients to higher doses of ionizing radiation compared to conventional X-rays. The Hounsfield unit (HU) is a relative quantitative measure of radiodensity used by radiologists to interpret CT images. It is based on the absorption coefficient of radiation within tissues, which is utilized during CT reconstruction to create a grayscale image. Distilled water at standard temperature and pressure is defined as 0 HU and air as -1000 HU. The scale can reach up to 1000 HU for bones, 2000 HU for dense bones and over 3000 HU for metals like steel or silver. The Hounsfield scale displays dense tissues with positive values as bright and less dense tissues with negative values as dark [DS19]. Given these advantages, CT scans remain the most relevant imaging modality for the development and evaluation of synthetic data in this research, underlining their importance in enhancing diagnostic capabilities and improving patient outcomes.

2.3 Machine learning

Having talked about the different medical imaging modalities and the challenges in pancreatic cancer diagnosis and monitoring, it is evident that machine learning techniques can play a crucial role in enhancing the accuracy and efficiency of detecting and predicting cancer recurrence [LST+16]. One of the most powerful tools in machine learning is the artificial neural network, a mathematical and computational model made up of numerous neurons, designed to simulate the structural and functional characteristics of biological neural networks. It is a self-adaptive system that alters its

internal structure based on external input, and is frequently used to model complex relationships between inputs and outputs.

To leverage these neural networks effectively, different learning approaches are employed. In supervised learning, the model is trained on a labeled dataset, meaning that each training example is paired with an output label. The goal is for the model to learn the mapping from inputs to outputs so that it can accurately predict the label of new, unseen data. In contrast, unsupervised learning deals with unlabeled data. The objective is to identify inherent patterns or structures in the input data [AAJM+20].

Conventional machine-learning techniques required significant engineering and domain expertise to design feature extractors that transformed raw data into suitable internal representations for pattern detection or classification. Representation learning methods, including deep learning, automate this process by discovering necessary representations from raw data. Deep learning, a subset of machine learning, involves neural networks with many layers and parameters, often referred to as deep neural networks. It uses multiple layers of nonlinear processing units for feature extraction and transformation, with lower layers learning simple features and higher layers learning complex features derived from these simpler ones [SS18]. Thus, deep learning uses multiple levels of non-linear modules to transform raw input into increasingly abstract representations, enabling the learning of complex functions. As a result of this, deep learning excels at finding intricate structures in high-dimensional data, making it applicable across various fields such as science, business and medicine [LBH15].

Backpropagation is a fundamental algorithm for training neural networks. It works by propagating the error from the output layer back through the network, adjusting the weights of the connections in order to minimize the error. This process involves two steps: We start with the forward pass, the input data is passed through the network to obtain the output. After that we have the backwards pass, the error is calculated based on the difference between the predicted output and the actual target. This error is then propagated backwards through the network, and the weights are updated using gradient descent to reduce the error in future predictions [Wer90].

Convolutional neural networks (CNNs) are a dominant class of deep learning methods in computer vision. They consist of multiple building blocks: convolution layers, pooling layers and fully connected layers. These models are designed to automatically learn spatial hierarchies of features through backpropagation. A convolution layer, a fundamental component of CNN architecture, performs feature extraction through a combination of linear and nonlinear operations, including the convolution operation and an activation function. [YNDT18]. Compared to fully connected networks, which means each neuron in a layer is connected to every neuron in the preceding layer, CNNs offer several advantages: they use local connections, which connect each neuron to only a small subset of neurons from the previous layer, thus reducing parameters and speeding up convergence; weight sharing, where groups of connections share the same weights to further decrease parameter count; and downsampling through pooling layers, which reduce data size while preserving important information and minimizing trivial features [LLY+21].

A fully convolutional network, as proposed by [LSD15], is designed to take input of arbitrary size

and produce correspondingly-sized output with efficient inference and learning. Ronneberger et al.[RFB15] enhanced this architecture by augmenting the standard contracting network, a series of convolutional and pooling layers that progressively reduce the spatial dimensions of the input image while increasing the depth of feature maps. This architecture was augmented with upsampling layers instead of pooling operators, thereby increasing the resolution of the output. To achieve localization, high-resolution features from the contracting path are merged with the upsampled output. In doing so, a convolutional layer can learn to generate a more precise output using this combined information. The model consist of a encoder-decoder model with skip connections, these connections enable the network to capture both the context and precise localizations of features. One significant modification in the proposed architecture by Ronneberger et al. is the large number of feature channels in the upsampling part. This enables the network to convey contextual information to higher-resolution layers. As a consequence the expansive path becomes nearly symmetric to the contracting path, forming a U-shaped architecture, as seen in Figure 1 [RFB15]. Additionally, the network avoids fully connected layers and only relies on the convolution of each layer.

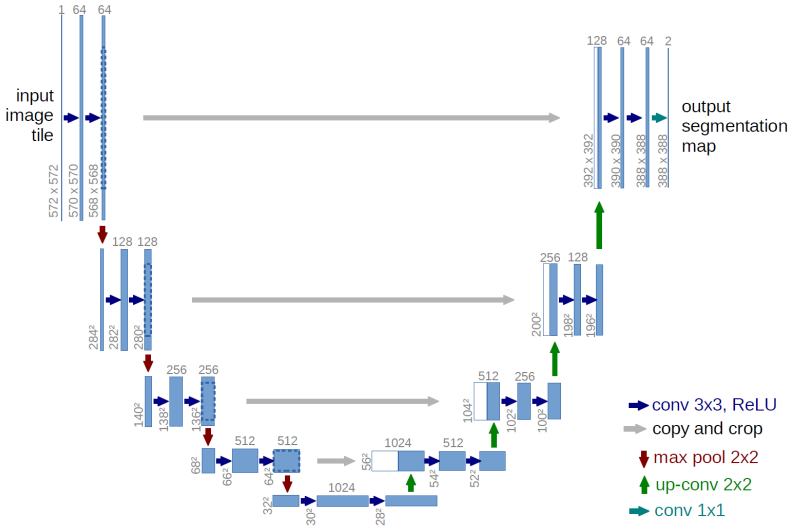


Figure 1: U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. Figure from Ronnenberger et al. [RFB15].

Given the potential of deep learning and other machine learning techniques to enhance the detection and prediction of complex patterns, they hold significant promise in improving cancer detection and monitoring, including for pancreatic cancer. CNNs, with their ability to efficiently learn spatial hierarchies and reduce dimensionality, are particularly effective for this purpose, offering advanced capabilities for accurate and efficient analysis of medical imaging data. Building on this foundation, the next section explores specific applications of machine learning in cancer detection, focusing on advancements in classification, segmentation, and detection as demonstrated in recent studies.

2.4 Cancer detection using machine learning

Yamashita et al. have illustrated how deep learning is applied to medical imaging for cancer detection, highlighting key advancements in classification, segmentation, and detection. For instance, deep learning models, especially CNNs, are frequently used to classify lesions in medical images, such as distinguishing between benign and malignant lung nodules in CT scans. These models require extensive labeled datasets and often involve specialized architectures for 3D imaging and time series data. Also, CNNs have been adapted to handle complex 3D structures and temporal data, improving diagnostic accuracy. Transfer learning and fine-tuning further enhance these models by leveraging pre-trained networks, which is crucial for effectively applying machine learning to cancer detection. This reflects the ongoing innovation and practical utility of machine learning techniques in advancing cancer diagnosis and monitoring [YNDT18].

Another study by Aytar et al. showcased the usage of synthetic data in enhancing machine learning models' performance by providing additional training data, thereby addressing the limitations of real data and improving the robustness and accuracy of the models in various applications [AG24]. In this study, Multi-Scale Gradients for Generative Adversarial Networks (MSG-GAN) was utilized to generate synthetic breast histopathology images, including both malignant and negatively labeled patches, to assist in cancer identification. The synthetic images were classified using the ResNet18 model through transfer learning, with the goal of comparing their performance to that of real data. The research demonstrates that synthetic data can closely mimic real data, with successful outcomes such as generating realistic synthetic images and potentially eliminating the need for manual visual breast cancer detection. The study concludes that the use of GANs in healthcare, for purposes such as data augmentation, dataset enrichment, and balancing, is highly beneficial and reliable, predicting a reduction in human intervention in disease detection and diagnosis and an acceleration of healthcare research [AG24].

2.5 Synthetic data

The field of machine learning has developed at an increasing rate in the last decade, it empowers intelligent computer systems to autonomously handle tasks, driving forward industrial innovation [Ng16]. These models are highly dependent on data. Despite this, real world data is often noisy and incomplete. Ensuring data quality is important when training ML models. If the models do not have quality training data, it may reproduce inaccurate results due to confusion and misinterpretation [PLW02]. Another problem is data scarcity, a significant part of the current AI challenge arises from a lack of adequate data: either there are too few available datasets, or manual labeling is excessively expensive or time consuming [BS19]. Furthermore, the inability to share datasets publicly due to confidentiality exacerbates this problem.

To summarize, real-world data presents various challenges, which is where synthetic data becomes significant. Synthetic data is artificially generated data that is annotated using algorithms or models that simulate the statistical properties and patterns found in real-world data [LSW+23]. It mimics-world data but is not obtained from actual measurements. It can provide a solution for problems that arise with real-world data. Most importantly, the data scarcity problem can be solved

by increasing the existing dataset with synthetic data samples. This can be done without the costly and time consuming acquisition of real-world data. As discussed in Section 1, this approach is particularly beneficial in the medical field, where getting large quality datasets is often prohibitively expensive and subject to stringent privacy regulations. It also allows researchers to create diverse situations for experimentation and validation without relying on a single dataset.

2.5.1 Applications

The application of synthetic data in healthcare is significant, enhanced by generative AI, medical imaging can benefit from improved data completion and image segmentation, aiding in disease diagnosis and treatment planning by providing healthcare professionals with detailed and comprehensive information from images [KXS+23]. This enables more precise diagnoses and personalized treatment strategies. Furthermore, clinical decision support systems utilized by generative AI can assist healthcare professionals in making better decisions and improving patient outcomes by delivering updated and accurate information [LSW+23].

Synthetic data is especially useful for doctors that utilize segmentation or classification systems. These systems need a large quantity of training data, potentially improving the overall quality of the diagnostic and treatment process. This can be achieved using synthetic data because it allows for the creation of diverse and comprehensive training datasets. It can be particularly beneficial for augmenting existing datasets, which might be limited in size or diversity. This augmentation helps address data imbalances, introduces variations that the model may not have encountered before, and improves the overall quality of the diagnostic and treatment process. Consequently, the integration of synthetic data not only expands the training set but also ensures that models are better equipped to handle real-world variations and complexities.

2.5.2 Evaluation metrics

Integrating synthetic data into medical systems can enhance the training of segmentation or classification models, ultimately improving diagnostic accuracy and treatment outcomes. To ensure the quality of such synthetic data, it is crucial to employ reliable evaluation metrics that can effectively assess the similarity between real and generated images. A method that quantitatively evaluates generated images is called the Fréchet Inception Distance (FID). FID measures the distance between the feature distributions of real images and images generated by an algorithm, as extracted by the Inception-v3 model, a convolutional neural network originally trained on 2D images [JRV+24]. Lower scores mean the two groups of images are more similar, with 0.0 being the perfect score that indicates that the two groups are identical [YZD21].

Next to that, a gray-level histogram is way to evaluate CT scans. It uses the intensity of diagonal pixels from all CT abdominal images of a complete scan. Figure 3 shows the resulting histogram, which displays three distinct peaks, a common characteristic in all CT abdominal scans. These peaks represent different regions, from left to right: the first peak corresponds to the intensity of black background pixels, the second peak represents low-intensity pixels from the external region surrounding the bronchial tree and the internal lung parenchyma, and the rightmost peaks correspond to high-intensity pixels representing fat, muscles, heart, bones and pulmonary nodules [CCC20].

This histogram provides a comprehensive visualization of the intensity distribution, aiding in the analysis and interpretation of the CT scans. The threshold refers to a specific intensity value used to differentiate between different regions in the CT abdominal images.

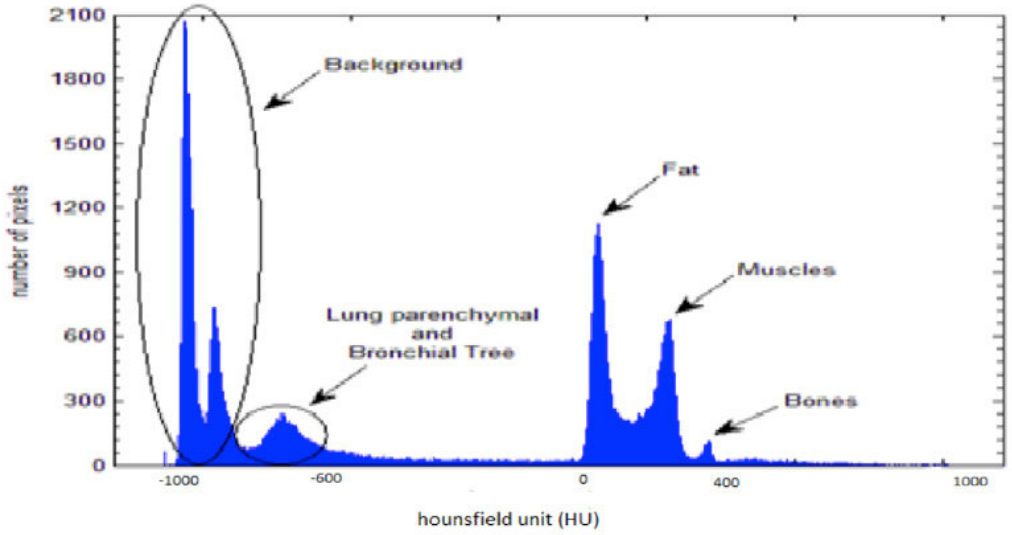


Figure 2: The histogram of Hounsfield units for the image [CCC20].

For the research of this thesis, a qualitative metric that can be useful is called TotalSegmentator [WBM+23]. TotalSegmentator, proposed by Wasserthal et al., is a deep learning segmentation model that can automatically segment 104 major anatomical structures in body CT images, including 27 organs, 59 bones, 10 muscles, and 8 vessels. Trained on a dataset of 1204 CT examinations using the nnU-Net algorithm [IJK+21], it achieved a DICE similarity coefficient of 0.943 on a diverse test set. The DICE score is a metric used to evaluate the similarity between two sets, often applied in tasks like image segmentation. It ranges from 0 to 1, where a score of 1 indicates perfect overlap between the sets and a score of 0 indicates no overlap [ZSZZ17]. This model can be utilized to evaluate the generated scans by attempting to segment the pancreas, thereby allowing us to determine the possibility of the model accurately generating realistic pancreatic structures. The TotalSegmentator model can be evaluated by calculating the DICE score between the actual segmentations provided by the MSD dataset [ARB+22] and the segmentations produced by TotalSegmentator, assessing the accuracy and reliability of the segmented pancreatic structures.

2.6 Generative AI techniques

In the introduction, we emphasized the urgent need for faster and more accurate screening methods for post-operative pancreatic cancer patients. This need can be addressed through advanced machine learning techniques, which require substantial amounts of high-quality data. Since obtaining such data can be expensive and time-consuming, leveraging generative AI to create synthetic data presents a promising solution. Generative AI encompasses a range of artificial intelligence techniques and models aimed at learning the underlying patterns and structures of a dataset to generate new data

points that could realistically belong to the original dataset. During training, a generative model seeks to estimate the data's probability distribution. Given that our data points originate from an unknown underlying distribution $x \sim p_{data}(x)$, we employ a model $p_\theta(x)$, which represents a family of parameterized probability distributions, to estimate p_{data} [PGK+23]. This newly created data can be in the form of text, audio, images or video. Here some prominent generative AI techniques will be discussed that are commonly used.

2.6.1 Data augmentation

Data augmentation is a technique used in machine learning to artificially increase the diversity of training data by applying various transformations like rotations, translations and flips. It can be seen as a set of techniques designed to modify existing real data rather than create entirely new synthetic data. In other words, it is data with minor alterations that the model's predictions should remain unaffected by. The main use case of data augmentation is preventing overfitting [SKF21], this happens when a model learns the details and noise in the training data to the extent that it negatively impacts the model's performance on new unseen data. Thus, data augmentation is a useful technique for improving the generalizability of a machine learning model. In the next section we will dive deeper into the benefits and applications of Generative Adversarial Networks (GANs) in generating synthetic data with enhanced realism and diversity, thereby complementing traditional data augmentation techniques.

2.6.2 GANs

A common technique for generating image data is the GAN. At its core, GANs capture the distributions of the training data making it possible the generate new samples from the learned distribution. It is a technique for semi-supervised and unsupervised learning [CWD+18]. It is a model that learns deep representations of data through a training process called adversarial learning. It accomplishes this by utilizing two neural networks, the generator and the discriminator. In Figure 3 the generator produces synthetic data samples, while the discriminator evaluates them to distinguish between real and generated samples. The discriminator can be seen as a function that maps an input image to a probability of it being real by evaluating individual samples and learning to distinguish real images from those generated by the generator.These two networks compete with each other and are trained simultaneously. The generator has no access to the real images and can only learn from the discriminator. On the other hand, the discriminator has access to both the real and generated data [CWD+18].

