ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

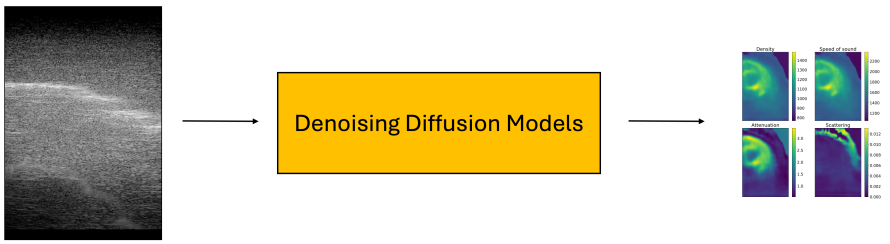
Institut für Integrierte Systeme
Integrated Systems Laboratory

Department of Information Technology and
Electrical Engineering

Autumn Semester 2024

USDatarecyclerV2: Denoising Diffusion Models for Ultrasound RF-Data Reconstruction

Master Project



Cyrill Knecht
cyrknech@student.ethz.ch

December 2024

Supervisors: Dr. Christoph Leitner, christoph.leitner@iis.ee.ethz.ch
Dr. Yawei Li, yawei.li@vision.ee.ethz.ch

Professor: Prof. Dr. Luca Benini, lbenini@iis.ee.ethz.ch

Acknowledgements

I am profoundly grateful to my supervisors, Yawei and Christoph, for their continous support, thoughtful advice, and invaluable expertise, which have been instrumental to the success of this project. I would also like to sincerely thank the Integrated Systems Laboratory (IIS) for hosting this innovative and highly interesting project.

Abstract

Medical Ultrasound (US) imaging is a vital diagnostic tool with a wide range of applications. The raw data from ultrasound imaging, known as Radio Frequency (RF) data, contains richer information than processed US images, enabling numerous valuable use cases. However, while extensive datasets of processed US images are readily available, raw RF data remains scarce.

This thesis builds upon an existing system designed to predict raw RF data from US images, focusing on enhancing its Deep Learning (DL) component through Denoising Diffusion DL models. The original pipeline employs a data-driven, physically-informed approach that integrates DL techniques with numerical simulations of ultrasound wave propagation. This hybrid, physically-informed, approach was retained, as End-to-End DL models have shown inferior performance in comparison.

In this work, various Denoising Diffusion DL Models and architectures are implemented, seamlessly integrated into the existing pipeline, and rigorously evaluated against the previously utilized Transformer-based DL model. All these models aim to predict key acoustic properties of tissue, such as density phantoms, based on input US images.

Additionally, the training process is enhanced by leveraging large-scale synthetic datasets generated from Computed Tomography (CT) scans. The impact of these larger datasets on model performance is systematically compared to that of smaller, existing datasets, providing valuable insights into the influence of dataset scale and quality on the overall pipeline's effectiveness.

Declaration of Originality

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor. For a detailed version of the declaration of originality, please refer to Appendix A

Cyrill Knecht,
Zurich, December 2024

Contents

List of Figures

List of Tables

List of Acronyms

CT . . . . . .Computed Tomography
CV . . . . . .Computer Vision

DDPM . . . . .Denoising Diffusion Probabilistic Model
DL . . . . . .Deep Learning

GAN . . . . . .Generative Adversarial Networks
GPU . . . . . .Graphics Processing Unit

I2I . . . . . .Image-to-Image
IIS . . . . . .Integrated Systems Laboratory

KL . . . . . .Kullback–Leibler

LDM . . . . . .Latent Diffusion Model

MAE . . . . . .Mean Absolute Error
MRI . . . . . .Magnetic resonance imaging

NN . . . . . .Neural Network

ODE . . . . . .Ordinary Differential Equation

RF . . . . . .Radio Frequency
RMSE . . . . . .Root Mean Square Error

SDE . . . . . .Stochastic Differential Equation

SLURM . . . . .Simple Linux Utility for Resource Management

SNR . . . . . .Signal-to-Noise Ratio

SSIM . . . . . .Structural Similarity Index Measure

TGC . . . . . .Time Gain Compensation

US . . . . . .Ultrasound

VAE . . . . . .variational autoencoder

Chapter 1

Introduction

1.1. Motivation

Medical US imaging [4] has been an essential tool in diagnostic medicine for decades due to its non-invasive nature, cost-effectiveness, and safety. Unlike imaging methods such as CT [5] and Magnetic resonance imaging (MRI) [6], ultrasound does not expose patients to ionizing radiation, making it safer for repeated use. Its portability further increases its utility, enabling its application in point-of-care scenarios, emergency settings, and resource-limited environments.

US-imaging works by capturing raw RF signals during the acquisition process and converting them into human-readable images. However, this transformation involves non-linear operations, such as envelope detection and log-compression, which discard essential details present in the RF data. RF data is inherently richer in information, as it retains the high-resolution signal characteristics necessary for advanced analysis. Applications relying on RF data include tissue characterization, muscle activity monitoring, and spatial alignment, demonstrating its significant value in medical imaging.

Despite its potential, RF data remains largely inaccessible. Proprietary US acquisition systems often restrict access to raw RF signals, and publicly available datasets predominantly contain only processed US images. In contrast, large-scale datasets of US images are widely available. To address this limitation, the predecessor project of this thesis introduced a **physically-informed data-driven pipeline** capable of predicting RF data from existing US images. This approach demonstrated the utility of combining DL with numerical simulations for realistic RF reconstruction.

Building on this foundation, this thesis investigates the use of **Denoising Diffusion Models**[7] to enhance the DL component of the existing pipeline. Diffusion models have recently emerged as the state-of-the-art in generative computer vision tasks due to their

ability to iteratively refine outputs, surpassing traditional approaches like Generative Adversarial Networks (GAN)[8] and VAE[9]. Their promising capabilities motivate their exploration in the Image-to-Phantom translation task, where intermediate representations, consisting of key acoustic properties of tissue and from now on also referred to as phantom maps, are predicted from US input images.

1.2. Objective

The primary objective of this thesis is to investigate the performance of **Denoising Diffusion Models** in the previously described physically-informed pipeline, also referred to as the **USDatarecycler Pipeline** to predict raw RF data from US images by leveraging . Specifically, the goals of the thesis were as follows:

1. **Adapting Diffusion Models:** Implement and adapt diffusion models for predicting phantom maps from ultrasound images.
2. **Training and Evaluation:** Train the models on a large synthetic dataset and evaluate their performance.
3. **Integrating Diffusion Models into the existing Pipeline:** Integrate the adapted diffusion models into the USDatarecycler Pipeline, ensuring compatibility with the numerical simulation component that reconstructs RF data.
4. **Testing on real-world data** Evaluate the models on external in vivo datasets, to assess their real-world generalization capabilities.

By addressing these objectives, this thesis explores the potential of diffusion models to improve the reconstruction accuracy and generalizability of RF data prediction.

1.3. Contributions

The contributions of this thesis, arising from its goals stated in the previous chapter were as follows:

- **Adaptation of Denoising Diffusion Models:** Implemented and adapted new diffusion model architectures building on the Fast-DDPM[10] and Marigold[11] models, to predict phantom maps, enabling realistic RF data reconstruction.
- **Integration into USDatarecycler Pipeline:** Integrated the best adapted models into the existing pipeline, maintaining compatibility with the physics-based numerical simulation component.

- **Improved Performance of Adapted Marigold Model** Introduced a fine-tuning strategy for the VAE component of the adapted Marigold model and optimized the original Marigold training for ultrasound data by changing to a two-step method.
- **Training and Evaluation on Synthetic and External Datasets:** Trained the newly implemented models on a newly produced, large scale, synthetic CT-Phantom 500k dataset and evaluated them against the baseline TransUNet[12] model. Tested the models on external datasets to demonstrate their generalizability.
- **Efficient Dataset Handling:** Developed a parallelized, optimized chunk-loading functionality to manage large-scale datasets efficiently on the Euler Cluster, addressing its specific hardware constraints.

1.4. Thesis Structure

The remainder of this thesis is organized as follows:

- **Chapter 2: Preliminaries / Background** Provides an overview of the US-Datarecycler pipeline, Denoising Diffusion Models, Ultrasound Imaging Modes, and Virtual Phantom Models. This chapter sets the foundation for understanding the used and adapted pipeline and the theoretical aspects relevant to this thesis.
- **Chapter 3: Related Work** Reviews relevant literature and existing models, focusing on Diffusion Models, I2I-translation, and general Semantic Segmentation in medical imaging. It includes detailed discussions on the class of Latent Diffusion Models[1], approaches to reduce sampling time, and key models such as Fast-DDPM[10], MedSegDiff-V2[13], Stable Diffusion[1], and Marigold[11].
- **Chapter 4: Implementation** Explains all implementations and architectures that were designed or adapted during the project.
- **Chapter 5: Experiments** Explains all experiments done in the following results section and the evaluation datasets and metrics that were used for the evaluation.
- **Chapter 6: Results** Presents the evaluation results of all models within the USDatarecycler pipeline. Also presents the results of the original TransUnet architecture trained on different datasets.
- **Chapter 7: Conclusion and Future Work** Summarizes the main contributions and findings of this thesis. It discusses the implications of using Denoising Diffusion Models for RF data prediction and outlines potential directions for further research.

Chapter 2

Preliminaries

To provide a clear understanding of the scope and context of this thesis, this chapter begins with a brief overview of the foundational work on which this research is built. It then introduces the principles of Denoising Diffusion Models, the architectural framework explored in this thesis in relation to the USDatarecycler pipeline. Finally, the chapter concludes with a concise description of Ultrasound Imaging Models and Virtual Phantom Models, offering additional context for the application of these concepts in this thesis. For more information on these topics one can also take a look at [3].

2.1. Original USDatarecycler Pipeline

The USDatarecycler Pipeline, shown in Figure 2.1, was developed as part of [3], aiming to reconstruct raw RF signals from US images. This process addresses a significant challenge: During US image creation, non-linear transformations such as envelope detection and log-compression result in the loss of critical information. The pipeline offers a combined physics-driven and data-driven approach to predict RF signals while ensuring alignment with fundamental acoustic principles.

2.1.1. Components

The USDatarecycler Pipeline consists of two main components:

1. **Deep Learning Model:**
   This component employs a deep learning framework to predict maps of acoustic properties, referred to as phantom maps, directly from US images. These phantom maps provide the necessary inputs for the numerical simulation, enabling RF signal reconstruction.

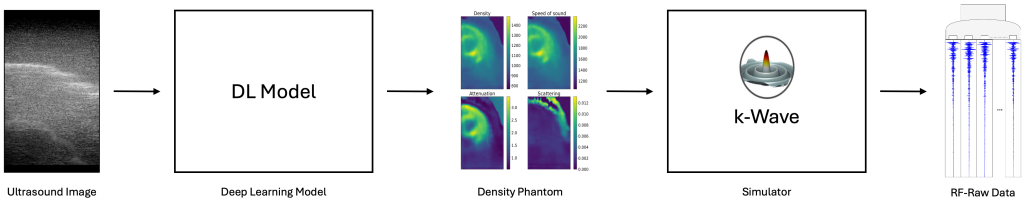


Figure 2.1.: Overview of the Pipeline introduced in USDatareyclerV1

2. **Physics-Informed Simulation (Simulator):**
   The second component uses the k-Wave toolbox [14] to numerically simulate ultrasound acquisition for a predicted imaging medium. By leveraging a physics-based simulation, it produces RF data that faithfully represents the physical properties of the medium. This simulation component ensures that the generated RF signals remain realistic and physically accurate, overcoming limitations inherent in purely data-driven models.

2.1.2. Model Training and Evaluation

The pipeline explored two different deep learning architectures, UNet[15] and TransUNet[12], to predict phantom maps from ultrasound images. The models' performance was evaluated by comparing predicted RF data with real RF signals using custom-designed metrics.

To further assess the effectiveness of the pipeline, its performance was compared with a baseline end-to-end model that directly predicted RF signals from US images. The inferior results of this end-to-end approach demonstrated the advantages of combining a numerical simulation with a deep learning model.

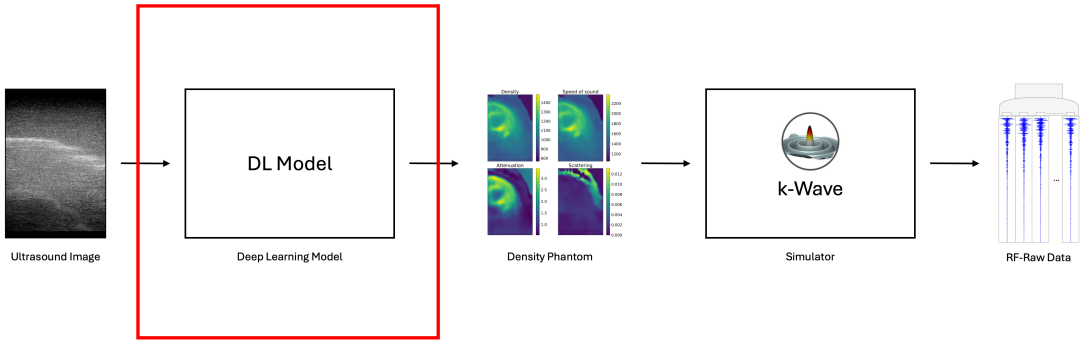


Figure 2.2.: Part of the USDatarecylcer Architecture that was changed in this work

2.2. Scope of this Thesis in the USDatarecycler Pipeline

This thesis focuses on advancing the first component of the USDatarecycler Pipeline, the Deep Learning model, as is shown in Figure 2.2. Specifically, the work involves the research, adaptation, and evaluation of state-of-the-art Denoising Diffusion models for predicting acoustic property maps, from now on also referred to as density phantoms or phantom maps, from US images.

Diffusion models have emerged as powerful generative modeling techniques, and their application in this domain presents an innovative approach to overcome the limitations of conventional deep learning architectures. Two adapted diffusion models, **Fast-DDPM** and **Marigold**, were explored in this thesis. Both models were tailored to the requirements of the USDatarecycler pipeline to improve the accuracy and fidelity of the predicted phantom maps. The best model was then integrated into the existing pipeline as shown in Figure 2.3.

By concentrating on the Deep Learning model within the pipeline, this work aimed to research if cutting-edge Diffusion models can enhance the reconstruction of RF signals and if they can improve performance on both synthetic and real-world datasets.

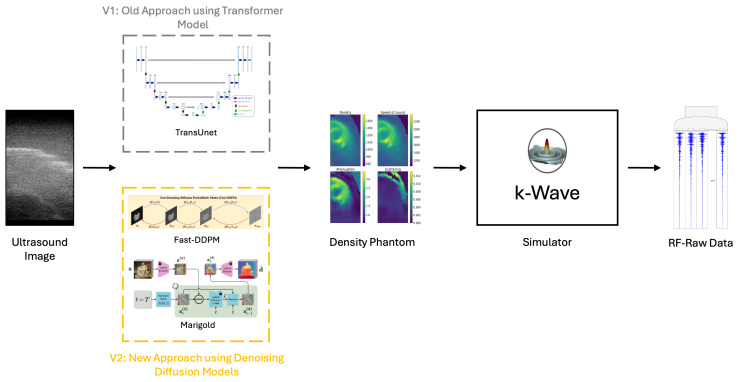


Figure 2.3.: USDatareyclerV1 vs USDatarecyclerV2

2.3. Denoising Diffusion Models

Denoising Diffusion Models [7] have emerged as the state-of-the-art in generative computer vision tasks, surpassing previous approaches such as GANs [8], VAEs [16], and other contenders. Their potential for high-quality image generation makes them particularly promising for the objectives of this thesis. This section provides an quick exploration of the core concepts, challenges, and model architectures related to diffusion models.

2.3.1. General Diffusion Approach

Denoising Diffusion Probabilistic Model (DDPM) operate by introducing noise to input data through a series of incremental steps, a process known as the **forward diffusion process**. At the final step, the data is transformed into pure noise. Conversely, the **reverse diffusion process** trains a neural network to iteratively denoise the input, step by step, until the original image is reconstructed. This reverse process is the cornerstone of generative capabilities, enabling the creation of high-fidelity images from pure noise, which is pivotal for various image generation tasks.

The forward diffusion process can be mathematically expressed as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}),$$

where $\mathbf{x}_t$ represents the noisy image at time step t, β_t is the noise variance at step t, and $\mathcal{N}$ denotes a Gaussian distribution. The parameter β_t is determined by the noise scheduler.

In the reverse diffusion process, the denoising network predicts the added noise $\epsilon_\theta(\mathbf{x}_t, t)$ at each step t, and the process is modeled as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(t)),$$

where $\mu_\theta(\mathbf{x}_t, t)$ is the predicted mean, and $\Sigma_\theta(t)$ is the predicted variance.

The primary training objective minimizes the difference between the actual noise ϵ and the predicted noise ϵ_θ, expressed as:

$$L_{\text{simple}} = E_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right].$$

This loss ensures the model learns to effectively de-noise inputs at every denoising time step.

2.3.2. Key Components of Diffusion Models

Two critical components define the performance of a diffusion model:

1. **The Deep Learning Backbone:** Most implementations employ variations of the U-Net architecture [15] for its ability to capture both local and global image features. However, recent advancements are moving towards transformer-based models, such as the Vision Transformer (ViT) [17], which excel in handling long-range dependencies and global context.
2. **The Noise Scheduler:** The noise scheduler determines the variance β_t added at each step of the forward process. A commonly used scheduler is the linear β_t schedule with 1000 steps. Alternatives, including cosine and exponential schedules, offer different trade-offs between model performance and convergence speed.

2.3.3. Computational Challenges

Despite their effectiveness, diffusion models face significant computational challenges:

- **High Training Costs:** Diffusion models require processing high-dimensional image data across numerous steps, leading to extended training times.
- **Slow Inference:** Generating an image from pure noise involves iterative denoising over hundreds or thousands of steps, which can be computationally prohibitive.

To address these issues, two key advancements have been explored:

- **Latent Diffusion Model (LDM):** These reduce training complexity by operating in a compressed latent space rather than pixel space, significantly lowering computational costs [1].
- **Accelerated Inference Techniques:** Methods to reduce the number of reverse steps while maintaining quality are discussed in the following subsections.

2.3.4. Latent Diffusion Models

LDMs are an extension of Denoising Diffusion Models that operate within a latent space rather than the original high-dimensional pixel space. This latent space is derived using a powerful encoder, typically based on pretrained computer vision backbones [18]. By mapping images to a compact latent representation, the computational burden of training and inference is significantly reduced.

The encoder compresses the input data into a lower-dimensional latent space, while a decoder reconstructs the final image after the denoising process. This approach allows the diffusion process to focus on essential features rather than redundant pixel-level details. Stable Diffusion [1, 19], one of the most prominent LDMs, exemplifies this strategy, achieving remarkable efficiency and quality in image synthesis tasks.

The architecture of the first rendition of Stable Diffusion [1] is depicted in Figure 2.4.

