Universiteit
Leiden
The Netherlands

Bachelor Computer Science

Synthetic Data Generation for Lung Tumor Detection:
A Comparison of StyleGAN3 and Latent Diffusion

Soham Chatterjee

Supervisors:
Prof. dr. K.J. Batenburg
Dr. D.M. Pelt

BACHELOR THESIS

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

01/07/2025

Abstract

This thesis investigates the generation of synthetic data for lung tumor detection by comparing StyleGAN3 and Latent Diffusion. The research addresses the challenge of dataset limitations in medical imaging, particularly for rare or atypical tumor types. Using the LIDC-IDRI dataset, I implement a custom GAN architecture, a StyleGAN3-based pipeline, and Latent Diffusion for generating synthetic CT images containing lung tumors. My comparison examines the image quality, training stability, and computational efficiency of these generative approaches. The study aims to provide insights into the strengths and limitations of each method for medical image synthesis. This work contributes to the field of synthetic medical data generation by offering an empirical evaluation framework for generative models in clinical AI applications.

Acknowledgements

I would first like to express my gratitude to Prof. dr. K.J. Batenburg for supervising this thesis and providing valuable guidance and feedback throughout the research process.
I would also like to acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.
For the Latent Diffusion training, this work was performed using the compute resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University.

Contents

1 Introduction

This chapter introduces the field of synthetic data generation for lung tumor detection in medical imaging, outlines the research questions guiding this thesis, and presents its key contributions. The chapter describes the clinical background and motivation of this work, followed by a clear problem statement. Next, it presents the research questions addressed in this comparative study of Generative Adversarial Networks (GANs) and diffusion models. The chapter concludes with a summary of the thesis contributions and an overview of the document structure.

1.1 Background and Motivation

Lung cancer remains the leading cause of cancer-related deaths worldwide, with approximately 1.8 million deaths reported in 2022 [ZXL+24]. Early-stage detection significantly improves survival rates, but prognosis worsens rapidly as the disease progresses [HLC+22]. In particular, stage 4 lung cancer continues to have one of the lowest survival rates in countries such as England [NHS23].
In response, researchers have increasingly applied artificial intelligence (AI) and deep learning to assist in the early detection of lung cancer. These approaches have demonstrated strong performance, achieving detection accuracies exceeding 90% in numerous studies [RA20, SDK+22, LMS+19]. However, performance tends to drop when dealing with rare or atypical tumor types that are underrepresented in existing datasets. This imbalance introduces bias and limits generalization [AAP+23]. One way to address this challenge is to supplement datasets with additional high-quality, diverse training samples. Unfortunately, collecting medical imaging data is difficult due to privacy regulations, the need for expert annotation, and limited access to clinical resources.
This has led to growing interest in the generation of synthetic data using generative models. The field gained momentum with the introduction of GANs in 2014 [GPAM+14], quickly becoming the standard for image synthesis, including in medical domains. GANs have been used to generate high-resolution CT scans that can augment training datasets and improve model robustness [MPS+23, PZM+24].
More recently, diffusion models have emerged as a powerful alternative. These models generate images through a gradual denoising process, which leads to greater stability during training and improved output diversity [HJA20]. Their ability to capture complex image distributions has made them particularly effective for producing realistic, high-fidelity images. Although their use in medical imaging is still in development, early work, such as Lung-DDPM, has demonstrated the potential of diffusion-based methods to outperform GANs in image quality and performance on downstream tasks, like nodule segmentation [JLB+25].

1.2 Problem Statement

Despite recent advances in deep learning and synthetic medical image generation, a significant challenge remains: effectively generating high-quality, diverse CT images. Current state-of-the-art (SOTA) generative methods, particularly GANs and emerging diffusion models, offer promising results. Still, their comparative strengths and weaknesses in the medical domain are poorly understood. In particular, there is a lack of systematic evaluation regarding the image realism and computational trade-offs of these methods. This thesis addresses that gap by empirically comparing

GAN-based and diffusion-based approaches for synthetic lung tumor data generation to inform future model selection in clinical AI pipelines.

1.3 Research Questions

This thesis aims to explore the effectiveness of synthetic data generation techniques in the context of lung tumor detection. The main research questions are as follows:

1. How do SOTA GAN-based and diffusion-based models compare in generating realistic and diverse CT images of lung tumors?
2. What are the computational efficiency and training stability trade-offs between the two generative approaches?

1.4 Contributions

The key contributions of this thesis are:

- A custom implementation and experimental analysis of a GAN-based model architecture tailored for lung CT image generation.
- A StyleGAN3-based pipeline for generating synthetic lung tumor images.
- A Latent Diffusion-based pipeline for generating synthetic lung tumor images.
- A comparative evaluation of the two approaches based on visual quality and computational cost.
- A discussion of challenges encountered during development and training.

1.5 Thesis Overview

This chapter has introduced the motivation and scope of this work. Chapter 2 provides a more detailed background on lung cancer detection, medical image synthesis, GANs, and diffusion models. Chapter 3 describes the datasets used and the preprocessing approaches. Chapter 4 presents the custom model and the SOTA models used and provides an evaluation framework. Chapter 5 reports the results of various experiments, including qualitative and quantitative comparisons of generated data and the computational efficiency of the models. Chapter 6 analyzes the results, limitations, and trade-offs of the two approaches. Chapter 7 summarizes the key findings and outlines directions for further research.
This bachelor thesis has been written at the Leiden Institute of Advanced Computer Science (LIACS) under the supervision of Prof. dr. K.J. Batenburg.

2 Key Concepts and Definitions

This section outlines the key concepts and definitions necessary for this thesis. It introduces the clinical context of lung cancer detection, the role of medical image synthesis, and the generative models employed for synthetic data generation. It also discusses the motivation and evaluation strategies for using synthetic data in medical imaging.

2.1 Lung Cancer Detection: Challenges and Approaches

While early-stage lung cancer disease is potentially curable, the majority of cases are detected too late for effective intervention [SMTK+25]. Low-dose computed tomography (LDCT) has emerged as an effective screening method, demonstrated by trials like NLST and NELSON, which showed 20–33% reductions in lung cancer mortality [MCFR21]. However, challenges such as high false-positive rates, inconsistent screening protocols, radiologist shortages, and the psychological burden on patients have limited widespread adoption. Screening programs must also navigate complexities around selecting high-risk individuals, managing indeterminate nodules, and minimizing overdiagnosis.
The increasing demand for radiologist expertise and high-volume screening interpretation has driven interest in AI-based detection systems. Studies show that deep learning models trained on LDCT can match or exceed radiologist performance, offering a potential solution to scalability and consistency issues in large-scale screening initiatives [MCFR21].

2.2 Acquisition of CT Scans

Computed Tomography (CT) acquires cross-sectional images by rotating an X-ray source and detector array around the patient, measuring the attenuation of X-rays through the body from multiple angles [Oht]. At each projection angle θ, the detector measures the transmitted intensity I after the X-ray beam traverses the object. According to the Beer–Lambert law:

$$I(\theta, t) \;=\; I_0 \exp\Big(-\int_{L(\theta,t)} \mu(x,y)\, ds\Big),$$

where I_0 is the incident intensity, $\mu(x, y)$ is the position-dependent linear attenuation coefficient, and the integral is taken along the ray path $L(\theta, t)$ indexed by detector position t. Equivalently, the measured projection data (sinogram) $p(\theta, t)$ can be written as

$$p(\theta, t) \;=\; -\ln\big[I(\theta, t)/I_0\big] \;=\; \int_{L(\theta,t)} \mu(x,y)\, ds,$$

which represents a line integral of $\mu(x, y)$ through the object at angle θ.
While the above describes the attenuation and reconstruction of a single cross-section in the (x, y) plane, clinical CT scanners collect a series of such slices along the z-axis. In a step-and-shoot system, the patient table is incremented between rotations so that each z-position yields its own sinogram $p(\theta, t)$. In modern helical (spiral) CT, the X-ray source and detector rotate continuously as the table moves, acquiring projections that cover an oblique path through (x, y, z). By reconstructing each slice, either independently or via a full 3D Radon inversion, the continuous attenuation map $\mu(x, y, z)$ is recovered over the entire volume.

Reconstruction algorithms (e.g., filtered backprojection or iterative reconstruction) invert this Radon transform to recover the continuous attenuation map $\mu(x, y)$ (and, when multiple slices are acquired, $\mu(x, y, z)$) throughout the scanned volume [Oht].
Once the continuous attenuation distribution $\mu(x, y, z)$ has been recovered, it is discretized into a regular grid of cubic (or nearly cubic) elements called voxels. Each voxel is assigned a single scalar value corresponding to the average attenuation coefficient within that small volume. In clinical CT, these attenuation values are typically converted to Hounsfield Units (HU) by

$$\mathrm{HU} \;=\; 1000 \times \frac{\mu_{\mathrm{voxel}} - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}}},$$

where μ_{water} is the attenuation coefficient of water. The final CT dataset, therefore, consists of a 3D array of voxels, each storing an HU value. Voxel dimensions (in-plane pixel size and slice thickness) define the spatial resolution and are determined by the detector geometry and acquisition parameters.
By discretizing the continuous attenuation field into a 3D voxel grid, CT data becomes a collection of scalar values on which all subsequent image processing, machine learning, or synthesis algorithms operate. In other words, any generative model or analysis pipeline in medical imaging takes as input a volumetric array of HU (or equivalent attenuation values), where each voxel is treated as an independent pixel with known spatial coordinates and intensity. This voxel-based representation is fundamental for modern image-based algorithms, as it defines both the data dimensionality and the physical meaning of each pixel value.

2.3 Medical Image Synthesis

Medical image synthesis is the computational generation of artificial medical images that mimic real clinical data. This field has evolved from traditional image processing to deep learning approaches, addressing data limitations in medical research and clinical practice.
Early approaches relied on statistical models and transformation algorithms. These methods included texture synthesis, registration-based techniques, and physics-based simulations [CT]. While useful for specific applications, they often failed to capture complex anatomical structures and pathological variations [SBD+].
Initial deep learning methods included autoencoders and variational autoencoders (VAEs), which learned compact representations of medical images. More advanced models, such as GANs and diffusion models, now demonstrate superior capabilities in generating high-fidelity images that preserve anatomical coherence [YWB].
Medical image synthesis typically involves several approaches [GSVV24]. Random-to-image generation creates novel medical images from noise. Image-to-image translation converts between imaging modalities, such as MRI to CT. Condition-to-image synthesis creates images based on specific parameters. Multimodal synthesis generates consistent images across multiple modalities. These methods enable applications across various medical imaging domains, including CT, MRI, X-ray, and ultrasound. The generative models described in subsequent sections represent current state-of-the-art techniques for medical image synthesis.

