HYBRID GAN-DIFFUSION FRAMEWORK FOR ULTRASOUND TO CT TRANSLATION: A RESIDUAL REFINEMENT APPROACH

by

Sserujja Abdallah Kulumba (200021350)
Oumar Ahmad Bargami (200021362)
Zakariaou Ibrahim (200021262)

Report on

EEE 4800 (Project and Thesis)

Department of Electrical and Electronic Engineering
Islamic University of Technology (IUT)
Gazipur, Bangladesh

October 2025

DECLARATION

We hereby declare that this thesis titled **"Hybrid GAN-DIFFUSION Framework for Ultrasound to CT Translation: A Residual Refinement Approach"** is our own original work and has not been submitted previously or concurrently for any other degree or qualification at this or any other institution. All sources of information and assistance used in this research have been duly acknowledged.

The team members all worked together in problem-solving, debugging and preparing the final manuscript

Sserujja Abdallah Kulumba	Oumar Ahmad Bargami	Zakariaou Ibrahim
200021350	200021362	200021262

APPROVAL PAGE

HYBRID GAN-DIFFUSION FRAMEWORK FOR ULTRASOUND TO CT TRANSLATION: A RESIDUAL REFINEMENT APPROACH

Submitted by:

Sserujja Abdallah Kulumba - 200021350
Oumar Ahmad Bargami - 200021362
Zakariaou Ibrahim - 200021262

in Consideration of Partial Fulfillment for the Requirements of the Degree of
BACHELOR OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING

Dr. Khondokar Habibul Kabir
Professor,
Department of Electrical and Electronic Engineering,
Islamic University of Technology (IUT),
Board Bazar, Gazipur-1704, Bangladesh

Date of Approval: _____

This thesis is dedicated to our parents and family members.
Their unconditional love, patience and encouragement have been our constant foundation.
Through long hours and complex challenges, their support never wavered,
and for that, we are eternally grateful.

ACKNOWLEDGMENTS

All praise and gratitude are due to Allah the Most Compassionate and Most Merciful, whose guidance enlightens our path and whose mercy sustains us in all that we do

We would like to thank our supervisor, Dr. Khondokar Habibul Kabir for his continuous advice, unwavering support, technical expertise and guidance from scratch throughout this thesis work. We are also deeply grateful to our teacher Mr. Md. Arefin Rabbi Emon, whose motivation and positive criticism provided an inspiring research environment.

We are indebted to the researchers behind the TRUSTED dataset. Their decision to make such a valuable and meticulously curated dataset publicly available was the catalyst for this entire project, and we are thankful for their contribution to the scientific community. We also thank the developers of PyTorch, MONAI, and Diffusers and we gratefully acknowledge their work.

We are grateful to our colleagues and peers for stimulating discussions and a supportive research atmosphere and in general the Electrical and Electronic Department

ABSTRACT

The translation of 3D ultrasound (US) to computed tomography (CT) is a significant challenge in medical imaging due to the vast domain gap and inherent US artifacts. This thesis introduces a novel, two-stage hybrid framework to address this, synergizing a custom Generative Adversarial Network (GAN) with a Denoising Diffusion Probabilistic Model (DDPM). The first stage confronts GAN instability and mode collapse with our proposed UNetSelfCritique3D generator, which uses an internal feature-matching loss to enforce structural consistency and produce a coarse but anatomically faithful CT volume. The second stage employs a conditional DDPM for residual refinement, using the coarse GAN output as a strong prior to synthesize missing high-frequency details. Validated on the clinical TRUSTED dataset, our baseline 3D Pix2Pix model achieved a Peak Signal-to-Noise Ratio (PSNR) of 11.34 dB and a Structural Similarity Index (SSIM) of 0.181. The UNetSelfCritique3D significantly improved performance to 20.57 dB PSNR and 0.468 SSIM, demonstrating effective mitigation of mode collapse. The final hybrid GAN-Diffusion model achieved the best results with 21.10 dB PSNR, 0.456 SSIM, and 0.547 Normalized Cross-Correlation (NCC), representing a substantial 86% improvement in PSNR over the baseline. This work establishes the first crucial performance benchmark for 3D US-to-CT translation and validates the feasibility of our hybrid approach. The proposed framework and code are made publicly available to foster reproducibility and accelerate future research.

Keywords: Medical Image Translation, Ultrasound, Computed Tomography, Generative Adversarial Networks, Diffusion Models, Deep Learning, Hybrid Models, Residual Learning.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Introduction

Medical imaging serves as the main diagnostic instrument which enables doctors to view internal body structures for improved medical evaluation and treatment planning. Two among the most common imaging modalities, Computed Tomography (CT) and Ultrasound (US), present complementary but inherently dissimilar perspectives. The CT system delivers absolute visual detail of body structures which makes it the most reliable instrument for medical surgery planning and diagnostic procedures. The method provides detailed results but produces ionizing radiation which creates an unacceptable danger to patients who need continuous monitoring and those who need multiple tests. Ultrasound provides risk-free operation and portable functionality with real-time feedback but its images become difficult to analyze because of noise and artifacts which hinder the interpretation of detailed anatomical structures. The dualism creates a major community problem which demands a solution to provide doctors with CT scan diagnostic capabilities while maintaining the advantages of ultrasound safety and mobility. The current study aims to solve this issue through the development of a new deep learning method which transforms US scans into CT-quality images for making high-quality imaging accessible to all while protecting patients.

1.2 Background

Contemporary medicine is closely linked to the ability to see inside the human body without performing any invasive procedures. The medical field depends on imaging technology to create essential maps which guide doctors during diagnosis and treatment planning and real-time surgical procedures. The two most common image display methods from the available options are Computed Tomography (CT) and Ultrasound (US) which provide different yet often complementary perspectives of the human body.

Recent advances in generative models—particularly Generative Adversarial Networks (GANs) [1] and Denoising Diffusion Models (DDMs) [2]—have shown impressive capabilities for

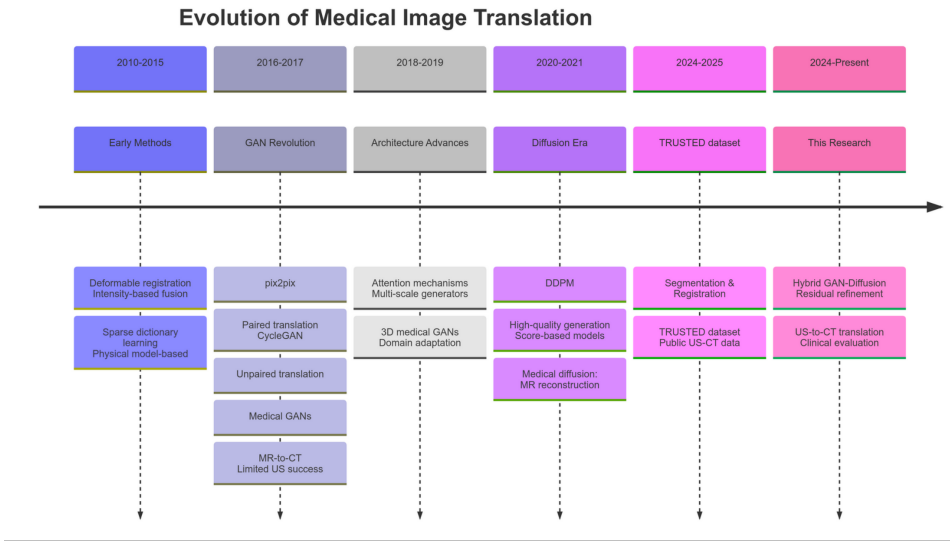


Figure 1.1: Pipeline of medical imaging with developing deep learning models.

medical image synthesis, shifting the field from image analysis toward image generation. Medical imaging itself has evolved through several landmark innovations. Sir Godfrey Hounsfield's invention of the CT scanner [3] in the 1970s transformed medicine by enabling detailed cross-sectional views of internal anatomy for the first time. Ultrasound technology later brought real-time, non-invasive imaging into routine clinical practice. The past decade has witnessed another transformation through artificial intelligence. As shown in Figure 1.1, deep learning has been extensively applied to various imaging pairs like CT-MRI and PET-CT. However, ultrasound-to-CT translation remains relatively unexplored due to unique technical challenges discussed in later chapters. This work addresses this gap. The invention of Generative Adversarial Networks (GANs) [1] and Denoising Diffusion Models (DDMs) [2] has made it possible that with excessive enough data modality exchange can be feasible. This inturn makes the foundation basement of our thesis as will be elaborated in subsequent sections of this book.

1.2.1 Physics of Computed Tomography (CT)

CT is a transmission imaging technique that reconstructs cross-sectional body images from X-ray projections captured at multiple angles. An X-ray source rotates around the patient, sending a fan-shaped beam through the body while detectors on the opposite side measure how much radiation passes through and detectors on the other side measure the strength of rays that pass through. This is repeated from several hundred distinct directions, and sophisticated mathematical algorithms, characteristically filtered back-projection-based, build these measurements of attenuation into a sophisticated 3D volumetric map. Voxel values are converted to the Hounsfield scale, the

standardized quantitative index of radiodensity with air at -1000 Hounsfield Units (HU) and water at 0 HU. Objective tissue differentiation is facilitated through this scale: bony density is bright with high HU, whereas soft tissues like muscle, fat, and viscera fall within a range at intermediate gray levels. The paramount benefit of CT is the excellent spatial and, particularly, contrast resolution, for bone, soft tissue, and vascular studies, yielding images with very good geometric fidelity, lofty signal-to-noise ratio, and clear, interpretible contrast. Through this, CT is the gold standard for many diagnostic functions, for example, the staging of malignancies, the evaluation of traumatic injuries, and the planning of surgery.

Its major drawback, however, is the use of ionizing radiation, posing an accumulating risk to patients, especially those requiring repeated follow-up scan or children. In addition, the bulky size, expense, and fixed-site location of the CT scanner restrict the use to dedicated departments of radiology.

1.2.2 Physics of Ultrasound (US)

Ultrasound (US) uses high-frequency sound to depict structures in the body. A transducer emits acoustic pulses and detects returning echoes when they strike interfaces between tissues of different acoustic impedances. Both the amplitude and time of flight of echoes are quantified by the transducer, with the luminance of a pixel corresponding to echo strength and its position by the time of return. Tissues that intensify sound considerably, such as renal capsule or diaphragm, are hyperechoic (bright), while fluid-filled structures such as cysts appear anechoic (dark) because they transmit sound with minimal reflection. US is famous for being nonionizing and safe because it does not employ ionizing radiation. It is highly portable, inexpensive, and provides real-time images, thus a valuable asset for dynamic assessments, biopsies guidance, and intraoperative navigation.

However, US imaging has certain inherent limitations. The most obvious is speckle noise, a gritty artifact caused by constructive and destructive interference of backscattered sound waves, that decreases image contrast and distorts fine anatomy. It is operator-dependent in its success and passes through bone and air. Highly reflective or attenuating structures like bowel gas or ribs can completely absorb the sound waves, creating acoustic shadows that cover the anatomy below them.

1.2.3 Complementary Nature of CT and US

The particular strengths and limitations of US and CT render them strong complements to each other. Take an image-guided surgery example: a surgeon wants the high-res, pre-op anatomical map provided by a CT scan but also requires the real-time, radiation-free feedback of

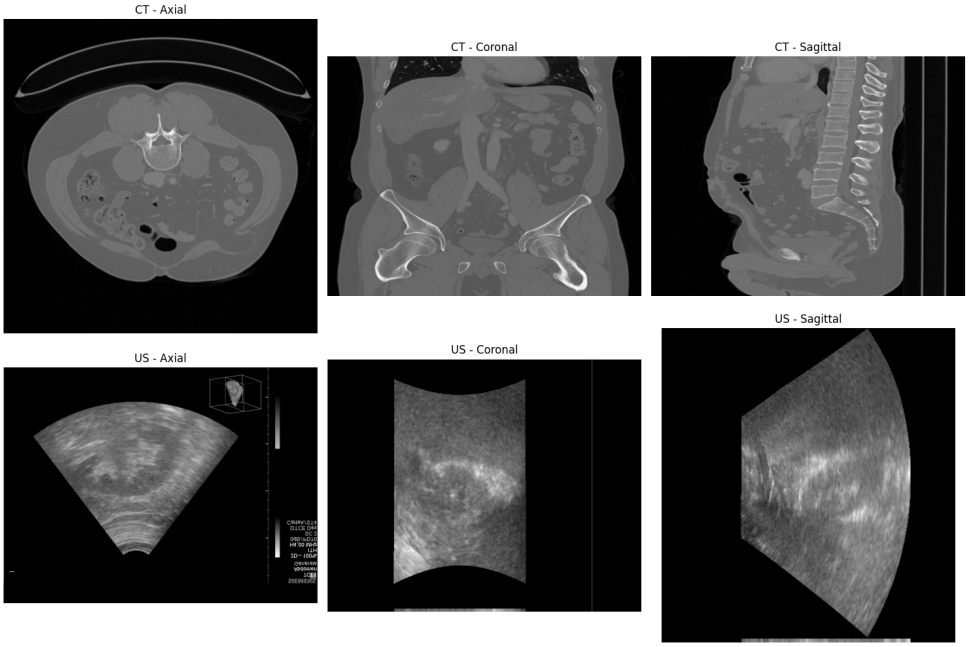


Figure 1.2: Comparison of 3D CT (first row) and 3D US (second row) views.

an ultrasound probe intraoperatively. Combining these two modalities is a common desire, but one which often requires advanced registration algorithms and does not always succeed. This leads to the intriguing question: suppose that we could infer the anatomical resolution of a CT scan from the real-time ultrasound data directly.

Locally, in Third World nations such as Bangladesh, Uganda, Cameroon or any other, access to advanced medical technology remains a huge barrier. While large cities may boast state-of-the-art CT scanners, these do not exist in rural and remote areas due to the prohibitive costs and heavy infrastructure they demand. The relatively cheaper and more portable ultrasound devices are infinitely more widespread. So there arises a need to find out ways to solve the issue and fortunately with rapid continuous developing technology the field has enhanced and binded the gap in a major shift towards an era of reduced reliance on high expense imaging equipment. With works such as this we make it possible for everyone to have easy access to both imaging modalities using AI-based approach hence leading to enhanced healthcare in our community.

This disparity necessitates with critical urgency the technologies that can augment the diagnostic capabilities of existing, accessible hardware. This research is a shift towards an era of reduced reliance on high-expense, centralized imaging equipment to a new AI-based approach that will be capable of maximizing the value of ubiquitous technology to directly address a leading healthcare issue in the nation.

1.3 Problem Statement (Non-Technical Viewpoint)

Imagine a child with a chronic disease of the kidneys that needs to be monitored once every six months to keep the organ healthy and the size proper. The process requires a CT scan, and with every scan, the child is exposed to a very slight dose of radiation. With repeated exposure over time, this accumulated exposure is a very serious medical concern. Now consider a doctor at a remote rural clinic that suspects that a patient is dealing with a complicated kidney issue. The doctor only has an ultrasound to use, and the resulting image is a blurred and hard-to-read photo. The nearest CT device is hundreds of miles away, a journey the patient cannot afford or is not healthy enough to endure.

This is the real-world problem. Patients, especially children and expectant mothers, are subjected to hazardous radiation for vital diagnostic insights. At the same time, millions of people living under resource constraints are deprived of quality diagnostics. The issue is that of safety and equity at the same time, and it is pressing to solve because it has a direct effect on patient safety and sustains quality inequities in the field.

1.3.1 Limitations of Existing Work

While researchers have attempted to solve this with AI, existing solutions have fallen short of real-world clinical needs:

- **2D Slice Focus:** Most of the previous work has been focused on translating 2D images (single slices), that do not contain the full 3D context required for surgeons to plan surgery or volumetric analysis.
- **Reliance on Synthetic Data:** The majority of the models are trained using synthetic ultrasound images. They are clean, perfect ultrasound images that lack the complex noise and artifacts that real clinical images contain, and thus the models underperform under real patient images.
- **Instability of Standard Models:** The standard models for this task that are used most extensively (standard GANs) are extremely unstable [4] and are prone to generate blurred, uninformative shapes that lose patient-specific information.

The inspiration behind this project comes from the pressing need to fill this safety and accessibility gap in medical imaging. The figures for healthcare inequity in developing countries and the lifetime costs of accumulative exposure to radiation are not mere theoretical statistics, but actual human beings. The impetus behind this research is the belief that deep learning can be a

democratizing element in medicine. The concept is to create a powerful AI that can be trusted to pierce the ”fog” of a noisy ultrasound to discover the clear anatomical reality below, as a CT might. The desire to build a clinically relevant, radiation-free solution to imaging is the impetus behind this work, serving a fundamental goal of the UN's Sustainable Development Goal 3 [5].

1.4 Research Gaps

The research gaps can be addressed as below:

1. There is a distinct lack of robust training and validation on real, paired, 3D clinical data, which presents a far greater challenge due to increased complexity, patient variability, and imaging artifacts [6].This needs a robust 3D generative model specifically designed to handle the high degree of noise, speckle, and artifacts inherent in clinical ultrasound volumes, which often leads to training instability and mode collapse in standard architectures.