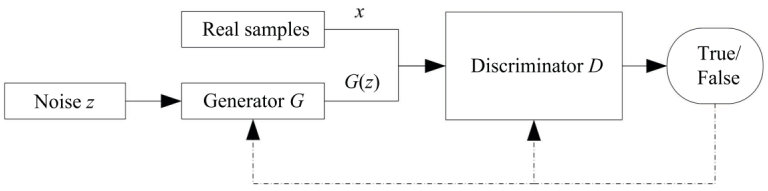


Figure 3: Computation procedure and structure of GAN. Figure from [WGD+17].

There are limitations that arise with the training of GANs [SC21]. One major issue is mode collapse,

which occurs when the generator learns to produce outputs that are too similar, leading to a lack of diversity in the samples. This essentially happens when the generator only generates a limited range of variations within the data distribution [KN+20]. Another limitation is non-convergence and instability, which can result from an improper network architecture, a poorly chosen objective function, or an unsuitable optimization algorithm. Selecting the right hyperparameters for these aspects can be challenging.

To address some of these challenges and improve the generation of medical images, a study from the University of Nebraska has developed a new GAN-based model, 3DGAUnet, specifically for generating CT images of pancreatic ductal adenocarcinoma (PDAC) tumors and pancreatic tissue [STB+23]. This approach addresses the lack of inter-slice connection data in 2D CT image synthesis models, thereby significantly improving efficiency and accuracy by maintaining contextual information between slices. The model was trained on two separate datasets, one with tumors and one with pancreas, and the synthetic data from both was blended together. Qualitative evaluation was performed using 2D cross-slices and 3D volume rendering [KPV13], while quantitative evaluation used FID values on 2D slices [HRU+17]. The 3DGAUnet model enhances the learning of shape and texture, demonstrating efficacy across diverse datasets and presenting a promising solution for data scarcity and accurate PDAC detection in medical imaging.

Another method using GANs, introduced by Pesaranghader et al. [PWH21], is the CT-SGAN model, which leverages a recurrent generative adversarial network to generate large-scale 3D synthetic CT scan volumes (size $\geq 224 \times 224 \times 224$) from a small dataset of chest CT scans. This model generates CT scan volumes incrementally by producing their constituent slices and slabs over time, addressing key challenges in medical training data. They evaluated the model using both qualitative measures, such as anatomical consistency, fidelity to real slices, and diversity and quantitative measures, including FID and IS scores. Additionally, they used 3D-SqueezeNet as a classifier for nodule detection, finding that classifiers trained with 10,000 synthetic CT scans performed better than those trained with 900 real CT scans.

These advancements highlight the potential of GAN-based models to overcome limitations in medical image generation, providing more diverse and accurate synthetic data for training purposes. However, despite their promise, GANs still face significant challenges, such as mode collapse, training instability, and the requirement for extensive computational resources. Continued research and development in this area promise to further enhance the quality and applicability of synthetic medical images in clinical practice.

2.6.3 Autoencoders

One of these alternatives is the autoencoder. An autoencoder (AE) is an unsupervised neural network model that compresses input data into a lower-dimensional representation through an encoder. It then reconstructs it back to its original form using a decoder, aiming to minimize the reconstruction error [BKG23]. The encoder component compresses and reduces the input into a lower dimension and outputs a new layer called code. This code layer gets reconstructed by the decoder to the original dimensionality of the input. It is especially useful in image data as this type of data has a high dimensionality. Autoencoders can effectively reduce the dimensionality

while preserving essential features, facilitating tasks such as compression, noise reduction, and feature extraction [BESS+24]. By introducing variational inference to an AE network, a new type of network was created, the variational autoencoder (VAE) [KMJRW14]. This is a generative model that utilizes Variational Bayes Inference, which allows the network to describe data generation using a probability distribution [KW+19]. Unlike traditional AEs, VAEs include an additional sampling layer alongside the encoder and decoder layers. Training a VAE involves encoding the input into a distribution over the latent space, from which a latent vector is sampled. This latent vector is then decoded to reconstruct the input. The reconstruction error is calculated and backpropagated through the network to update the model. Additionally, regularization is explicitly applied during training to prevent overfitting [BESS+24].
In conclusion, while AEs are effective in compressing high-dimensional data and preserving essential features, the introduction of variational inference in VAEs enhances their capabilities by allowing for probabilistic description of data generation. VAEs offer improved generative modeling and can be seen as a potential aiding tool for solving the data scarcity problem.

2.6.4 Diffusion models

Next is another very powerful generative model called the diffusion model. These models are probabilistic generative models that consist of two processes: the forward and the backward diffusion process. At their core, diffusion models operate by progressively adding Gaussian noise to the training data, effectively degrading it. The model then learns to reconstruct the original data by reversing this noising process.
A certain class of diffusion models that uses this process to generate data is called Denoising Diffusion Probabilistic Models (DDPMs) [HJA20]. In the forward diffusion process Gaussian noise is added into the input data. The reverse diffusion process, known as denoising, iteratively reverses the diffusion step by step to generate new sample data. So the process starts with a noisy input and progressively refines it to produce a high-quality sample, effectively mimicking the process of removing noise from the data [YZS+23]. Once trained, the diffusion model can generate new data by feeding randomly sampled noise through the learned denoising procedure.

Building on the foundational work of DDPMs, Dorjsembe et al. [DOX22] used a diffusion based model to create 3D MRI scans of brain tumors. The model they used is a 3D denoising diffusion probabilistic models (3D-DDPM). Ho et al. [HJA20] introduced the DDP model and with the use of the improvements and modifications of Nichol et al. [ND21], Dorjsembe et al. transformed it into the 3D DDPM model. The model was compared to two baseline models, 3D-α-WGAN [KHK19] and CCE-GAN [XSH21]. Using Multi-scale structural similarity (MS-SSIM) for quantitative evaluation they showed that the 3D-DDPM were more similar to the real data compared to the other two models. For qualitative evaluation they let experienced doctors classify the images and it also showed that the 3D-DDPM generated visually accurate images, while none of the baseline models were classified as real.

Another notable contribution in the field is the work by Kim et al. [KY22], who introduced a 4D image generation framework using a denoising diffusion probabilistic model to create intermediate temporal volumes between source and target images, incorporating 3D+time information. This model, which includes diffusion and deformation modules, learns spatial deformations and generates

intermediate frames along a continuous trajectory. Their results demonstrated effective generation of dynamic deformed images for 4D cardiac MR images, showing potential for clinical applications like analyzing anatomical changes.

Effectively by having a more clear training objective, diffusion models overcome the instability that comes with GANs, as discussed in 2.6.2 [BESS+24]. However, despite their advantages, diffusion models face notable challenges. Unlike GANs, which generate images in a single forward pass, diffusion models require multiple forward passes due to their iterative denoising process. This increased computational demand results in longer training and inference times, rendering diffusion models computationally less efficient compared to their generative counterparts [KAAL22]. It poses practical limitations for high-resolution image generation, particularly on consumer-grade GPUs. Thus, while diffusion models offer improvements in image stability and quality, their computational inefficiencies and memory requirements remain significant hurdles in their widespread adoption [PYG+23].

2.6.5 Latent diffusion model

Rombach et al. [RBL+22] have introduced a modification to the diffusion model that counters the problem that arises with vanilla diffusion models, called the Latent Diffusion Model (LDM). The model operates on the compressed latent space instead of the pixel-based space. As discussed in Section 2.6.4, diffusion models operate by progressively adding Gaussian noise to the training data, effectively degrading it. The model then learns to reconstruct the original data by reversing this diffusion process. This noising and denoising process happens in the lower dimensional latent space and so the method preserves perceptually important details while greatly reducing computational costs. This compressed image space is achieved by using an encoder-decoder architecture [RBL+22]. LDMs generate images through a process that begins with random noise inside the latent representation. This noise is iteratively refined over multiple steps. In each step, a small amount of Gaussian noise is added to the current image representation and the model is trained to predict this added noise. By subtracting the predicted noise from the current image, the model effectively denoises the image. This process is repeated numerous times, gradually transforming the initial random noise into a meaningful image representation. Once the denoising process is complete in this latent space, a decoder network is employed to map the refined latent representation back to the pixel space, producing the final generated image.

In Figure 4 you can see an overview of the latent diffusion model. The loss function of this model is:

$$L_{\text{LDM}} \coloneqq \mathbb{E}_{\mathbf{x},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t)\|_2^2\right]$$

The loss function, L_{LDM}, calculates the expected squared L2 norm difference between a noise sample ϵ and the noise predicted by the model, $\epsilon_\theta(z_t, t)$. Here, ϵ is drawn from a standard normal distribution $\mathcal{N}(0, 1)$. The goal is to train the model ϵ_θ to accurately predict the noise added to the data z_t at time t, thereby enhancing the model's ability to denoise or generate data effectively. In the denoising step, the model uses a UNet architecture seen in figure 1, which enhances a fully convolutional network with upsampling layers and skip connections to merge high-resolution features from the contracting path with the upsampled output, enabling precise localization and effective context capture [RFB15].

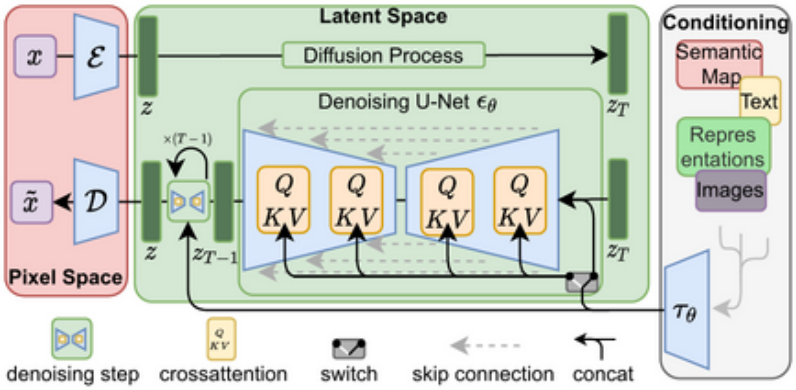


Figure 4: This schematic shows an overview of the latent diffusion approach, including encoder $\mathcal{E}$, decoder $\mathcal{D}$, and conditioning using a cross-attention mechanism. Figure from [RBL+22].

To conclude, diffusion models are a very powerful generative AI technique, particularly when integrated with autoencoders to leverage latent space effectively. They hold an advantage over GANs due to their enhanced stability, superior mode coverage and higher quality generated output.

3 Materials and methods

The research aims to assess whether a latent diffusion model originally designed for generating MRI scans can be adapted to accurately generate CT scans that contain the pancreas. Therefore, the research question is:

To what extent is it possible for an existing latent diffusion model, originally designed for MRI scans, to accurately generate the pancreas in synthetic CT scans?

We will address this research question through three specific sub-questions:

1. In Section 4.1, we will discuss the first sub-question:

   What are the effective methods for evaluating the quality of generated CT scans, and how do these methods compare in terms of reliability and comprehensiveness?

2. In Section 4.2, we will discuss the second sub-question:

   Is it possible to generate abdominal CT scans using the latent diffusion model that is initially designed for MRI scans?

3. In Section 4.3, we will discuss the third sub-question:

   To what extent does the surrounding context play a role in the generation of realistic representations of the pancreas, as analyzed through different cropping sizes of CT scans?

This method section encompasses several key components: the dataset description, model description, the experimental design and the experiments with the results. The experiment section, in turn, includes a detailed explanation of the experiment's purpose and rationale, data preprocessing techniques, training procedures, evaluation metrics used to validate the outcomes and a brief discussion of the findings.

3.1 Dataset description

The training data used for the latent diffusion model is sourced from the Medical Segmentation Decathlon (MSD) dataset [ARB+22]. The MSD challenge aimed to evaluate the capability of machine-learning algorithms in accurately segmenting a comprehensive set of prescribed regions of interest. These regions were defined by ten distinct datasets, each corresponding to a different anatomical structure and associated with at least one medical imaging task. This publicly available dataset encompasses a wide variety of labeled medical images across different anatomical structures and imaging modalities. For this research, Task_07 was used as it includes imaging data of the pancreas along with its corresponding segmentation labels. Task_07 of the MSD dataset comprises high-resolution CT scans of the abdominal region, annotated with precise segmentation labels of the pancreas. The dataset consists of 420 CT scans of which 281 scans have their segmentations. All the files in the dataset have dimensions $512 \times 512 \times Z$, where the Z value varies. Each image contains three planes: axial (transverse), coronal (frontal) and sagittal (side). In Figure 5 a slice from this dataset is shown with the green being the segmented pancreas [ARB+22].

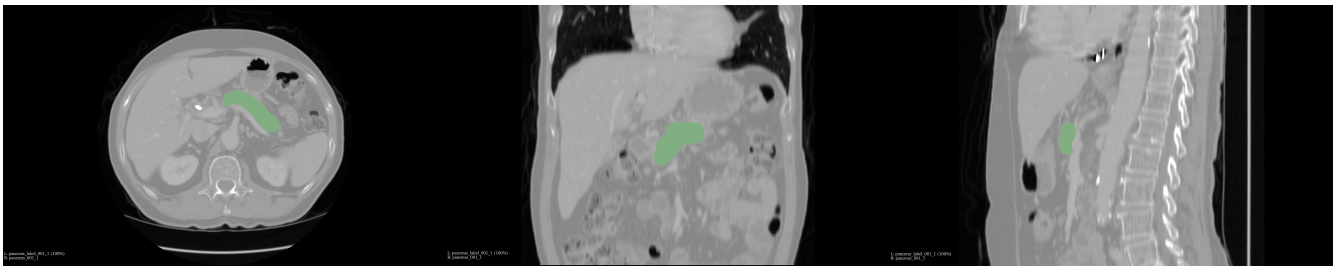


Figure 5: A slice from a scan from the MSD dataset with the green being the pancreas. The left image is the axial plane, the middle image is the coronal plane and the right image is the sagittal plan.

3.2 Model description

The model that was used in this research is called Brats mri generative diffusion and it is using the Latent diffusion model proposed by Rombach et al. [RBL+22]. The model is published by MONAI, a medical open network for artificial intelligence [PGK+23]. This model serves as a generator for creating images, it is trained as a 3D latent diffusion model, accepting Gaussian random noise inputs to produce image outputs. Figure 4 shows an overview of the latent diffusion model. At the left side of the figure you see the pixel space, you can see the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$. These are the perceptual compression models based on previous work [ERO21] and is comprised of an autoencoder trained on perceptual loss [JLO21] and a patch-based adversarial objective [DB16] [IZZE17]. Given an image $x \in \mathbb{R}^{H \times W \times D}$ with W being the width, H being height of the image

and D being the depth of the image, there is only one channel because the pixels of CT scans are grey-scale. The encoder $\mathcal{E}$ takes x as input and encodes it into a latent representation $z = \mathcal{E}(x)$. After this, the diffusion process takes place obtaining z_T. This is then inputted into the denoising step. The denoising step involves refining the latent representation z_T by progressively removing noise, a process that is learned by the neural network during the initial diffusion steps, ultimately producing a cleaner latent image ready for decoding. When the denoising step is finished, the decoder $\mathcal{D}$ decodes the image back from the latent space to pixel space, giving $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$, where $z \in \mathbb{R}^{h \times w \times d}$. The loss function that is used for this model is:

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

The loss function, $\mathrm{L}_{\mathrm{LDM}}$, is a Mean Squared Error (MSE) loss function that measures the expected squared L2 norm difference between a noise sample ϵ and the predicted noise $\epsilon_\theta(z_t, t)$ by the model, where ϵ is drawn from a standard normal distribution $\mathcal{N}(0, 1)$. It aims to train the model ϵ_θ to accurately predict the noise added to the data z_t at time t, thereby improving the model's ability to denoise or generate data. An overview of the dimension of the inputs and outputs of the used model are seen in Figure 6. The original dimensions of the input file of the MSD dataset is $512 \times 512 \times Z$, where Z varies between 37 and 751, with an average of 95. In the training of the diffusion model, this file gets resized to the input of the autoencoder $112 \times 128 \times 80$. After the encoder has processed the input, its output, which represents the latent space, has dimensions of $36 \times 44 \times 28$. Then the diffusion part takes place i.e. gradually adding Gaussian noise and the model learns to handle and reverse noise addition effectively. During inference, the process involves adding noise to the latent space and then using the neural network of the model to denoise it into something that looks like the images it was trained on. The input size for the decoder matches the output size of the encoder, and the decoder's output dimensions or the size of the generated samples of the model are $144 \times 176 \times 112$.

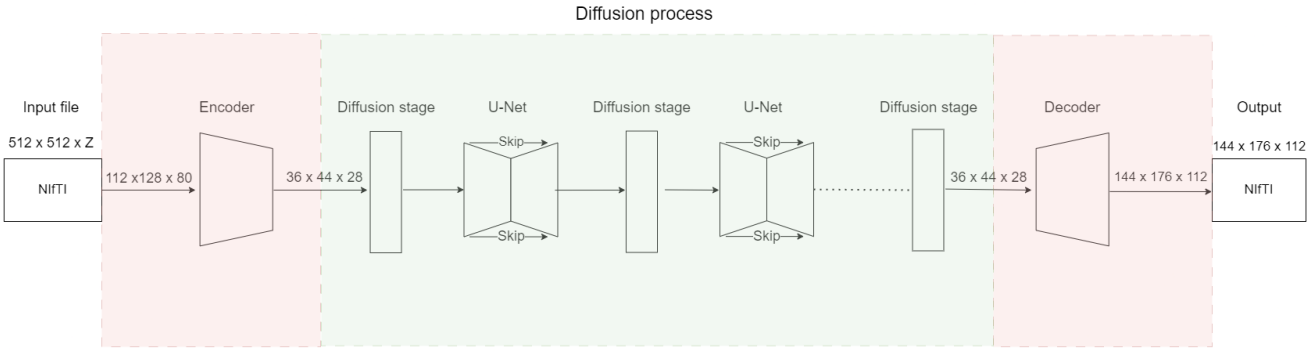


Figure 6: An overview of the model and its dimension of the inputs and outputs. The red part is the pre-trained autoencoder and the green part is the diffusion model that we trained.