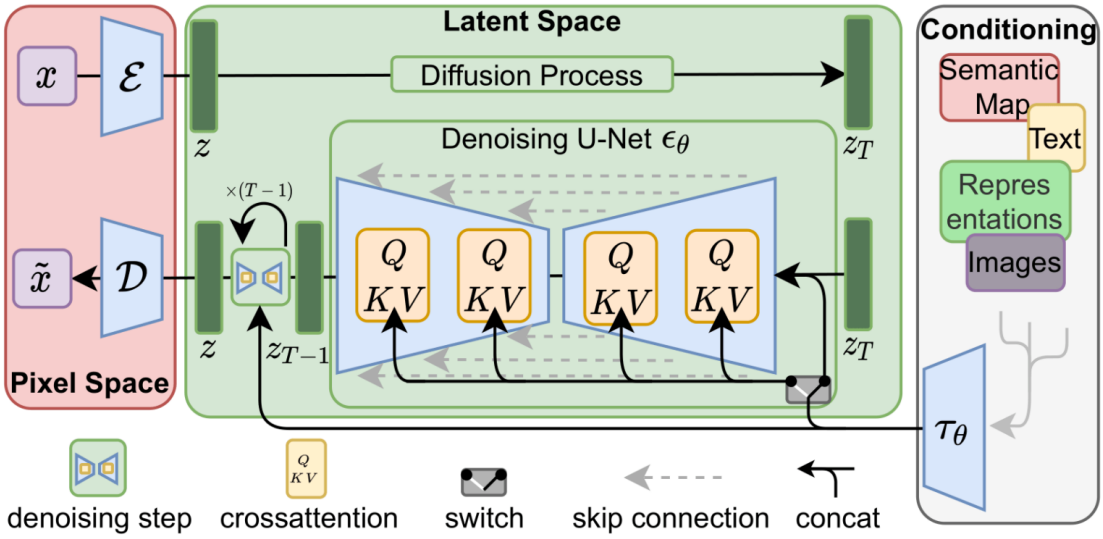


Figure 2.4.: Architecture of Stable Diffusion [1].

2.3.5. Approaches to Reduce Sampling Time in Diffusion Models

One of the primary limitations of diffusion models is the high computational cost associated with inference. Generating an image requires a large number of reverse diffusion steps, often taking several minutes per sample. This challenge is further compounded when multiple samples are generated for tasks like implicit ensembling, leveraging the probabilistic nature of diffusion models. Consequently, reducing the number of sampling steps has been a focal point of recent research, with two main approaches emerging:

Training-Free Approaches

Training-free methods aim to accelerate inference without retraining or modifying the original diffusion model. These approaches primarily focus on developing efficient solvers for the reverse diffusion process, often framed as solving Stochastic Differential Equation (SDE) or their deterministic counterparts, Ordinary Differential Equation (ODE).

Some of the most popular approaches are:

- **DDIM (Denoising Diffusion Implicit Models)** [20], which reformulates the reverse process to allow for fewer steps without significantly compromising image quality.
- **DPM-Solver** [21], a high-order ODE solver optimized for diffusion models.
- **Efficient Reverse Process Algorithms** [22], which explore conversion techniques to streamline the reverse diffusion process.
- **Reusing Attention Maps** [23], which leverages cached intermediate computations to enhance efficiency.

These approaches are model-agnostic, making them versatile solutions for reducing inference time.

Training-Based Approaches

Training-based methods address the inference time bottleneck by training a specialized model or distilling knowledge from the original diffusion model. Knowledge distillation involves using the original model as a teacher to guide the training of a student model capable of performing inference with significantly fewer steps while maintaining comparable performance.

Prominent works in this area include:

- **EM Diffusion** [24], which uses Expectation-Maximization to iteratively refine the student model.

- **Improved Distillation Techniques** [25], introducing strategies to better align the student's outputs with the teacher model.
- **Pagoda** [26], a framework for progressive distillation to reduce sampling steps incrementally.
- **SFDDM (Score Function Distillation for Diffusion Models)** [27], which directly distills score functions for accelerated sampling.
- **Score Compression Methods** [28], focusing on compact representations for faster diffusion processes.

While training-based approaches require additional computational effort upfront, they offer substantial long-term benefits in reducing inference time for practical deployments.

2.4. Ultrasound Imaging Modes

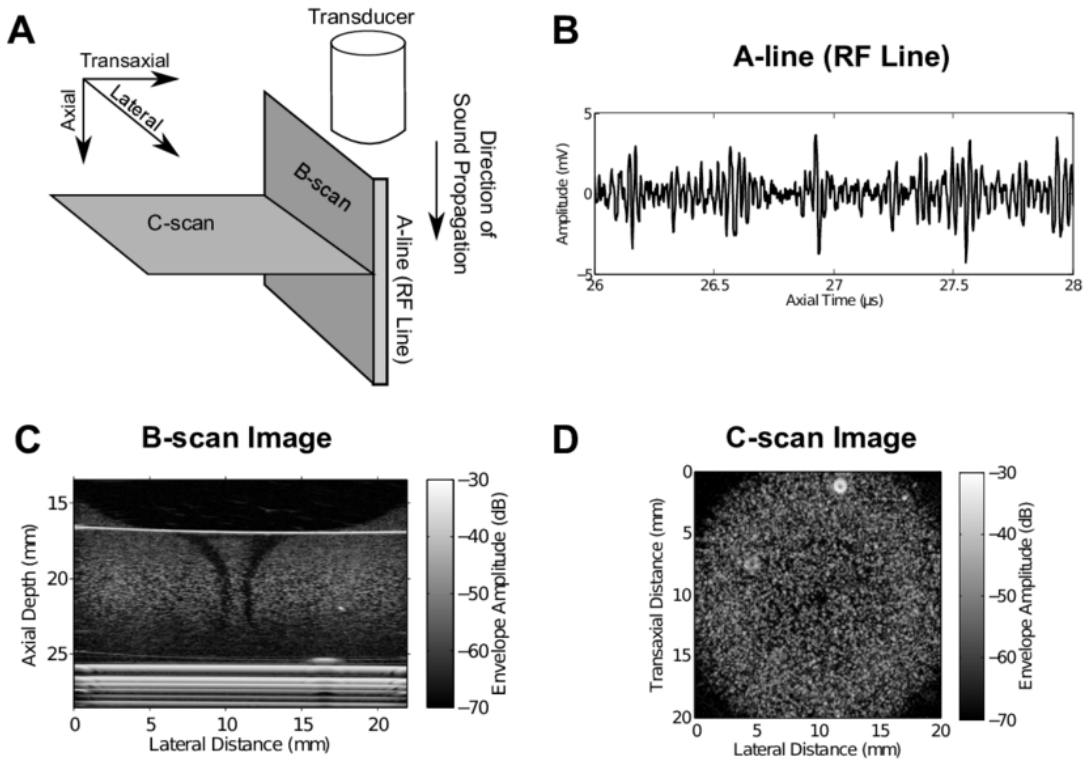


Figure 2.5.: Visualization of different Ultrasound Imaging Modes [2]

Ultrasound imaging encompasses several different modes, each suited to specific applications. Those modes are showcased in Figure 2.5.

The simplest mode is **A-mode** (Amplitude mode), where a single focused ultrasound beam is transmitted, and echoes are received on a single channel. The resulting plot represents the amplitudes of the reflected signals as a function of depth, making it a one-dimensional imaging technique.

The most commonly used mode, also the exclusively used mode in this thesis, is **B-mode** (Brightness mode). In B-mode imaging, raw ultrasound data is collected across multiple channels and processed to form a two-dimensional gray-scale image, where pixel brightness corresponds to the amplitude of the reflected signal. This mode is widely used in medical imaging due to its ability to provide detailed cross-sectional views of anatomical structures.

Another mode is **M-mode** (Motion mode), where a stationary transducer emits a sequence of ultrasonic pulses to capture either A-mode or B-mode images over time. This mode is particularly useful for visualizing and measuring motion, such as the movement of organs or blood flow.

Traditional B-mode imaging typically forms the ultrasound image line by line using focused ultrasound beams, optimizing lateral resolution. However, ultra-fast ultrasound imaging introduces an alternative approach by transmitting a single unfocused plane wave across the entire field of view at once. This method achieves significantly higher frame rates but at the cost of reduced lateral resolution.

2.5. Virtual Phantom Model

The virtual phantom model used in this work defines the relevant acoustic properties of the ultrasound imaging medium. It is represented as a two-dimensional grid G with height N_z and width N_x:

$$G = \{(i, j) \mid 1 \leq i \leq N_z, 1 \leq j \leq N_x\}$$

Here, the lateral direction is denoted as the x-axis and the axial direction as the z-axis. The grid discretizes the physical propagation medium with a uniform spacing Δ_{xz} in both directions. The transducer is placed at the top of the grid along the x-axis, transmitting pulses along the z-axis.

Each grid point (i, j) is assigned four acoustic properties that define the virtual phantom:

- **Density** $\rho(i, j)$ [kg/m^3]
- **Speed of Sound (SOS)** $c(i, j)$ [m/s]
- **Attenuation Coefficient** $\alpha(i, j)$ [dB/(MHz $\cdot$ cm)]
- **Scattering Parameter** $s(i, j)$ (unitless)

These properties are represented as four matrices of size $N_z \times N_x$, forming a $4 \times N_z \times N_x$ tensor referred to as the virtual phantom P. The four matrices are also referred to as the Phantom Channels throughout this work.

2.5.1. Scattering Model

The scattering parameter $s(i, j)$ introduces variability to the density map $\rho(i, j)$ to simulate diffusive scattering caused by sub-wavelength structures (scatterers). The density map influenced by scattering is denoted as $\rho_s(i, j)$, defined by:

$$\rho_s(i, j) = s^*(i, j) \cdot \rho(i, j)$$

where $s^*(i, j)$ is a scattering factor derived from a probabilistic model.

Scatterers are identified using a Bernoulli distribution with a probability υ_s, which specifies the density of scatterers in a given area. If a grid point is classified as a scatterer, its density value is sampled from a Gaussian distribution:

$$\upsilon_s(i, j) = \begin{cases} \mathcal{N}(\mu_s(i, j), \sigma_s(i, j)^2) & \text{if scatterer exists} \\ \rho(i, j) & \text{otherwise} \end{cases}$$

Here, $\mu_s(i, j) = \rho(i, j)$, and the scattering variance $\sigma_s(i, j)^2$ is proportional to both $s(i, j)$ and $\rho(i, j)^2$.

Experimentally, $s(i, j)$ is limited to the range $[0, 0.015625]$, ensuring that scattering does not introduce artifacts or distort the target's appearance.

2.5.2. Simplifications and Assumptions

To simplify the model and ensure computational efficiency:

- The attenuation power exponent and non-linearity parameter B/A are not modeled.
- Scattering effects are applied only to density, as changes in the speed of sound introduce computational overhead.
- Ultrasound acquisitions are modeled linearly, disregarding non-linear effects.

This approach balances realism and efficiency, enabling the generation of realistic virtual phantoms for the simulation of ultrasound acquisitions.

Chapter 3

Related Work

This thesis focuses on utilizing Denoising Diffusion Models[7] for translating medical US images to ultrasound density phantoms. Consequently, this chapter explores related models in the domains of I2I-Translation and Semantic Segmentation, as they closely align with the task at hand.

3.1. Diffusion Models for Medical Imaging

A comprehensive survey on Diffusion Models in medical imaging is presented in [29], summarizing the most prominent models up to the end of 2022. While diffusion models are predominantly applied for synthetic image generation, their utility extends to I2I translation and semantic segmentation tasks, which are of particular relevance to this thesis.

3.1.1. I2I Translation

Several models utilize conditional diffusion approaches for medical I2I translation [30, 31, 22]. These models integrate information from conditional input images to guide the generation of corresponding outputs. Additionally, non-medical I2I translation models, such as PITI [32], Palette [33], and DiffI2I [34], provide valuable insights applicable to this work.

Two models particularly relevant to this thesis are explored in detail:

BBDM: Image-to-Image Translation with Brownian Bridge Diffusion Models

BBDM [35] innovates by performing domain-to-domain translation without relying on standard conditional generation approaches. Operating in the latent space of a pretrained VQ-GAN [36], BBDM simplifies the generation process by targeting the conditional image during the forward diffusion process, effectively transforming the target image into the condition image. This streamlined approach enhances computational efficiency.

The loss function in BBDM is a simplified version of the Evidence Lower Bound (ELBO), enabling efficient training. Sampling is accelerated using DDIM [20], requiring only 200 steps. Notably, BBDM's training on a single NVIDIA 3090 GPU highlights its accessibility, leveraging the reduced complexity of latent space operations.

BBDM demonstrates superior performance compared to the standard Latent Diffusion Model in I2I translation tasks, emphasizing its effectiveness in latent space processing. Its latent-space-focused architecture aligns with the goals of this thesis, as the intermediate phantom maps predicted in the USDatarecycler Pipeline are analogous to domain translations in BBDM's framework.

Fast-DDPM: Fast Denoising Diffusion Probabilistic Models for Medical Image-to-Image Generation

Fast-DDPM [10] builds upon DDIM [20] to enhance both resource and time efficiency, achieving training and sampling with only 10 diffusion steps. This reduces training time by a factor of 5 and sampling time by a factor of 100 compared to conventional DDPM [37] models. This efficiency stems from task-specific noise schedulers, manually designed to optimize performance for specific tasks. During inference, the DDIM sampler operates only over these 10 steps, further accelerating the process.

Fast-DDPM's experiments on medical imaging datasets demonstrate its superiority over traditional DDPM models and GAN-based methods, with significantly reduced training and inference times. The insights provided by Fast-DDPM's noise scheduling and step reduction align with this thesis' focus on optimizing diffusion models for computational efficiency, particularly when adapting them for the US to phantom map translation task.

3.1.2. Semantic Segmentation

Semantic segmentation represents another Computer Vision (CV) task closely related to this thesis. Among the most promising models for medical imaging are the MedSegDiff series [38, 13], which utilize diffusion processes for segmentation tasks. MedSegDiff-V2, the latest iteration, is particularly noteworthy.

MedSegDiff-V2: Diffusion-Based Medical Image Segmentation with Transformers

MedSegDiff-V2 [13] extends the capabilities of MedSegDiff by incorporating a TransUNet [12] architecture instead of a standard U-Net [15]. Initial results with the TransUNet revealed suboptimal performance, prompting refinements in the conditioning mechanisms and architecture.

MedSegDiff-V2 employs two U-Nets: one for the diffusion model and another for encoding input images into conditioning representations. Conditioning mechanisms include an anchor conditioning at the start of the reverse diffusion process and a semantic conditioning during the bottleneck phase. The latter utilizes a Spectral-Space Transformer (SS-Former) with a Neural Band Pass Filter to fuse diffusion and condition embeddings in Fourier space. This innovative conditioning approach improves segmentation performance significantly.

The loss function combines the standard diffusion noise loss with Dice and Cross-Entropy (CE) losses applied at specific time steps, balancing segmentation accuracy and the diffusion process. MedSegDiff-V2's superior results across datasets like BraTS [39, 40, 41] and ISIC [42, 43] underscore its effectiveness.

The novel conditioning and architectural advancements of MedSegDiff-V2 offer valuable insights for adapting DL models in this thesis, particularly in integrating complex input-output mappings for accurate phantom map prediction.

3.2. Diffusion Models for Other Purposes

Diffusion models' versatility extends beyond I2I translation and segmentation. Two notable examples, Stable Diffusion [1] and Marigold [11], highlight their adaptability for diverse applications relevant to this thesis.

3.2.1. Stable Diffusion

Stable Diffusion [1] represents a groundbreaking text-to-image generative model, developed by Stability AI. It operates in a latent space, utilizing an auto-encoding variational framework to guide image generation from text prompts. A U-Net architecture with cross-attention mechanisms facilitates effective integration of textual and visual features.

The model's versatility extends to domain adaptation and inverse problem-solving tasks, offering insights into latent space optimization. Stable Diffusion's framework directly informs this thesis by demonstrating the potential of latent diffusion methods for generative and predictive tasks, such as translating US images into phantom maps.

3.2.2. Marigold: Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation

The Marigold [11] model adapts Stable Diffusion's framework for monocular depth estimation. By fine-tuning the pretrained U-Net component while keeping the Autoencoder frozen, Marigold excels at extracting depth-specific features. Its success underscores the adaptability of diffusion models for I2I-generation tasks in general.

For this thesis, Marigold's approach highlights the benefits of leveraging pretrained models and latent space representations for specialized tasks. The principles underlying Marigold's fine-tuning strategy align with the adaptation of diffusion models for accurate phantom map prediction, therefore it was chosen as one of the main models to adapt and specialize for the task of this thesis.

Chapter 4

Implementation

This chapter provides a detailed description of the implementations carried out during this thesis and explains the architectural decisions behind them.

For the Image-to-Phantom Translation task, two diffusion models were selected. The first model is an adaptation of Fast-DDPM [10], serving as a simpler baseline for non-pretrained diffusion models. It was chosen because, despite its straightforward design, it has demonstrated strong performance on related medical I2I tasks. Additionally, its inference speed-up approach offered potential for efficiently generating large amounts of synthetic RF data.

The second, more complex model is a variation of the Marigold model [11], evaluated in multiple configurations. The Marigold model was selected as a platform for further development because it delivered state-of-the-art results in its domain of monocular depth estimation, which shares similarities with I2I translation. Furthermore, it utilizes a robust pretrained backbone in the Stable Diffusion model, which was expected to provide significant advantages.

4.1. Adapted Fast-DDPM

The original Fast-DDPM model [10] is a denoising diffusion model designed with the principle that the number of denoising steps during training should match those used during inference. While this approach ensures fast performance, it also results in a relatively simple diffusion model. Fast-DDPM was chosen as a baseline to assess the potential of a less complex diffusion model for this task.

4.1.1. Architecture of the Adapted Fast-DDPM Model

The adaptations to the Fast-DDPM model were kept minimal, focusing only on modifying its input and output structures to accommodate the 4-channel phantom input instead of the original single-channel configuration. Two distinct architectures were implemented: the Multi-Channel and Single-Channel approaches, detailed below.

Multi-Channel Input and Output

In the Multi-Channel configuration, the model was adapted to accept 5 input channels—4 channels for the phantom and 1 for the ultrasound image—and to produce 4 output channels corresponding to the phantom predictions. This architecture is illustrated in Figure 4.1.

This approach was not included in the final evaluation, as early experiments showed significantly poorer performance compared to the Single-Channel configuration, described next.

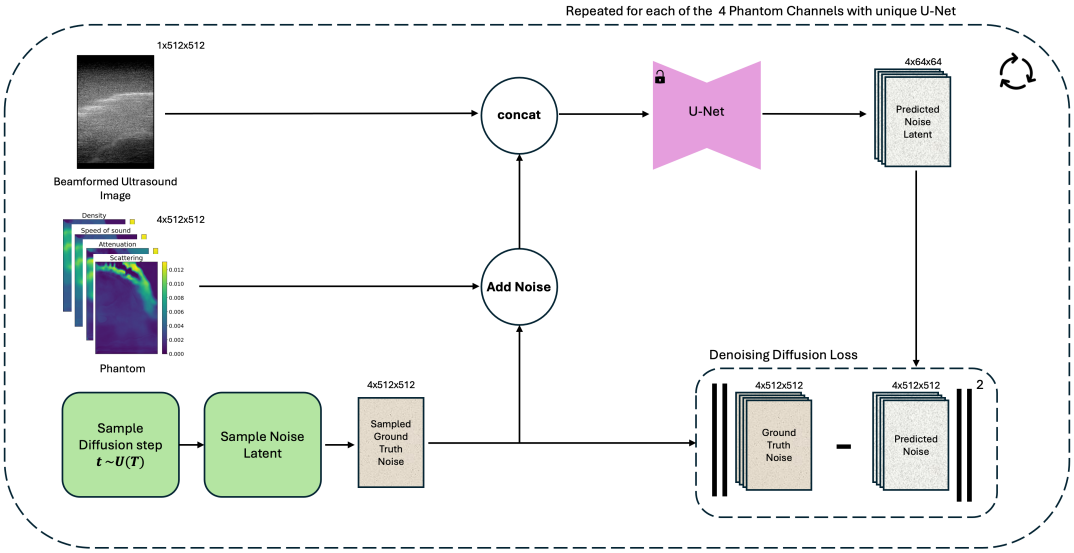


Figure 4.1.: Fast-DDPM Multi-Channel Training Process

Single-Channel Input and Output

Given the lackluster results of the Multi-Channel configuration, a Single-Channel variant was implemented. In this architecture, the original input structure of 2 input channels

and 1 output channel was retained. The strategy involved training four separate Fast-DDPM models, each responsible for predicting a single phantom channel, with the results combined to produce the complete phantom. This setup is illustrated in Figure 4.2.