2. A significant gap exists in achieving high-fidelity realism in the synthesized output. Existing single-stage generative models struggle to reconstruct both the coarse anatomical structure and the fine textural details of a CT scan from a US input simultaneously.

1.5 Research Question

The presence of the above gap opens a new scope of research arena and emulates the following technical research questions:

1. How can a 3D Generative Adversarial Network be architecturally modified with an internal regularization mechanism to ensure stable training and preserve structural consistency when translating from the noisy US domain to the CT domain?

2. How can a secondary refinement framework, such as a diffusion model, be integrated into a hybrid pipeline to learn the residual details and restore high-frequency textural information lost in the initial generative stage?

1.6 Problem Identification (Technical Statement)

The core technical challenge is to design and validate a stable, high-fidelity, 3D image-to-image translation framework for a cross-modality task with an exceptionally large domain gap. The problem requires learning a mapping from volumetric data governed by acoustic reflection principles (Ultrasound) to data governed by X-ray attenuation principles (Computed Tomography),

while overcoming the severe signal degradation (speckle, shadowing) in the source domain and the training pathologies (mode collapse) common to generative models.

1.7 Objectives (Technical Statement)

To address the above well defined research problem, the following SMART proposed methods is objectives are defined:

1. 1.To develop a robust preprocessing pipeline to register and align the unilateral US and bilateral CT volumes from the TRUSTED dataset [7], creating a suitable dataset for paired image translation.

2. Implement a baseline 3D Pix2Pix GAN model [8] to establish benchmark performance.

3. Design, implement and train the novel `UNetSelfCritique3D` GAN architecture to generate structurally coherent coarse CT volumes.

4. To construct and train a conditional 3D Denoising Diffusion Probabilistic Model (DDPM) capable of refining the coarse GAN output into a high-fidelity final volume.

5. To perform a comprehensive quantitative (using metrics like PSNR, SSIM) and qualitative (visual inspection) evaluation to validate the superiority of the proposed hybrid framework over baseline and intermediate models.

1.8 Scope (Boundary of the Study)

As any other research, there exists some boundaries and some are as follows:

- **Data Scope:** The research is exclusively trained and validated on the 59 paired 3D kidney volumes from 48 patients from TRUSTED dataset. Thus the model is specialized for the translation of transabdominal ultrasound of the kidney to non-contrast CT.

- **Out of Scope:**This research does not involve clinical trials, real-time implementation on ultrasound hardware, or validation on other organs or pathological cases not present in the training dataset.

1.9 Expected Outcome and Socio-Economic Benefits

- **Technical Deliverables:** A fully trained, two-stage hybrid generative model; the novel UNetSelfCritique3D algorithm; and a complete, reproducible software pipeline for 3D US-to-CT translation.

- **Healthcare Impact:** A successful US-to-CT translation model could enable "virtual CT" scans during interventions, reduce cumulative radiation exposure for patients with chronic conditions, and provide enhanced anatomical context in settings where CT is unavailable.

- **Socio-Economic Advantages:** The main technical challenge stands in creating and evaluating a steady high-precision three-dimensional image translation system which operates across different modalities with substantial domain differences. The objective requires developing a mapping system which converts volumetric data from ultrasound acoustic reflection principles to volumetric data based on X-ray attenuation principles of Computed Tomography while handling the severe signal deterioration (speckle and shadowing) from the source domain and the typical training issues (mode collapse) that affect generative models.

1.10 Justification of Novelty, Contribution and Importance

This work is novel for the following threemost important contributions:

1. **First Use of Real 3D Clinical Data:** 1.Unlike previous studies with 2D slices or with synthetic volumes, this is the first study to tackle the problem with actual, paired, 3D clinical volumes, with results that directly apply to practical application.

2. **Novel Self-Critique Architecture:** The proposed `UNetSelfCritique3D` architecture mitigates mode collapse to some and gives promising results as compared to normal GANs.

3. **First Hybrid GAN-Diffusion Pipeline:** The synergistic combination of a stabilized GAN for coarse structure and a diffusion model for fine detail is a new and powerful paradigm for cross-modality synthesis.

This work bridges an essential gap through the proof of a practical route to that previously deemed impractical. Its benefit is an eventual benchmark and a new methodology that will facilitate the building up to deployable systems at the bedside.

1.11 Organization of the Thesis

Seven chapters are included in this dissertation. Chapter 2 is the comprehensive literature review that applies to this dissertation. Chapter 3 describes the preprocessing of the data and the usage of the baseline model. Chapter 4 outlines the recently introduced UNetSelfCritique3D architecture. Chapter 5 describes the process of refinement using diffusion. Chapter 6 presents the qualitative and quantitative results. Finally, Chapter 7 concludes the dissertation and indicates the direction for future work.

Chapter 2

LITERATURE REVIEW

2.1 Overview

The foundation of any research is a major breakthrough to future innovations. In this chapter we delve into the exciting work of AI in medical imaging using the current research generative networks particularly Generative Adversarial Networks and Denoising Diffusion Models we explore how these two key innovations have successfully shifted the medical imaging arena in image-to-image translation which would have otherwise been impossible to achieve. By analyzing the existing work, we find out their key strengths, limitations and elaborate how our approach binds this gap.

2.2 Related Work

Since Goodfellow et al. [1] introduced Generative Adversarial Networks, they've become essential tools for medical image synthesis. Deep learning has dramatically improved cross-modality translation, especially through frameworks like Pix2Pix [8] and CycleGAN [9]. Researchers have successfully applied these methods across various medical imaging tasks—converting MRI to CT, PET to MRI [10] [11], and enhancing X-ray images through approaches like TomoGAN [12] and SCAN [13]. Furthermore, Nie et al. [14] demonstrated high-fidelity CT synthesis from MR using a 3D GAN architecture, while Wolterink [15] applied adversarial training for CT denoising with promising results. These successes have largely been confined to modalities with well-aligned anatomical structures and abundant paired datasets.

For ultrasound and CT specifically, early attempts like Yin et al. [16] used GANs to create ultrasound-like images from CT scans. However, their method generated synthetic ultrasound from anatomical models and only worked with 2D slices. Vedula et al. [17] took a different approach, using supervised CNNs to convert 2D ultrasound into synthetic CT images. While they tackled US-to-CT directly, they simulated ultrasound from CT volumes to create training pairs instead of using actual clinical scans. This sidesteps the acoustic challenges inherent in real ultrasound and limits how well the model handles real-world data. Recently, Noack et al.

[18] worked on thyroid screening using 3D ultrasound from the SegThy dataset. Though they used real ultrasound volumes, their method generated CT by assigning predetermined intensity values to segmented regions—essentially avoiding the harder problem of learning actual CT tissue appearance. Despite their limitations, these studies prove that GANs can learn the complex relationships between different medical imaging modalities.

Denoising Diffusion Probabilistic Models [2] currently represent the cutting edge for generating highly realistic images. Their strength lies in capturing fine textural details, which has made them popular for medical imaging tasks like creating synthetic training data and detecting anomalies. Unfortunately, their slow inference speed [19] makes them impractical as standalone synthesis models.

2.3 Limitations of Existing Work

Previous research differs from ours in several important ways. Much of it focuses on 2D image slices, losing the 3D structural information crucial for clinical use. Other studies rely on simulated ultrasound data, which sidesteps the real challenge of handling clinical noise and artifacts. As far as we know, no one has successfully developed a reliable system for translating actual 3D transabdominal ultrasound scans into CT volumes.

Our literature review identifies four major obstacles that have prevented US-to-CT synthesis from reaching clinical practice:

1. **2D Limitations:** Working with individual slices misses the volumetric context needed for surgical planning, radiation therapy, and organ volume measurements. These clinical tasks demand full 3D anatomical understanding.

2. **Synthetic Data Gap:** Training on simulated or heavily cleaned ultrasound data doesn't prepare models for the messy, artifact-filled images clinicians actually work with daily.

3. **GAN Instability:** The huge difference between ultrasound and CT makes standard GANs particularly unstable. This often causes mode collapse [4], where models generate blurry, generic outputs with no clinical value.

4. **Diffusion Model Speed:** While diffusion models create excellent images, their iterative generation process is too slow for practical 3D volume synthesis.

2.4 Research Contribution: Addressing the Gap

This thesis tackles these problems through two complementary innovations:

1. **Stabilizing 3D Generation:** Our UNetSelfCritique3D architecture directly addresses the instability problem (Limitation 3) through a self-critique mechanism that doesn't rely on discriminator gradients. By working with complete 3D volumes on real clinical data, it also solves the issues with 2D and synthetic approaches (Limitations 1 and 2).

2. **Balancing Speed and Quality:** Our hybrid pipeline finds the sweet spot between GAN efficiency and diffusion quality. The stabilized GAN handles fast coarse generation, while the diffusion model refines fine details. This avoids the impractical computational cost of pure diffusion methods (Limitation 4) while achieving the high-fidelity textures that GANs alone cannot produce.

Chapter 3

DATASET AND BASELINE MODEL IMPLEMENTATION

3.1 The TRUSTED Dataset: From Raw to Paired Volumes

The successful development of any data-driven model hinges on the quality of the training data. For the specific problem of 3D US-to-CT translation, the **TRUSTED (Tridimensional Renal Ultra Sound TomodEnsitometrie)** dataset is an invaluable and unique resource. It is the first publicly available dataset containing paired, co-registered 3D transabdominal ultrasound and CT images of human kidneys from 48 patients, with real clinical data that makes it the ideal testbed for developing a model with genuine clinical relevance. The paired nature of the dataset is perfectly suited for a supervised translation framework like Pix2Pix, while its 3D volumetric data enables the development of models that respect true anatomical continuity.

The TRUSTED dataset, as described in the original publication by Ndzimbong et al.[7], is a unique and valuable resource. However, as the dataset is multi-purpose its raw format presents several challenges for direct, paired image-to-image translation. The CT scans are bilateral, containing both kidneys within a large abdominal volume, whereas each 3D ultrasound acquisition is unilateral, focusing on only one kidney. This could be infeasible and unrealistic to use the raw dataset so a significant effort was required to align, register and create a dataset of precisely aligned one-to-one volumetric pairs.

This chapter details the implemented two-step process undertaken in this research: first, the development of a robust preprocessing pipeline to transform the raw dataset as illustrated in (Figure 3.1) into a format suitable for supervised 3D image translation (Figure 3.5). At later portions we will apply the refined paired dataset for implementation of the baseline 3D architecture.

3.1.1 Kidney ROI Extraction from Bilateral CT Scans

The first step was to isolate the relevant kidney portion from each bilateral CT volume. We used to US mask to findout corresponding kidney side in the CT mask by matching the kidney's region of

interest (ROI) of both volumes so the prefered side was identified and extracted. Forinastance the right kidney of patient ID "200" was easily identified by using that of US mask right side kidney matching with the bilateral CT mask. This ensured that each US volume had a corresponding CT volume of the exact same organ side thus eliminating extraneous anatomical information.

3.1.2 The Registration and Alignment Pipeline

The US and CT are taken with different instruments, probes, body orientations and random organ movement which can't otherwise deterministic so although the kidney is for patient 'X' these two 3D versions will exist in different coordinate systems despite being same organ. To overcome this one image modality had to be registered into the other, for our case we registered US into CT sapce (though the inverse would have aswell yielded the same results). A multi-stage registration pipeline was implemented to transform the US into CT coordinate space. By using the 3D surface meshes derived from the organ segmentation masks, providing a computationally efficient, fast and robust representation for alignment.

Step 1: Initialization (Coarse Alignment)

Direct fine-grained registration between 3D medical images (ultrasound and CT kidney meshes) can fail if the initial misalignment is too large. To address this, we performed a coarse pre-alignment using Principal Component Analysis (PCA), a statistical technique. It was first introduced by Hotelling et al. [20] in 1933 and later elaborated for modern data analysis in 2011 by Jolliffe et al. [21]. PCA identifies the principal axes of the kidney meshes, corresponding to the directions of maximum variance in their 3D geometry. Aligning these principal axes provides an estimate of the orientation of each mesh. This simple but critical procedure saves alot of computataion time and resources thus making it a perfect for pre-alignment.

In parallel, the centroids of the meshes were computed to determine their position in space. As depicted in Figure 3.1, this PCA-based coarse alignment reduced the initial misalignment, ensuring that subsequent fine-grained registration methods, such as iterative closest point (ICP), converge more reliably and avoid local minima.

Step 2: Rigid Refinement (Fine-Tuning)

Following the coarse alignment, the Iterative Closest Point (ICP) [22] algorithm was employed to iteratively refine the alignment of the US and CT kidney meshes. ICP was used as the second refiner though is constrained to a rigid transformation (rotation and translation) to meticulously fine-tune the overlay without altering the organ size or shape. During each iteration, corresponding points between the two meshes were identified based on proximity, and the optimal rigid

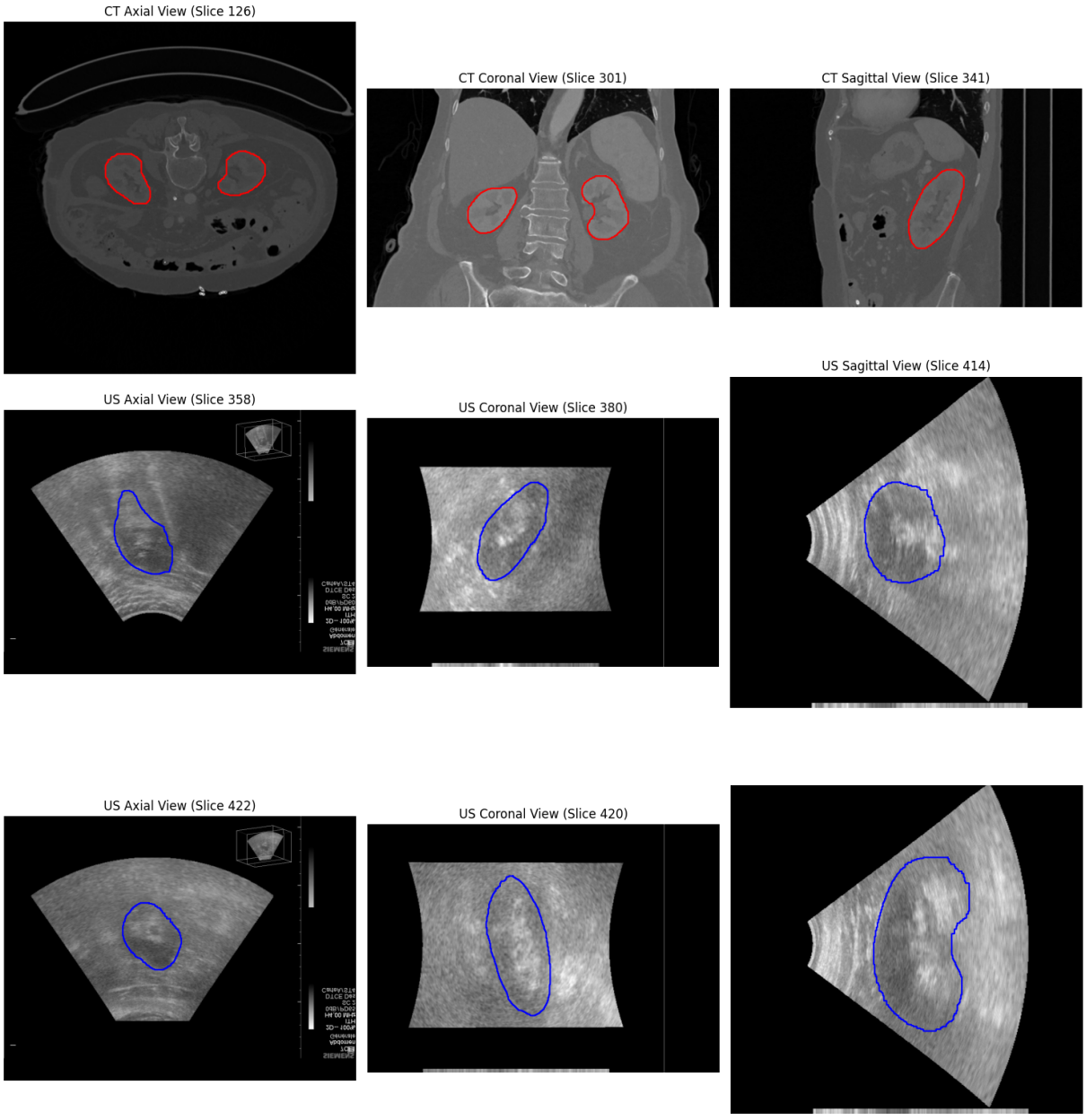


Figure 3.1: Different views of the TRUSTED dataset for patient ID “206”. Row 1: Bilateral CT with kidney region (red contours). Row 2: Right-side unilateral US. Row 3: Left-side unilateral CT.