3.3 Experimental design

This section outlines the experimental design used to address each sub-question. For each training session, a learning schedule was implemented. This schedule consists of a batch size, starting learning

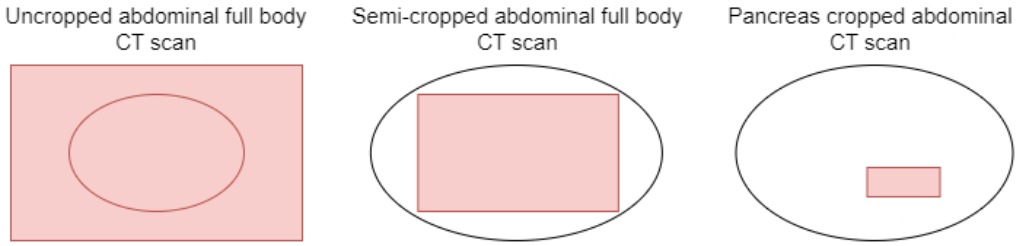
rate, milestones and a gamma parameter. The batch size is the number of training examples used in one iteration of updating the model's parameters, the starting learning rate defines the initial pace at which the model learns. Milestones are predefined points at which the learning rate is adjusted. The gamma parameter determines the factor by which the learning rate is multiplied at each milestone. The default learning schedule by MONAI for this architecture consists of a batch size of 4, a starting learning rate of 1×10^{-5}, the milestones were at 100 and 1000 epochs and the gamma was 0.1 [PGK+23]. The dimensions of the CT scans in the MSD dataset are $512 \times 512 \times Z$ [ARB+22]. We update the network's parameters according to ADAM optimization algorithm [KB14]. During training, the intermediate weights will be saved every 200 epochs to enable us to investigate the progression of the results. To answer the main question, a specific set of experiments was conducted for each sub question to gather data and draw conclusions. All the training sessions were done using an NVIDIA RTX A6000 GPU with 49,14 GB. Table 1 shows an overview of all the models that were trained.

Model	**Cropping size**	**Input Size**	**Learning schedule**	**Epochs**
Model 1	Full-body	512 × 512 × Z	Default	4600
Model 2	Pancreas	512 × 512 × Z	Default	1600
Model 3	Pancreas	512 × 512 × Z	Milestones	2600
Model 4	Pancreas	512 × 512 × Z	Starting learning rate	1000
Model 5	Pancreas	512 × 512 × Z	Gamma	1100
Model 6	Pancreas	240 × 240 × 155	Default	1800
Model 7	Semi full-body	240 × 240 × 155	Default	1800

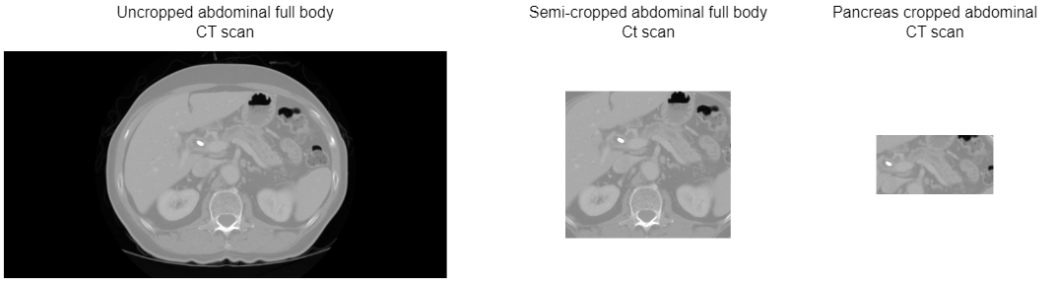
Table 1: Summary of different models that are trained.

4 Experiments

In Section 4.1, we answer sub-question 1 by studying the reliability and comprehensiveness of various evaluation metrics for assessing the quality of generated CT scans. This involves conducting experiments to determine the most suitable and reliable metrics for evaluating the generated data. In Section 4.2, we address sub-question 2 by evaluating the potential of the latent diffusion model to generate abdominal CT scans. This is done by training the model on a dataset of full-body CT scans and assessing its performance using the identified evaluation metrics. Lastly, in Section 4.3, we explore the role of surrounding context in generating realistic representations of the pancreas. This is done by training the model with different cropping sizes (full-body, pancreas, and semi-body crop) and comparing the results. These experiments will help determine the impact of context on the quality of the generated CT scans. In Section 4.3, we explore sub-question 3 by examining the impact of surrounding context on the generation of realistic representations of the pancreas. Answering this question involves training the model with different cropping sizes, as illustrated in Figure 7a (full-body, pancreas, and semi-body crop), and comparing the results to assess how context influences the quality of the generated pancreas in the CT scans.



(a) The three cropping sizes are highlighted, with the red marks indicating the regions used as input for training.



(b) The three cropping sizes seen in real CT scans.

Figure 7: Cropping sizes shown in a schematic and real manner.

4.1 Evaluation metrics

What are the effective methods for evaluating the quality of generated CT scans, and how do these methods compare in terms of reliability and comprehensiveness?

In this experiment, we will explore both quantitative and qualitative evaluation methods to assess the quality of generated CT scans. This experiment will be crucial for our research as it will determine the most effective and reliable techniques for evaluating synthetic medical images, which will directly impact the validation of our generation model and its practical applications.

In Section 4.1.1 we will discuss the FID score, a metric that measures the similarity between two sets of images based on the distribution of their feature representations as discussed in 2.5.2. In Section 4.1.2, we will discuss the grey-scale histogram. This method offers a detailed overview of the pixel intensity distribution, aiding in the identification of various tissue types and abnormalities, as discussed in 2.5.2. In Section 4.1.3, a qualitative metric called TotalSegmentator will be used. This is a model that can be used to segment organs or other tissues in CT scans [WBM+23].

4.1.1 FID score

We assessed the feasibility of using the FID score as a quantitative metric for evaluating the quality of synthetic abdominal CT scans by comparing them to real scans. The FID score, which relies on a CNN called Inception-v3 trained on 2D natural images, measures the similarity between two datasets, with lower scores indicating greater similarity.

We evaluated the feasibility of the FID score as a quantitative metric by checking if the results it provided correspond to visual interpretation. For both real and synthetic scans, we expect the FID score to be zero when the two datasets are identical, as this would indicate perfect alignment between the distributions of the two sets and confirm that the FID score algorithm functions correctly. We expect the FID score to be lower when comparing real CT scans to synthetic scans that visually resemble the real data more closely, indicating better alignment between the two datasets.

To begin, Table 2 shows the six different experiments. For experiment 1 and 2, we calculated the FID score for two sets of real abdominal CT scans that were identical, yielding a score of 0.0. This perfect score was expected, as the images in both sets were the same, confirming the FID score's sensitivity to identical images. Next, we compared two sets of real abdominal CT scans that were not identical but still comprised of real scans. This comparison resulted in a score of 852.516, reflecting the expected variability between different real images. For experiment 3 and 4 we evaluated the FID scores for synthetic full-body abdominal CT scans comparing two identical sets of synthetic scans. This produced a score of 0.0, again indicating no differences. However, comparing two different sets of synthetic scans from the same model yielded a low score of 1.405, suggesting minimal variation within the synthetic data. Similarly, for experiment 5 and 6, the model trained for 200 epochs produced FID scores of 0.0 for identical sets and 1.273 for different sets, further indicating minimal differences within this synthetic data. Finally, we applied the FID score in its intended context: comparing real abdominal CT scans with synthetic scans. When comparing a set of real CT scans to synthetic scans generated by the model trained for 200 epochs, the FID score was 4837.313. A similar comparison with synthetic scans from the model trained for 4600 epochs resulted in a slightly higher score of 4839.994. While we anticipated that the model trained for 4600 epochs would yield a lower FID score due to the visual resemblance of its synthetic scans to real CT scans, the results instead show a higher score. This unexpected outcome suggests that the FID score may not fully capture the visual improvements in the synthetic scans from the more extensively trained model.

This series of tests demonstrates that the FID score does not behave as anticipated, not aligning with visual assessments of similarity and not effectively quantifying the differences between real and synthetic images. This inconsistency underscores the limitation of the FID score, as it fails to effectively evaluate the quality of 3D medical images [YZD21]. As a result, the FID score was considered unsuitable for the experiments of this research and was therefore excluded from further analysis.

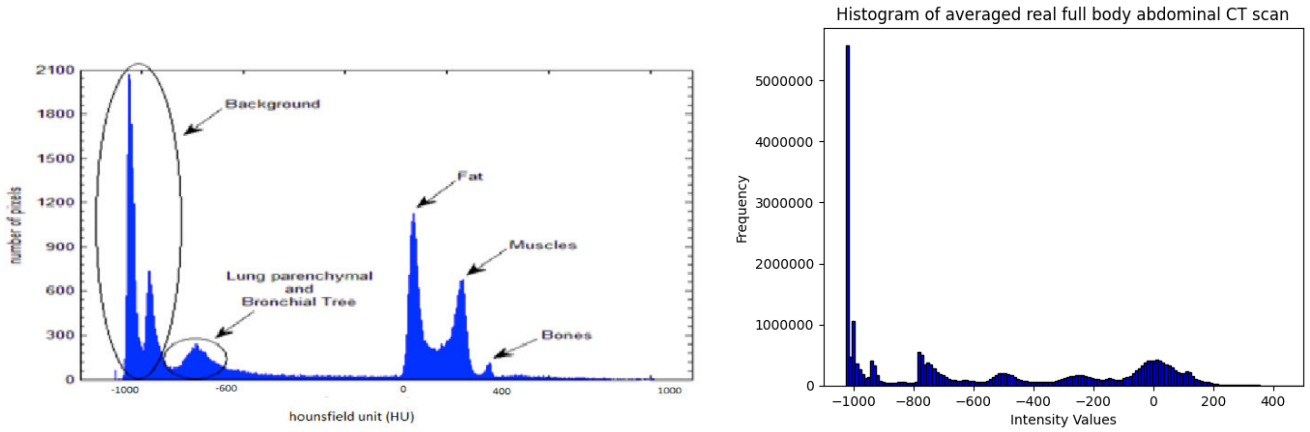
Experiment	Description	FID Score
1	Both sets with real CT scans	0.0
2	Both sets with real CT scans but not the same	852.516
3	Both sets with synthetic epoch 4600 CT scans	0.0
4	Both sets with synthetic epoch 4600 CT scans but not the same	1.405
5	Both sets with synthetic epoch 200 CT scans	0.0
6	Both sets with synthetic epoch 200 CT scans but not the same	1.273
7	One set with real CT scans, other set with synthetic epoch 200 scans	4837.313
8	One set with real CT scans, other set with synthetic epoch 4600 scans	4839.994

Table 2: FID score experiments for full-body scans

4.1.2 Grey-scale histogram

Another quantitative evaluation method that is explored is the grey-scale histogram. This metric provides a comprehensive overview of the pixel intensity distribution, which can help in identifying various tissue types and abnormalities. The X-axis represents the range of intensity values found in the 3D data array of the NIfTI file. These values typically correspond to the density of tissues in the CT scan, with different tissues having different intensity values. These intensity values are also known as Hounsfield units. The Y-axis represents the number of voxels (3D pixels) that have a specific intensity value. Essentially, it shows how many times each intensity value occurs in the entire 3D data array. Higher bars indicate that a particular intensity value is more common in the dataset. The grey-scale histogram of a real abdominal CT scan is seen in 8a and it shows three distinct peaks common in CT abdominal scans, with the first peak representing black background pixels, the second indicating low-intensity pixels from the bronchial tree and lung parenchyma, and the third corresponding to high-intensity pixels like fat, muscles, heart, bones, and pulmonary nodules [CCC20]. Figure 8b displays the average grey-scale histogram derived from four full-body abdominal CT scans. Looking at both histograms in Figure 8 reveals that the peaks are somewhat aligned. The double peak representing the background is prominently visible in Figure 8b, and the third peak, ranging from -800 HU to -600 HU, is also evident. Although the peak corresponding to fat, muscles, and bones around 0 HU is less distinct in Figure 8b, the averaging process has smoothed out the variations, resulting in a moderate alignment of these three peaks.
For our cropped pancreas CT scans, there is no widely established or reference histogram for the ground truth distribution. Therefore, we calculated the average grey-scale histogram from four real cropped pancreas CT scans from the MSD dataset [ARB+22], which serves as a reference for comparing the histograms of the synthetic scans. The same approach was taken for the semi-cropped pancreas scans. To assess whether the grey-scale histogram serves as a useful and valid evaluation metric for synthetic images, we computed the histogram for all three cropping sizes of the synthetic images. This analysis aims to determine if the model accurately captured the pixel distribution and generated the correct tissue types. The key factor is whether the peaks align with the ground truth; the specific intensity values are less important than the overall distribution pattern.

Figure 9 presents a comparison between the grey-scale histograms of pancreas-cropped regions from an average of real scans and a single synthetic scan. The distributions do not align closely, suggesting that the grey-scale histogram may not be the most suitable metric for evaluating the

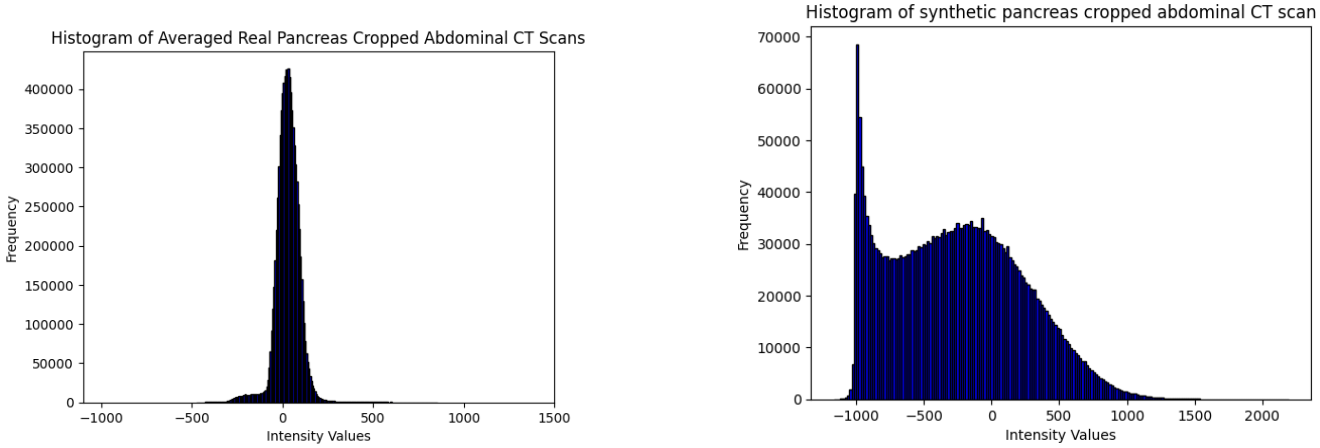


(a) The groundtruth grey-scale histogram of Hounsfield units for CT scans [CCC20].

(b) Average grey-scale histogram for real full-body CT scans.

Figure 8: Comparison of grey-scale histograms for full-body CT scans.

quality of images cropped to the pancreas.



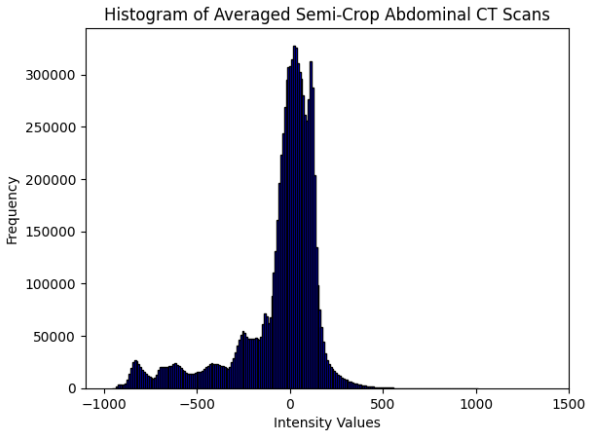
(a) The average grey-scale histogram of a real pancreas-cropped CT scan

(b) The grey-scale histogram of a synthetic pancreas-cropped CT scan.

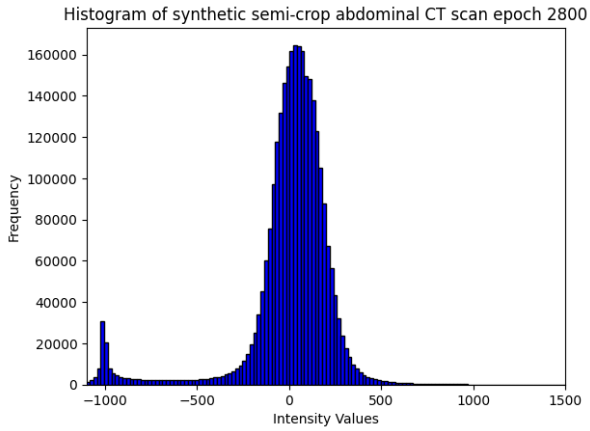
Figure 9: The grey-scale histograms of the real average pancreas-cropped and synthetic pancreas-cropped scans.