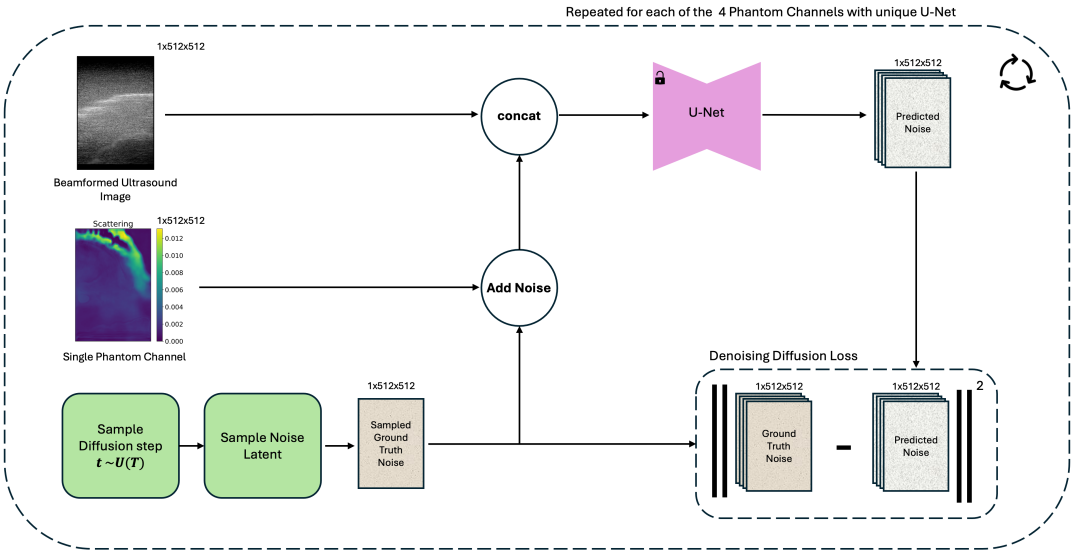


Figure 4.2.: Fast-DDPM Single-Channel Training Process

4.1.2. Training of the Adapted Model

Prototype versions of the adapted Fast-DDPM model were trained for both the Multi-Channel and Single-Channel configurations. These prototypes were run for 14 days on four NVIDIA RTX 4090 Graphics Processing Unit (GPU)s each.

By the time these experiments were underway, the Adapted Marigold model, discussed later in this chapter, had also reached the prototype phase. The results from the Adapted Marigold model were significantly more promising, leading to a decision to discontinue further investigation of the Fast-DDPM architecture. Consequently, no results from the adapted Fast-DDPM were included in the final evaluation.

4.2. Adapted Marigold

The Marigold model [11] was selected as the secondary diffusion model base for adaptation in this project due to its outstanding performance in the domain of monocular depth estimation. This task, which fundamentally involves I2I translation, closely aligns

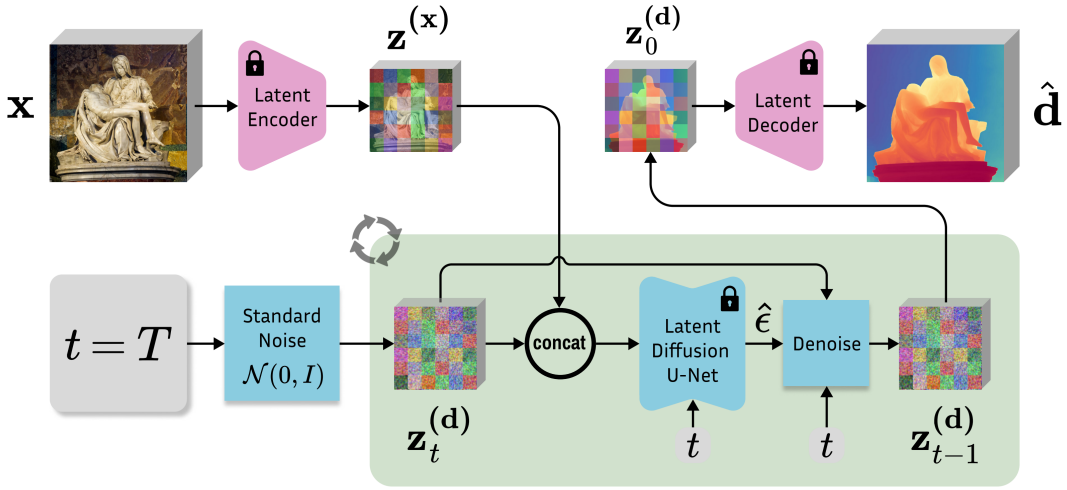


Figure 4.3.: Original Marigold Model Inference Architecture

with the Image-to-Phantom translation task addressed in this thesis. Marigold's demonstrated capabilities in handling such translation tasks made it a compelling candidate for exploring a more robust and complex approach.

The architecture of the original Marigold model for monocular depth estimation is shown in Figures 4.3 and 4.4. During training, the model uses a 3-channel RGB input image paired with a single-channel depth map. To ensure compatibility with the pretrained Stable Diffusion Autoencoder, the single-channel depth map is repeated across three channels to create a pseudo-RGB depth image. This step is essential, as the Autoencoder expects a standard 3-channel input format.

Both the 3-channel RGB input and the repeated 3-channel depth map are independently encoded into the latent space by the Autoencoder, resulting in two latent representations of size (4, 64, 64). These representations are concatenated along the channel dimension to form a combined latent input of size (8, 64, 64). This concatenated latent space serves as the input to the U-Net, where the diffusion process occurs. The U-Net processes this input and produces an output in the latent space with the original 4-channel structure. Finally, the Autoencoder decodes this 4-channel latent output back into a single-channel depth map, completing the translation from RGB input to depth output.

Notably, during this process, the Stable Diffusion Autoencoder remains frozen, meaning its weights are not updated. Instead, only the weights of the U-Net are trained, ensuring the pretrained encoding and decoding processes remain stable and unchanged.

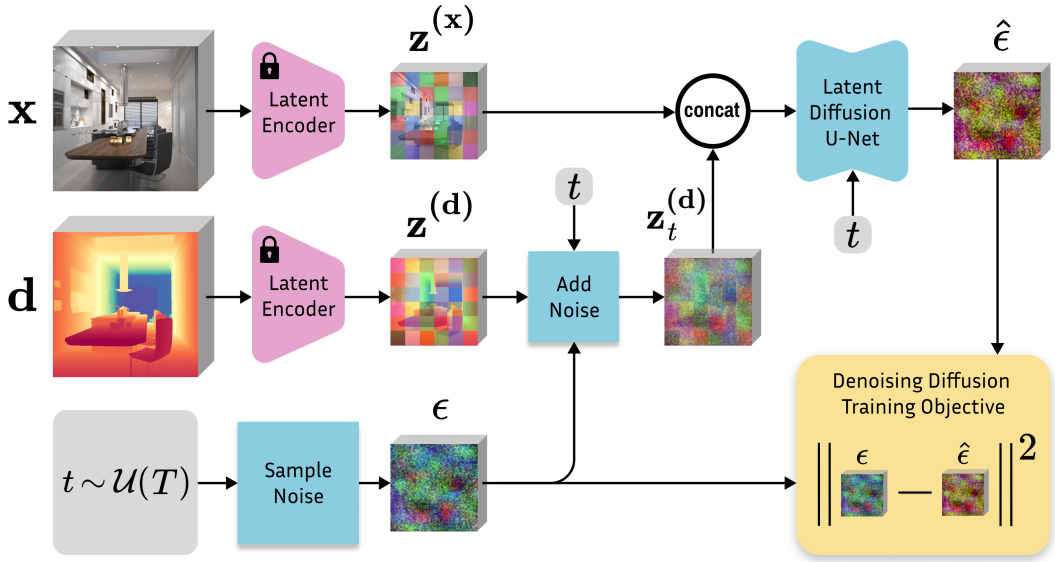


Figure 4.4.: Original Marigold Model Training Architecture

4.2.1. Architecture of the Adapted Marigold Model

Figure 4.5 illustrates the final adapted Marigold inference architecture. The chosen design employs a Single-Channel input variant, leveraging a two-step fine-tuning routine. The different architectural choices and corresponding ablations are detailed in the following sections.

Multi-Channel Input and Output

Initially, Marigold was adapted to handle the 4-channel nature of the Phantom ground truth images by modifying its input-output structures as shown in Figures 4.7 and 4.6. In this approach, the 4-channel Phantom image was split into four individual single-channel images. Each single-channel image was repeated across three channels to match the 3-channel input requirement of the Autoencoder. These repeated 3-channel images were independently encoded into the latent space by the Autoencoder.

The encoded latent representations were concatenated along the channel dimension to form a combined latent input of size (20, 64, 64). This expanded latent input was processed by the U-Net during the diffusion process, producing a latent representation of size (16, 64, 64). The 16-channel output was then split into four separate 4-channel latent spaces, each decoded independently into single-channel images. These single-channel outputs were concatenated to reconstruct the 4-channel Phantom image.

While this method allowed Marigold to handle the multi-channel phantom data at once, the increased complexity of the U-Net input-output structure posed challenges. The

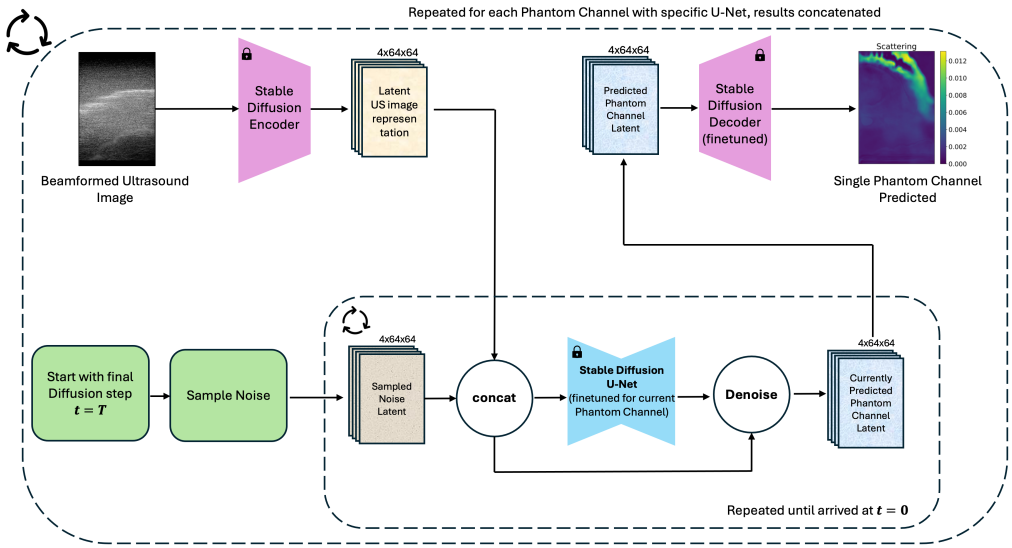


Figure 4.5.: Adapted Marigold Model Main Inference Architecture

larger latent space potentially diluted the U-Net's ability to focus on individual channel-specific details. These limitations prompted the exploration of an alternative, simpler, and ultimately more effective approach.

Single-Channel Input and Output

The final and more effective approach to managing the input-output structures closely adhered to the original Marigold architecture [11]. Instead of modifying the U-Net's input-output structure to accommodate multi-channel data, the Phantom Ground Truth was split into four separate single-channel images at the outset. Each channel was then assigned its own dedicated U-Net, which was trained independently while preserving the original input and output dimensions of the Marigold U-Net. This approach is illustrated in Figures 4.8 and 4.9.

This method offered several significant advantages. By maintaining the original input-output dimensions, the architectural complexity of the model was reduced, streamlining implementation and ensuring compatibility with the pretrained Stable Diffusion Autoencoder. Furthermore, training separate U-Nets for each channel allowed for a specialized focus on the unique characteristics of each channel. This channel-specific training ensured that each U-Net was better optimized for its respective channel, leading to improved performance in the Image-to-Phantom translation task.

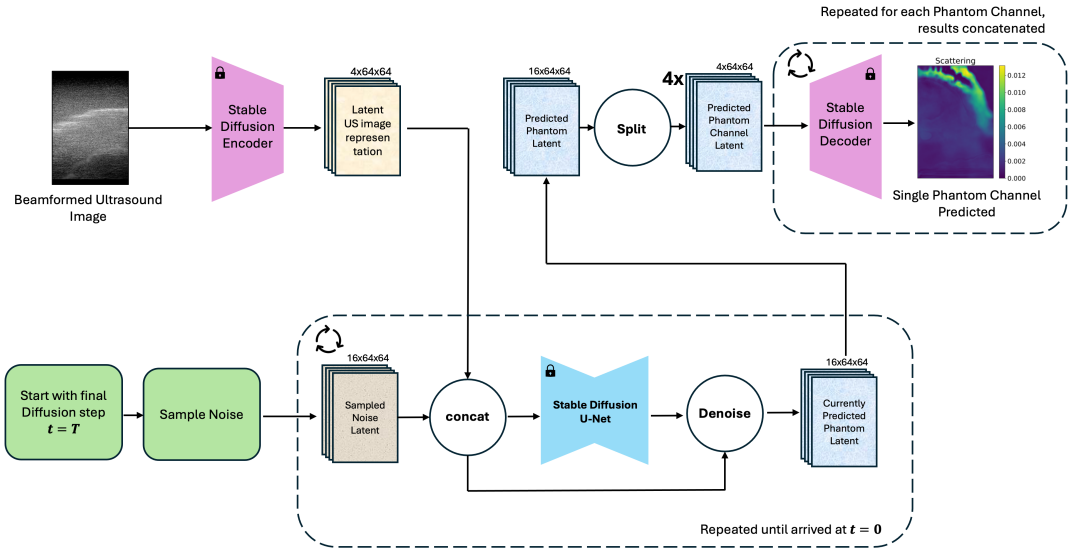


Figure 4.6.: Adapted Marigold Model Inference with Multi-Channel Input

After the individual U-Nets processed their respective single-channel inputs, the outputs were recombined to reconstruct the 4-channel Phantom image. This modular and focused approach effectively leveraged the strengths of the original Marigold framework while addressing the unique challenges of multi-channel data in a structured and efficient manner. Consequently, this approach was adopted as the final solution for the task.

4.2.2. Additional VAE Fine-Tuning

In the original Marigold architecture, the Stable Diffusion VAE remains frozen throughout the training process. However, as Stable Diffusion was primarily trained on natural images [44], which differ significantly from Phantom and Ultrasound data, this posed a limitation for our specific task. To address this, a two-step training routine was introduced as an iterative improvement to the architecture. This routine fine-tunes the VAE to better encode and decode data from our dataset, ensuring it is more effectively adapted for the Image-to-Phantom translation task.

The VAE fine-tuning was implemented in three distinct manners. The first approach involved fine-tuning the entire VAE, while the second approach focused on fine-tuning only the decoder. In the third approach, only the last layer of the decoder was fine-tuned. These methods are illustrated in Figures 4.10, 4.11, and 4.12. Their performance is compared in the results section.

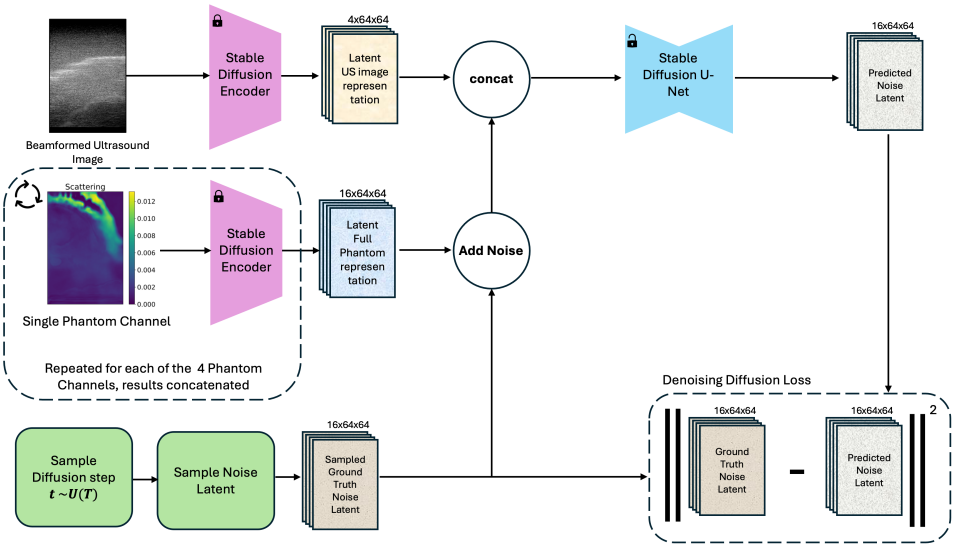


Figure 4.7.: Adapted Marigold Model Training with Multi-Channel Input

Following the fine-tuning process, the VAE is frozen again, and the standard training procedure for the U-Net and diffusion process is resumed. This two-step routine enables the model to retain the strengths of the original Stable Diffusion VAE while adapting to the specific characteristics of our dataset.

The reconstruction loss function used for fine-tuning the VAE is shown in Figure 4.13. To prevent catastrophic forgetting, the learning rate during fine-tuning was set significantly lower compared to the original training learning rate of the VAE.

Multi-VAE

Building on the success of the VAE fine-tuning, an extension was implemented to train separate VAEs for each Phantom channel, leading to the Multi-VAE approach. In the standard VAE fine-tuning, a single VAE is trained using data from all Phantom channels. While effective, this approach may limit the ability to specialize for the unique features of individual channels.

In the Multi-VAE approach, four distinct VAEs are fine-tuned, each trained exclusively on data from a specific Phantom channel. This specialization allows each VAE to adapt to the unique characteristics and features of its respective channel, theoretically enhancing the reconstruction quality for each channel.

The architecture implementing the Multi-VAE approach is shown in Figures 4.14 and 4.15. By enabling channel-specific fine-tuning at the VAE level, this strategy ensures a more

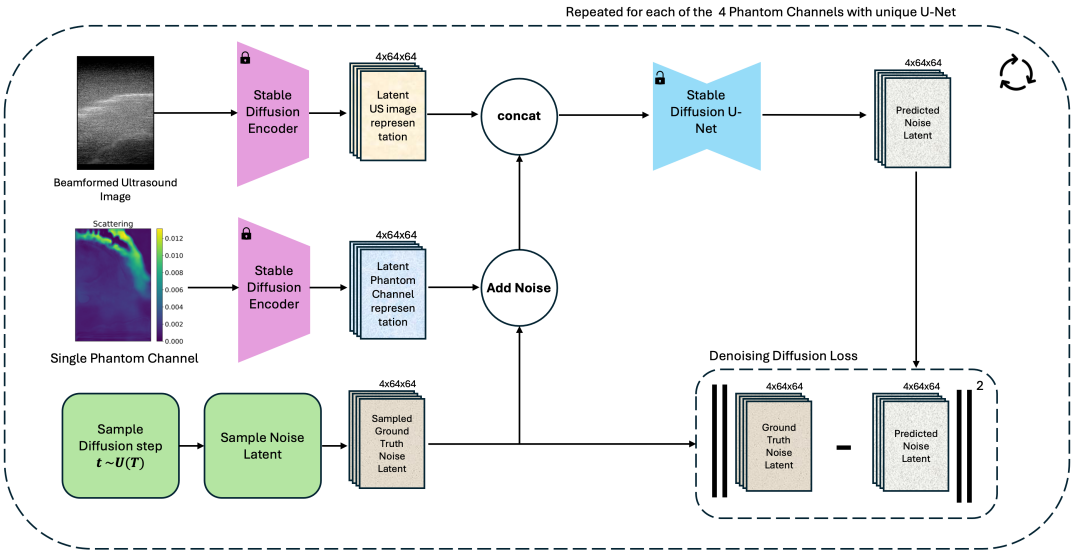


Figure 4.8.: Adapted Marigold Model Training with Single-Channel Input

precise encoding and decoding process, tailored to the distinct attributes of each Phantom channel. This design further improves the model's capacity to generate high-quality 4-channel outputs in the Image-to-Phantom translation task.