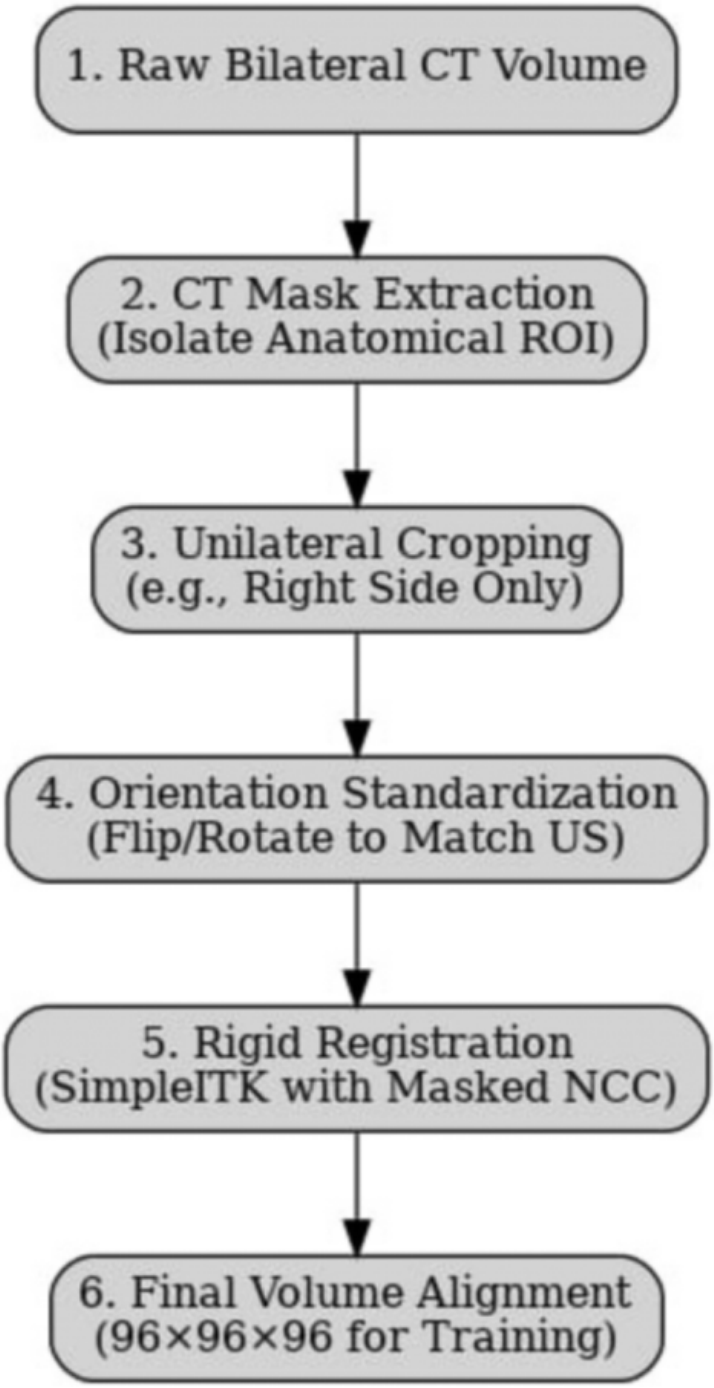


Figure 3.2: Step 1: PCA-based initialization for coarse alignment.

transformation that minimized the mean squared distance between these points was computed. This step ensured a precise registration, providing a reliable starting point for any subsequent non-rigid refinement or analysis.

Step 3: Affine Refinement (Final Adjustment)

To address more complex, non-linear differences caused by probe pressure or anatomical variation, a final affine refinement was applied. The affine transformation extended the rigid model by including anisotropic scaling and shearing, providing the best possible spatial correspondence between the US and CT volumes.

The TRUSTED dataset provides annotated masks for both CT and US images. Using the STAPLE algorithm [23], meshes were obtained and registration (Steps 2 and 3) was performed on these meshes, later applied to both images and masks. The final results are shown in Figure 3.3.

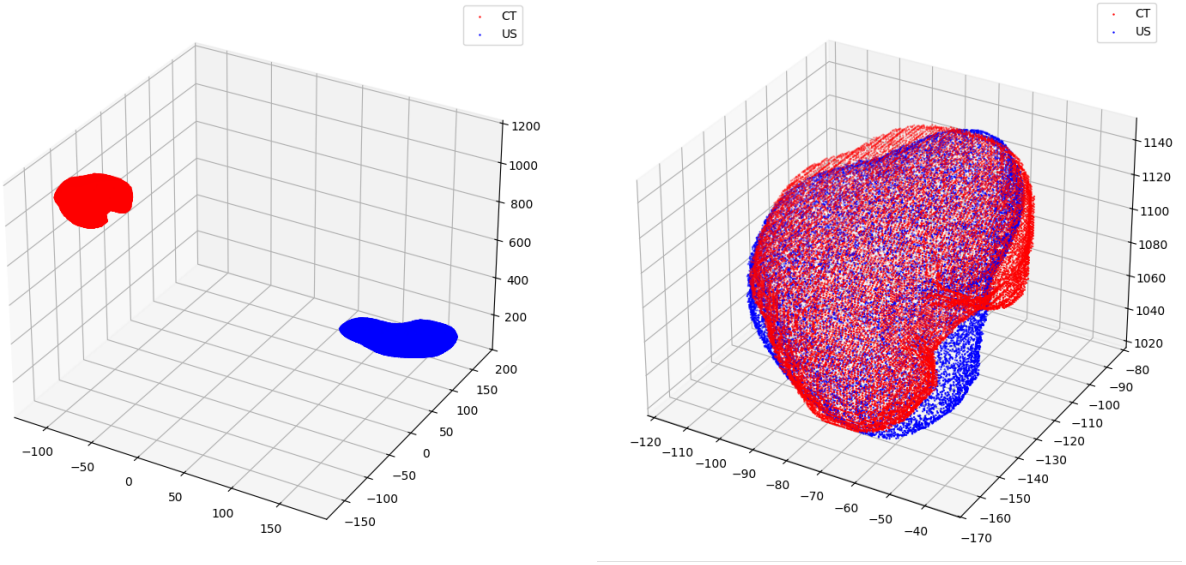


Figure 3.3: Figure 3.3 (a) shows the initial 3D graphical positions of individual kidneys for patient ID "200" which was possible using meshes generated from corresponding CT and US masks using STAPLE algorithm as shown in section 2.5. In fig3.3 (b) after registration we can see that the US (blue) is now registered and transferred to CT domain as observed from the scale ranges

The above was done on on the images and masks and transformation to get final paired transform which was used for Training, validation and evaluation processes.

3.1.3 Creation of the Final Paired Dataset

Upon completion of the registration pipeline, the final transformation matrix for each US–CT pair was saved. This matrix was applied to the original US intensity volumes and masks, warping

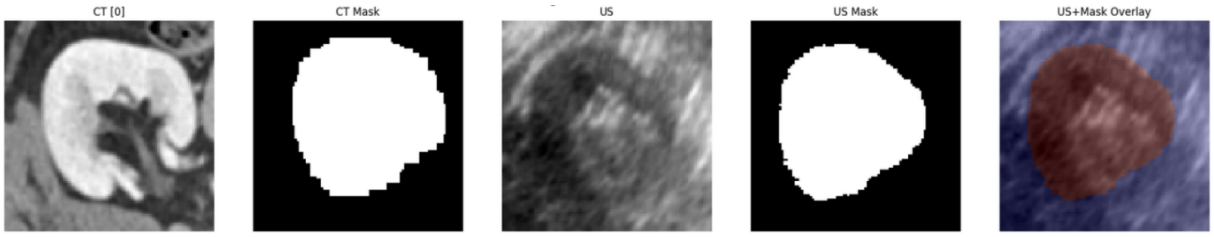


Figure 3.4: Final prepared TRUSTED dataset after registration and preprocessing.

them into the CT space. The result was a curated dataset of 59 perfectly co-registered pairs, each containing: (1) a 3D CT volume, (2) its CT mask, (3) a 3D US volume, and (4) its US mask. This processed dataset formed the foundation for all model training and evaluation.

3.2 Baseline Model: 3D Pix2Pix GAN

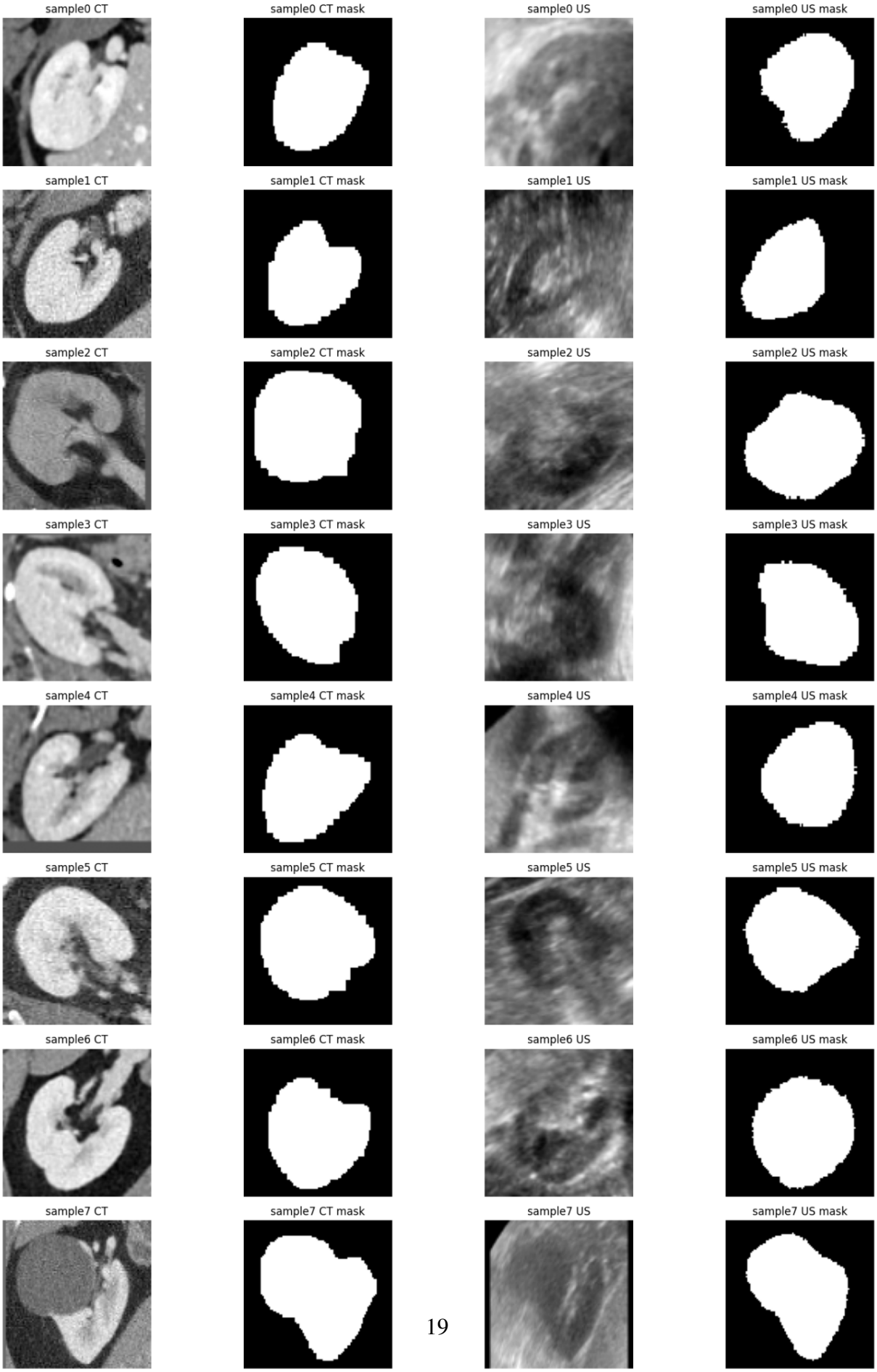
3.2.1 A Review of System Design

Image-to-image translation refers to the task of transforming an image from a source domain (e.g., ultrasound) to a target domain (e.g., CT) while preserving the underlying content. Early approaches relied on statistical methods or manual feature engineering, but the field was revolutionized by the advent of deep learning. Deep neural networks, particularly convolutional architectures, have demonstrated a remarkable ability to learn the complex, non-linear mappings required for such transformations directly from data.

3.2.2 Generative Adversarial Networks (GANs) in Medical Imaging

Among the most powerful image synthesis models are GANs [1]. A GAN consists of a Generator and a Discriminator locked in a competitive, zero-sum game. The Generator's goal is to create synthetic data (e.g., a fake CT image) that is indistinguishable from real data. The Discriminator's goal is to correctly identify whether a given image is real (from the training dataset) or fake (produced by the Generator). Through this adversarial training, the Generator becomes progressively better at producing realistic images to fool the Discriminator. Its more like cat and mouse game till ultimately Generator succeeds.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x,y)] + \mathbb{E}_{y\sim p_{data}(y),z\sim p_z(z)}[\log(1-D(G(y,z),y))] \quad (3.1)$$


sample0 CT
sample0 CT mask
sample0 US
sample0 US mask
sample1 CT
sample1 CT mask
sample1 US
sample1 US mask
sample2 CT
sample2 CT mask
sample2 US
sample2 US mask
sample3 CT
sample3 CT mask
sample3 US
sample3 US mask
sample4 CT
sample4 CT mask
sample4 US
sample4 US mask
sample5 CT
sample5 CT mask
sample5 US
sample5 US mask
sample6 CT
sample6 CT mask
sample6 US
sample6 US mask
sample7 CT
sample7 CT mask
sample7 US
sample7 US mask

3.2.3 The Pix2Pix Framework for Paired Translation

For tasks where paired images exist—meaning there is a direct one-to-one correspondence between a source and target image—the Pix2Pix framework is a highly effective conditional GAN (cGAN) architecture. The Generator in Pix2Pix is not given random noise but a specific conditioning image or the source image(US volume in our case). Its task is to learn the correlation and mapping from this specific input to its corresponding target (the CT volume). Such that the output is not only realistic but also a faithful translation of the input, to achieve this different losses have to be used to modulate the model in achieving desired output. The adversarial loss is combined with a direct reconstruction loss, such as an L1 or L2 distance between the generated image and the ground truth. This dual-objective function forces the Generator to produce images that are not only plausible to the Discriminator but also structurally similar to the target.

As this is the first work we had to establish a benchmark for performance, a 3D implementation of the Pix2Pix framework was chosen as the baseline model. This is a well-established architecture for paired image translation, and its performance on this task provides a clear measure of the problem's inherent difficulty as we will see in later chapters.

3.2.4 Generator Architecture: 3D U-Net with Residual Blocks

The generator as earlier explained has to learn the complex mapping from the US domain to the CT domain. Its architecture is based on a 3D U-Net [24], a design that has proven highly effective for image-to-image translation tasks due to its ability to simultaneously capture both high-level contextual information and low-level anatomical details through its symmetric encoder-decoder structure with skip connections. The U-Net architecture was originally proposed by Ronneberger et al. [25] in 2015 for biomedical image segmentation, demonstrating remarkable performance with limited training data. The concept of 3D medical imaging was explored by Çiçek et al. [24], replacing 2D convolutions thus enabling 3D volumetric data processing effectively. Their work demonstrated that 3D U-Net could learn volumetric representations while maintaining the original architecture's key strength: skip connections that preserve spatial information during downsampling. Simillar approach was adopted in this research to test the capability of a 3D model to implement the conversion using the default architecture. The sole component arrangement are described below.