Figure 10 shows a comparison between the grey-scale histograms of semi-cropped regions from an average of real scans and a single synthetic scan. It is seen that the peaks somewhat align, especially the peak at around an intensity value of 0 HU. This suggests that, for semi-cropped images, the synthetic images capture some of the key characteristics of the real data, though there are still discrepancies.

To conclude, the grey-scale histogram serves as a useful quantitative metric for evaluating synthetic images, particularly in the context of full-body CT scans, where it effectively reflects the distribution of pixel values and various tissue types. However, for pancreas-cropped scans, the grey-scale



(a) The average grey-scale histogram of a real semi-cropped CT scan



(b) The grey-scale histogram of a synthetic semi-cropped CT scan.

Figure 10: The grey-scale histograms of the real average semi-cropped and synthetic semi-cropped scans.

histogram proves less reliable due to significant discrepancies between the histograms of synthetic and real images. These discrepancies suggest that the grey-scale distribution does not accurately reflect the true characteristics of the synthetic images, making the histogram an ineffective metric for evaluating synthetic images at this cropping size. This is because the histogram fails to capture the nuanced differences in tissue types and structures specific to the cropped region, resulting in a poor alignment with the real data. Conversely, for semi-cropped images, there is a better alignment, highlighting the potential of this metric to evaluate models when sufficient context is provided.

4.1.3 TotalSegmentator

For a qualitative evaluation metric, we utilized the TotalSegmentator model [WBM+23], this segmentation model is pre-trained on full-body abdominal scans by Wasserthal et al. To assess the model's performance, we first tested it on three different real abdominal scans, where it successfully segmented the pancreas, seen in Figure 11.

To evaluate the scans with two additional cropping sizes using the TotalSegmentator model, we input three cropped pancreas scans and three semi-cropped scans into the model. Figure 12 shows the segmentation results on the pancreas-cropped scans performed by the TotalSegmentator model. The model segments something that does not resemble the pancreas in the first two scans, and it fails to do so in the last one. In contrast, Figure 13 shows the segmentations of the pancreas on semi-cropped scans. The model demonstrates significantly improved accuracy in capturing the shape of the pancreas across all three scans, compared to the segmentations on fully cropped scans. Here, accuracy refers to the model's ability to precisely match the true anatomical boundaries of the pancreas.

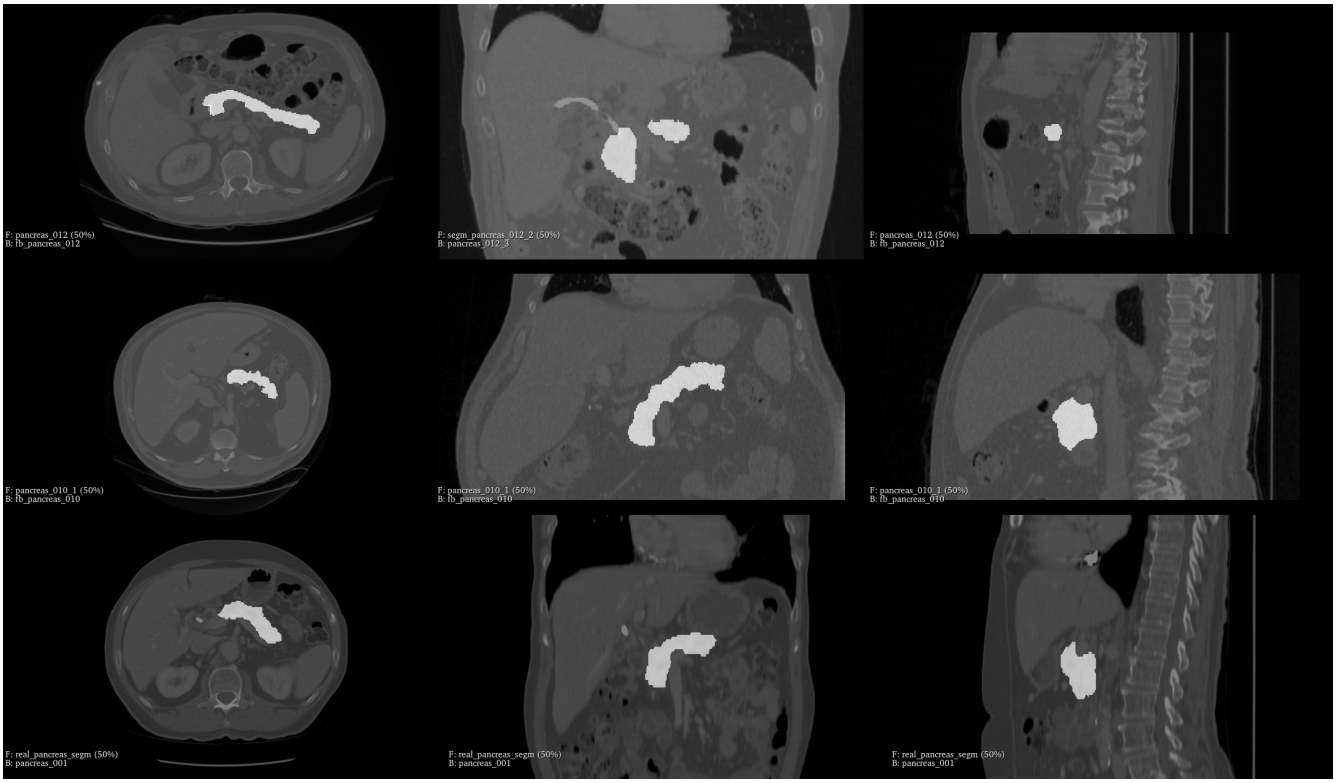


Figure 11: Segmentations of three real full-body abdominal scan using the TotalSegmentator with the left column being the axial plane, the middle column being the coronal plane and the right column being the sagittal plane. Each row represents a different scan

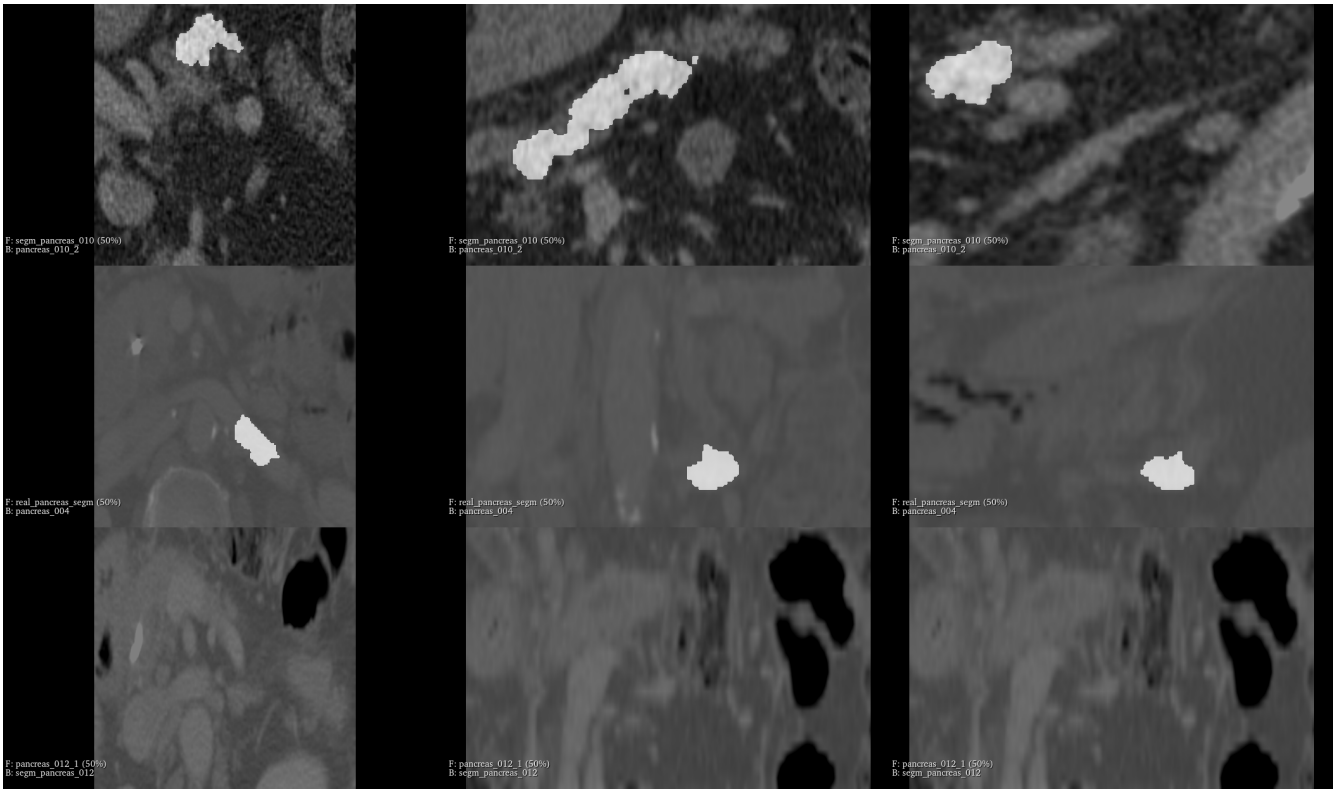


Figure 12: Segmentations of the real pancreas cropped scan using the TotalSegmentator.

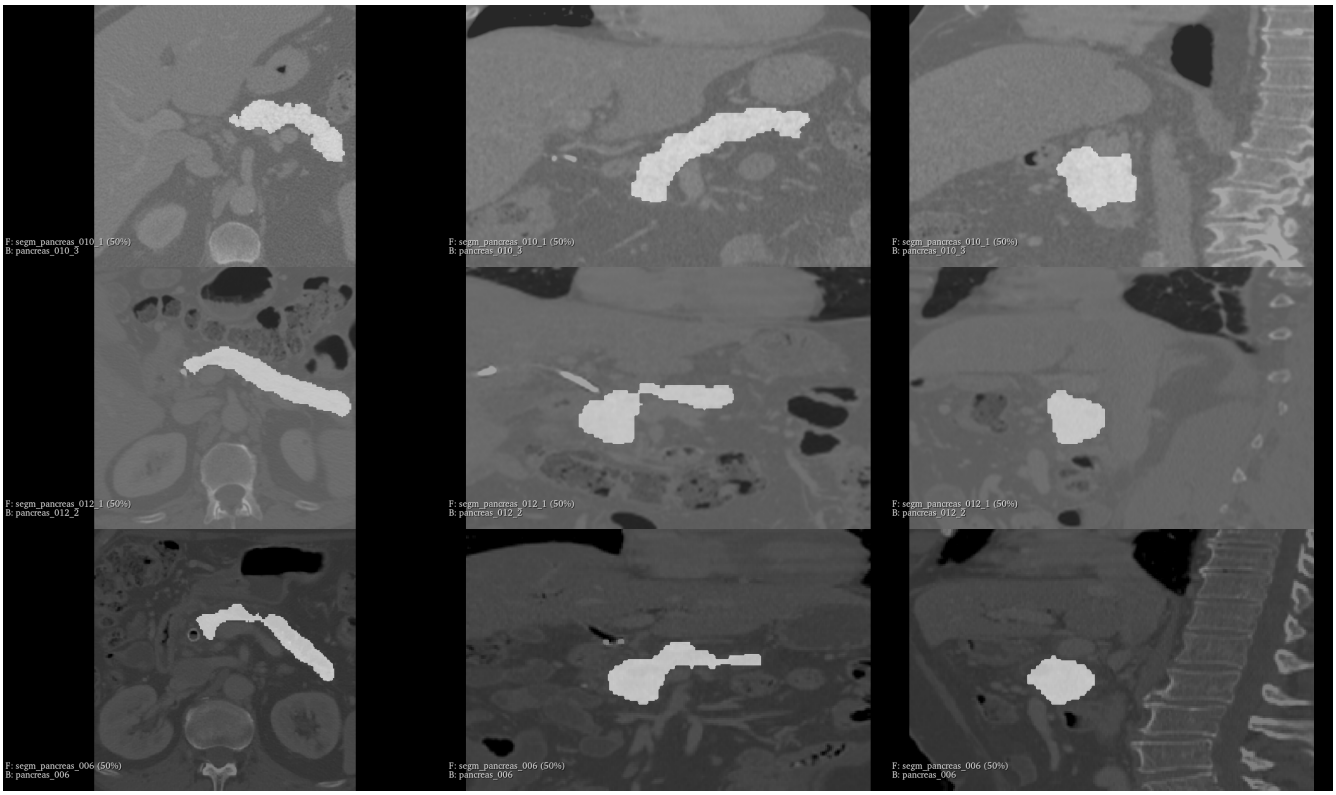


Figure 13: Segmentations of the real semi-crop abdominal scan using the TotalSegmentator.

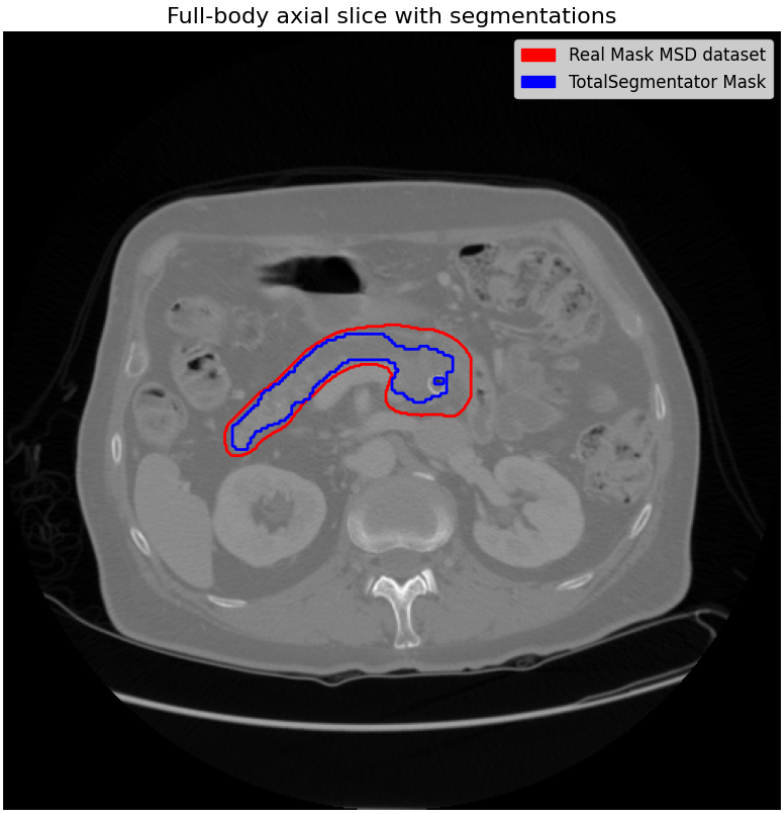


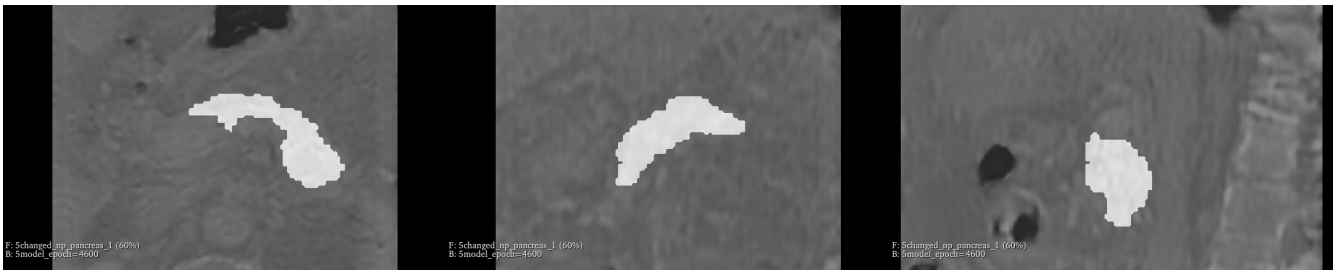
Figure 14: Segmentations of the MSD dataset and the TotalSegmentator.

To evaluate the performance of the TotalSegmentator on real scans, we calculated the DICE score for the full-body scans in the MSD dataset. By comparing the segmentations provided by the MSD dataset [ARB+22] with those generated by the TotalSegmentator, we found a DICE score of 0.724 for the full-body scans. Figure 14 illustrates the segmentations, with the MSD dataset segmentation shown in red and the TotalSegmentator segmentation in blue. As discussed in Section 2.5.2, the TotalSegmentator achieves a DICE score of 0.943 [WBM+23]. The lower DICE score observed in our comparison is likely due to the higher precision of the TotalSegmentator's segmentations. As shown in Figure 14, the TotalSegmentator's segmentation is entirely contained within the MSD segmentation, demonstrating that the TotalSegmentator performs well.