2.4 Generative Adversarial Networks

GANs, introduced by Goodfellow et al. [GPAM+14], are a class of generative models that consist of two neural networks: a generator G and a discriminator D. The generator maps random noise $z \sim p_z(z)$ to the data space to produce synthetic samples $G(z)$. The discriminator receives both real samples $x \sim p_{\text{data}}(x)$ and generated samples and outputs a probability $D(x) \in [0, 1]$ indicating whether the input is real or fake. Figure 1 illustrates the typical training workflow of a GAN.
The training objective of a GAN is formulated as a two-player minimax game:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))].$$

Here, the discriminator is trained to maximize its ability to distinguish real from fake, while the generator is trained to minimize this objective by producing samples that fool the discriminator [GPAM+14].
In practice, the generator is often optimized using a non-saturating heuristic loss to improve gradient flow:

$$\min_G -\mathbb{E}_{z \sim p_z(z)}[\log D(G(z))].$$

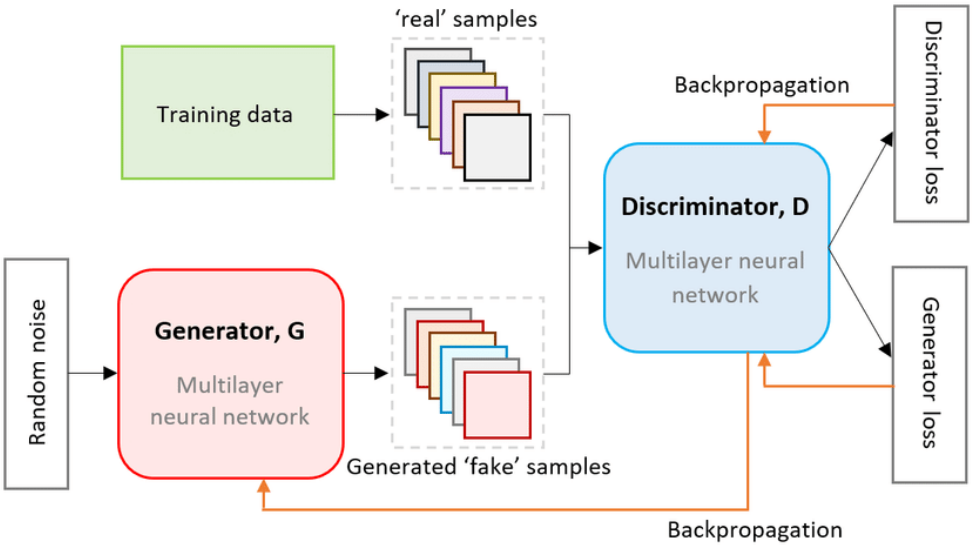

Figure 1: Overview of a basic GAN architecture. The generator learns to produce synthetic samples from random noise, while the discriminator attempts to distinguish real samples from generated ones. Both networks are trained simultaneously in a minimax game. Image reprinted from Little et al. [LEASS21].

2.5 Diffusion Models

Diffusion models, initially proposed by Sohl-Dickstein et al. [SDWMG15], and later refined by Ho et al. [HJA20], are a class of generative models that learn to synthesize data by reversing a gradual noising process. Figure 2 illustrates the typical training workflow of a diffusion model. Unlike GANs,

which rely on adversarial training, diffusion models use a likelihood-based approach that is more stable and easier to train.
The forward process, known as the diffusion or noise process, gradually adds Gaussian noise to the data over T time steps. This is defined as:

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

where β_t is a variance schedule controlling the amount of noise added at each step.
The model learns to reverse this process by estimating the noise at each step, typically using a neural network $\epsilon_\theta(x_t, t)$. The reverse denoising process is modeled as follows:

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

where μ_θ and Σ_θ are learned parameters. The training objective minimizes a reweighted variational bound, often simplified to:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right].$$

In recent work, diffusion models have outperformed GANs in generating high-resolution, diverse samples [DN21] and are gaining traction in medical imaging for tasks such as CT scan synthesis, tumor simulation, and data anonymization. Their ability to model complex data distributions with stable training dynamics makes them particularly promising in sensitive, high-stakes domains such as healthcare.

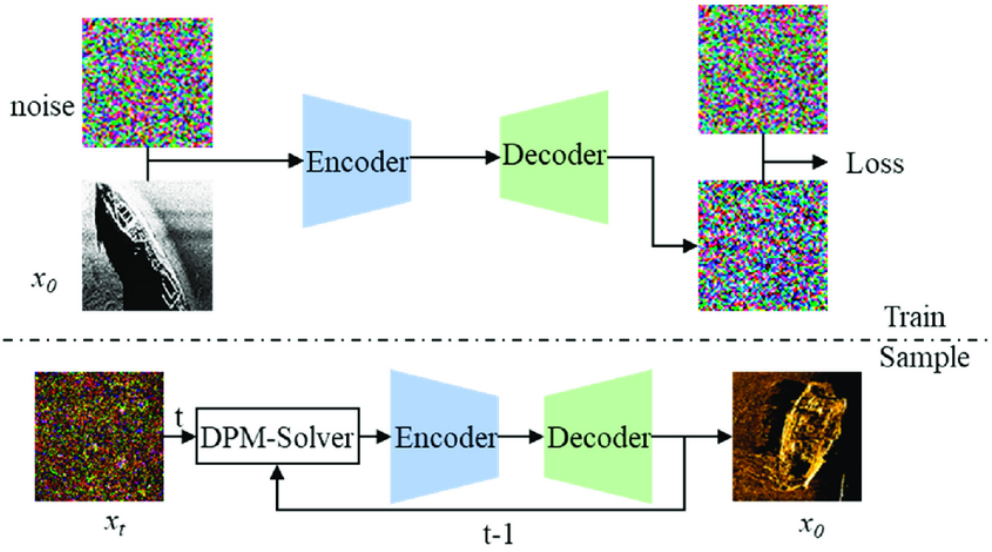

Figure 2: Overview of a diffusion model's forward and reverse processes. Noise is gradually added to the input data during training, and the model learns to denoise and reconstruct the original signal during the sampling process. Image reprinted from Yang et al. [YZZ+23].

2.6 Applications and Evaluation of Synthetic Medical Images

Synthetic data generation offers distinct advantages in medical contexts. Data augmentation through synthetic images can expand limited datasets, helping prevent overfitting and improving

model generalization. In lung cancer detection, synthetic CT scans can represent diverse tumor morphologies. Synthetic data can also model uncommon pathologies or anatomical variants that are difficult to capture in clinical datasets. Furthermore, synthetic images provide alternatives to real patient data, addressing ethical and legal restrictions on data sharing. In lung cancer screening specifically, synthetic samples can correct class imbalances by generating additional examples of underrepresented categories.

The potential applications of synthetic, unannotated CT images are illustrated in Figure 3. These include pretraining or domain adaptation for segmentation models, data augmentation in classification tasks, and radiologist training for rare case exposure. While annotations are not directly generated, these images support multiple downstream workflows where realistic structural variation is beneficial.

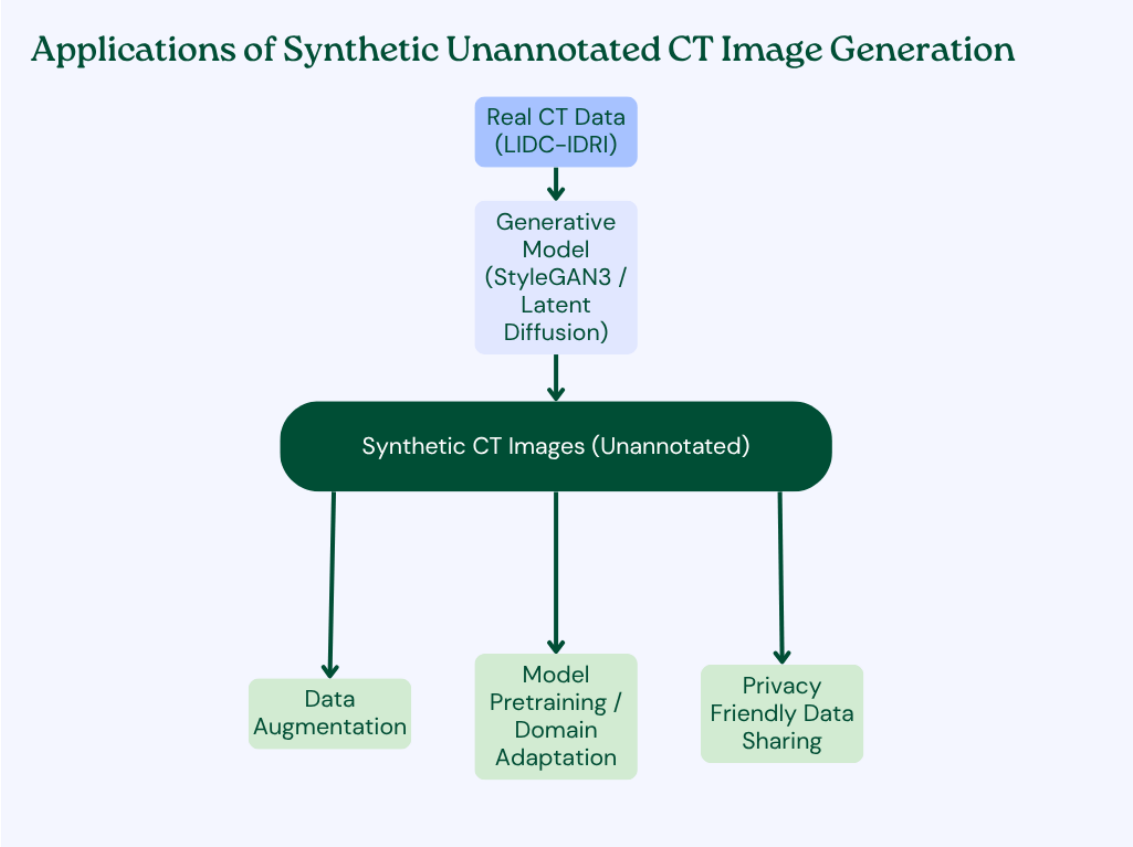

Figure 3: Applications of synthetic, unannotated axial CT images. These include training data augmentation, pretraining or domain adaptation, and synthetic case exposure for radiologist education.

The evaluation of synthetic medical images employs multiple complementary approaches. Visual assessment by radiologists evaluates anatomical accuracy, pathological plausibility, and overall realism. Statistical metrics provide quantitative measures, including Fréchet Inception Distance (FID) and precision/recall.

Despite their utility, synthetic images present limitations that require consideration. Synthetic data may not fully capture the diversity of real-world variations, creating distribution gaps. Generative

models may introduce subtle, unrealistic features or artifacts. The biological plausibility of synthetic images requires ongoing validation by medical experts to confirm clinical validity. In lung cancer detection, synthetic data must strike a balance between realism and utility while avoiding the introduction of misleading patterns that could impact diagnostic accuracy.