- **Encoder–Decoder Structure:** The network consists of an encoder which performs the downsampling process to extract relevant features from the input by progressively downsampling the input volume to capture the low and high contextual features which are obtained in early and latter layers respectively. The decoder does the inverse by upsampling these features to reconstruct the output volume. In order to obtain proper mapping of fine

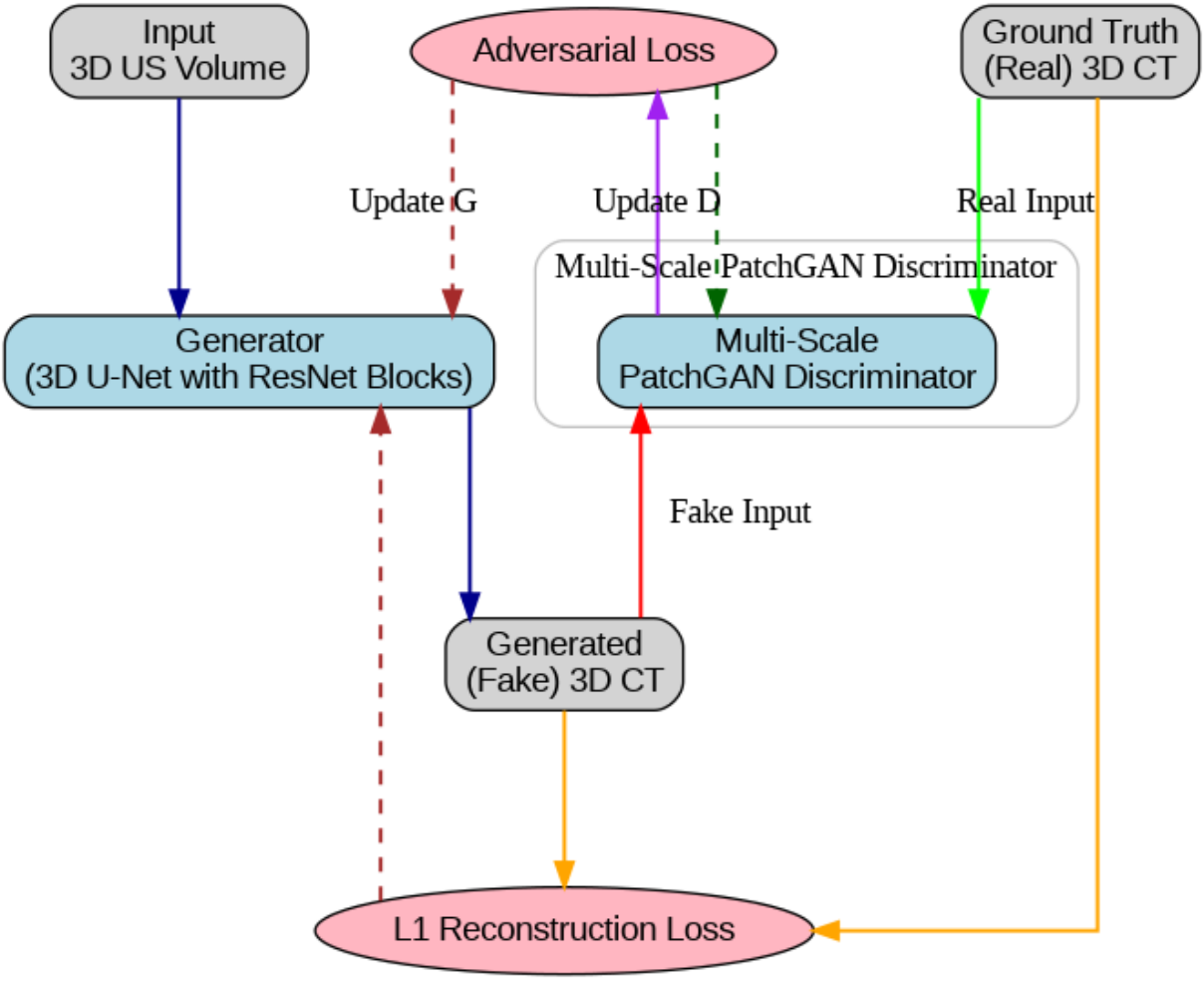


Figure 3.6: Block diagram of the baseline 3D Pix2Pix GAN architecture.

grained details obtained by encoders during downsampling, there has to be a link to the decorder which are the skip connections. These link corresponding layers in the encoder and decoder thus allowing fine-grained details from early layers to be passed directly to the reconstruction stage during upsampling which is very crucial for preserving anatomical boundaries.

- **Residual Bottleneck:** The bottleneck of the U-Net is where most compression of the feature representation takes place. The core idea is combining a series of residual blocks (ResNets) [26]. These block utilize a skip connection within the block itself hence allowing the network to learn a ”residual” mapping. This computationally eases the training of very deep networks and prevent the degradation of performance.

3.2.5 Discriminator Architecture: Multi-Scale PatchGAN

The discriminator's role is to distinguish between real CT volumes and synthetic ones (produced by the generator). The overall objective is not to be fooled by the generator but also provide necessary gradients (guidance) to the generator. In this work we used a Multi-Scale PatchGAN discriminator to provide a robust training signal.

- **PatchGAN Concept:** Traditional discriminators always classify the entire signal once which would mean classifying the entire 3D volume with a single real/fake label, the generator would undermine local texture and details. A PatchGAN discriminator outputs a 3D grid of predictions. Each value in this grid corresponds to a ”patch” of the input volume, classifying that local region as real or fake which outperforms the traditional discriminators.

- **Multi-Scale Enhancement:** In image conversions where source and destination have cordially less resemblance as in our case, a single discriminator would struggle to analyze entire 3D volume so concept of multi-scale arrised to make sure that at every step the entire volume can be investigated. In this model, three separate PatchGAN discriminators were used where one operates on full-resolution input, while the other two operate on versions of the input downsampled by factors of 2 and 4. This ensures that both global and local levels have an impact on the discriminator's loss which guides the generator towards realism.

3.3 Loss Function and Training Strategy

A Generative Adversarial Network training is a balancing act that is guided by carefully designing a set of objective functions that gets the best out of both generator and discriminator. However,

this is not as simple as it reads since the generator has to try to fool the discriminator and the discriminator must not allow itself to be fooled. This makes the training a delicate balancing act so the 3D Pix2Pix baseline model is trained using a composite loss function that is designed to enforce both realism and content fidelity. This learning is governed by two distinct loss functions, one for the discriminator and another combined for the generator.

3.3.1 Discriminator Loss (L_D)

The sole objective of Discriminator is to become an expert at distinguishing real CT volumes from those generated by the generator. Its loss is the sum of two components, typically calculated using a Binary Cross-Entropy (BCE)

- **Loss on Real Images:** The discriminator is fed a pair of a real US volume and its corresponding ground truth CT volume. It is trained to output a high value (close to 1), signifying "Real."
- **Loss on Fake Images:** The discriminator is fed a pair of a real US volume and the fake CT volume generated from it. It is trained to output a low value (close to 0), signifying "Fake."

The total discriminator loss is the average of these two losses. For our multi-scale architecture, the final L_D is the sum of the individual losses from the full-scale, half-scale, and quarter-scale discriminators.

3.3.2 Generator Loss (L_G)

The generator has a more complex, dual objective. It must not only create realistic images that can fool the discriminator but also ensure those images are a faithful translation of the input US volume. Therefore, its loss function is a weighted sum of these parts:

- **Adversarial Loss:** This loss measures how well the generator is fooling the discriminator. The generator wants the discriminator to see the fake CT it produced and classify it as "Real" (output a 1). The adversarial loss is therefore the BCE loss of the discriminator's output on the fake image with a target label of 1. This pressure forces the generator to learn the textures and characteristics of a real CT scan.
- **Reconstruction Loss:** To ensure the generated output is not just a random realistic CT but is structurally equivalent to the ground truth, a direct reconstruction loss is applied. We use the L1 Loss (Mean Absolute Error) between the generated CT volume and the ground truth CT volume. L1 loss is preferred over L2 (Mean Squared Error) in image-to-image

translation tasks as it has been shown to produce less blurry results and is more robust to outliers.

- **Feature Matching Loss:**To compare intermediate discriminator activations for real vs generated images, encouraging the generator to produce samples with similar internal statistics

The generator minimizes:

$$L_G = L_{\text{adv}} + \lambda\, L_{\text{rec}} + \beta\, L_{\text{FM}} \tag{3.2}$$

where λ and β are hyperparameters that control the balance between the two losses. A higher λ value (e.g., 100) places more importance on the L_1 reconstruction loss, strongly encouraging the generator to match the ground truth structure.

The generator and discriminator are trained in an alternating fashion, a process that is repeated for each batch of data across many epochs:

1. **Train the Discriminator:**
    - The generator's weights are frozen.
    - A batch of real US volumes is passed through the generator to create a batch of fake CT volumes.
    - The discriminator is trained for one step on the real CT volumes (with target labels of 1) and one step on the fake CT volumes (with target labels of 0).
    - The gradients are calculated, and the discriminator's weights are updated via backpropagation.
2. **Train the Generator:**
    - The discriminator's weights are frozen.
    - A batch of real US volumes is passed through the generator to create a batch of fake CT volumes.
    - These fake volumes are passed to the discriminator. The adversarial loss ($\mathcal{L}_{adv}$) is calculated based on how far the discriminator's predictions are from the "Real" target of 1.
    - The L1 reconstruction loss ($\mathcal{L}_{rec}$) is calculated between the fake CTs and the real CTs.
    - The total generator loss ($\mathcal{L}_G$) is computed, and its weights are updated via backpropagation.

This iterative, adversarial process continues, with the discriminator and generator progressively improving at their respective tasks. The entire model was implemented using the PyTorch framework [27] and trained using the Adam optimizer [28], a standard choice for GANs due to its adaptive learning rate capabilities.

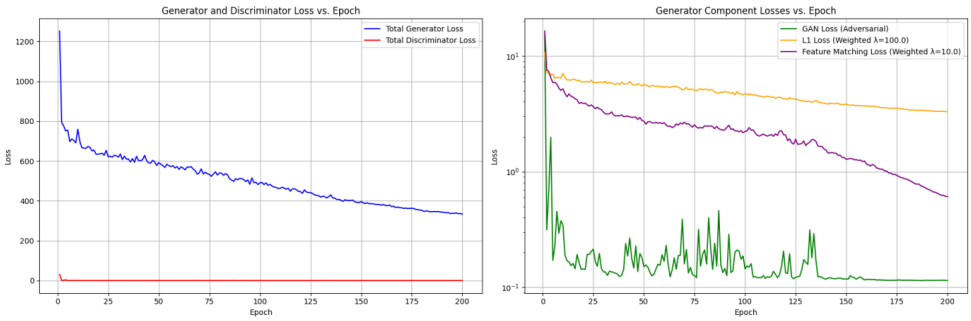


Figure 3.7: The training loss curves of the baseline model. From the above, the discriminator loss is seen to be constant throughout the training. This happens because discriminator becomes too confident of its predictions in such a way that Generator can't fool it thus stops giving desirable gradients to Generator leading to mode collapse.

3.4 Initial Experiments and Identified Limitations

The baseline 3D Pix2Pix model was trained on the prepared dataset until its loss curves converged. While the model successfully learned to produce outputs that were broadly kidney-shaped and possessed CT-like intensity profiles, its performance was ultimately hindered by a critical and well-known GAN failure mode.

3.4.1 The Problem of Mode Collapse

Upon qualitative inspection of the results, it became evident that the baseline model suffered from significant mode collapse. The generator, in its effort to find the easiest way to fool the discriminator, learned to produce a very limited variety of outputs. Regardless of the unique anatomical structure or noise pattern in the input US volume, the generator often produced a generic, "average-looking" kidney. The subtle but clinically important variations between patients—such as differences in the renal pelvis, vasculature, or overall morphology—were lost in translation.

This failure represents a fundamental inability of the baseline model to learn the true complexity of the US-to-CT mapping. The discriminator provided a strong enough signal to enforce general realism but was not sufficient to enforce a faithful, one-to-one structural translation for every unique input.

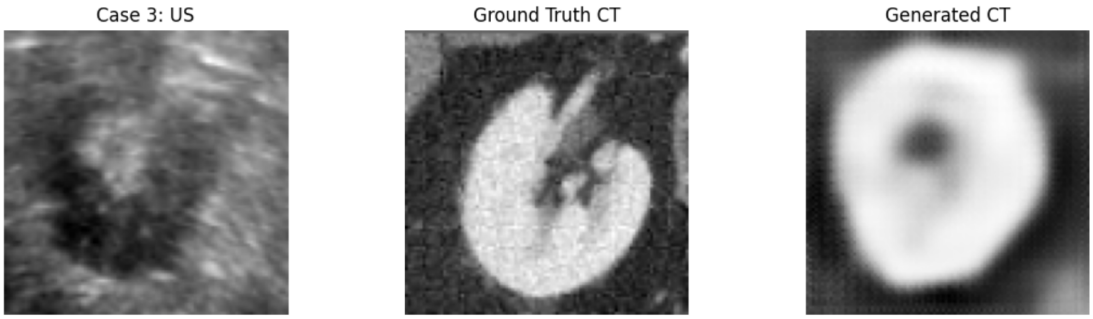


Figure 3.8: Evidence of Mode Collapse. A distinct US input slices on the left, GT CT and the nearly identical, and blurry CT output slices from the baseline Pix2Pix GAN.

This critical limitation of the baseline model serves as the primary motivation for the work presented in the next chapter. The need for a more sophisticated generator architecture which can have an internal sense of structural consistency and also guided by an external adversary loss. we sahall see how we overcome this obstacle in the next chapter.

Chapter 4

THE UNETSELFCRITIQUE3D GAN

The baseline 3D Pix2Pix GAN was trained as explored in previous chapter however, while the performance initially gives compelling and conceptually sound results the model under-performs in this task of US-to-CT translation as explicitly seen in Figure 3.8 where the output is a blurry image with less correlation to the target image leading to poor metric evaluation report (view chapter 6). This initially came as a surprise to us but further investigation showed the inherited problem of GANs known as model collapse. This is not merely an implementation flaw but a deeper, more fundamental limitation in the standard adversarial training dynamic where the discriminator as an external critic can effectively learn to identify what a realistic CT volume looks like in general since its work is abit easier in normal GAN architecture thus becomes over-confident. Well this is what we primarily train it to do but the problem with this is that while in this state it provides an insufficient and often unstable signal to enforce a faithful, structurally-specific translation for every unique input. The generator, in turn, exploits this by discovering a "safe" and outputs "a generic, blurry average" that consistently passes the discriminator's simple test with minimum loss.

To overcome the model collapse issue, researchers have investigated various strategies such as feature matching proposed by Salimans et al. [29] here, the generator is trained to match the statistics of intermediate features in the discriminator instead of directly maximizing the discriminator output. Metz et al. [30] introduced the unrolled GAN, which allows the generator to anticipate the discriminator's future updates promoting diversity in generated outputs also some work on spectral normalization [31] has been employed to stabilize discriminator training by constraining its Lipschitz constant thereby providing more reliable gradients to the generator. Additionally, progressive growing techniques [32] and self-attention mechanisms [13] have shown promising results in maintaining output diversity while improving image quality.

In our training we incorperated some of the well known techniques suggested but still got simillar results. We came up with an idea that what if we eliminate the need of generator relying on the external discriminator but rather be endowed with a mechanism to critique its own work which further omits the reliance on insufficient and unstable signals. This chapter introduces the UNetSelfCritique3D, a novel generator architecture designed with an intrinsic mechanism for self-assessment, ensuring that the structures it synthesizes are consistent with the structures it

analyzes.

4.1 Motivation: Overcoming the Limits of Traditional Discriminators

The core idea behind the UNetSelfCritique3D stems from a simple question: How can we ensure the generator pays attention to the input? The external discriminator penalizes unrealistic outputs but struggles to penalize outputs that are merely incorrect for a given input. We made various attempts by incorperating feature-matching loss, the results didn't show much improvement. The solution in our proposed architecture lies in creating an internal consistency check within the generator itself.

The U-Net architecture is naturally divided into two symmetric parts: an encoder that deconstructs the input image into a hierarchy of feature maps, and a decoder that reconstructs an output image from these features. The encoder acts as an analyzer, identifying anatomical structures at various scales within the input US volume. The decoder acts as a synthesizer, building the output CT volume layer by layer.

In a standard U-Net, these two paths are only linked by skip connections. We propose that they can be linked more powerfully through the loss function. The central hypothesis is this: a generator producing a faithful translation should exhibit feature-level symmetry. The features representing a specific anatomical structure during synthesis in the decoder should closely match the features that identified that same structure during analysis in the encoder. By enforcing this symmetry, we provide the generator with a powerful, stable, and internal training signal that directly combats mode collapse.

4.2 Architecture of the UNetSelfCritique3D Generator

The foundation of the UNetSelfCritique3D is the well-established 3D U-Net architecture [24], a related concept is the Attention U-Net [33], which learns to focus on salient regions chosen for its proven efficacy in dense, volumetric prediction tasks. Its structure is composed of:

- **The Encoder Path:** A series of convolutional blocks that progressively downsample the spatial dimensions of the input US volume. Each block doubles the number of feature channels, allowing the network to learn increasingly complex and abstract representations of the anatomy. The feature maps at each level (e.g., d_1, d_2, d_3, d_4) represent the "analysis" of the input at different scales.

- **The Bottleneck:** The deepest point of the network, where the feature representation is most abstract and spatially compressed.
- **The Decoder Path:** A series of transposed convolutional blocks that progressively upsample the feature representation, rebuilding the spatial dimensions to match the desired output size. From our design the feature maps at each level of the decoder (e.g., u_1, u_2, u_3, u_4) represent the "synthesis" of the output CT.
- **Skip Connections:** These connections link corresponding levels of the encoder and decoder (e.g., d_4 is concatenated with u_4). They are crucial for allowing the decoder to access fine-grained, low-level information from the encoder, which is essential for reconstructing sharp anatomical boundaries. We stil maintained this in our proposed architecture.