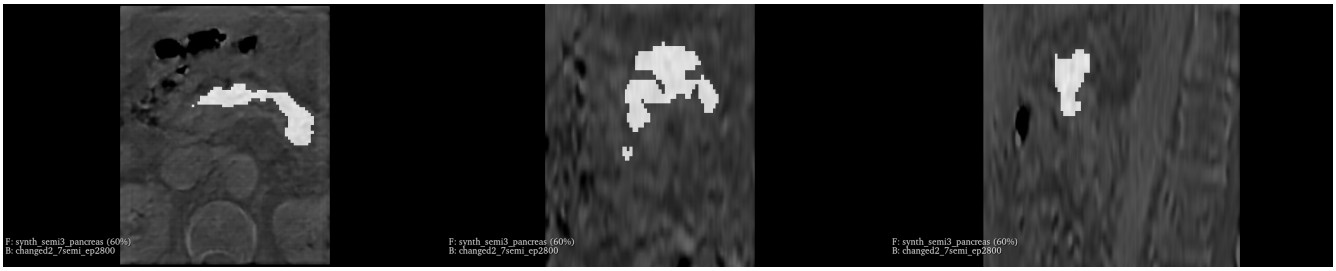
Subsequently, we inputted the synthetic scans to determine if the model could similarly segment the pancreas in these images. For the synthetic cropped pancreas scans, the model failed to produce a segmentation, demonstrating that the TotalSegmentator is not suitable for processing images that are cropped to the pancreas. Figure 15 presents the segmentations of synthetic images for both full-body and semi-cropped abdominal scans. When comparing the segmentations across different planes of these synthetic scans to the segmentation shown in Figure 11 and Figure 13, it is evident that the TotalSegmentator model captures the pancreas shape quite accurately for the full-body scans. However, verifying the precise location remains challenging. For the semi-cropped scans, the model's segmentation reasonably approximates the pancreas shape, though it is not as precise. Similar to the full-body scans, the exact positioning of the pancreas remains challenging in the

semi-cropped images as well.

Overall, the TotalSegmentator proves to be a valuable tool for evaluating the quality of synthetic scans, as it effectively assesses the anatomical accuracy of the generated images. However, it is not suitable for pancreas-cropped scans, as it fails to produce segmentations for these images.



(a) Segmentation of the synthetic full-body abdominal scan using the TotalSegmentator.



(b) Segmentation of the synthetic semi-crop abdominal scan using the TotalSegmentator.

Figure 15: Comparison of segmentation results for synthetic CT scans using the TotalSegmentator, with the left images being the axial plane, the middle being the coronal plane and the right being the sagittal plane.

In summary, we explored various evaluation metrics to assess the quality of synthetic abdominal CT scans that can be used for different cropping sizes. The FID score proved unreliable for evaluating the quality of synthetic abdominal CT scans, as it failed to reflect the visual improvements between models trained for different epochs, highlighting its unsuitability for assessing 3D medical images. The grey-scale histogram was explored as a quantitative evaluation method to assess whether the synthetic abdominal CT scans accurately captured the pixel intensity distribution and tissue types, by comparing the histograms of synthetic scans with those of real scans. The distributions of the full-body and semi-crop scans overlap reasonably with their real histogram but the pancreas cropped scans did not. Despite this, the grey-scale histogram is still a useful metric for assessing the quality of the generated scans. The TotalSegmentator model successfully segmented the pancreas in real full-body abdominal scans together with the synthetic full-body scan. The TotalSegmentator struggled with the real cropped pancreas scans and was not able to segment anything from the synthetic pancreas cropped scans. However, it performed well on real and synthetic semi-cropped scans. Overall, these evaluation methods, except the FID score, collectively provide a comprehensive assessment of the synthetic CT scans' quality and realism.

4.2 Feasibility of generating CT scans

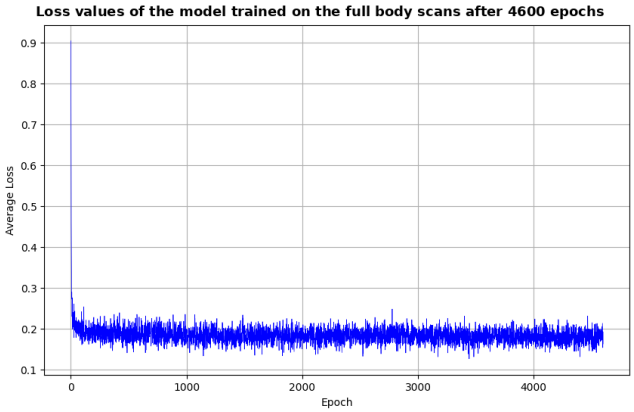
Is it possible to generate abdominal CT scans using the latent diffusion model that is initially designed for MRI scans?

This question was addressed by training the diffusion model on full-body abdominal CT scans and evaluating its performance using multiple methods: visual inspection, comparison of the grey-scale histogram of the synthetic scan with that of a real scan, segmentation of the pancreas using the TotalSegmentator and a qualitative rating by two clinicians of UMC Utrecht. The model was trained for 4600 epochs using the default learning schedule and the default input dimensions, as mentioned in Section 3.3.

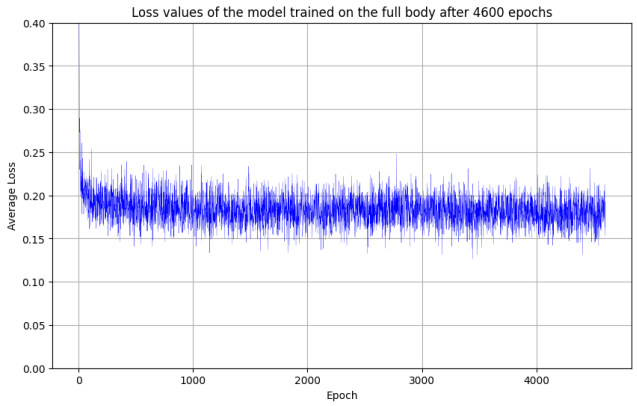
Visual inspection

Figure 17 illustrates the progression of the results over different training epochs. Notable improvements are evident in the first five rows, corresponding to epochs 200, 600, 1000, 1400, and 1800. During these epochs, the images demonstrate a clearer depiction of the spine in both the coronal (middle) and sagittal (left) planes, and the organs visible in the axial plane become more defined and robust. Additionally, the images generally exhibit reduced noise compared to earlier epochs.

However, a significant decline in image quality is observed in the second-to-last row, which represents epoch 4200, as seen in Figure 17. This decline is particularly pronounced in the coronal and sagittal planes, where the clarity and detail of the features are notably worse than those in the preceding epochs. This decline in image quality at epoch 4200 could suggest that the model is beginning to overfit, where it becomes too specialized to the training data and fails to generalize effectively. This overfitting often results in reduced clarity and increased noise in the generated images. However, the sample from epoch 4600, which appears to show improved quality, indicates that subsequent training may have allowed the model to recover from overfitting. This improvement could be due to stabilization in the later stages of training, which helped the model to generate more refined and accurate images.



(a) Loss values of the model trained for 4600 epochs on full-body abdominal scans.



(b) Loss values of the model trained for 4600 epochs on full-body abdominal scans with the Y-axis ranging from 0 to 0.4 for extra clarity.

Figure 16: Comparison of loss values for the model trained for 4600 epochs on full-body abdominal scans.

Figure 16 supports this observation, as it shows that the loss values plateau very quickly after epoch 250. This plateau suggests that the model's ability to learn from the data effectively has reached a limit. Despite this, in Figure 17 the improvements in the first five rows (epochs 200 to 1800) are evident, with a clearer spine, more robust organs, and less noise. Although there is a general decline in loss values initially, the improvements in image quality become minimal beyond epoch 1800, indicating diminishing returns with further training. Thus, while earlier epochs bring substantial enhancements in image quality, the minimal gains and observed decline in visual quality after epoch 1800 suggest that the additional training may lead to overfitting rather than further improvements.

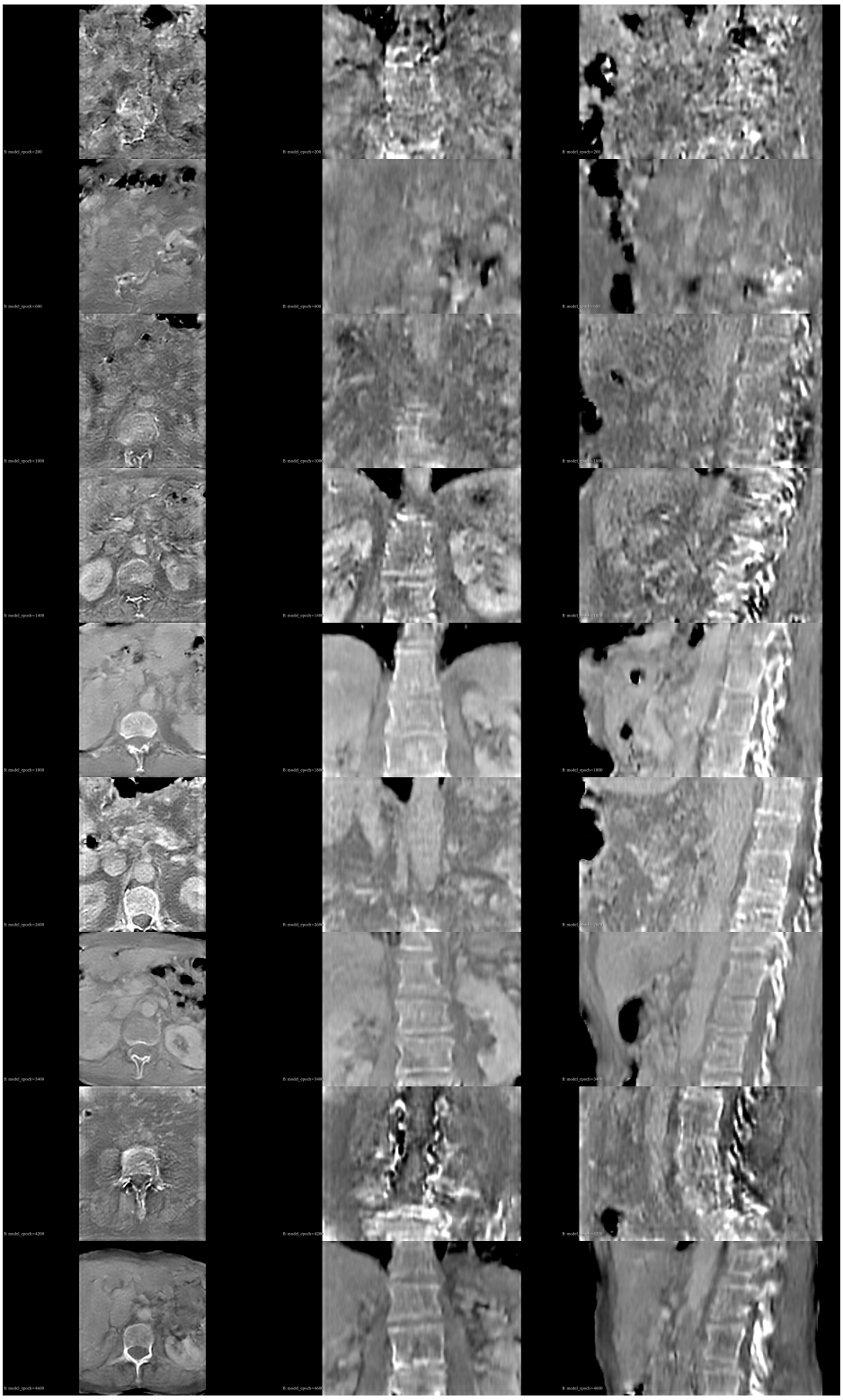
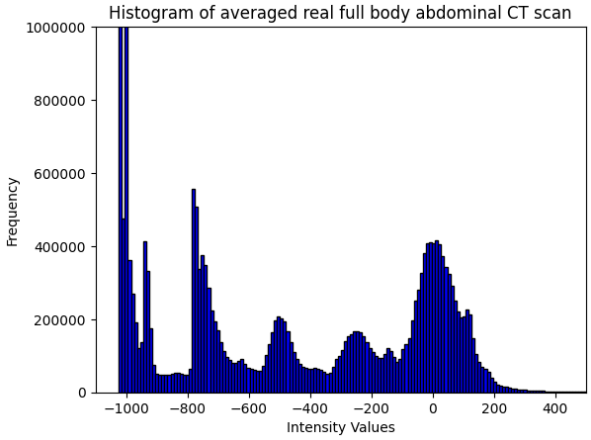


Figure 17: Progression of results of the full-body abdominal model. Every column stands for the three different planes: axial, coronal and sagittal respectively. Every row corresponds to a certain epoch: 200, 600, 1000, 1400, 1800, 2600, 3400, 4200, 4600 respectively.

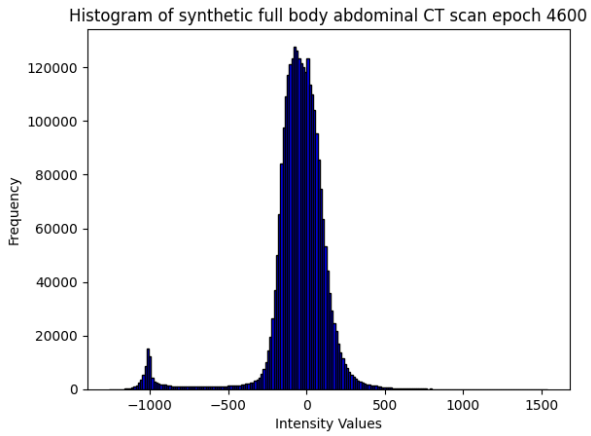
Grey-scale histogram

In Figure 8b, we observed the histogram of averaged real full-body scans. Figure 18a presents this distribution once more, but with a scaled-down Y-axis to emphasize the variation in other values. If we look at the grey-scale histogram of the synthetic sample in 18b and compare it to the average histogram seen in Figure 18a, we can see that the peaks present in the synthetic sample more or less align. In both histograms we can see a peak at an intensity of -1000, the reason why this peak is much smaller in Figure 18b is because the output of the synthetic samples is much more zoomed. This zoomed-in view emphasizes the details within the synthetic body scan, thereby reducing the relative proportion of the background in the histogram. Despite this, both histograms still exhibit double peaks at intensity values between 0 HU and 200 HU.

In summary, the histograms in Figures 8b and 18a show somewhat similar peak structures for both real and synthetic scans. The synthetic sample's zoomed-in view reduces the prominence of some peaks but still aligns well with the real data, particularly the double peaks between 0 HU and 200 HU.



(a) The average grey-scale histogram of a real full-body CT scan with the Y-axis ranging from 0 to 1000000.



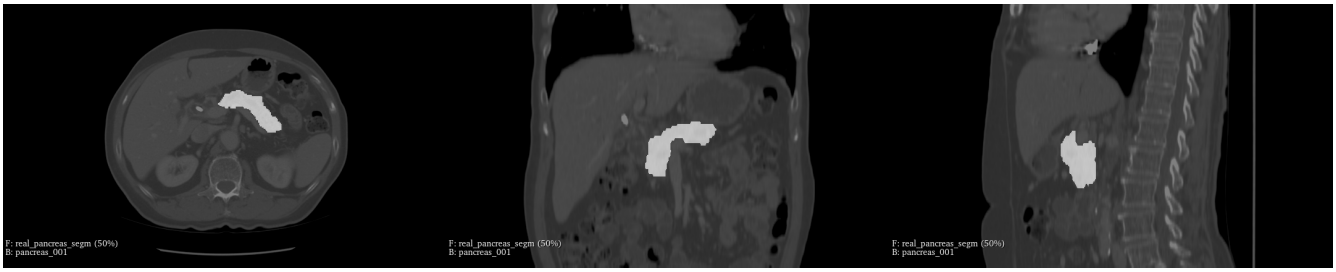
(b) The grey-scale histogram of a synthetic full-body CT scan after 4600 epochs of training.

Figure 18: Comparison of grey-scale histograms

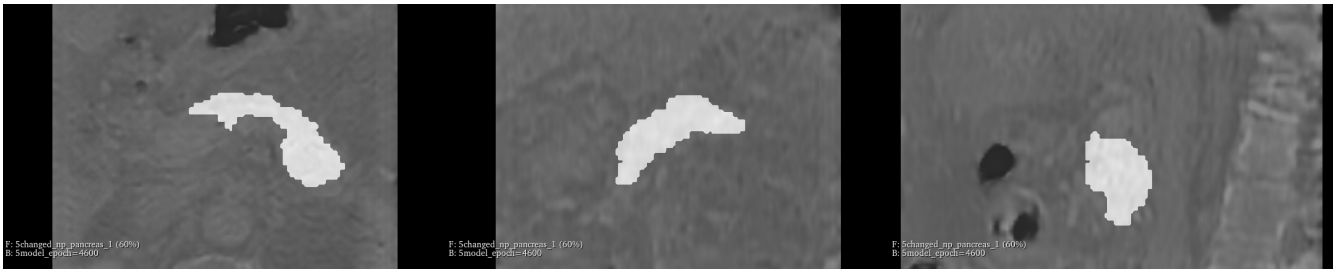
TotalSegmentator

As seen in Figure 19b, on full-body scans, the TotalSegmentator is able to segment something that looks like a pancreas. If we compare the shapes of the segmentation of the three different planes to those in Figure 19a, we can see similarities. Also the position of the segmentation seems to be similar. If we look at the rightmost (sagittal) plane of Figure 19b, we see that it is close to the spine, which is also the case with the segmentation of the real scan. It is difficult to discern the exact position of the segmentation of the middle (coronal) plane due to the limited contrast between the segmented region and the surrounding tissues.

Overall, the segmentation results in Figure 19b indicate that the TotalSegmentator successfully identifies and segments the pancreas in synthetic full-body scans, showing notable alignment with the segmentations from real scans. This suggests that the model is capable of generating synthetic CT scans of sufficient quality, as evidenced by the somewhat accurate segmentation results.



(a) Segmentation of the real full-body abdominal scan using the TotalSegmentator.



(b) Segmentation of the synthetic full-body abdominal scan using the TotalSegmentator.

Figure 19: Comparative segmentation of full-body abdominal scans: real (top) and synthetic (bottom) scans using the TotalSegmentator.