3 Data and Preprocessing

This section describes the dataset used in this study and the preprocessing pipeline employed to prepare the data for training the generative model. The focus is on extracting and balancing relevant 2D slices from volumetric CT scans while preserving tumor information.

3.1 Dataset: LIDC-IDRI

The experiments in this thesis use the publicly available LIDC-IDRI dataset, which comprises thoracic CT scans from over 1,000 patients, annotated by multiple radiologists [AIMB+15]. Each scan includes one or more nodules with varying levels of malignancy, annotated with detailed segmentation masks. Some samples of the dataset can be seen in Figure 4.

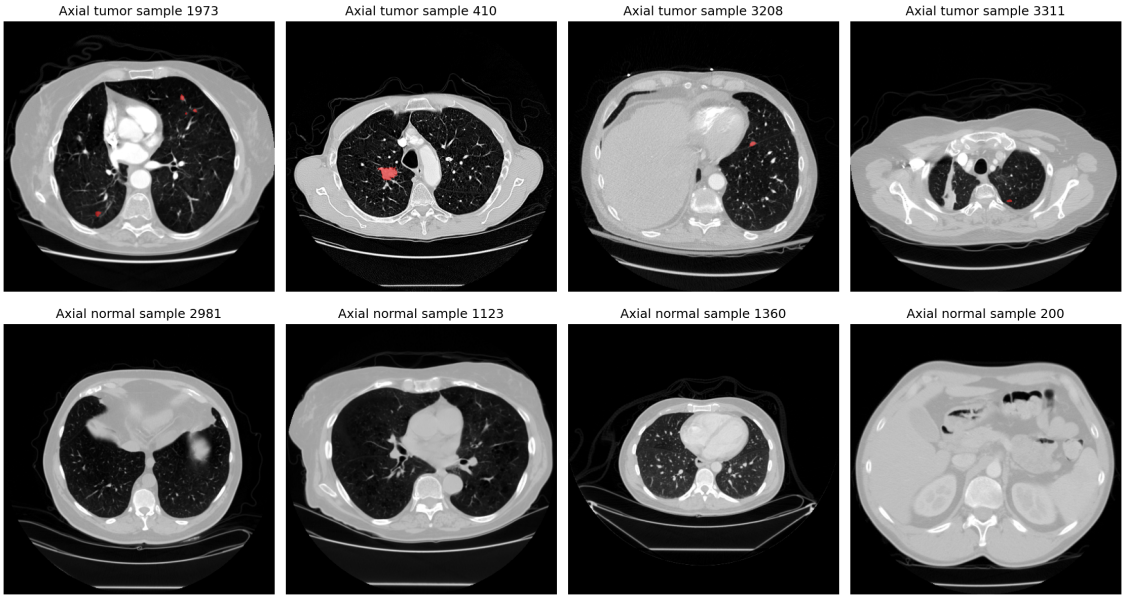

Figure 4: Representative axial CT slices from the LIDC-IDRI dataset. The top row shows tumor-containing slices with annotated nodule regions in red; the bottom row displays normal lung anatomy without nodules. The dataset exhibits high variability in anatomy, tumor size, and location, posing challenges for both detection and synthetic image generation.

3.2 Data Preprocessing Pipeline

The LIDC-IDRI dataset comprises three-dimensional CT volumes that require transformation into two-dimensional slices for training generative models. A consistent preprocessing approach was applied across all experiments.

The preprocessing begins with intensity normalization. CT images represent tissue radiodensity in Hounsfield Units (HUs), with air at -1000 HU, water at 0 HU, and soft tissues ranging from 30 to 80 HU. To maintain anatomical contrast, HU values were clipped to the range $[-1000, 400]$ and linearly scaled to $[0, 1]$. Segmentation masks were constructed by merging all radiologist annotations into a single binary mask per scan using a pixel-wise maximum operation.
While the implementation supports axial, sagittal, and coronal views, this study focuses exclusively on axial slices, as these represent the standard clinical view for assessing lung tumors. The pipeline extracts the image data and tumor mask for each slice.
Slices with low average intensity were filtered out to eliminate non-informative regions, such as those outside the thorax. All images and masks were resized to a standard resolution of 256×256 pixels, using bilinear interpolation for image data and nearest-neighbor interpolation for masks to preserve binary boundaries.
The pipeline implements a mechanism to maintain tumor visibility during resizing. The algorithm enforces a minimum pixel count for small nodules that might become indiscernible after resizing, preserving diagnostic relevance in the processed data.

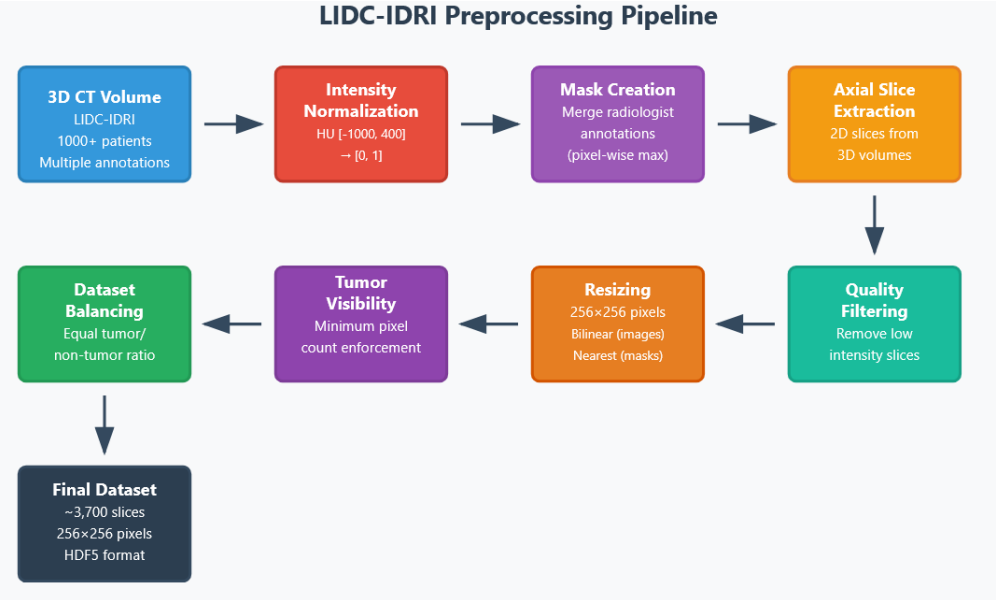

Figure 5: The data processing pipeline, from original LIDC-IDRI 3D medical images to 2D slices in HDF5 format.

3.3 Dataset Balancing

The LIDC-IDRI dataset naturally contains a higher proportion of non-tumor slices. For effective model training, the dataset was balanced to contain an equal ratio of tumor and non-tumor slices. This balancing was implemented through strategic sampling rather than data augmentation, guaranteeing all training samples maintained authentic anatomical structures. After balancing, approximately 3 700 slices were used in the experiments.
All processed data was stored in the HDF5 format, which enables efficient random-access reading during training. Slice-level metadata, including intensity statistics and nodule presence, were

preserved to support analysis and quality assessment. A visualization of this pipeline can be seen in Figure 5.

4 Methodology

This section details the custom implementations and experimental design for comparing GANs and diffusion models in lung tumor image synthesis. All experiments were conducted on axial CT slices from the LIDC-IDRI dataset as described in Section 3.

4.1 Custom GAN Implementation

The initial phase of this research involved developing a custom GAN architecture specifically tailored for CT lung tumor generation. This approach aimed to gain first-hand insights into the challenges of medical image synthesis while establishing a baseline for further comparisons.

4.1.1 Proposed Architecture

The final custom GAN implementation consists of a U-Net based generator and a PatchGAN discriminator, both adapted for the specific requirements of CT image synthesis with tumor annotations. Figure 6 presents an overview of this architecture.
The generator follows a U-Net structure with several modifications. It combines random noise (1 channel) and a one-hot encoded condition (2 channels), indicating whether the output should contain a tumor. The generator produces a two-channel output representing the CT image and the corresponding tumor mask simultaneously. An optional self-attention module at the bottleneck improves the coherence of generated structures. The U-Net structure maintains high-resolution information flow through skip connections between corresponding encoder and decoder layers.
The discriminator uses a PatchGAN architecture with spectral normalization. The network evaluates patches from the image conditioned on the tumor presence label. Spectral normalization stabilizes training by constraining the spectral norm of weight matrices. The number of downsampling layers automatically adapts to the input resolution.
The loss function combines several components. The adversarial loss follows the standard GAN formulation with label smoothing to improve stability. A reconstruction loss using the L1 norm measures the difference between real and generated CT images. Additionally, conditional mask loss applies different terms for images with and without tumors to ensure semantic consistency.

4.1.2 Development Process and Challenges

The development process followed an iterative approach with gradual complexity increases. Initial implementations faced several challenges.
Resolution scaling presented the first obstacle. Early experiments began with 64×64 resolution to reduce training time and memory requirements. However, this proved insufficient for capturing meaningful anatomical details. Successive iterations increased the resolution to 128×128 and ultimately 256×256, which provided better anatomical fidelity at the cost of longer training times and increased memory consumption.