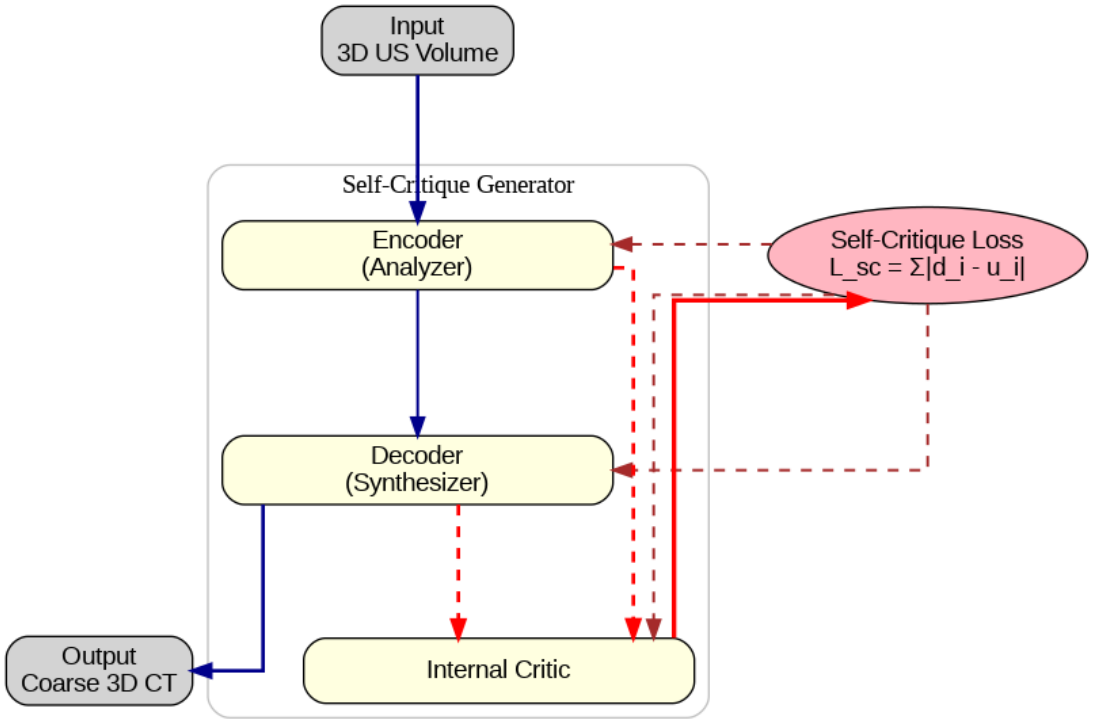


Figure 4.1: The UNetSelfCritique3D architecture. The diagram shows the standard U-Net shape with the encoder and decoder levels. Including arrows pointing from each corresponding feature map pair to "Self-Critique Loss," illustrating the comparison.

The key architectural innovation of the UNetSelfCritique3D is not a change in the layers themselves, but in their function. The model is designed to be simple and explicitly return the intermediate feature maps from both the encoder and decoder paths during a forward pass. This exposes the internal "thought process" of the network, allowing it to be harnessed directly by the loss function.

4.3 The Self-Critique Loss Function

The self-critique mechanism is formalized as a new component of the generator's total loss function. This Self-Critique Loss ($\mathcal{L}_{sc}$) is defined as the measure of dissimilarity between the corresponding feature maps of the encoder and decoder. We use the L1 distance (Mean Absolute Error) for its robustness and tendency to promote less blurry results compared to the L2 distance.

The loss is calculated as follows:

$$\mathcal{L}_{sc} = \sum_{i=1}^{N} \|d_i - u_i\|_1 \tag{4.1}$$

where d_i is the feature map from the i-th level of the encoder and u_i is the feature map from the corresponding i-th level of the decoder.

This self-critique loss is not used in isolation. It acts as a powerful regularizer, complementing the traditional GAN losses. The complete loss function for the generator ($\mathcal{L}_G$) is a weighted sum of four components:

1. **Adversarial Loss ($\mathcal{L}_{adv}$):** The standard GAN loss that encourages the generator to produce outputs that fool the discriminator, promoting realism.

2. **Reconstruction Loss ($\mathcal{L}_{rec}$):** An L1 loss between the final generated CT volume and the ground truth CT volume. This enforces pixel-level accuracy and global structural correctness.

3. **Feature Matching Loss ($\mathcal{L}_{fm}$):** Additional loss component for feature-level consistency.

4. **Self-Critique Loss ($\mathcal{L}_{sc}$):** The new loss term that enforces internal feature consistency, preventing mode collapse.

The total generator loss is therefore:

$$\mathcal{L}_G = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{fm}\mathcal{L}_{fm} + \lambda_{sc}\mathcal{L}_{sc} \tag{4.2}$$

The hyperparameters λ_{adv}, λ_{rec}, λ_{fm}, and λ_{sc} are weighting factors that balance the influence of each loss component on the final training objective.

4.4 Training and Implementation Details

The UNetSelfCritique3D model was trained using the Adam optimizer [28], a standard choice for GANs. The weighting parameters for the loss function were determined empirically to provide

a stable training dynamic, with the reconstruction and self-critique losses providing a strong, consistent gradient while the adversarial loss guided the refinement of textural realism.

During training, the inclusion of the $\mathcal{L}_{sc}$ term had an immediate and observable stabilizing effect. While the baseline model's loss was often erratic, indicative of the unstable cat-and-mouse game between the generator and discriminator, the UNetSelfCritique3D's loss curve demonstrated a much smoother and more consistent convergence. This stability is a direct result of the generator having a constant, internal objective to optimize for, independent of the discriminator's current state. The result is a model that not only converges more reliably but, as will be shown in Chapter 6, produces outputs of significantly higher quality and faithfulness to the input, effectively mitigating the mode collapse that plagued the baseline approach. However, further research to enhance these properties needs to be taken into account.

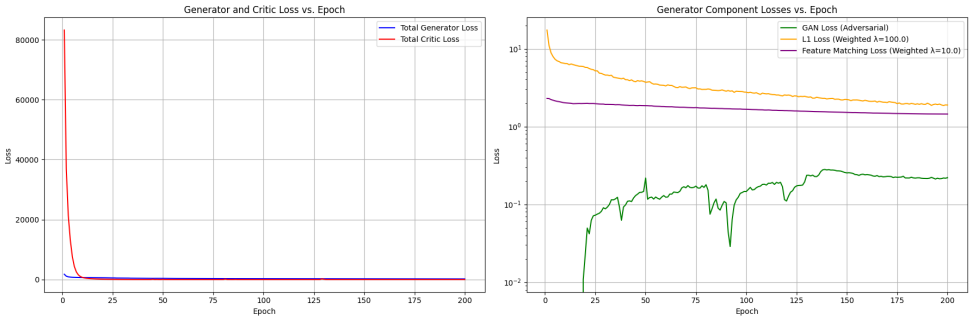


Figure 4.2: The training loss curves of UNetSelfCritique3D model.

Chapter 5

DIFFUSION-BASED RESIDUAL REFINEMENT

5.1 Denoising Diffusion Probabilistic Models (DDPMs)

A more recent class of generative models, Denoising Diffusion Probabilistic Models (DDPMs) [2], has emerged and demonstrated state-of-the-art performance in high-fidelity image synthesis. This concept is related to Latent Diffusion Models which operate in a compressed space [34]. Diffusion models operate via two processes. The forward process systematically and gradually adds Gaussian noise to a real image over a series of timesteps, until it becomes pure, unstructured noise. The reverse process then trains a neural network (typically a U-Net architecture) to reverse this process. At each timestep, the network learns to predict and remove a small amount of noise, gradually transforming a random noise input back into a clean, coherent image.

5.1.1 Conditional Diffusion Models for Guided Synthesis

This reverse process can be conditioned on additional information. In a conditional DDPM, the denoising network is given not only the noisy image and the current timestep but also a conditioning input, such as a class label or another image. This allows the model to generate a specific output guided by the condition, rather than a random sample from the data distribution.

5.1.2 Application of Diffusion Models as Refiners

The UNetSelfCritique3D architecture, as detailed in the previous chapter, to some extent effectively solves the critical problem of mode collapse, yielding a generator capable of producing structurally sound and patient-specific CT volumes. The resulting images are a significant leap forward from the baseline, capturing the coarse anatomy with impressive fidelity. However, the inherent nature of GANs, which are often optimized with perceptual and reconstruction losses, can lead to a characteristic smoothness or a lack of fine-grained textural detail. The outputs may not fully replicate the subtle textural patterns and high-frequency information characteristic of a

real CT scan as the translation is from artifact US to more detailed CT we needed to add some more filter to achieve an output which is as close to real CT as possible.

To elevate the synthesis from "plausible" to "indistinguishable," we introduce intergrated a second complementary stage to our framework which is the core idea of this chapter. This chapter details the design and integration of a Conditional Denoising Diffusion Probabilistic Model (DDPM), a powerful generative model tasked not with creating an image from scratch, but with performing a targeted, residual refinement of the GAN's output.

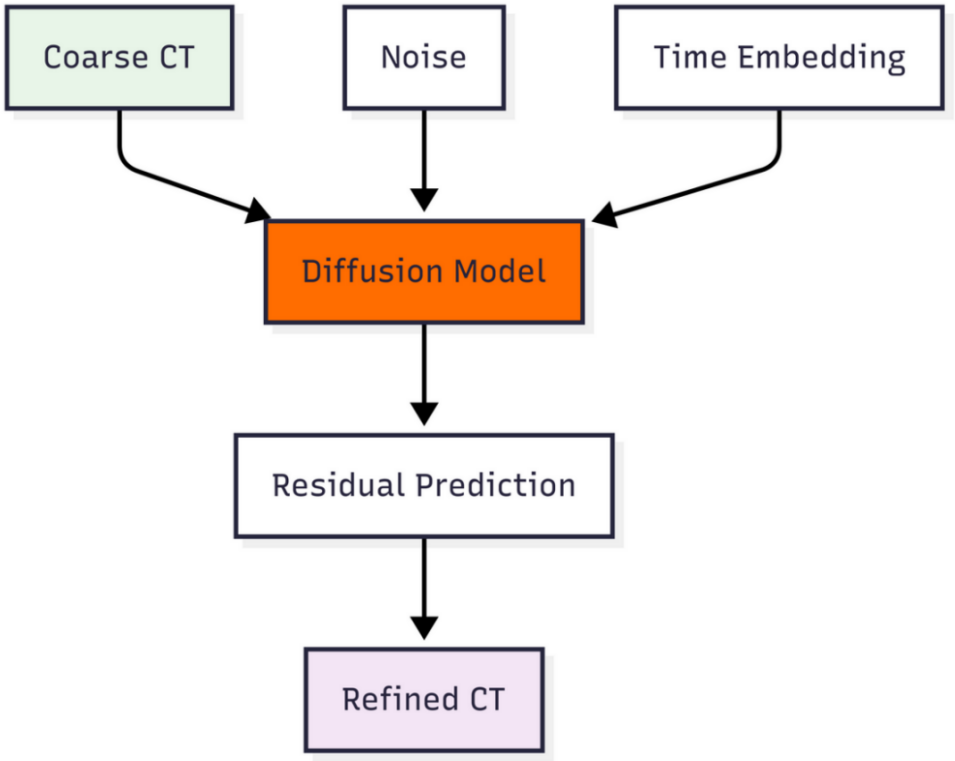


Figure 5.1: Flow Diagram of the DDPM.

5.2 Architecture of the Conditional 3D U-Net DDPM

The conditional 3D U-Net was the primary architecture employed in our design of the DDPM due to its ability to process information at different multiple scales hence specifically tailored for the iterative noise prediction task required by diffusion models. The core components of the model are explained below.

- **The U-Net Core:** In a denoising task, an architecture's ability to process information at multiple scales is essential and a foundational feature the model must recognize and remove

noise patterns that manifest differently at various levels of feature abstraction. This brings the U-Net srchitecture into consideration and thus we employed that as the fundamental block.

- **Time Embedding:** The current timestep t shows the awereness of the current noise margin which is vcrucial for a diffusion model in the denoising process. The model's behavior is dependent on whether it is removing a large amount of noise at an early step or refining tiny details at a late step. This awareness is achieved through a sinusoidal time embedding [35] mechanism. The integer timestep t is transformed into a high-dimensional vector, which is then projected and pasted into each residual block of the U-Net. This allows the network to modulate its features and condition its output on the current noise level.

- **Group Normalization:** Since we're working with 3D volumes, we implemented Group Normalization [36] within the network's residual blocks instead of the Batch Normalization as we're already working with smaller batches and Group Normalization is independent of batch size thus providing greater training stability for memory-intensive 3D volumes.

- **Conditioning Mechanism:** As the DDPM model isn't used as a generating model a primarily used we had to insert in a predefined condition to guide it during the refinement in order to keep the context and fidelity of the CT. In our case it worked as a refiner by further denoising the generated CT and inserting in higher contextual features which were other wise missed in the coarse CT. The goal thus was set with a conditioning mechanism. The "conditional" aspect of the model is key since denoising process isn't randomly but rather guided by the coarse CT scan. This conditioning input (`coarse_ct`) is concatenated with the noisy input x_{noisy} along the channel dimension at the very start of the U-Net making sure that at every denoising step the course anatomy is referenced thus keeping the model on track.

5.3 Rationale for a Hybrid Approach

The strategic choice to develop a hybrid GAN-Diffusion framework was made in order to capitalize on the unique and complementary advantages of each modeling domain. The proposed framework was made in such a way that we utilize the best from both models where a faster model like GANs produces a coarse but structurally correct output and the diffusion model then performs a "residual refinement," focusing only on adding the final layer of detail and realism. The traditional GAN's mode collapse issue was elevated by proposed UNetSelfCritique3D so enhancing the output we again extracted the exceptional fine detail and textural accuracy of

diffusion models in the same way avoiding the high computational cost due to iterative denoising process which requires hundreds or thousands of forward passes through the network to generate a single sample which at times may not aswell guarantee success in the output if the DDPM was solely used. Ultimately our proposed framework guided the input (US volume) into an accurate OUTPUT (CT volume) with a simple, effective and computationary stable structural. The motivation summery can be shortlisted briefly as

- **Generative Adversarial Networks (GANs):** GANs are computationally efficient at inference time as they only require a single forward pass to generate a full-volume output. They become the ideal candidates for producing a fast, structurally coherent "first draft" of the target image. Their primary weakness, as we've seen, is instability and a tendency to miss fine details. Our proposed model is even simpler as we dropped the discriminator's computations reducing alot of compuatations. The UNetSelfCritique3D outruns traditional GAN limitations.
- **Denoising Diffusion Models (DDPMs):** Diffusion models are capable of synthesizing images with unparalleled, high-fidelity detail and textural realism since they excel at capturing the true data distribution of image volumes correctly. Their primary weakness is computational cost caused by the iterative denoising process which requires hundreds or even thousands of sequential model evaluations, making them prohibitively slow for direct, end-to-end synthesis in many applications.

We build a synergistic pipeline by integrating them. The suggested GAN handles the "heavy lifting," quickly mapping the input US to a coarse CT volume that is subsequently used as input by the diffusion model as a strong prior. It then concentrates its significant power on the much easier, more limited task of learning the residual, which is the minute difference between the coarse GAN output and the ground truth. The diffusion model can achieve high fidelity in fewer denoising steps and with greater stability thanks to this "residual refinement," which is a far more manageable problem than creating a CT from scratch. This eliminates the need for costly computations and time.

5.4 The Full Hybrid Framework Pipeline

The complete inference pipeline integrates the pre-trained GAN and the trained diffusion refiner into a seamless, two-step process:

1. **Step 1: Coarse Synthesis.** An input 3D ultrasound volume is fed into the pre-trained UNetSelfCritique3D generator(which was trained). The generator performs a single, rapid

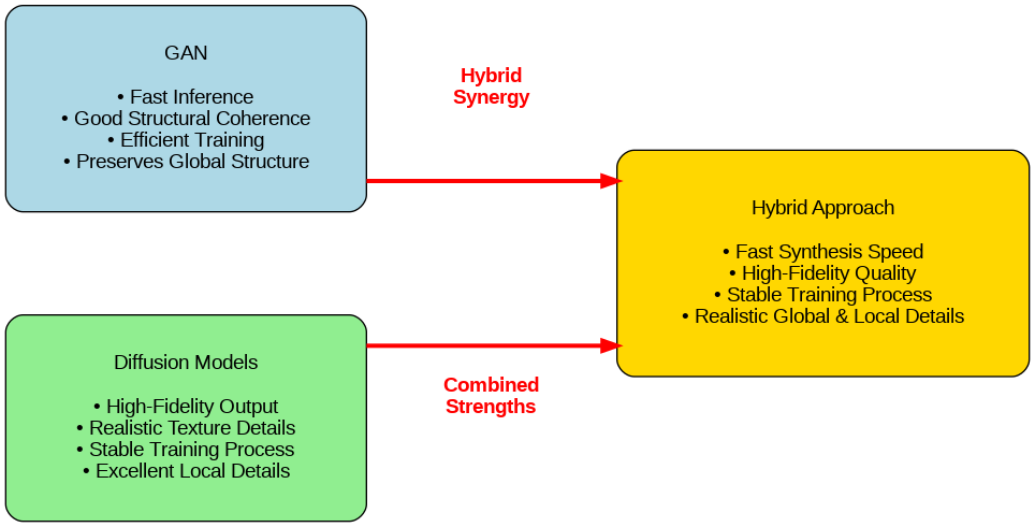


Figure 5.2: Hybrid GAN-Diffusion Conceptual Rationale.

forward pass to produce a `coarse_ct` volume. This volume is structurally accurate but may lack fine texture.