Rating UMCU

To evaluate the synthetic data qualitatively, an experiment was designed in which two clinicians from UMC Utrecht were asked to rate the synthetic samples. The experiment commenced with an explanation of the setup and presentation of a table outlining the criteria for rating the synthetic samples as seen in Table 3.

Rating	Description of the scan's visual features
0	Extremely noisy and no organs are distinguishable
1	Noisy and only a few organs are somewhat distinguishable
2	Somewhat noisy and some organs are faintly distinguishable
3	Somewhat realistic and some organs are distinguishable
4	Realistic and most organs are as distinguishable as in a real scan
5	Perfectly realistic and all the organs are as distinguishable as in a real scan

Table 3: Criteria for rating the quality of CT Scans

Initially, the raters first viewed a video of a real CT scan, which displayed the axial, coronal and sagittal planes to provide a comprehensive view of the 3D structures. Subsequently, six synthetic samples were presented, each beginning with a 3D video of the scan, followed by individual presentations of the axial, coronal and sagittal planes. Clinicians then rated each plane of the synthetic samples according to predefined criteria. The ratings for each plane were averaged to obtain a mean score for each plane and for the synthetic sample. The results are presented in Table 4.

Synthetic Sample	Axial Mean Rating	Coronal Mean Rating	Sagittal Mean Rating	Mean Sample Rating
1	2.5	1.5	2.0	2.0
2	1.0	0.5	0.5	0.7
3	3.0	3.0	1.5	2.5
4	0.5	2.5	1.0	1.3
5	1.5	1.5	1.0	1.3
6	0.5	0.5	0.0	0.3
Mean Rating	1.5	1.6	1.1	

Table 4: Mean ratings of the synthetic CT scan samples across the three planes.

The mean ratings indicate variability in the perceived quality of the synthetic samples, with some samples such as Synthetic Sample 3 receiving relatively higher ratings across all planes, suggesting better anatomical representation. Conversely, samples like Synthetic Sample 6 were rated poorly, particularly in the sagittal plane, indicating a lack of clarity or accuracy in those views. On average, the coronal plane received the highest mean rating (1.6), followed by the axial plane (1.5) and the sagittal plane (1.1), suggesting that the synthetic images were generally perceived as being of slightly better quality in the coronal and axial planes compared to the sagittal plane. These results highlight the varying degrees of effectiveness in generating synthetic scans, with some samples achieving closer alignment with expected anatomical structures than others.

The interobserver variability was assessed using Cohen's Kappa for the axial, coronal, and sagittal planes. Table 5 shows the results. These values indicate a slight to fair agreement between the two raters for the sagittal plane, while the axial and coronal planes show poor agreement.

Plane	Cohen's Kappa
Axial	-0.07
Coronal	-0.29
Sagittal	0.28

Table 5: Cohen's Kappa values for interobserver variability across different planes.

The evaluation of synthetic CT scans based on specific slices and overall impressions reveals significant challenges in accurately assessing image quality. While individual slices may present with poor resolution or clarity, the overall impression given by the video can still convey a more cohesive and recognizable anatomical structure. Furthermore, the axial plane appears to provide the most consistent anatomical recognition, particularly for the kidneys and spine, whereas the identification of the intestines remains notably difficult. This raises the question of whether mere recognizability of organs should be the primary benchmark for evaluation. It is possible to recognize an organ, yet the scan may still fail to be anatomically accurate or present the correct form and structure of the organs. Therefore, while recognition provides some indication of image quality, it is essential to also consider the anatomical correctness and overall coherence of the scans in the evaluation criteria.

To summarize and answer the second sub-question, the visual inspection reveals significant improvements in image quality up to epoch 1800, with clearer structures and reduced noise. However, the quality declined at epoch 4200, and subsequent improvements are minimal, suggesting that further training may lead to overfitting. The grey-scale histograms of the synthetic sample in Figure 18b and the real scan in Figure 18a show a slight peak alignment, with differences due to the zoomed-in synthetic samples. The TotalSegmentator segments a pancreas-like structure in full-body scans, similar in shape and position to real scans, though exact positions in the coronal plane are harder to discern due to limited contrast. Overall, the synthetic scans varied in quality, with the coronal (1.6) and axial (1.5) planes rated higher than the sagittal (1.1), and mean sample ratings ranging from 0.3 to 2.5, indicating some scans were closer to expected anatomical structures than others. To conclude, taking the results of the evaluation metrics into account, it is possible to generate abdominal CT scans using the latent diffusion model initially designed for MRI scans.

4.3 The role of surrounding context

To what extent does the surrounding context play a role in the generation of realistic representations of the pancreas, as analyzed through different cropping sizes of CT scans?

In Section 4.2, we presented the results when the diffusion model was trained on full-body abdominal CT scans. To further investigate, we trained the model on two additional crop sizes: one focused specifically on the pancreas, as discussed in 4.3.1, and another encompassing a semi-body region, as discussed in 4.3.2. Figure 20 shows the three cropping sizes in this research. Comparing the results of these three cropping sizes will allow us to evaluate the impact of different contexts on the visual quality of the generated CT scans, its ability to generate the pancreas and having the correct distribution of grey-scale pixels. By employing a larger cropping size, the model had access to more contextual information, whereas a smaller cropping size provided the model with less contextual data. This approach helped us understand how varying amounts of context influence the quality and realism of the generated pancreas in the synthetic scans.

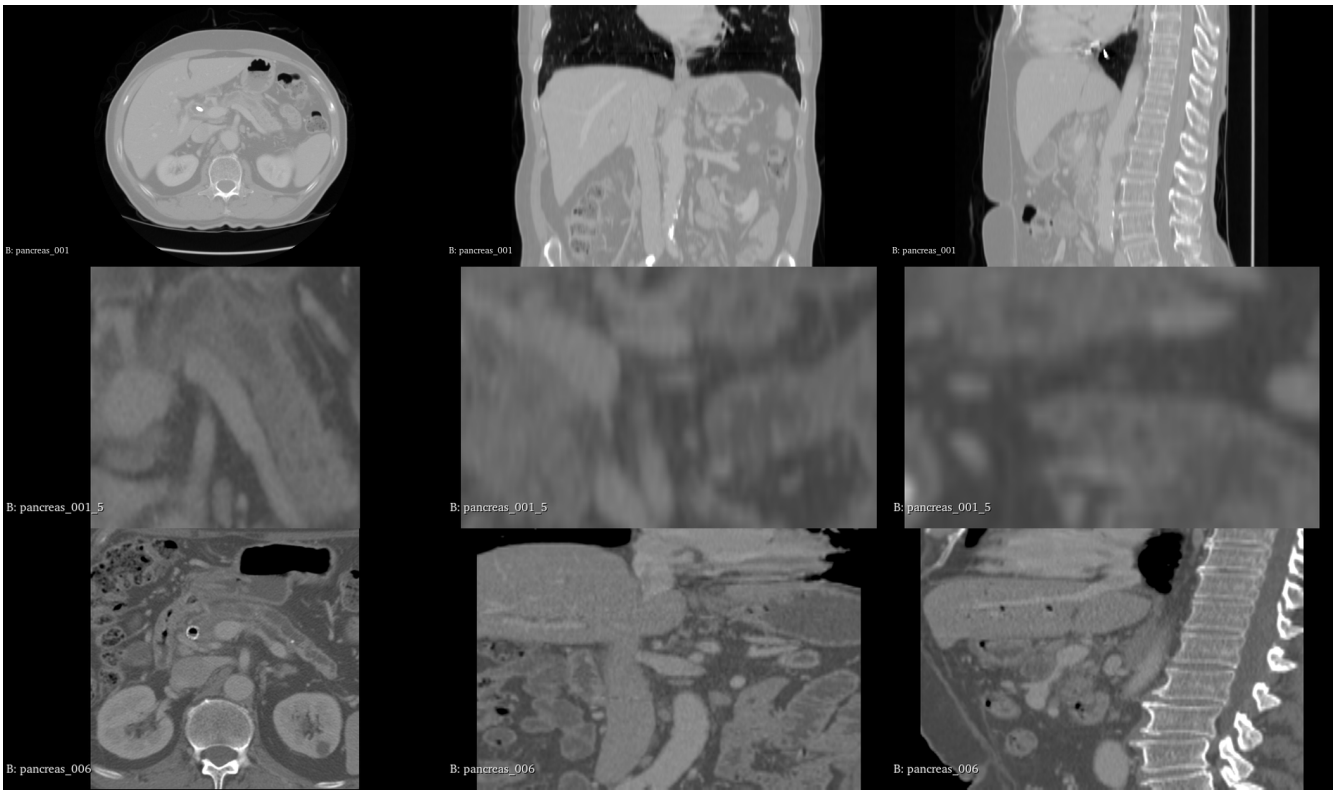


Figure 20: Slices of the real full-body (first row), cropped pancreas (second row) and semi-crop (third row) abdominal CT scans. The left images are the axial plane, the middle images are the coronal plane and the right images are the sagittal plan.

4.3.1 Pancreas cropped

Figure 21 show the pre-processing steps, in the first step the pancreas got segmented using the segmentations provided by the MSD dataset [ARB+22]. After the segmentation, the rectangle seen in blue, the image got cropped to the region that contained the pancreas. Subsequently it got resized back to $512 \times 512 \times Z$.

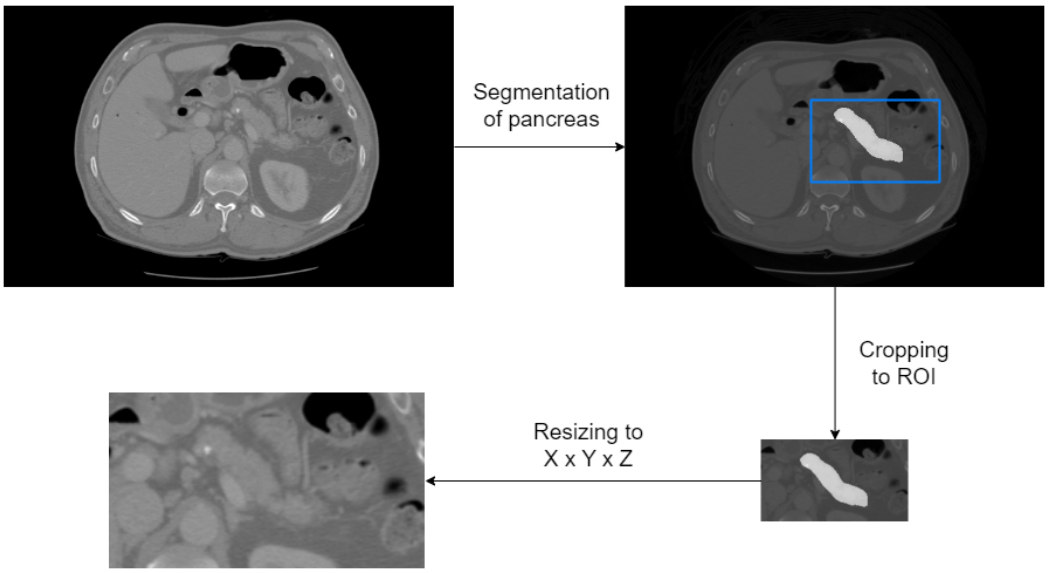


Figure 21: The pre-processing steps for cropping the images to the pancreas.

Training sessions cropped pancreas

Table 6 shows the different training sessions of the cropped pancreas scans. Model 1 has the default learning schedule, as discussed in 3.3, while the other four models deviate from this schedule. Each model has one parameter that deviates from the default learning schedule.

Model	Milestones	Starting learning rate	Gamma	Input dimensions	Number of epochs
1	100, 1000	10^{-5}	0.1	$512 \times 512 \times Z$	1600
2	700, 1400	10^{-5}	0.1	$512 \times 512 \times Z$	2600
3	100, 1000	10^{-4}	0.1	$512 \times 512 \times Z$	1000
4	100, 1000	10^{-5}	0.5	$512 \times 512 \times Z$	1100
5	100, 1000	10^{-5}	0.1	$240 \times 240 \times 155$	1800

Table 6: An overview of the different training session (models) of the cropped pancreas scans with different implemented learning schedules.

Model 1

The first row of Figure 23 shows a sample generated by model 1 after 1600 epochs of training using the default learning schedule. The results are noisy and lack the clarity needed for accurate analysis, indicating that the model struggles to produce high-quality images under these conditions. In Figure 24, it is observed that the loss value (dark blue) falls quickly and then stagnates after epoch 250, suggesting limited further improvement despite continued training.

Model 2

Given the sub-optimal results of model 1, we decided to retrain the diffusion model. In model 2, we adjusted the training milestones from 100 and 1000 to 700 and 1400. This change was made to extend the period during which the model trains at a higher learning rate, with the aim of enhancing its ability to capture more nuanced features from the data and ultimately improving the quality of the generated images. The second row of Figure 23, displays the generated sample produced by model 2 after 2600 epochs. Figure 24 shows the corresponding loss values (green) over time. Despite the introduction of new training milestones, the changes had little impact on both the visual quality of the generated image and the behavior of the loss values. The loss still decreased rapidly and then stagnated, indicating that changing the milestones did not significantly influence the model's performance.

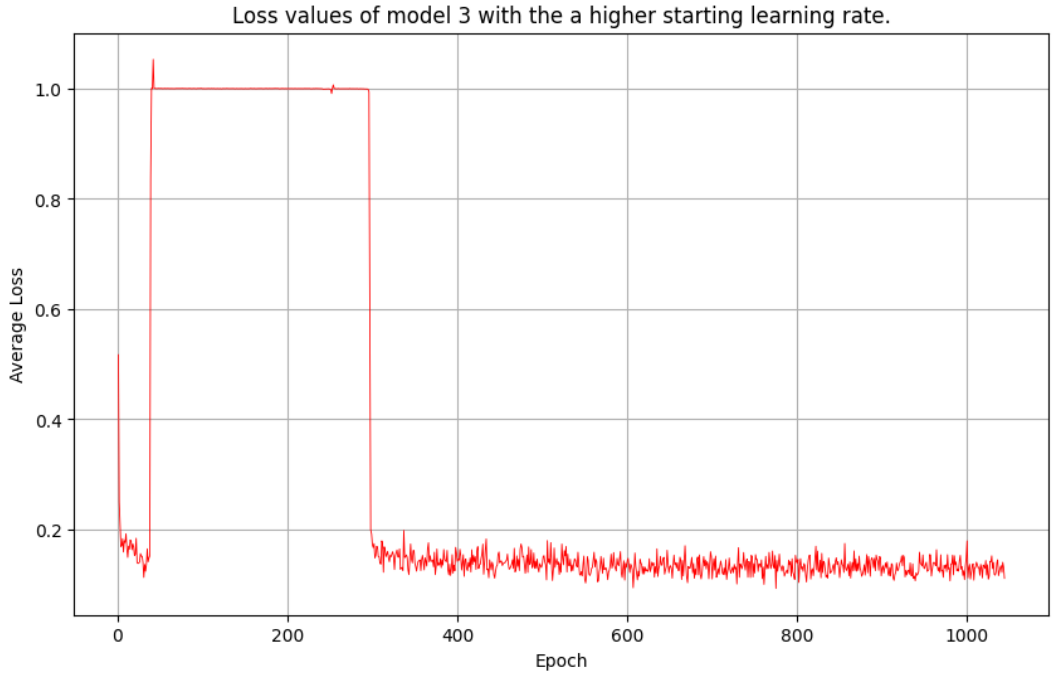


Figure 22: Loss values of model 3 with the a higher starting learning rate with the y-axis ranging from 0 to 1.0.

Model 3

We retrained the diffusion model on the cropped pancreas but this time with a higher starting learning rate. By increasing the learning rate, the model can adjust its weights more rapidly in the early stages of training. The model was trained for 1000 epochs. In the third row of Figure 23, we show the sample generated by the model after 1000 epochs. The visual quality has not improved compared to the other training sessions. Figure 22 takes a closer look at the loss values of model 3, also seen in red in Figure 24. The observed spike in loss values at the beginning of the training indicates that the model's performance was adversely affected by the initially high learning rate. This suggests that such a high learning rate was inappropriate and led to instability in the training process.

Model 4

For the next training on the cropped pancreas, the gamma parameter of the learning schedule was changed from 0.1 to 0.5. This was done to increase the rate at which the learning rate is reduced over time. The motivation behind this is that it can help in stabilizing training and potentially improve convergence by allowing the model to learn more effectively during later stages of training, especially if the previous setting led to an excessively aggressive decrease in learning rate. In the fourth row of Figure 23 we see that the results are still noisy. Figure 24 show the loss values (turquoise) over time. The loss values have not improved compared to the other models, this means increasing the gamma parameter does not significantly enhance the stability or quality of the generated images, indicating that the adjustment was not effective in addressing the noise and loss stagnation observed in previous models.

Model 5

For the final training of the diffusion model on the cropped pancreas, we decided to decrease the dimensions of the CT scans. Specifically, the cropped pancreas scans were resized to $240 \times 240 \times 155$ using bilinear interpolation, instead of the original $512 \times 512 \times Z$. This change aimed to address the issue of zoomed-in outputs produced by the model. The output appears more zoomed-in because the resizing process involves mapping the larger original dimensions $512 \times 512 \times Z$ into a smaller fixed size $112 \times 128 \times 90$, leading to a greater reduction in resolution and detail compared to resizing a smaller original image $240 \times 240 \times 155$ to the same size. This greater reduction in detail results in a more zoomed-in appearance as each voxel, or volume element, in the output represents a larger portion of the original image. The diffusion model was trained for 2800 epochs and had the default learning schedule, except the input dimensions of the scans. However, despite this adjustment, model 5, as shown in Figure 24, exhibits significantly higher loss values (violet) compared to the other models. Additionally, the quality of the generated samples did not improve, with the results still appearing noisy, as illustrated in the fifth row of Figure 23.