4.3. TransUNet

The original TransUNet architecture from [3] was utilized without any modifications. This ensured consistency with its established design, functionality, and performance characteristics. The original architecture is depicted in Figure 4.16.

In this project, the TransUNet model was retrained using the newly curated datasets. The results obtained from these retraining efforts served as a baseline for comparison with all other experiments conducted in this thesis. By establishing a consistent and reliable reference point, the retrained TransUNet enabled a comprehensive evaluation of the improvements introduced by the adapted Marigold architecture and other proposed methodologies.

4.4. Synthetic Datasets

In this thesis, multiple datasets were utilized to train and evaluate the models. The two primary datasets, the CT-Phantom 500k and CT-Phantom 50k datasets, were created by

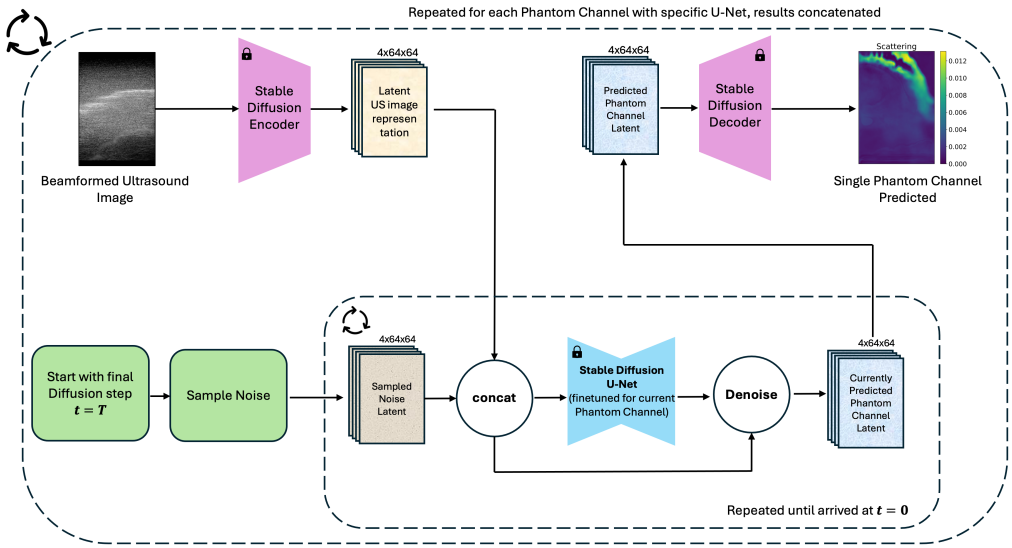


Figure 4.9.: Adapted Marigold Model Inference with Single-Channel Input

processing existing CT data. These datasets were generated by creating virtual phantoms based on CT scans and simulating corresponding beamformed US images using the k-Wave simulator [14] and the beamforming capabilities of dasIT [45]. Additionally, a small set of real-world US-RRF data pairings was included as external test sets, used exclusively for evaluation purposes.

The TransUNet model was trained on all available training datasets, whereas the Diffusion Models were exclusively trained on the primary CT-Phantom 500k dataset.

4.4.1. CT-Phantom Datasets

The CT-Phantom datasets, referred to as Phantom datasets for brevity, formed the backbone of training and evaluation for all models in this thesis. Each dataset sample, as illustrated in Figure 4.17, consists of a single-channel gray-scale US image paired with a corresponding 4-channel phantom image. The phantom images are suitable for simulation purposes and can be integrated into a physically informed model framework. Detailed descriptions of the phantom generation process and its underlying principles can be found in [3].

An overview of the dataset characteristics and structure is presented in Figure 4.18, while the generation process of a dataset sample is shown in Figure 4.19. This process is detailed in [3].

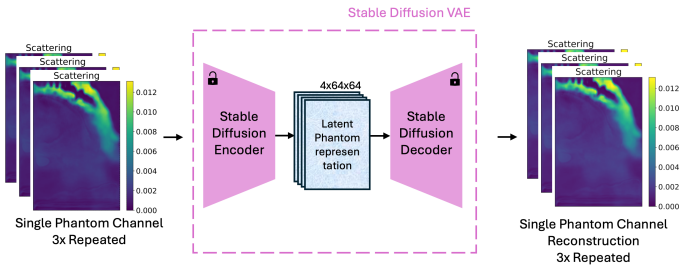


Figure 4.10.: Adapted Marigold Model VAE Full Fine-Tuning

Old 50k CT-Phantom Dataset

The Old 50k CT-Phantom Dataset was utilized in the first rendition of the USDatarecycler pipeline [3]. This dataset consists of 50,000 samples and was generated using the functionality of the USDatarecyclerV1 pipeline. However, during the dataset generation, a bug in the US beamforming process resulted in the corruption of a small percentage of US images.

50k CT-Phantom Dataset

The 50k CT-Phantom Dataset is a smaller subset of the 500k dataset, containing 50,000 samples. Unlike the Old 50k CT-Phantom Dataset, this dataset was generated after fixing the US beamforming bug, ensuring no corruption of US images. It was created as a random selection of samples from the 500k dataset and stored as a single chunk to facilitate ease of use in training and testing workflows. Its smaller size was chosen to evaluate the necessity of training on the larger 500k dataset.

500k CT-Phantom Dataset

The 500k CT-Phantom Dataset is the primary dataset used for training and evaluating models in this thesis. It comprises 500,000 samples, divided into training, validation,

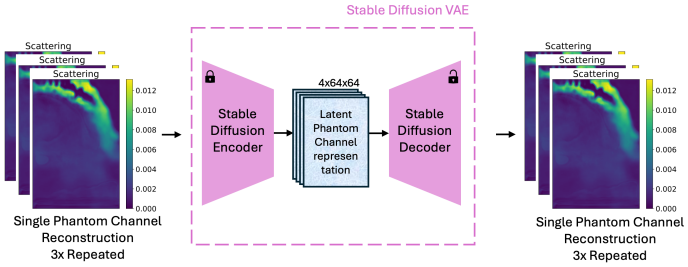


Figure 4.11.: Adapted Marigold Model VAE Decoder Fine-Tuning

and testing subsets. Due to its large size, the dataset was partitioned into 50 individual chunks, each containing 10,000 samples, to enable efficient data handling and storage. This chunk-based structure was instrumental in managing the dataset effectively within the constraints of computational resources.

4.5. Parallelized SLURM Chunk-Loader

The use of the Euler Cluster [46], which operates with SLURM [47] and enforces a disk space allocation cap per job, necessitated chunking the dataset and loading it dynamically. Since no existing solution adequately addressed these requirements, a custom chunk-loading functionality was implemented, as illustrated in Figure 4.20.

This implementation consists of a Chunkloader script that runs concurrently with standard model training or any process requiring data, using a PyTorch [48] Dataset object. Communication between the dataset class and the Chunkloader script occurs through a text file. Commands are written to this file by the dataset class and are subsequently queued and processed in a first-in-first-out (FIFO) manner by the Chunkloader script.

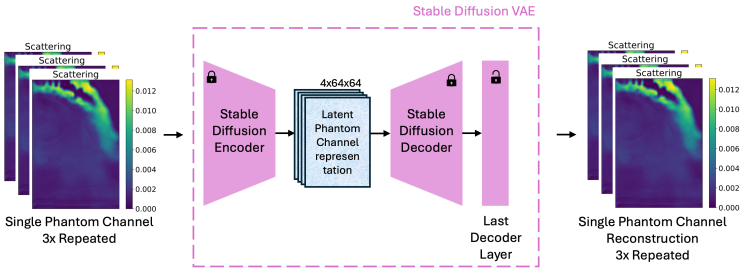


Figure 4.12.: Adapted Marigold Model VAE Decoder Last Layer Fine-Tuning

4.5.1. Storage Handling

The Chunkloading mechanism addresses the limited disk storage space available on compute nodes. To manage this, only a specific number of chunks is transferred from the dataset's main server location to the compute node storage. These chunks, stored as tar.gz archives, are extracted on the compute node and made ready for use. This storage workflow is depicted in Figure 4.21.

4.5.2. Functionality

The core functionality of the Chunkloader is built around a few essential commands, detailed below:

Initialization:

The initialization process, shown in Figure 4.22, begins with the Chunkloader loading a predefined number of initial chunks and creating the Command Handler text file. Once this file is established, the dataset class can operate seamlessly, and the chunked dataset is initialized and ready for use.

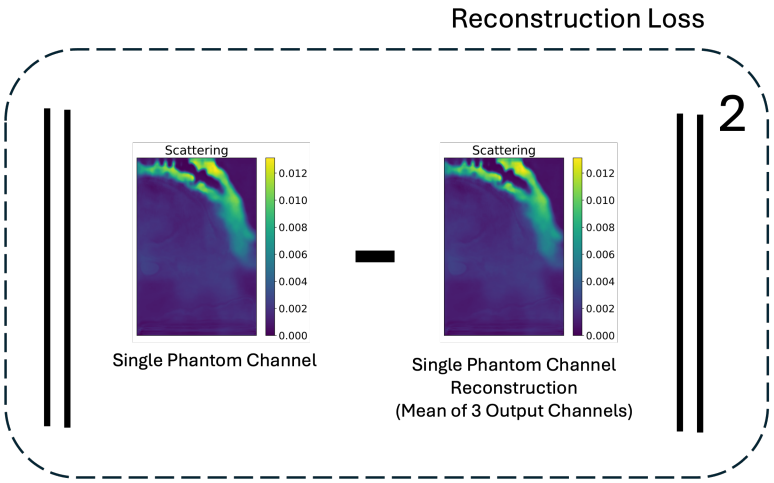


Figure 4.13.: Adapted Marigold Model VAE Loss Function for Fine-Tuning

Remove Command:

The remove command, illustrated in Figure 4.23, is central to the Chunk-loader's operation. This command removes specific chunks that are no longer required, freeing up memory and ensuring efficient use of available storage. After removing unnecessary chunks, new chunks are automatically loaded until a predefined number of chunks is available, maintaining the system's readiness for data processing.

Reset Command:

The reset command, shown in Figure 4.24, enables the clearing of all currently loaded chunks and re-initializes the Chunkloader. This functionality is particularly useful for restarting workflows or transitioning to a new dataset. By ensuring that the Chunkloader begins in a clean state, the reset command minimizes the risk of residual data interfering with subsequent operations.

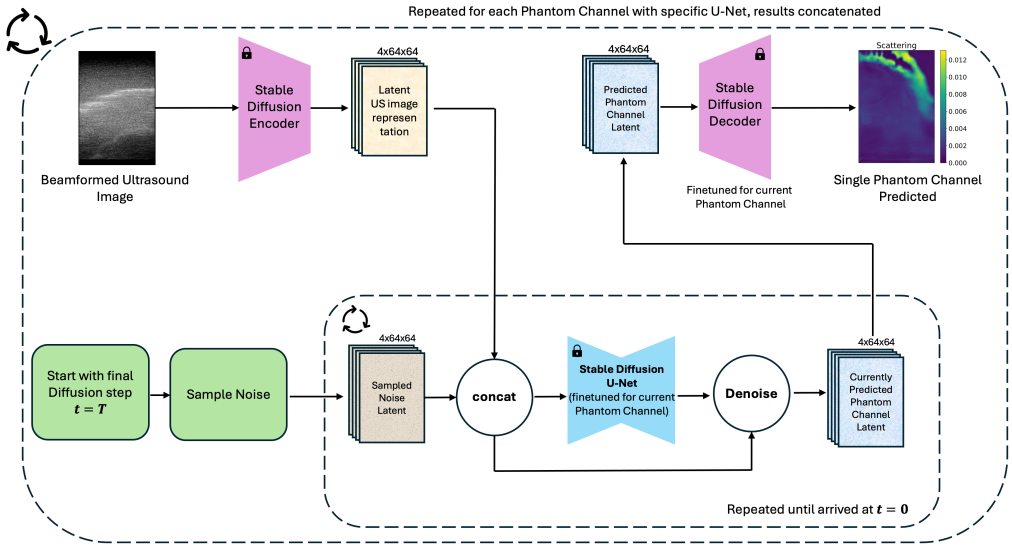


Figure 4.14.: Adapted Marigold Model Inference Architecture with Multi-VAE

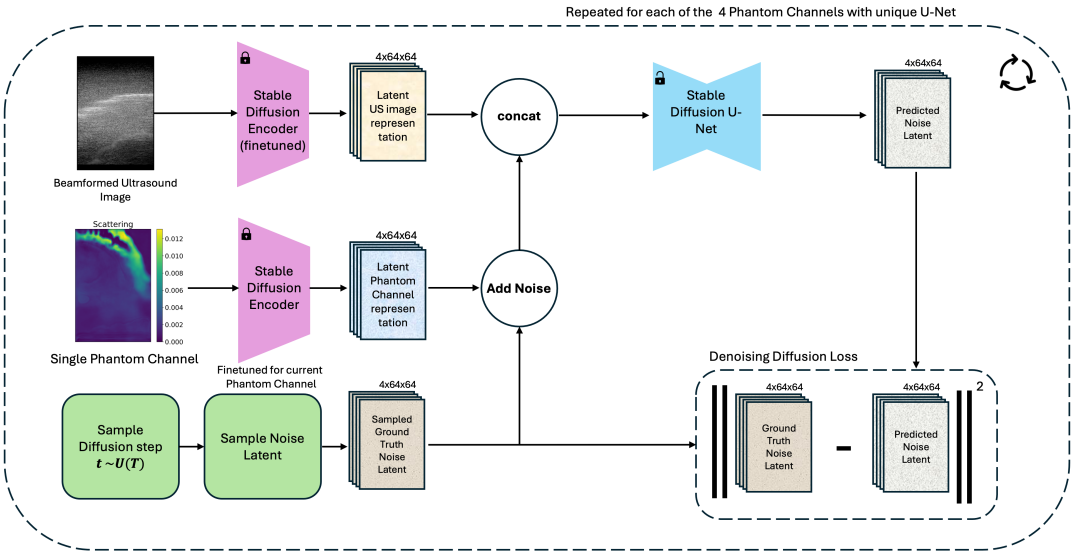


Figure 4.15.: Adapted Marigold Model Training Architecture with Multi-VAE

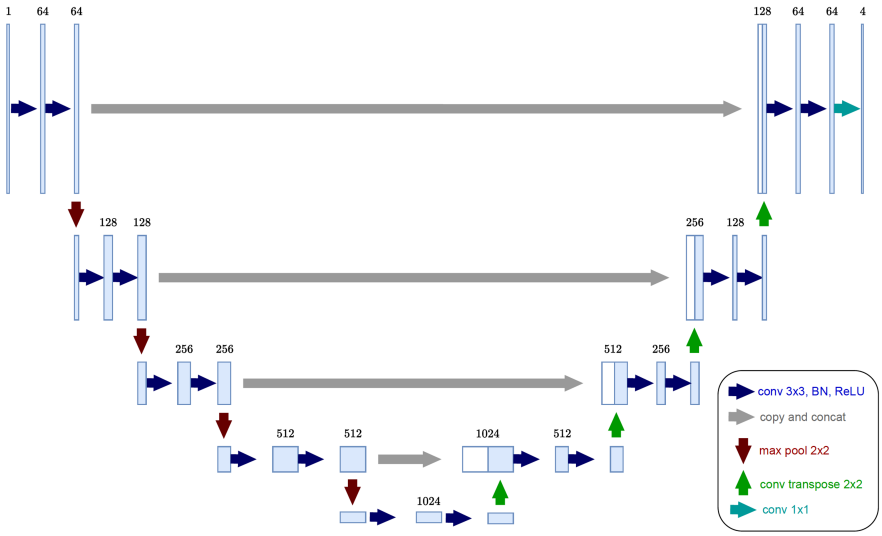


Figure 4.16.: TransUNet Model Architecture [3]

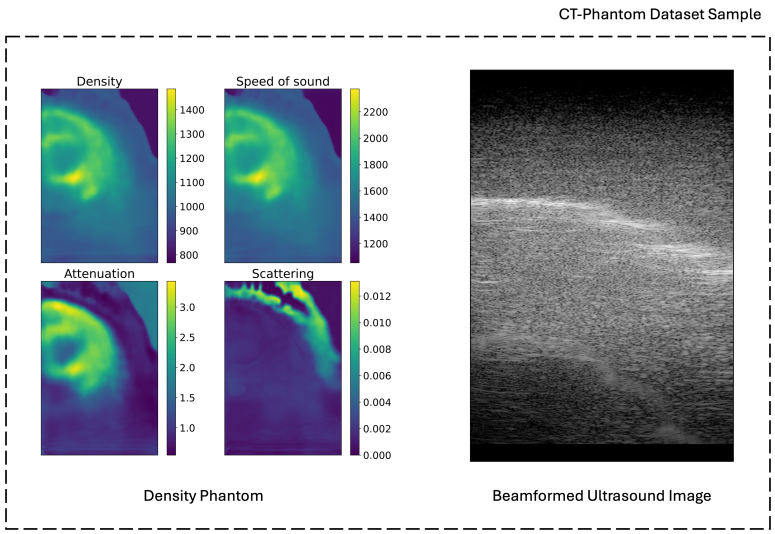


Figure 4.17.: Single Example Sample from a CT-Phantom Dataset

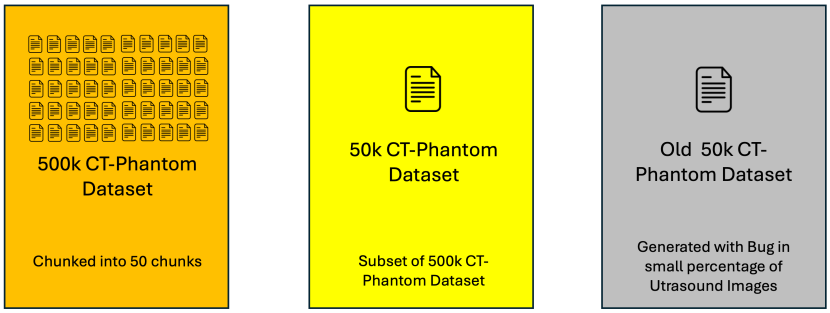


Figure 4.18.: Overview of the CT-Phantom Datasets

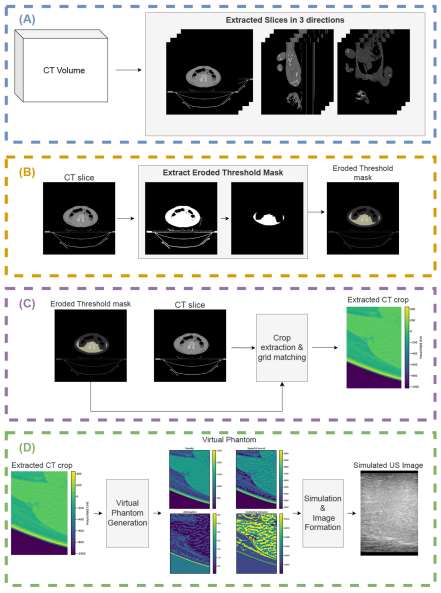


Figure 4.19.: Generation of a CT-Phantom Dataset Sample [3]