2. **Step 2: Iterative Refinement.** The diffusion process begins. A random Gaussian noise tensor of the same dimensions as the target volume is created. Then, for a set number of denoising timesteps (e.g., from T down to 1), the conditional U-Net is repeatedly called. At each step t, the network takes the current noisy volume and the `coarse_ct` as input and predicts the noise that was added at that step. This predicted noise is then subtracted from the noisy volume, yielding a slightly cleaner version, which becomes the input for step $t - 1$.

3. **Step 3: Final Output.** After the final denoising step, the result is the refined, high-fidelity CT volume. This volume retains the correct global structure from the GAN while incorporating the realistic, high-frequency details synthesized by the diffusion model.

5.5 Implementation and Training of the Diffusion Refiner

The conditional diffusion model is trained separately after the UNetSelfCritique3D generator has been fully trained and its weights are frozen. The training objective for the diffusion U-Net is simple and robust: to predict the noise that was added to an image.

For each training step:

1. A ground truth CT (`gt_ct`) and its corresponding `coarse_ct` (pre-generated by the GAN) are selected.

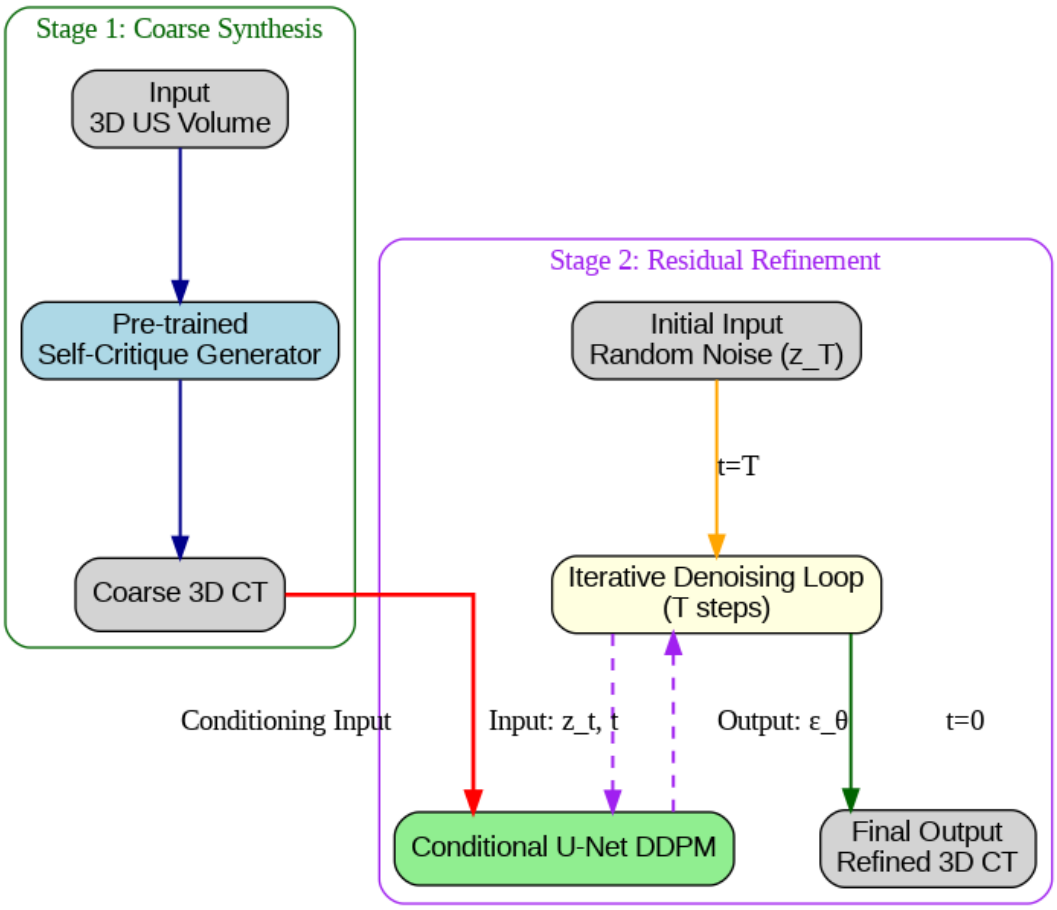


Figure 5.3: A detailed block diagram of the entire hybrid inference pipeline. Showing Input US to UNetSelfCritique3D, “Iterative Denoising Loop” with the Conditional DDPM.

2. A random timestep t is chosen.

3. The appropriate amount of Gaussian noise corresponding to timestep t is added to `gt_ct` to create a noisy version, x_{noisy}.

4. The U-Net is given x_{noisy}, the conditioning `coarse_ct`, and the time embedding for t.

5. The model outputs a `predicted_noise`.

6. A simple loss (e.g., Mean Squared Error) is calculated between the `predicted_noise` and the actual noise that was added in step 3. The network's weights are updated via backpropagation to minimize this loss.

By training on this objective across all possible timesteps and training samples, the network learns the complete denoising process, becoming an expert refiner ready to be deployed in the hybrid inference pipeline.

Chapter 6

RESULTS AND DISCUSSION

This chapter presents a comprehensive evaluation of the models developed throughout this thesis. The primary goal is to empirically validate the hypothesis that the proposed hybrid GAN-Diffusion framework offers a superior solution for 3D ultrasound-to-CT translation compared to both a standard baseline and the intermediate self-critiquing GAN. We present the results by carrying out both qualitative and qualitative assessments. The qualitative solely bases on visual inspection of the synthesized volumes where as the quantitative uses established image quality metrics. These assesements play a significant role in results analysis and thus the need to present both. The chapter concludes with an in-depth discussion interpreting these results, contextualizing their significance and acknowledging the limitations of the current work.

6.1 Quantitative Analysis

This involves the numerical evaluation of the model's performance using proved and acceptable and measurable statistical image-based similarity metrics. In this project, it is conducted to objectively assess the accuracy and fidelity of the generated CT images on all levels relative to the ground truth CT data.

6.1.1 Evaluation Metrics

The available literature relies majorly on five widely-used image quality metrics. Each metric captures a different aspect and specializes the correlation and similarity between the synthesized CT (CT_{synth}) and the ground truth CT (CT_{gt}). The objective was to provide a multi-faceted measure of performance and efficient evaluation as a measure of performance. Clear explantations and mathematical formulations behind these metrics is given (see Table: 6.1)

1. **Mean Absolute Error (MAE):** Also known as the L1 loss, MAE [37] measures the voxel-wise average absolute difference in the two volumes. It provides a direct and intuitive measure of reconstruction error. A lower MAE indicates a higher pixel-wise accuracy.

2. **Mean Squared Error (MSE):** MSE is another fundamental metric that measures the average of the squares of the errors between corresponding voxels. Here each difference is squared, MSE [37] places a significantly higher penalty on large errors (outliers) compared to MAE. This makes it particularly sensitive to instances where the synthesized image has voxels that are very different from the ground truth. A model with a low MSE is one that avoids making large, conspicuous errors. Like MAE, a lower MSE value indicates a better fit, with a score of zero representing a perfect reconstruction.

3. **Peak Signal-to-Noise Ratio (PSNR):** PSNR [38] is one of the most common metrics for evaluating the quality of reconstruction. By measuring the ratio between the maximum possible power of a signal (the dynamic range of the image intensities) and the power of the corrupting noise that affects its fidelity. A higher PSNR value [39] indicates a higher quality reconstruction with less error.

4. **Structural Similarity Index (SSIM):** The above metrics work well with voxel correspondance of both images however, they do not always align well with human visual perception. The metric used for this is SSIM [40]. This works by measuring the similarity between two images based on three components: luminance, contrast, and structure. as others its range is between -1 and 1, where 1 indicates a perfect match. SSIM is particularly valuable for this task as it can assess whether the anatomical structures in the synthesized volume are perceived similarly to the ground truth.

5. **Normalized Cross-Correlation (NCC)):** This is a similarity metric that quantifies the linear relationship between corresponding voxel intensities of two images — in this case, the synthesized CT and the ground-truth CT. It measures how well one image can be represented as a scaled and shifted version of the other, independent of absolute intensity differences.

The results as obtained are depicted in above table can show all the evaluations for each model. The remarks are:

- The **Baseline Model** establishes the lowest performance across all metrics, reflecting its poor reconstruction quality.

- The **UNetSelfCritique3D GAN** demonstrates a substantial improvement, with a significant reduction in MAE and a large increase in PSNR, SSIM and NCC. This improvement highlights the model's ability to escape mode collapse and produce sounding images near the ground truth.

Metric	Formula	Range	Interpretation
MAE	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$	$(0, \infty)$	Lower is better
MSE	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$	$(0, \infty)$	Lower is better
PSNR	$10\log_{10}\frac{MAX^2}{MSE}$	$(0, \infty)$ dB	Above 20 dB
SSIM	$\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$	$[-1, 1]$	Close to 1 is better
NCC	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$	$[-1, 1]$	Close to 1 is better

Table 6.1: Performance metrics

Model	MAE ↓	MSE ↓	PSNR (dB) ↑	SSIM ↑	NCC ↑
Baseline Pix2Pix	0.643	0.546	11.34	0.181	0.167
UNetSelfCritique3D	0.078	0.038	20.57	0.468	0.518
Hybrid GAN-Diffusion	**0.077**	**0.034**	**21.10**	**0.456**	**0.547**

Table 6.2: Quantitative comparison of the three models across all evaluation metrics.

- The **Hybrid GAN-Diffusion Model** The pipeline showed the best performance as compared to baseline and the GAN due to its magical ability to extract and utilize the two models. The model gave the best metrics across all evaluation as seen from the table.

6.2 Qualitative Analysis

Other than performance metrics the visual inspection is a fundamental factor in evaluation of medical imaging tasks. The fact that most metrics focus on voxel penalizing makes them prone to structural variance. The qualitative results presented below demonstrate a clear and dramatic progression in quality across the three models.

6.2.1 Comparative Analysis Framework

The evaluation was conducted on a held-out test set of patient scans from the preprocessed TRUSTED dataset. Three models were compared:

- **Baseline Model:** The 3D Pix2Pix GAN with a ResNet generator and a multi-scale discriminator, as detailed in Chapter 3.

- **UNetSelfCritique3D GAN:** The novel self-critiquing GAN architecture, trained with the combined adversarial, reconstruction, and self-critique losses, as detailed in Chapter 4.
- **Hybrid GAN-Diffusion Model:** The complete, two-stage framework where the output of the pre-trained UNetSelfCritique3D GAN is refined by the conditional DDPM, as detailed in Chapter 5.

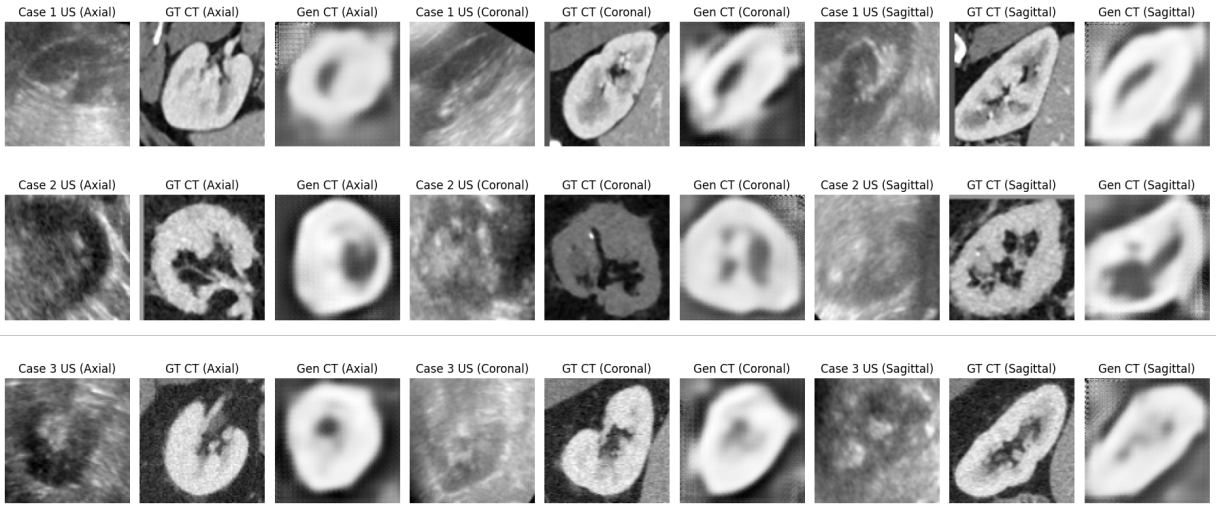


Figure 6.1: Showing all the 3 views of 3D US, CT and generated CT respectively.

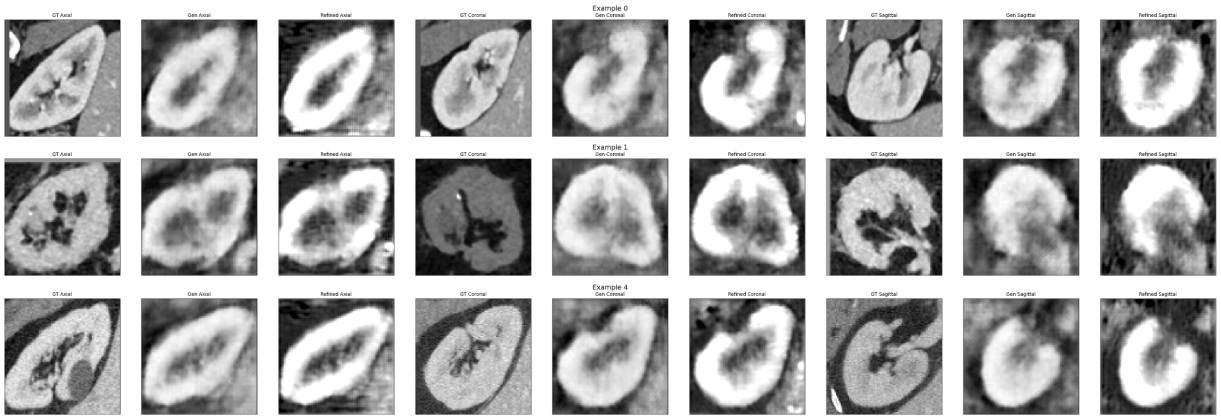


Figure 6.2: Showing all the 3D views of GT CT, Coarse CT and Refined CT respectively.

6.2.2 Visual Comparison Across Models

- **Baseline Model:** The outputs from the baseline Pix2Pix GAN consistently exhibit the signs of mode collapse and unstable training. As shown in Figure 6.1 showing the three different views of the first 3 sets of test data, the generated volume is blurry and lacks

fine anatomical definition. Critical internal structures like the renal pelvis and vasculature are poorly resolved or entirely absent. The overall shape is a generic approximation of a kidney, failing to capture the patient-specific morphology present in the ground truth.

- **UNetSelfCritique3D GAN:** The proposed GAN performs as expected underlaying a huge gap between its generations and those of the baseline Figure 6.2 shows that the outcome of similar first 3 sets of test data. We are able to observe a much more defined and structurally better-formed. The model to a significant extent avoids mode collapse, producing a kidney whose overall form and major inner structures are a faithful copy of the input US. This confirms that the inner feature-matching loss is effective at forcing the generator to preserve underlying anatomy. The output, though, still has a tell-tale ”GAN-like” smoothness and lacks the subtle textural realism of a real CT scan.

- **Hybrid GAN-Diffusion Model:** The final refinement process brings the output to an even higher level of realism. In Figure 6.2, the diffusion model enhances the structurally-correct output of the GAN by incorporating the lost high-frequency details. The final volume is more realistic with a real texture, harder edges, and more precise intensity distribution that is visually much more similar to the ground truth. The refinement process effectively reduces the remaining GAN artifacts and yields a level of detail that is beyond the capability of the GAN.