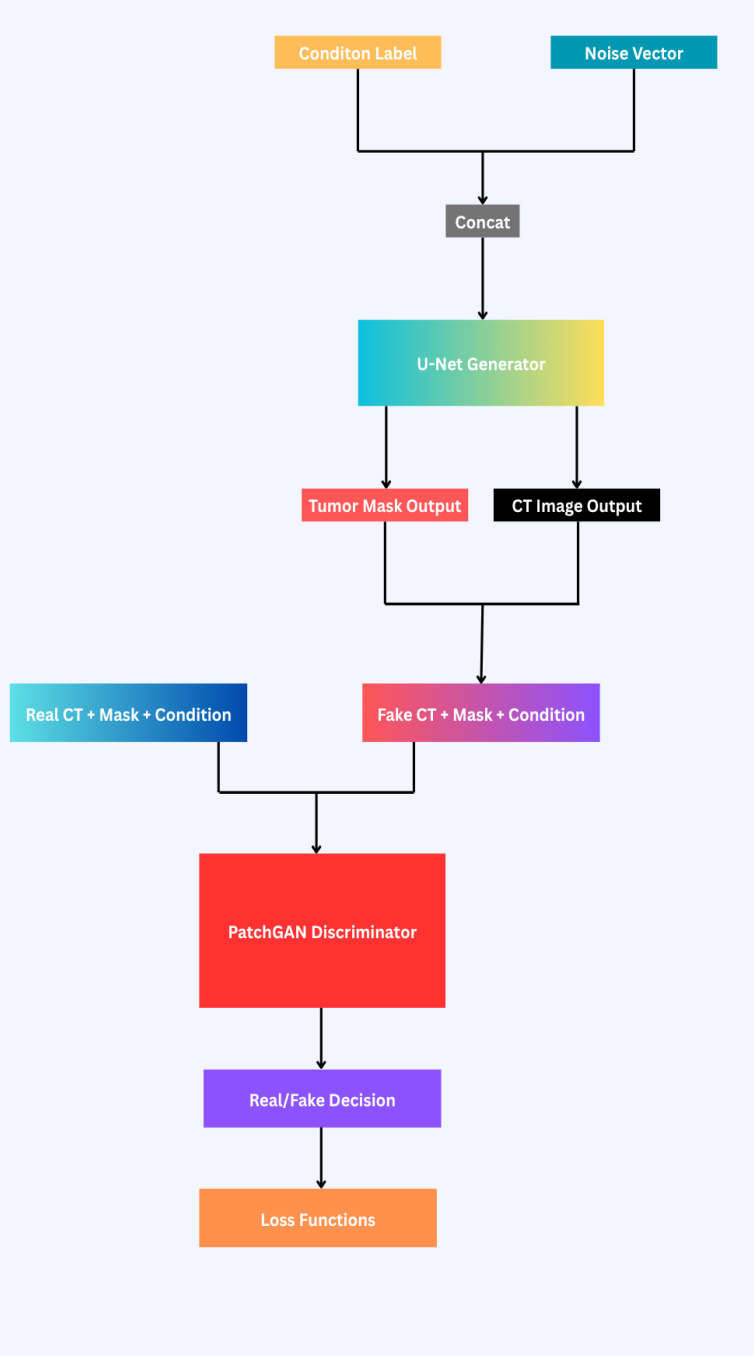

Figure 6: Overview of the custom GAN architecture for synthetic CT image generation. The generator takes a noise vector and condition label (tumor/no-tumor) and produces both a synthetic CT slice and a corresponding tumor mask. These outputs are evaluated by a PatchGAN discriminator, which is trained to distinguish real from fake samples. Loss functions include adversarial, reconstruction, and conditional mask losses.

View selection created a significant limitation in early experiments. The inadvertent use of lower-quality sagittal views due to incorrect indexing when extracting slices from the volumetric CT data resulted in the model training on data with suboptimal anatomical representation. This implementation error was identified by analyzing intermediate outputs and debugging the preprocessing pipeline. After correction, training exclusively on axial views substantially improved the learning process, as these views contain more consistent and diagnostically valuable patterns.

Training stability presented persistent challenges. Like many GAN implementations, the model suffered from instabilities, including mode collapse, where the generator produces limited variations regardless of input diversity. Several techniques were implemented to mitigate these issues, including label smoothing in the discriminator loss function, spectral normalization in discriminator layers, a two-timescale update rule (TTUR) with different learning rates for generator and discriminator, and a gradient penalty to enforce the Lipschitz constraint.

Computational constraints added another layer of complexity. Limited GPU memory (24GB VRAM on RTX 3090) restricted batch sizes for higher resolutions. To address this, mixed precision training was implemented using PyTorch's automatic mixed precision capabilities, allowing larger effective batch sizes and faster training while maintaining numerical stability.

4.1.3 Experimental Iterations

The custom GAN development progressed through several distinct iterations, each addressing specific limitations observed in previous versions.

The initial implementation used a simple DCGAN-like architecture operating on 64×64 resolution images. This version employed a basic generator with transposed convolutions and a conventional discriminator. The model processed CT slices and corresponding tumor masks as a two-channel input but struggled with generating coherent anatomical structures, as shown in Figure 7. This baseline implementation also used suboptimal sagittal view slices due to the preprocessing issue described earlier, further limiting its effectiveness.

The second iteration improved the conditional aspect by implementing explicit class conditioning. This modification allowed the generator to create images with or without tumors based on a binary condition vector. The resolution was increased to 128×128 to capture more anatomical detail, and the discriminator was improved with spectral normalization to enhance training stability. Despite these improvements, the generated images still exhibited anatomical inconsistencies and frequent artifacts at tissue boundaries.

The third iteration represented a significant architectural shift, implementing a Wasserstein GAN with gradient penalty (WGAN-GP) to address mode collapse issues observed in earlier versions. The WGAN framework replaces the traditional GAN minimax game with a different objective based on the Wasserstein distance between real and generated data distributions. Mathematically, the WGAN objective can be expressed as:

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}}[D(x)] - \mathbb{E}_{z \sim p_z}[D(G(z))]$$

where the discriminator D must be 1-Lipschitz continuous. The gradient penalty enforces this constraint by adding a regularization term:

$$\lambda \mathbb{E}_{\hat{x}} \left[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2 \right]$$

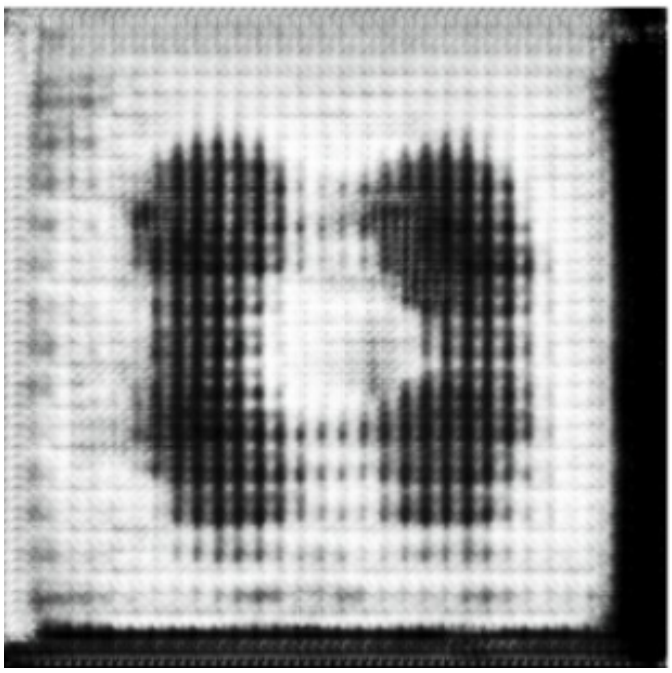

Figure 7: Low-quality outputs from the initial DCGAN implementation at 64×64 resolution, showing poor anatomical definition and artifacts.

where $\hat{x}$ is sampled uniformly along straight lines between pairs of real and generated samples. This approach provided greater training stability and resulted in improved image quality, as demonstrated in Figure 8. This iteration also introduced instance normalization in the generator for better feature normalization and incorporated an L1 reconstruction loss to improve anatomical fidelity.

The final iteration implemented a sophisticated U-Net generator with skip connections and a self-attention mechanism. This architecture preserved high-resolution spatial information through the encoding-decoding process, substantially improving detail preservation. Additional improvements included mixed precision training to handle the computational demands of higher resolution, a PatchGAN discriminator for better local structure assessment, and a composite loss function with separate terms for CT image and tumor mask generation. Training now used correctly preprocessed axial slices at 256×256 resolution, which provided much clearer anatomical structures for the

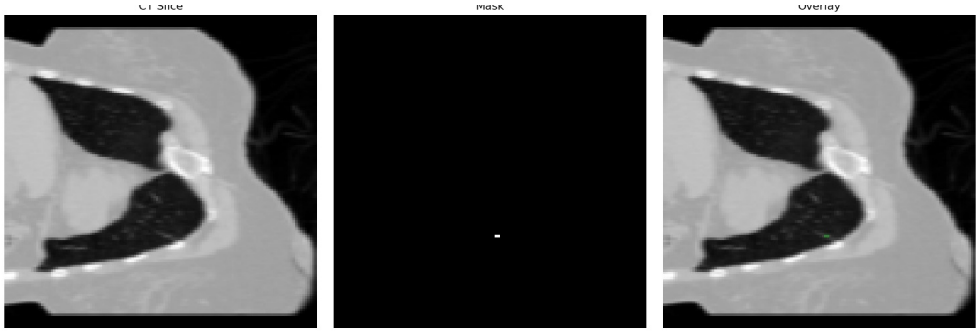

Figure 8: Medium-quality outputs from the WGAN-GP implementation showing improved anatomical structures but still lacking fine details.

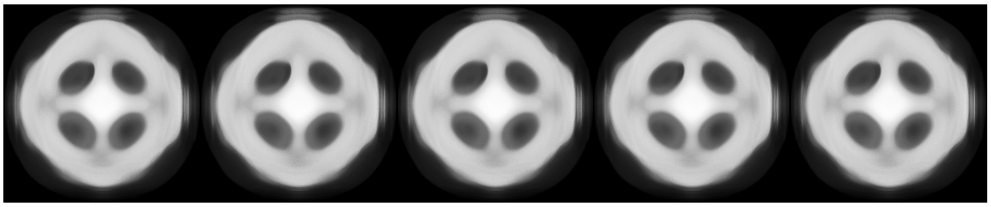

Figure 9: Evidence of model collapse in the final U-Net GAN implementation, where generated samples exhibit minimal variation despite different input noise vectors.

model to learn. This version trained on approximately 2 000 axial slices for 800 epochs at 256×256 resolution with a batch size of 8, using the Adam optimizer with different learning rates for the generator (1e-4) and discriminator (2e-4) to balance training dynamics.

4.1.4 Failure Analysis and Lessons Learned

Despite the progressive improvements, the custom GAN implementation ultimately encountered persistent limitations.

Model collapse emerged as a recurring issue, an example of which can be seen in Figure 9. After extended training (>500 epochs), the model exhibited signs of collapse, generating increasingly similar outputs regardless of input variation. Analysis revealed that the discriminator became too powerful, causing the generator gradient updates to become unstable. While techniques like spectral normalization and gradient penalties delayed this issue, they did not fully resolve it for the custom implementation.

Limited diversity in the generated images indicated that the model failed to capture the full distribution of tumor appearances, particularly in tumor morphology. This suggested difficulties in learning the complex distribution, possibly due to the relatively small dataset combined with the high dimensionality of the data.

Anatomical inconsistencies remained problematic even in the most successful generations, which sometimes contained implausible features such as asymmetric lung boundaries or unrealistic tumor placements. This highlighted the challenging nature of maintaining anatomical constraints in an adversarial training framework without explicit anatomical priors.

Several key lessons arose from the custom implementation experience. The critical importance of high-quality, consistent data preprocessing became evident, particularly for 3D medical imaging datasets. The need for balance between model complexity and training stability in GAN architectures was repeatedly demonstrated. Domain-specific architectural modifications, such as conditional generation and separate outputs for image and segmentation mask, proved valuable. The benefits of implementing state-of-the-art stabilization techniques were apparent even in baseline implementations.

These insights informed the second phase of experimentation, which employed state-of-the-art models with architectural innovations specifically designed to address the limitations observed in the custom implementation.

4.2 State-of-the-Art Models

While custom implementations provide valuable insights into the specific challenges of medical image synthesis, established architectures offer proven solutions to common generative modeling problems such as training instability and mode collapse. This section presents the implementation and adaptation of two advanced generative frameworks: StyleGAN3, which represents the cutting edge in GAN-based image synthesis, and a Latent Diffusion approach that utilizes a fundamentally different generative paradigm.