Looking at the results and loss values from the five training sessions with cropped pancreas scans, it became clear that the limited context provided by these images hindered the model's ability to learn accurate representations and generate quality images of the pancreas region. Consequently, we decided against further evaluating these samples with the TotalSegmentator and grey-scale

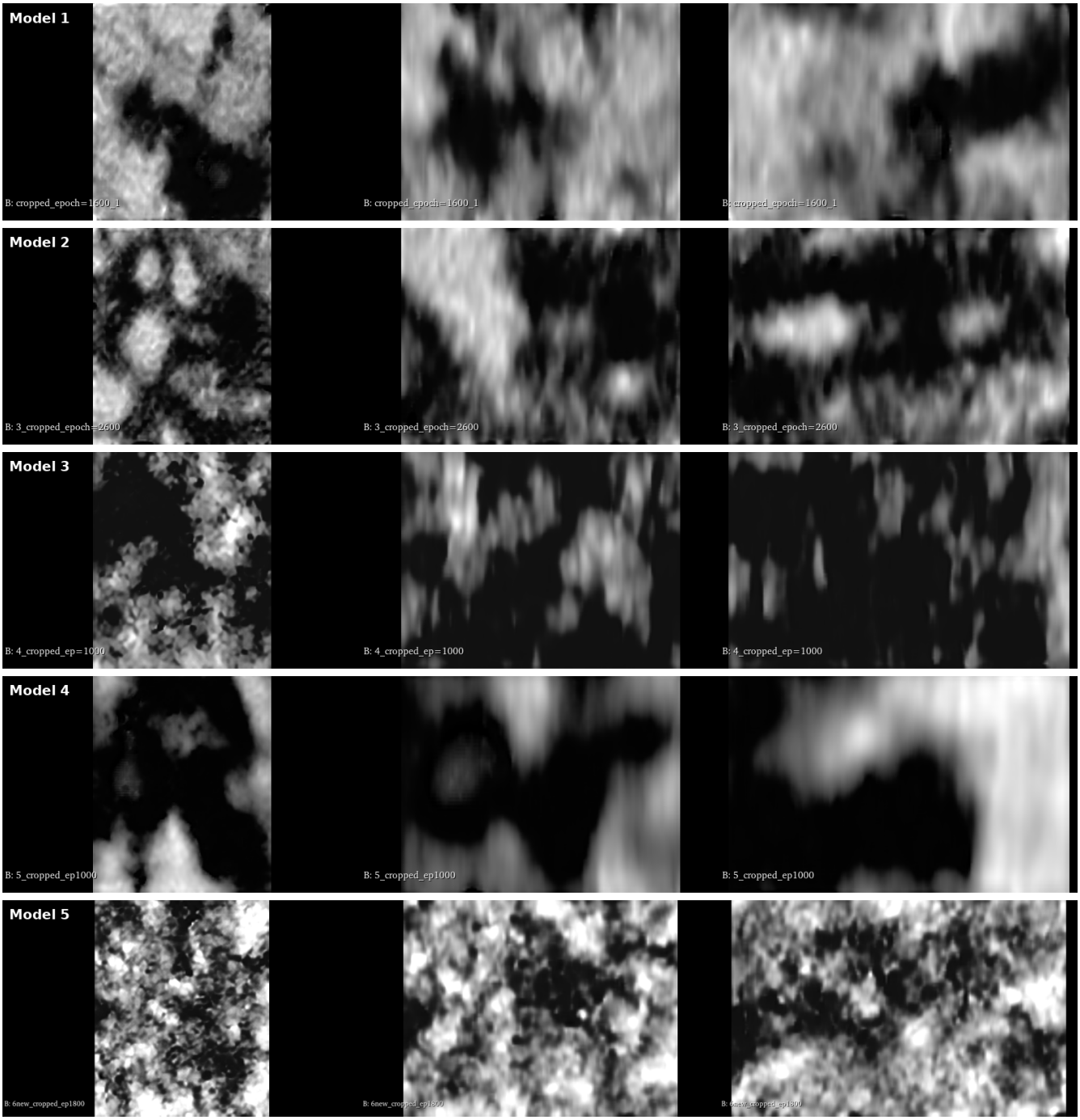


Figure 23: Results of the all the five training sessions (models) with every row being a model and each column being the axial, coronal and sagittal plane respectively.

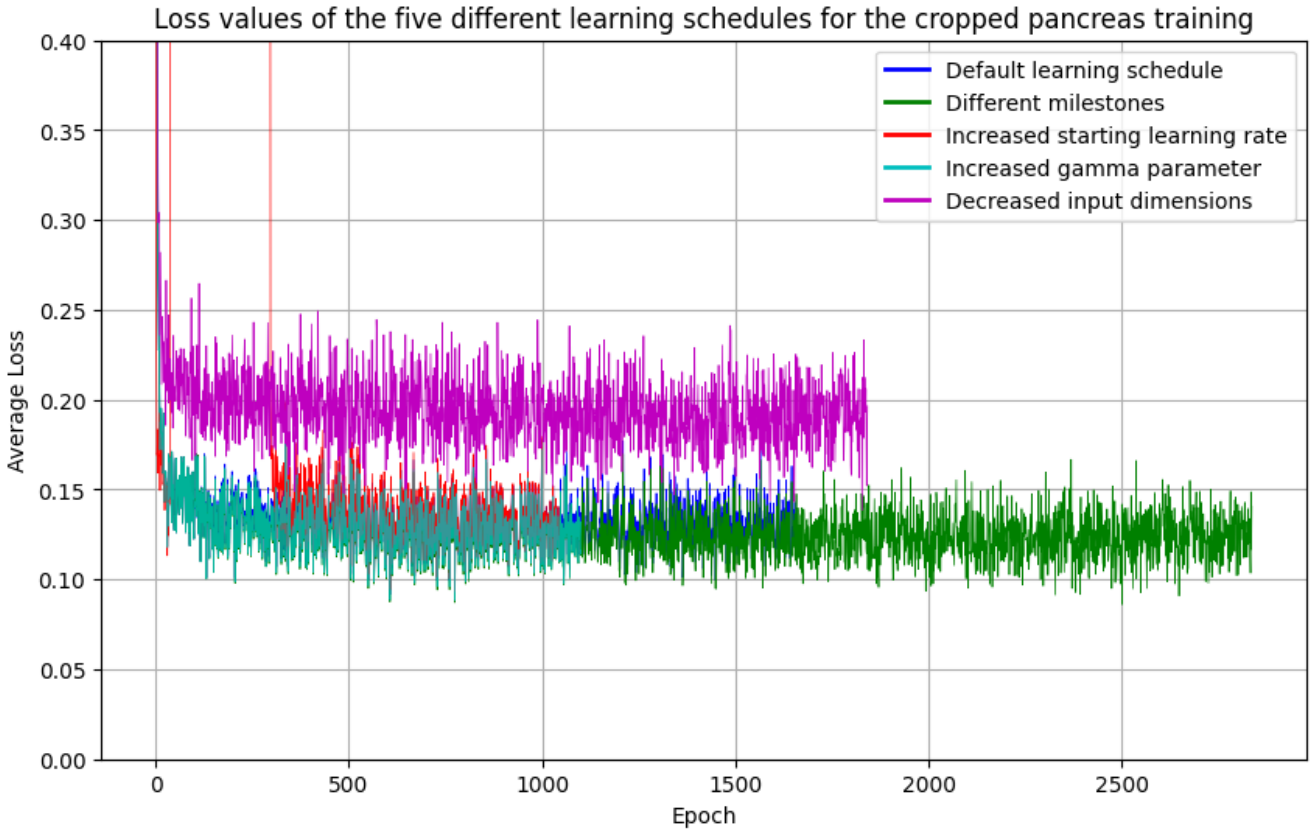


Figure 24: Loss values of the five different models during training with the y-axis ranging from 0 to 0.4.

histogram methods. To address this limitation, we turned to a different approach involving semi-body scans, which provide a broader context for the model.

4.3.2 Semi-crop

To test whether providing more contextual information improves the model's performance, we conducted a training session using semi-body scans. This approach involved defining the semi-body crop as the largest square that fits within the body segmentation. By increasing the crop size and incorporating more contextual data, we aimed to enhance the model's ability to accurately generate images, contrasting with the limitations observed in the previous cropped pancreas experiments. In Figure 25 the pre-processing steps are shown. In the first step, the trunk of the body is segmented using the TotalSegmentator model configured for body segmentation [WBM+23]. After that, the largest approximate cube that fit within the segmentation area was identified, as indicated by the red square in the third step of Figure 25. The image is cropped to this identified region and then resized to $240 \times 240 \times 155$ using bilinear interpolation. We chose this resizing to mitigate the zoomed-in output of the model, as explained in Section 4.3.1.

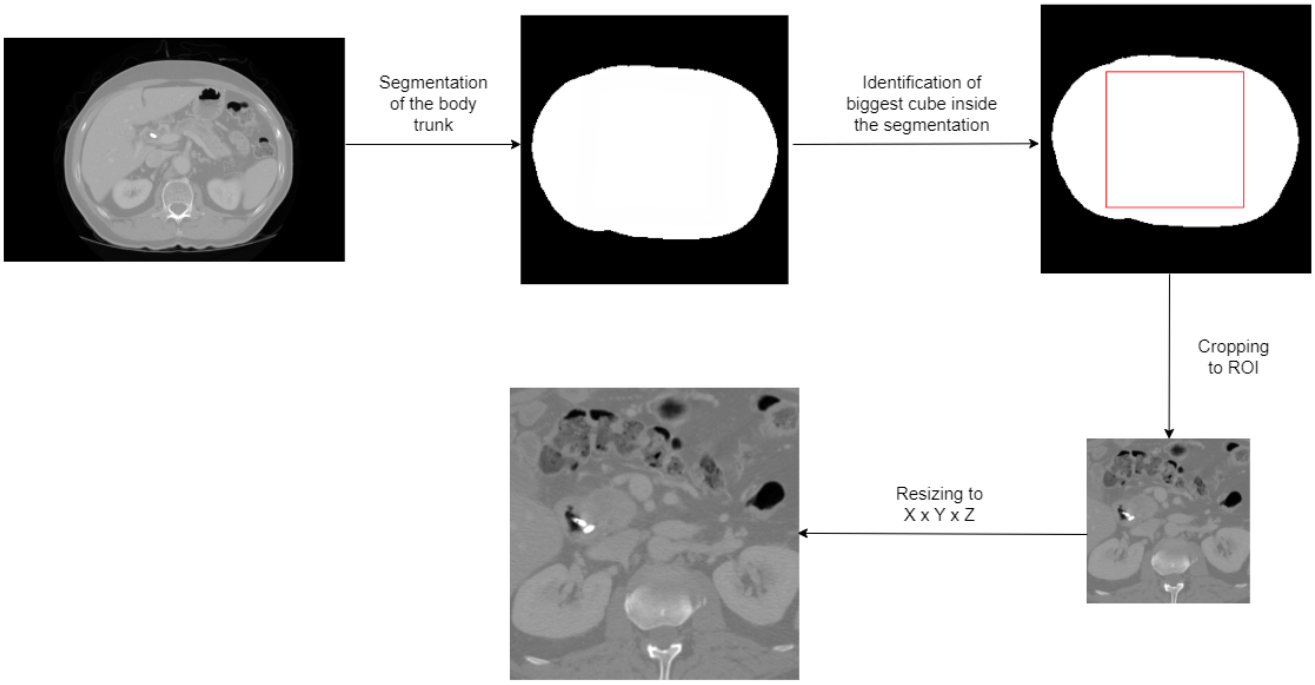


Figure 25: The preprocessing steps of the semi-cropped images.

Visual inspection

In Figure 26 the output of the model is shown. It can be seen that the model captures the basic structures of the scans reasonably well. The quality of the coronal plane (middle) seems to be the worst, whereas the axial plane (left) starts to resemble the real semi-cropped axial slice shown in the first column and third row of Figure 20. This might be because the axial plane has to most context compared to the coronal plane.

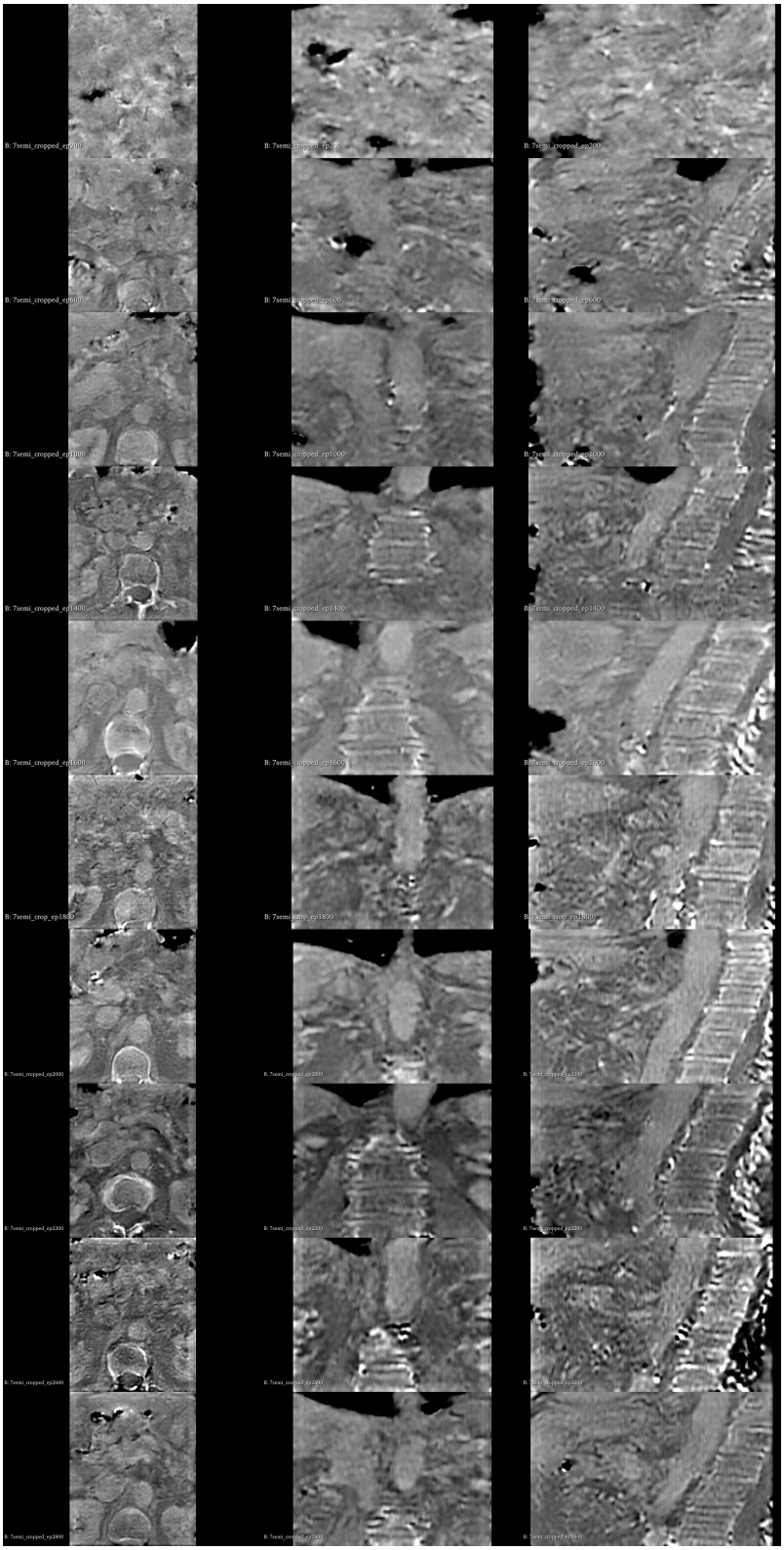
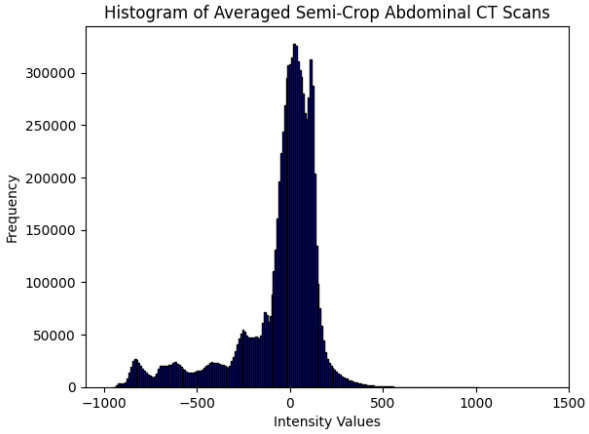


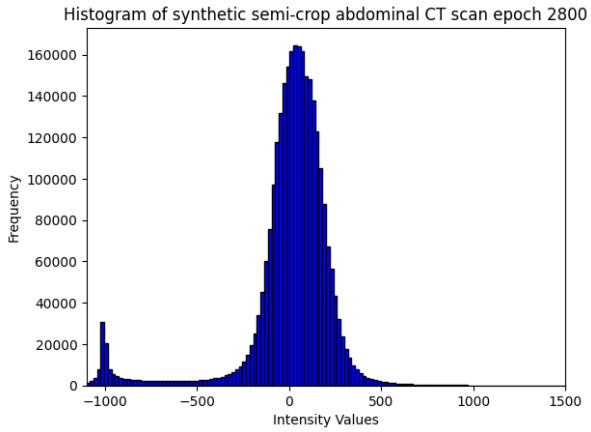
Figure 26: Progression of the results of the model on the semi-body crop. Every column stands for the three different planes: axial, coronal and sagittal respectively. Every row corresponds to a certain epoch: 200, 600, 1000, 1400, 1600, 1800, 2000, 2200, 2600, 2800 respectively.