6.2.3 Case-wise Comparison

In this section, we visualize the first two test cases showing histogram intensity between ground truth, coarse and refined CT highlighting their differencies and simillarities. From the figures we can see that the GT CT has higher intensity compared to the refined CT

6.3 Discussion

6.3.1 Interpretation of Results

Both the quantitative and qualitative results showed that the worst candidate was the baseline GAN followed by the UNetSelfCritique3D GAN and best was proposed hybrid model. The baseline provided the benchmark to weigh and measure the performance of future models and our proposed GAN successfully mitigated the mode collapse barrier which is the reason behind poor performance of baseline thus underlying a good cornerstone the GAN field. But the proposed GAN was later modified by incorporating a DDPM (Refiner) which handled the complexity of US-to-CT conversion problem thus creating realistic, high-fidelity texture through the residual

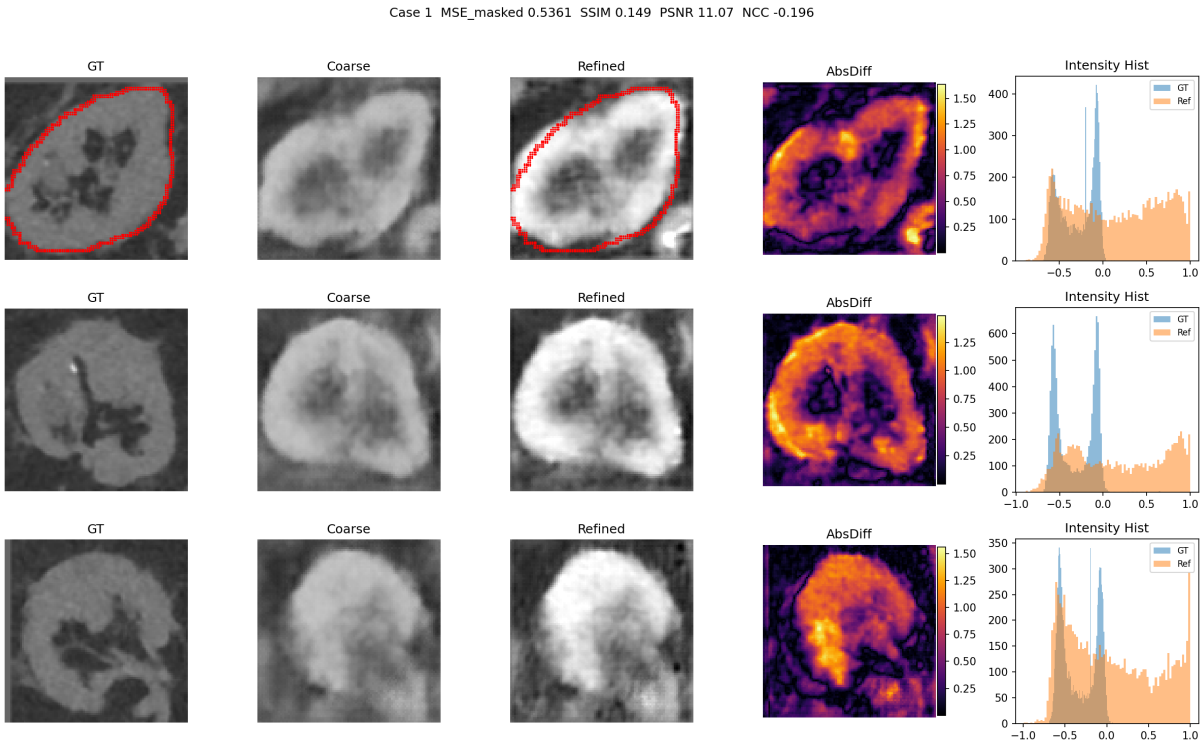


Figure 6.3: Histogram Intensity and Absolute Difference of case 1

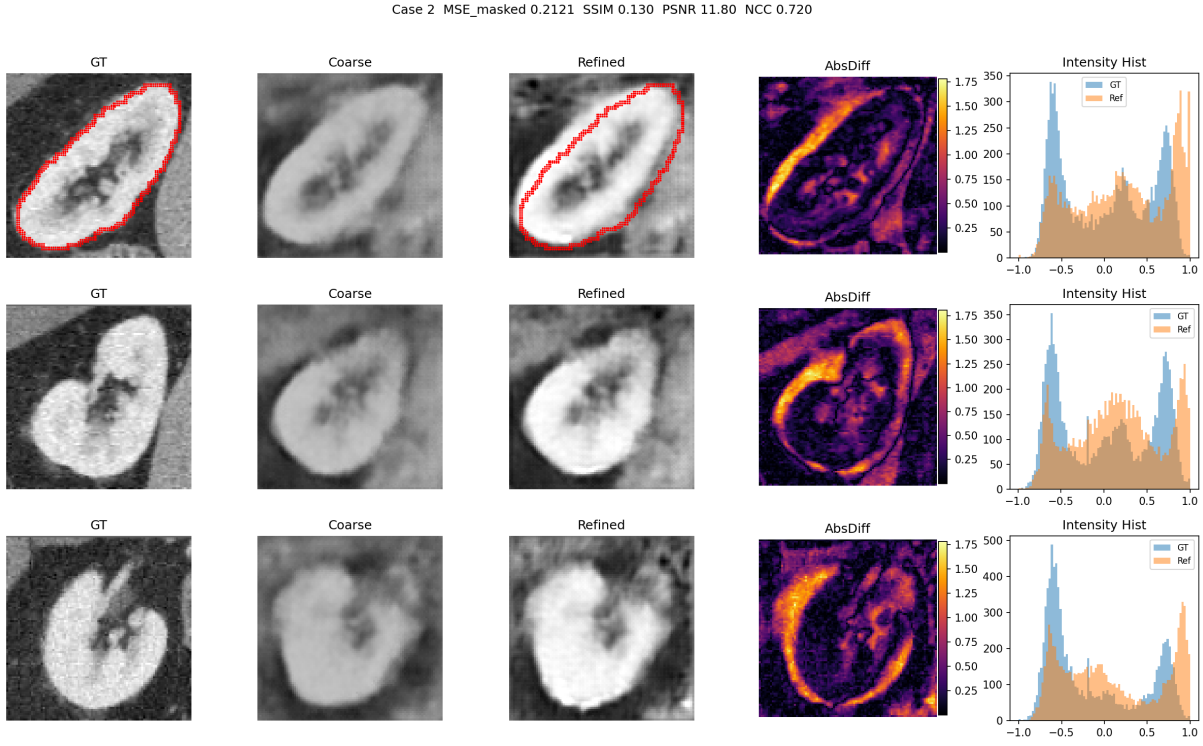


Figure 6.4: Histogram Intensity and Absolute Difference of case 2

mapping which is much simpler distribution of all the CT data resulting into an efficient and effective refinement process.

6.3.2 Limitations of Our Research

The positive results presented in this thesis lay a fundamental benchmark to the final destination. As much as the hybrid framework has shown a significant performance boost in 3D US-to-CT translation, there are several limitations prompted by a harsh and critical analysis. These limitations don't diminish the novelty of this work but rather, they point to the future course of this emerging and intricate field of study, making this project an essential starting point for further research.

1. **Residual Mode Collapse and Generator Stability:** While the UNetSelfCritique3D architecture successfully thwarted the immediate, catastrophic mode collapse of the baseline model, longer training revealed a more subtle, underlying instability. In later epochs, the model began converging on a less diverse set of anatomical the inherent tension within adversarial training can still, with sufficient time, push the generator into simplified solutions. Future work can explore other regularization techniques, i.e., other loss terms (e.g., contrastive losses) or more advanced discriminator networks, in an attempt to further anchor the generator in the subject-specific anatomical signature of the input.

2. **Fidelity Ceiling of the Diffusion Refiner:** The diffusion model undoubtedly enhanced the textural realism and sharpness of the GAN outputs. The ultimate synthesized volumes were not, nonetheless, flawlessly faithful and, in certain cases, continued to exhibit a degree of softness or blur compared to the ground truth CT (Figure 6.2). This implies that the diffusion model can be limited by the quality of coarse input it receives. If the output of the first stage of the GAN contains significant structural defects or lacks significant anatomical precursors, then the diffusion model's task is more one of "major correction" than "refinement," a significantly harder problem. Additional work must be done to explore more powerful diffusion architectures or alternative conditioning mechanisms that can provide the refiner with more informative information, perhaps incorporating features directly from the US input alongside the coarse CT

3. **Achieving Clinical Acceptability:** Clinical usefulness is the ultimate objective of medical image translation. Our findings are qualitatively striking and quantitatively robust but still fall short of the diagnostic quality that must be achieved for clinical acceptance. Subtle but important pathological features or anatomical structures need to be clearly visible and undistorted. This is a proof-of-concept study. In order to specify specific failure

modes and the performance levels required for medical use, radiologists and clinicians must rigorously validate the research in the specified, most significant next stage. Metrics like MSE, MAE, PSNR, and Normalized Cross-Correlation (NCC) generally need to be above some thresholds for clinical acceptability, which will need to be determined through clinical validation studies.

4. **Dataset Scale and Diversity:** As the first study to apply the TRUSTED dataset to this task, our study is inherently limited by its size. The model was trained on data from 48 patients, collected at a single institution using a single set of equipment. Its generalizability to diverse patient populations, various body habits, various ultrasound machines, and various sonographer skill levels is unknown. The next step required is a multi-center, large-scale study to develop a really robust and generalizable model.

5. **Pathology and Anatomical Extremes:** The TRUSTED dataset consists primarily of healthy kidneys. Consequently that makes our model an expert in translating healthy anatomy. Its performance on abnormal cases such as kidneys with large tumors, complex cysts, hydronephrosis or congenital anomalies is untested. These pathologies often present with unique and challenging signatures on ultrasound. Future research must focus on curating diverse datasets that include a wide spectrum of renal pathologies to train models that are not just anatomically aware but also pathologically sensitive.

Chapter 7

CONCLUSION AND FUTURE WORK

7.1 Summary of Findings

The study began with the establishment of a baseline via the application of a standard 3D Pix2Pix GAN, which, as anticipated, collapsed by experiencing severe mode collapse, as confirmation of the inadequacy of standard methods. In response, the first significant finding was that the inclusion of an internal self-criticism loss in our new UNetSelfCritique3D architecture could stabilize training, to a significant extent, and produce structurally correct anatomical translations. Admitting the textural limitations still being present in the GAN's output, we proceeded to construct a hybrid model. The second important discovery was that a conditional 3D DDPM, utilising the GAN's output as an effective prior, could effectively carry out residual refinement. The extensive testing confirmed this staged process, showing quantifiable progress at each stage and validating our key hypothesis that such a synergistic hybrid solution would be able to produce 3D CT volumes so much more realistic than baseline.

7.2 Statement of Contributions

This thesis makes several distinct and novel contributions to the field of medical image translation, establishing a foundation for future work:

1. **A Novel Framework for a New Problem:** We present the first end-to-end system that translates actual clinical 3D transabdominal ultrasound volumes into corresponding CT volumes, moving beyond previous work limited to 2D images or synthetic data.

2. **A Modified Generator Architecture:** Our UNetSelfCritique3D introduces a novel self-regularization mechanism within the GAN generator that effectively reduces mode collapse during cross-modality translation.

3. **A Validated Hybrid Methodology:** We demonstrate that combining GAN and diffusion models for residual refinement produces superior results, validating a practical approach to leveraging both generative architectures.

4. **A Benchmark and Data Pipeline:** We establish the first performance benchmark for this task and provide an open-source pipeline for processing the TRUSTED dataset, enabling future comparative research.

7.3 Future Research Directions

This work reveals several promising paths forward. Based on our findings, we identify five key areas for future investigation:

- **Stabilizing the Generator:** Our self-critique mechanism shows promise, but newer contrastive learning methods [41] and self-supervised techniques [42] could further reduce mode collapse and better preserve anatomical structures across diverse scans.
- **Smarter Diffusion Conditioning:** Currently, our diffusion model only sees the coarse CT output. Feeding it multi-scale features directly from the original ultrasound—using cross-attention [35] or feature pyramids [43]—could help it generate finer anatomical details.
- **Joint End-to-End Training:** While our two-stage pipeline works well, training the GAN and diffusion components together might discover better intermediate representations. Recent unified frameworks [44] suggest this could be computationally feasible.
- **More Diverse Data:** The biggest bottleneck remains data scarcity. We need larger, multi-center datasets that include varied pathologies—not just healthy kidneys. Collaborative efforts like the KiTS challenge [45] show how the community can tackle this together.
- **Clinical Validation:** Ultimately, these models must prove useful in real clinical settings. Working directly with radiologists to develop task-specific tools—for tumor segmentation or volume measurement—and ensuring interpretability [46] will be crucial for actual deployment.

This thesis demonstrates that high-quality 3D ultrasound-to-CT translation is challenging but achievable. While the path to radiation-free virtual CT imaging is long, our results suggest it's a worthwhile pursuit. We've established a foundation, proposed a working solution, and identified clear next steps for the field.

Chapter 8

DEMONSTRATION OF OUTCOME BASED EDUCATION (OBE)

The report illustrates how Outcome-Based Education (OBE) has been applied during the final year thesis project, *Ultrasound to Computed Tomography Image Translation using Deep Learning*. This documentation systematically maps project activities and outcomes to Washington Accord graduate attributes, demonstrating how the complex engineering challenge of medical image translation addresses the necessary knowledge profiles (K3-K8), complex engineering problem-solving features (P1-P7), and complex engineering activities (A1-A5). This comprehensive alignment confirms that the thesis work satisfies program outcomes and builds fundamental engineering competencies for professional practice.

8.1 Introduction

Outcome-Based Education (OBE) represents a paradigm shift in engineering education, emphasizing what learners can demonstrate upon course completion rather than what is taught. This documentation pertains to the OBE implementation of the thesis project titled *Ultrasound to Computed Tomography Image Translation using Deep Learning*, which addresses a major challenge in medical imaging. The project involves developing advanced deep learning systems that convert ultrasound images into CT-equivalent images to potentially reduce radiation exposure without compromising diagnostic quality. This multifaceted engineering project demonstrates achievement of specified program goals by combining theoretical knowledge with practical work, while considering societal, ethical, and professional obligations inherent in developing healthcare technology applications.

8.2 Course Outcomes (COs) Addressed

The following table shows the COs addressed in EEE 4800 for Project and Thesis.

Table 8.1: Course Outcomes Addressed

COs	**CO Statement**	**POs**	**EEE 4800**
CO1	Identify a contemporary real life problem related to electrical and electronic engineering by reviewing and analyzing existing research works.	PO2	✓
CO2	Determine functional requirements of the problem considering feasibility and efficiency through analysis and synthesis of information.	PO4	✓
CO3	Select a suitable solution and determine its method considering professional ethics, codes and standards.	PO8	✓
CO4	Adopt modern engineering resources and tools for the solution of the problem.	PO5	✓
CO5	Prepare management plan and budgetary implications for the solution of the problem.	PO11	✓
CO6	Analyze the impact of the proposed solution on health, safety, culture and society.	PO6	✓
CO7	Analyze the impact of the proposed solution on environment and sustainability.	PO7	✓
CO8	Develop a viable solution considering health, safety, cultural, societal and environmental aspects.	PO3	✓
CO9	Work effectively as an individual and as a team member for the accomplishment of the solution.	PO9	✓
CO10	Prepare various technical reports, design documentation, and deliver effective presentations for demonstration of the solution.	PO10	✓
CO11	Recognize the need for continuing education and participation in professional societies and meetings.	PO12	✓

8.3 Aspects of Program Outcomes (POs) Addressed

The following table shows the aspects addressed for certain Program Outcomes (POs) in EEE 4800 for Project and Thesis.

Table 8.2: Program Outcomes Aspects Addressed

PO	Statement	Different Aspects	Addressed
PO3	**Design/development of solutions:** Design solutions for complex electrical and electronic engineering problems and design systems, components or processes that meet specified needs with appropriate consideration for public health and safety, cultural, societal, and environmental considerations.	Public health	✓
		Safety	✓
		Cultural	✓
		Societal	✓
		Environmental	✓
PO4	**Investigation:** Conduct investigations of complex electrical and electronic engineering problems using research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of information to provide valid conclusions.	Design of experiments	✓
		Analysis and interpretation of data	✓
		Synthesis of information	✓
PO6	**The engineer and society:** Apply reasoning informed by contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to professional engineering practice and solutions to complex electrical and electronic engineering problems.	Societal	✓
		Health	✓
		Safety	✓
		Legal	✓
		Cultural	✓
PO7	**Environment and sustainability:** Understand and evaluate the sustainability and impact of professional engineering work in the solution of complex electrical and electronic engineering problems in societal and environmental contexts.	Societal	✓
		Environmental	✓
PO8	**Ethics:** Apply ethical principles embedded with religious values, professional ethics and responsibilities, and norms of electrical and electronic engineering practice.	Religious values	✓
		Professional ethics and responsibilities	✓
		Norms	✓
PO10	**Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large,	Comprehend and write effective reports	✓

8.4 Knowledge Profiles (K3–K8) Addressed

The following table shows the Knowledge Profiles (K3–K8) addressed in EEE 4800 for Project and Thesis.

Table 8.3: Knowledge Profiles Addressed

K	Knowledge Profile (Attribute)	Addressed
K3	A systematic, theory-based formulation of engineering fundamentals required in the engineering discipline	✓
K4	Engineering specialist knowledge that provides theoretical frameworks and bodies of knowledge for the accepted practice areas in the engineering discipline; much is at the forefront of the discipline	✓
K5	Knowledge that supports engineering design in a practice area	✓
K6	Knowledge of engineering practice (technology) in the practice areas in the engineering discipline	✓
K7	Comprehension of the role of engineering in society and identified issues in engineering practice in the discipline: ethics and the engineer's professional responsibility to public safety; the impacts of engineering activity; economic, social, cultural, environmental and sustainability	✓
K8	Engagement with selected knowledge in the research literature of the discipline	✓

The following table explains how the Knowledge Profiles (K3–K8) have been addressed in EEE 4800 (Project and Thesis).