4.2.1 StyleGAN3

StyleGAN3 is a recent creation that builds upon its predecessors by introducing several key improvements that make it particularly suitable for medical image synthesis. The architecture uses an alias-free generator design that eliminates unwanted artifacts and improves translation equivariance, creating more coherent anatomical structures [KAL+21]. Unlike traditional GANs that directly map latent vectors to images, StyleGAN3 uses a mapping network to transform the input latent code into an intermediate latent space ($\mathcal{W}$), followed by adaptive instance normalization (AdaIN) to control the generation process at different resolutions [KAL+21].
This study applied the StyleGAN3-T (translation equivariant) configuration directly to the medical imaging task without architectural modifications. The implementation utilized the same axial slice dataset as the final custom GAN implementation but benefited from StyleGAN3's inherently improved training dynamics.
The model's ability to preserve high-frequency details proved particularly valuable for tumor representation, as small nodules require precise preservation of local structures. StyleGAN3's translation equivariance also provided a consistent representation of lung anatomical features regardless of their position in the image, an important property for medical imaging applications.

4.2.2 Latent Diffusion Model

Latent Diffusion Models (LDMs) offer a compelling alternative to GANs by addressing their known challenges, such as training instability and limited diversity. Instead of operating directly in pixel space, LDMs first encode images into a lower-dimensional latent space using a pretrained autoencoder. The diffusion process is then applied in this latent space, which drastically reduces computational complexity while preserving semantic structure [RBL+21].
This approach allows for efficient training at high resolutions and enables the generation of detailed images with improved diversity. In this study, a Latent Diffusion pipeline was implemented using the CompVis framework. No conditioning was applied to the generation process, maintaining a direct comparison with the unconditioned StyleGAN3 outputs.
By learning to denoise latent representations over multiple timesteps, LDMs can capture complex anatomical patterns with stable training dynamics. Their ability to generate coherent CT-like textures and plausible structural variations makes them a strong candidate for medical image synthesis.

4.3 Evaluation Framework

To assess the quality and utility of the generated images, this study employs two quantitative evaluation metrics:

1. **Fréchet Inception Distance (FID):** FID quantifies the difference between the feature distributions of real and generated images using the activations of a pretrained Inception network. Lower FID values indicate higher similarity. For each generative model, 50 000 synthetic axial CT slices were generated and compared against 50 000 real slices. All images were resized and normalized per InceptionV3 input requirements.

2. **Precision and Recall (PR):** These metrics assess the generative model's ability to accurately represent the real data distribution (recall) and generate valid samples within the support of real data (precision), based on distances in the Inception feature space. PR50k computes precision and recall over 50 000 real and 50 000 generated samples, offering a complementary view to FID by disentangling fidelity and diversity.

With this combination of metrics, I was able to capture global anatomical realism using FID, while PR measured semantic accuracy of the images, which may be helpful for diagnostic applications.

4.4 Experimental Setup

The experiments were conducted using preprocessed axial slices from the LIDC-IDRI dataset, as described in Section 3. All slices were resized to 256×256 resolution. No tumor masks or segmentation labels were used in training.

4.4.1 Hardware

The GAN training and inference were performed on a single machine with two NVIDIA RTX 3090 GPUs (24×2 GB VRAM), Intel Xeon CPU, and 256 GB RAM. Mixed precision training was enabled where supported. Since the server was a shared server, full utilization was not possible. Instead, approximately 8.3GB of GPU VRAM was able to be reserved on a singular GPU, of which 6.5 GB was used effectively during training.
For Latent Diffusion training, the LIACS ALICE HPC Cluster was used, specifically the A100 node with 40 GB VRAM, 32 GB RAM, and 16 CPUs.

4.4.2 StyleGAN3 Configuration

The StyleGAN3-T variant was used with the default alias-free generator architecture and the StyleGAN2 discriminator. Training used the official NVIDIA implementation with the following configuration:

- Input resolution: 256×256
- Batch size: 8
- Generator latent dimensions: $z \in \mathbb{R}^{512}$, $w \in \mathbb{R}^{512}$
- Optimizer: Adam ($\beta = (0, 0.99)$, $\epsilon = 1\text{e}{-8}$)

- Learning rates: $G = 0.0025$, $D = 0.002$
- Loss: StyleGAN2 loss with R1 regularization ($\gamma = 8.0$)
- Data augmentations: x-flip, 90° rotation, translation, scaling, contrast, hue, brightness, saturation, anisotropy, etc.
- Adaptive Discriminator Augmentation (ADA): Enabled with target probability 0.6
- EMA decay: Half-life = 2.5 kimg

4.4.3 Latent Diffusion Configuration

The latent diffusion model and its first-stage autoencoder were trained using PyTorch Lightning with the following settings:

- **Trainer / Logging**
  - Accelerator: DDP on 1 GPU, precision = 32-bit, benchmark = true, `num_sanity_val_steps`=0
  - Max epochs: 7000
  - ModelCheckpoint: save top 3 (monitor = `val/loss_simple_ema`), every 3000 train steps
  - ImageLogger: batch_frequency = 3000, max_images = 8
- **Optimizer / LR**
  - Base learning rate: 5.0×10^{-6}
- **Diffusion**
  - Timesteps: 1000 (conditional = 1)
  - Noise schedule: linear $\beta_{\text{start}} = 8.5 \times 10^{-4} \rightarrow \beta_{\text{end}} = 1.2 \times 10^{-2}$
  - Scale factor: 0.31778
  - Unconditional (no conditioning stage)
- **Scheduler**
  - LambdaLinearScheduler: warm-up = 5000 steps; cycle = 100 000 steps
  - $f_{\text{start}} = 1 \times 10^{-6},\ f_{\text{max}} = 0.5,\ f_{\text{min}} = 0.1$
- **U-Net**
  - Image size = 64; in/out channels = 4; model channels = 320
  - Channel mults: [1, 2, 4, 4]; 2 res-blocks per level
  - Attention at resolutions [4, 2, 1], 8 heads; checkpointing enabled
- **First-Stage VAE**
  - AutoencoderKL (ckpt = `last.ckpt`; embed_dim = 4; double_z = true)

  - DDConfig: z_channels = 4; resolution = 256; in/out ch = 3; base ch = 64; ch_mult = [1,2,4]; 1 res-block; no attention; dropout = 0
  - Loss: LPIPSWithDiscriminator (disc_start = 50 001; kl_weight = 1e-6; disc_weight = 0.5)

- **Data (both stages)**
  - LungCTTrain / LungCTValidation from `dataset.zip`, image size = 256, RGB
  - Batch size = 8; num_workers = 12

- **Separate VAE Fine-Tuning**
  - Base LR = 4.5×10^{-6}
  - Checkpoint: every 2000 steps, top 5
  - ImageLogger: every 500 steps, max 4 images
  - Max epochs = 100; val_check_interval = 0.5
  - Batch size = 16; num_workers = 12

5 Experiments and Results

This section will present the results of the image generation for both methodologies. First, the quantitative findings will be reported. Then, a brief comparison of the training efficiency and speed will be presented, followed by a qualitative assessment of the models.

5.1 Quantitative Assessment of Generated Images

This section will go through the quantitative findings of both models and their progress throughout the training process. First, the quantitative findings will be discussed for both models.

5.1.1 StyleGAN3

StyleGAN3 achieved stable training dynamics across the full 25 000 kimg training duration as can be seen in Figure 10. The Fréchet Inception Distance (FID) consistently decreased during training, slowing down at 10 million images and then plateauing after approximately 20 million images. Final *fid50k* computed over 50 000 generated and real axial CT slices was **21.06**.

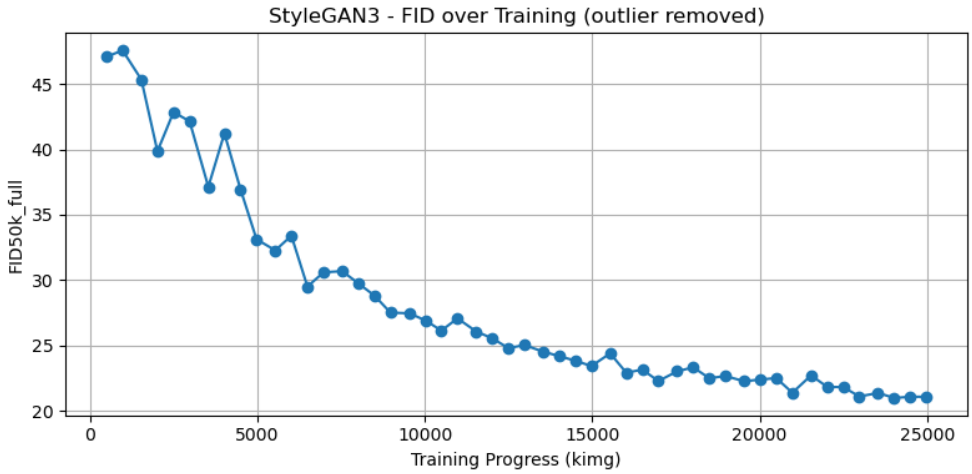

Figure 10: A graph showing the FID metrics over the 25M image training duration of the StyleGAN3. The first FID score of 500+ based solely on pure noise has been removed to maintain scaling consistency.

In addition, the `pr50k` metrics indicated **0.22** precision and **0.02** recall, suggesting that StyleGAN3 produced high-fidelity images with limited diversity. These values, also visible in Figure 11, reflect the generator's ability to synthesize anatomically realistic axial lung CT slices. However, the range of anatomical variation covered by the model remained narrow.

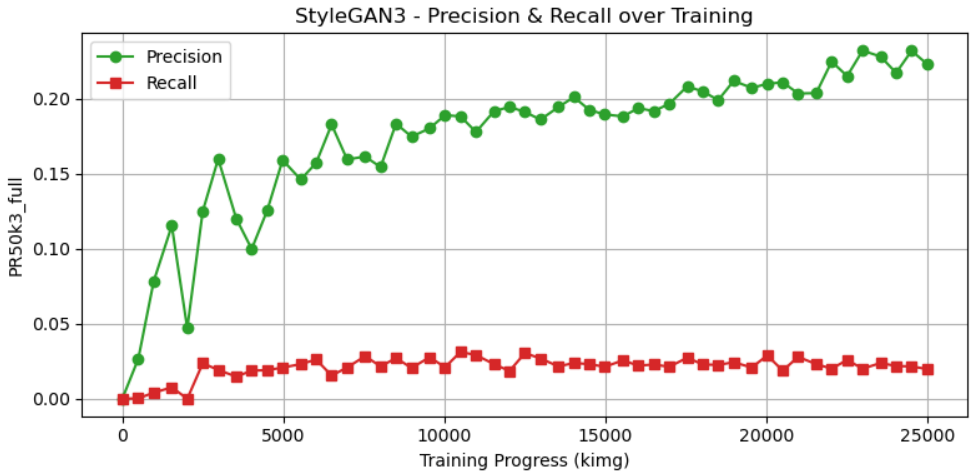

Figure 11: Precision and recall throughout training StyleGAN3. Precision improved steadily, indicating an increasing level of sample realism. Recall remained low, suggesting limited diversity across generated samples.