Figure 4.20.: Overview of Chunkloader Functionality

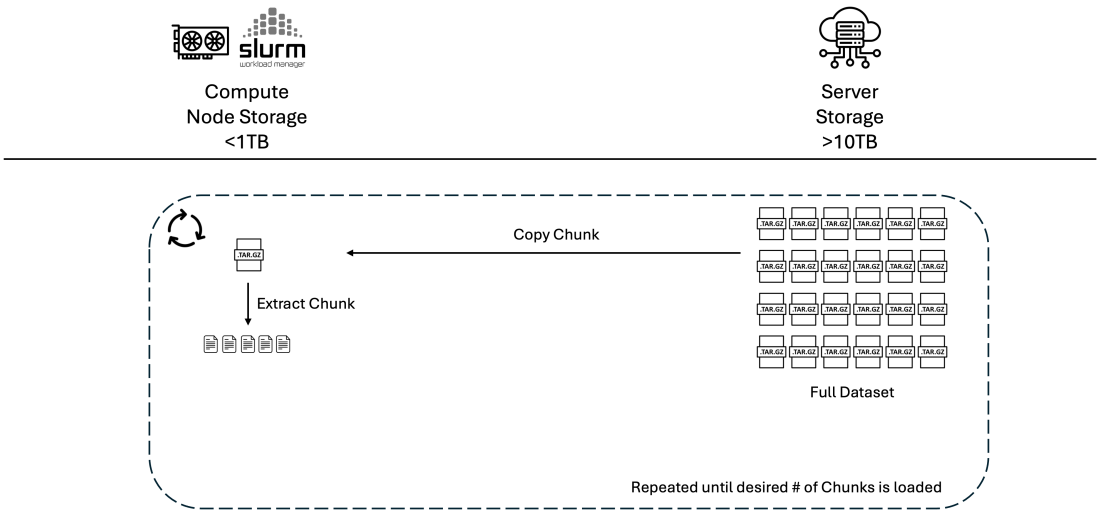


Figure 4.21.: Overview of Chunkloader Storage Usage

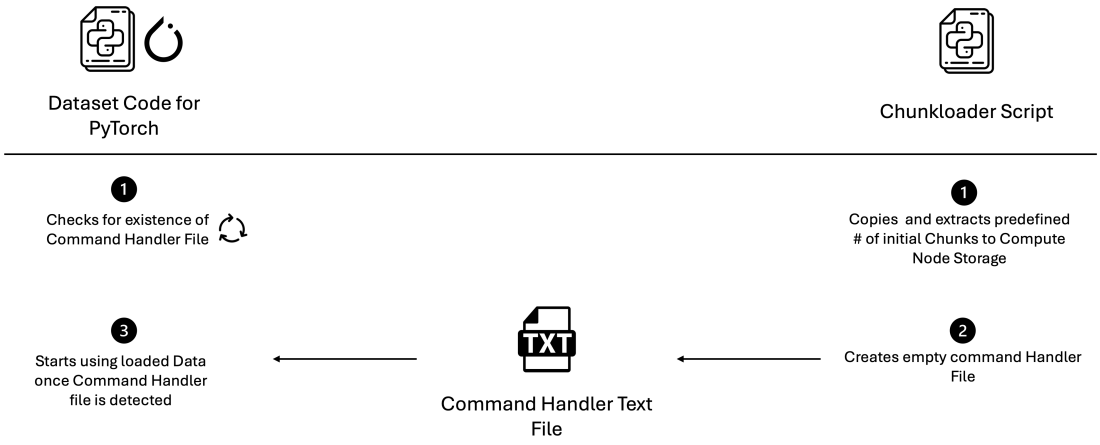


Figure 4.22.: Initialization of Chunkloader

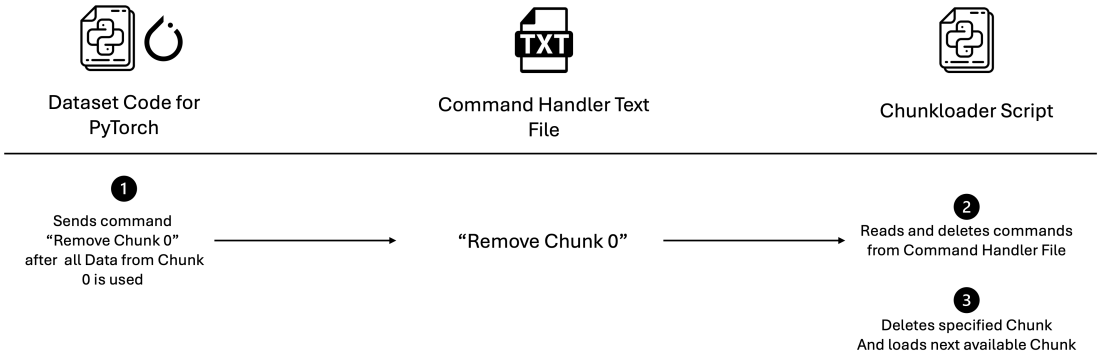


Figure 4.23.: Remove Command Chunkloader

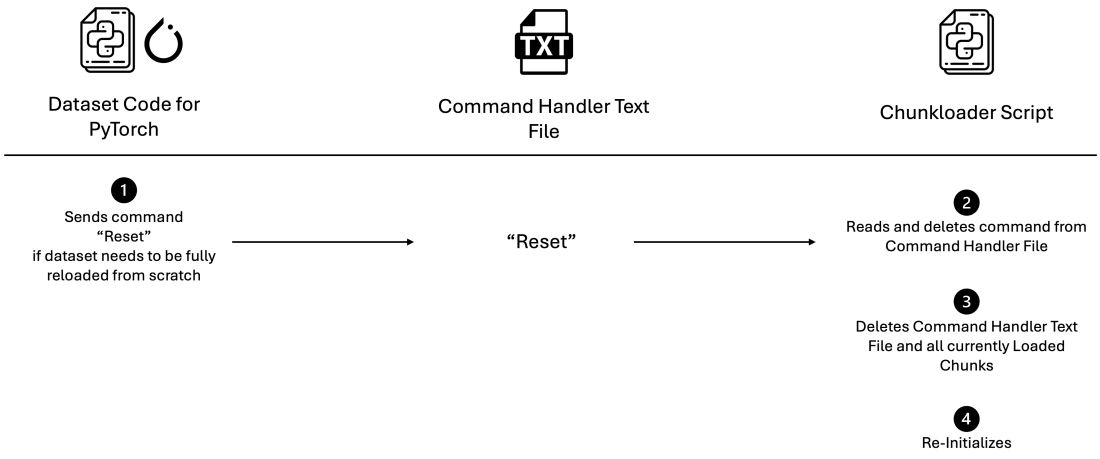


Figure 4.24.: Reset Command Chunkloader

Chapter 5

Experiments

This chapter presents a comprehensive overview of the experimental setups, evaluation datasets, metrics, and configurations employed in this thesis. Additionally, it details the experiments that were conducted.

5.1. Evaluation Datasets

The evaluation process involved using a combination of a primary synthetic test set and several real-world external evaluation datasets. While the external datasets were employed exclusively to assess the top-performing models within each group, ablation studies were conducted solely on the synthetic test set.

5.1.1. Synthetic Evaluation Dataset

The synthetic test set used for evaluation was the test split of the 50k CT-Phantom Dataset. This subset of the larger 500k CT-Phantom Dataset was selected for its smaller size, which simplified the evaluation process by eliminating the need for chunk-loading. Notably, the test split of the 50k dataset was derived after the train-test split of the 500k dataset, ensuring there was no overlap between the training data and the evaluation test set.

Due to time constraints, the evaluation was limited to the first 100 samples from this test set.

5.1.2. External Evaluation Datasets

In addition to the synthetic dataset, real-world US-RF data pairings were used for external evaluation. These datasets, described in more detail in [3], were excluded from training to assess the models' generalization capabilities on diverse real-world data, particularly in vivo tissue samples.

The datasets included data from virtual phantoms generated by different simulation methods, real physical phantoms, and in vivo tissue samples. This comprehensive evaluation ensured robust benchmarking of the models' performance in clinically relevant scenarios.

Overview of External Datasets

Verasonics System Recordings This dataset was acquired using the Verasonics Vantage 256 system equipped with a GE 9L-D transducer. It comprises recordings from two distinct sources: (1) the *CIRS GP Phantom Model*, which includes data from a physical phantom designed to replicate general ultrasound imaging conditions, and (2) *In Vivo MG Muscle Tissue*, which involves recordings of muscle tissue from the lower leg.

Plane-Wave Imaging Challenge in Medical Ultrasound (PICMUS) 2016 This dataset was specifically designed to evaluate beamforming strategies using plane-wave RF data. It includes three types of data: (1) *Physical Phantoms*, which provide real-world phantom recordings, (2) *Virtual Phantoms*, which simulate ultrasound imaging using the Field II simulator, and (3) *In Vivo Carotid Tissue*, capturing acquisitions of the carotid artery in vivo.

Challenge on Ultrasound Beamforming with Deep Learning (CUBDL) 2020 This dataset focuses on RF data from both phantom and in vivo acquisitions. It features 500 single plane-wave recordings of brachioradialis muscle tissue, collected from six distinct samples to ensure variability. These recordings were also acquired using the Verasonics Vantage 256 system, this time with the L10-5 transducer.

Evaluation Strategy

For consistency, only single plane-wave acquisitions with a zero-degree incident angle were used. A single beamforming method ensured compatibility between the original and predicted images, avoiding artifacts caused by variations in beamforming techniques.

Categorization of External Test Sets for Experiments

The external datasets were categorized as follows in the results section:

- **CUBDL In Vivo**: Brachioradialis muscle data from the CUBDL challenge.
- **PICMUS In Vivo**: In vivo carotid artery data.
- **PICMUS Phantom**: Physical and virtual phantom data from the PICMUS challenge.
- **Verasonics CIRS**: Data from the CIRS GP phantom model.
- **Verasonics Tissue**: In vivo muscle tissue data (MG muscle).

Generalization Considerations

Datasets such as PICMUS and CUBDL employed transducer and acquisition settings different from those used for training. Results on these datasets illustrate the models' ability to generalize across diverse imaging conditions. Simulation grids and parameters were heuristically defined to align with dataset specifications.

This external evaluation provided critical insights into the models' robustness and real-world applicability.

5.2. Evaluation Metrics

The evaluation metrics were mostly adopted from [3]. They are tailored to assess RF data predictions, reconstructed US images, and virtual phantom maps where applicable.

5.2.1. RF Metrics

RF data predictions were evaluated using normalized, Time Gain Compensation (TGC)-compensated signals. The normalized signal $\bar{R}$ is computed as:

$$\bar{R}[t, m] = \frac{R[t, m] - \mu}{\sigma}$$

where:

- $R[t, m]$: The raw RF signal at time step t and channel m.
- μ: The mean of $R[t, m]$ across all t and m.
- σ: The standard deviation of $R[t, m]$ across all t and m.

The following metrics were used:

- **Time Signal Root Mean Square Error (RMSE)**:

$$\text{Time Signal RMSE}(R, \hat{R}) = \sqrt{\frac{1}{MT} \sum_{m=1}^{M} \sum_{t=1}^{T} \Big(R[t, m] - \hat{R}[t, m] \Big)^2}$$

  where:

  - $\hat{R}[t, m]$: The predicted RF signal at time t and channel m.
  - T: The total number of time steps.
  - M: The total number of channels.

- **Time Signal Mean Absolute Error (MAE)**:

$$\text{Time Signal MAE}(R, \hat{R}) = \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=1}^{T} |R[t, m] - \hat{R}[t, m]|$$

- **Envelope RMSE**:

$$\text{Envelope RMSE}(R, \hat{R}) = \sqrt{\frac{1}{MT} \sum_{m=1}^{M} \sum_{t=1}^{T} \Big(E[t, m] - \hat{E}[t, m] \Big)^2}, \quad E[t, m] = |H(R)[t, m]|$$

  where:

  - $E[t, m]$: The envelope of the signal $R[t, m]$, derived using the Hilbert transform H.
  - $\hat{E}[t, m]$: The predicted envelope of the signal.

- **Signal-to-Noise Ratio (SNR) Predicted Average**:

$$\text{SNR}_{\text{Predicted Avg}} = \frac{1}{M} \sum_{m=1}^{M} \frac{\mu_m}{\sigma_m}, \quad \mu_m = \text{mean}(R[m]),\ \sigma_m = \text{std}(R[m])$$

  where:

  - μ_m: The mean of the signal for channel m.
  - σ_m: The standard deviation of the signal for channel m.

5.2.2. Image Metrics

Reconstructed US images were evaluated using the following metrics:

- **Image RMSE**:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(s_j - \hat{s}_j)^2}$$

  where:

  - s_j: The true pixel intensity at position j.
  - $\hat{s}_j$: The predicted pixel intensity at position j.
  - N: The total number of pixels in the image.

- **Structural Similarity Index Measure (SSIM)**:

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

  where:

  - $l(x, y)$: Luminance similarity between images x and y.
  - $c(x, y)$: Contrast similarity between images x and y.
  - $s(x, y)$: Structural similarity between images x and y.
  - α, β, γ: Weighting parameters for the components.

- **Mean Intensity Difference** (ΔI):

$$\Delta I = \frac{|I_t - I_p|}{I_t}, \quad I = \frac{1}{N}\sum_{j=1}^{N} s_j$$

  where:

  - I_t: The mean intensity of the true image.
  - I_p: The mean intensity of the predicted image.

- **Perceptual Loss**:

$$\text{Perceptual Loss} = \frac{1}{K}\sum_{k=1}^{K} \|\phi_k(s) - \phi_k(\hat{s})\|^2$$

  where:

  - $\phi_k(s)$: Feature maps extracted from layer k of a pre-trained Neural Network (NN) for the true image.

- $\phi_k(\hat{s})$: Feature maps extracted from layer k of a pre-trained NN for the predicted image.
- K: The total number of layers in the NN.

5.2.3. Patchwise Image Statistics

Patchwise statistics were used for external evaluation sets to provide more granular insights into performance. These include:

- **Average KL Divergence**:

$$\text{Kullback–Leibler (KL)}(h_p||h_t) = \sum_{l=1}^{D} h_p[l] \log\left(\frac{h_p[l]}{h_t[l]}\right)$$

  where:
  - h_p: Predicted intensity histogram.
  - h_t: True intensity histogram.
  - D: The total number of histogram bins.

- **Average Median Difference**:

$$\text{Average Median Diff} = \frac{1}{P}\sum_{p=1}^{P} |\text{Median}_p(s) - \text{Median}_p(\hat{s})|$$

  where:
  - P: The total number of patches.
  - $\text{Median}_p(s)$: Median intensity within patch p for the true image.
  - $\text{Median}_p(\hat{s})$: Median intensity within patch p for the predicted image.

- **Average Standard Deviation Absolute Difference**:

$$\text{Average Std Dev Abs Diff} = \frac{1}{P}\sum_{p=1}^{P} |\sigma_p(s) - \sigma_p(\hat{s})|$$

  where:
  - $\sigma_p(s)$: Standard deviation within patch p for the true image.
  - $\sigma_p(\hat{s})$: Standard deviation within patch p for the predicted image.

5.2.4. Phantom Metrics

For synthetic datasets with ground truth phantoms, pixel-wise RMSE was calculated for each acoustic property map to assess the models' performance on individual tissue properties. The evaluation was conducted for four key metrics:

- **Density RMSE ($\mathbf{RMSE_{Density}}$)**:

$$\mathrm{RMSE}_{\mathrm{Density}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{Density}_{\mathrm{True},i} - \mathrm{Density}_{\mathrm{Pred},i}\right)^2}$$

- **Speed of Sound RMSE ($\mathbf{RMSE_{SoS}}$)**:

$$\mathrm{RMSE}_{\mathrm{SoS}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{SoS}_{\mathrm{True},i} - \mathrm{SoS}_{\mathrm{Pred},i}\right)^2}$$

- **Attenuation RMSE ($\mathbf{RMSE_{Attenuation}}$)**:

$$\mathrm{RMSE}_{\mathrm{Attenuation}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{Attenuation}_{\mathrm{True},i} - \mathrm{Attenuation}_{\mathrm{Pred},i}\right)^2}$$

- **Scatter Variance RMSE ($\mathbf{RMSE_{ScatterVar}}$)**:

$$\mathrm{RMSE}_{\mathrm{ScatterVar}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{ScatterVar}_{\mathrm{True},i} - \mathrm{ScatterVar}_{\mathrm{Pred},i}\right)^2}$$

Where:

- N is the total number of pixels in the phantom map.
- True and Pred represent the ground truth and predicted values, respectively, for each metric.

These metrics provide a comprehensive understanding of model performance across different tissue properties.

5.3. Adapted Marigold Experiments

The Adapted Marigold models were trained on the 500k CT-Phantom dataset for a full training run. Each full training run required 14 days on a single NVIDIA RTX 4090 GPU, with models trained for approximately 178,000 steps using a batch size of 32 (achieved through gradient accumulation), corresponding to over 11 epochs. The trained models were then integrated into the complete USDatarecycler pipeline for evaluation.

5.3.1. Comparison of VAE Pre-training Architectures

Three VAE pre-training configurations were evaluated:

- Full VAE fine-tuning: All Autoencoder components were unfrozen.
- Decoder-only fine-tuning: Only the decoder was unfrozen.
- Last-layer decoder fine-tuning: Only the last decoder layer was unfrozen.

These configurations were compared against a baseline with no fine-tuning. Standard Diffusion U-Net training followed each VAE pre-training approach.

5.3.2. Impact of Multi-VAE Architecture

In the Multi-VAE setup, each phantom channel was assigned a dedicated U-Net and VAE. This setup was compared against a configuration employing a standard VAE fine-tuning using a single VAE to en- and decode all Phantom channels.

5.3.3. Impact of Learning Rate Schedule Lengths

The learning rate in the original Marigold model training follows an exponential decay schedule, starting with a brief warm-up phase. The schedule's primary impact is defined by the duration of the decay from the maximum to the minimum value. Since the minimum value is typically very small, minimal further learning occurs after reaching it. In this work, the duration of the decay, also referred to as the learning rate schedule length, was varied to assess its effect on the training process.

Three learning rate schedule lengths were tested:

- 25,000 steps: Original Marigold schedule.
- 50,000 steps: Intermediate decay length.
- 100,000 steps: Longest, main decay length.

5.3.4. Impact of Input Augmentation

Adapted Marigold models using the input augmentation techniques from [3] were compared against models trained without augmentation.

5.4. TransUNet Experiments

All experiments with TransUNet employed the architecture and hyperparameters defined in [3]. Training was conducted on two NVIDIA RTX 4090 GPUs over 14 days. Models trained on the 50k datasets were run for 95 epochs, while those trained on the 500k dataset required 19 epochs.

5.4.1. Comparison Between Models Trained on the Old and New 50k Phantom Dataset

A TransUNet trained on the old 50k CT-Phantom dataset (produced with a beamforming bug) was compared to a model trained on the corrected 50k CT-Phantom dataset.

5.4.2. Comparison Between Models Trained on the 50k and 500k Phantom Dataset

A TransUNet trained on the 50k CT-Phantom dataset was compared to one trained on the 500k CT-Phantom dataset.

5.4.3. Overview of TransUNet Model Performances

This section summarizes and compares the performance of all trained TransUNet models across the aforementioned experiments.