Grey-scale histogram

In Figure 27 we can see the two different distributions of grey-scale pixels. Because both the real and synthetic scans do not have a black background so the peak around the intensity of -1000 is less present in both Figure 27a and Figure 27b. The synthetic sample fails to capture the different grey-scales or tissues at around an intensity of 0-500, especially the bone tissue. This has an intensity around 400, which can be seen in Figure 3. Although the model seems to capture the structures of the bones, as seen in Figure 17. Therefore, while the model demonstrates proficiency in capturing structural details, it lacks accuracy in replicating the full range of grey-scale intensities.



(a) The grey-scale histogram of a real semi-cropped CT scan



(b) The grey-scale histogram of a synthetic semi-cropped CT scan after 2800 epochs of training.

Figure 27: Comparison of semi-cropped histograms

TotalSegmentator

By comparing the segmentation of real and synthetic scans of different cropping sizes, we can evaluate the extent to which the latent diffusion model can generate anatomically accurate pancreas structures in synthetic CT scans. As seen in Figure 29, we can see that using the semi-crop of the abdominal scans, the TotalSegmentator is able to segment something. Similarly to the segmentation of the full-body scan in Section 4.2, the shape of the segmentation masks of the three planes is very similar. Also, the position of the segmentation masks of all the planes overlap with the real segmentation masks in Figure 28. These results show that the model is still able to generate pancreatic tissue that is recognizable for the TotalSegmentator, even though the input has a smaller cropping size than what the TotalSegmentator is trained on, which is full-body abdominal scans.

The role of surrounding context to generate synthetic samples that correctly contain the pancreas is crucial yet not overly restrictive. While the cropped pancreas images were noisy and did not contain any structure, the semi-crop images were successful, the images contained a pancreas that was segmentable by the TotalSegmentator. This is likely because they retained enough contextual information. Similarly, the full-body images were successful in containing a segmentable pancreas by the TotalSegmentator, emphasizing the importance of sufficient surrounding context

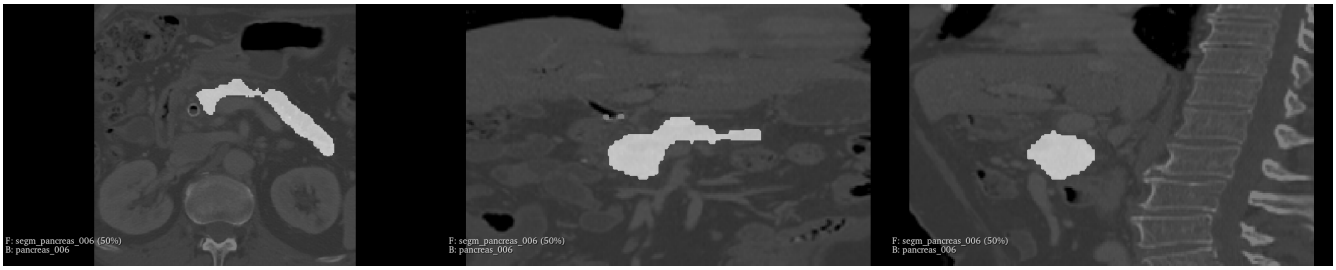


Figure 28: Segmentation of the real semi-crop abdominal scan using the TotalSegmentator.

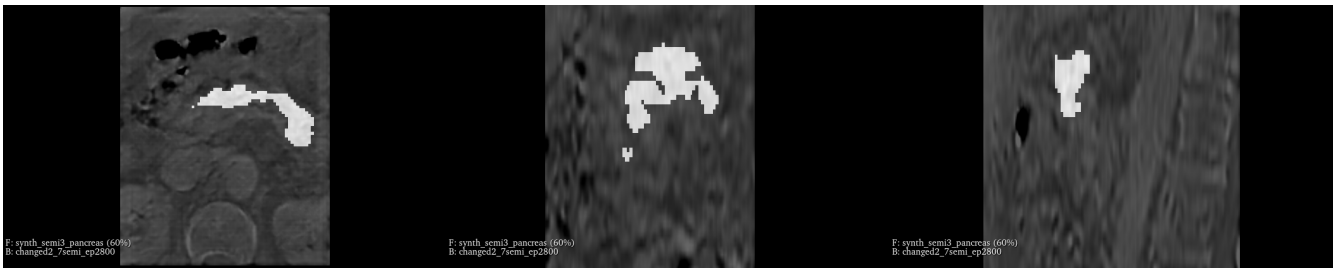


Figure 29: Segmentation of the synthetic semi-crop abdominal scan using the TotalSegmentator.

in generating accurate synthetic samples. The distribution of grey-scale pixels of the semi-crop images were a bit similar but not as similar as that of the distribution of the full-body. This shows that more context does result in a more similar grey-scale pixel distribution to that of a real CT scan. Next to that, we have seen that implementing different learning schedules (milestones, starting learning rate and the gamma parameter) on the cropped pancreas model does not result in better results. What we did found is that choosing dimensions of the training data that are closer to the input dimensions of the autoencoder, the output of the model is less zoomed-in.

To address the main research question, it is feasible for an existing latent diffusion model, originally designed for MRI scans, to accurately generate the pancreas in synthetic CT scans. However, the model's performance in generating accurate results diminishes when the input images lack sufficient contextual information or when the cropping size is too small, as seen with pancreas-cropped images. Using full-body and semi-cropped full-body images for training provides the necessary context to produce synthetic CT scans with accurate representations of the pancreas.

5 Discussion and future work

Exploring different training techniques for generating synthetic CT images can significantly impact the quality of the results. Investigating the impact of training full-body CT images with smaller dimensions may validate and refine the quality of the output by addressing excessive zooming in synthetic results. This method aims to potentially enhance the model's ability to produce more contextually accurate and detailed images, better resembling real full-body CT scans.

By training the autoencoder from scratch using a dataset that includes medical imaging scans, we can tailor the model specifically to the characteristics of medical data, potentially improving its performance and accuracy in generating synthetic scans. However, this approach can be resource-intensive, and limited computing resources may constrain the ability to fully leverage this potential. With more computing power, we could enhance the training process of the autoencoder, potentially leading to even more refined and accurate synthetic scan generation.

Evaluation metrics for synthetic data are also problematic, for instance, the FID score is inconsistent and fails to reflect improvements or capture distortion accurately [JRV+24]. Re-evaluating the FID score using an Inception-V3 network specifically trained on 3D medical images could offer a more reliable quantitative measure for synthetic data quality. Given that traditional FID metrics have limitations in assessing 3D medical imagery, adapting it for 3D contexts may provide a more accurate evaluation of synthetic image fidelity and realism. Expanding the application of the TotalSegmentator model to include the segmentation of additional basic organs, beyond the pancreas, will facilitate a more comprehensive evaluation of the quality of synthetic samples. This will enable a comprehensive evaluation of whether these synthetic scans can accurately represent a full-body scan and not just isolated organs.

We have discussed some problems synthetic data can solve, but it has its weaknesses. Firstly, generative models often lack explainability and may not handle corner cases or extreme outliers well, which can limit their trustworthiness and adoption, especially in critical fields [YZL+23]. Furthermore, generative models, being probabilistic, can produce misleading outputs and biases in training data can perpetuate societal biases [SBAG23, EWF22]. In medical contexts, synthetic data helps when real data cannot be shared, but risks of privacy leakage persist since the output mirrors the original data distribution. Additionally, healthcare systems lag in cybersecurity, making patient data vulnerable to attacks [CP21]. Mirksy et al. showed that synthetic data can be tampered with by malicious actors, affecting diagnoses [MMSE19], but recent advances offer solutions for detecting such manipulations [ZCC+24]. Despite these challenges, synthetic data remains a valuable tool if used with caution and proper safeguards. It holds potential for advancing the field by improving the accuracy, generalizability of synthetic medical images, thereby enhancing their utility in clinical and research settings. By generating high-quality synthetic medical images, these approaches can provide valuable supplementary data for training and validating models, addressing the challenge of limited access to diverse and comprehensive medical datasets.

6 Conclusion

To conclude, our research demonstrates that an existing latent diffusion model designed for MRI scans can be effectively adapted to generate accurate representations of the pancreas in synthetic CT scans. We found that the FID score is inadequate for evaluating CT images, while grey-scale histograms and the TotalSegmentator model offer more reliable metrics for assessing synthetic scan quality. Additionally, while different learning schedules did not enhance the results of the cropped pancreas scans, adjusting training data dimensions to match the autoencoder's input size improved the output for all cropping sizes by reducing image zooming. Our study also revealed that the accuracy of generated images is significantly affected by the amount of contextual information; larger cropping sizes and comprehensive image contexts, such as those provided by full-body and semi-cropped images, enhance the model's ability to produce accurate pancreatic representations. Conversely, the quality decreased with smaller cropping sizes or insufficient contextual information, as seen with pancreas-cropped images.

References

[AAJM+20] Mohamed Alloghani, Dhiya Al-Jumeily, Jamila Mustafina, Abir Hussain, and Ahmed J Aljaaf. A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and unsupervised learning for data science*, pages 3–21, 2020.

[AG24] Dilsat Berin Aytar and Semra Gunduc. Generation of synthetic data using breast cancer dataset and classification with resnet18. *arXiv preprint arXiv:2405.16286*, 2024.

[ARB+22] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[BESS+24] Staphord Bengesi, Hoda El-Sayed, Md Kamruzzaman Sarker, Yao Houkpati, John Irungu, and Timothy Oladunni. Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers. *IEEE Access*, 2024.

[BKG23] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374, 2023.

[BS19] Rohit Babbar and Bernhard Schölkopf. Data scarcity, robustness and extreme multi-label classification. *Machine Learning*, 108(8):1329–1351, 2019.

[CCC20] Ching-Hsue Cheng, Hsien-Hsiu Chen, and Tai-Liang Chen. A clinical decision-support system based on three-stage integrated image analysis for diagnosing lung disease. *Symmetry*, 12(3):386, 2020.

[CP21] Julie Anne Chua and C Pmp. Cybersecurity in the healthcare industry. *Physician Leadership Journal*, 8(1), 2021.

[CWD+18] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[DB16] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

[DOX22] Zolnamar Dorjsembe, Sodtavilan Odonchimed, and Furen Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical Imaging with Deep Learning*, 2022.

[DS19] Tami D DenOtter and Johanna Schubert. Hounsfield unit. 2019.

[EJ16] Aarthipoornima Elangovan and Thangaraja Jeyaseelan. Medical imaging modalities: a survey. In *2016 International Conference on emerging trends in engineering, technology and science (ICETETS)*, pages 1–4. ieee, 2016.

[ERO21] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[EWF22] Lauren Eskreis-Winkler and Ayelet Fishbach. You think failure is hard? so is learning from it. *Perspectives on Psychological Science*, 17(6):1511–1524, 2022.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[HRU+17] 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.

[IJK+21] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[IZZE17] 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.

[JLO21] Abdul Jabbar, Xi Li, and Bourahla Omar. A survey on generative adversarial networks: Variants, applications, and training. *ACM Computing Surveys (CSUR)*, 54(8):1–49, 2021.

[JRV+24] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024.

[KAAL22] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KHK19] Gihyun Kwon, Chihye Han, and Dae-shik Kim. Generation of 3d brain mri using auto-encoding generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 118–126. Springer, 2019.

[KMJRW14] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

[KN+20] Vandana Kushwaha, GC Nandi, et al. Study of prevention of mode collapse in generative adversarial network (gan). In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, pages 1–6. IEEE, 2020.

[KPV13] Ron Kikinis, Steve D Pieper, and Kirby G Vosburgh. 3d slicer: a platform for subject-specific image analysis, visualization, and clinical support. In *Intraoperative imaging and image-guided therapy*, pages 277–289. Springer, 2013.

[KW+19] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[KWIT16] Terumi Kamisawa, Laura D Wood, Takao Itoi, and Kyoichi Takaori. Pancreatic cancer. *The Lancet*, 388(10039):73–85, 2016.

[KXS+23] Murat Kuzlu, Zhenxin Xiao, Salih Sarp, Ferhat Ozgur Catak, Necip Gurler, and Ozgur Guler. The rise of generative artificial intelligence in healthcare. In *2023 12th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–4. IEEE, 2023.

[KY22] Boah Kim and Jong Chul Ye. Diffusion deformable model for 4d temporal medical image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 539–548. Springer, 2022.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[LL10] Po Sing Leung and Po Sing Leung. Overview of the pancreas. *The Renin-Angiotensin System: Current Research Progress in The Pancreas: The RAS in the Pancreas*, pages 3–12, 2010.

[LLY+21] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.

[LMA+16] Xiangrui Li, Paul S Morgan, John Ashburner, Jolinda Smith, and Christopher Rorden. The first step for neuroimaging data analysis: Dicom to nifti conversion. *Journal of neuroscience methods*, 264:47–56, 2016.

[Loh18] Erwin Loh. Medicine and the rise of the robots: a qualitative review of recent advances of artificial intelligence in health. *BMJ leader*, pages leader–2018, 2018.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[LST+16] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6(1):26286, 2016.

[LSW+23] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, and Wenqi Wei. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*, 2023.

[MMSE19] Yisroel Mirsky, Tom Mahler, Ilan Shelef, and Yuval Elovici. {CT-GAN}: Malicious tampering of 3d medical imagery using deep learning. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 461–478, 2019.

[Mos01] William W Moses. Trends in pet imaging. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 471(1-2):209–214, 2001.

[ND21] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[Ng16] Andrew Ng. What artificial intelligence can and can't do right now. *Harvard Business Review*, 9(11):1–4, 2016.

[PDJ21] Paula R Patel and Orlando De Jesus. Ct scan. 2021.

[PGK+23] WH Pinaya, MS Graham, E Kerfoot, PD Tudosiu, J Dafflon, V Fernandez, P Sanchez, J Wolleb, PF da Costa, A Patel, et al. Generative ai for medical imaging: extending the monai framework. preprint (2023). *arXiv preprint arXiv:2307.15208*, 2023.

[PLW02] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[PON+20] Stephen P Pereira, Lucy Oldfield, Alexander Ney, Phil A Hart, Margaret G Keane, Stephen J Pandol, Debiao Li, William Greenhalf, Christie Y Jeon, Eugene J Koay, et al. Early detection of pancreatic cancer. *The lancet Gastroenterology & hepatology*, 5(7):698–710, 2020.

[PWH21] Ahmad Pesaranghader, Yiping Wang, and Mohammad Havaei. Ct-sgan: computed tomography synthesis gan. In *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*, pages 81–93. Springer, 2021.

[PYG+23] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T Barron, Amit H Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. State of the art on diffusion models for visual computing. *arXiv preprint arXiv:2310.07204*, 2023.

[RBL+22] 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.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[SBAG23] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. Ai model gpt-3 (dis) informs us better than humans. *Science Advances*, 9(26):eadh1850, 2023.

[SC21] Divya Saxena and Jiannong Cao. Generative adversarial networks (gans) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, 54(3):1–42, 2021.

[SKF21] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8(1):101, 2021.

[SMZJ14] Rebecca Siegel, Jiemin Ma, Zhaohui Zou, and Ahmedin Jemal. Cancer statistics, 2014. *CA: a cancer journal for clinicians*, 64(1), 2014.

[SS18] Pramila P Shinde and Seema Shah. A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, pages 1–6. IEEE, 2018.

[STB+23] Yu Shi, Hannah Tang, Michael J Baine, Michael A Hollingsworth, Huijing Du, Dandan Zheng, Chi Zhang, and Hongfeng Yu. 3dgaunet: 3d generative adversarial networks with a 3d u-net based generator to achieve the accurate and effective synthesis of clinical tumor image data for pancreatic cancer. *Cancers*, 15(23):5496, 2023.

[WBM+23] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023.

[Wer90] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[WGD+17] Kunfeng Wang, Chao Gou, Yanjie Duan, Yilun Lin, Xinhu Zheng, and Fei-Yue Wang. Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, 4(4):588–598, 2017.

[XSH21] Shibo Xing, Harsh Sinha, and Seong Jae Hwang. Cycle consistent embedding of 3d brains with auto-encoding generative adversarial networks. In *Medical Imaging with Deep Learning*, 2021.

[YNDT18] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629, 2018.

[YZD21] Yu Yu, Weibin Zhang, and Yun Deng. Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, 2021.

[YZL+23] Zuhao Yang, Fangneng Zhan, Kunhao Liu, Muyu Xu, and Shijian Lu. Ai-generated images as data source: The dawn of synthetic era. *arXiv preprint arXiv:2310.01830*, 2023.

[YZS+23] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[ZCC+24] Giada Zingarini, Davide Cozzolino, Riccardo Corvi, Giovanni Poggi, and Luisa Verdoliva. M3dsynth: A dataset of medical 3d images with ai-generated local manipulations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13176–13180. IEEE, 2024.

[ZSZZ17] Jiachi Zhang, Xiaolei Shen, Tianqi Zhuo, and Hong Zhou. Brain tumor segmentation based on refined fully convolutional neural networks with a hierarchical dice loss. *arXiv preprint arXiv:1712.09093*, 2017.