5.1.2 Latent Diffusion

Due to time and resource constraints, calculating metrics over the model's training progress was not possible. However, FID was calculated over the final model checkpoint. The final model achieved an FID of **130.2**, indicating poor distribution compared to the original dataset. The FID was

calculated using a 1:1 ratio of real and fake images, with 500 steps for the latent diffusion sampler. This was done instead of a complete FID 50k metric, such as for StyleGAN3, due to time and resource constraints.

5.2 Model Training and Computational Efficiency Comparison

5.2.1 StyleGAN3

The StyleGAN3-T model was trained using a batch size of 8 on a single NVIDIA RTX 3090 GPU (24 GB VRAM). Peak GPU memory usage during training was approximately **6.5 GB**. Training progressed at an average speed of **86.5 images/second**, requiring a total of **25.95 days** (**622.9 hours**) to reach 25 000 kimg.
Training was stable throughout. Adaptive Discriminator Augmentation (ADA) helped regulate discriminator overfitting, especially in the early stages. Checkpoints and evaluation metrics were saved at regular intervals to monitor performance. CPU memory consumption remained low (approximately 1.5 GB), with no observed bottlenecks in data loading or maintenance.

5.2.2 Latent Diffusion

The LDM was trained with a batch size of 8 on a single NVIDIA A100 (MIG mode, 40 GB VRAM). Peak GPU memory usage was approximately **38.5 GB**. Training ran for **71.6 hours** with an average RAM usage of around **30 GB**. Checkpoints (9.7 GB) were saved at regular intervals. Training remained stable throughout and required minimal hyperparameter tuning (learning rate warm-up only). No data-loading or I/O bottlenecks were observed.

5.3 Qualitative Assessment of Generated Images

This section evaluates the qualitative assessments of the two different models separately.

5.3.1 StyleGAN3

The generated samples exhibit realistic lung anatomy with soft tissue textures, air-filled regions, and consistent symmetry. Throughout training, visual quality steadily improved, most notably in contrast, structural clarity, and noise reduction, as shown in Figure 12.

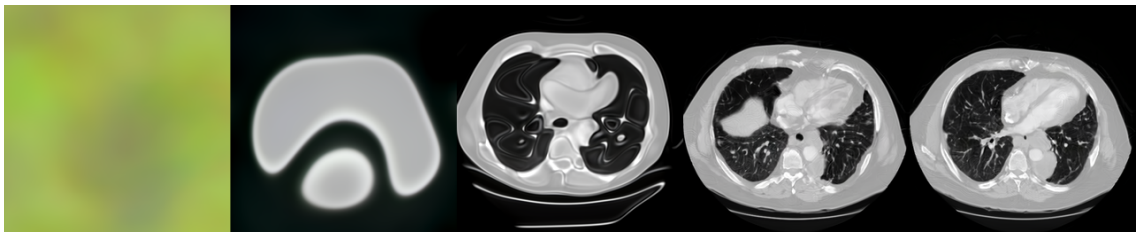

Figure 12: Progressive outputs from StyleGAN3 across training checkpoints (0kimg, 16kimg, 64kimg, 10 400kimg, 24 000kimg). The model improved steadily in structure and realism.

By 25 000 kimg, the model consistently produced high-fidelity outputs with sharp boundaries, realistic anatomical layouts, and minimal visual artifacts. Figure 13 shows a representative set of

outputs from the final generator. While tumor features are not explicitly modeled, some generated slices exhibit irregularities resembling nodules or lesions. However, the generated dataset still lacks sufficient structural variation, matching with the low recall observed in `pr50k`.

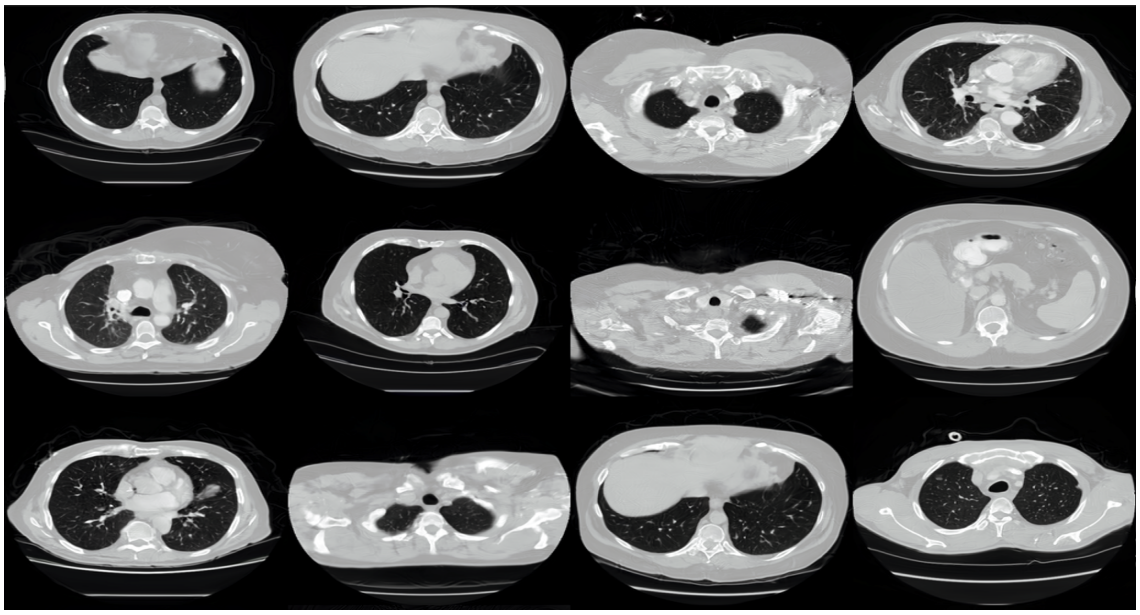

Figure 13: Grid of representative synthetic axial CT slices generated at 25 000 kimg. Images show consistent anatomical realism and visual coherence.

To assess the realism of generated images, two visual Turing tests were conducted. In the first one, non-expert participants, mostly university computer science students, were shown the image pair in Figure 14 and asked to identify which image was real, along with their confidence in their answer on a scale of 1 to 5 (excluding those who answered "I can't tell"). The results are summarized in Table 1.

Response Type	Count
Chose "Left is real"	**8**
Chose "Right is real"	12
Chose "I can't tell"	8

(a) Response distribution (N = 28)

Metric	Value
Mean Confidence	2.65
Std. Dev.	1.06

(b) Confidence statistics (excluding "I can't tell")

Table 1: Non-expert Turing test results on realism of synthetic lung CT images

The respondents were also given the option to provide an explanation for their reasoning. Most commonly among those who answered correctly, the reasoning was that the synthetic image looked too perfect, with a perfect wavy tissue structure.
The results indicate that non-expert participants were generally unable to reliably distinguish between real and synthetic CT images. With only 8 out of 28 identifying the real image correctly and 8 selecting "I can't tell," well below what would be expected from random guessing (50%). The mean confidence was low (2.65 ± 1.06), reflecting uncertainty even among those who made a

definitive choice. This suggests that the synthetic images generated by StyleGAN3 were visually convincing enough to fool non-experts in a side-by-side comparison.

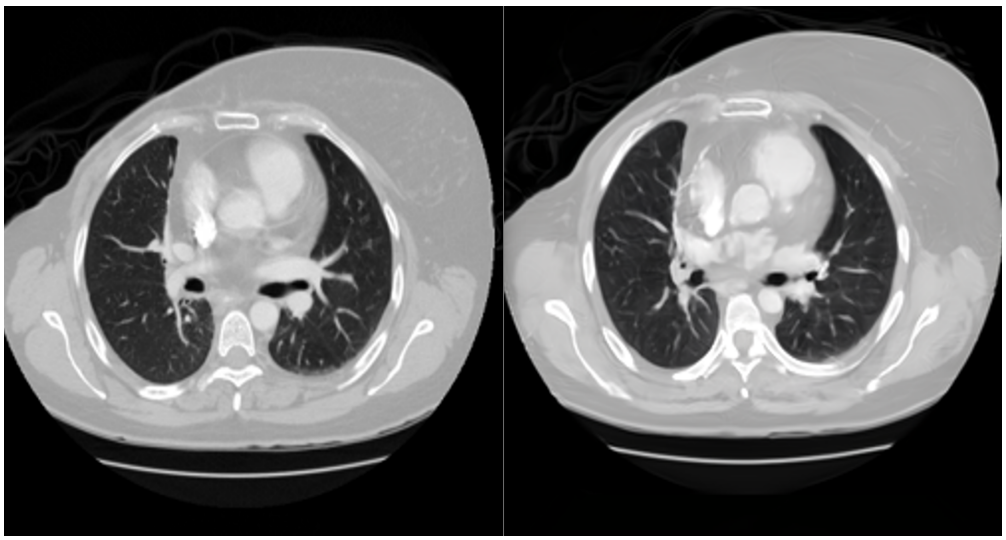

Figure 14: Comparison between a real axial CT slice (left) and a StyleGAN3-generated image (right). Structural realism is preserved in lung shape and texture.

Another, more thorough survey was conducted targeting individuals in the medical field, ranging from medical students to attending radiologists. All respondents were shown the same ten CT images, consisting of five real and five synthetic images, randomly selected from the datasets. Images were presented one at a time, with the ability to move forward and backward.
For each image, participants were required to classify it as "Real", "Synthetic", or "Not Sure". If they selected "Real" or "Synthetic," they were also asked to indicate their confidence on a scale from 1 to 5. Responses marked as "Not Sure" were considered incorrect in the calculation of the metrics. Respondents additionally self-reported their level of experience in interpreting chest CT scans. The results are summarized in Table 2.

Metric	Value
Mean Score	5.0/10.0
Median Score	5.0/10.0
Score Range	[5, 5]

(a) Response distribution (N = 2)

Metric	Value
Mean Confidence	3.12
Std. Dev.	0.60

(b) Confidence statistics (excluding "Not Sure")

Table 2: Expert Turing test results on realism of synthetic lung CT images

The average classification score across participants was 5.0/10.0, with a mean confidence score of 3.12 ± 0.60. Given that random guessing would yield an expected score of 5 out of 10, the average performance observed in the expert survey suggests that the synthetic images were often indistinguishable from real CT scans, even for trained observers. The mean confidence rating of 3.12 indicates moderate certainty, suggesting that participants found the task challenging and were not consistently confident in their assessments.