Chapter 6

Results

This chapter presents the results from the experiments described in the previous chapter. It begins with the outcomes of the Adapted Marigold models and their ablations, including experiments with different VAE pretraining structures, a Multi-VAE setup, different learning rate schedule lengths, and the impact of using input image augmentation. The chapter is ended by the results of the existing TransUNet models trained on various training datasets.

6.1. Adapted Marigold Results

This section covers the experiments and ablations performed on the Adapted Marigold models, showcasing results for different VAE pretraining setups, the Multi-VAE approach, variations in learning rate schedules, and input augmentation.

6.1.1. Comparison of VAE Pre-training Architectures

This subsection compares the results of Adapted Marigold models trained with different VAE pre-training setups, excluding the Multi-VAE architecture, which is covered separately.

As shown in Table 6.1, fine-tuning the VAE generally improves performance. The full VAE fine-tuning and decoder-only fine-tuning yield the best results overall. Specifically, the decoder-only fine-tuning achieves superior performance in the phantom metrics, while the full VAE fine-tuning excels in both RF and image metrics. Therefore VAE-finetuning was employed in the main Adapted Marigold model.

Table 6.1.: Marigold Results for Different VAE Pretraining Setups

Metric	No VAE	Full VAE	Dec. Only	Last Layer of Dec.
RF Metrics				
Time Signal RMSE ↓	1.403	1.416	1.414	**1.402**
Time Signal MAE ↓	0.535	**0.519**	0.524	0.531
Envelope RMSE ↓	1.062	**1.035**	1.045	1.062
SNR Predicted Avg ↑	24.664	26.687	**26.722**	24.846
Image Metrics				
Image RMSE ↓	44.894	**44.754**	44.289	45.392
SSIM ↑	0.133	0.135	0.134	**0.137**
Mean Intensity Difference ↓	44.894	44.754	**44.289**	45.392
Perceptual Loss ↓	0.076	**0.074**	**0.074**	0.077
Phantom Metrics				
Density RMSE ↓	**119.234**	120.334	119.671	120.499
SOS RMSE ↓	911.219	895.548	**892.341**	913.419
Attenuation RMSE ↓	0.625	0.617	**0.602**	0.658
Scatter Var RMSE ↓	615.63	553.693	**524.454**	635.588

6.1.2. Impact of Multi-VAE Architecture

This subsection examines the results of the Multi-VAE architecture compared to the baseline model.

As shown in Table 6.2, the Multi-VAE approach improves performance in the RF metrics but exhibits a trade-off with reduced performance in the Image and Phantom metrics. Given the relatively small performance differences and the significantly slower inference time of the Multi-VAE approach, it was not adopted for the main Marigold model.

6.1.3. Impact of Learning Rate Schedule Lengths

This subsection evaluates Adapted Marigold models trained with different learning rate schedule lengths.

As shown in Table 6.3, the 100k step learning rate schedule generally performs best, particularly in phantom metrics. The longer schedule allows the model to focus on learning coarser patterns, which seems beneficial for phantom tissue property predictions. However, for Ultrasound image reconstruction, the 50k schedule achieves the best results, as a solid middle ground. For the main Adapted Marigold model, the 100k schedule was ultimately chosen in order to predict phantoms as well as possible.

6.1.4. Impact of Input Augmentation

This subsection explores the effect of input augmentation on model performance.

Table 6.2.: Marigold Results with and without Multi-VAE

Metric	With Multi-VAE	Without Multi-VAE
RF Metrics		
Time Signal RMSE ↓	1.414	**1.402**
Time Signal MAE ↓	**0.522**	0.531
Envelope RMSE ↓	**1.057**	1.062
SNR Predicted Avg ↑	**26.469**	24.846
Image Metrics		
Image RMSE ↓	44.743	**38.475**
SSIM ↑	0.135	**0.137**
Mean Intensity Difference ↓	0.149	**0.137**
Perceptual Loss ↓	0.075	**0.071**
Phantom Metrics		
Density RMSE ↓	120.499	**119.234**
SOS RMSE ↓	119.671	**100.638**
Attenuation RMSE ↓	**0.611**	0.625
Scatter Var RMSE ↓	73.543	**68.477**

As seen in Table 6.4, input augmentation consistently improves performance across nearly all metrics. This is not surprising, as input augmentation is a standard practice in modern DL models. Therefore it was also applied to the main Adapted Marigold model.

6.1.5. Summary

In summary, the Adapted Marigold models demonstrate robust performance across various setups. Fine-tuning the VAE and incorporating input augmentation consistently enhance results. Using a Multi-VAE-approach improves performance on RF-prediction but as a trade-off worsens US-Image and Phantom predictions. Learning rate schedule lengths influence results differently for Ultrasound image reconstruction and phantom property prediction, with longer schedules generally yielding the best overall performance. Overall all the main adaptations to the original Marigold model training yield some benefits for the US to RF task.

6.2. TransUNet Results

This section presents the results for the TransUNet model across multiple training datasets. Comparisons are made between models trained on different versions of the Phantom Dataset, with an overview provided at the end.

Table 6.3.: Marigold Results for Different Learning Rate Schedule Lengths

Metric	25k LR Sched	50k LR Sched	100k LR Sched
RF Metrics			
Time Signal RMSE ↓	1.403	1.406	**1.402**
Time Signal MAE ↓	0.536	0.533	**0.531**
Envelope RMSE ↓	1.064	**1.060**	1.062
SNR Predicted Avg ↑	**25.812**	25.490	24.846
Image Metrics			
Image RMSE ↓	46.706	**45.141**	45.392
SSIM ↑	**0.148**	0.135	0.137
Mean Intensity Difference ↓	46.706	**45.141**	45.392
Perceptual Loss ↓	**0.004**	**0.004**	**0.004**
Phantom Metrics			
Density RMSE ↓	122.626	121.146	**120.499**
SOS RMSE ↓	**903.034**	915.401	913.419
Attenuation RMSE ↓	0.647	0.633	**0.625**
Scatter Var RMSE ↓	44.341	44.458	**45.444**

6.2.1. Comparison Between Models Trained on the Old 50k and New 50k Phantom Dataset

This experiment compares the performance of the TransUNet model trained on the Old 50k Phantom Dataset and the (New) 50k Phantom Dataset. Both models were trained under identical conditions, with the primary difference being the corrected beamforming process and updated sample collection in the newer dataset.

The results of this comparison, shown in Table 6.5, show that the (New) 50k Phantom Dataset provides clear improvements in most areas. While RF metrics remain largely consistent between the two datasets, notable gains are observed in image quality and phantom property predictions. Structural similarity, intensity differences, and other image-related metrics demonstrate better performance, reflecting the removal of corrupted samples in the newer dataset. Similarly, the model trained on the 50k dataset achieves more accurate predictions for tissue properties such as density, speed of sound, and attenuation. These improvements highlight the impact of correcting errors in virtual datasets, as their incorrectness can lead to model performance degradations as shown.

6.2.2. Comparison Between Models Trained on the 50k and 500k Phantom Dataset

This experiment examines the effects of dataset scale by comparing the performance of models trained on the 50k Phantom Dataset and the 500k Phantom Dataset. The 500k

Table 6.4.: Marigold Results with and without Input Augmentation

Metric	With Augmentation	Without Augmentation
RF Metrics		
Time Signal RMSE ↓	1.406	**1.402**
Time Signal MAE ↓	**0.530**	0.531
Envelope RMSE ↓	1.063	**1.062**
SNR Predicted Avg ↑	**25.652**	24.846
Image Metrics		
Image RMSE ↓	38.778	**38.475**
SSIM ↑	0.134	**0.137**
Mean Intensity Difference ↓	**0.099**	0.104
Perceptual Loss ↓	**0.076**	0.077
Phantom Metrics		
Density RMSE ↓	**117.017**	120.499
SOS RMSE ↓	**39.022**	45.444
Attenuation RMSE ↓	**0.625**	0.627
Scatter Var RMSE ↓	**71.219**	73.543

dataset includes additional samples that provide greater diversity and coverage. Table 6.6 provides a summary of the results.

The results demonstrate consistent improvements across all metrics when the model is trained on the larger dataset. RF metrics show better accuracy and higher signal-to-noise ratios, indicating improved generalization to unseen RF data. Similarly, image-related metrics such as structural similarity and intensity differences benefit from the increased diversity in the training data. The most significant gains are observed in phantom property predictions, with the larger dataset enabling more accurate estimations of density, speed of sound, attenuation and scattering. These improvements emphasize the value of a larger and more diverse dataset for training.

6.2.3. Overview of TransUNet Model Performances

This subsection provides an overview of the TransUNet model's performance across all datasets: the Old 50k Phantom Dataset, the 50k Phantom Dataset, and the 500k Phantom Dataset. The datasets differ in quality and size, with the Old 50k containing corrupted samples, the 50k correcting these issues, and the 500k significantly increasing the sample set.

The progression from the Old 50k dataset to the 50k dataset shows marked improvements in image and phantom metrics, highlighting the benefits of correcting beamforming errors and using higher-quality samples. The transition to the 500k dataset demonstrates the importance of dataset size, with substantial gains observed across all metrics. These

Table 6.5.: Comparison of metrics between the Old 50k CT-Phantom Dataset and the 50k Phantom Dataset.

Metric	Old 50k Phantom Dataset	50k Phantom Dataset
RF Metrics		
Time Signal RMSE ↓	**1.404**	1.411
Time Signal MAE ↓	**0.510**	0.517
Envelope RMSE ↓	**1.033**	1.035
SNR Predicted Avg ↑	**26.377**	25.285
Image Metrics		
Image RMSE ↓	41.683	**41.383**
SSIM ↑	0.135	**0.136**
Mean Intensity Difference ↓	0.102	**0.093**
Perceptual Loss ↓	**0.079**	**0.079**
Phantom Metrics		
Density RMSE ↓	104.550	**97.344**
SOS RMSE ↓	189.723	**178.828**
Attenuation RMSE ↓	0.538	**0.517**
Scatter Var RMSE ↓	60.784	**60.049**

results emphasize the critical role of both dataset quality and scale in achieving optimal model performance.

Table 6.7 shows the exact result metrics of all TransUNet models again in an overview.

6.3. Comparison between Adapted Marigold and TransUNet

This section presents a comparison between the best-performing TransUNet model, trained on the 500k dataset, and the Adapted Marigold model. The evaluation is conducted on two fronts: the standard synthetic test set and a range of external test sets. Additionally, visual comparisons are provided to complement the quantitative analysis.

6.3.1. Standard Synthetic Test Set Results

The results, summarized in Table 6.8, reveal that the TransUNet model achieves better performance on the majority of metrics for the synthetic test set. However, visual inspections, shown for example in Figure 6.1 and Figure 6.2, indicate notable differences. While TransUNet captures certain phantom structures effectively, its predictions exhibit over-smoothing, leading to a lack of detail compared to the Adapted Marigold model.

Table 6.6.: Comparison of metrics between the 50k Phantom Dataset and the 500k Phantom Dataset.

Metric	50k Phantom Dataset	500k Phantom Dataset
RF Metrics		
Time Signal RMSE ↓	1.411	**1.386**
Time Signal MAE ↓	0.517	**0.497**
Envelope RMSE ↓	1.035	**0.997**
SNR Predicted Avg ↑	25.285	**26.696**
Image Metrics		
Image RMSE ↓	41.383	**41.103**
SSIM ↑	0.136	**0.136**
Mean Intensity Difference ↓	0.093	**0.085**
Perceptual Loss ↓	**0.079**	0.080
Phantom Metrics		
Density RMSE ↓	97.344	**91.713**
SOS RMSE ↓	178.828	**169.111**
Attenuation RMSE ↓	0.517	**0.516**
Scatter Var RMSE ↓	60.049	**52.238**

For instance, Figure 6.3 provides a comprehensive view of the reconstructed phantoms and RF data predictions. The metrics-driven advantage of the TransUNet does not always translate into superior visual quality, as the Adapted Marigold model often retains more structural fidelity.

6.3.2. External Test Set Results

To evaluate generalization capabilities, the models were tested on a variety of external datasets, including recordings from the Verasonics system, data from the PICMUS and CUBDL challenges, and in vivo samples. These datasets were specifically chosen for their diversity and alignment with real-world imaging scenarios.

CUBDL In Vivo

As shown in Table 6.9, the Adapted Marigold model outperforms the TransUNet in nearly all metrics on the CUBDL dataset. This superiority is further evident in the visual results, as can be seen in Figure 6.4, where the Marigold predictions exhibit significantly more structural detail, whereas the TransUNet outputs are often noisier and less defined.

Table 6.7.: Overview of TransUNet model performance across all datasets.

Metric	Old 50k Dataset	50k Dataset	500k Dataset
RF Metrics			
Time Signal RMSE ↓	1.404	1.411	**1.386**
Time Signal MAE ↓	**0.510**	0.517	0.497
Envelope RMSE ↓	**1.033**	1.035	0.997
SNR Predicted Avg ↑	**26.378**	25.285	26.696
Image Metrics			
Image RMSE ↓	41.683	41.383	**41.103**
SSIM ↑	0.135	0.136	**0.136**
Mean Intensity Difference ↓	0.102	0.093	**0.085**
Perceptual Loss ↓	**0.079**	**0.079**	0.080
Phantom Metrics			
Density RMSE ↓	104.550	97.344	**91.713**
SOS RMSE ↓	189.723	178.828	**169.111**
Attenuation RMSE ↓	0.538	0.518	**0.516**
Scatter Var RMSE ↓	60.784	60.049	**52.238**

PICMUS In Vivo

For the PICMUS In Vivo dataset, the metrics present a mixed outcome, with no model emerging as a clear winner (Table 6.10). However, visual comparisons (Figure 6.5) indicate that the Adapted Marigold model more accurately reproduces underlying tissue structures compared to the overly smoothed predictions of TransUNet.

PICMUS Phantom

The results on the PICMUS Phantom dataset, summarized in Table 6.11, clearly favor the Adapted Marigold model across most metrics. Figure 6.6 further supports this finding, with Marigold's predictions closely resembling the ground truth in both ultrasound images and RF data.

Verasonics CIRS

The Verasonics CIRS dataset highlights the Adapted Marigold model's superiority in most metrics, as shown in Table 6.12. The visual results in Figure 6.7 further demonstrate the reduced randomness and enhanced structural accuracy of Marigold's predictions compared to TransUNet.

Table 6.8.: Comparison of metrics between Marigold Main and TransUNet 500k on Standard Synthetic Test Set

Metric	Marigold	TransUNet
RF Metrics		
Time Signal RMSE ↓	1.402	**1.386**
Time Signal MAE ↓	0.531	**0.497**
Envelope RMSE ↓	1.062	**0.997**
SNR Predicted Avg ↑	24.846	**26.696**
Image Metrics		
Image RMSE ↓	**38.475**	41.103
SSIM ↑	**0.137**	0.136
Mean Intensity Difference ↓	0.104	**0.081**
Perceptual Loss ↓	0.077	**0.08**
Phantom Metrics		
Density RMSE ↓	120.499	**91.713**
SOS RMSE ↓	913.419	**855.668**
Attenuation RMSE ↓	0.625	**0.516**
Scatter Var RMSE ↓	73.543	**52.238**

Verasonics Tissue

The Verasonics Tissue dataset results, shown in Table 6.13, heavily favor the Adapted Marigold model. This advantage is further supported by the visual results in Figure 6.8, where the Marigold predictions provide more accurate structural representations than TransUNet.

6.3.3. Discussion

The comparison between the TransUNet and Adapted Marigold models reveals complementary strengths. While TransUNet exhibits strong performance on synthetic datasets, its tendency to over-smooth predictions impacts visual quality. In contrast, the Adapted Marigold model excels in generalizing to external datasets, consistently producing visually superior results and demonstrating robustness across diverse imaging conditions. These findings underscore the importance of balancing metrics-driven optimization with visual inspection. An overview over the results of the Datasets visualized on two key metrics, the /glsus-Image RMSE and the RF Envelope RMSE can be seen in Figure 6.9. From this the, on average, better performance of the Adapted Marigold Model can be seen quite clearly.

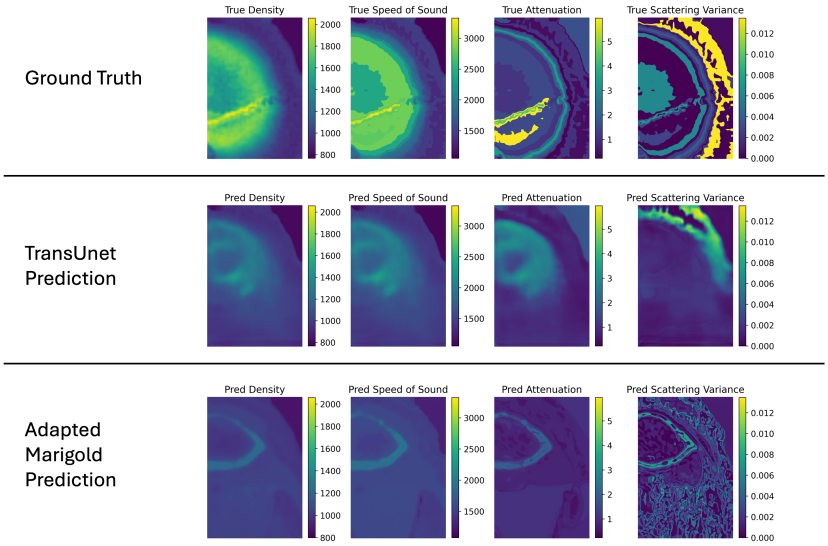


Figure 6.1.: Example Visual Phantom Prediction Results Comparison on Synthetic Test Set

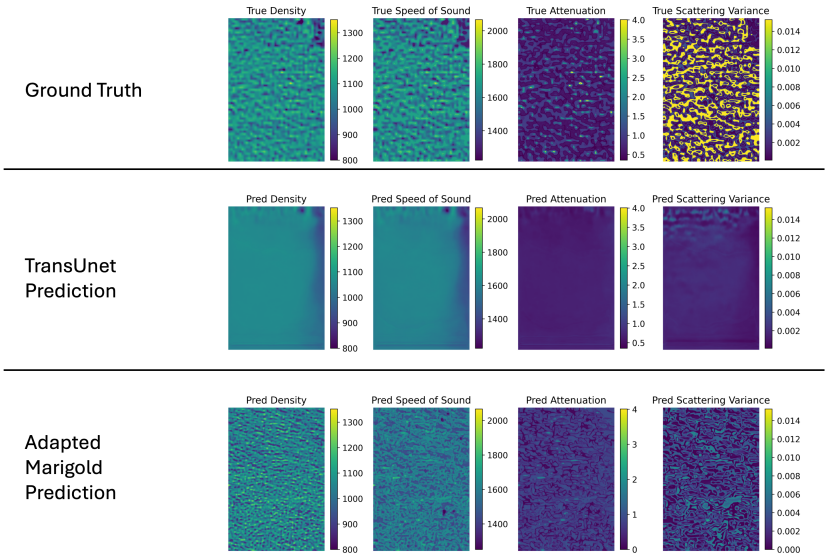


Figure 6.2.: Secondary Example Phantom Prediction Results Comparison on Synthetic Test Set