Table 8.4: Knowledge Profiles Justification

K	Explanation/Justification
K3	Applied fundamental engineering concepts including signal processing theory for medical image analysis, linear algebra for tensor operations in deep learning, calculus for optimization tasks, and information theory to assess image translation quality and information preservation across domains.
K4	Utilized specialized expertise in deep learning models (U-Net, GANs, Pix2Pix), medical imaging physics (ultrasound wave propagation, CT reconstruction algorithms), computer vision image quality evaluation, and current research in cross-modal medical image translation at the forefront of medical AI.
K5	Applied design knowledge to develop optimal neural network architectures for medical image translation, designed suitable loss functions combining adversarial, perceptual, and structural losses, and created evaluation systems that are both clinically relevant and technically sound.
K6	Implemented contemporary engineering tools including TensorFlow/Keras/PyTorch frameworks, medical image processing libraries (nibabel, SimpleITK), GPU-accelerated model training, data augmentation algorithms for medical images, and version control for reproducible research.
K7	Through the medical application context, understood engineering's role in healthcare: ethics in medical AI applications, patient safety implications of diagnostic errors, societal healthcare accessibility, environmental benefits of reduced radiation exposure, and sustainability through optimized medical workflows.
K8	Literature review and methodology development involved intensive engagement with research in medical image analysis (Medical Image Analysis, IEEE TMI journals), computer vision (CVPR, ICCV conferences), deep learning architectures, and clinical validation studies of AI in radiology.

8.5 Use of Complex Engineering Problems

The Project and Thesis course (EEE 4800) is based on solving a Complex Engineering problem by creating a medical image translation system that converts Ultrasound (US) images into Computed Tomography (CT) equivalents. This challenge requires combining expertise in medical imaging, deep learning, signal processing, and clinical applications. The solution involves developing novel neural network architectures that must be robust to domain shifts across imaging modalities, maintain anatomical consistency, remain diagnostically relevant, and operate on small paired datasets.

8.6 Socio-Cultural, Environmental, and Ethical Impact

Important implications considered include patient safety through accurate image translation to avoid misdiagnosis, accessibility through potential reduction of CT radiation exposure in developing countries, ethical issues related to sensitive medical information and algorithm transparency, and environmental concerns regarding reduced radiological exposure.

8.7 Attributes of Complex Engineering Problem Solving (P1–P7) Addressed

The following table shows the attributes of Complex Engineering Problem Solving (P1–P7) addressed in EEE 4800 for Project and Thesis.

Table 8.5: Complex Engineering Problem Solving Attributes

Attribute	**Range of Complex Engineering Problem Solving**	**Addressed**
Complex Engineering Problems have characteristic P1 and some or all of P2 to P7:		
Depth of knowledge required	**P1:** Cannot be resolved without in-depth engineering knowledge at the level of one or more of K3, K4, K5, K6 or K8 which allows a fundamentals-based, first principles analytical approach	✓
Range of conflicting requirements	**P2:** Involve wide-ranging or conflicting technical, engineering and other issues	✓
Depth of analysis required	**P3:** Have no obvious solution and require abstract thinking, originality in analysis to formulate suitable models	✓
Familiarity of issues	**P4:** Involve infrequently encountered issues	✓
Extent of applicable codes	**P5:** Are outside problems encompassed by standards and codes of practice for professional engineering	✓
Extent of stakeholder involvement and conflicting requirements	**P6:** Involve diverse groups of stakeholders with widely varying needs	✓
Interdependence	**P7:** Are high level problems including many component parts or sub-problems	✓

The following table explains how the attributes of Complex Engineering Problem Solving (P1–P7) have been addressed in EEE 4800 (Project and Thesis).

Table 8.6: Complex Engineering Problem Solving Justification

P	**Explanation/Justification**
P1	Solving medical image translation demands comprehensive understanding of deep learning architectures (K4), medical imaging physics (K3), and signal processing principles to build novel translation models that preserve anatomical structure while transforming medical imaging domains.
P2	The project balances conflicting requirements: translation accuracy versus computational efficiency, model complexity versus clinical deployability, image quality versus inference speed, and diagnostic accuracy versus generalization across patient populations.
P3	No standard solution exists for medical image translation, demanding abstract reasoning to design appropriate generative models, formulate optimal loss functions, and devise clinically relevant assessment metrics beyond traditional image quality measures.
P4	The challenge involves infrequently encountered issues in general engineering: domain transfer across diverse medical imaging modalities, processing 3D volumetric data, and ensuring clinical utility with limited annotated medical images.
P5	The project extends beyond medical imaging protocols to develop novel deep learning methods for cross-modal translation, requiring innovation beyond existing DICOM standards and traditional image processing guidelines.
P6	The solution addresses diverse stakeholders with varying needs: patients (safety, accessibility), clinicians (diagnostic accuracy), hospitals (cost-effectiveness), and regulatory agencies (safety standards, approval processes).
P7	The challenge decomposes into interconnected sub-problems: data preprocessing and normalization, neural architecture design, loss formulation, model training optimization, clinical validation, and deployment considerations in medical practice.

8.8 Attributes of Complex Engineering Activities (A1–A5) Addressed

The following table shows the attributes of Complex Engineering Activities (A1–A5) addressed in EEE 4800 for Project and Thesis.

Table 8.7: Complex Engineering Activities Attributes

Attribute	**Range of Complex Engineering Activities**	**Addressed**
Complex activities have some or all of the following characteristics:		
Range of resources	**A1:** Involve the use of diverse resources (people, money, equipment, materials, information and technologies)	✓
Level of interaction	**A2:** Require resolution of significant problems arising from interactions between wide-ranging or conflicting technical, engineering or other issues	✓
Innovation	**A3:** Involve creative use of engineering principles and research-based knowledge in novel ways	✓
Consequences for society and the environment	**A4:** Have significant consequences in a range of contexts, characterized by difficulty of prediction and mitigation	✓
Familiarity	**A5:** Can extend beyond previous experiences by applying principles-based approaches	✓

The following table explains how the attributes of Complex Engineering Activities (A1–A5) have been addressed in EEE 4800 (Project and Thesis).

Table 8.8: Complex Engineering Activities Justification

A	Explanation/Justification
A1	The project utilized diverse resources: high-performance computing clusters (GPUs), medical image datasets, deep learning frameworks (TensorFlow/PyTorch), medical image processing libraries (nibabel, ITK), and collaboration with medical professionals for clinical validation.
A2	Major challenges arose from interactions between technical requirements: balancing computational capabilities with model performance, balancing data augmentation with anatomical plausibility, and balancing diagnostic information preservation with achievement of technical metrics in translated images.
A3	The project involved innovative application of engineering principles through novel neural network architectures (U-Net, GANs, Pix2Pix) for medical imaging, novel loss functions integrating perceptual and structural measures, and research-driven knowledge from computer vision and medical AI.
A4	Significant implications exist for medical practice: potential effects on patient diagnostic processes, reduction in radiation exposure, possibility of CT-equivalent imaging in resource-limited settings, and ethical aspects of AI-powered diagnosis—all characterized by difficulty in predicting and mitigating adverse effects.
A5	Medical image translation extends beyond previous coursework experiences through principles-driven application in an emerging interdisciplinary field, requiring integration of knowledge from deep learning, medical physics, and clinical radiology in novel combinations.

REFERENCES

[1] I. Goodfellow et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2672–2680.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems 33*, 2020, pp. 6840–6851.

[3] G. N. Hounsfield, "Computerized transverse axial scanning (tomography): Part 1. Description of system," *The British Journal of Radiology*, vol. 46, no. 552, pp. 1016–1022, 1973.

[4] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, PMLR, 2017, pp. 214–223.

[5] United Nations, *Transforming our world: the 2030 Agenda for Sustainable Development*, General Assembly, A/RES/70/1, 2015.

[6] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[7] W. Ndzimbong et al., "TRUSTED: The Paired 3D Transabdominal Ultrasound and CT Human Data for Kidney Segmentation and Registration Research," *Scientific Data*, vol. 12, no. 1, p. 615, 2025.

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232.

[10] S. Hu, Y. Shen, S. Wang, B. Lei, et al., "Bidirectional mapping generative adversarial networks for brain mr to pet synthesis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 1, pp. 145–157, 2022. DOI: 10.1109/TMI.2021.3107013

[11] K. Armanious, C. He, S. Fischer, and T. Jiang, "MedGAN: Medical image translation using GANs," *Computerized Medical Imaging and Graphics*, vol. 79, p. 101 684, 2020.

[12] H. Shan et al., "3D TomoGAN for high-resolution 3D medical image synthesis," *Medical Image Analysis*, vol. 58, p. 101 529, 2019.

[13] H. Zhang, L. Ji, and Q. Wang, "SCAN: A Semantic-aware and Class-agnostic Network for Medical Image Synthesis," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, 2019, pp. 638–646.

[14] D. Nie, R. Trullo, C. Petitjean, S. Ruan, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018.

[15] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Generative adversarial networks for noise reduction in low-dose CT," *IEEE Transactions on Medical Imaging*, vol. 36, no. 12, pp. 2536–2545, 2017.

[16] S. Yin et al., "GAN-based synthetic medical image generation: A review," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3565–3577, 2020.

[17] S. Vedula, A. P. Miller, J. Podgurski, and M. A. L. Bell, "US-to-CT synthesis with a conditional generative adversarial network," in *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 11598, 2021, p. 115981M.

[18] T. Noack et al., "Ultrasound-to-CT translation for thyroid screening using 3D US volumes from the SegThy dataset," *arXiv preprint arXiv:2303.11929*, 2023.

[19] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8162–8171.

[20] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, p. 417, 1933.

[21] I. T. Jolliffe, *Principal Component Analysis*, 2nd. New York, NY, USA: Springer, 2011. DOI: 10.1007/978-1-4757-1904-8

[22] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[23] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[24] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 424–432.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 234–241.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, et al., *PyTorch: An imperative style, high-performance deep learning library*, 2019.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[30] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," *arXiv preprint arXiv:1611.02163*, 2017.

[31] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018.

[32] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[33] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[35] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017.

[36] Y. Wu and K. He, "Group normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[37] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.

[38] A. Hore and D. Ziou, "Image quality metrics: a survey," in *2010 20th International Conference on Pattern Recognition*, IEEE, 2010, pp. 2366–2369.

[39] U. Sara, M. Akter, and M. S. Uddin, "Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study," *Journal of Computer and Communications*, vol. 7, no. 3, pp. 8–18, 2019.

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[41] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, PMLR, 2020, pp. 1597–1607.

[42] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[43] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[44] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2021.

[45] N. Heller et al., "The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes," *arXiv preprint arXiv:1904.00445*, 2019.

[46] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

Appendix A

Detailed Model Architectures

This appendix provides the layer-by-layer specifications for the generative models used in this research. The architectures are designed for input volumes of size $1 \times 96 \times 96 \times 96$.

A.1 UNetSelfCritique3D Generator

The generator follows a 3D U-Net architecture with four downsampling and four upsampling levels. Skip connections concatenate features from the encoder to the decoder at corresponding levels.

Table A.1: Layer specification for the UNetSelfCritique3D Generator.

Layer	Kernel/Stride	Padding	Activation	Output Shape (C, D, H, W)
Encoder Path				
Input	-	-	-	(1, 96, 96, 96)
Conv3D	4x4x4 / 2	1	LeakyReLU	(64, 48, 48, 48)
Conv3D + BN	4x4x4 / 2	1	LeakyReLU	(128, 24, 24, 24)
Conv3D + BN	4x4x4 / 2	1	LeakyReLU	(256, 12, 12, 12)
Conv3D + BN	4x4x4 / 2	1	LeakyReLU	(512, 6, 6, 6)
Bottleneck				
Conv3D + BN	4x4x4 / 2	1	LeakyReLU	(512, 3, 3, 3)
TConv3D + BN	4x4x4 / 2	1	ReLU	(512, 6, 6, 6)
Decoder Path (with skip connections)				
TConv3D + BN	4x4x4 / 2	1	ReLU	(1024, 12, 12, 12)
TConv3D + BN	4x4x4 / 2	1	ReLU	(512, 24, 24, 24)
TConv3D + BN	4x4x4 / 2	1	ReLU	(256, 48, 48, 48)
TConv3D	4x4x4 / 2	1	Tanh	(1, 96, 96, 96)

TConv3D: Transposed Convolution, BN: Batch Normalization

A.2 Conditional 3D U-Net DDPM

The diffusion model is a 3D U-Net with ResNet blocks and sinusoidal time embeddings. The conditioning `coarse_ct` is concatenated to the input `x_noisy`.

Table A.2: Core components of the Conditional 3D U-Net DDPM.

Component	**Description**
Input	Concatenation of `x_noisy` (1 channel) and `coarse_ct` (1 channel). Shape: (2, 96, 96, 96).
Time Embedding	Sinusoidal position embedding projected to a 512-dimensional vector, then processed by linear layers and added into each ResNet block.
Downsampling	Series of ResNet blocks followed by a 3D convolution with stride 2.
Upsampling	Series of ResNet blocks followed by a 3D transposed convolution with stride 2.
Attention	Attention mechanisms are applied at lower-resolution feature maps (e.g., 12x12x12) to capture long-range dependencies.
Normalization	Group Normalization is used instead of Batch Normalization for stable training with small batch sizes.
Output	A final 3D convolution projects the feature maps back to a single channel, predicting the added noise.

Appendix B

Data Preprocessing Pipeline

The following algorithm outlines the automated pipeline developed to process the raw TRUSTED dataset into co-registered, paired 3D volumes suitable for training.

Algorithm 1 US-to-CT Registration and Alignment Pipeline

procedure REGISTERPAIR(US_volume, US_mask, CT_volume, CT_mask)
    US_mesh ← GenerateMesh(US_mask)
    CT_mesh ← GenerateMesh(CT_mask)

    ▷ Step 1: Coarse Alignment via Principal Axes
    T_pca ← ComputePCATransform(US_mesh, CT_mesh)
    US_mesh_coarse ← ApplyTransform(US_mesh, T_pca)

    ▷ Step 2: Rigid Refinement using ICP
    T_icp ← ComputeICPTransform(US_mesh_coarse, CT_mesh)

    ▷ Step 3: Final Affine Refinement
    T_affine ← ComputeAffineTransform(US_mesh_coarse, CT_mesh)

    ▷ Combine transformations in order
    T_final ← T_affine ∘ T_icp ∘ T_pca

    ▷ Apply final transform to original volumes
    US_volume_registered ← ResampleVolume(US_volume, T_final, reference=CT_volume)
    US_mask_registered ← ResampleVolume(US_mask, T_final, reference=CT_mask, interp='Nearest')

    return US_volume_registered, US_mask_registered
end procedure

Appendix C

Key Source Code Snippets

This section contains key Python snippets from the project, demonstrating the implementation of the novel loss function and the conditioning mechanism in the diffusion model.

C.1 Self-Critique Loss Function

The generator is modified to return encoder and decoder feature maps, which are then compared in the training loop to calculate the self-critique loss.

```
# In the training script:
def calculate_sc_loss(enc_maps, dec_maps, device):
sc_loss = 0.0
l1_loss = torch.nn.L1Loss().to(device)
# Compare corresponding feature maps
# (from outermost encoder to outermost decoder)
for enc_map, dec_map in zip(enc_maps, reversed(dec_maps)):
sc_loss += l1_loss(dec_map, enc_map)
return sc_loss
```

Listing C.1: Implementation of the Self-Critique Loss.

C.2 Conditional Diffusion Model Input

Conditioning is achieved by concatenating the coarse CT volume with the noisy input volume at the start of the denoising U-Net.

```
# In the Conditional UNet model for DDPM:
def forward(self, x_noisy, timestep, coarse_ct_condition):
# x_noisy: The noisy CT at step t
# timestep: The current timestep t
# coarse_ct_condition: The conditioning output from the GAN

# 1. Concatenate the noisy input and the condition
#    along the channel dimension.
```

```
#     Input shapes (B, 1, D, H, W) -> output (B, 2, D, H, W)
x_conditioned = torch.cat([x_noisy, coarse_ct_condition], dim=1)

# 2. Compute time embedding
time_emb = self.time_mlp(timestep)

# 3. Pass through the U-Net, injecting time_emb
predicted_noise = self.unet_core(x_conditioned, time_emb)

return predicted_noise
```

Listing C.2: Forward pass of the conditional denoising U-Net.

Appendix D

Hyperparameters and Environment

Table D.1: Training Hyperparameters for All Models.

Parameter	**UNetSelfCritique3D GAN**	**Conditional DDPM**
Optimizer	Adam	AdamW
Learning Rate	0.0002	0.0001
Adam Betas	(0.5, 0.999)	(0.9, 0.999)
Batch Size	1	1
Number of Epochs	300	200
Loss Weights (λ) for GAN:		
- Adversarial	1.0	N/A
- Reconstruction (L1)	100.0	N/A
- Feature Matching	10.0	N/A
- Self-Critique	10.0	N/A
Diffusion Parameters:		
Diffusion Timesteps (T)	N/A	1000
Noise Schedule	N/A	Linear
Inference Steps	N/A	200

Table D.2: Software and Hardware Environment.

Component	**Version / Specification**
Operating System	Ubuntu 20.04 LTS
GPU	NVIDIA A100 (40 GB)
CUDA Version	11.4
Python	3.8.10
PyTorch	1.12.1
MONAI	0.9.1
Diffusers (Hugging Face)	0.3.0
SimpleITK	2.2.0