However, for both visual Turing tests, it is essential to note that the narrow standard deviations in confidence ratings indicate limited spread, potentially reflecting centrality bias—participants may have tended toward mid-scale ratings (e.g., 3) when uncertain, rather than utilizing the full range of the confidence scale.

5.3.2 Latent Diffusion

The VAE exhibited excellent reconstruction fidelity, with outputs nearly indistinguishable from the original inputs, as shown in Figure 15 and Figure 16.

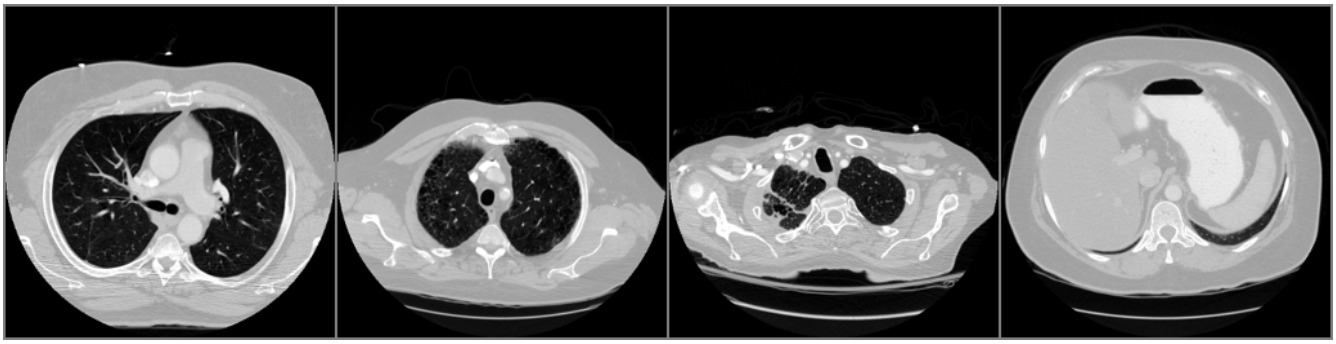

Figure 15: Original lung CT scan inputs provided to the VAE.

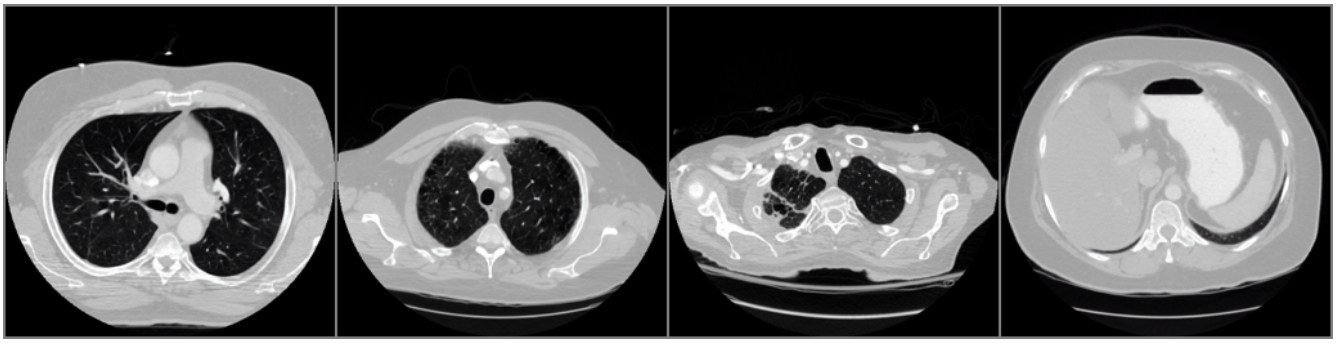

Figure 16: VAE reconstructions corresponding to the inputs in Figure 15.

The latent diffusion model learned to generate structurally plausible CT-like images early in training. By epoch six, it produced samples such as those in Figure 17. Although anatomically inaccurate, these outputs indicate that the latent space captures key structural priors. As training progressed, the outputs became increasingly realistic, as illustrated in Figure 18.

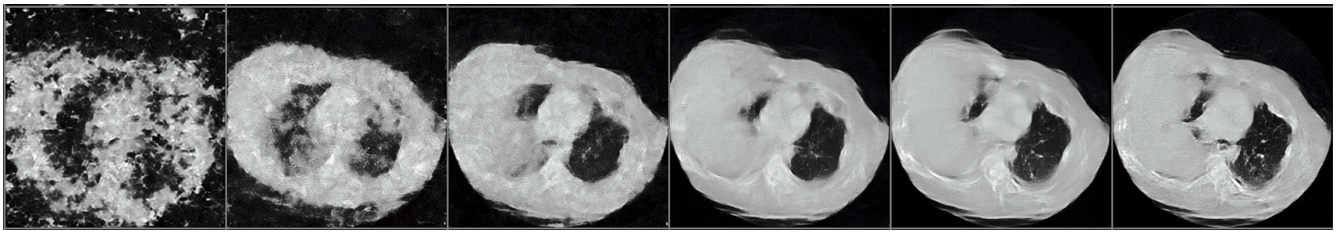

Figure 17: Generated progression row at epoch 6 showing initial structural learning, with 200 steps.

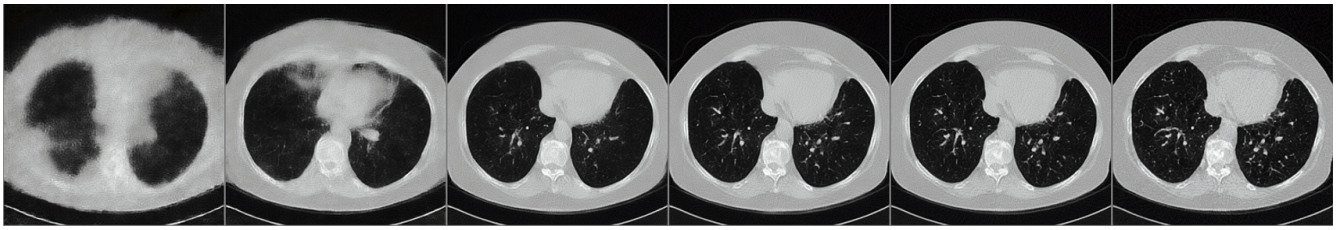

Figure 18: Generated progression row at epoch 222 showing improved anatomical plausibility, with 200 steps.

Despite the improvements, noticeable noise persists in the outputs. However, since the number of sampling steps is configurable post-training, this noise can be reduced at inference time by increasing the number of denoising steps, at the cost of slower generation. All samples shown here were generated with 200 steps.

6 Discussion

This section will discuss the outcomes of the experiments and their repercussions.

6.1 Comparative Analysis of StyleGAN3 vs. Latent Diffusion

The results show that StyleGAN3 requires more training steps to achieve the required model strength for generating plausible lung CT slices, only producing credible images after around 10 000 steps. In contrast, the latent diffusion model (LDM) can yield plausible outputs much sooner. StyleGAN3's tell-tale giveaway is a wavy structure in the lung tissue, while the LDM's is graininess. However, grain can be mitigated not only by increasing the DDIM step count (at the cost of time) but also by choosing alternative samplers (e.g., Euler–Ancestral or PNDM) or by tuning the guidance scale in a classifier-free setting. Finally, the quality of the pretrained VAE encoder in the LDM (its latent dimensionality and reconstruction error) imposes an upper bound on sharpness and may introduce blur unless a higher-capacity latent is used.

6.2 Trade-offs Between Quality and Computational Efficiency

Generative models in medical imaging must balance image quality, training stability, and computational cost. StyleGAN3 demonstrates high-quality generation of lung CT slices, achieving sharp

anatomical detail and plausible structures. However, these visual benefits come at a notable computational cost. Training StyleGAN3-T to convergence on 25 million images required approximately 26 days on a single NVIDIA RTX 3090, with peak GPU memory usage around 6.5 GB. Each 1kimg of training took on average 86.5 seconds, highlighting the resource intensity of state-of-the-art GAN training.
Despite these costs, StyleGAN3's training was stable with no signs of mode collapse or severe artifacts. This reliability is an advantage over earlier GAN architectures. However, the model's limited diversity, reflected in low PR recall, suggests that high fidelity does not necessarily imply full distribution coverage. The unconditioned nature of the generator also limits flexibility, as it cannot produce targeted outputs without retraining or architectural changes.
The LDM, by contrast, generates plausible anatomical images in only six epochs ($\approx$ 2h) but at the cost of much larger memory ($\approx$ 40GB on an A100 vs. 6.5GB for StyleGAN3). This stems from both the full UNet and VAE pipeline; choosing a coarser latent down-sampling (e.g., 16$\times$ vs. 8$\times$) or applying flash attention and gradient checkpointing can reduce VRAM needs. Stability was excellent thanks to a conservative learning rate, and the loss plateaued early. Quality is good, especially when using classifier-free guidance, but residual grain remains visible.

6.3 Limitations and Challenges

This section explores the various limitations and challenges that come with synthetic image generation using both StyleGAN3 and Latent Diffusion.

6.3.1 2D Representation, Lack of Context, and Unconditioned Generation

Both StyleGAN3 and the Latent Diffusion model are trained exclusively on 2D axial CT slices, omitting volumetric or multi-view information. Although the training set contains roughly half tumor-bearing slices, the models are unconditioned; there is no mechanism to specify whether a generated image should include a tumor or not. This slice-by-slice approach not only prevents guiding the generation with tumor masks or clinical labels but also breaks anatomical continuity across consecutive slices. Consequently, tasks requiring 3D consistency, such as mask-guided augmentation, tumor tracking, or volumetric segmentation, are not feasible with these models.

6.3.2 Limited Diversity and Mode Coverage

The synthetic images produced by StyleGAN3 also exhibited limited diversity, as indicated by low recall scores in the PR metric in Section 5.1. This suggests that the model did not fully capture the variability present in real-world lung anatomy and pathology. Over time, training improved image sharpness and consistency but did not significantly increase diversity, raising concerns about potential mode collapse or dataset bias.

Unfortunately, due to time and resource constraints, the diversity and mode coverage of the latent diffusion model could not be calculated similarly to StyleGAN3; therefore, a fair and accurate comparison is not possible.

6.3.3 Challenges of Medical Imaging Data

Medical imaging data presents intrinsic challenges for generative modeling. Clinical datasets are often small, imbalanced, and heterogeneous, particularly for rare pathologies. Unlike natural image domains, where millions of samples are readily available, annotated medical datasets are scarce due to concerns about privacy and high labeling costs. This scarcity is especially acute for rare tumor types or unusual anatomical variants, making it difficult for generative models to learn comprehensive distributions. As a result, models may overfit to dominant patterns, underrepresent rare cases, or generate unrealistic edge scenarios, even with SOTA technologies.