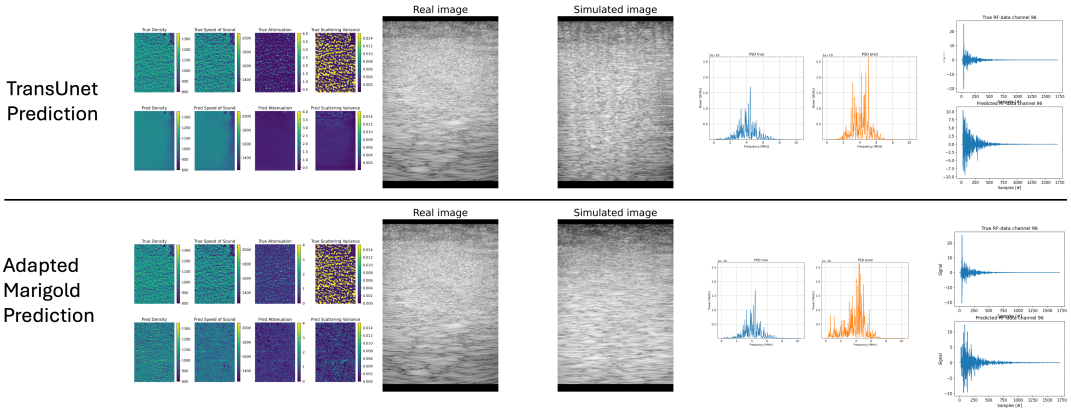


Figure 6.3.: Full Example Visual Results Comparison on Synthetic Test Set

Table 6.9.: Comparison of metrics between Marigold Main and TransUNet 500k on the CUBDL In Vivo External Dataset

Metric	Marigold	TransUNet
RF Metrics		
Time Signal RMSE ↓	1.415	**1.414**
Time Signal MAE ↓	**0.811**	0.883
Envelope RMSE ↓	**1.341**	1.439
SNR Predicted Avg ↑	**9.443**	7.457
Image Metrics		
Image RMSE ↓	**60.156**	79.852
SSIM ↑	**0.141**	0.125
Mean Intensity Difference ↓	**0.579**	0.915
Perceptual Loss ↓	0.214	**0.189**
Patchwise Image Statistics		
Average KL Divergence ↓	**1.237**	2.229
Average Median Diff ↓	**14.692**	21.142
Average Std Dev Abs Diff ↓	**9.850**	13.413

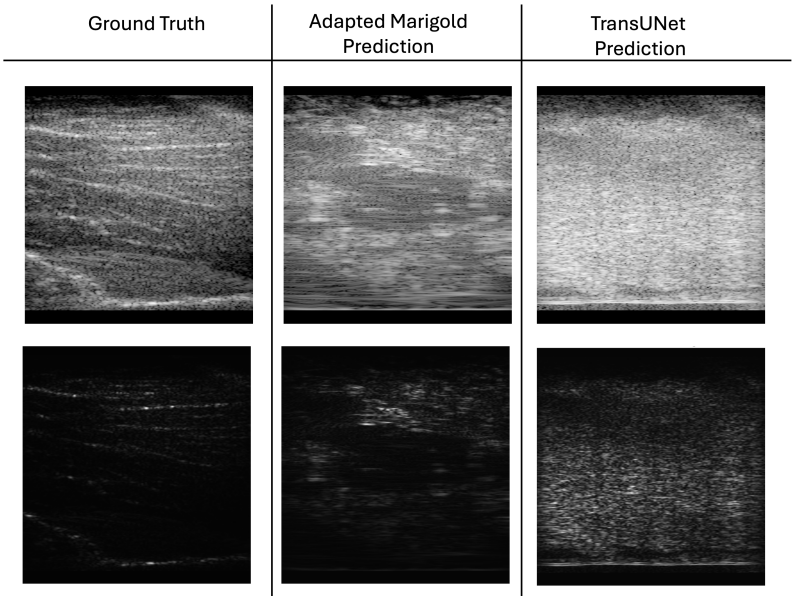


Figure 6.4.: Example US Image Reconstruction Results Comparison on CUBDL In Vivo Test Set

Table 6.10.: Comparison of metrics between Marigold Main and TransUNet 500k on the PICMUS in vivo External Dataset

Metric	**Marigold**	**TransUNet**
RF Metrics		
Time Signal RMSE ↓	**1.414**	1.417
Time Signal MAE ↓	0.778	**0.721**
Envelope RMSE ↓	**1.627**	1.715
SNR Predicted Avg ↑	32.894	**33.784**
Image Metrics		
Image RMSE ↓	65.939	**64.656**
SSIM ↑	**0.424**	0.281
Mean Intensity Difference ↓	0.492	**0.282**
Perceptual Loss ↓	**0.187**	0.214
Patchwise Image Statistics		
Average KL Divergence ↓	**2.439**	4.377
Average Median Diff ↓	**25.143**	34.687
Average Standard Deviation Abs Diff ↓	9.164	**9.088**

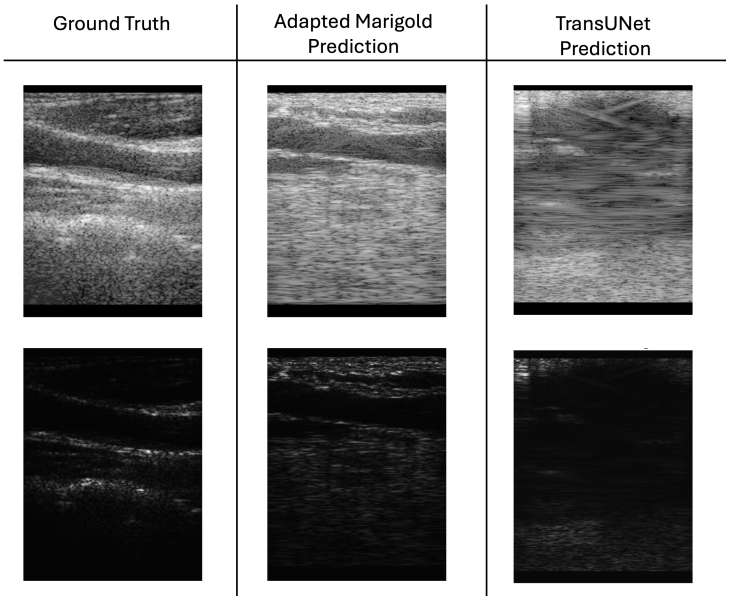


Figure 6.5.: Example US Image Reconstruction Results Comparison on PICMUS In Vivo Test Set

Table 6.11.: Comparison of metrics between Marigold Main and TransUNet 500k on the PICMUS Phantom External Dataset

Metric	**Marigold**	**TransUNet**
RF Metrics		
Time Signal RMSE ↓	1.416	**1.412**
Time Signal MAE ↓	**0.810**	0.910
Envelope RMSE ↓	**1.377**	1.378
Image Metrics		
Image RMSE ↓	**54.500**	67.116
SSIM ↑	0.386	**0.580**
Mean Intensity Difference ↓	**0.566**	0.744
Perceptual Loss ↓	0.218	**0.213**
Patchwise Image Statistics		
Average KL Divergence ↓	**2.145**	3.697
Average Median Difference ↓	**16.318**	26.317
Average Standard Deviation Abs Diff ↓	**9.109**	9.749

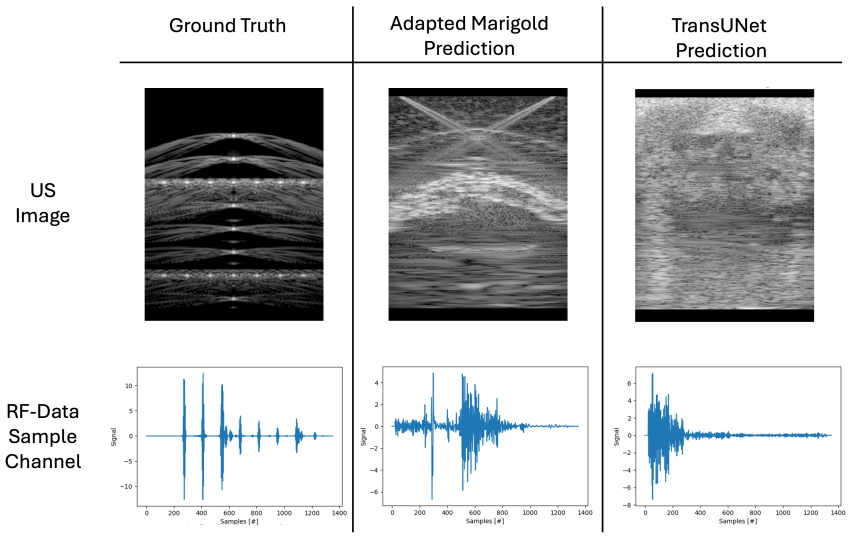


Figure 6.6.: Example US Image Reconstruction Results Comparison on PICMUS Phantom Test Set

Table 6.12.: Comparison of metrics between Marigold Main and TransUNet 500k on the Verasonics CIRS External Dataset

Metric	**Marigold**	**TransUNet**
RF Metrics		
Time Signal RMSE ↓	**1.415**	1.417
Time Signal MAE ↓	**0.971**	0.989
Envelope RMSE ↓	**1.041**	1.067
SNR Predicted Avg ↑	**6.707**	7.023
Image Metrics		
Image RMSE ↓	**71.553**	73.275
SSIM ↑	**0.168**	0.148
Mean Intensity Difference ↓	0.69	**0.683**
Perceptual Loss ↓	0.113	**0.107**
Patchwise Image Statistics		
Average KL Divergence ↓	**0.394**	0.964
Average Median Difference ↓	**8.183**	12.183
Average Standard Deviation Abs Diff ↓	8.172	**8.009**

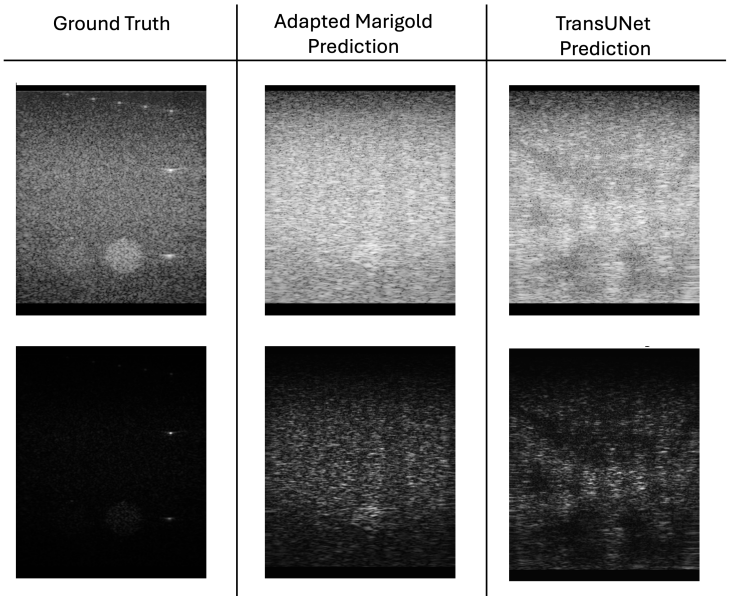


Figure 6.7.: Example US Image Reconstruction Results Comparison on Verasonics CIRS Test Set

Table 6.13.: Comparison of metrics between Marigold Main and TransUNet 500k on the Verasonics Tissue External Dataset

Metric	**Marigold**	**TransUNet**
RF Metrics		
Time Signal RMSE ↓	**1.413**	1.414
Time Signal MAE ↓	**0.862**	0.941
Envelope RMSE ↓	**1.32**	1.369
SNR Predicted Avg ↑	**13.78**	**13.78**
Image Metrics		
Image RMSE ↓	**44.311**	56.017
SSIM ↑	**0.142**	0.135
Mean Intensity Difference ↓	**0.109**	0.277
Perceptual Loss ↓	0.100	**0.078**
Patchwise Image Statistics		
Average KL Divergence ↓	0.125	**0.105**
Average Median Difference ↓	**1.436**	1.548
Average Standard Deviation Abs Diff ↓	**5.063**	8.090

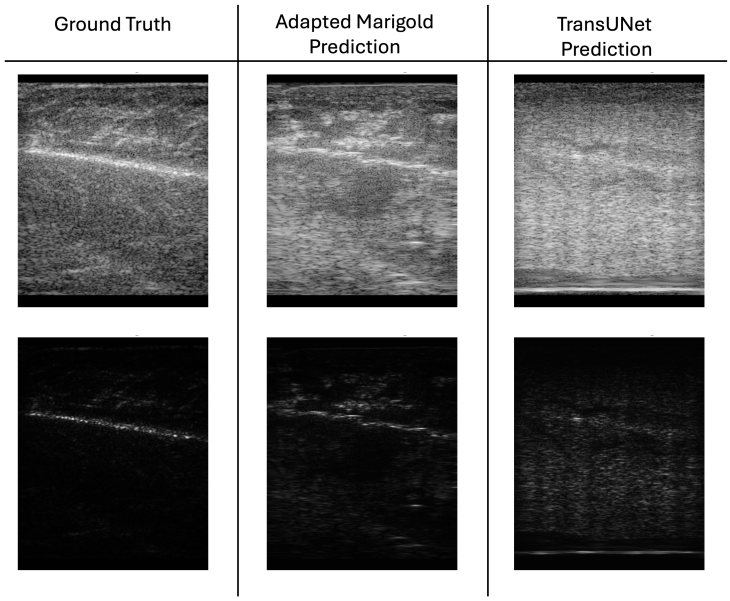


Figure 6.8.: Example US Image Reconstruction Results Comparison on Verasonics Tissue Test Set

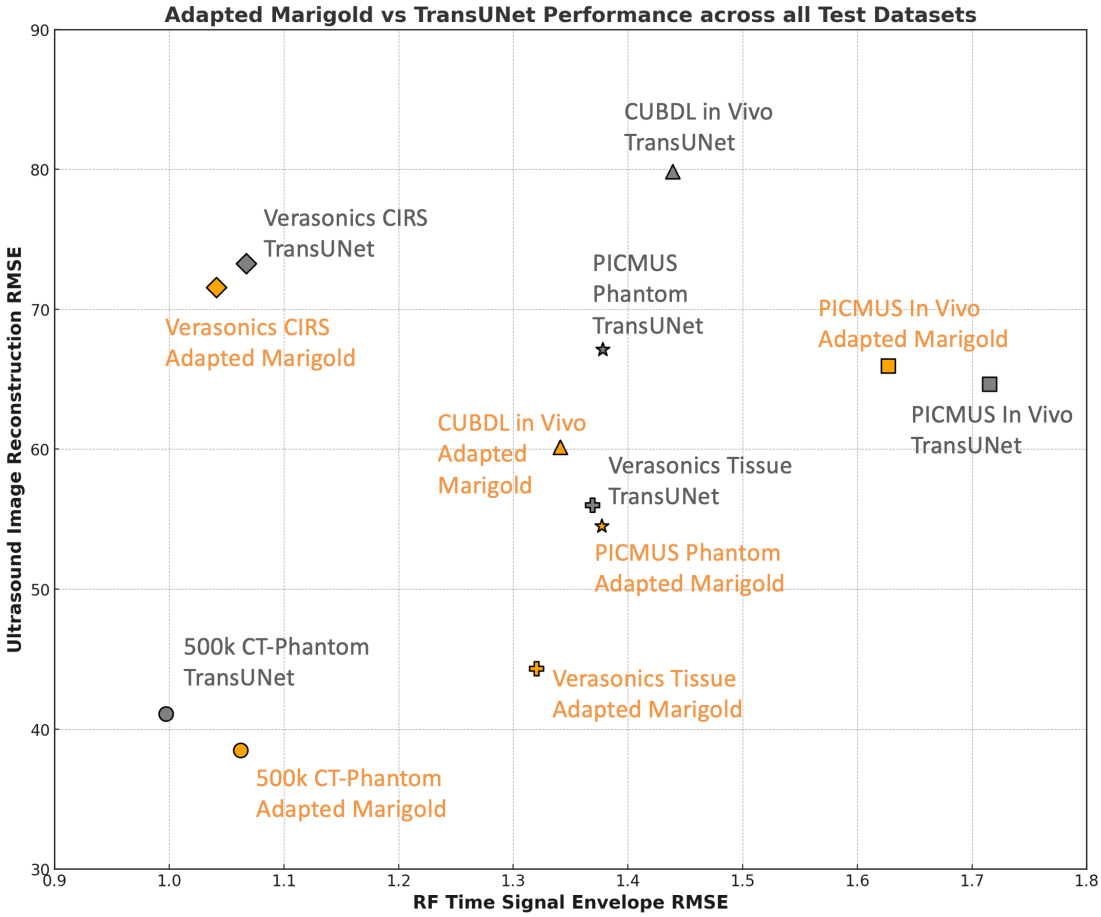


Figure 6.9.: Scatter Plot of Results over all Test Datasets

Chapter 7

Conclusion and Future Work

This thesis has demonstrated that state-of-the-art Denoising Diffusion Models, particularly the developed Adapted Marigold model, offer significant potential for Ultrasound RF data reconstruction. While the existing TransUNet model achieved slightly better quantitative performance on a synthetic test dataset, the Adapted Marigold model consistently outperformed TransUNet across almost all external datasets, demonstrating superior performance in real-world scenarios. This highlights the practical advantages of Denoising Diffusion Models, which excel at capturing the complexity and variability inherent in real-world imaging tasks.

The Adapted Marigold model's performance advantage on external datasets can be attributed to the nature of diffusion models, which iteratively refine predictions through a denoising process. This approach allows the model to focus on reconstructing fine-grained details and preserving structural integrity, making it particularly well-suited for handling the diverse and noisy conditions found in real-world ultrasound imaging. In contrast, the Transformer-based architecture of TransUNet, while effective for synthetic datasets with well-defined patterns, tends to produce over-smoothed results in more complex and variable scenarios, as it relies heavily on attention mechanisms that may struggle to capture the nuanced, localized details necessary for high-quality reconstructions.

This comparison underscores the complementary strengths of the two approaches and emphasizes the importance of considering both synthetic benchmarks and real-world evaluations. The Adapted Marigold model's robustness and ability to generalize across diverse datasets further validate the utility of the USDatarecycler pipeline. By integrating physically-informed simulations with data-driven methodologies, the pipeline effectively bridges the gap between synthetic data and real-world clinical applications, setting a strong foundation for advancing RF data reconstruction.

Future Work

Several avenues for future research and development emerge from this work. One promising direction is to leverage the synthetic RF data generated by the pipeline in real-world applications. Evaluating the performance of models trained on such an artificial dataset in practical scenarios would provide critical insights into their utility and limitations.

Another area of focus could be the optimization of the Adapted Marigold model's inference process. Currently, it lacks full parallelization, which impacts its computational efficiency. Parallelizing parts of the code or leveraging advanced Multi-GPU inference could significantly reduce inference time.

Expanding the diversity of the training datasets is another important direction, particularly by including more varied tissue types and imaging conditions. This would likely improve the model's generalization capabilities even more. Additionally, generating datasets that better mimic real-world variability in ultrasound imaging would further enhance robustness. Exploring hybrid architectures that combine the strengths of Denoising Diffusion Models and transformer-based models like TransUNet could yield improved performance by leveraging the quantitative strengths of TransUNet while retaining the visual quality benefits of diffusion models.

In conclusion, this work has established a strong foundation for using Denoising Diffusion Models in the field of ultrasound RF data reconstruction. By addressing the outlined challenges and exploring the proposed future directions, the potential for broader adoption of these models in medical imaging applications can be significantly enhanced.

Appendix A

Task Description

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. In consultation with the supervisor, one of the following three options must be selected:

I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used and cited generative artificial intelligence technologies[2].

I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used generative artificial intelligence technologies[3]. In consultation with the supervisor, I did not cite them.