6.3.4 Training Stability and Computational Demands

Despite StyleGAN3's architectural improvements, GAN training remains sensitive to hyperparameters and initialization. Achieving stable convergence requires extensive tuning and compute, which limits reproducibility and scalability.

For latent diffusion, the increased model complexity and iterative denoising process demanded far more memory and time. The combined VAE+UNet pipeline would not even initialize on a 15 GB VRAM card and required almost the full 40 GB of the A100 to run with a batch size of 8. Nevertheless, the diffusion model produced relatively realistic-looking results within only six epochs of training, or approximately 2 hours, and its progress appeared extremely stable, with the loss plateauing early on.

6.3.5 Evaluation Limitations

Quantitative evaluation relied on FID and Precision/Recall, which, while informative, do not fully reflect clinical realism or downstream task utility. Visual inspection can supplement these metrics, but without expert validation by radiologists, the clinical utility of the synthetic images remains speculative.

6.3.6 Ethical and Regulatory Considerations

Synthetic medical images may inherit biases from the training set and risk patient privacy violations if not properly anonymized. Complete ethical audits and compliance checks (e.g. GDPR, FDA) are beyond the scope of this bachelor's thesis; instead, I acknowledge that rigorous bias assessment is left for future work.

7 Conclusions and Further Research

In this section, a summary of the performed experiments will be given, and possible directions for further research will be explored.

7.1 Summary of Findings

This study evaluated the performance of StyleGAN3 in generating synthetic lung CT slices using axial views from the LIDC-IDRI dataset. The results demonstrate that StyleGAN3 can produce

anatomically coherent, high-fidelity images that resemble real CT scans, with minimal visual artifacts. The generated outputs displayed realistic lung fields, consistent tissue gradients, and tumor-like structures, despite the generator being unconditioned.
Quantitatively, the model achieved a Fréchet Inception Distance (FID) below 25 and a precision score of 0.22, indicating high visual fidelity. However, recall remained low at approximately 0.02, reflecting limited diversity in the generated dataset. This matches the visual inspection, which showed a tendency toward generating similar structural patterns, suggesting incomplete coverage of the real data distribution.
Training was stable and converged reliably, but required significant computational resources; approximately 26 days on a single high-end GPU. The evaluation framework combined both metric-based analysis and qualitative assessments, but lacked expert clinical validation.
These findings confirm that while StyleGAN3 excels in generating sharp and plausible medical images, it faces limitations in controllability and diversity. This motivates further exploration of alternative architectures such as diffusion models, which may offer complementary advantages in these areas.
The LDM generated plausible images in only six epochs ($\approx$ 2h) and supported classifier-free guidance and alternative samplers (DDIM, PNDM, Euler–Ancestral) to balance grain versus speed. However, it required $\approx$ 40 GB VRAM on an A100 (mitigated partially by coarser latents, flash attention, checkpointing), and its pretrained VAE imposed a blur ceiling. Loss plateaued early, suggesting the learning rate might be suboptimal. Quantitative metrics (FID, PR/recall) remain to be computed under identical preprocessing for a complete head-to-head comparison.
Overall, StyleGAN3 excels in sharpness and stability, but at a high cost and with limited controllability. In contrast, latent diffusion offers rapid convergence and flexible conditioning, albeit at the expense of memory and residual noise. Future work should standardize metrics across both models, optimize LDM sampling and VAE quality, and explore conditioning to harness the complementary strengths of GANs and diffusion for medical imaging.

7.2 Directions for Further Research

Several directions can extend and improve upon the findings of this work. First, conditioning the generative model on semantic attributes such as tumor presence, size, or anatomical region could enable targeted synthetic data generation for specific diagnostic tasks. This would increase the utility of synthetic images for augmenting underrepresented cases.
Second, transitioning from 2D to 3D modeling would improve spatial continuity and realism by incorporating volumetric context, which is crucial for many clinical applications. Techniques such as 3D GANs or slice-consistent diffusion models could help address this limitation; however, their performance-to-cost ratio should be further studied.
Third, future work should include radiologist-led evaluations, including structured visual Turing tests, to assess clinical plausibility. This would provide more meaningful validation than current quantitative metrics alone.
Additionally, using larger and more diverse datasets or applying domain adaptation techniques may reduce overfitting and improve generalization, particularly for rare or complex pathologies.
Lastly, a more in-depth evaluation of latent diffusion models is warranted: compute FID, precision/recall under the same preprocessing as StyleGAN3, experiment with classifier-free guidance

scales, and benchmark different samplers. This will clarify under which settings LDMs can match GAN fidelity and diversity without prohibitive compute.

References

[AAP+23] Anmol Arora, Joseph E. Alderman, Joanne Palmer, Shaswath Ganapathi, Elinor Laws, Melissa D. McCradden, Lauren Oakden-Rayner, Stephen R. Pfohl, Marzyeh Ghassemi, Francis McKay, Darren Treanor, Negar Rostamzadeh, Bilal Mateen, Jacqui Gath, Adewole O. Adebajo, Stephanie Kuku, Rubeta Matin, Katherine Heller, Elizabeth Sapey, Neil J. Sebire, Heather Cole-Lewis, Melanie Calvert, Alastair Denniston, and Xiaoxuan Liu. The value of standards for health datasets in artificial intelligence-based applications. *Nature Medicine*, 29(11):2929, 11 2023.

[AIMB+15] S G Armato III, G McLennan, L Bidaut, M F McNitt-Gray, C R Meyer, A P Reeves, B Zhao, D R Aberle, C I Henschke, E A Hoffman, E A Kazerooni, H MacMahon, E J R Van Beek, D Yankelevitz, A M Biancardi, P H Bland, M S Brown, R M Engelmann, G E Laderach, D Max, R C Pais, D P Y Qing, R Y Roberts, A R Smith, A Starkey, P Batra, P Caligiuri, A Farooqi, G W Gladish, C M Jude, R F Munden, I Petkovska, L E Quint, L H Schwartz, B Sundaram, L E Dodd, C Fenimore, D Gur, N Petrick, J Freymann, J Kirby, B Hughes, A V Casteele, S Gupte, M Sallam, M D Heath, M H Kuhn, E Dharaiya, R Burns, D S Fryd, M Salganicoff, V Anand, U Shreter, S Vastagh, B Y Croft, and L P Clarke. Data From LIDC-IDRI, 2015.

[CT] T F Cootes and C J Taylor. Statistical models of appearance for medical image analysis and computer vision *.

[DN21] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. 5 2021.

[GPAM+14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. 6 2014.

[GSVV24] Bernardo Gonçalves, Mariana Silva, Luísa Vieira, and Pedro Vieira. Abdominal MRI Unconditional Synthesis with Medical Assessment. 2024.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. 6 2020.

[HLC+22] Siyi He, He Li, Maomao Cao, Dianqin Sun, Fan Yang, Xinxin Yan, Shaoli Zhang, Yutong He, Lingbin Du, Xibin Sun, Ning Wang, Min Zhang, Kuangrong Wei, Lin Lei, Changfa Xia, Ji Peng, and Wanqing Chen. Survival of 7, 311 lung cancer patients by pathological stage and histological classification: a multicenter hospital-based study in China. *Translational Lung Cancer Research*, 11(8):1591–1605, 8 2022.

[JLB+25] Yifan Jiang, Yannick Lemaréchal, Josée Bafaro, Jessica Abi-Rjeile, Philippe Joubert, Philippe Després, and Venkata Manem. Lung-DDPM: Semantic Layout-guided Diffusion Models for Thoracic CT Image Synthesis. 2 2025.

[KAL+21] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-Free Generative Adversarial Networks. 6 2021.

[LEASS21] Claire Little, Mark Elliot, Richard Allmendinger, and Sahel Shariati Samani. Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study. 2021.

[LMS+19] S. K. Lakshmanaprabu, Sachi Nandan Mohanty, K. Shankar, N. Arunkumar, and Gustavo Ramirez. Optimal deep learning model for classification of lung cancer on CT images. *Future Generation Computer Systems*, 92:374–382, 3 2019.

[MCFR21] Katharina Martini, Guillaume Chassagnon, Thomas Frauenfelder, and Marie Pierre Revel. Ongoing challenges in implementation of lung cancer screening. *Translational Lung Cancer Research*, 10(5):2347, 5 2021.

[MPS+23] José Mendes, Tania Pereira, Francisco Silva, Julieta Frade, Joana Morgado, Cláudia Freitas, Eduardo Negrão, Beatriz Flor de Lima, Miguel Correia da Silva, António J. Madureira, Isabel Ramos, José Luís Costa, Venceslau Hespanhol, António Cunha, and Hélder P. Oliveira. Lung CT image synthesis using GANs. *Expert Systems with Applications*, 215, 4 2023.

[NHS23] NHS. Cancer Survival in England, cancers diagnosed 2016 to 2020, followed up to 2021 - NHS England Digital, 2023.

[Oht] Takumi Ohta. Basics of X-ray CT reconstruction Principles and applications of iterative reconstruction. *Rigaku Journal*, 39(1):2023.

[PZM+24] Vasileios C. Pezoulas, Dimitrios I. Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S. Tachos, and Dimitrios I. Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods, 12 2024.

[RA20] Diego Riquelme and Moulay A. Akhloufi. Deep Learning for Lung Cancer Nodules Detection and Classification in CT Scans, 12 2020.

[RBL+21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. 12 2021.

[SBD+] E Sizikova, A Badal, J G Delfino, M Lago, B Nelson, N Saharkhiz, B Sahiner, G Zamzmi, and A Badano. Synthetic Data in Radiological Imaging: Current State and Future Outlook.

[SDK+22] Imran Shafi, Sadia Din, Asim Khan, Isabel De La Torre Díez, Ramón del Jesús Palí Casanova, Kilian Tutusaus Pifarre, and Imran Ashraf. An Effective Method for Lung Cancer Diagnosis from CT Scan Using Deep Learning-Based Support Vector Network. *Cancers*, 14(21), 11 2022.

[SDWMG15] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. 3 2015.

[SMTK+25] Rebecca L Siegel Mph, — Tyler, B Kratzer, Angela N Giaquinto, — Hyuna, Sung Phd, — Ahmedin, and Jemal Dvm. Cancer statistics, 2025. 2025.

[YWB] Xin Yi, Ekta Walia, and Paul Babyn. Generative Adversarial Network in Medical Imaging: A Review.

[YZZ+23] Zhiwei Yang, Jianhu Zhao, Hongmei Zhang, Yongcan Yu, and Chao Huang. A Side-Scan Sonar Image Synthesis Method Based on a Diffusion Model. *Journal of Marine Science and Engineering*, 11(6), 6 2023.

[ZXL+24] Jialin Zhou, Ying Xu, Jianmin Liu, Lili Feng, Jinming Yu, and Dawei Chen. Global burden of lung cancer in 2022 and projections to 2050: Incidence and mortality estimates from GLOBOCAN. *Cancer Epidemiology*, 93:102693, 12 2024.