Title of paper or thesis:

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s): **First name(s):**

With my signature I confirm the following:
- I have adhered to the rules set out in the Citation Guide.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date **Signature(s)**

C. Knecht

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] E.g. ChatGPT, DALL E 2, Google Bard
[2] E.g. ChatGPT, DALL E 2, Google Bard
[3] E.g. ChatGPT, DALL E 2, Google Bard

Master Thesis at the Department of Information Technology and Electrical Engineering

Summer Semester 2024

Cyrill Knecht

Ultrasound Image Data Recycler

May 15, 2024

Advisors:	Dr. Christoph Leitner, ETZ J65, christoph.leitner@iis.ee.ethz.ch Dr. Yawei Li, ETF J78, yawei.li@vision.ee.ethz.ch
Supervisor:	Prof. Dr. Luca Benini, lbenini@iis.ee.ethz.ch Prof. Dr. Ender Konukoglu, kender@vision.ee.ethz.ch

Handout:	May, 2024
Due:	December, 2024

1. Project Outline

Medical ultrasound (US) imaging is a vital diagnostic tool and has many areas of application. The raw data from US imaging, known as radio-frequency (RF) data, contains more information than US images and has valuable use cases. Although large datasets of processed US images are widely available, raw RF data remains scarce.

In a Master thesis [1] we explored the first system designed to predict raw RF data from US images. We have developed a data-driven, physically-informed model that combines deep learning (DL) techniques with numerical simulations of ultrasound waves. Specifically, we employed a TransUNet neural network architecture [2] to predict the acoustic properties of the underlying tissue. The neural network is trained using a custom synthetic dataset generated by us. To create extensive and meaningful synthetic datasets of paired RF data and US images, we proposed four different strategies. These include using geometric features from standard images in the ImageNet dataset and information from medical CT images [3].

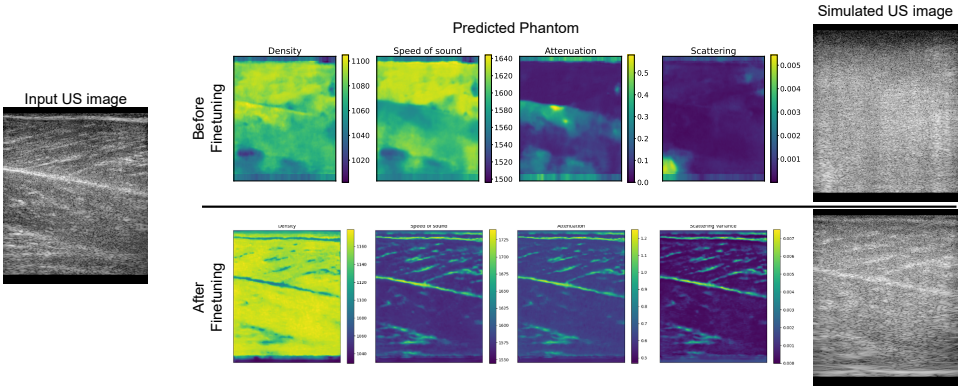


Figure 1: An example visual result, comparing the prediction of a medial gastrocnemius muscle ultrasound image before and after fine-tuning our model.

The performance of our model using different synthetic datasets was compared and evaluated for real RF data and we found that the model which generalizes best is trained on the CT phantom dataset. Additionally, we explored fine-tuning the trained models to predict RF data for medial gastrocnemius muscle images using a dedicated synthetic muscle phantom dataset created by us. We verifies that fine-tuning our model for prior-known tissue types significantly improves the predicted RF data. By comparing our physics-based model to an end-to-end DL model trained for RF data prediction, the benefit of introducing the numerical simulation into our predictions was highlighted.

In this project we build upon the results achieved by Reitsam 2024 [1] and aim to develop an even more robust system architecture with a particular focus on the machine learning model.

2. Research Questions

The research questions of this thesis are:

- What are other possible ML models for converting images to RF data. How large are the conversion losses and what are strategies to improve conversion?
- Quantification of how well your model generalises between different domains at all levels of the development.

2.1. Methodology

- **Workpackage I - Literature and Code review:** The aim of this task is to develop a general understanding of the data-driven elements of the system architecture on the one hand and the physics of ultrasound, the generation of ultrasound images, ultrasound beamforming and simulation on the other. In addition, the code repository developed in the work of Reitsam [1] will be evaluated, replicated and improved.

  Deliverables:

  - **D1:** Fine Grained Project Plan and Gantt Chart.
  - **D2:** Documentation of the literature review findings.
  - **D3:** Restructured Code Repository.

- **Workpackage II - Evaluation of ML Models and Building of Mock-ups:** Recently, diffusion models have been soaring in the fields of text-to-image generation, text-to-video generation, and image-to-image mapping. Diffusion models [4, 5] are trained on large-scale data and outperform generative adversarial networks (GAN) [6] in generating high-quality, realistic, and diverse outputs. In addition, Mamba architecture arises as a rival to Transformer architecture, which is claimed to achieve the modelling power of Transformers with a linearly increasing computational complexity [7]. One important possibility in this thesis will be to explore the utilization of state-of-the-art deep models such as diffusion models and Mamba for the mapping from ultrasound images to phantoms.

  Moreover, another important aspect in these investigations must be set on explainability. Why does one model work better than the other? And how can I quantify accuracy and losses? It will be important to also establish a statistical monitoring framework for quantification.

  Deliverables:

  - **D4:** An investigation of different model architectures and in particular DiT [5] and Mamba [7] for the ultrasound image-to-phantom mapping. Code to train, validate, and test the investigated deep architecture.

  - **D5:** An argumentation for a model of choice to be trained and integrated into the existing Ultrasound Data Recycler System Architecture.

- **Workpackage III - Integration, Training and Test:** Utilisation of the model of choice in the existing system architecture of the Ultrasound Data Recycler.

  Deliverables:

  - **D6:** Performance documentation of the implemented new model architecture.
  - **D7:** Trained model and report on explainability and accuracy.

- **Workpackage IV - Generation of Synthetic Training Data, and Transfer of the Network to a Real World Dataconversion Problem:** The aim of this task is to explore new realistic phantom creation methods using, e.g., Ultrasound, CT or MRI image databases. Furthermore, we want to apply our model to real world data. Are the results comparable to those obtained with synthetic data? If not, how can I adapt my methodology to increase accuracy. How robust is our model architecture to data coming from different source (e.g. instruments, probands,..)?

  Deliverables:

  - **D8:** Synthetic phantom data samples derived from realistic anatomical images.
  - **D9:** Retrained model on the new dataset.
  - **D10:** Application of the system architecture to real-world data and conversion of an image database

- **Workpackage VI - Report and Presentation:** Work on the final report and thesis presentation.

  Deliverables:

  - **D11:** Final report summarizing the entire project and clean and documented code repository.
  - **D12:** Prepared materials and presentation for the thesis run-through and defense.

3. Project Realization

3.1. Project Plan

Within the first week of the project you will be asked to prepare a project plan. This plan should identify the tasks to be performed during the project and sets deadlines for those tasks. The prepared plan will be a topic of discussion of the first week's meeting

between you and your advisers. Note that the project plan should be updated constantly depending on the project's status.

3.2. Meetings

Weekly meetings and reports must be held. The exact time and location of these meetings will be determined within the first week of the project in order to fit the student's and the assistant's schedule. These meetings will be used to evaluate the status and progress of the project. Besides these regular meetings, additional meetings can be organized to address urgent issues as well.

3.3. Report

Documentation is an important and often overlooked aspect of engineering. One final report has to be completed within this project. The common language of engineering is de facto English. Therefore, the final report of the work is preferred to be written in English. Any form of word processing software is allowed for writing the reports, nevertheless, the use of LaTeX with Tgif[1] or any other vector drawing software (for block diagrams) is strongly encouraged by the IIS staff.

Code Repository As many of our projects are heavily code-based, the documentation of the code and its repository is also considered important in the grading of your thesis. We suggest to follow coding standards and style guides when writing code, e.g. C [8], Python [9],...

Final Report The final report has to be presented at the end of the project and a digital copy need to be handed in. Note that this task description is part of your report and has to be attached to your final report. Our LaTeX template can be found here for download: https://iis-projects.ee.ethz.ch/index.php?title=Final_Report)

3.4. Presentation

There will be a presentation (15 min for the semester thesis, and 20 min for the MS thesis presentation followed by 5 min Q&A) at the end of this project in order to present your results to a wider audience. The exact date will be determined towards the end of the work. Additional tipps for preparing the presentation can be found here: https://iis-projects.ee.ethz.ch/index.php?title=Final_Presentation

[1] See: http://bourbon.usc.edu:8001/tgif/index.html and http://www.dz.ee.ethz.ch/en/information/how-to/drawing-schematics.html.

Place and Date _____ Signature Student C. Knecht

References

[1] M. Reitsma, "Ultrasound Image Recycling: Predicting Raw Data with Simulation-Driven Deep Learning," Zurich, May 2024.

[2] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," Feb. 2021, arXiv:2102.04306 [cs]. [Online]. Available: http://arxiv.org/abs/2102.04306

[3] H. Edgar, S. Daneshvari Berry, E. Moes, N. Adolphi, P. Bridges, and K. Nolte, "New Mexico Decedent Image Database (NMDID)," publisher: [object Object]. [Online]. Available: http://nmdid.unm.edu/

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[5] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[7] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," Dec. 2023, arXiv:2312.00752 [cs]. [Online]. Available: http://arxiv.org/abs/2312.00752

[8] Recommended C Style and Coding Standards. [Online]. Available: https://www.doc.ic.ac.uk/lab/cplus/cstyle.html

[9] PEP 8 – Style Guide for Python Code | peps.python.org. [Online]. Available: https://peps.python.org/pep-0008/

IIS - Digital Circuits and Systems Group
Semester/Master theses grading help sheet

Project:
Students:

Assistant _____
Date _____

Grade Aspects		Grading Modifiers							Grade Suggestion		
		Worse than Expected			GOOD	Better than expected			min	suggested	max
		-50	-25	-10		10	25	50			
Difficulty of project	Not all projects have the same difficulty. Sometimes during the course of the project, additional unexpected problems come up making the project much more challenging than originally intended. In other cases, projects may have to be simplified significantly during the course of the project to enable students to have at least some results	tasks were repeatedly simplified during project, extremely simple	Tasks had to be simplified during project, very simple	Simpler than usual	Well balanced project	More involved than usual	Very difficult		-50		25
Effort	A master thesis is a full time effort (40 hours per week), and a semester thesis is supposed to be a half-time work (20 hours/week). Some students spend significantly more (or less) time on the projects. This part is for their effort regardless of their efficiency and quality of results (which is separately handled)	Student was mostly absent	Did not spend enough time on the project	Spent less time than necessary	Student spent the required time for the project	Spend much more time on the project than expected	Student was constantly working on the project		-50		25
In/Dependence	A certain amount of guidance is expected in every project, and the students are expected to contribute to the project. This part rates deviations from the expectations, students that needed too much supervision, would not be able to continue without significant help.		had to be constantly told what to do	needed more supervison than expected	Some guidance was given, but student worked independently	Student worked completely independently			-25		10
Interest/Initiative	It is expected that students are engaged with the project, and contribute on their own accord to further the project. In some projects, the students are hard to motivate to work for the project and do very little if at all if they are not explicitly told to do by the supervisors.		showed very little interest in the project	was difficult to motivate for the project	Student was engaged with project and showed initiative	Made major additions to the project from own initiative			-25		10
Methodology	General approach to solving problems. Students are expected to use a systematic approach	Student did not follow any methodology at all	Larger flaws in the approach to the problem	not all aspects of the work was done systematically	The work was done systematically	The appraoch of the student was excellent, much better than what was expected			-50		10
Team Work	How did the student perform while working together with the assistants and other students. Were there major frictions and problems?		major problems working together as a team	problems in working together as a team	Student showed good team working skills	Was exceptionally good in team work			-25		10
Learning Effect	Some students posses less than average skills at the beginning of the project, however theya re able to learn and improve themselves during the project. This is meant to reward these students that improved more than expected during the course of the project.				Student showed development throughout the project	Remarkable development during the project	Extraordinary improvement during the course of the project		0		25
Results	A major part of the evaluation is what results were achieved. Note that, good results do not necessarily mean that the goals of the project were reached. For example it is a very good result if the work conclusively demonstrates that approach A is not suitable for realizing X. The quality of the results should be judged on how good they can be used further. The results are not good, if they are not substantiated, they can not help us in further projects and are unreliable. So the pratical value of the results are also a consideration, a PCB that can not be used due to design flaws is a bad result.	Project did not produce any results that can be used, results are demonstrably inferior to what has been achieved before	Weak results, the results can not be reliably used in other projects	Problems with some of the results, worse than expectations	Project achieved all its goals, and yielded good results	Project achieved much better results than was expected	Very good results	Breakthrough results, the project resulted in major new findings that can used further	-50		50
Documentation	Good documentation is an essential part of the project. It should contain, problem description, state of the artdescribing what has been done, and explaining why this is not good, list alternative approaches, justify the selected approach, explain implementation, describe the measurements, present results, compare to state of the art and provide a conclusion.	Documentation is unusable, missing major parts	Documentation is missing proper explanations for larger parts, it is not easy to follow what has been done	Incomplete/sloppy documentation, missing parts or insufficient explanations	Good Documentation, concepts are explained clearly and can be re-used for future works	Very good documentation, much better than what is expected	Excellent documentation, could be used as an example of how to write a report		-50		25
Presentation	Presentation exposes the project to the others. Good presentations should be rewarded		Very bad presentation, student unable to show what has been done	Problematic presentation, unable to convay the work/theory/results properly	Good Presentation	Excellent presentation			-25		10
Publications	If the results of the projecta are very good and the student project is contributing to the publications this should be rewarded. This should not be dependent on the acceptance of the paper, if the assistants see it good enough to be submitted it should be good enough for a reward				There are no direct plans to use the results in a scientific paper	Student is contributing to the results of a paper	Student is actively participating in the preparation of a scientific paper		0		25
					5.00			**GRADE**	1.50	**5.00**	7.25

Bibliography

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022.

[2] K. Mercado-Shekhar, "Developing high-frequency quantitative ultrasound techniques to characterize three-dimensional engineered tissues," Ph.D. dissertation, 01 2015.

[3] M. Reitsma, "Ultrasound image recycling: Predicting raw data with simulation-driven deep learning," Master Project, ETH Zurich, Department of Information Technology and Electrical Engineering, April 2024.

[4] T. Wagai, "Studies on the foundation and development of diagnostic ultrasound," *Proceedings of the Japan Academy, Series B, Physical and Biological Sciences*, vol. 83, no. 8, pp. 256–265, Dec 2007.

[5] C. J. Garvey and R. Hanlon, "Computed tomography in clinical practice," *BMJ*, vol. 324, no. 7345, pp. 1077–1080, May 2002.

[6] V. P. Grover, J. M. Tognarelli, M. M. Crossey, I. J. Cox, S. D. Taylor-Robinson, and M. J. McPhail, "Magnetic resonance imaging: Principles and techniques: Lessons for clinicians," *Journal of Clinical and Experimental Hepatology*, vol. 5, no. 3, pp. 246–255, Sep 2015, epub 2015 Aug 20.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[9] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, p. 307–392, 2019. [Online]. Available: http://dx.doi.org/10.1561/2200000056

[10] H. Jiang, M. Imran, L. Ma, T. Zhang, Y. Zhou, M. Liang, K. Gong, and W. Shao, "Fast-ddpm: Fast denoising diffusion probabilistic models for medical image-to-image generation," 2024.

[11] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," 2024. [Online]. Available: https://arxiv.org/abs/2312.02145

[12] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," 2021.

[13] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu, "Medsegdiff-v2: Diffusion based medical image segmentation with transformer," 2023.

[14] B. E. Treeby and B. T. Cox, "k-wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *Journal of Biomedical Optics*, vol. 15, no. 2, p. 021314, 2010. [Online]. Available: https://doi.org/10.1117%2F1.3360308

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[16] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, p. 307–392, 2019. [Online]. Available: http://dx.doi.org/10.1561/2200000056

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[18] M. Goldblum, H. Souri, R. Ni, M. Shu, V. Prabhu, G. Somepalli, P. Chattopadhyay, M. Ibrahim, A. Bardes, J. Hoffman, R. Chellappa, A. G. Wilson, and T. Goldstein, "Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks," 2023.

[19] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach, "Scaling rectified flow transformers for high-resolution image synthesis," 2024.

[20] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2022.

[21] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," may 2023, arXiv:2211.01095 [cs]. [Online]. Available: http://arxiv.org/abs/2211.01095

[22] Q. Lyu and G. Wang, "Conversion between ct and mri images using diffusion and score-matching models," 2022.

[23] R. Hunter, Łukasz Dudziak, M. S. Abdelfattah, A. Mehrotra, S. Bhattacharya, and H. Wen, "Fast sampling through the reuse of attention maps in diffusion models," 2024.

[24] S. Xie, Z. Xiao, D. P. Kingma, T. Hou, Y. N. Wu, K. P. Murphy, T. Salimans, B. Poole, and R. Gao, "Em distillation for one-step diffusion models," may 2024, arXiv:2405.16852 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2405.16852

[25] T. Yin, M. Gharbi, T. Park, R. Zhang, E. Shechtman, F. Durand, and W. T. Freeman, "Improved distribution matching distillation for fast image synthesis," may 2024, arXiv:2405.14867 [cs]. [Online]. Available: http://arxiv.org/abs/2405.14867

[26] D. Kim, C.-H. Lai, W.-H. Liao, Y. Takida, N. Murata, T. Uesaka, Y. Mitsufuji, and S. Ermon, "Pagoda: Progressive growing of a one-step generator from a low-resolution diffusion teacher," may 2024, arXiv:2405.14822 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2405.14822

[27] C. Hong, J. Huang, R. Birke, D. Epema, S. Roos, and L. Y. Chen, "Sfddm: Single-fold distillation for diffusion models," may 2024, arXiv:2405.14961 [cs]. [Online]. Available: http://arxiv.org/abs/2405.14961

[28] M. Zhou, H. Zheng, Z. Wang, M. Yin, and H. Huang, "Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation," may 2024, arXiv:2404.04057 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2404.04057

[29] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models for medical image analysis: A comprehensive survey," 2023.

[30] Y. Li, H.-C. Shao, X. Liang, L. Chen, R. Li, S. Jiang, J. Wang, and Y. Zhang, "Zero-shot medical image translation via frequency-guided diffusion models," *IEEE Transactions on Medical Imaging*, vol. 43, no. 3, p. 980–993, Mar. 2024. [Online]. Available: http://dx.doi.org/10.1109/TMI.2023.3325703

[31] M. Özbey, O. Dalmaz, S. U. Dar, H. A. Bedel, Şaban Özturk, A. Güngör, and T. Çukur, "Unsupervised medical image translation with adversarial diffusion models," 2023.

[32] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, and F. Wen, "Pre-training is all you need for image-to-image translation," 2022.

[33] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," 2022.

[34] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, R. Timotfe, and L. V. Gool, "Diffi2i: Efficient diffusion model for image-to-image translation," 2023.

[35] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "Bbdm: Image-to-image translation with brownian bridge diffusion models," 2023.

[36] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," 2021.

[37] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[38] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu, "Medsegdiff: Medical image segmentation with diffusion probabilistic model," 2023.

[39] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, c. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[40] ——, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[41] S. Bakas, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, 09 2017.

[42] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," 2018. [Online]. Available: https://arxiv.org/abs/1902.03368

[43] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, 2018. [Online]. Available: https://doi.org/10.1038/sdata.2018.161

[44] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: An open large-scale dataset for training next generation image-text models," 2022. [Online]. Available: https://arxiv.org/abs/2210.08402

[45] C. Leitner, "dasit: A delay-and-sum beamformer for plane-wave ultrasound," https://github.com/luuleitner/dasIT, 2024, accessed: 2024-12-20.

[46] S. C. G. at ETH Zurich, "Euler - high performance computing cluster," 2024, accessed: 2024-12-19. [Online]. Available: https://scicomp.ethz.ch/wiki/Euler

[47] M. Jette, A. Yoo, and M. Grondona, "Slurm: Simple linux utility for resource management," 07 2003.

[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019.